Enhanced Probability Plots for Identifying Differential Gene Expression and Methylation Patterns in Breast Cancer Cell Lines via RNA-Seq and MethylC-Seq

Zhaonan Sun, Han Wu, Shuang He, Tianhong He, and Yu Zhu

Department of Statistics, Purdue University, West Lafayette, USA

ABSTRACT

Next generation DNA sequencing (NGS) is poised to replace microarray as the primary technology for research in genomics, epigenomics and transcriptomics in the near future. Two NGS-based methods RNA-Seq and MethylC-Seq provide researchers with unprecedented power to map and quantify genome-wide gene expression levels and DNA methylation patterns. In this study, we adopt a simple model-based graphical tool called enhanced half-normal probability plot (EHNplot) for identifying differential gene expression and DNA methylation patterns in eight breast cancer cell lines using the RNA-Seq and MethylC-Seq data provided by the IDEA challenge. EHNplot combines the usual half-normal probability with Scheffe’s simultaneous prediction intervals and can be considered a conservative testing procedure with acceptable power. The application of the EHNplot method leads to the identification of lists of important genes responsible for differential gene expression and differential methylation in the eight breast cancer cell lines. In particular, gene signatures for separating the ER+ cell lines and the ER- cell lines have been constructed.

1 INTRODUCTION

The National Cancer Institute estimated that there were 207,090 newly diagnosed breast cancer cases and 39,840 breast cancer related deaths among American women in 2010. Due to increased public awareness, early detection, and improved treatments, both breast cancer incidence and mortality rates have been declining in the past decade. Nonetheless, breast cancer remains to be the most common form of cancer among women in the United States. With the advent of human genomics and its rapid advancing in the past decade, researchers are now able to investigate the entire genome, epigenome and transcriptome of a tissue cell such as various types of cancer cells, and gain more precise and broad knowledge about the cell’s biological features at the molecular and genetic levels. This knowledge can be exploited to develop new treatments for targeted diseases and tailor existing treatments to individual patients based on their unique genetic characteristics.

DNA microarray has been the predominant technology used for gene expression profiling, which is to simultaneously measure the activity of thousands of pre-specified target genes. Its application to breast cancer research has produced many exciting results. For example, four different molecular classes of breast cancer have been identified, which include the basal-like, luminal-A, luminal-B, and HER2-positive breast cancers (Perou et al. 2000). Furthermore, the use of gene expression profiling in breast cancer research has also led to the development of a number of gene signatures that have demonstrated verifiable prediction power on clinical outcome.

DNA methylation refers to the addition of a methyl group to DNA and typically occurs in a CpG dinucleotide in mammals. It is a major form of epigenetic modification. Abnormal DNA methylation affects genome stability, alters gene expression levels, and contributes to the development of cancer. Microarrays have also been used to profile DNA methylation patterns at a whole genome scale. Holm et al. (2010) found that the aforementioned four molecular classes demonstrate different methylation profiles in more than 800 cancer related genes. Van Der Auwera et al. (2010) established a significant correlation between methylation and expression for more than 4 thousand genes. Earlier, Rakyan et al. (2008) observed a ubiquitous negative correlation between methylation in a gene’s promoter region and its expression but a positive correlation between methylation in a gene’s body and its expression.

Notwithstanding its tremendous success, microarray or array-based technology for gene expression and DNA methylation profiling suffers from a number of major limitations including reliance on existing knowledge about genome sequence, high background levels, a limited dynamic range, limited genome coverage, and requirement of normalization when comparing data generated from different microarray experiments. To overcome these limitations, a new technology called next generation DNA sequencing (NGS) has been developed in the past several years, and is poised to replace microarray as the main technology for research in genomics, epigenomics and transcriptomics. For mapping and quantifying the entire transcriptome, NGS technology leads to a method termed RNA-Seq that provides genome-scale gene-expression measurements and single-base resolution for annotation; see Wang et. al (2008) for a review. In a nutshell, RNA-Seq first converts the entire population of transcripts of a cell to a library of cDNA fragments with adaptors attached to one or both ends. After amplification, single-end or pair-end sequencing method is used to sequence the molecules in a high throughput manner, which produces millions of reads of lengths between 30 to 400 bp. The reads are then mapped to a reference genome or assembled de novo to produce a genome wide transcription profile. Compared to microarray, RNA-Seq enjoys clear advantages including its single-base resolution, lower background levels, a large dynamic range, full genome coverage, high level of accuracy and reproducitity. Most importantly, RNA-Seq does not rely on existing knowledge of genome sequences, therefore, it can be used to discover novel genomic activity of known genomes as well as to study the transcription of entirely unknown genomes. The NGS technology combined with DNA bisulfite conversion leads to a new method for DNA methylation profiling at the genome scale. Following
2 METHODS

2.1 Data Description

2.1.1 Paired-End RNA-Seq Data

We use the RNA-Seq data of the eight breast cancer cell lines provided by the iDEA challenge. The original data contain both 50 bp paired-end reads and 100 bp single-end reads. In this study, we only use the 50 bp paired-end reads, and treat the two ends of each such read as two single reads. All these reads are then mapped to the 18974 annotated genes in the RefSeq RNA database. Table 1 gives a summary of the resulting data. For each gene, we calculate the number of reads with starting position falling into the gene region and refer to the resulting count as the RNA-Seq read count of the gene. 17568 genes have at least one count in one of the eight cell lines, and 14406 genes receive at least one count in all the eight cell lines. We use reads per kilobase of gene model per million mapped reads (RPKM) proposed by Mortazavi et al. (2008) to quantify the expression level of a gene. The histogram of log2 gene expression level of all the genes in all the eight cell lines is shown in Figure S1 in the supplementary document.

Table 1. Summary of Paired-end RNA-Seq Data

<table>
<thead>
<tr>
<th>Index</th>
<th>Cell Line</th>
<th>Type</th>
<th>Total Reads</th>
<th>Mapped to Refseq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MCF-10A</td>
<td>Non-tumorigenic</td>
<td>69417757</td>
<td>33657128</td>
</tr>
<tr>
<td>2</td>
<td>BT-20</td>
<td>ER-</td>
<td>79298225</td>
<td>37909839</td>
</tr>
<tr>
<td>3</td>
<td>MDA-MB-231</td>
<td>ER-</td>
<td>74889322</td>
<td>37182328</td>
</tr>
<tr>
<td>4</td>
<td>MDA-MB-468</td>
<td>ER-</td>
<td>35604176</td>
<td>17279583</td>
</tr>
<tr>
<td>5</td>
<td>MCF-7</td>
<td>ER+</td>
<td>32597943</td>
<td>15984595</td>
</tr>
<tr>
<td>6</td>
<td>T47D</td>
<td>ER+</td>
<td>78434015</td>
<td>35921563</td>
</tr>
<tr>
<td>7</td>
<td>BT-474</td>
<td>ER+</td>
<td>72046561</td>
<td>31758368</td>
</tr>
<tr>
<td>8</td>
<td>ZR-75-1</td>
<td>ER+</td>
<td>58635346</td>
<td>27079742</td>
</tr>
</tbody>
</table>

The median of the total number of reads a gene receives in all the eight cell lines is 495. Since the extremely low read counts are considered subject to much noise, we exclude those genes with total read count less than 50 in the eight cell lines. As a result, 13404 genes are left. Figure 1 shows the histogram of the log2 gene expression levels of the remaining genes.

2.1.2 MethylC-Seq Data

We use the MethylC-Seq data particulary the individual CpG site methylation data provided by the iDEA challenge. As discussed in the Introduction, methylation can occur in the body and promoter region of a gene, which have different implications for gene expression and regulation. We focus on gene promoter methylation in this study. The promoter region of a gene is typically located in the upstream of the gene. We quantify the promoter methylation level of a gene by the average methylation level of the first 100 CpG sites nearest to the starting position of the gene. If there are less than 100 CpG sites in the upstream region of a gene, the average methylation level of all the CpG sites in the upstream region is used.
By the quantification method defined above, we obtain the promoter methylation levels of all the genes from the RefSeq RNA database in the eight cell lines. The histogram of the promoter methylation levels of all the genes in the eight cell lines is constructed and included as Figure S2 in the supplementary document.

The histogram shows a high concentration of gene promoter methylation levels at 0, which is consistent with the previous observation that in general, gene promoter methylation is not as prevalent as gene body methylation. In this study, we are more interested in differential gene promoter methylation patterns. Therefore, genes with persistent low or high methylation levels across all the eight cell lines are not interesting and relevant. Among all the genes, we filter out those genes with total methylation level less than 5 percentage points or greater than 720 percentage points. After filtering, 10114 genes remains. The histogram of methylation level of the remaining genes is given in Figure 2. Notice that the concentration at 0 is much mitigated. The remaining near zero methylation levels indicate that some genes are unmethylated in some cell lines but not in others.

2.2 Model

Both the gene-expression and methylation data obtained in the previous section can be represented by a \( m \times n \) table in which the rows \( i = 1, 2, \ldots, m \) correspond to the eight cell lines in Table 1, and the columns \( j = 1, 2, \ldots, n \) correspond to the genes. In the gene expression data, \( n = 13404 \), whereas in the gene methylation data, \( n = 10114 \). Let \( y_{ij} \) denote either the expression or the methylation level of gene \( j \) in the \( i \)-th cell line. We postulate the following ANOVA model to facilitate the differential gene expression and methylation analysis:

\[
y_{ij} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \epsilon_{ij} \tag{1}
\]

for \( i = 1, \ldots, m \) and \( j = 1, \ldots, n \), where \( \mu \) is the grand mean, \( \alpha_i \) are the overall or main effect of the \( i \)-th cell line, \( \beta_j \) is the main effect of the \( j \)-th gene, \( (\alpha \beta)_{ij} \) is the interaction effect between the \( i \)-th cell line and \( j \)-th gene, and \( \epsilon_{ij} \) is the error term with mean 0 and variance \( \sigma^2 \). We further impose a set of constraints on the effects as follows.

\[
\alpha_1 = 0; \tag{2a}
\]

\[
\sum_j \beta_j = 0; \tag{2b}
\]

\[
(\alpha \beta)_{1j} = 0 \text{ for } j = 1, \ldots, n; \tag{2c}
\]

\[
\sum_j (\alpha \beta)_{ij} = 0 \text{ for } i = 2, \ldots, 8. \tag{2d}
\]

By the constraints above, we treat the first cell line MCF-10A, which is non-tumorigenic, as the baseline, and compare all the other seven cell lines to the baseline. The intercept \( \mu \) is the average expression or methylation level of MCF-10A. The cell line main effects \( \alpha_i \) (\( i = 2, \ldots, 8 \)) represent the overall difference between the \( i \)-th cell line and the baseline, respectively; and the gene main effects \( \beta_j \) (\( j = 1, \ldots, n \)) now become the deviation of the expression or methylation level of gene \( j \) in the baseline from the corresponding average level in the baseline. The interaction effects \( (\alpha \beta)_{ij} \) (\( i = 2, \ldots, 8; j = 1, \ldots, n \)) represent gene-specific differential expression or differential methylation of gene \( j \) in cell line \( i \) in comparison to the baseline. Clearly, the interaction effects are the most important inference target in statistical analysis. We want to emphasize that (1) is a general model. When applied to the gene expression data, the effects represent the influences of cell lines, genes, and their interaction on gene expression, whereas when applied to the gene methylation data, the effects represent the influences of cell lines, genes, and their interaction on gene methylation.

2.3 Enhanced Half Normal Probability Plot (EHNplot)

In the gene expression or methylation data, there are \( 8n \) observations (i.e., \( y_{ij} \)'s). Model (1) has \( 8n \) free parameters. Because there are no replicated observations for any cell line by gene combinations, the \( 8n \) effects can be estimated but not directly tested statistically, due to the lack of degree of freedom. The least squares estimates of the parameters under the imposed constraints admit
explicit forms given below:
\begin{align}
\hat{\mu} &= \bar{y}, & (3a) \\
\hat{\alpha}_i &= y_i - \bar{y}, & (3b) \\
\hat{\beta}_j &= y_{ij} - \bar{y}, & (3c) \\
\widehat{(\alpha\beta)}_{ij} &= (y_{ij} - \bar{y}) - (y_{i\cdot} - \bar{y})_i. & (3d)
\end{align}

Loh (1992) proposed a graphical method for identification of active effects in unreplicated factorial experiments based on the normal probability plot. Hamada and Balakrishnan (1998) modified Loh’s method by using the absolute effects and the half-normal plot instead. Assume that the genes are differentially expressed or methylated, that is, \((\alpha\beta)_{ij} = 0\) for \(1 \leq i \leq 8\) and \(1 \leq j \leq n\), the estimates \((\alpha\beta)_{ij}\) form a sample from the same normal distribution with mean 0 and variance \(2(n - 1)\sigma^2/n\). Then, the normal probability plot of \((\alpha\beta)_{ij}\) or the half-normal probability plot of the absolute values of \((\alpha\beta)_{ij}\) is expected to demonstrate an overall linear trend. Any effects that deviate from this linear trend are considered significant, indicating differentially expressed or methylated genes and cell lines. To judge if an effect is following the overall linear trend or not can become subjective. Loh (1992) proposed a method to incorporate Scheffe’s simultaneous prediction confidence interval or prediction band into the normal probability plot for objective identification of significant effects. Hamada and Balakrishna (1998) took the same approach in their modification of Loh’s method.

The main idea used to construct Scheffe’s prediction band is to separate all the effects according to their absolute values into the so-called inliers and outliers. According to Loh’s original proposal, effects with magnitude less than four times the median magnitude of all the effects are classified as inliers. These inliers together with their corresponding standard normal percentiles are used to fit a linear trend or line, and then Scheffe’s prediction band are constructed along the fitted linear line. The effects that fall out of the band are declared significant. The critical value involved in constructing Scheffe’s prediction band depends on how many largest effects among the so-called outliers we decide to test. The maximizer of \(\hat{\eta}z_{(k)} + S(\{|F_{(i)}|\}_{\gamma})^{1/2} (1 + w)^{1/2}\) as follows.
\begin{equation}
\hat{\eta}z_{(k)} + S(\{|F_{(i)}|\}_{\gamma})^{1/2} (1 + w)^{1/2}
\end{equation}
where \(w = \sum_{i=1}^{|I|} |z_i|^2/\sum_{i=1}^{|I|} |z_i|^2\) and \(|I|\) is the cardinality of the set \(I\). \(F_{(i)}\) is the percentile of \(F\) distribution with degree of freedom \(l\) and \(|I|\). 5. Generate the plot including the scatter plot of \(|(\alpha\beta)_{(i)}|\) versus \(z_{(i)}\), the fitted linear line \(\hat{\eta}z_{(k)} + z_{(k)}\), and the Scheffe’s prediction band given above.

An example of EHNplot can be found in Figure 5. After the EHNplot has been generated, the \(l\) effects under testing are checked to see if they are falling into the Scheffe’s prediction band from the largest to the smallest or in a top down manner. Once the smallest effect that is falling outside the Scheffe’s prediction band is identified, this effect and the effects larger than it are declared to be significant and chosen.

The performance of the EHNplot depends primarily on the width of the prediction band, which is controlled by two factors, the confidence level \(\gamma\) and the number of effects (i.e. \(l\)) we want to test simultaneously. The higher the confidence level, the wider is the prediction band; and the larger \(l\) is, again the wider the prediction band. For any given \(\gamma\), we propose a procedure to determine \(l\) as follows. Let \(l\) range from 1 to a pre-specified integer \(M\); for each fixed \(l\) in the range, generate the EHNplot and find the number of significant effects, denoted by \(s(l)\); and then generate the plot of \(s(l)\) versus \(l\). The maximizer of \(s(l)\) is set to be the chosen \(l\). The chosen number \(l\) represents the maximum number of significant effects the EHNplot method can identify. Figure 4 is an example of the plot of \(s(l)\) versus \(l\). According to this plot, the maximum number of significant effects is 577.

3 RESULTS
3.1 Differential Gene Expression
The gene expression data is fitted by Model (1) in which the response is log2 of gene expression level in RPMK, and the interaction effects are estimated by the formula given in (3d). Figure 3 is the histogram of the estimated interaction effects. The set confidence level \(\gamma\) to be 95%. The EHNplots are generated for \(l\) ranging from 1 to 1000 and the plot of \(s(l)\) versus \(l\) is given in Figure 4, which indicates that the maximum number of significant interaction effects is 557. Therefore, \(l\) is determined to be 577, and the EHNplot with \(l = 577\) is given in Figure 5. There are 277 genes involved in the 557 selected significant interactions. In other words, these selected genes have demonstrated differential gene expression at least in one of the seven cancer cell lines in contrast to the baseline (i.e. the non-tumorigenic cell line). The names of the selected genes together with their \(P\)-values are included in Table S1 in the supplementary document. Furthermore, we generate the heatmap based on the interaction effects of the selected 277 genes (Figure 6). Notice that the left four columns of the heatmap corresponding to the cell lines BT-474, ER-75-1, MCF7 and T47D, form the ER+ group, while the left three columns corresponding to the cell lines MB-468, BT-20, and MB-231 form the ER- group. Clearly, from the heatmap, clusters of genes exist and can be used as signatures to separate the seven cell lines.
3.2 Differential Methylation

The methylation data is fitted by Model (1) with the gene methylation level as the response, and the interaction effects are estimated by formula (3d). The histogram of the estimated interaction effects can be found in the supplementary document (Figure S3). Compared to the gene expression data, the estimated interaction effects from the methylation data show a more heavy-tailed distribution, indicating that gene methylation are susceptible to higher variability than gene expression. The EHNplot method is used to select significant interactions effects. At 95% confidence level, 276 interactions are identified, which involve 231 different genes. In other words, these 231 different genes are significantly differentially methylated in some of the cancer cell lines in comparison to the non-tumorigenic baseline. To save space, we include the EHNplot with \( l = 276 \) (Figure S4), the list of the selected 231 genes (Table S2), and the heatmap (Figure S5) for the interaction effects of the selected genes in the seven cell lines all in the supplementary document.

3.3 Gene Signatures for Separating ER+ and ER-

In Sections 3.1 and 3.2, we have selected genes that demonstrate overall differential gene expression and methylation patterns in comparison to the non-tumorigenic cell line (i.e. the baseline). It is also possible to investigate differential gene expression as well as differential methylation between different subgroups of cell lines using contrasts. It is well-known that ER positive and ER negative are directly related to the grade of breast cancer and its clinical and prognostic outcome. In what follows, we use contrast together with the EHNplot method to identify gene signatures that can separate these two groups of cell lines at the molecular level. For gene \( j (1 \leq j \leq n) \), we define

\[
\begin{align*}
c_{j-} &= \frac{1}{3} \left[ (\alpha \beta)_{2j} + (\alpha \beta)_{3j} + (\alpha \beta)_{4j} \right] \\
c_{j+} &= \frac{1}{4} \left[ (\alpha \beta)_{5j} + (\alpha \beta)_{6j} + (\alpha \beta)_{7j} + (\alpha \beta)_{8j} \right] \\
c_j &= c_{j-} - c_{j+}
\end{align*}
\]
We remark that $c_j$ represents the difference between the expression or methylation patterns of gene $j$ in the ER positive group versus in the ER negative group. Because we have obtained the estimates of the interactions $(\alpha,\beta)_{ij}$, $c_j$ can be estimated simply by plugging in the estimated interaction effects in (5c), and we denote the resulting estimate by $\hat{c}_j$. We apply the EHNplot method to $\hat{c}_j$ to identify genes that demonstrate significant differential expression or methylation patterns and use them as the gene signatures.

### 3.3.1 Signatures for Differential Expression between ER+ and ER-

For differential gene expression between the ER+ and ER- cell lines, 165 genes are identified using the estimated contrasts $\hat{c}_j$ and the EHNplot method (See Figure S6 in the supplementary document). The heatmap for the interaction effects of the selected 165 genes in the seven cell lines is generated and shown in Figure 7. Clearly, the heatmap shows that the 165 genes are further divided into four clusters. We label the four clusters from the top to the bottom as 1, 2, 3 and 4. Genes in Cluster 1 and 3 demonstrate significant interaction effects in the ER- cell lines but not in the ER+ cell lines, whereas genes in Cluster 2 and 4 behave exactly in a way opposite to those in Cluster 1 and 3. Therefore, potentially, any of these four clusters or their combinations can be used as gene signatures to separate ER+ cell lines from ER- ones. The list of the selected 165 genes together with the four clusters is given in the supplementary document (Tables S3 and S5).

### 3.3.2 Signatures for Differential Methylation between ER+ and ER-

For differential methylation between the ER+ and ER- cell lines, 74 genes are identified using the estimated contrasts $\hat{c}_j$ and the EHNplot method (See Figure S7 in the supplementary document). Similarly, the heatmap of the interaction effects of these genes in the seven cell lines are generated and shown in Figure 8. Again, four clusters of the select genes emerge, which can be used as gene signatures to separate the ER+ and ER- cell lines. The list of the 74 selected genes together with the four clusters is given in the supplementary document (Tables S4 and S6).

### 3.4 Joint Analysis for Differential Expression and Methylation

As discussed in the Introduction, it is reported in the literature that methylation level is negatively correlated with gene expression level in some breast cancer related genes. Compared the gene lists we have identified from differential gene-expression analysis and differential methylation analysis, we found only a few common genes. In order to further investigate the possible correlation between differential gene expression and methylation, we have carried out a naive joint analysis as follows. First, we identify 7856 genes present in both the filtered gene-expression and methylation data; and second, we classify the interaction effects of these genes with the seven cell lines into four groups according to the four combination of their signs. For example, the first group consists of those gene by cell line interactions that are positive for both gene expression and methylation. In other words, for the genes and cell line combinations in the first group, all the genes demonstrate increased expression and methylation levels. We then apply the EHNplot method to select significant genes from this group only. This is repeated for the other three groups. However, no particular interesting patterns emerge from this joint analysis. The correlations between differential gene expression and methylation in these four groups are not found to be significant, except for a small number of genes. This result makes us wonder if the correlation between methylation and differential gene expression can be established by large scale analysis. Perhaps it also suggests that more sensitive models need to be developed.

### 4 DISCUSSION

Among the 277 genes selected by the differential expression analysis in Section 3.1 (Table S1), some have been identified and discussed in the literature before. For example, CDKN2A, FAS, FGFR5, PRLR, and ROS1 are listed among the tumor genes by the website http://www.tumor-gene.org/ggdf.html, and CA9 and ORC6L are among the list of 162 genes identified in Van’t Veer LJ et al. (2002) when developing their gene signature for predicting
clonal outcome of breast cancer. CDKN2A acts mainly as a tumor suppressor gene and interacts with other breast cancer-related genes such as P53. Its mutation and aberrant expression is related to the development of a variety of cancers. FAS encodes one of several important proteins related to apoptosis and is a member of the tumor necrosis factor receptor superfamily. FGF5 is an oncogene and belongs to the FGF family involved in a variety of biological processes including tumor growth and invasion. PRLR is a member of the type I cytokine receptor family and may function to modulate the effects of prolactin in normal tissue and cancer. ROS1 is known to be a proto-oncogene and highly-expressed in a variety of tumor cell lines. CA9 may be involved in the control of cell proliferation and transformation and ORC6L encodes a protein that plays an essential role in coordinating chromosome replication and segregation with cytokinesis.

Among the 231 genes selected by the differential methylation analysis in Section 3.2 (Table S2), four genes (GRB7, MEIS2, NOTCH3, RB1) are also in the list of tumor genes in the same website given above, and two genes (BIRC5, EZH2) appear among the 162 genes in Van’t Veer LJ et al. (2002), and one gene (PPP1CC) is included in the 76 gene signature developed by Wang Y, et al. (2005) for predicting distant metastasis of lymph-node-negative primary breast cancer. GRB7 is adjacent to ERBB2, a proto-oncogene on chromosome 17, both of which are found to be co-amplified in breast cancer and known to be over-expressed in other types of cancers. RB1 is a tumor suppressor gene, and BIRC5 is a member of the inhibitor of apoptosis (IAP) gene family with different gene expression patterns in normal and tumor cells.

Among the 165 genes identified in the differential gene expression analysis for ER+ and ER- cell lines in Section 3.3.1 (Table S3), there are five genes (AR, CDKN2A, EGFR, ESR1, PRLR) appearing in the list of tumor genes cited above and one gene (CD44) appearing in the list of Wang Y, et al. (2005). AR is known to be related to various types of cancer particularly breast cancer. Some studies suggest a possible relationship between the length of the CAG repeat region in the AR gene but additional research is required to verify this relationship. EGFR is identified as an oncogene with over expression in a number of cancers including breast cancer, lung cancer and colon cancer, which leads to the development of several anticancer therapeutics directly targeting this gene. ESR1 encodes an estrogen receptor involved in breast cancer and its expression levels are directly related to the distinction between the ER+ and ER- cell lines. The read counts of this gene in the RNA-Seq data are 108 (MCF-10A), 430 (BT-20), 62 (MDA-MB-231), 161 (MDA-MB-468), 2969 (MCF-7) 5544 (T47D) 1609 (BT-474) and 2599 (ZR-75-1). Clearly, ESR1 is over-expressed in the ER+ breast cancer cell lines. The protein encoded by CD44 participates in a wide variety of cellular functions including tumor metastasis, splice variants of this gene play important roles in the process of a number of cancers, and the transcription of this gene is in part activated by Wnt signalling, which is linked to tumor development. Among the 74 genes identified by the differential methylation analysis for ER+ and ER- cell lines in Section 3.3.2 (Table S4), one gene (GH1) is found in the list of tumor genes and one gene (TNFSF10) is found in the gene signature of Wang et al. GH1 is a growth hormone gene and found to be associated with the development of breast cancer (Wagner, et al. 2005). TNFSF10 codes a protein that preferentially induces apoptosis in transformed and tumor cells, but does not appear to kill normal cells. Comparing the methylation level of this gene in the cell lines, we found that it was highly methylated in the ER+ cell lines and one of the ER- cell lines, but not in the others.

Even though the groups of genes identified in this study share important common genes with other groups identified by previous studies, the overlap is relatively limited. This phenomenon is not unique to our study. As a matter of fact, it is rather common as found in the literature that different studies often lead to seemingly different lists of important genes. For example, the three gene signatures, the 70-gene (Van’t Veer et al. 2002), the 76-gene (Wang et al. 2005), and the gene expression grade index (Haibe-Kains et al. 2008) signatures share only a small number of common genes. Nonetheless, these three signatures have shown similar prognostic performance. One explanation for the similar performance is that many genes are different in identity but are correlated in gene expression and relevant to the same underlying biological processes. Therefore, many novel genes identified in this study are expected to play critical roles in different types of breast cancer and can be used as hints for further study using traditional experiments. Their relevance needs to be verified using real life breast cancer patients data.

The EHNplot method is a simple model-based graphical tool for identifying differentially expressed and methylated genes. It can also be applied in general for large-scale testing in genomic data analysis. Simulation study shows that the EHNplot method controls the family wide type I error and thus is a conservative procedure, compared to other procedures. In fact, the lists of genes selected by using the EHNplot method are usually subsets of those identified by false discovery rate control procedures. The EHNplot method as implemented in this study is based on the normality assumption. This restriction can be easily removed by using other probability distribution plots. The theoretical property of the EHNplot method in large scale application has not been addressed in this paper and is currently under investigation. We will report the result in another publication in the near future.

REFERENCES


