Introduction to Cancer Genomics Workshop – 02/14/2019

In this workshop you will learn about the fundamentals of mutation identification in cancer. By the end of the course you should be able to generate a full R markdown for your assigned TCGA ID, including plots and identification of actionable mutations. Below is an outline of the course.

We will start off with a 20 minute introduction talking about historical and general workflows. This will be followed by a workshop designed to be completely interactive. There will be a pre-written R script that you can follow along with. For each line of code we will describe what it's doing and if necessary provide background information for specific sections. The workshop will also contain "bonus/challenge" questions that you are expected to try on your own. This will include applying what you have learnt to your own assigned TCGA IDs. We will walk around during this time in case there are questions. Lastly, we will also have homework questions that will be supplied at the end of the workshop.

Introduction (20 minutes):

Section I – Sequencing and Analysis Pipelines

1. The impact of next generation sequencing in cancer biology
2. The size of modern data sets (e.g. TCGA) and their current applications, e.g. in healthcare
3. Types of sequencing, e.g. WES and WGS, and also DNA vs RNA
    a. Pros and cons and how each modality can be integrated.
4. Basic bioinformatics workflow: from sample to identifying potentially targetable mutations. (the workshop will be based on the latter part)
5. Standard pipelines to go from sequence reads to actionable information:
    a. Nuances to consider
    b. Thresholding and filtering variants
6. What does a standard genome look like and how does it inform the analysis of cancer genomes – germline vs somatic mutations

Section II - **Basic vocabulary and concepts of mutations**

1. Classes of somatic mutations
    - Single nucleotide variants
        - Coding
        - Synonymous (Silent)
        - Non-Synonymous (Missense)
        - Nonsense
        - Noncoding (e.g. UTR)
        - Intronic
        - Intergenic
        - Splice site variants

- Small regional mutations
    - Insertion
    - Deletions
    - Duplications

2. Deciphering nomenclature of sequence variations
    a. DNA
    b. Protein

Section III – How to prioritize variants – "Driver" or "Passenger"

1. Coding vs non-coding – category prioritization
2. Variants in cancer-associated genes
3. Variant severity prediction algorithms
4. Druggable variants

Section IV - Workshop

- We will start off with basic data mining. You will learn how to access cbioportal's API through R and access all of the datasets available to the public
- From here we will download the breast cancer dataset relevant to this course
- Although this is not a course in R per se, you will learn how to manipulate/wrangle mutations within the context of a data-frame
- We will then study the mutations specific to the breast cancer dataset:
    - This includes how to find the most frequently mutated genes and types of mutations
    - How to look to see if there are any interesting patterns
- At this point we will show you how to plot and understand your data visually
- You will then learn how to further filter your variants by
    - Pathogenicity scores
    - Germline mutations
    - Specific mutation types
- We will take the above and replot the results
- Now we will show you how to query for specific variants and tabulate the results (e.g. BRCA gene variants)
- With this dataset we will show you how to cross-reference with existing data, for example to look for mutations that are possibly actionable: diagnostic, prognostic and or druggable.

- Based on the above you will now use your own assigned TCGA ID to compare to the overall breast cancer set and re-run through the same exercises above
- We will show you how to neatly package this into a report and list all of the potentially relevant mutations for your TCGA sample ID
- Finally, we will wrap-up with a few other ways to cross-reference your data and go over other potential datasets you can mine

Section V - Summary and further exercises

Here we will summarize and talk about the strengths and shortcomings of the current cancer genomics field. There will also be some "homework" or extra take-home exercises.