Exploring the Relation Between Lung Cancer and the TP53 Gene

BMI 5330 Final project
Jacob Stimes
August 9th, 2021

Introduction and Background

Most people will be impacted by cancer in one way or another throughout their lifetime, given that one out of every six deaths in the world is caused by some form of this disease, according to Roser, 2019. One form, lung cancer, is reported by the CDC in 2021 to be responsible for more deaths than any other type of cancer. Many people commonly know that doctors once encouraged smoking, although we have since learned that smoking is estimated to cause the majority of lung cancer, by instigating genetic changes (Walser, 2008). However, cases of lung cancer have developed in never-smokers too, suggesting that abstaining from smoking alone may not prevent the disease (Wakelee, 2007). Therefore, there are still merits to studying the disease and its genetic linkages, in both smokers and non-smokers, for the purpose of discovering other potential causes and insights about its development, prevention, and treatment.

Previous work by Mogi in 2011 has concluded that mutations in the tumor suppressor gene TP53 play a role in lung cancer (as well as other cancers). According to Mogi, this gene assists with DNA repair and warrants further investigation given its potential for assisting with pharmacological discoveries. The TP53 gene has been linked to other cancers by Olivier in 2010, suggesting that a deeper analysis of this gene's relation to lung cancer could assist with our understanding of other serious diseases, too. As recently as 2018, work by Li has shown that TP53 mutations can help inform prognoses of certain cancers.

Given the incidence and consequences of lung cancer worldwide and the perceived connection the condition has with the TP53 gene, I describe here my research into the genetic linkage between the TP53 gene and lung cancer, which specifically seeks to understand the range of variations in the gene, their associations with lung cancer and other phenotypes, and the gene's conservation status among other species. Some of the broader aims of this research were to:

- Assist others in learning more about this disease's cause: Lung cancer is one of the most common cancers: more than 200,000 new cases are expected to develop in Americans this year, with more than 100,000 Americans expected to die from the condition this year (American Cancer Society, 2021).
- Guide potential diagnostic and therapeutic methods: Current lung cancer treatments include invasive surgeries and harmful chemo and/or radiation therapies (Stinchcombe, 2008). Therefore, there is value in discovering insights that may guide the development of safer treatment options.

- Possibly uncover environmental or evolutionary factors surrounding the gene and disease: Non-smokers can also develop this condition, even though healthy, non-smokers are generally not screened for the disease (Mayo Clinic, 2021). Identifying any biomarkers that could indicate risk of the disease could help prioritize preventative treatments in the population.

Results

Overview of the TP53 gene

According to ClinGen, TP53 is a protein-coding gene that generates a tumor suppressor protein (Rehm, 2015). This gene plays an important role in regulating cell division. Given the intended function of the gene, it is no surprise that this gene has been linked to various forms of cancer in previous literature. ClinGen also reports linkages between the gene and Li-Fraumeni syndrome, a condition that predisposes those with it to increased risk of multiple forms of cancer.

From the UCSC Genome Browser (Kent, 2002), some insights can be gleaned regarding prominence of repeat elements and conservation with other species (see Figure 1 for gene visualization and Figures 2, 3, and 4 for some highlighted sequence conservation results. More conservation alignments can be viewed here: http://genome.ucsc.edu/cgi-bin/hgc?hgsid=1137781787_fSOEarwZLOkrG0B5ZbCepg2Umz0M&db=hg19&c=chr17&l=7582345&r=7601494&o=7582345&t=7601494&g=multiz100way&i=multiz100way).

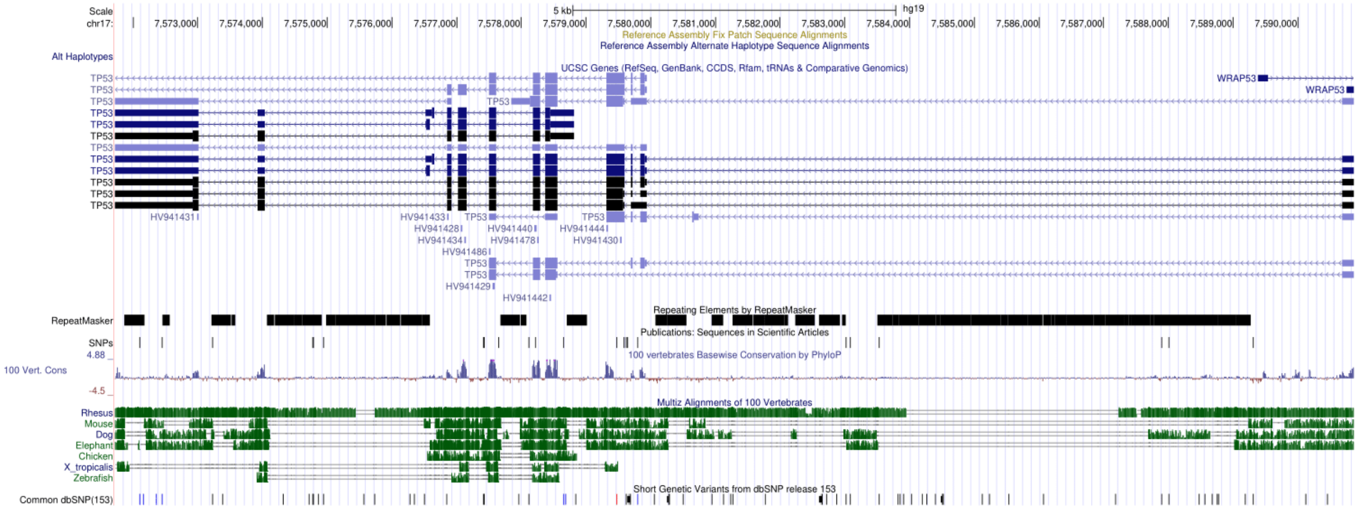
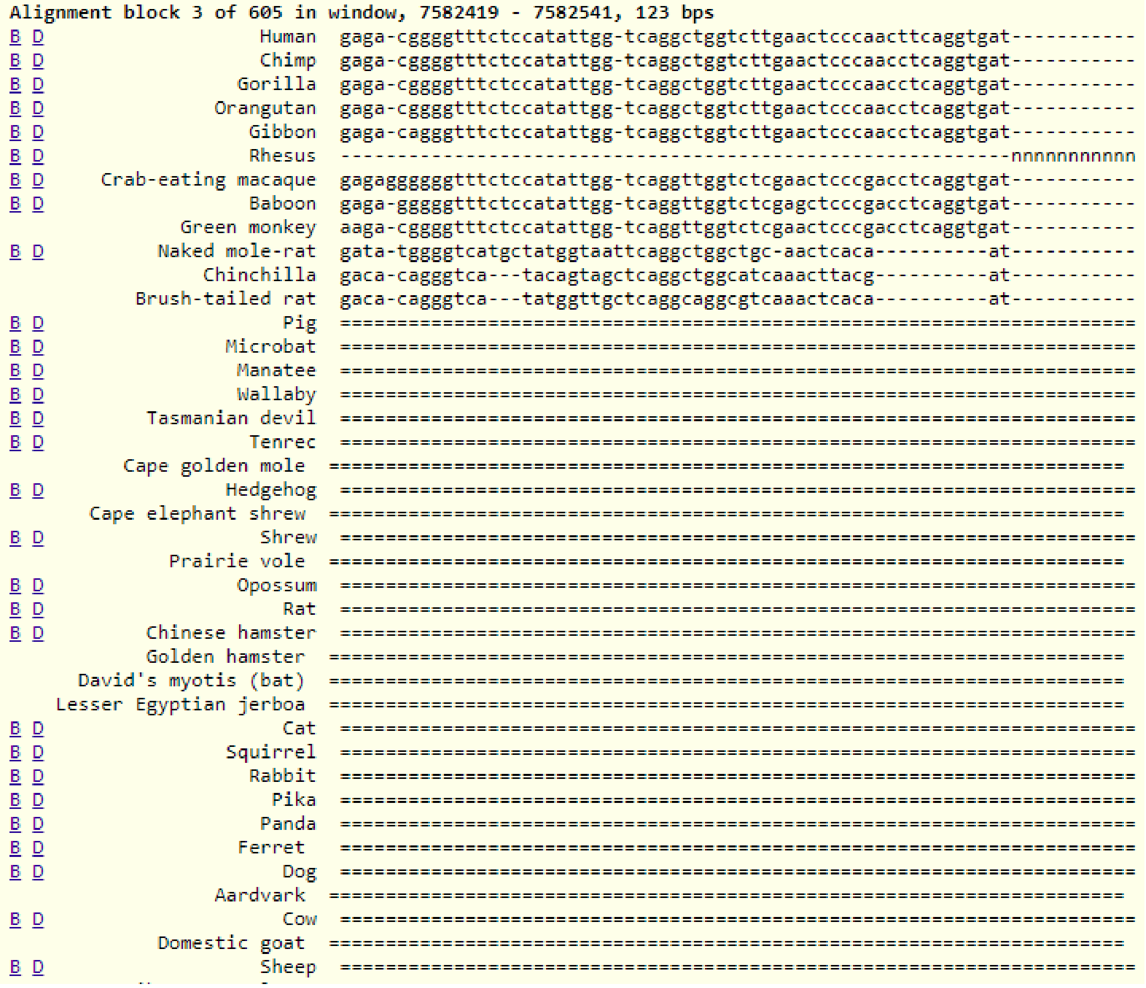
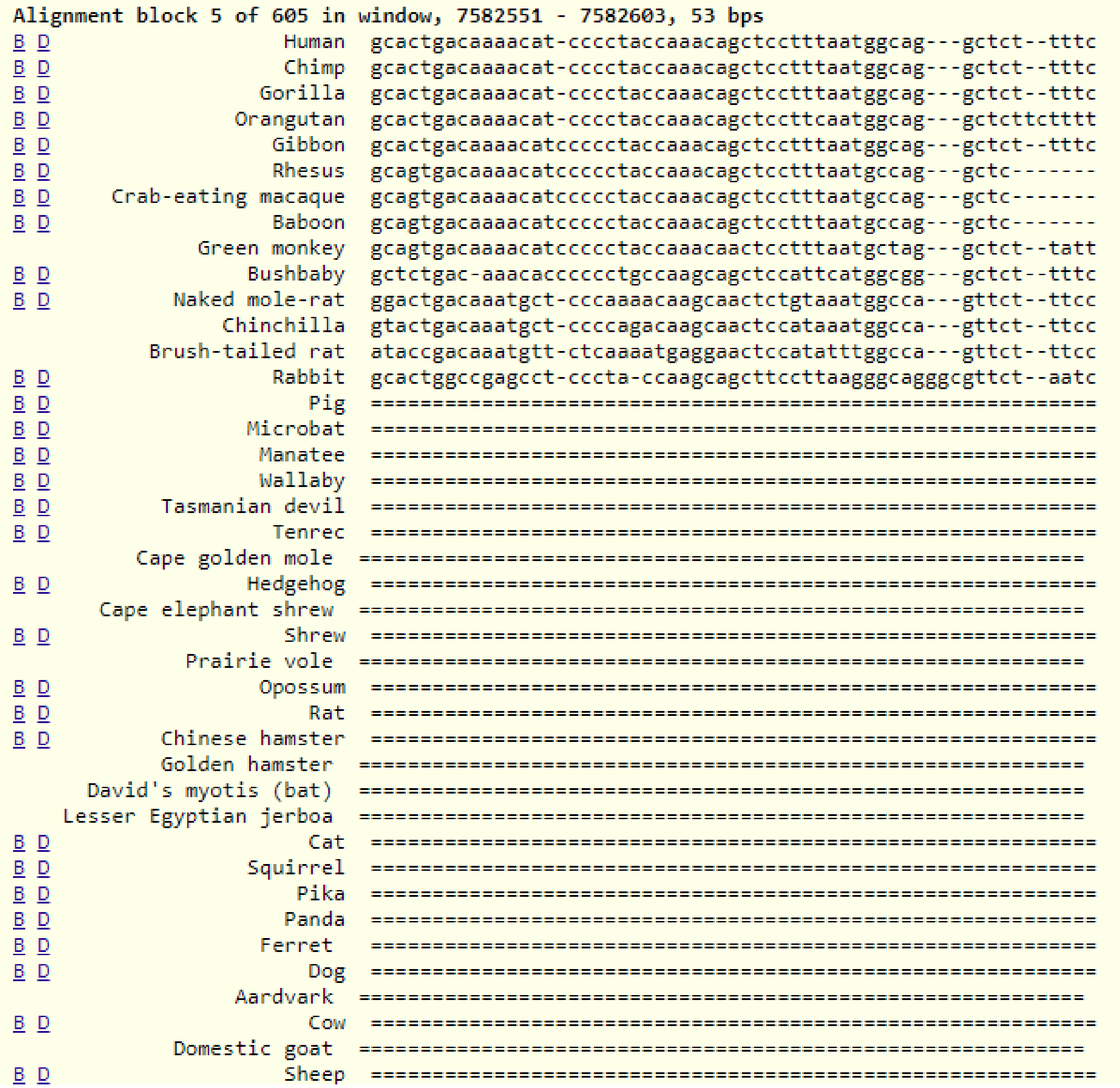


Figure 1: *TP53 genome annotation in UCSC Genome Browser (GRCh37/hg19 assembly). Multiple haplotypes are shown, aligned with repeat regions and conservation status with other species.*
http://genome.ucsc.edu



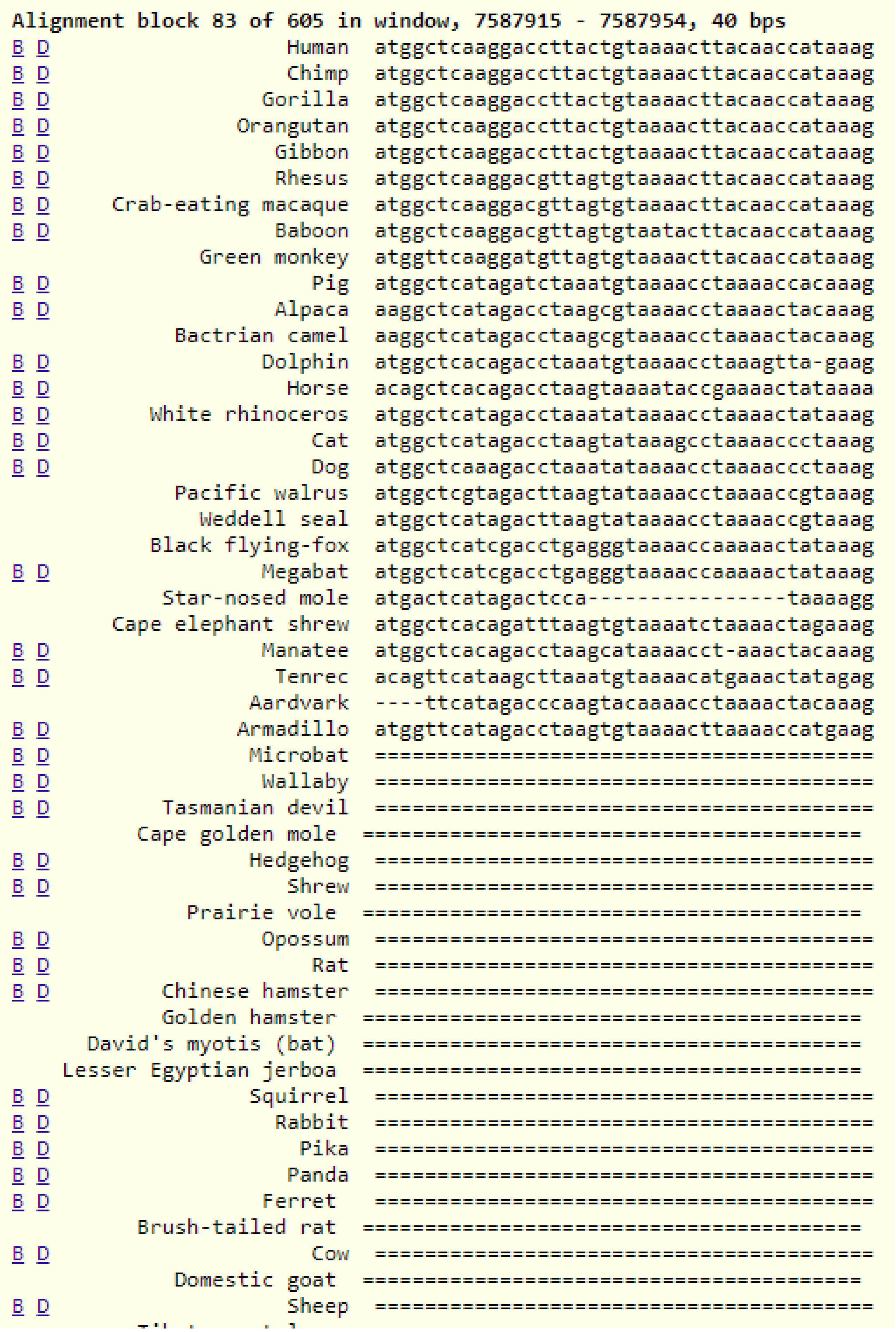
```
Alignment block 3 of 605 in window, 7582419 - 7582541, 123 bps
B D                   Human  gaga-cggggtttctccatattgg-tcaggctggtcttgaactcccaacttcaggtgat-----------
B D                   Chimp  gaga-cggggtttctccatattgg-tcaggctggtcttgaactcccaacctcaggtgat-----------
B D                 Gorilla  gaga-cggggtttctccatattgg-tcaggctggtcttgaactcccaacctcaggtgat-----------
B D               Orangutan  gaga-cggggtttctccatattgg-tcaggctggtcttgaactcccaacctcaggtgat-----------
B D                  Gibbon  gaga-cagggtttctccatattgg-tcaggctggtcttgaactcccaacctcaggtgat-----------
B D                  Rhesus  -----------------------------------------------------------nnnnnnnnnnn
B D     Crab-eating macaque  gagaggggggtttctccatattgg-tcaggttggtctcgaactcccgacctcaggtgat-----------
B D                  Baboon  gaga-gggggtttctccatattgg-tcaggttggtctcgagctcccgacctcaggtgat-----------
            Green monkey  aaga-cggggtttctccatattgg-tcaggttggtctcgaactcccgacctcaggtgat-----------
B D          Naked mole-rat  gata-tggggtcatgctatggtaattcaggctggctgc-aactcaca----------at-----------
                 Chinchilla  gaca-cagggtca---tacagtagctcaggctggcatcaaacttacg----------at-----------
           Brush-tailed rat  gaca-cagggtca---tatggttgctcaggcaggcgtcaaactcaca----------at-----------
B D                     Pig  ======================================================================
B D                Microbat  ======================================================================
B D                 Manatee  ======================================================================
B D                 Wallaby  ======================================================================
B D         Tasmanian devil  ======================================================================
B D                  Tenrec  ======================================================================
          Cape golden mole  ======================================================================
B D                Hedgehog  ======================================================================
       Cape elephant shrew  ======================================================================
B D                   Shrew  ======================================================================
              Prairie vole  ======================================================================
B D                 Opossum  ======================================================================
B D                     Rat  ======================================================================
B D         Chinese hamster  ======================================================================
            Golden hamster  ======================================================================
      David's myotis (bat)  ======================================================================
    Lesser Egyptian jerboa  ======================================================================
B D                     Cat  ======================================================================
B D                Squirrel  ======================================================================
B D                  Rabbit  ======================================================================
B D                    Pika  ======================================================================
B D                   Panda  ======================================================================
B D                 Ferret   ======================================================================
B D                     Dog  ======================================================================
                  Aardvark  ======================================================================
B D                     Cow  ======================================================================
             Domestic goat  ======================================================================
B D                   Sheep  ======================================================================
```

***Figure 2**: TP53 highlighted conservation sequence alignment from position 7582419-7582541.*



```
Alignment block 5 of 605 in window, 7582551 - 7582603, 53 bps
B D                   Human  gcactgacaaaacat-cccctaccaaacagctcctttaatggcag---gctct--tttc
B D                   Chimp  gcactgacaaaacat-cccctaccaaacagctcctttaatggcag---gctct--tttc
B D                 Gorilla  gcactgacaaaacat-cccctaccaaacagctcctttaatggcag---gctct--tttc
B D               Orangutan  gcactgacaaaacat-cccctaccaaacagctccttcaatggcag---gctcttcttt
B D                  Gibbon  gcactgacaaaacatcccctaccaaacagctcctttaatggcag---gctct--tttc
B D                  Rhesus  gcagtgacaaaacatcccctaccaaacagctcctttaatgccag---gctc-------
B D     Crab-eating macaque  gcagtgacaaaacatcccctaccaaacagctcctttaatgccag---gctc-------
B D                  Baboon  gcagtgacaaaacatcccctaccaaacagctcctttaatgccag---gctc-------
               Green monkey  gcagtgacaaaacatcccctaccaaacaactcctttaatgctag---gctct--tatt
B D                Bushbaby  gctctgac-aaacacccccctgccaagcagctccattcatggcgg---gctct--tttc
B D          Naked mole-rat  ggactgacaaatgct-cccaaaacaagcaactctgtaaatggcca---gttct--ttcc
                 Chinchilla  gtactgacaaatgct-ccccagacaagcaactccataaatggcca---gttct--ttcc
           Brush-tailed rat  ataccgacaaatgtt-ctcaaaatgaggaactccatatttggcca---gttct--ttcc
B D                  Rabbit  gcactggccgagcct-cccta-ccaagcagcttccttaagggcagggcgttct--aatc
B D                     Pig  ===========================================================
B D                Microbat  ===========================================================
B D                 Manatee  ===========================================================
B D                 Wallaby  ===========================================================
B D         Tasmanian devil  ===========================================================
B D                  Tenrec  ===========================================================
          Cape golden mole  ===========================================================
B D                Hedgehog  ===========================================================
       Cape elephant shrew  ===========================================================
B D                   Shrew  ===========================================================
              Prairie vole  ===========================================================
B D                 Opossum  ===========================================================
B D                     Rat  ===========================================================
B D         Chinese hamster  ===========================================================
            Golden hamster  ===========================================================
      David's myotis (bat)  ===========================================================
    Lesser Egyptian jerboa  ===========================================================
B D                     Cat  ===========================================================
B D                Squirrel  ===========================================================
B D                    Pika  ===========================================================
B D                   Panda  ===========================================================
B D                 Ferret   ===========================================================
B D                     Dog  ===========================================================
                  Aardvark  ===========================================================
B D                     Cow  ===========================================================
             Domestic goat  ===========================================================
B D                   Sheep  ===========================================================
```

Figure 3: *TP53 highlighted conservation sequence alignment from position 7582551-7582603.*



```
Alignment block 83 of 605 in window, 7587915 - 7587954, 40 bps
B D                   Human  atggctcaaggaccttactgtaaaacttacaaccataaag
B D                   Chimp  atggctcaaggaccttactgtaaaacttacaaccataaag
B D                 Gorilla  atggctcaaggaccttactgtaaaacttacaaccataaag
B D               Orangutan  atggctcaaggaccttactgtaaaacttacaaccataaag
B D                  Gibbon  atggctcaaggaccttactgtaaaacttacaaccataaag
B D                  Rhesus  atggctcaaggacgttagtgtaaaacttacaaccataaag
B D     Crab-eating macaque  atggctcaaggacgttagtgtaaaacttacaaccataaag
B D                  Baboon  atggctcaaggacgttagtgtaatacttacaaccataaag
               Green monkey  atggttcaaggatgttagtgtaaaacttacaaccataaag
B D                     Pig  atggctcatagatctaaatgtaaaacctaaaaccacaaag
B D                  Alpaca  aaggctcatagacctaagcgtaaaacctaaaactacaaag
             Bactrian camel  aaggctcatagacctaagcgtaaaacctaaaactacaaag
B D                 Dolphin  atggctcacagacctaaatgtaaaacctaaagtta-gaag
B D                   Horse  acagctcacagacctaagtaaatacсgaaaactataaaa
B D        White rhinoceros  atggctcatagacctaaatataaaacctaaaactataaag
B D                     Cat  atggctcatagacctaagtataaagcctaaaaccctaaag
B D                     Dog  atggctcaaagacctaaatataaaacctaaaaccctaaag
             Pacific walrus  atggctcgtagacttaagtataaaacctaaaaccgtaaag
               Weddell seal  atggctcatagacttaagtataaaacctaaaaccgtaaag
           Black flying-fox  atggctcatcgacctgagggtaaaaccaaaaactataaag
B D                 Megabat  atggctcatcgacctgagggtaaaaccaaaaactataaag
            Star-nosed mole  atgactcatagactcca----------------taaaagg
        Cape elephant shrew  atggctcacagatttaagtgtaaaatctaaaactagaaag
B D                 Manatee  atggctcacagacctaagcataaaacct-aaactacaaag
B D                  Tenrec  acagttcataagcttaaatgtaaaacatgaaactatagag
                   Aardvark  ----ttcatagacccaagtacaaaacctaaaactacaaag
B D               Armadillo  atggttcatagacctaagtgtaaaacttaaaaccatgaag
B D                Microbat  ========================================
B D                 Wallaby  ========================================
B D         Tasmanian devil  ========================================
          Cape golden mole  ========================================
B D                Hedgehog  ========================================
B D                   Shrew  ========================================
              Prairie vole  ========================================
B D                 Opossum  ========================================
B D                     Rat  ========================================
B D         Chinese hamster  ========================================
            Golden hamster  ========================================
      David's myotis (bat)  ========================================
    Lesser Egyptian jerboa  ========================================
B D                Squirrel  ========================================
B D                  Rabbit  ========================================
B D                    Pika  ========================================
B D                   Panda  ========================================
B D                  Ferret  ========================================
          Brush-tailed rat  ========================================
B D                     Cow  ========================================
             Domestic goat  ========================================
B D                   Sheep  ========================================
```

Figure 4: *TP53 highlighted conservation sequence alignment from position 7587915-7587954.*

TP53 Variants

Using the variant_analysis pipeline described in the Methods section, I generated data about TP53 variants. This analysis utilized the GENCODE human genome annotation (Frankish, 2018), 1000 Genomes variant data, dbSNP, and dbVar (specific assemblies & versions discussed below). This analysis produced the following output files, the contents of which are explained in Methods and interpretations discussed in Discussion:

- TP53_variants.csv
- TP53_variants_report.txt
- TP53_exons_variants.csv
- TP53_exons_variants_report.txt
- TP53_pathogenic_snps.csv
- TP53_pathogenic_svs.csv

All of these outputs are available at this URL:
https://github.com/jstimes/GeneAnalysis/tree/main/variant_analysis/data

In addition to the CSV and TXT files listed above, this analysis generated the allele frequency and variant-type distribution plots shown in figures 5, 6, 7, and 8.

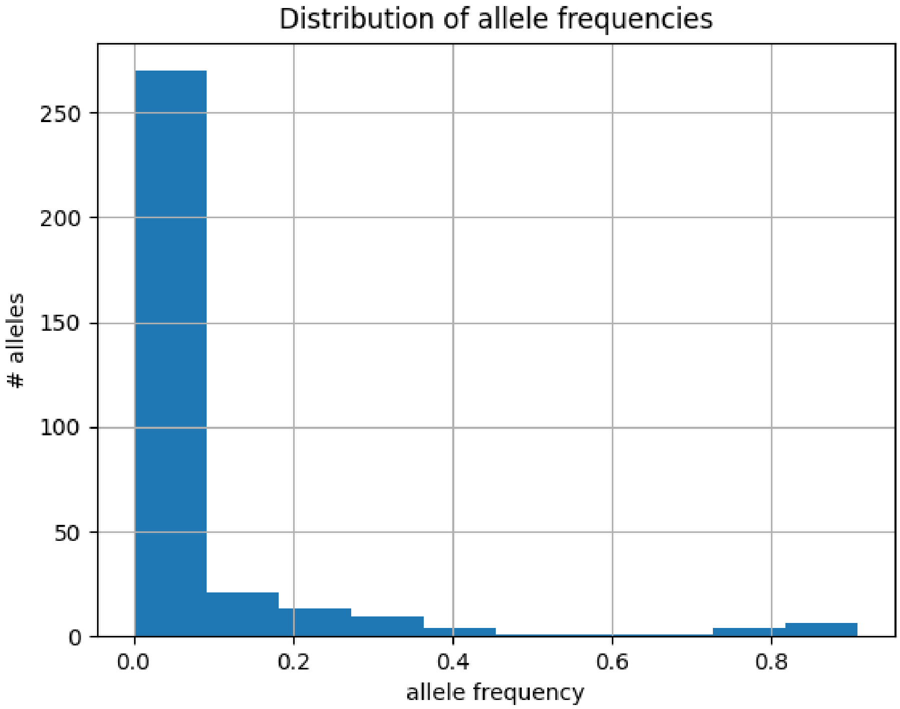


***Figure 5**: Allele frequency distribution for variants on the entire TP53 gene.*

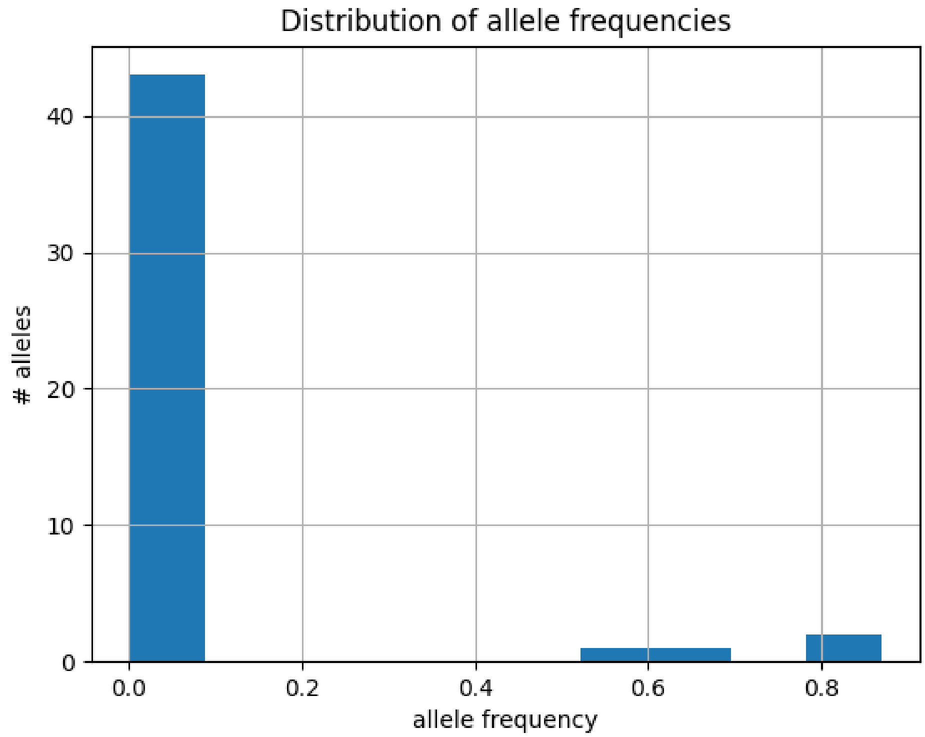


Figure 6: *Allele frequency distribution for variants on exons of the TP53 gene.*

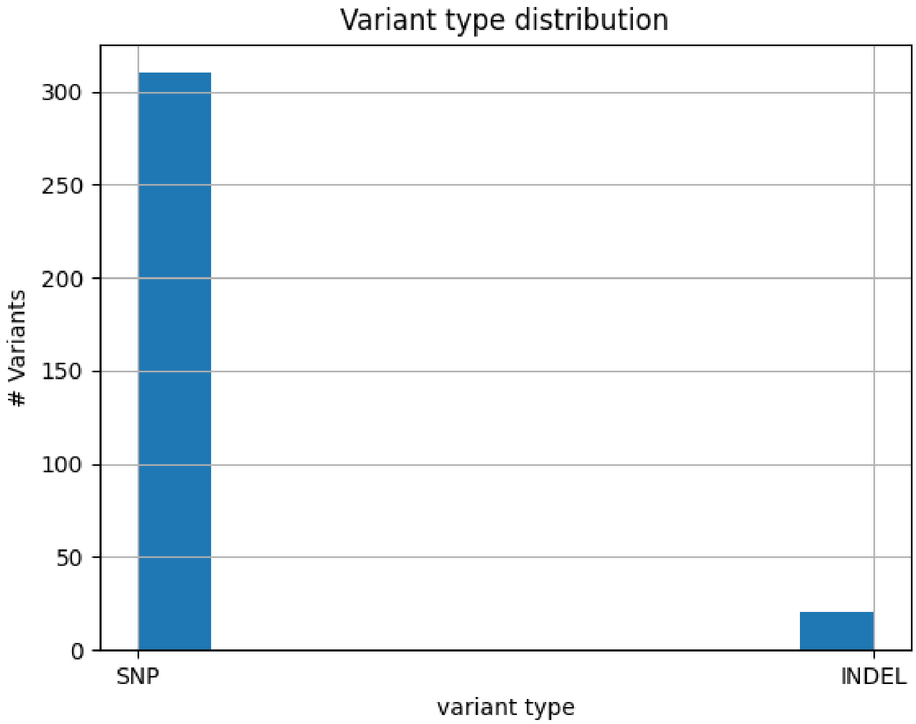


Figure 7: *Variant type distribution for variants on the TP53 gene.*

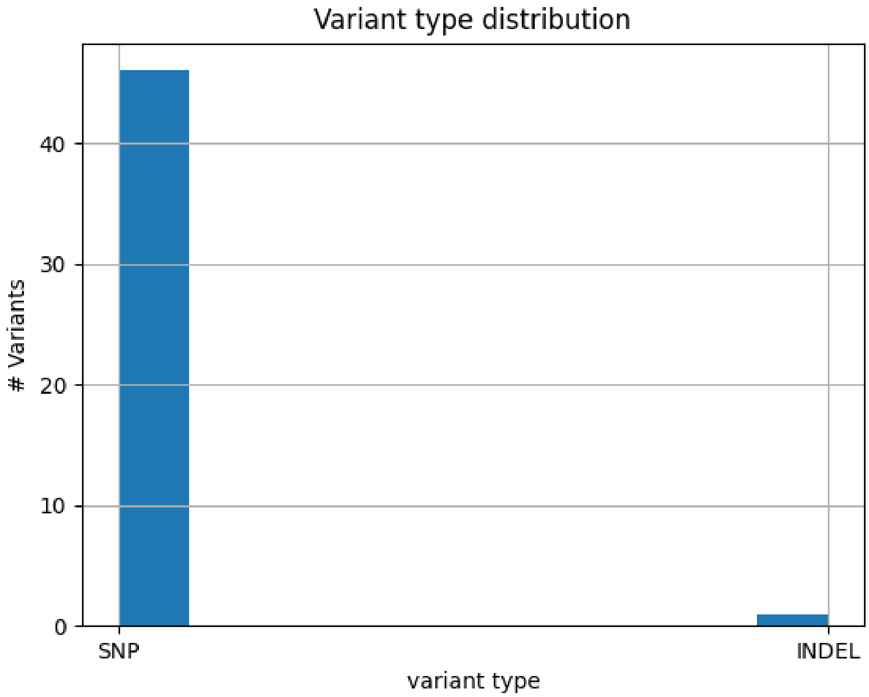


Figure 8: *Variant type distribution for variants on the exons of the TP53 gene.*

Clinical TP53 Variants

Similarly to the analysis above, some variant analysis was conducted using variant data linked to clinical outcomes, such as ClinVar (Landrum, 2017) and GWAS Catalog (Buniello, 2018). This analysis, described in the *clinical_data_analysis* subsection in Methods, produces the following output files (contents explained in Methods, interpretations provided in Discussion):

- TP53_clinvar_reported_conditions.csv
- TP53_gwas_data.csv
- TP53_simplified_gwas_data.csv

All these files can be found at this URL:
https://github.com/jstimes/GeneAnalysis/tree/main/clinical_data_analysis/data

Discussion

Phenotype Findings

Before getting into TP53-specific discussions, I want to elaborate on some lung cancer focused findings. First, it's important to understand that lung cancer could more accurately be described as a set of phenotypes - there are multiple sub-types of lung cancer, including lung adenocarcinoma, lung carcinoid, metastatic lung cancer, squamous cell, and large cell carcinoma (Markman, 2021). This is necessary to keep in mind because searching for only the condition “lung cancer” in both the genetic databases I used as well as the data outputs I generated will leave out relevant findings. In my variant discussions below, I generally considered all the sub-types listed above as being of the phenotype “lung cancer”.

Although this research is largely focused on findings pertaining to the TP53 gene, I also search by phenotype (lung cancer and sub-types) in databases such as ClinVar and GWAS catalog to understand what other kinds of genes and variants are associated with it. According to the GWAS Catalog, some other tumor suppressor genes, for example TP63, are also associated with lung cancer. I also found that TP53 is directly associated with other genes, for example one of its targets is the gene TP53BP1. One particular variant of TP53, VCV000012366.17, is linked with lung cancer according to ClinVar; however, it’s also linked to numerous other cancer-related diseases. All of these findings are somewhat expected: given that TP53 is a tumor suppressor gene, it’s likely it is not only associated with lung cancer, but any cancer that can form when tumor growth is not regulated. Likewise, given that there are multiple genes in the body responsible for regulating cell division, it’s expected that lung cancer is associated with multiple other genes as well.

TP53 Overview

According to the UCSC Genome Browser (*Figure 1),* there are more than 10 haplotypes of the TP53 gene. It is a protein coding gene, and combined with knowledge gleaned from ClinGen, we know this protein is p53. One thing that stood out to me in the Genome Browser was the long intron segments on the latter half of the haplotypes that appear to coincide with repeat regions. This prompted some brief exploration into prior literature on significance of these repeat regions with respect to TP53 function. Some research by Harris, 2009, suggests that p53 may, in addition to suppressing tumors, be used to reduce retroposition of repeat DNA elements. These repeat elements, particularly one known as L1, could prohibit a gene’s expression according to Harris. This could indicate that the presence of and magnitude of repeat elements in a gene could play some role in tumor development (if present in a gene responsible for regulating cell growth or division). I believe this highlights an opportunity for future research to further explore the connection between repeat elements and cancers/tumors.

Also from the UCSC Genome Browser we can make some interpretations about the conservation status amongst species for the TP53 gene. *Figure 1* shows some sequence overlaps with other species, and figures 2, 3, and 4 are some example alignment stretches. One can conclude that this gene has strong conservation with other primates, but not all vertebrate species. This finding could help inform researchers to study primates when seeking knowledge about incidence and treatment of lung cancer / TP53 related conditions.

TP53 Variants (Pathogenic and Clinical Associations)

The variant analysis I conducted yielded numerous interesting results. To begin, I’d like to compare the gene-wide variants and exon-only variants. Naturally, the gene-wide set of variants is far bigger than those only occurring in exons (330 vs 47). Interestingly, the exons-only variants have a higher mean allele frequency than the gene-wide mean allele frequency (0.067 for gene-wide, 0.071 for exons-only). I found this rather surprising actually. An initial assumption I had was that exons are less likely to acquire mutations because they are directly involved with protein coding, so adverse mutations would be selected against (whereas introns may accumulate mutations over time as they, as far as we know, don't affect the translated proteins as much). I believe this reinforces my takeaway from earlier about motivation to further study introns and repeat regions in relation to cancerous diseases.

Exploring the gene and exon variants further, I took a closer look at pathogenic variants. I further cleaned the data to only count variants as pathogenic if one of the 'significances' values included: "risk-factor", "likely-pathogenic", "pathogenic-likely-pathogenic", or "pathogenic" (since other values include "benign", "not likely", etc which are likely not actually pathogenic. I found that the only pathogenic variants in TP53, according to this filtering, occur in the exons. In other words, both the gene-wide and exon-only variants contain the same set of pathogenic variants. The average allele frequency of these is 0.122; the dbSNP IDs are: rs78378222, rs201744589, rs55819519, rs28934576, and rs1042522. This makes sense given that exons are directly involved with protein translation, and an issue in tumor suppression protein translation is more likely to lead to disease.

TP53 is also troubled by structural variants. According to dbVar (and the TP53 dbvar report generated from that data), there are 21 pathogenic structural variations in TP53, all of which are copy number variations except for one deletion/insertion. One aspect of this that I noticed is that the start/stop positions of these variants aren't always entirely enclosed just in chromosome 17 (the chromosome TP53 is on), suggesting that these variations likely cause issues with other genes. Another point of future study would be to study the widespread affect of these variations, not just their association with TP53.

Another insightful observation is the 'origins' field in the reports. There are pathogenic variants with both somatic and germline origins, which reinforces the research showing that lung cancer is not only caused by cancer. I believe being able to combine these results with other datasets (e.g. locations of individuals) could be a valuable further research effort to attempt to connect these germline mutations with environmental factors such as climate and smog levels.

Finally, looking at the clinically identified TP53 / condition associations in TP53_simplified_gwas_data.csv, it becomes even more clear that proper functioning of TP53 is essential not just to preventing lung cancer, but also numerous other conditions. The range of conditions covers not only many types of cancers, but cholesterol, blood pressure, and blood cell regulation conditions (also reinforces that simply searching for 'lung cancer' is insufficient since the lung-cancer-like phenotype here is "Esophageal squamous cell carcinoma"). On one hand, it seems overwhelming that there are so many implications and condition associations for TP53. However, I believe this also suggests that there are many diseases that could be studied to gain further insight into the mechanics of the TP53 gene.

Limitations and Divergence from proposal

Above, I listed some ways to continue this research going forward. Most of those followups are achievable by myself given my current knowledge of the subject and were constrained by time, however, some are limited due to the complexity and missing auxiliary datasets required (e.g. identifying linkages between intron repeat sequences and tumors, combining variant data with environmental datasets).

One way I diverged from my original project proposal was that I did not classify variant types at a very granular level. Some of the data reports, for example TP53_simplified_gwas_data.csv provided this fine-grained detail (for example "missense_variant") but most data sources I used did not provide variants types of that specificity and I did not have the time to determine these myself. Another divergence was lack of formal software testing that I hoped to have added. I relied simply on manual verification of scripts and analyses, but realize in a proper software project there should be automated tests for all significant logic. Finally, one other difference is that I did not generate much data about conservation as I did not fully understand what types of metrics and measurements are used in this sub-field of bioinformatics, and chose to devote more time to the pipeline development than this area of the project.

Methods

Variant analysis

Overview

As a part of this project, I developed a pipeline that generates variant data and synthesizes some figures and reports based on that data. It can be accessed from the following URL: https://github.com/jstimes/GeneAnalysis

It uses a genome annotation file and whole genome VCF file as input, and then, given a gene of interest, computes details about variants associated with that gene. This software utilizes the BEDTools software (Quinlan, 2010) and some open source python packages listed below.

Datasets

Human genome assembly: GRCh37
Genome annotation: From gencode, release 19
(http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/)
Genome variants: 1000 genomes phase 1 analysis
(https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/integrated_call_sets/)
External databases: dbsnp and dbvar via eutilities
(https://www.ncbi.nlm.nih.gov/snp/docs/eutils_help/).

Pipeline Steps

The file `run.sh` runs this analysis pipeline. Data inputs and outputs are expected to be loaded from / written to the 'data' folder. The input flags are as follows:

- -a: The genome **a**nnotation file to use.
- -c: The **c**hromosome number of interest
- -g: The **g**ene of interest
- -v: The variants VCF file to use.

<gene> and <chr> below refer to the values used for the -g and -c flags respectively.

The pipeline follows these steps:

1. Cleans pre-existing output from the 'data' folder.
2. Extracts the genome annotations just for the chromosome of interest.
    a. Written to *chr<chr>_annotation.gff3*
3. Extracts the genome annotations just for the gene of interest, and also generates bed files representing these annotations; one for the whole gene and one just for exons.
    a. Gene annotations written to *<gene>.gff3*
    b. Gene bed file written to *<gene>.bed*
    c. Exons bed file written to *<gene>_exons.bed*
4. Extracts the variants only on the chromosome of interest.
    a. Filtered VCF file written *<chr>_variants.vcf*
    b. Modified VCF file with the INFO column containing the dbSnp ref written to *chr<chr>_variants_adjusted.vcf*
    c. Variants on chromosome bed file written to *chr<chr>_variants.bed*
5. Computes gene and variant intersections.
    a. Whole gene and variant intersections written to *<gene>_variants.bed*
    b. Exon-only variant intersections written to *<gene>_exons_variants.bed*
6. Cleans the intersection data, joining it with dbSnp annotations where possible; generates plots and a text report based on this data.
    a. Whole-gene variant dataset written to *<gene>_variants.csv*
    b. Allele frequency distribution plot written to *<gene>_variants_af_distribution.png*
    c. Variant type distribution plot written to *<gene>_variants_variant_type_distribution.png*
    d. A text report summarizing some findings is written to *<gene>_variants_report.txt*
    e. Exon-only equivalents are written to files with the prefix *<gene>_exons_*
7. Generates another set of data based solely on dbsnp & dbvar data (i.e. not just variants defined in the VCF file that overlap with the gene's annotation; looks up all pathogenic dbsnp and dbvar entries associated with the gene of interest).
    a. Pathogenic dbsnp variant data written to *<gene>_pathogenic_snps.csv*
    b. Pathogenic dbvar variant data written to *<gene>_pathogenic_svs.csv*

Pipeline Outputs

***<gene>_variants.csv* & *<gene>_exons_variants.csv* columns:**

- *chr*: the chromosome of interest
- *start:* start position of this variant in the chromosome, in reference to the human genome assembly defined above.
- *stop*: stop position of this variant.
- *allele_frequency*: The reported allele frequency in the VCF file (AF=).
- *variant_type*: The variant type as reported in the VCF file.
- *dbsnp_id:* ID of the snp in dbSnp. This can be pasted into the dbSnp search bar for finding more information about the variant.
- *dbsnp_variant_type*: The variant type as reported by dbSnp. This is slightly more specific than the VCF variant type (e.g. this reports 'del' instead of 'indel').
- *significances*: This column includes any pathogenic metadata about the variant from dbSnp; empty if no metadata about this present in dbSnp. Values include:
    - 'pathogenic-likely-pathogenic', 'drug-response', 'likely-benign', 'likely-pathogenic', 'pathogenic', 'risk-factor', 'not-provided', 'uncertain-significance', 'benign-likely-benign', 'benign', 'conflicting-interpretations-of-pathogenicity'
- *origins:* dbSnp reported origin of variant. Values include:
    - 'unknown', 'maternal', 'somatic', 'germline'
- *diseases*: dbSnp reported diseases associated with this variant.

The columns of ***<gene>_pathogenic_snps.csv*** are a subset of those in ***<gene>_variants.csv.***

The columns of ***<gene>_pathogenic_svs.csv*** are similar to those of ***<gene>_pathogenic_snps.csv***, except *dbsnp_id* is replaced with *dbvar_id* and *variant_type* refers to the dbVar reported variant type.

Usage Instructions

Prerequisite: you must be using a UNIX environment to run this analysis.

1. Create a new directory and download the source code.
2. Ensure python3 is installed and install bedtools
    a. Python3 is likely already installed, but see these docs for any help:
        i. https://docs.python.org/3/using/unix.html
    b. For bedtools, you can either download and build the source as described below, or directly download the bedtools binary here:
        i. https://github.com/arq5x/bedtools2/releases/download/v2.30.0/bedtools.static.binary
        ii. Build from source:

```
wget -cd https://github.com/arq5x/bedtools2/archive/master.zip
unzip master.zip
cd bedtools2-master/
sudo yum -y install gcc-c++
sudo yum -y install zlib-devel
make
```

~~~
```
~~~

3. Install python3 package dependencies:
    a. If pip is not already installed, run:
        i. sudo apt install python3-pip
    b. Run `pip install x` for each of these packages:
        i. pandas
        ii. requests
        iii. biopython
        iv. numpy
        v. matplotlib
4. Change into the `variant_analysis` directory.
5. Download input data files to the 'data' directory:
    a. *gencode.v19.annotation.gff3*
        i. Download and uncompress this file (or a different version) from here:
        ii. http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/
    b. *ALL.wgs.integrated_phase1_v3.20101123.snps_indels_sv.sites.vcf*
        i. Download and uncompress this file (or a different version) from here:
        ii. https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/integrated_call_sets/
6. Mark the bash script as executable:
    ```
    chmod +x run.sh
    ```
7. Before you can run the script, you'll need to know your gene of interest and the chromosome number it is on. You can type in your gene name at UCSC Genome Browser to find its chromosome: http://genome.ucsc.edu/cgi-bin/hgGateway
8. If everything has been installed and the input data is present, you can run the pipeline using the following command:
    ```
    ./run.sh -a data/gencode.v19.annotation.gff3 -g <gene> -c <chromosome_number> -v data/ALL.wgs.integrated_phase1_v3.20101123.snps_indels_sv.sites.vcf
    ```
    Replacing <gene> and <chromosome_number> with your own choices, and remembering to update the data file paths if you downloaded different data to begin with.
9. The script will output "Finished" when it completes; some steps such as fetching results from dbsnp/dbvar can take some time to complete.

Clinical data analysis

Datasources

- *TP53_clinvar_results.txt* - ClinVar data for TP53 acquired by selecting the download link on this URL: https://www.ncbi.nlm.nih.gov/clinvar/?term=tp53%5Bgene%5D

- *gwas_data_2021-07-08.tsv* - from GWAS catalog downloads, v1.0: https://www.ebi.ac.uk/gwas/docs/file-downloads

ClinVar analysis
Given a set of ClinVar results related to the TP53 gene, organizes the results by condition. Output file is a CSV with first column 'condition' and second column is a ';'-delimited list of dbSNP IDs associated with this condition.

Run with:

```
python3 analyze_clinvar.py data/TP53_clinvar_results.txt
data/TP53_clinvar_reported_conditions.csv
```

This script generates a processed CSV file, *TP53_clinvar_reported_conditions.csv*, describing conditions associated with TP53 in ClinVar and the dbSNP IDs of variants associated with those conditions and TP53. The format is:
<condition>,<snp_id_1>;<snp_id_2>;...

GWAS Catalog analysis
Given the entire GWAS dataset, filters the data to just entries where TP53 is one of the reported genes. The data can be downloaded from the link above; the original file is too large to be hosted. Ensure it's in the 'data/' directory before running.

Run with:

```
python3 analyze_gwas.py data/gwas_data_2021-07-08.tsv data/TP53_gwas_data.csv
data/TP53_simplified_gwas_data.csv
```

This script generates two output files, each containing data that pertains to the gene of interest, in this case, TP53.

- *TP53_gwas_data.csv* contains the same columns used in the input file, except rows are filtered such that only rows where TP53 is a related gene are included.
- *TP53_simplified_gwas_data.csv* drops most columns and renames others such that its contents are just the most relevant features and is more similar to output files from the *variant_analysis* pipeline described above.

References

American Cancer Society. (2021, January 12). Lung Cancer Statistics: How Common is Lung Cancer? Retrieved June 21, 2021, from https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html

Buniello, A., Macarthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., . . . Parkinson, H. (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research, 47(D1). doi:10.1093/nar/gky1120

Cancer Treatment Centers of America. (2021, May 21). Types of Lung Cancer: Common, Rare and More Varieties. Retrieved June 21, 2021, from https://www.cancercenter.com/cancer-types/lung-cancer/types

CDC. (2021, June 08). Lung Cancer Statistics. Retrieved June 21, 2021, from https://www.cdc.gov/cancer/lung/statistics/index.htm

Frankish, A., Diekhans, M., Ferreira, A., Johnson, R., Jungreis, I., Loveland, J., . . . Flicek, P. (2018). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research, 47(D1). doi:10.1093/nar/gky955

Harris, C. R., Dewan, A., Zupnick, A., Normart, R., Gabriel, A., Prives, C., . . . Hoh, J. (2009). P53 responsive elements in human retrotransposons. Oncogene, 28(44), 3857-3865. doi:10.1038/onc.2009.246

Kent, W. J. (2002). The Human Genome Browser at UCSC. Genome Research, 12(6), 996-1006. doi:10.1101/gr.229102.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., . . . Maglott, D. R. (2017). ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 46(D1). doi:10.1093/nar/gkx1153

Li, V. D., Li, K. H., & Li, J. T. (2018). TP53 mutations as potential prognostic markers for specific cancers: Analysis of data from The Cancer Genome Atlas and the International Agency for Research on Cancer TP53 Database. *Journal of Cancer Research and Clinical Oncology, 145*(3), 625-636. doi:10.1007/s00432-018-2817-z

Markman, M. (2021, May 21). Types of Lung Cancer: Common, Rare and More Varieties. Retrieved August 9, 2021, from https://www.cancercenter.com/cancer-types/lung-cancer/types

Mayo Clinic. (2021, March 23). Lung cancer. Retrieved from https://www.mayoclinic.org/diseases-conditions/lung-cancer/diagnosis-treatment/drc-20374627

Mogi, A., & Kuwano, H. (2011). TP53 Mutations in Nonsmall Cell Lung Cancer. *Journal of Biomedicine and Biotechnology, 2011*, 1-9. doi:10.1155/2011/583929

Olivier, M., Hollstein, M., & Hainaut, P. (2009). TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology, 2*(1). doi:10.1101/cshperspect.a001008

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841-842. doi:10.1093/bioinformatics/btq033

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., . . . Watson, M. S. (2015). ClinGen — The Clinical Genome Resource. New England Journal of Medicine, 372(23), 2235-2242. doi:10.1056/nejmsr1406261

Roser, M., & Ritchie, H. (2015, July 03). Cancer. Retrieved June 21, 2021, from https://ourworldindata.org/cancer

Stinchcombe, T. E., & Socinski, M. A. (2009). Current Treatments for Advanced Stage Non-Small Cell Lung Cancer. *Proceedings of the American Thoracic Society, 6*(2), 233-241. doi:10.1513/pats.200809-110lc

Wakelee, H. A., Chang, E. T., Gomez, S. L., Keegan, T. H., Feskanich, D., Clarke, C. A., . . . West, D. W. (2007). Lung Cancer Incidence in Never Smokers. *Journal of Clinical Oncology, 25*(5), 472-478. doi:10.1200/jco.2006.07.2983

Walser, T., Cui, X., Yanagawa, J., Lee, J. M., Heinrich, E., Lee, G., . . . Dubinett, S. M. (2008). Smoking and Lung Cancer: The Role of Inflammation. *Proceedings of the American Thoracic Society, 5*(8), 811-815. doi:10.1513/pats.200809-100th