

Multiple Testing Procedures with Applications to Genomics

Part III. Applications to Genomics and Software Implementation

Sandrine Dudoit

PB HLTH 240D – Spring 2007

©Copyright 2007, all rights reserved

Outline: Part I. Motivation and Overview

- Multiple Hypothesis Testing Problems in Genomics.
- Multiple Hypothesis Testing Framework.
- Data Generating Distribution.
- Parameters.
- Null and Alternative Hypotheses.
- Test Statistics.
- Multiple Testing Procedures.
- Rejection Regions.

Outline

These lecture notes are based on the forthcoming book by Dudoit and van der Laan (2007).

Related articles and tech reports may be downloaded from Sandrine Dudoit's website

www.stat.berkeley.edu/~sandrine

and Mark van der Laan's website

www.stat.berkeley.edu/~laan.

Outline: Part I. Motivation and Overview

- Errors in Multiple Hypothesis Testing: Type I, Type II, and Type III errors.
- Type I Error Rates.
- Power.
- Unadjusted and Adjusted p -Values.
- Examples of Multiple Testing Procedures.

[Dudoit and van der Laan (2007, Chapter 1)]

Outline: Part II. Methodology

- Test Statistics Null Distribution [Chapter 2].
- Single-Step Multiple Testing Procedures for Controlling General Type I Error Rates, $\Theta(F_{V_n})$ [Chapter 4].
- Step-Down Multiple Testing Procedures for Controlling the Family-Wise Error Rate [Chapter 5].
- Augmentation Multiple Testing Procedures for Controlling Generalized Tail Probability Error Rates [Chapter 6].
- Resampling-Based Empirical Bayes Multiple Testing Procedures for Controlling Generalized Tail Probability Error Rates [Chapter 7].
- Appendix: Miscellaneous Mathematical and Statistical Results.

Outline: Part III. Applications to Genomics and Software Implementation

- Identification of Differentially Expressed and Co-Expressed Genes in High-Throughput Gene Expression Experiments [Chapter 9].
- Multiple Tests of Association with Biological Annotation Metadata [Chapter 10].
- HIV-1 Sequence Variation and Viral Replication Capacity [Chapter 11].
- Genetic Mapping of Complex Human Traits Using Single Nucleotide Polymorphisms: The ObeLinks Project [Chapter 12].
- Software Implementation [Chapter 13].

High-Throughput Gene Expression Experiments

- Identification of **differentially expressed** (DE) genes, i.e., genes whose expression measures are associated with possibly censored biological and clinical covariates and outcomes. Simultaneous test, for each gene, of the null hypothesis of no association between the expression measures and the covariates and outcomes.
- Identification of **co-expressed** (CE) genes, i.e., pairs (or sets) of genes with associated expression measures. Simultaneous test, for each gene pair, of the null hypothesis of no association (e.g., zero correlation) between their expression measures.

The multiple testing results can be used as a basis for further analyses, e.g., **clustering** and **graph theoretical analyses**.

Apo AI Dataset: Differential Expression

Callow et al. (2000). **Apo AI dataset**. The Apo AI microarray experiment was carried out as part of a study of **lipid metabolism** and **atherosclerosis** susceptibility in mice.

Apolipoprotein AI (Apo AI) is a gene known to play a pivotal role in high-density lipoprotein (HDL) **cholesterol metabolism** and mice with the Apo AI gene knocked-out have very low HDL cholesterol levels.

The goal of the experiment was to identify **differentially expressed** genes in the livers of Apo AI knock-out mice compared to inbred control mice.

The **treatment** group consists of 8 **inbred C57Bl/6 mice with the Apo AI gene knocked-out** and the **control** group consists of 8 **inbred C57Bl/6 mice**.

Apo AI Dataset: Differential Expression

For each of the 16 mice, **target samples** of complementary DNA (cDNA) were obtained from messenger RNA (mRNA) by reverse transcription and labeled using the red-fluorescent dye Cyanine 5 (Cy5).

The **reference sample** used in all hybridizations was prepared by pooling cDNA from the 8 control mice and was labeled with the green-fluorescent dye Cyanine 3 (Cy3).

Combined red- and green-labeled target cDNA samples were hybridized to **microarrays** with 6,384 spots ($= 4 \times 4 \times 19 \times 21$), including 257 probe sequences related to lipid metabolism.

Apo AI Dataset: Differential Expression

Pre-processing. Each of the 16 hybridizations produced a pair of 16-bit TIFF images, which were processed by seeded region growing segmentation and morphological opening background adjustment (Yang et al. (2002); R package `Spot`, experimental.act.cmis.csiro.au/Spot/index.php).

The resulting fluorescence intensity measures were normalized by within-print-tip-group loess robust local regression (Dudoit et al. (2002); Dudoit and Yang (2003); Yang et al. (2001); Yang and Paquet (2005); Bioconductor R package `marray`, www.bioconductor.org).

Among the 6,384 spots on the microarray, only those 5,548 spots corresponding to actual cDNA sequences are retained for subsequent analyses. The other 836 spots are either blank or control spots.

Apo AI Dataset: Differential Expression

Data. The data for each of the $n = 16$ mice consist of the following.

- Z , a binary **covariate/genotype** (1 for treatment vs. 0 for control).
- $X = (X(m) : m = 1, \dots, M)$, an $M = 5,548$ -dimensional **outcome/phenotype** vector of **microarray expression measures**.

The R package `ApoAI` provides microarray data objects, at various levels of processing, for the Apo AI experiment (www.stat.berkeley.edu/~sandrine/MTBook).

Apo AI Dataset: Differential Expression

Multiple testing question. Identify **differentially expressed** genes by testing, for each of the $M = 5,548$ probes, whether there is a **difference in mean expression measures** between knock-out and control mice.

Parameters of interest. For the m th probe, the parameter of interest is the **difference in mean expression measures** $\psi(m)$ between treatment and control mice, that is,

$$\psi(m) = E[X(m)|Z = 1] - E[X(m)|Z = 0], \quad m = 1, \dots, M.$$

Null hypotheses. Consider two-sided tests of the M null hypotheses of no differences in mean expression measures vs. the alternative hypotheses of different mean expression measures,

$$H_0(m) = I(\psi(m) = 0) \quad \text{vs.} \quad H_1(m) = I(\psi(m) \neq 0).$$

Apo AI Dataset: Differential Expression

Test statistics. Two-sample Welch t -statistics,

$$T_n(m) = \frac{\psi_n(m) - \psi_0(m)}{\sigma_n(m)} = \frac{\bar{X}_1(m) - \bar{X}_0(m) - 0}{\sqrt{\frac{\sigma_{0,n}^2(m)}{n_0(m)} + \frac{\sigma_{1,n}^2(m)}{n_1(m)}}},$$

where the null values $\psi_0(m)$ are zero and $n_k(m) = \sum_i I(Z_i = k)$, $\bar{X}_k(m) = \sum_i I(Z_i = k) X_i(m)/n_k(m)$, and $\sigma_{k,n}^2(m) = \sum_i I(Z_i = k) (X_i(m) - \bar{X}_k(m))^2 / (n_k(m) - 1)$ denote, respectively, the sample sizes, sample means, and sample variances for the expression measures of probe m in treatment ($k = 1$) and control ($k = 0$) mice (note that the sample sizes $n_k(m)$ may differ across probes m due to missing data).

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

Apo AI Dataset: Differential Expression

All multiple testing procedures single out 8 probes, out of 5,548 spotted probe sequences, as being differentially expressed between knock-out and control mice.

The negative t -statistics suggest that the probes are under-expressed in the Apo AI knock-out mice compared to the control mice.

The 8 most extreme probes actually correspond to only 4 distinct genes and 1 EST: ApoAI (2 copies), ApoCIII (2 copies), Steroidesaturase (2 copies), Catechol – O – methyltransferase (1 copy), and a novel EST (1 copy).

All changes were confirmed by real-time quantitative PCR (RT-PCR), as described in Callow et al. (2000).

Hyperlinked Supplementary Table 9.1.

Apo AI Dataset: Differential Expression

Multiple testing procedures.

- **FWER control:** Joint single-step maxT procedure (SS maxT) and minP procedure (SS minP), joint step-down maxT procedure (SD maxT) and minP procedure (SD minP), marginal single-step Bonferroni (1936) procedure (SS Bonferroni), marginal step-down Holm (1979) procedure (SD Holm), and marginal step-up Hochberg (1988) procedure (SU Hochberg).
- **gFWER control:** $gFWER(k)$ -controlling augmentation multiple testing procedure, based on FWER-controlling joint single-step maxT procedure, for an allowed number $k \in \{5, 10, 50, 100\}$ of false positives (gFWER(k) AMTP).
- **TPPFP control:** $TPPFP(q)$ -controlling augmentation multiple testing procedure, based on FWER-controlling joint single-step maxT procedure, for an allowed proportion $q \in \{0.05, 0.10, 0.25, 0.50\}$ of false positives (TPPFP(q) AMTP).
- **FDR control:** Marginal step-up Benjamini and Hochberg (1995) procedure (SU BH) and Benjamini and Yekutieli (2001) procedure (SU BY), TPPFP-based MTPs of van der Laan et al. (2004b) (TPPFP-based 1, TPPFP-based 2).

Apo AI Dataset: Differential Expression

The presence of ApoAI among the under-expressed genes is to be expected, as this is the gene that was knocked out in the treatment mice.

The ApoCIII gene, also associated with lipoprotein metabolism, is located very close to the ApoAI locus. Callow et al. (2000) showed that the down-regulation of ApoCIII is actually due to genetic polymorphism rather than lack of ApoAI. The presence of ApoAI and ApoCIII among the under-expressed genes thus provides a validation of the statistical methodology, if not a biologically novel finding.

Steroidesaturase is an enzyme that catalyzes one of the terminal steps in cholesterol synthesis.

Liver membrane-bound Catechol – O – methyltransferase was found to be a relevant factor in blood pressure regulation in rats (Tsunoda et al., 2003).

The novel EST shares sequence similarity to a family of ATPases.

Apo AI Dataset: Differential Expression

The Apo AI experiment is rather unusual, in the sense that 8 spotted probe sequences clearly stand out from the remaining 5,540 probes as being differentially expressed.

Such a **dichotomy in gene expression** is seldom observed in other applications of the microarray technology.

For example, in many cancer microarray studies, genes tend to exhibit a **continuum of change in expression measures** and it is difficult to identify distinct groups of genes.

Differences in **patterns of differential expression** likely reflect the **nature of the target samples** under investigation.

The Apo AI experiment compares relatively pure cell samples (hepatocytes), from wild-type and knock-out mice with an otherwise identical genetic background. In contrast, human cancer microarray studies typically assay samples composed of a variety of cell types, from genetically diverse individuals.

Apo AI Dataset: Differential Expression

Table 1: *Apo AI dataset: FWER-controlling non-parametric bootstrap-based single-step maxT MTP.*

Gene name	Spot ID	Estimate $\psi_n(m)$	t-statistic $T_n(m)$	Unadjusted p-value $P_{0n}(m)$	Adjusted p-value $\hat{P}_{0n}(m)$
1 Apoa1 ApoAI Apo AI, lipid-Img	2149	-3.17	-22.89	0.0000	0.0014
1 Sc5d Sterol-desaturase EST, Weakly similar to C-5 STEROL DESATURASE [Saccharomyces cerevisiae], lipid-UG	4139	-1.03	-13.02	0.0000	0.0136
1 Comt Catechol-O-methyltransferase CATECHOL O-METHYLTRANSFERASE, MEMBRANE-BOUND FORM, Brain-Img	5356	-1.86	-11.80	0.0000	0.0214
1 Apoa1 ApoAI EST, Highly similar to APOLIPOPROTEIN A-I PRECURSOR [Mus musculus], lipid-UG	540	-3.05	-11.69	0.0000	0.0224

Continued on next page ...

... continued from previous page

Gene name	Spot ID	Estimate $\psi_n(m)$	t-statistic $T_n(m)$	Unadjusted p-value $P_{0n}(m)$	Adjusted p-value $\hat{P}_{0n}(m)$
2 Apoc3 ApoCIII Apo CIII, lipid-Img	1739	-0.96	-10.81	0.0002	<u>0.0322</u>
1 EST est	1496	-0.99	-9.05	0.0000	0.0606
2 Apoc3 ApoCIII ESTs, Highly similar to APOLIPOPROTEIN C-III PRECURSOR [Mus musculus], lipid-UG	2537	-1.02	-8.74	0.0002	0.0694
2 Sc5d Sterol-desaturase similar to yeast sterol desaturase, lipid-Img	4941	-0.97	-7.29	0.0002	0.1346
4 Casp7 Caspase7 Caspase 7, heart-Img	954	-0.31	-4.70	0.0018	0.4840

Apo AI Dataset: Differential Expression

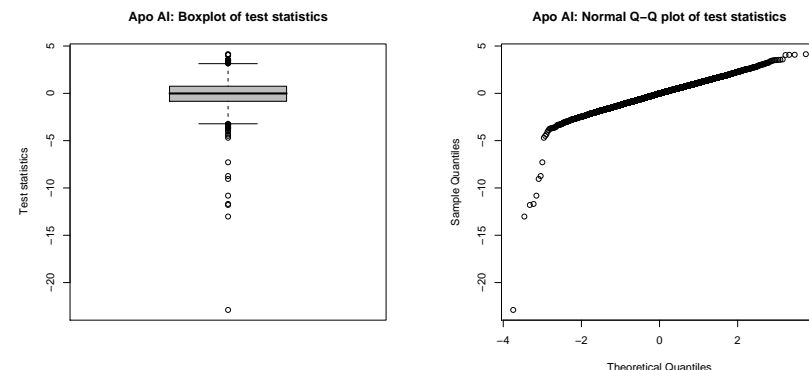


Figure 1: *Apo AI dataset: Test statistics.*

Apo AI Dataset: Differential Expression

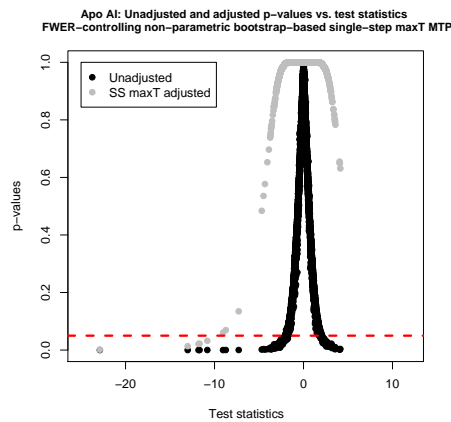


Figure 2: Apo AI dataset: FWER-controlling non-parametric bootstrap-based single-step maxT MTP, test statistics and p-values.

Apo AI Dataset: Differential Expression

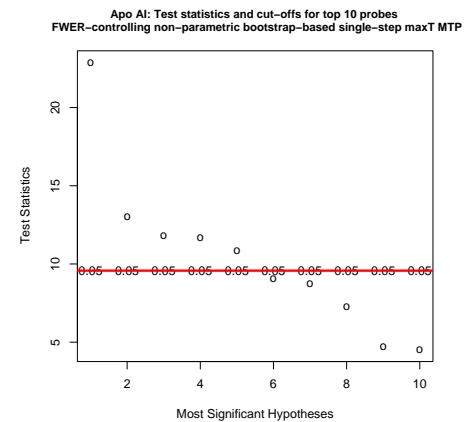


Figure 3: Apo AI dataset: FWER-controlling non-parametric bootstrap-based single-step maxT MTP, test statistics and cut-offs.

Apo AI Dataset: Differential Expression

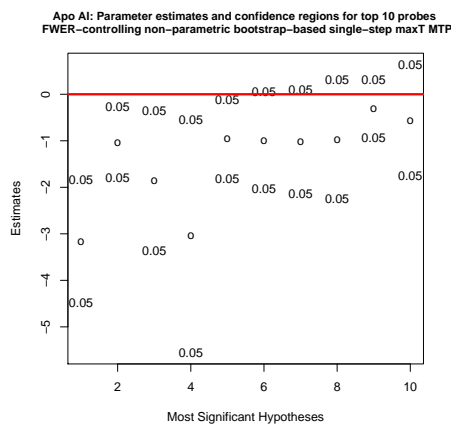


Figure 4: Apo AI dataset: FWER-controlling non-parametric bootstrap-based single-step maxT MTP, parameter estimates and confidence regions.

Apo AI Dataset: Differential Expression

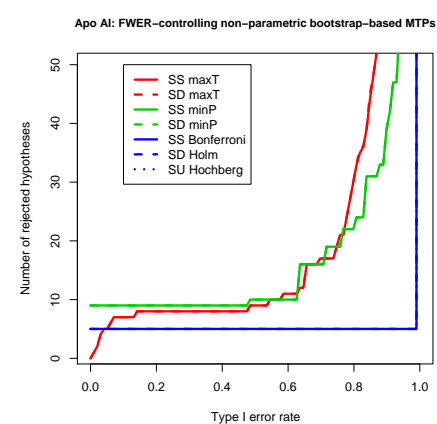


Figure 5: Apo AI dataset: FWER-controlling non-parametric bootstrap-based MTPs.

Apo AI Dataset: Differential Expression

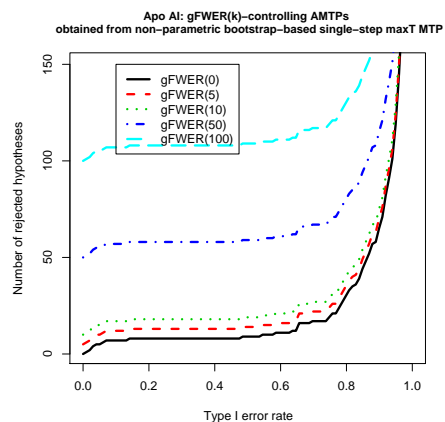


Figure 6: Apo AI dataset: gFWER-controlling non-parametric bootstrap-based AMTPs.

Apo AI Dataset: Differential Expression

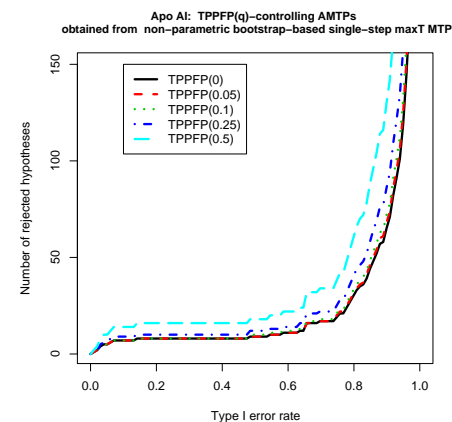


Figure 7: Apo AI dataset: TPPFP-controlling non-parametric bootstrap-based AMTPs.

Apo AI Dataset: Differential Expression

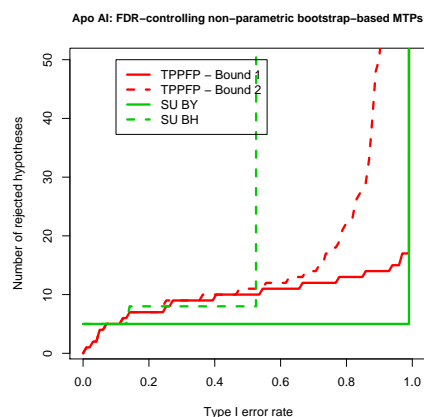


Figure 8: Apo AI dataset: FDR-controlling non-parametric bootstrap-based MTPs.

Cancer microRNA Dataset

In addition to passing genetic messages from DNA to the protein-making machinery of the cell, ribonucleic acids (RNA) serve many other cellular functions.

microRNAs (miRNA) are small, non-coding RNAs involved in gene regulation and developmental timing (microrna.sanger.ac.uk).

By binding to messenger RNA (mRNA), miRNAs regulate gene expression post-transcriptionally and affect the abundance of a wide range of proteins, in diverse biological processes.

By now, hundreds of miRNAs have been identified, in various multicellular organisms, including the fruitfly *Drosophila melanogaster* and humans, and many are evolutionary conserved.

Although the biological functions of miRNAs are still largely unknown, miRNAs are predicted to regulate up to 30% of all protein-coding genes.

Cancer microRNA Dataset

Each mammalian miRNA is believed to regulate approximately 200 genes and many genes have several target sites for one or several different miRNAs.

The large number of miRNA genes, their diverse expression patterns, and the abundance of miRNA targets, suggest the **involvement of miRNAs in a variety of diseases**, including cancers and viruses.

More than half of the known human miRNA genes are located in genomic regions related to **cancers**, such as, fragile sites, minimal regions of loss of heterozygosity, minimal regions of amplification, and common breakpoint regions.

miRNAs have also been implicated in several mammalian **viruses**, such as, the Epstein-Barr virus and the human immunodeficiency virus (HIV).

Cancer microRNA Dataset

Lu et al. (2005). Cancer microRNA dataset. Lu et al. (2005) used a bead-based flow cytometric profiling method to measure the levels of 217 known human miRNAs in $n = 186$ cell samples derived from cancerous and non-cancerous tissues.

These authors found that **predictors based on miRNA expression measures** are better able to distinguish developmental lineage, differentiation state, and cancer state, than the best corresponding predictors based on genome-wide mRNA expression measures from the same cells.

Cancer microRNA Dataset

Pre-processing. \log_2 -transform; exclude cell lines; exclude any miRNA with expression measures below a detection threshold of $\log_2 32 = 5$ in more than half of the $n = 186$ samples.

Data. The data for each of the $n = 186$ samples consist of the following.

- Y , a binary **cancer status outcome/phenotype** (1 for cancerous vs. 0 for non-cancerous).
- $X = (X(j) : j = 1, \dots, J)$, a $J = 155$ -dimensional **covariate vector** of real-valued **miRNA expression measures**.
- W , a 19-dimensional **tissue type** indicator vector — confounding variable.

The data are available at www.broad.mit.edu/cancer/pub/miGCM.

Cancer microRNA Dataset

Goals.

- Identify **differentially expressed** miRNAs, i.e., miRNAs whose expression measures are associated with cancer status (cancerous vs. non-cancerous).

Our approach is based on **tests for regression coefficients in logistic models** that relate cancer status Y to miRNA expression measures $X(j)$, while **adjusting** for the **confounding variable** tissue type W .

- Identify **co-expressed** miRNAs, i.e., pairs (groups) of miRNAs with correlated expression profiles across tissue samples.

Cancer microRNA Dataset: Differential Expression

Logistic regression model. For the j th miRNA, fit a logistic regression model that includes expression measure $X(j)$ and tissue type W as covariates,

$$\text{logit}(E[Y|X(j), W]) = \alpha(j) + \beta(j)X(j) + \gamma(j)W, \quad j = 1, \dots, J,$$

where $\text{logit}(z) = \log(z/(1 - z))$ is the logit function,

$\alpha(j)$ a baseline effect parameter,

$\beta(j)$ a **main effect parameter** for the expression measure $X(j)$ of the j th miRNA, and

$\gamma(j)$ a miRNA-specific 19-dimensional parameter vector adjusting for tissue type W .

Cancer microRNA Dataset: Differential Expression

Parameters of interest. For the j th miRNA, the parameter of interest is the **logistic regression coefficient** $\beta(j)$ for the expression measure $X(j)$.

Null hypotheses. Consider two-sided tests of the J null hypotheses of no association between the expression measures $X(j)$ and cancer status Y ,

$$H_0(j) = I(\beta(j) = 0) \quad \text{vs.} \quad H_1(j) = I(\beta(j) \neq 0).$$

Cancer microRNA Dataset: Differential Expression

Test statistics. t -statistics for logistic regression coefficients,

$$T_n(j) = \frac{\beta_n(j) - \beta_0(j)}{\sigma_n(j)},$$

where the null values $\beta_0(j)$ are zero and $\beta_n(j)$ are logistic regression parameter estimators with estimated standard errors $\sigma_n(j)$ (as implemented in the function `glm` from the R package `stats`, with the call `glm(Y ~ X(j) + W, family="binomial")`, using the binomial family and iteratively reweighted least squares (IWLS)).

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

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

Cancer microRNA Dataset: Differential Expression

Table 2: *Cancer miRNA dataset, differential expression: Tests for logistic regression coefficients.* Number of differentially expressed miRNAs (out of $J = 155$) for different nominal FWER levels α .

Nominal FWER, α	Number of miRNAs, R_n
0.05	90 (58%)
0.01	53 (34%)

All 90 miRNAs that are significantly differentially expressed at level $\alpha = 0.05$ have **negative test statistics** ($T_n(j) < -3.8$), suggesting **under-expression** in cancerous compared to non-cancerous tissues.

Cancer microRNA Dataset: Differential Expression

Table 3: Cancer miRNA dataset, differential expression: Tests for logistic regression coefficients. 53 most significantly differentially expressed miRNAs between cancerous and non-cancerous tissues (bootstrap-based single-step maxT adjusted p -values < 0.01).

Name	miRNA target sequence	Adjusted p -value	Test statistic
hsa – miR – 98	UGAGGUAGUAAGUUGUAUUGUU	0.0038	-4.88
hsa – miR – 28	AAGGAGCUCACAGUCUAUUGAG	0.0038	-4.79
hsa – miR – 196	UAGGUAGUUUCAUGUUGUUGG	0.0038	-4.79
hsa – miR – 30a	CUUUCAGUCGGAUUUUGCAGC	0.0038	-4.78
hsa – miR – 30e	UGUAAACAUCUUUGACUGGA	0.0038	-4.78
hsa – miR – 99a#	AACCCGUAGAUCCGAUCUUGUG	0.0038	-4.77
hsa – miR – 335	UCAAGAGCAAUAACGAAAAAUGU	0.0038	-4.72
hsa – let – 7e	UGAGGUAGGAGGUUGUAUAGU	0.0038	-4.69
hsa – miR – 23b#	AUCACAUUGCCAGGGAUACCAC	0.0038	-4.67
hsa – miR – 214	ACAGCAGGCACAGACAGGCAG	0.0038	-4.67
hsa – miR – 99b	CACCCGUAGAACCCACCUUGCG	0.0038	-4.67
hsa – miR – 30c	UGUAAACAUCUACACUCUCAGC	0.0038	-4.66
hsa – miR – 30b	UGUAAACAUCUACACUCAGC	0.0038	-4.66
hsa – miR – 338	UCCAGCAUCAGUGAUUUUGUUGA	0.0038	-4.65

Continued on next page ...

... continued from previous page

Name	miRNA target sequence	Adjusted p -value	Test statistic
hsa – miR – 103	AGCAGCAUUGUACAGGGCUAUGA	0.0038	-4.64
hsa – miR – 185	UGGAGAGAAAGGCAGUUC	0.0038	-4.63
hsa – miR – 151*	UCGAGGAGCUCACAGUCUAGUA	0.0038	-4.62
hsa – miR – 100#	AACCCGUAGAUCCGAUCUUGUG	0.0038	-4.61
hsa – miR – 20_(sub_1)	UAAAGUGCUUUAUAGUCAGGUAG	0.0038	-4.61
hsa – miR – 129*	AAGCCUUACCCCAAAAAGCAU	0.0038	-4.60
hsa – miR – 22#	AAGCUGCCAGUUGAAGAUCUGU	0.0038	-4.60
hsa – let – 7d#	AGAGGUAGUAGGUUGCAUAGU	0.0038	-4.58
hsa – miR – 107	AGCAGCAUUGUACAGGGCUAUA	0.0038	-4.58
rno – miR – 352	AGAGUAGUAGGUUGCAUAGUA	0.0038	-4.58
hsa – miR – 197	UUCACCACCUUCUCCACCCAGC	0.0038	-4.57
hsa – miR – 32	UAUUGCACAUUACUAAGUUGC	0.0038	-4.57
hsa – miR – 342	UCUCACACAGAAAUCGCACCCGUC	0.0038	-4.56
hsa – miR – 324 – 5p	CGCAUCCCCUAGGGCAUUGGUUGU	0.0038	-4.51
hsa – miR – 128b	UCACAGUGAACCCGUCUCUUC	0.0038	-4.51
hsa – miR – 126*	CAUUUAUACUUUUGGUACGGG	0.0038	-4.50
hsa – miR – 19b	UGUGCAAAUCCAUUGCAAACUGA	0.0038	-4.49
hsa – miR – 151_(sub_1)	ACUAGACUGAGGCUCCUUGAGG	0.0038	-4.49
hsa – miR – 199a*	UACAGUAGUCUGCACAUUGGUU	0.0038	-4.48
hsa – let – 7i	UGAGGUAGUAGUUUGUGCU	0.0038	-4.48
hsa – miR – 10b	UACCCUGUAGAACCGAAUUUGU	0.0038	-4.47
miR – 292 – 3p	AAGUGCCGCCAGGUUUUGAGUGU	0.0040	-4.46

Continued on next page ...

... continued from previous page

Name	miRNA target sequence	Adjusted p -value	Test statistic
hsa – miR – 136	ACUCCAUUUUGUUUGAUGAUGGA	0.0042	-4.45
mmu – miR – 10b	CCCUGUAGAACCCAAUUUGUGU	0.0042	-4.45
hsa – let – 7f	UGAGGUAGUAGAUUGUAUAGUU	0.0042	-4.44
hsa – miR – 302	UAAGUGCUUCCAUGUUUUGGUGA	0.0042	-4.43
mmu – let – 7g	UGAGGUAGUAGUUUGUACAGU	0.0042	-4.43
hsa – miR – 10a	UACCCUGUAGAUCCGAAUUUGUG	0.0042	-4.42
hsa – miR – 34b	AGGCAGUGUCAUUAAGCUGAUUG	0.0042	-4.42
hsa – miR – 92	UAUUGCACUUGUCCCGCCUGU	0.0042	-4.42
hsa – miR – 101	UACAGUACUGUGAUAACUGAAG	0.0044	-4.38
hsa – miR – 16	UAGCAGCACGUAAAUAUUGGCG	0.0046	-4.37
mmu – miR – 339	UCCCUGUCCUCCAGGAGCUCA	0.0046	-4.37
hsa – miR – 19a	UGUGCAAAUCUAUGCAAAACUGA	0.0046	-4.37
hsa – miR – 152	UCAGUGCAUGACAGAACUUGG	0.0052	-4.35
hsa – miR – 23a	AUCACAUUGCCAGGGAUUUCC	0.0052	-4.34
hsa – miR – 186	CAAAGAAUUUCUUUUUGGGUUU	0.0072	-4.30
rno – miR – 343	UCUCCUCUCCGUGUGCCAGU	0.0096	-4.29
hsa – miR – 140	AGUGGUUUUACCCUAUGGUAG	0.0096	-4.28

Located in minimal deleted regions, minimal amplified regions, and breakpoint regions involved in human cancers (Calin et al., 2004).

Cancer microRNA Dataset: Differential Expression

Our findings are in agreement with the original publication of Lu et al. (2005), the main distinctions being that the **single-step maxT procedure** takes into account the **joint distribution** of the test statistics and the **logistic regression model** allows adjusting for the **confounding variable** tissue type when comparing expression measures between cancerous and non-cancerous tissues.

Five of the **highly significant miRNAs** are located in minimal deleted regions, minimal amplified regions, and breakpoint regions **involved in human cancers** (Calin et al., 2004). Specifically, hsa – let – 7d and hsa – miR – 23b have been associated with urothelial cancer; hsa – miR – 22 with hepatocellular cancer; hsa – miR – 99a with lung cancer; hsa – miR – 100 with breast, cervical, lung, and ovarian cancers.

It would be of interest, as a follow-up analysis, to examine the **target sequences** of the differentially expressed miRNAs for the potential identification of common **motifs**.

Cancer microRNA Dataset: Co-Expression

Parameters of interest. $M = J(J - 1)/2 = 155 \times 154/2 = 11,935$ correlation coefficients for the expression measures of distinct pairs (j, j') of miRNAs,

$$\rho(j, j') = \text{Cor}[X(j), X(j')], \quad j = 1, \dots, J - 1, \quad j' = j + 1, \dots, J.$$

Null hypotheses. Consider two-sided tests of the M null hypotheses of no correlation in expression measures between pairs of miRNAs,

$$H_0(j, j') = \text{I}(\rho(j, j') = 0) \quad \text{vs.} \quad H_1(j, j') = \text{I}(\rho(j, j') \neq 0).$$

Cancer microRNA Dataset: Co-Expression

Table 4: *Cancer miRNA dataset, co-expression: Tests for correlation coefficients.* Number of co-expressed miRNA pairs (out of $M = 11,935$) for different nominal FWER levels α .

Nominal FWER, α	Number of miRNA pairs, R_n
0.05	8,916 (75%)
0.01	7,479 (63%)

Correlation coefficients found to be significantly different from zero at nominal FWER level $\alpha = 0.05$ range from 0.26 to 0.99, with median value 0.55.

Only 8% of all pairwise correlation coefficients are negative and none significantly so.

Cancer microRNA Dataset: Co-Expression

Test statistics. Difference statistics,

$$T_n(j, j') = \sqrt{n}(\rho_n(j, j') - \rho_0(j, j')),$$

where the null values $\rho_0(j, j')$ are zero and $\rho_n(j, j')$ are empirical correlation coefficients.

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

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

Cancer microRNA Dataset: Co-Expression

Table 5: *Cancer miRNA dataset, co-expression: Tests for correlation coefficients.* Twenty most significantly co-expressed pairs of miRNAs (bootstrap-based single-step maxT MTP).

Names		Correlation coefficient
hsa - miR - 106a#	hsa - miR - 17 - 5p#	0.99
mmu - miR - 200b	hsa - miR - 200b	0.99
mmu - miR - 200b	hsa - miR - 200c	0.99
hsa - miR - 107†	hsa - miR - 103	0.99
hsa - miR - 200b	hsa - miR - 200c	0.99
hsa - miR - 145‡	hsa - miR - 143‡	0.98
hsa - miR - 199a_(sub.1)	mmu - miR - 199b	0.98
hsa - miR - 17 - 5p	hsa - miR - 20_(sub.1)	0.97
hsa - miR - 19a#	hsa - miR - 19b#	0.97
hsa - miR - 29a	hsa - miR - 30a*	0.97
hsa - miR - 181a	hsa - miR - 181c	0.97
hsa - miR - 199a_(sub.1)	hsa - miR - 199a*	0.97
hsa - miR - 29b_(sub.2)	hsa - miR - 29c	0.97

Continued on next page ...

... continued from previous page

Names		Correlation coefficient
hsa – miR – 199a*	mmu – miR – 199b	0.96
hsa – miR – 200a	hsa – miR – 141	0.96
hsa – miR – 20_(sub.1) ‡	mmu – miR – 106a	0.96
hsa – miR – 106a	hsa – miR – 20_(sub.1) ‡	0.96
hsa – miR – 200a	hsa – miR – 200a	0.96
hsa – miR – 23b	hsa – miR – 23a	0.96
hsa – miR – 10a	hsa – miR – 10b	0.96

Several pairs are composed of miRNAs in the same family (e.g., hsa – miR – 10a and hsa – miR – 10b).

Up-regulated by the proto-oncogene c-MYC (O'Donnell et al., 2005).

† Increases cell growth in lung carcinomas (Cheng et al., 2005).

‡ Expressed at lower levels in cancerous and pre-cancerous tissues compared to normal colon tissues (Michael et al., 2003).

Cancer microRNA Dataset: Co-Expression

Several of the identified pairs are composed of miRNAs in the **same family** (e.g., hsa – miR – 10a and hsa – miR – 10b). The two most significantly correlated miRNAs are a pair of **paralogs**, hsa – miR – 17 – 5p (chromosome 17) and hsa – miR – 106a (chromosome X), which belong to **miRNA clusters** believed to be **up-regulated by the proto-oncogene c-MYC** (O'Donnell et al., 2005). hsa – miR – 19a, hsa – miR – 19b, and hsa – miR – 20 are also members of these paralogous miRNA clusters.

hsa – miR – 107 has been shown to **increase cell growth in lung carcinomas** (Cheng et al., 2005).

hsa – miR – 143 and hsa – miR – 145, located within 1.7 kb on human chromosome 5, were found to be **expressed at lower levels in cancerous and pre-cancerous tissues** compared to normal colon tissues (Michael et al., 2003).

It would be of interest to investigate the biological and medical implications of the identified **clusters of co-expressed miRNAs**.

Tests of Association with Biological Annotation Metadata

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

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

Tests of Association with Biological Annotation Metadata

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.

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**.

Tests of Association with Biological Annotation Metadata

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

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

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

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

Tests of Association with Biological Annotation Metadata

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

E.g. Gene Ontology (GO, www.geneontology.org) annotation. Gene pathway membership (e.g., Kyoto Encyclopedia of Genes and Genomes, KEGG, www.genome.ad.jp/kegg).

Gene regulation by particular transcription factors, presence or absence of certain motifs in gene control region (e.g., Transcription Factor DataBase, TRANSFAC, www.gene-regulation.com).

Exon/intron counts/lengths/nucleotide distributions.

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

Tests of Association with Biological Annotation Metadata

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

Let $X = (X(j) : j = 1, \dots, J) \sim P \in \mathcal{M}$ denote a J -dimensional random vector, with data generating distribution P belonging to a (possibly non-parametric) model \mathcal{M} . Suppose X contains G **gene-specific random variables** $(X(g) : g = 1, \dots, G)$.

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

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

Tests of Association with Biological Annotation Metadata

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

E.g. Vector of genome-wide mean transcript levels in a population of heat-shocked yeast cells.

Vector of regression coefficients relating survival to genome-wide transcript levels or DNA copy numbers in a population of cancer patients.

Tests of Association with Biological Annotation Metadata

The **association parameter** of interest is an M -vector

$$\psi = (\psi(m) : m = 1, \dots, M) = \rho(A, \lambda),$$

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

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

$$\psi(m) = \rho_m(A(\cdot, m), \lambda),$$

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

Tests of Association with Biological Annotation Metadata

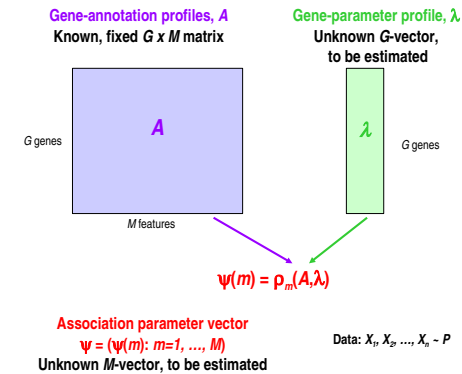


Figure 9: *Parameters for tests of association with biological annotation metadata.*

Tests of Association with Biological Annotation Metadata

As detailed in Dudoit and van der Laan (2007, Chapter 10) and Dudoit et al. (2007), this **general formulation** allows us to apply the **multiple hypothesis testing methodology** developed in Dudoit and van der Laan (2007) and related articles, to control a broad class of Type I error rates, defined as generalized tail probabilities (gTP), $gTP(q, g) = \Pr(g(V_n, R_n) > q)$, and generalized expected values (gEV), $gEV(g) = E[g(V_n, R_n)]$, for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n .

Unlike previously-proposed methods, these procedures take into account the **joint** distribution of the test statistics and provide Type I error control for **general data generating distributions** (with arbitrary dependence structures among variables), null hypotheses, and test statistics.

Tests of Association with Biological Annotation Metadata

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

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

Tests of Association with Biological Annotation Metadata

General gene-parameter profiles. Existing approaches typically consider binary gene-parameter profiles, e.g., vectors of indicators of differential expression. Our general definition of gene-parameter profiles allows consideration of a much broader class of testing problems, concerning arbitrary qualitative and quantitative parameters, such as, differences in mean expression levels or regression coefficients relating expression levels to clinical outcomes.

Estimated gene-parameter profiles. Existing approaches typically assume known gene-parameter profiles. For example, the list of 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 provides a more rigorous and general formulation of the statistical question.

Tests of Association with Biological Annotation Metadata

General tests of association. Common approaches to tests of association with GO annotation are typically limited to tests of independence in 2×2 contingency tables (e.g., based on the hypergeometric distribution, Fisher's exact test). Rows correspond to gene annotation with a given GO term (fixed binary gene-annotation profile) and columns to a gene property of interest, such as differential expression (treated as a fixed binary gene-parameter profile). Our approach allows consideration of a broader class of biological testing problems, while properly accounting for the fact that gene-parameter profiles are usually unknown and replaced by a random (i.e., data-driven) estimator.

Tests of Association with Biological Annotation Metadata

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

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

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

Tests of Association with Biological Annotation Metadata

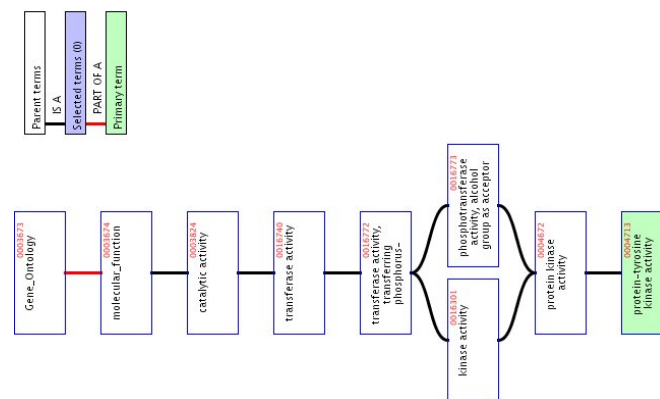


Figure 10: *Gene Ontology*. Portion of the DAG for the GO term *protein-tyrosine kinase activity* (GO:0004713), in the Molecular Function ontology. EBI QuickGO browser (www.ebi.ac.uk/ego).

Tests of Association with Biological Annotation Metadata

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 with each GO term m ,

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

$$g = 1, \dots, G, m = 1, \dots, M.$$

Detecting associations between GO annotation and other interesting features of a genome may be viewed as the multiple test of the null hypotheses of no **association** between **gene-annotation profiles** $A(\cdot, m)$ and a **gene-parameter profile** λ .

HIV-1 Sequence Variation and Viral Replication Capacity

The **protease** (PR) enzyme affects the reproductive cycle of the virus by breaking protein peptide bonds during replication.

The **reverse transcriptase** (RT) enzyme synthesizes double-stranded DNA from the virus' single-stranded RNA genome, thereby facilitating integration into the host's chromosome.

Because the **PR and RT regions** are **essential to viral replication**, many **antiretrovirals** (protease inhibitors and reverse transcriptase inhibitors) have been developed to target these specific genomic locations.

HIV-1 Sequence Variation and Viral Replication Capacity

Segal et al. (2004). HIV-1 dataset. Studying **genomic sequence variation** for the **human immunodeficiency virus type 1 (HIV-1)** could potentially give important insight into **genotype-phenotype associations** for the **acquired immune deficiency syndrome (AIDS)**.

- **Phenotype.** **Replication capacity (RC)** of HIV-1, which reflects the severity of the disease.
- **Genotypes.** **Codons/amino acids** in the protease and reverse transcriptase regions of the viral strand.

Goal. Relate HIV-1 protein sequence variation to viral replication capacity.

HIV-1 Sequence Variation and Viral Replication Capacity

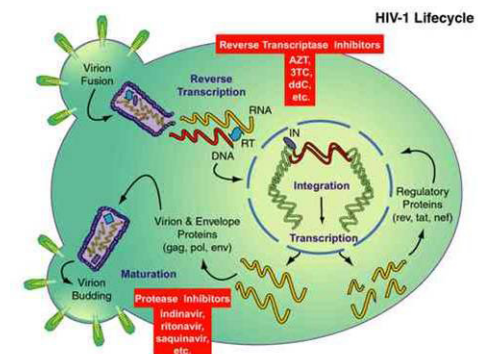


Figure 11: *HIV-1 lifecycle*. Diagram of the HIV-1 lifecycle and modes of action of protease and reverse transcriptase inhibitors.

HIV-1 Sequence Variation and Viral Replication Capacity

Data. The HIV-1 dataset comprises $n = 317$ records, linking viral replication capacity with PR and RT sequence data, from individuals participating in studies at the San Francisco General Hospital and the Gladstone Institute of Virology and Immunology. The data for each of the $n = 317$ patients consist of the following.

- Y , a continuous replication capacity outcome/phenotype.
- $X = (X(m) : m = 1, \dots, M)$, an $M = 96 + 186 = 282$ -dimensional covariate vector of binary codon genotypes in the PR (pr4–pr99) and RT (rt38–rt223) HIV-1 regions. Codons are recoded as binary covariates, with value of 0 corresponding to the wild-type codon, i.e., the most common codon among the $n = 317$ patients, and value of 1 for mutant codons, i.e., all other codons.

HIV-1 Sequence Variation and Viral Replication Capacity

Multiple testing question. Test for each of the $M = 282$ codon positions whether viral replication capacity Y is associated with the corresponding binary codon genotype $X(m) \in \{0, 1\}$.

Parameters of interest. For the m th codon position, the parameter of interest is the difference in mean replication capacity $\psi(m)$ for viruses with mutant and wild-type codons, that is,

$$\psi(m) = E[Y|X(m) = 1] - E[Y|X(m) = 0], \quad m = 1, \dots, M.$$

Null hypotheses. Consider two-sided tests of the M null hypotheses of no differences in mean RC vs. the alternative hypotheses of different mean RC,

$$H_0(m) = I(\psi(m) = 0) \quad \text{vs.} \quad H_1(m) = I(\psi(m) \neq 0).$$

HIV-1 Sequence Variation and Viral Replication Capacity

Test statistics. Two-sample pooled-variance t -statistics,

$$T_n(m) = \frac{\psi_n(m) - \psi_0(m)}{\sigma_n(m)} = \frac{\bar{Y}_{1,n}(m) - \bar{Y}_{0,n}(m) - 0}{\sigma_{p,n}(m) \sqrt{\frac{1}{n_0(m)} + \frac{1}{n_1(m)}}},$$

$$\sigma_{p,n}^2(m) = \frac{(n_0(m) - 1)\sigma_{0,n}^2(m) + (n_1(m) - 1)\sigma_{1,n}^2(m)}{n_0(m) + n_1(m) - 2},$$

where the null values $\psi_0(m)$ are zero, $n_k(m) = \sum_i I(X_i(m) = k)$ denotes the number of patients with codon genotype $X(m) = k \in \{0, 1\}$ at position m , and $\bar{Y}_{k,n}(m) = \sum_i I(X_i(m) = k) Y_i / n_k(m)$ and $\sigma_{k,n}^2(m) = \sum_i I(X_i(m) = k) (Y_i - \bar{Y}_{k,n}(m))^2 / (n_k(m) - 1)$ denote, respectively, the sample means and sample variances for the RC of patients with codon genotype $X(m) = k \in \{0, 1\}$ at position m . The pooled-variance estimators are denoted by $\sigma_{p,n}^2(m)$.

HIV-1 Sequence Variation and Viral Replication Capacity

Test statistics null distribution. Non-parametric bootstrap estimator of the null shift and scale-transformed test statistics null distribution, $B = 7,500$ samples.

Multiple testing procedures.

1. **FWER-controlling single-step maxT** procedure (SS maxT).
2. **$gFWER(k)$ -controlling augmentation** procedure, based on FWER-controlling single-step maxT procedure, for an allowed number $k \in \{10, 50\}$ of false positives ($gFWER(k)$ AMTP).
3. **$TPFP(q)$ -controlling augmentation** procedure, based on FWER-controlling single-step maxT procedure, for an allowed proportion $q \in \{0.10, 0.20, 0.50\}$ of false positives ($TPFP(q)$ AMTP).

HIV-1 Sequence Variation and Viral Replication Capacity

Table 6: *HIV-1 dataset*. t -statistics and sorted adjusted p -values for FWER-controlling single-step maxT procedure, $gFWER(k)$ -controlling augmentation procedure ($k = 5$), and $TPFP(q)$ -controlling augmentation procedure ($q = 0.10$).

Codon position	t -statistic	Adjusted p -values		
		SS maxT	$gFWER(k)$ AMTP	$TPFP(q)$ AMTP
pr32	-9.755	0.0001	0	0.0001
pr47	-9.579	0.0013	0	0.0013
pr34	-8.843	0.0087	0	0.0087
pr55	-8.150	0.0104	0	0.0104
pr90	-6.237	0.0396	0	0.0396
rt184	-6.162	0.0431	0.0001	0.0431
pr43	-6.118	<u>0.0444</u>	0.0013	<u>0.0444</u>
pr54	-5.539	0.0780	0.0087	0.0780
rt41	-5.225	0.0978	0.0104	0.0978
pr46	-5.224	0.0980	0.0396	0.0978
pr82	-4.521	0.1678	0.0431	0.0980
rt215	-4.479	0.1740	<u>0.0444</u>	0.1678
rt121	-4.070	0.2380	0.0780	0.1740

HIV-1 Sequence Variation and Viral Replication Capacity

The 13 codon positions with the smallest adjusted p -values all have negative t -statistics, suggesting that **mutant codons** (recoded as 1) are associated with **decreased viral replication capacity**.

The specific mutations observed in the present study are consistent with those found in the literature.

Protease: Vpr32I, Mpr46I, Ipr54V/L/T, Vpr82A/T/F/S, and Lpr90M increase the resistance of HIV-1 to various protease inhibitors.

Reverse transcriptase: Mrt41L, Mrt184V/I, and Trt215Y/F are related to azidothymidine (AZT) resistance.

HIV-1 Sequence Variation and Viral Replication Capacity

HIV: M=282 Codon Positions.

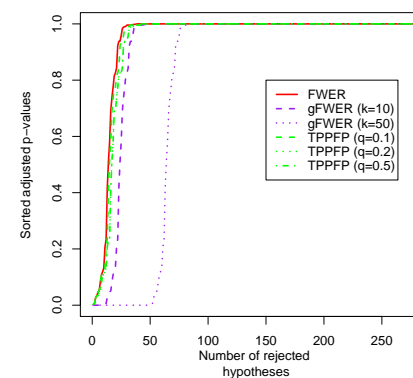


Figure 12: *HIV-1 dataset*. Sorted adjusted p -values for FWER-controlling single-step maxT procedure, $gFWER(k)$ -controlling augmentation procedure ($k \in \{10, 50\}$), and $TPFP(q)$ -controlling augmentation procedure ($q \in \{0.10, 0.20, 0.50\}$).

Genetic Mapping

A central question in **genetic mapping** is to relate **genotypes** at multiple genetic markers to **phenotypes**, i.e., relate DNA variation to biological and clinical outcomes.

In recent years, **single nucleotide polymorphisms** (SNP) have become the genetic markers of choice in genome-wide and candidate gene/pathway approaches for the genetic mapping of **complex traits**, such as, diabetes, multiple sclerosis, and obesity.

The biological question of **detecting genotype-phenotype associations** can be restated as a **multiple hypothesis testing** problem: the simultaneous test of the null hypotheses of no association between multilocus genotypes and one or more (possibly censored, qualitative or quantitative) phenotypes.

Genetic Mapping: The ObeLinks Project

Birkner et al. (2007). ObeLinks dataset. The goal of the ObeLinks Project is to identify genotypic combinations associated with human obesity (www.obelinks.org).

Obesity is a multifactorial disease, which, by definition, could be caused or influenced by multiple genetic and environmental factors, via gene-gene and/or gene-environment interactions.

The ObeLinks Project provides data on $n = 386$ morbidly obese patients, genotyped at 22 SNPs and for which 29 obesity-related phenotypes were measured (e.g., body mass index, glycemia, and insulinemia).

Genetic Mapping: The ObeLinks Project

The SNP genotype data are unphased, i.e., the parental origins of the two alleles at a given locus are generally unknown.

The most frequent SNP allele is denoted by 0 and referred to as the wild-type or major allele; the least frequent allele is denoted by 1 and referred to as the mutant or minor allele.

SNP genotypes are coded as (wt = 00) for the homozygous wild-type genotype, (ht = 01) for the heterozygous genotype, and (hm = 11) for the homozygous mutant genotype.

Genetic Mapping: The ObeLinks Project

The 22 SNPs are classified into three SNP sets, based on pathway membership and potential significance for obesity.

- **OB – IR:** 6 SNPs located in the coding or non-coding sequence of a gene for a protein involved in insulin signaling or the pathophysiology of insulin resistance (e.g., lipid transport).
- **OB – Signaling:** 8 SNPs located in the coding or non-coding sequence of a gene for an adiposity signal protein or its receptor or, more generally, for a protein involved in signal message transmission.
- **OB – ThermoG:** 8 SNPs located in the coding or non-coding sequence of a gene related to the thermogenesis process.

Note that, although biologically-founded, this classification is somewhat subjective, as SNPs could in principle be involved in any combination of the aforementioned three processes.

Furthermore, different SNPs may belong to the same gene.

Genetic Mapping: The ObeLinks Project

Three penetrance models are considered for the genotype-phenotype association.

- **Codominance model.** The three genotypes, (wt = 00), (ht = 01), and (hm = 11), are treated as having distinct effects on the phenotype, i.e., distinct penetrances.
- **Dominance model.** The model only distinguishes between two genotypes, (wt = 00) and (ht = 01 or hm = 11), i.e., (ht = 01) and (hm = 11) are assumed to have the same effect on the phenotype.
- **Recessive model.** The model only distinguishes between two genotypes, (wt = 00 or ht = 01) and (hm = 11), i.e., (wt = 00) and (ht = 01) are assumed to have the same effect on the phenotype.

Genetic Mapping: The ObeLinks Project

To assist in the detection of **gene interaction effects on phenotypes**, single-locus SNP genotypes are recoded into **multilocus composite genotypes** using **Galois lattices**.

The **genotype data structure** implied by the Galois lattice for a given genotype set is an $n \times M$ **binary matrix**, with rows corresponding to the n patients and columns to the M nodes in the lattice.

The (i, m) th entry of this matrix indicates whether patient i has the multilocus composite genotype corresponding to node m .

Genetic Mapping: The ObeLinks Project

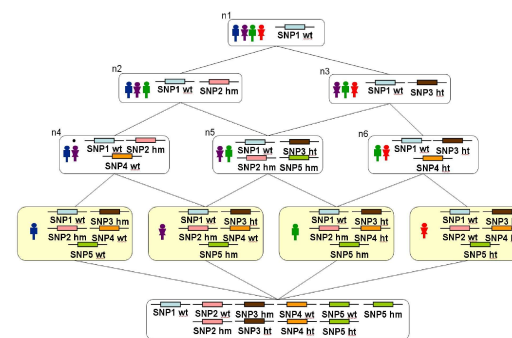


Figure 13: *Galois lattice for SNP genotypes*. The Hasse diagram represents the Galois lattice for a simple artificial example with $n = 4$ patients genotyped at 5 SNPs.

Genetic Mapping: The ObeLinks Project

Table 7: *Galois lattice for SNP genotypes*. The table represents the formal context $(\mathcal{O}, \mathcal{D}, \mathcal{I})$ for a simple artificial example with $n = 4$ patients genotyped at 5 SNPs.

Binary relation, \mathcal{I}	Objects, \mathcal{O}			
	o_1	o_2	o_3	o_4
(SNP1 = wt)	1	1	1	1
(SNP1 = ht)	0	0	0	0
(SNP1 = hm)	0	0	0	0
(SNP2 = wt)	0	0	0	1
(SNP2 = ht)	0	0	0	0
(SNP2 = hm)	1	1	1	0
(SNP3 = wt)	0	0	0	0
(SNP3 = ht)	0	1	1	1
(SNP3 = hm)	1	0	0	0
(SNP4 = wt)	1	1	0	0
(SNP4 = ht)	0	0	1	1
(SNP4 = hm)	0	0	0	0
(SNP5 = wt)	1	0	0	0
(SNP5 = ht)	0	0	0	1
(SNP5 = hm)	0	1	1	0

Genetic Mapping: The ObeLinks Project

Table 8: *ObeLinks dataset: Galois lattices for SNP genotype sets*.

SNP set	SNP genotype set	Penetrance model	Number of nodes, M
OB – IR	Codominant	Dominant	428
		Dominant	282
		Recessive	156
OB – Signaling	Codominant	Dominant	356
		Dominant	665
		Recessive	862
OB – ThermoG	Codominant	Dominant	9367
		Dominant	8113
		Recessive	3486

Genetic Mapping: The ObeLinks Project

Each of the **three sets of SNPs** (namely, OB – IR, OB – Signaling, and OB – ThermoG) is analyzed separately, under each of the **three penetrance models**, thus yielding **nine sets of SNP genotypes** and their corresponding **Galois lattices**.

For a given genotype set and obesity-related phenotype, one is interested in testing, for each node in the Galois lattice, the null hypothesis of no **association** between the **multilocus composite SNP genotype** and the **phenotype**.

Genetic Mapping: The ObeLinks Project

Data. The data for each of the $n = 386$ patients consist of the following.

- Y , one of 29 qualitative or quantitative obesity-related **outcomes/phenotypes**.
- $X = (X(m) : m = 1, \dots, M)$, an M -dimensional vector of binary **covariates/genotypes** for one of the nine SNP genotype sets,

$$X(m) = \begin{cases} 1, & \text{if patient has multilocus composite SNP genotype} \\ & \text{corresponding to } m\text{th node in Galois lattice} \\ 0, & \text{otherwise} \end{cases}$$

Genetic Mapping: The ObeLinks Project

Multiple testing question. For a given SNP genotype set and phenotype, identify genes associated with the phenotype by testing, for each of the M nodes in the Galois lattice, the null hypothesis of no **difference in mean phenotype** for the corresponding multilocus composite SNP genotypes.

Parameters of interest. The parameter of interest for the m th node is the **difference in mean phenotype** $\psi(m)$ for individuals possessing ($X(m) = 1$)/not possessing ($X(m) = 0$) the multilocus composite SNP genotype corresponding to node m . That is,

$$\psi(m) = E[Y|X(m) = 1] - E[Y|X(m) = 0], \quad m = 1, \dots, M.$$

Null hypotheses. Consider one-sided tests of the M null hypotheses of no elevated phenotype vs. the alternative hypotheses of elevated phenotype for individuals with genotypes corresponding to each node in the Galois lattice,

$$H_0(m) = I(\psi(m) \leq 0) \quad \text{vs.} \quad H_1(m) = I(\psi(m) > 0).$$

Genetic Mapping: The ObeLinks Project

Test statistics. **Two-sample pooled-variance t -statistics**,

$$T_n(m) = \frac{\psi_n(m) - \psi_0(m)}{\sigma_n(m)} = \frac{\bar{Y}_{1,n}(m) - \bar{Y}_{0,n}(m) - 0}{\sigma_{p,n}(m) \sqrt{\frac{1}{n_0(m)} + \frac{1}{n_1(m)}}},$$

$$\sigma_{p,n}^2(m) = \frac{(n_0(m) - 1)\sigma_{0,n}^2(m) + (n_1(m) - 1)\sigma_{1,n}^2(m)}{n_0(m) + n_1(m) - 2},$$

where the null values $\psi_0(m)$ are zero, $n_k(m) = \sum_i I(X_i(m) = k)$ denotes the number of patients possessing ($k = 1$)/not possessing ($k = 0$) the multilocus composite SNP genotype corresponding to node m , and $\bar{Y}_{k,n}(m) = \sum_i I(X_i(m) = k) Y_i / n_k(m)$ and $\sigma_{k,n}^2(m) = \sum_i I(X_i(m) = k) (Y_i - \bar{Y}_{k,n}(m))^2 / (n_k(m) - 1)$ denote, respectively, the sample means and sample variances for the phenotype of individuals possessing ($k = 1$)/not possessing ($k = 0$) the m th genotype. The pooled-variance estimators are denoted by $\sigma_{p,n}^2(m)$.

Genetic Mapping: The ObeLinks Project

Test statistics null distribution. Non-parametric bootstrap estimator of the null shift and scale-transformed test statistics null distribution, $B = 7,500$ samples.

Multiple testing procedures.

1. FWER-controlling single-step maxT procedure (SS maxT).
2. $gFWER(k)$ -controlling augmentation procedure, based on FWER-controlling single-step maxT procedure, for an allowed number $k \in \{10, 50\}$ of false positives ($gFWER(k)$ AMTP).
3. $TPFP(q)$ -controlling augmentation procedure, based on FWER-controlling single-step maxT procedure, for an allowed proportion $q \in \{0.10, 0.20, 0.50\}$ of false positives (TPFP(q) AMTP).
4. FDR-controlling step-up Benjamini and Hochberg (1995) procedure based on bootstrap unadjusted p -values (SU BH).

Genetic Mapping: The ObeLinks Project

Table 9: *ObeLinks* dataset: Phenotype distributions. Six-number summaries of the distributions of body mass index (BMI), glycemia, and insulinemia, for $n = 386$ patients participating in the ObeLinks Project.

	BMI	Glycemia	Insulinemia
Minimum	40.00	2.880	4.00
1st quartile	42.64	5.380	12.60
Median	45.14	6.050	17.60
Mean	47.55	7.057	20.62
3rd quartile	49.60	7.242	25.60
Maximum	87.24	21.100	70.20
# NA's	1.00	4.00	37.00

Genetic Mapping: The ObeLinks Project

Body mass index. Body mass index (BMI) is a simple function of height (in meters, m) and weight (in kilograms, kg), defined as

$$BMI = \frac{\text{Weight}}{\text{Height}^2} \text{ kg/m}^2.$$

This index is not gender-specific and represents a commonly-used indirect measure of fat mass, overweightedness, and obesity.

Individuals are classified as slim, normal, overweight, obese, and morbidly obese, if, respectively, $BMI < 18$, $BMI \in [18, 25)$, $BMI \in [25, 30)$, $BMI \in [30, 40)$, and $BMI \geq 40$.

Patients considered in the ObeLinks Project are morbidly obese, with BMI ranging from 40.00 to 87.24 and with an average BMI value of 47.55,

Genetic Mapping: The ObeLinks Project

Table 10: *ObeLinks* dataset: BMI, OB – IRCodominant SNP genotype set.

Node ID	Coverage	Adjusted p -values	
		SS maxT	SU BH
n104	78	0.0119	0.0214
(ENPP1 = wt, ABCC8, IVS15 – 3TC = hm)			
n105	63	0.0129	0.0214
(IRS1 = wt, ENPP1 = wt, ABCC8, IVS15 – 3TC = hm)			
n121	72	0.0299	0.0428
(ENPP1 = wt, ABCC8, Thr759Thr = wt, ABCC8, IVS15 – 3TC = hm)			
n126	30	0.0299	0.0428
(FABP2 = hm, IRS1 = wt, ENPP1 = wt, ABCC8, Thr759Thr = wt, ABCC8, IVS15 – 3TC = hm)			
n123	57	0.0379	0.0428
(IRS1 = wt, ENPP1 = wt, ABCC8, Thr759Thr = wt, ABCC8, IVS15 – 3TC = hm)			
n92	40	<u>0.0499</u>	0.0476
(ENPP1 = wt, PPI = wt, ABCC8, IVS15 – 3TC = hm)			
n102	90	0.0549	0.0476
(ABCC8, IVS15 – 3TC = hm)			
n125	38	0.0609	0.0476
(FABP2 = hm, ENPP1 = wt, ABCC8, Thr759Thr = wt, ABCC8, IVS15 – 3TC = hm)			
n47	35	0.0709	<u>0.0476</u>
(IRS1 = wt, ENPP1 = wt, PPI = wt, ABCC8, IVS15 – 3TC = hm)			
n103	75	0.0789	0.0514
(IRS1 = wt, ABCC8, IVS15 – 3TC = hm)			

Genetic Mapping: The ObeLinks Project

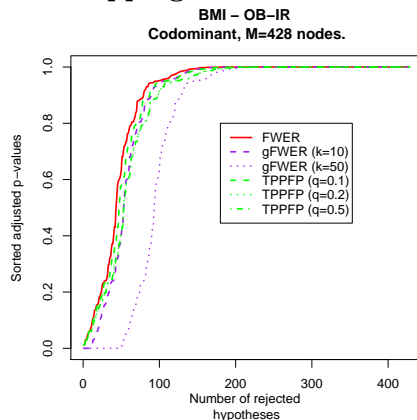


Figure 14: *ObeLinks* dataset: BMI, OB – IRCodominant SNP genotype set. Sorted adjusted p -values for FWER-controlling single-step maxT procedure, $gFWER(k)$ -controlling augmentation procedure ($k \in \{10, 50\}$), and $TPPFP(q)$ -controlling augmentation procedure ($q \in \{0.10, 0.20, 0.50\}$).

Genetic Mapping: The ObeLinks Project

BMI phenotype, OB – IRCodominant SNP genotype set. For a nominal Type I error level $\alpha = 0.05$, the FWER-controlling single-step maxT and FDR-controlling step-up Benjamini and Hochberg (1995) procedures identify, respectively, 6 and 9 nodes (out of $M = 428$) as significantly associated with BMI.

The identified SNPs tend to describe neighboring nodes of the Hass diagram for the Galois lattice, thereby suggesting that the corresponding genes belong to the same biological pathway.

Nodes n126 and n125 correspond to multilocus composite SNP genotypes with mutations in two genes known to be associated with obesity-related phenotypes: FABP2 homozygous mutant and ABCC8, IVS15 – 3TC homozygous mutant.

These results suggest a possible novel interaction between the two genes in their effect on BMI and glycemia.

Genetic Mapping: The ObeLinks Project

FABP2. The FABP2 gene codes for an intra-cellular fatty acid binding protein (FABP).

FABPs are divided into at least three distinct types: hepatic, intestinal, and cardiac FABPs. They form 14–15 kiloDalton (kDa) proteins and are thought to participate in the uptake, intra-cellular metabolism, and/or transport of long-chain fatty acids. FABPs may also be involved in the modulation of cell growth and proliferation.

The intestinal fatty acid binding protein 2 (FABP2) gene contains 4 exons and its product is an abundant cytosolic protein in epithelial cells of the small intestine. This gene has a common polymorphism at codon 54 that corresponds to an alanine-encoding allele and a threonine-encoding allele. The Thr-54 protein is associated with increased fat oxidation and insulin resistance.

Genetic Mapping: The ObeLinks Project

ABCC8, IVS15 – 3TC. The protein encoded by the ABCC8 gene is a member of the super-family of ATP-binding cassette (ABC) transporters. ABC proteins are divided into 7 distinct sub-families (ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, White) and transport various molecules across extra- and intra-cellular membranes.

The ABCC8 protein product is a member of the MRP sub-family, which is involved in multidrug resistance. This protein functions as a modulator of ATP-sensitive potassium channels and insulin release.

Genetic Mapping: The ObeLinks Project

Mutations and deficiencies in the **ABCC8** protein product have been observed in patients with **hyperinsulinemic hypoglycemia of infancy**, an autosomal recessive disorder of unregulated and high insulin secretion.

Mutations have also been associated with **non-insulin-dependent diabetes mellitus** (NIDDM), or **type 2 diabetes**, an autosomal dominant disease of defective insulin secretion.

The gene alias **SUR1** encodes the sulfonylurea receptor-1, a pancreatic regulatory subunit, which binds a widely-used class of **insulin-secreting drugs** and which has been associated with **hyperinsulinemia**.

Above descriptions based on Entrez Gene database
(www.ncbi.nlm.nih.gov/entrez).

Genetic Mapping: The ObeLinks Project

To date, mutations **FABP2 = hm** and **ABCC8, IVS15 – 3TC = hm** have only been studied singly in terms of their association with obesity-related phenotypes such as BMI.

Ongoing research efforts include investigating the **interaction between the two genes FABP2 and ABCC8**, as they both seem to have an important relationship with insulin sensitivity.

Software Implementation: Bioconductor R Package multtest

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

Please consult the package documentation (e.g., helpfiles, manuals) and the book chapters by Dudoit and van der Laan (2007, Section 13.1) and Pollard et al. (2005) for details.

Bioconductor R package: **multtest**.

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

URL: www.bioconductor.org.

Software Implementation: Bioconductor R Package multtest

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

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

Test statistics null distribution. Bootstrap null shift and scale-transformed; permutation (Chapter 2 in Dudoit and van der Laan, 2007).

Software Implementation: Bioconductor R Package `multtest`

Multiple testing procedures.

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

Software Implementation: SAS Macros

SAS macros are available to compute the following components of a MTP (Birkner et al., 2005):

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

Author: M. D. Birkner.

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

Software Implementation: Bioconductor R Package `multtest`

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

References

- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.
- Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.
- M. D. Birkner, K. S. Pollard, M. J. van der Laan, and S. Dudoit. Multiple testing procedures and applications to genomics. Technical Report 168, Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720-7360, 2005. URL www.bepress.com/ucbbiostat/paper168.
- M. D. Birkner, M. Courtine, M. J. van der Laan, K. Clément, J.-D. Zucker, and S. Dudoit. Statistical methods for detecting genotype/phenotype associations in the ObeLinks Project. Technical report, Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720-7360, 2007. (In preparation).

- C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, pages 3–62, 1936.
- G. A. Calin, C. Sevignani, C. D. Dumitru, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizu, S. Rattan, F. Bullrich, M. Negrini, and C. M. Croce. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci.*, 101(9):2999–3004, 2004.
- M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research*, 10(12):2022–2029, 2000.
- A. M. Cheng, M. W. Byrom, J. Shelton, and L. P. Ford. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Research*, 33(4):1290–1297, 2005.
- S. Dudoit and M. J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer, New York, 2007. (In preparation).
- S. Dudoit and Y. H. Yang. Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In G. Parmigiani, E. S.

March 19, 2007

Page 101

- Garrett, R. A. Irizarry, and S. L. Zeger, editors, *The Analysis of Gene Expression Data: Methods and Software*, pages 73–101. Springer, New York, 2003.
- S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139, 2002.
- S. Dudoit, M. J. van der Laan, and K. S. Pollard. Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 13, 2004. URL www.bepress.com/sagmb/vol3/iss1/art13.
- S. Dudoit, S. Keleş, and M. J. van der Laan. Multiple tests of association with biological annotation metadata. In D. Nolan and T. P. Speed, editors, *A Festschrift for David Freedman*, Institute of Mathematical Statistics, Lecture Notes – Monograph Series. 2007. (To appear).
- Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802, 1988.
- S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

March 19, 2007

Page 102

- J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz, and T. R. Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435(9):834–838, 2005. URL www.broad.mit.edu/cancer/pub/miGCM.
- M. Z. Michael, S. M. O'Connor, N. G. van Holst Pellekaan, G. P. Young, and R. J. James. Reduced accumulation of specific microRNAs in colorectal neoplasia. *Molecular Cancer Research*, 1(12):882–891, 2003.
- K. A. O'Donnell, E. A. Wentzel, K. I. Zeller, C. V. Dang, and J. T. Mendell. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, 435(7043):839–843, 2005.
- K. S. Pollard and M. J. van der Laan. Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference*, 125(1–2):85–100, 2004.
- K. S. Pollard, S. Dudoit, and M. J. van der Laan. Multiple testing procedures: The `multtest` package and applications to genomics. In R. C. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 15,

March 19, 2007

Page 103

- pages 249–271. Springer, New York, 2005. URL www.bioconductor.org/pubs/docs/mogr, www.bepress.com/ucbbiostat/paper164.
- M. R. Segal, J. D. Barbour, and R. M. Grant. Relating HIV-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 2, 2004. URL www.bepress.com/sagmb/vol3/iss1/art2.
- M. Tsunoda, J. Tenhunen, C. Tilgmann, H. Arai, and K. Imai. Reduced membrane-bound catechol-O-methyltransferase in the liver of spontaneously hypertensive rats. *Hypertension Research*, 26(11):923–927, 2003.
- M. J. van der Laan, S. Dudoit, and K. S. Pollard. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 14, 2004a. URL www.bepress.com/sagmb/vol3/iss1/art14.
- M. J. van der Laan, S. Dudoit, and K. S. Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 15, 2004b. URL www.bepress.com/sagmb/vol3/iss1/art15.

March 19, 2007

Page 104

Y. H. Yang and A. C. Paquet. Preprocessing two-color spotted arrays. In R. C. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter 4, pages 49–69. Springer, New York, 2005. URL www.bioconductor.org/pubs/docs/mogr.

Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty, editors, *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proceedings of SPIE*, pages 141–152, Bellingham, WA, May 2001. SPIE-International Society for Optical Engineering.

Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, 11(1):108–136, 2002.