Name: Avinash Chauhan

Date: 9/11/2022

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 National Cancer Institute 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 generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data.

Briefly skim the "Timeline & Milestones" page. When did TCGA publish their paper on breast cancer?
October 2012

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?
One way that "Data Use Certification" addresses its respective privacy concerns is the utility of non-identification, where the researchers may not contact the individuals from whom the data was collected from.

TCGA Cancers Selected for Study:

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

Open the breast ductal carcinoma page and read TCGA's provided background. List one interesting fact you found: One interesting fact I found was that about 10% of all cases of advanced breast cancer2 are invasive lobular breast carcinoma.

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,

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: ___________________

Date: ____________________

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 33 % 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: TP53 and PIK3CA had the highest rates of mutation, making up 60% of the most frequently mutated genes.

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 ultimately to provide a stratification mechanism among cohorts of sequenced cancer samples, allowing for the potential to discover major cancer-causing genome alterations.

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.)

One potential way is to isolate the data towards one particular cancer type and then utilize genomic sequencing methods to find commonalities that are inherent to that particular cancer. Additionally, we can highlight particular populations that have higher rates of particular cancers by segmenting according to various demographics (race, age, gender, etc.).

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

The benefits to TCGA are intuitive, in that it offers an unforeseen element of transparency. Even college qbio students have the opportunity to explore the data and develop novel insights based off the data. The drawbacks are largely privacy related, necessitating a cognizance to the ethics of this approach. Data security and privacy concerns could potentially arise.