BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(8)

PathAct: a novel method for pathway analysis using gene expression profiles

Kaoru Mogushi & Hiroshi Tanaka*

Department of Bioinformatics, Division of Medical Genomics, Medical Research Institute, Tokyo Medical and Dental University 24F M&D Tower Bldg., 1-5-45 Yushima, Bunkyo-ku, Tokyo, Japan; Hiroshi Tanaka – Email: tanaka@bioinfo.tmd.ac.jp; Phone: +81-3-5803-5839; *Corresponding author

Received April 14, 2013; Accepted April 16, 2013; Published April 30, 2013

Abstract:
We developed PathAct, a novel method for pathway analysis to investigate the biological and clinical implications of the gene expression profiles. The advantage of PathAct in comparison with the conventional pathway analysis methods is that it can estimate pathway activity levels for individual patient quantitatively in the form of a pathway-by-sample matrix. This matrix can be used for further analysis such as hierarchical clustering and other analysis methods. To evaluate the feasibility of PathAct, comparison with frequently used gene-enrichment analysis methods was conducted using two public microarray datasets. The dataset #1 was that of breast cancer patients, and we investigated pathways associated with triple-negative breast cancer by PathAct, compared with those obtained by gene set enrichment analysis (GSEA). The dataset #2 was another breast cancer dataset with disease-free survival (DFS) of each patient. Contribution by each pathway to prognosis was investigated by our method as well as the Database for Annotation, Visualization and Integrated Discovery (DAVID) analysis. In the dataset #1, four out of the six pathways that satisfied $p < 0.05$ and FDR < 0.30 by GSEA were also included in those obtained by the PathAct method. For the dataset #2, two pathways ("Cell Cycle" and "DNA replication") out of four pathways by PathAct were commonly identified by DAVID analysis. Thus, we confirmed a good degree of agreement among PathAct and conventional methods. Moreover, several applications of further statistical analyses such as hierarchical cluster analysis by pathway activity, correlation analysis and survival analysis between pathways were conducted.

Background:
Gene expression profiling by microarray analysis provides a huge amount of biological information and has been widely used in biological and clinical research. Since microarray technique simultaneously detects expression levels for more than ten thousand of genes, bioinformatics approaches for interpretation of such large-scale data are essential. The microarray data is often examined using the information of a pathway, which represents a series of biological reactions that causes a specific event such as signal transduction, cell proliferation, and drug metabolism. There are several pathway databases such as the KEGG PATHWAY database **[1],** BioCarta **[2]** and GenMAPP **[3].** In addition, Gene Ontology (GO) database **[4]** provides controlled vocabularies of various genes and has a hierarchical structure based on their functions. Recently, several interpretation tools such as gene set enrichment analysis (GSEA) **[5],** the Database for Annotation, Visualization and Integrated Discovery (DAVID) **[6],** GenMAPP **[3],** and GOMiner **[7]** have been developed and widely used in the microarray analysis. The majority of tools for pathway analysis detect pathway-level difference between two groups (e.g., cases and controls). The output results generated by these tools are typically given as a list of p-values and other software-specific information, which is not suitable for further analyses. If the activity (i.e., the degree of up- and down-regulation) of a pathway can be estimated in each samples, these information would be of great use for investigation of patient-specific characteristics of a disease and further development of personalized medicine. In this study, a novel method for pathway analysis, called PathAct, is introduced. PathAct can estimate individual pathway activity by conversion of gene expression data into quantitative values for both of each pathway and each sample. One of the most unique features is that the output data is given in matrix form, which can be used

for further analysis such as hierarchical clustering and other multivariate analysis methods.

The median polish (MP) algorithm **[8],** which is a core component in PathAct, is a suitable method for an additive decomposition of a two-dimensional data matrix. MP is known as an exploratory data analysis for extraction of both row-wide and column-wide trends from a two-dimensional matrix. MP is an iterative procedure that consists of the following four steps: (1) calculating median values of each row, (2) subtracting the median values from each row, (3) calculating median values of each column, and (4) subtracting the median values from each column. These steps are repeated using the residual matrix as a new data, and the median vectors for the row and the column are accumulated at each iteration. This procedure is iterated until the reduction of the sum of absolute residual is less than a specified value, or the maximum limit of iteration is exceeded. The MP method has been used for several bioinformatics tools including the robust multi-array average (RMA) method **[9],** one of the most well-known normalization methods for DNA microarray data.

Methodology:

PathAct algorithm

Application of the MP method to microarray data is formularized as follows. Suppose that a pathway p contains N_p genes, and M samples are obtained from the microarray experiment. Note that a gene may belong to multiple pathways. For the $N_p \times M$ gene expression matrix $\mathbf{G}_p$, the expression intensity of the i-th gene for the j-th individual is denoted as $\mathbf{G}_p[i, j]$. Using MP, $\mathbf{G}_p$ is decomposed into a gene effect $\mathbf{g}_p[i]$, an individual effect $\mathbf{a}_p[j]$, and an residual matrix $\mathbf{R}_p[i, j]$, $\mathbf{G}_p[i, j] = \mathbf{g}_p[i] + \mathbf{a}_p[j] + \mathbf{R}_p[i, j]$; The individual effect $\mathbf{a}_p$ is the pathway activity and is used for further analysis. The gene effect $\mathbf{g}_p$ reflects a degree of bias such as hybridization efficiency of each gene. When a database contains K pathways, these steps are repeated for K times to produce the $K \times M$ pathway activity matrix $\mathbf{A}$ by collecting $\mathbf{a}_p$ of each pathway. These procedures, namely PathAct, were implemented in R statistical language version 2.15.2 (http://www.r-project.org/). The program is freely available upon request. A dataset including 229 human pathways was obtained from the KEGG PATHWAY database using an R packages "KEGG.db" and "hgu133plus2.db". "KEGG.db" contains information about KEGG PATHWAY entry such as pathway IDs and names, whereas "hgu133plus2.db" is an annotation data for Affymetrix HG-133 plus 2.0 array including pathway information. When multiple probe sets correspond to a single gene, a probe set with the largest interquartile range (IQR) was selected as a responsible probe set for the gene. Termination conditions of each MP process were set to 1% as a change of absolute residual and ten times as the maximum iteration steps.

Application to clinical microarray datasets

To evaluate the feasibility of our pathway analysis method, two public microarray datasets were downloaded from the Gene Expression Omnibus (GEO) database (http:// www.ncbi.nlm.nih.gov/geo/). The dataset #1 (GSE19615) contains a total of 115 gene expression profiles obtained from tissue specimens of breast cancer patients **[10]** using HG-133 Plus 2.0 arrays. The data have clinicopathological information for estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) expressions. Triple-negative breast cancer (TNBC), which does not express ER, PR, and HER2, is associated with higher risk of distant metastasis and poor prognosis compared to other type of breast cancer (non-triple-negative breast cancer; NTNBC) **[11]**. Therefore, we aimed to identify pathways differently activated between TNBC and NTNBC. The gene expression profiles including 54,613 probe sets were converted into a pathway activity matrix using PathAct. Association between pathway activities and TNBC was determined by the Wilcoxon exact rank-sum test provided by "exactRankTests" package in R. Then, pathways with at least 1.1-fold increase or decrease between TNBC and NTNBC were further selected. The dataset #2 (GSE21653) contains a total of 266 gene expression profiles from breast cancer patients using HG-133 Plus 2.0 arrays **[12].** Among them, 252 patients had information for disease-free survival (DFS), which was calculated from the date of diagnosis until date of first relapse or date of death (when the relapse was not observed). These gene expression profiles were also converted into a pathway activity matrix using PathAct, and we investigated pathways associated with DFS using Cox proportional hazards model. In both datasets, pathways satisfying both $p < 0.05$ by Wald test and FDR < 0.30 were considered statistically significant.

Comparison with conventional pathway analysis methods

For the dataset #1, we investigated pathways associated with TNBC by gene set enrichment analysis (GSEA) **[5].** We used a collection of gene sets for KEGG pathways provided by MSigDB 3.1 (c2.cp.kegg.v3.1.symbols.gmt, available at http://www.broadinstitute.org/gsea/msigdb/). Pathways that satisfied both $p < 0.05$ and FDR < 0.30 were selected and were compared with the results obtained by PathAct. In the second analysis using dataset #2, as GSEA cannot conduct survival analysis, we first extracted genes related to DFS by Cox proportional hazards model using a cut-off value of $p < 0.05$ by Wald test. Then, the selected genes were analyzed by DAVID Functional Annotation Tool **[6]** version 6.7. We used the "KEGG_PATHWAY" category provided by DAVID for analysis of pathways that were overrepresented by the genes associated with DFS. Two separate analyses were performed for the genes up-regulated in poor prognosis patients (hazards ratio (HR) > 1 by the Cox regression) and for those down-regulated in poor prognosis patients (HR < 1). Pathways that satisfied both $p < 0.05$ and FDR < 0.30 by DAVID analysis were then selected.

Discussion:

Pathway-based analysis of TNBC dataset

Using the dataset #1, we calculated the pathway activity matrix of 229 KEGG pathways for the 115 patients by the PathAct method. We then identified 15 up-regulated and 13 down-regulated pathways in TNBC by comparing the pathway activity values between TNBC and NTNBC. Using the calculated pathway activity levels, we next performed a hierarchical cluster analysis using the above 28 pathways **(Figure 1A).** When the patients were classified into two major clusters, the left cluster contained four TNBC and 61 NTNBC patients, whereas the right cluster contained 26 TNBC and 24 NTNBC patients ($p < 0.001$ by Fisher's exact test). This indicates that the PathAct method can summarize the important molecular information in breast cancer, a part of which is necessary for classification of TNBC from NTNBC.

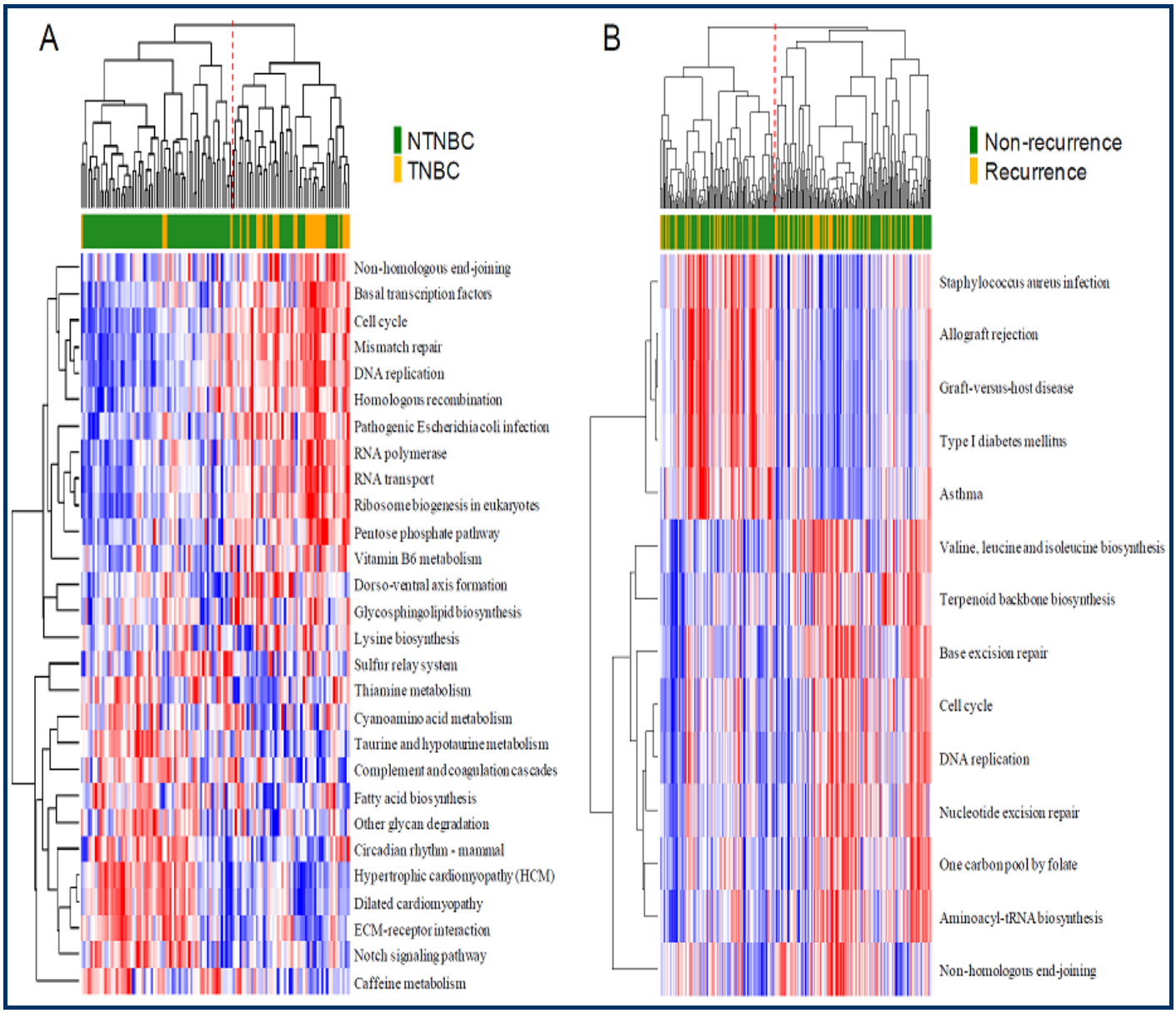

Figure 1: Hierarchical clustering of pathways using output data generated by PathAct. The pathway activity data was transformed into z-scores by setting the mean expression intensities to 0 and variances to 1 for all pathways. The Euclidean distance was used to calculate a similarity matrix among pathways or individuals, respectively. The red dashed lines indicate the two major clusters obtained by the hierarchical cluster analysis. **A**: a hierarchical cluster analysis using the selected 28 pathways for the dataset #1. **B**: a hierarchical cluster analysis using the selected 14 pathways for the dataset #2.

Pathways associated with TNBC

The pathways associated with cell proliferation (e.g., "DNA replication" and "Cell cycle") and transcription process (e.g., "Basal transcription factors" and "RNA polymerase") were significantly up-regulated in TNBC **Table 1 (see supplementary material).** This suggests that the cancer cells in TNBC are more aggressive than those in other type of cancer. On the other hand, the pathway "ECM-receptor interaction", related to cell adhesion, was down-regulated in TNBC **Table 1.** Because loss of cell adhesion promotes migration, invasion and metastasis of cancer cells, the down-regulation of these pathways suggested the higher metastatic ability of TNBC than NTNBC. It has been reported that TNBC are further classified into several subtypes such as basal-like, normal-like, and claudin-low subtypes **[13]**. Among them, the claudin-low subtype shows lower mRNA expression levels of cell cycle-related genes **[14].** As shown in **Figure 1A**, a part of TNBC patient's exhibits lower activity levels of cell cycle-related Pathways, which suggest the existence of subtypes among TNBC patients.

Pathway-based analysis of DFS in breast cancer

Using the dataset #2, we also obtained the pathway activity matrix of 229 KEGG pathways for the 266 patients by the PathAct method. Using the pathway activity levels, we selected 14 pathways that had significant association with DFS **Table 2 (see Supplementary material).** Among these pathways, nine showed hazard ratio (HR) > 1, and the other five had HR < 1. Similar to the analysis of the dataset #1, we conducted a

hierarchical cluster analysis using the selected 14 pathways **(Figure 1B).** When the patients were divided into the two clusters, the left cluster contained 27 recurrence and 80 recurrence-free patients, whereas the right cluster contained 56 recurrence and 89 recurrence-free patients (p = 0.030 by Fisher's exact test). This indicated that the patients in the right cluster had significantly higher risk of tumor recurrence.

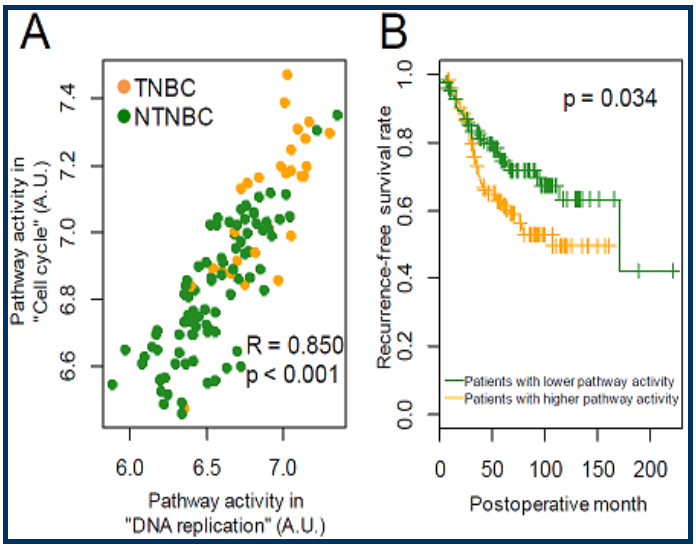

Figure 2: Application of PathAct method for further numerical analysis. **A**: correlation analysis of "DNA replication" and "Cell cycle" for the dataset #1 (p < 0.001 by Pearson's correlation test). **B**: analysis of correlation between DFS and the pathway activity in "Allograft rejection". Down-regulation of this pathway contribute to better prognosis (p = 0.034 by log-rank test).

Pathways associated with prognosis of breast cancer

The pathways associated with cell proliferation (e.g., "DNA replication", "Base excision repair", and "Cell cycle") were positively correlated with poor prognosis **Table 2.** Promotion of cell proliferation in cancer is a typical phenomenon in carcinogenesis and cancer progression. On the other hand, the five pathways for better prognosis contained "Asthma" and "Graft-versus-host disease", which did not seem to be associated with cancer. However, these pathways include genes related to immune response process such as antigen processing (e.g., MHC class II genes) and cytokines (e.g., IL2, 4, 5, and 9). This suggested that the immune response function in cancerous tissue was activated in patients with better prognosis compared with patients with poor prognosis. In fact, it has been reported that breast cancer patients with a higher number of tumor-infiltrating CD8(+) lymphocytes shows better prognosis **[15].** Therefore, activation of these pathways in patients with good prognosis is possibly caused by increased number of lymphocytes by lymphocytic infiltration.

Comparison to conventional methods

In order to validate the PathAct method, we compared the analysis results of the datasets #1 and #2 with other approaches. First, we performed GSEA for the dataset #1 and identified six pathways that were significantly altered in TNBC **Table 3 (see supplementary material).** Interestingly, four out of six pathways identified by GSEA were also included in those obtained by the PathAct method. Therefore, we confirmed that GSEA and PathAct had ability to detect common biological pathways in the microarray dataset. Next, we evaluated pathways associated with DFS in the dataset #2. Using Cox proportional hazards model, 1,487 probe sets were identified to be negatively correlated with DFS, whereas 1,081 probe sets had positive correlation with DFS. DAVID analysis was then performed for each of the selected gene sets **Table 4 (see supplementary material.** "Cell Cycle" and "DNA replication" pathways were significantly overrepresented in genes associated with poor prognosis, and these were also detected by the PathAct method. Similarly, 13 pathways were identified for genes correlated with good prognosis, and four pathways associated with immune response ("Asthma", "Graft-versus-host disease", "Allograft rejection", and "Type I diabetes mellitus") were also detected by the PathAct. Thus, we confirmed good agreement between PathAct and DAVID.

Quantitative analyses using pathway activities

A major advantage of PathAct method is that obtained pathway activity levels can be used for further statistical analysis. For example, the correlation analysis of "DNA replication" and "Cell cycle" for the dataset #1 was performed **(Figure 2A).** These two pathways showed significant correlation (p < 0.001 by Pearson's correlation test), and both pathways were up-regulated in TNBC compared with NTNBC. Another example is an analysis of correlation between DFS and the pathway activity in "Allograft rejection" **(Figure 2B).** When the median value of the pathway activity levels was chosen for a cut-off point, it was demonstrated that up-regulation of this pathway contribute to better prognosis (p = 0.034 by log-rank test). Because the "Allograft rejection" pathway contains many genes associated with immune response, this result infers that increased number of lymphocytes by lymphocytic infiltration could contribute to the better prognosis of the breast cancer patients.

Conclusion:

Pathway analysis plays an important role in interpreting genome-wide gene expression data. Several methods for pathway analysis have been proposed, but they are restricted to making comparisons between groups. Thus, the development of a flexible evaluation framework for individual patients is crucial for the advanced interpretation of microarray data. In this study, a novel approach for estimating individual pathway activity using the median polish algorithm was introduced. Using the clinical microarray datasets, the capability of the PathAct method was evaluated. The PathAct method could detect the similar pathways with those obtained by the conventional methods such as GSEA and DAVID. Moreover, because the processed data (i.e., the pathway activity matrix) are given as quantitative values for both of each pathway and each sample, they could be utilized for further statistical analysis including analysis of correlation and survival data. Therefore, PathAct is a promising tool for pathway-level investigation and interpretation of the comprehensive gene expression data.

Conflict of interest:

No conflict of interest was declared.

Acknowledgement:

The authors would like to thank Dr. Ken Miyaguchi for a critical reading of this manuscript.

Reference:
[1] Kanehisa M & Goto S, *Nucleic Acids Res.* 2000 **28**: 27 [PMID: 10592173]
[2] Nishimura D, *Biotech Softw Internet Rep.* 2001 **2**: 117
[3] Dahlquist KD *et al. Nat Genet.* 2002 **31**: 19 [PMID: 11984561]
[4] Ashburner M *et al. Nat Genet.* 2000 **25**: 25 [PMID: 10802651]
[5] Subramanian A *et al. Proc Natl Acad Sci USA.* 2005 **102**: 15545 [PMID: 16199517]
[6] Huang da W *et al. Nat Protoc.* 2009 **4**: 44 [PMID: 19131956]
[7] Zeeberg BR *et al. Genome Biol.* 2003 **4**: R28 [PMID: 12702209]
[8] Malo N *et al. Nat Biotechnol.* 2006 **24**: 167 [PMID: 16465162]
[9] Irizarry RA *et al. Biostatistics.* 2003 **4**: 249 [PMID: 12925520]
[10] Li Y *et al. Nat Med.* 2010 **16**: 214 [PMID: 20098429]
[11] Dent R *et al. Clin Cancer Res.* 2007 **13**: 4429 [PMID:17671126]
[12] Sabatier R *et al. Breast Cancer Res Treat.* 2011 **126**: 407 [PMID: 20490655]
[13] Lehmann BD *et al. J Clin Invest.* 2011 **121**: 2750 [PMID: 21633166]
[14] Prat A *et al. Breast Cancer Res.* 2010 **12**: R68 [PMID: 20813035]
[15] Mahmoud SM *et al. J Clin Oncol.* 2011 **29**: 1949 [PMID: 21483002]

Edited by P Kangueane
Citation: **Mogushi & Tanaka,** Bioinformation 9(8): 394-400 (2013)

BIOINFORMATION

open access

Supplementary material:

Table 1: List of pathways associated with TNBC using PathAct.

Pathway	p-value	FDR	Fold
Up-regulated pathways in TNBC			
DNA replication	<0.001	<0.001	1.272
Cell cycle	<0.001	<0.001	1.202
Mismatch repair	<0.001	<0.001	1.183
Ribosome biogenesis in eukaryotes	<0.001	<0.001	1.185
Vitamin B6 metabolism	<0.001	<0.001	1.457
RNA transport	<0.001	<0.001	1.115
Basal transcription factors	<0.001	<0.001	1.118
Pathogenic Escherichia coli infection	<0.001	<0.001	1.149
Homologous recombination	<0.001	0.002	1.105
Glycosphingolipid biosynthesis - lacto and neolacto series	<0.001	0.004	1.113
Pentose phosphate pathway	0.001	0.011	1.121
RNA polymerase	0.001	0.011	1.116
Dorso-ventral axis formation	0.001	0.012	1.135
Non-homologous end-joining	0.008	0.032	1.101
Lysine biosynthesis	0.016	0.051	1.200
Down-regulated pathways in TNBC			
Hypertrophic cardiomyopathy (HCM)	<0.001	<0.001	0.905
Notch signaling pathway	<0.001	<0.001	0.877
Circadian rhythm - mammal	<0.001	<0.001	0.851
Dilated cardiomyopathy	<0.001	0.001	0.905
Fatty acid biosynthesis	<0.001	0.001	0.826
ECM-receptor interaction	<0.001	0.003	0.832
Taurine and hypotaurine metabolism	0.001	0.008	0.854
Cyanoamino acid metabolism	0.001	0.008	0.834
Other glycan degradation	0.001	0.009	0.887
Sulfur relay system	0.002	0.012	0.893
Caffeine metabolism	0.006	0.027	0.871
Thiamine metabolism	0.010	0.038	0.894
Complement and coagulation cascades	0.011	0.040	0.904

Table 2: List of pathways associated with DFS by PathAct.

Pathway	HR (95% CI)	p-value	FDR
Negative correlation with DFS			
DNA replication	2.049 (1.235-3.399)	0.005	0.217
Base excision repair	3.130 (1.366-7.168)	0.007	0.217
Non-homologous end-joining	4.533 (1.491-13.78)	0.008	0.217
Aminoacyl-tRNA biosynthesis	3.353 (1.373-8.186)	0.008	0.217
Cell cycle	2.535 (1.257-5.112)	0.009	0.217
Nucleotide excision repair	2.818 (1.288-6.164)	0.009	0.217
Valine, leucine and isoleucine biosynthesis	2.536 (1.206-5.336)	0.014	0.279
Terpenoid backbone biosynthesis	2.203 (1.169-4.153)	0.015	0.279
One carbon pool by folate	2.593 (1.196-5.624)	0.016	0.279
Positive correlation with DFS			
Asthma	0.343 (0.159-0.737)	0.006	0.217
Graft-versus-host disease	0.526 (0.327-0.847)	0.008	0.217
Staphylococcus aureus infection	0.573 (0.378-0.868)	0.009	0.217
Allograft rejection	0.545 (0.345-0.862)	0.009	0.217
Type I diabetes mellitus	0.520 (0.304-0.891)	0.017	0.284

Table 3: TNBC-associated pathways analyzed by GSEA.

Pathway	Size	NES*	p-value	FDR
Aminoacyl trna biosynthesis	32	1.928	<0.001	0.063
Cell cycle	113	1.869	<0.001	0.069

RNA degradation	51	1.735	0.037	0.206
DNA replication	34	1.698	0.031	0.216
Basal transcription factors	33	1.696	0.020	0.175
Glycosphingolipid biosynthesis lacto and neolacto series	26	1.663	0.011	0.196

* NES, normalized enrichment score.

Table 4: List of pathways associated with DFS identified by DAVID.

Pathway	Size	%	p-value	FDR
Negative correlation with DFS				
hsa04110:Cell cycle	21	2.41	< 0.001	< 0.001
hsa03030:DNA replication	9	1.03	< 0.001	0.025
Positive correlation with DFS				
hsa05330:Allograft rejection	12	3.90	< 0.001	< 0.001
hsa05332:Graft-versus-host disease	12	3.90	< 0.001	< 0.001
hsa04940:Type I diabetes mellitus	12	3.90	< 0.001	< 0.001
hsa04612:Antigen processing and presentation	15	4.87	< 0.001	< 0.001
hsa05310:Asthma	10	3.25	< 0.001	< 0.001
hsa05320:Autoimmune thyroid disease	12	3.90	< 0.001	< 0.001
hsa04672:Intestinal immune network for IgA production	11	3.57	< 0.001	< 0.001
hsa05416:Viral myocarditis	12	3.90	< 0.001	< 0.001
hsa05322:Systemic lupus erythematosus	12	3.90	< 0.001	< 0.001
hsa04514:Cell adhesion molecules (CAMs)	13	4.22	< 0.001	< 0.001
hsa04640:Hematopoietic cell lineage	7	2.27	0.005	0.055
hsa00340:Histidine metabolism	4	1.30	0.017	0.149
hsa05340:Primary immunodeficiency	4	1.30	0.028	0.219