Published in final edited form as:

Stat Biosci. 2017 June ; 9(1): 73–90. doi:10.1007/s12561-016-9156-x.

Improving Hierarchical Models Using Historical Data with Applications in High-Throughput Genomics Data Analysis

Ben Li¹, Yunxiao Li¹, and Zhaohui S. Qin^{1,2,*}

¹Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA

²Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA

Abstract

Modern high-throughput biotechnologies such as microarray and next generation sequencing produce a massive amount of information for each sample assayed. However, in a typical high-throughput experiment, only limited amount of data are observed for each individual feature, thus the classical ‘large p , small n ’ problem. Bayesian hierarchical model, capable of borrowing strength across features within the same dataset, has been recognized as an effective tool in analyzing such data. However, the shrinkage effect, the most prominent feature of hierarchical features, can lead to undesirable over-correction for some features. In this work, we discuss possible causes of the over-correction problem and propose several alternative solutions. Our strategy is rooted in the fact that in the Big Data era, large amount of historical data are available which should be taken advantage of. Our strategy presents a new framework to enhance the Bayesian hierarchical model. Through simulation and real data analysis, we demonstrated superior performance of the proposed strategy. Our new strategy also enables borrowing information across different platforms which could be extremely useful with emergence of new technologies and accumulation of data from different platforms in the Big Data era. Our method has been implemented in R package “adaptiveHM”, which is freely available from <https://github.com/benliememory/adaptiveHM>.

Keywords

Bayesian hierarchical model; historical data; informative prior; 450K methylation array; bisulphite sequencing

1 Introduction

In the past decade, high-throughput technologies, including microarray and next generation sequencing, have played an important role in biomedical research and produced unprecedented amount of data for different tissues, conditions or species. In spite of huge volumes of total available data, the number of samples in a specific experiment remains

*To whom correspondence should be addressed. zhaohui.qin@emory.edu.

limited because of complex sample preparation and relatively high cost. Even for more cost-effective microarray technology, the number of samples of the same condition investigated in an experiment is much smaller than the number of genes examined in each sample. Hence, the datasets with such characteristics become typical instances for “large p , small n ” problems (Fan and Lv 2010).

A hierarchical model (Good 1965), which is conceptually related to regularization techniques (Hastie et al. 2009), could be a valuable statistical tool for addressing “large p , small n ” problems. A variety of efforts have been made by statisticians to show effectiveness of a hierarchical model in the analysis of microarray gene expression data (Kerr and Churchill 2001; Newton et al. 2001; Parmigiani et al. 2003; Smyth 2004). In addition, the genomics research community, facing massive datasets produced by high-throughput technologies, has enthusiastically embraced hierarchical models (Ji and Liu 2010). Examples of hierarchical model applications include Limma (Smyth 2004) for Microarray, edgeR (Robinson et al. 2010), DSS (Wu et al. 2013) for RNA-seq, TileMap (Ji and Wong 2005) for ChIP-chip, PICS (Zhang et al. 2011) for ChIP-seq, Minfi (Aryee et al. 2014) and DSS-single (Wu et al. 2015) for methylation array and whole genome bisulphite sequencing WGBS data.

The key benefit of the hierarchical model lies in the fact that it enables “borrowing” information from other features (e.g. genes/probes in gene expression microarray or CpG sites in methylation array) to stabilize and improve the inference results for individual feature. Such a strategy has been shown to be much more reliable over naïve inferences especially when the sample size is limited, thus leads to more accurate downstream analyses. However, a hierarchical model stabilizes inferences by “regressing” all the inferences toward their means, which inevitably brings over-correction problems — for genes whose intrinsic variances are far lower or higher from the mean level, the inferences from the hierarchical model could be biased. In fact, the over-correction is not unexpected since all the genes involved in a typical microarray study perform rather diversely and the exchangeability assumption usually does not hold. Therefore, “borrowing” information from all genes (including the ones with different properties) may not be the best strategy.

A crucial assumption made by a hierarchical model is that all features are considered exchangeable. That is to say, one is unable to distinguish any given feature from the others given the data observed since all features are regarded homogeneous. We believe this is a rather strong assumption and in reality it is often violated. Genes in the genomes are designed to carry out different tasks. For example, the diverse biological features of developmental genes, housekeeping genes, response genes are reflected in their expression profiles measured from many different conditions. As shown in the Figure 1 of Li et al. (2015), there are substantial differences in terms of means and variances from gene to gene. Given heterogeneity in high-throughput genomics data, it is counter-productive for a highly and stably expressed housekeeping gene to borrow information from developmental genes with a bimodal expression pattern across experiments.

Just like many other fields, the total amount of available genomic data is enormous and is still growing rapidly in the era of Big Data (Fan et al. 2014). For example, there is a massive

collection of publicly-available datasets (referred to as historical data hereafter) produced by the gene expression microarray technology. The collection of historical microarray data is so rich that for a given new experiment, often times one is able to find datasets under similar conditions or of the same cell/tissue type from the collection. Therefore, it is highly desirable if we can improve the statistical inference of the new experimental data by utilizing these historical data. Over the years, numerous strategies have been proposed to leverage historical data to help analysis of new experimental data under various scenarios and contexts. As early as in 2004, Kim and Park proposed to utilize historical data to obtain improved estimate of the sample variance under a student t -test framework for detecting differentially expressed (DE) genes (Kim and Park 2004). However, they use historical variances directly in an adjusted t -test without incorporating current information. Altman and colleagues presented the singular value decomposition (SVD) Augmented Gene expression Analysis Tool (SAGAT) to increase the discovery power of microarray experiments by using publicly available microarray datasets (Daigle et al. 2010). They only use co-expression information from historical data and do not utilize other historical information. Wu and colleagues utilize a database of historical experiments to adjust background for the DNA microarray (Sui et al. 2009) but the historical data are not explored in DE gene detection settings. Therefore, with further accumulation of historical data, a model-based method better incorporating both historical and current data for DE gene detection is demanding.

On the other hand, statisticians increasingly recognize the importance of incorporating historical data into the inference procedure. In particular, Bayesian framework has been identified as the ideal vehicle that can be utilized to achieve this goal. Among them, the power prior (Chen and Ibrahim 2006; Ibrahim et al. 2015), which has been used in various fields including clinical trials, health care, etc., has been proposed to construct informative priors from historical data to improve the inference for current data (Duan et al. 2006; Hobbs et al. 2011).

In a recent bioinformatics study, Li et al. proposed a Bayesian strategy to incorporate historical data to help detect DE genes in microarray experiment (Li et al. 2015). The main idea, behind the proposed method, named informative prior Bayesian test (IPBT), is the construction of gene-specific, informative prior for the variance of each gene using historical data. IPBT is perhaps the first method that incorporates historical data in a formal Bayesian framework to detect DE genes. Despite its significant improvement over hierarchical model-based methods demonstrated using both simulated as well as real benchmark datasets, the success of IPBT hinges on the availability of large quantity of high quality historical data produced from the same platform, which limits the applicability of IPBT.

It is highly desirable if we can utilize historical data generated from a different platform. This is not possible using IPBT because this method relies on an inexplicit assumption that the current data and historical data (for each gene's expression measure) share similar distribution. To overcome this limitation, in this work, we propose a novel strategy to incorporate historical data into the hierarchical model framework. The central idea is to "partially" utilize historical data: instead of numerical values, we only retain the order of the genes in the genome ranked by their variances estimated from historical data. Thus under a

hierarchical model framework, when borrowing strength from other genes, instead of all genes in the genome, our approach selects a subset of genes that have the closest variances according to historical data. To be specific, we propose two different approaches, a stratified hierarchical model and a sliding window hierarchical model. In the first approach, we decompose all genes into disjoint groups such that borrowing strength only occurs among genes in the same group. The gene groups are determined by historical data such that the expressions of genes within a group are exchangeable. In the second approach, instead of fixed windows, we use a sliding window approach to group neighboring genes. The rest of the paper is organized as follows. In Methods section, we describe the details for our two new approaches. In Results section, we first conduct simulation studies to compare our two new approaches with IPBT and other competing methods. Then we undertake similar performance comparison to heart and brain gene expression data from the global map of gene expression (Lukk et al. 2010). Subsequently we apply our two new approaches to DNA methylation array data to show our new strategy can be applied to data generated from different platforms. Lastly we conclude the paper with additional remarks and possible future extensions in Conclusion and Discussion section.

2 Methods

2.1 Basic statistical setup

Let X_{ijk} denotes the log-transformed gene expression value after proper preprocessing and normalization, where i denotes the gene, j denotes the condition (control group or treatment group), and k denotes the replicate with $i = 1, 2, \dots, I$, $j = 1, 2$, and $k = 1, 2, \dots, n$. The basic assumption for the gene expression value is:

$$X_{ijk} | \mu_{ij}, \sigma_i^2 \sim N(\mu_{ij}, \sigma_i^2) \quad (1)$$

where μ_{ij} denotes the mean for the i th gene in the j th group and σ_i^2 is the variance for the i th gene. The mean expression for each gene is compared between different groups to identify DE genes. That is to say, for i th gene, we test the hypotheses: $H_0: \mu_{i,1} = \mu_{i,2}$ versus $H_A: \mu_{i,1} \neq \mu_{i,2}$. A naïve and direct tool is t -test. Genes can be ranked based on their t statistics and DE genes could be identified by p-values or false discovery rates (FDR).

2.2 Hierarchical model (HM)

Due to the limited sample size, two sample t -test is often not appropriate to detect DE genes. HM has been recognized as a powerful method to improve the DE gene detection. For example, the popular DE gene detection algorithm Limma (Smyth 2004) adopts an empirical Bayes hierarchical model. For illustration purpose, here we adopt the HM notation used in the analysis of tiling array data from Ji and Wong (2005):

$$X_{ijk} | \mu_{i,j}, \sigma_i^2 \sim N(\mu_{i,j}, \sigma_i^2) \quad (2)$$

$$\mu_{ij}|\mu_0, \tau_0^2 \propto 1 \quad (3)$$

$$\sigma_i^2|\nu_0, \omega_0^2 \sim \text{Inv-}\chi^2(\nu_0, \omega_0^2) \quad (4)$$

where the mean parameter μ_{ij} is assumed to be uniform and variance parameter σ_i^2 is assumed to follow an inverse- χ^2 distribution with hyper-parameters ν_0 and ω_0^2 . Subsequent adjusted t -test is performed with an empirical Bayes estimator $\widehat{\sigma_{i,B}^2}$ for σ_i^2 . Essentially, the adjusted variance estimate for each gene is a weighted average of the gene's sample variance and mean of all genes' sample variances.

HM improves DE gene detection results by borrowing information from all genes. This implies the assumption that all genes are exchangeable, which cannot hold in most scenarios. Hence, we propose stratified hierarchical model and sliding window hierarchical model which apply HM structure to subset of genes which are considered exchangeable.

2.3 Stratified hierarchical model (stHM)

Let $X_{i(g)jk}$ denotes the k th replicate of log-transformed expression for the i th gene in group g ($g = 1, 2, \dots, G$) under condition j . We have:

$$X_{i(g)jk}|\mu_{i(g),j}, \sigma_{i(g)}^2 \sim N(\mu_{i(g),j}, \sigma_{i(g)}^2) \quad (5)$$

$$\mu_{i(g)j}|\mu_g, \tau_g^2 \propto 1 \quad (6)$$

$$\sigma_{i(g)}^2|\nu_g, \omega_g^2 \sim \text{Inv-}\chi^2(\nu_g, \omega_g^2) \quad (7)$$

where mean parameters $\mu_{i(g),j}$ and variance parameters $\sigma_{i(g)}^2$ for genes in the same group are assumed to follow the same distribution with hyper-parameters ν_g and ω_g^2 . Similarly, an empirical Bayes estimator $\widehat{\sigma_{i(g),B}^2}$ for $\sigma_{i(g)}^2$ is used for the subsequent adjusted t -test. The main difference between stHM and HM is that stHM “borrows” information only from genes in the same group instead of all genes in the same experiment. With appropriately identified groups, stHM “borrows” information from more similar genes and could alleviate the over-shrinkage suffered by HM. All genes in an experiment are divided into G disjoint subsets based on the order of their standard deviations estimated from the collection of historical data. More details about determining the number of groups (G) are discussed in section 2.5.

2.4 Sliding window hierarchical model (swHM)

In this approach, borrowing strength for each particular gene under the hierarchical framework is restricted to its “neighboring” genes, again determined by the standard deviations estimated from the historical data. Following the notations in (5)—(7), swHM can be described as:

$$X_{i(g_i)jk} | \mu_{i(g_i),j}, \sigma_i^2 \sim N(\mu_{i(g_i),j}, \sigma_{i(g_i)}^2) \quad (8)$$

$$\mu_{i(g_i)j} | \mu_{g_i}, \tau_{g_i}^2 \propto 1 \quad (9)$$

$$\sigma_{i(g_i)}^2 | \nu_{g_i}, \omega_{g_i}^2 \sim \text{Inv-}\chi^2(\nu_{g_i}, \omega_{g_i}^2) \quad (10)$$

The swHM strategy enables the identification of a group of more homogeneous genes to estimate the gene’s adjusted standard deviation at the cost of slightly more computation burden. To quantify the computational time difference, we apply stHM and swHM to a dataset containing 10000 genes and five control samples and five treatment samples 100 times on a MacBook Air laptop computer with 1.6 GHz i5 CPU and 4G RAM. The average running time is 0.49 seconds for stHM and 1.04 seconds for swHM.

2.5 Group identification

Our main purpose is to divide genes into subsets in which genes are consider homogeneous. A straightforward strategy is to use each gene’s mean expression level estimated from the current data to select subsets. This strategy has been used in methods developed to detect DE genes from RNA-seq data (Robinson and Smyth 2007; Wu et al. 2013). In our stHM and swHM, we use the order of standard deviation estimated from historical data to determine subsets. We conduct a simulation study to demonstrate the accuracy of standard deviations (SD) estimated by different strategies. We obtain parameters for underlying distribution from real data (566 normal samples) in the global gene expression map (Lukk et al. 2010) and generate two normal samples with these parameters. Ten historical samples are randomly chosen from these 566 samples. We calculate (a) sample SD, (b) standard HM SD, (c) stHM SD with groups identified by sample mean (d) swHM SD with groups identified by sample mean (e) stHM SD with groups identified by historical SD rank (f) swHM SD with groups identified by historical SD rank. We compare the ranks of all these SD estimates with true SD rank in Figure 1, which shows that HM merely shrinks SD estimate without changing the order of them and the standard existing approach of using mean estimated from existing data to choose subset genes marginally improves SD estimates. Using historical data’s SD rank could significantly improve SD estimates. Our two approaches (stHM and swHM) have similar performances with the aid of historical data.

We also examined how changing the number of groups could affect the estimation of standard deviations with six different settings (Figure 2). We use the same procedure to generate historical data and two current samples as the simulation in Figure 1. We use stHM to estimate SDs with groups identified from historical SD ranks in six different settings: (a) without grouping (standard HM) (b) 10 groups (c) 50 groups (d) 100 groups (e) 200 groups (f) 500 groups.

In our stHM and swHM, we use the order of standard deviation estimated from historical data to determine subsets. We define “Group Dividing Metric” (GDM) to help decide on the optimal number of groups:

$$\text{GDM} = \left[\sum_g \frac{\sum_{i(g)} (S_{i(g)} - \bar{S}_g)^2}{I(g)} \right] / G \quad (11)$$

where $S_{i(g)}$ is adjusted SD estimate from stHM or swHM for i th gene in group g , $\bar{S}_g$ is the mean of SD estimates in group g , $I(g)$ is the total number of genes in group g , G is current number of groups. $S_{i(g)}$ is obtained by applying classical hierarchical model within each group. For completeness, we list the empirical Bayes estimator for SD below (Ji and Wong 2005):

$$S_{i(g)} = \sqrt{\left(1 - \widehat{B}_g\right) sd_{i(g)}^2 + \widehat{B}_g \overline{sd}_g^2} \quad (12)$$

$$\widehat{B}_g = \frac{2/v}{1+2/v} \frac{I(g)-1}{I(g)} + \frac{1}{1+2/v} \left(\frac{2}{v}\right) \left(\overline{sd}_g^2\right)^2 \frac{I(g)-1}{S_g} \quad (13)$$

where $v = 2(K-1)$ and $S_g = \sum_{i(g)} (sd_{i(g)}^2 - \overline{sd}_g^2)^2$

One issue worth noting is that with increased group number (fewer genes in a group), the empirical Bayesian estimate might be inappropriate (negative value for estimated variance) when the expressions for all the genes in a group are too close. We avoid such inappropriate scenarios by using the group mean SD as the estimate for all the genes in the group.

In Figure 3 we show that, at various number of total genes (1,000, 5,000 and 10,000), the trend for GDM (blue lines) with different number of groups for stHM and swHM. We also use the correlation between true SDs and estimated SDs from our approaches as the indicator of accuracy for estimated SDs (red lines). For stHM, GDM decreases with the increase of group number, with rapid but short-lived descending at the beginning. For swHM, GDM is stably small with small window size and increases slowly with the increase of window size. Figure 3 illustrates that GDM and the accuracy of estimating SDs have strong negative correlation for both stHM and swHM. Therefore, GDM could be used to select the optimal group number or window size. Since different group number will lead to similar results when GDM is stable, our rule of thumb for choosing optimal parameter is to

find the “elbow point” in the curve of GDM for stHM and find the stable region for swHM. A conservative choice is to pick a number slightly larger than the “elbow point” for stHM and pick a moderate window size when GDM is low and relatively flat for swHM. For example, we can choose 15 to 20 groups for stHM and pick 50 as the window size for a dataset with 1000 genes. In our R package “adaptiveHM”, we provide users the option of using GDM to determine optimal parameters before detecting DE genes or detecting DE genes with user specified parameters directly.

2.6 Models for DNA methylation data

Log ratios of methylated to unmethylated intensities (M value) are more widely used than ratios of the methylated to the total of methylated and unmethylated intensities (beta value) for 450K methylation array because M value performs similarly as gene expression data measured by microarray and all the methods on gene expression microarray can be applied almost identically to M value estimated from 450K array (Aryee et al. 2014; Robinson et al. 2014). We also apply our new approaches on M value directly (formulas (5–7) for stHM and (8–10) for swHM) and do not explicitly write out the models again. However, for sequencing based DNA methylation profiling approach (BS-seq), the basic model assumption is completely different. Beta binomial model is widely used in analyzing BS-seq data (Feng et al. 2014; Xu et al. 2015). This paper mainly discusses the improvement of HM in array data, thus we only show one state-of-the-art beta binomial Bayesian hierarchical model (DSS) for differential methylated locus (DML) calling of BS-seq data (Feng et al. 2014) and our stratified strategy (stDSS) to explore the possibility to borrow information across platforms. For the sake of completeness, we here rewrite the distribution assumptions made in DSS as follows:

$$X_{ijk}|p_{ijk}, N_{ijk} \sim \text{Binomial}(N_{ijk}, p_{ijk}) \quad (14)$$

$$p_{ijk} \sim \text{Beta}(\mu_{ij}, \Phi_{ij}) \quad (15)$$

$$\Phi_{ij} \sim \log\text{-normal}(m_{0j}, r_{0j}^2) \quad (16)$$

where X_{ijk} , N_{ijk} is methylation reads and total reads for i th CpG site, j th group and k th replicate, respectively. p_{ijk} is the underlying true methylation proportion. μ_{ij} , Φ_{ij} are the mean and dispersion parameter for beta distribution, respectively. And m_{0j} , r_{0j}^2 are mean and variance parameter for the log-normal distribution.

For stDSS, we modified formulas (14–16) into (17–19)

$$X_{i(g)jk}|p_{i(g)jk}, N_{i(g)jk} \sim \text{Binomial}(N_{i(g)jk}, p_{i(g)jk}) \quad (17)$$

$$p_{i(g)jk} \sim \text{Beta}(\mu_{i(g)j}, \Phi_{i(g)j}) \quad (18)$$

$$\Phi_{i(g)j} \sim \text{log-normal}(m_{gj}, r_{gj}^2) \quad (19)$$

2.7 Advantage over IPBT

In a recent study, the novel IPBT approach has shown to improve DE gene detection by using historical data (Li et al. 2015). Although it has been shown that IPBT is reasonably robust when the historical data is rather noisy, its performance is somewhat sensitive to the quality of the corresponding historical data since IPBT uses historical SD directly. Here, we show that stHM and swHM can tolerate more noise in historical data because these methods only utilize semi-quantitative information from historical data—only the order of the estimated standard deviations is used.

Another important benefit is that stHM and swHM make it possible to utilize data obtained across different platforms. Different technologies or platforms have been widely used to quantify the same biological phenomenon. For example, gene expression can be measured by both microarray or RNA-seq. DNA Methylation can be measured by Illumina 450k methylation array or BS-seq. These examples share the characteristics that one platform/technology is cost efficient and has accumulated massive of data while the other platform/technology is more advanced but with limited data due to its high cost. These different platforms/technologies have different measurement results even if not considering normalization issue across different platforms. Hence, it is difficult to use IPBT directly when available historical data are generated from a different platform. On the other hand, using semi-quantitative information makes stHM and swHM feasible in that scenario because different measures from different platforms are supposed to reflect the same biological intrinsic characteristics.

3 Results

3.1 Simulation

A simulation study is conducted to compare stHM and swHM with IPBT and other well-established methods for detecting DE genes: (i) Student's *t*-test, (ii) SAM (R package 'siggenes'); (iii) Limma, (R package 'Limma'); (iv) Z test using the true variance; and (v) IPBT (R package 'IPBT').

Expressions for 1000 genes in k (ranging from 2 to 5) samples are simulated for both the control and treatment groups. 10% of the 1000 genes (i.e. 100 genes) are randomly selected as designated DE genes. Gene expression values in both the control and treatment groups are assumed to follow normal distributions. The distribution parameters are obtained from real data in the global gene expression map (Luk et al. 2010). We derive each gene's sample mean and variance from the 566 normal samples in the collection. For the treatment group, the mean and variance of a gene's expression value are assumed to be the same as their

counterparts in the control group except for the DE genes. The mean expression values for DE genes in the treatment group are two standard deviations higher. For historical data used by IPBT, stHM and swHM, ten normal samples are randomly chosen out of 566 (without replacement) from the global gene expression map.

We use the empirical FDR (Benjamini and Hochberg 1995; Tusher et al. 2001) to evaluate the performance for the top 100 genes ranked by test statistics. The simulation is repeated 500 times for each method. Figure 4(a) summarizes the distributions of the 500 FDRs for different methods by box plots. All methods using historical data clearly perform better than methods without using historical data except for the Z test with true variances (considered the gold standard). The methods using historical data and Z test have fairly close performances. We also use Receiver Operator Characteristic (ROC) curves to compare different methods. Figure 4(b) shows a typical example of ROC curve for one single simulation with sample size $k = 2$. The ROC curves again show that methods with historical data perform better than methods without historical data except for the Z test, and the performances of methods with historical data and Z test are similar.

We also repeat the simulation with a noisier historical data. Figure 5 shows that IPBT's performance started to deteriorate with noisier historical data while our new strategies maintain its performance advantage. Figure 4 and 5 together demonstrate that our new strategies could be almost as good as IPBT with accurate historical data and perform more robust than IPBT when the historical data becomes noisier.

3.2 Microarray

Datasets—We apply all the methods except Z test appeared in the simulation to real gene expression microarray data contained in the global map of gene expression (Lukk et al. 2010). The dataset contains 369 different tissues, cell lines and disease states, from 5372 human samples. In this study, we use brain and heart data to conduct the comparisons. The dataset was downloaded from arrayExpress (ID: E-MTAB-62), then processed and normalized by robust multiarray analysis (RMA)(Irizarry et al. 2003).

Heart data—Since we do not know the true DE genes in the real data, we use agreement as the performance measure which is defined to be the proportion of overlap between two lists (equal length) of genes. Two normal (out of 36) and two disease (out of 51) heart samples are randomly selected and we treat the remaining 34 normal heart samples as historical data.

We then apply stHM, swHM and competing methods on these data to obtain a list of top 1000 DE genes for each method, respectively. Above sampling and testing procedure is repeated five times. Then we calculate the agreement between every pair of the 1000 DE gene lists for each method. Figure 6(a) summarizes the agreement results, which shows that stHM and swHM have a higher agreement than methods that do not use historical data (t -test, SAM and Limma) but not as good as IPBT.

We also conduct performance comparison on each of the five testing sets individually. As different methods perform almost the same with sufficient large sample size, we define the true DE genes by applying t -test to all the available heart data. Figure 6 (b) and (c) shows the

performances of different methods by their FDRs and ROC curves. Methods with historical data perform similarly and much better than methods that do not use historical data.

The analysis procedure for heart data is also repeated for brain data, comparing two normal brain samples (out of 39) and two disease brain samples (out of 31). Figure 7 shows the corresponding results for brain data. Again, our new approaches perform similar to IPBT and much better than other methods.

3.3 DNA methylation data

Datasets—DNA methylation 450K array is an array-based technology measuring more than 485,000 CpG sites. On the other hand, BS-seq, covering the whole genome (around 28 million CpG sites), is considered as a better technology for measuring DNA methylation.

Here we use 50 liver cancer (LIHC) and matched normal control samples from The Cancer Genome Atlas (Cancer Genome Atlas 2012). Detailed barcodes of all these samples can be found in the supplementary file. For BS-seq data, we use data from liver and hippocampus samples from the Roadmap Epigenomics project (Bernstein et al. 2010) (GEO accession number GSE64577).

Analyze 450K array data using 450K array historical data—Similar to Section 3.2 on gene expression microarray data. We take two normal and two cancer data and treat them as being collected from the current experiment; and all other normal data are used as historical data. Figure 8 (a) shows that stHM and swHM achieve better agreements than Limma and *t*-test. FDR and ROC curves in Figure 8 (b) and (c) again illustrate that methods with historical data could benefit from historical data.

Analyze BS-seq data using 450K array historical data—It is more useful if the historical data generated from a different platform can be utilized effectively. Here we use 450K array data to group all the CpG sites. And then compare DSS and stDSS (We only include the CpG sites appeared in the 450K array). We adopted the same procedure as Wu et al. (2015) did to preprocess the BS-seq data. Since it is not possible to know which loci are *bona fide* DMLs, we use the FDR estimates from DSS to compare the number of DMLs identified after controlling FDRs. Table 1 shows that how many DMLs are identified when controlling FDR at 0.01, 0.05 and 0.10, respectively. There are about 420,000 CpG sites involved in the analysis after quality control with Minfi excluding low quality CpG sites. For stDSS, we include two different group schemes (100 groups, each group has about 4,200 CpG sites and 4,500 groups, each group has fewer than 100 CpG sites). We can see that with the help of historical data, more DMLs can be identified. With more groups, this advantage could be even more obvious.

4 Conclusion and Discussion

In this paper, we introduce two new approaches (stHM, swHM) to detect DE gene or DML by improving the state-of-the-art hierarchical model with the aid of historical data. Simulation studies show that these two new approaches outperform the standard HM as expected and are more robust than the IPBT, another method that utilizes historical data. The

real data analyses demonstrate that our new approaches could be applied to a variety of datasets such as gene expression microarray, methylation array. We further show that our new approaches make it possible to borrow data from different platforms. This feature could be extremely useful since more and more array and sequencing data measuring similar underlying biological phenomenon are accumulating but can hardly be effectively analyzed together. In summary, HM is the most efficient method when historical data is not available. However, IPBT could be a better alternative than HM with available high quality historical data. When historical data is only available from other platform or the historical data is noisy, we believe stHM and swHM are better choices and we highly recommend them.

Our main purpose of this paper is to introduce the framework of improved hierarchical model and to illustrate that the framework can be generally applied to different types of data. However, the methylation 450K data and BS-seq data have their own specific characteristics. It is possible to further tailor our framework to 450K methylation array and BS-seq to obtain a better performance. In addition, our idea can also be extended to detect differential methylated regions (DMR) and to discover the possibility to borrow information between gene expression microarray and RNA-seq technology.

Acknowledgments

This work was partially supported by National Institutes of Health grant P01GM085354. The authors want to thank Yikai Wang and members of the bioinformatics interest group at Emory University for their insightful and constructive comments and suggestions.

References

- Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* (Oxford, England). 2014; 30:1363–1369. DOI: 10.1093/bioinformatics/btu049
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, Statistical methodology*. 1995; 57:289–300.
- Bernstein BE, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*. 2010; 28:1045–1048. DOI: 10.1038/nbt1010-1045
- Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012; 487:330–337. DOI: 10.1038/nature11252 [PubMed: 22810696]
- Chen MH, Ibrahim JG. The Relationship Between the Power Prior and Hierarchical Models. *Bayesian Analysis*. 2006; 1:551–574.
- Daigle BJ Jr, et al. Using pre-existing microarray datasets to increase experimental power: application to insulin resistance. *PLoS Comput Biol*. 2010; 6:e1000718.doi: 10.1371/journal.pcbi.1000718 [PubMed: 20361040]
- Duan YY, Ye KY, Smith EP. Evaluating water quality using power priors to incorporate historical information. *Environmetrics*. 2006; 17:95–106. DOI: 10.1002/env.752
- Fan J, Han F, Liu H. Challenges of Big Data Analysis. *Natl Sci Rev*. 2014; 1:293–314. DOI: 10.1093/nsr/nwt032 [PubMed: 25419469]
- Fan J, Lv J. A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica*. 2010; 20:101–148. [PubMed: 21572976]
- Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res*. 2014; 42:e69.doi: 10.1093/nar/gku154 [PubMed: 24561809]

- Good, IJ. The Estimation of Probabilities: An Essay on Modern Bayesian Methods. M.I.T. Press; Cambridge, Mass: 1965.
- Hastie, T., Tibshirani, R., Friedman, JH. Springer series in statistics. 2. Springer; New York: 2009. The elements of statistical learning : data mining, inference, and prediction.
- Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*. 2011; 67:1047–1056. DOI: 10.1111/j.1541-0420.2011.01564.x [PubMed: 21361892]
- Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. *Stat Med*. 2015; 34:3724–3749. DOI: 10.1002/sim.6728 [PubMed: 26346180]
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003; 4:249–264. DOI: 10.1093/biostatistics/4.2.249 [PubMed: 12925520]
- Ji H, Liu XS. Analyzing 'omics data using hierarchical models. *Nature biotechnology*. 2010; 28:337–340. DOI: 10.1038/nbt.1619
- Ji H, Wong WH. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* (Oxford, England). 2005; 21:3629–3636. DOI: 10.1093/bioinformatics/bti593
- Kerr MK, Churchill GA. Experimental design for gene expression microarrays. *Biostatistics*. 2001; 2:183–201. DOI: 10.1093/biostatistics/2.2.183 [PubMed: 12933549]
- Kim RD, Park PJ. Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol*. 2004; 5:R70.doi: 10.1186/gb-2004-5-9-r70 [PubMed: 15345054]
- Li B, Sun Z, He Q, Zhu Y, Qin ZS. Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes. *Bioinformatics* (Oxford, England). 2015; doi: 10.1093/bioinformatics/btv631
- Luk M, et al. A global map of human gene expression. *Nature biotechnology*. 2010; 28:322–324. DOI: 10.1038/nbt0410-322
- Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of computational biology : a journal of computational molecular cell biology*. 2001; 8:37–52. DOI: 10.1089/106652701300099074 [PubMed: 11339905]
- Parmigiani, G., Garrett, ES., Irizarry, RA., Zeger, SL. *Statistics for biology and health*. Springer; New York: 2003. The analysis of gene expression data : methods and software.
- Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, Zhou X. Statistical methods for detecting differentially methylated loci and regions. *Front Genet*. 2014; 5:324.doi: 10.3389/fgene.2014.00324 [PubMed: 25278959]
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (Oxford, England). 2010; 26:139–140. DOI: 10.1093/bioinformatics/btp616
- Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* (Oxford, England). 2007; 23:2881–2887. DOI: 10.1093/bioinformatics/btm453
- Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*. 2004; 3 Article3. doi: 10.2202/1544-6115.1027
- Sui Y, Zhao X, Speed TP, Wu Z. Background adjustment for DNA microarrays using a database of microarray experiments. *Journal of computational biology : a journal of computational molecular cell biology*. 2009; 16:1501–1515. DOI: 10.1089/cmb.2009.0063 [PubMed: 19958080]
- Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*. 2001; 98:5116–5121. [PubMed: 11309499]
- Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*. 2013; 14:232–243. DOI: 10.1093/biostatistics/kxs033 [PubMed: 23001152]

- Wu H, et al. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res.* 2015; 43:e141.doi: 10.1093/nar/gkv715 [PubMed: 26184873]
- Xu T, et al. Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Res.* 2015; 43:2757–2766. DOI: 10.1093/nar/gkv151 [PubMed: 25722376]
- Zhang X, Robertson G, Krzywinski M, Ning K, Droit A, Jones S, Gottardo R. PICS: probabilistic inference for ChIP-seq. *Biometrics.* 2011; 67:151–163. DOI: 10.1111/j.1541-0420.2010.01441.x [PubMed: 20528864]

Figure 1. Standard deviation (SD) ranks between different strategies. True SD ranks V.S. (a) Sample SD rank (b) Standard HM SD rank (c) Sample mean stHM rank (d) Sample mean swHM rank (e) stHM SD rank (f) swHM SD rank.

Figure 2.

True SD V.S. stHM with different group numbers: (a) without grouping (standard HM) (b) 10 groups (c) 50 groups (d) 100 groups (e) 200 groups (f) 500 groups

Figure 3.

GDM versus accuracy for estimated SD with different group numbers on datasets with different genes. All scenarios have two control samples and two treatment samples. (a) stHM with 1000 genes (b) stHM with 5000 genes (c) stHM with 10000 genes (d) swHM with 1000 genes (e) swHM with 5000 genes (f) swHM with 10000 genes.

Figure 4. Simulation with accurate historical data (a) FDR (b) a typical ROC curve

Figure 5.

Simulation with inaccurate historical data (a) accurate historical data (b) historical data with 20% unbiased noise (c) historical data with 30% unbiased noise (d) historical data with 50% unbiased noise

Figure 6.

(a) Agreement for heart data (b) FDR for heart data (c) ROC curve for heart data Brain data

Figure 7.

(a) Agreement for brain data (b) FDR for brain data (c) ROC curve for brain data

Figure 8.

(a) Agreement for Methylation 450K array (b) FDR for Methylation 450K array (c) ROC curve for Methylation 450K array

Table 1

Number of DMLs identified when controlling FDR at 0.01, 0.05 and 0.10.

# of DML	FDR < 0.01	FDR < 0.05	FDR < 0.10
DSS	1,305	1,992	2,528
stDSS with 100 groups	1,312	2,032	2,567
stDSS with 4,500 groups	1,797	2,819	3,534