HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2015 April 07.

Published in final edited form as:

Nat Methods. 2014 June ; 11(6): 599–600. doi:10.1038/nmeth.2956.

TCGA-Assembler: An Open-Source Pipeline for TCGA Data Downloading, Assembling, and Processing

Yitan Zhu[1], **Peng Qiu**[2], and **Yuan Ji**[1,3,*]

[1] Center for Biomedical Research Informatics, NorthShore University HealthSystem, Evanston, IL 60201

[2] Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332

[3] Department of Health Studies, The University of Chicago, Chicago, IL 60637

To the Editor: The Cancer Genome Atlas (TCGA) has been generating multi-modal genomics, epigenomics, and proteomics data for thousands of tumor samples across more than 20 types of cancer. While the access to most level-1 and -2 TCGA data is restricted, the entire level-3 TCGA data as well as some level-1 clinical data (e.g., survival and drug treatments) are publicly available. Included in the public data are genome-wide measurements of different genetic characterizations, such as DNA copy number, DNA methylation, and mRNA expression for the same genes, providing unprecedented opportunities for systematic investigation of cancer mechanisms at multiple molecular and regulatory layers [1-3]. Few tools of integrative data mining for TCGA are present, partly due to lack of tools to acquire and assemble the large scale TCGA data. Specifically, the level-3 TCGA data are stored as hundreds of thousands of sample- and platform-specific files, accessible through HTTP directories on the servers of TCGA Data Coordinating Center (DCC) [4]. Navigating through all of the files manually is impossible. Although Firehose [5] nicely assemble and publish TCGA data, it does not share the program code for data assembly. Currently the community does not have access to open-source data retrieving tools for automatic and flexible data acquisition, hence severely hindering the progress in systemic data integration and reproducible computational analysis using TCGA data. To meet these challenges, we introduce TCGA-Assembler, a software package that automates and streamlines the retrieval, assembly, and processing of public TCGA data. TCGA-Assembler equips users the ability to produce Firehose-type of TCGA data, with open-source and freely available program script. TCGA-Assembler opens a door for the development of data-mining and data-analysis tools that generate fully reproducible results, including data acquisition.

[*] yji@health.bsd.uchicago.edu.

AUTHOR CONTRIBUTIONS

Y.Z., P.Q., and Y.J. initialized the original idea. Y.Z. and Y.J. designed and implemented the software, and Y.Z., P.Q., and Y.J. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Supplementary information is available in the online version of the paper (doi)

TCGA-Assembler consists of two modules (**Fig. 1a**), both written in R (http://www.r-project.org). Module A streamlines data downloading and quality check, and module B processes the downloaded data for subsequent analyses (**Supplementary Methods**). In particular, module A takes advantage of the informative naming mechanism of TCGA data file system (**Supplementary Fig. 1**) and applies a recursive algorithm to retrieve the URLs of all data files. By string matching on the URLs, module A allows users to download most of TCGA public data (**Supplementary Table 1**) across genomic features and cancer types. For each genomics feature (such as gene expression from RNA-Seq) a data matrix combining multiple samples (**Fig. 1b**) is produced, with rows representing genomics units (such as genes) and columns representing samples. Module B provides convenient and important data preprocessing functions, such as mega-data assembly, data cleaning, and quantification of various measurements. For users interested in integrative analysis [6], a mega data matrix (**Fig. 1c**) is required that matches different types of genomics measurements for the same genes across samples. Module B provides a function "*CombineMultiPlatfomData*" to fulfill this requirement (**Supplementary Methods**), which involves intricate data-matching steps to overcome the feature-labeling discrepancies caused by different lab protocols and biotechnologies in the experiments. Other data-processing functions are also provided to facilitate downstream analysis (**Supplementary Methods**).

Other big data tools for TCGA are available [5, 7, 8]. In particular, level-3 TCGA data can also be obtained from Firehose [5] at the MIT Broad Institute in the same format as in **Fig. 1b**, one for each cancer type and genomics platform. Module A of TCGA-Assembler not only provides the same type of data matrices, but also distributes R functions and associated computer program that produce the data matrices. Equipped with the open-source tool, users will be independent and control what and when TCGA data will be acquired locally. More importantly, quantitatively advanced users may integrate our open-source programs with downstream data analysis tools to realize reproducible and automated data analysis for TCGA. Unique to TCGA-Assembler is module B that provides critical functions for data cleaning and processing. For example, the mega data table (**Fig. 1c**) can be obtained with a single function, behind which substantial efforts have been directed to ensure the validity of process, such as to check and correct gene symbol discrepancies. Lastly, TCGA-Assembler is fully compatible with Firehose in that the data processing functions in Module B can directly process data files downloaded from Firehose. This compatibility is crucial to those who want to take advantage of both software pipelines.

TCGA-Assembler will remain freely available and open-source. In the future, more data processing and analysis functions will be continuously added to TCGA-Assembler based on user feedback and new research needs. The authors request acknowledgment of the use of TCGA-Assembler in published works.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

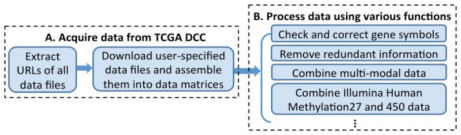
ACKNOWLEDGMENTS

This work was supported by grants from the US National Cancer Institute (R01 CA132897 and R01 CA163481).

REFERENCES

1. The Cancer Genome Atlas. https://tcga-data.nci.nih.gov/tcga/.
2. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61–70. [PubMed: 23000897]
3. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455(7216):1061–1068. [PubMed: 18772890]
4. Open-access HTTP directory on the data server of TCGA Data Coordination Center. https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/
5. Firehose website, https://confluence.broadinstitute.org/display/GDAC/Home
6. Xu, Y., et al. A Bayesian graphical model for integrative analysis of TCGA data.. IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS); 2-4 Dec. 2012; Washington, DC: p. 135-138.
7. Robbins DE, et al. A self-updating road map of The Cancer Genome Atlas. Bioinformatics. May 15; 2013 29(10):1333–40. doi: 10.1093/bioinformatics/btt141. Epub 2013 Apr 17. [PubMed: 23595662]
8. Gao J, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling. Apr 2.2013 6(269):l1. doi: 10.1126/scisignal.2004088.

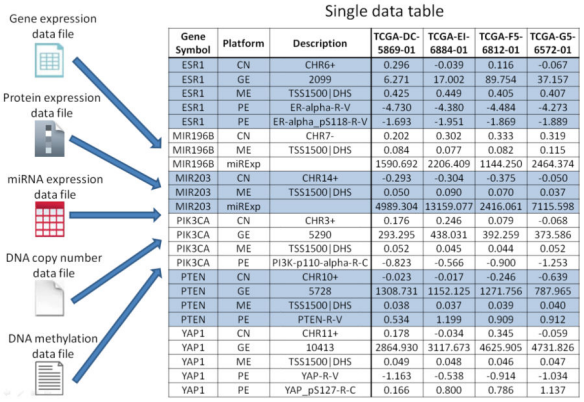
a



b

Composite.Element.REF	TCGA-AG-4001-01A-11-1933-20	TCGA-AG-3742-01A-12-1932-20	TCGA-AG-3582-01A-21-1932-20	TCGA-DC-6681-01A-13-1935-20	TCGA-EI-6882-01A-13-1936-20	TCGA-AG-3600-01A-21-1932-20	TCGA-AH-6547-01A-13-1935-20	TCGA-EI-6884-01A-13-1936-20
YWHAE|14-3-3_epsilon-M-C	-1.8843	-1.8550	-1.6834	-1.6465	-1.5266	-1.5456	-1.5488	-1.4178
EIF4EBP1|4E-BP1-R-V	-1.2975	-1.9200	-1.9261	-2.2372	-1.4419	-1.9354	-1.9977	-1.4680
EIF4EBP1|4E-BP1_pS65-R-V	-1.0005	-0.9388	-0.9430	0.7651	-1.0645	-0.8211	-0.7161	-1.0697
TP53BP1|53BP1-R-C	1.6575	1.3552	0.7499	0.5979	0.5280	0.9310	0.2020	0.8005
ACACA|ACC1-R-C	2.1955	1.8028	2.4295	1.1060	2.0590	2.8239	1.1217	1.7943
ACACA ACACB|ACC_pS79-R-V	-1.5132	-1.8540	-1.6152	-2.3356	-1.4146	-1.3199	-2.2345	-1.4082
NCOA3|AIB1-M-V	-0.7736	-0.2837	-1.2775	-1.1706	-0.8696	-1.0097	-1.1870	-0.8168
AKT1 AKT2 AKT3|Akt-R-V	1.7266	1.3942	1.2189	2.2763	1.6699	1.7089	0.9182	0.9438
AKT1 AKT2 AKT3|Akt_pS473-R-V	-2.6873	-2.2461	-1.9694	-2.4396	-2.8490	-2.4964	-1.8215	-2.2098

c



Single data table

Gene Symbol	Platform	Description	TCGA-DC-5869-01	TCGA-EI-6884-01	TCGA-F5-6812-01	TCGA-G5-6572-01
ESR1	CN	CHR6+	0.296	-0.039	0.116	-0.067
ESR1	GE	2099	6.271	17.002	89.754	37.157
ESR1	ME	TSS1500|DHS	0.425	0.449	0.405	0.407
ESR1	PE	ER-alpha-R-V	-4.730	-4.380	-4.484	-4.273
ESR1	PE	ER-alpha_pS118-R-V	-1.693	-1.951	-1.869	-1.889
MIR196B	CN	CHR7-	0.202	0.302	0.333	0.319
MIR196B	ME	TSS1500|DHS	0.084	0.077	0.082	0.115
MIR196B	miRExp		1590.692	2206.409	1144.250	2464.374
MIR203	CN	CHR14+	-0.293	-0.304	-0.375	-0.050
MIR203	ME	TSS1500|DHS	0.050	0.090	0.070	0.037
MIR203	miRExp		4989.304	13159.077	2416.061	7115.598
PIK3CA	CN	CHR3+	0.176	0.246	0.079	-0.068
PIK3CA	GE	5290	293.295	438.031	392.259	373.586
PIK3CA	ME	TSS1500|DHS	0.052	0.045	0.044	0.052
PIK3CA	PE	PI3K-p110-alpha-R-C	-0.823	-0.566	-0.900	-1.253
PTEN	CN	CHR10+	-0.023	-0.017	-0.246	-0.639
PTEN	GE	5728	1308.731	1152.125	1271.756	787.965
PTEN	ME	TSS1500|DHS	0.038	0.037	0.039	0.040
PTEN	PE	PTEN-R-V	0.534	1.199	0.909	0.912
YAP1	CN	CHR11+	0.178	-0.034	0.345	-0.059
YAP1	GE	10413	2864.930	3117.673	4625.905	4731.826
YAP1	ME	TSS1500|DHS	0.049	0.048	0.046	0.047
YAP1	PE	YAP-R-V	-1.163	-0.538	-0.914	-1.034
YAP1	PE	YAP_pS127-R-C	0.166	0.800	0.786	1.137

Figure 1.

TCGA-Assembler as a tool for acquiring, assembling, and processing public TCGA data. (**a**) Flowchart of TCGA- Assembler. Module **A** acquires data from TCGA DCC. Module **B** processes the obtained data using various functions. (**b**) Illustration of a data matrix file using protein expressions generated by Reverse Phase Protein Array (RPPA). Each row corresponds to a protein or phosphorylated protein, and each column corresponds to a sample. The first column shows the gene symbol (before "|") and the name of the protein antibody (after "|") used in RPPA. (**c**) Illustration of combining multi- modal data. After combination, a single mega data table is obtained, in which each column (except the first three columns) corresponds to a patient sample and each row corresponds to a genomic/ epigenomic feature. All multi-modal data of a gene are adjacent in the table and are indicated by alternating blue/white color. In the second column, GE represents gene expression, PE protein expression, ME DNA methylation, CN copy number, and miRExp miRNA expression. In the third column, the description of GE platform is the Entrez ID of gene; the description of PE platform is the name of the protein antibody used in RPPA assay; "TSS1500|DNS" for the description of ME platform indicates that the values are average methylation measurements of CpG sites that are within 1,500 nucleotide base pairs of transcription start site and are DNAse hypersensitive; the description of CN platform gives the chromosome ID and strand of a gene.