Name: Colin Yeo

Date: 9/10/22

TCGA Website Scavenger Hunt

QBIO Multi-omic Data Analysis

TCGA (Home Page):

The Cancer Genome Atlas (TCGA), founded in December of 2005, is a cancer genomics program hosted by the NCI and the National Human Genome Research Institute. The publicly available data from this project includes genomic, epigenomic, transcriptomic, and proteomic data. This data was collected from 20,000 different samples that span 33 different cancer types, including breast cancer, which we will be focusing on this semester.

Program History:

Describe one outcome or impact of TCGA: TCGA has deepened our scientists' understanding of cancer by noticing patterns of molecular alterations in different cancerous tissues.

Briefly skim the "Timeline & Milestones" page. When did TCGA publish their paper on breast cancer?

2012, October.

Because TCGA is a public dataset, and one of the first of its kind, they faced some initial concerns regarding the ethics of releasing health data to the public. Choose one of the papers in the "Ethics & Policies" section to skim. What is one way that your paper addresses these privacy concerns? The paper I selected implemented a tripartite responsibility system, that involves the rapid release of all sequence data about a certain size, and that requires active community-wide support of these efforts to continue.

TCGA Cancers Selected for Study:

List three criteria used to select which cancers to study: Poor prognosis, availability of samples meeting standards for patient consent, overall public health impact.

Open the breast ductal carcinoma page and read TCGA's provided background. List one interesting fact you found: drugs are being considered that inhibit blood vessel growth in cancer patients, cutting off blood supply to the tumor.

Publications by TCGA:

TCGA published (at least) one paper on each of their studied cancer types. These papers, called marker papers, include an early analysis of the data, including any molecular characterizations that were performed. Read the abstract of the 2012 breast ductal carcinoma cancer paper. List any genes you come across (these may be good starting points for your future analyses of this cancer):

TP53, PIK3CA, GATA3, MAP3K1, HER2, EGFR.

Using TCGA:

Go to the Genomic Data Commons (GDC) Data Portal via the link on TCGA home. This portal lets you view TCGA's data in a visual way. Let's explore this website. According to the Data Portal Summary, there are 72 projects in the GDC data portal. Now click on the "Projects" tab. Notice that not all projects in this data portal are TCGA-affiliated, though TCGA does make up 33 of the projects included.

Name: Colin Yeo

Date: 9/10/22

TCGA Website Scavenger Hunt

QBIO Multi-omic Data Analysis

Using TCGA (Continued)

Under the "Program" tab, select just TCGA studies. According to the graph at the top of the page, TP53 is the most mutated gene in TCGA projects, affecting approximately 35% of cases.

Return to the GDC Portal home page. Now click the breast image in the diagram to the right of the page. This directs you to the "Exploration" tab and automatically selects all primary sites associated with breast cancers. Now select TCGA as the program, and TCGA-BRCA as the as the project. This is the data we will be focusing on this semester.

The table on this page shows each patient along with their data. Feel free to explore the data files by clicking on any of the links provided.

Now explore the Cases, Genes, Mutations, and OncoGrid tabs above the pie charts. What is one takeaway from the plots provided here: There seem to be a number of "high-yield" targets/genes for the most common cases of breast cancer... TP53, PIK3CA, and the most common disease type is overwhelmingly ductal and lobular neoplasms.

As you can see, the GDC portal provides an overwhelming amount of information. Feel free to continue to explore it on your own time!

Discussion:

Think through the following questions, and record your answers below:

1. What is the goal of TCGA?

The goal of TCGA is to provide another source of data for scientists to explore cancer genomics and to advance the field of cancer medicine using real patient data.

2. What are some ways that we use TCGA's data for our own cancer research? (Think about the types of data available and brainstorm some research questions that can be proposed given that data.)

For our own cancer research, TCGA data may be useful in recognizing patterns between certain types of breast cancer, or recognizing common targets that may be worth pursuing in clinical therapeutics.

3. What are the benefits and drawbacks of TCGA or other large publicly available datasets?

TCGA's main benefit is that it makes data widely available for analysis by various groups of researchers, allowing for more studies and work to be done to make progress in the fight against diseases like cancer. However, because these datasets often use real individuals and their information may be considered private, acknowledging the privacy concerns that come with these datasets is extremely important.