Data Research Camp 2024

Lower grade glioma TCGA data

July 10, 2024

Introduction to TCGA[1]

The Cancer Genome Atlas (TCGA) is a landmark collection that maps the genomic profiles of 33 cancer types and subtypes, including 10 rare cancers, from bladder and breast to pancreatic and uterine. These "atlases," reveal the molecular features associated with cancer and help to inform everything from basic research to drug development and precision medicine. Moreover, all TCGA data are all in the public domain, allowing any researcher with an interest to access this information.

Is TCGA Still Collecting Data?

Despite the TCGA program closing in 2018, it continues to be an integral part of modern-day cancer genomics analyses. The latest version of the data, processed according to best-in-practice bioinformatics pipelines (which have continual updates), can be found at NCI's Genomic Data Commons (GDC). Moreover, researchers are actively characterizing and releasing data for newly sequenced whole genomes. This enables us to continue to use and build on TCGA to inform new research, leading to new data on the genes, proteins, pathways, and drivers underlying cancer.

Where Can I Find TCGA Data Today?

Nearly 2.5 petabytes of TCGA data are available through NCI's GDC. Visit NCI's Center for Cancer Genomics to learn more about the cancer types and criteria for TCGA's data set and see citations for seminal studies.

What Type of Data Are in TCGA?

In the TCGA, you'll find:

whole genome sequence (WGS),

whole exome sequence,

methylation,

RNA expression,

microRNA,

transposase-accessible chromatin with sequencing (ATAC-Seq),

reverse phase protein array (RPPA),

[1]DELIBERATELY TAKEN FROM https://datascience.cancer.gov/news-events/blog/cancer-genome-atlas-tcga-living-legacy-cancer-research

tissue slide images, and
clinical data sets.

How is TCGA Unique?

Team science led to the development of TCGA, with contributions from scientists at NCI and the National Human Genome Research Institute, along with thousands of researchers from institutions around the country. Together, this team developed TCGA's technology, tools, and resources and carried out characterization of thousands of samples. Importantly, more than 11,000 patients contributed their samples to science. This open-science framework continues today.

Open data sharing means that anyone can access these data sets (from the lab next door to one around the world). This broad data sharing helped expand the usefulness of the data in TCGA, as researchers look for new ways to limit bias and make the data more applicable to more people.

TCGA's collection also features "normal control" data. This means that patients gave blood or tissue samples taken from near the cancerous tumor, in addition to the tumor itself. Having normal samples offers a control, allowing researchers to examine the differences between normal and cancerous tissues.

How Has TCGA Influenced the Cancer Research Field?

Before TCGA, it was difficult for researchers to assemble all the bits and pieces of data on the numerous biological processes associated with cancer. Once TCGA came to fruition, it revolutionized research on the molecular mechanisms underlying cancer. If you use scholarly publications as a yardstick, you'll find that PubMed features more than 29,000 papers with mention of TCGA. Last year alone, there were over 5,000 TCGA citations. Tools are another good measure of how these data are impacting the field. Since TCGA's start, we now have countless tools to help you navigate and analyze these data, including many built by original TCGA team researchers. For example, cBioPortal provides an interface to help you analyze genetic and clinical data to study cancer and how it progresses over time. And of course, the most up-to-date version of the data (along with new visualization and cohort analysis tools) are available via NCI's GDC.

How Does the GDC Help Me Use TCGA Data?

The GDC's data portal gives you a full suite of web-based tools for studying TCGA data (along with other large-scale data sets), including building and comparing cohorts, examining mutation frequencies, visualizing gene expression clusters, and more. These analysis tools can help you access and use genomic data, no matter your skill level or experience. Alternatively, you can access these data in the cloud, letting you work with large data sets without needing to download and store those data.

And you don't have to start from scratch. Scientists use TCGA data to develop numerous pipelines and other methodologies for studying cancer. You can also find tools with shortcuts for processing and analyzing data from start to finish, such as the Multi-omics Pathway Workflow, or MOPAW, and BigQuery.

Why is TCGA Data Particularly Good for Data Science?

TCGA's tumor profiles provide a rich resource for exploring any number of topics of interest: from new drugs or biomarkers to new ways of diagnosing, preventing, and treating cancer. Perhaps one of the biggest impacts for TCGA data are "pan-cancer" studies, that is, examining multiple cancers at a time to reveal machinery that's similar among many tumors, no matter the tissue or organ of origin.

I Want to Integrate My Data With TCGA's Data for Analysis—What Format Do I Use?

The Cancer Genomics Cloud offers a detailed table of available TCGA data formats. Information on TCGA's metadata also is available, including a separate listing for TCGA's Genome Reference Consortium Human Build 38 (GRCH38) assembly.

What's Next for TCGA?

TCGA may be a "closed" program, but the data are continuing to have a major impact on the cancer research field. New tools and technologies, such as artificial intelligence (AI), have the potential to transform cancer treatment and care.

For example, researchers are using data to train and refine AI models for diagnosing, predicting, and tracking cancer. AI's especially promising for precision oncology—helping to predict how a patient will respond to treatment so clinicians can select the best treatment right from the start.

Importantly, with TCGA, researchers have the most vital commodity for moving the field forward—open access to data. By broadly sharing data, TCGA, and other data sets like it, give researchers the information they need to develop new and better ways of diagnosing, treating, and preventing cancer.

Prognostic Vs predictive biomarkers[2]

A variety of factors influence a patient's clinical outcome, including intrinsic characteristics of the patient, disease, or medical condition, and the effects of any treatments that the patient receives. Some of the intrinsic characteristics may be reflected as **prognostic biomarkers**, i.e., *biomarkers used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest*, and others as **predictive biomarkers**, i.e., *biomarkers used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent*. Prognostic biomarkers and predictive biomarkers cannot generally be distinguished when only patients who have received a particular therapy are studied. Some biomarkers are both prognostic and predictive. Prognostic biomarkers are often identified from observational data and are regularly used to identify patients more likely to have a particular outcome.

A case study of lower grade glioma

Here we introduce the TCGA protein data set that was used for the purpose of treatment selection in Ma *et al.* (2019) and Pedone et al. (2024). This dataset is provided in the rda file `LGGdata.rda`.

Analysis of TCGA data for treatment selection

Inherently a problem of outcome prediction, approaches to treatment selection can be enhanced through strategies that integrate prognostic with predictive characteristics of a candidate patient/disease.

[2]Taken from https://www.ncbi.nlm.nih.gov/books/NBK402284/

While intrinsic to the generalized linear model, the belief that prognostic features convey no additional information for treatment selection is actually a specific manifestation of the linear predictor. Ma et al. (2019) demonstrated that this premise fails for broader classes of modeling strategies based on predictive probability. Conceptually, they proposed an integrative treatment selection strategy in three parts. Given a discrete set of ordered response-levels describing a spectrum of possible clinical outcomes, the method first derives "baseline" predictive probability measures on the basis of the prognostic determinants of a particular patient's disease profile. Then it integrates the predictive features to adjust the baseline probability measures to reflect current knowledge pertaining to the effectiveness of each treatment option for the specific disease characteristics exhibited by the candidate patient. Finally, it provides rational to further elucidate the manner in which considerations of prognostic effects may alter treatment decisions derived from predictive features alone.

Lower grade glioma

Glioma is a type of cancer that develops in the glial cells of the brain. Glial cells support the brain's nerve cells and keep them healthy. Tumors are classified into grades I, II, III or IV based on standards set by the World Health Organization. For this study, TCGA studied lower grade glioma, which consists of grades II and III. Regardless of grade, as a glioma tumor grows, it compresses the normal brain tissue, frequently causing disabling or fatal effects. In 2010, more than 22,000 Americans were estimated to have been diagnosed and 13,140 were estimated to have died from brain and other nervous system cancers.

Measuring protein expression

Reverse-phase protein array (RPPA) is a high-throughput antibody-based targeted proteomics platform that can quantify hundreds of proteins in thousands of samples derived from tissue or cell lysates, serum, plasma, or other body fluids. Protein samples are robotically arrayed as microspots on nitrocellulose-coated glass slides. Each slide is probed with a specific antibody that can detect levels of total protein expression or post-translational modifications, such as phosphorylation as a

measure of protein activity. There are workflow protocols and software tools developed and optimized for RPPA that includes sample preparation, microarray mapping and printing of protein samples, antibody labeling, slide scanning, image analysis, data normalization and quality control. This is typically preformed by a bioinformatician. For the sake of this dataset, we will focus on level 3 data, i.e., data that have been already processed and are ready for analysis. Higher values correspond to highly active protein in that given sample.

TCGA data

We present a dataset obtained from the lower grade glioma (LGG) TCGA data portal, that combines clinical and level 3 protein expression data [3] (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). Baseline and follow-up information was collected for 411 patients. Exclusions included 18 patients with missing values of treatments and 68 patients who failed to contribute protein expression information. Among the remaining 325 patients, 196 received molecularly targeted therapies, 211 received adjuvant radiotherapy, and 79 received "conventional" cytotoxic therapies that involved neither targeted nor radiotherapy regimens. Ma et al. (2019) considered treatment selection of standard (neither targeted nor radiotherapy, n=79) versus advanced treatments (targeted or radiotherapy, n=246). Using the RECIST criteria (http://www.recist.com/), tumor response was categorized using the four standard ordinal-levels of progressive disease (PD), stable disease (SD), partial response (PR) and complete response (CR). Among those receiving standard therapy, only 9 patients achieved PR. Therefore, we combined PR and SD to formulate a new category of responders, which we abbreviate as PS, yielding three ordinal-levels of the outcome: CR, PS, and PD.

Matching

To account for potential select bias, we matched patients on the basis of the baseline covariates of tumor grade, gender, age and initial year of pathological diagnosis (IYPD). Specifically, 79 pairs of patients were produced using the R package of $MatchIt$ (default settings) (Ho et al., 2011); the resultant standardized mean differences were 0.000, -0.050, 0.051, and 0.162 for tumor grade,

[3] These are pre-processed RPPA data ready for analysis

gender, age and IYPD, respectively. Reasonably satisfactory matches were obtained, as all final standardized mean differences were below than the suggested cut-off value of 0.25 (Imai et al., 2008).

Note that the TCGA data set includes the short-term clinical outcome of tumor responses as well the overall survival data, and the elicited utility weight for PS reflects its relative importance in terms of long term benefits.

Pre-selection of prognostic and predictive features

To identify potential prognostic/predictive features among the 173 protein expressions were measured in the LGG data, Ma *et al.* (2019) fitted univariate logistic regression models with covariates of a protein, treatment, and their interaction (R package of $MASS$ (Venables and Ripley, 2002)). A protein was considered as a potential predictive (prognostic) feature given a p-value, obtained from Wald's test, was < 0.1 for the interaction (main) effect. With this criteria, 23 proteins were selected as potential predictive features and 5 as potential prognostic features. Two prognostic covariates, ACVRL1-R-C and HSP70-R-C, were utilized in the application given that they yielded the highest accuracy rate (78/158) in discriminant analysis using the 79 pairs of matched data with binary outcomes of PD/SD/PR as 0 and CR as 1. It is worthy noting that, to remove noise variables and enhance model performance, data pre-processing is commonly applied in high dimensional settings. We here describe a straightforward approach to pre-select some features for data analysis, and more advanced approaches can be applied.

Discussion

Note that what described so far concerns the dataset analyzed in Ma *et al.* (2019) and Pedone *et al.* (2024). We have also provided the R code to directly download the most recent protein and clinical dataset from TCGA that cotains additional data including expression levels of many more proteins

Given these datasets, there are many possible analyses of interest, including:

- Differential expression analysis
- Combine clinical characteristics with genomics data to inform treatment selection;
- study similarity and differences in the biology (e.g., in terms of protein networks) of subset of patients (respondent Vs non-respondent);
- Integrate with additional TCGA data (but it needs to download additional data, e.g., mutations, from TCGA);
- perform dimensional reduction;
- perform variable/protein selection with respect of a response of interest (e.g., treatment response).

It is important to note that Ma et al. (2019) and Pedone *at al.* (2024) did not attempt to develop prognostic/predictive biomarkers but rather use established molecular features for personalized treatment selection.

Some final remarks on practical aspects of treatment selection methods

Several aspects of the feature inputs, such as process or inter-observer reproducibility, need to be considered carefully before using the proposed methods for personalized selection with large scale genomic, imaging and clinical data. A checklist criteria has been developed by the US National Cancer Institute which addresses issues of specimens, assays, clinical trial design (McShane et al., 2013). Ma *et al.* (2019) matched patients with tumor grade, gender, age and initial year of pathological diagnosis. Although this matching seeks to reduce the potential impact of selection bias, the proposed method is perhaps most clinically useful when implemented with training data obtained from randomized clinical study.

Moreover, a patient's experience upon receiving a particular therapeutic strategy is often a complex synthesis of measures that describe both the extent of induced-harm as well as clinical benefit. Thus, patient response is often difficult to characterize in many cancer settings, especially

for multi-modal treatment strategies (see e.g. Hobbs et al., 2016). As a consequence, most methods for treatment selection implicitly assume that all patients should be treated with one of the therapeutic regimes under study (Ma et al., 2015; Geng et al., 2015). Yet, its important to note that future advances in statistical methods that facilitate formal prediction of harm-versus-benefit trade-offs may further elucidate sub-populations for which the absence of further clinical intervention provides the best option. This goal can be achieved only if prognostic features are integrated into the treatment selection process.

References

Geng, Y., Zhang, H. H., and Lu, W. (2015). On optimal treatment regimes selection for mean survival time. *Statistics in medicine* **34,** 1169–1184.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* **42,** 1–28.

Hobbs, B. P., Thall, P. F., and Lin, S. H. (2016). Bayesian group sequential clinical trial design using total toxicity burden and progression-free survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65,** 273–297.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171,** 481–502.

Ma, J., Hobbs, B. P., and Stingo, F. C. (2015). Statistical methods for establishing personalized treatment rules in oncology. *BioMed Research International* **2015,** 1–13.

Ma, J., Stingo, F. C., and Hobbs, B. P. (2019). Bayesian personalized treatment selection strategies that integrate predictive with prognostic determinants. *Biometrical Journal* **61,** 902–917.

McShane, L. M., Cavenagh, M. M., Lively, T. G., Eberhard, D. A., Bigbee, W. L., Williams, P. M.,

Mesirov, J. P., Polley, M.-Y. C., Kim, K. Y., Tricoli, J. V., et al. (2013). Criteria for the use of omics-based predictors in clinical trials. *Nature* **502,** 317–320.

Pedone, M., Argiento, R., and Stingo, F. C. (2024). Personalized treatment selection via product partition models with covariates. *Biometrics* **80(1),**.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

A Data Download and Preparation

In this appendix, we provide the `R` script used to directly download and process the latest data from the TCGA portal. This script ensures that users have access to the most recent and relevant data by querying, downloading, and preparing clinical and protein expression data for the TCGA-LGG project. The script automates the data querying steps and saves the datasets to CSV files, along with the session information to ensure reproducibility.

A.1 R Script

```
rm(list=ls())


# Install necessary libraries
if (!"BiocManager" %in% rownames(installed.packages()))
  install.packages("BiocManager")
if (!"TCGAWorkflow" %in% rownames(installed.packages()))
  BiocManager::install("TCGAWorkflow")
if (!"TCGAWorkflowData" %in% rownames(installed.packages()))
  BiocManager::install("TCGAWorkflowData")
if (!"TCGAbiolinks" %in% rownames(installed.packages()))
  BiocManager::install("TCGAbiolinks")
```

```
library(TCGAbiolinks)
library(TCGAWorkflow)
library(TCGAWorkflowData)
library(DT)
library(tidyverse)

# Download and prepare clinical data
query_clinical <- GDCquery(
  project = "TCGA-LGG",
  data.format = "bcr xml",
  data.category = "Clinical"
)

GDCdownload(query_clinical)
clinical_data <- GDCprepare_clinic(
  query = query_clinical,
  clinical.info = "patient"
  )

datatable(
  data = clinical_data,
  options = list(scrollX = TRUE, keys = TRUE),
  rownames = FALSE
)

# Extract treatment information
clinical_drug <- GDCprepare_clinic(
```

```
  query = query_clinical,
  clinical.info = "drug"
)

clinical_drug |>
  datatable(
    options = list(scrollX = TRUE, keys = TRUE),
    rownames = FALSE
  )

# Download and prepare protein expression data
query_protein <- GDCquery(
  project = "TCGA-LGG",
  data.category = "Proteome Profiling",
  data.type = "Protein Expression Quantification",
  sample.type = "Primary Tumor"
)

GDCdownload(query_protein)
protein_data <- GDCprepare(query_protein)

datatable(
  data = protein_data,
  options = list(scrollX = TRUE, keys = TRUE),
  rownames = FALSE
)

# Protein data name is a bit different from clinical
```

```
# data bcr_patient_barcode, but simply removing portion
# and analyte is safe here (see
# https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/ )

protein_datanew <- protein_data[,-c(1:4)] %>%
  rename_with(~ str_sub(., 1, -5), -1)
## > length(colnames(protein_datanew))
## [1] 430
## > length(unique(colnames(protein_datanew)))
## [1] 430

# Transpose the tibble keeping the names
data_matrix <- t(as.matrix(protein_datanew[,-1]))
colnames(data_matrix) <- protein_datanew$peptide_target
protein_data_new <- data.frame(data_matrix) %>%
  rownames_to_column(var = "bcr_patient_barcode")

# Save datasets to CSV files
write.csv(clinical_data, "clinical_data.csv", row.names = FALSE)
write.csv(clinical_drug, "clinical_drug.csv", row.names = FALSE)
write.csv(protein_data_new, "protein_data.csv", row.names = FALSE)

session_info <- sessionInfo()
writeLines(capture.output(session_info), "session_info.txt")
```

A.2 Datasets Description

The datasets provided `clinical_data.csv`, `clinical_drug.csv`, and `protein_data.csv` encompass comprehensive clinical and proteomic data from the TCGA-LGG project. More in detail:

- **Clinical Data**: This dataset contains detailed clinical information about the patients, including demographic details, diagnosis, treatment history, and follow-up data. It has 515 records and 75 features.
- **Treatment Data**: This subset of clinical data focuses specifically on the treatments administered to the patients. It includes information on various therapies, such as chemotherapy, radiotherapy, and molecularly targeted treatments. It has 695 records and 24 features.
- **Protein Expression Data**: This dataset provides quantitative measurements of protein expression levels in primary tumor samples. It offers insights into the molecular characteristics of the tumors. It has 487 records and 434 features.

The `bcr_patient_barcode` serves as the primary key that connects patients across the three datasets. This key allows for the linkage of patient data from the clinical, treatment, and protein expression datasets from the TCGA-LGG project.

These datasets collectively enable a multi-faceted analysis of clinical and molecular aspects of lower-grade gliomas. Note that these datasets provide raw data, as accessible from the TCGA portal. Therefore, no data were excluded or imputed, and no variables of interest were derived for analysis. Similarly, selection bias and pre-selection have not been addressed.

A.3 Session Info

The `session_info.txt` file provides the `R` session details and serves as a resource to ensure the reproducibility of the script.