Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 462549, 13 pages
http://dx.doi.org/10.1155/2015/462549

*Research Article*

# Identifying and Assessing Interesting Subgroups in a Heterogeneous Population

**Woojoo Lee,[1,2] Andrey Alexeyenko,[3] Maria Pernemalm,[4] Justine Guegan,[5] Philippe Dessen,[5] Vladimir Lazar,[5] Janne Lehtiö,[4] and Yudi Pawitan[1]**

[1]*Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden*
[2]*Department of Statistics, Inha University, Incheon 402-751, Republic of Korea*
[3]*Department of Microbiology, Tumour and Cell Biology, Bioinformatics Infrastructure for Life Sciences, Science for Life Laboratory, Karolinska Institutet, 17177 Stockholm, Sweden*
[4]*Department of Oncology and Pathology, Science for Life Laboratory, Karolinska Institutet, 17121 Solna, Sweden*
[5]*Genomics, Institut Gustave Roussy, F-94805 Villejuif, France*

Correspondence should be addressed to Yudi Pawitan; Yudi.Pawitan@ki.se

Received 10 November 2014; Revised 1 March 2015; Accepted 3 March 2015

Academic Editor: Kristel van Steen



Biological heterogeneity is common in many diseases and it is often the reason for therapeutic failures. Thus, there is great interest in classifying a disease into subtypes that have clinical significance in terms of prognosis or therapy response. One of the most popular methods to uncover unrecognized subtypes is cluster analysis. However, classical clustering methods such as $k$-means clustering or hierarchical clustering are not guaranteed to produce clinically interesting subtypes. This could be because the main statistical variability—the basis of cluster generation—is dominated by genes not associated with the clinical phenotype of interest. Furthermore, a strong prognostic factor might be relevant for a certain subgroup but not for the whole population; thus an analysis of the whole sample may not reveal this prognostic factor. To address these problems we investigate methods to identify and assess clinically interesting subgroups in a heterogeneous population. The identification step uses a clustering algorithm and to assess significance we use a false discovery rate- (FDR-) based measure. Under the heterogeneity condition the standard FDR estimate is shown to overestimate the true FDR value, but this is remedied by an improved FDR estimation procedure. As illustrations, two real data examples from gene expression studies of lung cancer are provided.

## 1. Introduction

Biological heterogeneity is common in many diseases; heterogeneity complicates clinical management, as it is often the reason for prognostic and therapeutic failures. Thus, there have been many attempts to classify a disease into subtypes with anticipation that different subgroups are associated with different clinical significance in terms of prognosis or therapy response (e.g., [1, 2]). A significant progress in designing efficient specific treatments can be achieved if novel clinically relevant subtypes are found.

One of the most popular methods for finding unrecognized subtypes is cluster analysis. However, classical clustering methods such as $k$-means clustering or hierarchical clustering are not guaranteed to produce clinically interesting subtypes because the main statistical variability could be dominated by genes not associated with interesting clinical phenotypes. Furthermore, it could be that prognostic factors shared within a subgroup do not have any important role in other subgroups. Thus, the association between prognostic factors and a clinical phenotype is attenuated and not detectable in the whole population. To address these problems, we extend the standard clustering algorithm to find interesting subgroups in the sense that within the subgroup we can find factors (in this paper: genes) strongly associated with the clinical phenotype. This idea can perhaps be illustrated more clearly as follows: suppose that $Y$ is an outcome (e.g., relapse) and $X$ is a randomized treatment;

it is common to search for a subgroup for which the treatment effect is largest. In effect we are searching for factors $Z$ that have significant interactions with $X$, such that a subgroup defined by $Z$ will have a large treatment effect on $Y$. A unique point in our current application is that both $X$ and $Z$ are given by the same set of gene expression data. Also, we allow complex subgroups to be discovered by a clustering method, which makes the process distinct from the standard interaction analysis.

Given a set of gene expression matrix, our goal of cluster analysis is to group patients and genes into subgroups that convey biological or clinical significance. This task can be translated to the biclustering problem. Biclustering methods attempt to simultaneously cluster both patients and genes with the goal of finding subsets of rows and columns in the expression matrix. Cheng and Church [3] firstly introduced biclustering to gene expression analysis. For reviewing the details of biclustering algorithms, see [4]. As Nowak and Tibshirani [5] noticed, however, most of biclustering algorithms tend to be dominated by groups of highly differentially expressed (DE) genes that may not be relevant to the biological process in question. In other words, irrelevant genes with strong signal can mask genes of highest biological relevance. Furthermore, iterative optimization methods adopted in biclustering algorithms depend on initial conditions. To overcome these limitations, we develop an extensive clustering search algorithm to find molecular subtypes (CAMS) based on clustering of patients with partially similar mRNA profile. CAMS is able to uncover the structures arising from relevant genes that may not be highly expressed but moderately expressed within each subtype.

CAMS produces many subtypes. For each subtype, $t$-statistics comparing two distinct phenotype groups (e.g., relapsed/not relapsed) are computed for whole genes and false discovery rate (FDR) estimate is used to correct for multiple comparisons. The number of genes having small FDR estimates (say, less than 0.1) is the basis for assessing the importance of the subtypes. In real data analysis, however, it is a common occurrence in heterogeneous populations that $P$ value distributions of the two-sample $t$-statistics show substantial shortage of small values compared to the uniform distribution [6]. If we ignore this effect we would miss potential discoveries by overestimating FDR. Since subtypes produced by CAMS still can be heterogeneous, it is crucial to study how the molecular heterogeneity of distinct subtypes affects the FDR estimate. In this paper, we introduce unobserved group (or latent group) variables into a simple model for gene expression and see how the heterogeneity induced by the unobserved group leads to the depletion of small $P$ values even when there are many significant signatures. Thus, without considering this underlying heterogeneity, the use of standard FDR estimate might hide promising discoveries. To resolve this problem, we develop an improved FDR estimation procedure to address the heterogeneity in a dataset.

In estimating FDR, the use of correct null density function is critical. Efron [7] considered three issues that substantially affect the null density estimate in computing FDR: (1) a large proportion of genuine but uninterestingly small effects, (2) hidden correlations, and (3) unobserved covariates. Many researchers have studied how they affect the standard FDR estimate [7–9]. In particular, possible connections between unobserved covariates and FDR have been explored in [6, 10]. Leek and Storey [6] showed numerically that the small $P$ values range from being inflated to depleted depending on the configuration of the unobserved covariates. They developed the so-called surrogate variable analysis (SVA) for capturing heterogeneity induced by the unobserved covariates and studied how SVA affects FDR estimate. Stegle et al. [10] considered a Bayesian method to account for hidden confounding variation in expression quantitative trait loci (QTLs) and showed that the method found additional expression QTLs in real datasets. However, their approaches were suggested to study the attenuated relationship by heterogeneity between a measured variable of interest and clinical outcomes, while we focus on finding submerged subtypes by heterogeneity. The novel contributions of this paper are (1) to explain how the heterogeneity induced by unobserved group leads to the depletion of small $P$ values analytically, (2) to analyze the bias of standard FDR estimates under the heterogeneity, and (3) to develop an improved FDR estimation procedure. With these in mind a FDR-based measure is considered to assess findings from a novel clustering procedure. This is illustrated using two datasets on lung cancer patients.

The rest of this paper is organized as follows. In Section 2, we describe the implementation details of CAMS. A brief review of notations and a standard FDR estimation method are given in Section 3, and it is analytically shown that the hidden subgroup in the population can induce a bias of standard FDR estimate in Section 4. We propose a FDR estimation procedure resolving the bias problem and show how to assess clustering results from CAMS with it in Sections 5 and 6. Section 7 includes two real data applications and is followed by concluding remarks.

## 2. Clustering Algorithm for Finding Molecular Subtypes

Consider a set of gene expression profiles from a group of cancer patients. The premise behind CAMS is that the novel molecular information on cancer heterogeneity is hidden in the gene expression profiles. To uncover the heterogeneity, CAMS implements a two-dimensional clustering "patients versus genes" extensively. The full algorithm is given in Algorithm 1.

We first explain the clustering steps of CAMS graphically in Figures 1(a) and 1(b). In the two figures, a set of gene expression profiles as a matrix with rows corresponding to genes and columns corresponding to patients is graphically represented. For illustrative purposes, we designed the following simple model.

(i) It has two observed groups: for example, relapse yes ($RY$) and relapse no ($RN$) groups.

(ii) It has two unobserved groups: the first two columns correspond to molecular subtype 1 (MS1) and the remaining two columns correspond to molecular

```
while r ∈ {1, 2, ..., n_perm} do
  Shuffle the genes
  Partition the genes into S disjoint subsets
  for i = 1 to S do
    Perform hierarchical clustering on the ith subset of the genes
    for the number of clusters (c) ∈ C do
      Cut the dendrogram from the hierarchical clustering to yield c clusters.
      for j = 1 to c do
        Take jth cluster (a subtype identifier)
        Run hierarchical clustering on patients using only the genes
        in jth cluster (Assume that this step yields K clusters).
        for k = 1 to K do
          Perform two-sample t-tests (e.g. relapse yes vs. no)
          with the individuals in kth cluster and the whole genes.
          Fit the null distribution of the two sample t-statistics
          with known functional forms.
          if
        (Normal approximation for the null distribution is acceptable) then
            Compute the FDR estimate based on the normal approximation.
          end if
          if
        (Normal approximation for the null distribution is not acceptable) then
            Compute the FDR estimate based on K permutations.
          end if
          Compute our proposed FDR-based measure (N01)
          to assess the resulting cluster.
          For a given subtype, compute the P-value of N01 by permuting group labels.
        end for
      end for
    end for
  end for
  r ← r + 1
end while
```

Algorithm 1: CAMS.

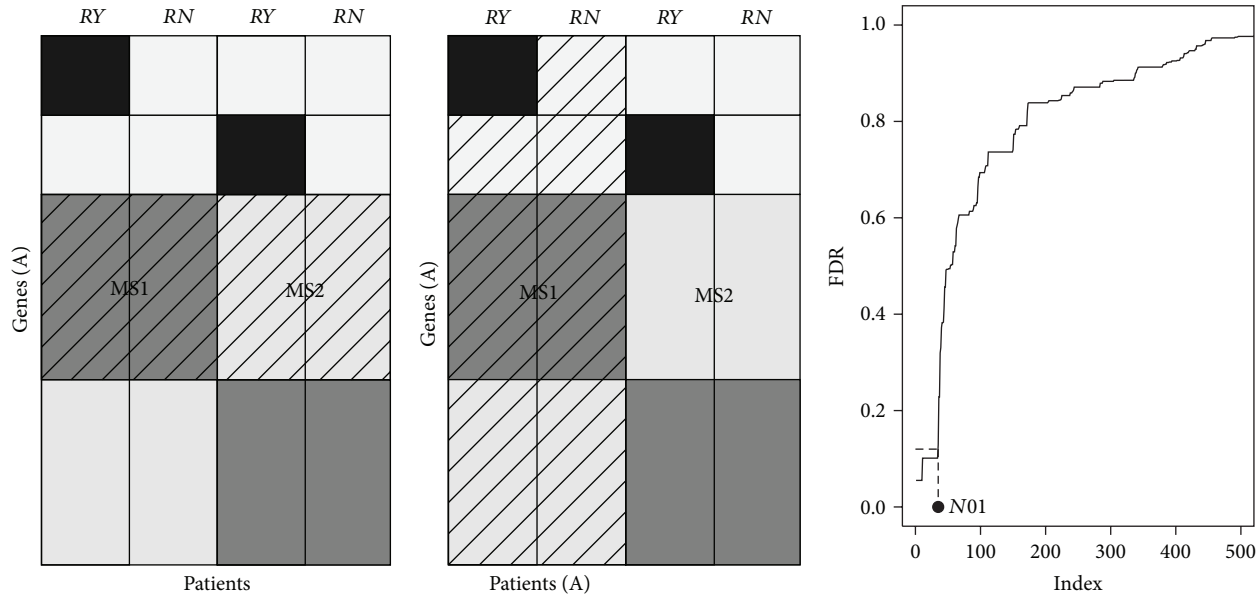
subtype 2 (MS2). This information is unknown to the researchers.

(iii) Some genes (marked in black) affect relapse within a MS.

The two key clustering steps of CAMS are as follows. Step I is clustering of genes. This step identifies several sets of genes having similar profiles across the patients. For example, in Figure 1(a), gene-set A (shaded region) is grouped and this will be used as a subtype identifier in next step. Step II is clustering of patients using gene-set A only. This step produces a subgroup of patients (individuals belonging to the shaded region of Figure 1(b)) with a common expression profile for gene-set A. Note that this subgroup is homogeneous in terms of the identified set of genes from the first step but can show distinct expression profiles between *RY* and *RN* on the other set of genes (e.g., genes marked in black). Thus, we hope that, within the subgroup of patients, good prognostic models can be constructed.

Technical description for CAMS is given as follows. In Step I, the set of gene probes on the microarray chip is grouped via hierarchical clustering. This is implemented using *hclust* in *R*. All the hierarchical clustering in this paper uses complete linkage method and Euclidean metric. This hierarchical clustering procedure is applied to disjoint subsets ($S$) of $m$ all available gene probes due to computational limit (e.g., $m = 41000$ gene probes in the lung cancer dataset). For example, if $S = 10$, our procedure makes 10 disjoint subsets of gene probes and each subset has $m/10$ gene probes sampled from the whole list. Then the whole list is systematically covered by applying the clustering to $S$ subsets sequentially. To allow various groupings of gene probes under different environments, we shuffle the whole list of gene probes several times. The number of clusters ($C$) from each subset $S$ varies on a vector of fixed numbers. For example, if $C = (2, 3, 4, 5, 6, 7, 8, 9, 10)$, then 9 different cluster analysis results are considered in the downstream analysis. Thus each gene probe could participate in different clustering solutions, from very large (>500 probes) to small sets. These clustering results can be used as subtype identifiers in next step.

In Step II, the same hierarchical clustering method is applied to cluster the patients by using each subtype identifier separately. Then the dendrogram is cut at the highest level where the clusters contain more patients than the threshold. Each subset of patients is treated as a candidate subtype.




(a) Step I (b) Step II (c) Step III

FIGURE 1: (a) Step I is clustering of genes. Genes (A) (shaded region) are grouped and will be used as a subtype identifier in the downstream analysis. (b) Step II is clustering of patients using Genes (A) (i.e., a gene-set obtained from Step I). Here, MS1 (a set of patients, shaded regions) is obtained as a subtype. (c) Step III shows how $N01$ is obtained from the FDR curve. We count the number of genes having FDR $< 0.1$. We repeat implementing (a), (b), and (c) across different clustering results extensively. Thus, no shaded columns in (b) will be covered subsequently.

When the clustering steps of CAMS are performed, only some of found cancer subtypes would be true discoveries. To assess whether subtypes are promising or not, $t$-statistics comparing two distinct phenotype groups (e.g., relapsed/not relapsed) within each subtype are computed for whole genes and the number of genes having small FDR estimates (say, less than 0.1) is calculated based on the $P$ values of the $t$-statistics. However, the effect of the molecular heterogeneity on this assessment has not been explored in detail. To deal with this issue, we first review a standard FDR estimation method below.

## 3. Notation and Standard FDR Estimation

In this section some basic notations are introduced to give a formal definition of FDR. For clarity and simplicity, we will limit our discussion to the most common problem of finding differentially expressed (DE) genes between two biological conditions. Let $z$ be a certain statistic to compare the mean log-expression level. The distribution of observed statistics $z$ follows a mixture model

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z), \tag{1}$$

where $\pi_0$ is the proportion of truly nondifferentially expressed (non-DE) genes and $f_0(z)$ and $f_1(z)$ are the density functions of $z$ for non-DE and DE genes, respectively.

Suppose we test $m$ genes with corresponding statistics $z_1, \ldots, z_m$. Let $P_1, \ldots, P_m$ be the ordered $P$ values from $m$ test statistics. For a fixed critical value $c$, we define the number of non-DE genes declared DE and the number of genes declared DE as

$$V(c) = \sum_i I_{(P_i \le c, i \in \text{Null})},$$
$$R(c) = \sum_i I_{(P_i \le c)}, \tag{2}$$

where $I(\cdot)$ is the indicator function. Then, the false discovery proportion (FDP) is defined as

$$\text{FDP}(c) = \frac{V(c)}{R(c)}, \tag{3}$$

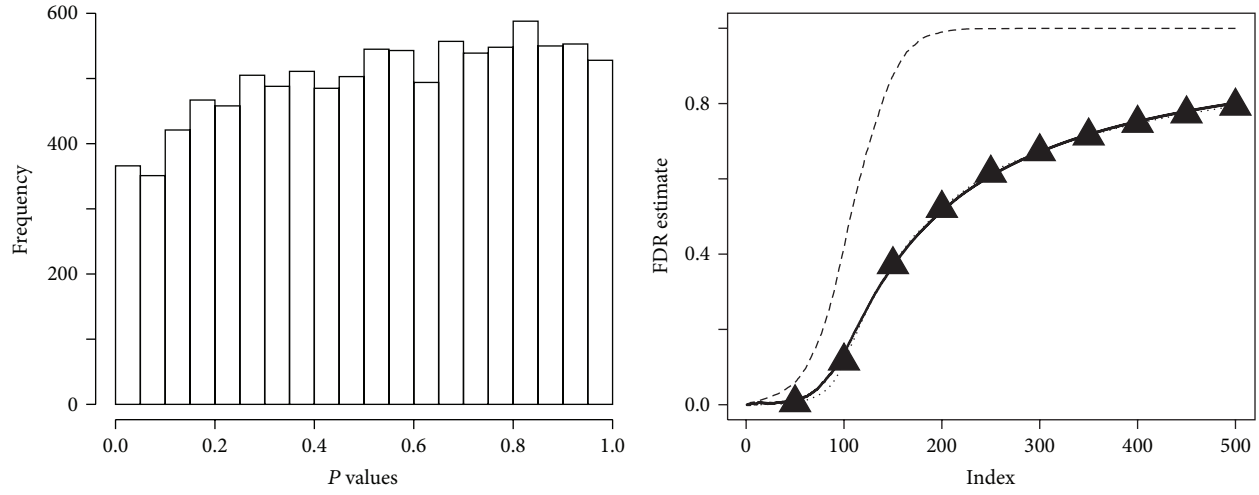
except in the case of $R(c) = 0$, in which case we just set $\text{FDP}(c) = 0$. The FDP is random proportion of false discoveries among the genes declared to be DE. The standard FDR is the marginal average of the FDP; namely, $\text{FDR}(c) = E(\text{FDP}(c))$.

The standard estimate of FDR [8, 11] as a function of the ordered $P$ values is given by

$$\widehat{\text{FDR}}(P_k) = \frac{m\hat{\pi}_0 P_k}{k}. \tag{4}$$

Monotonicity is imposed by taking the cumulative minimum over $\widehat{\text{FDR}}(P_i)$ $(i = k, \ldots, m)$. A common used formula for $\hat{\pi}_0$ is

$$\hat{\pi}_0 = \frac{(\text{Number of } P \text{ values} > \lambda)}{(m(1 - \lambda))} \tag{5}$$




(a) Histogram of $P$ values

(b) Comparison of FDR estimates

FIGURE 2: (a) $P$ value $= P(|T| \geq |t_{\text{obs}}|)$, where $T$ is a generic two-sample $t$-statistic and $t_{\text{obs}}$ is an observed $t$-statistic and (b) average from 50 simulations: true false discovery proportion (FDP) (solid), standard estimate (dashed), and proposed procedure (dotted). The dotted line coincides with the solid, so it is additionally marked with triangles.

for a certain choice of $\lambda$ [11]. Simple choices of $\lambda$ such as 0.5 or 0.75 are often used. Note that this standard estimation procedure does not consider the heterogeneity in population.

## 4. A Bias of the Standard FDR Estimate

Latent variables have been introduced for various purposes in multiple testing framework. Friguet et al. [12] and Leek and Storey [13] considered them as a source of dependence among genes. In this paper we introduce latent variables as a source of heterogeneity and design a latent group model leading to a depleted $P$ value distribution near 0. With practical applications in mind, we will adopt terminologies from two-sample microarray studies for cancer. Our toy model is already graphically represented in Section 2. There are two unobserved groups (molecular subtypes 1 (MS1) and 2 (MS2)) and two observed groups (relapse yes ($RY$) and relapse no ($RN$)). Genes affecting relapse within a MS are marked black and genes identifying MS are marked dark gray. More details to generate Figure 1 are as follows.

(i) For most genes, we choose one MS randomly with probability 0.5 and generate background effects from $N(\mu/2, 1)$. For other MS, we generate background effects from $N(0, 1)$. These genes are used to define specific molecular subtypes (MS1 and MS2) and are undiscriminating for the two observed groups ($RY$ and $RN$).

(ii) Some genes affect relapse within a MS. After choosing one MS with probability 0.5, we generate background effects from $N(\mu/2, 1)$. Then, we add signal effects generated from $N(\mu_0/2, 1)$ for $RY$ and $N(-\mu_0/2, 1)$ for $RN$, respectively. For other MS, we generate background effects from $N(0, 1)$.

Consider genes defining MS and highly expressed in MS1. Then, we have the following ANOVA representation:

$$Y_{ij} = \left(\frac{\mu}{2}\right) I_{(j\in \text{MS1})} + \varepsilon_{ij}, \tag{6}$$

where $i$ is the index for gene, $j$ is the index for patient, and $\varepsilon_{ij} \sim N(0, 1)$. For relapse-related genes within MS1, we have the following ANOVA representation:

$$Y_{ij} = \left(\frac{\mu}{2}\right) I_{(j\in \text{MS1})} + \left(\frac{\mu_0}{2}\right) I_{(j\in RY)} + \left(-\frac{\mu_0}{2}\right) I_{(j\in RN)} + \varepsilon^*_{ij}, \tag{7}$$

where $\varepsilon^*_{ij} \sim N(0, 2)$. In contrast to our model, Efron [7] considered the following model:

$$Y_{ij} = \left(\frac{\mu_{0i}}{2}\right) I_{(j\in RY)} - \left(\frac{\mu_{0i}}{2}\right) I_{(j\in RN)} + \varepsilon_{ij}, \tag{8}$$

where $\mu_{0i} \sim N(0, \sigma^2)$. Note that this model does not consider unobserved group, and as [7] pointed out, this model can lead to only a dilated null distribution that explains inflation of small $P$ values (i.e., false positives). In Figure 2(a), however, our latent group model shows the depletion of small $P$ values. Note that (6) dominates the overall shape of the $P$ value distribution because it has high proportion in the model.

We now see how the unobserved group in the population induces a bias of standard FDR estimate. As a first step, we compute two-sample $t$-statistic to compare $RY$ and $RN$. In $RY$, there are $n_y$ patients, where the half are from MS1 and the other from MS2. In $RN$, there are $n_n$ patients and it has the same structure. Thus, $RY$ and $RN$ groups consist of two normal distributions having different means. Consider the genes following (6). Let $\overline{RY}_i = \sum_j Y_{ij} 1_{(j\in RY)}/n_y$ and

$\overline{RN}_i = \sum_j Y_{ij} 1_{(j\in RN)}/n_n$. The $t$-statistic to test the null hypothesis (non-DE) is

$$z_i = \frac{\left(\overline{RY}_i - \overline{RN}_i\right)}{\left(\widehat{\sigma}_i \sqrt{1/n_y + 1/n_n}\right)}, \tag{9}$$

where

$$\widehat{\sigma}_i = \sqrt{\frac{\left(\sum_j \left(Y_{ij} - \overline{RY}_i\right)^2 1_{(j\in RY)} + \sum_i \left(Y_{ij} - \overline{RN}_i\right)^2 1_{(j\in RN)}\right)}{\left(n_y + n_n - 2\right)}}. \tag{10}$$

Note that, for large $n_y$ and $n_n$, we have

$$\widehat{\sigma}_i \longrightarrow_p \sqrt{1 + \frac{\mu^2}{16}}, \tag{11}$$

because

$$\begin{aligned}
&\frac{\sum_j \left(Y_{ij} - \overline{RY}_i\right)^2 1_{(j\in RY)}}{n_y} \\
&= \frac{\sum_j \left(Y_{ij} - \mu/4\right)^2 1_{(j\in RY)}}{n_y} + o_p(1) \\
&= \frac{\sum_{(j\in RY \bigcap \mathrm{MS1})} \left(Y_{ij} - \mu/4\right)^2}{n_y} \\
&\quad + \frac{\sum_{(j\in RY \bigcap \mathrm{MS2})} \left(Y_{ij} - \mu/4\right)^2}{n_y} + o_p(1) \\
&\longrightarrow_p 0.5\left(1 + \frac{\mu^2}{16}\right) + 0.5\left(1 + \frac{\mu^2}{16}\right) \\
&= \left(1 + \frac{\mu^2}{16}\right), \\
&\frac{\sum_j \left(Y_{ij} - \overline{RN}_i\right)^2 1_{(j\in RN)}}{n_n} \\
&= \frac{\sum_j \left(Y_{ij} - \mu/4\right)^2 1_{(j\in RN)}}{n_n} + o_p(1) \\
&\longrightarrow_p \left(1 + \frac{\mu^2}{16}\right).
\end{aligned} \tag{12}$$

Meanwhile, the numerator in (9) is

$$\begin{aligned}
&\sqrt{n}\left(\overline{RY}_i - \overline{RN}_i\right) \\
&= \sqrt{n}\left(\frac{\sum_j Y_{ij} 1_{(j\in RY)}}{n_y} - \frac{\sum_j Y_{ij} 1_{(j\in RN)}}{n_n}\right) \\
&= \sqrt{n}\left(\frac{\sum_{(j\in RY \bigcap \mathrm{MS1})} \left(Y_{ij} - \mu/2\right)}{n_y} + \frac{\sum_{(j\in RY \bigcap \mathrm{MS2})} Y_{ij}}{n_y}\right. \\
&\qquad \left. - \frac{\sum_{(j\in RN \bigcap \mathrm{MS1})} \left(Y_{ij} - \mu/2\right)}{n_n} - \frac{\sum_{(j\in RN \bigcap \mathrm{MS2})} Y_{ij}}{n_n}\right) \\
&\longrightarrow_d N(0, 4).
\end{aligned} \tag{13}$$

Thus, we have for large $n$

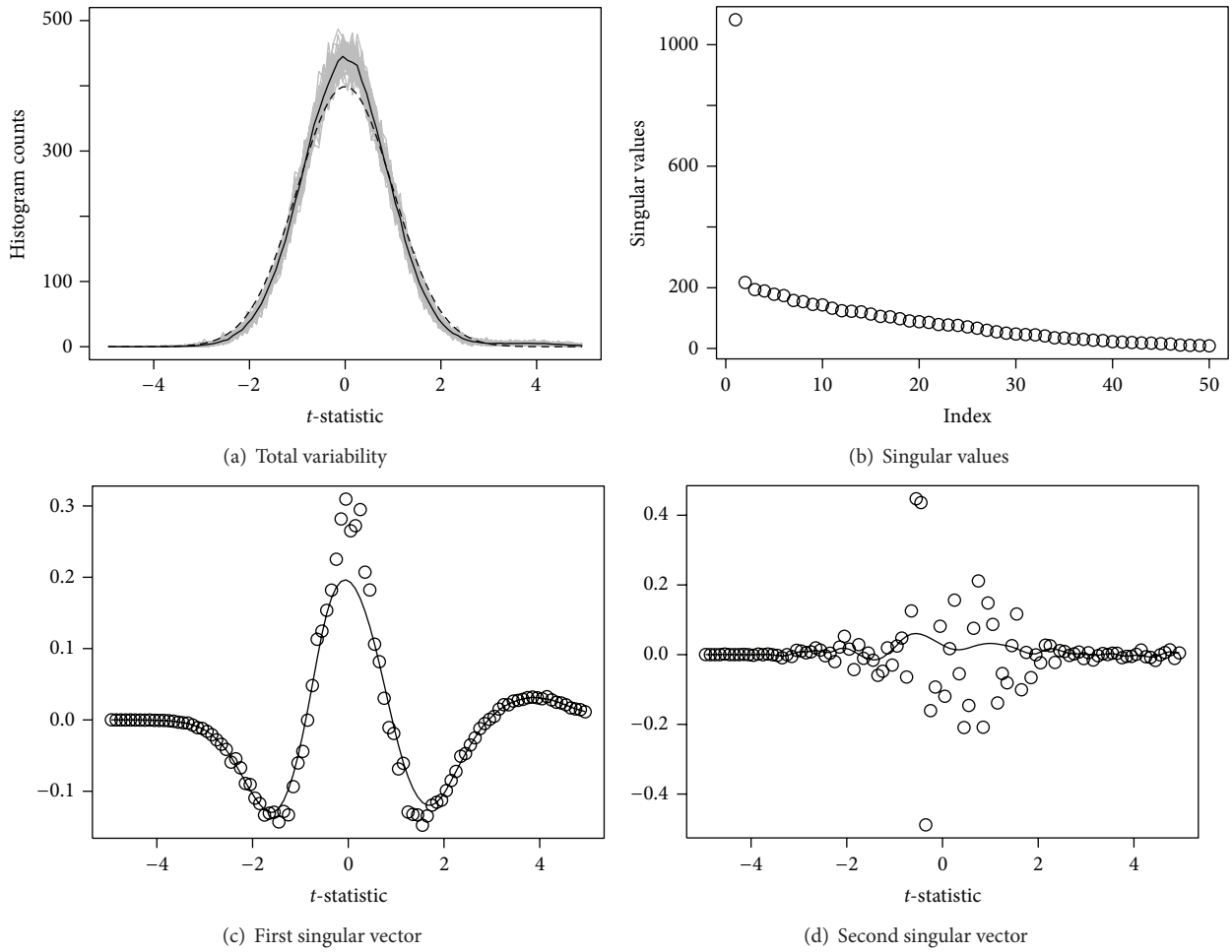
$$z_i \longrightarrow_d N\left(0, \frac{1}{\left(1 + \mu^2/16\right)}\right). \tag{14}$$

Since $1 + \mu^2/16 > 1$ for any $\mu \neq 0$, the use of standard Gaussian distribution for $z_i$ leads to inflated $P$ values. Thus, in (4), $\widehat{\pi}_0$ is overestimated and $R(c)$ is smaller than it should be. Subsequently, (4) overestimates FDR. If the strength of background signal $\mu$ becomes larger, the degree of depletion of small $P$ values becomes more severe because (14) will be more concentrated at 0 as $\mu$ increases. Consequently, the heterogeneity induced by the unobserved group makes the $t$-statistics conservative and leads to upward bias of standard FDR estimate as shown in Figure 2(b). In our simulation, we use 10,000 genes and 60 patients, with 30 belonging to each MS. The proportion of genes defining specific MS is 0.99. Within each MS, the number of $RY$ and $RN$ is assumed to be same for simplicity and we use $\mu = 2$ and $\mu_0 = 3$.

## 5. Proposed FDR Estimation Procedure

While performing CAMS, we want to assess whether clustering results are informative or not with respect to a measure based on FDR. Thus, in computing FDR estimate, the population heterogeneity should be addressed properly. Furthermore, when many datasets are considered simultaneously, it is desirable to have a fast and stable algorithm to compute FDR estimate. Reflecting these aspects, we propose a new FDR estimation procedure.

Our starting point is Pawitan et al.'s FDR estimation procedure [9] because it is computationally flexible to accommodate new changes. A similar permutation-based approach to deal with the dependence in computing FDR estimates was developed by [14]. Pawitan et al. [9] explored the variation pattern of the null distribution of test statistics using the singular value decomposition (SVD) when there are correlations between genes. To check the validity of the SVD analysis in our problem, it is needed to confirm whether the main variation pattern of permutation distribution can represent that of sampling distribution.




(a) Total variability (b) Singular values

(c) First singular vector (d) Second singular vector

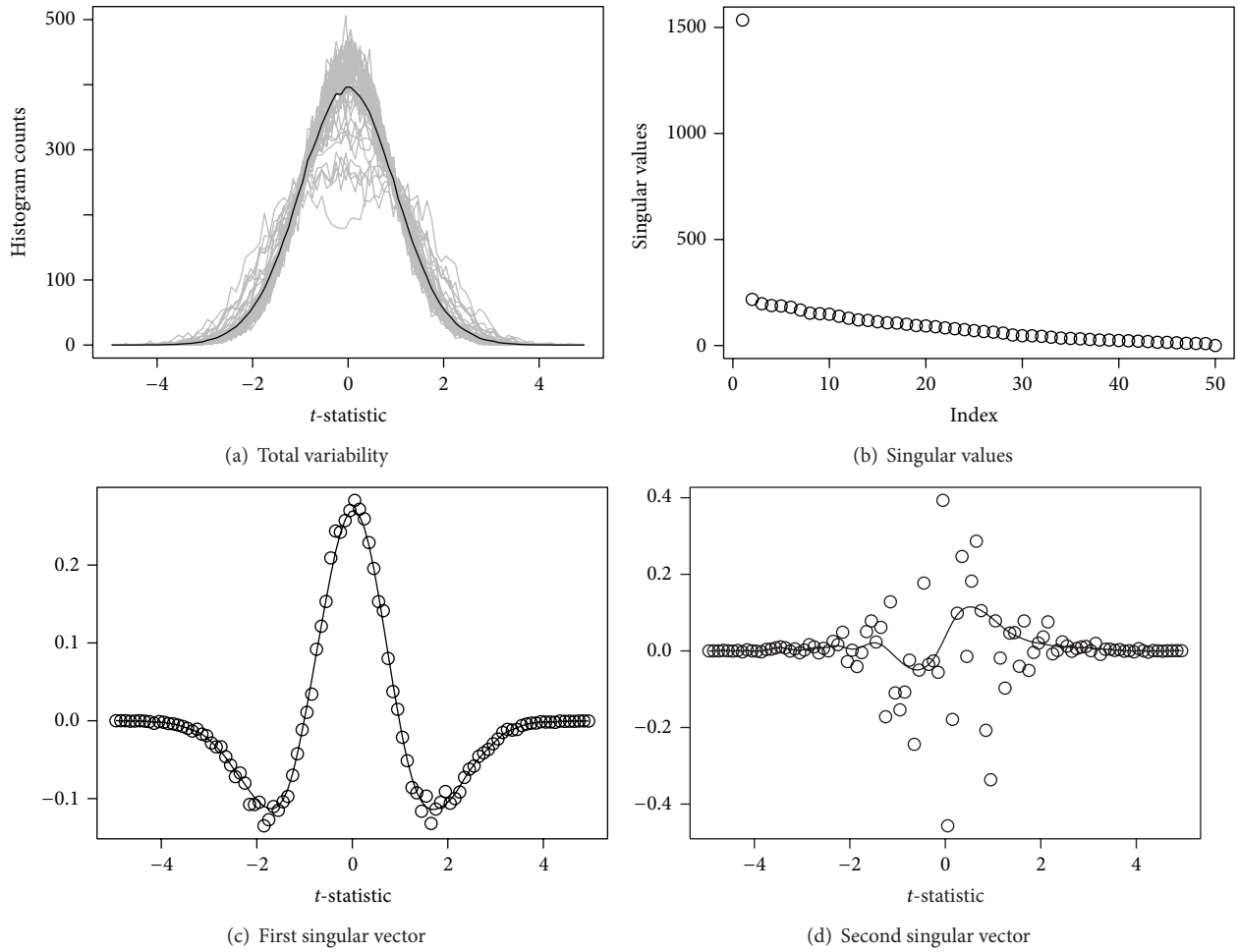
FIGURE 3: (a) Each simulation contributes a single gray line. The solid black line is the average of 50 simulations, and dashed line is the expected histogram-count vector from $N(0, 1)$. (b) shows singular values from the singular value decomposition (SVD) of $Y$, and the dots in (c) and (d) are the components of the singular vectors generated by the SVD, and the solid lines are robust smoothing curves.

*5.1. The Validity of SVD Analysis.* We firstly demonstrate the variation pattern of the sampling distributions from the latent group model through the SVD analysis. We partition the range of the observed statistics into $B$ equispaced bins with width $\Delta$. Let the histogram-count vector $\mathbf{y} = (y_1, \ldots, y_B)$ be the number of statistics that fall into each bin. Each simulation contributes a single count vector $\mathbf{y_i}$. Let $\boldsymbol{\eta}$ be the expected histogram-count vector from standard Gaussian distribution and the $B \times K$ matrix $Y$ the matrix of centered count vectors $\mathbf{y_i} - \boldsymbol{\eta}$. $K$ is the number of simulations and 50 is used in our example.

Figure 3(a) shows the total variability of sampling distributions; the solid line is $\bar{\mathbf{y}}$ and the dashed line is $\boldsymbol{\eta}$. The solid line has higher peak and smaller width than the dashed line, so this is consistent with our analytical findings. To see the variability of $\mathbf{y_i} - \boldsymbol{\eta}$, we perform the SVD of $Y$. The variation is dominated by one large singular value, associated with the pattern seen in the plot of the first singular vector. A consequence of this pattern is that the sampling distribution tends to have a leptokurtic shape compared to the standard Gaussian distribution. Subsequent singular vectors do not have large contributions to the variation.

In practice, we cannot create real data as in simulation. To circumvent this problem, we use permutation to generate the null distribution, but we first check the variability pattern of the distributions from permutation. Let $X$ be a microarray data matrix, let $\mathbf{g} = (g_1, \ldots, g_n)$ be the vector of group labels, and let $g^*$ be a random rearrangement of $g$. With each permuted dataset $(X, g^*)$, we compute test statistics. So each permutation contributes a single count vector $\mathbf{y_i^*}$. Let $\bar{\mathbf{y}}^*$ be the mean vector of $\mathbf{y_i^*}$ over $K$ permutations and the $B \times K$ matrix $Y^*$ the mean-corrected matrix of count vector $\mathbf{y_i^*}$. The SVD results of $Y^*$ are reported in Figure 4.

Figure 4(a) shows the total variability of the distribution over permutations, and the solid line is the average of the permuted null distributions. In Figure 4(b), the first singular value is dominating others and Figure 4(c) shows that the pattern of the first singular vector from $Y^*$ is very close to that from $Y$. This implies that the main variation of permuted distributions explains that of the sampling distributions well,




(a) Total variability

(b) Singular values

(c) First singular vector

(d) Second singular vector

Figure 4: (a) Each permutation contributes a single gray line. The solid black line is the average of 100 permutations. (b) shows singular values from the singular value decomposition (SVD) of $Y^*$, and the dots in (c) and (d) are the components of the singular vectors generated by the SVD, and the solid lines are robust smoothing curves.

so the SVD analysis for permuted data is valid under our latent group model.

Since we have the validity of the SVD analysis, Pawitan et al.'s method [9] can be adopted to correct the overestimation by unobserved group. We assume that the observed statistics $z$ follow a mixture model (1). They suggested to fit

$$\mathbf{y} \sim \text{Poisson}\left(m\Delta f\left(z\right)\right), \tag{15}$$

where

$$f\left(z\right) = \pi_0\left(\phi_0\left(z\right) + b\phi_1\left(z\right)\right) + \left(1-\pi_0\right) f_1\left(z\right), \tag{16}$$

where $f_0(z) = \phi_0(z) + b\phi_1(z)$, $\phi_0(z)$ is the average of null distributions over permutations, and $\phi_1(z)$ is the first singular vector of $Y^*$. In this paper, the parameter $b$ captures the variation of the null distribution due to the heterogeneity by unobserved group. The original computing procedure is given as follows.

(1) Perform $K$ permutations of group labels. Each permuted dataset generates a histogram-count vector $\mathbf{y}^*$.

(2) Compute the predictor $\phi_0$ from the average vector $\overline{\mathbf{y}}^*$ by scaling so that it integrates to 1.

(3) Construct a matrix $Y^*$ from the $\mathbf{y}^*$s. Compute the predictor $\phi_1$ from the smoothed first singular vector $u_1$.

(4) Since $f_1$ is unknown, the regression is performed in two steps. First, fit the reduced model $\mathbf{y} \sim \text{Poisson}(\mu = m\Delta f)$, where

$$f = \beta_0\phi_0 + \beta_1\phi_1, \tag{17}$$

and compute the residual vector $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}}$. Estimate $f_1$ by smoothing the residual vector $\mathbf{r}/m\Delta$ as a function of $z$.

(5) Fit the full model (16)

$$f = \beta_0\phi_0 + \beta_1\phi_1 + \beta_2 f_1, \tag{18}$$

and reestimate the full set of coefficients $(\beta_0, \beta_1, \beta_2)$.

The coefficient of $\phi_0$ becomes the estimate for $\pi_0$. Given estimates of parameters, $P$ values inflated by the heterogeneity are corrected by using the following definition:

$$P \text{ value} = \int_{|z|\geq|z_{\text{obs}}|} \widehat{f_0}(z)\, dz, \tag{19}$$

where $\widehat{f_0}(z)$ is the null density estimate corrected by the first singular vector. For the FDR estimate, we have

$$\widehat{\text{FDR}}(c) = \frac{m\widehat{\pi}_0 \int_{|z|>c} \widehat{f}_0(z)\, dz}{\sum_i I(|z_i| > c)}. \tag{20}$$

(Strictly speaking, this is an FDP estimate rather than an FDR estimate.) Simulation studies show that this estimate has a negligible bias (Figure 2(b)). It may be possible to improve the null density estimate further using the second singular vector in some cases, but we will not attempt this here.

*5.2. Improved Algorithm for Many Datasets.* CAMS generates many subtypes. Since not all the subtypes are meaningful, it is needed to assess each of them quickly. In particular, for the subtypes showing the depletion of small $P$ values, it is desirable to apply our FDR procedure to address such depletions.

One measure to assess subtypes from CAMS is the number of genes having FDR $<$ $c$, where $c$ is a suitably chosen small value; we use $c = 0.1$ in our examples and call this measure $N01$. Figure 1(c) shows that $N01$ is obtained by counting the number of genes with FDR $< 0.1$. To compute $N01$ for many datasets, the previous procedure becomes

(1) computationally intensive: the permutation step takes long time;

(2) unstable: some null density estimates have negative values.

To increase computational speed, we note that $\widehat{f_0}(z)$ from the SVD analysis is empirically well approximated by $N(0, (1 - b/\sqrt{2})^2)$ for $|b| < 0.2$, which can be checked before the permutation step. But this approximation does not seem to be reliable when $|b| > 0.2$. Furthermore, the null density estimates often have negative values when $b$ is large and this leads to a numerical problem in estimating FDR. Thus, we propose a more stable algorithm to find good approximation to the null density. The main idea is to pick up a few vectors $\mathbf{y}_i^*$ that are closest to the histogram counts of the observed test statistics $\mathbf{y}$ with respect to a certain metric. To emphasize goodness of fit at the center of the distribution, we use

$$\text{Dist}(\mathbf{y}, \mathbf{y}_i^*) = (\mathbf{y} - \mathbf{y}_i^*)^T W_{\mathbf{y}} (\mathbf{y} - \mathbf{y}_i^*), \tag{21}$$

where $W_{\mathbf{y}} = \text{Diag}(\mathbf{y})$ as a distance measure. This distance measure gives larger weights to the central part of the histogram. We find top 5 curves that minimize (21) and use their average as $\widehat{f_0}(z)$. For simplicity we estimate $\pi_0$ with (5). The resulting procedure thus becomes as follows.

(i) Before the permutation step, approximate $\widehat{f_0}(z)$ by $N(0, (1 - b/\sqrt{2})^2)$ and obtain $\phi_0$ and $\phi_1$ using known functional forms [15]. After fitting (16) as described in Steps 4 and 5 in the previous section, check whether $|b| < 0.2$ or not. If $|b| < 0.2$, compute FDR estimate.

(ii) In the case of $|b| > 0.2$, perform $K$ permutations and compute (21) for each permuted dataset. Find the top 5 curves that minimize (21) and take the average as the null density estimate. Estimate $\pi_0$ using formula (5).

## 6. Assessing the Clustering Results from CAMS

CAMS can generate a practically unlimited number of candidate subtypes by permuting the gene probes for doing extensive search. If a subtype is depleted in small $P$ values, it is desirable to assess it with $N01$. To see the proportion of subtypes requiring such assessment, we define the ratio of low $P$ value areas as

$$\text{Ratio}(\lambda) = \frac{\sum_i^m 1_{(P_i \leq \lambda)}}{m\lambda}, \tag{22}$$

where $\lambda = 0.2$ is used in practice. The denominator corresponds to the expected number of $P$ values less than $\lambda$ when the null hypothesis holds. We regard Ratio $< 1$ as indicating the targeted situation (the depletion of small $P$ values). When the whole set of patients shows the depletion, we often observe high proportion of potential subtypes with Ratio $< 1$, so it is safe to use $N01$ as a default assessment measure.

We provide an implementation of the proposed method as an *R* package at http://fafner.meb.ki.se/personal/yudpaw/. Two necessary inputs for the implementation are gene expression data matrix and corresponding group vector (a clinical outcome such as disease outcome, e.g., relapse indicator). To enable further analysis when there is auxiliary information such as survival time, the software stores the following results:

(i) genes defining a cancer subtype,

(ii) patient IDs that belong to a subtype,

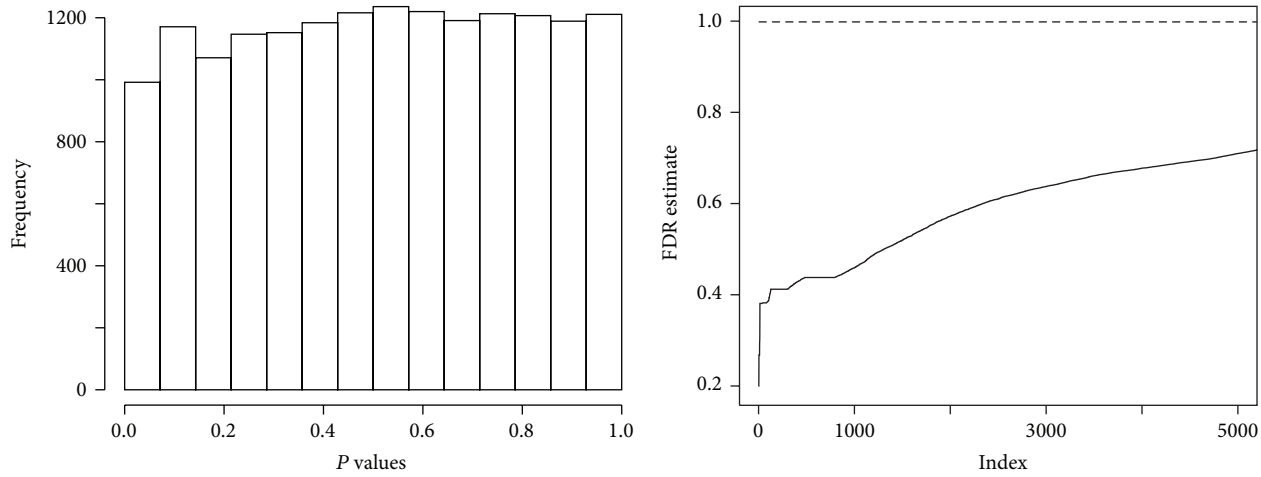
(iii) $N01$ and respective $P$ value.

Note that we may have high $N01$ by chance because several optimization procedures (e.g., the biclustering procedure) are performed before computing $N01$. To address this point, we randomly permute group labels of each subtype $N_p$ times and compute $N01$ based on the permuted data ($N01_{\text{perm}}$).

Then, we compute a standardized statistic of $N01$ for $i$th subtype:

$$z_i = \frac{N01_i - \overline{N01_i}}{s_i}, \tag{23}$$

where $\overline{\text{N01}_i}$ and $s_i$ are the mean and standard deviation of $N01_i$ and $N01_{\text{perm}}$'s. $N_p = 50$ is used in practice. Likewise, we standardize $N01_{\text{perm}}$. This standardization enables us to have precise estimate for $P$ value and reasonable resolution for estimating FDR. After stacking all the standardized statistics




(a) Histogram of $P$ values (b) FDR

FIGURE 5: (a) $P$ value distribution of two-sample $t$-statistics for detecting differentially expressed genes from a lung cancer data comparing relapse versus no relapse and (b) the corresponding false discovery rate (FDR) estimate. In (b), the dashed line is the standard FDR estimate and the solid line is from our proposed procedure. The $x$-axis in (b) denotes the ranking of genes where higher ranking corresponds to higher statistical significance.

in a vector $z_{\text{perm}}$, the $P$ value of $N01$ for $i$th subtype is defined as

$$P\text{-value}_i = \frac{\sum_k I\left(z_i \le z_{\text{perm},k}\right)}{K}, \tag{24}$$

where $K$ is the length of $z_{\text{perm}}$ and $z_{\text{perm},k}$ is the $k$th element in $z_{\text{perm}}$. Thus, the subtype with large $P$ value (24) will not be considered as an interesting cancer subtype even though it has high $N01$.

To find clinical implication of the subtype, we evaluate the prognostic signature in the subgroup of patients using the logistic regression with L1 penalty. We first classify patients belonging to the subgroup into good and poor prognosis groups based on cross validated probabilities of being relapsed patients from the logistic regression. Then, the strength of the prognostic signatures from the logistic regression is assessed by computing the survival difference between good and poor prognosis groups and the area under the operating characteristic curve (AUC).

## 7. Real Data Analysis

*7.1. Chemores Data Example.* Lung cancer is one of the most prevalent and deadliest cancers. Human lung cancers are classified into two major subtypes, small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC, which accounts for around 80% of all primary lung cancers, is a known heterogeneous group and its prognosis is generally poor [16]. In the current clinical practice, it is difficult to perform histopathological classification with small biopsies [17]. In order to improve the selection of patients who most likely will benefit from adjuvant chemotherapy (ACT), there is an urgent need to establish new diagnostic tools.

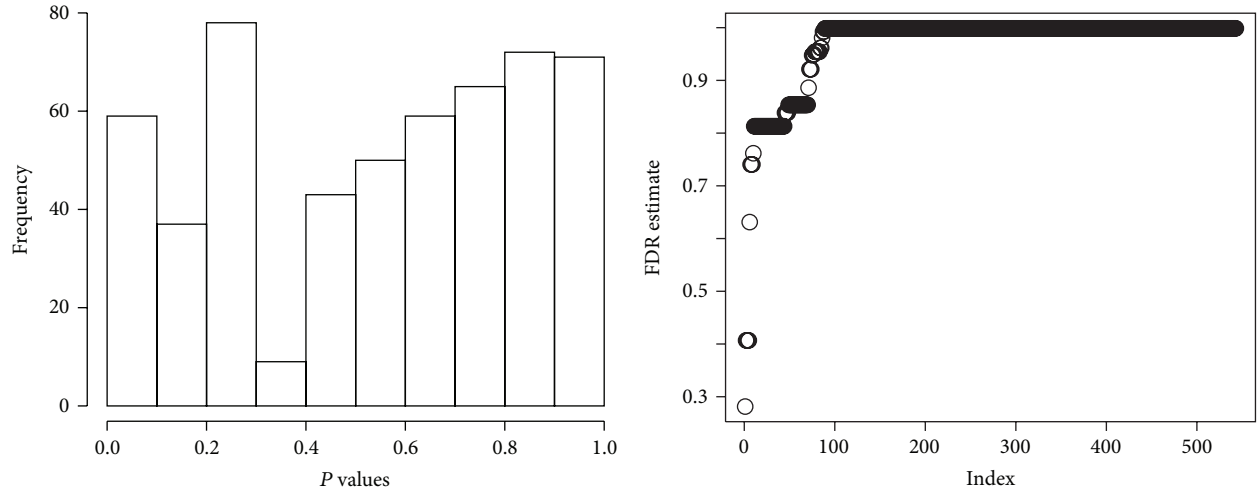
In this view, a study was organized by the Chemores initiative, which became an EU funded (FP6) Integrated Project involving 19 academic centers, organizations for cancer research, and research-oriented biotechnology companies in 8 European countries. Tissue samples from a cohort of 123 patients who underwent complete surgical resection between 30 January 2002 and 26 June 2006 are analyzed. All the patients belong to NSCLC and 59 patients experienced a relapse. This group of patients represents a heterogeneous group of lung cancers. We assayed the samples for gene expression, performed using dual-color human array from Agilent containing 41000 gene probes; a dye-swap of tumor versus normal lung tissue from same individual was employed for each sample and the log-ratio values were combined by averaging (the dataset is available at http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1132). Figure 5(a) shows the depletion in small $P$ values of the two-sample $t$-statistics for the 41000 gene probes. Figure 5(b) shows the corresponding pessimistic standard FDR estimate by [11] (dashed line). Thus, to take into account the heterogeneity issue properly, CAMS is needed here. Two inputs for implementing CAMS are a gene expression matrix and a relapse indicator. Table 1 shows a summary of output. The first column of this output contains unique names of subtypes. The second and third columns tell how many genes are involved in defining each subtype and the number of patients in the subtype. The $P$ values in the last column are computed using (24). The full lists of genes and patients can be identified by SubtypeID.

To reduce the computation time further, we consider filtering out uninteresting cases in the first stage. We compute $N01$ through the FDR based on the normal approximation only. We call this $N01_0$. If $N01_0$ is small, we skip the remaining procedure and go to search for next subtype. Figure 6(a) shows histogram of $P$ values for $N01$ after filtering out the uninteresting cases having $N01_0 \le 2$. The standard FDR estimate is given in Figure 6(b), showing some interesting

Table 1: The output from *R* package.

| Subtype_ID | Genes_in_clusters | Patients_in_subtype | *N*01 | *P* value* |
|---|---|---|---|---|
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 6 | 1535 | 69 | 17 | 0.020 |
| 7 | 124 | 28 | 23 | 0.020 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

*The *P* value of *N*01 for *i*th subtype is computed by using (24).




(a) Histogram of *P* values (b) FDR curve

Figure 6: (a) Histogram of *P* values for *N*01 from a lung cancer data and (b) the corresponding false discovery rate (FDR) estimate.

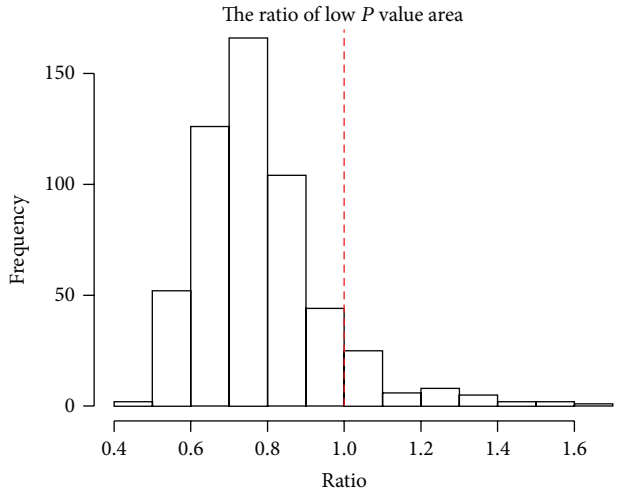
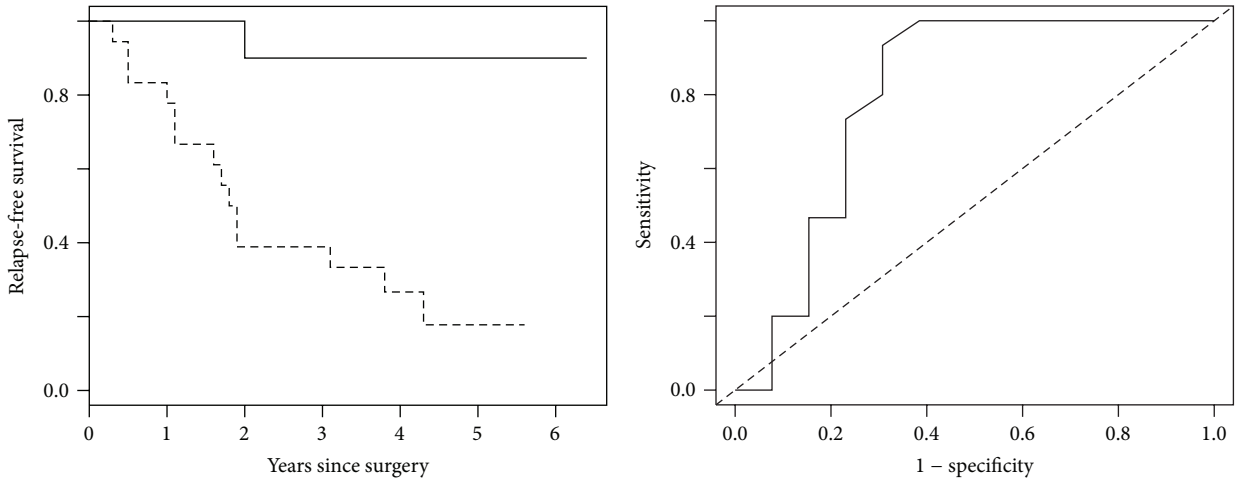



Figure 7: The proportion of subtypes showing depleted *P* value distribution from our clustering results is 0.908 (left side of vertical dashed line).

subtypes. In this analysis, we compute the proportion of subtypes showing depleted distributions with (22) and it is 0.908 (Figure 7). Therefore, *N*01 is essential in assessing the quality of each subtype.

From the top list of subtypes, one promising subtype is further analyzed using survival information to compute the appropriate prognostic signature for that subtype. To deal with large number of predictors (genes) we use logistic regression with L1 penalty [18] where the relapse status is the response variable. The cross validated probability of being a relapsed patient is computed from the leave-one-out cross validation, and the poor prognosis group is defined as the patients having the probability ≥0.5. To assess the strength of the prognostic signatures from the logistic regression, we compute the survival difference between good and poor prognosis groups. In Figure 8(a), the Kaplan-Meier curves of relapse-free survival show big difference between those two groups. Figure 8(b) shows operating characteristic curves for identifying relapse during follow-up. The area under the curve (AUC), computed under leave-one-out cross validation, is 0.806.

*7.2. Bild et al.'s Data Example.* As another application, we use lung cancer data by Bild et al. [19]. Their research purpose was to identify gene expression signatures of human cancers that reflect the activity of a given pathway. The gene expression dataset for lung cancer consists of 53 squamous cell carcinomas (SCC) and 58 adenocarcinomas (AC), so we expect that the group of patients represents a heterogeneous group. Among 58 relapsed patients, 26 and 32 patients belong to SCC and AC, respectively. The expression




(a) Kaplan-Meier Curve (b) ROC curve

Figure 8: (a) Kaplan-Meier curves of good and poor prognosis groups for a promising subtype and (b) receiver operating characteristic (ROC) curve.

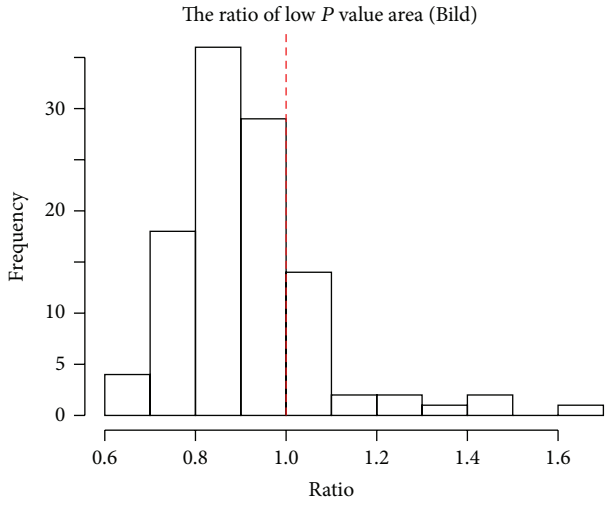



Figure 9: The proportion of subtypes showing depleted $P$ value distribution from our clustering results is 0.798 (left side of vertical dashed line).

dataset was obtained using Human U133 2.0 plus arrays (Affymetrix) containing 56475 gene probes. It is available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3141. For the downstream analysis, we normalized the dataset for each patient to have zero mean after taking logarithm. The same procedures as described in analyzing Chemores data were applied to the normalized data. The proportion of subtypes showing depleted distributions is 0.798, so $N01$ is crucial in assessing the quality of each subtype. See Figure 9. Likewise in the previous section, further survival analysis can be done, but we omit the results here for brevity.

## 8. Discussion and Conclusions

In this paper, we proposed an extensive clustering algorithm to find cancer subtypes and have addressed the heterogeneity issue induced by the unobserved group to assess the resulting subtypes appropriately. The unobserved group creates a serious conservative bias problem when standard FDR estimation is used, but our proposed FDR estimation method resolves it. SVD is used as a tool for discovering the effect of heterogeneity on the null distribution of the test statistics. In particular, when many datasets are considered simultaneously, we develop a much faster and more stable FDR estimation algorithm than the method in [9].

Although we focus only on the heterogeneity issue in this paper, Efron's three issues [7] should be considered simultaneously in high-throughput data analysis. It is difficult, however, to distinguish genes with small effects from correlation effects because both can produce similarly wide distributions of the test statistic. We also expect that there is some confounding between the heterogeneity effect and the above two effects. Thus, careful joint approaches for dealing with the three issues are required. For example, Pawitan et al. [8] showed that it is possible to get less bias by estimating $\pi_0$ and $f_1(z)$ using a joint estimation method. This issue needs further investigation.

Recently, several biclustering algorithms have been proposed for gene expression data, and a comparative study was performed in [20]. They pointed out that performance on synthetic datasets did not always correlate with that on real datasets and no algorithm is uniformly the best under different environments. Considering this point, CAMS is also expected to have its own weakness and strength. Thus, it is needed to study when CAMS performs well compared to other biclustering methods. On the one hand, it is possible to embed existing biclustering algorithms into CAMS with some modification. Then, we can compare performances of various biclustering methods when subtypes are assessed by $N01$.

In addition to the above issues, there are still many scientific questions to be considered here. For example, should two similarly constructed clusters be combined or

remained separate? How can we assign an independent test sample to newly constructed subtypes? A practical method for dealing with these scientific problems will require further research.

## Conflict of Interests

The authors declare that they have no competing interests.

## Authors' Contribution

Yudi Pawitan, Woojoo Lee, and Andrey Alexeyenko conceived the study and wrote the paper. Woojoo Lee and Andrey Alexeyenko performed data analysis. All authors read and approved the final paper.

## Acknowledgments

This work is supported by research grants from the European Union under the Chemores project and the Swedish Research Council and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A1061332). The support by BILS (Bioinformatics Infrastructure for Life Sciences) is also gratefully acknowledged.

## References

[1] C. M. Perou, T. Sørile, M. B. Eisen et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[2] T. Sørlie, C. M. Perou, R. Tibshirani et al., "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10869–10874, 2001.

[3] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00)*, pp. 93–103, San Diego, Calif, USA, August 2000.

[4] A. Tanay, R. Sharan, and R. Shamir, "Biclustering algorithms: a survey," in *Handbook of Bioinformatics*, A. Srinivas, Ed., Chapman & Hall, New York, NY, USA, 2004.

[5] G. Nowak and R. Tibshirani, "Complementary hierarchical clustering," *Biostatistics*, vol. 9, no. 3, pp. 467–483, 2008.

[6] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genetics*, vol. 3, no. 9, article e161, 2007.

[7] B. Efron, "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 96–104, 2004.

[8] Y. Pawitan, K. R. K. Murthy, S. Michiels, and A. Ploner, "Bias in the estimation of false discovery rate in microarray studies," *Bioinformatics*, vol. 21, no. 20, pp. 3865–3872, 2005.

[9] Y. Pawitan, S. Calza, and A. Ploner, "Estimation of false discovery proportion under general dependence," *Bioinformatics*, vol. 22, no. 24, pp. 3025–3031, 2006.

[10] O. Stegle, L. Parts, R. Durbin, and J. Winn, "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies," *PLoS Computational Biology*, vol. 6, no. 5, 2010.

[11] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[12] C. Friguet, M. Kloareg, and D. Causeur, "A factor model approach to multiple testing under dependence," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1406–1415, 2009.

[13] J. T. Leek and J. D. Storey, "A general framework for multiple testing dependence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 48, pp. 18718–18723, 2008.

[14] N. Meinshausen, "False discovery control for multiple tests of association under general dependence," *Scandinavian Journal of Statistics. Theory and Applications*, vol. 33, no. 2, pp. 227–237, 2006.

[15] B. Efron, "Microarrays, empirical bayes and the two-groups model," *Statistical Science*, vol. 23, no. 1, pp. 1–22, 2008.

[16] M. Tsuboi, T. Ohira, H. Saji et al., "The present status of postoperative adjuvant chemotherapy for completely resected non-small cell lung cancer," *Annals of Thoracic and Cardiovascular Surgery*, vol. 13, no. 2, pp. 73–77, 2007.

[17] P. T. Cagle, T. C. Allen, S. Dacie et al., "Revolution in lung cancer new challenges for the surgical pathologist," *Archives of Pathology and Laboratory Medicine*, vol. 135, no. 1, pp. 110–116, 2011.

[18] J. J. Goeman, "$L_1$ penalized estimation in the Cox proportional hazards model," *Biometrical Journal*, vol. 52, pp. 70–84, 2010.

[19] A. H. Bild, G. Yao, J. T. Chang et al., "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, no. 7074, pp. 353–357, 2006.

[20] K. Eren, M. Deveci, O. Küçüktunç, and Ü. V. Çatalyürek, "A comparative analysis of biclustering algorithms for gene expression data," *Briefings in Bioinformatics*, 2012.