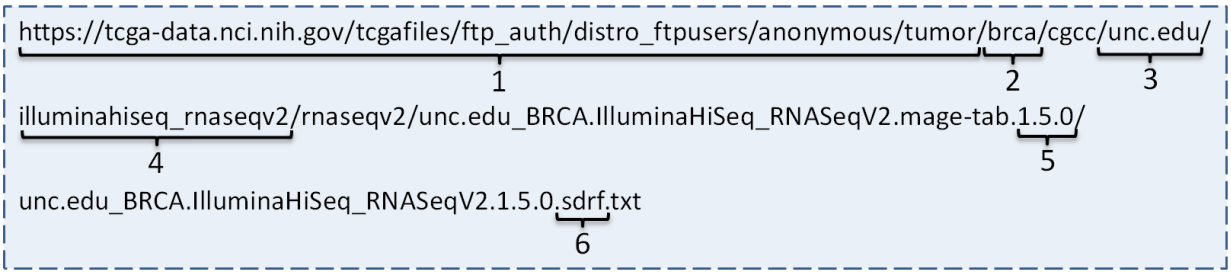
Supplementary Figure 1. Illustration of file URLs on TCGA DCC data servers using the Sample and Data Relationship Format (SDRF) file of BRCA RNA-seq data as an example. The six items in the URL are 1) *URL* of the root directory including public TCGA data, 2) *cancer type*, 3) *institution* that generated the data, 4) *assay platform,* 5) *version number,* and 6) *file type*. Here, the cancer type is breast invasive carcinoma (BRCA), the institution is the University of North Carolina, the assay platform is the Illumina HiSeq2000 system and the RNA-seq version 2 data pipeline as explained in the Supplementary Methods, the version number is 1.5.0, and the file type is SDRF. These six items are used in a string matching method to uniquely retrieve the URL of this file.



Supplementary Figure 2. Illustration of a data matrix file using protein expressions generated by Reverse Phase Protein Arrays (RPPAs). 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.

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-6884-01A-13-1936-20
YWHAE|14-3-3_epsilon-M-C	–1.8843	–1.8550	–1.6834	–1.6465	–1.4178
EIF4EBP1|4E-BP1-R-V	–1.2975	–1.9200	–1.9261	–2.2372	–1.4680
EIF4EBP1|4E-BP1_pS65-R-V	–1.0005	–0.9388	–0.9430	0.7651	–1.0697
TP53BP1|53BP1-R-C	1.6575	1.3552	0.7499	0.5979	0.8005
ACACA|ACC1-R-C	2.1955	1.8028	2.4295	1.1060	1.7943
ACACA ACACB|ACC_pS79-R-V	–1.5132	–1.8540	–1.6152	–2.3356	–1.4082
NCOA3|AIB1-M-V	–0.7736	–0.2837	–1.2775	–1.1706	–0.8168
AKT1 AKT2 AKT3|Akt-R-V	1.7266	1.3942	1.2189	2.2763	0.9438
AKT1 AKT2 AKT3|Akt_pS473-R-V	–2.6873	–2.2461	–1.9694	–2.4396	–2.2098

Supplementary Table 1. Summary of public TCGA data that can be acquired and processed by TCGA-Assembler. Entries are the numbers of patient samples measured by different assay platforms and the numbers of patients with de-identified clinical information (accurate as of August, 2013).

Cancer type	Number of patient samples (including both tumor and normal tissue samples)					Number of patients with clinical information
	RNA-seq	DNA methylation	DNA copy number	Protein expression	miRNA-seq	
Adrenocortical carcinoma [ACC]	0	81	0	0	0	0
Bladder urothelial carcinoma [BLCA]	167	225	314	127	211	157
Breast invasive carcinoma [BRCA]	1032	1161	2011	410	1113	969
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	118	178	279	0	136	74
Colon adenocarcinoma [COAD]	412	552	897	332	420	431
Lymphoid neoplasm diffuse large B-cell lymphoma [DLBC]	0	20	36	0	17	16
Esophageal carcinoma [ESCA]	0	78	125	0	12	18
Glioblastoma multiforme [GBM]	169	432	1106	214	0	576
Head and neck squamous cell carcinoma [HNSC]	341	435	824	212	411	349
Kidney chromophobe [KICH]	91	67	132	0	91	64
Kidney renal clear cell carcinoma [KIRC]	552	893	1058	454	614	503
Kidney renal papillary cell carcinoma [KIRP]	131	217	297	0	171	126
Acute myeloid leukemia [LAML]	179	194	392	0	188	200
Brain lower grade glioma [LGG]	272	278	514	0	299	230
Liver hepatocellular carcinoma [LIHC]	152	183	289	0	187	127
Lung adenocarcinoma [LUAD]	506	654	1062	237	545	456
Lung squamous cell carcinoma [LUSC]	460	586	930	195	514	392
Ovarian serous cystadenocarcinoma [OV]	420	631	1174	412	485	576
Pancreatic adenocarcinoma [PAAD]	41	79	124	0	53	53
Prostate adenocarcinoma [PRAD]	220	275	431	0	248	165
Rectum adenocarcinoma [READ]	155	188	314	130	162	168
Sarcoma [SARC]	51	95	160	0	84	72
Skin cutaneous melanoma [SKCM]	312	348	661	207	285	275
Stomach adenocarcinoma [STAD]	318	414	706	0	381	320
Thyroid carcinoma [THCA]	552	580	1012	226	566	481
Uterine corpus endometrial carcinoma [UCEC]	501	590	1013	200	544	479
Uterine carcinosarcoma [UCS]	0	58	106	0	0	10
Total	**7154**	**9492**	**15967**	**3356**	**7737**	**7287**

TCGA-Assembler: Open-Source Software for Retrieving and Processing TCGA Data (Supplementary Methods)

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

1. Center for Biomedical Research Informatics, NorthShore University HealthSystem, Evanston, Illinois, USA
2. Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA
3. Department of Health Studies, The University of Chicago, Chicago, Illinois, USA

* Correspondence to: yji@health.bsd.uchicago.edu

Note: The R code in Supplementary Methods is for illustration purposes. To test TCGA-Assembler and its full functionalities, please refer to the Quick Start Guide and Full User Manual at http://health.bsd.uchicago.edu/yji/TCGA-Assembler.htm and the step-by-step instructions therein.

1. MOTIVATION

The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga/) is supported by the National Cancer Institute and the National Human Genome Research Institute to chart the molecular landscape of tumor samples for more than 20 types of cancer[1,2]. TCGA has been generating multi-modal genomics, epigenomics, and proteomics data for thousands of cancer patients, providing unprecedented opportunities for researchers to systematically study cancer mechanisms at molecular and regulatory layers. There are different levels of data access for TCGA. While the access to most of the level-1 and -2 data is restricted, the entire level-3 TCGA data as well as some de-identified patient clinical information (e.g., survival and drug treatments) are publicly available. Level-3 TCGA data are normalized measurements of genomics and epigenomics features, e.g., DNA copy number, DNA methylation, mRNA expression, miRNA expression, and protein expression. Such a wealth of information enables biologists, clinicians, and quantitative geneticists to address various important research questions. For example, the level-3 data can be used to infer the effects of multiple types of transcriptional regulators, such as DNA methylation and transcription factors, on gene expression. This type of investigation requires matched data from multiple samples measured by all the relevant biological assays. We summarize the public TCGA data and report in Supplementary Table 1 the number of available samples measured by different assay platforms, and the numbers of patients with clinical information for the different cancer types.

A challenge preceding any TCGA data analysis is data acquisition and processing. It is a challenge for two reasons. First, the current TCGA data storage and management are not organized to facilitate downstream analysis, making data acquisition and assembly an intimidating task for most cancer researchers. Second, there are no open-source tools in the community that supports user-friendly means to access and navigate TCGA data. To meet this challenge, we introduce TCGA-Assembler, an open-source and freely available pipeline that automates and streamlines the data acquisition, assembly, and processing, TCGA-Assembler is programmed using R, and is freely available at http://health.bsd.uchicago.edu/yji/TCGA-Assembler.htm. Supplementary Table 1 shows the number of samples and the number of patients with clinical information that TCGA-Assembler can retrieve and process for different assay platforms and cancer types.

The level-3 TCGA data are stored under the open-access HTTP directory on the data servers of TCGA Data Coordinating Center (DCC, https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/). The DCC provides data in the form of hundreds of thousands of data files, each with a different web address, i.e. Uniform Resource Locator (URL), and consisting of measurements of one type of genomics/epigenomics feature for a patient sample across the genome. TCGA-Assembler provides functions that streamline the downloading and processing of these data files to form data matrices for different genomic data modalities and different cancer types. In a data matrix, each row represents a feature and each column corresponds to a sample (see Supplementary Figure 2).

2. TCGA-Assembler

Figure 1a in the main text summarizes the data acquisition and processing procedures of TCGA-Assembler, which includes two modules. Module A streamlines data downloading and assembly and module B processes the assembled data matrices and prepares them for subsequent analyses.

2.1. Module A: Data Acquisition and Assembly

Module A performs two steps to acquire and assemble TCGA data. TCGA systematically names data files and directories using key words, such as cancer type, institution name, assay platform, and file type. These key words are directly incorporated in the URLs of the files. Such a file/directory naming mechanism allows us to use string-matching algorithms to identify the URLs of the files of interest. In module A, the first step is to extract the URLs of all files in the open-access directory on TCGA data server, which is realized by an important function called *TraverseAllDirectories* (see user manual for its details). These URLs correspond to all public TCGA data files and are stored locally, for example, in the file *DirectoryTraverseResult_Jan-30-2014.rda* included in the package, which was generated on Jan. 30, 2014. In the second step, module A downloads and assembles user-specified data files. Take the RNA-Seq data of breast invasive carcinoma (BRCA) samples as an example. To get these data, module A first downloads an index file of BRCA RNA-seq data called the SDRF (Sample and Data Relationship Format) file using string matching on URLs (see Supplementary Figure 1). With the information in the SDRF file, the URLs of relevant data files corresponding to BRCA RNA-Seq samples are located from the master list of URLs obtained in the first step. Then the data files are downloaded through the identified URLs and assembled into data matrices. These two steps correspond to the two blocks in module A of Figure 1a in the main text. During this process, module A also checks whether the probe/assay information is consistent across all data files of the same assay platform in order to ensure data consistency.

To illustrate this process, we still use BRCA RNA-seq data as an example. TCGA possesses 1,032 BRCA patient samples, which have been RNA-sequenced by the Illumina HiSeq2000 system, and the recorded sequence data have been processed by the RNA-seq version 2 pipeline that uses the Mapsplice alignment algorithm and the RSEM algorithm to generate expression values[3,4]. For each patient sample, six individual files are generated by this pipeline and stored at TCGA DCC site. The files are 1) un-normalized expression values of genes and 2) of isoforms, 3) normalized expression values of genes and 4) of isoforms, and 5) expression quantifications for exons and 6) junctions. Therefore, there are a total of 1,032×6=6,192 files for the BRCA samples from the RNA-seq assay alone. Manually downloading all of these files from their corresponding URLs will be an error-prone, tedious and non-reproducible task. Instead, TCGA-Assembler uses a single command in R (see below) to automatically download and combine these data files into six data matrix files, one for each type of measurement on all 1,032 patient samples.

RNASeqRawData = DownloadRNASeqData(traverseResultFile = "./DirectoryTraverseResult_Jan-30-2014.rda", saveFolderName = "./SuppTest1/", cancerType = "BRCA", assayPlatform = "RNASeqV2", dataType = c("exon_quantification", "rsem.isoforms.results", "rsem.isoforms.normalized_results", "junction_quantification", "rsem.genes.normalized_results", "rsem.genes.results"));

There are five input arguments in this function call. The first argument *traverseResultFile* is the path of a file containing a master list of URLs of all data files under the open-access directory on TCGA DCC data server. The master list is generated using the aforementioned function *TraverseAllDirectories* and only needs to be generated once for downloading data of various cancer types and assay platforms. The second argument *saveFolderName* is a user-specified local directory to store the downloaded contents. The third, fourth, and fifth arguments are key words that specify the cancer type, assay platform, and data type of interest, respectively. They are specified by users from a list of options (see user manual for details). This command downloads all expression data files of BRCA samples produced by the RNA-seq version 2 pipeline and combines them into data matrices that are stored in tab-delimited .txt files under the local directory *SuppTest1*. On a computer with 2.5MBps Internet speed and 2.93GHz CPU, downloading and combining the six data files of one sample takes on average 25 seconds, or about 7.2 hours to complete the entire process for all 1,032 BRCA samples. The time will be shorter on computers with faster Internet connection and CPUs. More importantly, this function can be executed as a background process with minimum technical intervention from users and multiple instances of the program can be launched simultaneously to download different data. For advanced technical users, the scripts behind the command are made available and can be modified to fit specific needs.

2.2. Module B: Data Processing and Integration

TCGA-Assembler goes beyond data acquisition by providing unique and important data processing functions as indicated by module B in Figure 1a of the main text. These functions achieve initial but critical data preprocessing goals, such as correcting errors in the gene symbols, which can severely damage subsequent analyses if not corrected. We will elaborate on a few of these functions below.

(1) **Check and correct gene symbols.** Many level-3 TCGA data files use gene symbols as identifiers. Unfortunately, some gene symbols used in TCGA do not match the official gene symbols assigned by the HUGO Gene Nomenclature Committee (HGNC)[5]. For example, in some data files gene symbols

MARCH5 and FEB6 were incorrectly converted to 5-MAR and 6-FEB, respectively. These conversions could have been caused by the auto-correction function in software such as Excel. To correct these types of errors automatically, TCGA-Assembler provides a function to identify and correct invalid gene symbols as well as map aliases and outdated symbols to official HGNC gene symbols. This function keeps gene identifiers consistent among data produced using different assay platforms and by different research centers, thus facilitating downstream integrative analysis of multi-modal and multi-cancer TCGA data.

(2) **Combine multi-modal data.** A unique and important feature of TCGA data is the matched measurements of multiple genomics and epigenomics features over thousands of patient samples, which allow for integrative data analyses across multiple data platforms. To prepare data for such analyses, TCGA-Assembler provides functions that can match and combine the multi-modal data by patient samples and genomics/epigenomics features. Matching patient samples across platforms is realized by merging the data of common samples measured by all assay platforms. The combined data are then sorted by genes such that data of the multiple genomics/epigenomics features of a gene are stacked next to each other. For example, the copy number, mRNA expression, DNA methylation, and protein expression of a gene will be put together as adjacent rows. Figure 1b in the main text gives an illustration of the combination outcome, which is a single mega data table.

(3) **Combine HumanMethylation27 and HumanMethylation450 data.** TCGA uses two Illumina Infinium arrays to measure DNA methylation in patient samples, i.e. the HumanMethylation27 BeadChip and HumanMethylation450 BeadChip. The HumanMethylation27 BeadChip measures DNA methylation at about 27,000 CpG sites. The HumanMethylation450 BeadChip measures DNA methylation at more than 450,000 CpG sites, including ~90% of the CpG sites measured by HumanMethylation27 BeadChip. For many cancer types, such as BRCA, both chips have been used to generate data for a large number of samples. Data from both groups may be included in downstream statistical analysis for increased power. TCGA-Assembler provides a function that combines HumanMethylation27 data and HumanMethylation450 data. The function first identifies the CpG sites measured by both arrays and combines the data of these common CpG sites. Quantile normalization is then performed on the combined data to eliminate any systematic differences (if any) between the two assays.

(4) **Summarize DNA methylation in different genomic regions.** DNA methylation in different regions of genes may lead to specific epigenetic effects. For example, DNA methylation in promoter regions can repress the expression of genes. TCGA-Assembler provides a function to summarize the methylation level in different regions of genes by calculating an average methylation value of CpG sites in the region, such as TSS1500 (within 1500 nucleotide base pairs of a transcription start site), TSS200, 5'UTR, 1st Exon, 3'UTR, and CpG sites hypersensitive to DNAse or not. The region-specific methylation data can facilitate research of different interests.

In addition to the above functions, we have implemented many other data processing functions in module B. Examples include drawing boxplots of data matrices after they have been processed to identify and remove sample outliers, extracting subsets of data according to user-specified tissue types, e.g. recurrent solid tumor or blood derived normal cell, and calculating gene-level DNA copy numbers. Details of these functions are provided in the user manual.

3. EXAMPLE of Incorporating TCGA-Assembler with Downstream Data Analysis

3.1. Software Installation and Preparation

TCGA-Assembler is built on R (http://www.r-project.org/) and requires R packages *HGNChelper*, *RCurl*, *httr*, *stringr*, *digest*, *bitops*, and their dependents. Assume that a recent R version is installed (version 2.15.1 or later). The following command installs the packages

install.packages(c("HGNChelper", "RCurl", "httr", "stringr", "digest", "bitops"), dependencies=T)

Remark: R can be downloaded and installed from http://www.r-project.org/. Another way to install the R packages is that in R GUI (Graphical User Interface), go to *Packages* menu and click on *Install package(s)*. Select the best CRAN mirror site for you. And then select the package and click *ok* to install.

Caution1: If you want to copy and paste the commands in this file to R console to run the examples, please use Adobe Reader (freely available from http://get.adobe.com/reader/) or Adobe Acrobat to open and review this pdf file. Other pdf viewers may add additional characters to the end of lines when you copy and paste the code.

Caution2: Depending on which R packages are already installed on users' computers, occasionally users could experience R errors complaining about conflicts of different packages. We did not experience any in our testing, but do not rule out potential errors due to system setup, or clashing with existing R packages.

Caution3: When using TCGA-Assembler, avoid reading or writing files in Dropbox or similar cloud-based folders, which may produce error message. Make sure that you have read and write access to the working directory.

To download and use TCGA-Assembler, go to http://health.bsd.uchicago.edu/yji/TCGA-Assembler.htm. Click *Download Software* and unzip the downloaded file to your desired file directory on the local computer. For example, in our own test, we unzipped the package and created the folder

/Users/zhuy/TCGA-Assembler/

for the unzipped files. Then we set the Present Working Directory (PWD) of R using

setwd("/Users/zhuy/TCGA-Assembler")

Users should use

setwd("foo")

where *foo* is the directory that stores unzipped TCGA-Assembler files on user's computer. To start using TCGA-Assembler, load all the functions in modules A and B into the working space using the following commands.

```
source("./Module_A.r");
source("./Module_B.r");
```

3.2. Incorporation with Downstream Data Analysis

Being open-source and freely available software, TCGA-Assembler can be seamlessly integrated with downstream data analysis scenario. Here, we present a simple example to demonstrate the possibility of incorporating functions in TCGA-Assembler with analysis scripts to realize reproducible data retrieval and analysis. The example is about the identification of differentially expressed genes using t-tests. We note that the use of t-tests is an arbitrary choice for illustrative purposes. More sophisticated analysis tools could be similarly integrated with TCGA-Assembler. The proposed simple example consists of these steps. As a prerequisite, we assume that users have installed TCGA-Assembler and gone through steps in Sections 3.1.

First, download gene expression data of six BRCA samples (including three tumor samples and three normal tissue samples) using the following command. The last input argument of the command allows the function to download data of samples specified by TCGA sample barcodes (refer to user manual for details).

```
RNASeqRawData = DownloadRNASeqData(traverseResultFile =
"./DirectoryTraverseResult_Jan-30-2014.rda", saveFolderName = "./SuppTest2/", cancerType = "BRCA",
assayPlatform = "RNASeqV2", dataType = "rsem.genes.normalized_results", inputPatientIDs =
c("TCGA-A1-A0SB-01",  "TCGA-A1-A0SD-01", "TCGA-A1-A0SE-01", "TCGA-A7-A0D9-11",
"TCGA-A7-A0DB-11", "TCGA-A7-A13E-11"));
```

Second, process the downloaded data for basic quality control and extract normalized count data generated by the RNA-seq version 2 pipeline, using module B function *ProcessRNASeqData*.

```
GeneExpData = ProcessRNASeqData(inputFilePath =
"./SuppTest2/BRCA__unc.edu__illuminahiseq_rnaseqv2__rsem.genes.normalized_results__Jan-30-2014.txt",
outputFileName = "READ__illuminahiseq_rnaseqv2__GeneExp", outputFileFolder = "./SuppTest2", dataType =
"GeneExp", verType = "RNASeqV2");
```

The first input argument of the command is the path of the data file produced in the previous step.

Third, extract data of primary solid tumors and data of solid normal tissues using module B function *ExtractTissueSpecificSamples*.

```
TumorData = list(Des = GeneExpData$Des, Data = ExtractTissueSpecificSamples(inputData =
GeneExpData$Data, tissueType = "TP", singleSampleFlag = FALSE));

NormalData = list(Des = GeneExpData$Des, Data = ExtractTissueSpecificSamples(inputData =
GeneExpData$Data, tissueType = "NT", singleSampleFlag = FALSE));
```

Fourth, use t-test to check 10 genes to see whether they are significantly differentially expressed between primary solid tumors and solid normal tissues.

```
for (i in 61:70)
{
   print(TumorData$Des[i, ]);
   TestResult_i = t.test(x = TumorData$Data[i, ], y = NormalData$Data[i, ]);
   print(TestResult_i);
}
```

4. DISCUSSION and CONCLUSION

Besides TCGA DCC, which is the original source of TCGA publicly accessible data, level-3 TCGA data can also be obtained from Firehose at the MIT Broad Institute (https://confluence.broadinstitute.org/display/GDAC/Home). Although primarily a data analysis infrastructure, Firehose provides level-3 TCGA data matrices, one for each cancer type and assay platform. Unlike TCGA-Assembler, Firehose is not open-source and its computer program is not publicly available. Advantages of TCGA-Assembler include that users can execute it in real time to capture the most recent updates from TCGA as new data are still being produced and that advanced technical users can adapt the scripts and functions to facilitate their own data pipelines. More importantly, TCGA-Assembler provides various unique functions for data processing, such as combining data across

different genomics platforms. These data processing functions are also Firehose-compatible, which allows users who already possess Firehose data to perform data manipulation, such as removing redundant data entries in the data matrix files from Firehose.

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. Graphical user interfaces will also be added to the pipeline to enable navigation of the software, particularly for biological researchers who are less familiar with R.

REFERENCES

[1] Cancer Genome Atlas Network. Nature **490**, 61–70 (2012).
[2] Cancer Genome Atlas Research Network. Nature **455**, 1061–1068 (2008).
[3] Wang, K. et al. Nucleic Acids Res. **38**(18), e178 (2010).
[4] Li, B. & Dewey, C.N. BMC Bioinformatics, **12**:323 (2011).
[5] Wain, H.M. et al. Genomics **79**(4), 464–470 (2002).