

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
16 December 2004 (16.12.2004)

PCT

(10) International Publication Number
WO 2004/109540 A2

(51) International Patent Classification⁷: **G06F 17/18**

Heights, OH 44143 (US). **ISHWARAN, Hemant**
[CA/US]; 22025 Rye Road, Shaker Heights, OH 44122
(US).

(21) International Application Number:
PCT/US2004/017504

(74) Agents: **VARMA, Anita** et al.; Ropes & Gray LLP, One
International Place, Boston, MA 02110-2624 (US).

(22) International Filing Date: 1 June 2004 (01.06.2004)

(25) Filing Language: English

(81) Designated States (*unless otherwise indicated, for every
kind of national protection available*): AE, AG, AL, AM,
AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN,
CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI,
GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE,
KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD,
MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG,
PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM,
TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM,
ZW.

(26) Publication Language: English

(30) Priority Data:
60/474,456 30 May 2003 (30.05.2003) US

(71) Applicants (*for all designated States except US*): **CASE
WESTERN RESERVE UNIVERSITY** [US/US]; 10900
Euclid Avenue, Cleveland, OH 44106-7219 (US). **THE
CLEVELAND CLINIC FOUNDATION** [US/US]; 9500
Euclid Ave., Cleveland, OH 44195 (US).

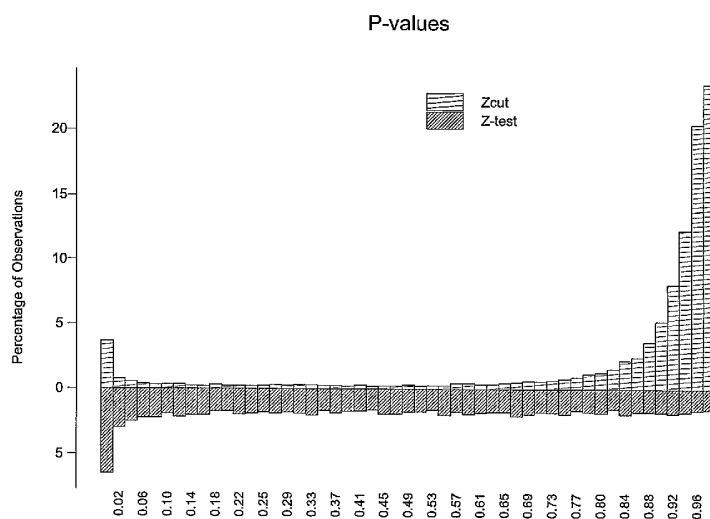
(84) Designated States (*unless otherwise indicated, for every
kind of regional protection available*): ARIPO (BW, GH,
GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM,
ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),
European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,
FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI,

(72) Inventors; and

(75) Inventors/Applicants (*for US only*): **RAO, Jonna-
gadda, Sunil** [US/US]; 474 Wordsworth Court, Richmond

[Continued on next page]

(54) Title: A STOCHASTIC VARIABLE SELECTION METHOD FOR MODEL SELECTION



(57) Abstract: A method of identifying differentially-expressed genes includes deriving an analysis of variance (ANOVA) or analysis of covariance (ANCOVA) model for expression data associated with a number of genes; from the ANOVA or ANCOVA model, deriving a linear regression model defined at least in part by an observation vector representative of an observed subset of the gene-expression data, a design matrix of regressor variables, a vector of regression coefficients representing gene contribution to the observation vector, and a measurement error vector; and to the linear regression model, applying a hierarchical selection algorithm to designate a subset of the regression coefficients as significant regression coefficients, the selection algorithm representing at least one of the observation vector, the design matrix, and the measurement error vector as being hierarchically dependent on parameters having predetermined probabilistic properties, wherein the designated subset corresponds to a respective subset of the genes identified as differentially expressed.

WO 2004/109540 A2



SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Published:

— *without international search report and to be republished upon receipt of that report*

TITLE: A STOCHASTIC VARIABLE SELECTION METHOD FOR MODEL SE-
LECTION

CROSS-REFERENCE TO RELATED APPLICATIONS

This application incorporates by reference in entirety, and claims priority
5 to and benefit of, U.S. Provisional Patent Application No. 60/474,456, filed
on 30 May 2003.

STATEMENT REGARDING FEDERAL FUNDING

Work described herein was funded, in part, by the National Institutes of
Health, Grant No. K25-CA89867. The United States government has cer-
tain rights in the invention.

10

BACKGROUND

The systems and methods described herein pertain to statistical data anal-
ysis in general, and to model selection in particular.

Multiple hypothesis testing—also known by model selection, data classi-
15 fication, multiple detection, and other descriptors—has become important
in an age of increasingly sophisticated data-gathering technologies. The
high data-throughput nature of instruments at our disposal not only cre-
ates an enormous wealth of information, but also poses a data-analysis
challenge, because of the large multiple testing or classification problems
20 involved.

A commonly-used approach has been to focus on one type of model se-

lection error at the expense of other types. A drawback of such a method is that more subtle features of the data are missed, features that might give analysts more insight into how to test various hypotheses on collected data.

5 SUMMARY OF THE INVENTION

It is desirable to devise a new method for performing model selection, data classification, significance testing, or hypothesis testing—especially in high-dimensional or large-scale data contexts: a new method that strikes a balance between tolerances for various error types.

- 10 In one aspect, the invention is a method of analyzing data, seeking to discern structure within the data that would then allow it to fit one or more possible models or hypotheses. Data being analyzed can be, among other things: mass spectroscopy data; gene expression data; protein expression data; genotyping data; metabolite production data; virus data; other
- 15 types of biomolecular data; data collected from a large-scale—and possibly intractable—system, network or array; or data collected from any system wherein a type of classification decision has to be made based on, among other things, information contained in the data, or possibly even based on external information.
- 20 Accordingly, the methods and systems described herein can access and process data by casting an analysis or classification problem into, for example, the context and language of analysis of variance (ANOVA) or analysis of covariance (ANCOVA). An ANOVA or ANCOVA formulation for data analysis may be recast as a regression problem comprising an ob-
- 25 servation vector, a design matrix comprising a plurality of regressors, a

vector of regression coefficients that are representative of the model, and a measurement error vector.

According to one practice, the systems and methods described herein apply a hierarchical model selection algorithm to classify the regression coefficients (which can serve as model parameters), into any number of groups, where the number of groups depends on the context. According to one exemplary embodiment, the classification, or model selection, problem involves only two groups. For such an embodiment, the selection algorithm seeks to identify the subset of the model parameters (equivalently, the subset of the regression coefficients) that are "significant." This is also known as binary hypothesis testing, consisting of a null hypothesis and an alternative hypothesis. An example of this occurs in the analysis of differential gene expression data, wherein data is collected from two samples of gene arrays—one serving as a control sample and the other a treatment sample. The control sample may represent healthy tissue, while the treatment sample may be, for example, representative of cancerous tissue. In this embodiment, a goal then is to identify the subset of the genes that express themselves differently (alternative hypothesis) in the two samples, thus giving the investigator a clue as to which genes are "significant" in the particular cancer represented by the treatment group. The null hypothesis in this context would be that a gene does not differentially express itself.

The systems and methods described herein also apply, in various embodiments, to data analysis contexts involving more than two groups; a goal therein includes classifying the model parameters into one of a plurality of possible models or groups. In a multigroup embodiment the groups indicate, for example, stages of disease, and interest lies in finding genes that differentially express in various patterns across the stages. For example,

genes that differentially express throughout a disease process such as cancer might be termed tumor progression genes and those that differentially express primarily during the early or primarily during the late stages may be termed early and late hit-and-run genes, respectively. In this setup, a particular stage (typically the most benign) acts as a baseline or reference from which stagewise differential gene expression is ascertained.

It is commonly known that statistical methods applied to classification problems involve various types of errors. In the case of two-way classification, for example, there are, among others, two types of errors: false discovery (otherwise known as false alarm) and false non-discovery (also known as a "miss" in certain contexts). The systems and methods described herein provide a mechanism by which a balance can be struck between tolerance for one type of error versus the other.

Prior art methods have focused primarily on keeping false discovery rates (that is, false alarm rates) in check. In the context of differential gene expression, this has led to falsely dismissing some genes as differentially inexpressive, even though they do contribute to the treatment sample.

The systems and methods described herein, however, strike a balance between two types of errors: false discovery and false non-discovery. By identifying a subset of the data having intermediate features that are less readily classifiable than other data points that have distinctly prominent features, the methods and systems described herein flag the analyst to pay closer attention to the intermediate data points, and use greater care in classifying them. In an embodiment involving differential gene expression data, for example, there are some genes whose expression levels are not as prominent, and hence they are not as clearly identifiable or classifiable,

as others with distinctly prominent differential gene expression features. These intermediate genes sometimes do associate with the treatment sample; yet, they are missed by current methods, which focus primarily on false discovery rates.

- 5 The systems and methods described herein remedy the problem by applying a hierarchical selection algorithm to the ANOVA- or ANCOVA-based regression formulation of the data, to classify the regression coefficients into the number of groups under consideration in any given context. The selection algorithm defines a hierarchy of variables (referred to also as hyperparameters) with appropriately chosen probabilistic properties, to represent the regression coefficients, the observation vector, the design matrix, the measurement error vector, or any combination thereof. The probabilistic properties of the hyperparameters are chosen in a way so as to properly model the number of classification groups in the context of interest.
- 10
- 15 The systems and methods described herein further may involve a Bayesian approach, for example in the selection algorithm. One example of a hierarchical Bayesian approach is the parametric stochastic variable selection (P-SVS) algorithm.

- The systems and methods described herein also may employ statistical techniques, such as appropriate scaling or mean subtraction, to decorrelate model parameters or simplify their respective, effective probabilistic complexities.
- 20

- Posterior statistics of the model parameters are used by the systems and methods described herein, to define test statistics that may be used to perform the classification of data. For example, from the posterior density,
- 25

the systems and methods disclosed herein may use the posterior mean of the regression coefficients (model parameters), conditioned on the observation vector, to test for significance of the model parameters (in the case two hypotheses) or to perform multi-group classification. In an alternative
5 embodiment, the conditional median can be used as a test statistic.

The systems and methods described herein highlight how the posterior mean vector of the model parameters is simply a ridge estimate from a generalized ridge regression of the observation vector on the design matrix. Furthermore, the posterior mean may comprise a weighted average
10 of shrunken posterior estimates of the model parameters.

Another distinctive feature of the systems and methods described herein is a graphical method that facilitates the proper identification and/or classification of the intermediate values of the data. By plotting the posterior variance against the posterior mean of the model parameters, the systems
15 and methods disclosed herein highlight a feature of the intermediate data points that may otherwise be missed by other methods known in the art. Those of ordinary skill in the art know, or can readily ascertain having read this specification, that variations of the plot just described are possible; for example, the posterior mean may be replaced by the posterior median.
20 Moreover, a generalization exists of the two-group BAM shrinkage plot extended to multiple groups. In one embodiment, a particular group is designated as a baseline, and individual two dimensional plots of pairwise BAM test statistic values (also called Bayes test statistics)—which are group effect test statistics against the designated baseline group—are
25 plotted. According to one practice, a number of two-dimensionals plots are generated based, at least in part, on the total number of groups under study. Such variations and/or generalizations do not depart from the

scope of the claimed subject matter.

Another feature of the systems and methods described herein is a post-processing of the data whereby the distribution of the posterior mean (or median) is approximated by a probability mixture model. In one embodiment involving classification over only two groups, the mixture may be, for example, a two-point probability density mixture. Alternatively, it may be, for example, that the two densities are normal with the same mean but differing variances, corresponding to the two different hypotheses. Using collected data, a fit is made to the theoretical mixture probability, and its parameters are determined.

The systems and methods described herein may be implemented on a computer system, with data on a computer-readable format such as a compact disk, a hard drive, a flash card, or a network server. Computer instructions can be written, either using commercially-available software packages (such as SAS) or using in-house custom-designed code, to implement the methods and systems described herein. Alternatively, in one embodiment, the data and/or the computer software that implements the methods and systems described herein may be hosted on a server, downloadable by an end user. In yet another embodiment, the data and the instructions may be downloadable from, and/or the software executable on, a distributed computing environment comprising one or more computers, network servers, or other data processing platforms. The various computerized configurations configured to implement the systems and methods described herein are readily apparent to those of ordinary skill in the art, after having read this specification.

Further features and advantages of the systems and methods described

herein will be apparent from the following description of illustrative embodiments, and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

The following figures depict certain illustrative embodiments of the systems and methods described herein, in which like reference numerals refer to like elements. These depicted embodiments are to be understood as illustrative, and not as limiting in any way.

Figure 1 depicts p-values from gene simulation model using $p = 10000$ genes with a mean group separation of 1.5 and 0 for 10% and 90% of genes respectively.

Figure 2a shows estimated values for $|E(\mu_j|Y)|$ versus $\text{Var}(\mu_j|Y)$ from the gene simulation model of Figure 1.

Figure 2b depicts absolute mean differences $|\bar{Y}_{j,2} - \bar{Y}_{j,1}|$ versus $\text{Var}(\mu_j|Y)$ for gene j .

Figure 3a shows BAM test statistics $\mu_{j,n}^*$ versus $Z_{j,n}$ from simulations in Figure 2.

Figure 3b depicts a BAM test statistic by group plot.

Figure 3c depicts a Z-test by group plot.

Figure 4a shows the total number of misclassified observations from gene simulation model $p = 25000$.

Figure 4b shows a close-up view for smaller α values.

Figure 5a depicts the total number of misclassified observations from the gene simulation model used in Figure 4, but with smaller signal-to-noise ratio.

Figure 5b plots the BAM test statistics against Z-tests similar to Figure 3.

5 **Figure 6a** shows gene mean values $\bar{Y}_{j,l}$ versus gene sample variances $\hat{\sigma}_{j,l}^2$.

Figure 6b shows group mean differences $\bar{Y}_{j,2} - \bar{Y}_{j,1}$ versus standard errors.

Figure 7a shows the total number of misclassified observations from the Poisson gene simulation model.

10 **Figure 7b** shows the total number of misclassified observations from variance stabilized data.

Figure 8a shows the gene mean values $\bar{Y}_{j,l}$ versus gene standard deviations $\hat{\sigma}_{j,l}$ from colon cancer data.

15 **Figure 8b** shows the group mean differences $\bar{Y}_{j,2} - \bar{Y}_{j,1}$ versus standard errors.

Figure 8c shows genes found significant using Zcut with $\alpha = 0.0005$.

Figure 9 depicts estimated values for $|E(\mu_j|Y)|$ versus $\text{Var}(\mu_j|Y)$ using weighted regression.

Figure 10 shows genes found significant using Zcut with $\alpha = 0.0005$.

20 DETAILED DESCRIPTION OF CERTAIN ILLUSTRATED EMBODIMENTS

To provide an overall understanding of the systems and methods described herein, certain illustrative practices and embodiments will now be described. However, it will be understood by one of ordinary skill in the art that the systems and methods described herein can be adapted and
5 modified and applied in other applications, and that such other additions, modifications, and uses will not depart from the scope hereof.

OVERVIEW

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill
10 in the art to which the systems and methods described herein pertain.

The embodiments described herein comprise statistical data analysis methods, and the systems that implement those methods, which may be used in various contexts and applications. Although by no means limited to large-scale systems, in terms of scope, the systems and methods described
15 herein are suitable for applications wherein a large body of data is collected, and there is a need to discern structure within the data for the purpose of classification, significance detection, or both. Accordingly, biomedical applications, such as gene array data analysis, wherein the problem of interest includes identifying which genes contribute to disease, for exam-
20 ple, are natural contexts for classification and model selection.

Generally, prior art methods have had shortcomings in properly classifying data points that fall within some intermediate or "shady region" within the collected data set. This has been so at least in part because prior art methods have focused on keeping in check one type of classification error (false discovery or false alarm) at the expense of another classification
25

error (false non-discovery).

The systems and methods described herein address these shortcomings, and improve upon the prior art by providing not only a theoretical (and hence objective) benchmark to classify data points that fall within some
5 intermediate range for a particular context of interest, but also provide a graphical tool for a data analyst to use to develop rules of thumb, and to cultivate a better intuition for, the distribution of data, especially the intermediate-range data whose classification has caused trouble for prior art methods.

- 10 Employing a series of pre-processing, analysis, and post-processing methods, the systems and methods described herein not only address the concerns of prior approaches (in which false discovery rate was paramount in importance), but also disclose a method to control other sources of misclassification (a false non-discovery, for example).
- 15 To appreciate the scope of applicability of the systems and methods described herein, various contexts are described from which data can be obtained and to which the methods and systems described herein may be applied.

GENERATING DATA FOR ANALYSIS

- 20 Statistical methods described herein may be used to analyze a wide variety of data, including data generated by studies in the physical, biological or social sciences. While the statistical methods described herein are widely applicable, the methods will often provide particularly substantial

advantages when applied to large data sets (e.g. data sets having greater than about 100, 1000 or 10,000 data points. Furthermore, methods disclosed herein may be particularly beneficial when there is reason to believe that it would be of interest to decrease the rate of false negative categorization
5 in the data analysis.

In certain embodiments, statistical methods described herein may be applied to biomolecular expression data. The term "biomolecules" is used herein to refer to any of the various molecules produced by or present in a cell, a sample containing cells, a non-cellular or substantially non-cellular biological sample (e.g. a sample of extracellular material, such as hair) or organisms that have non-cellular states, such as viruses. "Expression data" is data that reflects the amount (e.g., relative or absolute, measured directly or indirectly) of one or more biomolecules present in a sample. Data sets that include expression data may also include other
10 information, such as, for example, information about post-translational modifications on proteins. Biomolecules may be, for example, genomic DNA, RNAs (both mRNAs and other RNAs), lipids, carbohydrates, or other metabolites. Where the biomolecules to be studied are RNA transcripts, the data may be referred to as "gene expression data" or "DNA
15 expression data". Where the biomolecules to be studied are proteins, the data may be referred to as "protein expression data".

A variety of methods for obtaining biomolecular expression data are well-known in the art. Several examples are described briefly here.

GENE EXPRESSION DATA

25 In various aspects, the systems and methods described herein are suitable

for analyzing gene expression data. Gene expression data may be gathered in any way that, in view of this specification, is available to one of skill in the art. Although many statistical methods provided herein are powerful tools for the analysis of data obtained by highly parallel data collection
5 systems, methods disclosed herein are also useful for the analysis of data gathered by more traditional methods.

Many gene expression detection methodologies employ a polynucleotide probe which forms a stable hybrid with that of the target gene. If it is expected that the probes will be perfectly complementary to the target sequence,
10 stringent conditions will often provide superior results. Lesser hybridization stringency may be used if some mismatching is expected, for example, if variants are expected with the result that the probe will not be completely complementary. Conditions may be chosen which rule out nonspecific/adventitious bindings, that is, which minimize noise.

15 Gene arrays are often used for highly parallel observation of many genes simultaneously. An array may comprise a set of spatially addressable probes that hybridize to several, and even thousands or more different transcripts. Such arrays are often classified as microarrays or macroarrays, and this classification depends on the size of each position on the array.

20 Probes in an array may be immobilized on or in a solid or semisolid support. Oligonucleotides can be bound to a support by a variety of processes, including lithography, and where the support is solid, it is common in the art to refer to such an array as a "chip", although this parlance is not intended to indicate that the support is silicon or has any particular conductive properties. Nucleic acid probes may be spotted onto a substrate
25 in a two-dimensional matrix or array, or a pre-fabricated array may be

purchased from a manufacturer, such as Affymetrix (Santa Clara, CA). Samples of nucleic acids (e.g. RNAs or cDNAs derived from cells of interest) can be labeled and then hybridized to the probes. Nucleic acids, comprising the labeled or tagged sample nucleic acids bound to probe nucleic acids, can be detected once the unbound portion of the sample is washed away. In addition, a variety of corporations will perform microarray analysis on a contractual basis, such as Cellomics, Inc. (Pittsburgh, PA), GenHunter Corp. (Nashville, TN) and Affymetrix.

In other embodiments, the sample nucleic acid is not labeled. In this case, hybridization can be determined, e.g., by plasmon resonance, as described, e.g., in Thiel et al. (1997) *Anal. Chem.* 69:4948.

When using commercially available microarrays, adequate hybridization conditions are provided by the manufacturer. When using non-commercial microarrays, adequate hybridization conditions can be determined based on the following hybridization guidelines, as well as on the hybridization conditions described in the numerous published articles on the use of microarrays.

Theory and practice of nucleic acid hybridization is described, e.g., in S. Agrawal (ed.) *Methods in Molecular Biology*, volume 20; and Tijssen (1993) *Laboratory Techniques in biochemistry and molecular biology- hybridization with nucleic acid probes*, e.g., part I chapter 2 "Overview of principles of hybridization and the strategy of nucleic acid probe assays," Elsevier, New York provide a basic guide to nucleic acid hybridization.

Resultant hybridization patterns on an array may be visualized or detected in a variety of ways, with the particular manner of detection being

chosen based on the particular label of the target nucleic acid, where representative detection means include scintillation counting, autoradiography, fluorescence measurement, colorimetric measurement, light emission measurement, light scattering, and the like.

- 5 One method of detection includes an array scanner that is commercially available from Affymetrix (Santa Clara, Calif.) or Agilent. Additional scanning devices are described in, for example, U.S. Patent Nos. 5,143,854 and 5,424,186.

10 When fluorescently labeled probes are used, the fluorescence emissions at each site of a transcript array can be detected by scanning confocal laser microscopy. In one embodiment, a separate scan, using the appropriate excitation line, is carried out for each fluorophore used. Alternatively, a laser can be used that allows simultaneous specimen illumination at wavelengths specific to more than one fluorophore and emissions from more
15 than one fluorophore can be analyzed simultaneously. Fluorescence laser scanning devices are described in Schena et al., 1996, *Genome Res.* 6:639-645 and in other references cited herein. Alternatively, the fiber-optic bundle described by Ferguson et al., 1996, *Nature Biotech.* 14:1681-1684, may be used to monitor mRNA abundance levels.

- 20 Following the data gathering operation, the data will typically be reported to a data analysis system. To facilitate data analysis, the data obtained by the reader from the device will typically be analyzed using a digital computer. Typically, the computer will be appropriately programmed for receipt and storage of the data from the device, as well as for analysis and
25 reporting of the data gathered, e.g., subtraction of the background, deconvolution of multi-color images, flagging or removing artifacts, verifying

that controls have performed properly, normalizing the signals, interpreting fluorescence data to determine the amount of hybridized target, normalization of background and single base mismatch hybridizations, and the like. Various analysis methods that may be employed in such a data
5 analysis system, or by a separate computer are described herein.

While the above discussion focuses on the use of arrays for the collection of gene expression data, such data may also be obtained through a variety of other methods, that, in view of this specification, are known to one of skill in the art. The following are examples (not intended to be comprehensive)
10 of alternative methods for gathering gene expression data.

A method for high throughput analysis of gene expression is the serial analysis of gene expression (SAGE) technique, first described in Velculescu et al. (1995) *Science* 270, 484-487. Among the advantages of SAGE is that it has the potential to provide detection of all genes expressed in a given
15 cell type, whether previously identified as genes or not, provides quantitative information about the relative expression of such genes, permits ready comparison of gene expression of genes in two cells, and yields sequence information that can be used to identify the detected genes. Thus far, SAGE methodology has proved itself reliable in detecting expression
20 of regulated and nonregulated genes in a variety of cell types (Velculescu et al. (1997) *Cell* 88, 243-251; Zhang et al. (1997) *Science* 276, 1268-1272 and Velculescu et al. (1999) *Nat. Genet.* 23, 387-388.

Gene expression data may be gathered by RT-PCR. mRNA obtained from a sample is reverse transcribed into a first cDNA strand and subjected to
25 polymerase chain reaction (PCR). House keeping genes, or other genes whose expression is fairly constant can be used as internal controls and

controls across experiments. Following the PCR reaction, the amplified products can be separated by electrophoresis and detected. Taqman (TM) fluorescent probes, or other detectable probes that become detectable in the presence of amplified product may also be used to quantitate PCR products. By using quantitative PCR, the level of amplified product will correlate with the level of RNA that was present in the sample. The amplified samples can also be separated on an agarose or polyacrylamide gel, transferred onto a filter, and the filter hybridized with a probe specific for the gene of interest. Numerous samples can be analyzed simultaneously by conducting parallel PCR amplification, e.g., by multiplex PCR.

Transcript levels may also be determined by dotblot analysis and related methods (see, e.g., G. A. Beltz et al., in *Methods in Enzymology*, Vol. 100, Part B, R. Wu, L. Grossmam, K. Moldave, Eds., Academic Press, New York, Chapter 19, pp. 266-308, 1985). In one embodiment, a specified amount of RNA extracted from cells is blotted (i.e., non-covalently bound) onto a filter, and the filter is hybridized with a probe of the gene of interest. Numerous RNA samples can be analyzed simultaneously, since a blot can comprise multiple spots of RNA. Hybridization is detected using a method that depends on the type of label of the probe. In another dotblot method, one or more probes of one or more genes characteristic of disease D are attached to a membrane, and the membrane is incubated with labeled nucleic acids obtained from and optionally derived from RNA of a cell or tissue of a subject. Such a dotblot is essentially an array comprising fewer probes than a microarray.

Techniques for producing and probing nucleic acids are further described, for example, in Sambrook et al., "Molecular Cloning: A Laboratory Manual" (New York, Cold Spring Harbor Laboratory, 1989).

PROTEIN EXPRESSION DATA

In various aspects, the systems and methods described herein are suitable for analyzing protein expression data. Protein expression data may be gathered in any way that, in view of this specification, is available to one of skill in the art. Although many analytical methods provided herein are powerful tools for the analysis of protein data obtained by highly parallel data collection systems, many such methods are also useful for the analysis of data gathered by more traditional methods.

Immunoassays are commonly used to quantitate the levels of proteins in samples, and many other immunoassay techniques are known in the art. The invention is not limited to a particular assay procedure, and therefore is intended to include both homogeneous and heterogeneous procedures. Exemplary immunoassays which can be conducted according to the invention include fluorescence polarization immunoassay (FPIA), fluorescence immunoassay (FIA), enzyme immunoassay (EIA), nephelometric inhibition immunoassay (NIA), enzyme linked immunosorbent assay (ELISA), and radioimmunoassay (RIA). An indicator moiety, or label group, can be attached to the subject antibodies and is selected so as to meet the needs of various uses of the method which are often dictated by the availability of assay equipment and compatible immunoassay procedures. General techniques to be used in performing the various immunoassays noted above are known to those of ordinary skill in the art.

Protein levels may be detected by a variety of gel-based methods. For example, proteins may be resolved by gel electrophoresis, preferably two-dimensional electrophoresis comprising a first dimension based on pI and a second dimension of denaturing PAGE. Proteins resolved by electrophore-

sis may be labeled beforehand by metabolic labeling, such as with radioactive sulfur, carbon, nitrogen and/or hydrogen labels. If phosphorylation levels are of interest, proteins may be metabolically labeled with a phosphorus isotope. Radioactively labeled proteins may be detected by autoradiography, or by use of a commercially available system such as the PhosphorImager (TM) available from Molecular Dynamics (Amersham). Proteins may also be detected with a variety of stains, including but not limited to, Coomassie Blue, Ponceau S, silver staining, amido black, SYPRO dyes, etc. Proteins may also be excised from gels and subjected to mass spectroscopic analysis for identification. Gel electrophoresis may be preceded by a variety of fractionation steps to generate various subfractionated pools of proteins. Such fractionation steps may include, but are not limited to, ammonium sulfate precipitation, ion exchange chromatography, reverse phase chromatography, hydrophobic interaction chromatography, hydroxylapatite chromatography and any of a variety of affinity chromatography methods.

Protein expression levels may also be measured through the use of a protein array. For example, one type of protein array comprises an array of antibodies of known specificity to particular proteins. Antibodies may be affixed to a support by, for example, the natural interaction of antibodies with supports such as PVDF and nitrocellulose, or, as another example, by interaction with a support that is covalently associated with protein A (see for example U.S. Pat. No. 6,197,599), which binds tightly to the constant region of IgG antibodies. Antibodies may be spotted onto supports using technology similar to that described above for spotting nucleic acid probes onto supports. In another example, an array is prepared by coating a surface with a self-assembling monolayer that generates a matrix of positions where protein capture agents can be bound, and protein capture agents

range from antibodies (and variants thereof) to aptamers, phage coat proteins, combinatorially derived RNAs, etc. (U.S. Pat. No. 6,329,209). Proteins bound to such arrays may be detected by a variety of methods known in the art. For example, proteins may be metabolically labeled in the sample with, for example, a radioactive label. Detection may then be accomplished using devices as described above. Proteins may also be labeled after being isolated from the sample, with, for example, a cross-linkable fluorescent agent. In one example, proteins are desorbed from the array by laser and subjected to mass spectroscopy for identification (U.S. Pat. No. 6,225,047). In another variation, the array may be designed for detection by surface plasmon resonance. In this case, binding is detected by changes in the surface plasmon resonance of the support (see, for example, Brockman and Fernandez, American Laboratory (June, 2001) p.37).

METABOLITE EXPRESSION DATA

- 15 In various aspects, the systems and methods described herein are suitable for analyzing metabolite expression data. Metabolites, such as lipids, carbohydrates, cofactors, amino acids and other types of molecules found in organism may be measured in a number of different ways, and such methods tend to vary depending on the type of compound to be analyzed. High throughput systems for metabolite analysis include liquid chromatography (e.g. reverse phase HPLC) coupled to a mass spectrophotometer and gas chromatography coupled to a mass spectrophotometer. Other traditional methods for analyzing metabolites may be employed, such as enzymatic assays, selective extractions and others known in the art.
- 20
- 25 Methods disclosed herein may also be used to analyze biomolecular data that may not typically be characterized as expression data. Biomolecular

arrays are available, for example, from Santa Clara, CA-based Affymetrix. A survey of biomolecular arrays appears in "The Array of Today: Biomolecule arrays become the 21st century's test tube," by J. D. Cortese, *The Scientist*, 14[17]:25, Sep. 4, 2000.

- 5 In one embodiment, biomolecular arrays are ordered molecular arrays that simplify discrimination of surface-bound biomolecules through the spatial control of ligand presentation. According to one practice, first, photolithography is used to spatially direct the synthesis of a matrix of biological ligands. A high-affinity binding partner is then applied to the matrix,
10 which binds at locations defined by the ligand array. An atomic force microscope (AFM) may then be used to detect the presence and organization of the high-affinity binding partner.

- A number of high throughput systems for genotype analysis are now available. An array of single-base extension tags (SBE-TAGs) has been described
15 for the highly parallel analysis of single nucleotide polymorphisms present in a human genotype. Hirschhorn et al., *Proc Natl Acad Sci U S A*. 2000 97(22):12164-69. SBE-TAGS may also be used for the analysis of other organisms. Similar arrays have been used to identify pathogens (e.g. viruses, protozoans or bacteria), for example, by hybridizing samples of genomic
20 or transcribed nucleic acids to an array containing probes corresponding to different genomic regions of a variety of related pathogenic and non-pathogenic organisms. Borucki et al. *Vet Microbiol* 2003 92(4):351-62. Alkhalidi et al. *J AOAC Int* 2002 85(4):906-10. Wilson et al. *Mol Cell Probes* 2002 Apr;16(2):119-27.

25 TISSUE DATA

In various aspects, the systems and methods described herein are suitable for analyzing tissue array data, whereby multiple tissue cores are affixed, typically in the form of an array, to a chip, in a manner not unlike how genes are affixed to a gene array; a typical size of a tissue array is about
5 300 to about 400 cores, although other array sizes may be used in various alternative embodiments, without departing from the scope of the systems and methods disclosed herein.

Tissue arrays facilitate rapid evaluation, typically in parallel and in situ, of the role of normal and abnormal genes or gene products from hundreds
10 of tissue samples. For example, studies involving tumors at various stages and disease grades improve the evaluation of the diagnostic, prognostic, and therapeutic characteristics of cancer gene candidates. In the absence of tissue arrays, such evaluation is frequently difficult, because it is typically conducted serially. Tissue microarrays provide a high-throughput
15 means for the study of, for example, and without limitation, gene dosage and protein expression patterns in a large number of individual tissues for rapid and comprehensive molecular profiling of cancer and other diseases, without usurping limited tissue supplies.

In one practice, a tissue array application includes searching for oncogene
20 amplifications in tumor tissue panels. Large-scale studies involving tumors encompassing differing stages and grades of disease facilitate a more efficient validation of putative markers that may ultimately correlate genotypes with phenotypes.

In one exemplary embodiment, hundreds of substantially cylindrical formalin fixed and paraffin-embedded tissue cores are densely and precisely
25 arrayed into a typical paraffin block. From this, up to a few hundred se-

rial sections may be produced and placed on individual glass microscopic slides. An adhesive-coated tape system (e.g., produced by Instrumedics, Hackensack, New Jersey) may be useful in maintaining array formatting in the transfer of cut ribbons to the glass slides. Moreover, the tissue array
5 technology is applicable to other medical research disciplines in which, for example, paraffin-embedded tissues are used, including, but not limited to, structural, developmental, and metabolic studies.

In general, by employing tissue arrays, various observations may be made about the tissues; for example, immunohistochemistry (IHC), morphology
10 and protein expression along a particular pathway of interest. In particular, tests that may be performed using tissue arrays may include morphologic brightfield examination (H&E), protein expression studies (immunohistochemistry), DNA copy number analysis (fluorescence in situ hybridization FISH), or a combination thereof. Also, in situ PCR and tra-
15 ditional histologic special stains may be applied to tissue array slides. In some cases, mRNA level expression studies may also be performed.

The studies listed above may be done separately, or, more typically and effectively, substantially in parallel (for example, to concurrently compare gene amplification and expression in the same tissue). The systems and
20 methods described herein are suitable for performing statistical analyses of data obtained from tissue arrays.

VIRAL DATA

In various aspects, the systems and methods disclosed herein are suitable for analyzing viral data, whereby, for example, genomic sequences from
25 a viral bank of viral genomes are printed onto an array, in a manner not

unlike how genes are affixed to a gene array. These genomic sequences typically represent highly conserved sequences of viral DNA, which are characteristic of, or substantially specific to, a particular virus. This aspect is not necessarily associated with viral gene expression per se.

- 5 According to one practice, an unknown viral sample (e.g., from nasal lavage or sputum) is hybridized to the virus chip, and the pattern of hybridization is observed. Within the context of the systems and methods described herein, this analysis includes a one-way ANOVA. Of interest here is, among other things, which hybridizations are significantly different from zero;
10 typically, this is not a 2-way or M -way ANOVA. Studies of viral data by using the systems and methods described herein have potential applications in bio-terrorism applications.

- An exemplary embodiment, in the context of differential gene expression data, is now described as a two-way model selection problem. This is for illustrative purposes only, and should not be construed as limiting the scope
15 of the invention in any way. Those of skill in the art will appreciate that the methods and systems described herein readily apply to other data analysis contexts, including multi-way classification problems.

- DNA microarrays open up a broad new horizon for investigators interested in studying the genetic determinants of disease. The high throughput nature of these arrays, where differential expression for thousands of genes can be measured simultaneously, creates an enormous wealth of information, but also poses a challenge for data analysis because of the large multiple testing problem involved. The solution has generally been
20 to focus on optimizing false-discovery rates while sacrificing power. The drawback of this approach is that more subtle expression differences will
25

be missed, which might give investigators more insight into the genetic environment necessary for a disease process to take hold.

Disclosed herein is a method for detecting differentially expressed genes based on a high-dimensional model selection technique, Bayesian ANOVA
5 for Microarrays (BAM), which strikes a balance between false rejections and false nonrejections. The approach is based at least in part on a weighted average of generalized ridge regression estimates that provides the benefits of using shrinkage estimation combined with model averaging. A graphical tool based on the amount of shrinkage is developed to visualize
10 the trade-off between low false-discovery rates and finding more genes. Simulations are used to illustrate BAMs performance, and the method is applied to a large database of colon cancer gene expression data. A working hypothesis in the colon cancer analysis is that large differential expressions may not be the only ones contributing to metastasis; in fact, moderate changes in expression of genes may be involved in modifying the
15 genetic environment to a sufficient extent for metastasis to occur. A functional biological analysis of gene effects found by BAM, but not by other false-discovery-based approaches, supports the hypothesis.

1. Introduction. DNA microarray technology allows researchers to measure the expression levels of thousands of genes simultaneously over different time points, different experimental conditions, or different tissue samples. It is the relevant abundance of the genetic product that provides surrogate information about the relative abundance of the cells' proteins. The differences in protein abundance are what characterize genetic differences between different samples. In the preparation of a DNA microarray sample, DNA or RNA molecules labeled with fluorescent dye are hy-
20
25

bridized with a library of complementary strands fixed on a solid surface.

Generally, there are two major branches of chip technologies. Oligonucleotide arrays contain gene-specific sequences, called probes, about 20 bases long for each gene. The resulting fluorescence intensity from the hybridization process gives information about the abundance of the corresponding sample mRNA (a precursor to the cells proteins). The other type of array involves complementary DNA (cDNA), which can be spotted on nylon filters or glass slides. Complex mRNA probes are reverse-transcribed to cDNA at two sites for each gene and labeled with red control or green test fluorescent dyes. The ratio of red/green intensities represents the amount of RNA hybridizing at each site.

Although many analysis questions may be of interest, a commonly posed question asks for the detection of differentially expressing genes between experimental states, for example, between control samples and treatment samples, or between normal tissue samples and diseased tissue samples. Current approaches involve Bayes and empirical Bayes mixture analysis, and multiple hypothesis testing approaches with corrections designed to control the expected false discovery rate (FDR). The FDR is defined as the false-positive rate among all rejected (null) hypotheses; that is, the total number of rejected hypotheses where the null is in fact true, divided by the total number of rejected hypotheses. Researchers have provided a sequential p-value method to control the expected FDR. (Note: what others sometimes call the FDR is referred to herein as the expected FDR, a slight departure in terminology.) Given a set of independent hypothesis tests and corresponding p values, the method provides an estimate k , such that if one rejects those hypotheses corresponding to $P_{(1)}, P_{(2)}, \dots, P_{(k)}$, the k -ordered observed p values, then the FDR is, on average, controlled un-

der some pre-chosen level. For convenience, this is referred to herein as the BH method.

ANOVA-based models are another way to approach the problem. The first ANOVA methods were developed to account for ancillary sources of variation when making estimates of relative expression for genes. More recently, an efficient approach casting the differential detection problem as an ANOVA model and testing individual model effects with FDR corrections has been developed. With all of these FDR applications, the methods work well by ensuring that an upper bound is met; however, a side effect is often a high false nondiscovery rate (FNR). The FNR is defined as the proportion of nonrejected (null) hypotheses that are incorrect. Researchers have shown that the BH method cannot simultaneously optimize the expected FDR and the expected FNR, implying that the method has low power. (The simulations in Sec. 6 of this disclosure also confirm this behavior.) An oracle threshold value was described by researchers, which improves on BH in the case where p-value distributions are independent two-point mixtures. It has been recognized that power for the BH approach could be improved. Typically, however, there is a price to be paid in terms of increased computation and some subjectiveness. Moreover, the largest relative power gains observed by researchers is typically realized when large proportions of genes are truly differentially expressed, a property that might not hold in some disease problems, because the number of genes expressed compared with dimension are expected to be small.

1.1. Larger Models Versus False Discovery.

The systems and methods described herein and employed for detecting

gene expression changes use a Bayesian ANOVA for microarrays (BAM) technique that provides shrinkage estimates of gene effects derived from a form of generalized ridge regression and are suitable for high-dimensional model selection in linear regression problems. A key feature of the BAM
5 systems and methods is that its output permits different model estimators to be defined, and each can be tailored to suit the various needs of the user. For example, for analysts concerned with false discovery rates, the systems and methods described herein construct an estimator that focuses on the FDR.

- 10 Also disclosed herein is a graphical method, based on the amount of shrinkage, which can be used to visualize the trade-off between a low FDR and finding more genes. This device can be used to select α cutoff values for model estimators. Selecting an appropriate α value is critical to the performance of any method used to detect differentially expressing genes.
15 Simply relying on using preset α values, especially conventional values used in traditional problems, can be a poor strategy as such values may sometimes be magnitudes off from optimal ones. The case study example of Section 7 illustrates how small α may be in practice.

As a consequence of shrinkage and model averaging, the BAM method
20 will be shown to strike a good balance between identifying large numbers of genes and controlling the number of falsely identified genes. This kind of property may be of great importance in the search for a colon cancer metastatic signature, a topic that will be touched upon more in Section 7 as one illustration of applicability of the systems and methods described
25 herein.

Very little is currently known about the genetic determinants of colon can-

cer metastasis, although it is generally agreed that a genetic signature is complex. In fitting in with this general philosophy, the hypothesis adopted herein is that genes with large differential expressions may not be the only ones contributing to metastasis — that in fact, more moderate changes in
5 expression of some genes might be sufficient to trigger the process.

Proving this hypothesis directly is difficult. A more reasonable surrogate hypothesis is that the genes showing more moderate changes in expression provide a suitable milieu for other metastatic events—accompanied by other genes showing much larger expression differences. This general
10 principle may be at play in many diseases other than colon cancer; therefore, increased power in detecting differentially expressed genes becomes more important, and the systems and methods described herein address this need.

To fix ideas about BAM, consider Table 1 and Figure 1 which present re-
15 sults from a gene simulation model involving $p = 10000$ genes (specific details are provided in Section 6). Figure 1 plots the histogram of the p-values from “Zcut” method (Section 3.3) against the p-values derived from individual Z-tests which use a pooled estimate for the standard error (Section 5.1). The figure shows the effect shrinkage plays in BAM; here it has
20 the effect of pushing the p-value distribution for Zcut apart, helping to more clearly delineate expressive genes. Each gene had 5 observations per group. The top portion depicts p-values derived from the Zcut estimator, and the bottom portion includes data from standard Z-tests.

Differences in procedures can be carefully compared by Table 1, which
25 records the number of genes detected as expressing, the FDR and FNR, and the Type-I and Type-II error rates. The *Type-I error rate* is the propor-

- tion of genes rejected given they are non-expressing (the null), while the *Type-II error rate* is the proportion of non-rejected genes given they are expressing (the alternative). Values are tabulated at an $\alpha = 0.05$ value. It should be noted how Zcut leads to a reduced FDR, while at the same time
- 5 seeking to maintain high power. Table 1 also records the results of the BH method applied to the p-values from the Z-tests. Also recorded are the results from the "FDRmix" procedure (Section 4), a hybrid BH procedure. Table 1 shows that both BH and FDRmix lead to a smaller number of identified genes than Zcut or Z-test. This is at least in part because both pro-
- 10 cedures attempt to control the FDR, which typically results in fewer genes being found significant. Here, the BH method has the lowest FDR, slightly smaller than its target $\alpha = 0.05$ value. Although FDRmix is far off from the α target value, it does what its supposed to, which is to reduce the FDR of Zcut while maintaining power.
- 15 Care should be exercised in directly comparing FDR and FNR values, or for that matter Type-I and Type-II error rates, for the different procedures at the same α value, since an α target value means something different for each procedure. Looking at the different rates individually will does not, by itself, explain how the procedures perform overall. Thus, to be able to
- 20 compare procedures on a more equal basis, an overall measure of performance is defined herein, "TotalMiss," which is also recorded in Table 1. This is the total number of false rejections and false non-rejections, i.e. the number of misclassified genes for a procedure, which can be thought of as a measure of total risk. In terms of the total risk, Table 1 shows that Zcut is
- 25 the clear winner here. A more detailed study of how TotalMiss varies with α will be presented in Section 6. This kind of analysis may be important in assessing the robustness of a procedure. Since α values can vary considerably depending upon the data, procedures that are robust will be those

that exhibit uniformly good risk behavior.

1.2. Organization of the Remainder of the Disclosure.

This disclosure is organized as follows. Section 2 presents an overview of a stochastic variable selection algorithm that involves shrinkage of ordinary
5 least squares estimators. The BAM procedure is introduced in Section 3. There the Zcut estimator and a shrinkage plot, which can be used as a graphical device for setting α values, are introduced. Theorems 1 and 2 of Section 3.4 discuss the robustness and adaptiveness of the BAM method and provide explanations for the Zcut method. Section 4 introduces the
10 FDRmix procedure. Section 5 discusses some more commonly known procedures. The performance of BAM is studied via simulations in Section 6 and compared in detail to the more standard methods of Section 5. In Section 7 the colon cancer problem is discussed and a detailed analysis of the data based on BAM is presented. Section 8 concludes with a discussion,
15 including a description of how the systems and methods described herein may be extended to analyze more than two groups.

2. The P-SVS procedure. An approach employed by the systems and methods described herein includes recasting the problem of finding differentially expressing genes as a problem of determining which factors
20 are significant in a Bayesian ANOVA model. This is referred to herein as the BAM method. As one can generally rewrite an ANOVA model as a linear regression model, the task of finding expressive genes can be conceptualized as a variable selection problem. This connection is employed to advantage by the systems and methods described herein to select vari-
25 ables in high-dimensional regression problems called P-SVS (Parametric

Stochastic Variable Selection). Section 3 outlines the details of the BAM approach and how it relates to P-SVS. A general description of the P-SVS procedure is now presented.

The P-SVS method is a hybrid version of the spike and slab models. It includes a Bayesian hierarchical approach for model selection in linear regression models,

$$Y_i = x_i^T \beta_0 + \epsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where Y_i is the response value, ϵ_i are i.i.d $N(0, \sigma_0^2)$ measurement errors and $\beta_0 = (\beta_{1,0}, \dots, \beta_{p,0})^T$ is the unknown (true) p -dimensional vector of coefficients for the covariates x_i . To answer the question of which $\beta_{j,0}$ are non-zero, the P-SVS approach works by modeling (1) as a hierarchical model

$$\begin{aligned} (Y_i | \beta, \sigma^2, x_i) &\stackrel{\text{iid}}{\sim} \text{Normal}(x_i^T \beta, \sigma^2), & i = 1, \dots, n \\ (\beta_j | \gamma_j, \tau_j^2) &\stackrel{\text{iid}}{\sim} \text{Normal}(0, \gamma_j \tau_j^2), & j = 1, \dots, p \\ (\gamma_j | \gamma^*, \lambda) &\stackrel{\text{iid}}{\sim} (1 - \lambda) \delta_{\gamma^*}(\cdot) + \lambda \delta_1(\cdot) \\ (\tau_j^{-2} | b_1, b_2) &\stackrel{\text{iid}}{\sim} \text{Gamma}(b_1, b_2) \\ \lambda &\sim \text{Uniform}[0, 1] \\ (\sigma^{-2} | a_1, a_2) &\sim \text{Gamma}(a_1, a_2). \end{aligned} \quad (2)$$

- A key feature of the model is the choice for the priors of τ_j^2 and γ_j which are calibrated so that the variance $\nu_j^2 = \gamma_j \tau_j^2$ for a coefficient β_j has a bimodal distribution. A large value for ν_j^2 occurs when $\gamma_j = 1$ and τ_j^2 is large and will induce large values for β_j , thus identifying the covariate as potentially informative. Small ν_j^2 occur when $\gamma_j = \gamma^*$ (chosen to be a small value). In this case the value for β_j will become near zero, signifying that β_j is

- unlikely to be informative. The value for λ in (2) controls how likely it is for γ_j to be 1 or γ^* , and thus it controls how many β_j are non-zero, and so it controls the complexity of the model. Generally, the choice of hyperparameters for priors is important to the behavior of ν_j^2 , and hence P-
- 5 SVS's ability to properly select variables. The values for hyperparameters $\gamma^*, a_1, a_2, b_1, b_2$ need not be tuned for each data set, and may be kept fixed.

Let $\gamma = (\gamma_1, \dots, \gamma_p)^T$. This is a way of encoding models in binary strings (if $\gamma_j = 1$ select β_j). One approach used in spike and slab variable selection looks to the posterior behavior of γ to identify the "best model", for example by identifying the γ which occurs with highest posterior probability. However, in very large variable selection problems information is best processed differently as the information contained in γ is too coarse (if p is very large, a high frequency model may not even be found). Variables are selected by considering the magnitude of their posterior mean values. Motivation to use the posterior mean to select variables stems from its interpretation as an adaptive weighted average of generalized ridge estimates. Such values are reliable as they are produced by model averaging in combination with shrinkage, two methods that improve model selection. It can easily be seen that the posterior mean is a weighted ridge estimate. Let $\beta = (\beta_1, \dots, \beta_p)^T$. From (2), the posterior mean can be written as

$$E(\beta|Y) = \iint \beta \pi(d\beta|\nu^2, \sigma^2, Y) \pi(d\gamma, d\tau^2, d\lambda, d\sigma^2|Y),$$

where $\nu^2 = (\nu_1^2, \dots, \nu_p^2)^T$, $\tau^2 = (\tau_1^2, \dots, \tau_p^2)^T$ and $Y = (Y_1, \dots, Y_n)^T$. Elementary calculations show that

$$(\beta|\nu^2, \sigma^2, Y) \sim \text{Normal}(\Sigma^{-1}X^TY, \sigma^2\Sigma^{-1}), \quad (3)$$

where X is the $n \times p$ design matrix, $\Sigma = \sigma^2\Gamma^{-1} + X^TX$ and Γ is the $p \times p$ di-

agonal matrix with diagonal values $\nu_j^2 = \gamma_j \tau_j^2$. Notice that the conditional posterior mean for β is

$$E(\beta|\nu^2, \sigma^2, Y) = \Sigma^{-1} X^T Y = (\sigma^2 \Gamma^{-1} + X^T X)^{-1} X^T Y,$$

which is the the ridge estimate from a generalized ridge regression of Y on X with weights $\sigma^2 \Gamma^{-1}$. Small values for diagonal elements of Γ have the effect of shrinking coefficients. Thus, the posterior mean for β can now seen to be a weighted average of ridge shrunken estimates where the adaptive
 5 weights are determined from the posteriors of γ , τ^2 and λ .

This shift from high frequency models selected by γ to models selected on the basis of individual performance of variables is corroborated by other researchers. It has been shown that the high frequency model can be suboptimal even in the simplest case when the design matrix is orthog-
 10 onal. Under orthogonality, the optimal predictive model is not the high frequency model, but the median probability model defined as the model consisting of variables whose overall posterior probability is greater than or equal to $1/2$.

3. Bayesian ANOVA for microarrays. The BAM systems and methods described herein apply the variable selection device by recasting the microarray data problem in terms of ANOVA, and hence as a linear regression model. Consider the case wherein $l = 2$ is the number of groups of samples; extensions to allow for more groups are discussed later in section 8. For group l , let $Y_{j,k,l}$ denote the k -th expression value, $k = 1, \dots, n_{j,l}$, for gene $j = 1, \dots, p$. Group $l = 1$ will correspond to the control group, while $l = 2$ represents the treatment group. For example, in the colon cancer study of Section 7, the control group represents Duke B-survivor

colon tissue samples while the treatment group are metastasized colon cancer samples. Microarray data, as in our colon cancer study, will often be collected from balanced experimental designs with fixed group sizes $n_{j,1} = N_1$ and $n_{j,2} = N_2$. In such settings, $Y_{1,k,l}, \dots, Y_{p,k,l}$ will typically represent the p gene expressions obtained from a microarray chip for a specific individual k with a tissue type from group l (i.e. either an individual k from the control group or an individual k from the treatment group). However, since more complex experimental designs are possible the problem is approached herein more generally by allowing for unbalanced data. A key question of general interest is which genes are expressing differently over the two groups. Let $\epsilon_{j,k,l}$ be i.i.d $N(0, \sigma_0^2)$ measurement errors. The problem may be formulated as the ANOVA model,

$$Y_{j,k,l} = \theta_{j,0} + \mu_{j,0}I\{l = 2\} + \epsilon_{j,k,l}, \quad j = 1, \dots, p, \quad k = 1, \dots, n_{j,l}, \quad l = 1, 2, \quad (4)$$

where those genes that are expressing differentially correspond to indices j where $\mu_{j,0} \neq 0$ (if genes turn on then $\mu_{j,0} > 0$, otherwise if they turn off $\mu_{j,0} < 0$).

Observe that (4) has $2p$ parameters. Many of the $\theta_{j,0}$ parameters will be non-zero since they estimate the overall mean for a gene j . However, since our interest is only in $\mu_{j,0}$, the gene-treatment effect, it is convenient to force $\theta_{j,0}$ to be near zero so that the effective dimension of the problem is p . A useful strategy is to replace $Y_{j,k,l}$ by its centered value

$$Y_{j,k,l} - \bar{Y}_{j,l}, \quad (5)$$

where $\bar{Y}_{j,l} = \sum_{k=1}^{n_{j,l}} Y_{j,k,l} / n_{j,l}$ is the mean for gene j over group l . This will also reduce the correlation between the Bayesian parameters θ_j and μ_j and greatly improves the calibration of P-SVS. Section 3.2 will explain

this point in more detail.

Extensions to (4) are possible. For example, additional covariates may be included in the model, which is useful in experiments where external data (such as clinical information of individuals providing tissue samples) is collected alongside gene expressions. Thus (4) may be extended to the important class of ANCOVA (analysis of covariance) models. In another extension, genes may have different variability in addition to different mean expression levels. In such cases, (4) may be extended to allow for differing gene measurement error variances σ_j^2 . Often σ_j^2 is related to the mean gene expression. In settings where this relationship is simple, for example, where the variance might be linear in the mean, this behavior is often handled by applying a variance stabilizing transform. Section 6.1 will study how well this approach works. On the other hand, the relationship between σ_j^2 and the mean expression may be complex, as is the case for the colon cancer study of Section 7. For such problems a specialized method will be developed.

The expressions $Y_{j,k,l}$ are typically the end product of some form of probe-level standardization across samples (i.e. chips). Standardization for the colon cancer data involved standardizing chips to a gamma distribution with a fixed shape parameter. Further gene-level pre-processing is done as discussed below. With so much processing of the data it is natural to question the assumption of normality in (4). In fact, Theorem 1 of Section 3.4 will show that this assumption is not necessary due to the effect of a central limit theorem, as long as $\epsilon_{j,k,l}$ are independent with suitably behaved moments and if group sizes $n_{j,l}$ are relatively large. Some empirical evidence of this effect will be presented in Section 6 for Poisson data.

- 3.1. Linear regression and pre-processing the data. To harness P-SVS in the BAM method, (4) is rewritten as a linear regression model (1), that is, using a single index i . This is accomplished by stringing observations (5) out in order, starting with the observations for gene $j = 1$, group $l = 1$, followed by values for gene $j = 1$, group $l = 2$, followed by observations for gene $j = 2$, group $l = 1$, etc. Notationally, the following parameter adjustments are also made. In place of β_j , coefficients (θ_j, μ_j) are used. Hierarchical prior variances for (θ_j, μ_j) are now denoted by $(\nu_{2j-1}^2, \nu_{2j}^2)$. Analogous notational changes are applied to γ and τ^2 in (2).
- 10 The second step to using P-SVS requires a data pre-processing step which rescales the observations by the square-root of the sample size divided by the mean square error and by transforming the design matrix so that its columns each satisfy $\sum_i x_{i,j}^2 = n$. This calibrates the conditional variance $(\Gamma^{-1} + X'X/\sigma^2)^{-1}$ for β in (3) to have diagonal elements nearly zero or one
- 15 in low to moderate correlation problems. Variances of zero correspond to non-significant variables, while variances of one correspond to informative covariates. Since the conditional variance of one represents a good lower bound for significant variables, one can then use a standard $N(0, 1)$ distribution to assess whether a specific regression coefficient β_j should be
- 20 considered informative, and hence included in the model.

Let $\hat{\sigma}_n^2$ denote the usual unbiased estimator for σ_0^2 from (4),

$$\hat{\sigma}_n^2 = \frac{1}{n - 2p} \sum_{j,k,l} \left(Y_{j,k,l} - \bar{Y}_{j,1} I\{l = 1\} - \bar{Y}_{j,2} I\{l = 2\} \right)^2,$$

where $n = \sum_{j=1}^p n_j$ are the total number of observations and $n_j = n_{j,1} + n_{j,2}$ is the total sample size for gene j . To calibrate the data, replace the

observations (5) with rescaled values

$$\tilde{Y}_{j,k,l} = (Y_{j,k,l} - \bar{Y}_{j,1}) \times \sqrt{n/\hat{\sigma}_n^2}.$$

The effect of this scaling will be to force σ^2 to be approximately equal to n and it will rescale posterior mean values so they can be directly compared to a limiting normal distribution. Columns are also rescaled so they have second moments equal to one. Thus after pre-processing, (4) may be rewritten as

$$\tilde{Y} = \tilde{X}\tilde{\beta}_0 + \tilde{\epsilon},$$

where $\tilde{Y}$ is the vector of the n strung out $\tilde{Y}_{j,k,l}$ values, $\tilde{\beta}_0$ is the new regression coefficient under the scaling, $\tilde{\epsilon}$ is the new vector of measurement errors and $\tilde{X}$ is the $n \times (2p)$ design matrix. Here $\tilde{X}$ is defined so that its $2j - 1$ column consists of zeroes everywhere except for the values $\sqrt{n/n_j}$ placed along the n_j rows corresponding to gene j , while column $2j$ consists of zeroes everywhere except for values $\sqrt{n/n_{j,2}}$ placed along the $n_{j,2}$ rows corresponding to gene j for group $l = 2$.

3.2. Ridge regression estimates. Because of the simple nature of the design matrix $\tilde{X}$ the conditional distribution for β may be explicitly determined. Adjusting to the new notation, a little bit of algebra using (3) shows that

$$((\theta_j, \mu_j)^T | v_{2j-1}^2, v_{2j}^2, \sigma^2, Y) \sim \text{Normal}((\hat{\theta}_{j,n}, \hat{\mu}_{j,n})^T, \hat{\Sigma}_{j,n}^{-1}),$$

where

$$\begin{pmatrix} \hat{\theta}_{j,n} \\ \hat{\mu}_{j,n} \end{pmatrix} = \frac{n\hat{\Sigma}_{j,n}^{-1}}{\hat{\sigma}_n\sigma^2} \begin{pmatrix} n_j^{-1/2} \sum_{k,l} (Y_{j,k,l} - \bar{Y}_{j,1}) \\ n_{j,2}^{-1/2} \sum_k (Y_{j,k,2} - \bar{Y}_{j,1}) \end{pmatrix}, \quad (6)$$

and

$$\hat{\Sigma}_{j,n} = \begin{pmatrix} n/\sigma^2 + 1/v_{2j-1}^2 & n/\sigma^2 \times \sqrt{n_{j,2}/n_j} \\ n/\sigma^2 \times \sqrt{n_{j,2}/n_j} & n/\sigma^2 + 1/v_{2j}^2 \end{pmatrix}.$$

Recall that as a consequence of the pre-processing step, σ^2 is approximately equal to n . Also, since $\hat{\theta}_{j,n}$ is nearly zero due to the centering (5), it is expected that $v_{2j-1}^2 \approx 0$. Thus,

$$\hat{\Sigma}_{j,n}^{-1} \approx \begin{pmatrix} 0 & 0 \\ 0 & v_{2j}^2/(v_{2j}^2 + 1) \end{pmatrix}.$$

This implies that the posterior correlation between θ_j and μ_j are nearly zero. Thus, the centering method (5) has two important roles: (i) it reduces the dimension of the parameter space and (ii) it reduces correlation between parameters.

Now for genes that are expressing differently, it is expected that v_{2j}^2 is large, and thus the conditional mean for μ_j is approximately

$$\hat{\mu}_{j,n} \approx \frac{\sqrt{n_{j,2}}}{\hat{\sigma}_n} (\bar{Y}_{j,2} - \bar{Y}_{j,1}) \quad (7)$$

- 5 and the conditional variance for μ_j is approximately $v_{2j}^2/(v_{2j}^2 + 1) \approx 1$. As the conditional variance represents a lower bound for the posterior variance $V(\mu_j|Y)$ this suggests that the posterior mean $E(\mu_j|Y)$ is to be compared to a $N(0, 1)$ distribution to test whether μ_j is essentially zero. However, Theorem 1 of Section 3.4 will show that it is more appropriate to
- 10 use the theoretical limiting variance for $\hat{\mu}_{j,n}$ as our lower bound; here equal to $n_j/n_{j,1}$. Thus, $E(\mu_j|Y)$ is to be compared to a $N(0, n_j/n_{j,1})$ distribution to test whether $\mu_{j,0}$ is non-zero. This is the basis of the Zcut procedure.

Theorem 1 justifies this strategy more rigorously. In the meantime, some evidence for this approach may be seen in Figure 2(a) where the posterior absolute means and posterior variances from the earlier gene simulation model are plotted. In this "shrinkage plot" the lower variance bound of one is represented by a thin dashed horizontal line, while the thick dashed horizontal line represents the theoretical variance $n_j/n_{j,1} = 2$. Circles indicate genes whose true means $\mu_{j,0} = 0$, while crosses are used for $\mu_{j,0} \neq 0$. Notice how large values for $|E(\mu_j|Y)|$ have posterior variances near one, which then jump up to the theoretical variance $n_j/n_{j,1} = 2$ as $|E(\mu_j|Y)|$ becomes moderate in size, finally dropping down to about zero as $|E(\mu_j|Y)|$ becomes small.

Figure 2(b) shows that the genes with large posterior variances are those with group mean differences that are intermediate in size. Thus, the jump is seen because the BAM systems and methods described herein shrink the posterior means while inflating the variance for these intermediate values, making it harder to reject the null $\mu_{j,0} = 0$. To better classify those genes with moderate values for $|E(\mu_j|Y)|$ the variance is adjusted to equal $n_j/n_{j,1}$. As illustration, consider the two vertical lines in Figure 2(a). The thin dashed line is the value for the 99.5 percentile from a standard $N(0, 1)$ distribution, while the thick dashed line denotes the same percentile from a $N(0, 2)$ distribution. Observe how the adjustment to the variance helps to avoid misclassifying genes.

It is instructive to work out what the posterior conditional variance is without the use of the centering method (5). It is expected that $\hat{\theta}_{j,n}$ be non-zero and that v_{2j-1}^2 be large. If $\mu_{j,0}$ is non-zero, then v_{2j}^2 will be large, and

hence

$$\hat{\Sigma}_{j,n}^{-1} \approx \begin{pmatrix} n_j/n_{j,1} & -\sqrt{n_j n_{j,2}/n_{j,1}} \\ -\sqrt{n_j n_{j,2}/n_{j,1}} & n_j/n_{j,1} \end{pmatrix}.$$

Note the resulting non-zero posterior correlation of $-\sqrt{n_{j,2}/n_j}$ between θ_j and μ_j .

3.3. The Zcut procedure. Formally, is referred to herein as the Zcut procedure, or simply, Zcut, includes the following method for identifying parameters $\mu_{j,0}$ which are not zero (and hence genes j which are expressing). Identify a gene j as differentially expressing if

$$|E(\mu_j|Y)| \geq z_{\alpha/2} \sqrt{n_j/n_{j,1}},$$

where $z_{\alpha/2}$ is the $100 \times (1 - \alpha/2)$ percentile of a standard normal distribution. The value $E(\mu_j|Y)$ is obtained by averaging Gibbs sampled draws
5 for μ_j .

In practice an appropriate way to select an α value for Zcut is employed. A technique used in Section 7 includes selecting α on the basis of a shrinkage plot, like Figure 2(a). In Section 7, α is chosen to coincide with the vertical
10 quantile so that most of the observations to its right will have posterior variance, $\text{Var}(\mu_j|Y)$, nearly equal to one. This removes many intermediate values which could inflate the FDR.

3.4. Robustness and adaptiveness. The Zcut technique is justified by way of a central limit theorem. This analysis will justify the adjustment to the variance used in Figure 2(a), identifying it with the asymptotic variance

for the conditional posterior mean. These results hold even when measurement errors are not normally distributed, assuming that appropriate moment conditions are satisfied. In the following theorem note carefully the degeneracy of the limiting distribution. This is a consequence of the
 5 centering method (5).

THEOREM 1. Assume that (4) represents the true model where $\epsilon_{j,k,l}$ are independent random variables such that $E(\epsilon_{j,k,l}) = 0$, $E(\epsilon_{j,k,l}^2) = \sigma_0^2$ and $E(|\epsilon_{j,k,l}^3|)$, $E(\epsilon_{j,k,l}^4) \leq C_0$ for some fixed constant $C_0 < \infty$. If $\sigma^2/n \rightarrow 1$ and $n_{j,1} \rightarrow \infty$ and $n_{j,2} \rightarrow \infty$ for each j such that $n_{j,2}/n_{j,1} \rightarrow r_{j,0} < \infty$, then under the null hypothesis $\mu_{j,0} = 0$, keeping (v_{2j-1}^2, v_{2j}^2) fixed,

$$(\hat{\theta}_{j,n}, \hat{\mu}_{j,n})^T \xrightarrow{d} Z_j \Sigma_j^{-1} (\sqrt{r_{j,0}}, \sqrt{1+r_{j,0}})^T, \quad j = 1, \dots, p,$$

where

$$\Sigma_j = \begin{pmatrix} 1 + 1/v_{2j-1}^2 & \sqrt{r_{j,0}/(1+r_{j,0})} \\ \sqrt{r_{j,0}/(1+r_{j,0})} & 1 + 1/v_{2j}^2 \end{pmatrix}$$

and Z_j are independent Normal(0, 1) random variables.

PROOF OF THEOREM 1. With a little bit of rearrangement in (6) it can be shown that

$$(\hat{\theta}_{j,n}, \hat{\mu}_{j,n})^T = M_{j,n} \sqrt{n_{j,2}} (\bar{Y}_{j,2} - \bar{Y}_{j,1}), \quad \text{where} \quad M_{j,n} = \frac{n \hat{\Sigma}_{j,n}^{-1}}{\hat{\sigma}_n \sigma^2} \begin{pmatrix} \sqrt{n_{j,2}/n_j} \\ 1 \end{pmatrix}^T. \quad (8)$$

Because the third moment of $\epsilon_{j,k,l}$ is bounded a Liapounov Central Limit Theorem may be applied to each of the group averages $\bar{Y}_{j,l}$. Use the fact that the averages are independent, and that $\sqrt{n_{j,2}/n_{j,1}} \rightarrow \sqrt{r_{j,0}}$, to deduce

that under the null, $\sqrt{n_{j,2}}(\bar{Y}_{j,2} - \bar{Y}_{j,1}) \overset{d}{\rightsquigarrow} \text{Normal}(0, \sigma_0^2(1+r_{j,0}))$. Meanwhile, some algebraic manipulation shows that

$$\hat{\sigma}_n^2 = \frac{1}{n-2p} \sum_{j,k,l} \epsilon_{j,k,l}^2 - \frac{1}{n-2p} \sum_{j,l} n_{j,l} \bar{\epsilon}_{j,l}^2,$$

where $\bar{\epsilon}_{j,l} = \sum_{k=1}^{n_{j,l}} \epsilon_{j,k,l} / n_{j,l}$. A bounded fourth moment implies that the first term on the right converges in probability to σ_0^2 , while for the second term, a second moment ensures that $\bar{\epsilon}_{j,l}^2 \xrightarrow{P} 0$ for each j and l . Hence it follows that $\hat{\sigma}_n^2 \xrightarrow{P} \sigma_0^2$. Therefore, since $\sigma^2/n \rightarrow 1$, it follows that

$$M_{j,n} \xrightarrow{P} \sigma_0^{-1} \Sigma_j^{-1} \left(\sqrt{r_{j,0}/(1+r_{j,0})}, 1 \right)^T.$$

Putting the pieces together and appealing to Slutsky's Theorem gives the desired result. Note that the degeneracy of the limit poses no problem by applying the Cramér Wold device. $\square$

Most of the conditions of Theorem 1 are likely to hold in practice. Relax-
 5 ing the assumption that errors are normally distributed and i.i.d is an especially helpful feature. The condition that $\sigma^2/n \rightarrow 1$ is quite realistic and understandable due to the rescaling employed by the systems and methods described herein. It holds quite accurately in many problems. For example, in the simulations presented in Figure 2(a), σ^2/n had a posterior
 10 mean of 0.96 with a posterior standard deviation of 0.04. The value for $r_{j,0}$ appearing in the theorem represents the limiting ratio of the group sizes. In the previous simulation $r_{j,0} \approx n_{j,2}/n_{j,1} = 1$. If gene j is really expressing, i.e. $\mu_{j,0} \neq 0$, then it is expected that v_{2j}^2 to be large, and hence Theorem 1 and an argument similar to (7) shows that $\hat{\mu}_{j,n}$ can be approx-
 15 imated in distribution by a $\text{Normal}(0, 1 + r_{j,0})$, or roughly a $\text{Normal}(0, 2)$,

- under the null. Notice how this matches up with the adjusted variance used in the simulation. In general, under the conditions of Theorem 1, $\hat{\mu}_{j,n}\sqrt{n_{j,1}/n_j}$ can be approximated by a standard normal distribution under the null when v_{2j}^2 is large. Integrating over the hyperparameters yields
- 5 $\mu_{j,n}^* := E(\mu_j|Y)\sqrt{n_{j,1}/n_j}$, which can be thought of as a "Bayes test statistic." Under the null it has a standard normal distribution. This is the rationale for the Zcut rule defined in Section 3.3. Theorem 1 shows that BAM is robust to underlying model assumptions and provides an explanation for the Zcut rule.

Theorem 2 below translates BAM's ability to adaptively shrink coefficients into a statement about risk performance. Shrinkage of coefficients implies shrinkage of test statistics. BAM's adaptiveness allows it to shrink the Bayes test statistics $\mu_{j,n}^*$ for coefficients $\mu_{j,0} = 0$ while allowing $\mu_{j,n}^*$ from coefficients $\mu_{j,0} \neq 0$ to have values similar to frequentist Z-tests

$$Z_{j,n} = \frac{\bar{Y}_{j,2} - \bar{Y}_{j,1}}{\hat{\sigma}_n \sqrt{1/n_{j,1} + 1/n_{j,2}}}. \quad (9)$$

- 10 This has a direct impact on performance. Since both $\mu_{j,n}^*$ and $Z_{j,n}$ are compared to a Normal(0,1) in assessing significance, Zcut will produce p-values that match up closely with Z-tests for non-zero coefficients, while p-values from the zero coefficients will be much larger. Recall that this p-value effect was seen in Figure 1. Figure 3a illustrates the effect in terms of
- 15 test statistics. Expressed genes are indicated by a cross, while non-expressed genes are identified by circles (these are the values mostly near zero on the x-axis which have been shrunk by BAM). This translates into power for Zcut and a low FDR.

Theorem 2 below quantifies this adaptiveness in terms of risk behavior. By

way of introducing some notation, let $\hat{d}_{j,\alpha} \in \{0, 1\}$ represent the classification rule for some two-sided test at a fixed α level. A value $\hat{d}_{j,\alpha} = 1$ means the null $\mu_{j,0} = 0$ is rejected, otherwise if $\hat{d}_{j,\alpha} = 0$ the null is accepted. Let

$$R_{j,\alpha} = \Pr\{\hat{d}_{j,\alpha} = 1, \mu_{j,0} = 0\} + \Pr\{\hat{d}_{j,\alpha} = 0, \mu_{j,0} \neq 0\}.$$

This is the expected risk for gene j . The total expected risk for all genes is $R(\alpha) = \sum_{j=1}^p R_{j,\alpha}$. The average $R(\alpha)/p$ is the misclassification rate. Write $R_B(\alpha)$ and $R_F(\alpha)$ for the total expected risk using $\mu_{j,n}^*$ and $Z_{j,n}$ respectively for some fixed α value.

- 5 THEOREM 2. Assume that (4) represents the true model where $\epsilon_{j,k,l}$ are i.i.d Normal(0, σ_0^2) variables. Suppose that $\sigma^2 = n$. Then for each $0 < \delta < 1/2$ there exists values (v_{2j-1}^2, v_{2j}^2) for $j = 1, \dots, p$ such that $R_B(\alpha) < R_F(\alpha)$ for each $\alpha \in [\delta, 1 - \delta]$.

PROOF OF THEOREM 2. Let $I_0 = \{j : \mu_{j,0} = 0\}$ denote the indices for the zero coefficients. Define p_0 to be the cardinality of I_0 . Since errors are assumed to be normally distributed, (9) implies that $Z_{j,n} \stackrel{D}{=} Z_j C_n$, where $C_n = \sigma_0/\hat{\sigma}_n$ and Z_j are independent Normal($\mu_{j,0}/S_{0,n}$, 1) variables where $S_{0,n} = \sigma_0 \sqrt{1/n_{j,1} + 1/n_{j,2}}$. It follows that

$$\begin{aligned} R_F(\alpha) &= \sum_{j \in I_0} \Pr\{|Z_{j,n}| \geq z_{\alpha/2}\} + \sum_{j \in I_0^c} \Pr\{|Z_{j,n}| < z_{\alpha/2}\} \\ &= p_0 \Pr\{|N(0, 1)| \geq C_n^{-1} z_{\alpha/2}\} + \sum_{j \in I_0^c} \Pr\{|Z_j| < C_n^{-1} z_{\alpha/2}\}. \end{aligned}$$

Recall that $(\hat{\theta}_{j,n}, \hat{\mu}_{j,n})^T$ can be written as (8). One can show that (remember

$\sigma^2 = n$):

$$\hat{\sigma}_n M_{j,n} \rightarrow (0, v_{2j}^2 / (1 + v_{2j}^2))^T, \quad \text{as } v_{2j-1}^2 \rightarrow 0.$$

Because $\mu_{j,n}^* = \hat{\mu}_{j,n} \sqrt{n_{j,1}/n_j}$, use (8) and the definition of $Z_{j,n}$ to deduce that for each $0 < \xi_j < 1$ we can find a v_{2j-1}^2 and v_{2j}^2 such that $\mu_{j,n}^* = Z_{j,n} \xi_j$. In particular, this means that for each $0 < \delta_1 < 1$ and $0 < \delta_2 < 1$, (v_{2j-1}^2, v_{2j}^2) may be found such that $\mu_{j,n}^* = Z_{j,n} \delta_1$ for $j \in I_0$ and $\mu_{j,n}^* = Z_{j,n} \delta_2$ for $j \in I_0^c$. Therefore,

$$\begin{aligned} R_F(\alpha) - R_B(\alpha) &= p_0 \left(\Pr\{|N(0, 1)| \geq C_n^{-1} z_{\alpha/2}\} - \Pr\{|N(0, 1)| \geq \delta_1^{-1} C_n^{-1} z_{\alpha/2}\} \right) \\ &\quad + \sum_{j \in I_0^c} \left(\Pr\{|Z_j| < C_n^{-1} z_{\alpha/2}\} - \Pr\{|Z_j| < \delta_2^{-1} C_n^{-1} z_{\alpha/2}\} \right). \end{aligned}$$

Both sums on the right-hand side are continuous functions of α and each has a minimum and maximum over $\alpha \in [\delta, 1 - \delta]$. In particular, the second sum may be made arbitrarily close to zero uniformly over $\alpha \in [\delta, 1 - \delta]$ by choosing δ_2 close to one, while the first sum remains positive and uniformly bounded away from zero over $\alpha \in [\delta, 1 - \delta]$. Thus, for a suitable δ_2 , $R_F(\alpha) - R_B(\alpha) > 0$ for each $\alpha \in [\delta, 1 - \delta]$. $\square$

Theorem 2 shows that over a wide range of α values, with suitably selected values of (v_{2j-1}^2, v_{2j}^2) , the expected total risk for Zcut is less than that for the classification rule derived from Z-tests. The conditions of the theorem are fairly reasonable. The restriction to normally distributed errors is made for convenience since a central limit theorem as in Theorem 1 generally applies in practice. Section 6 will provide several simulations verifying Zcut's better risk performance over the use of Z-tests.

As mentioned earlier, there exists a generalization of the two-group BAM shrinkage plot extended to multiple groups. In one embodiment, a particular group is designated as a baseline, and individual two-dimensional (2-D) plots of pair-wise BAM test statistic values (also called Bayes test statistics) which are group effect test statistics against the designated baseline group are plotted. Accordingly, a number of these 2-D plots are generated based, at least in part, on the total number of groups under study. Note that the posterior variance on these is typically not actually plotted, but rather a cutoff rule based on the posterior variance coalescing to a value of 1 is used as a way to indicate which genes are being significantly turned on or off on these 2-D plots.

For example, and without limitation, if groups A, B, C and D are of interest, and group A is designated to be the baseline, the systems and methods described herein generate the following 3 BAM test statistic by group plots:

BA vs CA
CA vs DA
BA vs DA,

where the horizontal and vertical axes depict the BAM test statistic values for the group effect of comparing some group relative to the baseline group A. Moreover, as shown in Figure 3b, points are colored (e.g., magenta for genes that are significantly turned on or off across substantially all groups, and cyan and green for hit-and-run genes) based on whether a gene has a posterior variance coalesced to about 1: something analogous to what is done in the initial two-group shrinkage plot.

This can be further enhanced by choosing one of the generated plots and overlaying primarily significant genes with respect to the omitted group using a different plotting symbol but maintaining the original color scheme. Note that this overlay strategy (or a similar one suitably adapted to the
5 number of groups) is resorted to typically when there are more than 3 groups being studied.

In multigroup problems, typically more than one parameter is estimated for each gene. In the example above, there are 3 parameter estimates of gene effect for each gene. Traditional least square test statistics (Z-test) are
10 typically positively correlated, and this correlation induces errors in classifying genes as being differentially expressed or not (i.e., a mistake made with classifying one of the gene-specific effects, increases the likelihood of making mistakes on the others for that gene). BAM shrinkage reduces this correlation to about zero, and hence mitigates the effects of what is
15 referred to herein as "regression to the mean." This combined with the original shrinkage of the test statistic point estimates themselves typically yields uniformly superior risk performance in multigroup problems.

As an illustration of this, consider a multigroup analysis of colon cancer data (a problem to be discussed later in this specification). In this con-
20 text, patterns of differential gene expression across four groups are sought: BSurvivors (the Stage 2 survivors), C (Stage 3), D (Stage 4) and METS (tumors metastasized to the liver). Figures 3b and 3c depict a BAM test statistic by group plot and a Z-test by group plot, respectively. In an embodiment shown in Figures 3b and 3c, the BSurvivors group is selected as
25 a baseline and two other stages are chosen for plotting (even though the actual analysis was run with all 4 stages). That is, the overlaying plotting option is bypassed for illustration purposes, graphical clarity, and expedi-

ency.

Figure 3b depicts a BAM test statistic (also known as Bayes test statistic) by group plot. Gene-specific BAM test statistic estimates are shown. In this plot, the METS vs BSurvivors effect is plotted against the D vs BSurvivors effect. Clearly, a different plot emerges from the one generated using Z-test. Shrinkage is evident along the coordinate axes (i.e. shrinkage of the corresponding Z-test values shown in Figure 3c) as well as in the marked non-elliptical shape of the plot (shrinkage of the correlation between gene-specific Z-test values). Coloring here is used to indicate different types of patterns of interest in this multigroup framework: magenta indicates genes significantly turned on or off across the stages and cyan and green indicate hit-and-run genes. Apriori, obvious hit-and-run genes would be ones in cyan and green that hug either coordinate axes (i.e., early or late hit-and-run). These are clearly visible in the Figure. Contrast this with Figure 3c.

Figure 3c depicts a Z-test by group plot. The analysis is substantially similar to that presented for Figure 3b, except the use of standard least square estimates. Clearly hit and run genes (potentially in quadrants 2, 4, 6 and 8) are not visible since they would be ones hugging the coordinate axes in these quadrants. Additionally, note the elliptical shape of the plot indicating positive correlation between the gene-specific effect estimates. Also note that quadrants here were generated by using the simple ± 2 cutoff rule for statistical significance and are numbered going clockwise starting from 1 in the bottom right-hand quadrant.

4. Generating the null: FDRmix. Values estimated from BAM may also

be used in a hybrid BH procedure as a method for controlling the FDR. This new model selection procedure is referred to herein as FDRmix. Like the Zcut procedure, FDRmix selects models based on posterior values for μ_j . In this approach, $T_j = E(\mu_j|Y)$, the posterior mean value for μ_j , is used
 5 as the test statistic in selecting models.

To implement FDRmix, the conditional density for T_j under the null hypothesis $\mu_{j,0} = 0$, i.e., $F_0(dT_j)$, is determined. Although an exact derivation difficult, an accurate and simple approximation for F_0 may be derived by considering $\hat{\mu}_{j,n}$ given earlier in (6). Suppose that the null is true. Although the posterior should identify gene j as having a mean of zero, there will still be a positive posterior probability of a large ν_{2j}^2 and a resulting misclassification. Equation (7) identifies T_j under this scenario. Suppose that the data is balanced with fixed group sizes so that $n_{j,1} = N_1$ and $n_{j,2} = N_2$. Then, under the null, conditioning on the event $A_j = \{\nu_{2j}^2 \text{ is large}\}$,

$$(T_j|A_j, \mu_{j,0} = 0) \approx \frac{\sqrt{N_2}}{\hat{\sigma}_n}(\bar{Y}_{j,2} - \bar{Y}_{j,1}) \approx \text{Normal}(0, (N_1 + N_2)/N_1).$$

On the other hand, if the posterior correctly identifies that gene j is not expressing, i.e. that $A_j^c = \{\nu_{2j}^2 \approx 0\}$, then (6) suggests that

$$(T_j|A_j^c, \mu_{j,0} = 0) \approx \frac{\sqrt{N_2}}{\hat{\sigma}_n(1 + 1/\nu_{2j}^2)}(\bar{Y}_{j,2} - \bar{Y}_{j,1})$$

for some small value of ν_{2j}^2 . Under the null this also has a normal distribution with mean zero, but unlike the previous case it is not so clear what the theoretical variance is.

These arguments suggest that the null distribution of T_j may be approxi-

mated using the two-point normal mixture,

$$F_0(dT_j) \approx (1 - \Pi_0)\phi(T_j|V_1) + \Pi_0\phi(T_j|V_2),$$

where $\phi(\cdot|V)$ denotes a normal density with mean zero and variance V . It is anticipated that $V_2 = (N_1 + N_2)/N_1$ but the values for V_1 and $\Pi_0 = \Pr\{A_j|\mu_{j,0} = 0\}$ are unknown. All of these values, however, may be estimated by fitting a two-point normal mixture to simulated data. Thus, to estimate V_1 , V_2 and Π_0 data from the model (4) may be simulated, where $\mu_{j,0} = 0$ for all $j = 1, \dots, p$ (typically p is chosen to be some large number, e.g., 25000). Notice this simulation requires little, if anything, more than knowing the sample sizes N_1 and N_2 and the value for σ_0^2 , which may be estimated accurately from the original data. The BAM systems and methods described herein are then applied to the simulated data and a two-point mixture model is fitted to the averaged values for μ_j collected from the Gibbs output. The results from fitting the mixture can now be used to derive p-values which are then analyzed in the usual way by the BH method to determine which hypothesis to reject. To compute the p-values suppose that T_j^o is the estimated value for $E(\mu_j|Y)$ from the original (non-simulated) data. Then its p-value P_j may be approximated by

$$P_j = 2\Pr\{|T_j^o| < T_j|\mu_{j,0} = 0\} \approx 2 \left(1 - \Pi_0\Phi\left(|T_j^o|/\sqrt{V_1}\right) - (1 - \Pi_0)\Phi\left(|T_j^o|/\sqrt{V_2}\right) \right),$$

where $\Phi(\cdot)$ denotes a standard normal cdf and the expectation on the left is over the null.

- 5. Comparison procedures.** BAM was tested against several different model selection procedures. These included what are called "Bayes Exch,"
- 5 "Z-test," "Bonf," and "BH." Below is a brief description of these methods.

5.1. Z-test. Here genes are identified by Z-test statistics defined earlier in (9). Z-test identifies gene j as expressing if $|Z_{j,n}| \geq z_{\alpha/2}$.

5.2. Bonf. The Bonf procedure is a Bonferroni correction to Z-test. Thus gene j is identified as expressing if $|Z_{j,n}| \geq z_{\alpha/(2p)}$.

5.3. BH. The BH procedure was also applied. The p-values used were based on the test statistics $Z_{j,n}$ under a two-sided test. Thus, if

$$P_j = 2\Pr\{\text{Normal}(0, 1) > |Z_{j,n}|\} ,$$

- 5 then gene j is considered expressing if $P_j \leq P_{(k)}$, where $P_{(k)}$ is the k -th ordered p-value where $k = \max\{j : P_j \leq j\alpha/p\}$ and $\alpha > 0$ is some pre-chosen target FDR. Although the original BH procedure was designed to control the expected FDR assuming independent null hypotheses, it has been shown that the method applies under certain types of dependencies.
- 10 For example, the method computed from dependent Z-tests such as $Z_{j,n}$ will control the expected FDR if applied to those $Z_{j,n}$ for which the null is true. Thus, when many nulls are true, this approximately controls the expected FDR.

5.4. Bayes Exch. Also implemented was a simple Bayesian exchangeable model. To model (4), the data $Y_{j,k,l}$ is replaced with sufficient statistics $\bar{Y}_{j,2} - \bar{Y}_{j,1}$. Conditional on $\hat{\sigma}_n^2$ this generally yields a $\text{Normal}(\mu_j, \hat{\sigma}_n^2(1/n_{j,1} + 1/n_{j,2}))$ distribution. Thus, to identify genes the empirical hierarchical model

was used

$$\begin{aligned}(\bar{Y}_{j,2} - \bar{Y}_{j,1} | \mu_j, \hat{\sigma}_n^2) &\sim \text{Normal}(\mu_j, \hat{\sigma}_n^2(1/n_{j,1} + 1/n_{j,2})) \\(\mu_j | \tau_0^2) &\sim \text{Normal}(0, \tau_0^2) \\(\tau_0^{-2} | t_1, t_2) &\sim \text{Gamma}(t_1, t_2),\end{aligned}$$

where $t_1 = t_2 = 0.0001$ was selected to ensure a non-informative prior for τ_0^2 . This extends the models considered by other researchers, by allowing for a hyperprior on τ_0^2 .

Observe that given the data Y the hyperparameter τ_0^2 and the estimate $\hat{\sigma}_n^2$,

$$(\mu_j | Y, \hat{\sigma}_n^2, \tau_0^2) \sim \text{Normal}(\tau_0^2(\bar{Y}_{j,2} - \bar{Y}_{j,1}) / (S_{j,n}^2 + \tau_0^2), S_{j,n}^2 \tau_0^2 / (S_{j,n}^2 + \tau_0^2)),$$

where $S_{j,n}^2 = \hat{\sigma}_n^2(1/n_{j,1} + 1/n_{j,2})$. Thus, a reasonable procedure identifies gene j as expressing if

$$|\bar{Y}_{j,2} - \bar{Y}_{j,1}| \geq z_{\alpha/2} S_{j,n} \sqrt{1 + S_{j,n}^2 / \hat{\tau}_0^2}, \quad (10)$$

where $\hat{\tau}_0^2$ is some estimate for τ_0^2 . In all examples considered, $\hat{\tau}_0^2 = E(\tau_0^2 | Y, \hat{\sigma}_n^2)$.

- 5 Notice that when $\hat{\tau}_0^2 \rightarrow \infty$ the limit of the test (10) corresponds to a standard Z-test. However, whenever $\hat{\tau}_0^2 < \infty$, the Bayesian test (10) is more conservative.

6. Simulations. To assess the procedures, they were tested simulated data from the ANOVA model (4). Means for genes were set to have a group mean separation of $\mu_{j,0} = 1.5$ and $\mu_{j,0} = 0$ for 10% and 90% of parameters respectively. This corresponds to a model in which 10% of

genes are expressing and represents a fairly realistic scenario for microarray data. For convenience $\theta_{j,0}$ were set to zero for all j . Thus, gene simulation model was characterized by

$$Y_{j,k,l} = \mu_{j,0}I\{l = 2\} + \epsilon_{j,k,l}, \quad j = 1, \dots, p, \quad k = 1, \dots, n_{j,l}, \quad l = 1, 2,$$

where $\epsilon_{j,k,l}$ were taken to be i.i.d $N(0, \sigma_0^2)$ variables. For the variance, σ_0^2 were set to one. Group sizes were fixed at $n_{j,1} = n_{j,2} = 5$ and number of genes at $p = 25000$ (the simulation reported in Table 1 and Figure 1 used the same configuration as above but with $p = 10000$).

- 5 All model estimators reported were based on a range of preset α levels. For Zcut, Bayes Exch, Z-test and Bonf, a preset α value meant that the corresponding test was computed based on the appropriate quantile for a standard normal (for example, Z-test was computed using $z_{\alpha/2}$ as a cutoff while Bonf used $z_{\alpha/(2p)}$). For FDRmix and BH the α value used was the
- 10 target FDR. To fairly compare the procedures under the same α values the total number of misclassified observations, TotalMiss, discussed in the Introduction, were recorded. Also recorded were the FDR, FNR and the Type-I and Type-II error rates.

The results of the simulations are reported in Table 2 and Figure 4. From

15 these the following general observations may be made:

1. Z-test has the best total risk performance for small α values, but its good risk behavior is only local. For example, once α increases the value for TotalMiss increases rapidly and procedures like Zcut and FDRmix quickly have superior risk performance. Furthermore, if α
- 20 becomes very small, its TotalMiss values also shoot up. Thus, trying

Table 1: Results from gene simulation model ($\alpha = 0.05$).

	Detected	TotalMiss	FDR	FNR	Type-I	Type-II
Zcut	481	675	0.162	0.063	0.009	0.597
FDRmix	416	702	0.142	0.067	0.007	0.643
Z-test	1106	792	0.406	0.039	0.049	0.343
BH	148	862	0.034	0.087	0.001	0.857

Table 2: Gene simulation model $p = 25000$ with 10% of genes expressing. Estimators based on α levels of 0.05, 0.01, 0.001 respectively.

	Detected	TotalMiss	FDR	FNR	Type-I	Type-II
Zcut	1286	1700	0.189	0.061	0.011	0.583
	822	1824	0.088	0.072	0.003	0.700
	281	2231	0.021	0.090	0.000	0.890
FDRmix	1148	1724	0.162	0.064	0.008	0.615
	719	1885	0.072	0.076	0.002	0.733
	72	2430	0.014	0.097	0.000	0.972
Bayes Exch	498	2048	0.046	0.083	0.001	0.810
	78	2424	0.013	0.097	0.000	0.969
	3	2497	0.000	0.099	0.000	0.999
Z-test	2808	2006	0.412	0.038	0.051	0.340
	1245	1709	0.182	0.062	0.010	0.593
	453	2081	0.038	0.084	0.001	0.826
Bonf	18	2482	0.000	0.099	0.000	0.993
	10	2490	0.000	0.099	0.000	0.996
	2	2498	0.000	0.099	0.000	0.999
BH	415	2115	0.036	0.085	0.001	0.840
	116	2388	0.017	0.096	0.000	0.954
	11	2489	0.000	0.099	0.000	0.996

- to find the small region of α values where Z-test does well is a tricky proposition. No theory exists to select this region, which makes Z-test unreliable. Overall, Zcut and FDRmix have the best total risk behavior. For small to moderate α values Zcut is better, while FDRmix is better for large α . BH has poorer performance for small to moderate α , only approaching the same performance as Zcut and FDRmix for α as large as 0.2. This confirms its inability to simultaneously optimize the FDR and FNR. Bayes Exch tracks BH in terms of total risk over small to moderately large α .
- 5
- 10
- 15
- 20
- 25
2. FDRmix does not achieve target α values but does further reduce the FDR for Zcut as designed. The FDR for BH is better, but is generally smaller than its target value. When α becomes small its FDR drops to zero and it becomes too conservative.
 3. Among all the procedures tested Zcut has Type-II error rates closest to those observed for Z-test.
 4. The worst procedure overall is Bonf whose TotalMiss plot is flat. Alternatives to Bonferroni corrections have been attempted for microarrays. These often involve nonparametric stepdown adjustments to raw p-values using permutation distributions of the test statistics. Modest power gains have been seen but coarseness of the permutation distributions limit the usefulness of these approaches to situations where a large enough number of samples are available. To illustrate, an adjusted p-value method was applied, and it was found that with α equal to 0.05, 0.01 and 0.001, 18, 10 and 2 significant genes were found, respectively. This performance mirrors that of the Bonferroni method. Since step-down adjustment methods also attempt to control family wise error rates, it's not surprising at how conservative they were found to be.

Interestingly, although Figures 4a-4b show that Z-test may be quite volatile near its optimal α value, the plots suggest that Z-test nevertheless outperforms Zcut for smaller α values. Thus, Z-test's volatility may not be a problem as long as one adopts the strategy of choosing a small α value. In fact, this can be a bad strategy as the TotalMiss values for Z-test can sometimes rise rapidly even at near zero values for α . To illustrate this behavior the signal to noise ratio of the previous simulation was reduced by setting $\mu_{j,0} = 0.5$ for the 10% of genes that were expressing. All other values of the simulation were kept the essentially the same.

- Figure 5a records the TotalMiss values for Z-test and Zcut from the analysis by depicting the total number of misclassified observations from the gene simulation model used in Figure 4, but with smaller signal-to-noise ratio. The plot clearly shows that Zcut has superior risk performance over Z-test even at near zero α values. Notice also how the TotalMiss values for Z-test increase rapidly due to overfitting, while the risk performance for Zcut stabilizes. The better risk performance by Zcut is due, at least in part, to its adaptive shrinkage ability as predicted by Theorem 2.

Figure 5b shows the extent to which BAM is able to shrink test statistics compared to the Z-test. The figure plots the BAM test statistics against Z-tests similar to Figure 3a. An O denotes Zcut, and "x" denotes Z-test. It is instructive to compare this to Figure 3a to see how much shrinkage is actually going on.

- 6.1. Unequal Variances: Poisson Gene Simulation Model. Untransformed gene expression data often fail to satisfy the assumptions of an equal variance model. A pattern often seen is that the variance for a gene expression is

related to its mean expression. Often the relationship is linear, but sometimes it may be quite complex, such as in the case of colon cancer (Section 7). To deal with these problems, it is generally recommended that the data be transformed by a variance stabilizing transformation.

To study how well the BAM systems and methods described herein perform under the scenario wherein variances are proportional to the mean, and also to see how robust they are in light of an assumption of normality, data from the Poisson model

$$Y_{j,k,l} = \xi_{j,k,l} + \epsilon_{j,k,l}, \quad j = 1, \dots, p, \quad k = 1, \dots, n_{j,l}, \quad l = 1, 2,$$

- 5 were simulated, where $\xi_{j,k,l}$ are independent $\text{Poisson}(\mu_{j,l})$ variables, independent of $\epsilon_{j,k,l}$, the $N(0, \sigma_0^2)$ measurement errors. The variances $\sigma_0^2 = 0.01$ were set to a small value. This slightly smooths the data, although at the same time the variance for a gene expression will be proportional to its mean. Figure 6(a) shows gene mean values $\bar{Y}_{j,l}$ plotted against gene sample variances $\hat{\sigma}_{j,l}^2$. The left-hand side is group $l = 1$, while the right-hand side is group $l = 2$.

- Expressing genes are identified by a cross, while non-expressing genes by a circle.. Gene group sizes $n_{j,1} = 21$ and $n_{j,2} = 34$ were used, and the number of genes was set at $p = 60000$. These sample sizes were selected
 15 to match those of the colon cancer data set. About 90% of the genes were set to have equal means over both groups, so that for these genes $\mu_{j,l} = \mu_j$, where μ_j was drawn randomly from a uniform distribution on $(1, 3)$. For the remaining 10% of the genes (the expressors), the group one means $\mu_{j,1}$ were sampled independently from a uniform distribution on $(4, 5)$. Then
 20 the group two means $\mu_{j,2}$ were set to this value, incrementing by either +1

or by -1 randomly with equal probability. This corresponds to genes for the treatment turning on or off.

The better performance of the methods on the untransformed data can be explained by Figure 6(b), which shows group mean differences $\bar{Y}_{j,2} - \bar{Y}_{j,1}$ versus standard errors. The left-hand side includes the untransformed data, whereas the right-hand side is computed from the variance-stabilized data. On the left-hand side the untransformed data is plotted for the group mean difference $\bar{Y}_{j,2} - \bar{Y}_{j,1}$ for each gene versus the corresponding standard error $(\hat{\sigma}_{j,1}^2/n_{j,1} + \hat{\sigma}_{j,2}^2/n_{j,2})^{1/2}$, where

$$\hat{\sigma}_{j,l}^2 = \frac{1}{n_{j,l} - 1} \sum_{k=1}^{n_{j,l}} (Y_{j,k,l} - \bar{Y}_{j,l})^2,$$

is the sample variance for gene j over group l .

It is seen that even though the standard error is not constant over the group mean differences (compare this to the square-root transformed plot on the right-hand side), its value still remains fairly constant, only becoming elevated for large group mean differences, where it remains relatively small in magnitude compared to the mean effect. Thus, even though in the untransformed data the gene sample variances are proportional to their means (Figure 6(a)), once the sample size is taken into account when computing the standard error, the effect of unequal variances gets washed out. It seems better to simply not transform the data, since a side effect is that means are compressed, leading to smaller models.

The results of the Poisson simulation are reported in Figure 7, which plots the total number of misclassified observations as α is varied. Figure 7(a)

is from the untransformed data, while Figure 7(b) are results from the data under a square-root variance stabilizing transformation. The following conclusions may be drawn:

1. The variance stabilizing transform reduces the size of models across essentially all estimators with fewer genes being identified as expressing. In essentially all cases the value for TotalMiss increases under the transformation.
2. Estimators generally exhibit the same patterns of behavior as in the previous simulation.

6.2. Dependence: Correlations Between Genes. The simulations up to this point have generally assumed that genes are independent, but in practice it is more realistic to expect gene expressions to be correlated. A gene's expression across different samples may be safely assumed to be independent in experimental designs in which for a fixed j , the values of $Y_{j,k,l}$ represent expressions measured over different tissue types l for different samples k . Within a given sample k , however, different genes can cluster into small groups — often a manifestation of gene proximities along biological pathways. This has been termed "clumpy dependence," and is the most likely scenario of dependence for these types of experimental designs. To study how the procedures perform under this form of dependence data were simulated according to the model

$$Y_{j,k,l} = \mu_{j,0}I\{l = 2\} + \zeta_{m_j,k,l} + \epsilon_{j,k,l}, \quad j = 1, \dots, p, \quad k = 1, \dots, n_{j,l}, \quad l = 1, 2$$

- 10 where $\epsilon_{j,k,l}$ are i.i.d $N(0, \sigma_0^2)$ measurement errors, independent of $\zeta_{m_j,k,l}$. The same dimensions and group sizes were used as in the earlier set of

- simulations: $p = 25000$, $n_{j,1} = n_{j,2} = 5$. Variables $\zeta_{m_j,k,l}$ were introduced to induce dependence amongst genes into small "blocks." Here, $m_j = \lceil j/B \rceil$, where $B = 50$ and $\lceil j/50 \rceil$ represents the first integer greater than or equal to $j/50$. Thus, for each fixed k and l , there were 500 variables $\zeta_{1,k,l}, \dots, \zeta_{500,k,l}$.
- 5 These each induce a block of B dependent observations. For example, "block m " includes those $Y_{j,k,l}$ where $\lceil j/B \rceil = m$, which are dependent since they share the common value $\zeta_{m,k,l}$. Variables across different values for k and l were assumed independent. That is, $\zeta_{j,k,l}$ was independent of $\zeta_{j,k',l'}$ if $k \neq k'$ or $l \neq l'$. This ensured that a gene's expression was independent across samples.
- 10

Table 3: Dependent gene simulation model. Estimators based on α levels of 0.05, 0.01, 0.001.

	Detected	TotalMiss	FDR	FNR	Type-I	Type-II
Zcut	1299	1702	0.191	0.061	0.011	0.581
	807	1868	0.109	0.074	0.004	0.712
	286	2230	0.029	0.089	0.000	0.889
FDRmix	1146	1738	0.166	0.065	0.008	0.618
	686	1939	0.094	0.077	0.003	0.751
	33	2467	0.000	0.099	0.000	0.987
Z-test	2762	1979	0.404	0.039	0.049	0.343
	1269	1698	0.184	0.062	0.010	0.586
	473	2078	0.055	0.084	0.001	0.821
BH	403	2129	0.029	0.086	0.001	0.845
	66	2435	0.000	0.098	0.000	0.974
	4	2496	0.000	0.099	0.000	0.999

The first 2500 (10%) of genes were set to have non-zero $\mu_{j,0}$ coefficients and the remaining 90% to have zero coefficients. This ensured that blocks were made up entirely of expressing genes or entirely of non-expressing genes. All $\zeta_{m_j,k,l}$ variables were drawn from a $\text{Normal}(0, \eta_0^2)$ distribution.

The values $\eta_0^2 = 19$ and $\sigma_0^2 = 1$ were selected to induce a correlation of $\rho_0 = 0.95$ between observations within the same block. So that the signal-to-noise ratio would be similar to the earlier simulation, $\mu_{j,0} = 1.5/\sqrt{1 - \rho_0}$ were selected for the expressing genes.

- 5 Table 3 presents the results in a style similar to Table 2 for direct comparison. Values reported in Table 3 were obtained by averaging over 100 independent simulations. For brevity, only the results for Zcut, FDRmix, Zcut and BH are reported. The values for TotalMiss and Detected were rounded to the nearest integer.
- 10 Comparing Table 3 to Table 2 the following observations may be made:
 1. The FDR procedures, FDRmix and BH, start to break down as α decreases. The number of detected genes drops rapidly and the FDR becomes essentially zero. FDR procedures will have high variability in dependence scenarios and are less reliable for low α than for
 15 high α . Techniques are available for correcting FDR methods under dependence.
 2. On the other hand, Zcut and Z-test have performance measurements similar to those seen in Table 2. Overall, clumpy dependence seems to have a minimal effect on them. This is partly explained in that
 20 these procedures classify genes based on test statistics that are affected by dependence mainly through $\hat{\sigma}_n^2$. As long as block sizes B are relatively small compared to p (a likely scenario with microarray data), the effect of dependence is marginal since $\hat{\sigma}_n^2$ is a robust estimator for the variance, $\sigma_0^2 + \eta_0^2$.

7. Metastatic signatures for colon cancer. A practical illustration of the BAM systems and methods is, as discussed above, detection of a metastatic gene expression signature for colon cancer. This problem is of significant medical importance. Colorectal cancer is the second leading cause of cancer mortality in the adult American population, accounting for 140,000 new cases annually and 60,000 deaths. The current clinical classification of colon cancer is based on the anatomic extent of the disease at the time of resection. It is known that this classification scheme is highly imperfect in reflecting the actual underlying molecular determinants of colon cancer behavior. For instance, upwards of 20% of patients whose cancers metastasize to the liver are not given life saving adjuvant chemotherapy based on the current clinical staging system. Thus, there is an important need for the identification of a molecular signature that will identify tumors that metastasize. In addition, this signature will no doubt suggest new targets for the development of novel therapies.

The gene expression considered herein is from the Ireland Cancer Center at Case Western Reserve University. The Center has collected a large database of gene arrays on liver metastasized colon cancers (METS), modified Duke's B1 and B2 survivors (B-survivors) as expressed by the Astler-Coller-Dukes staging system (Cohen et al., 1997), and normal colon tissue samples. The B-survivors represent an intermediate stage of tumor development with metastasis to the liver being seen as the disease worsens. As of February, 2002, there were 76 total samples available for analysis made up of 34 METS, 21 B-survivors and 21 normals.

The gene arrays used in compiling the database were EOS Biotechnology 59K-on-one chips, which were Affymetrix chips modified for using a smaller subset of eight more sensitive probes for each gene than do the

original Affy chips which use 20 probes per gene. An advantage of such a system is the ability to assay many more genes on each chip. In fact, yields of up to about 60,000 pieces of genetic information are capable of being processed on each chip. Generally, the genes are treated as independent,
 5 realizing that there is still an open debate as to the total number of actual genes in the human genome.

The analysis was based on the data for the liver METS and B-survivor tissue samples (normal tissue samples were excluded as a primary interest was understanding cancer evolving from the intermediate B-survivor
 10 stage). Using the earlier notation, group sizes were $n_{j,1} = 21$ for B-survivors (the control group) and $n_{j,2} = 34$ for the liver METS (treatment group). In total there were $n_j = 55$ samples for each of $p = 59341$ probe sets.

Figures 8(a) and 8(b) plot different summary values for the data. Figure 8(a) shows the gene mean values $\bar{Y}_{j,l}$ versus gene standard deviations $\hat{\sigma}_{j,l}$ from
 15 colon cancer data. Primarily mean values less than about 30 are plotted, to help zoom in plot. Here the mean for a specific gene is plotted against its standard deviation in order to identify any potential mean-variance relationship. The figure reveals a pattern that at first glance appears to be linear. However, careful inspection shows that although standard deviations
 20 are generally increasing with the mean expression for a gene, one can also see a broad strip of values where standard deviations are essentially constant. Thus, the pattern here is one of overdispersed variances.

This raises the question of how to handle the unequal variances. One approach includes trying different transformations in the hope of stabilizing
 25 the variance. Consider the right-hand side of Figure 8(b), which is the result of the log-transformation $\log(Y_{j,k,l} + 1 + \Delta)$, where $\Delta = |\min\{Y_{j,k,l}\}|$; the

- constant $1 + \Delta$ has been added to ensure that gene expressions were strictly larger than one so that logs for small values were not unduly inflated. The figure shows the group mean differences $\bar{Y}_{j,2} - \bar{Y}_{j,1}$ versus standard errors. The left-hand side is computed from untransformed data, while the right-
- 5 hand side is computed from the log-transformed data. For untransformed data, mean differences less than about 5 in absolute value are plotted to help zoom in plot. As may be seen, the transformation has helped to stabilize the standard error when compared to the mean differences between the two tissue groups.
- 10 As was discussed in Section 6.1, the stability of the standard error is an important criterion for assessing validity of inference based on an equal variance model. However, as also discussed, the gain in stabilizing standard errors by a transformation may sometimes be offset by the manner in which mean values are nonlinearly transformed. Evidence that this is
- 15 happening for the data can be seen from Figure 8(c), which shows genes found significant using Zcut with $\alpha = 0.0005$. Crosses denote genes detected from untransformed data, while circles are from log-transformed data. In other words, in the figure are plotted the genes detected by Zcut on both the untransformed and log-transformed data using the two group
- 20 ANOVA model of Section 3 (Zcut based on an $\alpha = 0.0005$ cutoff criteria). As can be seen, the log-transform appears to be encouraging Zcut to pick genes with very small mean differences; thus, the effect of the transform is to overly enhance the signal-to-noise ratio for these genes. Such problems with the log-transform have been noticed elsewhere, and some researchers
- 25 argue against the use of log-transforms in analysis of gene expression data.

However, some form of variance stabilization is called for here as the left-hand side of Figure 8(b) reveals that the standard errors for the untrans-

)

formed data are quite variable, certainly more than what was seen in the Poisson simulations of Section 6.1. Global transformations to the data lead to undesirable results. Therefore, the classical method of weighted regression was relied on to deal with heteroscedasticity. In this approach, genes
5 were grouped into $C = 2$ clusters, depending on whether the standard error for a gene was less than or greater than about the 99th percentile for standard errors. This allows treatment of the small group of genes with very high variability differently from the remaining genes. For each group, gene expressions $Y_{j,k,l}$ were then rescaled by dividing by the square root
10 of the group mean-square error (the unbiased estimator for the variance). The BAM systems and methods described herein were then applied to the newly rescaled data. It should be noted that this method may be applied for any number of clusters C , although it is generally recommended that the data undergo as little transformation as possible in trying to reduce
15 variability.

Figure 9 depicts estimated values for $|E(\mu_j|Y)|$ versus $\text{Var}(\mu_j|Y)$ using weighted regression. Values for $|E(\mu_j|Y)|$ less than about 10 are plotted to help zoom in plot. The figure contains the shrinkage plot of the posterior absolute means and posterior variances for μ_j from the Gibbs output
20 using the weighted regression approach. Once again a variance of one is represented by a thin dashed horizontal line, while the thick dashed horizontal line represents the asymptotic value $n_j/n_{j,1} = 55/21$.

The two vertical lines in the figure represent the $100 \times (1 - \alpha/2)$ percentile from the appropriate normal distributions for $\alpha = 0.0005$. The thin
25 dashed line denotes the uncorrected $N(0, 1)$ distribution, while the thick dashed line is from the adjusted $N(0, n_j/n_{j,1})$ distribution. The value for $\alpha = 0.0005$, which we used for the cutoff value in the analysis, was chosen

here by eye-balling Figure 9 and choosing the vertical line that makes almost all observations to its right have posterior variance of roughly one. This removes intermediate values that could inflate the FDR; the same technique was used to select the value of $\alpha = 0.0005$ in Figure 8(c).

Table 4: Number of colon cancer genes detected using weighted regression approach.

Zcut	FDRmix	Bayes Exch	Z-test	Bonf	BH
3743	3334	3433	4576	1369	2960

- 5 Figure 10 shows genes found significant using Zcut with $\alpha = 0.0005$. Crosses denote genes detected from original data, while circles denote using a weighted regression approach. Group mean differences and standard errors plotted are computed from the original data, similar to Figure 8(c). Mean differences less than about 5 in absolute value are plotted
- 10 to help zoom in plot. The figure shows the genes identified by Zcut from the original data and for the new data formed by the weighted regression approach. Observe how the weighting allows Zcut to pick up moderate to high expression genes as desired, while allowing genes with small expressions but high standard errors to be dropped off. Zcut is also able to pick
- 15 up genes with smaller expressions which have small standard errors, but it does not pick up very small mean expression values, as observed with the log-transform. Thus, the new Zcut is more flexible, but there is still a large overlap in detected genes as desired. Of the 3743 genes identified by Zcut over the weighted data, 2864 (76.5%) were also identified over the
- 20 original data. Table 4 records the number of genes detected by each of the methods for the weighted data (α value of 0.0005).

Interestingly, substantially all of the genes picked by BH are contained

Table 5: Grouping of genes found by Zcut and not by BH by biological function

Functional Group	Number of genes
Transcription factor	3
DNA binding	1
Cell adhesion	3
G-protein signaling	2
STAT cascade	3
TGF receptor ligand	2
Growth factor receptor ligand	1
Oncogene	2
Intracellular signaling	4
Signal transduction	6
Cell communication	15

in the list of genes picked by Zcut. These are clearly the genes showing large differential expression between B-survivors and METS. It is informative to look at the list of non-overlapping genes between these two methods. There were 783 genes in this non-overlapping set. This number was further reduced to 779 after removing genes indicating potential liver or spleen contamination. Following this, two quality control analyses were run independently to assess differences in samples (some genes were prepared in San Francisco and others in Cleveland). Genes showing differences in sites were excluded, as were genes showing variability between tumors on the same patient.

In the end, there were 476 genes in this non-overlapping set. Of these, 193 were stripped away from further analysis since they were expressed sequence tags (EST's). This left 283 genes from which EOS probe sets were

mapped to Affymetrix U133A and U133B human gene chips. Once this conversion was done, the information was entered into GeneSpring, a software used to extract functional information about genes printed on Affymetrix chips. The remaining genes were then categorized into different functional groups. The findings were quite interesting. Table 5 provides a breakdown for groupings which represent important functional pathway steps identified in the metastasis process. These include genes which are transcription factors and genes involved in DNA binding, cell adhesion and various signaling, cell communication or cascade pathways. In fact, some of the cascade pathways have been identified as potential sources of possible gene-to-gene interactions due to dimerizations produced during a particular step in a cascade.

8. Discussion. It has been shown how to cast the problem of looking for differentially expressed genes in microarrays as a high dimensional subset selection problem using BAM, a powerful new Bayesian model selection technique. Although the BAM systems and methods were considered here in detail in the context of the two-way ANOVA model, the method permits extensions to more than two groups. Extensions to ANCOVA models are also possible for use in study designs where gene expressions are measured under various experimental conditions.

An important feature of BAM is that it can be used differently depending upon the need of the user. For example, Zcut, a model estimator computed from posterior means, amounts to a weighted combination of ridge regression estimators and strikes a balance between low FDR values and higher power. If the FDR is of greater importance, this disclosure has provided a novel technique for estimating the distribution of gene effects under the

- null hypothesis using a two-component mixture model, leading to what has been termed herein as the FDRmix procedure. The shrinkage effect of BAM provides the power gains seen for the model estimators herein. BAM adaptively shrinks test statistics for intermediate and small sized effects, whereas larger effects yield test statistics similar to Z-test. This is the reason why even though more gene effects are detected as significant using the systems and methods described herein than, say, BH, a source of conservatism is built-in minimizing the potential of too many falsely rejected null hypotheses.
- 10 The significance of the increased power was evident attempting to determine a metastatic genetic signature of colon cancer. Known to be a very complex disease process, Zcut detected over 700 more significant genes than BH did. Many of these genes in fact turned out to be potentially implicated in the metastasis of B-Survivor solid tumors from the colon to the liver. It was interesting that no currently known genes with obvious involvement in colon cancer metastasis were part of the non-overlapping list. The implications of this work becomes clearer by noting that omitting potentially important genes at the level of initial filtering may lead to further derived discriminant rules based on these filtered subsets of genes that may end up leaving out valuable information.

The colon cancer example also illustrated the difficulties in finding global variance stabilizing transformations for the data. Complex non-linear relationships between the gene effect variances and means can result in an adverse affect on the mean when applying such transformations. This was further illustrated via Poisson simulations. As an alternative, a weighted regression approach was presented. This approach is not limited to the BAM technique and can be used with standard methods as well.

8.1. Extensions to More than 2 Groups.. Model (4) can be easily extended to the case involving more than two groups. By way of illustration, an outline of the case involving three groups is now presented. As before, assume that group $l = 1$ is the control group. Groups $l = 2$ and $l = 3$ represent two distinct treatments. If $Y_{j,k,l}$ are the expressions, then testing for treatment effect can be formulated using the ANOVA model,

$$Y_{j,k,l} = \theta_j + \mu_{j,2}I\{l = 2\} + \mu_{j,3}I\{l = 3\} + \epsilon_{j,k,l},$$

where $j = 1, \dots, p$, $k = 1, \dots, n_{j,l}$, $l = 1, 2, 3$, and $\epsilon_{j,k,l}$ are iid $N(0, \sigma^2)$. Testing whether genes differ in groups 2 and 3 from the control corresponds to testing whether $\mu_{j,2}$ and $\mu_{j,3}$ differ from zero. As before many of the θ_j parameters will be non-zero; therefore, the dimension of the parameter space is reduced from $3p$ to $2p$ by centering the responses. Applying the appropriate rescaling as part of the pre-processing of the data, $Y_{j,k,l}$ is replaced with

$$\tilde{Y}_{j,k,l} = (Y_{j,k,l} - \bar{Y}_{j,1}) \times \sqrt{n/\hat{\sigma}_n^2},$$

where

$$\hat{\sigma}_n^2 = \frac{1}{n - 3p} \sum_{j,k,l} \left(Y_{j,k,l} - \bar{Y}_{j,1}I\{l = 1\} - \bar{Y}_{j,2}I\{l = 2\} - \bar{Y}_{j,3}I\{l = 3\} \right)^2,$$

and $\bar{Y}_{j,l}$ is the mean for gene j over group l .

8.2. Differing measurement errors: heteroscedasticity.. One can also extend the ANOVA model (4) to handle differing measurement error variances σ_j^2 for $j = 1, \dots, p$. Suppose that genes can be bunched into C clusters based on their variances. Let $i_j \in \{1, \dots, C\}$ indicate the variance cluster gene j belongs to. Now modify (4) by replacing $\epsilon_{j,k,l}$ with variables, say

$\tilde{\epsilon}_{j,k,l}$, where $\tilde{\epsilon}_{j,k,l}$ are independent with a $N(0, \sigma_{i_j}^2)$ distribution. To handle heteroscedastic variances a weighted regression is applied by reweighting observations by the inverse of their standard deviation. Thus, (4) can now be written as

$$\sigma_{i_j}^{-1/2} Y_{j,k,l} = \theta_j^* + \mu_j^* I\{l=2\} + \epsilon_{j,k,l}, \quad j = 1, \dots, p, \quad k = 1, \dots, n_{j,l}, \quad l = 1, 2,$$

where $\theta_j^* = \sigma_{i_j}^{-1/2} \theta_j$ and $\mu_j^* = \sigma_{i_j}^{-1/2} \mu_j$ denote the new parameters. The C distinct variances $\sigma_1^2, \dots, \sigma_C^2$ are unknown, but they can be estimated accurately when p is large, and thus we can accurately approximate $\sigma_{i_j}^{-1/2} Y_{j,k,l}$. Hence, if $\hat{\sigma}_c^2$ is our estimator for σ_c^2 (we use the usual unbiased estimator)

- 5 simply replace the data $Y_{j,k,l}$ with rescaled data $\hat{\sigma}_{i_j}^{-1/2} Y_{j,k,l}$ and apply the various methods as before.

- The systems and methods described herein may be implemented on a computer system configured for analyzing statistical data, wherein the computer system executes software or firmware instructions to perform
- 10 steps characteristic of the BAM systems and methods disclosed herein. In one embodiment, the computer system includes a database on a computer-readable storage medium comprising the data. Additionally, or alternatively, the computer system may include a computer-readable storage medium comprising instructions for causing a computer apparatus to execute the
- 15 steps of the BAM statistical analysis methods described herein. In one embodiment, a network server provides access to instructions for causing a computer system to execute the steps associated with the BAM methods described herein. The server may distribute data fed into, or processed out of the BAM techniques, across a data network.

- 20 The contents of all references, patents and published patent applications

cited throughout this Application, as well as their associated figures are hereby incorporated by reference in entirety.

Many equivalents to the specific embodiments and the aspects and practices associated with the systems and methods described herein exist. Accordingly, it will be understood that the invention is not to be limited to the embodiments, methods, and practices disclosed herein, but is to be understood from the following claims, which are to be interpreted as broadly as allowed under the law.

10

What is claimed is:

CLAIMS:

1. A method of identifying differentially-expressed genes, comprising:
 - (a) deriving an analysis of variance (ANOVA) or analysis of covariance (ANCOVA) model for expression data associated with a plurality of genes;
 - (b) from the ANOVA or ANCOVA model, deriving a linear regression model defined at least in part by an observation vector representative of an observed subset of the gene-expression data, a design matrix of regressor variables, a vector of regression coefficients representing gene contribution to the observation vector, and a measurement error vector; and
 - (c) to the linear regression model, applying a hierarchical selection algorithm to designate a subset of the regression coefficients as significant regression coefficients, the selection algorithm representing at least one of the observation vector, the design matrix, and the measurement error vector as being hierarchically dependent on parameters having predetermined probabilistic properties, wherein the designated subset corresponds to a respective subset of the genes identified as differentially expressed.
2. The method of claim 1, wherein the parameters are chosen to distinguish statistical properties of the significant regression coefficients.
3. The method of claim 1, wherein the observation vector is at least in part expressed as a sum of:

- (a) a vector obtained by multiplicatively applying the design matrix to the regression coefficient vector; and
- (b) the measurement error vector.

- 5 4. The method of claim 1, wherein the selection algorithm includes a Bayesian selection algorithm.
- 10 5. The method of claim 4, wherein the Bayesian selection algorithm uses a posterior distribution of the regression coefficient vector, conditioned on the observation vector, to identify the significant subset of the regression coefficients.
- 15 6. The method of claim 5, wherein the Bayesian selection algorithm uses a posterior median of the regression coefficient vector, conditioned on the observation vector, to identify the significant subset of the regression coefficients.
- 20 7. The method of claim 5, wherein a posterior variance of the regression coefficient vector, conditioned on the observation vector, is plotted against a posterior mean of the regression coefficient vector, conditioned on the observation vector, to graphically estimate a threshold of comparison parameter used by the selection algorithm.
- 25 8. The method of claim 5, wherein the Bayesian selection algorithm uses a posterior mean of the regression coefficient vector, conditioned on the observation vector, to identify the subset of the regression co-

efficients that are significant.

- 5
9. The method of claim 8, wherein the posterior mean is estimated at least in part by using Gibbs sampling.
10. The method of claims 8, wherein the posterior mean of the regression coefficient vector includes a ridge estimate from a generalized ridge regression of the observation vector on the design matrix.
- 10 11. The method of claim 10, wherein the posterior mean of the regression coefficient vector includes a weighted average of shrunk posterior estimates of the regression coefficient vector.
- 15 12. The method of claims 8, wherein an absolute value of the posterior mean of at least a portion of the regression coefficient vector is entry-wise compared to an entry-wise predetermined threshold value to designate the significant regression coefficients, a regression coefficient being significant if the absolute value of its posterior mean is equal to or exceeds the respective predetermined coefficient.
- 20
13. The method of claim 12, wherein the predetermined threshold value is associated with a theoretical limiting variance vector corresponding to entry-wise variances of the conditional mean of the regression coefficient vector.
- 25

14. The method of claim 8, wherein, under a null hypothesis indicating substantially no differential gene effect, the distribution of the posterior mean vector is approximated by a two-point mixture probability density.
- 5
15. The method of claim 14, wherein the two-point probability density includes a sum of:
- (a) a first normal probability density of zero mean and a first variance, the first normal probability density being scaled by a first probability value; and
- 10
- (b) a second normal probability density of zero mean and a second variance, the second normal probability density being scaled by a second probability value, the second probability value being equal to 1 minus the first probability value.
- 15
16. The method of claim 15, wherein the first variance and the second variance are distinct, and one of the first variance and the second variance corresponds to the significant regression coefficients.
- 20
17. The method of claim 1, wherein the selection algorithm includes a parametric stochastic variable selection (P-SVS) algorithm.
- 25
18. The method of claim 17, wherein a mean of the observation vector is subtracted from the observation vector to create a substantially zero-mean observation vector.

19. The method of claim 18, including, prior to applying the selection algorithm, causing the posterior variance of the regression coefficient vector to be one of a pre-determined set of values by:
- (a) an appropriate scaling of the observation vector;
 - 5 (b) transforming the design matrix to cause each column of the design matrix to have a respective appropriate squared sum value; and
 - (c) rescaling at least one column of the design matrix by an appropriate design scaling value.
- 10
20. The method of claim 19, including a weighted regression step wherein a subset of the observation vector is re-weighted by a standard deviation inverse associated with the subset of the observation vector.
- 15 21. The method of claim 19, wherein the pre-determined set is associated with distinct values, each of the distinct values corresponding to a respective hypothesis about the data.
- 20 22. The method of claim 19, wherein the value of the appropriate scaling includes a square root of the ratio: size of the observation vector divided by a measurement error variance.
- 25 23. The method of claim 19, wherein the appropriate squared sum value includes the observation vector size.

24. The method of claim 19, wherein the appropriate design scaling value causes the columns of the design matrix to have a predetermined second moment.
- 5 25. The method of claim 24, wherein the predetermined second moment is about one.
26. A computer apparatus for analyzing statistical data, comprising a set of instructions for causing the apparatus to execute the method
10 of claim 1.
27. The apparatus of claim 26, further comprising a database on a computer-readable storage medium comprising the data.
- 15 28. A computer-readable storage medium comprising instructions for causing a computer apparatus to execute the method of claim 1.
29. A network server providing access to instructions for causing a computer apparatus to execute the method of claim 1.
20
30. The server of claim 29, wherein the server distributes the instructions across a data network.
31. A method of analyzing statistical data, comprising:
25 (a) deriving an analysis of variance (ANOVA) or analysis of covariance (ANCOVA) model for the data;

- (b) from the ANOVA or ANCOVA model, deriving a linear regression model defined at least in part by an observation vector representing the data, a design matrix of regressor variables, a regression coefficient vector, and a measurement error vector; and
- 5 (c) to the linear regression model, applying a hierarchical selection algorithm to designate a subset of the regression coefficients as significant regression coefficients, the selection algorithm representing at least one of the observation vector, the design matrix, and the measurement error vector as being hierarchically dependent on a plurality of parameters that have predetermined probabilistic properties.
- 10
32. The method of claim 31, wherein the statistical data comprises biomolecular data.
- 15
33. The method of claim 32, wherein the biomolecular data comprises biomolecular array data.
34. The method of claim 33, wherein the biomolecular array data is selected from the group consisting of: gene expression data, protein expression data, metabolite production data, genotype data, tissue data, viral data, and a combination thereof.
- 20
35. A method of identifying differentially-produced biomolecules, comprising:
- 25

- 5 (a) deriving an analysis of variance (ANOVA) or analysis of covariance (ANCOVA) model for biomolecular production data associated with a plurality of biomolecules;
- (b) from the ANOVA or ANCOVA model, deriving a linear regression model defined at least in part by an observation vector representing the data, a design matrix of regressor variables, a regression coefficient vector, and a measurement error vector; and
- 10 (c) to the linear regression model, applying a hierarchical selection algorithm to designate a subset of the regression coefficients as significant regression coefficients, the selection algorithm representing at least one of the observation vector, the design matrix, and the measurement error vector as being hierarchically dependent on a plurality of parameters that have predetermined probabilistic properties.

15

36. A method of mass-spectroscopic data, comprising:

- (a) deriving an analysis of variance (ANOVA) or analysis of covariance (ANCOVA) model for mass-spectroscopic data;
- 20 (b) from the ANOVA or ANCOVA model, deriving a linear regression model defined at least in part by an observation vector representing the data, a design matrix of regressor variables, a regression coefficient vector, and a measurement error vector; and
- 25 (c) to the linear regression model, applying a hierarchical selection algorithm to designate a subset of the regression coefficients as significant regression coefficients, the selection algorithm representing at least one of the observation vector, the design matrix, and the measurement error vector as being hierarchically dependent on a plurality of parameters that have predetermined

probabilistic properties.

37. A method of analyzing statistical data, comprising:

- 5 (a) deriving an M -way analysis of variance (ANOVA) or analysis of covariance (ANCOVA) model for the data, M being at least one;
- 10 (b) from the ANOVA or ANCOVA model, deriving a linear regression model defined by an observation vector representing the data, a design matrix of regressor variables, a regression coefficient vector, and a measurement error vector; and
- 15 (c) to the linear regression model, applying a hierarchical selection algorithm to classify each of the regression coefficients into a number of groups, the selection algorithm representing at least one of the observation vector, the design matrix, and the measurement error vector as being hierarchically dependent on a plurality of parameters that have predetermined probabilistic properties.

38. The method of claim 37, wherein M is at least 3.

20

39. The method of claim 38, wherein the selection algorithm includes a Bayesian selection algorithm.

25 40. The method of claim 39, wherein the Bayesian selection algorithm uses a posterior distribution of the regression coefficient vector, conditioned on the observation vector, to identify the significant subset

of the regression coefficients.

41. The method of claim 40, including:

- 5 (a) designating a first group, from among the M groups, as a baseline reference group; and
- (b) plotting Bayes test statistics associated with a second group, distinct from the first group, versus Bayes test statistics associated with the first group.

- 10 42. The method of claim 41b, including plotting Bayes test statistics associated with a third group, distinct from each of the first and second groups, versus Bayes test statistics associated with the first group.

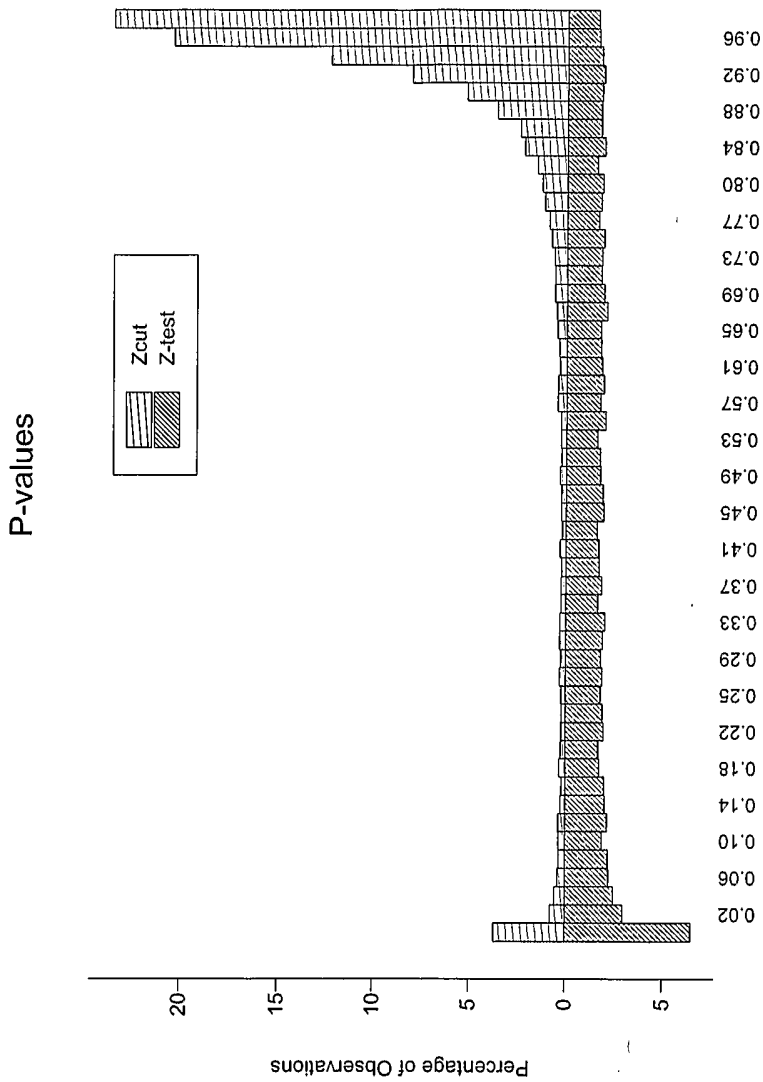


Figure 1

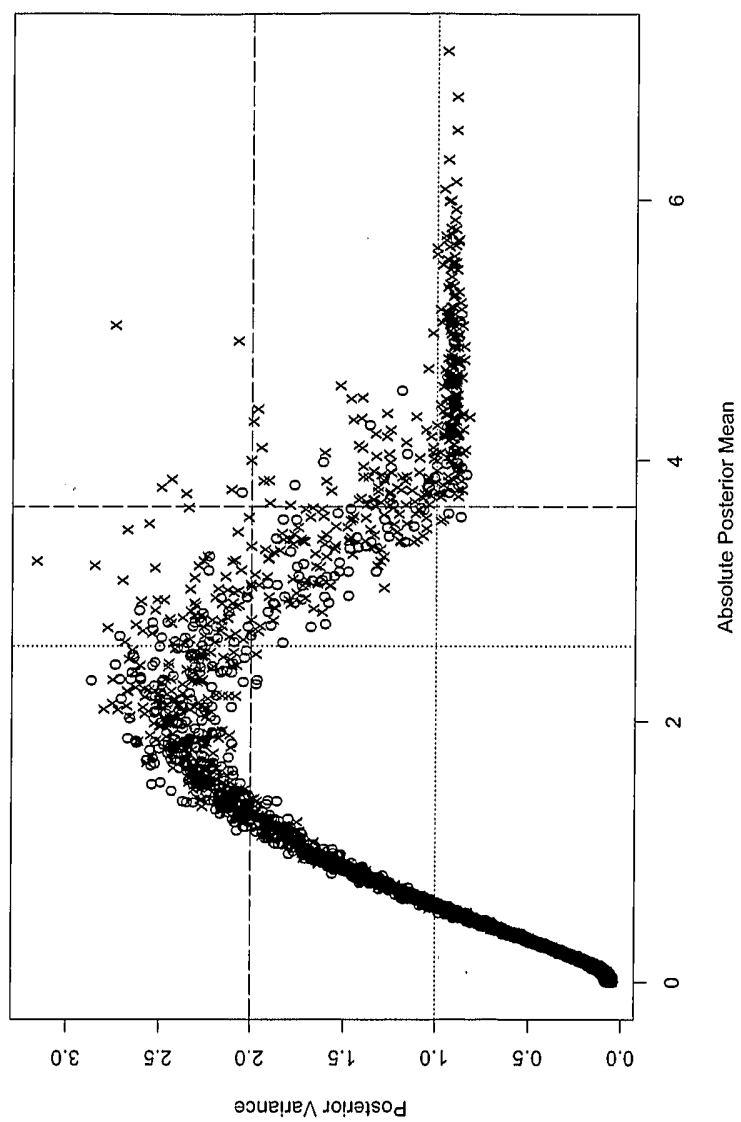


Figure 2a

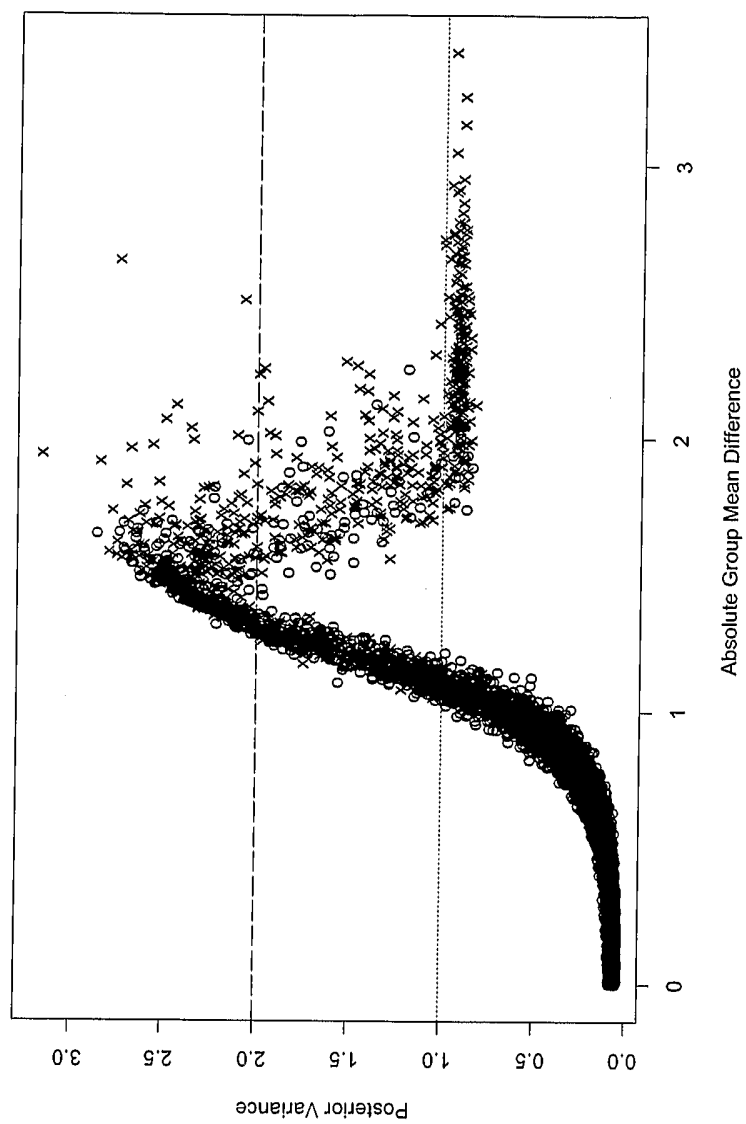


Figure 2b

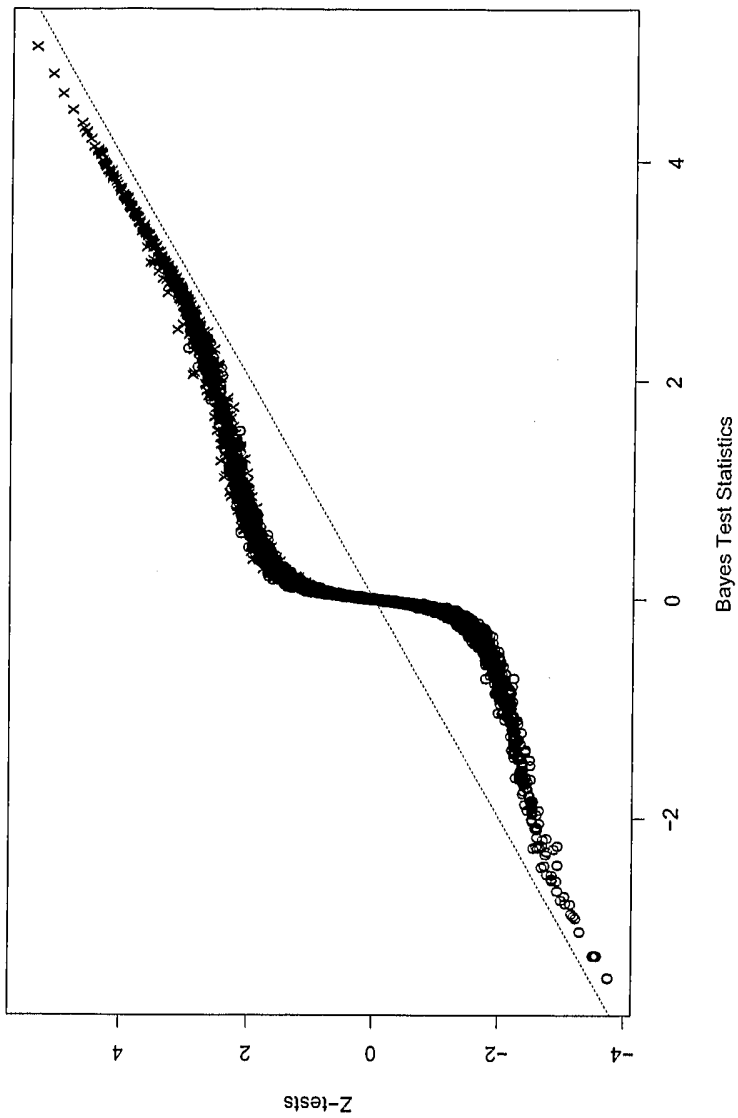


Figure 3a

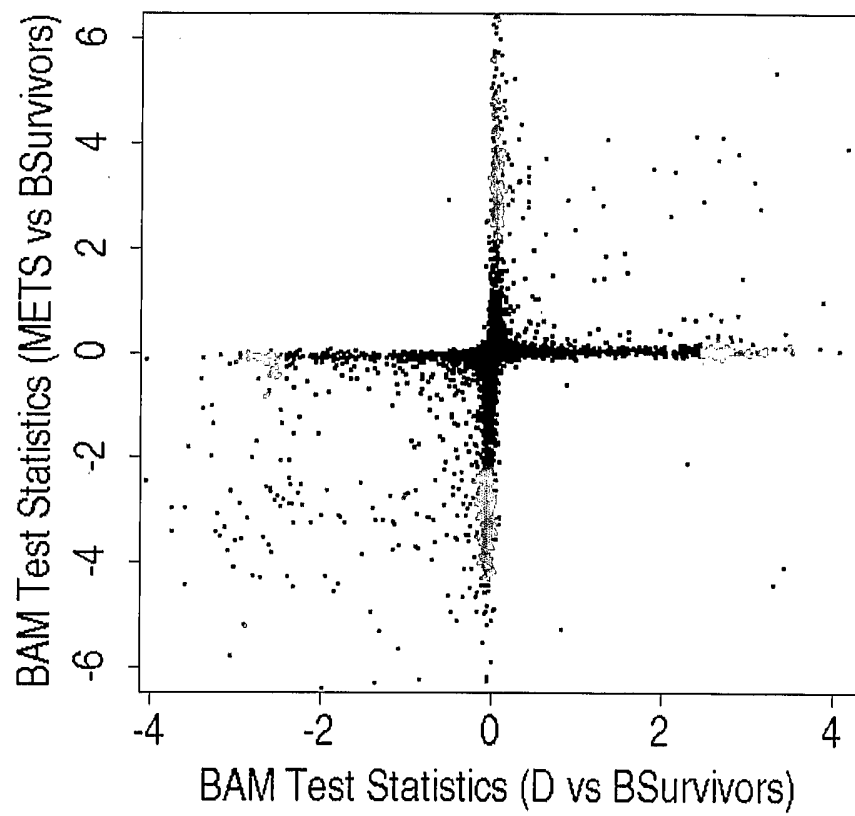


Figure 3b

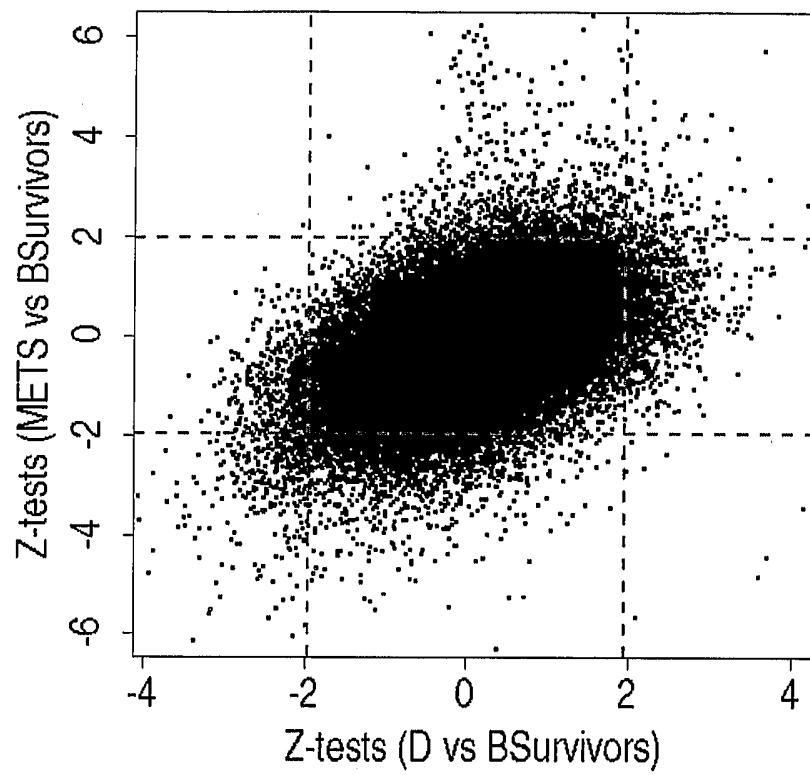
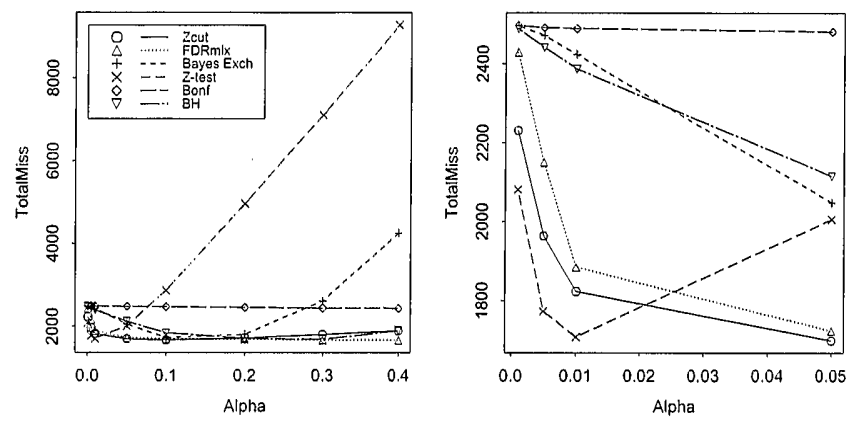


Figure 3c



Figures 4a, 4b

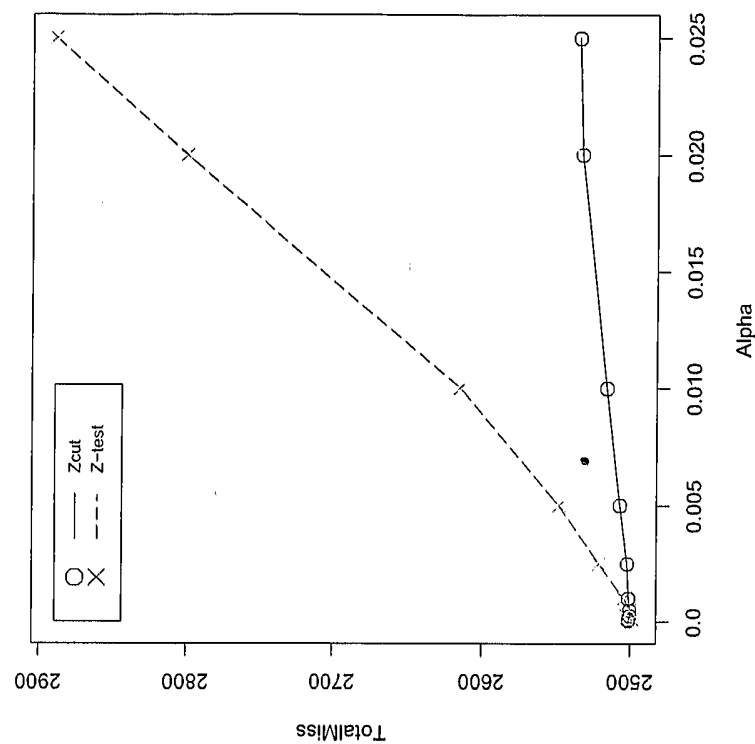


Figure 5a

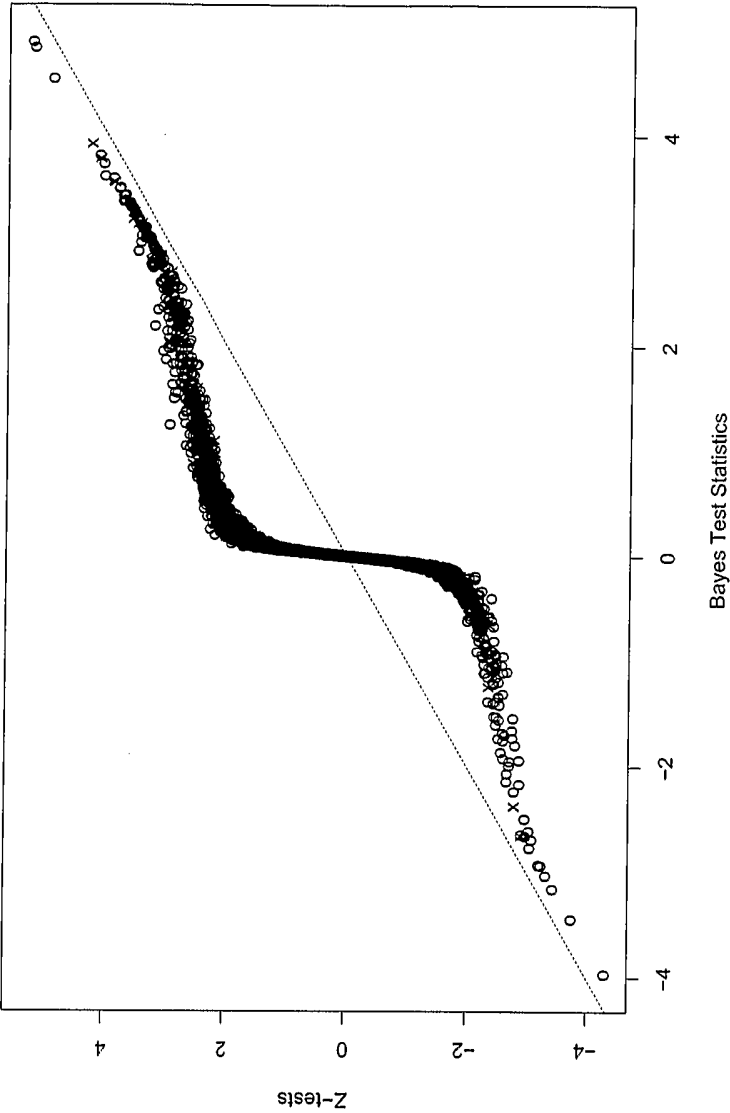


Figure 5b

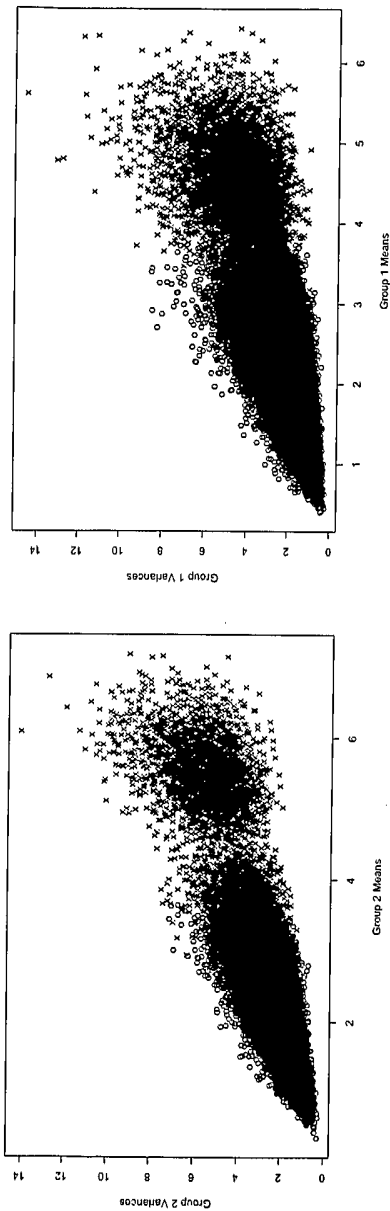


Figure 6a

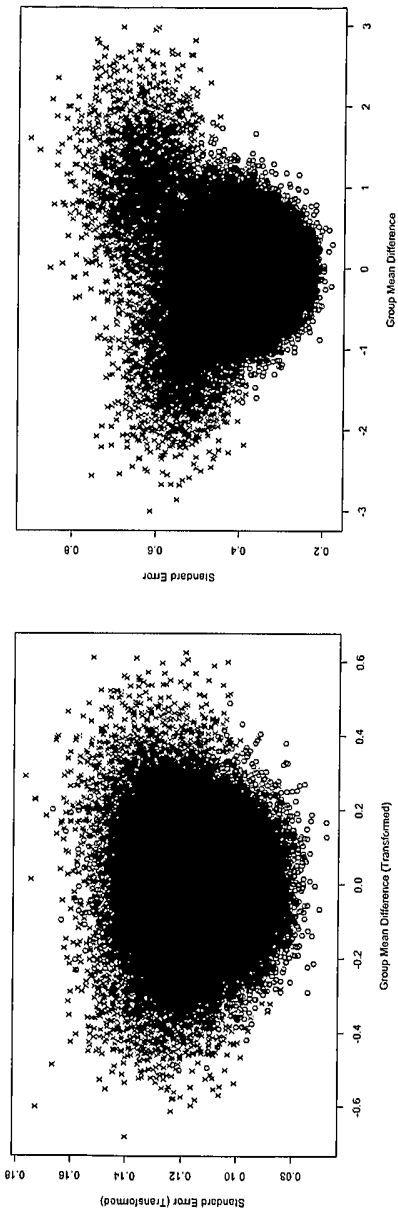


Figure 6b

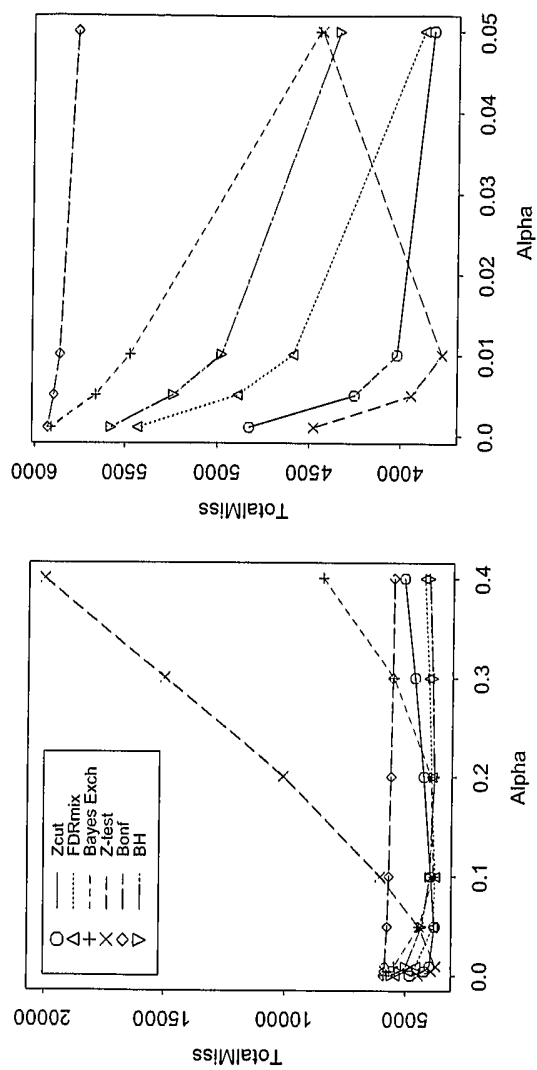


Figure 7a

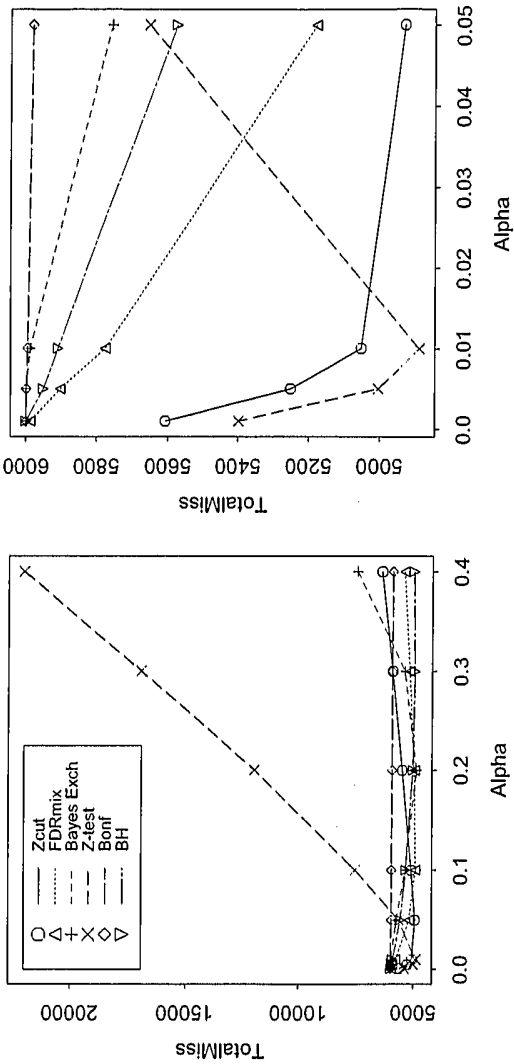


Figure 7b

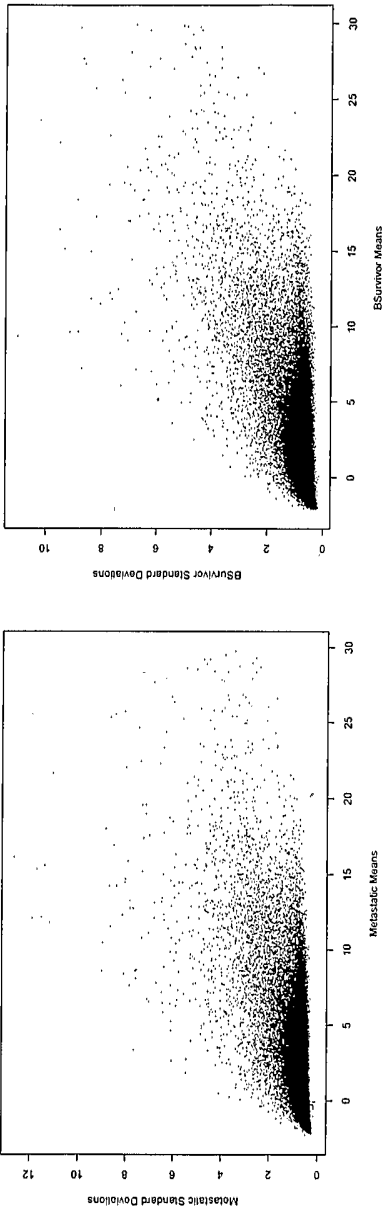


Figure 8a

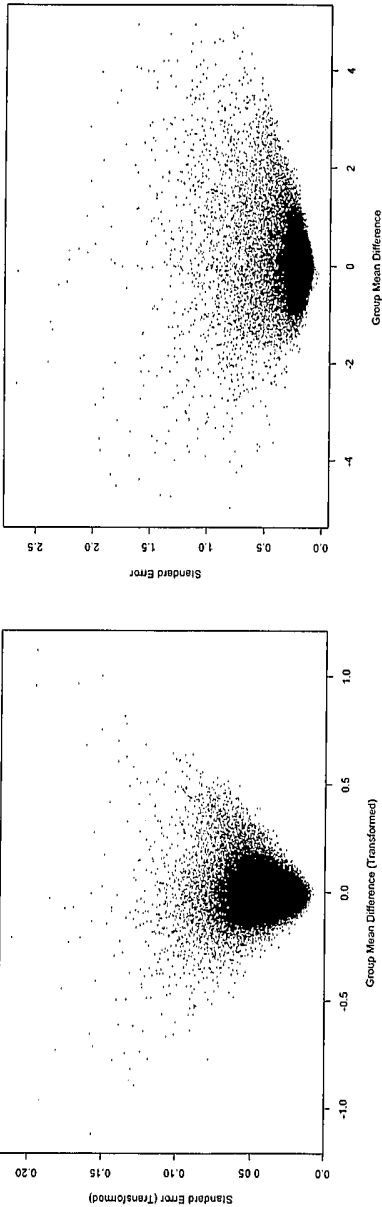


Figure 8b

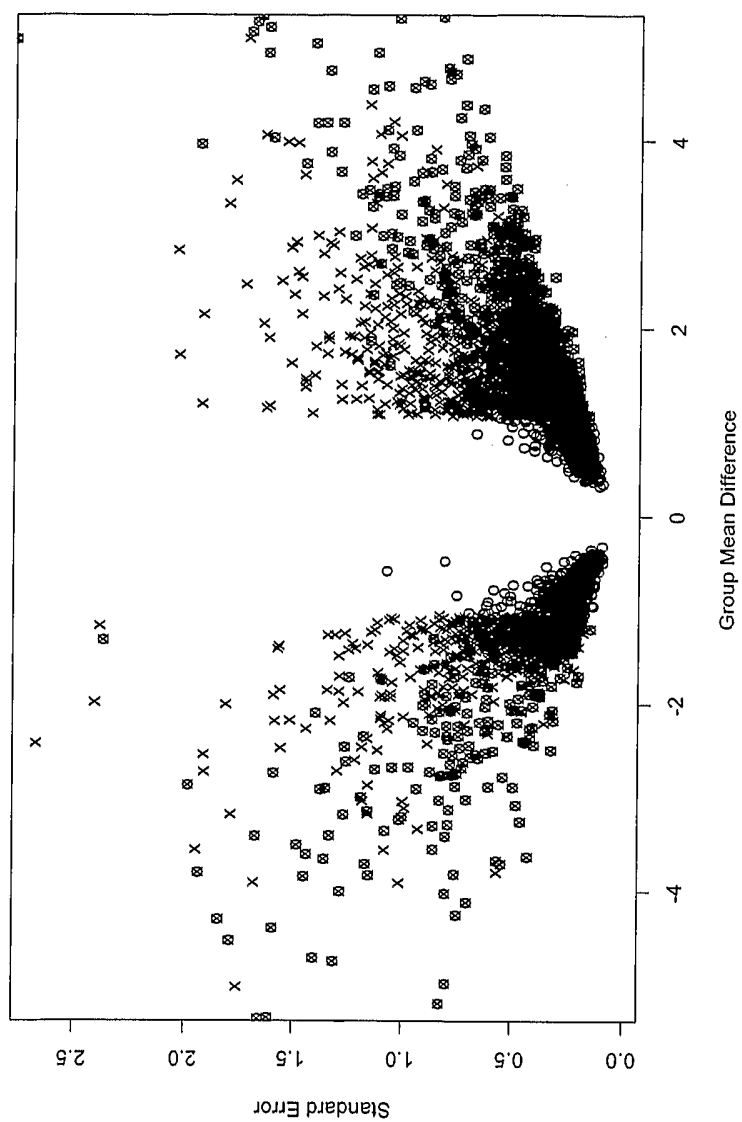


Figure 8c

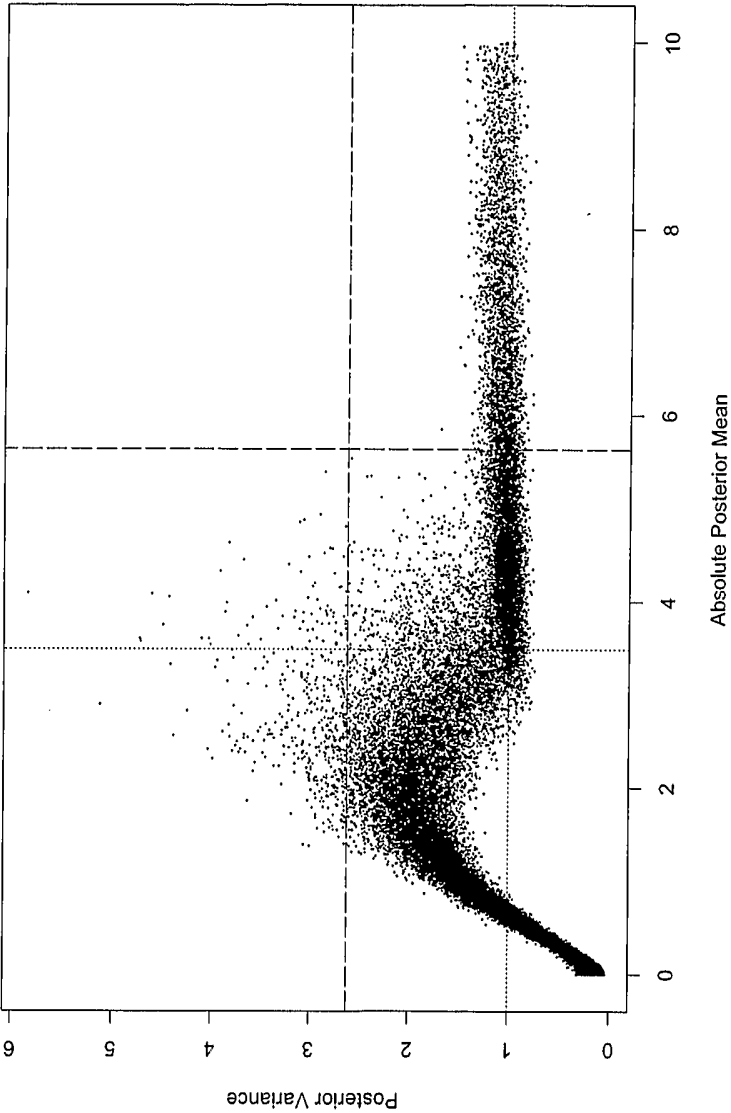
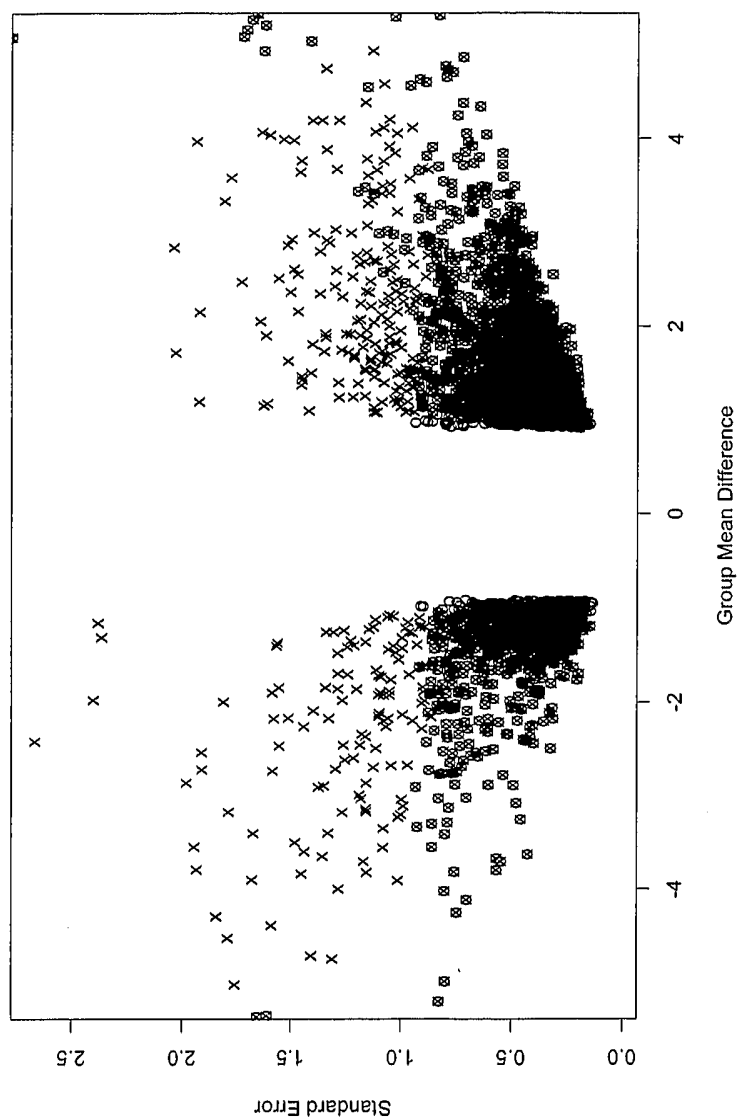


Figure 9

**Figure 10**