RESEARCH ARTICLE

An empirical Bayes model for gene expression and methylation profiles in antiestrogen resistant breast cancer

Jaesik Jeong[1], Lang Li[2], Yunlong Liu[2], Kenneth P Nephew[3], Tim Hui-Ming Huang[4], Changyu Shen[1*]

Abstract

Background: The nuclear transcription factor estrogen receptor alpha (ER-alpha) is the target of several antiestrogen therapeutic agents for breast cancer. However, many ER-alpha positive patients do not respond to these treatments from the beginning, or stop responding after being treated for a period of time. Because of the association of gene transcription alteration and drug resistance and the emerging evidence on the role of DNA methylation on transcription regulation, understanding of these relationships can facilitate development of approaches to re-sensitize breast cancer cells to treatment by restoring DNA methylation patterns.

Methods: We constructed a hierarchical empirical Bayes model to investigate the simultaneous change of gene expression and promoter DNA methylation profiles among wild type (WT) and OHT/ICI resistant MCF7 breast cancer cell lines.

Results: We found that compared with the WT cell lines, almost all of the genes in OHT or ICI resistant cell lines either do not show methylation change or hypomethylated. Moreover, the correlations between gene expression and methylation are quite heterogeneous across genes, suggesting the involvement of other factors in regulating transcription. Analysis of our results in combination with H3K4me2 data on OHT resistant cell lines suggests a clear interplay between DNA methylation and H3K4me2 in the regulation of gene expression. For hypomethylated genes with alteration of gene expression, most (~80%) are up-regulated, consistent with current view on the relationship between promoter methylation and gene expression.

Conclusions: We developed an empirical Bayes model to study the association between DNA methylation in the promoter region and gene expression. Our approach generates both global (across all genes) and local (individual gene) views of the interplay. It provides important insight on future effort to develop therapeutic agent to re-sensitize breast cancer cells to treatment.

Background

The term epigenetics in general refers to heritable pattern of gene expression that is mechanistically regulated through processes other than alteration in the primary DNA sequences [1,2]. Epigenetics has implications in both our understanding of gene regulation in complex organisms such as mammals and clinical investigation on various diseases such as cancer [3,4]. It is now clear that epigenetic events can occur at both the DNA level (i.e. DNA methylation) and chromatic level (i.e. histone

modifications), resulting in an intricate process of interactions that ultimately lead to the alteration of gene expression [5-7].

DNA methylation is a process that adds a methyl group to the cytosine ring via a co-valent bond, using S-adenosyl-methionine as the methyl donor and DNA methyltransferases (DNMTs) as the catalytic enzyme [5]. In mammals, DNA methylation is mostly common on cytosines that precede a guanosine (the CpG dinucleotide). Two features characterize the distribution of the CpG dinucleotides in the genome. First, the overall frequency of the CpG dinucleotides is substantially less than one would expect from probabilistic calculations, which is likely due to a depletion process induced by

* Correspondence: chashen@iupui.edu
[1]Department of Medicine/Division of Biostatistics, Indiana University, Indianapolis, IN, USA
Full list of author information is available at the end of the article

methylation over time [8]. Second, the distribution of CpG dinucleotides in the genome is highly asymmetric with a high concentration of DNA segments 200bp to several kb in length called "CpG islands", residing in the promoter region and first exon for approximately 60% of genes [6]. A striking feature that distinguishes CpG islands from CpG dinucleotides is that under normal conditions, CpG islands generally lack DNA methylation, whereas CpG dinucleotides are typically methylated (i.e. 80%) [2]. While the relationship between CpG island methylation and gene silencing is well established, the mechanisms underlying this phenomena are less clear but thought to include physical blocking of transcription factor binding [9,10] and/or recruitment of transcriptional repressors to the methylated sites [11].

A more complete understanding of the DNA methylation in carcinogenesis is beginning to emerge. A general observation is that the level and pattern of DNA methylation in cancer cells is the opposite of their normal counterparts. The cancer methylome is characterized by global hypomethylation of DNA, which is linked primarily to repeated DNA sequences becoming hypomethylated. Hypomethylation may contribute to carcinogenesis by promoting tumor formation or progression in a number of possible ways, including affecting transposable element activation, DNA/chromosomal rearrangements, tumor suppressor gene or oncogene copy number, and/or altered chromosome conformation. In contrast to normal cells, increased methylation of CpG islands is a common occurrence in cancer, and is associated with epigenetic silencing during all phases of the cancer process, including tumor initiation, progression and drug resistance. Aberrant CpG island methylation is associated with silencing of genes involved in control of the cell cycle, apoptosis and drug sensitivity, as well as tumor suppressor genes.

Although the above phenomena are well documented in all cancers and recognized as playing an important role in almost every aspect of carcinogenesis, the mechanistic nature of the relationship between methylation and regulation of gene expression remains incompletely understood, including the heterogeneity of the relationship among genes, the interaction of methylation at different sites and the involvement of other epigenetic events.

In the clinical setting, a critical issue for cancer treatment is acquired drug resistance, where patients initially respond to chemotherapy but cease to respond after repeated exposure to the same drug. Potentially, epigenetic alterations, such as DNA methylation, are likely to play an important role in acquired drug resistance, as suggested by several studies [12-15], though much work is yet to be done to gain a clear insight into this phenomenon. Based on our experience in studies of

hormone-therapy resistance in breast cancer, antiestrogen resistance is accompanied by dramatic alterations in the expression level of many genes, and alteration of DNA methylation may be one of the causes.

In this article, we focused on understanding the association between CpG island methylation and gene expression in breast cancer. In particular, we attempted to gain a better understanding of differences in DNA methylation and gene expression between hormone-therapy-sensitive and -resistant cell lines. We considered two breast cancer cell lines that are resistant to tamoxifen and fulvestrant, respectively. These are two clinically important therapeutic agents that target estrogen receptor alpha (ER-alpha), a nuclear receptor that primarily mediates genomic regulation of gene transcription and non-genomic activation of various kinase pathways [16]. It is well known that ER-alpha is a key protein implicated in the majority of breast cancers. Although both tamoxifen and fulvestrant are antagonists of ER-alpha, their mechanisms of action differ markedly [17]. Tamoxifen functions as a competitive agent of E2 (the ligand that stimulates ER-alpha), blocking E2 binding to ER-alpha. In spite of this antagonistic action, tamoxifen-bound ER-alpha is capable of regulating gene transcription through genomic/non-genomic actions. On the other hand, fulvestrant directly inhibits the process through which ER-alpha executes genomic regulation function, rapidly inducing cytoplasm aggregation and ER-alpha degradation [17].

Based on their different mechanisms of action, the transition to a resistant state by constant exposure to these agents likely involves both similar and distinct molecular alterations. The aim of this study is to identify at both the individual gene as well as genome level the regulation status in both DNA methylation and gene expression by comparing drug-resistant cell lines to drug-sensitive cell lines. This study provides important insight on the search of potential targets for epigenetic therapy to re-sensitize tumor cells to hormone or chemo-therapy. Toward this goal, we developed an empirical Bayes statistical model to integrate gene expression and DNA methylation data. Advantages of such a model include (i) consideration of probe-probe variation, (ii) easily interpretable confidence of the detections and (iii) straight forward false discovery rate (FDR) control/estimate [18,19].

Methods
Experiment
The Human Genome U133A 2.0 Array was used for gene expression analysis. We restricted our analysis to probes with at least two "present" calls among four replicates. Differential methylation hybridization (DMH) was done using customized 60-mer oligonucleotide
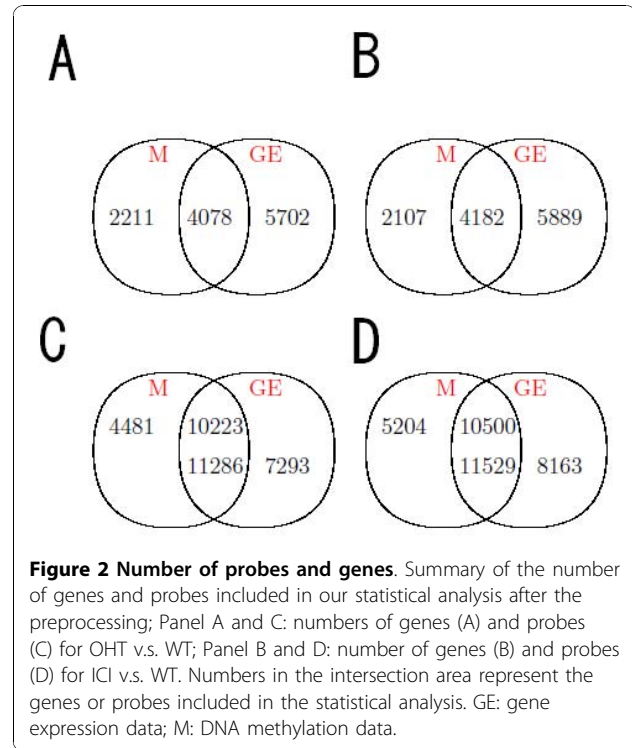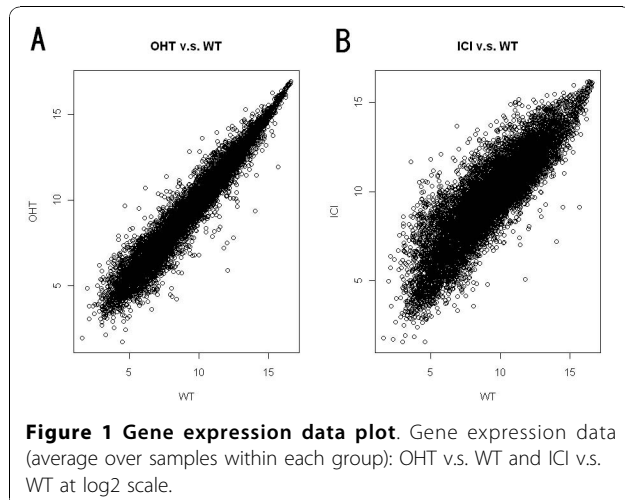
microarrays, which contain ~44,000 CpG-rich fragments from ~12,000 promoters of defined genes [20]. Microarray Analysis Suite (MAS) version 5.0 was used for pre-processing. Experimental details were described in [20]. The data discussed in this paper have been deposited in NCBI's Gene Expression Omnibus http://www.ncbi.nlm.nih.gov/geo/ and are accessible through GEO Series accession number (GSE5840 for gene expression and GSE25519 for methylation).

Data

We compared gene expression and DNA methylation in tamoxifen-(OHT; 4-hydroxytamoxifen) and fulvestrant (ICI 182,780)-resistant MCF7 breast cancer cell lines to the parental, wild type (WT) MCF7 cell line [20]. There were four replicates for gene expression data and no replicate for the methylation data. The structure for the gene expression data is given in Additional file 1. We restricted our analysis to probes with at least two "present" calls among four replicates in gene expression and our focus in methylation is on promoter region probes. Then, for WT v.s. OHT, the gene expression data have a total of 11286 probes that cover 4078 genes, while methylation data have a total of 10223 probes and 4078 genes. For WT v.s. ICI, there are 11529 probes for gene expression and 10500 methylation probes. In addition, there are 4182 genes in common. The gene expression data plots (average over replicates within each group) are given in Figure 1 as an illustration. We also summarized the number of genes and probes included in our statistical analysis after the preprocessing (Figure 2). Probe intensities are log2 transformed before analysis for both gene expression and methylation data.

The Model

We consider an empirical Bayes model to correlate alteration of gene expression and DNA methylation in

Figure 1 Gene expression data plot. Gene expression data (average over samples within each group): OHT v.s. WT and ICI v.s. WT at log2 scale.

Figure 2 Number of probes and genes. Summary of the number of genes and probes included in our statistical analysis after the preprocessing; Panel A and C: numbers of genes (A) and probes (C) for OHT v.s. WT; Panel B and D: number of genes (B) and probes (D) for ICI v.s. WT. Numbers in the intersection area represent the genes or probes included in the statistical analysis. GE: gene expression data; M: DNA methylation data.

WT and drug-resistant cell lines. The marginal model for gene expression (G_{ijkl}) is given by:

$$G_{ijkl} = \mu_{il} + b_{ij} + \epsilon_{ijkl},$$
$$i = 1, \cdots, I, \; j = 1, \cdots, J_i, \; k = 1, \cdots, K, \; l = 1, 2 \quad (1)$$

where i indexes gene, j indexes probe in a gene, k indexes replicates, and l denotes WT ($l = 1$) and resistant group ($l = 2$). Therefore, μ_{il} is the average expression level for gene i in cell line l, b_{ij} is the added effect of probe j for each gene i, and ε_{ijkl} is the error term. We consider the following distributions for each component in the model:

$$b_{ij} \sim N(0, \sigma^2), \; \epsilon_{ijkl} \sim N(0, \delta^2), \; (\mu_{i1}, \mu_{i2})^t \sim N((\mu_1, \mu_2)^t, \Sigma_1). \quad (2)$$

Marginal model for methylation (M_{ihl}) is given in a similar fashion:

$$M_{ihl} = \eta_{il} + a_{ih} + d_{ihl}, \; i = 1, \cdots, I, \; h = 1, \cdots, H_i, \; l = 1, 2 \quad (3)$$

Similarly, η_{il} is gene effect in each group, a_{ij} is the probe effect in each gene, and d_{ihl} is the error term. However, no replicate is here in contrast to GE data. Again, we assume the same distributions for components in Eq. (3):

$$a_{ih} \sim N(0, \omega^2), \; d_{ihl} \sim N(0, \tau^2), \; (\eta_{i1}, \eta_{i2})^t \sim N((\eta_1, \eta_2)^t, \Sigma_2). \quad (4)$$

To integrate both marginal models into an unified model, we assume

$$(\mu_{i1}, \mu_{i2}, \eta_{i1}, \eta_{i2})^t \sim N(\mu, \Sigma) \tag{5}$$

where $\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma^* \\ \Sigma^* & \Sigma_2 \end{pmatrix}$.

Models (1) - (5) can be represented in a simple linear model format [21,22]:

$$D_i = \Delta_i \beta_i + \epsilon_i, \, i = 1, \cdots, I \tag{6}$$

where $\beta_i = (\mu_{i1}, \mu_{i2}, \eta_{i1}, \eta_{i2}, b_1, \cdots, b_{J_i}, a_1, \cdots, a_{H_i})^t$, $D_i = (G_i^t, M_i^t)^t$. G_i and M_i are gene expression and methylation data for gene i, respectively. The distribution assumption for each component in the unified model (Eq. (6)) is:

$$\beta_i \sim N(\mu_i^*, \Sigma_{pi}), \epsilon_i \sim N(0, \Sigma_{e_i}) \tag{7}$$

where $\mu_i^* = (\mu, 0)^t$,

$$\Sigma_{pi} = \begin{pmatrix} \Sigma & 0 & 0 \\ 0 & \sigma^2 I_{J_i} & 0 \\ 0 & 0 & \omega^2 I_{H_i} \end{pmatrix}, \quad \Sigma_{e_i} = \begin{pmatrix} \delta^2 I_{N_i} & 0 \\ 0 & \tau^2 I_{M_i} \end{pmatrix}.$$

Here $N_i = 2K \times J_i$ and $M_i = 2H_i$. As we can see, our model has hierarchical structure for each gene i:

$$D_i \mid \beta_i \sim N(\Delta_i \beta_i, \Sigma_{e_i}), \quad \text{prior of } \beta_i \sim N(\mu_i^*, \Sigma_{pi}) \tag{8}$$

Thus, posterior distribution of $\beta_i|D_i$ follows $N(K, K_*)$ where $K = K^*(\Delta_i^T \Sigma_{e_i}^{-1} D_i + \Sigma_{pi}^{-1} \mu_i^*)$ and $K^* = (\Delta_i^T \Sigma_{e_i}^{-1} \Delta_i + \Sigma_{pi}^{-1})^{-1}$. Inference is based on the posterior distribution above.

More details about statistical modeling are provided in Additional file 2.

Estimation

Expectation-Maximization (EM) algorithm [23,24] is widely used to obtain maximum likelihood estimates when there are unobserved variables. Basically, EM algorithm consists of two iterative steps: Expectation and Maximization. In the E-step, expectation of complete-data log likelihood conditional on data and current value of parameters is calculated. In the M-step, parameters are updated by the value that maximizes the expectation from E-step. Here we briefly describe EM

algorithm applied to our case. More details about E-step and M-step are given in Additional file 2.

E-step
For iteration k, we obtain $E(\beta_i|D_i, \theta^{(k-1)})$ and $Cov(\beta_i|D_i, \theta^{(k-1)})$ that allows us to compute

$$Q(\theta; \theta^{(k-1)}) \equiv E[\log L_c(\theta) \mid D_i, \theta^{(k-1)}].$$

Here $\log L_c(\theta) = \sum_{i=1}^{I} \{ \log[D_i, \beta_i \mid \theta] \}$ is complete-data log likelihood function where θ is the parameter vector.

M-step
In this step, we update θ by values that maximize the target function, $Q(\theta; \theta^{(k-1)})$ given in the E-step.

Inference on Relationship between Gene Expression and Methylation
Inference on relationship between gene expression and methylation can be made by using posterior distribution of $(\mu_{i1}, \mu_{i2}, \eta_{i1}, \eta_{i2})^t$ conditional on data and parameter estimates, where μ_{i1} and η_{i1} are mean parameters of gene expression and methylation in WT and μ_{i2} and η_{i2} are mean parameters of gene expression and methylation in the resistant cell line. Since we are interested in the correlation of the differentiation of gene expression and DNA methylation, we focus on the posterior distribution of

$$(\mu_{GE_i}, \eta_{M_i})^t = (\mu_{i2} - \mu_{i1}, \eta_{i2} - \eta_{i1})^t, \tag{9}$$

which can be easily calculated through a linear transformation of $(\mu_{i1}, \mu_{i2}, \eta_{i1}, \eta_{i2})^t$.

To characterize the correlation of gene expression and DNA methylation for each gene, we first divide the two-dimensional sample space of $\left(\mu_{GE_i}, \eta_{M_i} \right)$ into nine categories by applying two thresholds to each of the μ_{GE_i} and η_{M_i} dimensions. The nine categories represent the combination of three levels of alteration in gene expression and DNA methylation: up-regulation, no change, down-regulation. For instance, the north-east region will be "up-regulation in both expression and DNA methylation". The thresholds are chosen to be $\pm C \cdot \sigma$, where σ is the standard deviation of the posterior mean of μ_{GE_i} or η_{M_i} across all genes. In our application, we chose $C = 1.5$. We then calculate for each gene the posterior probability of each of the nine regions, which characterizes the correlation of gene expression and DNA methylation for each gene in a probabilistic manner. Based on these probabilities, we will assign

each gene to one of the nine categories. See result section for details.

Results and Discussion

Association between Gene Expression and Methylation Status

The main output from our model is a joint posterior distribution of the difference of expression and methylation levels between drug-resistant cell lines and WT for each gene. Such a distribution provides us with a probabilistic measure on the strength of the association for each gene. On the other hand, the center (or the mean) of the prior distribution of the difference of expression and methylation provides us with a global view of the association across genes. In the following discussion, up and down regulation is always in reference to the WT.

To facilitate the understanding of the association between gene expression and DNA methylation at the individual gene level, we used four thresholds (two for gene expression and two for methylation) to categorize each gene into nine categories (Figure 3). In the figure, NG, UP, and DN stand for no change, up regulation, and down regulation in gene expression. Similarly, NM, HO, and HR stand for no change in methylation, hypomethylation and hypermethylation. For example, the upper-left cell is the category for down regulation of gene expression and hypermethylation. The thresholds are determined based on the posterior distribution (see Method section). Given the thresholds, one can calculate posterior probability that each gene falls into one category. We then assign each gene to one of the nine categories based on these probabilities. We consider two ways of categorization. In the first method, a gene is assigned to the category with the maximum probability. The limitation of this method is that when the probability mass is evenly distributed among the several categories, the maximum is not a substantially

dominating number, which still entails a lot of uncertainty. Therefore, in the second method, we apply a threshold to the maximum probability for gene assignment. Genes with maximum probability below the threshold are not assigned to any category. In the context of FDR, applying a cutoff value to the posterior probability means that the false discovery rate is controlled at α = $D(\kappa) = K$, where $K = \sum_i I_{[P_i^s \geq \kappa]}$ i.e., the number of genes whose posterior pass the threshold k and $D(\kappa)$ is the summation of one minus the posterior probabilities of these genes:

$$D(\kappa) = \sum_i \{1 - P_i^s\} I_{[P_i^s \geq \kappa]} \qquad (10)$$

where I is the indicator function and P_i^s is the posterior probability of gene i belonging subcategory s [19]. We illustrate these definitions through two examples. Figure 4 is the contour plots of the posterior distributions of two genes (*CDH3, KIAA0478*) from the OHT vs. WT comparison superseded on nine categories. It is well known that gene *CDH3*, which acts as a tumor suppressor gene, is hypomethylated in breast cancer [6]. This is visually confirmed by Figure 4 by observing that most of the probability mass is concentrated on the region of no expression change but reduced methylation. It can be shown that the probability of this category is 0.8. On the other hand, *KIAA0478* concentrates on the region of no expression and methylation change with a probability of 0.63. To characterize all genes assigned into one of the nine categories, we apply three assignment rules: maximum probability, maximum probability over 0.6 and maximum probability over 0.7. Table 1 and 2 summarize the results for OHT v.s. WT and ICI v.s. WT, respectively. The gene lists obtained from using three different cutoff values are given in Additional file 3.
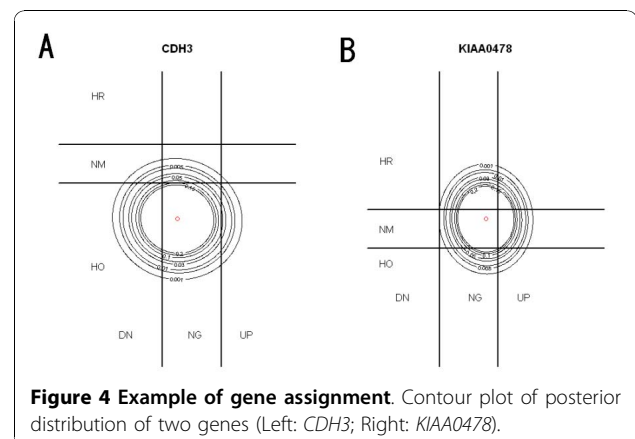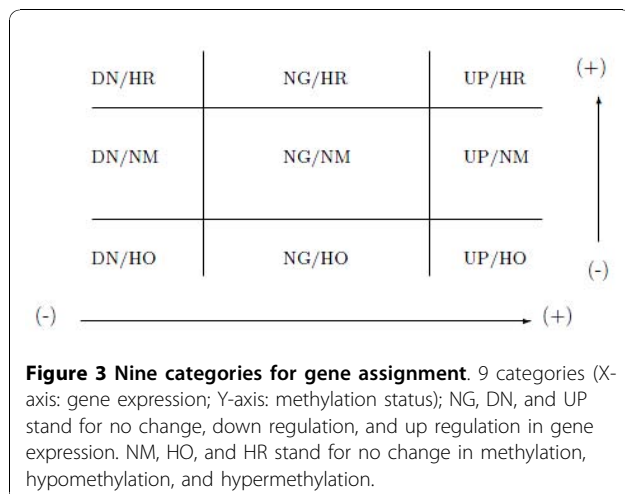

Figure 3 Nine categories for gene assignment. 9 categories (X-axis: gene expression; Y-axis: methylation status); NG, DN, and UP stand for no change, down regulation, and up regulation in gene expression. NM, HO, and HR stand for no change in methylation, hypomethylation, and hypermethylation.

Figure 4 Example of gene assignment. Contour plot of posterior distribution of two genes (Left: *CDH3*; Right: *KIAA0478*).

Table 1 Gene assignment to nine category

Methylation	Gene expression		
	DOWN	No change	UP
HYPER	0/0/0	1/0/0	0/0/0
No change	101/45/17	2284/1382/786	176/63/35
HYPO	48/11/6	1331/744/496	137/46/24

Three values in each cell present the number of genes belonging to each category using three different assignment rules: maximum probability, maximum probability over 0.6, and maximum probability over 0.7. (OHT v.s. WT)

Not surprisingly, the category NG/NM contains most of the genes. The other feature is that very few genes are hypermethylated, similar to what was observed in the original report of the experiments [20]. While the reason for this is not clear, one possibility is that hypomethylation and up regulation of the corresponding gene(s) may provide the drug-resistant cells with a survival and growth advantage. Among those hypomethylated genes with expression alteration, the majority are up-regulated, consistent with what is known regarding promoter methylation and gene expression. The hypomethylated, down-regulated genes suggest other mechanisms are involved in regulating expression in addition to DNA methylation, such as the repressive histone methylation [25-31].

With hypomethylated genes only, the Venn diagram for the overlap of genes with expression alteration between the two drug-resistant cell lines is shown in Figure 5 (based on maximum probability criterion). It can be seen that the overall overlap is small, suggesting distinct sets of hypomethylated genes between OHT and ICI resistant breast cancer cells. The number of overlapping genes that are up-regulated in both cell lines (i.e., 20) is disproportionly higher than the overlapping genes for down regulation patterns (i.e., 2), suggesting that the association of DNA hypomethylation and up regulation in both cell lines may share a common gene set, at least more common than the other association patterns. More details about overlaps for different cutoffs are given in Additional file 4. At the global level, the Pearson correlation coefficients

Table 2 Gene assignment to nine category

Methylation	Gene expression		
	DOWN	No change	UP
HYPER	0/0/0	1/1/0	0/0/0
No change	86/20/9	2309/1518/922	237/114/64
HYPO	50/12/7	1350/811/526	149/61/40

Three values in each cell present the number of genes belonging to each category using three different assignment rules: maximum probability, maximum probability over 0.6, and maximum probability over 0.7.(ICI v.s. WT)
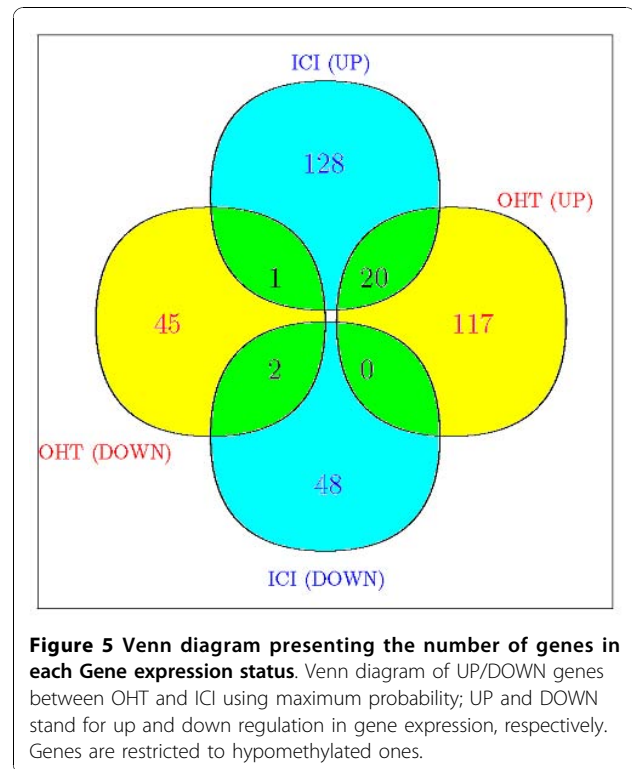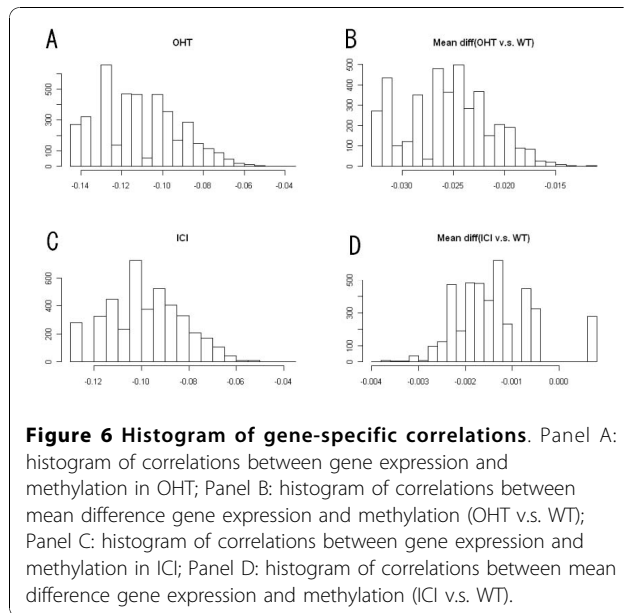

Figure 5 Venn diagram presenting the number of genes in each Gene expression status. Venn diagram of UP/DOWN genes between OHT and ICI using maximum probability; UP and DOWN stand for up and down regulation in gene expression, respectively. Genes are restricted to hypomethylated ones.

(across genes) between DNA methylation and gene expression are -0.17 for WT, -0.19 for OHT and -0.16 for ICI. Hence, there seems to be some relationship, in general, between DNA methylation and gene expression, and although the magnitude is not strong, this observation is consistent with the general understanding that increased promoter DNA methylation correlates with a lower level of gene activity. From another angle, if we examine the correlation between methylation and gene expression alterations (i.e., $corr(\mu_{i2} - \mu_{i1}, \eta_{i2} - \eta_{i1})$ in Eq. (9)) when comparing OHT/ICI to WT, the correlation coefficients across genes are -0.04 for OHT and 0.01 for ICI, implying methylation and gene regulation are not strongly associated in this model system, and this relationship may be highly gene-specific. The local level (gene-specific) view is given in the following histogram (Figure 6). Panels A and C present the histograms of correlations between gene expression and methylation in OHT and ICI. Panels B and D present distribution of correlations between gene expression mean difference and methylation mean difference obtained from OHT v.s. WT and ICI v.s. WT, respectively. Panels A and C show the distribution of correlations between gene expression and methylation ranging from -0.15 to 0, implying heterogeneity in gene-specific correlations.

Figure 6 Histogram of gene-specific correlations. Panel A: histogram of correlations between gene expression and methylation in OHT; Panel B: histogram of correlations between mean difference gene expression and methylation (OHT v.s. WT); Panel C: histogram of correlations between gene expression and methylation in ICI; Panel D: histogram of correlations between mean difference gene expression and methylation (ICI v.s. WT).

Histone Methylation

It is well known that the interplay of histone modification and DNA methylation affects the transcriptional regulation [25-29]. To examine the involvement of histone methylation in the association of alterations of DNA methylation and gene expression, we analyzed some in-house histone methylation data. The data were generated by chromatin-immunoprecipitation and high-throughput sequencing (ChIP-seq). The experimental protocol followed the same line of procedure reported previously [32,33]. We will focus our discussion on the dimethylation on lysine residue 4 on H3 (H3K4me2) on OHT MCF7 cell lines. Our data include 26443 genes with two replicates. We first compared H3K4me2 levels in OHT between genes with DNA hypomethylation and those without alteration in DNA methylation, i.e. the third row versus the second row in Table 1 (maximum probability rule is used to assign genes to a category). The fold change is 1.10 (95% CI: 1.00-1.20). Therefore, DNA hypomethylation is associated with enhanced H3K4me2 in this setting. This observation is intuitively appealing as both DNA hypomethylation and H3K4me2 were found to be related to gene activation [27,28,30] and [31].

We next compared the H3K4me2 levels for (i) genes in the UP/HO category versus those in the DN/HO category; (ii) UP/NM category versus DN/NM category (Table 1). The fold changes are 1.56 (95% CI 0.89-2.72) and 1.24 (95% CI: 0.82-1.88), respectively. Therefore, consistent with previous findings [27,28,30,31], our results show that H3K4me2 is likely to be associated with transcriptional activation. Moreover, there seems to be a higher level of H3K4me2 change in genes with

DNA hypomethylation than those without DNA methylation change, suggesting an interaction between H3K4me2 and DNA methylation in regulating gene expression.

Gene Ontology Analysis

Although it is likely that other genetic or epigenetic activities in addition to DNA methylation are involved in the regulation of genes, our finding is consistent with the general observation that hypomethylation leads to up regulation. Furthermore, we conducted a gene ontology analysis on these genes and the results are shown in Table 3. Several functional categories are over-represented in both cell lines, such as cell death, connective tissue development and function and cellular development. On the other hand, distinct functional categories are associated with each cell line. Taken together, these observations suggest while common mechanisms for switching of genes through DNA demethylation at promoter regions are shared by OHT- and ICI-resistant cells, unique processes are also associated with development of acquired resistance to the distinctly different antiestrogens. These distinct functional and molecular changes associated with the acquisition of resistance to the two different classes of antiestrogens include signaling and growth regulatory processes.

Conclusions

In this article, we developed an empirical Bayes model to study the association between altered DNA methylation in the promoter region and gene expression by comparing WT with OHT and ICI resistant MCF7 breast cancer cell lines. Our statistical model incorporates various sources of variations that generate probabilistic characterization of such an association. The model structure also allows a natural incorporation of other epigenetic processes to investigate their regulatory roles in acquired antiestrogen resistance.

Our models are characterized by a hierarchical structure that has been shown to be more efficient and stable than analysis of individual gene separately [34]. It also allows one to estimate the correlation between gene expression and DNA methylation at the level of individual genes. However, our models induce a marginally positive correlation between probes of the same gene, which might not hold for all genes and all microarray platforms. A small simulation study (data not shown) suggests that the inference on gene level quantity μ_{il} and η_{il} is relatively robust when probes are actually negatively correlated. Finally, given the complexity of our model it is not possible to use standard diagnostic tools to check model assumptions. Nevertheless, it is still possible to examine posterior quantities of latent variable that is conditional on the data and parameter at

Table 3 Gene Ontology Analysis

OHT	ICI
Cell Death	Cell Death
Connective Tissue Development & Function	Connective Tissue Development & Function
Cellular Development	Cellular Development
Cellular Compromise	Immune Cell Trafficking
Infectious Disease	Nervous System Development & Function
Post-Translational Modification	Organismal Development
Carbohydrate Metabolism	Cellular Movement
Cell-To-Cell Signaling and Interaction	Hematological System Development & Function
Cellular Assembly and Organization	Cellular Function and Maintenance
Cell Morphology	Hematological Disease

This table includes top 10 over-represented functional categories for genes with hypomethylation and up regulation of gene expression with maximum probability

their estimated values. See Additional file 5, 6, 7, 8, 9, 10 for details. Consistent with original publication of the data [20], our results showed that almost all DNA methylation alterations were in the direction of reduction when resistant cell lines were compared with wild type, suggesting a homogenous pattern of DNA methylation during the acquisition of drug resistance. Furthermore, the OHT and ICI cell lines shared similar yet holded unique association patterns. It is noted that a proportion of genes are hypomethylated with down regulation of gene expression, suggesting the involvement of other genetic and epigenetic factors in the regulation process.

Although there exists a weak correlation between DNA methylation at promoter regions and gene expression for the three cell lines studied, the correlation of methylation and gene expression alterations, when comparing OHT/ICI to WT at the global level, is essentially 0. This implies that the relation between alterations in DNA methylation at promoter region and gene expression is gene-specific and, likely due to the involvement of other factors.

Additional material

Additional file 1: Table S1 – Gene expression data structure. This table shows the gene expression data structure in both group: wild type and antiestrogen resistant group.

Additional file 2: Table S2 – Gene lists which are hypomethylated and up-regulated. This table presents gene lists obtained by using three different cutoff values with the OHT data set.

Additional file 3: Figure S1 – Venn diagram presenting gene overlap. Each figure is obtained by using three different cutoff values. Based on each cutoff values, the status of each gene is determined in both data sets. The number of gene overlaps which were obtained by using gene expression data in both data set is calculated. The numbers in each Venn diagram presents the number of common genes in both data sets.

Additional file 4: Figure S2 – Histogram of residuals. This histogram is based on standardized residuals obtained by using estimates in our model; Top: residuals histogram of gene expression in ICI; Bottom: residuals histogram of methylation in ICI.

Additional file 5: Figure S3 – Q-Q plot of residuals. Each Q-Q plot is based on standardized residuals obtained by using parameter estimates in our model in ICI; Left: this plot is obtained by using gene expression residuals; Right: this plot is obtained by using methylation residuals.

Additional file 6: Figure S4 – Histogram of gene effect. Each histogram is based on estimated gene effect in our model in ICI; Top: these plots are obtained by using estimated gene effect of each group in gene expression (Left:WT and Right:ICI); Bottom: these plots are obtained by using estimated gene effect of each group in methylation (Left:WT and Right:ICI).

Additional file 7: Figure S5 – Q-Q plot of gene effect. Each Q-Q plot is based on estimated gene effect in our model in ICI; Top: these plots are obtained by using estimated gene effect of each group in gene expression (Left:WT and Right:ICI); Bottom: these plots are obtained by using estimated gene effect of each group in methylation (Left:WT and Right:ICI).

Additional file 8: Figure S6 – Histogram of added probe effect. Each histogram is based on estimated probe effect in our model in ICI; Left: these plots are obtained by using estimated added probe effect in gene expression; Right: these plots are obtained by using estimated added probe effect in methylation.

Additional file 9: Figure S7 – Q-Q plot of added probe effect. Each Q-Q plot is based on estimated probe effect in our model in ICI; Left: these plots are obtained by using estimated added probe effect in gene expression; Right: these plots are obtained by using estimated added probe effect in methylation.

Additional file 10: Supplementary materials – Details about modeling and estimation. This file includes details about marginal modeling and parameter estimation. Also, the exact form of parameter estimators are given.

Author details
[1]Department of Medicine/Division of Biostatistics, Indiana University, Indianapolis, IN, USA. [2]Department of Medical and Molecular Genetics, Indiana University, Indianapolis, IN, USA. [3]Department of Medical Science/ Simon Cancer Center, Indiana University, Bloomington, IN, USA. [4]Department of Molecular Virology, Immunology and Medical Genetics/Human Cancer Genetics, The Ohio State University, Columbus, OH, USA.

Authors' contributions
JJ and CS conceived and designed the statistical model, and drafted the manuscript. JJ implemented the algorithm and carried out simulation studies and real data analysis. LL and YL participated in algorithm and

manuscript revision. KPN and THH generated the experimental data, provided biological input and revised the manuscript. All authors read and approved the final manuscript.

References
1. Baylin SB, Herman JG: DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet* 2000, **16**:168-174.
2. Bird A: DNA methylation patterns and epigenetic memory. *Gene Dev* 2002, **16**:6-21.
3. Herman JG: Hypermethylation of tumor suppressor genes in cancer. *Semin Cancer Biol* 1999, **9**:359-367.
4. Jones PA, Laird PW: Cancer-epigenetics comes of age. *Nat Genet* 1999, **21**:163-167.
5. Herman JG, Baylin SB: Gene silencing in cancer in association with promoter hypermethylation. *New Engl J Med* 2003, **349**:2042-2054.
6. Hinshelwood RA, Clark SJ: Breast cancer epigenetics: normal human mammary epithelial cells as a model system. *J Mol Med* 2008, **86**:1315-1328.
7. Yuan G, Ma P, Zhong W, Liu JS: Statistical assessment of the global regulatory role of histone acetylation in Saccharomyces cerevisiae. *Gen Biol* 2006, **7**:R70.
8. Rideout WMIII, Coetzee GA, Olumi AF, Jones PA: 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p 53 genes. *Science* 1990, **249**:1288-1290.
9. Iguchi-Ariga SM, Schaffner W: CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. *Gene Dev* 1989, **3**:612-619.
10. Molloy PL, Watt F: DNA methylation and specific protein-DNA interactions. *Philos Trans R Soc Lond B* 1990, **326**:267-275.
11. Clouaire T, Stancheva I: Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatic? *Cell Mol Life Sci* 2008, **65**:1509-1522.
12. Li M, Balch C, Montgomery JS, Jeong M, Chung JH, Yan P, Huang TH, Kim S, Nephew KP: Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. *BMC Med Genom* 2009, **2**:2:34.
13. Ottaviano YL, Issa JP, Parl FF, Smith HS, Baylin SB, Davidson NE: Methylation of the estrogen receptor gene CpG island marks loss of estrogen receptor expression in human breast cancer cells. *Cancer Res* 1994, **54**:2552-2555.
14. Das PM, Singal R: DNA methylation and cancer. *J Clin Oncol* 2004, **22**:4632-4642.
15. Dwivedi RS, Qiu YY, Devine J, Mirkin BL: Role of DNA methylation in acquired drug resistance in neuroblastoma tumors. *Proc Indian Nat Sci Acad* 2003, **69**:111-120.
16. Anderson E: The role of estrogen and progesterone receptors in human mammary development and tumorigenesis. *Breast Cancer Res* 2002, **4**:197-201.
17. Howell A, Abram P: Clinical development of fulvestrant ('Faslodex'). *Cancer Treat Rev* 2005, **31**:S3-9.
18. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Met* 1995, **57**:289-300.
19. Newton MA, Noueiry A, Sarkar D, Ahlquist P: Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 2004, **5**:155-176.
20. Fan M, Yan PS, Hartman FC, Chen L, Paik H, Oyer SL, Salisbury JD, Cheng AS, Li L, Abbosh PH, Huang TH, Nephew KP: Diverse gene expression and DNA methylation profiles correlate with differential adaptation of breast cancer cells to the antiestrogens Tamoxifen and Fulvestrant. *Cancer Res* 2006, **66**:11954-11966.
21. Smith AFM: A general Bayesian linear model. *J R Stat Soc B* 1973, **35**:67-75.
22. Wang CS, Rutledge JJ, Gianola D: Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genet Sel Evol* 1994, **26**:91-115.
23. McLachlan GJ, Krishnan T: *The EM Algorithm and Extensions* Wiley; 2007.
24. Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B Met* 1977, **39**:1-38.
25. Cedar H, Bergman Y: Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet* 2009, **10**:295-304.
26. Jones PA, Baylin SB: The epigenomics of cancer. *Cell* 2007, **128**:683-692.
27. Zhang Y, Reinberg D: Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Gene Dev* 2001, **15**:2343-2360.
28. Strahl BD, Ohba R, Cook RG, Allis CD: Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in tetrahymena. *Proc Natl Acad Sci* 1999, **96**:14967-14972.
29. Rea S, Elsenhaber F, O'Carroll D, Strahl BD, Sun Z, Schmid M, Opravil S, Mechtler K, Ponting CP, Allis CD, Jenuwein T: Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* 2000, **406**:593-599.
30. Li B, Carey M, Workman J: The role of chromatin during transcription. *Cell* 2007, **128**:707-719.
31. Lee MG, Villa R, Trojer P, Norman J, Yan KP, Reinberg D, Di CL, Shiekhattar R: Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. *Science* 2007, **318**:447-450.
32. Lee TI, Johnstone SE, A YR: Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat Protoc* 2006, **1**:729-748.
33. Feng W, Liu Y, Wu J, Nephew KP, Huang TH, Li L: A Poisson mixture model to identify changes in RNA polymerase II binding quantity using high-throughput sequencing technology. *BMC Genom* 2008, **9**:S2-S23.
34. Ji H, Liu S: Analyzing omics data using hierarchical models. *Nat Biotechnol* 2010, **28**:337-340.