SCIENTIFIC REPORTS

OPEN

BCIP: a gene-centered platform for identifying potential regulatory genes in breast cancer

Received: 14 December 2016
Accepted: 20 February 2017
Published: 22 March 2017

Jiaqi Wu[1,*], Shuofeng Hu[1,*], Yaowen Chen[1,2], Zongcheng Li[1,3], Jian Zhang[1], Hanyu Yuan[1], Qiang Shi[1], Ningsheng Shao[1] & Xiaomin Ying[1]

Breast cancer is a disease with high heterogeneity. Many issues on tumorigenesis and progression are still elusive. It is critical to identify genes that play important roles in the progression of tumors, especially for tumors with poor prognosis such as basal-like breast cancer and tumors in very young women. To facilitate the identification of potential regulatory or driver genes, we present the Breast Cancer Integrative Platform (BCIP, http://omics.bmi.ac.cn/bcancer/). BCIP maintains multi-omics data selected with strict quality control and processed with uniform normalization methods, including gene expression profiles from 9,005 tumor and 376 normal tissue samples, copy number variation information from 3,035 tumor samples, microRNA-target interactions, co-expressed genes, KEGG pathways, and mammary tissue-specific gene functional networks. This platform provides a user-friendly interface integrating comprehensive and flexible analysis tools on differential gene expression, copy number variation, and survival analysis. The prominent characteristic of BCIP is that users can perform analysis by customizing subgroups with single or combined clinical features, including subtypes, histological grades, pathologic stages, metastasis status, lymph node status, *ER/PR/HER2* status, *TP53* mutation status, menopause status, age, tumor size, therapy responses, and prognosis. BCIP will help to identify regulatory or driver genes and candidate biomarkers for further research in breast cancer.

Breast cancer is a frequently diagnosed carcinoma and is the leading cause of cancer death among females worldwide. An estimated 1,676,600 cases were diagnosed and 521,900 deaths occurred in 2012, accounting for 25% of the total cancer cases and 15% of all the cancer deaths among females[1]. Breast cancer is a heterogeneous disease with a high degree of diversity in morphology, histology, pathological features, and molecular alterations, such as gene mutations and abnormal expression[2]. Researches based on gene expression (GE) patterns have classified breast cancer into distinct subgroups corresponding to different prognostic outcomes and therapeutic responses[3–6]. A number of studies have focused on the recognition of biomarkers and characterization of gene function in particular breast cancer subgroups[7–10].

Triple-negative breast cancer (TNBC) is a highly aggressive subtype of breast cancer and the vast majority is basal-like phenotype[11]. Due to its high genetic heterogeneity, TNBC does not possess a common genetic mutation and thus lacks effective targeted therapies[12]. However, several studies based on GE have identified several critical genes that may be potential druggable targets for the treatment of TNBC[13–15]. For example, *MELK* has been characterized as an oncogenic kinase essential for basal-like breast cancer (BBC) via a kinome-wide screening, integrative analysis with multiple GE datasets, and further *in vitro* and *in vivo* experiments[14]. *BCL11A* has also been reported to be a novel TNBC oncogene by *in silico* analysis of several microarray datasets and subsequent experimental validations[15]. These studies suggest that GE profiles are important resources for regulatory gene and biomarker identification in breast cancer. Differential expression analysis, copy number variation (CNV) analysis, survival analysis, and co-expression analysis on multiple credible and qualified datasets are effective approaches for recognizing novel regulatory genes and biomarkers.

[1]Beijing Institute of Basic Medical Sciences, Beijing 100850, China. [2]Department of Obstetrics and Gynecology, Fuzhou General Hospital of Nanjing Military Command, Fuzhou, Fujian 350025, China. [3]Translational Medicine Center of Stem Cells, 307-Ivy Translational Medicine Center, Laboratory of Oncology, Affiliated Hospital, Academy of Military Medical Sciences, Beijing 100071, China. [*]These authors contributed equally to this work. Correspondence and requests for materials should be addressed to X.Y. (email: yingxm@bmi.ac.cn)

In order to help researchers use gene expression profiles, some databases and tools have been developed[16–19]. However, integrative platforms combined with multi-omics data and customized analysis tools for breast cancer are still lacking. In this study, we developed BCIP, which provides differential expression analysis, copy number variation analysis, survival analysis, co-expression analysis, microRNA (miRNA) regulation analysis, and pathway analysis for query genes. To ensure the reliability of the analysis, we collected and obtained GE profile data on 9,381 samples from 29 datasets with strict quality control and uniform processing. We also incorporated CNV information on 3,035 samples, 324,219 miRNA-target interactions, 286 KEGG pathways, and data from tissue-specific gene functional networks of mammary gland and mammary epithelium. In order to facilitate researchers' analysis of the specific subgroups they focused on, we developed a comprehensive and flexible interface that permits users to customize subgroups with single or combined clinical features of interest, including subtypes, grades, stages, metastasis status, lymph node status, prognosis, age, tumor size, *ER/PR/HER2* status, *TP53* mutation status, menopause status, and therapy response. BCIP will be a valuable tool for the identification of regulatory or driver genes in breast cancer.

Methods

Data collection and processing.

We initially retrieved and collected data from NCBI Gene Expression Omnibus[20] (GEO), European Genome-phenome Archive of EMBL European Bioinformatics Institute[21] (EMBL-EBI), and The Cancer Genome Atlas[22] (TCGA) with the following criteria: (1) gene microarray or high-throughput sequencing data of RNAs extracted from primary breast tumor or adjacent normal tissues; (2) the sample size of each dataset is no less than 50; (3) clinical information were provided together with the dataset, mainly including subtypes, histological grades, pathologic stages, ER/PR/HER2 status, and prognosis; (4) the dataset was available for download before Jan 1, 2016, which was the latest date we collected the datasets. Finally, we obtained a preliminary collection of 86 independent datasets. To assure adequate specimens in subgrouping, we assessed the sample number demand and removed the datasets with less than 100 samples. The rest 30 datasets include 27 datasets from GEO (measured by Affymetrix microarray), 2 datasets from TCGA (measured by Agilent microarray and Illumina HiSeq), and 1 dataset of the METABRIC from EBI (measured by Illumina HT-12 v3 microarray).

Then we performed quality control, normalization and duplicate removing on all the 30 datasets. Quality control was carried out by *simpleaffy* and *affyPLM* R packages on each of the 27 GEO datasets independently. The raw data of each dataset were then normalized, summarized, and log-transformed using robust multi-array average (RMA) function of *affy* R package. The probe-based expression was converted into GE profiles, and the gene containing multiple probes was represented by the probe with the largest interquartile range across the samples. For the METABRIC dataset, we deleted 12 samples since 8 samples were duplicated in the discovery and validation sets and 4 were represented twice in the validation set. We used the processed expression matrix data of METABRIC directly[23]. For the TCGA Agilent and RNA-Seq data, we removed 22 samples without matched clinical information and used level_3 log2 normalized data from TCGA directly.

Furthermore, the tumor purity of the samples profiled on Affymetrix platforms was detected through a robust method, ESTIMATE, which uses the ESTIMATE-based tumor purity score developed by Affymetrix data to evaluate tumor purity[24]. This method was not applied to predict the tumor purity of the samples profiled on Affymetrix Human Genome U133B Array because of the insufficiency of the gene signatures intersection. Depending on the results of tumor purity estimation (Supplementary Figure S1), we eliminated one dataset with the lowest mean tumor purity, in order to reduce noises caused by diverse tumor purity. Finally, GE profiles of a total of 26,339 genes from 9,381 samples of 29 datasets were available for transcriptome analysis.

Sample subgrouping features.

We compiled a series of clinical features along with each sample for sample subgrouping. For samples with some clinical features that were not initially provided (mainly $ER^{-/+}$/$PR^{-/+}$/$HER2^{-/+}$, TNBC, and PAM50 subtypes) in certain collected datasets, we defined these features using a computational method based on GE profiles. Expressions of *ER*, *PR*, and *HER2* were respectively fitted by a Gaussian bimodal distribution model and the parameters were estimated via EM algorithm using *Mclust* function in *mclust* R package. The expression status for *ER*, *PR*, and *HER2* were discriminated as positive (ER^+, PR^+, and $HER2^+$) or negative (ER^-, PR^-, and $HER2^-$). On the basis of the identification of *ER*, *PR*, and *HER2* positive or negative expression status, we classified samples into TNBC or non-TNBC subtype[6]. Samples defined as ER^-, PR^- and $HER2^-$ status were identified as TNBC and otherwise non-TNBC. Molecular classification for PAM50 subtypes was provided in some datasets, and if not, we classified the patients into the five intrinsic breast cancer subtypes using the 50-gene subtype classifier, PAM50[4]. The feature of the prognosis status was classified into good/poor prognosis using the median survival time as the delimitation.

Copy number variation.

We obtained CNV data for 28,678 genes of 3,035 samples from METABRIC and TCGA. Both the DNA microarray platforms of the 2 datasets are Affymetrix Genome-Wide Human SNP Array 6.0. The numerical values of CNV were processed, summarized, and normalized relying on the relative intensity of probe hybridization on the arrays. Segmented data were converted to the gene level matrix using GISTIC 2.0[25], which were annotated for gene content based on hg19/GRCh37 for the TCGA data. For the METABRIC data, we generated a patient-by-gene CNV matrix through the processed segment data by matching the overlap of the segments with the gene regions whose annotations and coordinates were given by hg18/Ensembl 54. For more accurate and reliable analysis, we set the gain/loss threshold to 0.1 and −0.1, respectively. When the CNV value of a gene is greater than 0.1, the gene is defined as copy number gain. When the CNV value of a gene is smaller than −0.1, the gene is defined as copy number loss.

Statistical analysis.

All the statistical analysis were performed using R programming platform. An unpaired *t* test was used for differential GE analysis in Transcriptome Analysis for 2 subgroups. One-way analysis of variance (ANOVA) was used for more than 2 subgroups if GE satisfied the assumption of a normal distribution, and if not, the non-parametric test (Kruskal–Wallis test) was used to assess statistical differences among these subgroups. The *survfit* function of *survival* R package was used for survival analysis. Kaplan-Meier curves and log-rank test were used to assess survival differences. In Transcriptome Survival Analysis, we classified patients into 2 groups according to an optimal GE cutoff value based on the Cutoff Finder application[26]. This program will traverse the GE values of all patients and the optimal cutoff value can minimizes the p value of survival differences. In CNV Survival Analysis, patients are separated into 2 groups according to their CNV status (gain/loss) of the query gene. Hazard ratio (HR) was calculated using Cox proportional hazards regression model. Co-expression analysis was performed using *cox* function of *WGCNA* R package. The correlation of GE was evaluated by Pearson correlation coefficient (PCC) as well as false discovery rate (FDR) adjusted p-value. The genes with absolute PCC ≥ 0.3 and adjusted p-value ≤ 0.05 were considered co-expressed in Co-expression Analysis.

Database schema and implementation.

BCIP was implemented based on the Apache HTTP server 2.2 with MySQL 5.1.73 at the back end and the PHP 5.5.31, HTML, and JavaScript at the front end. All the computing programs were completed with R 3.2.3 and dependent packages.

Results

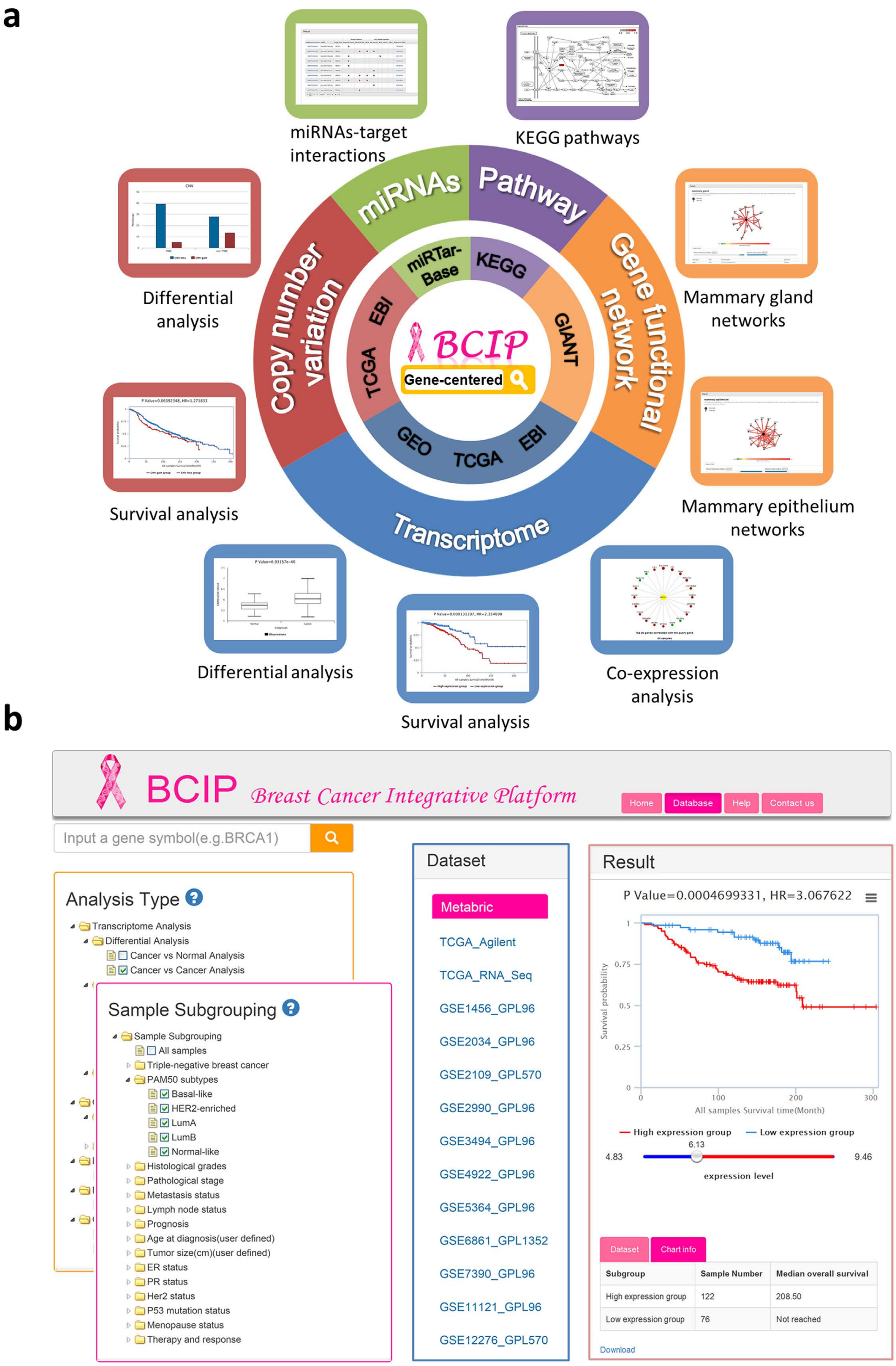
Overview of BCIP.

BCIP is a gene-centered platform that provides (1) differential expression analysis, survival analysis, and co-expression analysis based on transcriptome data; (2) differential analysis and survival analysis based on CNVs; (3) miRNA regulation analysis on miRNA-target interactions; (4) KEGG pathway analysis; and (5) network analysis on mammary tissue-specific gene function networks (Fig. 1a). BCIP provides a user-friendly interface consisting of four panels: *Analysis Type*, *Sample Subgrouping*, *Dataset*, and *Result* (Fig. 1b). A gene symbol can be input in the text field where we provide a fuzzy matching function. Users can then select any of 5 analytical categories in the *Analysis Type* panel, including Transcriptome Analysis, Copy Number Variation Analysis, MicroRNA-target Interaction Analysis, Pathway Analysis, and Gene Functional Network Analysis. After selecting analytical category, users can customize subgroups with single or combined clinical features of interest in the *Sample Subgrouping* panel. BCIP provides a total of 15 clinical features, including TNBC and non-TNBC subtypes, PAM50 subtypes, histological grades, pathologic stages, metastasis status, lymph node status, *ER*/*PR*/*HER2* status, *TP53* mutation status, menopause status, age, tumor size, therapy responses, and prognosis. The *Dataset* panel provides all of the available datasets for the selected options in the *Analysis Type* and *Sample Subgrouping*. Finally, the *Result* panel returns corresponding graphical and tabular presentation and analysis results after choosing from the above options.

Transcriptome Analysis.

We collected GE data of breast cancer tissue samples from publicly available databases of GEO, EMBL-EBI and TCGA and obtained 86 datasets. After excluding the datasets with insufficient samples (less than 100) or low tumor purity, we finally retained 29 datasets with the GE profiles of 9,381 samples (Fig. 2a and Supplementary Table S1). The GE profiles are used for differential expression analysis, survival analysis, and co-expression analysis.

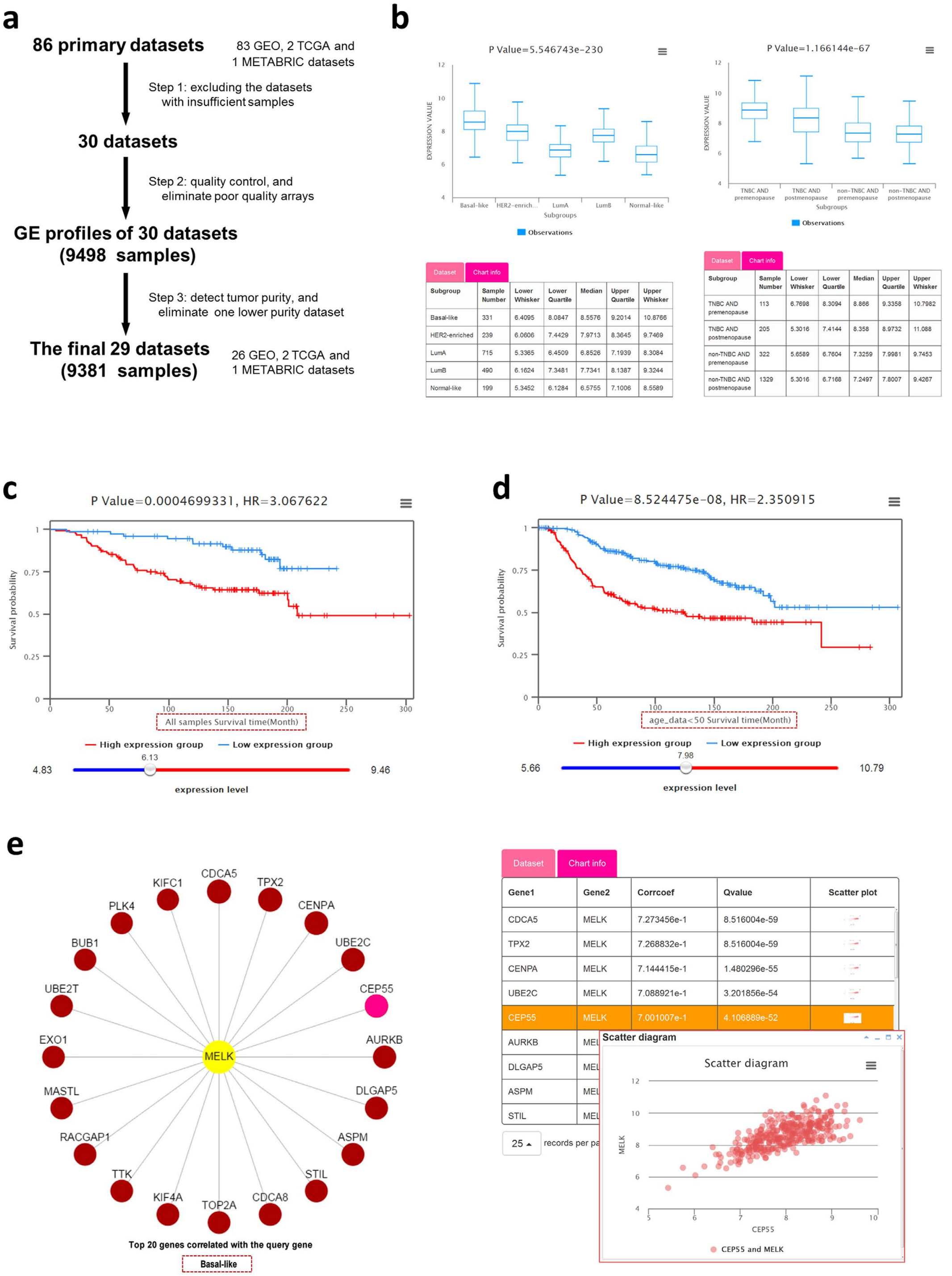
- **Differential expression analysis.** There are 2 options for differential expression analysis in *Analysis Type*: *cancer vs normal, cancer vs cancer*. The *cancer vs normal* option is designed to show the expression difference between the tumor and normal tissues. The *cancer vs cancer* option supports differential analysis in tumor samples among user-defined subgroups. Users are allowed to customize specific subgroups with single or combined clinical features of interest. Differential expression analysis results will be illustrated with a box plot in the *Result* panel. Dataset and chart information are also presented below the graph. For example, *MELK*, a recently reported oncogenic kinase in BBC[14], has the highest expression level in basal-like subtype in the METABRIC dataset (Fig. 2b, left panel). When both TNBC and non-TNBC in the Triple-negative breast cancer group and pre-menopause/post-menopause in the Menopause status group are selected, BCIP will divide samples into 4 subgroups according to the combination of 2 groups (Fig. 2b, right panel).
- **Survival analysis.** Survival analysis is provided to investigate the association of gene with clinical prognosis. BCIP offers 5 survival types, including overall survival (OS), disease-specific survival (DS), disease-free survival (DFS), recurrence-free survival (RFS), and distant metastasis-free survival (DMFS). Users can perform survival analysis in the specific subgroup customized with single or combined clinical features of interest. Patients in the specific subgroup are divided into 2 groups according to an optimal cutoff of the GE levels. The optimal cutoff is determined by the Cutoff Finder that maximizes the survival differences between 2 groups[26]. Notably, the cutoff value can be flexibly moved through a slider bar. Kaplan-Meier survival curves will be redrawn dynamically with the change of the cutoff value, together with the p-value and HR. Figure 2c and d show the overall survival analysis results of *MELK* in all patients of the GSE7390 dataset and in the samples younger than 50 years old of the METABRIC dataset, respectively.
- **Co-expression analysis.** Co-expression analysis is widely used to provide clues for potential associations among genes, proteins, and other biomolecules in various carcinomas[27–29]. Given a gene, BCIP displays the top 20 co-expression genes in the form of a circle dot where positive and negative correlations are represented by red and green, respectively (Fig. 2e). All of the co-expression genes with absolute PCC ≥ 0.3 and adjusted p-value ≤ 0.05 are presented in descending order of PCCs in the table below the graph. Scatter plots of GE levels of the co-expressed genes and the query gene are provided when clicking on the dots or the last column of the table. Notably, BCIP permits users to investigate co-expression relationship in any specified subgroup that users customized with the clinical features of interest.



Subgroup	Sample Number	Median overall survival
High expression group	122	208.50
Low expression group	76	Not reached

Figure 1. Schematic diagram showing the architecture of BCIP. (**a**) Data sources and applications of BCIP. The same color system was used to characterize the correspondence, while the inner light-colored donut chart represents the data sources and the outer deep-colored donut chart indicates the corresponding applications (analysis types). (**b**) Overview of the data portal of BCIP. The panel of 4 modules includes: *Analysis Type, Sample Subgrouping, Dataset,* and *Result*.

Copy Number Variation Analysis. CNVs exist pervasively in human genomes and contribute to the diversity and susceptibility of numerous diseases[30]. It may be an important factor in cancer occurrence and development. A series of studies and attempts have been carried out to explore the impact of CNV on breast cancer[31–33]. We collected and incorporated CNVs information of 3,035 tumor samples from METABRIC and TCGA. BCIP provides differential analysis and survival analyses for CNV data.

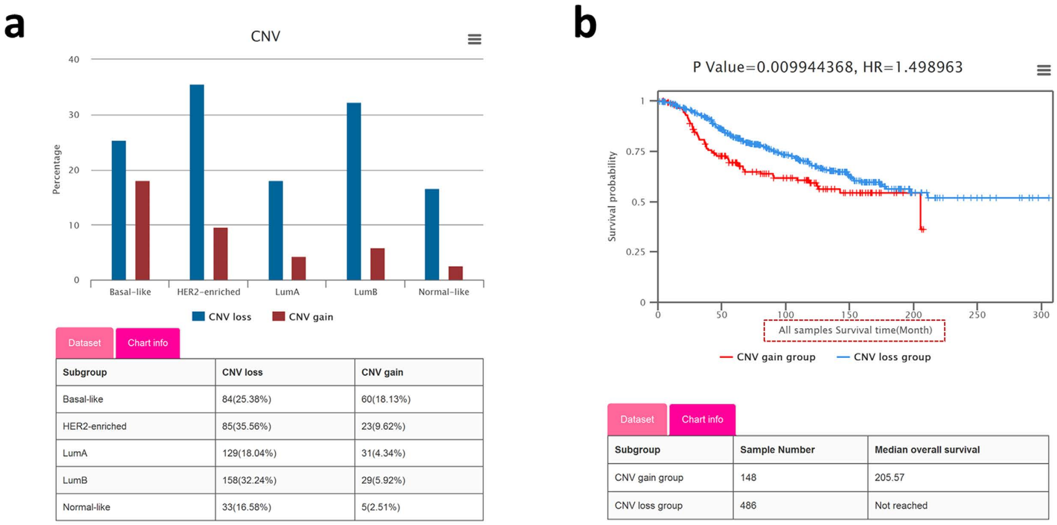


Subgroup	Sample Number	Lower Whisker	Lower Quartile	Median	Upper Quartile	Upper Whisker
Basal-like	331	6.4095	8.0847	8.5576	9.2014	10.8766
HER2-enriched	239	6.0606	7.4429	7.9713	8.3645	9.7469
LumA	715	5.3365	6.4509	6.8526	7.1939	8.3084
LumB	490	6.1624	7.3481	7.7341	8.1387	9.3244
Normal-like	199	5.3452	6.1284	6.5755	7.1006	8.5589

Subgroup	Sample Number	Lower Whisker	Lower Quartile	Median	Upper Quartile	Upper Whisker
TNBC AND premenopause	113	6.7698	8.3094	8.866	9.3358	10.7982
TNBC AND postmenopause	205	5.3016	7.4144	8.358	8.9732	11.088
non-TNBC AND premenopause	322	5.6589	6.7604	7.3259	7.9981	9.7453
non-TNBC AND postmenopause	1329	5.3016	6.7168	7.2497	7.8007	9.4267

Gene1	Gene2	Corrcoef	Qvalue	Scatter plot
CDCA5	MELK	7.273456e-1	8.516004e-59	
TPX2	MELK	7.268832e-1	8.516004e-59	
CENPA	MELK	7.144415e-1	1.480296e-55	
UBE2C	MELK	7.088921e-1	3.201856e-54	
CEP55	MELK	7.001007e-1	4.106889e-52	
AURKB	MEL			
DLGAP5	MEL			
ASPM	MEL			
STIL	MEL			

Figure 2. Visualization of transcriptome data information and analytic results. (**a**) Flow chart of the transcriptome data processing procedures. (**b**) Box plots showing differential gene expression of *MELK* in breast cancer PAM50 subtypes (left) and other subgroups (right): TNBC AND premenopause, TNBC AND postmenopause, non-TNBC AND premenopause, and non-TNBC AND postmenopause. The tables display detailed information of the dataset and each subgroup of the chart. (**c,d**) The low (blue curve) and high (red curve) levels of *MELK* expression groups are correlated with the overall survival times in all patients of the GSE7390 dataset (p-value = 0.00047, HR = 3.067622) or in the patients younger than 50 years old of the METABRIC dataset (p-value = 8.524475e-08, HR = 2.350915). (**e**) The circle dots show the top 20 genes co-expressed with *MELK* in basal-like subtype patients of the METABRIC dataset. Positive and negative correlation is respectively distinguished by red and green circles. The table lists co-expressed genes with PCC ≥ 0.3 and adjusted p-value ≤ 0.05. Scatter plot showing the expression status of *MELK* and *CEP55*.



Subgroup	CNV loss	CNV gain
Basal-like	84(25.38%)	60(18.13%)
HER2-enriched	85(35.56%)	23(9.62%)
LumA	129(18.04%)	31(4.34%)
LumB	158(32.24%)	29(5.92%)
Normal-like	33(16.58%)	5(2.51%)

Subgroup	Sample Number	Median overall survival
CNV gain group	148	205.57
CNV loss group	486	Not reached

Figure 3. Visualization of copy number variations (CNVs) based analyses reveals CNV status and survival correlation. (**a**) Histogram depicting the percentage of samples with *MELK* copy number losses/gains in each PAM50 subtype of the METABRIC dataset. (**b**) Kaplan-Meier plot showing the disease-specific survival rate comparison between patients with *MELK* copy number losses (blue curve) and copy number gains (red curve), and the tables showing the sample numbers and median survival times of the 2 groups.

- **Differential analysis between copy number gain/loss.** The proportions of samples with CNV (copy number gain and loss) of the query gene in any user-defined subgroups are respectively displayed in a histogram. A table matching the histogram provides details of the sample numbers and the proportions of copy number gain/loss in the corresponding subgroups. For example, *MELK* has a much higher (more than 2 times) proportion of samples with copy number gain in basal-like subtype than other PAM50 subtypes (Fig. 3a), which provides a clue to explain its high expression in BBC.
- **Survival analysis with CNV data.** Survival analysis based on gene CNVs will shed light on the correlation between prognostic outcomes and CNV status. BCIP provides 5 survival types (OS, DS, DFS, RFS, and DMFS), permitting users to perform analysis in specific subgroups customized with single or combined clinical features of interest. Samples are separated into 2 groups according to their CNV status (gain/loss) of the query gene. The Kaplan-Meier plot shows that the samples with *MELK* copy number gain have shorter disease-specific survival times than those with copy number loss in the METABRIC dataset (Fig. 3b).

MicroRNA-target Interaction Analysis.

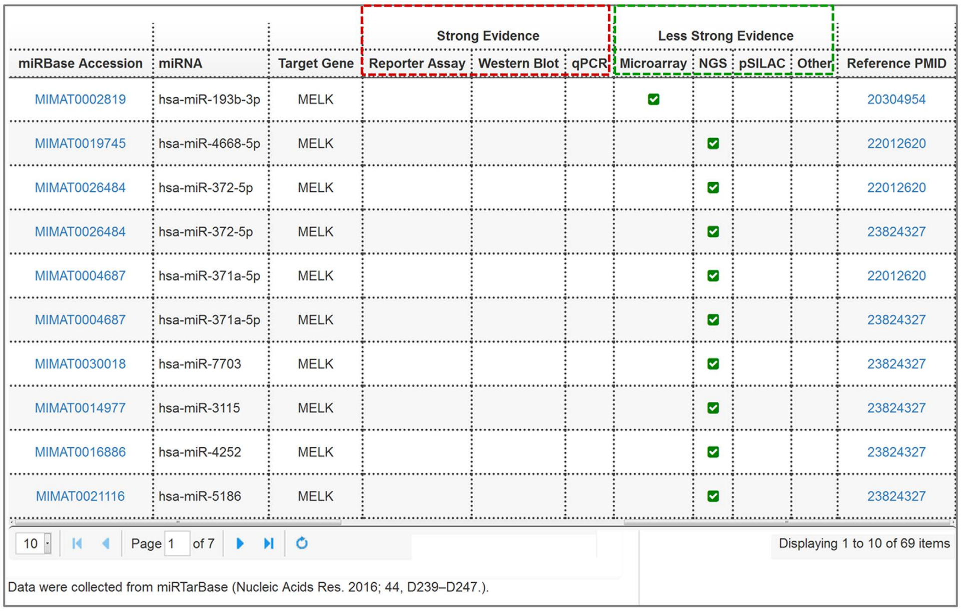
miRNAs are small non-coding RNAs that can regulate protein-coding messenger RNAs (mRNAs) at the post-transcriptional level[34]. The pivotal role of miRNAs is known as a modulator participating in various biological processes. Numerous studies suggest that dysregulation of miRNAs may contribute to the initiation and progression of cancers[35], and miRNAs can be regarded as diagnostic signatures or therapeutic biomarkers in breast cancer[36,37]. To facilitate researchers' investigations into potential regulation mechanisms of query genes, BCIP provides MicroRNA-target Interaction Analysis, which illustrates miRNAs targeting the query gene. There is a total of 324,219 miRNA-target interactions between 2,619 miRNAs and 14,884 target genes from miRTarBase[38] that are maintained in BCIP. All of these interactions are experimentally validated.

- **miRNA-target interactions.** The results for miRNA-target interactions analysis are presented as a table and list the mature miRNAs that target the input gene and the corresponding experiment types (Fig. 4). Experiment types of reporter assay, Western blot, real-time quantitative PCR (qPCR) are regarded as strong evidence, while microarray, next-generation sequencing (NGS), and pulsed stable isotope labeling by amino acids in cell culture (pSILAC) are regarded as less strong evidence for the interactions between miRNAs and genes[38]. In addition, external links to the miRBase database and PubMed have been embedded in the table for the added convenience in retrieving corresponding information.

Pathway Analysis.

It is known that some biological pathways involved in metabolism, apoptosis, and signal transduction play critical roles in cell proliferation and differentiation during tumorigenesis and cancer development. Understanding which pathways a gene participates in will be of great help for researchers in characterizing its functions in breast cancer. We have collected 22,455 linked entries between 6,755 genes and 286 human pathways from the KEGG database[39].

- **KEGG pathways.** A table consisting of pathway classes, pathway IDs, and pathway names are presented in the KEGG pathways analysis results to depict pathway information (Fig. 5). The thumbnail images in the last column can pop-up pathway maps when clicked. The pathway maps contain molecular interaction and reaction networks in which the query gene was involved. The query gene is highlighted in red.



			Strong Evidence			Less Strong Evidence				
miRBase Accession	miRNA	Target Gene	Reporter Assay	Western Blot	qPCR	Microarray	NGS	pSILAC	Other	Reference PMID
MIMAT0002819	hsa-miR-193b-3p	MELK				☑				20304954
MIMAT0019745	hsa-miR-4668-5p	MELK					☑			22012620
MIMAT0026484	hsa-miR-372-5p	MELK					☑			22012620
MIMAT0026484	hsa-miR-372-5p	MELK					☑			23824327
MIMAT0004687	hsa-miR-371a-5p	MELK					☑			22012620
MIMAT0004687	hsa-miR-371a-5p	MELK					☑			23824327
MIMAT0030018	hsa-miR-7703	MELK					☑			23824327
MIMAT0014977	hsa-miR-3115	MELK					☑			23824327
MIMAT0016886	hsa-miR-4252	MELK					☑			23824327
MIMAT0021116	hsa-miR-5186	MELK					☑			23824327

Displaying 1 to 10 of 69 items

Data were collected from miRTarBase (Nucleic Acids Res. 2016; 44, D239–D247.).

Figure 4.   The miRNA-target interactions table lists the experimentally validated mature miRNAs that target *MELK*. The red box and green box respectively represent the strong experiment evidence (reporter assay, Western blot, and qPCR) and less strong evidences (microarray and next-generation sequencing experiments).

Gene Functional Network Analyses.

Each biomolecule is located in complex biological networks and exerts its functions together with other related molecules. Notably, gene expression as well as gene-gene functional relationships in the complex biological regulation network may be tissue-specific[18,40]. Identifying a gene's functional partner in specified tissue can facilitate researchers to infer gene functions and molecular mechanisms. Here we provide mammary tissue-specific gene functional networks analysis. Data on both mammary epithelium and mammary gland gene functional networks were collected and processed from the GIANT webserver[40].

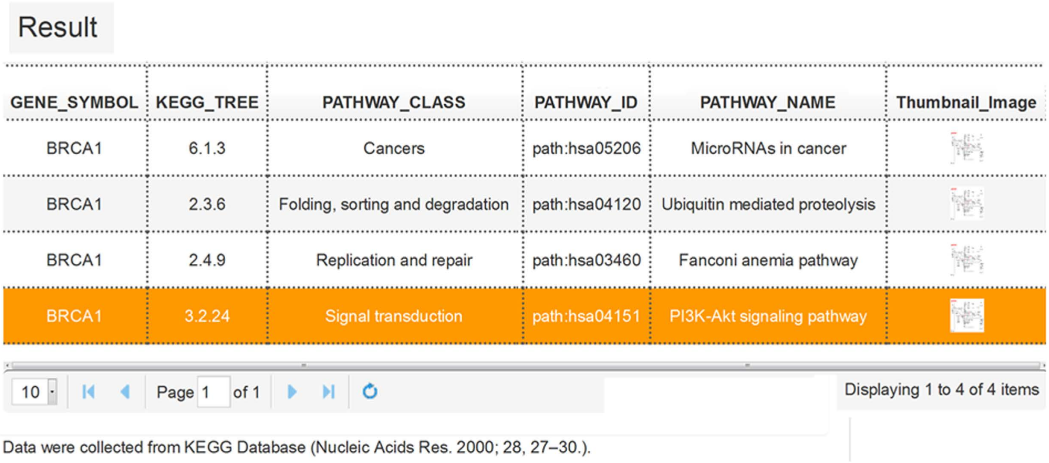
- **Mammary epithelium and gland networks.** Users are allowed to search gene functional relationships in mammary epithelium and gland gene functional networks. The dynamic network shows a subset of the entire network, and the nodes and edges are controlled by the slider bars of the maximum number of genes and the minimum relationship confidence (Fig. 6). A table below the network lists the details of the top 50 functional related genes in descending order of the average edge score, which reflects the relationship strength between the 2 genes. Clicking on any gene leads to a new table showing the corresponding detail about that gene.

Case Study and Discussion

For a query gene, BCIP helps to demonstrate its potential as a biomarker or regulatory gene in breast cancer. Take *MELK*, a promising therapeutic target of BBC reported recently, as an example to demonstrate the utility and advantage of BCIP[14]. Differential expression analysis shows that *MELK* has a much higher expression level in tumors than adjacent normal tissues across all of the available datasets (Supplementary Figure S2) and has the highest expression level in basal-like subtype among PAM50 subtypes across all of the datasets (Supplementary Figure S3). When subgroups are customized with tumor grades, we found that higher *MELK* expression level was significantly associated with higher histological grades among all of the datasets (Supplementary Figure S4). Survival analysis shows that overexpression of *MELK* is strongly correlated with poor prognosis (Supplementary Figure S5–9). These *in silico* results indicate that *MELK* might play roles in BBC.

To provide clues of the possible molecular mechanism of *MELK* in BBC, we further analyzed the co-expression genes and regulatory miRNAs of *MELK*. In basal-like subtype of the METABRIC dataset, *MELK* was significantly co-expressed with 78 genes with PCC > 0.6, including *CDCA5*, *TPX2*, and *CEP55* (Fig. 2e). Several studies have demonstrated that *TPX2* and *CEP55* are critical molecules for breast cancer migration, invasion, cell proliferation, and metastasis[41–43]. *CDCA5* has been reported to play a crucial role in human lung carcinogenesis and has the potential of being a therapeutic target for oral squamous cell carcinoma[44,45]. These results may be valuable clues for the investigation of potential function and molecular mechanism of *MELK* in breast cancer. Additionally, we found miRNAs targeting *MELK*, including hsa-miR-193b-3p and hsa-miR-372-5p (Fig. 4). Previous studies have shown that miR-193b represses cell proliferation and regulates cyclin D1 in melanoma[46], and miR-372 suppresses tumor proliferation, invasion, and migration in various tumor types[47,48]. This result indicates that *MELK* might be regulated by miR-193b-3p and miR-372-5p in breast cancer.

We have developed BCIP, a user-friendly, open-access, integrative analysis platform that integrates almost 10,000 tumor and normal tissue samples of breast cancer. It will facilitate the identification of potential biomarkers and regulatory genes in breast cancer. Compared with other bioinformatics resources and analysis tools, BCIP has 3 unique characteristics: (i) BCIP incorporates multiple analysis types, including differential expression



Result

GENE_SYMBOL	KEGG_TREE	PATHWAY_CLASS	PATHWAY_ID	PATHWAY_NAME	Thumbnail_Image
BRCA1	6.1.3	Cancers	path:hsa05206	MicroRNAs in cancer	
BRCA1	2.3.6	Folding, sorting and degradation	path:hsa04120	Ubiquitin mediated proteolysis	
BRCA1	2.4.9	Replication and repair	path:hsa03460	Fanconi anemia pathway	
BRCA1	3.2.24	Signal transduction	path:hsa04151	PI3K-Akt signaling pathway	

Displaying 1 to 4 of 4 items

Data were collected from KEGG Database (Nucleic Acids Res. 2000; 28, 27–30.).

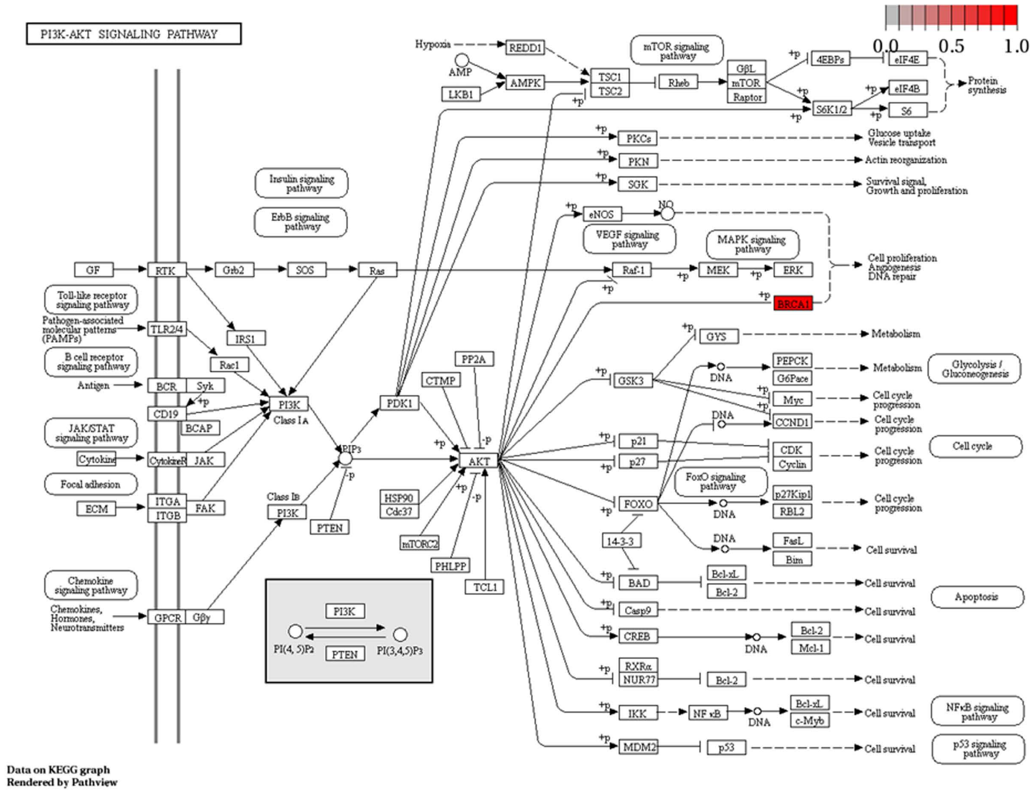
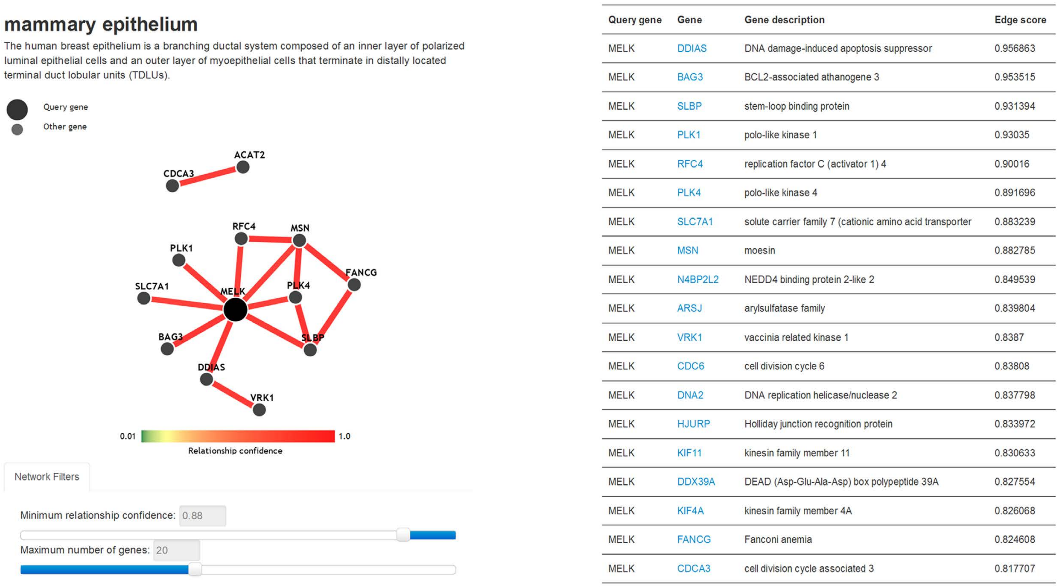


Figure 5. KEGG pathway analysis results and pathway map visualization of *BRCA1*. The table lists all of the pathways involving gene *BRCA1*, which consists of the KEGG tree, pathway class, pathway ID, and pathway name. The schematic representation showing the PI3K-Akt signaling pathway map[39], which depicts the molecular interaction and reaction network. Gene *BRCA1* is highlighted in red.

analysis, copy number variation and survival analysis, gene co-expression analysis, miRNA regulation analysis, KEGG pathways presentation, and mammary tissue-specific gene functional network analysis. All of these analysis tools help to sketch an overview of a gene in breast cancer. (ii) It provides dozens of datasets that are screened from publicly available databases, selected with strict quality control and processed with uniform normalization methods. Users can observe the consistency of the analysis results across multiple datasets, which will be helpful to evaluate the robustness of analysis results. (iii) BCIP permits users to perform analysis in specific breast cancer subgroups that are customized with single or combined clinical features of interest, including molecular subtypes, therapy response, and various clinical features.



mammary epithelium

The human breast epithelium is a branching ductal system composed of an inner layer of polarized luminal epithelial cells and an outer layer of myoepithelial cells that terminate in distally located terminal duct lobular units (TDLUs).

Minimum relationship confidence: 0.88

Maximum number of genes: 20

Query gene	Gene	Gene description	Edge score
MELK	DDIAS	DNA damage-induced apoptosis suppressor	0.956863
MELK	BAG3	BCL2-associated athanogene 3	0.953515
MELK	SLBP	stem-loop binding protein	0.931394
MELK	PLK1	polo-like kinase 1	0.93035
MELK	RFC4	replication factor C (activator 1) 4	0.90016
MELK	PLK4	polo-like kinase 4	0.891696
MELK	SLC7A1	solute carrier family 7 (cationic amino acid transporter	0.883239
MELK	MSN	moesin	0.882785
MELK	N4BP2L2	NEDD4 binding protein 2-like 2	0.849539
MELK	ARSJ	arylsulfatase family	0.839804
MELK	VRK1	vaccinia related kinase 1	0.8387
MELK	CDC6	cell division cycle 6	0.83808
MELK	DNA2	DNA replication helicase/nuclease 2	0.837798
MELK	HJURP	Holliday junction recognition protein	0.833972
MELK	KIF11	kinesin family member 11	0.830633
MELK	DDX39A	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39A	0.827554
MELK	KIF4A	kinesin family member 4A	0.826068
MELK	FANCG	Fanconi anemia	0.824608
MELK	CDCA3	cell division cycle associated 3	0.817707

Figure 6. Visualization of the mammary epithelium-specific gene functional network. The dynamic network displays the front portion of the entire network of *MELK* with 50 other genes. The number of genes in the network is controlled by the slider bar of the minimum relationship confidence and the maximum number of genes. The table lists the top 50 related genes in accordance with the priority of average edge score.

Lots of studies have been done to identify biomarkers and to uncover molecular mechanisms of tumorigenesis, cell invasion, and metastasis in breast cancer. However, many tumors with high invasion and poor outcomes, such as TNBC or basal-like tumors, still lack well-defined molecular biomarkers and therapy targets due to the high heterogeneity. BCIP serves as a convenient and efficient platform to identify biomarkers, characterize potential functions and mechanisms of genes in breast cancer. Researchers can find clues for subsequent experiments and clinical analysis. In our future work, we will continue incorporating newly available, credible data into BCIP and provide reliable supports for researchers of breast cancer.

References

1. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA. Cancer J. Clin.* **65,** 87–108 (2015).
2. Polyak, K. Heterogeneity in breast cancer. *J. Clin. Invest.* **121,** 3786–3788 (2011).
3. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* **98,** 10869–10874 (2001).
4. Parker, J. S. *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).
5. Reis-Filho, J. S. & Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet* **378,** 1812–1823 (2011).
6. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121,** 2750–2767 (2011).
7. Rivenbark, A. G., O'Connor, S. M. & Coleman, W. B. Molecular and Cellular Heterogeneity in Breast Cancer: Challenges for Personalized Medicine. *Am. J. Pathol.* **183,** 1113–1124 (2013).
8. Jin, Y. *et al.* FOXC1 is a critical mediator of EGFR function in human basal-like breast cancer. *Ann. Surg. Oncol.* **21,** 758–766 (2014).
9. Cheng, G. *et al.* Higher levels of TIMP-1 expression are associated with a poor prognosis in triple-negative breast cancer. *Mol. Cancer* **15,** 30 (2016).
10. Bastien, R. R. L. *et al.* PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. *BMC Med. Genomics* **5,** 44 (2012).
11. Perou, C. M. Molecular Stratification of Triple-Negative Breast Cancers. *The Oncologist* **16,** 61–70 (2011).
12. Abramson, V. G., Lehmann, B. D., Ballinger, T. J. & Pietenpol, J. A. Subtyping of triple-negative breast cancer: Implications for therapy. *Cancer* **121,** 8–16 (2015).
13. Wang, Y. *et al.* CDK7-Dependent Transcriptional Addiction in Triple-Negative Breast Cancer. *Cell* **163,** 174–186 (2015).
14. Wang, Y. *et al.* MELK is an oncogenic kinase essential for mitotic progression in basal-like breast cancer cells. *eLife* **3,** e01763 (2014).
15. Khaled, W. T. *et al.* BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nat. Commun.* **6,** 5987 (2015).
16. Kapushesky, M. *et al.* Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.* **38,** D690–D698 (2010).
17. Rhodes, D. R. *et al.* ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform. *Neoplasia N. Y. N* **6,** 1–6 (2004).
18. Liu, X., Yu, X., Zack, D. J., Zhu, H. & Qian, J. TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics* **9,** 271 (2008).
19. Zilliox, M. J. & Irizarry, R. A. A Gene Expression Barcode for Microarray Data. *Nat. Methods* **4,** 911–913 (2007).
20. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41,** D991–D995 (2013).
21. Lappalainen, I. *et al.* The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* **47,** 692–695 (2015).
22. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **19,** A68–A77 (2015).

23. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352 (2012).
24. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4,** 2612 (2013).
25. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
26. Budczies, J. *et al.* Cutoff Finder: A Comprehensive and Straightforward Web Application Enabling Rapid Biomarker Cutoff Optimization. *PLOS ONE* **7,** e51862 (2012).
27. Ingthorsson, S. *et al.* HER2 induced EMT and tumorigenicity in breast epithelial progenitor cells is inhibited by coexpression of EGFR. *Oncogene* **35,** 4244–4255 (2016).
28. Zhang, X.-H. *et al.* Coexpression of VEGF-C and COX-2 and its association with lymphangiogenesis in human breast cancer. *BMC Cancer* **8,** 4 (2008).
29. Ma, L. & Jiang, T. Clinical implications of Ezrin and CD44 co-expression in breast cancer. *Oncol. Rep.* **30,** 1899–1905 (2013).
30. Ueno, T. *et al.* Genome-wide copy number analysis in primary breast cancer. *Expert Opin. Ther. Targets* **16,** S31–S35 (2012).
31. Krepischi, A. C. *et al.* Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res*. **14,** R24 (2012).
32. Masson, A. L. *et al.* Expanding the genetic basis of copy number variation in familial breast cancer. *Hered. Cancer Clin. Pract*. **12,** 15 (2014).
33. Suehiro, Y. *et al.* Germline copy number variations associated with breast cancer susceptibility in a Japanese population. *Tumour Biol.* **34,** 947–952 (2013).
34. Ambros, V. The functions of animal microRNAs. *Nature* **431,** 350–355 (2004).
35. van Schooneveld, E. *et al.* Dysregulation of microRNAs in breast cancer and their potential role as prognostic and predictive biomarkers in patient management. *Breast Cancer Res.* **17,** 21 (2015).
36. Tashkandi, H., Shah, N., Patel, Y. & Chen, H. Identification of new miRNA biomarkers associated with HER2-positive breast cancers. *Oncoscience* **2,** 924–929 (2015).
37. Bertoli, G., Cava, C. & Castiglioni, I. MicroRNAs: New Biomarkers for Diagnosis, Prognosis, Therapy Prediction and Therapeutic Tools for Breast Cancer. *Theranostics* **5,** 1122–1143 (2015).
38. Chou, C.-H. *et al.* miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* **44,** D239–D247 (2016).
39. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. **28,** 27–30 (2000).
40. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47,** 569–576 (2015).
41. Yang, Y. *et al.* TPX2 promotes migration and invasion of human breast cancer cells. *Asian Pac. J. Trop. Med*. **8,** 1064–1070 (2015).
42. Wang, Y., Jin, T., Dai, X. & Xu, J. Lentivirus-mediated knockdown of CEP55 suppresses cell proliferation of breast cancer cells. *Biosci. Trends* **10,** 67–73 (2016).
43. Inoda, S. *et al.* Cep55/c10orf3, a tumor antigen derived from a centrosome residing protein in breast carcinoma. *J. Immunother.* **32,** 474–485 (2009).
44. Nguyen, M.-H. *et al.* Phosphorylation and activation of cell division cycle associated 5 by mitogen-activated protein kinase play a crucial role in human lung carcinogenesis. *Cancer Res.* **70,** 5337–5347 (2010).
45. Tokuzen, N., Nakashiro, K., Tanaka, H., Iwamoto, K. & Hamakawa, H. Therapeutic potential of targeting cell division cycle associated 5 for oral squamous cell carcinoma. *Oncotarget* **7,** 2343–2353 (2015).
46. Chen, J. *et al.* MicroRNA-193b Represses Cell Proliferation and Regulates Cyclin D1 in Melanoma. *Am. J. Pathol.* **176,** 2520–2529 (2010).
47. Huang, X., Huang, M., Kong, L. & Li, Y. miR-372 suppresses tumour proliferation and invasion by targeting IGF2BP1 in renal cell carcinoma. *Cell Prolif.* **48,** 593–599 (2015).
48. Wu, G. *et al.* Mir-372 down-regulates the oncogene ATAD2 to influence hepatocellular carcinoma proliferation and metastasis. *BMC Cancer* **14,** 107 (2014).

Acknowledgements

This work was supported by a grant from China National High Technology Research and Development Program 2014AA020604 (X. Ying). The authors gratefully acknowledge the technical assistance for platform construction from Mingming Su, Honglei Li, Jiemin Zou, and Xinlei Zhang. This study makes use of data generated by Molecular Taxonomy of Breast Cancer International Consortium, the authors greatly thank to the METABRIC group for the approval of the data usage. The authors gratefully acknowledge the copyright permission of KEGG for pathway map image used in our article.

Author Contributions

J.W. and S.H. carried out the majority of the data analysis and platform construction, participated in the platform and web page design, and drafted the manuscript. Y.C. participated in providing technical assistance to platform construction. Z.L. and J.Z. provided guidance for data processing procedures, and participated in part of data analysis. H.Y. and Q.S. contributed to the data collection and arrangement from publicly available databases. N.S. provided guidance to the study. X.Y. conceived the study and platform construction, contributed to the web page design, and revised the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: The authors declare no competing financial interests.

How to cite this article: Wu, J. *et al.* BCIP: a gene-centered platform for identifying potential regulatory genes in breast cancer. *Sci. Rep.* **7**, 45235; doi: 10.1038/srep45235 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.