Dissecting the consistently across all stages widespread deregulation of gene expression in pancreatic ductal adenocarcinoma for early diagnosis and therapy.

Aristeidis Sionakidis[1], Panagiotis Nikolaos Lalagkas[2], Andigoni Malousi[3], and Ioannis S. Vizirianakis[4,5]

[1]Institute of Genetics and Cancer, University of Edinburgh, Scotland, UK
[2]Department of Biological Sciences, University of Massachusetts, USA
[3]Laboratory of Biological Chemistry, School of Medicine, Aristotle University of Thessaloniki, Greece
[4]Laboratory of Pharmacology, School of Pharmacy, Aristotle University of Thessaloniki, Thessaloniki, Greece
[5]Department of Life and Health Sciences, University of Nicosia, Nicosia, Cyprus

Keywords: PDAC, gene expression, bioinformatics, pharmacogenomics, precision medicine
Corresponding authors: Aristeidis Sionakidis, A.Sionakidis@sms.ed.ac.uk; Ioannis S. Vizirianakis, ivizir@pharm.auth.gr

1 Supplementary Methods

We searched the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/) for "pancreatic ductal adenocarcinoma" and filtered results for organism (*Homo sapiens*) and study type "Expression profiling by array" (see Figure 1). We then examined the results one by one to filter out studies where patients (or the corresponding samples) have undergone pharmacological treatment prior to measurement, and studies where human cell lines, cultures or xenografts were used instead of patient samples. We also excluded studies involving samples from very particular categories of patients (e.g. patients with known mutational status on genes) and tissue biopsy studies which did not have any information on the stage of the tumors. Regarding blood biopsy studies, we used the same exclusion criteria with the exception of tumor stage filtering. We therefore identified 10 studies [1–10] that satisfied our criteria. One study [3] had both blood and tumor tissue samples. The blood samples from this study were only used in the blood samples analysis, while the tumor tissue samples from this study were only used in the tumor tissue analysis. Overall, samples from 7 studies [1–7] were used for the tumor tissue analysis and samples from 4 studies [3, 8–10] were used for the blood samples analysis. Certain studies [1, 3, 4, 6, 7] used adjacent normal tissue for control samples. Controls in three of the liquid biopsy studies [8–10] were healthy non-cancer subjects. Control samples (normal adjacent tissue) from a study [3] which had both circulating tumor cells and tissue cells were only used in the tissue samples analysis. Circulating tumor cell data from this study were only used in the blood samples analysis.

The following sections give information on the tools that we used for our analysis (see section 2), the phenotypic data preprocessing steps (see section 1.1.1), the imputation and annotation methods we used on the gene expression matrix of each study (see sections 1.1.2, 1.1.3), the way the samples from each study were normalised (see section 1.1.4), the production of diagnostic plots (see section 1.1.5), differential gene expression analysis (see section 1.2), active subnetwork analysis (see section 1.3), miRNA enrichment analysis (see section 1.3) and further pharmacogenomic analysis steps (see section 1.5).

1.1 Data preprocessing

1.1.1 Sample filtering

The published gene expression data were downloaded from GEO using the R package *GEOquery*[11]. Regarding tumor tissue samples, stage was determined using the American Joint Committee on Cancer (AJCC) [12]. For six [1–5, 7, 13] out of seven studies AJCC classification information was available. For one study [6] AJCC staging was not available, but samples had been characterised with TNM information. We used the TNM information to classify the samples with respect to the AJCC system using the 8th edition's criteria [14]. One study [1, 13] had samples from multiple sources (i.e. hospitals, institutions) some of which had undergone treatment (neoadjuvant or adjuvant). All samples originating from these sources were discarded as it was not clear which ones had undergone treatment and

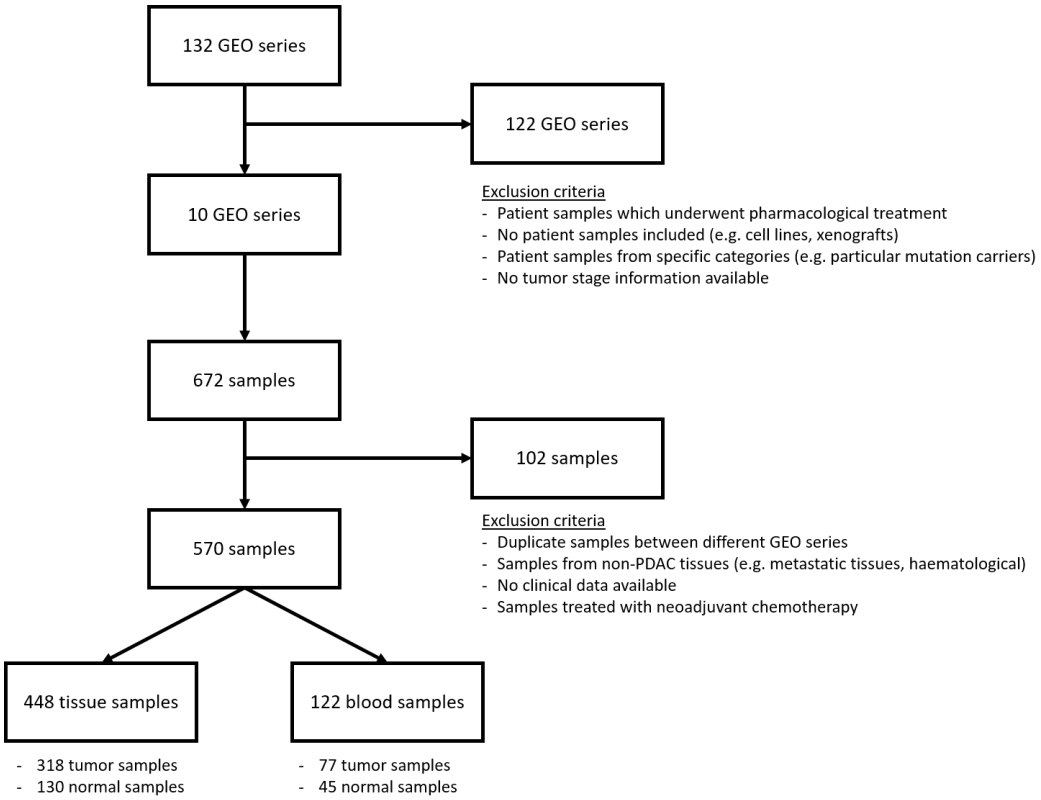

Figure 1: **Search strategy.** The series of steps following the search of the term "pancreatic ductal adenocarcinoma" in the NCBI GEO database (filtered for organism: *Homo sapiens* and study type: "Expression profiling by array").

which ones had not. For all tumor tissue studies we only kept samples resected from the tumor site and healthy control samples (either healthy subjects or normal adjacent tissue). One study [2] had metastatic samples from the patients' livers; these were excluded from the analysis. Furthermore, ten tumor site samples from this study were also excluded because they originally came from another study [3] of our pool of studies.

1.1.2 Gene expression matrices: Missing values and imputation

None of the blood sample studies had missing values in the corresponding gene expression matrices, while two [2, 6] of the tumor tissue sample studies had missing values. In one study [2], only 15% of the probes had recorded expression values; we therefore only kept these probes. In the study by Yang et al. [6] there were only 23 missing values which were imputed using 10-nearest neighbor imputation using the *impute* [15] package which was specifically designed to perform imputation in microarray data. We then removed all the samples that we also excluded at the previous step (see section 1.1.1) from the expression matrices.

1.1.3 Gene expression matrices: Annotation

We used NCBI Entrez [16] nomenclature as our consistent annotation across studies. We discarded the rows that mapped to multiple different Entrez identifiers. When two or more rows mapped to the same unique Entrez identifier, these were averaged into one probe. In this manner, the final version of each gene expression matrix consisted of one row per Entrez identifier (i.e., one row per gene). Eight [1–3, 5, 7–10, 13] out of our ten studies included Entrez annotation information in their experiment's feature data. One study [4] had RefSeq [17] annotation available and we used the *org.Hs.eg.db* [18] package map these identifiers to the corresponding Entrez identifiers. Some probes mapped to multiple RefSeq identifiers, however it is possible that these RefSeq identifiers map to the same Entrez identifier. We therefore discarded only the probes with RefSeq identifiers which mapped to multiple or none Entrez identifiers. We then proceeded with averaging the Entrez-annotated probes as described before. One study [6] only had probe sequence information available. We followed the pipeline described by Ensembl [19] (`https://www.ensembl.org/index.html`) to annotate these probes with RefSeq gene identifiers using the NCBI nucleotide BLAST® [20] online interface (`https://blast.ncbi.nlm.nih.gov/Blast.cgi`). We then proceeded with mapping the RefSeq identifiers to Entrez identifiers as described previously.

1.1.4 Global expression matrices and normalisation

We then unified the expression matrices from the tumor tissue studies into one tumor tissue global expression matrix and the expression matrices from the blood sample studies into one blood sample global expression matrix. We did not perform the join based only on the common genes among the studies, but kept all the genes from each study. We also generated a second kind of global expression matrices which was the only kind that was used for downstream analysis: the normalised global expression matrix. For these matrices, we firstly normalised the samples of each study separately by converting each sample's expression value for a gene to the corresponding z-score (standard score), by subtracting the mean expression value for this gene across samples from each sample's expression value and then dividing by the standard deviation of expression values for this gene across samples. Then, we proceeded with joining the expression matrices as described previously.

1.1.5 Diagnostic plots

The diagnostic plots we generated were based on subsets of the global expression matrices (filtered for the genes that are present in all studies). We used the two versions of the global expression matrix for each category of samples (tumor tissue, blood samples) to produce diagnostic plots in order to examine how comparable the expression value scales of samples are across studies. We generated three kinds of plots: multidimensional scaling plots, expression boxplots and sample distance heatmaps. We therefore generated six plots in total for each type of samples. Multidimensional scaling plots project the gene expression

data to two dimensions and plot the first two principal components, expression boxplots show how the expression values of each sample compare with those of the other samples and the sample distance heatmaps demonstrate how "close" or "far" the samples are from each other in terms of high-dimensional Manhattan distance (L1-distance). The Manhattan distance is preferable to the Euclidean distance when the vectors (samples) are of particularly high-dimensional as in the case of gene expression samples[21]. Supplementary Figures 1-6 display the multidimensional scaling plots, expression boxplots and heatmaps, for both tumor tissue samples and blood samples. After the normalisation is applied the samples are scattered in the two-dimensional projection in a manner which is not study-specific (i.e. driven by batch effects). That is also illustrated on the boxplots of expression and the heatmaps of sample distances.

1.2 Differential Gene Expression Analysis

Differential Gene Expression Analysis (DGEA) was carried out by fitting a linear model for the expression values of each gene, adjusted for AJCC classification (normal, stage 1, stage 2, stage 3, stage 4) and study using the *limma*[22] package in R. We performed DGEA twice: once on the tumor tissue samples and once on the blood samples. Staging information was not available in the blood samples case, where the linear models were only adjusted for their status (normal/tumor). In both cases we used the global normalised gene expression matrix as input which was formed as described in section 1.1.4. We carried out seven comparisons in the tissue samples analysis: 1) stage 1 vs. normal, 2) stage 2 vs. normal, 3) stage 3 vs. normal, 4) stage 4 vs. normal, 5) stages 1 & 2 vs. stages 3 & 4, 6) stage 1 vs. stage 2 and 7) stage 1 vs. stage 4. In the blood samples analysis we compared tumors to normal samples. DGEA with *limma* requires a gene expression matrix with genes in rows and samples in columns, a design matrix with all predictors of interest (categorical predictors are divided into dummy variables), and a contrast matrix which specifies the groups to be compared. The design matrix with the predictor variables is generated by *limma's* `model.matrix` function. A design matrix has dimension $n \times d$, where n is the number of samples in the gene expression matrix to be analysed and d is the number of experimental conditions (predictor variables). Categorical variables are broken down to dummy variables, while continuous variables can be directly used. In our case no continuous variables were used. Design matrices can also have an additional column for an intercept term, which is especially useful when continuous predictors are used, but in our case, where only categorical predictors are used, it can be ignored (set to 0) as results are essentially the same in both cases[23]. A separate linear model is fit for each gene with the expression values as the predicted continuous outcome variable and the set of selected variables in the design matrix as predictors. We used *limma's* robust Empirical Bayes algorithm[24] to smooth standard errors and account for genes with large variances, by decreasing those genes' prior distribution effect on model hyperparameter estimation and increasing the respective effect of other genes. We used the Benjamini-Hochberg p-value (BH-adjusted p-value) correction[25]

to account for multiple testing. We selected genes with an adjusted p-value below 0.05 for downstream analysis. In typical *limma* output, the log-fold change statistic reported for each gene for a particular comparison represents the numerical difference between the coefficients estimated for the two contrasts (responders - non-responders). In our case, since the data have been normalised prior to being subjected to DGEA, the corresponding statistic represents the numerical difference between the coefficients of the two contrasts measured in standard deviations from the gene's mean expression across samples (s.d. units). A volcano plot was generated using the *EnhancedVolcano*[26] package to visualise the results of DGEA.

1.2.1 Gene signature metascores

After identifying our signature of 820 genes, we produce two separate metascores: a) the **URS (Up-Regulated Score)** and b) the **DRS, Down-Regulated Score**. These two scores consist of the average expression (excluding missing values) of each sample on the list of the up-regulated and the down-regulated genes respectively. The scores are then an indication of whether a sample's gene expression with respect to the down-regulated or up-regulated genes of the signature is high or low. The scores were calculated for all samples (tumors and normal samples). We then assessed whether normal samples are have statistically significantly different *URS* and *DRS* scores when compared to samples from different stages and tumor blood samples. We used independent t-tests to examine the association, a p-value threshold of 0.05 for statistical significance and Cohen's d to estimate the effect size of the association. Normal samples from tissue samples and blood samples were combined and tested against tumor samples as a whole. The same scores were produced for TCGA data and the same tests were performed. However, due to the fact that there were no normal samples in the TCGA data, the comparison was made between samples of dead and alive subjects to see whether the *URS* and the *DRS* are also associated with vital status.

1.3 Active subnetwork enrichment analysis

For the identification of enriched biological networks we used the *pathfindR*[27] package. The approach offered by the package incorporates active subnetwork identification and subsequent enrichment analysis using the identified active subnetworks. For a given list of significantly differentially expressed genes, an active subnetwork is defined as a group of interconnected genes in a protein-protein interaction network (PIN) that predominantly consists of significantly differentially expressed genes. In other words, active subnetworks define distinct response-related sets of interacting genes. We selected the BioGRID[28] protein interaction network (PIN) offered by the package. We performed enrichment analysis within all gene sets available in the package: BioCarta[29], Gene Ontology[30,31] - Biological Processes (GO-BP), Gene Ontology - Cellular Components (GO-CC), Gene Ontology - Molecular Functions (GO-MF), Kyoto Encyclopedia of Genes and Genomes[32–34] (KEGG) and Reactome[35,36]. We used the list of genes (gene symbols) we identified through DGEA,

accompanied by the corresponding BH-adjusted p-values and expression changes as input. We used the "greedy" algorithm that is offered by the package for the active subnetwork search, which is based on the work of Chuang et al.[37], for its simplicity and speed. We used the default options for search depth (1), maximum depth (1), minimum gene set size (10), maximum gene set size (300).

By default, the identified subnetworks are filtered for subnetworks with a score higher than the given quantile threshold (default is 0.80) and subnetworks in which genes from our pool of significant genes are present in a proportion higher than 2%. We run the active subnetwork search for ten iterations and focused on the pathways which had BH-adjusted enrichment p-values lower than 0.05. Subsequently, hierarchical clustering is performed on the enriched terms to produce grouped results and highlight key pathways with different biological functions.

For each remaining active subnetwork, and using the genes included in each of these subnetworks, separate pathway enrichment analyses are performed via one-sided hypergeometric testing. The enrichment tests use the genes in the BioGRID PIN as the gene pool (i.e., as background genes for the hypergeometric test). Next, the p-values obtained from the enrichment tests are adjusted (we used the BH correction here). Pathways with adjusted p-values larger than the given threshold (we set this to 0.05) are discarded. The remaining significantly enriched pathways per all filtered subnetworks are then aggregated by keeping only the lowest adjusted p-value for each pathway if a pathway was found to be significantly enriched in the enrichment analysis of more than one subnetwork. This whole process of active subnetwork search and enrichment analysis is repeated 10 times (10 iterations).

1.3.1 Clustering of enriched terms

After the output from `run_pathfindR` was generated, we proceeded with clustering (hierarchical clustering, distance metric: "average") the enriched terms in order to identify the most representative terms in groups of similar pathways. Firstly, using the list of input genes in each enriched pathway, a kappa statistics matrix with the pairwise kappa statistics (i.e., a chance-corrected measure of co-occurrence between two sets of categorized data), between the enriched pathways is calculated. Subsequently, hierarchical clustering is performed, defining distance as "1 - kappa statistic" and the optimal number of clusters is selected by maximizing the average silhouette width. The representative term in each cluster is then determined as the term with the lowest BH-adjusted p-value (calculated during the enrichment step previously).

1.4 miRNA enrichment analysis

miRNA enrichment analysis was performed online using the Mienturnet® tool [38], `http://userver.bio.uniroma1.it/apps/mienturnet/`). We supplied the tool with the list of significantly ($p.adj < 0.05$) differentially expressed genes derived from the

comparisons of the four different stages to normal tissue samples and the comparison of tumors to normal samples from liquid biopsies (5 lists in total). We removed the miRNAs and miRNA host-genes with $p.adj < 0.05$ before submitting the lists to the tool so that they do not affect the enrichment result. Enriched miRNAs with $p.adj < 0.05$ were considered to be significantly enriched.

1.5 Drug-gene interactions and Circos plots

Cancer driver status of genes information was downloaded from COSMIC (Catalogue Of Somatic Mutations In Cancer) [39]. Interactions between approved drugs and genes were downloaded from the DrugBank [40] database (https://go.drugbank.com/). Drugs were classified into categories using the Anatomic Therapeutic Chemical (ATC) classification system (https://www.who.int/tools/atc-ddd-toolkit/atc-classification, August 2020 version). We mapped our lists of differentially expressed genes from all stages and blood samples to approved drugs with which pharmacological interaction is documented. Circular plots with pharmacogenomic links and additional annotation were then generated using Circos [41]. The supporting configuration files (Perl scripts) and required text files are available in the repository (see section 2).

2 Software, code and data availability

Preprocessing and analysis of the gene expression data, was perfomed in R [42] (http://www.r-project.org – version 4.1.1). We used packages from both CRAN (http://cran.r-project.org/) and Bioconductor (version 3.13, http://www.bioconductor.org/). All relevant code to reproduce the analysis is available at a GitHub repository (https://github.com/lalagkaspn/pdac_omics). Code for the preprocessing and analysis of the tumor tissue biopsy data is included in the "GSE_tumor_stage.R" script. Code for blood biopsy studies is included in the "GSE_blood_samples.R". The "Pathway_ontology_driver_genes.R" script contains code for the active subnetwork analysis of the differentially expressed genes and finally the "Circos.R" script contains the code for the preparation of all text files that required by the Circos [41] software to produce pharmacogenomic plots. For every analytical process which included a random component (e.g. enrichment analysis), we used the random number generator version (`RNGversion`) "4.0.2" and random seed (`123`), for consistency and reproducibility. miRNA enrichment analysis was performed using Mienturnet, an online tool (http://userver.bio.uniroma1.it/apps/mienturnet/). The input lists for Mienturnet are available in the repository. Interactions between approved drugs and genes were downloaded from the DrugBank [40] database (https://go.drugbank.com/). Drugs were classified into categories using the Anatomic Therapeutic Chemical (ATC) classification system (https://www.who.int/tools/atc-ddd-toolkit/atc-classification, August 2020 version).

References

1. Stratford JK, Bentrem DJ, Anderson JM, et al. A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. PLoS medicine 7 2010;7.
2. Broeck AVD, Vankelecom H, Eijsden RV, Govaere O, and Topal B. Molecular markers associated with outcome and metastasis in human pancreatic cancer. Journal of experimental & clinical cancer research : CR 1 2012;31.
3. Sergeant G, Eijsden R van, Roskams T, Duppen VV, and Topal B. Pancreatic cancer circulating tumour cells express a cell motility gene signature that predicts survival after surgery. BMC cancer 2012;12.
4. Yang S, He P, Wang J, et al. A Novel MIF Signaling Pathway Drives the Malignant Character of Pancreatic Cancer by Targeting NR3C2. Cancer research 13 2016;76:3838–50.
5. Janky R, Binda MM, Allemeersch J, et al. Prognostic relevance of molecular subtypes and master regulators in pancreatic ductal adenocarcinoma. BMC cancer 1 2016;16.
6. Yang MW, Tao LY, Jiang YS, et al. Perineural Invasion Reprograms the Immune Microenvironment through Cholinergic Signaling in Pancreatic Ductal Adenocarcinoma. Cancer research 10 2020;80:1991–2003.
7. García-García AB, Gómez-Mateo MC, Hilario R, et al. mRNA expression profiles obtained from microdissected pancreatic cancer cells can predict patient survival. Oncotarget 62 2017;8:104796–805.
8. Sakai Y, Honda M, Matsui S, et al. Development of novel diagnostic system for pancreatic cancer, including early stages, measuring mRNA of whole blood cells. Cancer science 4 2019;110:1364–88.
9. Irigoyen A, Jimenez-Luna C, Benavides M, et al. Integrative multi-platform meta-analysis of gene expression profiles in pancreatic ductal adenocarcinoma patients for identifying novel diagnostic biomarkers. PloS one 4 2018;13.
10. Caba O, Prados J, Ortiz R, et al. Transcriptional profiling of peripheral blood in pancreatic adenocarcinoma patients identifies diagnostic biomarkers. Digestive diseases and sciences 11 2014;59:2714–20.
11. Davis S and Meltzer P. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 2007;14:1846–7.
12. Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA: a cancer journal for clinicians 2 2017;67:93–9.
13. Stratford JK, Yan F, Hill RA, et al. Genetic and pharmacological inhibition of TTK impairs pancreatic cancer cell line growth by inducing lethal chromosomal instability. PloS one 4 2017;12.

14. Cong L, Liu Q, Zhang R, et al. Tumor size classification of the 8th edition of TNM staging system is superior to that of the 7th edition in predicting the survival outcome of pancreatic cancer patients after radical resection and adjuvant chemotherapy. Scientific Reports 2018 8:1 1 2018;8:1–9.

15. Hastie T, Tibshirani R, Narasimhan B, and Chu G. impute: Imputation for microarray data. R package version 1.66.0. 2021.

16. Bethesda (MD): National Library of Medicine (US) NCfBI. Gene [Internet]. 2004. URL: https://www.ncbi.nlm.nih.gov/gene/.

17. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research D1 2016;44:D733–D745.

18. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.13.0. 2021.

19. Allen JD, Wang S, Chen M, et al. Probe mapping across multiple microarray platforms. Briefings in Bioinformatics 5 2012;13:547.

20. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, and Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Research suppl_2 2008;36:W5–W9.

21. Aggarwal CC, Hinneburg A, and Keim DA. On the Surprising Behavior of Distance Metrics in High Dimensional Space.

22. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 2015;43:e47.

23. A guide to creating design matrices for gene expression experiments. 2020. URL: https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/designmatrices.html (visited on 12/22/2021).

24. Phipson B, Lee S, Majewski IJ, Alexander WS, and Smyth GK. ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. The annals of applied statistics 2 2016;10:946.

25. Yekutieli D and Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference 1999;82:171–96.

26. Blighe K, Rana S, and Lewis M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version 1.10.0. 2021. URL: https://github.com/kevinblighe/EnhancedVolcano.

27. Ulgen E, Ozisik O, and Sezerman OU. PathfindR: An R package for comprehensive identification of enriched pathways in omics data through active subnetworks. Frontiers in Genetics SEP 2019;10:858.

28. Oughtred R, Rust J, Chang C, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein science : a publication of the Protein Society 1 2021;30:187–200.

29. Nishimura D. BioCarta. https://home.liebertpub.com/bsi 3 2004;2:117–20.

30. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. Nature genetics 1 2000;25:25.

31. Carbon S, Douglass E, Good BM, et al. The Gene Ontology resource: enriching a GOld mine. Nucleic acids research D1 2021;49:D325–D334.

32. Kanehisa M and Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 1 2000;28:27–30.

33. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein science : a publication of the Protein Society 11 2019;28:1947–51.

34. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, and Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic acids research D1 2021;49:D545–D551.

35. Griss J, Viteri G, Sidiropoulos K, Nguyen V, Fabregat A, and Hermjakob H. ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. Molecular & cellular proteomics : MCP 12 2020;19:2115–24.

36. Gillespie M, Jassal B, Stephan R, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Research D1 2022;50:D687–D692.

37. Chuang HY, Lee E, Liu YT, Lee D, and Ideker T. Network-based classification of breast cancer metastasis. Molecular systems biology 2007;3.

38. Licursi V, Conte F, Fiscon G, and Paci P. MIENTURNET: An interactive web tool for microRNA-target enrichment and network-based analysis. BMC Bioinformatics 1 2019;20:1–10.

39. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research D1 2019;47:D941–D947.

40. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research D1 2018;46:D1074–D1082.

41. Krzywinski MI, Schein JE, Birol I, et al. Circos: An information aesthetic for comparative genomics. Genome Research 2009.

42. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.