Identification of differentially expressed molecular functions associated with breast cancer using Gibbs sampling

GANG ZHOU¹, MING-QIAN LU², DAO-JUN LI², BAO-AN GAO³ and RONG GUO²

¹Internal Medicine; Departments of ²Oncology and ³Respiration, The First College of Clinical Medical Science, China Three Gorges University, Yichang Central People's Hospital, Yichang, Hubei 443003, P.R. China

Received February 29, 2016; Accepted July 27, 2017

DOI: 10.3892/ol.2017.7158

Abstract. The aim of the present study was to identify differentially expressed molecular functions (DEMFs) for breast cancer using the Gibbs sampling approach. Molecular functions (MFs) were obtained on the basis of the Bayesian Approach for Geneset Selection package. Subsequently, MFs were converted into Markov chains (MCs) prior to calculating their probabilities, utilizing the MC Monte Carlo algorithm. DEMFs were identified with probabilities ≥0.8 and the gene compositions were studied. Finally, a co-expression network was constructed via the empirical Bayes method and a pathway enrichment analysis of genes in DEMFs was performed. A total of 396 MFs were identified and all transformed to MCs. With the threshold, 2 DEMFs (structural molecule activity and protein heterodimerization activity) were obtained. The DEMFs were comprised of 297 genes, 259 of which were mapped to the co-expression network. These 297 genes were identified to be enriched in 10 pathways, and ribosome was the most significant pathway. The results of the present study revealed 2 DEMFs (structural molecule activity and protein heterodimerization activity) which may be associated with the pathological molecular mechanisms underlying breast cancer, based on Gibbs sampling.

Introduction

Breast cancer is the most common type of cancer diagnosed in females and the second leading cause of cancer mortality in females worldwide, following lung cancer (1). Breast cancer is characterized by a distinct metastatic pattern involving the regional lymph nodes, bone marrow, lungs and liver (2). A family history of breast cancer and other factors (including female sex, elderly age and ionizing radiation) can increase the risk of developing breast cancer (3). In addition, 5-10% of breast cancer cases are caused by inherited gene mutations (4). Several gene markers have been identified to predict the response to therapeutic regimens, including receptor tyrosine-protein kinase ErbB-2 and Stearoyl-CoA desaturase-1 (5-7). However, the pathological molecular mechanisms underlying breast cancer remain unknown. Individual anticancer therapies and the identification of molecular functions (MFs), on the basis of genes identified as significant, may enable this challenge to be solved.

Gene ontology (GO) is intended to enable the annotation of homologous gene and protein sequences in multiple organisms using a common vocabulary that results in the ability to query, and retrieve genes and proteins on the basis of shared biology (8). The ontology covers three categories: Cellular component, biological process and MF (8). Studying MFs enables analysis of the elemental activities of a gene at the molecular level. MF is defined as the biochemical activity (including specific binding to ligands or structures) of a gene product and describes only what happens, without specifying where or when the event occurs (9). However, a limited number of studies have been performed with the aim of identifying the roles of significant MFs in the progression of a disease.

The objective of the present study was to identify differentially expressed MFs (DEMFs) for breast cancer using the Gibbs sampling approach. To achieve the objectives, MFs were evaluated according to the Bayesian Approach for Geneset Selection (BAGS) package and DEMFs were identified using Gibbs sampling on the basis of the Markov chain (MC) Monte Carlo (MCMC) algorithm. Subsequently, additional analyses were performed for DEMFs, including construction of a co-expression network and pathway enrichment analysis.

Materials and methods

Gene expression data recruitment and pre-processing. The publicly available microarray gene expression profile of breast cancer was recruited from EMBL-EBI ArrayExpress (www.ebi.ac.uk/arrayexpress; accessing no. E-GEOD-10780) (10). It was presented on the A-AFFY-44-Affymetrix GeneChip Human Genome U133 Plus 2.0 platform (HG-U133_Plus_2; Affymetrix; Thermo Fisher Scientific, Inc., Waltham, MA,
To control the quality of the expression data on the probe-level, standard procedures were performed, including background correction (11), normalization (12), probe correction (13) and summarization (11). Briefly, background correction and normalization were carried out using the Robust Multi-array Average algorithm and quantile based algorithm, to eliminate the influence of nonspecific hybridization. Furthermore, the Micro Array Suite algorithm was used to revise perfect match and mismatch values, and the median polish method was used to summarize expression values. Subsequently, the preprocessed probe-level dataset were converted into expression measures and screened using the feature filter method to discard duplicated genes (14). A total of 20,102 genes were obtained.

MFs identification. To identify MFs for breast cancer, the ‘AnnotationMFGO’ function of the BAGS package was utilized. The BAGS package provides functions to perform statistical identification of gene functional classes that behave in a distinct manner between the phenotypes of interest for datasets under cross-sectional or time series designs (15,16). The normalized data set was inputted into the package and used to identify enriched MFs, but only MFs from GO with ≥5 ge were distinguished from the acceptance rate, which was the average of the acceptance probability over iterations.

Gibbs sampling. In the present study, Gibbs sampling was used to investigate the significance and functions of MFs in breast cancer. Gibbs sampling, typically used as a means of statistical inference, particularly Bayesian inference, is a MCMC algorithm for obtaining a sequence of observations that are approximated from a specified multivariate probability distribution (17-19).

MC. To implement the Gibbs sampler procedure, the MFs were transformed to a data set with functional class expression measurements that were MCAs. An MC is a sequence of random variables where the distribution of each random variable depends only on the value of the previous random variable (20). For example, a sample m was taken from a distribution depending on a parameter vector o ε O of length T, with prior distribution g(o_1, …, o_T). It may be that T was large and that numerical integration to identify the marginal densities of the o_i may be computationally expensive. An alternative method of calculating the marginal densities was to create an MC on the space O by repeating these two steps: The first step, selecting a random index 1 ≤ j ≤ T; the second step, endowed a new value for o_j according to g(o_1, …, o_{j-1}, o_j, …, o_T).

Posterior inference. Following the transformation into MCs, posterior inferences f were defined to identify the probability distributions of MFs from breast cancer (21). The posterior distribution was represented by q(x, y), which was the probability of parameter x provided an observation y. Provided that prior information is available about the parameters p(x), the posterior distribution may be developed through the normalized product of the prior and a sampling distribution (likelihood), q(y | x). The algorithm associated with the objective (target) density f and the conditional density q produces an MC (X_(t)^(o)). Provided x_(t) denoted the value of X^(o), y = q(x) it was defined Y_t ∼ q(y|x^(o)),

\[ x_{t+1}^{(i)} = \begin{cases} Y_t & \text{with probability } \rho(x^{(i)}, Y_t), \\ X_t^{(i)} & \text{with probability } 1 - \rho(x^{(i)}, Y_t) \end{cases} \]

Where q(x) was an MF, the probability was represented by q(y, x), which was calculated as follows:

\[ \rho(x, y) = \min \left\{ f’(y') q(x | y') \middle/ f’(y) q(y | x) \right\} \]

The distribution q was termed the instrumental (or proposal or candidate) distribution and the probability q(y, x) was calculated as follows:

\[ \bar{\rho} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \rho(X_t^{(i)}, Y_t) = \int \rho(x, y) f’(x | y) q(y | x) dx dy \]

T was the length of the observation vector. Furthermore, this algorithm satisfied the balance condition (22), f(x)K(y | x) = f(y)K(x | y), from which it was deduced that f was the stationary distribution of the MC [X^(o)] by integrating each side of the equality in x; K was transition kernel.

DEMFs. In the present study, probabilities of MFs were evaluated using the posterior distributions on the basis of Gibbs sampling. If the posterior probability of an MF was ≥0.05, this indicated the reliability of the MF (23). However, there was not a standard for high frequencies. Based on previous study (15), if the probability of an MF was ≥0.8, the MF was differentially expressed. To additionally investigate the properties of DEMFs, the gene compositions of DEMFs were studied, and those genes involved in the DEMFs were considered to be differentially expressed genes (DEGs).

Co-expression network. The empirical Bayes (EB) approach was selected to construct a co-expression network for the DEGs of breast cancer (24). An l-by-n matrix of expression values was created, where l was the number of genes (or probes) under consideration and n was the total number of microarrays over all conditions. These values should be normalized and the E obtained. The conditions array of length was n. The members of this array should use values in 1-F, where F was the total number of conditions. Using ‘ebPatterns’ function to define equal co-expression/differential co-expression (DC) classes.

Using E and the conditions array, the intra-group associations for all s = l(l-1)/2 gene pairs were calculated and the resulting D matrix of associations were s-by-F. Initialization of the hyper-parameters was performed to identify the component normal mixture model that fit the associations of D following transformation. Expectation Maximization calculations were not conducted and instead the initial estimates of the hyper parameters were used to generate posterior probabilities of DC (25). Ultimately, the co-expression network was visualized by Cytoscape 3.5.1 (http://www.cytoscape.org/).

Topological analysis. To understand the functionality of complex systems of gene signatures in the co-expression
network, the biological importance of genes was characterized using indices of topological analysis. Degree was counted by summing up the number of the adjacent genes (26). The degree \( D(v) \) of a node \( v \) was defined as follows:

\[
D(v) = \sum_{j} a_{ij}
\]

Pathway enrichment analysis. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis for DEGs was performed by using the online tool Database for Annotation, Visualization and Integrated Discovery (DAVID) (27). KEGG pathways with \( P<0.01 \) were selected on the basis of Expression Analysis Systematic Exploried (EASE) test implemented in DAVID (28). The calculated formula of EASE was as follows:

\[
P = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}}
\]

Where \( n = a'+b+c+d \) was the number of background genes, \( a' \) was the gene number of one gene set in the gene lists, \( a'+b \) was the number of genes in the gene list including at least one gene set, \( a'+c \) was the gene number of one gene list in the background genes; \( a' \) was replaced with \( a = a'-1 \).

Results

Identification of MFs. Under the condition of GOs with \( \geq 5 \) genes, a total of 396 MFs were obtained based on the BAGS package. Of all the MFs, 32 possessed between 100 and 200 genes, 34 possessed between 200 and 300 genes, and 24 terms possessed \( \geq 300 \) genes. DNA binding contained 561 genes, receptor activity contained 542 genes, and transferase activity transferring phosphorus containing groups included 401 genes. Transmembrane receptor activity with 390 genes and substrate specific transporter activity of 378 genes were markedly MF terms for breast cancer. The MFs with \( \geq 200 \) genes are listed in Table I.

Identification of DEMFs. Prior to evaluating the probabilities of MFs, using Gibbs sampling on the basis of the MCMC algorithm, the 396 MFs in expression data structure were converted to MCs. Fig. 1 illustrates the association between probability distribution and each MF, and lists the top five MFs with high probabilities, including: Structural molecule activity, 0.967; protein heterodimerization activity, 0.889; metalloendopeptidase activity, 0.576; structural constituent of the cytoskeleton, 0.550; integrin binding, 0.539. In addition, an MF with probability \( \geq 0.8 \) was defined as DEMF, thus two DEMFs were obtained (structural molecule activity and protein heterodimerization activity) for breast cancer. The results of the present study demonstrated that structural molecule activity was comprised of 222 genes; whereas, protein heterodimerization activity consisted of 75 genes. Genes in the DEMFs were denoted as DEGs in the present study; therefore, a total of 297 DEGs were identified, all of which may serve functions in the progression of breast cancer.

The co-expression network for 297 DEGs was constructed using the EB approach and subsequently visualized using Cytoscape 3.5.1 (Fig. 2). There were 259 nodes and 985 edges in the co-expression network. Subsequently, topological analysis was implemented for the co-expression network to identify significant genes. As presented in Fig. 2, the top 10 nodes with higher degrees were ribosomal protein (RP)L14 (degree, 49), RPL27A (degree, 45), RPLP2 (degree, 45), CDC42 small effector 2 (degree, 40), RPS2 (degree, 40), actin like 7A (degree, 35), RPL3 L (degree, 35), RPL4J (degree, 35), RPS18 (degree, 35) and topoisomerase (DNA) II β (degree, 34).

KEGG enrichment analysis for DEGs was conducted to gain additional insight into DEMFs and to identify the
associated biological pathways. A total of 10 significant pathways with P<0.01 were identified (Table II). Ribosome (P=1.27x10^{-64}), hypertrophic cardiomyopathy (P=1.09x10^{-7}) and dilated cardiomyopathy (P=3.02x10^{-7}) were identified as the three most significant pathways in breast cancer. When calculating the counts for each pathway, the ribosome pathway was found to be the most significant, followed by hypertrophic cardiomyopathy and dilated cardiomyopathy.

**Table II. Pathways with P<0.01, based on differentially expressed genes from differentially expressed molecular functions.**

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Count</th>
<th>P-value</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ribosome</td>
<td>55</td>
<td>1.27x10^{-64}</td>
<td>RPL18, RPL19, RPL14, RPL13, RPL15, RPLP2, RPS2, RPS3, RPLP0, RPLP1, FAU, RPL11, RPS4X, RPS18, RPS19, RPL41, RPS16, RPS15, RPS12, RPL3L, RPS4Y1, RPS10, RPS11, RPL27A, RPL35, RPS15A, RPL37, RPL38, RPL39, RPS25, RPL30, RPS27, RPS28, RPS29, RPL7, RPL6, RPL31, RPL9, RPL34, RPL8, RPL3, RPS21, RPS23, RPS24, RPS9, RPL27, RPL23A, RPL24, RPS6, RPS5, RPS8, RPS28, RPS29, RPS7, RPL22</td>
</tr>
<tr>
<td>Hypertrophic cardiomyopathy</td>
<td>15</td>
<td>1.09x10^{-7}</td>
<td>ACTB, LAMA2, ACTC1, DES, MYL2, MYL3, DMD, MYBPC3, MYH7, MYH6, TTN, TPM2, TPM1, TPM4, TGF2B</td>
</tr>
<tr>
<td>Dilated cardiomyopathy</td>
<td>15</td>
<td>3.02x10^{-7}</td>
<td>ACTB, LAMA2, ACTC1, DES, MYL2, MYL3, DMD, MYBPC3, MYH7, MYH6, TTN, TPM2, TPM1, TPM4, TGF2B</td>
</tr>
<tr>
<td>Tight junction</td>
<td>17</td>
<td>1.25x10^{-6}</td>
<td>ACTB, PPP2R1B, PPP2R1A, MYL5, MYL2, EPB41, VAPA, MYLPF, MYH7, ACTN2, MYH6, ACTN3, LLGL1, MYL9, PPP2CA, MYH11, SPTAN1</td>
</tr>
<tr>
<td>Focal adhesion</td>
<td>20</td>
<td>4.41x10^{-6}</td>
<td>EGFR, COL4A4, ACTB, CAV1, COL4A2, MYL5, MYL2, ERBB2, MYLPF, ACTN2, ACTN3, MYL9, LAMA2, LAMA1, LAMA4, LAMC3, BCL2, COMP, LAMC1, LAMB1</td>
</tr>
<tr>
<td>Small cell lung cancer</td>
<td>10</td>
<td>6.72x10^{-4}</td>
<td>COL4A4, LAMA2, LAMA1, LAMA4, COL4A2, LAMC3, BCL2, TP53, LAMC1, LAMB1</td>
</tr>
<tr>
<td>ECM-receptor interaction</td>
<td>10</td>
<td>6.72x10^{-4}</td>
<td>COL4A4, LAMA2, LAMA1, LAMA4, COL4A2, LAMC3, COMP, AGRN, LAMC1, LAMB1</td>
</tr>
<tr>
<td>Cardiac muscle contraction</td>
<td>8</td>
<td>7.21x10^{-3}</td>
<td>ACTC1, MYL2, MYL3, MYH7, MYH6, TPM2, TPM1, TPM4</td>
</tr>
<tr>
<td>ABC transporters</td>
<td>6</td>
<td>8.50x10^{-3}</td>
<td>ABCG8, ABCG5, TAP2, TAP1, ABCG4, ABCG1</td>
</tr>
<tr>
<td>Regulation of actin cytoskeleton</td>
<td>14</td>
<td>9.35x10^{-3}</td>
<td>EGFR, ACTB, MYL5, MYL2, MYLPF, ACTN2, ARPC4, ARPC5, ACTN3, MYL9, ARPC1B, ARPC3, ARPC2, MSN</td>
</tr>
</tbody>
</table>

ECM, extracellular matrix; ABC transporter, ATP-binding cassette transporter. The Expression Analysis Systematic Explored (EASE) test was used for analysis.

Figure 1. Probabilities for 396 MFs. The top five MFs with high probabilities were structural molecule activity, 0.967; protein heterodimerization activity, 0.889; metalloendopeptidase activity, 0.576; structural constituent of the cytoskeleton, 0.550; integrin binding, 0.539. MF, molecular function.

Figure 2. Co-expression network on the basis of differentially expressed genes from differentially expressed molecular functions. RPL, ribosome protein L; TOP2B, topoisomerase II β; RPS, ribosomal protein S; ACT17A, actin-like 7A; KRT2, keratin 2; CDC42SE2, CDC42 small effector 2.
was enriched by the maximum numbers of 55 DEGs, a number of which exhibited a high degree (including RPL14, RPL27A and RPLP2), thus it may be a potential biomarker for the detection and treatment of breast cancer.

Discussion

In the present paper, two DEMFs were identified for breast cancer, including structural molecule activity and protein heterodimerization activity. Structural molecule activity is defined as the action of a molecule that contributes to the structural integrity of a complex or assembly within or outside a cell (29). A previous study demonstrated that structural molecule activity was among the top ten gain/loss spatial functional hotspots of hepatocellular carcinoma with enriched functions (29). In breast cancer, Hong et al (30) reported that the dominant function of the identified protein was involved in the structural molecule activity. It has been suggested that extracellular matrix protein expression and nuclear gene expression are associated with a negative feedback regulatory mechanism of breast cancer, and differential genes actively participated in structural molecule activity (31). In addition, ~9.3% of upregulated genes in breast tumors were identified to belong to categories of structural molecule activity (32). In the present study, the DEMF, structural molecule activity, was associated with breast cancer, which validated the feasibility of Gibbs sampling to identify DEMFs for breast cancer.

Protein heterodimerization activity interacts selectively and non-covalently with a non-identical protein to form a heterodimer (33). Previous studies had revealed that Protein heterodimerization activity was involved in a number of cancer types, including nasopharyngeal carcinoma (33), esophageal cancer (34) and hepatocellular carcinoma (35). Topoisomerase IIα is a prognostic biomarker and a novel therapeutic target of nasopharyngeal carcinoma, which was identified to be associated with tumor aggressiveness and significantly enriched in protein heterodimerization activity (33). However, there are a limited number of studies that have focused on the association between breast cancer and protein heterodimerization activity directly. The Erb-b2 receptor tyrosine kinase (ErbB) 2/Erb-b3 heterodimer functioned as an oncogenic unit to drive breast tumor cell proliferation (36). The receptor tyrosine kinase Erb-b2 receptor tyrosine kinase 2 was demonstrated to be overexpressed in ~25% breast cancer cases and to serve a signaling function, via self-constitutive phosphorylation and the trans-activation of heterodimerization partners (37). The results of the present study revealed that the heterodimerization activity of proteins or genes were identified and associated with the formation of breast cancer.

In the present study, the gene compositions of DEMFs were subsequently studied and the results identified that the ribosome was the most significant pathway. In the ribosome pathway, RPL14, RPL27A and RPLP2 were the three DEGs that exhibited the highest degrees. Ribosomes, the organelles that catalyze protein synthesis, consist of a small 40S subunit and a large 60S subunit. RPL14, RPL27A and RPLP2 belong to members of ribosomal protein L (RPL) family. A number of human pathological conditions, not associated with genetic alterations of oncogenes or tumor suppressors, are associated with an increased risk of developing cancer and may be characterized by quantitative and/or qualitative changes in ribosomes (38). Therefore, in the present study it was considered that DEG-associated analysis demonstrated the two DEMFs were associated with breast cancer.

The results of the present study identified two DEMFs (structural molecule activity and protein heterodimerization activity) utilizing Gibbs sampling, and validated the associations between these DEMFs and breast cancer. Thus, the results of the present study may provide a novel insight into the underlying pathological mechanisms of breast cancer.

References