Data Resources and Machine Learning for Transcriptomics Data Analysis

Bahar TERCAN
Institute for Systems Biology

Asim LEBLEBICI
Dokuz Eylul University

Introduction

Different types of omics data are analyzed individually or integratively to understand the cancer biology and for better decision-making on cancer patients' diagnosis and prognosis. The analyses include but are not limited to classification of tumor (sub)types, clustering of samples, predicting prognosis and drug response, and the understanding of the information flow between different data types.

The omics data types and the relevant fields are listed in Table 1. A genome is the entire DNA of an organism. Genomics relates to all genes in contrast to genetics which considers only a limited number of genes. Transcriptomics relates to mRNAs, non-coding RNAs, and small RNAs. It is a snapshot of the samples or cell's current situation. Although the active elements are proteins, transcriptomics data can be used as a proxy to protein expression. Proteomics is the omics approach that focuses on proteins' structure, location, quantity, modifications, and functions in tissue and cell. The Human Protein Atlas (Fernandes, 2004), which started with the end of the Human-Genome Project, and The Cancer Proteome Atlas of MD Anderson Cancer Center are the major data portals created for this concept (Li et al., 2017). RNA expression levels may not always correlate with protein expression levels, activity, and post-translational modifications for various reasons; therefore, it has an important place in the holistic approach. Lipidomics is an omics approach that aims to describe lipids and the functions of lipid-forming building blocks. Metabolomics shows the genomic and transcriptome makeup in practice. Phenomics emerges as a result of the system formed by all omics structures. The phenotype (external structure) describes the entirety of the observable characteristics of a living thing. It depends on the genes that govern enzyme and protein synthesis, namely its genotype (hereditary structure) and the effects of the environmental conditions in which it lives.

Table 1. Omics Data Types and Relevant Fields in Systems Biology.

Omics	Relevant field
Genomics	DNA
Transcriptomics	RNA
Proteomics	Protein
Lipidomics	Lipid
Metabolomics	Metabolite
Phenomics	Phenotype

In the rest of this chapter, we give a detailed description of the data resources and analyses in different types of transcriptomics data that is bulk (microarray and RNA sequencing) and single-cell RNA sequencing (scRNA-seq) data. We also mention drug and perturbation datasets.

Microarray Data Analysis

The expression of thousands of genes can be measured by microarray technology at a time. Known gene sequences are placed on a glass slide (chip) and a sample is placed in contact with this glass slide, complementary base pairings produce light that identifies gene expression in the sample (*Microarray Technology*, n.d.). Microarray data analysis starts with the biological question or hypothesis and is followed by experimental design. The data is collected, RNA is extracted, fluorescent labeling is performed. The image is acquired after microarray hybridization. Following the image analysis, data preprocessing and normalization, further statistical/machine learning analysis is performed to investigate the biological question (Leung & Cavalieri, 2003).

RNA sequencing (RNA-seq) Data Analysis

RNA sequencing is performed using next-generation sequencing and counts the discrete sequence reads (Hitzemann et al., 2013). The raw RNA-seq data is stored in FASTQ files and for each read the file has an ID, read sequence, and a quality score (Chu & Corey, 2012). The low-quality reads are filtered, and the rest of the reads are mapped to the reference genome (if the reference genome is available). After splice junction detection and gene/isoform expression quantification are done, further analysis can be performed to relate the transcriptomics data to relevant phenotype(s) and answer biological questions (Chen et al., 2011). RNA sequencing does not require a model organism unlike microarray platforms (Young et al., 2012).

Single Cell RNA Sequencing Data Analysis

scRNA-seq data enables researchers to understand the tumor heterogeneity and perform analyses at the cell level which provides higher resolution compared to bulk RNA sequencing (RNA seq) data analysis. In bulk RNA-seq data, each sample is an average

expression level of all cells in the sample and represented by an expression profile. In scRNA-seq data, each cell is represented by an expression profile and different cell types like immune and tumor cells can be analyzed individually or in their cluster. Clustering analysis can be performed on single-cell data to find different cell groupings and use the signature genes for each cluster to give a clue about the biological processes that are going on in the sample.

Publicly Available Data Resources

Gene-Expression Omnibus (GEO) (Edgar et al., 2002), was originally developed to host gene expression studies but now it also provides access to other types of high throughput data like protein expression, methylation, and copy number variation(CNV). Users can find the datasets by either entering GEO Accession ID or searching for keywords from the web interface. R Bioconductor GEOquery package (Davis & Meltzer, 2007) allows users to get data from GEO and parses it into R data structures.

The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015) is a publicly available multi-omics data platform that consists of gene, exon, miRNA and protein expression, CNV, loss of heterozygosity (LOH) mutations, single nucleotide polymorphism (SNP), and DNA methylation data together with clinical features of over 20,000 samples from 33 cancer types. TCGA provides researchers to do multi-omics data analysis to characterize cancer types and subtypes, and find biomarkers for diagnosis and prognosis of cancer patients (Hoadley et al., 2014), (Zhou et al., 2020), (Liu et al., 2018), (Berger et al., 2018). The TCGA data can be retrieved from the GDC portal.

The Expression Atlas is located under European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI). It contains microarray, RNA-seq, proteomics data that meet various criteria.

In DDBJ (DNA Data Bank of Japan) center sequencing data is being collected in a joint consortium with GenBank at the NCBI and with the European Nucleotide Archive (ENA) at the EBI. Sequencing data is being collected in a joint consortium with GenBank at the NCBI and with the ENA at the EBI. The name of the common mechanism in this framework is International Nucleotide Sequence Database Collaboration (INSDC) (Fukuda et al., 2021).

Some example databases for publicly available transcriptomics data can be found in Table 2.

Table 2. Publicly Available Bulk Transcriptomics Databases

Bulk Transcriptome	Database description	Link
	ArrayExpress: Archive of Functional Genomics Data	www.ebi.ac.uk/arrayexpress
DDBJ	Biological database that collects DNA sequences	www.ddbj.nig.ac.jp
Expression Atlas	Gene expression pattern data	www.ebi.ac.uk/gxa
GEO	Public functional genomics data repository	www.ncbi.nlm.nih.gov/geo
HCMDB	Expression data of cancer metastasis	hcmdb.i-sanger.com
GDC Data Portal	TCGA - The Cancer Genome Atlas Program	portal.gdc.cancer.gov

Some of the R and Python libraries that can be used to retrieve data from Array Express, Expression Atlas, GDC Data portal - TCGA, and GEO are listed in Table 3.

Table 3. R and Python Package for Accessing Genomic Data Portal

Data Portal	R Package	Python Package
ArrayExpress	ArrayExpress	arrayexpress
Expression Atlas	ExpressionAtlas	Geneexpatlas
GDC Data Portal	TCGAbiolinks	Pytcga
GEO	GEOquery	GEOparse

Table 4 shows some of the databases that host single-cell RNA-sequencing data.

Table 4. Publicly Available Single-Cell Transcriptome Databases

Single-Cell Transcriptome	Database description	Link
GEO	Public functional genomics data repository	www.ncbi.nlm.nih.gov/geo
PanglaoDB	scRNA sequencing experiments from mouse and human	panglaodb.se
SC2disease	Single-cell transcriptome for human diseases database	easybioai.com/sc2disease
scRNASeqDB	Gene expression profiling scRNA-seq	bioinfo.uth.edu/scrnaseqdb
	Single-Cell Expression Atlas	www.ebi.ac.uk/gxa/sc
Tabula Sapiens	Human transcriptome reference at single-cell resolution	tabula-sapiens-portal.ds.czbiohub.org

Some of the R and Python environments that can be used to analyze scRNA-seq data are listed in Table 5.

Table 5. Single-Cell Data Analysis Packages

Package	Environment	Link
Bioconductor	R	www.bioconductor.org
SEURAT	R	satijalab.org/seurat
scanpy	Python	scanpy.readthedocs.io
scanpy	R	theislab.github.io/scanpy-in-R
scRNA-tools	R — Python	www.scrna-tools.org

The Drug Databases

Different types of drug databases keep drug-related information like targeted pathways/ genes or drug screening results. We mention some of the most up-to-date drug and cancer dependency databases.

The Dependency Map (DEPMAP) portal consists of CRISPR (Ledford, 2015) and RNA interference (RNAi) (Hannon, 2002) screens, Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al., 2019) multi-omics data, and drug response screening datasets like profiling relative inhibition simultaneously in mixtures (PRISM) (Corsello et al., 2020), Cancer Therapeutics Response Portal (CTRP) (Rees et al., 2016) and the Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al., 2013) to detect cancer vulnerabilities. Using these datasets, researchers can relate mutation and/or gene expression to drug or gene intervention response, detect genes that are commonly essential for cell lines or specifically essential to a particular subset of cell lines (Copeland, 2012), (Shimada et al., 2021).

Connectivity Map (CMAP) (Lamb et al., 2006) (Subramanian et al., 2017) and the Library of Integrated Network-based Cellular Signatures (LINCs) (Keenan et al., 2018), provide gene expression after a chemical compound perturbation. These resources have been used for prioritizing drug candidates and detecting the drugs that can be repurposed (Dudley et al., 2011), (Gottlieb et al., 2011).

A list of drug databases can be found in Table 6.

Table 6. Drug-Related Resources

Drug Portals	Database Description	Link
DSEA	Drug Set Enrichment Analysis	dsea.tigem.it
Gene2drug	Pathway-based Rational Drug Repositioning	gene2drug.tigem.it
DRUGBANK	Database for Drug and Drug Target Info	go.drugbank.com
DGIdb	The drug-gene interaction database	www.dgidb.org
ChEMBL	Bioactive molecules with drug-like properties database	www.ebi.ac.uk/chembl
PubChem	Collection of chemical information	pubchem.ncbi.nlm.nih.gov
PharmGKB	Pharmacogenomics knowledge resource	www.pharmgkb.org
STITCH	Interaction networks of chemicals and proteins	stitch.embl.de
LINCS L1000	Gene expression profiles for small molecules and drugs	lincsproject.org/LINCS

Analyses Performed in Cancer Research

Gene IDs/Symbol Conversion

Gene ids (EntrezID, gene name, EnsembleID, etc.) obtained as a result of the analyzes may differ. Different gene enrichment tools may require different gene name inputs. That's why there are some packages and online tools for different notations. Tools such as DAVID and UCSC Gene ID Converter can be used online and bioMart, AnnotationDBi, and ClusterProfiler packages as R packages (Roy, 2020). Table 7 shows examples of Gene ID mapping tools.

Table 7. Some Examples of Gene ID Mapping Tools.

Gene ID mapping tools	Link
HGNC	www.genenames.org
AnnotationDbi	www.bioconductor.org/packages/release/bioc/html/AnnotationDbi.html
biomaRt	bioconductor.org/packages/release/bioc/html/biomaRt.html
org.Hs.eg.db	bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html
clusterProfiler	guangchuangyu.github.io/software/clusterProfiler

SYNGO DEFINING THE SYNAPSE	www.syngoportal.org/convert.html
DAVID Bioinformatics Database	david.ncifcrf.gov/conversion.jsp
g:Profiler	https://biit.cs.ut.ee/gprofiler/convert
UCSC Gene ID Converter	www.biotools.fr/human/ucsc_id_converter

Feature Selection/Reduction and Visualization

In transcriptomics data analysis, the number of genes (features) is very high (in thousands) compared to the number of samples causing the curse of dimensionality. There are also housekeeping genes that are almost equally expressed in every cell obscuring the difference among samples/cells.

Feature selection means picking a subset of informative genes for further analysis, and it is performed using statistical tests like t-test between two groups (for exp., cancer vs normal). Feature reduction is performed to map the features into a lower-dimensional space that can capture the variance in the dataset like Principal Component Analysis (PCA) or Multidimensional Scaling (MDS). After picking or forming 2 or 3 dimensions (features), we can visualize the data in a lower space.

Classification Analysis

Classification analysis can be performed to predict healthy versus cancer tissues and different subtypes of cancer. Different subtypes are treated differently, and the prognosis may also be different, so it is important to know/predict which subtype the sample/patient belongs to. The classification analysis algorithms like decision trees, logistic regression, k-nearest neighbor algorithm (KNN), support vector machines (SVM), random forest, and artificial neural networks can be used in transcriptomics data classification. Many classification algorithms internally have feature selection mechanisms that can detect discriminative genes between subclasses or the labels of interest.

Clustering Analysis

Clustering analysis shows which samples are similar in terms of their expression profile and which genes are grouped in terms of their expression pattern over samples. Different genes can be grouped and enriched with a biologically meaningful unit like a pathway or biological process term. Similarly, similar samples are clustered together according to the gene expression profiles implying that they have shared biological events and may show similar prognosis or drug response. The clustering algorithms like hierarchical clustering, k-means, self-organizing maps (SOM) can be used for clustering transcriptomics data.

Regression Analysis

In cancer research, regression analysis is performed to predict the numerical value of a relevant phenotype like drug response. Some of the regression analysis algorithms are linear regression, support vector regression (SVR), and random forest regression.

Differential Expression Analysis

Given two groups of samples (before and after drug treatment, healthy vs. cancer), differential expression analysis can be performed to get the differentially expressed genes between two conditions. T-test and Wilcoxon test are commonly used for microarray data. There are some frequently used methods baySeq, DESeq2, EBSeq, edgeR, limma-voom, NOISeq, sleuth, and TCC-GUI for RNA-seq data analysis.

After getting the differentially expressed, either each gene is searched individually, or gene set enrichment is performed to get biological differences between the two groups.

Gene Set Enrichment Analysis

Gene Set Variation Analysis-GSVA (Hänzelmann et al., 2013) and single-sample Gene Set Enrichment Analysis-ssGSEA (Sweet-Cordero et al., 2005) are methods which are used for gene set enrichment analysis within the gene expression data (without a comparison group) and map the gene expression profile into a functional annotation profile.

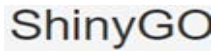
Gene Ontology (GO) (Harris et al., 2004) is a large human and machine-readable knowledge base, defined from different perspectives regarding the functions of genes. Gene ontology has been defined to cover three areas: biological processes (GO-BP), molecular function (GO-MF), and cellular components (GO-CC) (Ashburner et al., 2000; Gene Ontology Consortium, 2021).

Panther database, which is a part of the gene ontology database, is a biological database created to describe the functions of gene-protein families (Thomas et al., 2003).

Kyoto Encyclopedia of Genes and Genomes (KEGG) database maps genes, chemicals, and drugs to functional elements (pathways). The database is kept-up-to date and is a free online resource accessible to all researchers. It contains submodules such as genes, pathways, ligands, and drugs (Kanehisa & Goto, 2000).

Gene set enrichment tools are listed in Table 8.

Table 8. Functional Gene Set Enrichment Analysis Tools

Enrichment Tools	Description	Link
	Database for Annotation Visualization & Integrated Discovery	david.ncifcrf.gov
	A suite of gene list enrichment analysis tools	maayanlab.cloud/Enrichr
	Kyoto Encyclopedia of Genes and Genomes	www.genome.jp/kegg
	Online maps of metabolic and signaling pathways	www.biocarta.com
	An ontology-based pathway database coupled with data analysis tools	www.pantherdb.org/pathway
	Gene Set Enrichment Analysis	www.gsea-msigdb.org/gsea/index.jsp
	The Gene Ontology Resource	www.geneontology.org
	An annotation and analysis resource	metascape.org/gp
	ConsensusPathDB-human	cpdb.molgen.mpg.de
	Gene Set Clustering based on Functional annotation	github.com/genescf
	The Molecular Signatures Database MSigDB	www.gsea-msigdb.org/gsea/msigdb
	Network of Cancer Genes & Healthy Drivers	ncg.kcl.ac.uk
	Web server for functional enrichment analysis	biit.cs.ut.ee/gprofiler/gost
	Portal for gene list enrichment analysis	toppgene.cchmc.org
	GO enRIchment anaLysis and visuaLizAtion tool	cbl-gorilla.cs.technion.ac.il
	GO Enrichment Analysis	bioinformatics.sdstate.edu/go
	Integrated Differential Expression and Pathway analysis	bioinformatics.sdstate.edu/idep
	Intelligent prioritization and exploratory visualization of biological functions for GSEA	kobas.cbi.pku.edu.cn
	Web Gene Ontology Annotation Plot	wego.genomics.cn

	The functional Enrichment analysis tool	www.funrich.org
	Network augmented Gene Set Enrichment Analysis	www.inetbio.org/ngsea
	Collections of genes and variants associated with human diseases	www.disgenet.org
	Database of biological pathways	www.wikipathways.org/index.php/WikiPathways
	Enrichment Analysis Utilizing Active Subnetworks	github.com/egeulgen/pathfindR
	Visualization of Functional Enrichment Result	github.com/YuLab-SMU/enrichplot
	Protein-Protein Interaction Networks	string-db.org
	Network Data Integration Analysis and Visualization	cytoscape.org
	Visual web editor for cancer pathways and genomic data	www.pathwaymapper.org
	Collect and disseminate biological pathway and interaction data	www.pathwaycommons.org
	Visualization and analysis of cancer genomics data sets	www.cbioportal.org
	Gene Expression Profiling Interactive Analysis	gepia.cancer-pku.cn
	Network-Based Gene Enrichment Analysis	net-ge2.biocomp.unibo.it
	Gonet tool for interactive GO analysis	tools.dice-database.org/GOnet/
	An integrated data-mining platform for comprehensive analysis of cancer transcriptome	ualcan.path.uab.edu/home

Drug Response Data Analysis

To find the best treatment for an individual patient, the drug response can be predicted given patient gene expression data or other genetic attributes. The computational drug response analysis was performed for predicting the Area Under Curve (AUC), half-maximal inhibitory concentration (IC50), and half-maximal effective concentration (EC50) for cell line or patient sample to each drug and to relate the best possible drug

treatment to genetic characteristics like gene expression profiles and mutation status.

Chapter Summary

This chapter aims to provide introductory material to the researchers who are new to bioinformatics and computational cancer research domain and aim to work on transcriptomics data, particularly. We provide basic information about different types of omics data and more detailed explanations on transcriptomics data for cancer research. We mention the publicly available datasets and tools. We explain different analyses performed to analyze bulk and single-cell RNA sequencing transcriptomics data. We also touch upon the functional annotation tools and drug response databases that relate the analyses results to phenotypes.

Acknowledgments

Bahar Tercan has been supported by Personalized Cancer Models to Discover and Develop New Therapeutic Targets (NCI P01 CA077852) and Molecular Determinants of Cancer Therapeutic Response (NCI U01 CA217883) projects. Asim Leblebici has been supported by TUBITAK 2214-A International Research Fellowship Program for Ph.D. Students. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and TUBITAK.

Declaration of Interest

The authors declare no competing interests.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: a tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29.

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., Liu, Y., Fan, H., Shen, H., Ravikumar, V., Rao, A., Schultz, A., Li, X., Sumazin, P., Williams, C., Mestdagh, P., Gunaratne, P. H., Yau, C., Bowlby, R., … Akbani, R. (2018). A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell*, *33*(4), 690–705.e9.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527.

Chen, G., Wang, C., & Shi, T. (2011). Overview of available methods for diverse RNA-Seq data analyses. *Science China. Life Sciences*, *54*(12), 1121–1128.

Chu, Y., & Corey, D. R. (2012). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*, *22*(4), 271–274.

Copeland, R. (2012). Faculty Opinions recommendation of The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. In *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature*. https://doi.org/10.3410/f.14264142.15777280

Corsello, S. M., Nagari, R. T., Spangler, R. D., Rossen, J., Kocak, M., Bryan, J. G., Humeidi, R., Peck, D., Wu, X., Tang, A. A., Wang, V. M., Bender, S. A., Lemire, E., Narayan, R., Montgomery, P., Ben-David, U., Garvie, C. W., Chen, Y., Rees, M. G., … Golub, T. R. (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, *1*(2), 235–248.

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. In *PLOS ONE* (Vol. 12, Issue 12, p. e0190152). https://doi.org/10.1371/journal.pone.0190152

Davis, S., & Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, *23*(14), 1846–1847.

Dudley, J. T., Deshpande, T., & Butte, A. J. (2011). Exploiting drug-disease relationships for computational drug repositioning. *Briefings in Bioinformatics*, *12*(4), 303–311.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210.

Fernandes, P. (2004). HUPO (Human Proteome Organization). In *Dictionary of Bioinformatics and Computational Biology*. https://doi.org/10.1002/9780471650126.dob0945

Fukuda, A., Kodama, Y., Mashima, J., Fujisawa, T., & Ogasawara, O. (2021). DDBJ update: streamlining submission and access of human data. *Nucleic acids research*, 49(D1), D71–D75. https://doi.org/10.1093/nar/gkaa982

Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research, 49*(D1), D325–D334.

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., 3rd, Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J., Aguet, F., Weir, B. A., Rothberg, M. V., Paolella, B. R., … Sellers, W. R. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, *569*(7757), 503–508.

Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology, 7*, 496.

Hannon, G. J. (2002). RNA interference. *Nature*, *418*(6894), 244–251.

Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, *14*, 7.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., … Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, *32*(Database issue), D258–D261.

Hitzemann, R., Bottomly, D., Darakjian, P., Walter, N., Iancu, O., Searles, R., Wilmot, B., & McWeeney, S. (2013). Genes, behavior, and next-generation RNA sequencing. In *Genes, Brain, and Behavior* (Vol. 12, Issue 1, pp. 1–12). https://doi.org/10.1111/gbb.12007

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D. M., Niu, B., McLellan, M. D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A. A., Van't Veer, L. J., Lopez-Bigas, N., … Stuart, J. M. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, *158*(4), 929–944.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*(1), 27–30.

Keenan, A. B., Jenkins, S. L., Jagodnik, K. M., Koplev, S., He, E., Torre, D., Wang, Z., Dohlman, A. B., Silverstein, M. C., Lachmann, A., Kuleshov, M. V., Ma'ayan, A., Stathias, V., Terryn, R., Cooper, D., Forlin, M., Koleti, A., Vidovic, D., Chung, C., … Pillai, A. (2018). The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *Cell Systems*, *6*(1), 13–24.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., & Golub, T. R. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, *313*(5795), 1929–1935.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29.

Ledford, H. (2015). CRISPR, the disruptor. *Nature*, *522*(7554), 20–24.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M. G., Haag, J. D., Gould, M. N., Stewart, R. M., & Kendziorski, C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. In *Bioinformatics* (Vol. 29, Issue 16, pp. 2073–2073). https://doi.org/10.1093/bioinformatics/btt337

Leung, Y. F., & Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis. *Trends in Genetics: TIG*, *19*(11), 649–659.

Li, J., Akbani, R., Zhao, W., Lu, Y., Weinstein, J. N., Mills, G. B., & Liang, H. (2017). Explore, Visualize, and Analyze Functional Cancer Proteomic Data Using the Cancer Proteome Atlas. In *Cancer Research* (Vol. 77, Issue 21, pp. e51–e54). https://doi.org/10.1158/0008-5472.can-17-0369

Liu, Y., Sethi, N. S., Hinoue, T., Schneider, B. G., Cherniack, A. D., Sanchez-Vega, F., Seoane, J. A., Farshidfar, F., Bowlby, R., Islam, M., Kim, J., Chatila, W., Akbani, R., Kanchi, R. S., Rabkin, C. S., Willis, J. E., Wang, K. K., McCall, S. J., Mishra, L., … Laird, P. W. (2018). Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell*, *33*(4), 721–735.e8.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Microarray Technology. (n.d.). Retrieved November 25, 2021, from https://www.genome.gov/genetics-glossary/Microarray-Technology

Rees, M. G., Seashore-Ludlow, B., Cheah, J. H., Adams, D. J., Price, E. V., Gill, S., Javaid, S., Coletti, M. E., Jones, V. L., Bodycombe, N. E., Soule, C. K., Alexander, B., Li, A., Montgomery, P., Kotz, J. D., Hon, C. S.-Y., Munoz, B., Liefeld, T., Dančík, V., … Schreiber, S. L. (2016). Correlating chemical sensitivity and basal gene expression reveals the mechanism of action. *Nature Chemical Biology*, *12*(2), 109–116.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Shimada, K., Bachman, J. A., Muhlich, J. L., & Mitchison, T. J. (2021). shinyDepMap,

a tool to identify targetable cancer genes and their functional connections from Cancer Dependency Map data. *eLife*, *10*. https://doi.org/10.7554/eLife.57116

Sturn, A., Quackenbush, J., & Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics*, *18*(1), 207–208.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., … Golub, T. R. (2017). A Next-Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, *171*(6), 1437–1452.e17.

Su, W., Sun, J., Shimizu, K., & Kadota, K. (2019). TCC-GUI: a Shiny-based application for differential expression analysis of RNA-Seq count data. *BMC Research Notes*, *12*(1), 133.

Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J. J., Ladd-Acosta, C., Mesirov, J., Golub, T. R., & Jacks, T. (2005). An oncogenic KRAS2 expression signature was identified by cross-species gene-expression analysis. *Nature Genetics*, *37*(1), 48–55.

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, *43*(21), e140.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research*, *13*(9), 2129–2141.

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, *19*(1A), A68–A77.

Wong, P. S., Tamano, K., & Aburatani, S. (2021). Improvement of Free Fatty Acid Secretory Productivity in by Comprehensive Analysis on Time-Series Gene Expression. *Frontiers in Microbiology*, *12*, 605095.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., Ramaswamy, S., Futreal, P. A., Haber, D. A., Stratton, M. R., Benes, C., McDermott, U., & Garnett, M. J. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, *41*(Database issue), D955–D961.

Young, M. D., McCarthy, D. J., Wakefield, M. J., Smyth, G. K., Oshlack, A., & Robinson, M. D. (2012). Differential Expression for RNA Sequencing (RNA-Seq) Data: Mapping, Summarization, Statistical Analysis, and Experimental Design. In *Bioinformatics for High Throughput Sequencing* (pp. 169–190). https://doi.org/10.1007/978-1-4614-0782-9_10

Zhou, L., Huang, W., Yu, H.-F., Feng, Y.-J., & Teng, X. (2020). Exploring TCGA database for identification of potential prognostic genes in stomach adenocarcinoma. *Cancer Cell International, 20*, 264.

About The Authors

Bahar TERCAN is a postdoctoral fellow at Institute for Systems Biology, Seattle, WA, US. She received her Ph.D. from the Medical Informatics Department at the Middle East Technical University in Turkey. Her research interests include applying statistical and machine learning methods to multi-omics and drug response data for patient classification and personalized medicine.

Email: bahar.tercan@isbscience.org, ORCID: 0000-0002-5332-264X.

Asim LEBLEBICI is a Ph.D. candidate in the Department of Translational Oncology at Dokuz Eylul University in Izmir, Turkey. He is a visiting scholar at the Institute for Systems Biology, Seattle, WA, US with support of the TUBITAK 2214/A-International Research Fellowship Program for Ph.D. Students. His main research areas of interest are biostatistics, bioinformatics, artificial intelligence, and health applications. He is currently working on gene expression changes in cancer progression using microarray and RNA-seq data.

Email: asim.leblebici@isbscience.org, ORCID: 0000-0002-5197-6631

To Cite This Chapter

Tercan, B., & Leblebici, A. (2021). Data resources and machine learning for transcriptomics data analysis. In M. Ozaslan & Y. Junejo (Eds.), *Current Studies in Basic Sciences, Engineering and Technology 2021*(pp.70–85). ISRES Publishing

Copyright © 2021 by ISRES Publishing