Review Article

The landscape of cancer drug development through transcriptome analysis

Pramod Kumar Maurya, Ashutosh Mani*
Department of Biotechnology Motilal Nehru National Institute of Technology Allahabad, INDIA
*Corresponding author: amani@mnnit.ac.in

Abstract
Transcriptomics analysis reveals the study of complete transcriptome or the set of all RNA transcripts existing present in a cell. Currently, transcriptome analysis is developing as an important systems biology tool for studying the gene expression cell for the development of biomarkers and therapeutic targets related to a particular disease like cancer. This article summarizes different aspects and prospects of cancer therapy from the transcriptomic analysis viewpoint.

Keywords: Transcriptome analysis, cancer, drug discovery, drug targetting

Received on: 11.11.2025
Accepted on: 24.12.2025
Published on 26.12.2025

The genome wide transcriptomics analysis provides rational approach for targeted therapy by focusing the leading molecules involved in different signaling pathways of cell proliferation, differentiation and migration, as the complete transcriptome profiling provide the expression patterns of gene which helps to determine current cellular state and compare the gene expressions signatures during different cellular states [1,2].Since 1990s, after completion of human genome project and microarray innovations, the advances in the study of whole transcriptome began which have totally transformed the transcriptomics field [3].

The genome-wide transcriptome profiling provides comprehensive view of current cellular state and help us to understand how cellular state changes during normal to disease conditions like cancer as well as also during different treatments like drugs [4].

Several transcriptome analysis techniques are available for profiling and analyzing expression of genes as well as to obtain functional insights. There are mainly two types of general approaches that involve different high throughput techniques for transcriptome profiling at large scale of analysis including hybridization-based techniques such as microarray technology and RNA sequencing-based techniques such as RNA-seq (RNA sequencing), SAGE (serial analysis of gene expression) and MPSS (massively parallel signature sequencing). These techniques have been used to analyze highly complex transcriptome of different prokaryotic and eukaryotic genomes [5–7].

Microarray hybridization technique

Microarray technology is also known as GeneChip, DNA/RNA Chip and belongs to hybridization based technique for transcriptome profiling. A single microarray chip can identify the gene expression of whole genome on a large scale as well as differentially expressed genes with high accuracy. In this method, large number of cDNA probes (of known sequence) are fixed on a chip and the labeled sample is hybridized against it the gene expressions in sample are identified by hybridization signal straight of each probe are reference sequence [8].

The excess sample is washed off followed by scanning the micro array chip using laser and micro array chip analyzed for fluorescent signal strength by using appropriate software [9].

RNA sequencing technique

Since the last decade, RNA sequencing has become a rapidly growing approach for trancriptome analysis and it also provides quantitative measure of gene expression in

an organism or cell. NGS is an application of high throughput sequencing (HTS) technology with advent of relatively larger detection capacity and lower cost as well as in small time. RNA-seq is the sequencing technique to expand gene and genome database and ESTs tags, of species as well as to identify functional genes to develop molecular markers, to explore the spatial expression of any specific cells or tissues which is useful in diagnose disease, perform drug screening and explore drug mechanism. It a high resolution, high throughput and highly sensitive technique used to study whole transcriptome of several species. RNA-seq efficiently identify the differentially expressed alleles of transcripts pertaining to particular cellular state of disease like cancer without any reference genome information and has various applications like clinical diagnosis, disease mechanism and drug development [10,11].

High throughput transcriptomics techniques are highly sophisticated so, they need significant interpretation and analysis in order to generate meaningful outcome or information [12].

Data preprocessing

Microarray data are produced as raw high resolution image files which are further subjected to spectral analysis to obtain the processed intensity and need statistical analysis to determine the significance of obtained signal while RNA-seq data are produced as short raw RNA seq reads which are further subjected to alignment with reference genome to connect sequence read abundance with expression of certain genes and further statistical analysis is needed to convert quantitative measure of each transcript into differential gene expression [13–15].

While calculating differences in gene expression, the log transformation and data normalization are very essential for efficient comparison as well as to prove distribution of data. The comparison between normal and experimental conditions is usually calculated by using t-test which determines the statistically significant differentially expressed genes. P-value is calculated for screening the significant gene expression signatures. If the sample size is very large then FDR (false discovery rate) value is preferred rather than P-value [16,17].

Target selection

Recent advancement in high throughput omics technology such as microarray technology and NGS (next-generation sequencing) usually produce large amount of data from a single sample and the data produced is not meaningful in its raw format. There is need for data mining tools, algorithms and biomedical literature along with statistical analysis tools to identify several aspects of drug discovery such as candidate drug target, drug target-disease associations, disease relevant drug target and efficacy, druggability, safety of targets. Omics data produced using high throughput technologies is usually analyzed using bioinformatics tool to identify drug targets, identify new therapeutic targets along with their functions, to identify disease relevant genes, to build gene regulatory network and protein-protein interaction network [18].

The genes which are differentially expressed between healthy and cancer patients or the genes which are differentially expressed under a drug response or the genes which are co-expressed with other genes that are associated with cancer progression and metastasis or the genes which affect cancer progression on manipulation; are the some suitable drug targets which are interconnected with a cascade of intracellular signaling pathways followed by a specific cellular activity or reaction related to cancer progression and metastasis [19].

Revolution in omics have generated vast amount of gene expression data whose analysis helpful in identifying the candidate targets of disease like cancer by comparing the gene expression signatures

of normal and cancer or disease conditions in order to obtain the genes responsible for the development of disease, using several bioinformatics methods or tools [4,20]. The therapeutic value and efficacy of a target can be determined by studying or analyzing the gene expression signatures generated due to drug or disease in human cell lines or in *in-vivo* models and these transcriptomics signatures are important for selection of candidate drug target to identify efficiency of target during drug development [21].

Currently several databases such as The Cancer Genome Atlas (TCGA), Cancer Cell Lines Encyclopedia (CCLE) and Genomics an Drug Sensitivity in Cancer (GDSC) are generally used to analyze the regulation of coding and non-coding RNA in tumor progression with respect to drugs by using clinical drug-response data of various cell lines and tumor samples. These databases provide information on omics data of various aspects of normal and tumor samples such as methylation pattern, survival, mutation and change in copy number [22,23].

Single cell whole transcriptome analysis using high throughput techniques reveal information about gene expression of several molecular signaling pathways in cancer which may act as therapeutic targets for cancer as the inhibitors against them work as anti-cancer agents. Thus, the gene expression analysis reveal about intermediate targets of such pathways. In actual sense, the blocking of such pathways may work as anticancer therapy. The potent target is usually selected with properties to block metastasis and cell proliferation with less damage to cellular metabolism [24–26].

Validated categories of targets

Pharmacologically a target is a protein or any other bio-molecule which exhibits or allows interaction with drug like compounds such as small organic molecules, antibodies, therapeutic proteins [27].Since past two decades, several targets have been identified and characterized to successful target classes. The particular class of proteins which are more amenable as validated drug targets are G-protein coupled receptors (GPCRs) [28,29],nuclear hormone receptors [30], ion channels [31], proteases [32], Kinases [33], phosphatase and other enzymes target classes.

More often, the drug targets which have involvement in certain biological process that is critical to disease like cancer are of first choice. The successful target classes are also categorized according to their mechanisms of action such as enzymes, ion channels, transport proteins, receptor, metabolites, proteins, DNA, RNA and targets of monoclonal antibodies. Like proteins RNA also has been identified as drug target in anticancer therapy as RNA has important role in regulation of transcription and translation, catalysis, RNA splicing, peptide bond generation and protein transport. In recent researches non-coding RNA which have specialized functions, are gaining importance as suitable drug targets [34,35].

Approaches for target selection

Identification of drug target requires several molecular level aspects of a specific disease like cancers such as molecular mechanism of disease, involved metabolic pathways, analysis of gene sequences, protein structures and protein interaction network. Recent advances in bio-informatics has put forth several computation tools and approaches to identify novel drug targets from omics data and information [36,37].

Gene based approach

In gene-based approach, a particular class of drug target is selected at first, and then data mining is performed to screen sequence database in order to determine all the possible members of that class including new protein coding genes and new members [38]. Further functional annotation is done for functional annotation of genes and proteins [39].

Disease based approach

In disease based approach various attributes of a particular disease is usually selected such as etiology of disease, specific therapeutic categories in disease and various methods including gene expression analysis and gene linkage analysis. At very first, the gene expression signatures in various disease stages is compared with healthy conditions to identify the differentially expressed genes and then filter them to select candidate gene which are central to disease metabolism as well as have significant therapeutic values. Further the predictive disease model is built to model a disease mechanism, to identify specific drug compounds to inhibit the disease process and it helps to identify critical pathways in disease as well as disease related genes, proteins that regulate cellular process, hence identify putative therapeutic targets. Finally the pharmacogenomics information such as drug response, drug action, disease pathophysiology and study of SNPs helps in validating the role of drug targets in disease [40].

Gene regulatory network based approach

In gene regulatory network (GRN)-based approach, a endogenous metabolic regulatory network is reconstructed in which the drug target interact with several interactors in signaling network and also participates in several signaling pathways. The inhibition of drug target may block the activity of targets in these signaling pathways. In summary, these gene regulatory networks provide a comprehensive understanding of cellular signaling regulation and gene regulatory mechanisms related to a particular disease [41,42].

PPI network based approach

Protein expression in normal and disease condition is compared to obtain DEGs for identification of drug targets. PPI data can be used to identify drug targets where the protein interaction map helps to identify novel pathways and functional complexes of uncharacterized proteins. Once the pathway associated with disease is identified in the network then key nodes in the pathway network is identified as disease specific proteins and determined as drug targets. The experimental data is integrated with available several database using system biology approach to characterize the disease related pathways and ultimately the drug targets [43,44].

***In-silico* screening of small molecules**

Drug discovery approach contains various stages including target selection, hit identification, lead optimization and clinical studies in which hit identification and lead optimization are the steps of drug identification using computational tools and techniques related to docking studies [45].

Structure-based virtual screening is a high throughput screening (HTS) based computational technique for docking of a library of small molecules into active site of a protein which may be receptor or enzyme with several conformations for each molecule based on selected parameter. The top rank solutions of active compounds according to choice are moved forward for testing as hit identification. Virtual screening is an *in-silico* based high throughput screening method which offers a quick access of large library having millions of compounds and screen ligands on the basis of docking or scoring scheme against a biological target to reduce the number of ligands required for testing regarding early hits [46].

Databases of ligands

Several private and publicly database of ligands are available which contains collection of small molecules and compounds as well. Most of the publically available databases have been developed according to the requirement of academics to access quickly as well as purchase the chemicals. There are various ligand databases which have different attributes of the compound collection [47].

PubChem database

Pubchem (http://pubchem.ncbi.nlm.nih.gov/was developed in 2004 and currently containing three subcategories of primary ligand database: substance, compound and Bioassay. It is the largest publicly available database of molecular structures of about 92 millions of compounds.

ChEMBL database

ChEMBL (http://www.ebi.ac.uk/chembl) was launched in 2009 and is manually aerated publicly available chemical database of about 1.5 millions of distinct bioactive compounds with accompanying information regarding druglike properties such as functional assay and binding data and ADMET (Absorption, Distrtribution, Metabolism, excretion and toxicity) data, derived from multiple screening resources, literature, Pubchem bioassay and GSK (Glaxo Smith Kline) data respository.

BindingDB database

Binding DB database initially (http://www.bindingdb.org/bind/index.jsp) was launched in 1995 and till May 2017 it found to contain about 1.3 millions of data of interaction of 600,622 drug or small molecule with about 7100 proteins targets. Binding DB is a publicly available database of drug target binding data derived from literature scientific, selected Pubchem confirmation bioassays, finding from enzyme inhibition and kinetics, NMR, binding assays.

ZINC database

ZINC (http://zinc.docking.org/) database was launched in 2004 is a curated database of 120 million of publicly and available drug-like chemical compounds specially curated for virtual screening and can be purchased by pharmaceutical companies, biotech companies and research universities. Currently the compounds in Zinc database are linked to its biological target, processes and also with the providing company for purchasing of reagents.

ChemSpider database

Chemspider (http://www.chemspider.com/) is a database of about 67 million chemical structures from hundreds of data sources. The database is not available free for public use while one can download about 5000 chemical structures with their respective properties.

DrugBank database

Drug Bank (http://www.drugbank.ca/) is a publicly available comprehensive resource of chemi-informatics and bioinformatics with detailed information of drug and its drug target data (Sequence, structure and pathways). Currently DrugBank contains 9591 drug entries out of which 2037 drugs are FDA-approved small molecules, 24 drugs are FDA-approved protein/peptide drugs, 96 drugs are nutra ceuticals and more than 6000 drugs are experimental drugs.

GRAC database

GRAC (http://www.guidetoppharmacology.org/about.jsp) database is a publicly available database of about 8611 ligands from distinct resources such as approved drug, monoclonal antibodies, testing compounds, compounds against Alzheimer disease target phase I candidate's scientific literature.

Commercial ligand databases

Several chemical or ligand databases are commercially available as screening library for virtual screening analysis. Some of commercially available ligand databases are ChemBridge (http://www.chembridge-com/index.php), Maybridge (http://www.maybridge.com/), ChemDiv (http://www.chemdiv.com/products/screening-libraries/) and Life Chemicals (http://www.lifechemicals.com/).

Molecular docking studies

The molecular docking studies are used to analyze or score the interaction between a small molecules and a protein target at

atomic level by identifying the number of favorable intermolecular interactions (hydrophobic or hydrogen-bonds) between them. The ligands are ranked according to calculated free energy for favorable binding to the target receptor.

Several computation tools or docking programs are available to perform molecular docking studies such as Autodock (http://autodock.scrpps.edu/), Autodock Iva (http://viva.scripps.edu/), DOCK (http://dock.compbio.ucst.edu/), flip DOCK (http://flipdock.scripps.edu/) and glide (http://www.schrodinger.com/glide).

Conclusion

Transcriptomics based high throughput technologies have revolutionized the biomedical research in lieu of biomarker identification and target validation in the field of drug discovery. It significantly approaches towards the selection of proteins as a biomarker or therapeutic target associated with certain kind of disease like cancer. Several researchers and pharmaceutical companies have focused in drug discovery strategies by analyzing the interaction of drug with functional and regulatory proteins in which bioinformatics analysis have played significant role to explore many aspects of drug discovery. The analysis of cancer transcriptomics datasets, with innovation of bio-informatics and available algorithms usually reveal Network resulted as gene perturbations which help to identify candidate targets as well as biomarkers in cancer and provide several aspects of drug discovery approaches such as target identification, drug screening and development; leading to the development of rational approach for analyzing omics data in lieu of drug development.

References:

[1] Rhodes D R and Chinnaiyan A M 2005 Integrative analysis of the cancer transcriptome *Nature Genetics* **37** S31–7

[2] Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, Edfors F, Sanli K, Feilitzen K von, Oksvold P, Lundberg E, Hober S, Nilsson P, Mattsson J, Schwenk J M, Brunnström H, Glimelius B, Sjöblom T, Edqvist P-H, Djureinovic D, Micke P, Lindskog C, Mardinoglu A and Ponten F 2017 A pathology atlas of the human cancer transcriptome *Science* **357**

[3] Schena M, Shalon D, Davis R W and Brown P O 1995 Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray *Science* **270** 467–70

[4] Wechsler-Reya R J 2003 Analysis of gene expression in the normal and malignant cerebellum *Recent Prog Horm Res* **58** 227–48

[5] Anon Massively Parallel Signature Sequencing (MPSS)

[6] Yamamoto M, Wakatsuki T, Hada A and Ryo A 2001 Use of serial analysis of gene expression (SAGE) technology *J Immunol Methods* **250** 45–66

[7] Dong Z and Chen Y 2013 Transcriptomics: Advances and approaches *Sci. China Life Sci.* **56** 960–7

[8] Gabig M and Wegrzyn G 2001 An introduction to DNA chips: principles, technology, applications and analysis *Acta Biochim Pol* **48** 615–22

[9] Chavan P, Joshi K and Patwardhan B 2006 DNA Microarrays in Herbal Drug Research *Evid Based Complement Alternat Med* **3** 447–57

[10] Hrdlickova R, Toloue M and Tian B 2017 RNA-Seq methods for transcriptome analysis *Wiley Interdiscip Rev RNA* **8**

[11] Picelli S 2017 Single-cell RNA-sequencing: The future of genome biology is now *RNA Biol* **14** 637–50

[12] Jiang Z, Zhou X, Li R, Michal J J, Zhang S, Dodson M V, Zhang Z and Harland R M 2015 Whole transcriptome analysis with sequencing: methods, challenges and potential solutions *Cell Mol Life Sci* **72** 3425–39

[13] Durinck S 2008 Pre-processing of microarray data and analysis of differential expression *Methods Mol Biol* **452** 89–110

[14] Reilly M and Valentini D 2009 Visualisation and pre-processing of peptide microarray data *Methods Mol Biol* **570** 373–89

[15] Tao Z, Shi A, Li R, Wang Y, Wang X and Zhao J 2017 Microarray bioinformatics in cancer- a review *J BUON* **22** 838–43

[16] Hatfield G W, Hung S-P and Baldi P 2003 Differential analysis of DNA microarray gene expression data *Mol Microbiol* **47** 871–7

[17] Li X, Cooper N G F, O'Toole T E and Rouchka E C 2020 Choice of library size normalization and statistical methods for differential gene expression analysis in balanced two-group comparisons for RNA-seq studies *BMC Genomics* **21** 75

[18] Singh A J, Ramsey S A, Filtz T M and Kioussi C 2018 Differential gene regulatory networks in development and disease *Cell Mol Life Sci* **75** 1013–25

[19] Allison D 2002 STATISTICAL METHODS FOR MICROARRAY RESEARCH FOR DRUG TARGET IDENTIFICATION

[20] Kozian D H and Kirschbaum B J 1999 Comparative gene-expression analysis *Trends Biotechnol* **17** 73–8

[21] Walker M G 2001 Pharmaceutical target identification by gene expression analysis *Mini Rev Med Chem* **1** 197–205

[22] Cho W C S 2010 Omics Approaches in Cancer Research *An Omics Perspective on Cancer Research* ed W C S Cho (Dordrecht: Springer Netherlands) pp 1–9

[23] Tomczak K, Czerwińska P and Wiznerowicz M 2015 The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge *Contemp Oncol (Pozn)* **19** A68–77

[24] Van Loo P and Voet T 2014 Single cell analysis of cancer genomes *Current Opinion in Genetics & Development* **24** 82–91

[25] Saadatpour A, Lai S, Guo G and Yuan G-C 2015 Single-Cell Analysis in Cancer Genomics *Trends in Genetics* **31** 576–86

[26] Chambers D C, Carew A M, Lukowski S W and Powell J E 2019 Transcriptomics and single-cell RNA-sequencing *Respirology* **24** 29–36

[27] Drews J and Ryser S 1997 Classic drug targets *Nature Biotechnology* **15** 1350–1350

[28] Hébert T E and Bouvier M 1998 Structural and functional aspects of G protein-coupled receptor oligomerization *Biochem Cell Biol* **76** 1–11

[29] Nambi P and Aiyar N 2003 G protein-coupled receptors in drug discovery *Assay Drug Dev Technol* **1** 305–10

[30] Ruau D, Duarte J, Ourjdal T, Perrière G, Laudet V and Robinson-Rechavi M 2004 Update of NUREBASE: nuclear hormone receptor functional genomics *Nucleic Acids Res* **32** D165–7

[31] Bennett P B and Guthrie H R E 2003 Trends in ion channel drug discovery: advances in screening technologies *Trends Biotechnol* **21** 563–9

[32] Docherty A J, Crabbe T, O'Connell J P and Groom C R 2003 Proteases as drug targets. *Biochem Soc Symp* 147–61

[33] Cohen P 2002 Protein kinases--the major drug targets of the twenty-first century? *Nat Rev Drug Discov* **1** 309–15

[34] Bull S C and Doig A J 2015 Properties of Protein Drug Target Classes *PLoS One* **10**

[35] Dutt R and Madan* V G and A K 2018 Emerging Molecular Targets for Anti-Cancer Drug Design *Current Chemical Biology* **12** 88–99

[36] Jiang Z and Zhou Y 2005 Using bioinformatics for drug target identification from the genome *Am J Pharmacogenomics* **5** 387–96

[37] Benson J D, Chen Y-N P, Cornell-Kennon S A, Dorsch M, Kim S, Leszczyniecka M, Sellers W R and Lengauer C 2006 Validating cancer drug targets *Nature* **441** 451–6

[38] Smith C 2004 Drug target identification: a question of biology *Nature* **428** 225–31

[39] Gabaldón T and Huynen M A 2004 Prediction of protein function and pathways in the genome era *Cell Mol Life Sci* **61** 930–44

[40] Meltzer P S 2001 Spotting the target: microarrays for disease gene discovery *Curr Opin Genet Dev* **11** 258–63

[41] Madhamshettiwar P B, Maetschke S R, Davis M J, Reverter A and Ragan M A 2012 Gene regulatory network inference: evaluation and application to ovarian

cancer allows the prioritization of drug targets *Genome Medicine* **4** 41

[42] Xie Y, Wang R and Zhu J 2014 Construction of breast cancer gene regulatory networks and drug target optimization *Arch Gynecol Obstet* **290** 749–55

[43] Feng Y, Wang Q and Wang T 2017 Drug Target Protein-Protein Interaction Networks: A Systematic Perspective *BioMed Research International* **2017** e1289259

[44] Amala A and Emerson I A 2019 Identification of target genes in cancer diseases using protein–protein interaction networks *Netw Model Anal Health Inform Bioinforma* **8** 2

[45] Shoichet B K 2004 Virtual screening of chemical libraries *Nature* **432** 862–5

[46] Lyne P D 2002 Structure-based virtual screening: an overview *Drug Discovery Today* **7** 1047–55

[47] Wishart D S 2012 Chapter 3: Small Molecules and Disease *PLOS Computational Biology* **8** e1002805