University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2006 *Paper* 202

Multiple Tests of Association with Biological Annotation Metadata

Sandrine Dudoit* Sunduz Keles†

Mark J. van der Laan‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

†Dept. of Biostatistics & Medical Informatics and Dept. of Statistics, University of Wisconsin, Madison, keles@biostat.wisc.edu

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper202

Multiple Tests of Association with Biological Annotation Metadata

Sandrine Dudoit, Sunduz Keles, and Mark J. van der Laan

Abstract

We propose a general and formal statistical framework for the multiple tests of associations between known fixed features of a genome and unknown parameters of the distribution of variable features of this genome in a population of interest. The known fixed gene-annotation profiles, corresponding to the fixed features of the genome, may concern Gene Ontology (GO) annotation, pathway membership, regulation by particular transcription factors, nucleotide sequences, or protein sequences. The unknown gene-parameter profiles, corresponding to the variable features of the genome, may be, for example, regression coefficients relating genome-wide transcript levels or DNA copy numbers to possibly censored biological and clinical outcomes and covariates. A generic question of great interest in current genomic research, regarding the detection of associations between biological annotation metadata and genome-wide expression measures, may then be translated into the multiple tests of hypotheses concerning association measures between gene-annotation and gene-parameter profiles. A general and rigorous formulation of the statistical inference question allows us to apply the multiple testing methodology developed in Dudoit and van der Laan (2006) and related articles, to control 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. Resampling-based single-step and stepwise multiple testing procedures, that take into account the joint distribution of the test statistics, are provided to control Type I error rates defined as tail probabilities for arbitrary functions of the numbers of false positives and rejected hypotheses.

The proposed statistical and computational methods are illustrated using the acute lymphoblastic leukemia (ALL) microarray dataset of Chiaretti et al. (2004), with

the aim of relating GO annotation to differential gene expression between B-cell ALL with the BCR/ABL fusion and cytogenetically normal NEG B-cell ALL. The sensitivity of the identified lists of GO terms to the choice of association parameter between GO annotation and differential gene expression demonstrates the importance of translating the biological question in terms of suitable gene-annotation profiles, gene-parameter profiles, and association measures. In particular, the results show the limitations of binary gene-parameter profiles of differential expression indicators, which are still the norm for combined GO annotation and microarray data analyses. Procedures based on such binary gene-parameter profiles tend to be conservative and lack robustness with respect to the estimator for the set of differentially expressed genes.

WWW companion: www.stat.berkeley.edu/∼sandrine/Docs/Papers/DFF06/DFF.html

Contents

1 Introduction

1.1 Motivation

Experimental data, such as microarray gene expression measures, gain much in relevance from their association with biological annotation *metadata*, i.e., data on data, such as, GenBank sequences, Gene Ontology terms, KEGG pathways, and 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. This includes accessing and pre-processing the data, making inference from these data, and summarizing and interpreting the results.

In this context, an important class of statistical problems involves testing for associations between known fixed features of a genome and 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. Fixed features typically consist of gene annotation metadata, that reflect current knowledge on gene properties, such as, nucleotide and protein sequences, regulation, and function. Variable features often consist of gene expression measures, that reflect cellular type and/or state under particular conditions. The fixed and variable features define, respectively, *gene-annotation profiles* and *gene-parameter profiles*; the parameter of interest then corresponds to *measures of association between known gene-annotation profiles and unknown gene-parameter profiles*.

For instance, for the yeast *Saccharomyces cerevisiae* (in short, *S. cerevisiae*), one may be interested in detecting associations between the vector of mean transcript (i.e., mRNA) levels for all (approximately 6,500) genes under heat-shock conditions and *Gene Ontology* (GO) annotation for these genes. The reader is referred to the Gene Ontology Consortium website (`www.geneontology.org`) and to Section 4, below, for more information on gene ontologies, and to the *Saccharomyces Genome Database* (SGD) website (`www.yeastgenome.org`), for details on *S. cerevisiae*. In this example, the population of interest may consist of all heat-shocked yeast cells from well-defined cultures of a particular strain of *S. cerevisiae* (e.g., strain S288C). For each of the three gene ontologies (BP, CC, and MF, as described in Section 4.1), each gene is annotated with a fixed set of GO terms (i.e., this set is constant across population units for a given version of the GO Database). Thus, for each GO term, one may define a gene-annotation profile as a known, fixed binary vector indicating for each gene whether it is annotated or not with the particular GO term. The transcript levels, however, vary among population units and the gene-parameter profile, i.e., the vector of genome-wide mean transcript levels in the population of heat-shocked yeast cells, is unknown and may be estimated, for example, from a microarray experiment involving a sample of yeast cells from the population. The association parameter of interest, between GO annotation and transcript levels, is then a vector of association measures (e.g., two-sample t-statistics) between the known

binary gene-annotation profiles and the unknown continuous gene-parameter profile.

Similar inference questions arise in many other contexts and involve a variety of definitions for the gene-annotation profiles, the gene-parameter profiles, and the association parameters of interest. For example, in cancer microarray studies, one may seek associations between GO gene-annotation profiles and a gene-parameter profile of regression coefficients relating (censored) patient survival data to genome-wide transcript levels or DNA copy numbers. Furthermore, gene-annotation profiles need not be binary or even polychotomous, and may correspond to pathway membership, regulation by particular transcription factors, nucleotide sequences, and protein sequences.

Note that, for the sake of illustration, we focus on gene-level features. However, our proposed methodology is generic and may be applied to other types of features, such as those concerning gene isoforms and proteins. For instance, as in alternative splicing microarray analysis, one may collect data at the finer level of gene isoforms, where one gene may have multiple isoforms (Blanchette et al., 2005). In this context, *isoform-parameter profiles* may refer to the distribution of isoform microarray expression measures in a well-defined population, while *isoform-annotation profiles* may consist of intron/exon counts/lengths/nucleotide distributions. One may also consider protein-level features, where, for example, *protein-parameter profiles* correspond to antibody microarray expression measures and *protein-annotation profiles* refer to protein function, domain structure, and post-translational modification (e.g., from Swiss-Prot; `www.expasy.org/spro`).

1.2 Contrast with other approaches

Existing approaches for tests of association with biological annotation metadata focus primarily on relating microarray gene expression measures and GO annotation. Relevant articles and software packages include: `FatiGO` from the `BABELOMICS` suite (Al-Shahrour et al. (2004, 2005); `www.babelomics.org`); `GOstat` (Beissbarth and Speed (2004); `gostat.wehi.edu.au`); `Ontologizer` (Grossmann et al. (2006); `www.charite.de/ch/medgen/ontologizer`); McCarroll et al. (2004); `GSEA-P` (Mootha et al. (2003), Subramanian et al. (2005); `www.broad.mit.edu/gsea/doc/doc_index.html`); Tian et al. (2005). Methods proposed thus far suffer from a number of limitations, related, to a large extent, to the absence of a clear and precise statement of the statistical inference question. As a result, the analyses often lack statistical rigor and tend to be ad hoc and dataset-specific.

One of our main contributions is the systematic and precise translation of a general class of biological questions into a corresponding class of multiple hypothesis testing problems. A key step in this process is the proper definition of the gene-annotation profiles, gene-parameter profiles, and association parameters of interest. This general formulation then allows us to apply the multiple testing methodology developed in Dudoit and van der Laan (2006) and related articles (Birkner et al., 2005; Dudoit et al., 2004a,b; Keleş et al., 2004;

van der Laan et al., 2004a,b, 2005; van der Laan and Hubbard, 2005; Pollard et al., 2005a,b; Pollard and van der Laan, 2004; Rubin et al., 2005), to control a broad class of Type I error rates, defined as generalized tail probabilities (gTP), $gTP(q, g) = Pr(g(V_n, R_n) > q)$, for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n.

We wish to emphasize the crucial and often ignored distinction between: (i) the definition of a *parameter* of interest, measuring the association between gene-annotation and gene-parameter profiles, i.e., the statistical formulation of the biological question; (ii) making *inferences*, i.e., deriving estimators of and testing hypotheses concerning this parameter, based on a sample drawn from the population under consideration. Most methods proposed to date focus on (ii), without providing a clear statement of the question being answered in (i), that is, various estimation and testing approaches are proposed for an undefined parameter of interest.

Due to its general and rigorous statistical framework, 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 Gene Ontology metadata and implemented in the software packages listed on the "Gene Ontology Tools" webpage (`www.geneontology.org/GO.tools.shtml`).

1. *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, intron/exon counts/lengths/nucleotide distributions, mean transcript levels.

2. *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.

3. *Estimated gene-parameter profiles.* Existing approaches typically assume known gene-parameter profiles. For example, the list of differentially expressed 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, as in Section 3, provides a more rigorous and general formulation of the statistical question.

4. *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). As in Table 2, rows correspond to gene annotation with a given GO term (fixed binary gene-annotation profile) and columns to an "interesting" gene property, such as differential expression (treated as a fixed binary gene-parameter profile). The approach proposed in Section 3 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.

1.3 Outline

This article proposes a general and formal statistical framework for multiple tests of association with biological annotation metadata, using the multiple testing methodology of Dudoit and van der Laan (2006) and related articles.

Section 2 provides an introduction to multiple hypothesis testing. Section 3 presents the proposed statistical framework for multiples tests of association with biological annotation metadata and discusses in detail the main components of the inference problem, namely, the gene-annotation profiles, the gene-parameter profiles, and the association parameters. Multiple testing procedures for tests of association between gene-annotation profiles and gene-parameter profiles are outlined. Section 4 gives an overview of the Gene Ontology (GO) and R software for analyzing GO annotation metadata (e.g., for assembling GO gene-annotation profiles). The proposed statistical and computational methods are illustrated in Section 5, using the acute lymphoblastic leukemia (ALL) microarray dataset of Chiaretti et al. (2004), with the aim of relating GO annotation to differential gene expression between B-cell ALL with the BCR/ABL fusion and cytogenetically normal NEG B-cell ALL. Finally, Section 6 summarizes our findings and outlines ongoing work.

2 Overview of multiple hypothesis testing

This section introduces a general statistical framework for multiple hypothesis testing and summarizes in turn the main ingredients of a multiple testing problem, including: the data generating distribution; the parameters of interest; the null and alternative hypotheses; the test statistics; rejection regions (i.e., cutoffs) for the test statistics; Type I and Type II errors; Type I error rates and power; the test statistics null distribution; multiple testing procedures; adjusted p-values.

The reader is referred to our earlier articles and book for further detail on the multiple testing methodology, its software implementation, and its application to a variety of testing problems in biomedical and genomic research (Birkner et al., 2005; Dudoit and van der Laan, 2006; Dudoit et al., 2004a,b; Keleş et al., 2004; van der Laan et al., 2004a,b, 2005; van der Laan and Hubbard, 2005; Pollard et al., 2005a,b; Pollard and van der Laan, 2004; Rubin et al., 2005).

2.1 Null and alternative hypotheses

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

Let $\mathcal{X}_n \equiv \{X_i : 1, \ldots, n\}$ denote a *random sample* of n independent and identically distributed (i.i.d.) random variables from a *data generating distribution* P, i.e., $X_i \overset{i.i.d.}{\sim} P$, $i = 1, \ldots, n$. Suppose that the data generating distribution P is an element of a particular *statistical model* $\mathcal{M}$, i.e., a set of possibly non-parametric distributions, $P \in \mathcal{M}$. Let P_n denote the corresponding *empirical distribution*, which places probability $1/n$ on each realization of X.

In order to cover a broad class of testing problems, specify M pairs of null and alternative hypotheses in terms of a collection of M *submodels*, $\mathcal{M}(m) \subseteq \mathcal{M}$, $m = 1, \ldots, M$, for the data generating distribution P. The M *null hypotheses* and corresponding *alternative hypotheses* are defined as

$$H_0(m) \equiv \mathrm{I}(P \in \mathcal{M}(m)) \qquad \text{and} \qquad H_1(m) \equiv \mathrm{I}(P \notin \mathcal{M}(m)), \tag{1}$$

respectively. Here, $\mathrm{I}(\cdot)$ is the indicator function, equaling one if the condition in parentheses is true and zero otherwise.

In many testing problems, the submodels concern *parameters*, i.e., functions $\Psi(P) = \psi = (\psi(m) : m = 1, \ldots, M)$ of the data generating distribution P, and each null hypothesis may refer to a single parameter, $\psi(m) = \Psi(P)(m) \in I\!R$.

This general submodel representation covers tests of means, quantiles, correlation coefficients, and regression coefficients in linear and non-linear models (e.g., logistic, survival, time-series, and dose-response models).

Let $\mathcal{H}_0 = \mathcal{H}_0(P) \equiv \{m : H_0(m) = 1\}$ and $\mathcal{H}_1 = \mathcal{H}_1(P) \equiv \mathcal{H}_0^c(P) = \{m : H_1(m) = 1\}$ denote, respectively, the sets of $h_0 \equiv |\mathcal{H}_0|$ *true null hypotheses* and $h_1 \equiv |\mathcal{H}_1| = M - h_0$ *false null hypotheses*, i.e., *true positives*.

2.2 Test statistics and rejection regions

The goal of a *multiple testing procedure* (MTP) is to accurately estimate, i.e., *reject*, the set $\mathcal{H}_1$ of true positives, while probabilistically controlling false positives.

The decisions to reject or not the null hypotheses are based on an M–vector of *test statistics*, $T_n = (T_n(m) : m = 1, \ldots, M)$, that are functions $T_n(m) = T(m; \mathcal{X}_n)$ of the data $\mathcal{X}_n$. A broad class of testing problems may be addressed using difference statistics and t-statistics (Equations (20) and (21), respectively). Denote the typically unknown (finite sample) *joint distribution* of the test statistics T_n by $Q_n = Q_n(P)$.

A MTP provides *rejection regions* $\mathcal{C}_n(m)$, i.e., sets of values for each test statistic $T_n(m)$ that lead to the decision to reject the corresponding null hypothesis $H_0(m)$, $m = 1, \ldots, M$. In other words, a MTP produces a random (i.e., data-driven) subset $\mathcal{R}_n$ of rejected hypotheses that estimates the set $\mathcal{H}_1$ of true positives,

$$\mathcal{R}_n \equiv \{m : T_n(m) \in \mathcal{C}_n(m)\} = \{m : H_0(m) \text{ is rejected}\}. \tag{2}$$

2.3 Errors in multiple hypothesis testing

2.3.1 Type I and Type II errors

In any testing situation, two types of errors can be committed: a *false positive*, or *Type I error*, is committed by rejecting a true null hypothesis ($\mathcal{R}_n \cap \mathcal{H}_0$), and a *false negative*, or *Type II error*, is committed when the test procedure fails to reject a false null hypothesis (i.e., a true positive) ($\mathcal{R}_n^c \cap \mathcal{H}_1$).

The main decisions and errors in a multiple testing problem are summarized in Table 1, below, where the numbers of rejected hypotheses, Type I errors, and Type II errors are defined as

$$\begin{aligned} R_n &\equiv |\mathcal{R}_n| = \sum_{m=1}^{M} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m)), \\ V_n &\equiv |\mathcal{R}_n \cap \mathcal{H}_0| = \sum_{m \in \mathcal{H}_0} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m)), \\ \text{and} \qquad U_n &\equiv |\mathcal{R}_n^c \cap \mathcal{H}_1| = \sum_{m \in \mathcal{H}_1} \mathrm{I}(T_n(m) \notin \mathcal{C}_n(m)), \end{aligned} \tag{3}$$

respectively. Note that both U_n and V_n depend on the unknown data generating distribution P through the unknown set of true null hypotheses $\mathcal{H}_0 = \mathcal{H}_0(P)$. Therefore, the numbers $h_0 = |\mathcal{H}_0|$ and $h_1 = |\mathcal{H}_1| = M - h_0$ of true and false null hypotheses are *unknown parameters*, the number of rejected hypotheses R_n is an *observable random variable*, and the entries in the body of the table, U_n, $h_1 - U_n$, V_n, and $h_0 - V_n$, are *unobservable random variables* (that depend on the unknown data generating distribution P through $\mathcal{H}_0(P)$).

Ideally, one would like to simultaneously minimize both the number of Type I errors and the number of Type II errors. Unfortunately, this is not feasible and one seeks a *trade-off* between the two types of errors. A standard approach is to specify an acceptable level α for a suitably defined Type I error rate and derive testing procedures, i.e., rejection regions, that aim to minimize a Type II error rate, i.e., maximize power, within the class of tests with Type I error rate at most α.

2.3.2 Type I error rates and power

When testing multiple hypotheses, there are many possible definitions for the Type I error rate and power of a test procedure. Accordingly, we define a *Type I error rate* as an arbitrary 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. Likewise, *power* may be defined as a parameter $\vartheta_n = \vartheta(F_{U_n,R_n})$ of the joint distribution F_{U_n,R_n} of the numbers of Type II errors $U_n = |\mathcal{R}_n^c \cap \mathcal{H}_1|$ and rejected hypotheses R_n.

Type I error rates of particular interest are *generalized tail probability* (gTP) error rates,

$$gTP(q,g) \equiv Pr(g(V_n, R_n) > q), \tag{4}$$

and *generalized expected value* (gEV) error rates,

$$gEV(g) \equiv E[g(V_n, R_n)], \tag{5}$$

for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n and user-supplied bounds q.

Generalized tail probability error rates include as special cases the following commonly-used Type I error rates.

- The *generalized family-wise error rate* (gFWER), corresponding to $g(v, r) = v$ and $q \in \{0, \ldots, (h_0 - 1)\}$, is the probability of at least $(q + 1)$ Type I errors,

  $$gFWER(q) \equiv Pr(V_n > q) = 1 - F_{V_n}(q). \tag{6}$$

  When $q = 0$, the gFWER reduces to the usual *family-wise error rate* (FWER), controlled by the classical Bonferroni procedure.

- The *tail probability for the proportion of false positives* (TPPFP) *among the rejected hypotheses*, corresponding to $g(v, r) = v/r$ and $q \in (0, 1)$, is defined as

  $$TPPFP(q) \equiv Pr\left(\frac{V_n}{R_n} > q\right) = 1 - F_{V_n/R_n}(q), \tag{7}$$

  with the convention that $V_n/R_n \equiv 0$ if $R_n = 0$.

The generalized expected value error rate for $g = v/r$ corresponds to the *false discovery rate* (FDR), i.e., the expected proportion of false positives among the rejected hypotheses,

$$FDR \equiv E\left[\frac{V_n}{R_n}\right] = \int q dF_{V_n/R_n}(q), \tag{8}$$

again with the convention that $V_n/R_n \equiv 0$ if $R_n = 0$.

2.4 Test statistics null distribution

As discussed in Section 3.4, below, a key feature of our proposed multiple testing procedures is the *test statistics null distribution* used to obtain rejection regions (i.e., cut-offs) for the test statistics, confidence regions for the parameters of interest, and adjusted p-values. Indeed, whether testing single or multiple hypotheses, one needs the (joint) distribution of the test statistics in order to derive a procedure that probabilistically controls Type I errors. In practice, however, the true distribution $Q_n(P)$ of the test statistics is unknown and replaced by a null distribution Q_0. The choice of a suitable null distribution is crucial, in order to ensure that (finite sample or asymptotic) control of the Type I error rate under the *assumed null distribution* does indeed provide the desired control under the *true distribution*.

2.5 Multiple testing procedures

Having identified a suitable test statistics null distribution Q_0 (or estimator thereof, Q_{0n}), there remains the main task of specifying rejection regions $\mathcal{C}_n(m)$ for each null hypothesis $H_0(m)$. As detailed in Dudoit and van der Laan (2006) and as summarized in Section 3.4, below, we have developed resampling-based single-step and stepwise multiple testing procedures for controlling a broad class of Type I error rates, in testing problems involving general data generating distributions (with arbitrary dependence structures among variables), null hypotheses (defined in terms of submodels for the data generating distribution), and test statistics (e.g., t-statistics, χ^2-statistics, F-statistics). Procedures that take into account the *joint* distribution of the test statistics are provided to control Type I error rates defined as tail probabilities and expected values for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n.

2.6 Adjusted p-values

As in the case of single hypothesis testing, one can report the results of a multiple testing procedure in terms of the following quantities: rejection regions for the test statistics, confidence regions for the parameters of interest, and adjusted p-values.

Adjusted p-values, for the test of multiple hypotheses, are defined as straightforward extensions of unadjusted p-values, for the test of individual hypotheses. Consider any multiple testing procedure $\mathcal{R}_n(\alpha) = \mathcal{R}(T_n, Q_0, \alpha)$, with rejection regions $\mathcal{C}_n(m;\alpha) = \mathcal{C}(m; T_n, Q_0, \alpha)$. Then, the *adjusted p-value* for null hypothesis $H_0(m)$ is defined as

$$\begin{aligned} \widetilde{P}_{0n}(m) &\equiv \inf\{\alpha \in [0,1] : \text{Reject } H_0(m) \text{ at nominal MTP level } \alpha\} \quad (9) \\ &= \inf\{\alpha \in [0,1] : m \in \mathcal{R}_n(\alpha)\} \\ &= \inf\{\alpha \in [0,1] : T_n(m) \in \mathcal{C}_n(m;\alpha)\}, \qquad m = 1, \ldots, M. \end{aligned}$$

That is, $\widetilde{P}_{0n}(m)$ is the *smallest nominal Type I error level* (e.g., gFWER, TPPFP, or FDR) of the *multiple hypothesis testing procedure* at which one would reject $H_0(m)$, given T_n.

For example, the adjusted p-values for the classical FWER-controlling marginal Bonferroni procedure are $\widetilde{P}_{0n}(m) = \min(MP_{0n}(m), 1)$. Adjusted p-values for FWER-controlling joint single-step common-cut-off maxT Procedure 1 are given in Equation (25).

As in single hypothesis tests, the smaller the adjusted p-value $\widetilde{P}_{0n}(m)$, the stronger the evidence against the corresponding null hypothesis $H_0(m)$. Specifically, for a multiple test at nominal Type I error level α, one has two equivalent representations for the set of rejected hypotheses, in terms of rejection regions for the test statistics and in terms of adjusted p-values,

$$\mathcal{R}_n(\alpha) = \{m : T_n(m) \in \mathcal{C}_n(m;\alpha)\} = \{m : \widetilde{P}_{0n}(m) \leq \alpha\}. \quad (10)$$

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).
- 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 number of rejected hypotheses and tolerable false positive rate for a particular experiment and available resources.

3 Statistical framework for multiple tests of association with biological annotation metadata

Sections 3.1 – 3.3 introduce the main components of our approach to multiple tests of association with biological annotation metadata, namely: the gene-annotation profiles A, the gene-parameter profiles λ, and the association measures $\psi = \rho(A, \lambda)$ between gene-annotation and gene-parameter profiles. We stress that the choice of a suitable association parameter ψ is perhaps the most important and hardest aspect of the inference problem, as this parameter represents the statistical translation of the biological question of interest. Once the association parameter ψ is appropriately and precisely defined, one can rely on a variety of statistical methods to estimate and test hypotheses concerning this parameter. Section 3.4 describes how the multiple testing methodology of Dudoit and van der Laan (2006) and related articles may be used to detect associations between gene-annotation and gene-parameter profiles.

Note that, for the sake of illustration, we focus on gene-level features. However, as mentioned in Section 1.1, the methodology is generic and may be applied to other types of features, such as those concerning gene isoforms and proteins.

3.1 Gene-annotation profiles

Gene-annotation profiles refer to features of a genome that are assumed to be known and constant among units in a population of interest. Such features typically consist of gene annotation metadata, that reflect current knowledge

on gene properties, such as, nucleotide and protein sequences, regulation, and function.

Specifically, 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. Thus, row $A(g,\cdot) \equiv (A(g,m) : m = 1,\ldots,M)$ denotes an M–dimensional gene-specific feature vector for the gth gene, $g = 1,\ldots,G$, and column $A(\cdot,m) \equiv (A(g,m) : g = 1,\ldots,G)$ denotes a G–dimensional *gene-annotation profile* for the mth feature, $m = 1,\ldots,M$.

In many applications, the element $A(g,m)$ is a binary indicator, coding the YES/NO answer to the mth question, among a collection of M questions, one may ask about gene g. For example, $A(g,m)$ could indicate whether gene g is annotated with a particular GO term m, among M terms in one of the three ontologies (BP, CC, or MF), i.e., whether gene g is an element of the node corresponding to the mth term in the GO directed acyclic graph (DAG). Other gene-annotation profiles of interest may refer to intron/exon counts/lengths/nucleotide distributions, gene pathway membership (e.g., from the Kyoto Encyclopedia of Genes and Genomes, KEGG; `www.genome.ad.jp/kegg`), or gene regulation by particular transcription factors. Regarding transcription regulation, one could use data from the Transcription Factor DataBase (TRANSFAC; `www.gene-regulation.com`) to generate gene-annotation profiles as follows. For a given transcription factor binding motif, a binary gene-annotation profile could consist of indicators for the presence or absence of the motif in the upstream control region of each gene. A continuous gene-annotation profile could be based on the position weight matrix of the binding motif.

Note that the aforementioned features are only *fixed in time* for a given version/release of the corresponding database(s), i.e., such biological data are constantly evolving as our knowledge of the roles of genes and proteins is accumulating and changing. The dynamic nature of biological annotation metadata is an important issue in terms of software design (Section 4.2; Gentleman et al. (2005)). Note also that the gene-annotation profiles are not restricted to be binary or even polychotomous and, in particular, could be continuous gene-parameter profiles, suitably estimated from previous studies.

The main point, regarding the formulation of the statistical inference question, is that gene-annotation profiles are *known* and *constant among population units*.

3.2 Gene-parameter profiles

Gene-parameter profiles are generally unknown and concern the distribution of variable features of a genome in a well-defined population. Gene-specific variables of interest reflect cellular type and/or state under particular conditions and include microarray measures of transcript levels and comparative genomic hybridization (CGH) measures of DNA copy numbers.

Specifically, let X be a J–dimensional random vector, containing G gene-specific random variables $(X(g) : g = 1,\ldots,G)$. In addition to the G gene-specific variables, X may include various biological and clinical covariates (e.g.,

age, sex, treatment, timepoint) and outcomes (e.g., survival time, response to treatment, tumor class). Let P denote the data generating distribution for the random J–vector X and suppose that P belongs to a (possibly non-parametric) model $\mathcal{M}$.

Let the parameter mapping $\Lambda : \mathcal{M} \to I\!R^G$ define a G–dimensional *gene-parameter profile*, $\Lambda(P) = \lambda = (\lambda(g) : g = 1, \ldots, G)$, where each $\lambda(g) = \Lambda(P)(g) \in I\!R$ is a gene-specific real-valued parameter. For example, $\lambda(g)$ could be the mean expression measure $E[X(g)]$ of gene g or a regression coefficient in a model relating an outcome component of X to the expression measure $X(g)$ of gene g, $g = 1, \ldots, G$.

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. The sample is assumed to consist of n independent and identically distributed (i.i.d.) copies of $X \sim P$, $\mathcal{X}_n = \{X_i : 1, \ldots, n\}$, corresponding to n randomly sampled population units.

3.3 Association measures for gene-annotation and gene-parameter profiles

Let the parameter mapping $\Psi : \mathcal{M} \to I\!R^M$ specify an M–dimensional *association parameter vector*,

$$\Psi(P) = \psi = (\psi(m) : m = 1, \ldots, M) \equiv \rho(A, \Lambda(P)), \tag{11}$$

defined in terms of an *association measure* $\rho : I\!R^{G \times M} \times I\!R^G \to I\!R^M$, *known fixed gene-annotation profiles* A, and an *unknown gene-parameter profile* $\lambda = \Lambda(P)$.

The choice of a suitable association parameter is subject matter-dependent and requires careful consideration. For instance, for Gene Ontology annotation, it is desirable that the association parameter reflect the structure of the GO directed acyclic graph (Section 4.1). In principle, the dimension of the association parameter vector ψ could differ from the number M of features under consideration. In addition, one could accommodate several gene-parameter profiles λ.

The various quantities in the inference problem are summarized in Figure 1; examples of association parameters are given next and in Section 5.

3.3.1 Univariate association measures

In the simplest case, one could define the M association parameters univariately, i.e., define $\psi(m)$ based only on the mth gene-annotation profile $A(\cdot, m)$, $m = 1, \ldots, M$. Specifically, for the mth feature, let

$$\Psi(P)(m) = \psi(m) \equiv \rho_m(A(\cdot, m), \Lambda(P)), \tag{12}$$

where $\rho_m : I\!R^G \times I\!R^G \to I\!R$ provides a measure of association (e.g., correlation coefficient) between the G–dimensional gene-annotation profile $A(\cdot, m)$ and gene-parameter profile $\lambda = \Lambda(P)$. In many situations, the same association

measure ρ_m may be used for each of the M features.

Continuous gene-annotation profiles and continuous gene-parameter profiles. For continuous gene-annotation and gene-parameter profiles, one may use as association measure the Pearson correlation coefficient between two G–vectors. That is,

$$\psi(m) = \frac{\sum_{g=1}^{G}(A(g,m) - \bar{A}(m))(\lambda(g) - \bar{\lambda})}{\sqrt{\sum_{g=1}^{G}(A(g,m) - \bar{A}(m))^2}\sqrt{\sum_{g=1}^{G}(\lambda(g) - \bar{\lambda})^2}}, \tag{13}$$

where $\bar{A}(m) \equiv \sum_g A(g,m)/G$ and $\bar{\lambda} \equiv \sum_g \lambda(g)/G$ denote, respectively, the averages of the G components of the gene-annotation profile $A(\cdot, m)$ and gene-parameter profile λ.

Binary gene-annotation profiles and binary gene-parameter profiles. For binary gene-annotation and gene-parameter profiles, one may build 2×2 contingency Table 2 and use as association measure the χ^2-statistic (or corresponding p-value) for the test of independence of rows and columns. That is,

$$\psi(m) = \frac{G(g_{00}(m)g_{11}(m) - g_{01}(m)g_{10}(m))^2}{(g_{00}(m) + g_{01}(m))(g_{00}(m) + g_{10}(m))(g_{11}(m) + g_{01}(m))(g_{11}(m) + g_{10}(m))}, \tag{14}$$

where $g_{kk'}(m) \equiv \sum_g \mathrm{I}(A(g,m) = k)\mathrm{I}(\lambda(g) = k')$, $k, k' \in \{0, 1\}$. Note that in this context *the χ^2-statistic $\psi(m)$ is a parameter*, i.e., it is a function of the data generating distribution P, via the gene-parameter profile $\lambda = \Lambda(P)$, and is *unknown* and *constant* among population units.

Binary gene-annotation profiles. For binary gene-annotation profiles, one may consider association parameter vectors of the form

$$\psi = A^\top \lambda. \tag{15}$$

That is, the association parameter for the mth feature is the sum,

$$\psi(m) = \sum_{g=1}^{G} A(g,m)\lambda(g) = \sum_{g=1}^{G} \mathrm{I}(A(g,m) = 1)\lambda(g),$$

of the parameters $\lambda(g)$ for genes g that have the property of interest, i.e., such that $A(g,m) = 1$. Such an association parameter is considered by Tian et al. (2005), to relate continuous microarray differential expression gene-parameter profiles to binary pathway gene-annotation profiles. The following standardized association parameters (corresponding to association measures based on two-sample t-statistics) may also be considered,

$$\psi(m) = \frac{\bar{\lambda}_1(m) - \bar{\lambda}_0(m)}{\sqrt{\frac{v[\lambda]_1(m)}{A_1(m)} + \frac{v[\lambda]_0(m)}{A_0(m)}}}, \tag{16}$$

where, for the mth feature, $A_k(m) \equiv \sum_g \mathrm{I}(A(g,m)=k)$, $\bar{\lambda}_k(m) \equiv \sum_g \mathrm{I}(A(g,m)=k)\lambda(g)/A_k(m)$, and $v[\lambda]_k(m) \equiv \sum_g \mathrm{I}(A(g,m)=k)(\lambda(g)-\bar{\lambda}_k(m))^2/A_k(m)$ denote, respectively, the numbers, averages, and variances of annotated ($k=1$) and unannotated ($k=0$) gene-parameters $\lambda(g)$.

In commonly-encountered combined GO annotation and microarray data analyses, a binary gene-parameter profile could indicate whether genes are differentially expressed or not in two populations of cells, a continuous gene-parameter profile could consist of coefficients for the regression of a (censored) clinical outcome on gene expression measures, and binary gene-annotation profiles could denote whether genes are annotated or not with particular GO terms (Section 5; Al-Shahrour et al. (2004, 2005); Beissbarth and Speed (2004); Grossmann et al. (2006)).

3.3.2 Multivariate association measures

More generally, the mth association parameter could be based on the entire gene-annotation matrix A or a subset of columns thereof, that is, $\Psi(P)(m) = \psi(m) \equiv \rho_m(A, \Lambda(P))$, for an association measure $\rho_m : \mathbb{R}^{G\times M} \times \mathbb{R}^G \to \mathbb{R}$. Association parameters of interest include: linear combinations of association parameters for several features, partial correlation coefficients, χ^2-statistics for higher-dimensional contingency tables (e.g., with one dimension corresponding to a gene-parameter profile λ and other dimensions to several gene-annotation profiles $A(\cdot,m)$), (contrasts of) regression coefficients of a gene-parameter profile λ on several gene-annotation profiles $A(\cdot,m)$.

In the case of Gene Ontology annotation, the association parameter ψ should preferably reflect the structure of the GO directed acyclic graph, by taking into account, for instance, annotation information for ancestor (i.e., less specific) or offspring (i.e., more specific) terms (Section 4.1). Specifically, let $\mathcal{P}(m)$ denote the set of (immediate) parents of a term m. As the genes annotated by the child term m are subsets of the genes annotated by the parent terms $\mathcal{P}(m)$, then $A(g,m)=1$ implies $A(g,p)=1$ for $p \in \mathcal{P}(m)$.

Following the causal inference literature (van der Laan, 2006; van der Laan and Robins, 2003), an association parameter of interest for GO term m is the *marginal causal effect parameter*, defined as

$$\psi(m) = E[E[\lambda|A(\cdot,m)=1, A(\cdot,\mathcal{P}(m))]] - E[E[\lambda|A(\cdot,m)=0, A(\cdot,\mathcal{P}(m))]], \tag{17}$$

where $A(\cdot,\mathcal{P}(m))$ denotes the submatrix of gene-annotation profiles for parent terms $\mathcal{P}(m)$ and the expected values are defined with respect to the empirical distribution of $\{(A(g,m), A(g,\mathcal{P}(m)), \lambda(g)) : g=1,\ldots,G\}$.

In the special case of binary gene-parameter profiles (for differential expression), the so-called parent-child method of Grossmann et al. (2006) takes into account the structure of the GO DAG by testing for associations between gene-annotation and gene-parameter profiles using hypergeometric p-values computed conditionally on the annotation status of parent terms.

One could also consider Boolean combinations of annotation indicators for multiple features, that is, a transformed gene-annotation matrix whose columns are Boolean combinations of the columns of the original gene-annotation matrix. Such an approach would be particularly relevant in the context of transcription regulation, where individual features correspond to single transcription factor binding motifs and Boolean combinations to binding modules for multiple transcription factors.

3.4 Multiple hypothesis testing

3.4.1 Null and alternative hypotheses

Certain biological annotation metadata analyses may involve the *two-sided tests* of the M null hypotheses of no association between the gene-annotation profiles $A(\cdot, m)$ and a gene-parameter profile λ, i.e., tests of

$$H_0(m) \equiv \mathrm{I}(\psi(m) = \psi_0(m)) \qquad \text{vs.} \qquad H_1(m) \equiv \mathrm{I}(\psi(m) \neq \psi_0(m)). \tag{18}$$

Other analyses may call for the *one-sided tests* of

$$H_0(m) \equiv \mathrm{I}(\psi(m) \leq \psi_0(m)) \qquad \text{vs.} \qquad H_1(m) \equiv \mathrm{I}(\psi(m) > \psi_0(m)). \tag{19}$$

The M–vector $\psi_0 = (\psi_0(m) : m = 1, \ldots, M)$, of *null values* for the association parameter ψ, is determined by the biological question. For example, if $\psi(m) = \rho_m(A(\cdot, m), \lambda)$ is the Pearson correlation coefficient between the gene-annotation profile $A(\cdot, m)$ and the gene-parameter profile λ, then one may set $\psi_0(m) = 0$.

Note that in many situations, the same association measure ρ_m is used for each of the M features and one only has a single, common null value $\psi_0(m)$.

3.4.2 Test statistics

As in Chapter 1 of Dudoit and van der Laan (2006), consider the general situation where, given a random sample $\mathcal{X}_n$ from the data generating distribution P, one has an *asymptotically linear estimator* $\psi_n = \hat{\Psi}(P_n)$ of the association parameter vector $\psi = \Psi(P)$, with M–dimensional vector *influence curve* $IC(X \mid P)$. Let $\hat{\Sigma}(P_n) = \sigma_n = (\sigma_n(m, m') : m, m' = 1, \ldots, M)$ denote a consistent estimator of the covariance matrix $\Sigma(P) = \sigma = (\sigma(m, m') : m, m' = 1, \ldots, M)$ of the vector influence curve $IC(X \mid P)$. For example, σ_n could be a bootstrap-based estimator of the covariance matrix σ or could be computed from an estimator $IC_n(X)$ of the influence curve $IC(X \mid P)$.

Each null hypothesis $H_0(m)$ may then be tested using a (unstandardized) *difference statistic*,

$$T_n(m) \equiv \sqrt{n}\,(\psi_n(m) - \psi_0(m))\,, \tag{20}$$

or a (standardized) *t-statistic*,

$$T_n(m) \equiv \sqrt{n}\frac{\psi_n(m) - \psi_0(m)}{\sqrt{\sigma_n(m, m)}}. \tag{21}$$

Let $Q_n = Q_n(P)$ denote the typically unknown (finite sample) joint distribution of the M–vector of test statistics $T_n = (T_n(m) : m = 1, \ldots, M)$, under the data generating distribution P.

A broad range of association parameters ψ and corresponding estimators ψ_n satisfy the above conditions. In particular, suppose $\lambda_n = \hat{\Lambda}(P_n)$ is an asymptotically linear estimator of the gene-parameter profile $\lambda = \Lambda(P)$, based on a random sample $\mathcal{X}_n$ from P. Let $\psi_n \equiv \rho(A, \lambda_n)$ denote the corresponding *resubstitution*, or *plug-in*, estimator of the association parameter vector $\psi = \rho(A, \lambda)$. Then, if the function $\rho(A, \lambda)$ is differentiable with respect to λ, the resubstitution estimator ψ_n is also asymptotically linear. One can therefore handle tests where the gene-parameter profiles λ are (functions of) means, variances, correlation coefficients, and regression coefficients, and where the association measures ρ are correlation coefficients, two-sample t-statistics, and χ^2-statistics. Examples are provided in Section 5, in the context of tests of association between differential gene expression in ALL and GO annotation.

Certain testing problems may call for other test statistics T_n, such as F-statistics, χ^2-statistics, and likelihood ratio statistics.

3.4.3 Test statistics null distribution

As detailed in Chapter 2 of Dudoit and van der Laan (2006), a key feature of our proposed multiple testing procedures is the *test statistics null distribution* (rather than data generating null distribution) used to obtain rejection regions (i.e., cut-offs) for the test statistics, confidence regions for the parameters of interest, and adjusted p-values. In practice, the true distribution $Q_n(P)$ of the test statistics T_n is unknown and replaced by a null distribution Q_0. The choice of a suitable null distribution is crucial, in order to ensure that (finite sample or asymptotic) control of the Type I error rate under the *assumed null distribution* does indeed provide the desired control under the *true distribution*. This issue is particularly relevant for large-scale testing problems, such as those involving gene annotation metadata, which concern high-dimensional multivariate distributions, with complex and unknown dependence structures among variables.

Chapter 2 of Dudoit and van der Laan (2006) provides a general characterization for a proper test statistics null distribution, in terms of *null domination* conditions for the joint distribution of the $\mathcal{H}_0$–specific test statistics $(T_n(m) : m \in \mathcal{H}_0)$. This general characterization leads to the explicit proposal of two test statistics null distributions $Q_0 = Q_0(P)$: the asymptotic distribution of the vector of null value shifted and scaled test statistics and the asymptotic distribution of the vector of null quantile-transformed test statistics.

Specifically, the original null distribution of Dudoit et al. (2004b), van der Laan et al. (2004b), and Pollard and van der Laan (2004) is defined as the asymptotic distribution of the M–vector Z_n of *null value shifted and scaled test statistics*,

$$Z_n(m) \equiv \sqrt{\min\left(1, \frac{\tau_0(m)}{Var[T_n(m)]}\right)}(T_n(m) - E[T_n(m)]) + \lambda_0(m), \qquad (22)$$

where $\lambda_0(m)$ and $\tau_0(m)$ are, respectively, user-supplied upper bounds for the means and variances of the $\mathcal{H}_0$–specific test statistics. In this construction, the location null values $\lambda_0(m)$ are chosen such that the joint distribution of $(Z_n(m) : m \in \mathcal{H}_0)$ is asymptotically stochastically greater than that of $(T_n(m) : m \in \mathcal{H}_0)$. The scale null values $\tau_0(m)$ are chosen to prevent a degenerate limit for the false null hypotheses ($m \in \mathcal{H}_1$); an important issue for power considerations. For a broad class of testing problems, such as the test of single-parameter null hypotheses using t-statistics (Equation (21)), the null values are $\lambda_0(m) = 0$ and $\tau_0(m) = 1$ and the null distribution is an M–variate Gaussian distribution, with mean vector zero and covariance matrix equal to the correlation matrix of the vector influence curve. That is, $Q_0 = N(0, \sigma^*)$, where $\sigma^* = \Sigma^*(P) \equiv Cor[IC(X|P)]$. For testing the equality of K population mean vectors using F-statistics, the null values are $\lambda_0(m) = 1$ and $\tau_0(m) = 2/(K-1)$, under the assumption of equal variances in the different populations.

The second and most recent proposal of van der Laan and Hubbard (2005) is defined as the asymptotic distribution of the M–vector Z_n of *null quantile-transformed test statistics*,

$$Z_n(m) \equiv \dot{Q}_{0,m}^{-1} Q_{n,m}(T_n(m)), \tag{23}$$

where $\dot{Q}_{0,m}$ are user-supplied marginal test statistics null distributions that satisfy the marginal null domination condition $\liminf_n \ \dot{Q}_{0,m}^{-1} Q_{n,m}(z) \geq z$. This latest proposal has the advantage that the marginal test statistics null distributions may be set to the optimal, i.e., most powerful, null distributions one would use in single hypothesis testing (e.g., permutation null distributions, Gaussian or other parametric null distributions).

In practice, the test statistics null distribution $Q_0 = Q_0(P)$ is unknown, as it depends on the unknown data generating distribution P. Resampling procedures are provided to conveniently obtain consistent estimators of the null distribution and the corresponding test statistic cut-offs, parameter confidence regions, and adjusted p-values.

We stress the generality of the aforementioned test statistics null distributions: Type I error control does not rely on restrictive assumptions such as subset pivotality and holds for general data generating distributions (with arbitrary dependence structures among variables), null hypotheses (defined in terms of submodels for the data generating distribution), and test statistics (e.g., t-statistics, χ^2-statistics, F-statistics).

3.4.4 Multiple testing procedures

Having identified a suitable test statistics null distribution Q_0 (or estimator thereof, Q_{0n}), there remains the main task of specifying rejection regions (i.e., cut-offs) for the test statistics, confidence regions for the parameters of interest, and adjusted p-values. One can apply the multiple testing methodology developed in Dudoit and van der Laan (2006) and related articles to control a broad class of Type I error rates, defined as generalized tail probabilities, $gTP(q, g) = Pr(g(V_n, R_n) > q)$, and generalized expected values,

$gEV(g) = E[g(V_n, R_n)]$, for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n.

An overview of available MTPs is provided in Chapter 3 of Dudoit and van der Laan (2006). Core methodological Chapters 4 – 8 discuss the following main approaches for deriving rejection regions.

Chapter 4. *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 (Dudoit et al., 2004b; Pollard and van der Laan, 2004). Error rates of the form $\theta(F_{V_n})$ include the generalized family-wise error rate, $gFWER(q) = 1 - F_{V_n}(q) = Pr(V_n > q)$.

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

Chapter 6. *Augmentation multiple testing procedures* (AMTP) for controlling *generalized tail probability* error rates, $gTP(q, g) = Pr(g(V_n, R_n) > q)$, for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n, based on an initial gFWER-controlling procedure (Dudoit et al., 2004a; van der Laan et al., 2004a). Error rates treated in detail include the gFWER, with $g(v, r) = v$, and TPPFP, with $g(v, r) = v/r$.

Chapter 7. *Resampling-based empirical Bayes procedures* for controlling *generalized tail probability* error rates. The special case of TPPFP control is discussed in detail in van der Laan et al. (2005).

These multiple testing procedures are implemented in the Bioconductor R package multtest (Pollard et al. (2005b); `www.bioconductor.org`).

3.4.5 FWER-controlling single-step common-cut-off maxT procedure

For the purpose of illustration, we focus on control of the family-wise error rate, using the single-step maxT procedure, a common-cut-off procedure exploiting the joint distribution of the test statistics. We rely on the *bootstrap* to yield a consistent estimator Q_{0n} of the null value shifted and scaled test statistics null distribution Q_0 and corresponding single-step maxT cut-offs and adjusted p-values. The method is summarized below for convenience; details are given in Chapter 4 of Dudoit and van der Laan (2006) and in Dudoit et al. (2004b).

Procedure 1 [Single-step common-cut-off maxT procedure] *Given an M–variate test statistics null distribution Q_0, the* single-step common-cut-off maxT procedure *is based on the distribution of the* maximum test statistic $\max_m Z(m)$*, for a random M–vector $Z = (Z(m) : m = 1, \ldots, M) \sim Q_0$. For controlling the FWER at nominal level $\alpha \in [0, 1]$, the common cut-off $c(Q_0, \alpha)$*

is defined as the $(1-\alpha)$*–quantile of the distribution of* $\max_m Z(m)$*, that is,*

$$c(Q_0, \alpha) \equiv \inf \left\{ z : Pr_{Q_0} \left(\max_m Z(m) \leq z \right) \geq (1-\alpha) \right\}. \tag{24}$$

The adjusted p-value $\widetilde{p}_{0n}(m)$ *for null hypothesis* $H_0(m)$ *is the probability, under* Q_0*, that* $\max_m Z(m)$ *exceeds the corresponding observed test statistic* $t_n(m)$*, that is,*

$$\widetilde{p}_{0n}(m) = Pr_{Q_0} \left(\max_m Z(m) \geq t_n(m) \right), \qquad m = 1, \ldots, M. \tag{25}$$

Procedure 2 [Bootstrap-based single-step common-cut-off maxT procedure]

1. *Given* B *(non-parametric or model-based) bootstrap samples of the data* $\mathcal{X}_n$*, obtain an* $M \times B$ *matrix of test statistics,* $\mathbf{T}_n^B = \left(T_n^B(m,b) \right)$*, with rows corresponding to the* M *null hypotheses and columns to the* B *bootstrap samples.*

2. *Compute row means and variances of the matrix* $\mathbf{T}_n^B$*, to yield estimates of the means,* $E[T_n(m)]$*, and variances,* $Var[T_n(m)]$*, of the test statistics under the true data generating distribution* P*. That is, compute*

$$\begin{aligned} E[T_n^B(m,\cdot)] &\equiv \frac{1}{B} \sum_{b=1}^{B} T_n^B(m,b) \\ \text{and} \qquad Var[T_n^B(m,\cdot)] &\equiv \frac{1}{B} \sum_{b=1}^{B} (T_n^B(m,b) - E[T_n^B(m,\cdot)])^2. \end{aligned}$$

3. *Obtain an* $M \times B$ *matrix,* $\mathbf{Z}_n^B = \left(Z_n^B(m,b) \right)$*, of null value shifted and scaled bootstrap statistics* $Z_n^B(m,b)$*, by row-shifting and scaling the matrix* $\mathbf{T}_n^B$ *using the bootstrap estimates of* $E[T_n(m)]$ *and* $Var[T_n(m)]$ *and the user-supplied null values* $\lambda_0(m)$ *and* $\tau_0(m)$*. That is,*

$$Z_n^B(m,b) \equiv \sqrt{\min \left(1, \frac{\tau_0(m)}{Var[T_n^B(m,\cdot)]} \right)} \left(T_n^B(m,b) - E[T_n^B(m,\cdot)] \right) + \lambda_0(m). \tag{26}$$

For t-statistics defined as in Equation (21), the null values are $\lambda_0(m) = 0$ *and* $\tau_0(m) = 1$*.*

4. *Estimate the null distribution* Q_0 *by the empirical distribution* Q_{0n} *of the* B *columns of matrix* $\mathbf{Z}_n^B$*.*

5. *Compute the maximum statistic,* $\max_m Z_n^B(m,b)$, $b = 1, \ldots, B$, *for each bootstrap dataset, i.e., each column of the matrix* $\mathbf{Z}_n^B$.

6. *For controlling the FWER at nominal level* $\alpha \in [0,1]$, *the bootstrap single-step maxT common cut-off* $c(Q_{0n}, \alpha)$ *is defined as the* $(1-\alpha)$*–quantile of the empirical distribution of the* B *maxima* $\{\max_m Z_n^B(m,b) : b = 1, \ldots, B\}$, *that is,*

$$c(Q_{0n}, \alpha) \equiv \inf\left\{ z : \frac{1}{B}\sum_{b=1}^{B} \mathrm{I}\left(\max_m Z_n^B(m,b) \leq z\right) \geq (1-\alpha) \right\}. \quad (27)$$

7. *The bootstrap single-step maxT adjusted p-value* $\widetilde{p}_{0n}(m)$ *for null hypothesis* $H_0(m)$ *is the proportion of maxima* $\{\max_m Z_n^B(m,b) : b = 1, \ldots, B\}$ *that exceed the corresponding observed test statistic* $t_n(m)$, *that is,*

$$\widetilde{p}_{0n}(m) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{I}\left(\max_m Z_n^B(m,b) \geq t_n(m)\right), \qquad m = 1, \ldots, M. \quad (28)$$

4 The Gene Ontology

4.1 Overview of 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, cellular components, and molecular functions. The ontologies specify terminologies and relationships among terms. They are organism-independent and can be applied even as our knowledge of the roles of genes and proteins is accumulating and changing. The GO Consortium and other organizations supply *mappings* between GO terms and genes in various organisms. Detailed documentation is available in the "Gene Ontology Documentation" webpage (`www.geneontology.org/GO.contents.doc.html`).

4.1.1 The three gene ontologies: BP, CC, and MF

The GO Consortium provides three ontologies, each consisting of a structured network of terms describing gene products.

- *Biological Process* (BP or P). The Biological Process ontology refers to series of biological events that are accomplished by one or more ordered assemblies of molecular functions. Examples of broad BP terms are *cellular physiological process* (`GO:0050875`) and *signal transduction* (`GO:0007165`);

examples of more specific BP terms are *pyrimidine base metabolism* (GO:0006206) and *alpha-glucoside transport* (GO:0000017).

- *Cellular Component* (CC or C). The Cellular Component ontology refers to subcellular structures, with the proviso that the components be part of some larger object, which may be an anatomical structure (e.g., *rough endoplasmic reticulum* (GO:0005791), *nucleus* (GO:0005634)) or a gene product group (e.g., *ribosome* (GO:0005840)).

- *Molecular Function* (MF or F). The Molecular Function ontology refers to tasks or activities performed by individual (or assembled complexes of) gene products. Examples of broad MF terms are *catalytic activity* (GO:0003824), *transporter activity* (GO:0005215), and *binding* (GO:0005488); examples of narrower MF terms are *adenylate cyclase activity* (GO:0004016) and *Toll binding* (GO:0005121).

A gene product may be used in one or more biological processes, may be associated with one or more cellular components, and may have one or more molecular functions.

Example: Gene product *ABL1_HUMAN***.** The *Homo sapiens* gene product *Splice Isoform IA of Proto-oncogene tyrosine-protein kinase ABL1* (*ABL1_HUMAN*) can be described by the following terms in each of the three gene ontologies (AmiGO browser; Last updated 2006-02-14; www.godatabase.org/cgi-bin/amigo/go.cgi?view=details&search_constraint=gp&session_id=6973b1139030258&gp=P00519).

- *Biological Process*: *regulation of progression through cell cycle* (GO:0000074); *S-phase-specific transcription in mitotic cell cycle* (GO:0000115); *mismatch repair* (GO:0006298); *regulation of transcription, DNA-dependent* (GO:0006355); *DNA damage response, signal transduction resulting in induction of apoptosis* (GO:0008630).

- *Cellular Component*: *nucleus* (GO:0005634).

- *Molecular Function*: *DNA binding* (GO:0003677); *protein-tyrosine kinase activity* (GO:0004713); *protein binding* (GO:0005515).

4.1.2 GO directed acyclic graphs

For each of the three gene ontologies, GO terms are organized in a *directed acyclic graph* (DAG), where a *directed* graph has one-way edges and an *acyclic* graph has no path starting and ending at the same vertex. Each GO term is associated with a single vertex, or node, in the DAG. The words *term*, *node*, and *vertex*, may therefore be used interchangeably.

For a given GO term, an *ancestor* refers to a less specialized term; an *offspring* refers to a more specialized term. A *parent* is an immediate/direct ancestor of a term; a *child* is an immediate/direct offspring of a term. A *root node* has

no parents, i.e., no incoming edges; a *leaf node* has no children, i.e., no outgoing edges. In a DAG, a child may have several parents.

Every GO term must obey the so-called *true path rule*: if a (child) term describes a gene product, then all its immediate parent and more distant ancestor terms must also apply to the gene product.

The DAG structure of GO terms and corresponding true path rule are germane to the definition of a suitable association measure between gene-annotation profiles and gene-parameter profiles (Section 3.3). Furthermore, as discussed in Sections 4.2 – 4.5, in the context of Bioconductor annotation software, the true path rule is also relevant when assembling gene-annotation matrices.

4.1.3 GO software tools

Many software tools have been developed to deal with GO annotation metadata. The "Gene Ontology Tools" webpage (`www.geneontology.org/GO.tools.shtml`) provides a list of consortium and non-consortium software for searching and browsing the three gene ontologies, for annotating genes and gene products using GO, and for combined GO and gene expression microarray data analysis.

For instance, the `AmiGO` browser (`www.godatabase.org`) allows: searching for a GO term and viewing all gene products annotated with this term; searching for a gene product and viewing all its associated GO terms; browsing the ontologies to view relationships among terms and gene products annotated with a given term.

The `QuickGO` browser (`www.ebi.ac.uk/ego`), developed by the European Bioinformatics Institute (EBI), also permits searches and graphical displays of the Gene Ontology by GO term, GO term identifier (ID), gene product, and other identifiers.

Software packages developed as part of the Bioconductor Project are discussed in Sections 4.2 – 4.5.

Example: GO term *protein-tyrosine kinase activity*. To get a sense of the information provided by the GO Consortium, consider the Molecular Function ontology and the GO term *protein-tyrosine kinase activity*, with GO term ID `GO:0004713`.

Go to the `AmiGO` browser (`www.godatabase.org`), enter the GO term ID `GO:0004713` in the `Search GO` box, select `Exact Match`, select `Terms`, and click on the `Submit Query` button. There are two main options for displaying information on a GO term: a "tree view" and a "graphical view". Click on the small tree-like icon (top-left corner of the table) to display the tree view with all ancestors (i.e., less specific terms) of the GO term *protein-tyrosine kinase activity*. Click on the `Graphical View` button to display the portion of the MF DAG corresponding to the GO term. Additional information may be obtained by clicking on the hyperlinked text `protein-tyrosine kinase activity`.

The GO term *protein-tyrosine kinase activity* has one (immediate) parent, *protein kinase activity* (`GO:0004672`), which itself has two parents, *kinase activity* (`GO:0016301`) and *phosphotransferase activity, alcohol group as acceptor*

(GO:0016773). Altogether, the term *protein-tyrosine kinase activity* has 7 ancestors. According to the true path rule, any gene annotated with the GO term *protein-tyrosine kinase activity* should also be annotated with all of its less specific ancestor terms.

The portion of the MF DAG for the GO term *protein-tyrosine kinase activity* is displayed in Figures 2 and 3 using, respectively, the AmiGO and QuickGO browsers (note the different ordering of nodes in these two representations: for AmiGO, the offspring nodes are at the top of the graph, while for QuickGO, they are at the bottom of the graph).

4.1.4 GO gene-annotation matrices

For each of the three gene ontologies, one may define a $G \times M$ binary gene-annotation matrix A, indicating for each gene g whether it is annotated or not with each GO term m,

$$A(g,m) \equiv \begin{cases} 1, & \text{if gene } g \text{ is annotated with GO term } m \\ 0, & \text{otherwise} \end{cases}, \qquad (29)$$
$$g = 1, \ldots, G,\ m = 1, \ldots, M.$$

Section 4.5 provides sample R code for assembling GO gene-annotation matrices using Bioconductor annotation metadata packages.

As detailed in Section 3, detecting associations between GO annotation and other interesting features of a genome may be viewed as the multiple tests of the null hypotheses of no association between a gene-parameter profile $\lambda = \Lambda(P)$ and gene-annotation profiles $A(\cdot, m)$. The multiple testing methodology proposed in Dudoit and van der Laan (2006) and related articles is well-suited to handle the complex and unknown dependence structure among test statistics implied by the DAG structure of GO terms. The methods are illustrated in Section 5, for tests of association between differential gene expression in ALL and GO annotation.

4.2 Overview of R and Bioconductor software for GO annotation metadata analysis

As discussed in Gentleman et al. (2005), the *Bioconductor Project* (www.bioconductor.org) provides R packages for accessing and performing statistical inference with GO annotation metadata. The packages include: a general annotation software package (annotate); packages for graph theoretical analyses (e.g., graph, Rgraphviz); a GO-specific metadata package for navigating the three GO DAGs (GO); an Entrez Gene[1]-specific metadata package, providing bi-directional mappings between Entrez Gene IDs and GO term IDs (humanLLMappings; www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene); various Affymetrix chip-specific

[1]N.B. The LocusLink database has been superseded by the Entrez Gene database.

metadata packages, providing bi-directional mappings between Affymetrix probe[2] IDs and GO term IDs (e.g., hu6800, hgu95av2; www.affymetrix.com); a package for annotating and generating HTML reports for Affymetrix chip data (annaffy).

Bioconductor metadata packages are updated regularly to reflect the evolving nature of biological annotation metadata; it is therefore crucial to keep track of *version* numbers. For information on Bioconductor software, please consult the "Documentation" (www.bioconductor.org/docs) and "Workshops" (www.bioconductor.org/workshops) sections of the Bioconductor Project website, in addition to the standard R help facilities (e.g., help function, manuals, etc.).

In order to run through the examples below, one needs to install and load the following Bioconductor packages: annotate, GO, hgu95av2. The annotation metadata used in the examples correspond to the following package versions.

```
> library(annotate)
> library(GO)
> library(hgu95av2)
>
> packageDescription("annotate")$Version
[1] "1.8.0"
> packageDescription("GO")$Version
[1] "1.10.0"
> packageDescription("hgu95av2")$Version
[1] "1.10.0"
```

Accessing and analyzing annotation metadata from databases such as GenBank (www.ncbi.nlm.nih.gov/Genbank), GO (www.geneontology.org), and PubMed (www.pubmed.gov), presupposes the ability to perform the following essential bookkeeping task: *mapping between different identifiers (ID) for a given gene/probe*. Bioconductor annotation metadata packages consist of *environment* objects that provide *key-value mappings* between different sets of gene/probe identifiers.

For instance, in the hgu95av2 annotation metadata package, for the Affymetrix chip series HG-U95Av2, the hgu95av2PMID environment provides mappings from Affymetrix probe IDs (keys) to PubMed IDs (values); similarly, the hgu95av2GO environment provides mappings from Affymetrix probe IDs (keys) to GO term IDs (values).

Example: Affymetrix probe ID 1635_at. As of Version 1.10.0 of the hgu95av2 package, the Affymetrix probe with ID 1635_at corresponds to the gene with symbol ABL1 and long name v-abl Abelson murine leukemia viral oncogene homolog 1, located on the long arm of chromosome 9. This probe maps to one GenBank accession number, one Entrez Gene ID, 14 distinct GO term IDs, and 160 distinct PubMed IDs.

[2]N.B. In the context of Affymetrix oligonucleotide chips, we use the shorter term *probe* to refer to a *probe-pair-set*, i.e., a collection of perfect match (PM) and mismatch (MM) *probe-pairs* that map to a particular gene.

```
> probe <- "1635_at"
> get(probe, env=hgu95av2SYMBOL)
[1] "ABL1"
> get(probe, env=hgu95av2GENENAME)
[1] "v-abl Abelson murine leukemia viral oncogene homolog 1"
> get(probe, env=hgu95av2MAP)
[1] "9q34.1"
> get(probe, env=hgu95av2ACCNUM)
[1] "U07563"
> get(probe, env=hgu95av2LOCUSID )
[1] 25
> unique(names(get(probe, env=hgu95av2GO)))
 [1] "GO:0000074" "GO:0000115" "GO:0000166" "GO:0003677" "GO:0004713"
 [6] "GO:0005515" "GO:0005524" "GO:0005634" "GO:0006298" "GO:0006355"
[11] "GO:0006468" "GO:0007242" "GO:0008630" "GO:0016740"
> length(get(probe, env=hgu95av2PMID))
[1] 160
```

The remainder of this section gives a brief overview of two main types of Bioconductor annotation metadata packages: the GO package (Section 4.3) and the hgu95av2 package for the Affymetrix chip series HG-U95Av2 (Section 4.4). Section 4.5 illustrates how these two packages may be used to assemble a GO gene-annotation matrix.

4.3 The GO annotation metadata package

The GO package provides environment objects containing key-value pairs for mappings between GO term IDs, GO terms, GO term ancestors, GO term parents, GO term children, GO term offspring, and Entrez Gene IDs. The `GO()` command lists all environments available in the GO package.

```
> GO()

Quality control information for  GO
Date built: Created: Fri Sep 30 03:02:24 2005

Mappings found for non-probe based rda files:
        GOALLLOCUSID found 9556
        GOBPANCESTOR found 9888
        GOBPCHILDREN found 4989
        GOBPOFFSPRING found 4989
        GOBPPARENTS found 9888
        GOCCANCESTOR found 1612
        GOCCCHILDREN found 578
        GOCCOFFSPRING found 578
        GOCCPARENTS found 1612
```

```
GOLOCUSID2GO found 70818
GOLOCUSID found 8017
GOMFANCESTOR found 7334
GOMFCHILDREN found 1403
GOMFOFFSPRING found 1403
GOMFPARENTS found 7334
GOOBSOLETE found 1032
GOTERM found 18834
```

For information on any of the GO environments, use the `help` function, e.g., `help(GOTERM)` or `? GOBPPARENTS`. For instance, the environment `GOTERM` provides mappings from *GO term IDs* (keys) to *GO terms* (values); the environments `GOBPPARENTS`, `GOCCPARENTS`, and `GOMFPARENTS`, provide ontology-specific mappings from *GO term IDs* (keys) to *GO term parent IDs* (values). The environments `GOALLLOCUSID`, `GOLOCUSID2GO`, and `GOLOCUSID`, provide mappings between *GO term IDs* and *Entrez Gene IDs* and are used in Section 4.5, below, to assemble an Entrez Gene ID-by-GO term ID gene-annotation matrix for the MF gene ontology.

Example: GO term ID `GO:0004713`**.** Let us use the GO package to obtain information on (all) ancestors, the (immediate) parents, the (immediate) children, and (all) offspring of the term corresponding to the GO term ID `GO:0004713`.

```
> ## List all GO IDs
> GOID <- ls(env = GOTERM)
> length(GOID)
[1] 18834
> GOID[1:10]
 [1] "GO:0000001" "GO:0000002" "GO:0000003" "GO:0000004" "GO:0000006"
 [6] "GO:0000007" "GO:0000009" "GO:0000010" "GO:0000011" "GO:0000012"
>
> ## Get information on GO term corresponding to GO ID GO:0004713
> GOID <- "GO:0004713"
> term <- get(GOID,env=GOTERM)
> class(term)
[1] "GOTerms"
attr(,"package")
[1] "annotate"
> slotNames(term)
[1] "GOID"       "Term"       "Synonym"    "Secondary"  "Definition"
[6] "Ontology"
> term
GOID = GO:0004713
Term = protein-tyrosine kinase activity
Synonym = protein tyrosine kinase activity
Definition = Catalysis of the reaction: ATP + a protein tyrosine = ADP
```

```
     + protein tyrosine phosphate.
Ontology = MF
>
> ## Get GO IDs of parents
> parents <- get(GOID,env=GOMFPARENTS)
> parents
         isa
"GO:0004672"
> mget(parents,env=GOTERM)
$"GO:0004672"
GOID = GO:0004672
Term = protein kinase activity
Definition = Catalysis of the transfer of a phosphate group, usually
     from ATP, to a protein substrate.
Ontology = MF

>
> ## Get GO IDs of ancestors
> ancestors <- get(GOID,env=GOMFANCESTOR)
> ancestors
[1] "all"        "GO:0003674" "GO:0003824" "GO:0016740" "GO:0016772"
[6] "GO:0016773" "GO:0016301" "GO:0004672"
>
> ## Get GO IDs of children
> children <- get(GOID,env=GOMFCHILDREN)
> children
[1] "GO:0004714" "GO:0004715" "GO:0004716"
>
> ## Get GO IDs of offspring
> offspring <- get(GOID,env=GOMFOFFSPRING)
> offspring
 [1] "GO:0004714" "GO:0004715" "GO:0004716" "GO:0005020" "GO:0005021"
 [6] "GO:0005023" "GO:0005010" "GO:0005011" "GO:0005017" "GO:0005003"
[11] "GO:0005006" "GO:0005007" "GO:0005008" "GO:0005009" "GO:0008288"
[16] "GO:0005018" "GO:0005019" "GO:0005004" "GO:0005005" "GO:0008313"
[21] "GO:0004718"
```

As already noted in the example on p. 23 and Figures 2 and 3, the term corresponding to the GO term ID `GO:0004713` is *protein-tyrosine kinase activity*, in the Molecular Function ontology. It has one (immediate) parent term, *protein kinase activity*.

4.4 Affymetrix chip-specific annotation metadata packages: The hgu95av2 package

The Bioconductor Project provides *Affymetrix chip-specific annotation metadata packages* for the main chip series for the human, mouse, rat, and other genomes, e.g., HU-6800, HG-U133, HG-U95, MG-U74, and RG-U34 series. These packages, built using the infrastructure package AnnBuilder, contain environment objects for mappings between Affymetrix probe IDs and other types of gene/probe identifiers.

Note that analogous packages are not supplied for two-color spotted microarrays, as there is no standard microarray design for this type of platform and specialized annotation metadata packages may have to be created for each microarray facility (e.g., using AnnBuilder). Once annotation metadata packages are available to provide mappings between different sets of gene/probe identifiers, the tools in annotate and related packages may be used in a similar manner for any type of microarray platform.

Consider the hgu95av2 package, for the Affymetrix chip series HG-U95Av2. This package provides the following environments.

```
> ? hgu95av2
> hgu95av2()

Quality control information for  hgu95av2
Date built: Created: Tue Oct  4 21:31:35 2005

Number of probes: 12625
Probe number missmatch: None
Probe missmatch: None
Mappings found for probe based rda files:
         hgu95av2ACCNUM found 12625 of 12625
         hgu95av2CHRLOC found 11673 of 12625
         hgu95av2CHR found 12145 of 12625
         hgu95av2ENZYME found 1886 of 12625
         hgu95av2GENENAME found 11418 of 12625
         hgu95av2GO found 9942 of 12625
         hgu95av2LOCUSID found 12203 of 12625
         hgu95av2MAP found 12109 of 12625
         hgu95av2OMIM found 9881 of 12625
         hgu95av2PATH found 3928 of 12625
         hgu95av2PMID found 12086 of 12625
         hgu95av2REFSEQ found 12008 of 12625
         hgu95av2SUMFUNC found 0 of 12625
         hgu95av2SYMBOL found 12159 of 12625
         hgu95av2UNIGENE found 12118 of 12625
Mappings found for non-probe based rda files:
         hgu95av2CHRLENGTHS found 25
```

```
hgu95av2ENZYME2PROBE found 643
hgu95av2GO2ALLPROBES found 5480
hgu95av2GO2PROBE found 3890
hgu95av2ORGANISM found 1
hgu95av2PATH2PROBE found 155
hgu95av2PFAM found 10439
hgu95av2PMID2PROBE found 98214
hgu95av2PROSITE found 8249
```

For more information on any of these environments, use the `help` function, e.g., `help(hgu95av2GO)` or `? hgu95av2GO`. We focus on the three environments related to GO: `hgu95av2GO`, `hgu95av2GO2ALLPROBES`, and `hgu95av2GO2PROBE`. The HG-U95Av2 chip contains 12,625 probes (keys in the `hgu95av2GO` environment), with the first 10 Affymetrix probe IDs listed below.

```
> ## List all Affymetrix IDs
> AffyID <- ls(env = hgu95av2GO)
> length(AffyID)
[1] 12625
> AffyID[1:10]
 [1] "1000_at"   "1001_at"   "1002_f_at" "1003_s_at" "1004_at"   "1005_at"
 [7] "1006_at"   "1007_s_at" "1008_f_at" "1009_at"
```

4.4.1 Probes-to-most specific GO terms mappings: The `hgu95av2GO` environment

The `hgu95av2GO` environment contains key-value pairs for the mappings from *Affymetrix probe IDs* (keys) to *GO term IDs* (values). Each Affymetrix probe ID is mapped to a list of one or more elements, where each element corresponds to a particular GO term and is itself a list of the following three elements.

- `"GOID"`: A GO term ID corresponding to the Affymetrix probe ID (key).
- `"Evidence"`: A code for the evidence supporting the association of the GO term to the Affymetrix probe.
- `"Ontology"`: An abbreviation for the name of the ontology to which the GO term belongs: BP (Biological Process), CC (Cellular Component), or MF (Molecular Function).

Note that only the *directly associated terms* or *most specific terms* (i.e., not their less specific ancestor terms) a probe is annotated with are returned as values in `hgu95av2GO`. The `GO` package (Section 4.3) may be used to obtain more information on the GO term IDs, e.g., GO term, (all) ancestors, (immediate) parents, (immediate) children, and (all) offspring.

Example: GO terms directly associated with Affymetrix probe ID `1635_at`. Let us obtain GO annotation information for the probe with Affymetrix

ID 1635_at, corresponding to the ABL1 gene. The code below shows that probe 1635_at is directly annotated with 14 distinct GO terms (the same GO term ID may be returned multiple times with a different evidence code). As already noted in the example on p. 22, one of these terms, with GO term ID GO:0004713, is *protein-tyrosine kinase activity*, in the Molecular Function ontology.

```
> probe <- "1635_at"
> probe2GO <- get(probe, env = hgu95av2GO)
> length(probe2GO)
[1] 14
> unique(names(probe2GO))
 [1] "GO:0000074" "GO:0000115" "GO:0000166" "GO:0003677" "GO:0004713"
 [6] "GO:0005515" "GO:0005524" "GO:0005634" "GO:0006298" "GO:0006355"
[11] "GO:0006468" "GO:0007242" "GO:0008630" "GO:0016740"
> probe2GO[[5]]
$GOID
[1] "GO:0004713"

$Evidence
[1] "TAS"

$Ontology
[1] "MF"

> get(probe2GO[[5]]$GOID, env=GOTERM)
GOID = GO:0004713
Term = protein-tyrosine kinase activity
Synonym = protein tyrosine kinase activity
Definition = Catalysis of the reaction: ATP + a protein tyrosine = ADP
     + protein tyrosine phosphate.
Ontology = MF
```

The hgu95av2GO environment (and analogous environments for other chip series) may be used to assemble an Affymetrix probe ID-by-GO term ID gene-annotation matrix, row by row. This may entail, however, a number of data processing steps. Firstly, only the most specific terms a probe is annotated with are returned as values in hgu95av2GO. One therefore needs to add all ancestor (less specific) terms in order to comply with the true path rule. Secondly, several probes may correspond to the same gene, i.e., several Affymetrix probe IDs may map to the same Entrez Gene ID according to the environment hgu95av2LOCUSID. Thirdly, the hgu95av2GO environment returns GO terms for all three gene ontologies at once. One may need to separate terms according to membership in the BP, CC, and MF ontologies (e.g., using the GOTERM environment from the GO package).

Alternately, one may assemble an Affymetrix probe ID-by-GO term ID gene-annotation matrix, column by column, using the hgu95av2GO2ALLPROBES environment described below.

4.4.2 GO terms-to-directly annotated probes mappings: The hgu95av2GO2PROBE environment

The hgu95av2GO2PROBE environment provides key-value pairs for the mappings from *GO term IDs* (keys) to *Affymetrix probe IDs* (values). Values are vectors of length one or greater depending on whether a given GO term ID is mapped to one or more Affymetrix probe IDs. The value names are evidence codes for the GO term IDs.

Note that the probes a particular GO term is mapped to are only those associated *directly* with the GO term (vs. indirectly via its immediate children or more distant offspring). For a list of all probes associated directly or indirectly with a particular GO term, one may use the hgu95av2GO2ALLPROBES environment.

Example: Affymetrix probes directly associated with GO term ID GO:0004713. In the following example, 205 distinct Affymetrix probe IDs are associated directly with the GO term *protein-tyrosine kinase activity* (GO:0004713). The Affymetrix probe IDs include 1635_at, corresponding to the ABL1 gene.

```
> GOID <- "GO:0004713"
> GO2Probes <- get(GOID, env = hgu95av2GO2PROBE)
> length(unique(GO2Probes))
[1] 205
> GO2Probes[1:10]
       <NA>         <NA>         <NA>         <NA>         <NA>          TAS
  "1635_at"  "1636_g_at"  "1656_s_at"  "2040_s_at"  "2041_i_at"   "39730_at"
        IEA          IEA          IEA          TAS
  "1084_at" "35162_s_at"    "1564_at"     "854_at"
> is.element("1635_at", GO2Probes)
[1] TRUE
```

4.4.3 GO terms-to-all annotated probes mappings: The hgu95av2GO2ALLPROBES environment

The hgu95av2GO2ALLPROBES environment provides key-value pairs for the mappings from *GO term IDs* (keys) to *Affymetrix probe IDs* (values). Values are vectors of length one or greater depending on whether a given GO term ID is mapped to one or more Affymetrix probe IDs. The value names are evidence codes for the GO term IDs.

Note that, in accordance with the true path rule, the probes a particular GO term is mapped to are associated either *directly* with the GO term or *indirectly* via any of its immediate children or more distant offspring. The main difference between the hgu95av2GO2PROBE and hgu95av2GO2ALLPROBES environments is that the former considers only the GO term itself, while the later considers the GO term and any of its descendants. Thus, the Affymetrix probe IDs returned by hgu95av2GO2PROBE are a subset of the probe IDs returned by

hgu95av2GO2ALLPROBES.

Example: Affymetrix probes directly or indirectly associated with GO term ID GO:0004713**.** In the following example, 319 distinct Affymetrix probe IDs (some with multiple evidence codes) are associated either directly or indirectly with the GO term ID GO:0004713. This list of 319 Affymetrix probe IDs indeed includes the list of 205 probe IDs associated directly with the term GO:0004713.

```
> GOID <- "GO:0004713"
> GO2AllProbes <- get(GOID, env = hgu95av2GO2ALLPROBES)
> length(GO2AllProbes)
[1] 370
> length(unique(GO2AllProbes))
[1] 319
> sum(is.element(GO2Probes,GO2AllProbes))
[1] 205
```

The hgu95av2GO2ALLPROBES environment immediately yields an Affymetrix probe ID-by-GO term ID gene-annotation matrix, column by column. However, as with the hgu95av2GO environment, a number of data processing steps may be required, concerning, for example, uniqueness of Entrez Gene IDs and membership in the BP, CC, and MF ontologies.

4.5 Assembling a GO gene-annotation matrix

This section provides R code for assembling an Entrez Gene ID-by-GO term ID gene-annotation matrix A, column by column. Specifically, rows correspond to (unique) Entrez Gene IDs mapping to probes on the HG-U95Av2 chip and columns to terms in the Molecular Function ontology that map directly or indirectly to at least 10 Entrez Gene IDs for the HG-U95Av2 chip.

In practice, it may not be desirable to build the full $G \times M$ gene-annotation matrix, as this matrix could potentially be very large and sparse (padded with zeros). Rather, we assemble a (smaller) *gene-annotation list*, that provides, for each GO term ID, a list of Entrez Gene IDs annotated with the GO term.

```
> ## List all Affymetrix IDs for HG-U95Av2 chip
> AffyID <- ls(env=hgu95av2GO)
> length(AffyID)
[1] 12625
>
> ## Get all unique Entrez Gene IDs for HG-U95Av2 chip
> LLID <- as.character(unique(unlist(mget(AffyID, env=hgu95av2LOCUSID))))
> length(LLID)
[1] 9085
>
> ## Get MF GO IDs
```

```
> GOID <- ls(env=GOTERM)
> O <- unlist(lapply(mget(GOID, env=GOTERM), function(z) z@Ontology))
> table(O)
O
  BP   CC   MF
9888 1612 7334
> MFID <- GOID[O=="MF"]
>
> ## For each MF GO ID, get all Entrez Gene IDs for genes annotated directly or indirectly w
> allMFLLID <- mget(MFID, env=GOALLLOCUSID)
>
> ## For each MF GO ID, get HG-U95Av2-specific Entrez Gene IDs for genes annotated directly
> MFLLID <- lapply(allMFLLID, function(z) intersect(z, LLID))
> numMFLLID <- unlist(lapply(MFLLID, length))
> summary(numMFLLID)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   0.000    1.000    1.000    9.539    1.000 6762.000
>
> ## Retain only MF GO IDs that map to at least 10 Entrez Gene IDs for the HG-U95Av2 chip
> MFAnnotList <- MFLLID[numMFLLID > 9]
> length(MFAnnotList)
[1] 466
> summary(unlist(lapply(MFAnnotList, length)))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   10.0    16.0    27.5   132.2    70.0  6762.0
> MFAnnotList[1]
$"GO:0000146"
 [1] "4620"  "4621"  "4624"  "4625"  "4640"  "4643"  "4644"  "4646"  "4647"
[10] "4650"  "58498"

>
> ## Get Affymetrix IDs for probes annotated with GO term ID GO:0004713
> is.element("GO:0004713",names(MFAnnotList))
[1] TRUE
> length(MFAnnotList["GO:0004713"][[1]])
[1] 180
```

5 Tests of association between GO annotation and differential gene expression in ALL

5.1 Acute lymphoblastic leukemia dataset of Chiaretti et al. (2004)

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) and Gene Ontology (GO) annotation metadata.

5.1.1 Bioconductor experimental data R package ALL

The ALL dataset is available in the Bioconductor experimental data R package ALL (Version 1.0.2, Bioconductor Release 1.7). The main object in this package is `ALL`, an instance of the class *exprSet*. The `exprs` slot of `ALL` provides a matrix of 12,625 Affymetrix *expression measures* (chip series HG-U95Av2) for each of 128 ALL cell samples. The `phenoData` slot contains 21 *phenotypes*, i.e., responses and covariates, for each of the 128 cell samples. Phenotypes of interest include: `ALL$BT`, the type and stage of the cancer, i.e., B-cell ALL or T-cell ALL, of stage 1, 2, 3, or 4; `ALL$mol.biol`, the molecular class of the cancer, i.e., BCR/ABL, NEG, ALL1/AF4, E2A/PBX1, p15/p16, or NUP-98.

The expression measures have been obtained using the three-step robust multichip average (RMA) pre-processing method, implemented in the Bioconductor R package affy (Bolstad et al., 2005), and have been subjected to a base 2 logarithmic transformation. For greater detail on the ALL dataset, please consult the ALL package documentation.

5.1.2 The BCR/ABL fusion

A number of recent articles have investigated the prognostic relevance of the *BCR/ABL fusion* in adult ALL of the B-cell lineage (Gleissner et al., 2002). The BCR/ABL fusion is the molecular analogue of the *Philadelphia chromosome*, one of the most frequent cytogenetic abnormalities in human leukemias. This *t(9;22) translocation* leads to a head-to-tail fusion of the `v-abl Abelson murine leukemia viral oncogene homolog 1` (`ABL1`) from chromosome 9 with the 5' half of the `breakpoint cluster region` (`BCR`) on chromosome 22 (Figure 4). The `ABL1` proto-oncogene encodes a cytoplasmic and nuclear protein tyrosine kinase that has been implicated in processes of cell differentiation, cell division, cell adhesion, and stress response. Although the BCR/ABL fusion protein, encoded by sequences from both the `ABL1` and `BCR` genes, has been extensively studied, the function of the normal product of the `BCR` gene is not clear. The `BCR/ABL` proto-oncogene has been found to be highly-expressed in chronic myeloid leukemia (CML) and acute myeloid leukemia (AML) cells (Mukhopadhyay et al., 2002).

An interesting question is therefore the identification of genes that are differentially expressed between B-cell ALL with the BCR/ABL fusion and cytogenetically normal NEG B-cell ALL.

In order to address this quesion, we consider the expression measures of the $n = 79$ B-cell ALL cell samples (`ALL$BT` equal to `B`, `B1`, `B2`, `B3`, or `B4`), of the BCR/ABL or NEG molecular types (`ALL$mol.biol` equal to `BCR/ABL` or `NEG`).

5.1.3 Gene filtering

Many of the genes represented by the 12,625 probes are not expressed in B-cell lymphocytes. Accordingly, as in von Heydebreck et al. (2004), we only retain the 2,391 probes that meet the following two criteria: (i) fluorescence intensities greater than 100 (absolute scale) for at least 25% of the 79 cell samples; (ii) interquartile range (IQR) of the fluorescence intensities for the 79 cell samples greater than 0.5 (log base 2 scale).

Furthermore, different probes may correspond to the same gene, i.e., map to the same Entrez Gene ID, according to the environment `hgu95av2LOCUSID` from the `hgu95av2` package. In order to obtain one expression measure per gene, we choose to average the expression measures of multiple probes mapping to the same gene.

These various pre-processing steps lead to $G = 2,071$ genes with unique Entrez Gene IDs.

5.1.4 Reduced ALL dataset: Genotypes and phenotypes of interest

The combined genotypic and phenotypic data for the $n = 79$ B-cell ALL cell samples of the BCR/ABL and NEG molecular types may be summarized by the set $\mathcal{XY}_n \equiv \{(X_i, Y_i) : i = 1, \ldots, n\}$, of n pairs of G–dimensional gene expression profiles $X_i = (X_i(g) : g = 1, \ldots, G)$, $G = 2,071$, and cancer class labels $Y_i \in \{NEG, BCR/ABL\}$. Among the $n = 79$ B-cell ALL cell samples, there are $n_{BCR/ABL} \equiv \sum_i \mathrm{I}(Y_i = BCR/ABL) = 37$ BCR/ABL and $n_{NEG} \equiv \sum_i \mathrm{I}(Y_i = NEG) = 42$ NEG samples.

5.2 Multiple hypothesis testing framework

Our primary question of interest is the identification of genes that are differentially expressed (DE) between BCR/ABL and NEG B-cell ALL. A subsequent question involves relating differential gene expression to GO annotation.

As detailed below, GO annotation metadata for the filtered list of $G = 2,071$ unique genes from the HG-U95Av2 chip may be summarized by binary gene-annotation profiles.

The gene-parameter profiles of interest concern differential gene expression between BCR/ABL and NEG B-cell ALL, i.e., the association between microarray gene expression measures and cancer class. Continuous gene-parameter profiles of unstandardized and standardized measures of differential expression are estimated, respectively, by (unstandardized) differences of empirical means and (standardized) two-sample t-statistics. Binary gene-parameter profiles, indicating whether genes are differentially expressed or not, are estimated by imposing cut-off rules on two-sample t-statistics or adjusted p-values.

The following association measures between GO gene-annotation profiles and DE gene-parameter profiles are considered: two-sample t-statistics for tests of association between binary GO gene-annotation profiles and continuous DE gene-parameter profiles; χ^2-statistics for tests of association between binary GO gene-annotation profiles and binary DE gene-parameter profiles.

Significant associations between differential gene expression and GO annotation are identified by applying FWER-controlling bootstrap-based single-step maxT Procedure 2.

5.2.1 Gene-annotation profiles

Gene Ontology annotation metadata for the HG-U95Av2 chip series are obtained as described in Sections 4.2 – 4.5, from the following Bioconductor R packages: the GO-specific metadata package GO (Version 1.10.0, Bioconductor Release 1.7) and the Affymetrix chip-specific metadata package hgu95av2 (Version 1.10.0, Bioconductor Release 1.7).

For each of the three gene ontologies, binary gene-annotation matrices A_{BP}, A_{CC}, and A_{MF}, are assembled for the GO terms annotating at least 10 of the $G = 2,071$ filtered genes (sample R code provided in Section 4.5). Specifically, for gene ontology $o \in \{BP, CC, MF\}$, $A_o = (A_o(g,m) : g = 1,\ldots,G;\ m = 1,\ldots,M_o)$ is a $G \times M_o$ matrix, with element $A_o(g,m)$ indicating whether gene g is annotated or not by GO term m and such that $\sum_g A_o(g,m) \geq 10$ for each term m. The numbers of terms considered in each gene ontology are $M_{BP} = 367$, $M_{CC} = 81$, and $M_{MF} = 185$.

5.2.2 Gene-parameter profiles

Definition of gene-parameter profiles Consider a data structure $(X, Y) \sim P$, where $X = (X(g) : g = 1,\ldots,G)$ is a $G = 2,071$–dimensional vector of microarray gene expression measures and $Y \in \{NEG, BCR/ABL\}$ is a cancer class label. Let $\pi_k \equiv Pr(Y = k)$ denote the proportion of cancers of class $k \in \{NEG, BCR/ABL\}$. Define conditional G–dimensional mean vectors and $G \times G$ covariance matrices for the expression measures of class $k \in \{NEG, BCR/ABL\}$ cancers by

$$\mu_k \equiv E[X|Y = k] \qquad \text{and} \qquad \sigma_k \equiv Cov[X|Y = k],$$

respectively.

Gene-parameter profiles, concerning differential gene expression between BCR/ABL and NEG B-cell ALL, may be specified in various ways. *Continuous* DE gene-parameter profiles may be defined in terms of the following *unstandardized* and *standardized* measures of differential gene expression between BCR/ABL and NEG B-cell ALL,

$$\begin{aligned} \lambda^d(g) &\equiv \mu_{BCR/ABL}(g) - \mu_{NEG}(g) \qquad (30) \\ \text{and} \quad \lambda^t(g) &\equiv \frac{\mu_{BCR/ABL}(g) - \mu_{NEG}(g)}{\sqrt{\frac{\sigma_{BCR/ABL}(g,g)}{\pi_{BCR/ABL}} + \frac{\sigma_{NEG}(g,g)}{\pi_{NEG}}}}. \end{aligned}$$

Absolute values of $\lambda^d(g)$ and $\lambda^t(g)$ may be used for measuring two-sided differential expression, i.e., either over- or under-expression in BCR/ABL compared to NEG B-cell ALL.

Binary DE gene-parameter profiles may be defined in terms of indicators for two-sided and one-sided differential expression.

$$\begin{aligned} \lambda^{\neq}(g) &\equiv \mathrm{I}(\mu_{BCR/ABL}(g) \neq \mu_{NEG}(g)) = \mathrm{I}(\lambda^d(g) \neq 0) = \mathrm{I}(\lambda^t(g) \neq 0), \\ \lambda^{+}(g) &\equiv \mathrm{I}(\mu_{BCR/ABL}(g) > \mu_{NEG}(g)) = \mathrm{I}(\lambda^d(g) > 0) = \mathrm{I}(\lambda^t(g) > 0), \\ \text{and} \quad \lambda^{-}(g) &\equiv \mathrm{I}(\mu_{BCR/ABL}(g) < \mu_{NEG}(g)) = \mathrm{I}(\lambda^d(g) < 0) = \mathrm{I}(\lambda^t(g) < 0). \end{aligned} \tag{31}$$

Estimation of gene-parameter profiles The above DE gene-parameter profiles may be estimated as follows, based on the dataset $\mathcal{XY}_n$ of gene expression measures for the $n = 79$ B-cell ALL cell samples of the BCR/ABL and NEG molecular types.

The resubstitution estimators of the continuous gene-parameter profiles of Equation (30) are based, respectively, on differences of empirical means and two-sample Welch t-statistics (up to the multiplier $1/\sqrt{n}$). That is,

$$\begin{aligned} \lambda_n^d(g) &\equiv \mu_{BCR/ABL,n}(g) - \mu_{NEG,n}(g) \\ \text{and} \quad \lambda_n^t(g) &\equiv \frac{1}{\sqrt{n}} \frac{\mu_{BCR/ABL,n}(g) - \mu_{NEG,n}(g)}{\sqrt{\frac{\sigma_{BCR/ABL,n}(g,g)}{n_{BCR/ABL}} + \frac{\sigma_{NEG,n}(g,g)}{n_{NEG}}}}, \end{aligned} \tag{32}$$

where $\mu_{k,n}(g) \equiv \sum_i \mathrm{I}(Y_i = k) X_i(g)/n_k$ and $\sigma_{k,n}(g,g) \equiv \sum_i \mathrm{I}(Y_i = k)(X_i(g) - \mu_{k,n}(g))^2/(n_k - 1)$ denote, respectively, the empirical means and variances of the gene expression measures for cancers of class $k \in \{NEG, BCR/ABL\}$.

Estimating the two-sided binary gene-parameter profile $\lambda^{\neq}$ of Equation (31) involves the two-sided tests of the G null hypotheses $H_0(g) = \mathrm{I}(\mu_{BCR/ABL}(g) = \mu_{NEG}(g))$, of no differences in mean gene expression measures between BCR/ABL and NEG B-cell ALL. Likewise, estimating the one-sided binary gene-parameter profiles λ^+ and λ^- involves, respectively, the one-sided tests of the G null hypotheses of no over-expression ($H_0(g) = \mathrm{I}(\mu_{BCR/ABL}(g) \leq \mu_{NEG}(g))$) and no under-expression ($H_0(g) = \mathrm{I}(\mu_{BCR/ABL}(g) \geq \mu_{NEG}(g))$) in BCR/ABL compared to NEG B-cell ALL. For single-step common-cut-off maxT Procedure 1, adjusted p-values produce the same gene rankings as the test statistics defined in Equation (32). Simple and naive estimators of the three sets of differentially expressed genes (i.e., true positives), represented by the gene-parameter profiles $\lambda^{\neq}$, λ^+, and λ^-, are therefore given, respectively, by the sets of genes with the largest γG values of $|\lambda_n^t(g)|$, $\lambda_n^t(g)$, and $-\lambda_n^t(g)$. That is,

$$\begin{aligned} \lambda_{n,\gamma G}^{\neq}(g) &\equiv \mathrm{I}\left(\sum_{g'=1}^{G} \mathrm{I}(|\lambda_n^t(g)| \geq |\lambda_n^t(g')|) \geq (1-\gamma)G \right), \\ \lambda_{n,\gamma G}^{+}(g) &\equiv \mathrm{I}\left(\sum_{g'=1}^{G} \mathrm{I}(\lambda_n^t(g) \geq \lambda_n^t(g')) \geq (1-\gamma)G \right), \\ \text{and} \quad \lambda_{n,\gamma G}^{-}(g) &\equiv \mathrm{I}\left(\sum_{g'=1}^{G} \mathrm{I}(-\lambda_n^t(g) \geq -\lambda_n^t(g')) \geq (1-\gamma)G \right). \end{aligned} \tag{33}$$

Analogous estimators may also be based on other test statistics, such as unstandardized difference statistics λ_n^d. More sophisticated estimators, that translate the proportion γ of rejected hypotheses into a Type I error rate such as the gFWER, TPPFP, or FDR, could be based on adjusted p-values for the multiple tests of the G null hypotheses $H_0(g)$. For example, one could estimate $\lambda^{\neq}$ by

$$\lambda_{n,\alpha}^{\neq}(g) \equiv \mathrm{I}\left(\widetilde{P}_{0n}^{\neq}(g) \leq \alpha\right), \tag{34}$$

where $\widetilde{P}_{0n}^{\neq}(g)$ are adjusted p-values for a suitably chosen multiple testing procedure, such as, FWER-controlling single-step maxT Procedure 1 or a TPPFP-controlling augmentation multiple testing procedure (Chapter 6, Dudoit and van der Laan (2006); van der Laan et al. (2004a)). One-sided binary gene-parameter profiles λ^+ and λ^- could be estimated likewise.

5.2.3 Association measures for gene-annotation and gene-parameter profiles

The association between continuous DE gene-parameter profiles, as in Equation (30), and binary GO gene-annotation profiles may be measured by two-sample Welch t-statistics (or corresponding p-values). Specifically, given a continuous G–vector x and a binary G–vector y, define the following association measure,

$$\rho^t(x,y) \equiv \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{\frac{v[x]_1}{y_1} + \frac{v[x]_0}{y_0}}}, \tag{35}$$

where $y_k \equiv \sum_g \mathrm{I}(y(g) = k)$, $\bar{x}_k \equiv \sum_g \mathrm{I}(y(g) = k)x(g)/y_k$, and $v[x]_k \equiv \sum_g \mathrm{I}(y(g) = k)(x(g) - \bar{x}_k)^2/(y_k - 1)$, $k \in \{0,1\}$.

The association between binary DE gene-parameter profiles, as in Equation (31), and binary GO gene-annotation profiles may be measured by χ^2-statistics (or corresponding p-values) for the test of independence of rows and columns in a 2×2 contingency table, such as Table 2. Specifically, given binary G–vectors x and y, define the following association measure,

$$\rho^{\chi}(x,y) \equiv \frac{G(g_{00}g_{11} - g_{01}g_{10})^2}{(g_{00} + g_{01})(g_{00} + g_{10})(g_{11} + g_{01})(g_{11} + g_{10})}, \tag{36}$$

where $g_{kk'} \equiv \sum_g \mathrm{I}(x(g) = k)\mathrm{I}(y(g) = k')$, $k, k' \in \{0,1\}$.

Given an association measure[3] $\rho : \mathbb{R}^{G \times M} \times \mathbb{R}^G \to \mathbb{R}^M$, a $G \times M$ GO gene-annotation matrix A, and a G–dimensional DE gene-parameter profile $\lambda = \Lambda(P)$, the M–dimensional association parameter vector $\psi = \Psi(P)$ of primary interest is defined as

$$\psi \equiv \rho(A, \lambda). \tag{37}$$

[3]N.B. For ease of notation, ρ^t and ρ^{χ}, defined in Equations (35) and (36) as real-valued association measures, may also refer loosely to $\mathbb{R}^M$–valued association measures, defined as $\rho^t(X,y) \equiv (\rho^t(X(\cdot,m),y) : m = 1,\ldots,M)$ and $\rho^{\chi}(X,y) \equiv (\rho^{\chi}(X(\cdot,m),y) : m = 1,\ldots,M)$ for $X \in \mathbb{R}^{G \times M}$ and $y \in \mathbb{R}^G$.

The corresponding resubstitution estimator $\psi_n = \hat{\Psi}(P_n)$ is simply obtained by replacing the gene-parameter profile λ by an estimator thereof $\lambda_n = \hat{\Lambda}(P_n)$, that is,

$$\psi_n \equiv \rho(A, \lambda_n). \tag{38}$$

5.2.4 Null and alternative hypotheses

For the t-statistic-based association measure ρ^t of Equation (35), the identification of GO terms m that are significantly (positively or negatively) associated with BCR/ABL vs. NEG differential gene expression involves the two-sided tests of the M null hypotheses $H_0(m) = \mathrm{I}(\psi(m) = \psi_0(m))$ against the alternative hypotheses $H_1(m) = \mathrm{I}(\psi(m) \neq \psi_0(m))$, with null values $\psi_0(m) = 0$. In some contexts, one may be interested in identifying positive (negative) associations, i.e., in the one-sided tests of the M null hypotheses $H_0(m) = \mathrm{I}(\psi(m) \leq \psi_0(m))$ $(H_0(m) = \mathrm{I}(\psi(m) \geq \psi_0(m)))$ against the alternative hypotheses $H_1(m) = \mathrm{I}(\psi(m) > \psi_0(m))$ $(H_1(m) = \mathrm{I}(\psi(m) < \psi_0(m)))$.

For the χ^2-statistic-based association measure ρ^χ of Equation (36), the identification of GO terms m that are significantly (positively or negatively) associated with BCR/ABL vs. NEG differential gene expression involves the one-sided tests of the M null hypotheses $H_0(m) = \mathrm{I}(\psi(m) \leq \psi_0(m))$ against the alternative hypotheses $H_1(m) = \mathrm{I}(\psi(m) > \psi_0(m))$. A natural choice for the null values is the mean of the $\chi^2(1)$-distribution, $\psi_0(m) = 1$.

5.2.5 Test statistics

One-sided and two-sided tests of null hypotheses concerning any of the association parameters defined above may be based on (unstandardized) difference statistics $T_n(m)$, defined as in Equation (20).

For one-sided tests, large values of the test statistics $T_n(m)$ provide evidence against the corresponding null hypotheses $H_0(m)$, that is, rejection regions are of the form $\mathcal{C}_n(m) = (c_n(m), +\infty)$. For two-sided tests, large values of the absolute test statistics $|T_n(m)|$ provide evidence against the corresponding null hypotheses $H_0(m)$.

5.2.6 Multiple testing procedures

For the purpose of illustration, we focus on control of the family-wise error rate, using single-step maxT Procedure 1, based on the non-parametric bootstrap null value shifted test statistics null distribution (null shift values $\lambda_0(m) = 0$ and no scaling). The main steps are outlined in Procedure 2.

Bootstrap-based single-step maxT adjusted p-values $\widetilde{P}_{0n}(m)$ are computed as in Equation (28). Let $O_n(m)$ denote indices for the ordered adjusted p-values, so that $\widetilde{P}_{0n}(O_n(1)) \leq \ldots \leq \widetilde{P}_{0n}(O_n(M))$. GO terms with adjusted p-values less than or equal to α are declared significantly associated with differential gene expression at nominal FWER level α. That is, the list of GO terms found to be

associated with differential gene expression is

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

where $R_n(\alpha) \equiv |\mathcal{R}_n(\alpha)|$ denotes the number of identified GO terms.

5.2.7 Summary of testing scenarios

This section summarizes our approach for identifying GO terms associated with BCR/ABL vs. NEG differential gene expression. For each of the three gene ontologies (i.e., BP, CC, and MF), we consider the following three types of testing scenarios, each corresponding to a different association parameter $\psi = \rho(A, \lambda)$ for GO annotation and BCR/ABL vs. NEG differential gene expression. Scenarios $\mathsf{MT}[t, t]$ and $\mathsf{MT}[d, t]$ are very similar and correspond, respectively, to *continuous* gene-parameter profiles of *standardized* and *unstandardized* measures of differential gene expression. In contrast, Scenario $\mathsf{MT}[\neq, \chi]$ corresponds to a *binary* gene-parameter profile of differential gene expression indicators.

1. **Scenario $\mathsf{MT}[t, t]$: Association parameter $\psi^{t,t} \equiv \rho^t(A, |\lambda^t|)$, for standardized continuous DE gene-parameter profile λ^t.** Consider the two-sided tests of

   $$H_0^{t,t}(m) \equiv \mathrm{I}(\psi^{t,t}(m) = \psi_0^{t,t}(m)) \quad \text{vs.} \quad H_1^{t,t}(m) \equiv \mathrm{I}(\psi^{t,t}(m) \neq \psi_0^{t,t}(m)),$$

   where the association parameter vector of interest is defined as $\psi^{t,t} \equiv \rho^t(A, |\lambda^t|)$, based on Equations (30) and (35), and the null values are $\psi_0^{t,t}(m) \equiv 0$. The continuous DE gene-parameter profile λ^t is estimated by λ_n^t, as in Equation (32), and the association parameter $\psi^{t,t}$ is estimated by the resubstitution estimator $\psi_n^{t,t} \equiv \rho^t(A, |\lambda_n^t|)$, as in Equation (38). The test statistics are defined as (unstandardized) difference statistics,

   $$T_n^{t,t}(m) \equiv \sqrt{n}(\psi_n^{t,t}(m) - \psi_0^{t,t}(m)),$$

   and the null hypotheses $H_0^{t,t}(m)$ are rejected for large absolute values of $T_n^{t,t}(m)$.

2. **Scenario $\mathsf{MT}[d, t]$: Association parameter $\psi^{d,t} \equiv \rho^t(A, |\lambda^d|)$, for unstandardized continuous DE gene-parameter profile λ^d.** Consider the two-sided tests of

   $$H_0^{d,t}(m) \equiv \mathrm{I}(\psi^{d,t}(m) = \psi_0^{d,t}(m)) \quad \text{vs.} \quad H_1^{d,t}(m) \equiv \mathrm{I}(\psi^{d,t}(m) \neq \psi_0^{d,t}(m)),$$

   where the association parameter vector of interest is defined as $\psi^{d,t} \equiv \rho^t(A, |\lambda^d|)$, based on Equations (30) and (35), and the null values are $\psi_0^{d,t}(m) \equiv 0$. The continuous DE gene-parameter profile λ^d is estimated by λ_n^d, as in Equation (32), and the association parameter $\psi^{d,t}$ is estimated by the resubstitution estimator $\psi_n^{d,t} \equiv \rho^t(A, |\lambda_n^d|)$, as in Equation (38). The test statistics are defined as (unstandardized) difference statistics,

   $$T_n^{d,t}(m) \equiv \sqrt{n}(\psi_n^{d,t}(m) - \psi_0^{d,t}(m)),$$

and the null hypotheses $H_0^{d,t}(m)$ are rejected for large absolute values of $T_n^{d,t}(m)$.

3. **Scenario MT[$\neq, \chi$]: Association parameter $\psi^{\neq,\chi} \equiv \rho^{\chi}(A, \lambda^{\neq})$, for binary DE gene-parameter profile $\lambda^{\neq}$.** Consider the one-sided tests of

$$H_0^{\neq,\chi}(m) \equiv \mathrm{I}(\psi^{\neq,\chi}(m) \leq \psi_0^{\neq,\chi}(m)) \quad \text{vs.} \quad H_1^{\neq,\chi}(m) \equiv \mathrm{I}(\psi^{\neq,\chi}(m) > \psi_0^{\neq,\chi}(m)),$$

where the association parameter vector of interest is defined as $\psi^{\neq,\chi} \equiv \rho^{\chi}(A, \lambda^{\neq})$, based on Equations (31) and (36), and the null values are $\psi_0^{\neq,\chi}(m) \equiv 1$ (the mean of the $\chi^2(1)$-distribution). The following two types of estimators $\lambda_n^{\neq}$ are considered for the binary DE gene-parameter profile $\lambda^{\neq}$: $\lambda_{n,\gamma G}^{\neq}$, with numbers of DE genes $\gamma G = 20$, 50, 100 (Equation (33)); $\lambda_{n,\alpha}^{\neq}$, defined in terms of adjusted p-values for FWER-controlling permutation-based single-step maxT Procedure 1 ($B = 1,000$ permutations of the cancer class labels) and nominal level $\alpha = 0.05$ (Equation (34)). Given an estimator $\lambda_n^{\neq}$ of $\lambda^{\neq}$, the association parameter $\psi^{\neq,\chi}$ is estimated by the resubstitution estimator $\psi_n^{\neq,\chi} \equiv \rho^{\chi}(A, \lambda_n^{\neq})$, as in Equation (38). The test statistics are defined as (unstandardized) difference statistics,

$$T_n^{\neq,\chi}(m) \equiv \sqrt{n}(\psi_n^{\neq,\chi}(m) - \psi_0^{\neq,\chi}(m)),$$

and the null hypotheses $H_0^{\neq,\chi}(m)$ are rejected for large values of $T_n^{\neq,\chi}(m)$.

For each of the three testing scenarios, the test statistics null value shifted null distribution Q_0 is estimated as in Procedure 2, with $B = 5,000$ non-parametric bootstrap samples of the data $\mathcal{XY}_n$ and $Z_n^B(m,b) = T_n^B(m,b) - E[T_n^B(m,\cdot)]$ (i.e., null shift values $\lambda_0(m) = 0$ and no scaling). Bootstrap-based single-step maxT adjusted p-values $\widetilde{P}_{0n}(m)$ are computed as in Equation (28) for one-sided testing Scenario MT[$\neq, \chi$]. For two-sided testing Scenarios MT[t, t] and MT[d, t], adjusted p-values are computed based on absolute values of $Z_n^B(m,b)$ and $T_n(m)$.

In what follows, the G–dimensional gene-parameter profiles λ correspond to the $G = 2,071$ genes with unique Entrez Gene IDs, obtained as described in Section 5.1. For each of the three gene ontologies, binary gene-annotation matrices are assembled for the GO terms annotating at least 10 of the $G = 2,071$ genes of interest: $G = 2,071 \times M_{BP} = 367$ gene-annotation matrix A_{BP} for the BP ontology, $G = 2,071 \times M_{CC} = 81$ gene-annotation matrix A_{CC} for the CC ontology, and $G = 2,071 \times M_{MF} = 185$ gene-annotation matrix A_{MF} for the MF ontology.

5.3 Results

5.3.1 Differentially expressed genes between BCR/ABL and NEG B-cell ALL

In order to identify differentially expressed genes between BCR/ABL and NEG B-cell ALL, two-sided tests of the G null hypotheses $H_0(g) = \mathrm{I}(\mu_{BCR/ABL}(g) =$

$\mu_{NEG}(g)$) are performed using the two-sample t-statistics $\lambda_n^t(g)$ of Equation (32) and FWER-controlling bootstrap-based single-step maxT Procedure 2. Adjusted p-values $\widetilde{P}_{0n}^{\neq}(g)$ are obtained using the `MTP` function from the `multtest` package (Version 1.8.0, Bioconductor Release 1.7), with $B = 5,000$ non-parametric bootstrap samples and other arguments set to their default values.

Figure 5 displays a normal quantile-quantile plot of the test statistics $\lambda_n^t(g)$ (Panel (a)) and a plot of the sorted bootstrap-based single-step maxT adjusted p-values $\widetilde{P}_{0n}^{\neq}(g)$ (Panel (b)). A handful of genes stand out in terms of their large absolute test statistics and small adjusted p-values.

For control of the FWER at nominal level $\alpha = 0.05$, Procedure 2 identifies 16 differentially expressed genes, i.e., 16 genes with $\widetilde{P}_{0n}^{\neq}(g) \leq \alpha$. Table 3 provides the test statistics, adjusted p-values, and various identifiers for these 16 genes. A more detailed hyperlinked table is posted on the website companion (Supplementary Table 1; `www.stat.berkeley.edu/~sandrine/Docs/Papers/DFF06/DFF.html`).

Only two of the 16 identified genes have a negative test statistic (`MX1` and `TPD52L2`), suggesting that most DE genes tend to be *over-expressed* in cell samples with the BCR/ABL fusion. The gene showing the most over-expression in BCR/ABL cell samples, as measured by the t-statistics λ_n^t, is the `ABL1` gene (`v-abl Abelson murine leukemia viral oncogene homolog 1`), located on the long arm of chromosome 9 (9q34.1). As mentioned in Section 5.1, the BCR/ABL phenotype is indeed defined in terms of the `ABL1` gene.

Furthermore, many of the DE genes seem to be related to apoptosis or oncogenesis. For example, the `Kruppel-like factor 9` (`KLF9`) gene encodes a transcription factor that binds GC-box elements in gene promoter regions. The Krüppel-like factor (KLF) family is comprised of highly-related zinc-finger proteins, that are important components of the eukaryotic cellular transcriptional machinery and that take part in a wide range of cellular functions (e.g., cell proliferation, apoptosis, differentiation, and neoplastic transformation). In particular, KLFs have been linked to various cancers (Kaczynski et al., 2003). The intron-less gene `AHNAK nucleoprotein (desmoyokin)` (`AHNAK`), located on the long arm of chromosome 11 (11q12.2), encodes an unusually large protein ($\approx$ 700kDa) that is typically repressed in cell lines derived from human neuroblastomas and several other types of tumors (Shtivelman et al., 1992). Yet another example, the `caspase 8, apoptosis-related cysteine peptidase` (`CASP8`) gene, encodes a key enzyme at the top of the apoptotic cascade and has been linked to neuroblastoma (Banelli et al., 2002). Likewise, other genes listed in Table 3, including `MX1`, `FYN`, `ACTN1`, `FHL1`, and `TRAM2`, appear to be related to the molecular biology of cancer. For further detail, the interested reader is invited to consult Supplementary Table 1 and follow links to PubMed and other databases.

Our results are in general agreement with those of von Heydebreck et al. (2004), slight differences being due, most likely, to our preliminary gene filtering, which involves averaging the expression measures of multiple probes mapping to the same Entrez Gene ID.

5.3.2 GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL

Figure 6 displays, for each of the three gene ontologies and each of the three testing scenarios, plots of the sorted adjusted p-values, $\widetilde{P}_{0n}(O_n(1)) \leq \ldots \leq \widetilde{P}_{0n}(O_n(M))$, for FWER-controlling bootstrap-based single-step maxT Procedure 2 ($B = 5,000$ bootstrap samples). The smaller the adjusted p-values, the less conservative the procedure and the longer the list $\mathcal{R}_n(\alpha) = \{m : \widetilde{P}_{0n}(m) \leq \alpha\}$ of identified GO terms at any given nominal Type I error level α. Table 4 summarizes the results in terms of the numbers $R_n(\alpha) = |\mathcal{R}_n(\alpha)|$ of GO terms found to be significantly associated with BCR/ABL vs. NEG differential gene expression at different nominal FWER levels α.

In general, 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 differential gene expression at nominal FWER level $\alpha \in \{0.05, 0.10, 0.20\}$. The adjusted p-values for Scenarios $\mathsf{MT}[t,t]$ and $\mathsf{MT}[d,t]$ (red and blue plotting symbols), corresponding, respectively, to standardized and unstandardized continuous DE gene-parameter profiles, are similar: for the BP and MF gene ontologies, Scenario $\mathsf{MT}[t,t]$ seems to be slightly more conservative than Scenario $\mathsf{MT}[d,t]$, however, this does not hold for the CC ontology. Scenario $\mathsf{MT}[\neq,\chi]$, with four different estimators of the binary DE gene-parameter profile $\lambda^{\neq}$, tends to be more conservative than either Scenario $\mathsf{MT}[t,t]$ or $\mathsf{MT}[d,t]$. Furthermore, the choice of parameter γG, for the number of genes called differentially expressed, can have a substantial impact on the adjusted p-values for Scenario $\mathsf{MT}[\neq,\chi : \gamma G]$. There are some indications, especially for the CC ontology, that larger values of the parameter γG lead to larger numbers of identified GO terms. Note that for Scenario $\mathsf{MT}[\neq,\chi]$, the p-value-based estimator $\lambda^{\neq}_{n,\alpha}$, with $\alpha = 0.05$, and the naive estimator $\lambda^{\neq}_{n,\gamma G}$, with $\gamma G = 20$, yield very similar results (green and purple plotting symbols). Indeed, when applied to the entire dataset of $n = 79$ cell samples, permutation-based single-step maxT Procedure 1 identifies 20 genes as being differentially expressed between BCR/ABL and NEG B-cell ALL at nominal FWER level $\alpha = 0.05$. In other words, $\lambda^{\neq}_{n,0.05}$ and $\lambda^{\neq}_{n,20}$ yield the same estimate of the binary gene-parameter profile $\lambda^{\neq}$ for the set of DE genes. Minor discrepancies between the results of Scenarios $\mathsf{MT}[\neq,\chi : \alpha = 0.05]$ and $\mathsf{MT}[\neq,\chi : \gamma G = 20]$ are due to the fact that while the estimators $\lambda^{\neq}_{n,0.05}$ and $\lambda^{\neq}_{n,20}$ coincide on the full dataset, they may differ on bootstrap samples of these data.

Next, the three testing scenarios are compared in terms of the contents of the lists $\mathcal{R}_n(\alpha)$ of identified GO terms. Specifically, let $\mathcal{O}_n(r) \equiv \{O_n(1), \ldots, O_n(r)\}$ denote the set of indices corresponding to the r smallest adjusted p-values for a given gene ontology and testing scenario. Measures of agreement between testing scenarios are provided by the numbers of common GO terms among sets of ordered GO terms $\mathcal{O}_n(r)$ of various cardinality r, i.e., by the cardinality of the intersections between sets $\mathcal{O}_n(r)$ for different testing scenarios. Figure 7 displays plots of numbers of common GO terms for pairs of testing scenar-

ios. As expected, there is substantial overlap between the GO terms identified by Scenarios $\mathsf{MT}[t,t]$ and $\mathsf{MT}[d,t]$ for continuous DE gene-parameter profiles (blue plotting symbols in top panels). This suggests that, for the ALL dataset, standardized (λ^t) and unstandardized (λ^d) continuous measures of differential gene expression have similar properties. In contrast, there is much less overlap between the GO terms identified by Scenario $\mathsf{MT}[\neq,\chi]$, for binary DE gene-parameter profiles, and either Scenario $\mathsf{MT}[t,t]$ or $\mathsf{MT}[d,t]$. For example, for the MF gene ontology, among the top 10 GO terms $\mathcal{O}_n(10)$ identified by each testing scenario, 6 are common to Scenarios $\mathsf{MT}[t,t]$ and $\mathsf{MT}[d,t]$, whereas at most 3 are common to Scenarios $\mathsf{MT}[t,t]$ and $\mathsf{MT}[\neq,\chi]$. Again, note the near perfect agreement between Scenarios $\mathsf{MT}[\neq,\chi:\alpha=0.05]$ and $\mathsf{MT}[\neq,\chi:\gamma G=20]$ (purple plotting symbols in lower panels). Figure 7 again illustrates the lack of robustness of Scenario $\mathsf{MT}[\neq,\chi:\gamma G]$ to the choice of parameter γG.

Moreover, examine graphical summaries of the joint distributions of the estimated continuous DE gene-parameter profile λ_n^t and the gene-annotation profiles $A(\cdot,m)$ for the top two GO terms $m\in\{O_n(1),O_n(2)\}$ identified according to each testing scenario. Figure 8 displays conditional boxplots of λ_n^t given $A(\cdot,m)$, that is, boxplots of the unannotated and annotated estimated gene-parameter profiles, $(\lambda_n^t(g):A(g,m)=0)$ and $(\lambda_n^t(g):A(g,m)=1)$, respectively. While the boxplots reveal clear differences (non-overlapping notches) between unannotated and annotated profiles for some of the GO terms (e.g., MF term `GO:0003735`), the differences can be subtle for other terms (e.g., MF term `GO:0003924`). Not surprisingly, the most extreme differences are seen with Scenarios $\mathsf{MT}[t,t]$ and $\mathsf{MT}[d,t]$, and, to a lesser extent, with Scenario $\mathsf{MT}[\neq,\chi:\alpha=0.05]$ for the CC ontology. The boxplots again illustrate differences between Scenario $\mathsf{MT}[\neq,\chi]$ and either Scenario $\mathsf{MT}[t,t]$ or $\mathsf{MT}[d,t]$.

Tables 5, 6, and 7 report various p-value-based measures of association between the estimated DE gene-parameter profiles λ_n^t and $\lambda_{n,\alpha}^{\neq}$ and the gene-annotation profiles $A(\cdot,m)$ for the top two GO terms $m\in\{O_n(1),O_n(2)\}$ identified according to each testing scenario, in the BP, CC, and MF gene ontologies, respectively. The transformation to the $[0,1]$ p-value scale allows a more direct comparison of the various testing scenarios. The tables again highlight the differences between Scenario $\mathsf{MT}[\neq,\chi]$, for binary DE gene-parameter profiles, and either Scenario $\mathsf{MT}[t,t]$ or $\mathsf{MT}[d,t]$, for continuous DE gene-parameter profiles. As expected, Scenarios $\mathsf{MT}[t,t]$ and $\mathsf{MT}[d,t]$ tend to identify GO terms with small p-values $P_{0n}^{t,t}(m)$ for t-tests of association between estimated continuous gene-parameter profiles λ_n^t and gene-annotation profiles $A(\cdot,m)$. In contrast, and also as expected, Scenario $\mathsf{MT}[\neq,\chi]$ tends to identify GO terms with small p-values $P_{0n}^{\neq,\chi}(m)$ for χ^2-tests of association between estimated binary gene-parameter profiles $\lambda_{n,\alpha}^{\neq}$ and gene-annotation profiles $A(\cdot,m)$. Furthermore, the tables corroborate our earlier observation that Scenario $\mathsf{MT}[\neq,\chi]$ tends to be more conservative than either Scenario $\mathsf{MT}[t,t]$ or $\mathsf{MT}[d,t]$. Indeed, some of the GO terms with small p-values $P_{0n}^{t,t}(m)$ for continuous gene-parameter profiles have very large p-values $P_{0n}^{\neq,\chi}(m)$ for binary gene-parameter profiles (e.g., MF term `GO:0003735` in Table 7). Such terms are likely to be identified by Sce-

narios $\mathsf{MT}[t,t]$ and $\mathsf{MT}[d,t]$, but "missed" by Scenario $\mathsf{MT}[\neq,\chi]$. The converse phenomenon is not as striking. However, one should keep in mind that Scenario $\mathsf{MT}[\neq,\chi]$ depends on the choice of estimator for the binary DE gene-parameter profile $\lambda^{\neq}$, i.e., on parameters such as α and γG. In particular, with certain values of α (or γG), binary Scenario $\mathsf{MT}[\neq,\chi]$ may become more similar to either continuous Scenario $\mathsf{MT}[t,t]$ or $\mathsf{MT}[d,t]$. Column $A_1(m)$ in Tables 5, 6, and 7 suggests that, compared to Scenario $\mathsf{MT}[\neq,\chi]$, Scenarios $\mathsf{MT}[t,t]$ and $\mathsf{MT}[d,t]$ tend to identify GO terms annotating a larger number of genes (this observation also holds for the top 20 terms identified according to each testing scenario; data not shown).

Figure 9 displays a scatterplot matrix of the 50 smallest adjusted p-values, based on Scenario $\mathsf{MT}[t,t]$, for each of the three gene ontologies. The plots indicate that more terms tend to be identified in the BP ontology compared to either the CC or MF ontologies, and fewer terms tend to be identified in the MF ontology compared to either the BP or CC ontologies. Note that comparisons based on adjusted p-values take into account differences in the numbers of tested hypotheses, $M_{BP} = 367$, $M_{CC} = 81$, and $M_{MF} = 185$, for each ontology.

Tables 8, 9, and 10 list the 20 GO terms with the smallest adjusted p-values for Scenario $\mathsf{MT}[t,t]$, applied to the BP, CC, and MF gene ontologies, respectively. Figures 10, 11, and 12 display portions of the directed acyclic graphs for the top 20 GO terms in each ontology. The figures suggest that GO terms associated with BCR/ABL vs. NEG differential gene expression tend to concentrate in certain branches of the DAGs, i.e., differential expression is associated with related properties of gene products. While it is known that many of the effects of the BCR/ABL fusion are mediated by tyrosine kinase activity, the MF GO term *protein-tyrosine kinase activity* (`GO:0004713`) does not appear to be significantly associated with differential gene expression between BCR/ABL and NEG B-cell ALL (adjusted p-value of 0.8890 for Scenario $\mathsf{MT}[t,t]$).

For illustration purposes, we further investigate two of the GO terms from Tables 8 and 10: GO term *anti-apoptosis* (`GO:0006916`), with ninth smallest adjusted p-value for Scenario $\mathsf{MT}[t,t]$ applied to the BP gene ontology, and GO term *structural constituent of ribosome* (`GO:0003735`), with the smallest adjusted p-value for Scenario $\mathsf{MT}[t,t]$ applied to the MF gene ontology. Tables 11 and 12 list genes directly or indirectly annotated with GO terms `GO:0006916` and `GO:0003735`, respectively. Figure 13 displays mean-difference plots of the average expression measures in BCR/ABL and NEG cell samples for genes annotated with GO terms `GO:0006916` and `GO:0003735`.

Panel (a) in Figure 13 indicates that genes annotated with BP GO term *anti-apoptosis* (`GO:0006916`) tend to be over-expressed in BCR/ABL compared to NEG cell samples. Among these 21 genes, only `SOCS2` is significantly differentially expressed between BCR/ABL and NEG B-cell ALL (nominal FWER level $\alpha = 0.05$, Table 3). However, a brief survey of the literature reveals that several of the genes in Table 11 interact with the `BCR/ABL` proto-oncogene. For instance, Kirchner et al. (2003) investigate mechanisms for the `BCR/ABL`-mediated activation of the transcription factor NF-κB/Rel encoded by the `NFKB1` gene. Their findings suggest that NF-κB/Rel may be a potential target for molecular

therapies of leukemia. Mukhopadhyay et al. (2002) demonstrate that ectopic expression of `BCR/ABL` interferes with the `tumor necrosis factor` (`TNF`) signaling pathway through the down-regulation of TNF receptors. The `TNF` gene encodes a multifunctional proinflammatory cytokine involved in the regulation of a wide spectrum of biological processes, including cell proliferation, differentiation, apoptosis, lipid metabolism, and coagulation. The `TNF` gene has been implicated in a variety of diseases, including autoimmune diseases, insulin resistance, and cancer.

As seen in Table 12, 22 of the 24 genes annotated with MF GO term *structural constituent of ribosome* (`GO:0003735`) code for ribosomal proteins. Although none of the 24 annotated genes are identified as being significantly differentially expressed between BCR/ABL and NEG B-cell ALL (nominal FWER level $\alpha = 0.05$, Table 3), Panel (b) in Figure 13 suggests that these genes tend to be under-expressed in BCR/ABL cell samples.

6 Discussion

We have proposed a general and formal statistical framework for multiple tests of association with biological annotation metadata. A key component of our approach is the systematic and precise translation of a generic biological question into a corresponding multiple hypothesis testing problem, concerning association measures between known gene-annotation profiles and unknown gene-parameter profiles. This general and rigorous formulation of the statistical inference question allows us to apply the multiple testing methodology developed in Dudoit and van der Laan (2006) and related articles, to control 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.

The flexibility of our approach was illustrated using the ALL microarray dataset of Chiaretti et al. (2004), with the aim of relating GO annotation to differential gene expression between BCR/ABL and NEG B-cell ALL. This analysis demonstrates the importance of selecting a suitable DE gene-parameter profile λ and measure ρ for the association between this gene-parameter profile and GO gene-annotation profiles A. Indeed, for the ALL dataset, the choice of gene-parameter profile for measuring differential expression between BCR/ABL and NEG B-cell ALL had a large impact on the list of identified GO terms. Testing scenarios based on binary DE gene-parameter profiles (Scenario $\mathsf{MT}[\neq, \chi]$) tended to be more conservative than scenarios based on continuous DE gene-parameter profiles (Scenarios $\mathsf{MT}[t, t]$ and $\mathsf{MT}[d, t]$), with little overlap between the lists of identified GO terms. Furthermore, testing scenarios based on binary gene-parameter profiles were sensitive to the somewhat arbitrary DE/non-DE gene dichotomization, that is, Scenario $\mathsf{MT}[\neq, \chi : \gamma G]$ lacked robustness with respect to the choice of parameter γG for the number of genes called differentially expressed according to the estimator $\lambda^{\neq}_{n,\gamma G}$. In contrast, continuous gene-parameter profiles based on standardized and unstandardized measures of

differential gene expression lead to very similar results (Scenarios $\mathsf{MT}[t,t]$ and $\mathsf{MT}[d,t]$).

Our results on the ALL microarray dataset clearly show the limitations of binary gene-parameter profiles of differential expression indicators, which are still the norm for combined GO annotation and microarray data analyses. Our proposed statistical framework, with general definitions for the gene-annotation and gene-parameter profiles, allows consideration of a much broader class of inference problems, that extend beyond GO annotation and microarray data analysis. Gene-annotation profiles may be continuous or polychotomous and may correspond, for example, to intron/exon counts/lengths/nucleotide distributions, gene pathway membership, or gene regulation by particular transcription factors. Likewise, gene-parameter profiles may be continuous or polychotomous and may correspond, for example, to regression coefficients relating possibly censored biological and clinical outcomes to genome-wide transcript levels, DNA copy numbers, and other covariates.

This first application of our proposed methodology only considered control of the family-wise error rate using single-step common-cut-off maxT Procedure 1, based on the non-parametric bootstrap null value shifted test statistics null distribution. Adjusted p-values tended to be quite large, with only a handful of GO terms identified as being significantly associated with BCR/ABL vs. NEG differential gene expression. Joint stepwise augmentation and empirical Bayes procedures could be used for control of a broader and more biologically relevant class of Type I error rates, defined as tail probabilities, $gTP(q,g) = Pr(g(V_n, R_n) > q)$, for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n (Dudoit and van der Laan, 2006; Dudoit et al., 2004a; van der Laan et al., 2004a,b, 2005). Error rates based on the proportion V_n/R_n of false positives (e.g., TPPFP and FDR) are especially appealing for large-scale testing problems, compared to error rates based on the number V_n of false positives (e.g., gFWER), as they do not increase exponentially with the number M of tested hypotheses. More powerful analyses may also be achieved with the new null quantile-transformed test statistics null distribution of van der Laan and Hubbard (2005). The multiple testing methodology developed in Dudoit and van der Laan (2006) and related articles is particularly well-suited to handle the variety of parameters of interest and the complex and unknown dependence structures among test statistics (e.g., implied by the DAG structure of GO terms) that are likely to be encountered in these and other high-dimensional inference problems in biomedical and genomic research.

Ongoing efforts include consideration of more general and biologically pertinent multivariate association measures ρ. For instance, for GO annotation metadata, the association parameter for a given GO term could take into account the structure of the DAG by considering the gene-annotation profiles of offspring or ancestor terms. We are also interested in developing better numerical and graphical methods for representing and interpreting the multiple testing results, e.g., the lists of GO terms and associated adjusted p-values. Finally, we are planning on implementing the proposed methods in an R package to be

released as part of the Bioconductor Project.

Software and website companion

The multiple testing procedures proposed in Dudoit and van der Laan (2006) and related articles (Birkner et al., 2005; Dudoit et al., 2004a,b; Keleş et al., 2004; van der Laan et al., 2004a,b, 2005; van der Laan and Hubbard, 2005; Pollard et al., 2005a,b; Pollard and van der Laan, 2004; Rubin et al., 2005) 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 (Pollard et al. (2005b); `www.bioconductor.org`).

The experimental data (ALL) and annotation metadata (annaffy, annotate, GO, hgu95av2) packages used in the analysis of Section 5 may also be obtained from the Bioconductor Project website.

The website companion to this article provides additional tables, figures, code, and references: `www.stat.berkeley.edu/~sandrine/Docs/Papers/DFF06/DFF.html`.

Acknowledgment

We are most grateful to Robert Gentleman for many stimulating discussions on statistical and computational methods for the analysis of biological annotation metadata.

Table 1: *Type I and Type II errors in multiple hypothesis testing.* This table summarizes the different types of decisions and errors in multiple hypothesis testing. The number of rejected hypotheses is $R_n \equiv |\mathcal{R}_n| = \sum_{m=1}^{M} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m))$, the number of Type I errors is $V_n \equiv |\mathcal{R}_n \cap \mathcal{H}_0| = \sum_{m \in \mathcal{H}_0} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m))$, and the number of Type II errors is $U_n \equiv |\mathcal{R}_n^c \cap \mathcal{H}_1| = \sum_{m \in \mathcal{H}_1} \mathrm{I}(T_n(m) \notin \mathcal{C}_n(m))$.

		Null hypotheses		
		non-rejected	**rejected**	
Null hypotheses	**true**	$\lvert\mathcal{R}_n^c \cap \mathcal{H}_0\rvert$	$V_n = \lvert\mathcal{R}_n \cap \mathcal{H}_0\rvert$ (Type I)	$h_0 = \lvert\mathcal{H}_0\rvert$
	false	$U_n = \lvert\mathcal{R}_n^c \cap \mathcal{H}_1\rvert$ (Type II)	$\lvert\mathcal{R}_n \cap \mathcal{H}_1\rvert$	$h_1 = \lvert\mathcal{H}_1\rvert$
		$M - R_n$	$R_n = \lvert\mathcal{R}_n\rvert$	M

Table 2: *Association parameters for binary gene-annotation profiles and gene-parameter profiles.* Given a binary gene-annotation profile $A(\cdot, m)$ and a binary gene-parameter profile λ, one may build a 2×2 contingency table, with rows corresponding to the gene-annotation profile and columns to the gene-parameter profile. Cell counts are defined as $g_{kk'}(m) \equiv \sum_g \mathrm{I}(A(g,m) = k)\mathrm{I}(\lambda(g) = k')$, $k, k' \in \{0, 1\}$. For instance, $g_{11}(m)$ corresponds to the number of genes scored as one for both the gene-annotation profile and the gene-parameter profile, i.e., the number of genes possessing both features of interest.

		Gene-parameter profile, λ		
		1	**0**	
Gene-annotation profile, $A(\cdot, m)$	**1**	$g_{11}(m) = \sum_{g=1}^{G} A(g,m)\lambda(g)$	$g_{10}(m) = \sum_{g=1}^{G} A(g,m)(1-\lambda(g))$	$A_1(m) = \sum_{g=1}^{G} A(g,m)$
	0	$g_{01}(m) = \sum_{g=1}^{G}(1-A(g,m))\lambda(g)$	$g_{00}(m) = \sum_{g=1}^{G}(1-A(g,m))(1-\lambda(g))$	$A_0(m) = \sum_{g=1}^{G}(1-A(g,m))$
		$G\bar{\lambda} = \sum_{g=1}^{G} \lambda(g)$	$G(1-\bar{\lambda}) = \sum_{g=1}^{G}(1-\lambda(g))$	G

Table 3: *Differentially expressed genes between BCR/ABL and NEG B-cell ALL.* This table provides the Affymetrix probe IDs, Entrez Gene IDs (`hgu95av2LOCUSID` environment in `hgu95av2` package), gene symbols (`hgu95av2SYMBOL` environment), gene names (`hgu95av2GENENAME` environment), test statistics $\lambda_n^t(g)$ (Equation (32)), and adjusted p-values $\widetilde{P}_{0n}^{\neq}(g)$ (Equation (28)) for the 16 genes found to be significantly differentially expressed between BCR/ABL and NEG B-cell ALL, at nominal FWER level $\alpha = 0.05$, according to bootstrap-based single-step maxT Procedure 2, with two-sample t-statistics $\lambda_n^t(g)$ and $B = 5,000$ bootstrap samples. A more detailed hyperlinked table, including information on gene function, chromosomal location, links to GenBank, Entrez Gene, NCBI Map Viewer, UniGene, PubMed, AmiGO, and KEGG, is provided on the website companion (Supplementary Table 1).

Probe ID	**Entrez Gene ID**	**Symbol**	$\lambda_n^t(g)$	$\widetilde{P}_{0n}^{\neq}(g)$
1635_at	25	ABL1	8.44	0
v-abl Abelson murine leukemia viral oncogene homolog 1				
40202_at	687	KLF9	6.33	0
Kruppel-like factor 9				
37027_at	79026	AHNAK	5.71	0.0014
AHNAK nucleoprotein (desmoyokin)				
39837_s_at	168544	ZNF467	5.45	0.0034
zinc finger protein 467				
33774_at	841	CASP8	5.29	0.0042
caspase 8, apoptosis-related cysteine peptidase				
37014_at	4599	MX1	-5.23	0.0050
myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)				
2039_s_at	2534	FYN	5.21	0.0050
FYN oncogene related to SRC, FGR, YES				
39329_at	87	ACTN1	4.97	0.0096
actinin, alpha 1				
32542_at	2273	FHL1	4.96	0.0102
four and a half LIM domains 1				
40051_at	9697	TRAM2	4.59	0.0268
translocation associated membrane protein 2				
38032_at	9900	SV2A	4.54	0.0308
synaptic vesicle glycoprotein 2A				
39319_at	3937	LCP2	4.50	0.0346
lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa)				
33232_at	1396	CRIP1	4.46	0.0368
cysteine-rich protein 1 (intestinal)				

Continued on next page ...

... *continued from previous page*

Probe ID	**Entrez Gene ID**	**Symbol**	$\lambda_n^t(g)$	$\widetilde{P}_{0n}^{\neq}(g)$
36591_at	7277	TUBA1	4.37	0.0444
tubulin, alpha 1 (testis specific)				
38994_at	8835	SOCS2	4.35	0.0466
suppressor of cytokine signaling 2				
40076_at	7165	TPD52L2	-4.33	0.0480
tumor protein D52-like 2				

Table 4: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL.* This table reports, for each of the three gene ontologies and each of the three testing scenarios, the numbers $R_n(\alpha) = |\mathcal{R}_n(\alpha)|$ of GO terms found to be significantly associated with BCR/ABL vs. NEG differential gene expression at different nominal FWER levels α.

	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**		

Table 5: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, top two BP GO terms.* This table provides association measures between the estimated DE gene-parameter profiles λ_n^t and $\lambda_{n,\alpha}^{\neq}$ and the gene-annotation profiles $A(\cdot, m)$ for the top two BP GO terms $m \in \{O_n(1), O_n(2)\}$ identified according to each of the three testing scenarios. $A_1(m) \equiv \sum_g A(g, m)$: Number of genes directly or indirectly annotated with GO term m (out of $G = 2,071$ genes, `GOALLLOCUSID` environment in `GO` package). $P_{0n}^{t,t}(m)$: Nominal unadjusted p-value for the two-sample t-test comparing the unannotated and annotated estimated continuous DE gene-parameter profiles, $(\lambda_n^t(g) : A(g, m) = 0)$ and $(\lambda_n^t(g) : A(g, m) = 1)$, respectively (`t.test` function from the R `stats` package, with default argument values). $P_{0n}^{\neq,\chi}(m)$: Unadjusted p-value for the χ^2-test of independence between the estimated binary DE gene-parameter profile $\lambda_{n,\alpha}^{\neq}$, $\alpha = 0.05$, and the gene-annotation profile $A(\cdot, m)$ (`chisq.test` function from the R `stats` package, with arguments `simulate.p.value = TRUE`, `correct=FALSE`). $\widetilde{P}_{0n}(m)$: Bootstrap-based single-step maxT adjusted p-value, according to which the top two GO terms are identified for each testing scenario.

BP

Scenario	GO term	$A_1(m)$	$P_{0n}^{t,t}(m)$	$P_{0n}^{\neq,\chi}(m)$	$\widetilde{P}_{0n}(m)$
MT$[t, t]$	GO:0008152	1076	0	0.1704	0.0262
	GO:0044237	1045	0	0.1824	0.0428
MT$[d, t]$	GO:0006091	98	0	0.6172	0.0366
	GO:0000226	14	0.0018	1	0.0582
MT$[\neq, \chi : \alpha = 0.05]$	GO:0008361	27	0.0553	0.0035	0.0828
	GO:0016049	27	0.0553	0.0010	0.0828
MT$[\neq, \chi : \gamma G = 20]$	GO:0008361	27	0.0553	0.0020	0.2078
	GO:0016049	27	0.0553	0.0020	0.2078
MT$[\neq, \chi : \gamma G = 50]$	GO:0048522	87	0.0356	0.0120	0.1860
	GO:0048518	96	0.0439	0.0145	0.2338
MT$[\neq, \chi : \gamma G = 100]$	GO:0050793	24	0.0854	0.0175	0.1458
	GO:0007155	59	0.0006	0.1109	0.1980

Table 6: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, top two CC GO terms.* Details in Table 5 caption.

CC

Scenario	**GO term**	$A_1(m)$	$P_{0n}^{t,t}(m)$	$P_{0n}^{\neq,\chi}(m)$	$\widetilde{P}_{0n}(m)$
MT$[t,t]$	GO:0005840	25	0	1	0.0056
	GO:0030529	77	0	0.6387	0.0138
MT$[d,t]$	GO:0005840	25	0	1	0.0040
	GO:0005830	11	0	1	0.0052
MT$[\neq,\chi:\alpha=0.05]$	GO:0005578	10	0.0167	0.0775	0.4940
	GO:0031012	10	0.0167	0.0815	0.4940
MT$[\neq,\chi:\gamma G=20]$	GO:0005578	10	0.0167	0.1069	0.3500
	GO:0031012	10	0.0167	0.0975	0.3500
MT$[\neq,\chi:\gamma G=50]$	GO:0005576	54	0.0009	1	0.0078
	GO:0005615	31	0.0480	0.2509	0.0078
MT$[\neq,\chi:\gamma G=100]$	GO:0005576	54	0.0009	1	0.0488
	GO:0005615	31	0.0480	0.2439	0.1280

Table 7: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, top two MF GO terms.* Details in Table 5 caption.

MF

Scenario	**GO term**	$A_1(m)$	$P_{0n}^{t,t}(m)$	$P_{0n}^{\neq,\chi}(m)$	$\widetilde{P}_{0n}(m)$
MT$[t,t]$	GO:0003735	24	0	1	0.0024
	GO:0003723	143	0	0.4068	0.1168
MT$[d,t]$	GO:0003735	24	0	1	0.0022
	GO:0003723	143	0	0.3968	0.0784
MT$[\neq,\chi:\alpha=0.05]$	GO:0004930	10	0.2241	0.0065	0.0366
	GO:0003924	34	0.6501	0.0395	0.7046
MT$[\neq,\chi:\gamma G=20]$	GO:0004930	10	0.2241	0.0025	0.0168
	GO:0003924	34	0.6501	0.0495	0.6210
MT$[\neq,\chi:\gamma G=50]$	GO:0004930	10	0.2241	0.0040	0.4108
	GO:0030246	22	0.8582	0.1919	0.4794
MT$[\neq,\chi:\gamma G=100]$	GO:0005509	69	0.0004	0.1399	0.3140
	GO:0004930	10	0.2241	0.0025	0.3262

Table 8: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, top 20 BP GO terms.* This table lists the 20 GO terms with the smallest adjusted p-values for Scenario $\mathsf{MT}[t,t]$ applied to the BP gene ontology. $A_1(m) \equiv \sum_g A(g,m)$: Number of genes directly or indirectly annotated with GO term m (out of $G = 2,071$ genes, `GOALLLOCUSID` environment in `GO` package). $\widetilde{P}_{0n}(m)$: Bootstrap-based single-step maxT adjusted p-value for Scenario $\mathsf{MT}[t,t]$.

BP, Scenario $\mathsf{MT}[t,t]$

GO term ID	**GO term**	$A_1(m)$	$\widetilde{P}_{0n}(m)$
GO:008152	*metabolism*	1076	0.0262
GO:044237	*cellular metabolism*	1045	0.0428
GO:009058	*biosynthesis*	187	0.0750
GO:044238	*primary metabolism*	1002	0.0750
GO:044249	*cellular biosynthesis*	169	0.0862
GO:006091	*generation of precursor metabolites and energy*	98	0.0928
GO:019882	*antigen presentation*	15	0.1098
GO:030333	*antigen processing*	14	0.1444
GO:006916	*anti-apoptosis*	21	0.1564
GO:043066	*negative regulation of apoptosis*	26	0.1692
GO:043069	*negative regulation of programmed cell death*	26	0.1692
GO:007154	*cell communication*	390	0.1754
GO:006457	*protein folding*	52	0.1910
GO:007165	*signal transduction*	351	0.1946
GO:000226	*microtubule cytoskeleton organization and biogenesis*	14	0.2302
GO:006082	*organic acid metabolism*	65	0.2538
GO:006163	*purine nucleotide metabolism*	29	0.2820
GO:007155	*cell adhesion*	59	0.2822
GO:007028	*cytoplasm organization and biogenesis*	10	0.2976
GO:019752	*carboxylic acid metabolism*	63	0.3108

Table 9: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, top 20 CC GO terms.* Details in Table 8 caption.

CC, Scenario MT[t, t]

GO term ID	**GO term**	$A_1(m)$	$\widetilde{P}_{0n}(m)$
GO:0005840	*ribosome*	25	0.0056
GO:0030529	*ribonucleoprotein complex*	77	0.0138
GO:0005830	*cytosolic ribosome (sensu Eukaryota)*	11	0.0144
GO:0043234	*protein complex*	334	0.0778
GO:0005886	*plasma membrane*	200	0.1316
GO:0005829	*cytosol*	78	0.2204
GO:0005737	*cytoplasm*	578	0.2304
GO:0005887	*integral to plasma membrane*	125	0.2338
GO:0031226	*intrinsic to plasma membrane*	125	0.2338
GO:0019866	*inner membrane*	37	0.2574
GO:0005743	*mitochondrial inner membrane*	28	0.2636
GO:0005746	*mitochondrial electron transport chain*	11	0.2692
GO:0000502	*proteasome complex (sensu Eukaryota)*	26	0.2714
GO:0000323	*lytic vacuole*	28	0.2866
GO:0005764	*lysosome*	28	0.2866
GO:0005576	*extracellular region*	54	0.3130
GO:0005773	*vacuole*	29	0.3172
GO:0005622	*intracellular*	1152	0.3350
GO:0043228	*non-membrane-bound organelle*	218	0.3524
GO:0043232	*intracellular non-membrane-bound organelle*	218	0.3524

Table 10: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, top 20 MF GO terms.* Details in Table 8 caption.

MF, Scenario MT[t, t]

GO term ID	**GO term**	$A_1(m)$	$\widetilde{P}_{0n}(m)$
GO:0003735	*structural constituent of ribosome*	24	0.0024
GO:0003723	*RNA binding*	143	0.1168
GO:0048037	*cofactor binding*	11	0.1518
GO:0051082	*unfolded protein binding*	47	0.2210
GO:0016853	*isomerase activity*	28	0.2348
GO:0016491	*oxidoreductase activity*	89	0.3476
GO:0005509	*calcium ion binding*	69	0.3496
GO:0015399	*primary active transporter activity*	57	0.4314
GO:0004872	*receptor activity*	101	0.4518
GO:0004871	*signal transducer activity*	242	0.4566
GO:0016765	*transferase activity, transferring alkyl or aryl (other than methyl) groups*	10	0.4570
GO:0016860	*intramolecular oxidoreductase activity*	13	0.4636
GO:0016614	*oxidoreductase activity, acting on CH-OH group of donors*	18	0.4734
GO:0016616	*oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor*	18	0.4734
GO:0043169	*cation binding*	230	0.5002
GO:0005489	*electron transporter activity*	47	0.5420
GO:0005386	*carrier activity*	73	0.5502
GO:0004888	*transmembrane receptor activity*	59	0.5690
GO:0003824	*catalytic activity*	635	0.5826
GO:0003676	*nucleic acid binding*	449	0.6718

Table 11: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, BP GO term GO:0006916.* This table lists genes directly or indirectly annotated with GO term *anti-apoptosis* (out of $G = 2,071$ genes, GOALLLOCUSID environment in GO package). The term *anti-apoptosis* (GO:0006916) has the ninth smallest adjusted p-value for Scenario $\mathsf{MT}[t, t]$ applied to the BP gene ontology (Table 8).

BP GO:0006916

Probe ID	Symbol	Name
1237_at	IER3	immediate early response 3
1295_at	RELA	v-rel reticuloendotheliosis viral oncogene homolog A, nuclear factor of kappa light polypeptide gene enhancer in B-cells 3, p65 (avian)
1377_at	NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)
1564_at	AKT1	v-akt murine thymoma viral oncogene homolog 1
1830_s_at	TGFB1	transforming growth factor, beta 1 (Camurati-Engelmann disease)
1852_at	TNF	tumor necrosis factor (TNF superfamily, member 2)
1997_s_at	BAX	BCL2-associated X protein
277_at	MCL1	myeloid cell leukemia sequence 1 (BCL2-related)
31536_at	RTN4	reticulon 4
32060_at	BNIP2	BCL2/adenovirus E1B 19kDa interacting protein 2
33284_at	MPO	myeloperoxidase
36578_at	BIRC2	baculoviral IAP repeat-containing 2
38578_at	TNFRSF7	tumor necrosis factor receptor superfamily, member 7
38771_at	HDAC1	histone deacetylase 1
38994_at	SOCS2	suppressor of cytokine signaling 2
39097_at	SON	SON DNA binding protein
39378_at	BECN1	beclin 1 (coiled-coil, myosin-like BCL2 interacting protein)
39436_at	BNIP3L	BCL2/adenovirus E1B 19kDa interacting protein 3-like
40570_at	FOXO1A	forkhead box O1A (rhabdomyosarcoma)
595_at	TNFAIP3	tumor necrosis factor, alpha-induced protein 3
641_at	PSEN1	presenilin 1 (Alzheimer disease 3)

Table 12: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, MF GO term GO:0003735.* This table lists genes directly or indirectly annotated with GO term *structural constituent of ribosome* (out of $G = 2,071$ genes, GOALLLOCUSID environment in GO package). The term *structural constituent of ribosome* (GO:0003735) has the smallest adjusted p-value for Scenario $\mathsf{MT}[t,t]$ applied to the MF gene ontology (Table 10).

MF GO:0003735

Probe ID	Symbol	Name
2016_s_at	RPL10	ribosomal protein L10
31511_at	RPS9	ribosomal protein S9
31546_at	RPL18	ribosomal protein L18
31955_at	FAU	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived)
32221_at	MRPS18B	mitochondrial ribosomal protein S18B
32315_at	RPS24	ribosomal protein S24
32394_s_at	RPL23	ribosomal protein L23
32433_at	RPL15	ribosomal protein L15
32437_at	RPS5	ribosomal protein S5
33117_r_at	RPS12	ribosomal protein S12
33485_at	RPL4	ribosomal protein L4
33614_at	RPL18A	ribosomal protein L18a
33661_at	RPL5	ribosomal protein L5
33668_at	RPL12	ribosomal protein L12
33674_at	RPL29	ribosomal protein L29
34316_at	RPS15A	ribosomal protein S15a
36358_at	RPL9	ribosomal protein L9
36572_r_at	ARL6IP	ADP-ribosylation factor-like 6 interacting protein
36786_at	RPL10A	ribosomal protein L10a
39856_at	RPL36AL	ribosomal protein L36a-like
39916_r_at	RPS15	ribosomal protein S15
41152_f_at	RPL36A	ribosomal protein L36a
41214_at	RPS4Y1	ribosomal protein S4, Y-linked 1
41746_at	NHP2L1	NHP2 non-histone chromosome protein 2-like 1 (S. cerevisiae)

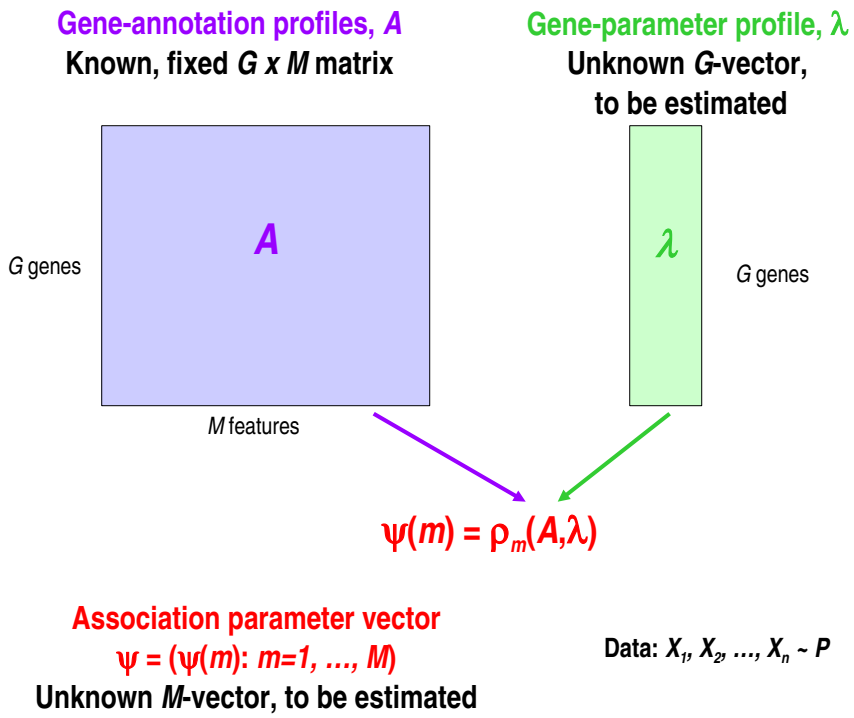


Figure 1: *Association parameters for gene-annotation profiles and gene-parameter profiles.*

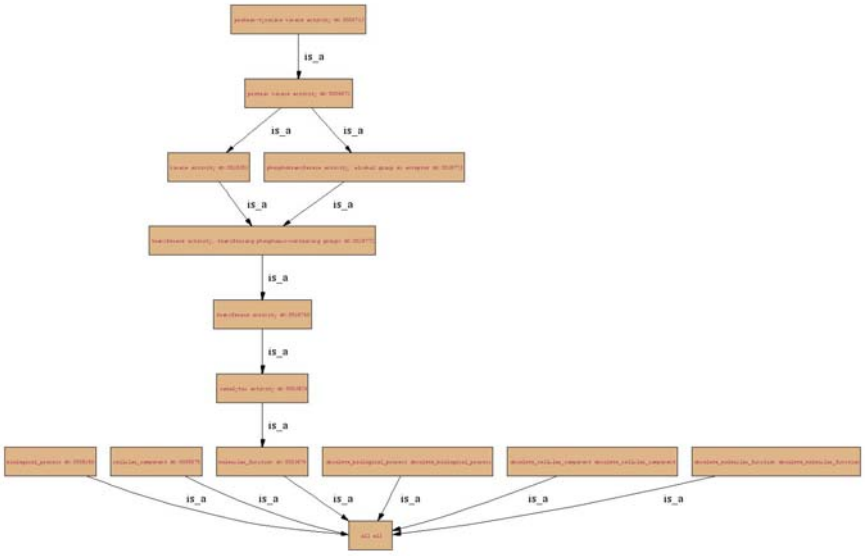


Figure 2: *DAG for MF GO term* `GO:0004713`*,* `AmiGO`*.* Portion of the directed acyclic graph for the GO term *protein-tyrosine kinase activity* (`GO:0004713`), in the Molecular Function ontology. This display, obtained using the `AmiGO` browser (Last updated 2006-02-14; `www.godatabase.org`), shows the nodes corresponding to all (less specific) ancestors of the term *protein-tyrosine kinase activity*.

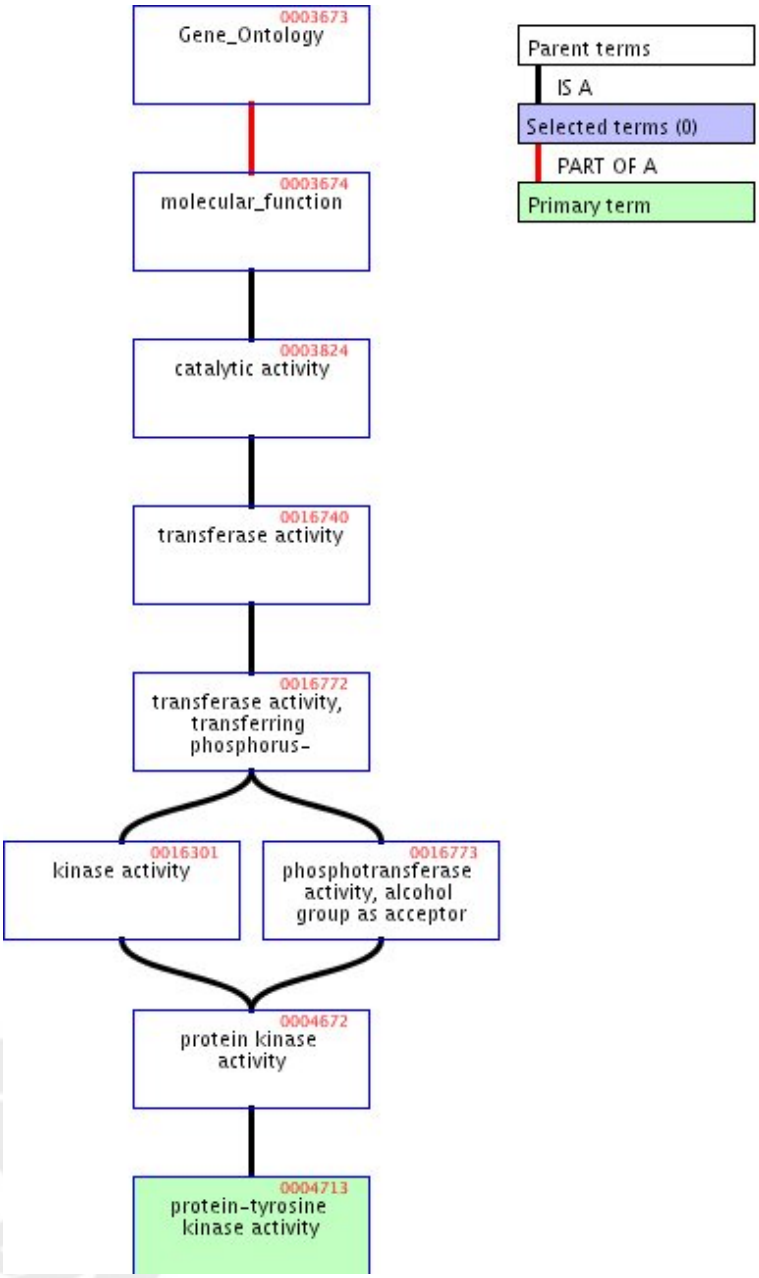
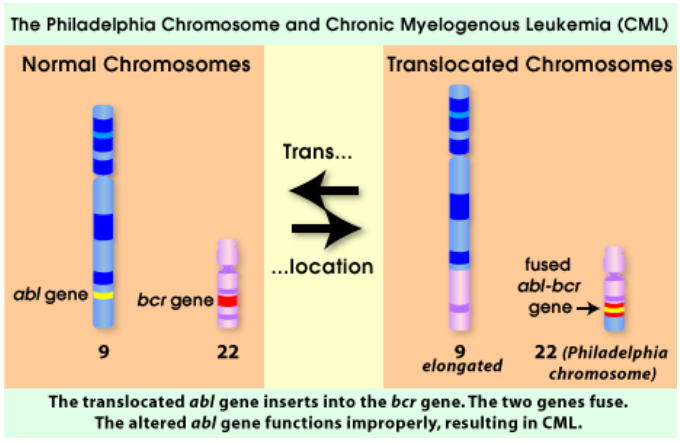
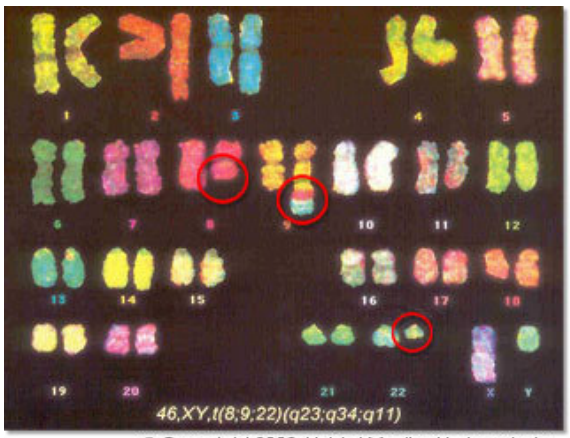


Figure 3: *DAG for MF GO term* `GO:0004713`*,* `QuickGO`*.* Portion of the directed acyclic graph for the GO term *protein-tyrosine kinase activity* (`GO:0004713`), in the Molecular Function ontology. This display, obtained using the EBI `QuickGO` browser (Last updated 2001-03-30 04:29:44.0; `www.ebi.ac.uk/ego`), shows the nodes corresponding to all (less specific) ancestors of the term *protein-tyrosine kinase activity*.

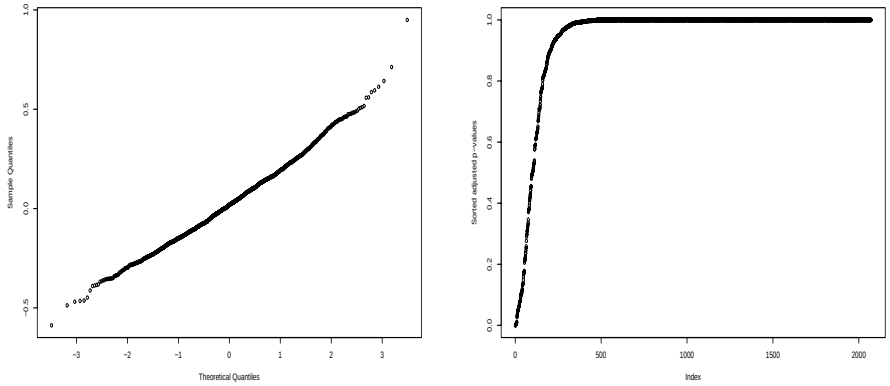


Panel (a): t(9;22) translocation



Panel (b): Karyotype

Figure 4: *The Philadelphia chromosome and the BCR/ABL fusion.* The BCR/ABL fusion is the molecular analogue of the Philadelphia chromosome. This t(9;22) translocation leads to a head-to-tail fusion of the `v-abl Abelson murine leukemia viral oncogene homolog 1` (`ABL1`) from chromosome 9 with the 5' half of the `breakpoint cluster region` (`BCR`) on chromosome 22. (Figure obtained from the Genetic Science Learning Center, The University of Utah; `gslc.genetics.utah.edu/units/disorders/karyotype/reciprocal.cfm`).



Panel (a): Test statistics Panel (b): Adjusted p-values

Figure 5: *Differentially expressed genes between BCR/ABL and NEG B-cell ALL.* Panel (a): Normal quantile-quantile plot of two-sample t-statistics $\lambda_n^t(g)$. Panel (b): Plot of sorted bootstrap-based single-step maxT adjusted p-values $\widetilde{P}_{0n}^{\neq}(g)$.

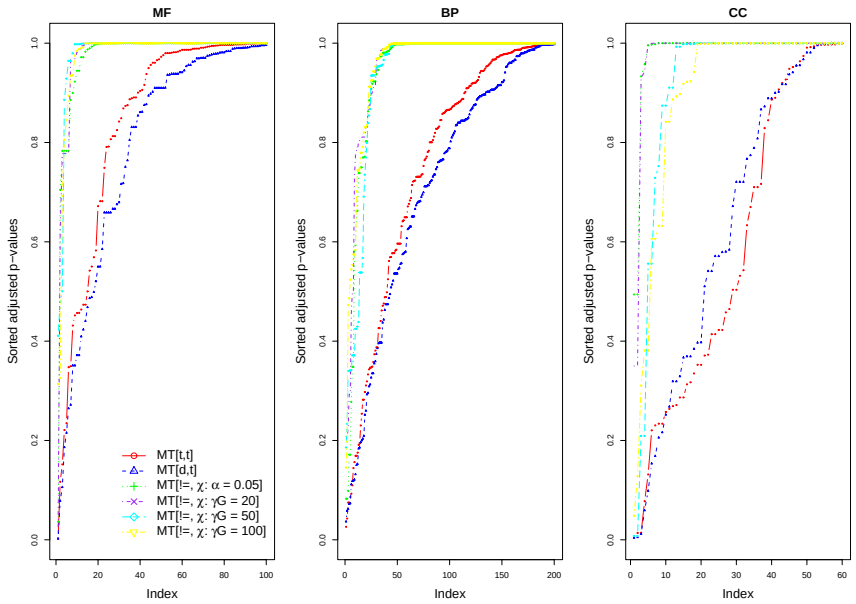


Figure 6: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, adjusted p-values.* Plots of sorted bootstrap-based single-step maxT adjusted p-values $\widetilde{P}_{0n}(m)$, for each of the three gene ontologies and each of the three testing scenarios.

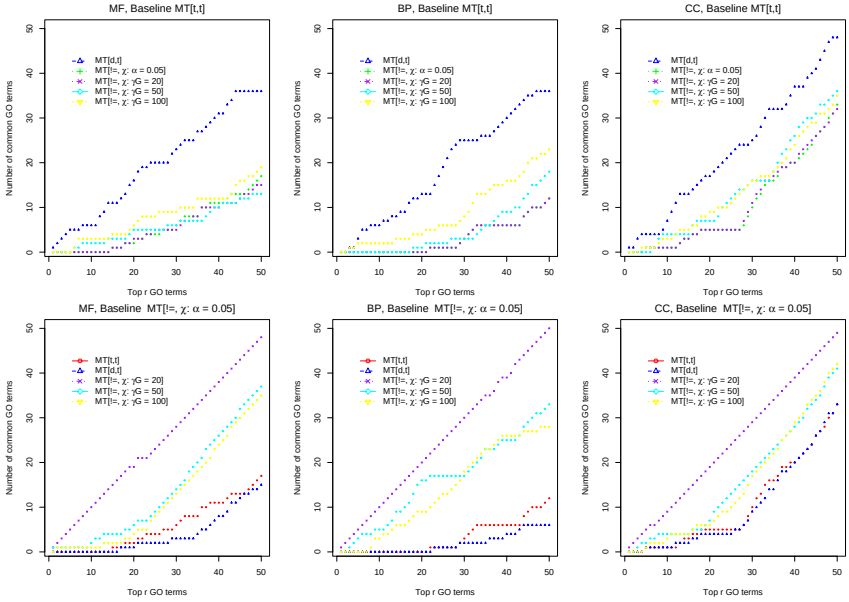


Figure 7: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, common terms between testing scenarios.* Plots of numbers of common GO terms among sets of ordered GO terms $\mathcal{O}_n(r)$ of various cardinality r for pairs of testing scenarios. Scenario $\mathsf{MT}[t,t]$ is used as the baseline in the top panels and Scenario $\mathsf{MT}[\neq, \chi : \alpha = 0.05]$, with adjusted p-value-based estimator $\lambda^{\neq}_{n,\alpha}$, $\alpha = 0.05$, for the binary DE gene-parameter profile $\lambda^{\neq}$, is used as the baseline in the bottom panels. For example, the blue curve in the top left panel is a plot of $|\mathcal{O}_n^{d,t}(r) \cap \mathcal{O}_n^{t,t}(r)|$ vs. r for the MF gene ontology, i.e., of the overlap between the r most significant MF GO terms according to Scenarios $\mathsf{MT}[d,t]$ and $\mathsf{MT}[t,t]$.

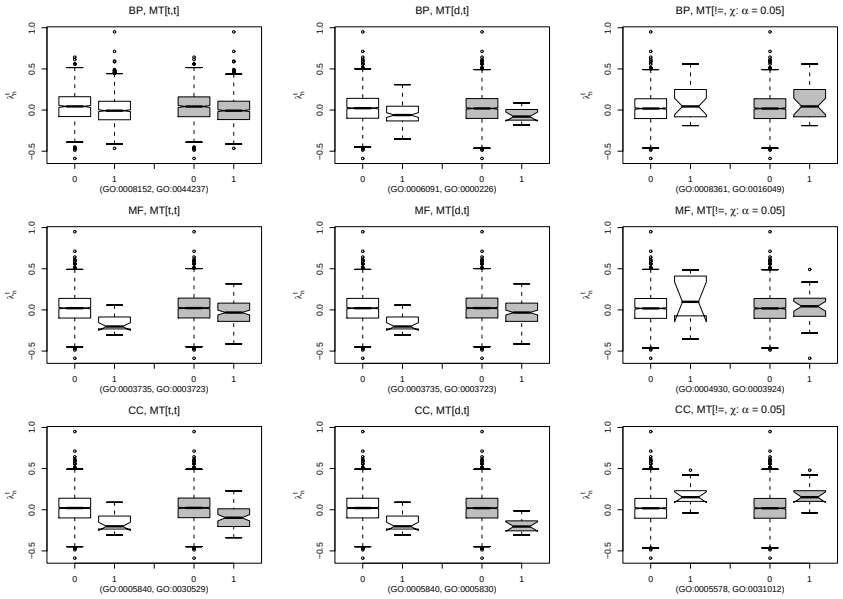


Figure 8: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, conditional distribution of λ_n^t given A.* Conditional boxplots of the estimated continuous DE gene-parameter profile λ_n^t given the gene-annotation profiles $A(\cdot, m)$ for the top two GO terms $m \in \{O_n(1), O_n(2)\}$ identified according to each of the three testing scenarios. Rows correspond to gene ontologies and columns to testing scenarios. In each panel, the white and gray boxplots correspond, respectively, to the GO terms with the smallest and second smallest adjusted p-values; boxplots for unannotated and annotated estimated gene-parameter profiles, $(\lambda_n^t(g) : A(g, m) = 0)$ and $(\lambda_n^t(g) : A(g, m) = 1)$, are labeled as 0 and 1, respectively. Non-overlapping notches (informally) represent large differences in medians.

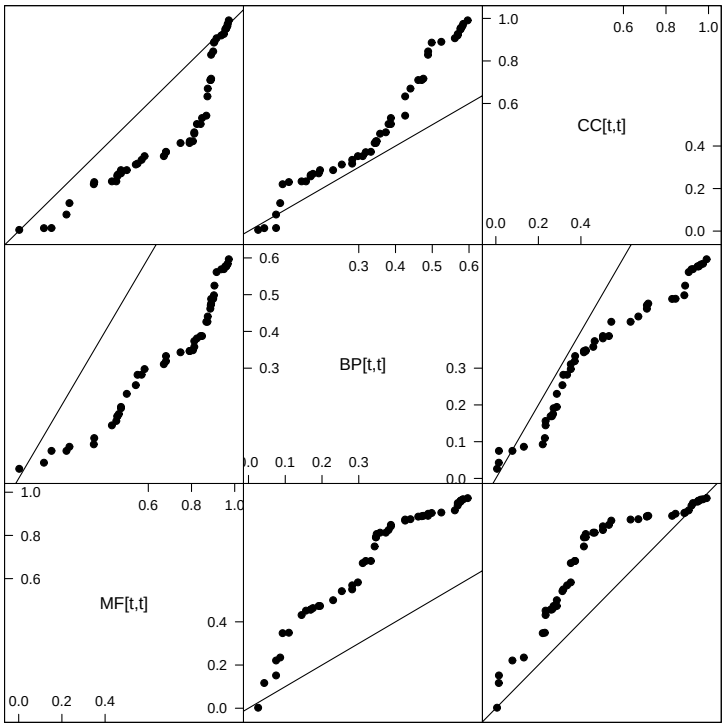


Figure 9: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, comparison of adjusted p-values for the three gene ontologies.* Scatterplot matrix of the 50 smallest adjusted p-values for each of the three gene ontologies, based on Scenario $\mathsf{MT}[t,t]$. The identity line is drawn for reference.

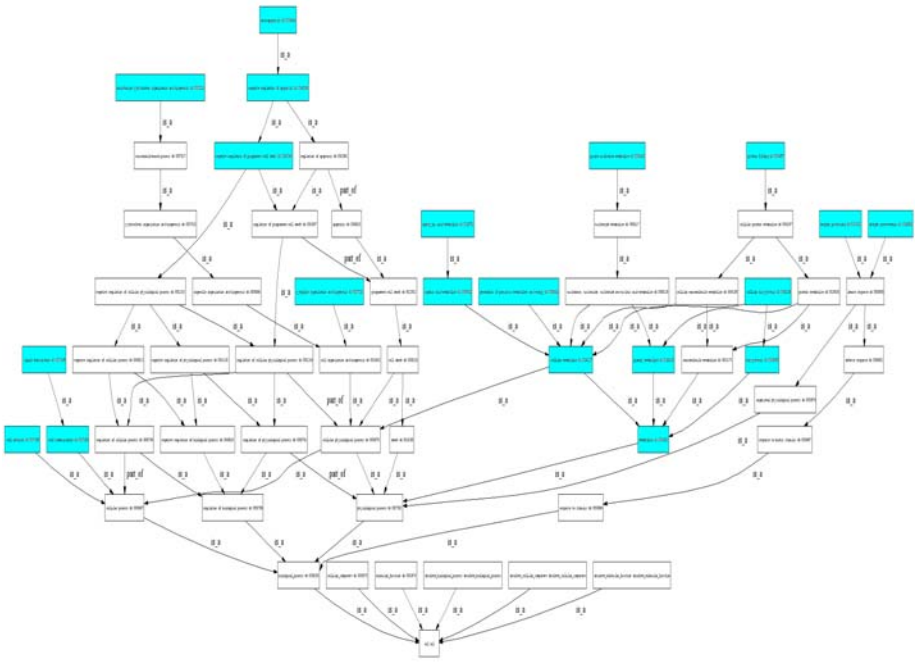


Figure 10: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, DAG for top 20 BP GO terms.* Portion of the directed acyclic graph for the 20 GO terms with the smallest adjusted p-values for Scenario $\mathsf{MT}[t,t]$ applied to the BP gene ontology (`AmiGO`). Nodes for the top 20 terms are shaded in turquoise.

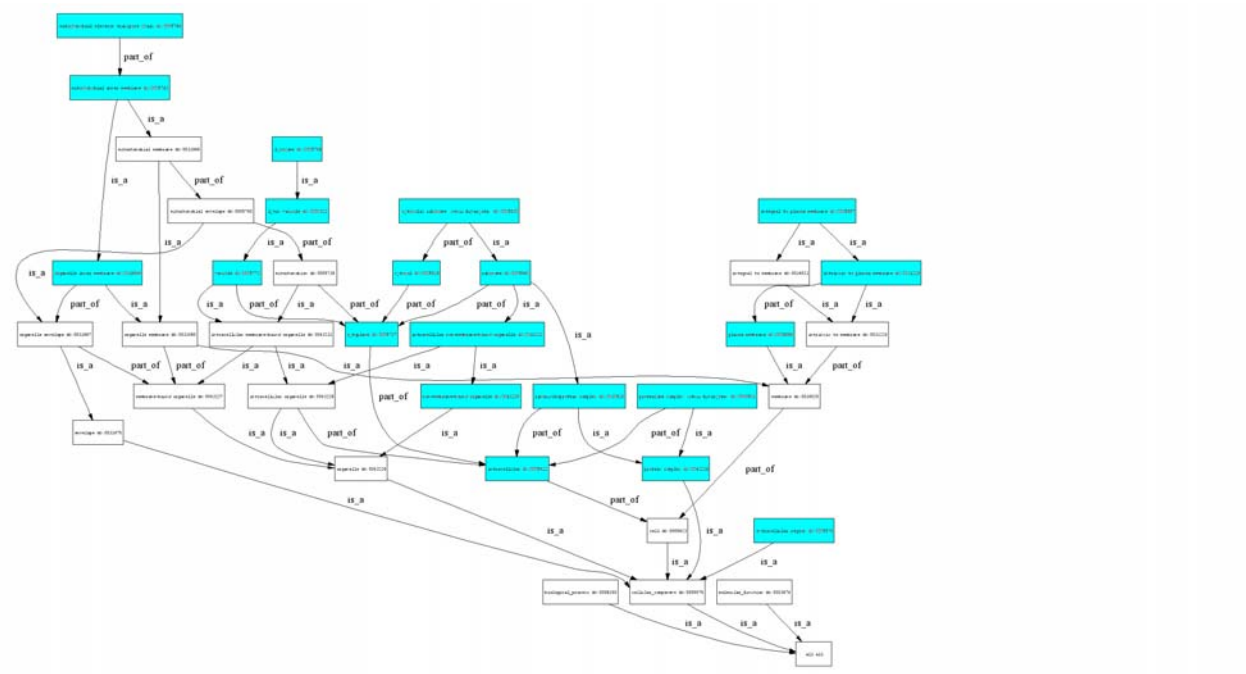


Figure 11: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, DAG for top 20 CC GO terms.* Portion of the directed acyclic graph for the 20 GO terms with the smallest adjusted p-values for Scenario $\mathsf{MT}[t,t]$ applied to the CC gene ontology (`AmiGO`). Nodes for the top 20 terms are shaded in turquoise.

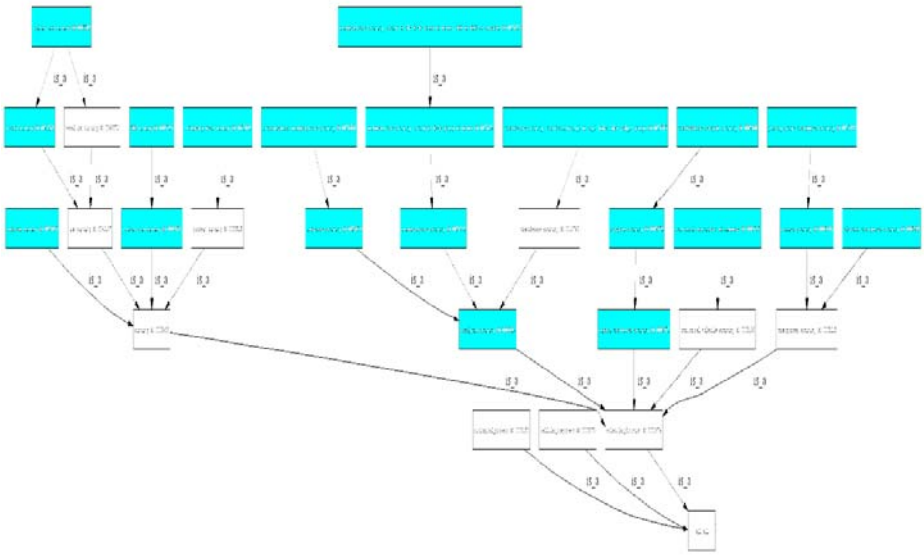
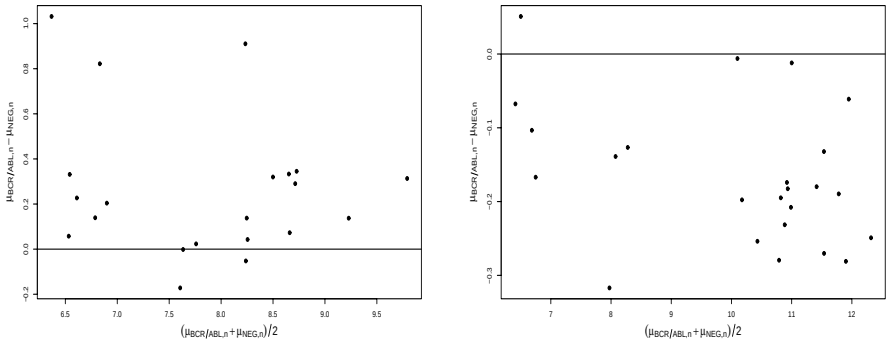


Figure 12: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, DAG for top 20 MF GO terms.* Portion of the directed acyclic graph for the 20 GO terms with the smallest adjusted p-values for Scenario $\mathsf{MT}[t,t]$ applied to the MF gene ontology (`AmiGO`). Nodes for the top 20 terms are shaded in turquoise.



Panel (a): BP `GO:0006916` Panel (b): MF `GO:0003735`

Figure 13: *GO terms associated with differential gene expression between BCR/ABL and NEG B-cell ALL, BP GO term `GO:0006916` and MF GO term `GO:0003735`.* This figure displays mean-difference plots of average expression measures in BCR/ABL and NEG cell samples, i.e., plots of $\mu_{BCR/ABL,n}(g) - \mu_{NEG,n}(g)$ vs. $(\mu_{BCR/ABL,n}(g) + \mu_{NEG,n}(g))/2$, for genes directly or indirectly annotated with GO terms `GO:0006916` (Panel (a)) and `GO:0003735` (Panel (b)). The term *anti-apoptosis* (`GO:0006916`) has the ninth smallest adjusted p-value for Scenario $\mathsf{MT}[t,t]$ applied to the BP gene ontology (Tables 8 and 11) and the term *structural constituent of ribosome* (`GO:0003735`) has the smallest adjusted p-value for Scenario $\mathsf{MT}[t,t]$ applied to the MF gene ontology (Tables 10 and 12).

References

F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatic*, 20(4):578–580, 2004. URL `fatigo.bioinfo.cipf.es`.

F. Al-Shahrour, P. Minguez, J. M. Vaquerizas, L. Conde, and J. Dopazo. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Research*, 33: W460–W464, 2005. URL `www.babelomics.org`.

B. Banelli, I. Casciano, M. Croce, A. Di Vinci, I. Gelvi, G. Pagnan, C. Brignole, G. Allemanni, S. Ferrini, M. Ponzoni, and M. Romani. Expression and methylation of CASP8 in neuroblastoma: Identification of a promoter region. *Nature Medicine*, 8(12):1333–1335, 2002.

T. Beissbarth and T. P. Speed. GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004. URL `gostat.wehi.edu.au`.

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, 2005. URL `www.bepress.com/ucbbiostat/paper168`.

M. Blanchette, R. E. Green, S. E. Brenner, and D. C. Rio. Global analysis of positive and negative pre-mRNA splicing regulators in *Drosophila*. *Genes & Development*, 19(11):1306–1314, 2005.

B. M. Bolstad, R. A. Irizarry, L. Gautier, and Z. Wu. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 2: Preprocessing High-density Oligonucleotide Arrays, pages 13–32. Statistics for Biology and Health. Springer-Verlag, New York, 2005.

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 and Applications to Genomics*. Springer, 2006. (In preparation).

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, 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.

R. C. Gentleman, V. J. Carey, and J. Zhang. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 7: Meta-data Resources and Tools in Bioconductor, pages 111–133. Statistics for Biology and Health. Springer-Verlag, New York, 2005.

B. Gleissner, N. Gokbuget, C. R. Bartram, B. Janssen, H. Rieder, J. W. Janssen, C. Fonatsch, A. Heyll, D. Voliotis, J. Beck, T. Lipp, G. Munzert, J. Maurer, D. Hoelzer, E. Thiel, and German Multicenter Trials of Adult Acute Lymphoblastic Leukemia Study Group. Leading prognostic relevance of the BCR-ABL translocation in adult acute B-lineage lymphoblastic leukemia: a prospective study of the German Multicenter Trial Group and confirmed polymerase chain reaction analysis. *Blood*, 99(5):1536–1543, 2002.

S. Grossmann, S. Bauer, P. N. Robinson, and M. Vingron. An improved statistic for detecting over-representated gene ontology annotations in gene sets. Technical report, Computational Molecular Biology, Max Planck Institute for Molecular Genetics, 2006. (Accepted, *RECOMB 2006*).

J. Kaczynski, T. Cook, and R. Urrutia. Sp1- and Krüppel-like transcription factors. *Genome Biology*, 4(2):206, 2003.

S. Keleş, M. J. van der Laan, S. Dudoit, and S. E. Cawley. Multiple testing methods for ChIP-Chip high density oligonucleotide array data. Technical Report 147, Division of Biostatistics, University of California, Berkeley, 2004. URL www.bepress.com/ucbbiostat/paper147.

D. Kirchner, J. Duyster, O. Ottmann, R. M. Schmid, L. Bergmann, and G. Munzert. Mechanisms of Bcr-Abl-mediated NF-κB/Rel activation. *Experimental Hematology*, 31(6):504–511, 2003.

S. A. McCarroll, C. T. Murphy, S. Zou, S. D. Pletcher, C-S Chin, Y. N. Jan, C. Kenyon, C. I. Bargmann, and H. Li. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genetics*, 36(2):197–204, 2004.

V. K. Mootha, C. M. Lindgren, K-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. H. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003.

A. Mukhopadhyay, S. Shishodia, J. Suttles, K. Brittingham, B. Lamothe, R. Nimmanapalli, K. N. Bhalla, and B. B. Aggarwal. Ectopic expression of protein-tyrosine kinase Bcr-Abl suppresses tumor necrosis factor (TNF)-induced NF-κB activation and IκBα phosphorylation. Relationship with

down-regulation of TNF receptors. *Journal of Biological Chemistry*, 277(34): 30622–30628, 2002.

K. S. Pollard, M. D. Birkner, M. J. van der Laan, and S. Dudoit. Test statistics null distributions in multiple testing: Simulation studies and applications to genomics. *Journal de la Société Française de Statistique*, 146(1–2):77–115, 2005a. URL `www.stat.berkeley.edu/~sandrine/Docs/Papers/SFdS05/SFdS.html`. Numéro double spécial *Statistique et Biopuces*.

K. S. Pollard, S. Dudoit, and M. J. van der Laan. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 15: Multiple Testing Procedures: the multtest Package and Applications to Genomics, pages 249–271. Statistics for Biology and Health. Springer-Verlag, New York, 2005b. URL `www.bepress.com/ucbbiostat/paper164`.

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.

D. B. Rubin, M. J. van der Laan, and S. Dudoit. Multiple testing procedures which are optimal at a simple alternative. Technical Report 171, Division of Biostatistics, University of California, Berkeley, 2005. URL `www.bepress.com/ucbbiostat/paper171`.

E. Shtivelman, F. E. Cohen, and J. M. Bishop. A human gene (AHNAK) encoding an unusually large protein with a 1.2-μm polyionic rod structure. *Proc. Natl. Acad. Sci.*, 89(12):5472–5476, 1992.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, 102(43): 15545–15550, 2005. URL `www.broad.mit.edu/gsea/doc/doc_index.html`.

L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park. Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci.*, 102(38):13544–13549, 2005.

M. J. van der Laan. Statistical inference for variable importance. *International Journal of Biostatistics*, 2(1):Article 2, 2006. URL `www.bepress.com/ijb/vol2/iss1/2`.

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`.

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`.

M. J. van der Laan and A. E. Hubbard. Quantile-function based null distribution in resampling based multiple testing. Technical Report 198, Division of Biostatistics, University of California, Berkeley, 2005. URL `www.bepress.com/ucbbiostat/paper198`.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York, 2003.

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`.