Name: ____Chloe Liu______

Date: ___________________

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: Cancers of different tissues can share the same alterations and be biologically more similar to each other than to other tumors of the same tissue of origin

Briefly skim the "Timeline & Milestones" page. When did TCGA publish their paper on breast cancer? 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? TCGA policies require that all PIs contributing annotated biospecimens provide documentation to the Biospecimen Core Resource and the Project Team that their IRBs have approved the use of data and specimens for TCGA studies

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: ___________Men can also have breast cancer___________

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 and GATA3

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 __67__ 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 _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: __ductal and lobular neoplasms is the most common disease type;____

____

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?

___To provide a publicly available dataset for researchers to have multi-omic data analysis on various cancers_____

____

____

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.)

___Predict or model the mutation rate of TP35/other genes, and the frequency of different kinds of mutation________

___Does a particular gene affect a particular kind of breast cancer subtype more than the others?_______________

____

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

__TCGA has a clear classification of the statistics of different cancer subtypes, and provide us with data from______

different -omics aspects. But TCGA lacks the normal sequencing data that can be compared with the tumor

data. We need to search for other datasets such as GTEx for more data.

____