International Journal of Biomedical Data Mining
ISSN: 2090-4924

Xu et al., Biomedical Data Mining 2015, 4:1
DOI: 10.4172/2090-4924.1000115

Research Article Open Access

Identifying DNA Methylation Variation Patterns to Obtain Potential Breast Cancer Biomarker Genes

Xu L[1,2], Mitra-Behura S[3,4], Alston B[5,6], Zong Z[7] and Sun S[8]*

[1]*Liberal Arts and Science Academy in Austin, Texas, USA*
[2]*Harvard University, USA*
[3]*DeBakey High School for Health Professions in Houston, Texas, USA*
[4]*Massachusetts Institute of Technology, USA*
[5]*St. John's High School in Houston, Texas, USA*
[6]*Southern Methodist University, USA*
[7]*Department of Computer Science, Texas State University, Texas, USA*
[8]*Department of Mathematics, Texas State University, Texas, USA*

Abstract

Patterns of DNA methylation in human cells are crucial in regulating tumor growth and can be indicative of breast cancer susceptibility. In our research, we have pinpointed genes with significant methylation variation in the breast cancer epigenome to be used as potential novel biomarkers for breast cancer susceptibility. Using the statistical software package R, we compare DNA methylation sequencing data from seven normal individuals with eight breast cancer cell lines. This is done by selecting CG sites, or cytosine-guanine pairings, at which normal cell and cancer cell variation patterns fall in different ranges, and by performing upper one-tailed chi-square tests. These selected CG sites are mapped to their corresponding genes. Using the ConsensusPath Database software, we generate genetic pathways with our data to study biological relations between our selected genes and tumorigenic cellular mechanisms. Using breast cancer-related genes from the PubMeth and GeneCards databases, we have discovered 26 potential biomarker genes, which are biologically linked to genes known to be associated with breast cancer. Our results have numerous implications for early screening and detection measures for breast cancer susceptibility. Furthermore, novel treatments may be developed as more research is conducted exploring the biomarker genes' association with stimulating tumorigenesis.

Keywords: DNA methylation; Breast cancer; Variation; Biomarker

Introduction

Breast cancer is especially prevalent in developed countries, due to correlations between tumorigenesis and factors such as old age, obesity, and lack of physical activity. In developing countries, many cancers are strongly linked to viral infections such as hepatitis B, hepatitis C, and human papillomavirus and are harder to treat because of the dearth of healthcare. Every year, more than 14 million new cases of cancer occur globally, causing more than 8 million deaths in the world and constituting about 14.6% of all human deaths [1,2]. The financial costs of cancer are over 1 trillion dollars annually, showing the necessity for improvements in cancer screening for early detection and effective treatment [2].

One of the most significant challenges of treating cancer is detection at an early stage [3]. While there exist screening tests to identify specific cancers and medical imaging to identify cancerous tumors throughout the body, such testing is expensive and is thus often reserved for those with higher tumorigenesis susceptibility, due to family history and environmental factors. Leaving much of the population at risk of missing early tumor detection, current medical screening is carefully distributed based on cost-benefit optimization. Most cancers have much higher cure rates when detected early, especially breast cancer; early prevention is much more effective than chemotherapy and surgery in the late stages of cancer [3-5]. Therefore, there is an urgent need to find early detection methods for breast cancer patients. Recently, research on cancer cell line DNA data using novel DNA sequencing techniques has identified potential biomarker genes that can predict a patient's likelihood of developing breast cancer. These genes can be detected within patient DNA before any tumorigenesis, allowing for preventative measures and screening to be conducted earlier [6]. With the advent of next generation sequencing techniques [7,8], many new methodologies to analyze cancer epigenomics have arisen, providing faster and cheaper sequencing of epigenetic data. We use publicly available epigenomic data from these sequencing techniques to examine the relationship between DNA methylation variation patterns and breast cancer tumorigenesis, aiming to pinpoint potential biomarkers for breast cancer genesis. DNA methylation is the biochemical process in which a methyl group is added to the 5' positions of the cytosine bases of eukaryotic DNA. Methylation suppresses the expression of genes by interfering with DNA transcription in a human genome [9]. Methylation on gene promoters may lead to suppression of gene expression. Hypermethylation is characterized by an increase in methylation of CG sites (i.e., cytosine and guanine pairings), while hypomethylation is a decrease in methylation. These terms are defined relative to reference DNA methylation levels.

The role of DNA methylation in gene expression makes it crucial in the regulation of many cellular processes, such as embryonic development, genomic imprinting, X-chromosome inactivation, and the preservation of chromosome stability after replication [10-12]. Methylation on or near gene promoters will vary depending on cell types, indicating the role of methylation in cell differentiation. Many

***Corresponding author:** Sun S, Department of Mathematics, Texas State University, Texas, USA, Tel: 512-245-3422; E-mail: ssun5211@yahoo.com

Received June 22, 2015; **Accepted** September 07, 2015; **Published** September 20, 2015

Citation: Xu L, Mitra-Behura S, Alston B, Zong Z, Sun S (2015) Identifying DNA Methylation Variation Patterns to Obtain Potential Breast Cancer Biomarker Genes. Biomedical Data Mining 4: 115. doi:10.4172/2090-4924.1000115

Copyright: © 2015 Xu L, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

diseases, such as cancers, lupus, muscular dystrophy, and birth defects have been linked to differential methylation and defective imprinting. Abnormal methylation often leads to either under-expression or over-expression of the affected gene, as well as any genes indirectly affected by the methylated gene, which can lead to several human diseases (e.g., cancers) [13-17]. In particular, genes that regulate the cell cycle by inducing apoptosis can be affected by differential methylation, leading to greater susceptibility of tumorigenesis [6]. Differential methylation of tumor suppressor genes has resulted in the development of cancer, as tumor suppressor genes are often silenced due to hypermethylation. However, genomes of cancer cells have been shown to be overall hypomethylated, with exceptions of hypermethylation at genes involved in cell cycle regulation, tumor cell regulation, and DNA repair. Studying DNA methylation in the context of cancer is important because abnormal methylation events, including hypomethylation and hypermethylation, can serve as biomarkers indicating susceptibility to the development of cancer in a patient [6,18,19].

In order to accurately identify DNA methylation biomarkers, it is crucial to obtain methylation signals at the single CG site level in an entire human genome. During the last several years, the next generation sequencing technologies make this important task of methylation sequencing possible. The methylation data utilized in our project have been sequenced using the reduced representative bisulfite sequencing (RRBS) protocol [20], which is a cost and time efficient technique for analyzing genome-wide methylation profiles on a single nucleotide level. After conducting quality assessment and preprocessing steps using available software packages MethyQA [21] and BRAT [22], we obtain methylation levels and sequencing coverage at each CG site. In particular, we use methylation sequencing data for eight breast cancer cell lines (BT20, BT474, MCF10A, MCF7, MDAMB231, MDAMB468, T47D, and ZR751) [23] and seven normal samples from the ENCODE project (encodeproject.org) [24]. We then use statistical methods to study the methylation variation patterns in these normal and cancer epigenomic data and then identify potential DNA methylation biomarkers. In this paper, we focus on studying methylation variation rather than the average methylation level because methylation levels across different cancer samples are often heterogeneous (or have a large amount of variation). The average-based statistical analysis may not identify the genes with strong heterogeneous methylation patterns.

In identifying potential DNA methylation biomarkers for breast cancer and their relationships to known breast cancer oncogenes, our research allows for more thorough determination of breast cancer susceptibility based on the methylation patterns and functions of certain genes. Novel cancer treatments can be pioneered based on these specific biomarkers and how they interact with known oncogenes, and these new potential biomarkers can be added to the breast cancer gene database, further contributing to fully understanding the complex gene interactions that lead to breast cancer tumorigenesis and to one day eradicating it.

Methods

The preprocessing step of both the breast cancer cell line and normal sample sequencing data begin with the quality assessment step using the software package MethyQA [21]. In particular, MethyQA generates basic and informative diagnostic plots for the FASTQ format raw sequencing reads. The quality assessment plots include figures for sequencing quality scores, per sequencing GC content, and so on. We then use the method mentioned in the MethyQA to check the bisulfite conversion rate by examining the methylation levels of the cytosine sites that are not paired with a guanine site (i.e., non-CG cytosine sites). The checking results show that the bisulfite conversion rate is very high, so the bisulfite-treatment has been done properly. After the quality assessment step, we use the *trim* function provided in the software package BRAT [22] to remove the bases with a quality score less than 20. Trimmed reads are aligned to the human reference genome version 19 (hg19). Aligned reads are then processed using the *acgt-count* function of BRAT to generate methylation levels. The *acgt-count* function can generate the counts of "A", "C", "G", and "T" at each base, and then produce a methylation level that is the ratio of the count of "C" (or the number of methylated reads) to the count of "C" and "T" (or the total number of reads covering a cytosine site). The methylation level at each CG site ranges from 0 to 1. We then use *R*, a statistical computing and graphics programming language [25], to analyze the methylation levels of breast cancer cell lines and normal samples.

Data I: Breast cancer cell lines

The methylation sequencing data for both the breast cancer cell lines and normal individuals are sorted and selected by choosing CG sites that have at most two missing values across the cell lines. To pinpoint the methylation variation patterns of breast cancer cell lines, we compute the standard deviation and mean of the methylation levels at the CG site level across all cell lines. It is evident that most CG sites are either fully methylated or unmethylated, and some CG sites contain much more methylation variation than others (Figure 1).

Comparing the standard deviation and mean of the methylation level for all CG sites, we find that partially methylated sites tend to have a relatively higher standard deviation. This finding indicates that differential or heterogeneous methylation patterns exist at those partially methylated CG sites. We study partially methylated CG sites represented above as points with $0.4 \leq \text{mean} \leq 0.6$ and standard deviation ≥ 0.3 (Figures 2 and 3).

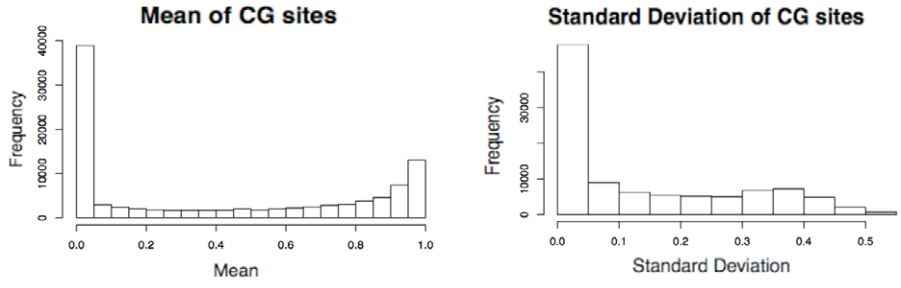

Figure 1: Histograms for mean and standard deviation of the methylation levels.

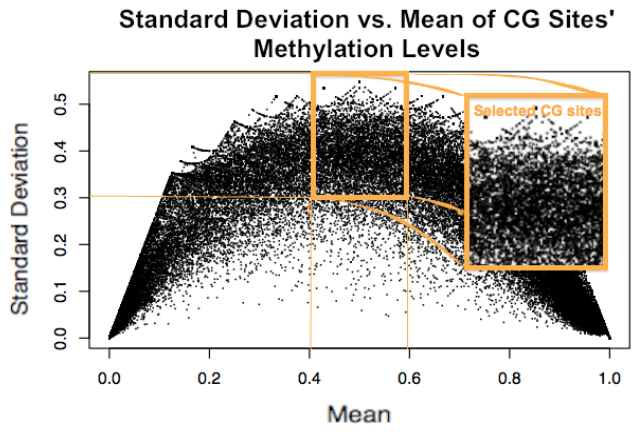

Figure 2: Standard deviation vs. mean methylation level of selected CG sites (orange).

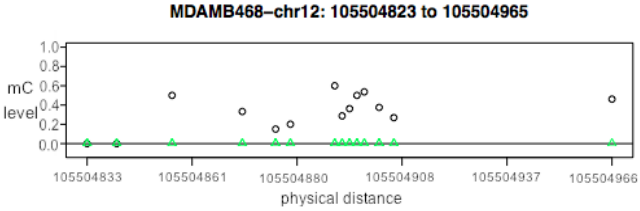

Figure 3: Methylation level plot for MDAMB468 chromosome 12. Green triangles indicate CG sites; black circles represent the methylation levels at each of these CG sites.

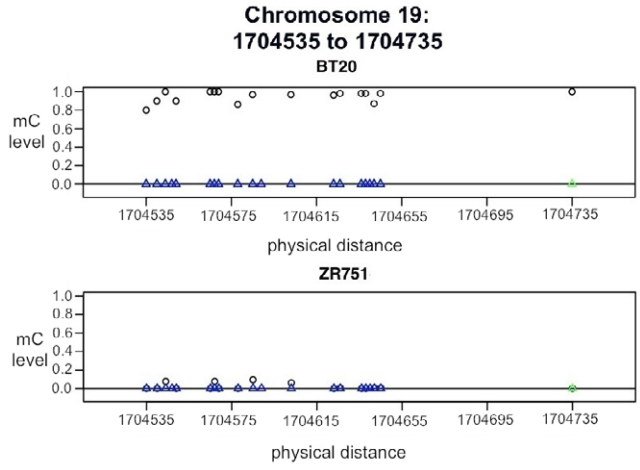

Figure 4: Methylation level of selected CG sites on chromosome 19 for BT20 and ZR751. Blue triangles indicate grouped CG sites; green triangles are CG sites; black circles represent methylation levels.

With the selected CG sites, we generate groups by placing any CG sites that are within 20 base pairs of another one in the same group. Then, we focus on groups with more than 10 CG sites as they are indicative of genes of interest due to their high methylation variation within a small portion of a chromosome. We plot chromosome location and methylation level for each selected group (Figure 3). This allows us to identify specific CpG islands with high methylation variation across cell lines; these CpG islands may influence breast cancer tumorigenesis.

By plotting the CG sites and methylation levels for all groups across the eight breast cancer cell lines, we can visualize groups with higher variance based on differences in methylation patterns between cell lines. In Figure 4, the BT20 cell line is almost fully methylated on all of the CG sites within the interval (1704535 to 1704735 base) on chromosome 19, while the ZR751 cell line is unmethylated on nearly all CG sites. This variation in methylation across the breast cancer cell lines is indicative of many differential methylation sites within a concentrated region of a chromosome, and is thus a possible gene or promoter of interest.

Data II: Normal data from the ENCODE

We obtain RRBS data of seven non-tumorigenic cells from the Encyclopedia of DNA Elements Projects (ENCODE): 0203015, h12529n, h12803n, kidney0111002, n00204, n30, and n41 [24]. These datasets provide information about methylation signals of non-tumorigenic human cells and serve as a standard of comparison for the eight breast cancer cell lines in our analysis. We compute the standard deviation and mean of the methylation levels at particular CG sites across all seven samples (Figure 5).

Studying the standard deviation and mean of the methylation levels for all normal samples, we find that most CG sites have either no methylation or full methylation, and their standard deviations are much smaller (e.g., less than 0.1) than the standard deviations of cancer cell lines (Figure 5 for normal data and Figure 2 for cancerous data). Therefore, we choose to study the CG sites whose standard deviations are ≤ 0.1 (Figure 5).

Statistical analysis: Chi-square test

We use chi-square tests [26] to compare variation levels of the normal data with the cancer cell line data. This allows us to select significant CG sites to map to genes that are potential methylation biomarkers for breast cancer prediction. For chi-square analysis, we have disregarded the MCF10A cell line because MCF10A is not as tumorigenic as the other cell lines. Based on the hypothesis that the seven normal cells display less methylation variation than cancer cells, we compare the methylation-level variation of breast cancer cell lines with the normal cells using a chi-square statistical test. Finally, we select the CG sites that are shown to have significantly large variation, i.e., the CG sites with p-values less than a significance level.

First, we find each CG site at which there are at most two missing methylation values in both the cancer cell lines and normal individual data. At each CG site, we test the following hypotheses: $H_0: \sigma = \sigma_0^2$

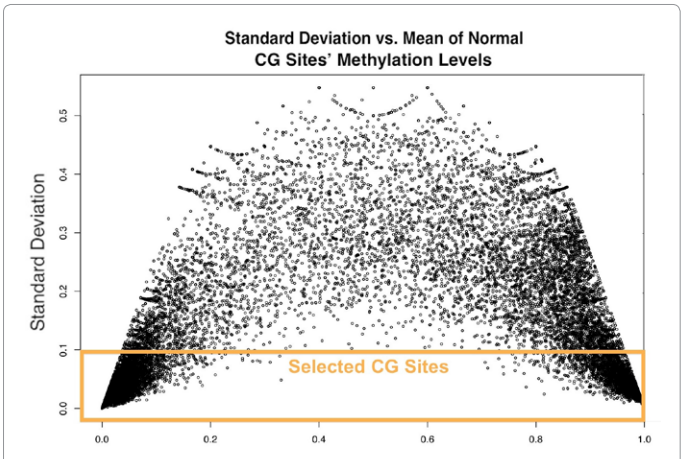

Figure 5: Standard deviation vs. mean methylation of select normal CG sites.

versus $H_a: \sigma > \sigma_0^2$, for an upper one-tailed test. In these hypotheses, σ_0^2 is the population variance of the non-tumorigenic samples, and σ^2 is the population variance of all samples including both tumorigenic cell lines and non-tumorigenic samples.

We assume that at each CG site the normal samples have more stable or similar methylation levels. Therefore, they have a smaller population variance than the cancer cell lines. We compare the variance of both the normal cell data and the cancer cell line data to find CG sites where this is true: if the breast cancer cell lines have higher variance than the normal cells at a certain CG site, it indicates that there are more methylation events occurring at this site, which may be either directly or indirectly related to breast cancer tumorigenesis.

Our null and alternative hypotheses are defined in terms of the population variance (or population standard deviation), and our chi-square test statistic is: $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$, where n in this case is fourteen from our seven breast cancer cell lines and seven normal cell data, s^2 is the sample variance of all the data, and σ_0^2 is the population variance of normal samples. The ratio s^2/σ_0^2 compares the sample standard deviation to the target standard deviation. The more this ratio deviates from 1, the more likely we are to reject the null hypothesis H_0. If our test statistic falls within the critical region of the chi-square distribution, we reject our null hypothesis, H_0, that the variance is equal to a specified value, σ_0^2. That is, we reject H_0 if $\chi^2 > \chi^2_{1-\alpha, n-1}$ for our upper one-tailed chi-square test. $\chi^2_{1-\alpha, n-1}$ is the critical value of the chi-square distribution with n - 1 degrees of freedom [26]. Small p-values indicate that cancer methylation-level variation is significantly larger than normal methylation variation. We select a significance level, α, of 0.01 to ensure a more conservative selection of significant CG sites.

Because statistical tests are conducted for a large number of CG sites, there is a high probability of type I error. Performing more than one chi-square test means the probability of making one mistake must be raised to the m^{th} power to find the true probability of error where m is the number of chi-square tests performed. This type of error requires that we correct for multiple testing. In this situation, we have used the Bonferroni approach to correct for multiple testing [27]. The Bonferroni approach is a single-step conservative approach in which equivalent adjustments are made to each p-value. The numbers of CG sites selected to do the chi-square test and the CG sites that are selected after the multiple testing corrections are given in Table 1. After performing the Bonferroni correction, we are able to accurately identify the genes and promoter regions that are significant within our dataset.

Chromosome	Number of CG sites input into statistical analysis	Number of CG sites selected after multiple testing correction
1	60705	14025
2	101491	41630
3	30219	7723
4	26076	6490
5	29345	8301
6	26751	6716
7	36989	8613
8	28083	7325
9	32794	6997
10	31829	8913
11	36581	7840
12	29895	7370
13	14811	3831
14	21360	5314
15	20613	5627
16	37291	6065
17	42875	9430
18	14016	4316
19	50023	7777
20	23110	5814
21	11004	1856
22	21307	4444
X	16010	1514

Table 1: The numbers of CG sites selected to do the chi-square test and the CG sites that are selected after the multiple testing corrections.

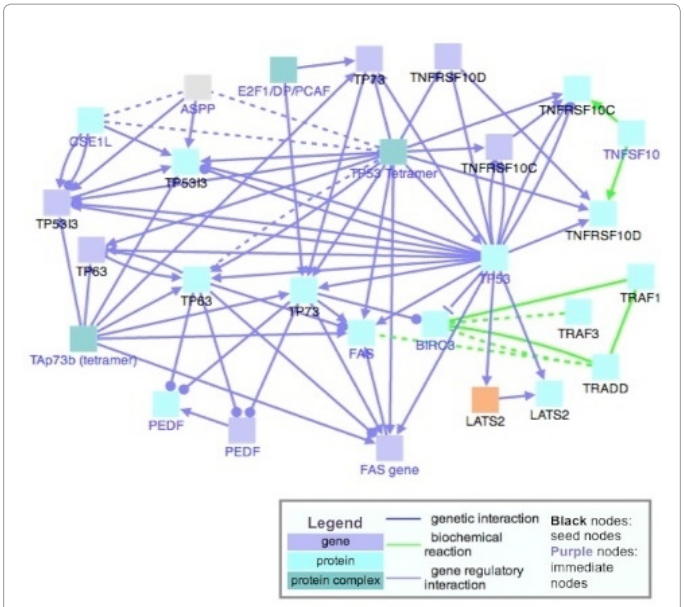

Figure 6: Tumor suppressor genetic pathways. This genetic pathway contains proteins and genes already known to be significant to cancer or tumorigenesis, including important tumor suppressors p53, p63 and various TNFs.

Results

We map the CG sites selected from our chi-square analysis to the corresponding genes and promoters. Using the genetic annotation code written by Dr. Sun's lab, we have generated a list of significant genes and promoter regions related to breast cancer. We then use these genes to conduct genetic pathway analysis. We use the list of genes that have at least one significant CG site to do genetic pathway analysis, which is used to pinpoint well-known cancer-associated genes.

Genetic pathway analysis

A genetic pathway is a collection of genes with direct and indirect interactions, including RNA and protein product interactions. We visualize genetic pathway (or network) with the ConsensusPathDB (CPDB) software [28-31]. The genetic pathway allows us to understand gene interactions and the impact of methylation errors on directly and indirectly related genes. The large number of interactions between protein products and genes indicates the significance of these genes and their influence on tumorigenesis. These genetic pathways also indicate important biological functions in cancer cells. Our pathway analysis results include the tumor suppressor genetic pathway that involves TP53, TP63, and TNF genes (Figure 6), as well as the breast cancer and estrogen reception pathway that involves ESR1 and GREB1, BCAR3, and TAF1 genes (Figure 7). Because the genes that are associated with tumorigenesis have been found in our dataset, the viability of our selection methods is confirmed.

TP53 in particular is known to be a major tumor suppressor gene, and this map (Figure 6) clearly shows the influence it has, along with other tumor suppressor genes such as TP63 and TP73, in tumor

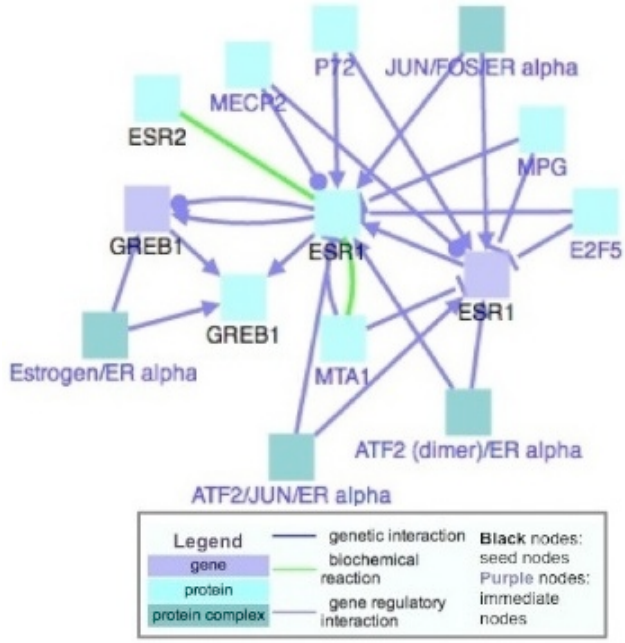

Figure 7: Genetic pathway contains estrogen reception genes. This genetic pathway contains proteins known to be significant in breast cancer expression and estrogen reception, particularly GREB1 and ESR1.

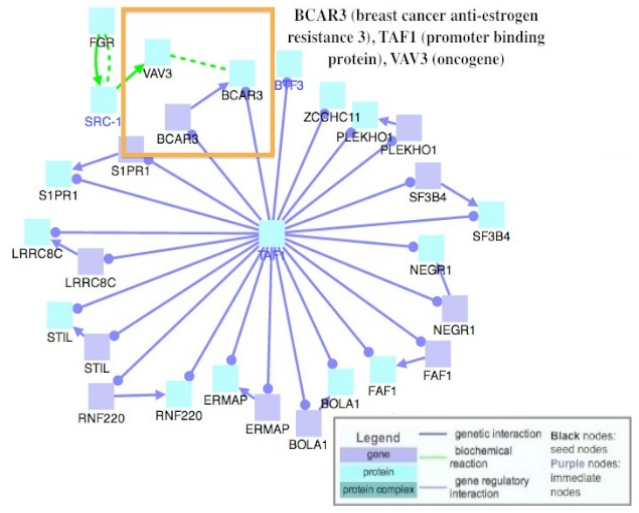

Figure 8: Genetic pathway linked to other genes. This genetic pathway indicates the link between seemingly non-cancer related genes and genes that are strongly linked to breast cancer tumorigenesis. Indirect linkages that may be overlooked in clinical trials can be seen clearly in genetic pathways such as the one shown above.

suppression, as nearly every other gene on the map is connected to it. Several tumor necrosis factors, such as TNFs, can also be seen in the gene linkage; it would make sense that these are related to the tumor suppressor genes as these factors are also involved in cell cycle regulation. Here we can see how methylation of genes that are in the same genetic pathway as major tumor suppressor genes can also contribute to breast cancer tumorigenesis. These genes include the FAS gene, encoding a cell surface death receptor acting as an apoptosis antigen, as well as CSE1, encoding the chromosome segregation 1-like protein and playing a role in cell proliferation and apoptosis. The information of these genes can be found from the GeneCards web (http://www.genecards.org/).

This genetic pathway network contains many genes linked directly to important breast cancer genes, particularly the ESR1 and GREB1 genes (Figure 7). ESR1 functions as an estrogen receptor and transcription factor. It is linked to nearly all of the genes in this genetic pathway network. As a transcription factor, ESR1 is likely influential in the expression of the other genes as shown in Figure 7. The GREB1 gene (growth regulation by estrogen in breast cancer) is a protein-coding gene, which may play a role in estrogen-stimulated cell proliferation. It is an early response gene in the estrogen receptor-regulated pathway and is already proven to be strongly linked to breast cancer.

We have found that ESR1 and GREB1 have large methylation variation, and they are important in breast cancer studies. The genes linked both directly and indirectly to these two genes, such as MTA1 (metastasis associated gene 1) and E2F5 (transcription factor crucial to the control of cell cycle and action of tumor suppressor proteins), may also be used as potential biomarkers for breast cancer tumorigenesis [6]. The relation of the other genes in this diagram to breast cancer development may be less evident. MPG (encoding DNA-3-methyladenine glycosylase) and ATF2 (encoding a cAMP-responsive element binding protein activating transcription factor) are both related to ESR1 and ESR2. MECP2 (encoding the methyl CpG binding protein 2) is capable of binding specifically to methylated DNA and repressing transcription from methylated gene promoters. However, some genes that are less obviously related to breast cancer genes have been overlooked in many databases [32], though they are indirectly or directly related to significant genes associated with breast cancer. Visualizing these genetic pathways provides a biological context for the list of potential biomarker genes that we have generated.

The genetic pathway network in Figure 8 shows genes that are more indirectly linked to known tumorigenic genes. The BCAR3 gene is a breast-cancer related gene, which is also influenced by the TAF1 promoter-binding protein. The TAF1 gene is also controlled by many other genes as shown in the diagram, which may then indirectly affect the BCAR3 gene. Furthermore, the BCAR3 protein product is also influenced by the VAV3 oncogene. This pathway shows us that even errors in gene expression seemingly unrelated to cancer can create enough of an effect on genes directly related to cancer genesis to eventually cause cancer in what is essentially a chain reaction.

Database contributions

We compare our list of selected genes with genes that are related to breast cancer as reported at the PubMeth [32] and GeneCards databases (genecards.org). The genes we found that are not previously contained in the databases can be researched further for possible linkage to breast cancer tumorigenesis. We have selected these genes based on the number of CG sites at which methylation variation is found to be significant with the chi-square testing. To be very selective, we choose genes that have at least one hundred significant CG sites from our chi-square tests. These 88 genes, we believe, are likely to be linked to breast cancer tumorigenesis, and uncharacteristic methylation patterns for these specific genes can be used as potential biomarkers in diagnosing breast cancer susceptibility. We have generated a genetic pathway diagram of these 88 genes (Figure 9) using the CPDB software to compare biological relationships between our selected genes and genes known to be linked to breast cancer. We use the 100 genes most strongly associated with breast cancer genesis from the PubMeth and GeneCards databases for our genetic pathway comparison (Figure 9).

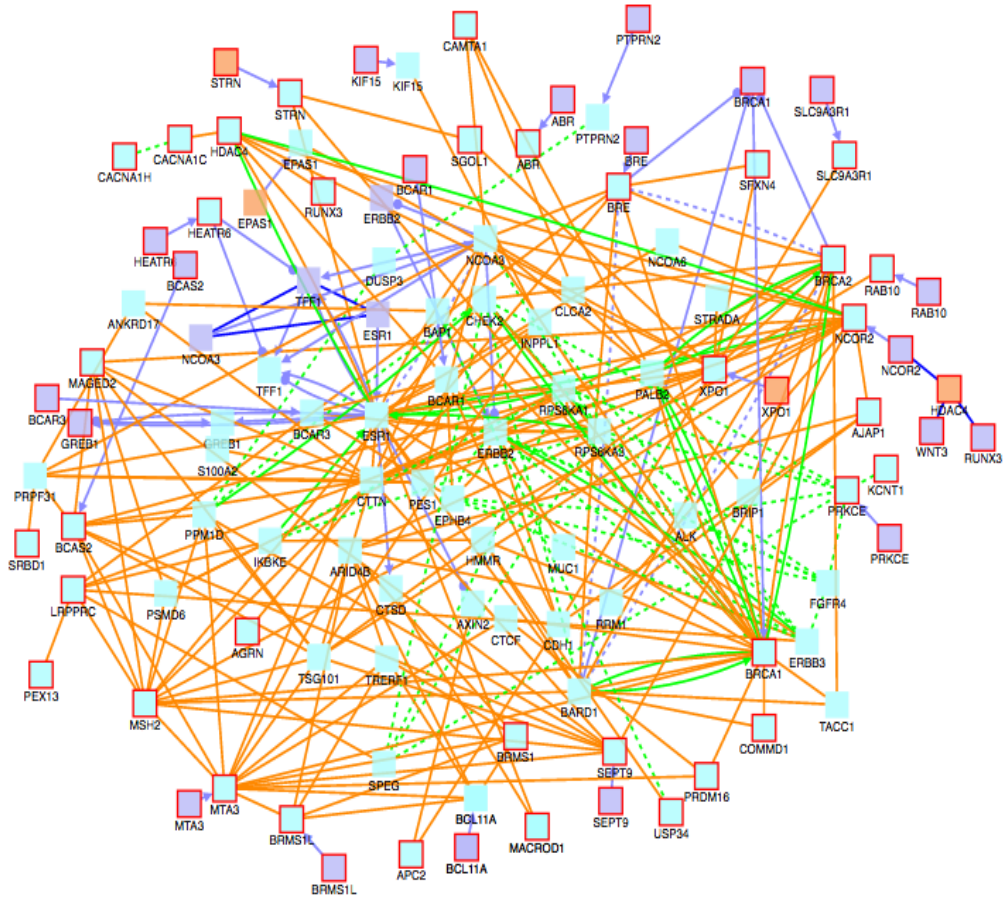

Figure 9: Genetic pathway related to 88 selected genes. The genes (boxed in red) that are significant in over 100 CG sites from the chi-square analysis are linked to genes associated with breast cancer. These are considered novel potential biomarkers.

In Figure 9, we can see several genes selected by our analysis (boxed in red) connected to the genes from the PubMeth and GeneCards databases. From the list of 88 genes selected based on significant methylation variation, we find 26 genes to be connected in genetic pathways to genes already linked to breast cancer, listed in Table 2.

These 26 genes, which are our pinpointed potential biomarkers for breast cancer, are biologically related to known breast cancer-related genes through genetic pathways. Several of these genes are linked to cell cycle regulation and tumor suppression, and others are directly linked to genes (such as BRCA1) that are the most strongly linked to breast cancer. These genes may be added to many prominent breast cancer gene databases, such as GeneCards (genecards.org) and PubMeth [32]. Furthermore, in studying their cellular functions and the genes that they are linked to, future research may be conducted to investigate the effect of these biomarkers on breast cancer tumorigenesis. The results of our research may also be used for epigenetic screening of patients by studying the methylation patterns of these biomarker genes.

Conclusion and Discussion

There is a significant need for easily implemented early detection and screening methods for cancer, as the five-year survival rate drops drastically in late-stage breast cancers [3-5]. We uncover a number of genes that can be used to determine a patient's breast cancer susceptibility by comparing methylation variation patterns of breast cancer data with normal cell data. We are able to pinpoint these genes by comparing methylation sequencing (RRBS) data from breast cancer cell lines with normal sample data from the ENCODE database. First, we compare the standard deviation of methylation levels of cancer and normal CG sites; if the cancer methylation standard deviation for each CG site is within a certain range, that CG site is mapped to a gene of interest. Then, we perform upper one-tailed chi-square tests to determine whether the variance of the normal cell methylation data and the variance of both normal and cancer cell methylation data are significantly different. We select only the genes with a large number of CG sites of significance.

Because our goal is to identify genes with large methylation variation patterns, variance and chi-square tests are obvious and convenient methods for us to use. It is not proper to use an F-test to compare the variances of the cancerous and normal samples because the F-test will fail to identify those important CG sites whose mean methylation jumps from ~0 to ~1 or ~1 to ~0 between normal and cancerous samples and the variability remains low in both normal and tumor cells. Besides the parametric method we used, non-parametric methods may also be suitable to analyze our data. The chi-square test can help us identify CG sites and genes with large methylation variation. However, the methylation variation patterns of CG sites within a long gene may be very different and it is important to investigate these variation differences within a long gene. We are working on studying the DNA methylation variation and heterogeneity patterns within genes in detail in another project.

Selected gene	Related breast cancer-linked gene	Description of selected gene function
AGRN (375790)	PPM1D	Iaminin G, Kazal type serine protease inhibitor, epidermal growth factor domains, modulates calcium homeostasis
AJAP1 (55966)	ARID4B	Adherens junction associated protein, playing a role in cell adhesion and migration
CAMTA1 (23261)	NCOA3	Calmodulin binding transcription activator 1, acts as a transcriptional activator and possible tumor suppressor
PRDM16 (63976)	CTCF	PR-domain zinc finger 16, binds DNA and acts as transcriptional regulator, with CEBPB regulates differentiation of my- oblastic precursors into adipose cells
RUNX3 (864)	BCAR1	Runt-related transcription factor 3, transcription factor to activate or suppress transcription, tumor suppressant (can be deleted or silenced in cancer)
WNT3A (89780)	ESR1 (indirect link)	Wingless-type MMTV integration site family member 3A, regulate cell patterns in embryogenesis, implicated in oncogenesis
BCL11A (53335)	ARID4B	B-cell CLL/Lymphoma 11A zinc finger protein, important in leukemogenesis and hematopoiesis, B-cell proto-oncogene
BRE (9577)	BRCA1	Brain and reproductive organ-expressed tumor necrosis factor modulator, homeostasis and cellular differentiation, death receptor-associated anti-apoptotic protein, regulate TNF- alpha signaling with interaction with TNFRSF1A
COMMD1 (150684)	BRCA1	Copper metabolism domain containing 1, regulate copper homeostasis, sodium intake and NF-kappa-B degradation
HDAC4 (9759)	ESR1 (indirect link)	Histone deacetylase 4, critical in transcriptional regulation and cell cycle progression, represses transcription
LRPPRC (10128)	ERBB3	Leucine-rich pentatricopeptide repeat containing, transcriptional regulation of nuclear and mitochondrial genes, role in cytoskeleton organization and vesicular transport
MSH2 (4436)	GREB1, ERBB2, TSG101, PSMD6	MutS homolog 2, component of post-replication DNA mismatch repair system, associated with hereditary nonpolyposis colon cancer
MTA3 (57504)	CTCF, TRERF1	Metastasis-associated 1 family member 3, maintain epithelial architecture and transcriptional repression
PEX13 (5194)	BCAR3 (indirect)	Peroxisomal biogenesis factor 13, peroxisomal membrane protein, deficiency leads to peroxisome biogenesis disorders
PRKCE (5581)	ERBB3, RPS6KA3	Protein kinase C epsilon, major receptors for phorbol esters (tumor promoters), regulate cell adhesion, motility, migration and cell cycle
USP34 (736)	BAP1, CDH1	Ubiquitin specific peptidase 34, regulate WNT pathway and processing of ubiquitinated proteins
XPO1 (7514)	BRCA2, CDH1,CTSD	Exportin 1, mediates nuclear export of cellular proteins, viruses (influenza A, HIV) use it to export spliced RNAs from the nucleus, causing further mutation
PTPRN2 (5799)	DUSP3	Protein tyrosine phosphatase receptor type N polypeptide 2, regulate cell cycle and oncogenic transformation
KCNT1 (57582)	ALK (indirect)	Potassium channel subfamily T member 1, sodium-activated potassium channel regulating homeostasis and developmental signaling pathways
MACROD1 (28992)	ESR1, CTSD	MACRO domain containing 1, plays a role in estrogen signaling by binding androgen receptors to amplify response, im- portant in carcinogenesis by activating ESR1 transcription
CACNA1C (775)	BCAR1 (indirect)	Calcium channel voltage-dependent L-type alpha 1C subunit, mediate calcium ion entry and calcium-dependent processes such as cell motility, gene expression and apoptosis
NCOR2 (9612)	ESR1	Nuclear receptor co-repressor, mediates transcriptional silencing, prevents basal transcription, cancer-associated
CACNA1H (8912)	BCAR1 (indirect)	Calcium channel voltage-dependent T-type alpha 1H subunit, mediate calcium ion entry and calcium-dependent processes such as cell motility, gene expression and apoptosis
ABR (29)	NCOA3 (indirect)	Active BCR-related gene, contains GTP-ase activating domain, plays a role in morphogenesis
SEPT9 (10801)	BCAS2, PPM1D	Septin 9, septin family involved in cytokinesis and cell cycle control, possible ovarian tumor suppressant
APC2 (10297)	AXIN2, CDH1	Adenomatosis polyposis coli 2, promotes rapid degeneration of CTNNB1 and may act as a tumor suppressor, WNT signaling

Table 2: 26 Novel biomarkers for breast cancer susceptibility. In the first column, the number below each gene symbol is the Entrez gene ID (a series of gene ID numbers).

By selecting abnormal standard deviations to isolate significant genes within the cancer cell lines, we find that some of the genes we have selected include the most prevalent genes seen specifically in breast cancer such as: TP53, TP63, TNFs, ESR1, GREB1, and BCAR3. From these similarities, we know that the methods we have used to compare the methylation levels of the CG sites are meaningful and produce valid results. Then, chi-square tests allow us to more accurately identify potential biomarker genes in breast cancer. For further isolation, after comparing our set of genes with genes linked to breast cancer from the ENCODE and GeneCards databases, we have obtained a set of genes that are candidates for biomarkers of breast cancer. Using the CPDB software, we have generated genetic pathways linking our candidate genes to the 100 genes that are most strongly linked to breast cancer tumorigenesis to explore biological mechanisms. From this analysis, we have pinpointed 26 genes that may be used as potential biomarkers for breast cancer susceptibility.

Ultimately, our results have many implications for the future research of breast cancer detection and treatment. In identifying several biomarkers for breast cancer tumorigenesis susceptibility, a patient's likelihood of developing tumors may be determined by high-throughput sequencing. The epigenome sequencing is a viable method of detection as RRBS (reduced representative bisulfite sequencing) is a cost-efficient method for profiling sequences of the genome with high GC contents. Furthermore, the 26 genes we identified as potential biomarkers, in contributing to the PubMeth and GeneCards breast cancer-related gene databases, may also be the basis for novel breast cancer prevention and treatment research.

Our results are useful for further breast cancer research. Researchers and medical doctors can perform DNA methylation analyses on their patients' epigenomic data and identify genes prevalent in the data. If these genes overlap with the genes we have discovered to be highly associated with cancer, that patient may be more likely to develop breast cancer. Further research can also be done on the relationships between these new potential biomarkers and other genes that allow further understandings of the genetic pathways involved in breast tumorigenesis. In the future, we plan to connect our identified DNA methylation variation patterns to specific cancer gene expression patterns. By understanding methylation variation patterns, we can predict regions of the human genome that are more likely linked to breast cancer. Using more selective or advanced statistical methods, we plan to show more linkages between different genes involved in cancer and find which genes are the main causal factors in cancer genesis.

Acknowledgements

This research was conducted during and after the students' participation in the 2014 Mathworks Honors Summer Math Camp at Texas State University. We give our special thanks to the Mathworks Honors Summer Math Camp program for giving us this opportunity to conduct research together. This work was supported by Dr. Shuying Sun's start-up funds and the Research Enhancement Program provided by Texas State University. We appreciate three anonymous reviewers' thoughtful comments and questions, which help us improve this manuscript greatly.

References

1. de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, et al. (2012) Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. The Lancet Oncology 13: 607-615.
2. Stewart BW, Wild CP (2014) World Cancer Report. IARC.
3. Houssami N, Given-Wilson R, Ciatto S (2009) Early detection of breast cancer: overview of the evidence on computer-aided detection in mammography screening. Journal of medical imaging and radiation oncology 53:171-176.
4. Domingo L, Jacobsen KK, Euler-Chelpin MV, Vejborg I, Schwartz W, et al. (2013) Seventeen-years overview of breast cancer inside and outside screening in Denmark. Acta Oncol 52:48-56.
5. Shetty MK (2010) Screening for breast cancer with mammography: current status and an overview. Indian journal of surgical oncology 1: 218-223.
6. Yang X, Yan L, Davidson NE (2001) DNA methylation in breast cancer. Endocr Relat Cancer 8: 115-127.
7. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9: 387-402.
8. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46.
9. Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. Genes & development 25: 1010-1022.
10. Feinberg AP, Cui H, Ohlsson R (2002) DNA methylation and genomic imprinting: insights from cancer into epigenetic mechanisms. Semin Cancer Biol 12: 389-398.
11. Plass C, Soloway PD (2002) DNA methylation, imprinting and cancer. European journal of human genetics 10: 6-16.
12. Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, et al. (2011) DNA methylation profiles of human active and inactive X chromosomes. Genome Res 21: 1592-1600.
13. Cao Y, Li Y, Zhang N, Hu J, Yin L, et al. (2015) Quantitative DNA hypomethylation of ligand Jagged1 and receptor Notch1 signifies occurrence and progression of breast carcinoma. American journal of cancer research 5: 1621-1634.
14. Cho YH, McCullough LE, Gammon MD, Wu HC, Zhang YJ, et al. (2015) Promoter Hypermethylation in White Blood Cell DNA and Breast Cancer Risk. Journal of Cancer 6: 819-824.
15. Perez-Janices N, Blanco-Luquin I, Torrea N, Liechtenstein T, Escors D, et al. (2015) Differential involvement of RASSF2 hypermethylation in breast cancer subtypes and their prognosis. Oncotarget
16. Pouliot MC, Labrie Y, Diorio C, Durocher F (2015) The Role of Methylation in Breast Cancer Susceptibility and Treatment. Anticancer Res 35: 4569-4574.
17. Ullah F, Khan T, Ali N, Malik FA, Kayani MA, et al. (2015) Promoter Methylation Status Modulate the Expression of Tumor Suppressor (RbL2/p130) Gene in Breast Cancer. PLoS One 10: e0134687.
18. Jones PA, Baylin SB (2007) The epigenomics of cancer. Cell 128: 683-692.
19. Jones PA, Buckley JD (1990) The role of DNA methylation in cancer. Adv Cancer Res 54: 1-23.
20. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, et al. (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc 6: 468-481.
21. Sun S, Noviski A, Yu X (2013) MethyQA: a pipeline for bisulfite-treated methylation sequencing quality assessment. BMC Bioinformatics 14:259.
22. Harris EY, Ponts N, Levchuk A, Roch KL, Lonardi S (2010) BRAT: bisulfite-treated reads analysis tool. Bioinformatics 26: 572-573.
23. Sun Z, Asmann YW, Kalari KR, Bot B, Eckel-Passow JE, et al. (2011) Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. PLoS One 6: e17490.
24. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.
25. R Development Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
26. Cochran WG (1952) The X2 test of goodness fit. Annals of Mathematical Statistics 25: 315-345.
27. Shaffer JP (1995) Multiple Hypothesis-Testing. Annu Rev Psychol 46:561-584.
28. Kamburov A, Stelzl U, Lehrach H, Herwig R (2013) The ConsensusPathDB interaction database: 2013 update. Nucleic acids research, 41(Database issue): D793-800.
29. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, et al. (2011) ConsensusPathDB: toward a more complete picture of cell biology. Nucleic acids research 39(Database issue): D712-717.
30. Pentchev K, Ono K, Herwig R, Ideker T, Kamburov A (2010) Evidence mining and novelty assessment of protein-protein interactions with the ConsensusPathDB plugin for Cytoscape. Bioinformatics 26: 2796-2797.
31. Kamburov A, Wierling C, Lehrach H, Herwig R (2009) ConsensusPathDB--a database for integrating human functional interaction networks. Nucleic acids research 37(Database issue):D623-628.
32. Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, et al. (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. Nucleic acids research 36(Database issue):D842-846.