A Pathway-Based Classification of Breast Cancer Integrating Data on Differentially Expressed Genes, Copy Number Variations and MicroRNA Target Genes

Hae-Seok Eo, Jee Yeon Heo, Yongjin Choi, Youngdon Hwang, and Hyung-Seok Choi*

Breast cancer is a clinically heterogeneous disease characterized by distinct molecular aberrations. Understanding the heterogeneity and identifying subgroups of breast cancer is essential to improving diagnoses and predicting therapeutic responses. In this paper, we propose a classification scheme for breast cancer which integrates data on differentially expressed genes (DEGs), copy number variations (CNVs) and microRNAs (miRNAs)-regulated mRNAs. Pathway information based on the estimation of molecular pathway activity is also applied as a postprocessor to optimize the classifier. A total of 250 malignant breast tumors were analyzed by k-means clustering based on the patterns of the expression profiles of 215 intrinsic genes, and the classification performances were compared with existing breast cancer classifiers including the BluePrint and the 625-gene classifier. We show that a classification scheme which incorporates pathway information with various genetic variations achieves better performance than classifiers based on the expression levels of individual genes, and propose that the identified signature serves as a basic tool for identifying rational therapeutic opportunities for breast cancer patients.

INTRODUCTION

Breast cancer is a clinically heterogeneous disease, so breast tumors which show a similar histology may entail markedly different prognoses and treatment responses (Henderson and Patek, 1998; Perou et al., 2000; van’t Veer et al., 2002). This clinical heterogeneity can be explained by molecular differences between tumors (Sorlie et al., 2001); as such, several promising techniques for coping with the problem of heterogeneity in breast cancer have been introduced (Ross et al., 2006). Recently, microarray-based gene expression profiling has been extensively used in the evaluation of breast cancer samples (Kim et al., 2012; Puustai et al., 2007).

The precise definition of breast cancer subtypes is still a subject of debate (Sontrop et al., 2011). In a seminal study by Perou et al. (2000), a molecular taxonomy of breast cancers was introduced as luminal, HER2-enriched (ERBB2), basal and normal breast-like subtypes based on their intrinsic gene set. More recent studies, however, have suggested that the fourth subtype, i.e., the normal breast-like group, is a less clear subtype (Desmedt et al., 2008; Parker et al., 2009). Despite these debates, it is widely accepted that breast cancer subtypes are associated with differences in terms of prognosis and treatment response.

For a more reliable and precise classification of breast cancer tumors, which is important for the proper selection of therapy, several intrinsic gene sets have been introduced (Ross, 2009). BluePrint, a commercialized multi-gene predictor, was developed, enabling rationalization in patient selection for either chemotherapy or hormone therapy prescription (Krijgsman et al., 2011). More recently, another molecular subtyping profile based on 798 probe-sets was developed using 327 breast cancer patient specimens (Kao et al., 2011). These excellent studies adopted microarray-based analysis and selected marker genes by scoring each individual gene (Kao et al., 2011; Krijgsman et al., 2011).

In this study, we describe a pathway-based classification of breast cancer which integrates data on differentially expressed genes (DEGs), copy number variations (CNVs) and microRNA (miRNA)-regulated mRNAs. Abnormal gene expression in cancer cells may be due to genomic or epigenomic alterations such as CNVs and miRNAs. Therefore, these multidimensional approaches are needed to identify and refine cancer-related genes (Donahue et al., 2012; Taylor et al., 2010). Pathway information was incorporated in a condition-specific manner (Pathway Activity inference using Condition-responsive genes, PAC) (Chuang et al., 2007; Lee et al., 2008). In a training set of 327 tumors, a 215-gene signature, which is not encoded as individual genes, but rather as subnetworks within larger biological pathways, was identified. Finally, the classification performances of the gene signature were compared with those of BluePrint and Kao et al. (2011) using an independent data set.

Keywords: breast cancer, classification, copy number variation, differentially expressed gene, microRNA, pathway
Breast Cancer Sub-Classification
Hae-Seok Eo et al.

(Dedeuwaerder et al., 2011; Lu et al., 2008; Richardson et al., 2006; Turashvilli et al., 2007).

MATERIALS AND METHODS

Breast tumor samples
The 327 patient samples with clinical data (GSE20685) were used for identifying DEGs and training our classification model (Kao et al., 2011). The validation data set (n = 250), which was used for evaluating the classification performances of the classifiers, was also derived from four independent data sets (GSE3744, GSE5460, GSE5764, GSE20711) (Dedeuwaerder et al., 2011; Lu et al., 2008; Richardson et al., 2006; Turashvilli et al., 2007). All samples were characterized by use of the Affymetrix Human Genome U133 Plus 2.0 array and normalized using the Robust Multi-array Average (RMA) method in R/BioConductor (http://www.bioconductor.org).

Identification of DEGs
The 866 candidate genes were selected from the literature and converted into Affymetrix probe identifiers using the DAVID GENE ID conversion tool. Centroids for each subtype were computed using the 327 training set. During cross-validation, we performed $k$-means clustering with the centroids and identified the probes that best discriminate between the four molecular subtypes. In each iteration of the cross-validation, an SD, which describes the difference for each probe over a series of samples, was increased by 0.1 and adopted that shows high prediction accuracy with a minimal set of genes (Fig. 2).

Identification of CNV-driven DEGs
The 516 breast invasive carcinoma samples containing clinical information with matched normal blood samples were downloaded from the TCGA. CNV and gene expression data across all samples were characterized by use of the Affymetrix SNP 6.0 array and Agilent 244K Custom Gene Expression G4502A-07-3 array, respectively. GISTIC analysis was then performed on the 516 CNV tumor data to identify any CNV regions that were significantly amplified or deleted.

Identification of putative miRNA target genes
The 22 data sets, which contain both 1,265 miRNA and 1,568 gene expression profiles, were collected from the GEO. Additional information related to the samples is available in Table S3. For the identification of miRNA-regulated miRNAs related in breast cancer, we calculated Pearson's correlations for all combinations of miRNA-mRNA pairs over a series of samples in each data set and compiled co-expressed miRNA-mRNA pairs that show significant correlations. (For a description of the association analysis between miRNA-mRNA pairs in detail, please refer to Cho et al., 2011)

Incorporation of pathway information
A total of 225 control samples, which were used in the calculation of molecular pathway activity, were collected from seven independent data sets (GSE5764, GSE7904, GSE8977, GSE-10780, GSE10810, GSE15043, GSE17907) (Chen et al., 2010; Gu et al., 2009; Karnoub et al., 2007; Pedraza et al., 2010; Richardson et al., 2006; Sircoulomb et al., 2010; Turashvilli et al., 2007). 1,340 pathway information were also downloaded from the Pathway Commons database. We integrated the expression and pathway data sets by overlaying the expression values of each gene on its corresponding protein in each pathway, and searched for subnetworks whose activities between tumors and controls were highly discriminative (Chuang et al., 2007; Lee et al., 2008).

RESULTS

Overview of a classification scheme for breast cancer
861 breast cancer-related genes were collected by integrating the results from three different genetic analyses, including the identification of DEGs, CNV-driven DEGs, and putative miRNA target genes (data not shown). Subsequently, unnecessary genes containing redundant information, which may lead to decreased classification performance, were sequentially eliminated by incorporating pathway information with different threshold values (Chuang et al., 2007; Lee et al., 2008). (See Fig. 1 for an overview of the analyses performed.) To derive the intrinsic gene list, a 327-tumor training set (GSE20685) with clinical data, which deposited in the Gene Expression Omnibus (GEO) (www.ncbi.nlm.gov/geo) (Barrett et al., 2011), was used.

Identification of differentially expressed genes (DEGs)
866 candidate genes, which had previously been adopted in the prognosis, grading and classification of breast cancers, were selected from the literature, and the mean expression value for each candidate gene in each subtype, called a centroid, was computed using the 327-tumor training set. For the development of optimized centroids, only those tumors which showed the highest correlation with each other within a given subtype were adopted for this calculation.

Subtype predictions were performed by using $k$-means clustering (Kaufmann and Rousseeu, 1990) with the calculated centroids. For the selection of probes with the most discriminating power, a standard deviation (SD) that describes the difference for each probe over a series of samples was calculated. During cross-validation, a SD was iteratively increased by 0.1, and a minimal set of probes that showed high prediction accuracy was chosen for the identification of DEGs. In this study, a 646-gene signature comprising 1,340 probes with an SD of 0.8 showed improved accuracy, with 82.42% for breast cancer subtyping compared to 866-gene signature comprising 2,373 probes with an SD of 0.0 (Fig. 2).

Identification of CNV-driven DEGs
516 breast invasive carcinoma samples, which contain both CNV and gene expression data, were obtained from The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov) for the identification of CNV-driven DEGs. CNV frequency and amplitude were examined to determine the significance of CNVs using GISTIC analysis (Beroukhim et al., 2007), and 324 significant CNV regions comprising 327 genes were identified from the tumor samples. Among these significant CNV regions, chromosomal amplification (1q21.3, 6p21.3, 7p11.2, 8q24.3) and deletion (8q23.3, 11q22, 22q13) have been proposed as loci for breast cancer susceptibility genes (Hawthorn et al., 2010).

Subsequently, by calculating Pearson's correlation between the DNA copy number and gene expression level, 73 CNV-driven genes, which differentially expressed between tumor and normal samples, were selected from the 327 candidate genes (see Supplementary Table S2). ERBB2, MYC, GSTT1, PIK3CA and CWF19L2, which have also been reported to be related to chances of developing breast cancer (Figueiredo et al., 2007; Loi et al., 2010; Nordgard et al., 2008; Oliveira et al., 2010; Qu et al., 2008), were included in the CNV-driven genes.