This task involves fetching the latest TCGA data for somatic variants found in the genomes of colon cancer patients, cleaning it up, and performing some analysis on it.  You may use whatever environment you are comfortable with.  In addition to answering the questions, please share any scripts you created (or representative screenshots of work not involving scripts) so we can see the process you used to solve them.

The highlighted text gives some context behind the questions.

1. Download data

Download a zip file of the MAF data files from here:

<https://drive.google.com/file/d/14FE4uC0ooYcCLJvoULc35COIT8yO1pqF/view?usp=sharing>

Download a tsv file of the CDR clinical data from here:

<https://drive.google.com/file/d/1s-KcWj1RmoDkHBJokHx8eshYcMdLiuPV/view?usp=sharing>

FYI - Here is some information about how these data were originally downloaded from the primary sources, and some information about the file contents.

Get MAF data from GDC - go to <https://portal.gdc.cancer.gov/>

Click Repository button.

Under Files tab, Data Format = maf, Access = open; under Cases tab, Project = TCGA-COAD

Click Add All Files to Cart (should be 457 files, 83MB). Download manifest.

Use the GDC Data Transfer tool(  <https://gdc.cancer.gov/access-data/gdc-data-transfer-tool>  )   to download them using the manifest just created. (creates 457 directories, each containing 1 compressed MAF plus some auxiliary files).

The .maf.gz files are gzip-compressed TSV tables, with some comment lines beginning with #, followed by a header row, followed by the actual data.  There is one mutation per line.  Each MAF file was generated from two samples, a tumor and a normal sample from the same patient.  The sample names are given in the Tumor\_Sample\_Barcode and Matched\_Norm\_Sample\_Barcode columns.  We will be referring to the Participant field of the barcode; see here for more information: <https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/>

The annotations.txt file gives info about the quality of the sample and how it was processed.

Get Clinical data from the PanCan CDR paper <https://gdc.cancer.gov/about-data/publications/PanCan-Clinical-2018>

Excel file lives here:

[TCGA-CDR-SupplementalTableS1.xlsx](https://api.gdc.cancer.gov/data/1b5f413e-a8d1-4d10-92eb-7c4ae739ed81)

<https://api.gdc.cancer.gov/data/1b5f413e-a8d1-4d10-92eb-7c4ae739ed81>

Save the “TCGA-CDR” tab (the first one) of the Excel file as a TSV file.

Note that the second tab of the Excel file, “TCGA-CDR\_Notes”, describes the fields in the first tab.

1. Filter out irregular samples, ones that have an annotation other than category = General.

Many forms of chemotherapy intentionally cause DNA damage.  While the intent when enrolling patients in the TCGA study was to include only patients who had never been treated before, in some cases it was later discovered that the patient had received treatment.  Such patients will show different patterns of mutations than patients that have never been treated. Also, some samples were preserved using the FFPE method rather than the flash-frozen method, which also induces its own characteristic type of mutation.  We want to exclude these types of samples from our analysis. A row where the category listed in the annotations file is not General indicates that there is something irregular with the sample which is why it would make sense to discard the sample for our analysis.

The category=General provides information about the issues related to the data processing of the sample. The particular issue raised with these samples is not relevant for our purposes.  To generate MAFs for TCGA, the outputs of four different mutation calling algorithms were merged; the annotations with category=General indicate that for these samples only three algorithms were used.

For instance, keep the sample in this directory

mafs/ffe712b7-8eb1-4d6c-bb69-7bb86eef3a5f$ cut -f 5 annotations.txt

category

General

But discard this sample:

mafs/ff94da29-7fef-45ad-b73c-6c6f8acd8bfa$ cut -f 5 annotations.txt

category

Prior malignancy

General

This should discard approximately 60 samples.

How many samples need to be discarded?

1. Merge the MAFs you will keep into a single file.  Keep just one header, and discard the comment lines that begin with #.

There will be around 220,000 lines.  How many lines of data are there?

1. There are biological replicates in this data set - while most patients have only a single tumor sample, some patients have two or three.  These need to be deduplicated to avoid messing up the statistics.

For patients with two or more tumor samples, keep just the one that alphabetically sorts last.  Typically, this sample is the one that was sequenced most recently.  In some cases, the reason for the later sample was poor quality in the first sample.

 For instance, patient 2674 has three tumor samples represented in Tumor\_Sample\_Barcode:

1. TCGA-A6-2674-01A-02D-1953-10
2. TCGA-A6-2674-01A-02D-A270-10
3. TCGA-A6-2674-01B-04D-A270-10

Keep only the mutations associated with TCGA-A6-2674-01B-04D-A270-10, and discard data for the other two tumor samples.

There should be around 175,000 lines remaining.  How many lines of data remain?

1. Enrich the data for severe mutations.  Keep only lines where either PolyPhen score contains possibly\_damaging or probably\_damaging, or where the SIFT score contains deleterious or deleterious\_low\_confidence, and remove the rest.

Some changes to the DNA do not cause any change in which amino acid it codes for, or changes the amino acid to one that is chemically similar (eg as far as acid vs base, and hydrophilic vs hydrophobic).  Other DNA changes cause an amino acid to be substituted that is chemically very different.  Or, if a length of DNA is inserted or deleted that is not a multiple of three, this causes a frameshift, and all DNA downstream from that point will likely map to very different amino acids.  PolyPhen and SIFT are two models that attempt to guess the impact of such changes on the structure and function of the resulting protein.

There should be around 70,000 lines remaining.  How many lines of data remain?

1. Find which genes are mutated in the most patients.

If a gene is frequently mutated in cancer patients, more often than one would expect by chance, this is evidence that the mutation may offer a growth advantage to the cancer, ie positive natural selection.

The Hugo\_Symbol column gives the common gene name.   If a gene is mutated more than once in a single patient, count it just once.  The gene most frequently mutated in these patients is KRAS.

How many patients is it mutated in? What are the four next most frequent genes?

1. Determine how the KRAS mutation status affects the Progression Free Interval, which is how long after the first treatment the patient avoids having a recurrence of the cancer.

The CDR file contains Progression Free Interval (PFI) information for all TCGA projects.  To line it up to the MAF file, bcr\_patient\_barcode in the CDR file matches the first part of the Tumor\_Sample\_Barcode in the MAF file.  For instance, the line in the CDR file with bcr\_patient\_barcode = TCGA-3L-AA1B matches to the line in the MAF file where Tumor\_Sample\_Barcode = TCGA-3L-AA1B-01A-11D-A36X-10 .

The CDR file is missing information from a handful of the patients in our set.  Discard patients that are missing this data.

For TCGA, clinical data tended to become available later than molecular data such as DNA mutations.  In some cases it never showed up.

How many patients are missing from the CDR file?

The Progression-Free Interval is how long the patient goes after the initial diagnosis until the cancer comes back.  If you check with the patient on some date and find that the cancer has in fact returned, then you can calculate the exact Progression-Free Interval.  However, if you check with the patient and they do not have cancer at that date, then you have no way of knowing whether the cancer will never return (yay!), or whether it will return but at some future date; in this case, the Progression-Free Interval is said to be “censored” - all you know is that the return (if any) is later than the time you checked, you do not know the exact time.

The PFI.time and PFI fields give the Progression-Free Interval time and whether or not that time is exact vs censored.  The PFI.time field gives the date (in days since initial cancer diagnosis) of the patient’s followup visit with their doctor.  And, if the PFI field = 1, the cancer has returned, and the PFI.time gives an exact time.  If the PFI field = 0, the cancer has not returned by the date in PFI.time, and might or might not return in the future; the PFI.time represents a censored time.

Check how mutations in KRAS affect the Progression Free Interval. First, plot the Kaplan-Meier survival curves for patients with and without a KRAS mutation. Then, test whether the difference is statistically significant.  What does this say about patients who have a mutation in KRAS?