Estimating the Arm-Wise False Discovery Rate in Array Comparative Genomic Hybridization Experiments

Daniel P. Gaile[a b c 1], Elizabeth D. Schifano[d], Jeffrey C. Miecznikowski[abc], James J. Java[a], Devin McQuaid[e], Jeffrey M. Conroy[e], Norma J. Nowak[e]

[a]Department of Biostatistics, University at Buffalo, Buffalo, NY 14214-3000, USA
[b]New York State Center of Excellence in Bioinformatics and Life Sciences, Buffalo, NY 14203-1199, USA
[c]Department of Biostatistics, Roswell Park Cancer Institute, Buffalo, NY 14263-0001, USA
[d]Department of Statistical Science, Cornell University, Ithaca, NY 14853-3801, USA
[e]Cancer Genetics, Roswell Park Cancer Institute, New York 14263-0001, USA

Short title:

aFDR

[1] Corresponding Author. Department of Biostatistics, School of Public Health and Health Professions, 249 Farber Hall, University at Buffalo, 3435 Main Street, Buffalo NY 14214-3000, USA. Tel:+1(716) 881 8955. e-mail:dpgaile@buffalo.edu

Abstract

Array Comparative Genomic Hybridization (aCGH) is an array-based technology which provides simultaneous spot assays of relative genetic abundance (RGA) levels at multiple sites across the genome. These spot assays are spatially correlated with respect to genomic location and, as a result, the univariate tests conducted using data generated from these spot assays are also spatially correlated. In the context of multiple hypothesis testing, this spatial correlation complicates the question of how best to define a 'discovery' and consequently, how best to estimate the false discovery rate (FDR) corresponding to a given rejection region.

One can quantify the number of discoveries as the total number of spots for which the spot-based univariate test statistic falls within a given rejection region. Under this spot-based method, separate but correlated discoveries are identified. We show via simulation study that the method of Benjamini and Hochberg (1995) can provide a reasonable estimate of the spot-wise FDR, but these results require that the simulated spot assays are categorized as true or false discoveries in a particular way. However, laboratory researchers may actually be interested in estimating a 'regional' FDR, rather than a 'local' spot-wise FDR. We describe an example of such circumstances, and present a method for estimating the (chromosome) arm-wise False Discovery Rate. In this framework, one can quantify the number of discoveries as the total number of chromosome arms for which at least one spot-based test statistic falls into a given rejection region. Defining the discoveries in this way, both the biological and testing objectives coincide. We provide results from a series of simulations which involved the analysis of preferentially re-sampled spot assay values from a real aCGH dataset.

Introduction

The advent of array-based technologies has promoted the rapid growth of statistical research related to multiple hypothesis testing methods. Multiple testing procedures are designed to simultaneously test $m > 1$ hypotheses while controlling an error rate. Benjamini and Hochberg (1995) proposed controlling the false discovery rate (FDR), the expected proportion of Type I errors among the rejected hypotheses, to account for the inherent multiplicity problem of simultaneous testing. Ten years after the publication of this seminal work, estimation of false discovery rates remains an open and very active area of research with the bulk of the developments occurring in the context of expression microarray experiments. Dudoit et al. (2003) provides a comprehensive comparison and review of commonly used error rates in expression microarrays, including the FDR and family-wise error rate (FWER) among others. Recently FDR methodologies have been applied to the analysis of Array Comparative Genomic Hybridization (aCGH) (Nigro et al., 2005; Lai and Zhao, 2005), a technology which provides spatially correlated test statistics.

Array CGH datasets consist of observed values for a set of simultaneous spot assays designed to quantify relative genomic abundance (RGA) levels at multiple sites across the genome. The biological relevance of such datasets follows from the fact that deletions and amplifications of portions of the genome affect the expression of tumor-suppressor genes and oncogenes, respectively. The deletions and amplifications that provide a selective advantage for cell growth will proliferate, and ultimately result in tumor formation. Array CGH provides a means to detect and quantify genomic regions of abnormal RGA levels, such as chromosomal losses/gains or localized deletions/amplifications, relative to a reference sample. Although aCGH has recently been implemented with oligonucleotide, including Single Nucleotide Polymorphism (SNP) chips (Lucito et al., 2003; Barrett et al., 2004; Bignell et al., 2004; Huang et al., 2004), we focus only on Bacterial Artificial Chromosome (BAC) arrays in this manuscript. A typical BAC aCGH experiment involves hybridizing differentially labeled DNA from a tumor and reference sample to an array of mapped sequences (BACs). Ideally, the relative hybridization intensity of the tumor and reference signals at each BAC is proportional to the RGA of those sequences in the tumor and reference genomes. The surrogate measure for DNA relative abundance is thus the $\log_2$ fluorescence ratio, i.e., the $\log_2\left(\frac{\text{Tumor}}{\text{Control}}\right)$ value, at each BAC. A more complete discussion on aCGH can be found in Pinkel et al. (1998) and Snijders et al. (2001). For information on aCGH and its applications in cancer, see Pinkel and Albertson (2005).

Array CGH datasets are often analyzed with the goal of identifying genomic regions for which RGA is associated with a clinical outcome such as disease-free survival (progressors vs. non-progressors). In such a setting, one might consider conducting the following set of tests:

Hypothesis Set 1

H_{0i} : Relative genomic abundance of the i^{th} BAC **is not** related to clinical outcome.

versus

H_{1i} : Relative genomic abundance of the i^{th} BAC **is** related to clinical outcome.

for $i = 1, \ldots, I$

where I is the total number of BACs. The spatial correlation of the BAC assays complicates the question of how best to define a 'true discovery' in the context of this multiple hypothesis testing. One obvious approach is to define a true discovery as any BAC for which there exists a relationship, regardless of strength, to the clinical outcome. We will refer to this as the 'standard' definition of a true discovery as it is the definition that currently dominates the literature. For a given analysis, the actual proportion of false discoveries for a specified rejection region is the number of false discoveries (i.e., the number of BACs for which the null hypothesis is true and the univariate test statistic falls within the rejection region) divided by the total number of univariate test statistics which fall within the rejection region[2]. We show via simulation study that if one defines true discoveries in the standard way then the method of Benjamini and Hochberg (1995), abbreviated BH henceforth, can provide a reasonable estimate of the spot-wise (i.e., BAC-wise) FDR. To our knowledge we are the first to provide such a study for aCGH data.

Not every BAC considered to be a discovery under the standard definition would be considered a discovery by a researcher. Candidate tumor-suppressor genes and oncogenes are believed to reside near the BACs with the strongest relationship to the clinical outcome. This follows from the assumption that BACs are related to clinical outcome via their correlation to the gene[3], and this correlation is inversely related to genomic distance. Thus, candidate gene identification is often obtained through database (e.g., GenBank, http://www.ncbi.nih.gov/Genbank/index.html) queries for known genes residing in the regions determined by the set of BACs which are 'most related' to the clinical outcome. Genomic regions spanned by the set of BACs which are weakly related to clinical outcome, but lie in close proximity to the set of 'most related' BACs are often not interrogated. Dismissal of such regions from further consideration is justified by the assumption that the weak relationship with clinical outcome is wholly explained by this close proximity of the 'most related' BACs to the putative tumor-suppressor genes or oncogenes. This set of non-interrogated weakly related BACs constitutes a set of true discoveries under the standard definition but are considered redundant and hence ignorable discoveries in practice.

Although we will show via simulation study that BH procedure can provide a reasonable estimate of the BAC-wise FDR, we argue the appropriateness of estimating this quantity. We claim that the definition of a true discovery associated with Hypothesis Set 2, which we will refer to as the 'biological' definition, coincides with the goals and practices of a typical aCGH researcher. Clearly this biological definition is not equivalent to the standard definition, hence, an estimate of the FDR corresponding to the standard definition may not be suitable in the aCGH context. Furthermore, we provide simulation results which illustrate that the standard FDR estimates can dramatically overestimate the FDR associated with Hypothesis Set 2.

Hypothesis Set 2

H_{0i} : The i^{th} BAC **is not** one of the closest flanking BACs
to a gene for which relative genomic abundance is related to clinical outcome

versus

H_{1i} : The i^{th} BAC **is** one of the closest flanking BACs
to a gene for which relative genomic abundance is related to clinical outcome

for $i = 1, \ldots, I$

Unfortunately, the literature has yet to provide the statistical machinery required to formally test the hypotheses in Hypothesis Set 2. In this manuscript we provide a partial solution to the problem by proposing a hypothesis set that can be evaluated formally and is of interest to the typical aCGH researcher. Namely, we propose consideration of Hypothesis Set 3 and argue that estimates of the FDR associated with discoveries defined in this context should be obtained, possibly in conjunction with those associated with the standard definition.

[2] If no univariate test statistics fall within the rejection region then define the true proportion of false discoveries as 0.

[3] By correlated, we mean that abnormal RGA for the BAC is correlated to abnormal RGA for the gene.

Hypothesis Set 3

$H_{0\ell}$: The ℓ^{th} chromosome arm **does not** contain a gene for which relative genomic abundance is related to clinical outcome

versus

$H_{1\ell}$: The ℓ^{th} chromosome arm contains **at least one** gene for which relative genomic abundance is related to clinical outcome

for $\ell = 1, \ldots, L$

The hypotheses in Hypothesis Set 3 can be formally evaluated by noting that, under reasonable assumptions, Hypothesis Set 3 is equivalent to Hypothesis Set 4; the set of hypotheses in Hypothesis Set 4 can be tested using our approach, provided the distributions of the L minimum p-values corresponding to the L chromosome arms can be estimated. We will consider a case where these estimates are obtained via permutation and provide a simulation study which demonstrates that the arm-wise FDR can be controlled under such conditions.

Hypothesis Set 4

$H_{0\ell}$: The ℓ^{th} chromosome arm **does not** contain a BAC for which relative genomic abundance is related to clinical outcome

versus

$H_{1\ell}$: The ℓ^{th} chromosome arm contains **at least one** BAC for which relative genomic abundance is related to clinical outcome

for $\ell = 1, \ldots, L$

Array CGH is not the only technology which produces spatially correlated test statistics. The problem of estimating the FDR in the context of quantitative trait loci (QTL) mapping is closely related to the problem which we consider. QTL mapping involves the estimation of association between identity by descent (IBD) status for a given marker and phenotype where the observed/estimated IBD values are spatially correlated as a consequence of the underlying biological processes (e.g., meiosis). Typical aCGH analyses involve the estimation of association between RGA for a given region (e.g., BAC) and a clinical outcome where the observed/estimated RGA levels are spatially correlated as a consequence of the underlying biological processes. In a recent publication, Benjamini and Yekutieli (2005) provide a detailed discussion concerning the estimation of FDR in QTL. Benjamini and Yekutieli (2005) also provide simulation results which demonstrate that if one defines true discoveries in the standard way then the BH method can provide a reasonable estimate of the marker-based FDR. Their findings are consistent with the results which we will present below.

Lee et al. (2002) proposed controlling an interval-wise error rate (IWER) for QTL analyses, implementing an FDR approach for least-squares regression interval mapping. The interval-wise error rate is analogous to the arm-wise error rate which we consider. Lee et al. (2002) introduce the FDRm, which is the interval-wise false discovery rate and is calculated with respect to the maximum F-statistic for each marker interval. A critical component of their algorithm involves the use of permutation-based density estimates for the maximum F-statistics. The method which we propose to control the arm-wise FDR utilizes permutation-based density estimates for the minimum p-values and is similar to their approach. Although Lee et al. (2002) apply their proposed methodology to a real dataset, they did not extend their work to include simulations to demonstrate that the method actually controlled the FDRm. Furthermore, Lee et al. (2002) engaged in direct comparisons of FDRm and standard FDR estimates while omitting any discussion of the fact that these methods estimated the FDRs for different sets of hypotheses.

In this manuscript, we propose the aFDR method to control the arm-wise FDR associated with Hypothesis Set 4. We also present a simulation study in which we consider the comparison of BAC assay values for two clinical subgroups (e.g., progressors and non-progressors). The BAC assay values were assigned to each sample in a manner designed to mimic underlying selective pressures for true biomarkers. The results of our study support the following claims:

1. The BH method controls the BAC-wise FDR, provided one defines true discoveries in the standard way.

2. The standard FDR estimates can dramatically overestimate the arm-wise FDR.

3. Our proposed algorithm controls the arm-wise FDR.

The underlying correlation structure of the test statistics is central to the issues we consider in this manuscript. We took great care to preserve this structure by constructing simulated datasets using preferentially re-sampled BAC assay values from a real aCGH dataset. These observed BAC assay values were sampled in a manner designed to preserve the correlation structure of the test statistics within each chromosome arm.

aFDR: a method to estimate the arm-wise false discovery rate

We propose to evaluate the hypotheses given in Hypothesis Set 4 as follows: 1) evaluate the complete set of univariate test statistics for association between RGA and the clinical outcome variable of interest, 2) calculate the complete set of p-values associated with the set of test statistics, 3) determine the minimum p-value observed on each chromosome arm, 4) adjust the observed minimum p-values such that they are uniformly distributed under the null hypothesis, and 5) obtain estimates of the arm-wise FDR using the adjusted minimum p-values as inputs to the BH procedure.

Let P_i represent the univariate p-value corresponding to the test statistic for the i^{th} hypothesis test given in Hypothesis Set 1. Define $P_{\ell(1)}$ to be the minimum p-value on the ℓ^{th} chromosome arm, i.e., $P_{\ell(1)} = \min_{i \in C_\ell} P_i$, where C_ℓ is the set of indices for all BACs located on the ℓ^{th} chromosome arm. Let F_ℓ represent the cumulative distribution function of $P_{\ell(1)}$ under the null hypothesis given in Hypothesis Set 4. If we define $U_\ell = F_\ell(P_{\ell(1)})$ then by the *Probability Integral Transformation Theorem* (Casella and Berger, 2002, page 54), $U_\ell \sim \mathrm{U}[0,1]$ provided the null hypothesis given in Hypothesis Set 4 is true.

We estimate F_ℓ via permutation, whereby $\widehat{F}_\ell$ is the empirical cumulative distribution function corresponding to $\{p^*_{\ell(1),1}, \ldots, p^*_{\ell(1),b}, \ldots, p^*_{\ell(1),B}\}$, the set containing the minimum p-value of the ℓ^{th} chromosome arm for each of the B dataset permutations. We thus define $\widehat{u}_\ell = \widehat{F}_\ell(p_{\ell(1)})$, where $\widehat{u}_\ell$ and $p_{\ell(1)}$ are the observed counterparts to U_ℓ and $P_{\ell(1)}$, respectively. Under the null hypothesis given in Hypothesis Set 4, we may consider $\widehat{u}_\ell \sim \mathrm{U}[0,1]$ and apply the BH procedure to these L adjusted minimum p-values to account for multiple testing.

Below is the proposed aFDR algorithm for controlling the arm-wise FDR.

```
The aFDR Algorithm:
Let $L$ be the total number of chromosome arms and let $B$ be the total number of permutations:

  1. Obtain estimates of $F_\ell$, $\ell = 1, \ldots, L$.
       – Permute data $B$ $(= 10,000)$ times
       – Obtain arm-wise minimum $p$-values: $p^*_{\ell(1),b}$, $b = 1, \ldots, B$, $\ell = 1, \ldots, L$
       – Let $\widehat{F_\ell}$ be the empirical cdf corresponding to
         $\{p^*_{\ell(1),1}, \ldots, p^*_{\ell(1),b}, \ldots, p^*_{\ell(1),B}\}$, $\ell = 1, \ldots, L$
  2. Obtain observed arm-wise minimum $p$-values: $p_{\ell(1)}$, $\ell = 1, \ldots, L$.
  3. Calculate $\widehat{u}_\ell = \widehat{F_\ell}(p_{\ell(1)})$.
       – In practice, use the 'rank-it': $\widehat{u}_\ell = \frac{\text{rank}(p_{\ell(1)})}{(B+1)+1}$, where $p_{\ell(1)}$ is ranked
         with respect to the set $\{p_{\ell(1)}, p^*_{\ell(1),1}, \ldots, p^*_{\ell(1),B}\}$, $\ell = 1, \ldots, L$, $b = 1, \ldots, B$.
  4. Perform BH procedure on $\widehat{u}_\ell$ values, $\ell = 1, \ldots, L$.
```

Simulation Study

We consider the comparison of BAC assay values for two clinical subgroups (e.g., progressors and non-progressors). Simulated datasets consisted of BAC assay values for n samples from each subgroup. Assignment of BAC assay values to the samples was designed to mimic underlying selective pressures for M true biomarkers by preserving the true correlation structure of the test statistics within each chromosome arm (described in greater detail below). The Wilcoxon Rank Sum test was selected to serve as the univariate test statistic for the hypotheses given in Hypothesis Set 1; this choice of test statistic was made primarily to satisfy computational considerations, and is frequently used in array-based analyses.

Description of Data

The dataset on which the simulations were based consists of aCGH data for $N = 143$ Wilms' tumor samples. The 143 Wilms' tumor samples were assayed with BAC arrays developed at Roswell Park Cancer Institute (RPCI) (Cowell and Nowak, 2003; Nowak et al., 2005). These arrays provided estimates of RGA for 5620 RPCI-11 BAC clones located across 39 autosomal chromosome arms[4]. The raw fluorescence values for the tumor samples were processed according to algorithms described in Gaile et al. (2006).

Selectively Weighted Observations

Each of the simulated datasets were constructed by preferentially re-sampling chromosome arms based on BAC assay values at the locations of ten putative biomarkers. The BACs which were selected to serve as putative biomarkers satisfied two properties: 1) evidence of a relatively high frequency of abnormal RGA was observed in the 143 samples for the given BAC, and 2) the BAC was in close proximity to a biomarker which has been discussed in the literature.

[4] Chromosome arms with insufficient BAC coverage were not considered in the analysis.

Let $\mathbf{X}_{(I\text{x}N)}$ represent the data matrix of the $\log_2\left(\frac{\text{Tumor}}{\text{Control}}\right)$ values, where the X_{ij}^{th} element is the $\log_2\left(\frac{\text{Tumor}}{\text{Control}}\right)$ value of the i^{th} BAC from the j^{th} tumor sample, $i = 1,\ldots,I,\ j = 1,\ldots,N$. We derive from $\mathbf{X}_{(I\text{x}N)}$ the simulated data matrix $\mathbf{W}_{(I\text{x}2n)}$ of $\log_2\left(\frac{\text{Tumor}}{\text{Control}}\right)$ values, where the rows again correspond to the I BACs but the columns are such that the first n correspond to the tumor samples for clinical subgroup A (e.g. progressors) and the second n columns correspond to the tumor samples for clinical subgroup B (e.g. non-progressors). Let $\beta(m)$ be the function that maps the m^{th} biomarker, $m = 1,\ldots,M$, to its BAC index i, and similarly let $\eta(m)$ be the function that maps the m^{th} biomarker to its chromosome arm index ℓ. Thus $X_{\beta(m)j}$ is the m^{th} biomarker $\log_2\left(\frac{\text{Tumor}}{\text{Control}}\right)$ value for the j^{th} tumor sample.

For a prespecified number of samples within each subgroup n, number of true biomarkers M, and selection weights w_m corresponding to the m^{th} biomarker ($m = 1,\ldots,M$), we propose the algorithm below to create simulated aCGH datasets which preserve intra-arm correlations.

```
Algorithm for Simulating Data:
Let $J = \{1,\ldots,N\}$, the set of all tumor sample indices.
For each chromosome arm containing a biomarker, $\eta(m),\ m = 1,\ldots,M$:

  1. Define sampling weights for each tumor sample and create index groups

       – Obtain $v_{j'}^{A}$ using the ‘rank-it’ method: $v_{j'}^{A} = \frac{\text{rank}(X_{\beta(m)j'})}{N+1}$, $j' \in J$, where
         ranks are evaluated with respect to $\{X_{\beta(m)j'} \mid j' \in J\}$.
       – Sample $n$ indices without replacement from $J$ according to weights
         $\rho_{j'}^{A} = (1 - v_{j'}^{A})^{w_m}$ to form index set $A$, $A \subset J$ and $|A| = n$.
       – Define $J_B = J \backslash A (= J \cap A^c)$, the set of all indices not contained in set
         $A$, and obtain $v_{j*}^{B}$ via the ‘rank-it’ method: $v_{j*}^{B} = \frac{\text{rank}(X_{\beta(m)j*})}{(N-n)+1}$, $j* \in J_B$,
         where ranks are evaluated with respect to $\{X_{\beta(m)j*} \mid j* \in J_B\}$.
       – Sample $n$ indices without replacement from $J_B$ according to weights
         $\rho_{j*}^{B} = (v_{j*}^{B})^{w_m}$ to form index set $B$, $B \subset J_B$ and $|B| = n$.

  2. Assign chromosome arms $\eta(m)$ to the simulated subgroups of size $n$

       – Assign the $n$ chromosome arms $\eta(m)$ indexed by set $A$ to the simulated
         subgroup A.
       – Assign the $n$ chromosome arms $\eta(m)$ indexed by set $B$ to the simulated
         subgroup B.

For remaining chromosome arms, randomly sample $2n$ chromosome arms without
replacement from $J$; assign $n$ arms to subgroup A and $n$ arms to subgroup B.
```

Simulation Parameters

Table 1 enumerates the parameter values considered in the simulation study, for a total of 48 possible parameter combinations. For each combination of parameters, 2800 replications were performed. Signal strengths (or selection weights), w_m, were chosen such that the probability of the minimum adjusted p-value across true biomarkers being less than or equal to α was approximately 0.90 for each sample size and number of true biomarker combination (i.e., $\mathrm{P}\left(\min_{(\text{true biomarkers})} p^{FWER} \leq \alpha\right) \approx 90\%$, for $\alpha = \{0.05, 0.50, 0.95\}$, corresponding to HIGH, MED, LOW signal strengths, respectively). This was accomplished using the permutation-based procedure of Westfall and Young (1993) to estimate the FWER. For example, in the LOW signal strength, one would expect that in about 90% of the replicates, at least one biomarker would be found significant at a FWER of .95. The LOW signal strength can be thought to represent an extremely underpowered experiment, where the copy number of the biomarker is truly different between groups, but the sample size is not large enough to detect it when the FWER is reasonably controlled.

Simulation Example

As an illustrative example, consider one simulation dataset consisting of a 20 progressors and 20 non-progressor ($n = 20$); for each sample, we have $\log_2\left(\frac{\text{Tumor}}{\text{Control}}\right)$ values obtained for 5620 BACs. This dataset was simulated as described above by preferentially re-sampling chromosome arms based on spot assay values at $M = 5$ known BACs (candidate biomarkers) from the Wilms' tumor aCGH dataset. This simulated dataset, as such, preserves the spatial correlation with respect to genomic location by chromosome arm.

For the i^{th} BAC, consider the following hypothesis test: H_{0i} : there is no difference (between progressors and non-progressors) in RGA at the i^{th} BAC, versus H_{1i} : there is a difference (between progressors and non-progressors) in RGA at the i^{th} BAC. We conduct this test for $i = 1, 2, \ldots, 5620$, using the Wilcoxon Rank Sum test statistic and obtain p-values, p_i, $i = 1, 2, \ldots, 5620$.

Figure 1 displays the $-\log_{10} p$-values by chromosomal location for the dataset under consideration. In Figure 1(a), chromosome 1 (arm q) contains one of the five biomarkers, where the $-\log_{10} p$-value of the biomarker is indicated by the red diamond. Noting that all $-\log_{10} p$-values of the BACs located on 1q are colored in light red, Figure 1(a) provides a representative example of the spatial correlation often found in aCGH experiments.

If we conduct each of the Wilcoxon Rank Sum tests at level $\alpha = 0.05$, then under the complete null hypothesis and under the assumption of independent tests, we would expect to find (erroneously) $5620 * .05 = 281$ significant results. To account for multiple testing, we apply the BH algorithm (using the R package `multtest`) to the 5620 observed p-values and obtain a threshold value which corresponds to a nominal FDR.

Consider this example in the context of Hypothesis Set 1, where each spot assay is a potential discovery. Here, one quantifies the total number of discoveries as the total number of

Variable	Description	Values
n	# Samples Within Group	$10, 20, 35, 50$
M	# True Biomarkers	$1, 2, 5, 10$
w_m	Signal Strength	Low, Medium, High

Table 1: Simulation Parameters.

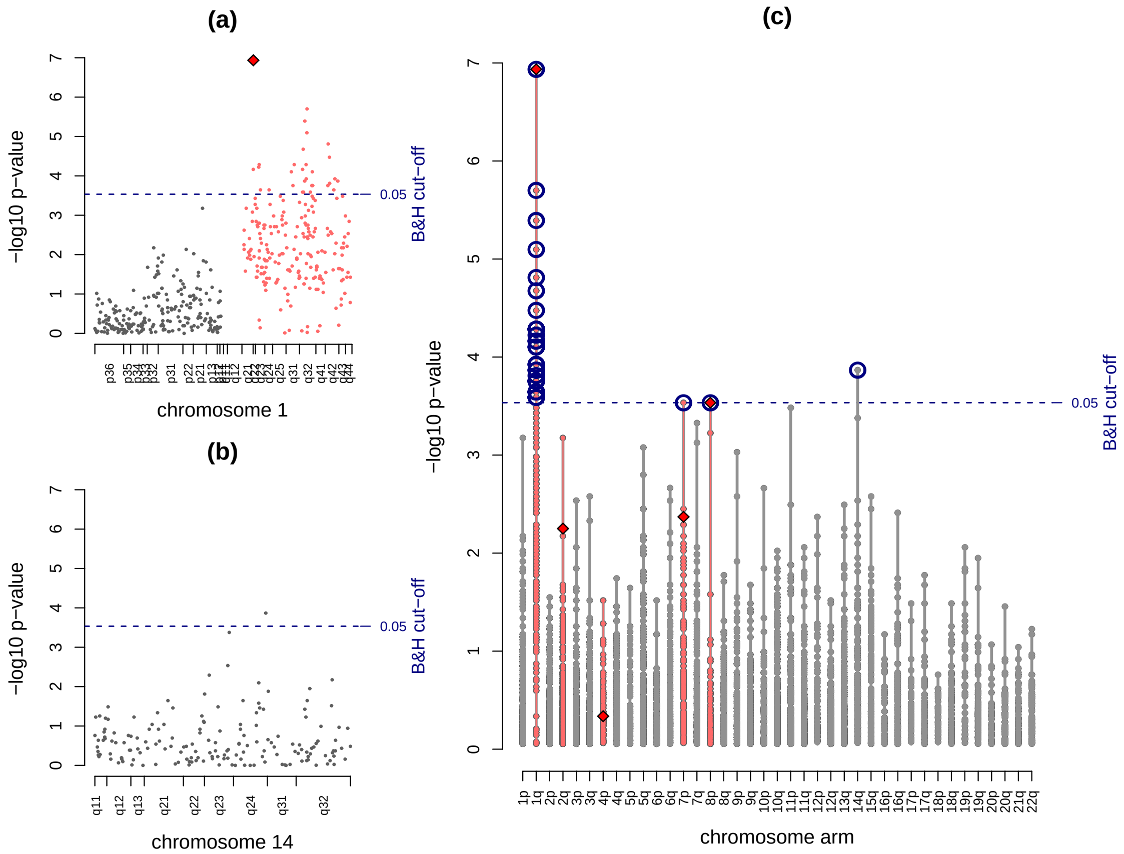

Figure 1: Genomic location plotted against $-\log_{10} p$-values with BH significance threshold at 0.05 for (a) chromosome 1, (b) chromosome arm 14q, (c) all 39 chromosome arms considered in the analysis. Red diamonds indicate $-\log_{10} p$-values of the biomarkers, and light solid circles represent all $-\log_{10}(p\text{-values})$ corresponding to BACs located on the same chromosome arm as a biomarker. The 31 blue-circled points represent all BAC-wise discoveries at BH significance threshold at 0.05.

BACs for which the test statistic, or p-value, falls within a given rejection region. In Figures 1(a-c), all BACs with $-\log_{10} p$-values greater than or equal to the BH significance threshold at 0.05 are called discoveries. The 31 blue-circled points in Figure 1(c), a condensed plot of the $-\log_{10} p$-values by chromosome arm, represent all BAC-wise discoveries at BH significance threshold of 0.05.

As indicated in by red diamonds in Figure 1(c), the true biomarkers reside on chromosome arms 1q, 2q, 4p, 7p and 8p. Thus, all BAC-wise discoveries located on these five chromosome arms are the standard true discoveries in the context of Hypothesis Set 1; the RGA of these BACs is related to progressor vs. non-progressor status. All but one of the 31 discoveries are standard true discoveries, as a discovery was also found on chromosome arm 14q (see Figure 1(b) and 1(c)). In the standard framework, the actual proportion of false discoveries for this dataset is consequently $1/31 = 0.032$.

In the biological framework, the red diamonds on chromosome arm 1q and 8p are the only true BAC-wise discoveries. Thus, in the context of Hypothesis Set 2, the actual proportion of false discoveries for this dataset is $29/31 = 0.935$.

Notice that if we define a discovery as a chromosome arm that contains at least one BAC for which RGA is related to progressor vs. non-progressor status and use the BAC-wise BH threshold, then there are 4 total discoveries with chromosome arm 14q as the only false discovery. Thus, Hypothesis Set 4 with a BAC-wise BH threshold yields an actual false discovery proportion of $1/4 = .25$ for this dataset.

To achieve BH FDR control in this example, we must have FDR $\leq .05$. It is important to note that the proportion of false discoveries is not equivalent to the FDR. The proportion of false discoveries is a value associated with one dataset, whereas the FDR is the *expected* proportion of false discoveries in the population of datasets. The BH procedure is designed to control the FDR, not the proportion of false discoveries. Thus, looking at this single dataset alone will not indicate whether the FDR is controlled, but the high proportions of false discoveries associated with Hypothesis Set 2 and Hypothesis Set 4 (using a BAC-wise BH threshold) in this example dataset are typical of those found in the simulation study. As such, it would appear that the BH procedure provides reasonable estimates of the BAC-wise FDR if using the standard definition of a true discovery, but not if using the biological definition. It also appears that one can not expect to control the arm-wise FDR using a BAC-wise BH threshold. The simulation results (below) corroborate these findings, which provided motivation for us to develop a method to control the the arm-wise FDR.

Simulation Results

Figure 2 provides simulation results to assess the BH FDR controlling procedure for the BAC-wise FDR. Each of the lines in Figures 2(a) and (b) represents one of the 48 possible parameter combinations, where the average proportion of false discoveries over 2800 replications is plotted against the nominal BAC-wise FDR according to the BH algorithm. This average proportion of false discoveries serves as an estimate of the FDR. Figure 2(a) indicates that the BAC-wise FDR using the standard definition was controlled under all conditions which were simulated. Figure 2(b) indicates that the BAC-wise FDR using the biological definition was not controlled under all conditions which were simulated. The number of true discoveries (biomarkers), M, had the most influence in estimation of the FDR, with the 10 true discoveries yielding the most conservative estimate (Figures 2(a) and 2(a1)). This observation is in agreement with the BH procedure's conservative estimate of the FDR. Figures 2(a2) and 2(a3) exhibit slight differences in FDR estimation, as a consequence of w_m values and n values, respectively. Selection weight, w_m, had the most effect in FDR estimation (Figures 2(b) and 2(b2)). Figures 2(b1) and 2(b3) indicate minor differences in FDR estimation, as a consequence of M values and n values, respectively.

Figure 3 provides simulation results to assess the BH and aFDR procedures control the arm-wise FDR. The lines in Figures 3(a) and 3(b) are defined as in Figures 2(a) and 2(b). Figure 3(a) indicates that the arm-wise FDR was not controlled by the BH procedure under all conditions which were simulated. Figure 3(b) indicates that the arm-wise FDR was controlled by the proposed aFDR method under all conditions which were simulated. Figures 3(a1) and 3(a3) show minimal effect of M values and n values in estimation of the arm-wise FDR. Figure 3(a2) indicates that the breakdown in control of the arm-wise FDR is related to selection weight. Figures 3(b2) and 3(b3) show minimal effect of selection weight and sample size on the control of the arm-wise FDR when the aFDR method is employed. Figure 3(b1) indicates that the number of true discoveries, M, had the most influence in estimation of the FDR, with the 10

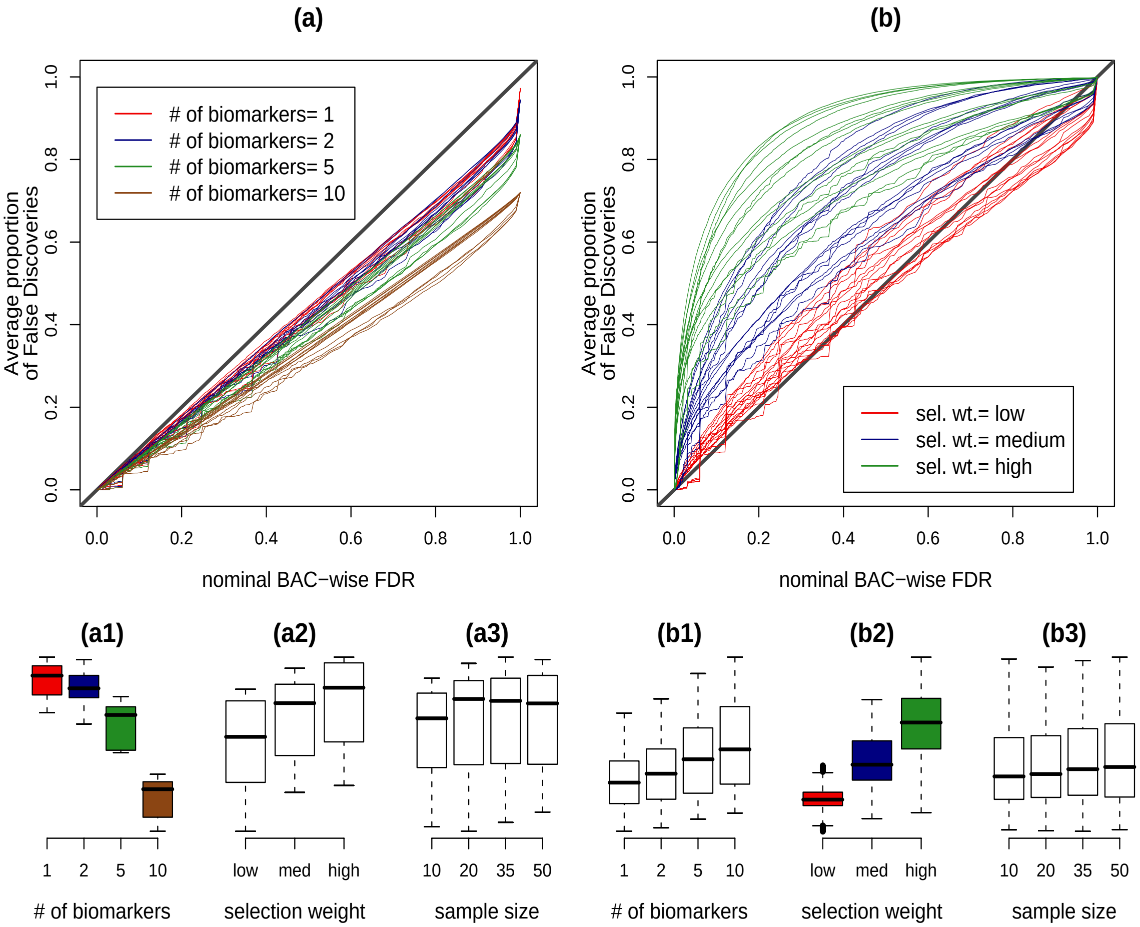

Figure 2: Nominal BAC-wise FDR plotted against BAC-wise FDR estimates each of 48 possible parameter combinations according to the (a) standard definition; red, blue, green, and brown correspond to 1,2,5,10 biomarkers, respectively, and (b) biological definition; red, blue, and green correspond to LOW, MED, and HIGH selection weights (i.e., biomarker signal strengths), respectively. Box plots (a1)-(a3) and (b1)-(b3) illustrate the difference in nominal FDR and the estimated FDR (for standard and biological definitions, respectively) based on (1) number of biomarkers, (2) selection weights, and (3) number of samples within the clinical subgroup. Vertical positions for each of the boxplots correspond to the estimated average proportion of false discoveries.

true discoveries yielding the most conservative estimate. As mentioned above, this observation is in agreement with the conservative nature of BH procedure.

Discussion

We have proposed a simple method for the estimation of the arm-wise False Discovery Rate in aCGH experiments and we have presented a simulation study which demonstrates that our

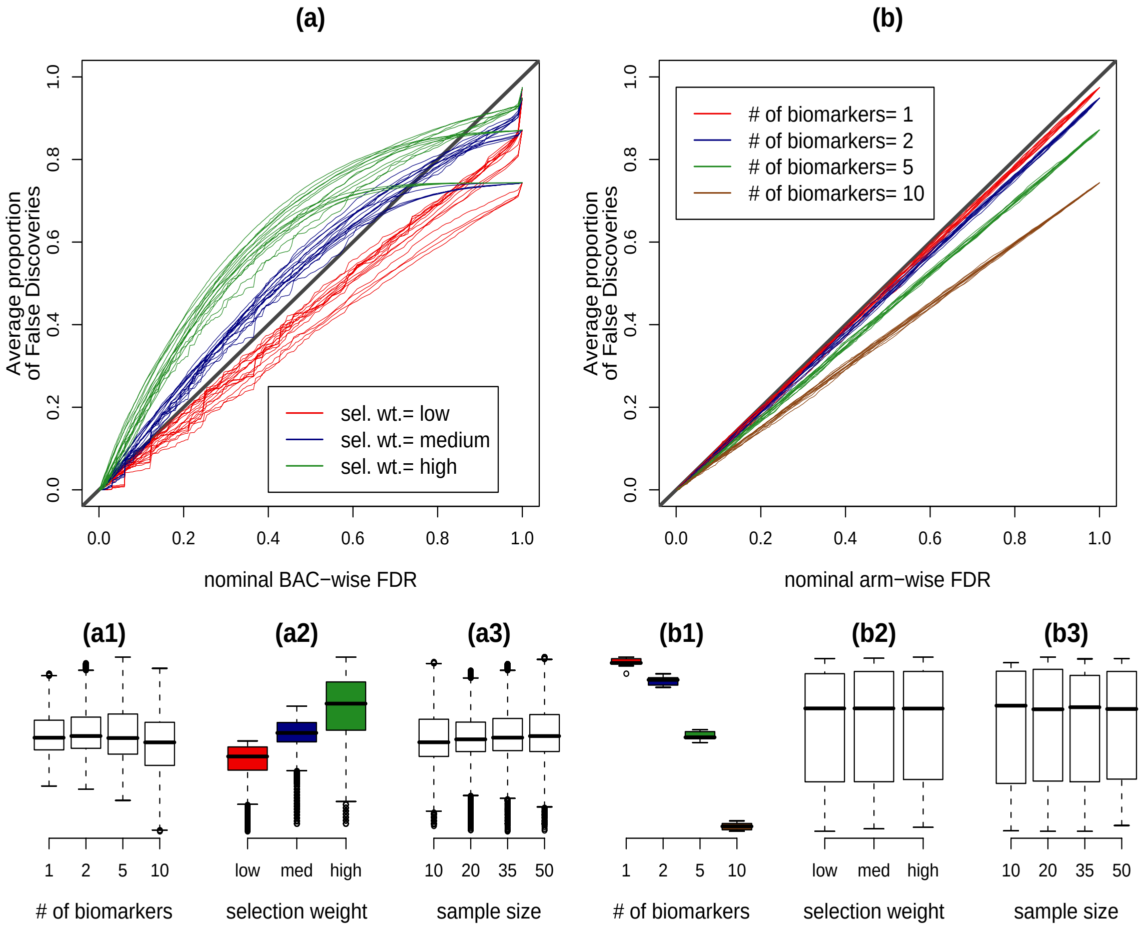

Figure 3: (a) Nominal BAC-wise FDR plotted against estimated arm-wise FDR for each of 48 possible parameter combinations; red, blue, and green lines correspond to LOW, MED, and HIGH selection weights (biomarker signal strengths), respectively. (b) Nominal arm-wise FDR plotted against estimated arm-wise FDR for each of 48 possible parameter combinations; red, blue, green, and brown lines correspond to the 1, 2, 5, 10 biomarkers, respectively. Box plots (a1)-(a3) and (b1)-(b3) illustrate the difference in nominal FDR and estimated FDR (for BAC-wise and arm-wise FDR, respectively) based on (1) number of biomarkers, (2) selection weights, and (3) number of samples within the clinical subgroup. Vertical positions for each of the boxplots correspond to the estimated average proportion of arm-wise false discoveries.

method exerts proper control. We have also provided evidence that the method of Benjamini and Hochberg (1995) can provide reasonable estimates of the BAC-wise FDR but that it can fail to control the arm-wise FDR. A direct comparison of the BH and aFDR methods is inappropriate as they are designed to control different quantities. Since neither method exerts control over both quantities (i.e., BAC-wise and arm-wise FDRs) then the method of choice should be based upon which of the two measures the researcher would like to control.

From a computational standpoint the simulations which we performed were rather involved and precluded running the analysis across several different aCGH datasets or employing different test statistics. However, the proposed method is not limited to Wilcoxon Rank Sum test statistics; other test-statistics could easily be accommodated as the procedure only relies on p-values. We chose the Wilcoxon Rank Sum test statistic for this analysis, not only because of its commonality and appealing lack of distributional assumptions, but also because of its computational efficiency. In particular, the p-values corresponding to each rank could be calculated and stored prior to running the analysis, so that a simple look-up procedure could be used during the actual analysis process. The proposed method can also be adapted to handle more than two clinical subgroups; any number of subgroups can be considered, as long as one can permute the data in the appropriate manner.

In the construction of the simulated datasets, the chromosome arms that did not contain a true biomarker were randomly selected and assigned to a sample. The chromosome arms that did contain the biomarkers, though preferentially placed in a particular subgroup, were randomly assigned to samples within the subgroups. Thus the simulation scheme incorporated the *intra*-arm correlation structure, but did not incorporate the *inter*-arm correlation structure. The inter-arm correlation was not introduced into our datasets, as it would further complicate the definition of a true discovery.

We have proposed a method to estimate the FDR associated with arm-wise partitions of the genome. However, the method can be applied across any set of partitions. As long as the genome is partitioned prior to the analysis or is partitioned in a way that depends upon the observed multivariate set of p-values, then inference conducted in the manner which we propose should be valid. Currently we are developing a method to fit a segmented mixture model to the aCGH $\log_2$ fluorescence ratios and applying out method to estimate the segment-wise FDR. In essence, our new approach segments the genomic locations using the observed $\log_2$ fluorescence ratios, but does not utilize information about population membership (e.g., progressors vs. non-progressors). This approach is rather involved and we have chosen to address it in a future manuscript.

An obvious extension of the work presented here is to examine the behavior of the aFDR algorithm in the context of QTL mapping. Another possible extension is to pre-cluster gene expression data and then apply the method to estimate the cluster-wise FDR.

Acknowledgments

This work was supported by grants from the National Cancer Institute (CA 27469, to the GOG Administrative Office; CA 37517, to the GOG Statistical Office; and CA16056–27–RPCI, Cancer Center Support Grant to Roswell Park Cancer Institute).

References

Barrett, M. T., Scheffer, A., Ben-Dor, A., Sampas, N., Lipson, D., Kincaid, R., Tsang, P., Curry, B., Baird, K., Meltzer, P. S., Yakhini, Z., Bruhn, L., and Laderman, S. (2004), "Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA," *Proc Natl Acad Sci U S A*, 101, 17765–17770.

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

Benjamini, Y. and Yekutieli, D. (2005), "Quantitative Trait Loci Analysis using the False Discovery Rate," *Genetics*, 171, 783–790.

Bignell, G. R., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigorova, M., Jones, K. W., Wei, W., Stratton, M. R., Futreal, P. A., Weber, B., Shapero, M. H., and Wooster, R. (2004), "High-resolution analysis of DNA copy number using oligonucleotide microarrays," *Genome Res*, 14, 287–295.

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, Duxbury Press.

Cowell, J. K. and Nowak, N. J. (2003), "High-resolution analysis of genetic events in cancer cells using bacterial artificial chromosome arrays and comparative genome hybridization." *Adv Cancer Res*, 90, 91–125.

Dudoit, S., Shaffer, J., and Boldrick, J. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103.

Gaile, D. P., Miecznikowski, J. C., Choe, S. E., and Halfon, M. S. (2006), "Putative null distributions corresponding to tests of differential expression in the Golden Spike dataset are intensity dependent. Technical Report 06-01." Tech. rep., University at Buffalo, State University of New York.

Huang, J., Wei, W., Zhang, J., Liu, G., Bignell, G. R., Stratton, M. R., Futreal, P. A., Wooster, R., Jones, K. W., and Shapero, M. H. (2004), "Whole genome DNA copy number changes identified by high density oligonucleotide arrays," *Hum Genomics*, 1, 287–299.

Lai, Y. and Zhao, H. (2005), "A statistical method to detect chromosomal regions with DNA copy number alterations using SNP-array-based CGH data." *Comput Biol Chem*, 29, 47–54.

Lee, H., Dekkers, J. C. M., Soller, M., Malek, M., Fernando, R. L., and Rothschild, M. F. (2002), "Application of the false discovery rate to quantitative trait loci interval mapping with multiple traits," *Genetics*, 161, 905–914.

Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J. A., Rostan, S., Nguyen, K. C. Q., Powers, S., Ye, K. Q., Olshen, A., Venkatraman, E., Norton, L., and Wigler, M. (2003), "Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation," *Genome Res*, 13, 2291–2305.

Nigro, J. M., Misra, A., Zhang, L., Smirnov, I., Colman, H., Griffin, C., Ozburn, N., Chen, M., Pan, E., Koul, D., Yung, W. K. A., Feuerstein, B. G., and Aldape, K. D. (2005), "Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma," *Cancer Res*, 65, 1678–1686.

Nowak, N. J., Snijders, A., Conroy, J. M., and Albertson, D. (2005), "The BAC Resource: Tools for Array CGH and FISH." *Current Protocols in Human Genetics*, pp. 1–34.

Pinkel, D. and Albertson, D. G. (2005), "Array comparative genomic hybridization and its applications in cancer," *Nat Genet*, 37 Suppl, 11–17.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B. M., Gray, J. W., and Albertson, D. G. (1998), "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays," *Nat Genet*, 20, 207–211.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D., and Albertson, D. G. (2001), "Assembly of microarrays for genome-wide measurement of DNA copy number," *Nat Genet*, 29, 263–264.

Westfall, P. H. and Young, S. S. (1993), *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*, Wiley.