Identification of Four Hub Genes Involved in Breast Cancer Based on Robust Rank Aggregation and WGCNA Methods

Rongqin Ke (rke@hqu.edu.cn)
Huaqiao University - Quanzhou Campus: Huaqiao University

Jinbao Yin
Huaqiao University - Quanzhou Campus: Huaqiao University

Primary research

Keywords: breast cancer, RRA, WGCNA, hubgenes

DOI: https://doi.org/10.21203/rs.3.rs-70354/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Background: Further elucidation of the molecular mechanisms of the occurrence, development and prognosis of breast cancer remains an urgent need. Identifying hub genes involved in these pathogenesis and progression can potentially help to unveil these mechanisms and provide novel therapeutic targets for breast cancer.

Methods: In this study, we systematically integrated robust rank aggregation (RRA), functional enrichment analysis, protein-protein interaction (PPI) networks construction and analysis, weighted gene co-expression network analysis (WGCNA), DNA methylation analyses and genomic mutation analyses, GSEA and GSVA to identify potential hub genes that are highly associated with breast cancer.

Results: We identified a total of 512 robust DEGs that were significantly associated with breast cancer based on RRA analysis and functional enrichment analysis. CENPL, ISG20L2, MRPL3 and LSM4 were identified as four potential hub genes for breast cancer through the WGCNA analysis and literate search. These four hub genes were upregulated in breast cancer tissues and associated with tumor progression. ROC and Kaplan-Meier indicated these four hub genes all showed good diagnostic performance and prognostic values for breast cancer. Methylation analyses and genomic mutation analyses suggested that the abnormal up-regulation of these genes are likely resulted from hypomethylation and gene mutations. Moreover, GSEA and GSVA for single potential hub genes revealed they were all tightly related to the proliferation of tumor cells.

Conclusion: We identify four genes (CENPL, ISG20L2, MRPL3, and LSM4) that are likely playing key roles in the molecular mechanism of occurrence and development of breast cancer. They may become potential therapeutic targets for breast cancer patients with further studies. Keywords: breast cancer, RRA, WGCNA, hub genes

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the latest manuscript can be downloaded and accessed as a PDF.

Tables

Due to technical limitations, tables is only available as a download in the Supplemental Files section.

Figures
Figure 1

Workflow of our study. Gene expression profiles of eight breast cancer related datasets were downloaded from the GEO database and subjected for differential expression analysis. RRA algorithm was applied to integrate the DEGs in these eight datasets to search for robust DEGs. WGCNA was used to identify hub genes associated with breast cancer in TCGA_BRCA dataset. Subsequently, novel key genes were validated based on multiple datasets and databases. Moreover, GSEA and GSVA for single hub genes were performed to reveal their potential biological functions and mechanism in breast cancer based on METABRIC dataset.
Figure 2

Identification of robust DEGs between breast cancer and normal breast tissue in 8 datasets downloaded GEO database based on RRA analysis. The Heatmap shows the 20 most significant up-regulated genes and down-regulated genes according to adjusted P values. Each row in the figure represents one gene and each column is one dataset. Red shows up-regulation and green signifies down-regulation. The numbers in the heatmap squares indicates fold changes (breast cancer.Vs.Normal breast tissue) in each data set that conducted in the “limma” R package.
GO enrichment analysis and KEGG pathways analysis of 512 DEGs. a GO terms of biological process (BP). b GO terms of cellular component (CC). c GO terms of molecular function (MF). d KEGG pathways.
Figure 4

Identification of candidate gene module and 102 hub genes for breast cancer based on TCGA_BRCA dataset through WGCNA. a Left, analysis of the scale-free fitting indices for various soft-thres holding powers(β), red line indicated Scale Free Topology Model Fit, signed R^2 is 0.90. Right, mean connectivity analysis of various soft-thres holding powers (β value range 1-20). b Left, his to gram shows the frequency distribution of the k(namely connection)when β =5.Right, checking the scale-free topology when β = 5, the figure shows that log10(k) and log10(p(k)) are negatively correlated (correlation coefficient 0.97), denoting that the gene scale-free network that we constructed is guaranteed. c Clustering dendro grams of genes based on dissimilarity to pological over lap calculation formula(1 - TOM) and merged gene set modules. Seven weighted gene co-expression network modules were constructed and shown in different colors. d Heat map of the correlation between module eigengenes and breast cancer samples traits (Tumor). The numbers in each square of heatmap indicates the Pearson correlation coefficient (up)and P value(down). e Scatter plot of gene significance for “Tumor” and module membership in the blue module. f Scatter plot of gene significance for “Tumor” and module membership in the brown module.
Figure 5

Correlation analysis of four novel key genes with clinicopathological variables in breast cancer based on METABRIC data set. a CENPL, b ISG20L2, c LSM4 and d MRPL3. abs(correlation): absolute value of the Spearman correlation coefficient.

Figure 6
Methylation level analyses and genetic alterations of novel hub genes for breast cancer. The methylation levels of CENPL, ISG20L2, LSM4, and MRPL3 in breast cancer and normal tissues were examined using DiseaseMeth 2.0 database based on 450k (Illumina Infinium Human Methylation 450 Bead Chip) platform. eGeneticalterations of CENPL, ISG20L2, MRPL3, and LSM4 were further examined in cBiportal database, four hub genes were altered in 570 (26%) of 2173 breast cancer patients, and CENPL and ISG20L2 altered the most (20%) with gene amplification as the main alteration type.

Figure 7

The diagnostic value analysis and validation of four novel hub genes in breast cancer. ROC curves analysis for CENPL, ISG20L2, LSM4 and MRPL3 based on: a TCGA dataset, b GEO dataset. ROC, receiver operating characteristic; AUC, area under the ROC curve.

Figure 8

The prognostic value analysis of four novel hub genes in breast cancer based on METABRIC dataset (a-d) and TCGA_BRCA dataset (e-h), Expression levels of CENPL, ISG20L2, LSM4 and MRPL3 are significantly associated with the OS of patients in breast cancer (all P<0.05, HR<1).
Figure 9

Gene set enrichment analysis (GSEA) of potential hub genes in the METABRC dataset. Tumor cell proliferation related gene-sets were significantly enriched in the high-expression group of each hubgene. a CENPL, b ISG20L2, c LSM4 and d MRPL3.
Figure 10

Clustering heatmaps of differentially expressed pathways derived from GSVA of each hub genes in the METABRIC dataset. a CENPL, b ISG20L2, c LSM4 and d MRPL3.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table.pdf
- SupplementaryMaterial.pdf