

## Invited Mini Review

## Exploring cancer genomic data from the cancer genome atlas project

Ju-Seog Lee\*

Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA

**The Cancer Genome Atlas (TCGA) has compiled genomic, epigenomic, and proteomic data from more than 10,000 samples derived from 33 types of cancer, aiming to improve our understanding of the molecular basis of cancer development. Availability of these genome-wide information provides an unprecedented opportunity for uncovering new key regulators of signaling pathways or new roles of pre-existing members in pathways. To take advantage of the advancement, it will be necessary to learn systematic approaches that can help to uncover novel genes reflecting genetic alterations, prognosis, or response to treatments. This minireview describes the updated status of TCGA project and explains how to use TCGA data. [BMB Reports 2016; 49(11): 607-611]**

## INTRODUCTION

Human Genome Project (HGP) was successfully completed in 2003 and began a new era of genome-based medicine (1-3). Success of HGP motivated development of new technologies for genome-wide copy number alteration analysis, gene expression profiling, and better sequencing methods that can comprehensively characterize entire genomes at low cost. The comprehensive genome map has improved our understanding of complex genetic networks involved in development of human disease and allowed us to uncover functions of associated genetic elements in molecular level. There are two major branches in genomics. Structural genomics generally collect and catalogues all genetic and epigenetic elements by massive sequencing. On the contrary, functional genomics aims to uncover the functional roles of genetic or epigenetic elements in context of different biological systems. Microarrays is another most important technology for genomics that is designed to capture the expression patterns of coding/non-

coding genes, alterations in copy number and methylation status in entire genome simultaneously. Because functional activity of genes is well reflected in gene expression patterns, microarray technology has been extensively used to generate expression profile data from diseased tissues for identifying disease-associated genes and from cell lines for characterizing newly discovered genes.

In recent years, by using publicly available genomic and epigenomic data, many genome-wide analyses uncovered novel genes associated with human diseases and genes with unexpected roles in different cellular context (4-6). Thus, in-depth analysis of multiple genomic data will undoubtedly reveal novel insights into the regulation of many signaling pathways or novel key regulators of the pathways.

In this review, I will provide a short description of major progress on cancer genomics, particularly in The Cancer Genome Atlas (TCGA) project. Furthermore, I will provide description on the data generated by different platforms and analytical tools that have been developed through the progression of TCGA projects.

## THE CANCER GENOME ATLAS (TCGA)

The Cancer Genome Atlas (TCGA) Project is a multi-institutional innovative research program supported by the National Institutes of Health. TCGA was launched to facilitate the comprehensive understanding of the cancer genetics using state-of-art genomic technologies and analysis tools to catalogue all of the potential cancer drivers, identify robust prognostic and predictive biomarkers and novel druggable therapeutic targets, and uncover molecular subtypes of tumors that are different in prognosis and response to treatments. With use of several different technical platforms, TCGA currently collects and maintains many different genome-wide data including expression of coding and non-coding RNA, somatic mutations, copy-number alteration, and epigenomic data like promoter methylation. In addition to genomic and epigenomic data, it collects proteomic data by using state-of-art technology reverse phase protein arrays (RPPA). The project plans to collect multi-platform data from hundreds of tissues per each cancer type and share the data with any investigators who are interested in genome-based medicine or those who are interested in studying function of genes without any restriction

\*Corresponding author. Tel: +1-713-834-6151; Fax: +1-713-563-4235; E-mail: jlee@mdanderson.org

<https://doi.org/10.5483/BMBRep.2016.49.11.145>

Received 16 August 2016

**Keywords:** Clinical significance, Genomics, Methylation, Proteomics, The cancer genome atlas

in use of the data. In 2005, a pilot study (phase I) started aiming to test the feasibility of ideas and develop the research infrastructure by characterizing few selected cancer types that are understudied: lung squamous cell cancer, glioblastoma, and ovarian cancers (7-10). Phase 2 study was started in 2009 and expanded to additional cancer types (33 cancer types).

New oncogenes and tumor suppressor genes were identified through analysis of TCGA data. Some of findings are unexpected and showed significant association with clinical outcomes. For instance, genomic level analysis showed that non-hypermutated adenocarcinomas from the colon and rectum are almost indistinguishable at molecular level (11). Alterations in the FGFR kinase genes are very common in lung squamous cell cancer, while *KRAS* and *EGFR* mutations are most commonly activated oncogenes in lung adenocarcinoma (10). Thus, TCGA data provides an unprecedented opportunity for exploring human genome to uncover previously unrecognized genetic disruptions in smaller scale and gene-focused studies.

TCGA has established a pipeline for collecting and processing tissues from numerous source sites (tissue banks at hospitals), generation of high quality genomic and proteomic data, and distribution and analysis of the data. Most importantly, major bodies for data generation and analysis are consisted of the Genome Characterization Centers (GCCs), Genome Sequencing Centers (GSCs) and Genome Data Analysis Centers (GDACs). The GCCs aim to identify all genomic alterations in the tumors in each cancer type. Each GCC uses most advanced platform technologies to generate mRNA and miRNA expression data, DNA methylation data, and copy number alteration data. The genetic changes identified by the GCCs are further characterized by the GSCs that perform large-scale genomic sequencing using the latest sequencing technologies to identify small genomic changes that could play a role in cancer. All of the data generated by the GCCs and GSCs on the multiple genomic platform technologies from thousands of tissue samples are transferred to GDAC through Data Coordinating Center (DCC). The GDACs are responsible for analysis of the data and development of new bioinformatics tools that can facilitate use of TCGA data by the entire research community.

## TYPES OF DATA GENERATED FROM TCGA PROJECT

Six different platform data are currently generated from GCC and GSC and available to general public. These include somatic mutation data, mRNA and miRNA expression data, DNA methylation data, copy number alteration data, and proteomic data.

### Whole exome sequencing data

Majority of mutation data were generated by whole exome sequencing using second-generation DNA sequencing instruments (mostly Illumina and ABI SOLiD). Whole exome

sequencing analysis is carried out by sequencing the DNA coding for protein products, but not DNA sequences that do not directly code for proteins. However, about 10% of samples in TCGA project underwent whole genome sequencing, which sequences every base-pair of DNA and that can reveal any alteration in regulatory regions of genome.

### mRNA expression data

mRNA expression profile data were first generated by using microarray technologies from Affymetrix or Agilent, but RNA sequencing (RNA-seq) technology from Illumina was used in later stage of TCGA project. RNA-seq technology has several advantages over microarray platform as it can quantify rare and common transcripts, alternative splicing, previously unrecognized transcripts, gene fusions, as well as non-coding RNAs. It can also quantify distribution of somatic mutations and edited RNAs (12).

### microRNA expression data

microRNA (miRNA) is a small non-coding RNA (~22 nucleotides in size) that regulates other genes through post-transcriptional manner (13). miRNA expression profile data were generated by directly sequencing small molecule RNAs using RNA-seq technology from Illumina. These data were separately processed and maintained from data from mRNA-seq data as their biological and molecular characteristics are different from coding RNAs.

### DNA methylation data

DNA methylation is an epigenetic mark which is frequently associated with transcriptional activity of genes. TCGA DNA methylation data were initially generated by using Illumina 27K DNA methylation array (HumanMethylation27 containing 27,578 probes in 14,495 genes). Later, it was replaced by 450K methylation arrays (HumanMethylation450 containing 485,512 probes covering 99% RefSeq genes).

### DNA copy number alteration data

Copy number alteration is probably most frequent genetic events during the course of tumor development. Copy number data were generated by using Affymetrix SNP 6.0 arrays containing 1.8 million genetic markers, including more than 906,600 single nucleotide polymorphisms (SNPs) and more than 946,000 probes for the detection of copy number variation.

### Reverse-phase protein array (RPPA) data

RPPA is an antibody-based quantitative methods assessing hundreds of protein markers in thousands samples in a cost-effective, sensitive and high-throughput manner (14). This technology has been extensively validated for both cell line and patient samples, and its applications range from building reproducible prognostic models to assessing underlying biology associated with prognosis. Current RPPA data from TCGA

project include expression and modification of ~200 proteins.

In addition to genomic and proteomic data, TCGA data also include slide images for histopathology and details on patients information such as tumor stages, races, potential etiology, treatments and survival.

## WHERE TO GET TCGA DATA

All of genomic, proteomic, and clinical data from TCGA project were available from TCGA data portal site. However, as of July 15<sup>th</sup>, 2016, the TCGA Data Portal is no longer operational and all TCGA data now resides at the Genomic Data Commons (GDC, <https://gdc-portal.nci.nih.gov/>). While a vast majority of TCGA data in the GDC are publically available without restriction, meaning that no authentication or authorization is necessary to access it, some of the data are controlled access, meaning that special authorization process is necessary to access the data. Access to controlled data is typically granted by program-specific Data Access Committees (<https://gdc.nci.nih.gov/access-data/obtaining-access-controlled-data>). Public availability of the data is ruled by the NIH Genomic Data Sharing Policy (<https://gds.nih.gov/>). Open access data typically includes the data that cannot identify individuals such as high level genomic and proteomic data as well as most clinical and all biospecimen data elements. Controlled data includes individually identifiable data such as low level genomic sequencing data, germline variants, SNP6 genotype data, and certain clinical data elements.

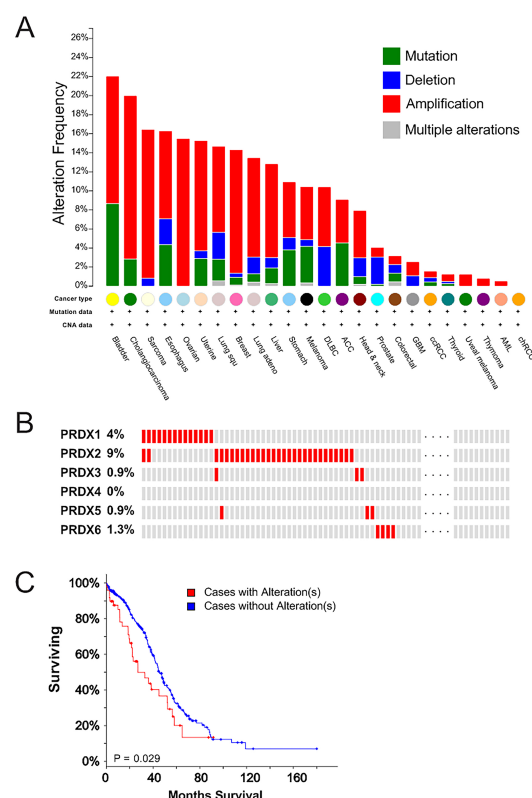
Processed high level data are also available from UCSC Cancer Genomics Browser (<https://genome-cancer.ucsc.edu/>). It offers more user-friendly processed data and limited visualization tools are also available. Histology information is also available from The Cancer Digital Slide Archive, CDSA (<http://cancer.digitalslidearchive.net/>), which provides the interactive tools for viewing and annotating diagnostic and tissue slide images from TCGA project (15). In addition to genomic, proteomic, and clinical data, TCGA also offers radiological imaging data from TCGA patients through The Cancer Imaging Archive, TCIA (<http://www.cancerimagingarchive.net>) in order to stimulate imaging phenotype-genotype study (16).

## HOW TO ANALYZE TCGA DATA

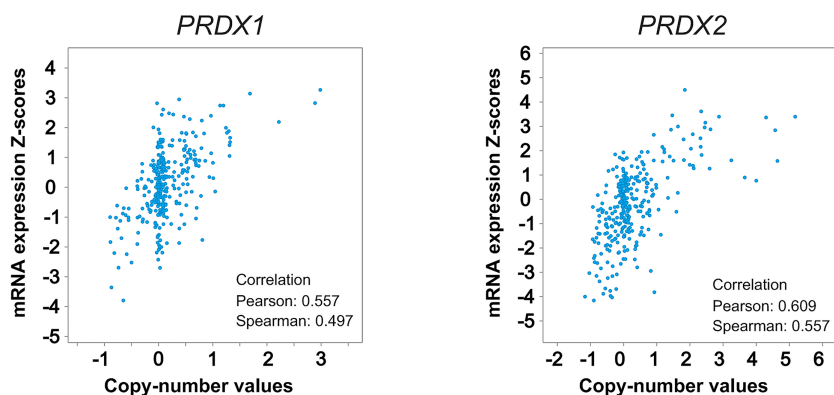
Comprehensive genomic data from large number of patients would undoubtedly improve our knowledge in understanding of cancer-related genes and their clinical relevance. However, analysis of such “big data” would require substantial skills in computational tools, statistics, and programming languages. Thus, it would be necessary to develop easy-to-use and intuitive genomic tools that can help researchers or clinicians in analysis and interpretation of all the data types in a meaningful way. TCGA provides intuitive web-based tools.

**The cBioPortal for Cancer Genomics** (<http://cbioportal.org>)

offers probably best web-based tool for beginners who have limited experience in analysis of genomic data and only wish to analyze limited number of genes (17). The cBioPortal is an open-access resource developed by investigators at the Memorial Sloan-Kettering Cancer Centre (MSKCC). It allows users to search gene(s) of interest in certain cancers or all cancers in TCGA data and provides a flexible interface to multiple data sets and easy-to-use visualization options. The cBioportal offers unique analysis and visualization tools such as MEMo (Mutual Exclusivity Modules) analysis, correlation plots for expression and copy number alteration or methylation of genes, assessing clinical relevance of genes by Kaplan-Meier plots, co-expression analysis, network analysis. In addition, it also offers highly useful OncoPrint diagrams that are an intuitive diagram of genomic alterations such as somatic mutations and copy number alterations across a set of samples. Mutationmapper provides summary diagram of all mutations on a linear protein map and links to protein 3D structure database to examine potential effects of mutations. More importantly, all analyzed data can be downloaded in table



**Fig. 1.** Visualization of analyzed data. (A) The spectrum of genetic alteration in PRDX genes in different cancer types. (B) Genetic alterations of PRDX genes in ovarian cancer. (C) Kaplan-Meier plot of patients with ovarian cancer stratified according to genetic alteration of PRDX genes.



**Fig. 2.** Scatter plots between mRNA expression and copy number alteration of PRDX1 and PRDX2 in ovarian cancer.

format for further analysis.

**The Broad GDAC Firehose** (<http://gdac.broadinstitute.org/>) is a web portal site that has been developed by the Broad Institute, aiming to deliver automated analyses of the TCGA data to general users. It provided preprocessed annotated data and association analysis across all types of data including clinical data. For example, it can provide list of genes whose copy number alteration, methylation, mRNA expression, and mutations are significantly correlated with tumor stages, survival of patients, sex, ages, or ethnic groups. Expression of genes of interest across all cancer types can be also easily assessed in firebrowse (<http://firebrowse.org/>).

**PROGeneV2** (<http://watson.compbio.iupui.edu/chirayu/proggene/database/>) provide survival analysis of patients from multiple cohorts in database (18). Users can choose either single gene or set of genes to estimate their association with prognosis of patients. Because typical molecular biologists would not have good background on statistics that is necessary to run survival analysis, it would be useful tool for them.

**Mexpress** (<http://mexpress.be/>) provides easy-to-use data visualization tool of the TCGA data including mRNA expression, DNA methylation, and clinical data (19). In addition, it also provide the correlation among data sets.

**The Cancer Proteome Atlas (TCPA)** ([http://app1.bioinformatics.mdanderson.org/tcpa/\\_design/basic/index.html](http://app1.bioinformatics.mdanderson.org/tcpa/_design/basic/index.html)) is data portal for proteomic data from TCGA project (20). It provides correlation analysis between proteins and association of proteins with prognosis of patients. In addition to TCGA data, it also provide data from established cancer cell lines.

## EXPLORATION OF GENOMIC DATA

Analysis tools from TCGA project developed to make that basic scientists without training in informatics, statistics, and

clinical knowledge can analyze the data and interpret the results. The potential involvement of genes of interest in cancer development can be easily assessed. For example, genetic alterations of peroxiredoxin family in all cancer types can be assessed through cBioPortal (Fig. 1A) and alterations of individual genes in certain cancer type (i.e., ovarian cancer) are visualized in oncoprint format (Fig. 1B). Furthermore, the clinical relevance of alteration is estimated and displayed in Kaplan-Meier plots (Fig. 1C). Clinical association of genes of interest can be further validated by using tools in PROGeneV2. Correlation between different genomic data is also readily visualized through cBioPortal and Firehose (Fig. 2).

## CLOSING REMARK

TCGA is an unprecedented powerful public resource of cancer genomic data providing researchers with a great opportunity to increase present knowledge on cancer. Multi-layer analyses performed on different platforms reflecting distinct biological characteristics provide a better understanding of cancer biology, leading to improvement in patient stratification, identification of novel prognostic or predictive markers, and finding novel potentially druggable therapeutic targets. The translation of genomic knowledge into biological insights will move these new findings to the next level and guide to a new era in data-driven molecular biology.

## ACKNOWLEDGEMENTS

This study was supported in part by National Institutes of Health grants CA150229, 2016 cycle of Institutional Research Grants from The University of Texas MD Anderson Cancer Center, 2016 cycle of Sister Institute Network Grant from The University of Texas MD Anderson Cancer Center, and a grant from The University of Texas MD Anderson Cancer Center Duncan Family Institute for Cancer Prevention and Risk Assessment.

## REFERENCES

1. Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945
2. Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860-921
3. Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. *Science* 291, 1304-1351
4. DeNicola GM, Chen PH, Mullarky E et al (2015) NRF2 regulates serine biosynthesis in non-small cell lung cancer. *Nat Genet* 47, 1475-1481
5. Park YY, Kim K, Kim SB et al (2012) Reconstruction of nuclear receptor network reveals that NR2E3 is a novel upstream regulator of ESR1 in breast cancer. *EMBO Mol Med* 4, 52-67
6. Saha SK, Parachoniak CA, Ghanta KS et al (2014) Mutant IDH inhibits HNF-4alpha to block hepatocyte differentiation and promote biliary cancer. *Nature* 513, 110-114
7. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061-1068
8. International Cancer Genome Consortium (2010) International network of cancer genome projects. *Nature* 464, 993-998
9. Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609-615
10. Cancer Genome Atlas Research Network (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489, 519-525
11. Cancer Genome Atlas Research Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330-337
12. Han L, Diao L, Yu S et al (2015) The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. *Cancer Cell* 28, 515-528
13. Yates LA, Norbury CJ and Gilbert RJ (2013) The long and short of microRNA. *Cell* 153, 516-519
14. Spurrier B, Ramalingam S and Nishizuka S (2008) Reverse-phase protein lysate microarrays for cell signaling analysis. *Nat Protoc* 3, 1796-1808
15. Gutman DA, Cobb J, Somanna D et al (2013) Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Am Med Inform Assoc* 20, 1091-1098
16. Clark K, Vendt B, Smith K et al (2013) The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 26, 1045-1057
17. Cerami E, Gao J, Dogrusoz U et al (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2, 401-404
18. Goswami CP and Nakshatri H (2014) PROGeneV2: enhancements on the existing database. *BMC Cancer* 14, 970
19. Koch A, De Meyer T, Jeschke J and Van Criekinge W (2015) MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data. *BMC Genomics* 16, 636
20. Li J, Lu Y, Akbani R et al (2013) TCPA: a resource for cancer functional proteomics data. *Nat Methods* 10, 1046-1047