WEEK 5 ASSIGNMENT

Introduction to Computational Biology – BIOL 509000 | Fall 1 2020

Christina Morgenstern

Question #1: On the GDC Data Portal home page, there is a graphical representation of the cases within the data repository, given by primary site. What are the top 5 (by number of cases) primary tissue sites? For each of these top five, how many cases fall within that category?

The top five primary tissue sites affected are:

1. Lung (12.025 cases)
2. Breast (9.114 cases)
3. Bone Marrow (8.996 cases)
4. Colorectal (8.073 cases)
5. Nervous System (3.850 cases)

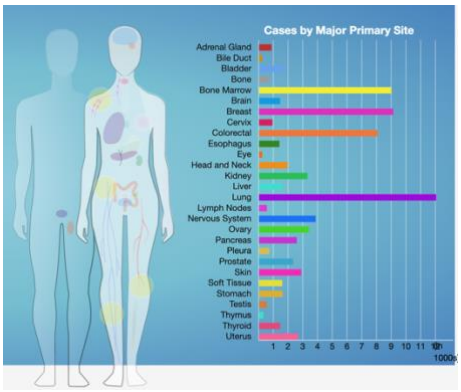

Figure 1. Graphical representation of cases by major primary site on the GDC Data Portal home page.

Question #2: How many cases are available within this Project? How many primary site(s) are represented? Name the primary site(s).

If project ID is TARGET-AML (and Cancer Gene Census is True) 17 cases are listed. A single primary site is represented: Hematopoietic and reticuloendothelial systems

Deleting the default filter Cancer Gene Census is True reports 2.146 cases from Hematopoietic and reticuloendothelial systems (2.076) and 70 cases from Unknown .

Question 3: Consider some of the clinical information available associated with these cases. Propose a criterion by which you could classify these cases into two (or more) categories which could drive a subsequent analysis to determine if transcriptome-level differences exist between these two (+) groups. Then, use the search criteria to determine how many samples (of the total samples for this single primary site) would fall into each of these categories.

One could assess transcriptome-level differences between gender (male, female), race (white, black or African American, Asian, Unknown) or ethnicity (non Hispanic or Latino, Hispanic or Latino, Unknown).

Considering gender, the following number of cases fall into either category:

Male: 10 cases
Female: 7 cases

Considering the full dataset, the following number of cases for gender were found:

Male: 1.063 cases
Female: 970 cases
Unknown: 19 cases

Question #4: How many disease types are listed in this section? For all the disease types, how many different tissue sites are reported?

This information is more easily accessed through the exploration tab.
There are 21 disease types for *CCND3* listed:

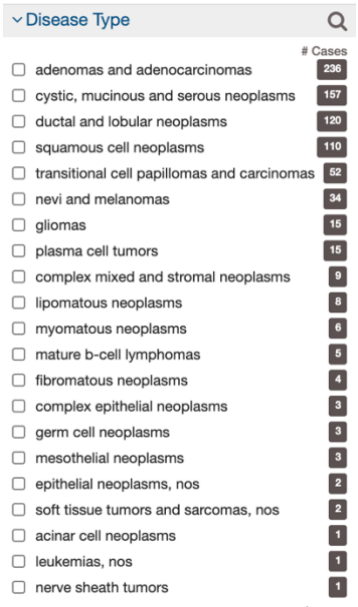

Figure 2. Disease types listed for cases with associated *CCND3* mutations.

There are 36 primary sites for *CCND3* listed:

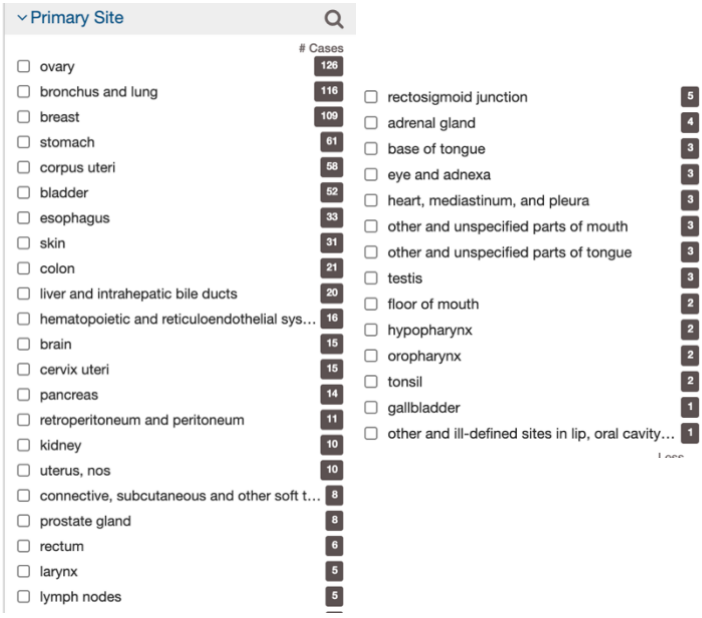

Figure 3. Primary sites listed for cases with associated *CCND3* mutations.

Question #5: Consider the primary site for which cases contain a mutant version of this gene. What are the top five primary sites, and how many cases fall in each category?

There are 100 cases with a mutation.

The top five primary sites with a mutant version of *CCND3* are:

1. Corpus uteri (31 cases)
2. Hematopoietic and reticuloendothelial systems (16 cases)
3. Bladder (12 cases)
4. Bronchus and Lung (8 cases)
5. Breast (5 cases)

Question #6: How many of these cases are reported as female? As male?

Female: 61
Male: 38
Missing entry: 1

Question #7: How many of the reported mutations are missense? Frameshift? Result in a new stop codon?

Out of 37 mutations:
Missense: 34
Frameshift: 2
Stop gained: 1

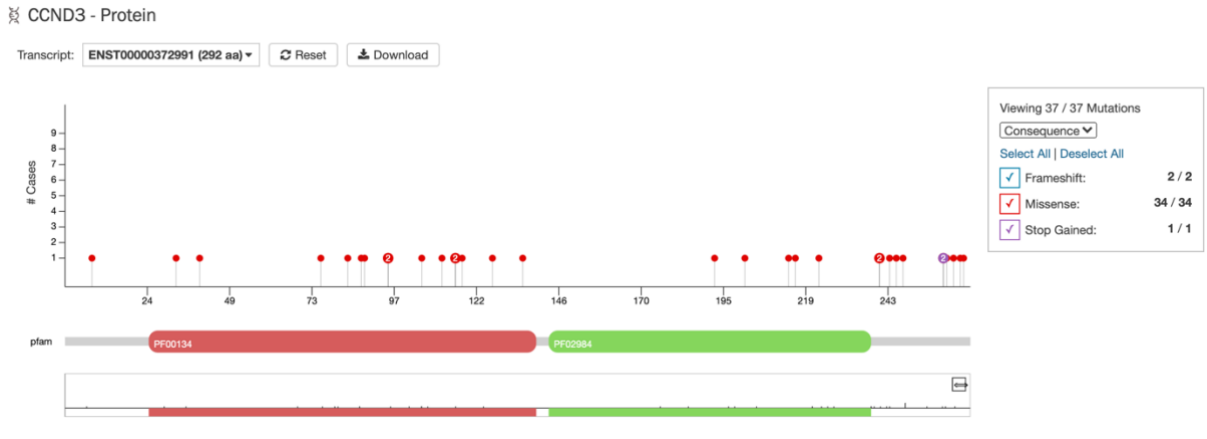

Figure 4. Lollipop chart of CCND3 protein.

Question #8: For each of these high impact mutations, report the "AA Change" that is reported.

High impact mutations in CCDN3 (see Fig. 3 for labels)
1: Q260*
2: R271Pfs*53
3: P284Cfs*12

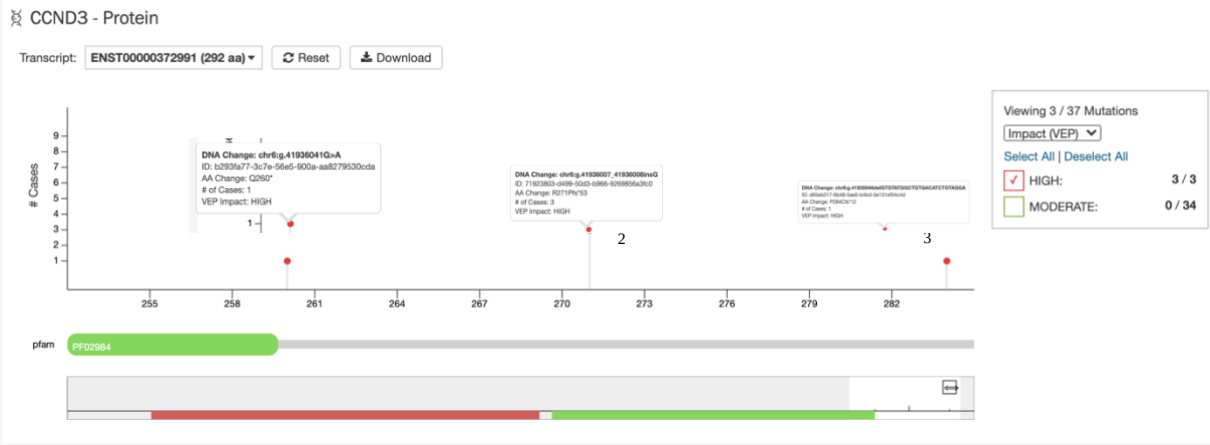

Figure 5. Lollipop chart of CCND3 protein with high impact mutations shown.

Question #9: What external repository is reported as an external reference for this mutation?

The COSMIC (Catalogue Of Somatic Mutations In Cancer) is reported as external repository for high impact mutation 1.

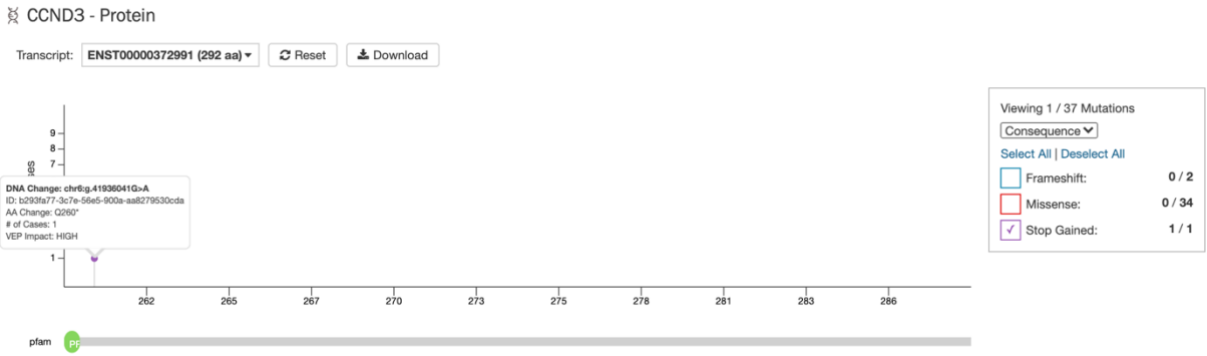

Figure 6. Lollipop chart of CCND3 protein with stop gained mutation.

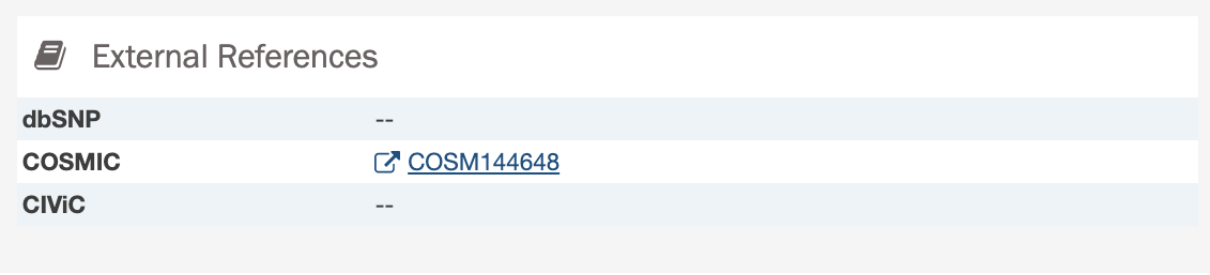

Figure 7. External repository listed for identified Stop gained mutation.

Question #10: What is the project name and disease type listed for the sample that had this mutation?

The project name is: TCGA-BLCA
The disease types listed: Adenomas and Adenocarcinomas, Epithelial Neoplasms, NOS, Squamous Cell Neoplasms, Transitional Cell Papillomas and Carcinomas.

Figure 8. Project name and disease type associated with stop gained mutation.