Screening of breast cancer biomarkers based on bioinformatics analysis

Yuehong Xu
Chongqing Medical University
Changchun Niu (bright_star2000@163.com)
UCAS Chongqing Hospital

Article

Keywords:

Posted Date: May 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1640218/v1

Abstract

Breast cancer is one of the most common malignancies in women, its incidence is increasing year by year, and it seriously threatens the life and health of women. Therefore, it is necessary to find more sensitive and specific biomarkers to reveal the initial disorders and underlying mechanisms of the disease. Bioinformatics analysis of breast cancer gene chip data was performed to find expression characteristic gene profiles to screen potential biomarkers of breast cancer. GEO 2R was used for differential gene expression analysis of breast cancer chips screened in the comprehensive gene expression database, R language was used for gene function annotation GO enrichment analysis and KEGG pathway enrichment analysis, and String database was used for protein-protein interactions analysis .Finally, survival analysis was performed on the five most significantly different genes through the GEPIA online analysis website, and the differential gene expression levels in normal tissue and breast cancer tissue were verified by boxplot. Through differential analysis of cancer samples from breast cancer patients and normal breast samples, a total of 1174 differential genes were obtained, including 973 down-regulated genes and 201 up-regulated genes. The differentially expressed genes were enriched in 56 different GO subsets. By using the GEPIA online analysis website to conduct survival analysis on the five most significantly differentially expressed genes, we found that the most significantly differentially expressed gene HEPN1 has a significant relationship with the prognosis of patients. The expression levels of five genes, HEPN1, GPD1, C14orf180, TUSC5 and PLIN4 in normal breast tissue and breast cancer tissue were verified by boxplots. The boxplots showed that HEPN1, GPD1, C14orf180, TUSC5 and the expression level of PLIN4 gene in breast cancer tissue was significantly lower than that in normal breast tissue. It indicates that HEPN1, GPD1, C14orf180, TUSC5 and PLIN4 genes are expected to be potential diagnostic biomarkers for breast cancer patients. However, additional research is required to demonstrate our findings and motivate the clinical importance of HEPN1 in breast cancer.

Introduction

Breast cancer is the malignant tumor with the highest incidence in women worldwide, posing a serious threat to women's life and health, with 41,760 deaths reported in 2019 [1, 2]. Since the 21st century, the incidence of breast cancer has remained high, showing an increasing trend, and the incidence of the population is younger [3]. For breast cancer patients, there are various treatments such as surgery, radiotherapy, endocrine therapy, chemotherapy, and targeted therapy, but 40% of breast cancer patients experience tumor recurrence, and 60%-70% of them have distant metastasis[4, 5]. Drug resistance, recurrence and metastasis of breast cancer are still the main reasons for treatment failure in most patients. Due to the extensive existence of individual differences, it is difficult to predict the response of different patients to treatment, causing patients to bear a large number of unnecessary adverse drug reactions and a huge economic burden. However, its exact etiology and pathogenesis remain unclear. Therefore, finding more sensitive and specific markers to reveal the initial dysregulation and underlying mechanisms of the disease is an important issue to be solved in current breast cancer research.

With the development of bioinformatics and the generation of big data such as genomics and transcriptomics, the use of bioinformatics and computer science methods to analyze these data to study the relationship between multiple biomolecules has become one of the important research methods to elucidate the mechanism of disease and predict therapeutic targets. Compared with traditional breast cancer diagnostic methods, molecular markers have unique advantages, such as strong specificity and easy dynamic monitoring[6]. In this study, differentially expressed genes (DEGs) were identified using gene chips of breast cancer in the gene expression omnibus (GEO) database, and were systematically analyzed using KEGG, String database and R language program package. The functions of differential genes and their roles in disease-related signaling pathways provide ideas for the identification of novel diagnostic biomarkers for breast cancer and to reveal the initial dysregulation and underlying mechanisms of breast cancer.

Materials And Methods

Data acquisition and preprocessing. The GEO database of the National Center for Biosciences (http://www.ncbi.nlm.nih.gov/geo) is an open access data platform. It contains the largest collection of microarray and high-throughput sequencing gene expression profiling data to date. Using "breast cancer" as the search term, enter the GEO database to search for the published breast cancer gene chip dataset, and obtain the dataset GSE29431. Using the Affymetrix Human Genome U133 Plus 2.0 Array platform, breast cancer tissues were used as samples, 54 patients with breast cancer were enrolled, and 12 normal tissues were used as controls. The chip data of each sample is preprocessed, such as completion of missing values or taking the mean value when the gene corresponds to multiple probes. Background correction, normalization, and summarization were then performed.

Analysis of differential genes. Gene differential expression analysis was performed using R language limma package (limma package is an R language package based on bio-conductor specially used for processing expression profile chip data). The thresholds for differential gene identification were taken as $p < 0.05$ and $|\log2fold(FC)| \geq 1$.

GO and KEGG enrichment analysis. We performed GO and KEGG enrichment analysis on differential genes using the GO enrichment and KEGG enrichment functions of the Cluster Profiler package (http://bioconductor.org/packages/release/bioc/html/clusterProfer.html) in Bioconductor. GO functional enrichment analysis was carried out from three aspects, including biological process (BP), cellular component (CC) and molecular function (MF). And select the top 20 significantly enriched pathways for visualization through the R language ggplot2 package.

Protein-protein interaction network analysis. The String database (http://string-db.org) was used to evaluate protein-protein interactions (PPIs) in functional protein association networks. The 1174 differentially expressed genes were imported into the String database, protein interaction analysis was performed, and a PPI network was constructed. The lines between nodes represent the interaction relationship between proteins.

Survival analysis. Survival analysis was performed on the five most significantly differentially expressed genes (HEPN1, GPD1, C14orf180, TUSC5 and PLIN4) by using the GEPIA online analysis website.

Validation of differential gene expression levels. We show the expression levels of HEPN1, GPD1, C14orf180, TUSC5 and PLIN4 genes in normal and breast cancer tissues by boxplots.

Results

Analysis of differential genes. We downloaded the gene expression profiles of GSE29431 from the GEO database, including 2 groups (54 breast cancer patients and 12 healthy controls). Subsequently, we identified 1174 differential genes from the GSE29431 dataset, of which 201 genes were up-regulated and 973 genes were down-regulated. We selected the top 50 differentially expressed genes to draw a heat map (as shown in Figure 1). The differential expression analysis test data of the top 50 genes are shown in Table 1. The five genes with the most significant differences included HEPN1, GPD1, C14orf180, TUSC5 and PLIN4.

Table 1. Analysis of differential expression of genes (top 50) in breast cancer patients and controls

ID	logFC	t	P.Value	adj.P.Val
HEPN1	-3.563188222	-22.13471403	1.55E-41	3.37E-37
GPD1	-2.713780204	-23.08382717	1.56E-33	1.66E-29
C14orf180	-3.374748954	-22.93544532	2.29E-33	1.66E-29
TUSC5	-3.454638472	-22.37055701	1.00E-32	5.45E-29
PLIN4	-5.182630648	-22.28482354	1.26E-32	5.47E-29
CA4	-3.299847481	-31.20950032	1.87E-32	6.79E-29
ITGA7	-3.314514472	-21.83632669	4.16E-32	1.29E-28
S100B	-2.331809407	-21.43467393	1.23E-31	3.35E-28
KANK3	-3.211710907	-20.69160424	9.58E-31	2.32E-27
PPP1R1A	-3.852232407	-20.36115288	2.43E-30	5.28E-27
LOC101930114	-3.580563389	-19.22917812	6.39E-29	1.19E-25
LGALS12	-3.334919046	-19.21897031	6.58E-29	1.19E-25
TIMP4	-5.414165648	-18.93437471	1.53E-28	2.56E-25
NPR1	-2.428254537	-18.77385175	2.47E-28	3.84E-25
MRAP	-2.241332148	-18.228354	1.29E-27	1.87E-24
GLYAT	-3.212384722	-18.00599469	2.55E-27	3.47E-24
LOC101926960	-3.549222685	-17.87255737	3.85E-27	4.93E-24
LVRN	-3.797365787	-17.74402506	5.75E-27	6.94E-24
PDE2A	-3.161842944	-17.43208876	1.53E-26	1.75E-23
HSPB7	-1.831890639	-17.34736685	1.99E-26	2.17E-23
RBP4	-4.290525972	-17.1636739	3.57E-26	3.70E-23
CCDC69	-3.058627231	-16.50641612	2.96E-25	2.93E-22
BHMT2	-2.185756444	-16.36903537	4.64E-25	4.39E-22
ITIH5	-3.955621991	-16.34871543	4.96E-25	4.50E-22
LOC284825	-2.863705259	-16.27938304	6.23E-25	5.42E-22
TNMD	-4.397284435	-16.06722762	1.25E-24	1.05E-21
CRYAB	-3.924310676	-16.04821999	1.34E-24	1.06E-21
DGAT2	-4.714381435	-16.04294971	1.36E-24	1.06E-21
ATOH8	-1.282317713	-16.01662954	1.48E-24	1.11E-21

SLC19A3	-4.344914898	-15.98273405	1.66E-24	1.20E-21
PPARG	-4.184964389	-15.92893918	1.99E-24	1.39E-21
FHL1	-3.669314611	-15.91549098	2.08E-24	1.41E-21
DEFB132	-4.285100028	-15.85294922	2.56E-24	1.69E-21
SLC7A10	-2.34257113	-15.67744214	4.61E-24	2.95E-21
GPIHBP1	-3.056934028	-15.63768802	5.26E-24	3.27E-21
GPR146	-3.201143694	-15.5722904	6.56E-24	3.97E-21
SGCG	-4.21236887	-15.42520563	1.08E-23	6.35E-21
ALDH1L1	-1.489132056	-15.37155843	1.30E-23	7.42E-21
ANGPTL8	-2.107141019	-15.30395015	1.63E-23	9.09E-21
PEAR1	-2.284923824	-15.25421788	1.93E-23	1.05E-20
AIFM2	-2.762923602	-15.07341198	3.59E-23	1.90E-20
COPG2IT1	-3.603184907	-15.0071905	4.51E-23	2.33E-20
CIDEA	-1.822927454	-14.97614279	5.02E-23	2.54E-20
SYN2	-1.466667389	-14.94006728	5.68E-23	2.81E-20
ACSM5	-2.329249593	-14.8459675	7.87E-23	3.80E-20
KLB	-4.562398463	-14.78047133	9.88E-23	4.67E-20
MYOM1	-3.097678028	-14.70083207	1.30E-22	6.03E-20
LOC102723493	-1.891032102	-14.69012763	1.35E-22	6.13E-20
TMEM37	-1.741200139	-14.60747178	1.80E-22	8.01E-20
PCK1	-4.59206037	-14.3630843	4.26E-22	1.85E-19

GO enrichment analysis of differentially expressed genes. The results of GO analysis showed that the differentially expressed genes were enriched into 56 different GO subsets. The most significantly enriched subsets of each were extracellular structure organization, extracellular matrix, and cofactor binding (Figure 2). The GO subsets enriched in the top 20 are shown in Table 2.

Figure 2. GO enrichment analysis “Count” is the number of genes enriched, and “p. value.adjust” is the corrected P value. **(a)** Biological processes (**b)** Cellular components **(c)** Molecular functions

Table 2. GO enrichment analysis (top 20)

Term	Count	P-Value	Category
extracellular matrix	86	3.16E-15	CC
collagen-containing extracellular matrix	76	1.83E-14	CC
extracellular structure organization	70	2.66E-11	BP
multicellular organismal homeostasis	68	9.65E-07	BP
urogenital system development	64	7.54E-12	BP
regulation of lipid metabolic process	63	1.44E-07	BP
regulation of vasculature development	63	1.03E-06	BP
renal system development	61	3.83E-12	BP
regulation of angiogenesis	58	1.54E-06	BP
cell-cell junction	58	1.33E-05	CC
extracellular matrix organization	57	5.23E-08	BP
adherens junction	57	0.0016235	CC
response to acid chemical	56	8.08E-08	BP
kidney development	54	8.68E-10	BP
fatty acid metabolic process	54	1.03E-06	BP
cofactor binding	54	0.002207086	MF
actin cytoskeleton	51	0.0016235	CC
cell adhesion molecule binding	51	0.014555902	MF
apical part of cell	45	0.0016235	CC
organic acid catabolic process	42	8.24E-06	BP

“Term” is the name of the GO subset, “Count” is the number of enriched genes, “P-Value” is the P value, and “Category” is the type of subset belonging

KEGG pathway enrichment analysis of differentially expressed genes. The differentially expressed genes were enriched in signaling pathways such as PI3K-Akt, Focal adhesion, and proteoglycan in cancer (Table 3), with statistical significance. For the above KEGG pathways that meet the requirements, the R language is used for partial visualization (Figure 3).

Table 3. KEGG pathway enrichment analysis of differentially expressed genes

Term	Count	P-Value
PI3K-Akt signaling pathway	46	0.008276276
Focal adhesion	31	0.006678947
Proteoglycans in cancer	31	0.007120448
ECM-receptor interaction	25	1.70E-06
PPAR signaling pathway	22	4.86E-06
Carbon metabolism	22	0.005118641
AMPK signaling pathway	21	0.008276276
Relaxin signaling pathway	21	0.016820746
Insulin resistance	19	0.012446009
AGE-RAGE signaling pathway in diabetic complications	18	0.012446009
Regulation of lipolysis in adipocytes	16	0.000392322
Glycerolipid metabolism	14	0.006678947
Pyruvate metabolism	13	0.002854637
Malaria	12	0.008276276
Propanoate metabolism	10	0.006678947
Fatty acid degradation	10	0.025282878

“Term” is the name of the KEGG pathway, “Count” is the number of genes, and “P-Value” is the P value

Protein-protein interaction analysis. The 1174 differentially expressed genes were imported into the String database, and protein-protein interactions (PPIs) were analyzed, and a PPIs network was constructed. The lines between nodes represent the interaction relationship between proteins. (Figure4)

Survival analysis. By using the GEPIA online analysis website, we performed a survival analysis of the five most significantly different genes, HEPN1, GPD1, C14orf180, TUSC5 and PLIN4, and the results are shown in Figure 5. The abscissa in the figure represents the survival time, the ordinate represents the overall survival rate, the red indicates the overall survival rate when the gene is highly expressed, and the blue indicates the overall survival rate when the gene expression is low. The results showed that the decrease in the expression of gene HEPN1 significantly decreased the overall survival rate of breast cancer patients ($P<0.05$), and the expression changes of other genes had no significant effect on the overall survival rate of breast cancer patients. Therefore, it is speculated that the low expression of HEPN1 may play an important role in the prognosis and development of breast cancer patients. The

higher the expression of HEPN1 gene, the better the prognosis of patients. It indicates that the HEPN1 gene may be a potential biomarker for predicting the prognosis of breast cancer patients.

Validation of differential gene expression levels. We showed the expression levels of 5 differential genes in normal and breast cancer tissues through boxplots (Figure 6). It can be seen from Figure 6 that the expression levels of HEPN1, GPD1, C14orf180, TUSC5 and PLIN4 genes in breast cancer tissues were significantly lower than those in normal tissues. These five differential genes are expected to be potential biomarkers of breast cancer.

Discussion

Breast cancer is one of the most common malignant tumors in women, and its incidence is increasing year by year and tends to be younger, and it is a serious threat to women's life and health. Therefore, finding more sensitive and specific markers to reveal the initial disorders and underlying mechanisms of the disease is currently Important issues to be addressed in breast cancer research. In this study, we used bioinformatics analysis of breast cancer gene chip data to find expression characteristic gene profiles.

Through the differential analysis of breast cancer samples and normal samples, a total of 1174 differential genes were obtained, including 973 down-regulated genes and 201 up-regulated genes. This further shows that the pathogenesis of breast cancer is extremely complex, which is the result of the interaction of many genes and/or protein molecules. GO enrichment analysis showed that the differential genes mainly included BP, CC, MF involved in extracellular matrix, extracellular structure organization, homeostasis of multicellular organisms, regulation of lipid metabolism process, regulation of vascular development, cell-cell connection, etc. Among them, the most significantly enriched subsets of BP, CC and MF were extracellular structure organization, extracellular matrix and cofactor binding, respectively. It can be seen that these enriched genes or protein molecules play an important role in the occurrence and development of breast cancer. The KEGG pathway enrichment analysis found that the differential genes were mainly enriched in PI3K-Akt, Focal adhesion, proteoglycan in cancer and other signaling pathways, and the most significant enriched pathway was the PI3K-Akt pathway.

By using the GEPIA online analysis website, we found that the gene HEPN1, has a significant relationship with the prognosis of patients. HEPN1, full name hepatocellular carcinoma down-regulated 1, is expressed in the liver and encodes a short peptide mainly confined to the cytoplasm. Transient transfection studies have shown that expression of this gene significantly inhibits cell growth and may play a role in apoptosis. The expression of this gene is down-regulated or lost in hepatocellular carcinoma (HCC), suggesting that the loss of this gene is involved in the carcinogenesis of hepatocytes. Relevant studies have shown that the silencing of HEPN1 is related to the aggressive biological behavior of hepatocellular carcinoma and pituitary growth hormone adenomas[7, 8].

Through boxplots, we verified the expression levels of the five most significantly different genes, HEPN1, GPD1, C14orf180, TUSC5 and PLIN4, and found that the expression levels of these genes in breast cancer patients were significantly lower than those in normal controls, indicating that these genes may be a

diagnostic biomarker for breast cancer. GPD1, the full name of glycerol-3-phosphate dehydrogenase 1, encodes a member of the NAD-dependent glycerol-3-phosphate dehydrogenase family and plays a key role in carbohydrate and lipid metabolism. Studies have shown that GPD1 is down-regulated in breast cancer tissues compared with normal tissues, and GPD1 protein may be a potential protein biomarker for predicting breast cancer[9]. In addition, studies have shown that GPD1 overexpression can enhance the anticancer activity of metformin, and patients with increased GPD1 expression in tumor cells may respond better to metformin treatment[10]. The full name of C14orf180 gene is chromosome 14 open reading frame 180, and the protein it encodes may be a component of membrane. Studies have shown that C14orf180 is a novel regulator of lipid storage and possible differentiation of adipocytes[11]. The full name of TUSC5 gene is tumor suppressor candidate 5. Studies have shown that this gene is related to the progression of liver cancer[12], colorectal cancer [13] and breast cancer[14]. The full name of PLIN4 gene is perilipin 4, which is mainly expressed in adipose tissue. Some studies have shown that compared with normal breast tissue, the expression level of PLIN4 gene in breast cancer tissue is lower, and the expression of other genes in the PLIN family except PLIN3 is also significantly reduced in breast cancer tissue. The study also found that high expression of PLIN1 may predict longer overall survival in breast cancer patients, while overexpression of PLIN2 indicates poorer overall survival in breast cancer patients[15].

In conclusion, the five most significantly different genes HEPN1, GPD1, C14orf180, TUSC5 and PLIN4 screened by microarray data analysis of breast cancer in the comprehensive gene expression database in this study may be potential diagnostic biomarkers for breast cancer. In addition, the differentially expressed gene HEPN1 may be a potential biomarker for predicting the prognosis of breast cancer patients. There are some shortcomings in this study. Firstly, the data set we analyzed is a single data set, and the sample size involved is small; secondly, we have not performed clinical verification on the obtained differential genes. Additional research is required to demonstrate our findings and motivate the clinical importance of HEPN1 in breast cancer.

Declarations

Data availability

The datasets analyzed during the current study are available in the GEO repository, [http://www.ncbi.nlm.nih.gov/geo].

Data availability

The datasets analyzed during the current study are available in the GEO repository, [http://www.ncbi.nlm.nih.gov/geo].

References

1. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
2. DeSantis, C.E., et al., *Breast cancer statistics, 2019.* CA Cancer J Clin, 2019. **69**(6): p. 438-451.
3. Chen, W., et al., *Cancer statistics in China, 2015.* CA Cancer J Clin, 2016. **66**(2): p. 115-32.
4. Telang, N., *Putative cancer-initiating stem cells in cell culture models for molecular subtypes of clinical breast cancer.* Oncol Lett, 2015. **10**(6): p. 3840-3846.
5. Zhang, J., et al., *MicroRNA-138 modulates metastasis and EMT in breast cancer cells by targeting vimentin.* Biomed Pharmacother, 2016. **77**: p. 135-41.
6. Shimomura, A., et al., *Novel combination of serum microRNA for detecting breast cancer in the early stage.* Cancer Sci, 2016. **107**(3): p. 326-34.
7. Peng, H., et al., *Silencing of HEPN1 is responsible for the aggressive biological behavior of pituitary somatotroph adenomas.* Cell Physiol Biochem, 2013. **31**(2-3): p. 379-88.
8. Moh, M.C., et al., *HEPN1, a novel gene that is frequently down-regulated in hepatocellular carcinoma, suppresses cell growth and induces apoptosis in HepG2 cells.* J Hepatol, 2003. **39**(4): p. 580-6.
9. Yoneten, K.K., et al., *Comparative Proteome Analysis of Breast Cancer Tissues Highlights the Importance of Glycerol-3-phosphate Dehydrogenase 1 and Monoacylglycerol Lipase in Breast Cancer Metabolism.* Cancer Genomics Proteomics, 2019. **16**(5): p. 377-397.
10. Xie, J., et al., *GPD1 Enhances the Anticancer Effects of Metformin by Synergistically Increasing Total Cellular Glycerol-3-Phosphate.* Cancer Res, 2020. **80**(11): p. 2150-2162.
11. Kerr, A.G., et al., *Epigenetic regulation of diabetogenic adipose morphology.* Mol Metab, 2019. **25**: p. 159-167.
12. Wang, S., et al., *Dysregulation of miR484-TUSC5 axis takes part in the progression of hepatocellular carcinoma.* J Biochem, 2019. **166**(3): p. 271-279.
13. Yue, N., et al., *MicroRNA-1307-3p accelerates the progression of colorectal cancer via regulation of TUSC5.* Exp Ther Med, 2020. **20**(2): p. 1746-1751.
14. Chen, X. and J. Chen, *miR-3188 Regulates Cell Proliferation, Apoptosis, and Migration in Breast Cancer by Targeting TUSC5 and Regulating the p38 MAPK Signaling Pathway.* Oncol Res, 2018. **26**(3): p. 363-372.
15. Zhang, X., L. Su, and K. Sun, *Expression status and prognostic value of the perilipin family of genes in breast cancer.* Am J Transl Res, 2021. **13**(5): p. 4450-4463.

Figures

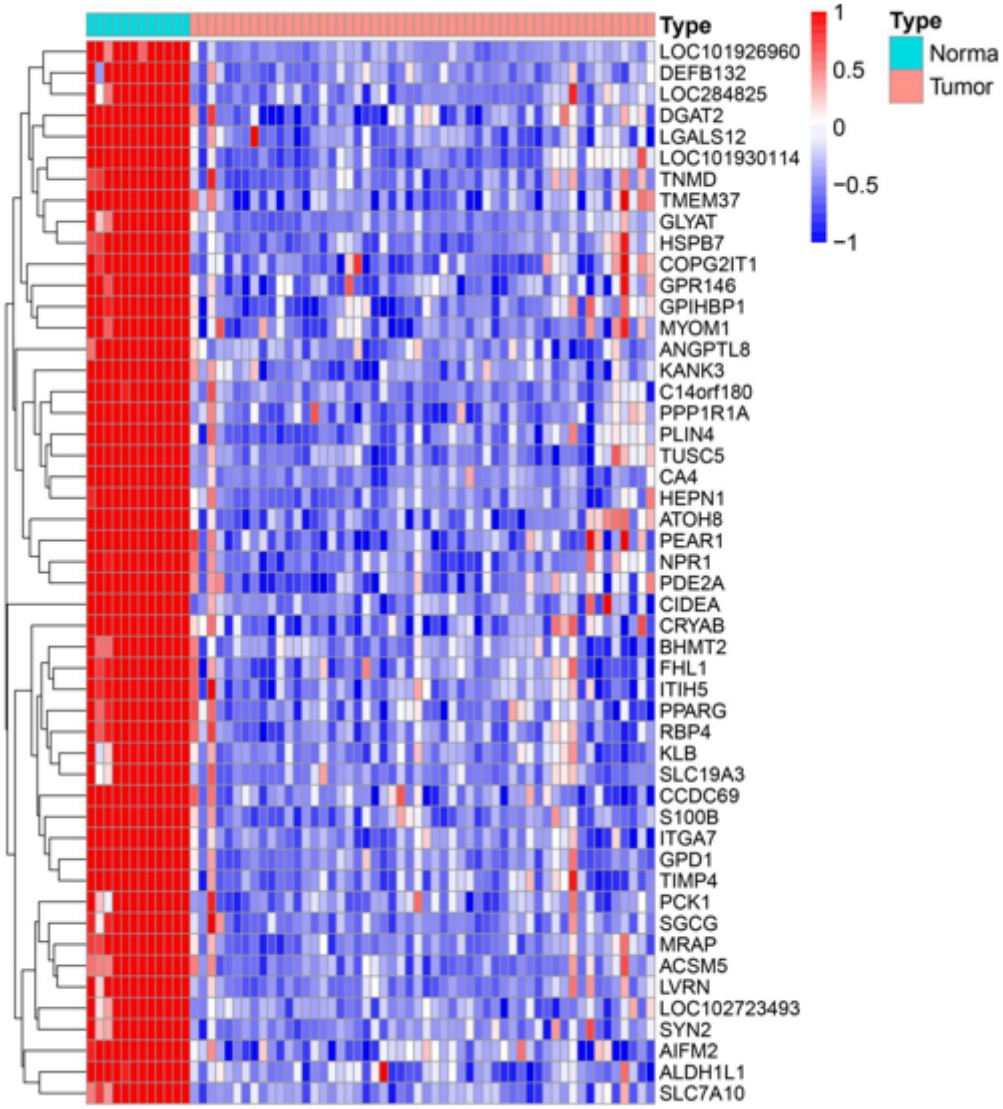

Figure 1

Heat map of the top 50 differentially expressed genes

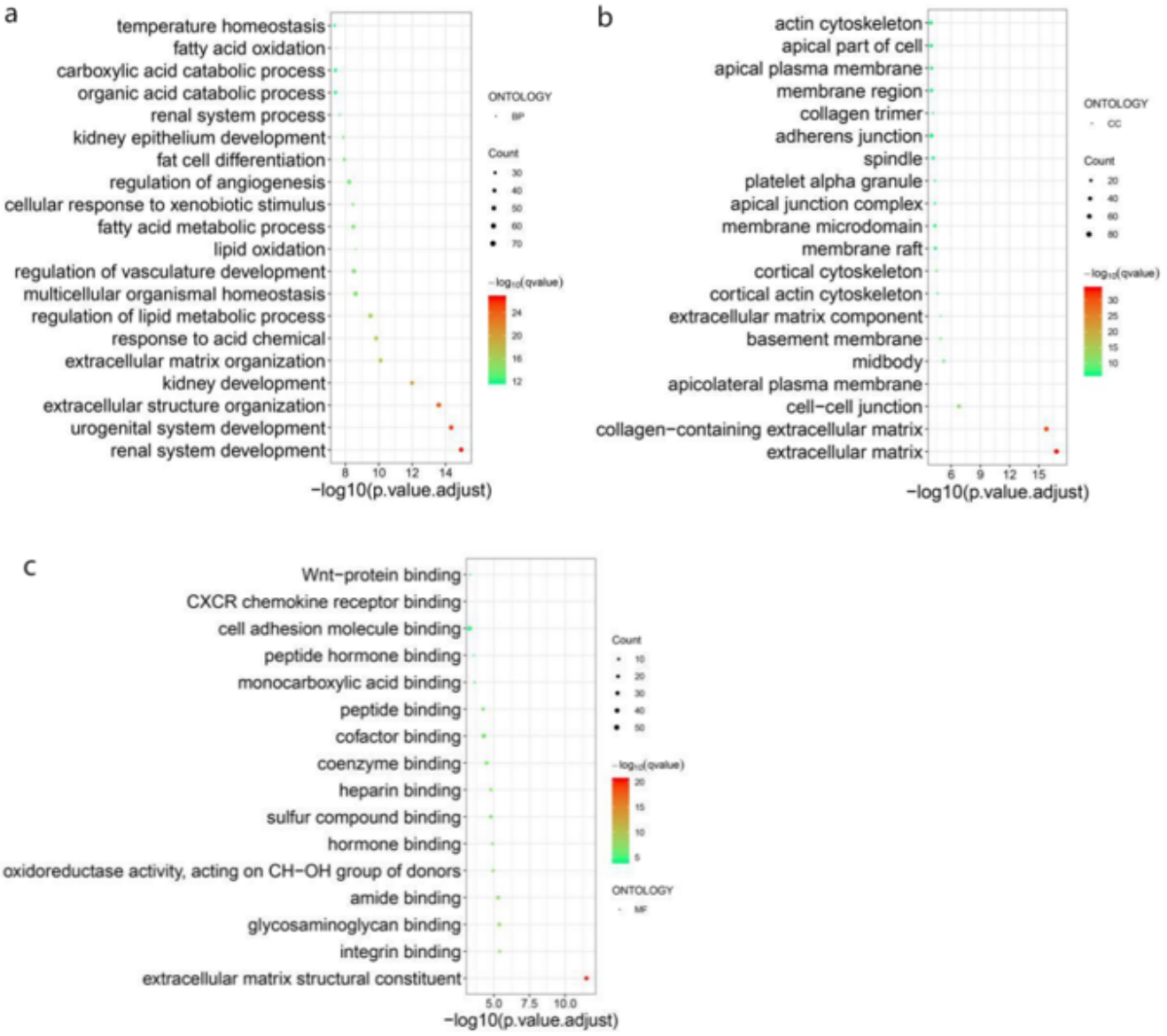

Figure 2

GO enrichment analysis “Count” is the number of genes enriched, and “p. value.adjust” is the corrected P value. **(a)** Biological processes (**b**) Cellular components **(c)** Molecular functions

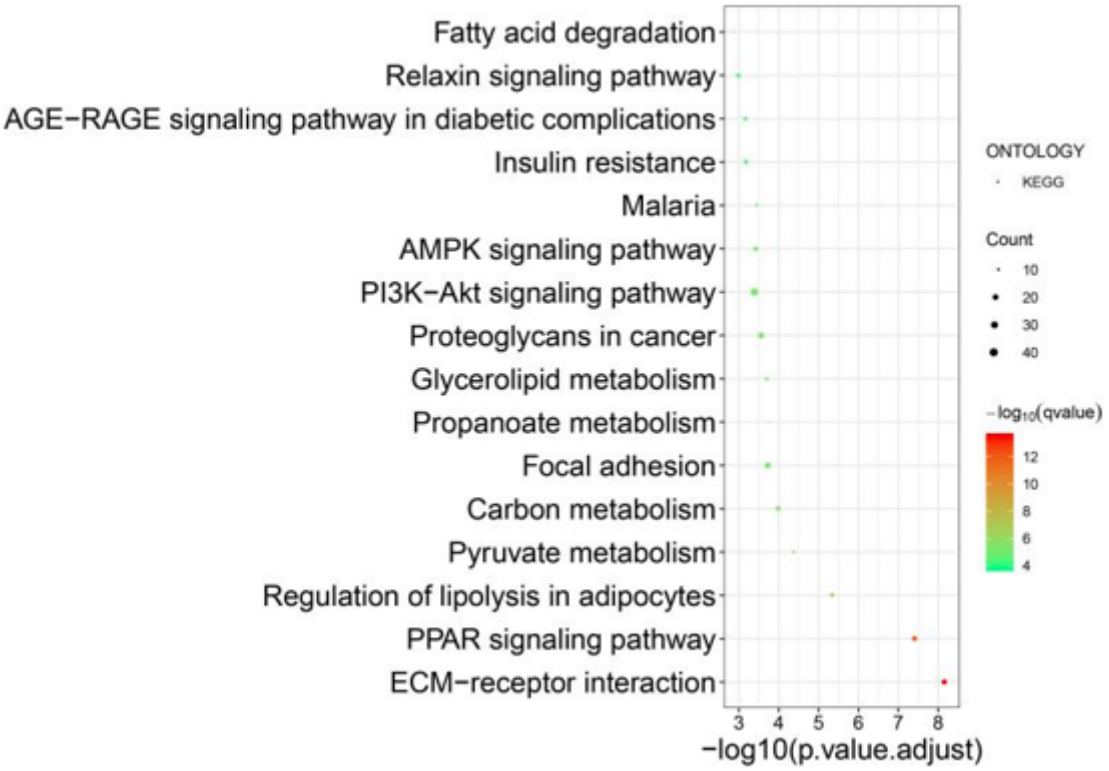

Figure 3

Partial visualization of KEGG pathway enrichment analysis

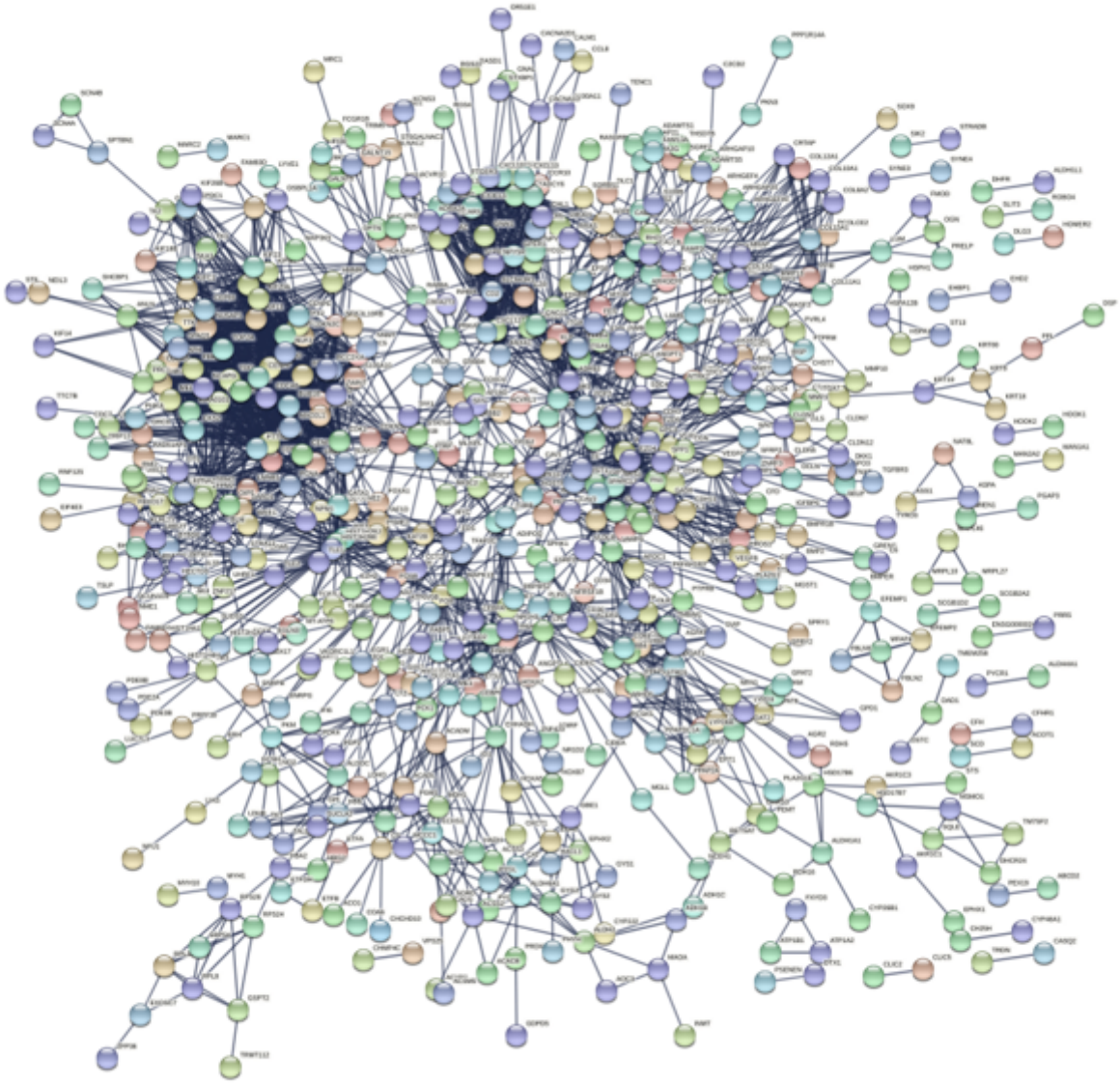

Figure 4

Protein-protein interaction network

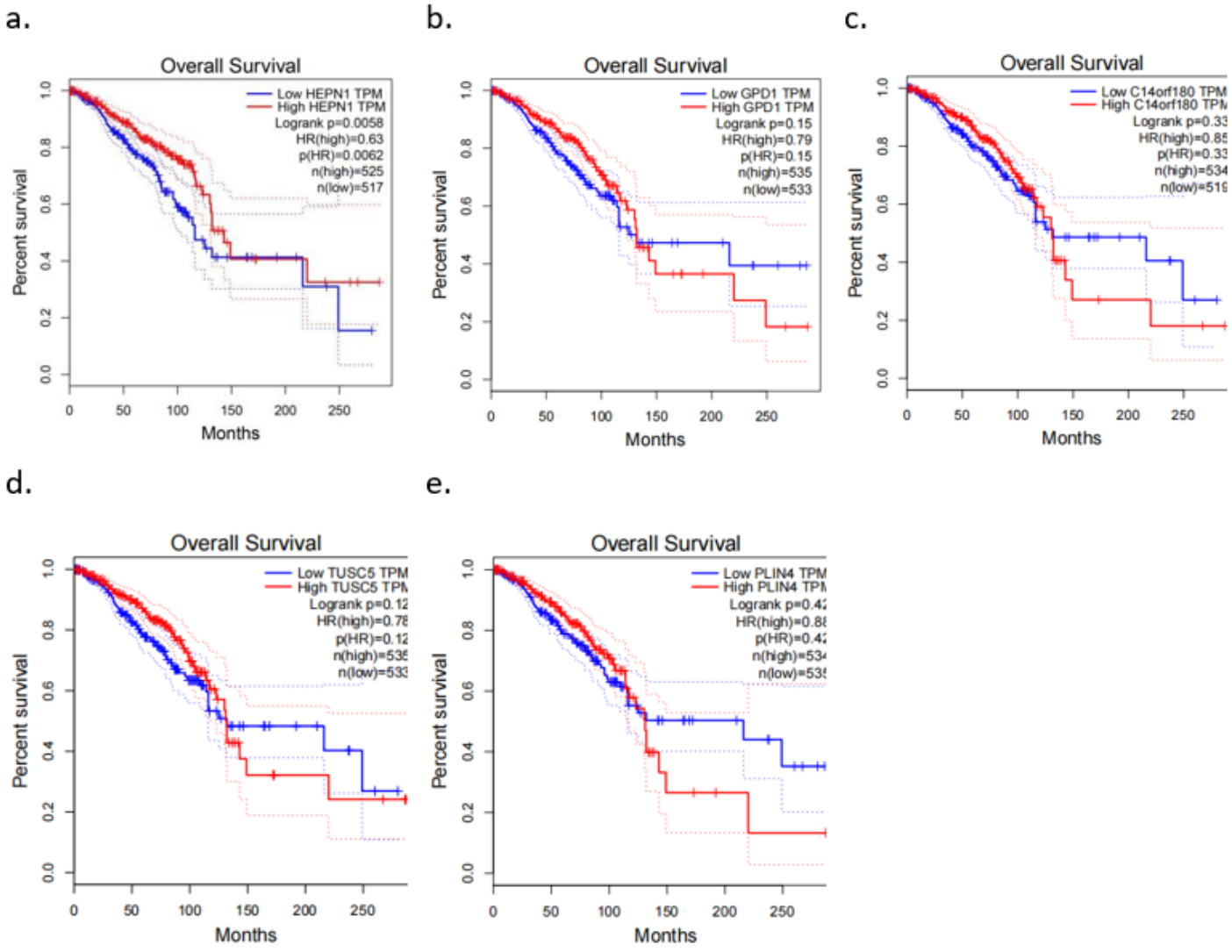

Figure 5

Survival analysis of genes with the most significant differences **(a)** HEPN1 gene survival analysis. **(b)** GPD1 gene survival analysis **(c)** C14orf180 gene survival analysis **(d)** TUSC5 gene survival analysis **(e)** PLIN4 gene survival analysis

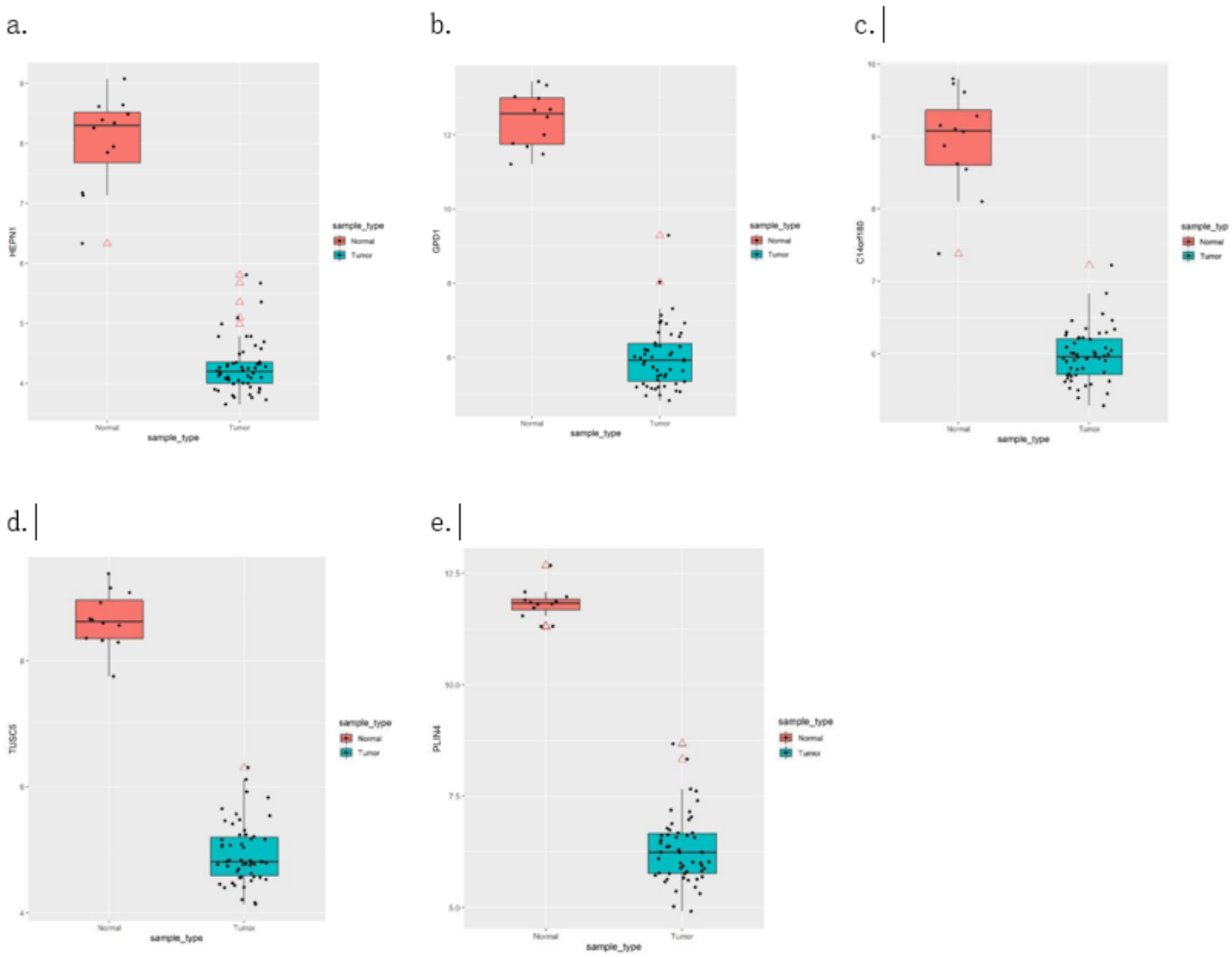

Figure 6

Expression levels of differentially expressed genes in breast cancer tissues and normal tissues. **(a)** HEPN1 gene expression level **(b)** GPD1 gene expression level **(c)** C14orf180 gene expression level **(d)** TUSC5 gene expression level **(e)** PLIN4 gene expression level