# Sources for research

This document is based on a review of 65 papers. The various tools, data sets and methods are written down in each section.

# Tools

|  |  |
| --- | --- |
|  | **Source** |
| **Swiss Target Prediction** (a database for predicting the potential targets of small molecules based on their chemical structure; http://www.swisstargetprediction.ch/) | s41598-025-87286-z |
| **HERB (Herbal Medicine Database)** (a database for herbal medicine-related bioactive compounds and their targets; http://herb.ac.cn/) | s41598-025-87286-z |
| **ETCM (Encyclopedia of Traditional Chinese Medicine)** (a database of TCM with detailed information about the chemical components and therapeutic targets; http://www.tcmip.cn/ETCM/) | s41598-025-87286-z |
| **BATMAN (Bioinformatics Analysis Tool for Molecular Mechanism of Traditional Chinese Medicine)** (a bioinformatics tool designed to predict the targets of TCM compounds and explore the relationship between TCM and diseases; http://bionet.ncpsb.org.cn/batman-tcm/#/home) | s41598-025-87286-z |
| **HIT (Herb Ingredients’ Targets Database)** a database focused on herb-target interactions, providing information on the molecular interactions between herbs and their target proteins; http://www.badd-cao.net:2345/) | s41598-025-87286-z |
| **PharmMapper** a web-based tool for predicting potential protein targets of small molecules, based on their molecular structure; http://www.lilab-ecust.cn/pharmmapper/) | s41598-025-87286-z |
| The **STRING** interaction network analysis tool (https://string-db.org/) was used to create Protein-Protein Interaction (PPI) networks | s41598-025-87286-z  s41598-025-95232-2 |
| We used the online platform **CB-Dock2** (https://cadd.labshare.cn/cb-dock2/php/index.php) for this predictive analysis | s41598-025-87286-z |
| A web resource for analyzing cancer OMICS data, the  **UALCAN platform** (The University of ALabama at Birmingham CANcer data analysis portal) (http://ualcan.path.uab.edu/index.html) | s41420-024-02058-4 |
| Trajectory analysis was executed using **Monocle 2 (v2.22)** to determine cell differentiation pathways and progression states. | s41419-024-07205-4 |
| **CytoTRACE** was utilized to assess the transcriptional diversity of malignant cells based on their differentiation or stemness status (https://cytotrace.stanford.edu). | s41419-024-07205-4 |
| **STRING** https://version-11-5.string-db.org/ | s41416-024-02875-5 |
| On a practical note, there are several software  libraries and tools that are regularly utilized for graph machine  learning tasks; some of the more popular ones include: **PyTorch Geometric (PyG)** | s41416-024-02706-7 |

|  |  |
| --- | --- |
| The **Kaplan-Meier plotter database (www.kmplot.com**)  was deployed to elucidate the association between  mRNA levels of each pivotal DEG and the prognostic  outcomes of patients afflicted with TNBC. | s12967-024-05843-y |
| **Gene Expression Profiling Interactive Analysis (GEPIA**;  http://gepia.cancer-pku.cn/) represents a sophisticated  interactive web service dedicated to the analysis of RNA  sequencing expression data, incorporating 9,736 tumor  and 8,587 normal samples derived from the Cancer  Genome Atlas (TCGA) and the Genotype-Tissue Expression  (GTEx) projects. | s12967-024-05843-y |
| After identifying the shared DEGs, gene ontology  (GO) enrichment analysis and Kyoto Encyclopedia of  Genes and Genomes (KEGG) pathway enrichment analysis  were conducted using the **Database for Annotation,**  **Visualization, and Integrated Discovery (DAVID, version**  **12.0).** | s12967-024-05843-y |
| Furthermore**, Metabolite Set Enrichment Analysis**  **(MSEA) was performed via MetaboAnalyst 6.0** (https://metascape.org/gp/index.html) , aimed at elucidating  metabolic pathways distinctly altered in TNBC patients  in comparison to HC subjects. | s12967-024-05843-y |
| **Using the online tool TIDE** (http://tide.dfci.harvard.edu/), we calculated the scores for samples in both the high and  low-risk groups | s12885-025-14053-8 |
| **Tumor Immunotherapy Gene Expression Resource (TIGER)** and **ICBatlas** are comprehensive resources for integrative analysis of the transcriptome profiles related to tumor immunology. | s41746-024-01043-6 |
| **The Cancer Immunome Atlas** is a web-accessible database that characterizes the intratumoral immune landscapes and the cancer antigenomes of 20 solid cancers. | s41746-024-01043-6 |
| Exclusive and shared proteins were visualized using Venn diagrams (https://bioinformatics.psb.ugent.be/webtools/Venn/, accessed on July 30, 2024). | s41598-025-95232-2 |
| Proteomic results were filtered using **Perseus analysis software** (version 2.0.9.0)33. Proteins were retained in the abundance matrix only if they had nonzero values in at least 90% of samples from at least one experimental group | s41598-025-95232-2 |
| Alignment metrics were evaluated using samtools, **Alfred and Bamds**t (https://github.com/shiquan/bamdst) and all files successfully passed this QC step. | s41598-025-94707-6 |
| Using **the Genome Analysis Toolkit** (GATK; v4.2.0.0)66, we implemented several procedures following GATK best practices. | s41598-025-94707-6 |
| Following analysis **from GEPIA2 webtool**, we observed a positive and strong correlation between WTAP and a signature of 17 EMT genes, listed in Supplementary Table 1 (Fig. 5B). | s41420-024-02058-4 |

|  |  |
| --- | --- |
| To target TIME interactions with specific tumour phenotypes and alterations in TNBC we applied **TMArQ t**o a large TMA data set comprising over 200 TNBCs from a molecularly very well-profiled, population-representative, TNBC cohort of patients --> Staaf, J. et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nat. Med. 25, 1526–1533. https:// doi. org/ 10. 1038/ s41591- 019- 0582-4 (2019).  **https://github.com/StaafLab/TMArQ**  **https://github.com/StaafLab/tcgaBrca** | s41598-024-72306-1 |
| Based on the set of potential pyroptosis drugs obtained, we proposed to use a **biological factor-regulated neural networkmodel (BFReg-NN) t**o learn and obtain excellent compound medications | s41467-024-51980-9 |
| **Weighted gene co-expression network analysis (WGCNA)**  The accessible R package of scWGCNA (v0.0.0.9) for ST was used for coexpression analysis (https://github.com/smorabit/hdWGCNA). The Pearson Correlation Coefficient was used to assess similarities between paired mRNAs based on gene expression matrices. These similarities were then converted into adjacency matrices. | s41419-024-07205-4 |
| We subsequently employed **the “glmnet” R package** with least absolute shrinkage and selection operator (LASSO) Cox regression to mitigate gene collinearity and reduce the number of candidate gene | s12885-025-14053-8 |
| For African TNBC samples, raw paired-end reads underwent trimming using trimmomatic (v0.39)62 and were subsequently aligned to the human reference **genome GRCh38 from NCBI using the Burrows-Wheeler Aligner (BWA)-MEM algorithm (v 0.7.15)** | s41598-025-94707-6 |
| For instance, Wang et al. introduced **multi-omics graph convolutional networks (MORONET),** a multi-omics integration learning framework that employs GCNs for omics-specific learning and **incorporates a view correlation discovery network (VCDN)** to unearth intricate cross-omics correlations within the label space. | s12859-024-05989-y |
| **Li et al. crafted a multi-omics integration model based on graph convolutional network (MoGCN),** a model designed for cancer subtype classification. This model initially utilizes AEs and the **similarity network fusion (SNF**) method for dimensionality reduction of the original features and for constructing **the patient similarity network (PSN),** respectively. Subsequently, both the vector features and the PSN are fed into the GCN for further training and evaluation. Bo Yang et al. proposed **multi-reconstruction graph convolutional network (MRGCN)** for the integrative representation of multi-omics data. | s12859-024-05989-y |
| **differential sparse canonical correlation analysis network (DSCCN)**  https://github.com/hyr0771/DSCCN. | s12859-024-05749-y |
| **DIABLO:** an integrative approach for identifying keymolecular drivers from multi-omics assays | s12967-025-06446-x |
| **Boruta** features selection | s13369-024-09341-7 |
| **BASIS:** high-performance bioinformatics platform | s44303-024-00025-3 |
| **SOTIP** https://github.com/TencentAILabHealthcare/SOTIP  Multiomics spatial data | s41467-022-34867-5 |
| **ENCODE** https://genomebrowser.wustl.edu/encode/ | s41598-025-94707-6 |

# Data

|  |  |
| --- | --- |
|  | **Source** |
| The gene expression data and clinical data for TNBC were obtained from TCGA database through **UCSC XENA website** (http://xena.ucsc.edu/). | s41598-025-87286-z |
| Similarly, to provide a reference for selecting cell lines for subsequent cellular experiments, we first downloaded relevant data from the **CCLE** (https://sites.broadinstitute.org/ccle/) database | s41598-025-87286-z |
| We conducted a pan-cancer analysis using the **TISCH2** (http://tisch.comp-genomics.org) database, in order to observe the effective genes expression across different types of cancer. | s41598-025-87286-z |
| Subsequently, we used the **CCLE** (https://sites.broadinstitute.org/ccle/) database to investigate the differential expression of the effective gene across various breast cancer subtypes. | s41598-025-87286-z |
| **Kyoto Encyclopedia of Genes and Genomes (KEGG)** is a database that helps understand biological processes and disease mechanisms through genome, chemical, and pathway annotations. | s41598-025-87286-z |
| The targets of the effective gene were searched in the **PDB database** (http://www.wwpdb.org/) and saved as PDB format, and the ligands were stored in mol2 format for PA. | s41598-025-87286-z |
| For comparative purposes, CD3 IHC data from the TNBC cohort was analysed using **QuPath20** (https://qupath.github.io). | s41598-024-72306-1 |
| **University of Alabama at Birmingham Cancer (UALCAN)** data analysis portal, which hosts a collection of publicly available cancer omics, including protein data from the **Clinical Proteomic Tumor Analysis Consortium (CPTAC)** | s41598-024-54732-3 |
| **MSigDB database** (https://www.gsea-msigdb.org/gsea/msigdb/) | s41419-024-07205-4 |
| Pathway enrichment analysis in specific spots was performed using the **KEGG database via the DAVID website** (https://david.ncifcrf.gov). | s41419-024-07205-4 |
| The TNBC RNA sequencing **data of FUSCC** were downloaded  from The National Omics Data (http://www.biosino.org/node). | s13058-023-01642-3 |
| Other novel CDK inhibitors, such as dinaciclib,  PF-06873600, and trilaciclib, have been analyzed in  **clinical trials to assess the antitumor activity of TNBC**  (https://clinicaltrials.gov/). | s13045-022-01341-0 |
| were downloaded from the online analysis website **MEXPRESS** | s12967-024-05946-6 |
| Furthermore**, the Human Protein**  **Atlas (HPA) database** (https://www.proteinatlas.org)  provides an invaluable open-access repository of immunohistochemical  images, documenting a broad spectrum  of immune response observations across both neoplastic  and normal tissues | s12967-024-05843-y |

|  |  |
| --- | --- |
| Concurrently, **UALCAN**  (http://ualcan.path.uab.edu) emerges as an extensive,  intuitive web portal tailored for the analysis of cancer  OMICS data. | s12967-024-05843-y |
| Since the inauguration of **the Human Genome Project**  in 1990, the **Cancer Genome Anatomy Project** (1997),  **Cancer Genome Atlas** (2006), **Human Cell Atlas** (2016),  and many other projects have sought to build a database  or atlas of the landscape of human cancers. | s12943-024-01941-z s12943-024-01941-z |
| The validation set **GSE135565 (GPL570)**, comprising 84  TNBC samples, was sourced from the Gene Expression  Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo) . | s12935-025-03648-7 |
| A total of 1,660 genes involved in 86 metabolic pathways  were downloaded from the **Kyoto Encyclopedia of**  **Genes and Genomes (KEGG)** database (https://www.genome.jp/kegg/) | s12885-025-14053-8 |
| **Cancer Cell Line Encyclopedia (CCLE)** | s10911-023-09540-2 |
| In addition, themedical images (MRI,CT, digital histopathology, etc.) of some of these patients can be downloaded from **The Cancer Imaging Archive (TCIA)** database, enabling the multimodality analysis of immunotherapy studies. | s41746-024-01043-6 |
| After data processing, relative abundance values for all identified proteins were obtained. The database used was **UP000005640**, available at https://www.uniprot.org/proteomes/UP000005640 (taxonomy Homo sapiens), downloaded on March 9, 2023, and containing 82,861 proteins. | s41598-025-95232-2 |
| Clinicopathological characteristics for the **Sweden Cancerome Analysis Network – Breast (SCAN-B**) discovery and validation cohorts are provided in Table 1, with detailed patient characteristics available in Supplementary Data 1. | s41467-025-58158-x |
| Human triple-negative breast cancer (TNBC) **MIBI-TOF dataset13**:  Raw data were downloaded from Spatial Omics DataBase102 (https://gene.ai.tencent.com/SpatialOmics/dataset?datasetID=47) | s41467-025-57029-9 |
| All clinical material and data presented herein are available from the respective publications or upon request from the authors (see Methods section). RNAseq from MAS98.12 and MAS98.12PR are available **at Gene Expression Omnibus (GSE265955)**, while normalized RPPA data are included as Supplementary Table S7. The PDX models MAS98.12 and MAS98.12PR are available upon reasonable request. | s41416-024-02875-5 |
| Transcriptomics data from metastatic lesions in the **MET500** cohort of solid cancers, was retrieved from https://xenabrowser.net, while clinical data was taken from Supplementary Table S1 in Robinson et al. | s41416-024-02875-5 |
| A comprehensive search of the **PubMed electronic database**  was conducted using the keywords detailed in the Supplementary Materials to identify studies on omics integration published between 2018 and 2024. | s12967-025-06446-x |

|  |  |
| --- | --- |
| Our study included one multiomics cohort of patients with TNBC breast cancer who underwent surgery and adjuvant chemotherapy at **Fudan University Shanghai Cancer Center (FUSCC).** This cohort consisted of 465 patients, of whom 360 had RNA-seq data (n = 81 for the LAR subtype; n = 279 for the non-LAR subtype). Further details about this cohort have been described in our previous study [6]. | s12935-024-03313-5 |

# Studies

|  |  |
| --- | --- |
| Genomic data for the TNBC cohort supporting the conclusions of this article are available in an open repository  as described in the original study (see https://data.mendeley.com/datasets/2mn4ctdpxp/3). | s41598-024-72306-1 |
| All code for data analysis associated with the current submission is available at  https://github.com/AbhibhavS/BreastCancer-MultiOmics. | s41389-024-00521-6 |

# Methods

|  |  |
| --- | --- |
| Several transcriptome-based cell-type quantification methods for immuno-oncology have to date been reported, of which **CIBERSORT** is one of the most commonly used | s41598-024-72306-1  s12885-025-14053-8 |
| We sketched the immune profile using the TIMER deconvolution model and the suggested different algorithms (**CIBERSORT, XCELL**) in TIMER2.0. | s41419-024-07205-4 |
| log2 fold change **(LFC)**  **(pydeseq2)** | s41416-024-02875-5  s12967-024-05843-y |
| **cox regression analysis** and **LASSO** based. **Least Absolute Shrinkage and Selection Operator (LASSO)** | s12935-025-03648-7 |
| To address errors arising from inconsistent sample sizes across classifications, the **Synthetic Minority Oversampling Technique (SMOTE)** was used. SMOTE algorithm is a classic method to solve unbalanced dataset, its full name is Synthetic Minority Over-sampling Technique. SMOTE algorithm is based on the principle of balancing | s12905-024-03231-8 |
| The immune infiltration of tumors within different risk subgroups was assessed via the **ESTIMATE** algorithm, **CIBERSORT** algorithm, and **ssGSEA** algorithm. | s12885-025-14053-8 |
| For example, Vanguri et al.47 and Chowell et al. employed **Response Evaluation Criteria in Solid Tumors (RECIST)** | s41746-024-01043-6 |
| Two commonly employed methods for encoding are one-hot encoding and **BLOcks SUbstitution Matrix (BLOSUM) encoding** (Table 3). | s41746-024-01043-6 |
| Among them, **BLOSUM** is more prevalent as it offers insights into the homologies between protein sequences. | s41746-024-01043-6 |
| In addition, personalized sequencing encoding techniques utilizing ML algorithms have also gained popularity.  **These include byte pair encoding, skip-gram encoding, principal component analysis (PCA) encoding and physicochemical properties (PCP) encoding.** | s41746-024-01043-6 |
| A total of 19 mixed subtype samples and 15 compartmentalized subtype samples were utilized. We did not perform the dimensionality reduction step and retained all proteins for **running scNiche**. | s41467-025-57029-9 |
| Using **weighted gene coexpression network analysis (WGCNA)**, we identified 1230 genes showing significant and coordinated alterations in the MPS2 subtype (Supplementary Fig. S2a). | s41421-024-00715-7 |
| We followed most of the **PRISMA 2020 checklist** to define the inclusion/exclusion criteria and structure the paper. | s12967-025-06446-x |
| **Metabolite set enrichment analysis (MSEA)** | s12967-024-05843-y |
| Subsequently, advanced multivariate statistical analyses, namely principal component analysis (PCA) and **orthogonal partial least squaresdiscriminant analysis (OPLS-DA)** were conducted using  Simca-P14.1 software. | s12967-024-05843-y |
| Recently, ML models, particularly joint dimension reduction algorithms such as **negative matrix factorization (NMF), PCA, singular value decomposition (SVD), canonical correlation analysis (CCA),** have emerged as powerful tools for encoding data from diverse platforms into a shared latent space, thereby enabling effective batch effect removal | s41746-024-01043-6 |
| **Seahorse measurements** | s41419-024-07037-2 |