A survey and evaluation of Web-based tools/databases for variant analysis of TCGA data

Zhuo Zhang,* Hao Li,* Shuai Jiang, Ruijiang Li, Wanying Li, Hebing Chen and Xiaochen Bo

Corresponding authors: Xiaochen Bo, Beijing Institute of Radiation Medicine, Beijing 100850, China. Tel.: +86 10 6693 1207; Fax: +86 10 6693 1207; E-mail: boxc@bmi.ac.cn; Hebing Chen, Beijing Institute of Radiation Medicine, Beijing 100850, China. Tel.: +86 10 6693 1207; Fax: +86 10 6693 1207; E-mail: chb-1012@163.com

*These authors contributed equally to this work.

Abstract

The Cancer Genome Atlas (TCGA) is a publicly funded project that aims to catalog and discover major cancer-causing genomic alterations with the goal of creating a comprehensive ‘atlas’ of cancer genomic profiles. The availability of this genome-wide information provides an unprecedented opportunity to expand our knowledge of tumourigenesis. Computational analytics and mining are frequently used as effective tools for exploring this byzantine series of biological and biomedical data. However, some of the more advanced computational tools are often difficult to understand or use, thereby limiting their application by scientists who do not have a strong computational background. Hence, it is of great importance to build user-friendly interfaces that allow both computational scientists and life scientists without a computational background to gain greater biological and medical insights. To that end, this survey was designed to systematically present available Web-based tools and facilitate the use TCGA data for cancer research.

Key words: The Cancer Genome Atlas, cancer, bioinformatics tools, databases, survey

Introduction

Cancer continues to be a key field of interest for human geneticists, despite the complexities involved. Moreover, despite the frequency of cancer diagnoses, scientists still do not know the causes for many cancers, or how best to treat them. More recently, high-throughput DNA sequencing [1–3] has revolutionized the study of cancer, and the use of sequencing data to assist in diagnosis is generally referred to as precision medicine [4, 5]. Thus, advances in our understanding of the cancer genome have the potential to improve precision medicine for individuals. In particular, massive efforts to undertake parallel next-generation sequencing (NGS) have revolutionized most facets of scientific discovery, and they are also responsible for many advances in the application of genomic information to human health, particularly in the field of oncology. Regarding the latter, the potential utility of these data encompasses early detection, diagnosis, prognosis ascertainment, recurrence detection, risk assessment and treatment selection for many cancers. The Cancer Genome Atlas (TCGA) project [6] represents a significant advance in cancer genomics with its aim to provide a comprehensive catalog of key genomic changes that occur in major
cancer types [7, 8]. In addition, these data facilitate more effective diagnoses, treatments and prevention. Thus, this project has remarkable potential for scientists who study cancer, and many achievements with these data have already been published [9–14]. Comprehensive genomic data from a large number of patients would undoubtedly improve our knowledge and understanding of cancer-related genes and their clinical relevance.

Currently, analyses of TCGA data are complex, with multiple steps involved (Figure 1) [15]. Moreover, to obtain meaningful biological results, each step of an analysis needs to be carefully considered, with specific tools applied to certain experimental models. To develop relevant and realistic exploration tools for available data, coordination between experimentalists and computational scientists is needed. However, life scientists may find it difficult to use many of the computational tools that have been developed by computational scientists and which require data preparation and installation and use of packaged software. This problem is further complicated by the fact that some software are platform- or operating system-specific. Conversely, computer scientists may face challenges in performing experimental validations to confirm predictions based on data analysis. Fortunately, there are Web-based tools that provide sophisticated computational solutions to help bridge this gap between wet-lab scientists and the many in silico tools available for the analysis of cancer genomic data. It is apparent that the appropriate choice of tools is not a trivial task, especially for inexperienced users. To the best of our knowledge, a comprehensive review of all available Web-based TCGA data analysis tools has not been reported. Such a review would be tremendously helpful for researchers with an interest in analyzing cancer genomic data, as it could potentially provide a guide for selecting analytical tools for a particular application. Therefore, we initiated this survey of existing Web-based tools/databases to compile a comprehensive list of programs that can perform variant analysis of TCGA data. Nonpublic tools and business tools were excluded from this survey.

A total of 61 online analysis tools for cancer genome data were surveyed, including 32 which are primarily based on TCGA data. We have listed the functions, characteristics and suitable research areas for each. In addition, we have classified these complex tools into three categories based on their different uses of cancer genome data to facilitate their application by scientists lacking relevant data analysis experience. In addition, five case studies are described from a user’s perspective, which illustrate the major international cancer research areas and apply our review to the selection of these tools. It is anticipated that these efforts will enable researchers to select and use publicly available analysis tools.

The present article is structured as follows. First, the TCGA database is introduced as a resource for understanding cancer genome data, and this is important for researchers who initially access this database. Next, all of the publicly available online analysis tools and their classifications are described. Finally, five cancer genome research questions with case studies are presented and discussed, and general recommendations for tool selection and prioritization according to the different types of cancer research are presented.

### Variant data types within TCGA

To provide a comprehensive analysis of cancer genome profiles, TCGA applied high-throughput technologies based on microarray data of nucleic acids and proteins and NGS methods that provide global analyses of nucleic acids to generate genomic, transcriptomic, epigenomic and clinical data for several cancer types. To date, there are >10 000 cases of 33 tumor types available, with 20 cancer types each having >200 cases. The TCGA Data Portal is no longer operational, and all TCGA data have been centralized at the Genomic Data Commons (GDC) (https://gdc.nci.nih.gov/). The data can be downloaded for academic use.

The identifier (ID) types listed at the GDC include: file universally unique identifier (UUID), file submitted ID (file name), case UUID, case submitted ID (case ID) and project ID. These ID types provide good identification and cataloging of a large amount of data (Table 1). The data types for each cancer include: somatic mutations, copy numbers, gene expression, microRNA (miRNA) expression, DNA methylation, reverse protein phase array (RPPA), and clinical information. Each data type includes raw and processed data that are available for public download, except for the raw sequencing files (Table 2). Somatic mutations are identified based on exome sequencing data, with exome sequencing able to detect single-nucleotide variants that are categorized as nonsynonymous or synonymous. Nonsynonymous single-nucleotide variants cause single amino acid substitutions, which may lead to altered protein function(s) or truncated proteins. Copy number alterations are generally the most frequent genetic events that occur during tumor development, and they have been determined with the Affymetrix SNP (Single Nucleotide Polymorphism) 6.0 array, which detects gains and losses in the genome. Gene expression and miRNA expression are determined with RNA sequencing (RNAseq) and miRNA sequencing analyses, respectively. The abundances of transcripts, isoforms, novel transcripts, gene fusions and noncoding RNAs can be extracted from the sequencing data. DNA methylation is determined by using the illumina platform, which provides single-nucleotide resolution of Cpgs across the vast majority of CpG islands and promoters in the genome. DNA methylation profiling provides information regarding epigenetic changes that have occurred in the genome. Protein expression is determined with RPPA [16], which is an array-based method of detecting proteins at nanogram levels. Validated antibodies are used to determine protein levels, as well as the levels of phosphorylated proteins. This analysis allows activated proteins to be detected, which would not be able to be inferred from RNA expression data. Clinical data are listed for each patient with standard metrics such as patient age, patient gender and time to death or last known contact date. For each cancer, there are specific stratification parameters. For instance, Gleeson scores are provided for prostate cancer, and Breslow index values are provided for melanoma. Overall survival, as well as progression-free survival, can be calculated and stratified according to cancer-specific staging. Generated data are also categorized not only by data type but also by data level. Raw, nonnormalized data (Level I), processed data (Level II) and segmented/interpreted data (Level III) apply to individual samples, while summarized data (Level IV) refer to analyses across sample sets. Levels III and IV data are freely available from publicly accessible databases; yet, access to lower level data (e.g. Levels I and II) requires specific permissions to be acquired and granted. Overall, each data type is comprehensive in its coverage of the genome, and it is ideal for scientists who are studying cancer to obtain an integrated analysis of TCGA data.

### Overview and categories of public Web-based tools for analyzing TCGA data

Owing to the large amount of genomic data available, specialized Web-based tools have been developed to aid clinicians and researchers in their analysis and interpretation of available data types in a meaningful way. Here, we have
Data types provided by TCGA

| DNA                              | RNA                              | Protein                           |
|----------------------------------|----------------------------------|-----------------------------------|
| Somatic mutation                | DNA methylation                  | Reverse protein phase array       |
| DNA copy number                  | Gene expression                  | Clinical data                     |
| microRNA                         |                                  |                                   |

Data preprocessing

Tumor subtype

Integrated Multi-Dimensional Analysis

Differential expression

Correlation analysis

Survival analysis

System network

Biomarker discovery

Figure 1. Overview of common analysis and some applications for multidimensional data available from TCGA.

Table 1. ID types within TCGA

| ID type            | Description                                                                 | Example                                                                 |
|--------------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------|
| File UUID          | ID of data in TCGA                                                          | 00a2364d-7385-4fa8-8562-b4f19548505a                                     |
| File Submitted ID  | ID of data uploaded to TCGA                                                 | 147f470-7440-42b8-8e3a-4e28b654916e-beta-value                         |
| Case UUID          | Sample/case ID in TCGA                                                      | 942c0088-c9a0-428c-a879-e16f8c5bdfb8                                   |
| Case Submitted ID  | ID of sample/case uploaded to TCGA, which is commonly used to represent sample/case | TCGA-CJ-4642                                                             |
| Project ID         | Project ID which sample/case belongs to                                     | TCGA-BRCA                                                               |
Auxiliary analysis.

as follows: (1) Global analysis; (2) Target analysis; and (3) divided our systematic exploration into three main categories
guish and guide the selection of these available tools, we have
tools are presented in alphabetical order. To further distin-
org/).

background. Gene expression across all cancer types can also be
easily assessed at the Firebrowse Web portal (http://firebrowse.

Broad GDAC Firehose (http://gdac.broadinstitute.org/) is a Web
portal site developed by the Broad Institute to perform auto-
mated analyses of TCGA data for general users. Preprocessed
annotated data and association analysis across all types of data,
including clinical data, are provided. For example, it can provide
a list of genes whose copy number alterations, methylation sta-
tus, mRNA expression and mutations significantly correlate
with tumor stage and patient survival, gender, age and ethnic
background. Gene expression across all cancer types can also be
easily assessed at the Firebrowse Web portal (http://firebrowse.

Global analysis

Global analysis tools allow users to examine the overall features
of cancer genomes, and they can be a valuable resource for sci-
entists who have just started to study cancer genomic data.
There are two types of global analysis tools: type I and type II.
The former only provides a global analysis, while the latter pro-
vides selected target analysis in addition to global analysis.

Type I

Broad GDAC Firehose

Cancer Landscapes

Regulome Explorer

TCGA Mbatch

Table 2. Description of data types and their access level

| Data type                     | Description                              | Access Level |
|-------------------------------|------------------------------------------|--------------|
| Aligned Reads                 | Raw sequencing data                      | Controlled   |
| Raw Simple Somatic Mutation   | Raw mutation information data            | Controlled   |
| Annotated Somatic Mutation    | Annotated mutation information data      | Controlled   |
| Aggregated Somatic Mutation   | Aggregated mutation information data     | Controlled   |
| Masked Somatic Mutation       | Transformed mutation information data    | Open         |
| Gene Expression Quantification| Gene expression data                     | Open         |
| Copy Number Segment           | Copy number information data             | Open         |
| Masked Copy Number Segment    | Transformed copy number information data | Open         |
| Methylation Beta Value        | Methylation data                         | Open         |
| Isoform Expression Quantification| Mature miRNA expression data           | Open         |
| miRNA Expression Quantification| miRNA expression data                   | Open         |
| Biospecimen Supplement        | Biospecimen information                  | Open         |
| Clinical Supplement           | Clinical information                     | Open         |

Table 3 provides a detailed list of the Web-based tools that
represent the main resources currently available for analyzing
TCGA data. Many useful indices are also indicated to facilitate
the selection of tools according to a particular need.
Furthermore, an enumeration of all back-end databases used,
as well as main analysis content, uniform resource locator
(URL), visualization type, download, batch query and applica-
tion programming interface (API) availability, is presented.
In the sections of each category below, and in Tables 3–4, the
tools are presented in alphabetical order. To further distin-
guish and guide the selection of these available tools, we have
divided our systematic exploration into three main categories
as follows: (1) Global analysis; (2) Target analysis; and (3)
Auxiliary analysis.

In Table 4, an additional 29 online resources are provided. In
these tools, TCGA data are not the major analysis object, and
many of the tools do not access TCGA data unless an upgraded
version is used.

Cancer Landscapes [17] is a Web-based tool that derives data
networks by using a newer data-driven modeling method that
is based on generalized sparse inverse covariance selection.
This tool integrates genetic, epigenetic and transcriptional data
from multiple cancers. Users are provided with interactive Web
content that visualizes constructed network models based on
statistical optimization.

canEvolve

The Web portal, canEvolve [18], stores functional genomics and
other large-scale data on cancer, including gene and miRNA ex-
pression profiles and copy number changes. This tool provides
users with easy access to information and analysis results
derived from primary, integrative and network analyses of
ongenomic data that are generated by using various func-
tional genomics platforms. The algorithms used for the analysis
pipelines were selected based on the creators’ experience in cre-
ating and using such tools to generate biologically relevant
hypotheses.

Regulome Explorer

Regulome Explorer [19] is a Web tool that integrates associ-
ations between clinical and molecular features of TCGA data.
This tool enables users to search and visualize analytical data
that are filtered according to user-specified parameters. All data
types are mapped to a circos plot with genomic coordinates.
There are other views available, which can be used to evaluate
associations, including graphs and tables. Two-dimensional
distributions of feature pairs (identified by association analysis)
are also provided. Correlation of features is represented as
edges between corresponding nodes.

TCGA Mbatch

TCGA Mbatch (http://bioinformatics.mdanderson.org/tcgam
batch/) allows the user to assess and quantify the presence of
any batch effects in a given TCGA data set via algorithms such
as hierarchical clustering and principal component analysis.
The results from these algorithms are then presented graphic-
ally as both simple and interactive diagrams. If significant batch
effects are observed in the data, the user has the option to
download data that have been computationally corrected ac-
cording to methods such as Empirical Bayes (ComBat), Median
Polish and analysis of variance.
| Name             | Databases                | Batch queries | Mutation analysis | Correlation analysis | Differential expression analysis | Pathway analysis | Kaplan–Meier plots | Pan-cancer analysis | Visualization type                  | Download | API          | URL                               |
|------------------|--------------------------|---------------|-------------------|----------------------|----------------------------------|------------------|---------------------|---------------------|-------------------------------------|----------|--------------|-----------------------------------|
| BCMD             | TCGA TCGA                | No            | No                | No                   | No                               | No               | No                  | No                  | No Image Matrix Histogram Networks Matrix | No No    | No           | http://tcga.lbl.gov:9999/         |
| Broad GDAC Firehose Cancer Landscapes | TCGA                | No            | No                | Yes                  | Yes                              | Yes              | Yes                 | Yes                 | No gyro car                            | Yes No   | No           | http://gdac.broadinstitute.org/    |
| Cancer3D         | TCGA CCLE                | No            | Yes               | No                   | No                               | No               | No                  | No                  | No Genomic coordinates Network Scatter plots/box plots 3D structure | Yes No   | http://www.cancer3d.org             |
| canEvolve        | TCGA ICGC GEO GEO GSK GTEx | Yes           | No                | Yes                  | Yes                              | Yes              | Yes                 | No                  | No Heatmap Network Plots Networks Matrix Heatmaps Image | Yes No   | http://www.canevolve.org/          |
| cbioportal       | TCGA CCLE                | Yes           | Yes               | Yes                  | No                               | No               | Yes                 | Yes                 | No Networks Plots Networks Matrix Heatmaps | Yes Yes  | http://cbioportal.org             |
| CDSA             | TCGA                     | No            | No                | No                   | No                               | No               | No                  | No                  | No Image                          | No No    | http://cancer.digitalsldearchive.net/ http://cancerlandscapes.org |
| CELLX            | TCGA CCLE GEO GEO GSK GTEx | Yes           | Yes               | Yes                  | No                               | No               | Yes                 | No                  | No Heatmap Matrix                      | Yes No   | http://www.cellx.sourceforge.net   |
| GDISC            | TCGA                     | No            | No                | Yes                  | No                               | No               | No                  | Yes                 | No Matrix Box plots | Yes No    | http://gdisc.bme.gatech.edu/       |
| GEPIA            | TCGA GTEx                | Yes           | No                | Yes                  | Yes                              | No               | No                  | No                  | No Matrix Box plots/box plots violin plots/dot plots Heatmap Matrix | Yes Yes  | http://gepia.cancer-pku.cn/         |
| IntOGen          | TCGA ICGC                | Yes           | Yes               | No                   | No                               | No               | No                  | Yes                 | No Heatmap Matrix                     | Yes No   | http://www.intogen.org/search      |
| KMplotter        | TCGA GEO GEO EGA EGA     | Yes           | No                | Yes                  | No                               | No               | Yes                 | No                  | No Matrix Heatmaps | Yes No    | http://kmplot.com/analysis/        |
| MethHC           | TCGA                     | Yes           | No                | Yes                  | No                               | No               | Yes                 | No                  | No Heatmaps | Yes No    | http://methhc.mbc.ntcu.edu.tw      |
| MEXPRESS OASISPRO | TCGA TCGA                | No            | No                | Yes                  | No                               | No               | No                  | No                  | Yes Genomic coordinates | Yes No    | http://mexpress.be/ http://tinyurl.com/oasispro |
| OncoScape        | TCGA CCLE                | Yes           | No                | Yes                  | Yes                              | No               | No                  | No                  | No Heatmap Pathway maps Matrix         | Yes No   | http://oncoscape.nki.nl/           |
| Name                  | Databases          | Batch queries | Mutation analysis | Correlation analysis | Differential expression analysis | Pathway analysis | Kaplan–Meier plots | Pan-cancer analysis | Visualization type                      | Download | API | URL                                      |
|-----------------------|--------------------|---------------|-------------------|----------------------|-------------------------------|------------------|-------------------|---------------------|----------------------------------------|-----------|----|------------------------------------------|
| PathwayMapper         | TCGA               | No            | No                | No                   | No                            | Yes              | No                | No                   | Line plots               | Yes       | Yes | http://pathwaymapper.org                |
| PROGgeneV2            | TCGA               | Yes           | No                | No                   | No                            | Yes              | No                | No                   | Pathway maps              | Yes       | No  | http://www.compbio.iupui.edu/proggene  |
| Regulome Explorer     | TCGA               | No            | No                | Yes                  | Yes                           | No               | Yes               | No                   | Circos                   | Yes       | No  | http://explorer.canceregulome.org/all_pairs/|
| TANRIC                | TCGA               | No            | Yes               | Yes                  | Yes                           | No               | Yes               | No                   | Heatmaps                 | Yes       | No  | http://ibl.mdanderson.org/tanric/design/basic/index.html |
| TCGA Clinical Explorer| TCGA               | No            | Yes               | No                   | No                            | Yes              | No                | No                   | Matrix                   | Yes       | No  | http://genomeportal.stanford.edu/pan-tcga/ |
| TCGA Mbatch           | TCGA               | No            | No                | No                   | No                            | No               | No                | No                   | Matrix                   | Yes       | No  | http://explorer.canceregulome.org/all_pairs/|
| TCGA NG-CHM           | TCGA               | No            | No                | Yes                  | No                            | Yes              | No                | No                   | Heatmaps                 | Yes       | No  | http://bioinformatics.mdanderson.org/chm |
| TCGA SpliceSeq        | TCGA               | No            | No                | No                   | No                            | No               | No                | No                   | Matrix                   | Yes       | No  | http://bioinformatics.mdanderson.org/TCGASpliceSeq/ |
| TCGA4U                | TCGA               | Yes           | Yes               | No                   | Yes                           | No               | Yes               | No                   | Heatmap                  | Yes       | No  | http://www.tcga4u.org:8888               |
| TCIA                  | TCGA               | No            | No                | No                   | No                            | No               | No                | No                   | Image                    | Yes       | Yes | http://www.cancerimagingarchive.net    |
| TCPA                  | TCGA               | No            | No                | Yes                  | Yes                           | No               | Yes               | No                   | Networks, Heatmaps        | Yes       | No  | http://www.tcpaportal.org/tcpa/         |
| UALCAN                | TCGA               | Yes           | No                | No                   | Yes                           | No               | Yes               | No                   | Heatmaps                 | Yes       | No  | http://ualcan.path.uab.edu/tutorial.html |
| UCSC Xena             | TCGA, GDC, ICGC, GTEx, TARGET TOIL | No | Yes | No | No | No | Yes | Yes | Scatter plot, Heatmap, 3D structure | Yes       | No  | http://xena.ucsc.edu/getting-started/   |
| Vanno                 | TCGA               | No            | Yes               | No                   | No                            | No               | No                | No                   | Matrix, 3D structure       | Yes       | No  | http://cgtsc.gcu.edu.tw/vanno            |
| Wanderer              | TCGA               | No            | No                | Yes                  | Yes                           | No               | No                | No                   | Genomic coordinates       | Yes       | Yes | http://maplab.cat/wanderer             |
| Zodiac                | TCGA               | Yes           | No                | No                   | No                            | No               | Yes               | No                   | Matrix, Circular network  | No        | No  | http://www.compgenome.org/zodiact2/     |
**Table 4. Additional databases and Web servers**

| Name                  | Content                                           | URL                                      |
|-----------------------|---------------------------------------------------|------------------------------------------|
| AnimalTFDB 2.0        | Animal transcription factors                      | http://bioinfo.life.hust.edu.cn/AnimalTFDB/ |
| ArrayMap              | A resource for genomic copy number profiles of human tumors | http://www.arraymap.org                  |
| BloodSpot             | Gene expression profiles and transcriptional programs for healthy and malignant hematopoiesis | www.bloodspot.eu                         |
| BrecAN-DB             | Break point profiles of cancer genomes            | http://brecandb.igp.res.in               |
| Cancer RNA-Seq Nexus  | Phenotype-specific transcriptome profiling         | http://syslab4.nchu.edu.tw/CRN           |
| ccnSAR                | Cancer research and drug discovery                | http://cansar.icr.ac.uk/                 |
| CGWB                  | Cancer cell metabolism gene                       | http://bioinfo.mc.vanderbilt.edu/ccmGDB   |
| ChimerDB 3.0          | A computational platform to integrate clinical tumor mutation profiles with the reference human genome | https://cgwb.nci.nih.gov/                 |
| ChIPBase v2.0         | Transcriptional regulatory networks of noncoding RNAs and protein-coding genes | http://rana.sysu.edu.cn/chipbase/         |
| CMPD                  | Cancer mutant proteome database                   | http://cgbc.cgu.edu.tw/cmpd             |
| COSMIC                | Somatic mutations in human cancer                 | http://cancer.sanger.ac.uk               |
| dbDEMC 2.0            | Differentially expressed miRNAs in human cancer   | http://www.picb.ac.cn/dbDEMC             |
| DBTSS                 | Transcriptome, epigenome and genome sequence variation data  | http://dbtss.hgc.jp/                     |
| DiseaseMeth           | Human disease methylation database                | http://bioinfo.hrbru.edu.cn/diseasemeth/  |
| DriverDbv2            | Human cancer driver gene                          | http://nsg.ym.edu.tw/driverdb           |
| LNCEditng             | A database for functional effects of RNA editing in IncRNAs | http://bioinfo.life.hust.edu.cn/LNCEditng/ |
| IncRNAASNp            | SNPs in IncRNAs                                   | http://bioinfo.life.hust.edu.cn/IncRNAASNp/   |
| miRTarBase 2016       | miRNA database                                    | http://miRTarBase.mbc.ntu.edu.tw/        |
| Mutagene              | Cancer genetic heterogeneity                       | https://www.ncbi.nlm.nih.gov/projects/mutagene/ |
| MutationAligner       | Recurrent mutation hot spots                      | http://www.mutationaligner.org           |
| mutLBSgeneDB          | Mutated ligand-binding site gene Database          | http://chaobilbioinfo.org/mutLBSgeneDB   |
| NetGestalt            | Multidimensional omics data                       | http://www.netgestalt.org                |
| Oncotator             | Cancer variant annotation tool                    | http://www.broadinstitute.org/oncotator/  |
| PhosphoSitePlus       | Protein posttranslational modifications           | http://www.phosphosite.org/              |
| POSTAR                | Posttranscriptional regulation                    | http://postar.ncrnlab.org/               |
| RBP-Var               | Functional variants involved in regulation mediated by RNA-binding proteins | http://www.rbp-var.biols.ac.cn/          |
| WebGestalt 2017       | Enrichment analysis                               | http://www.webgestalt.org                |
| YM500v2               | MiRNAs for human cancer                           | http://nsg.ym.edu.tw/ym500/              |

**TCGA Next-Generation Clustered Heatmaps**

TCGA Next-Generation Clustered Heatmaps (TCGA NG-CHM) (http://bioinformatics.mdanderson.org/chm) is a tool that creates interactive large-scale visualizations of data based on a classic heat map approach. The user is able to zoom and pan across a heatmap, alter its color scheme, generate production quality PDFs and access rows, columns and individual heatmap entries that are related to statistics, databases and other information. TCGA NG-CHM also provides pathway and gene ontology (GO) information, chromosomal interactive ideograms, rapid recoloring, high-resolution graphics output and links to public information resources (e.g. cBioPortal) regarding genes, proteins, pathways and drugs.

**The Cancer Proteome Atlas**

The Cancer Proteome Atlas (TCPA) [20] is a portal for accessing proteomic data available from TCGA project, which includes extensively validated antibodies for nearly 200 proteins and phosphoproteins. Correlation analyses can be performed between proteins and for associations between proteins and patient prognosis. In addition to TCGA data, TCPA can also access data from established cancer cell lines and can provide validation of findings from TCGA RPPA data through independent sample cohorts.

**Type II**

**MethHC**

MethHC [21] is a database that integrates a large collection of DNA methylation data and mRNA/miRNA expression profiles in human cancers, and also identifies correlations between DNA methylation and mRNA/miRNA expression data from TCGA. The methylation data span gene regions [e.g. promoter, enhancer, 5’ untranslated region (UTR), first exon, gene body and 3’ UTR] and CpG islands (e.g. regions, shelves and shores). MethHC also provides methylation patterns of different cancers with hierarchical clustering graphs. Users can easily obtain 250 hypermethylated genes, 250 hypomethylated genes and 250 of the most differentially methylated genes for particular cancer types.

**Omics Analysis System for Precision Oncology**

Omics Analysis System for Precision Oncology (OASISPRO) [22] is an online platform that is designed to mine quantitative omics information from TCGA. This tool can effectively visualize patients’ clinical profiles and other omics data and can evaluate prediction performance by using held-out test sets. OASISPRO is also rather unique in that is uses a machine learning method.
OncoScape

OncoScape [23] is an R package software for cancer gene prioritization that has a Web portal for interactive analyses. OncoScape can access five complementary data types across 11 different cancers to identify new candidate cancer genes and explore cancer aberrations by using a fusion of genomic data. For example, with this tool, molecular profiling data of two groups of samples can be compared to identify genes that exhibit significant differences. OncoScape can also perform analyses of gene expression, DNA copy number, DNA methylation, mutation and short hairpin RNA (shRNA) knock-down data. Users can explore candidate genes for each cancer type and upload their own gene list to obtain a detailed aberration profile. OncoScape can provide box plots that show log changes in gene expression (e.g. copy number data) for tumor and normal samples, and can provide an overview of the prioritization scores in genomic regions and pathway diagrams.

TCGA Clinical Explorer

TCGA Clinical Explorer [24] enables the cancer research community and others to explore clinically relevant associations inferred from TCGA data. With its accessible Web and mobile interfaces, users can examine queries and test hypotheses regarding genomic/proteomic alterations across a broad spectrum of malignancies. This tool also summarizes TCGA clinical parameters and translates these data into a list of clinically relevant cancer drivers, including genes, miRNAs and proteins. All analyses include 25 cancer types and 18 clinical parameters. Users can query TCGA data in multiple ways, including searching for clinically relevant gene/protein/miRNAs by name, cancer type or clinical parameter; profiling genomic/proteomic changes according to clinical parameters in a cancer type; and testing two-hit hypotheses.

TCGA SpliceSeq

TCGA SpliceSeq [25] investigates cross-tumor and tumor-normal alterations in mRNA splicing patterns of TCGA RNASeq data. Percent Spliced In (PSI) values for splice events derived from 33 different types of tumor samples, including available adjacent normal samples, have been loaded into this tool. As a result, users can investigate the splicing pattern of a gene of interest in a variety of tumor types. TCGA SpliceSeq also provides knowledge discovery via genome-wide PSI splice event searches to locate significant splice variations among tumor types, or between tumor and normal tissue, and these splicing data can be downloaded for integrative analyses.

Target analysis

Target analysis is the category of public Web-based tools that is most often used by researchers. These tools allow researchers to investigate a target of interest with in-depth analyses of gene(s) and miRNAs.

Cancer3D

Cancer3D [26] is a public database that analyzes cancer missense mutations in the context of protein structures. It also allows users to explore two different cancer-related problems at the same time, e.g. drug sensitivity/biomarker identification and prediction of cancer drivers. In addition, somatic missense mutations from TCGA and Cancer Cell Line Encyclopedia (CCLE) can be mapped onto >24 300 structures, as well as onto 1300 potential novel protein domains.

cBioPortal

The cBioPortal [27] for Cancer Genomics offers one of the best Web-based tools for beginners who have limited experience analyzing genomic data and only want to analyze a limited number of genes. The cBioPortal is an open-access resource that was developed at the Memorial Sloan Kettering Cancer Center (MSKCC) for the visualization, analysis and download of large-scale cancer genomics data sets. It allows users to search gene(s) of interest in certain cancers or among all cancers in TCGA data, while providing a flexible interface for working with multiple data sets and easy-to-use visualization options. The cBioPortal also offers correlation plots for expression and copy number alterations or methylation of genes, an ability to assess clinical relevance of genes with Kaplan-Meier plots, co-expression analysis and network analysis. Additionally, the portal facilitates interactive explorations of custom data sets with access to OncoPrinter and MutationMapper Web tools. OncoPrint diagrams provide intuitive diagrams of genomic alterations such as somatic mutations and copy number alterations across a set of samples, while MutationMapper provides a summary diagram of mutations on a linear protein map that has links to a database of three-dimensional (3D) protein structures for the user to examine the potential effects of the mutations identified.

Gene Expression Profiling Interactive Analysis

Gene Expression Profiling Interactive Analysis (GEPIA) [28] is a Web-based tool that rapidly delivers customizable functionalities based on TCGA and GTEx data. GEPIA provides key interactive and customizable functions that include differential expression analysis, profiling plotting, correlation analysis, patient survival analysis, similar gene detection and dimensionality reduction analysis.

IntOGen

IntOGen [29] is a Web platform that can identify cancer drivers across tumor types and perform a systematic analysis of the most up-to-date large data sets of tumor somatic mutations. The IntOGen pipeline integrates the results of tumor genome studies conducted with different mutation-calling workflows, and it is scalable to hundreds of thousands of tumor genomes. This tool can also compute the frequency of mutation for individual genes and/or pathways within a project or cancer site, detect a subset of novel candidate drivers and download driver mutations from previous studies.

KMplotter

KMplotter is an online tool that draws survival plots, which can be used to assess the relevance of gene expression levels on clinical outcome for treated and untreated cancer patients. Data are derived from gene expression, relapse-free survival and overall survival data that are downloaded from Gene Expression Omnibus (GEO) (Affymetrix microarrays only), European Genome-phenome Archive (EGA) and TCGA. Specifically, survival analyses can be performed for mRNAs from four cancer types (breast, ovarian, lung and gastric) and for miRNAs from two cancer types (breast and liver) [30].

MEXPRESS

MEXPRESS [31] is a straightforward and easy-to-use Web tool that integrates and visualizes gene expression, DNA methylation and clinical TCGA data on a single-gene level. It also provides correlation among data sets, has a unique set of features that are easy to use, and it can integrate visualizations of
different data types for hundreds of samples. Currently, the developer of this tool is also looking into updating MEXPRESS to use the new repository of TCGA data.

PROGeneV2
PROGeneV2 [32] is a tool that allows researchers to use publicly available data to study prognostic implications of genes of interest in multiple cancers. For example, this tool can be used to generate plots of survival analysis data according to gene expression profiles of target genes in selected data sets from multiple cancers. Furthermore, either single genes or sets of genes can be used to estimate their association with prognosis of patients. This tool can also provide survival analyses for miRNA and PROGmiRV2 [33], and its usage is similar to that of PROGeneV2.

TANRIC
TANRIC [34] is an open-access resource for investigating the function and clinical relevance of long noncoding RNAs (lncRNAs) in cancer. TANRIC provides three analysis modules that enable users to examine the function and underlying mechanisms of lncRNAs. It can characterize the expression profiles of lncRNAs in large patient cohorts of up to 20 cancer types, including TCGA, CCLE and other independent data sets. Users can examine whether lncRNAs exhibit differential expression profiles between tumors and normal samples, or among tumor subgroups. Possible correlations between lncRNAs and patient survival time can also be identified, while correlations between lncRNAs and various molecular data for protein-coding and miRNA genes can be explored.

TCGA4U
TCGA4U [35] is a tool that provides visualizations of the relationship between cancer genomics alterations and clinical data. This Web tool can apply four types of data (somatic mutation, DNA methylation, gene expression and copy number variants) for specific genes or gene lists to five types of cancer (lung squamous cell carcinoma, breast invasive carcinoma, colon adenocarcinoma, lung adenocarcinoma and rectum adenocarcinoma). By using specific genes and gene lists to analyze genomic alterations and characterize the molecular characteristics of cancers, cancer genomic mining is performed with the following outputs: potential driver genes are identified, GO term maps are generated and survival analyses are conducted.

UALCAN
UALCAN [36] is an interactive Web portal for researchers to facilitate the study of gene expression variation and survival associations across tumors. All data are from the TCGA database. It can help researchers identify survival associations that involve any gene of interest, across different cancer types as well as cancer subtypes as defined by various clinicopathologic features. The analysis results can be downloaded in several formats. Thus, this online tool can aid cancer biologists and clinicians in the identification of novel diagnostic and therapeutic targets, and investigate the gene expression and its disease association in any particular cancer.

UCSC Xena
UCSC Xena (http://xena.ucsc.edu/getting-started/) is a new tool that has been developed by the UCSC Cancer Browser, and it can analyze and visualize a user’s private functional genomics and data sets in the context of public and shared genomic/phenotypic data sets. The Xena platform consists of a set of federated data hubs and the Xena browser. The latter integrates across the hubs, thereby providing one location at which to analyze and visualize data. There is a large public Xena hub that currently hosts an expanding set of searchable data from several large consortiums, including TCGA, GDC, International Cancer Genome Consortium (ICGC), Genotype-Tissue Expression (GTEx), Therapeutically Available Research to Generate Effective Treatments (TARGET) and Scalable and Efficient Workflow Engine (TOIL). Dynamic Kaplan–Meier survival analyses can also be performed to assess survival according to certain parameters, and these data can be presented as visual spreadsheets, scatter plots and bar graphs.

Wanderer
Wanderer [37] is a public Web server that is able to explore and interpret gene-associated expression profiles and DNA methylation for all of the cancer types available at TCGA. This tool also provides normal–tumor paired comparisons in the form of graphs and comprehensive tables.

Zodiac
Zodiac [38] is a search engine and computational tool that obtains multiple features of gene networks, including copy number, gene expression, methylation, mutation, miRNA and some protein expression data, to describe molecular interactions for approximately 200 million pairs of genes. Zodiac then integrates existing knowledge about cancer genetic interactions with a Bayesian graphical model of TCGA data to produce updated and data-enhanced knowledge. The results are organized into a comprehensive database that allows customized searches to be performed. Zodiac also provides data processing and analysis tools that allow users to customize prior networks and update genetic pathways of interest. Furthermore, this tool can be used to identify gene interactions, to discover potential drug targets, and to identify potential genetic aberrations such as gene fusions.

Auxiliary analysis
The third category of public Web-based tools translates TCGA data into an online resource that is easily accessed, browsed and downloaded. These data can help users complement their experimental results, or they can provide additional proof and explanation of their research for comprehensive biological discoveries.

BCMD
BCMD [39] is a platform that can be used to represent and characterize tumor histology, and it can additionally provide an integrated analysis with clinical outcome. Data and intermediaries for a number of tumor types are available, and it has an interface that allows for panning and zooming of whole-mount tissue sections with or without overlaid segmentation results for quality control.

CDSA
CDSA [40] provides interactive tools for viewing and annotating diagnostic and tissue slide images of different tumor types from TCGA project. Currently, it hosts >20 000 whole-slide images from 22 cancer types. This searchable resource provides users with an opportunity to identify and explore sets of images according to particular genomic, pathologic or clinical criteria. Thus, CDSA represents a valuable resource for the fields of imaging and pathology.
Patterns in global alteration profiles

Various alteration phenotypes have been observed in cancer cells. One of the most conspicuous of these is the mutation phenotype [46], where tumor cells exhibit an abnormally high mutation burden. Somatic mutation patterns have been described for: malignant melanoma [47], small cell lung carcinoma [48], acute lymphoblast leukemia [49], colorectal cancer [10], kidney cancer [50] and lung cancer [51]. These studies have demonstrated the value of whole-genome sequencing for obtaining global alteration profiles and analyzing the patterns observed.

Broad GDAC Firehose is a good Web-based tool for exploring global alteration profiles. In this portal, the cancer type for mutation analysis can be directly specified, and a wealth of content analysis data can be selected. The latter includes aggregate analysis, correlation analysis with mutation and several mutation analysis methods including MutSig v2.0 (Figure 2A). The online results give users access to both standard data packages (right column), and standard analyses suite (left column). Analyses results may also be accessed from the unified reports. Furthermore, the results of an analysis can be downloaded in a PDF format, and this online tool has an interactive API for fine-grained querying of results via the Web. Another tool, Cancer Landscapes, can provide a high-performance statistical network modeling of multiple human cancers. Tumors are used to represent different cancer types and shapes represent different types of data. Users first select one of the multicancer modes for further analysis. The system then loads the model where different data types and cancers are represented as specific shapes and colors. Users can click on nodes to view the details of a local network and associated pathways (Figure 2B). In this exploration view, users can switch between different data types, adjust the optimization parameters and organize the network.

Stratification of cancer patients

It is necessary for cancers to be properly classified to achieve effective clinical management and meaningful laboratory
investigations of underlying cancer mechanisms. While tumors may appear similar when examined with conventional diagnostic methods, they may look markedly different from a molecular viewpoint, and this can lead to differences in outcome and treatment response. Therefore, the molecular features of tumors can be used to stratify patients to support more accurate clinical and therapeutic decisions.

Molecular stratification of tumors has been an important area of cancer research over the past few decades [58–61], and the studies performed have underscored the heterogeneous and complex nature of cancer subgroups. Molecular subtypes can be identified through different data types, including gene expression, copy number, DNA methylation and mutation data. Moreover, an integrated analysis is needed based on the different cancer characteristics. Currently, there are no tools that can directly provide stratification because of the complexity of this analysis. As a result, scientists need to combine many data types and clinical features for a comprehensive assessment.

Figure 2. Two explorations of global alteration profile patterns as provided by publicly accessible Broad GDAC Firehose and Cancer Landscape Web tools. (A) This window view displays the user interface of Broad GDAC Firehose where users can choose a specific mutation analysis method. (B) This window provides network modeling of multiple cancers and data sets as indicated by the data sets and data types that were selected at the far right in Cancer Landscapes.
OASISPRO can identify genes that are strongly associated with tumor stage by applying user-selected machine learning algorithms to omic data and evaluating prediction performance by using held-out test sets (Figure 4). However, OASISPRO only focuses on the classification of clinical phenotypes, and it cannot synthesize a variety of data types. Users have to strictly follow the settings of the tool for step-by-step selection. In addition, OASISPRO can only use a single clinical feature parameter for analysis. Thus, OASISPRO would be useful for preliminary analyses and scientific hypotheses.

Correlation with multiple molecular features

Studies of correlations among multiple molecular features can provide valuable insight into complex biological systems. Individual data sets that include genomic, epigenomic, transcriptomic or proteomic information are highly informative, and the integration of these data sets offers an exciting potential to answer many long-standing questions. For example, integrated analyses of transcriptomic, proteomic and metabolomic data have helped researchers better understand global regulatory processes and complex metabolic networks in cancer [62, 63].

Many tools can provide correlation analyses for various molecular features. In fact, more than half of the tools included in our study can conduct a correlation analysis. However, the major function of Regulome Explorer is to perform correlation analyses. Users can select a data set to load and get the genome-level view for the correlation between different data types. This tool provides both circos plots and network representations of correlations between multi-omics features, and it includes nine data types (Figure 5). It can map multi-omics features onto genomic locations for further systems biology analyses. Moreover, the parameters of a correlation can be adjusted according to a filter panel that is presented on the right side of the Web server and both network maps and detailed data tables of correlations are provided.

Survival analysis

Identification of prognostic biomarkers, which may include genes, polymorphisms, mutations, micromolecules or epigenetic regulators, represents a major advance in the field of cancer genomics. Cancer research predominantly focuses on specific patient populations for biomarker identification. Gene signatures have been developed specific for prognostication in particular subtype of a cancer, for instance, a subgroup of population treated with a specific drug. To date, gene signatures of prognostic importance have been reported for breast cancer [64, 65], colon cancer [66, 67], liver cancer [68], lung cancer [69, 70] and pancreatic cancer [71]. Generally, the primary end point of prognostic assessment is survival analysis, and patient groups are divided into good or bad prognosis groups based on weighted or unweighted expression of individual genes or groups of genes. This type of analysis provides a rationale for mechanistic studies, followed by therapeutic targeting.

Web-based tools can be used to identify and expand prognostic biomarker targets in different cancers based on the publicly available data these tools have compiled. In addition to providing easy-to-perform prognostic analyses for multiple cancers, they can also be important hypothesis-generating tools for researchers working on topics related to cancer. Here, PROGgeneV2 and KMplotter were selected to perform test analyses. Users can select gene(s), cancer type, survival measure and the data set for specific parameter settings. The results of the survival analysis conducted by PROGgeneV2 are presented in a KM plot (Figure 6), while KMplotter could not provide results because of an insufficient number of TCGA samples. These results demonstrate that the parameters and data sources for Web-based tools are not exactly the same, as the number of lung adenocarcinoma samples obtained from TCGA differed between the two analysis programs. Therefore, users need to carefully consider the data being subjected to analysis and select appropriate parameters.

Usage advice

Our study has identified three categories of all online TCGA analysis tools. The user can make preliminary screening according to their own work needs. All tools in each category have their unique features that we described above. It can also be identified based on different cancer genomic research questions as we described in case studies. Finally, the user need to combine their study, such as data sources, data types, analytical methods and research purposes, to determine the specific tool for further analysis. The following are specific suggestions for different analysis of TCGA data.

Mutation analysis

There are 10 online tools (Broad GDAC Firehose, Cancer3D, cbioportal, CELLX, IntOGen, TANRIC, TCGA Clinical Explorer, TCGA4U, UCSC Xena and Vanno) that can perform mutation analysis. In general, we recommend cbioportal because this tool contains a variety of cancer types and multiple visualizations, and it is powerful but easy to use.

Correlation analysis

There are 17 online tools (Broad GDAC Firehose, Cancer Landscapes, canEvolve, cbioportal, CELLX, GDISC, GEPIA, MethHC, MEXPRESS, OASISPRO, Regulome Explorer, TANRIC, TCGA Clinical Explorer, TCGA NG-CHM, TCPA, Wanderer and Zodiac) that can perform correlation analysis. In general, we recommend Broad GDAC Firehose from Broad institute of MIT and Harvard, which has a variety of analysis algorithms available to users.

Differential analysis

There are 12 online tools (Broad GDAC Firehose, canEvolve, cbioportal, CELLX, GEPIA, MEXPRESS, OncoScape, TANRIC, TCGA4U, TCPA, UALCAN and Wanderer) that can perform differential analysis. In general, we recommend GEPIA, an analysis tool for gene expression profiling. Differential analysis is this tool’s main analysis function, and the online analysis interface is simple and easy to understand.

Pathway analysis

There are eight online tools (Broad GDAC Firehose, Cancer Landscapes, canEvolve, MethHC, OncoScape, PathwayMapper, Regulome Explorer and TCGA NG-CHM) that can perform pathway analysis. We recommend Broad GDAC Firehose and OncoScape; the former has a variety of analysis methods, and the latter is more simple and intuitive.

Survival analysis

There are 16 online tools (Broad GDAC Firehose, Cancer Landscapes, canEvolve, cbioportal, CELLX, GDISC, GEPIA, KMplotter, OASISPRO, PROGgeneV2, TANRIC, TCGA Clinical Explorer, TCGA4U, TCPA, UALCAN and UCSC Xena) that can
perform survival analysis. If users want to perform this single analysis, we recommend PROGgeneV2, which has a wide range of data sources and adjustable parameters for survival analysis.

Pan-cancer analysis
There are eight online tools (Broad GDAC Firehose, Cancer Landscapes, cibiportal, IntOGen, Regulome Explorer, TCGA NGCHM, UCSC Xena and Zodiac) that can perform pan-cancer analysis. In general, we recommend cibiportal and Cancer Landscapes. The former has a large number of samples from pan-cancer studies and powerful analytical capabilities. The latter has combined pan-cancer model for analysis.

Discussion
The functionalities of a cancer can be better characterized by integrating information from different modalities. TCGA data were collected by using a number of different modalities, and data for several tumor types are available. Consequently, TCGA data represents a valuable resource for researchers to advance
their understanding of various cancers and to facilitate the realization of precision medicine in oncology. Multilayer analyses performed on different platforms reflect distinct biological characteristics, and these provide a better understanding of cancer biology. As a result, improvements in patient stratification, identification of novel prognostic or predictive markers and the identification of novel therapeutic targets can be achieved. However, integrating information from different modalities to obtain a comprehensive analysis remains a prodigious challenge [72].

Figure 4. Views of interface windows in OASISPRO. (A) The stepwise selection of parameters for conducting a classification of clinical phenotypes is shown. (B) This window presents the input variables and results obtained from a representative analysis.
Figure 5. A representative window of the results provided by Regulome Explorer for a correlation analysis. This figure displays the main user interface, including the option for using multiple data types.

Figure 6. A representative survival plot generated with PROgeneV2. TP53 gene expression was applied to a lung adenocarcinoma data set from TCGA.
Many bioinformatics tools that are compatible with TCGA data have been developed for basic scientists who do not have extensive training in informatics, statistics or clinical knowledge. Correspondingly, the wealth of available tools for analysis and interpretation of data reflects the importance of TCGA and the dynamic nature of the field of data analysis. Therefore, the goal of this review was to provide a comprehensive introduction to publicly available Web-based resources and tools to help researchers select the appropriate tool for their needs. Thus, we organized these resource tools into three categories: global analysis, target analysis and auxiliary analysis. In addition, we provided five case studies, which demonstrate classic analysis methods along with corresponding tools. However, none of these tools completely replaces advanced computational and statistical methodologies. Moreover, it remains the responsibility of cancer researchers to understand this vast amount of data and translate it into testable hypotheses and novel diagnostic and therapeutic options for the clinic. To this end, it is our hope that the current survey will afford researchers the confidence needed to extend their current knowledge of cancer genomics and its complex details and networks to identify new approaches and targets for cancer treatment and prevention.

Key Points

- TCGA provides unprecedented opportunities to increase our knowledge of cancer and facilitate the realization of precision medicine in oncology.
- The most comprehensive and currently available Web servers and resources that assist with TCGA data analysis are enumerated.
- The tools are classified based on their different analysis modes to help researchers select the appropriate tool for their work.
- Case studies are provided, which further illustrate the roles of TCGA data analysis in five predominant areas of cancer research.

Supplementary Data

Supplementary data are available online at https://academichub.com/bib.

Funding

This work was supported by grants from the Major Research Plan of the National Key R&D Program of China (grant number 2016YFC0901600), the National Natural Science Foundation of China (grant number U1435222).

References

1. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860–921.
2. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304–51.
3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:431–45.
4. Roychowdhury S, Iyer MK, Robinson DR, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med 2011;3(111):111ra121.
5. Garraway LA. Genomics-Driven Oncology: framework for an Emerging Paradigm. J Clin Oncol 2013;31(15):1806–14.
6. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013;45:1113–20.
7. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–8.
8. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474:609–15.
9. Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRα, IDH1, EGFR, and NF1. Cancer Cell 2010;17:98–110.
10. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7.
11. Brennan CW, Verhaak RG, McKenna A, et al. The somatic genomic landscape of glioblastoma. Cell 2013;155(2):462–77.
12. Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. Cell 2015;161:1681–96.
13. Ceccarelli M, Barthel FP, Malta TM, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell 2016;164:550–63.
14. Cancer Genome Atlas Research Network, Albert Einstein College of Medicine, Analytical Biological Services, et al. Integrated genomic and molecular characterization of cervical cancer. Nature 2017;543:378–84.
15. Schadt EE, Linderman MD, Sorenson J, et al. Computational solutions to large-scale data management and analysis. Nat Rev Genet 2010;11(9):647–57.
16. Spurrier B, Ramalingam S, Nishizuka S. Reverse-phase protein lysate microarrays for cell signaling analysis. Nat Protoc 2008;3(11):1796–808.
17. Kling T, Johansson P, Sanchez J, et al. Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. Nucleic Acids Res 2015;43(15):e98.
18. Samur MK, Yan Z, Wang X, et al. canEvolve: a web portal for integrative oncogenomics. PLoS One 2013;8(2):e56228.
19. Madhavan S, Gusev Y, Natarajan TG, et al. Genome-wide multi-omics profiling of colorectal cancer identifies immune determinants strongly associated with relapse. Front Genet 2013;4:236.
20. Li J, Lu Y, Akbani R, et al. TCPA: a resource for cancer functional proteomics data. Nat Methods 2013;10(11):1046–7.
21. Huang WY, Hsu SD, Huang HY, et al. MethHC: a database of DNA methylation and gene expression in human cancer. Nucleic Acids Res 2015;43:D856–61.
22. Yu KH, Fitzpatrick MR, Pappas L, et al. Omics analysis system for precision oncology (OASISPRO): a web-based omics analysis tool for clinical phenotype prediction. Bioinformatics 2017;34(2):319–20.
23. Schlicker A, Michaut M, Rahman R, et al. OncoScape: exploring the cancer aberration landscape by genomic data fusion. Sci Rep 2016;6(1):28103.
24. Lee H, Palm J, Grimes SM, et al. The Cancer Genome Atlas Clinical Explorer: a web and mobile interface for identifying clinical-genomic driver associations. Genome Med 2015;7(1):112.
25. Ryan M, Wong WC, Brown R, et al. TCGASpliceSeq a compendium of alternative mRNA splicing in cancer. Nucleic Acids Res 2016;44(D1):D1018–22.
26. Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res* 2015;43(D1):D968–73.

27. Cerami E, Gao J, Dogrusoz U, et al. The CBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;2(5):401–4.

28. Tang Z, Li C, Kang B, et al. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res* 2017;45(W1):W98–102.

29. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, et al. IntOGen: mutations identifies cancer drivers across tumor types. Nat Methods 2013;10(11):1081–2.

30. Lanzczyk A, Nagy A, Bottai G, et al. miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. *Breast Cancer Res Treat* 2016;160(3):439–46.

31. Koch A, De Meyer T, Jeschke J, et al. MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data. *BMC Genomics* 2015;16(1):636.

32. Goswami CP, Nakshatri H. PROGgeneV2: enhancements on the existing database. *BMC Cancer* 2014;14:570.

33. Goswami CP, Nakshatri H. PROGmiR: a tool for identifying prognostic miRNA biomarkers in multiple cancers using publicly available data. J Clin Bioinform 2012;2(1):23.

34. Li J, Han L, Roebuck P, et al. TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res* 2015;75(18):3728–37.

35. Huang ZZ, Duan HL, Li HM. Identification of gene expression pattern related to breast cancer survival using integrated TCGA datasets and genomic tools. *Biomed Res Int* 2015;2015:878546.

36. Chandrothakar DS, Bashel B, Balasubramanya SAH, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* 2017;19(8):649–58.

37. Diez-Villanueva A, Mallona I, Peinado MA. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenetics Chromatin* 2015;8:22.

38. Zhu Y, Xu Y, Helseth DL, et al. Zodiac: a comprehensive depiction of genetic interactions in cancer by integrating TCGA data. *J Natl Cancer Inst* 2015;107.

39. Chang H, Han J, Borowsky A, et al. Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association. *IEEE Trans Med Imaging* 2013;32:670–82.

40. Gutman DA, Cobb J, Somanna D, et al. Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Am Med Inform Assoc* 2013;20(6):1091–8.

41. Ching KA, Wang K, Kan Z, et al. Cell Index Database (CELLX): a web tool for cancer precision medicine. *Pac Symp Biocomput* 2015;10–19.

42. Spainhour JCG, Lin J, Qiu P. GDISC: a web portal for integrative analysis of gene-drug interaction for survival in cancer. *Bioinformatics* 2017;33:1426–8.

43. Bahceci I, Dogrusoz U, La KC, et al. PathwayMapper: a collaborative visual web editor for cancer pathways and genomic data. *Bioinformatics* 2017;33(14):2238–40.

44. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 2013;26(6):1045–57.

45. Huang Pj, Lee CC, Tan BC, et al. Vanno: a visualization-aided variant annotation tool. *Hum Mutat* 2015;36(2):167–74.

46. Stephens PJ, McBride DJ, Lin ML, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 2009;462(7276):1005–10.

47. Pleasance ED, Cheetham RK, Stephens PJ, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 2010;463(7278):191–6.

48. Pleasance ED, Stephens PJ, O’Meara S, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 2010;463(7278):184–90.

49. Holmfeldt L, Wei L, Diaz-Flores E, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet* 2013;45:242–52.

50. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 2013;499:43–9.

51. Seo JS, Ju YS, Lee WC, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res* 2012;22(11):2109–19.

52. De Keersmaecker K, Atak ZK, Li N, et al. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat Genet* 2013;45:186–90.

53. Ding L, Ley TJ, Larson DE, et al. Clonal evolution in relapsed acute myeloid leukemia revealed by whole-genome sequencing. *Nature* 2012;481(7382):506–10.

54. Nik-Zainal S, Alexandrov LB, Wedge DC, et al. Mutational processes molding the genomes of 21 breast cancers. Cell 2012;149(5):979–93.

55. Stephens PJ, Tarpey PS, Davies H, et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature* 2012;486(7403):400–4.

56. Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science* 2008;321(5897):1807–12.

57. Kan Z, Zheng H, Liu X, et al. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res* 2013;23:1422–33.

58. Hurst CD, Alder O, Platt FM, et al. Genomic subtypes of non-invasive bladder cancer with distinct metabolic profile and female gender bias in KDM6A mutation frequency. *Cancer Cell* 2017;32(5):701–15.e707.

59. Study identifies subtypes of pediatric high-grade gliomas. *Cancer Discov* 2017;7:1359–60.

60. Jusakul A, Cutcutache I, Yong CH, et al. Whole-genome and epigenomic landscapes of etiologically distinct subtypes of cholangiocarcinoma. *Cancer Discov* 2017;7:1116–35.

61. Northcott PA, Buchhalter I, Morrissey AS, et al. The whole-genome landscape of medulloblastoma subtypes. *Nature* 2017;547:311–17.

62. Jiang N, Hjorth-Jensen K, Hekmat O, et al. In vivo quantitative phosphoproteomic profiling identifies novel regulators of castration-resistant prostate cancer growth. *Oncogene* 2015;34(21):2764–76.

63. Mach N, Ramayo-Caldas Y, Clark A, et al. Understanding the response to endurance exercise using a systems biology approach: combining blood metabolomics, transcriptomics and miRNomics in horses. *BMC Genomics* 2017;18(1):187.
64. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160–7.
65. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351(27):2817–26.
66. Tan IB, Tan P. Genetics: an 18-gene signature (ColoPrint(R)) for colon cancer prognosis. Nat Rev Clin Oncol 2011;8(3):131–3.
67. Yi JM, Dhir M, Van Neste L, et al. Genomic and epigenomic integration identifies a prognostic signature in colon cancer. Clin Cancer Res 2011;17(6):1535–45.
68. Budhu A, Forgues M, Ye QH, et al. Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. Cancer Cell 2006;10(2):99–111.
69. Lu Y, Lemon W, Liu PY, et al. A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. PLoS Med 2006;3(12):e467.
70. Hsu YC, Yuan S, Chen HY, et al. A four-gene signature from NCI-60 cell line for survival prediction in non-small cell lung cancer. Clin Cancer Res 2009;15(23):7309–15.
71. Sergeant G, van Eijden R, Roskams T, et al. Pancreatic cancer circulating tumour cells express a cell motility gene signature that predicts survival after surgery. BMC Cancer 2012;12(1):527.
72. Hoadley KA, Yau C, Wolf DM, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 2014;158:929–44.