INTRODUCTION

In any population there is variation that occurs through inheritance, and new variation that occurs through mutations. There are a number of different factors that contribute to these mutations such as radiation, tobacco smoke, and UV light, but one factor of particular interest is mobile elements. Mobile elements are DNA sequences that can move around the genome and affect the activity of nearby genes. For my independent study, I will aid the Layer lab by identifying and characterizing mobile elements (looking at deletions) in many different individuals from a number of species. I will also aim to extrapolate any patterns that exist within species regarding mobile elements, and any larger patterns at play. I will be contributing in multiple ways. First, by identifying significant alignments within sub populations of the species I have data on. Secondly, by creating a highly customizable python program and guide that will be stored in the associated project github repository so other researchers can do variant analysis and research on any species they wish. This second contribution is my main focus. Additionally, I will continue to catalog these mobile elements in species and ask questions such as are there traits that are affected more by mobile elements than others in future work.

My timeline was flexible and I have completed the following. First, I obtained the necessary data I needed and identified the tools needed to process it. Those tools included knowledge of different python libraries used to process vcf and fasta data, Blast, AWS, samtools (https://samtools.github.io/hts-specs/VCFv4.2.pdf), and vcf command line tools. Next, I did data exploration to identify what cleaning needs to be done to the data set, and I identified the apis I needed to do analysis on what variants are associated with what sub genuses and populations. After this, I worked on applying these tools to the data to consolidate the amount of data I had to do analysis on. It was very important to do this to resolve the limitations in computing power I had. I also had to make sure I was extracting relevant and useful data so my results were of

high quality. Finally, I worked on finding any patterns that exist in the data regarding structural variants in sub populations for each species.

I referenced a number of different academic papers in order to guide my work. Some reference material I drew ideas from is given below. It is a published paper on a project where researchers worked on finding mobile elements in salmon and how they played a role in the domestication of salmon and salmon breeding selection (Bertolotti et al., 2020). I did a somewhat similar analysis on pigs which can be found later on in the paper. I also referenced this paper for further knowledge on the subject matter (Yiwei et al.,2022).

CONSOLIDATED GUIDE - GITHUB

The github link below contains a consolidated guide written in python on the processes to do effective genomic analysis. Additionally, it contains further resources for this type of analysis and explains how to find if there exist significant alignments in certain populations. I hope to add to it as time permits and as feedback is received to make it more comprehensive.

Repository Web Link: https://github.com/rmrychecky/sv_processes

Repository Cloning Link:

- HTTPS: https://github.com/rmrychecky/sv_processes.git
- SSH: git@github.com:rmrychecky/sv_processes.git

PROCESSES

My first step was to obtain the necessary data to do analysis on. The data I accessed was through the Layer Lab's amazon storage. For each species, I worked with a structural variant file which was in the form of a vcf file and a reference file which was in the form of a fasta file. Over the course of the semester, I applied my processes to 5 species: humans, cows, chickens,

horses, and pigs. Once I obtained these files for each species my first step was to extract actual sequences of deletions across the samples. In order to do this I first had to filter the vcf files to just contain deletions as these were the structural variant types I was interested in. In order to do this I utilized samtools (https://samtools.github.io/hts-specs/VCFv4.2.pdf). A toolset in this library called bcftools allowed me to filter out everything except for variations of type “deletion” with command line instruction:

```
bcftools view -i 'SVTYPE="DEL"' $vcf > $output
```

At this point I was able to read the filtered structural variant and reference files into python. I wrote a function to open and process the vcf file by reading in each record and creating a dictionary whose keys were each chromosome and whose associated values were lists of the pairs of the starting and ending positions of the deletions at that chromosome. Next, I wrote a script to plot all the deletion lengths of the variants. Since I am trying to analyze the similarity of the extracted sequences I plotted a histogram of the lengths of each deletion and found a peak for each species as shown in the table below.

Species	Deletion Length Peak
Human	250-350
Cow	250-350
Chicken	50-150
Horse	200-300

Pig	250-350

Table 1: Deletion Lengths by Species

This helped me reduce the amount of data I had to go on to process, increasing the efficiency of my program. After filtering out the sequences that didn't fit in the desired length range I used the line below in order to obtain the actual sequences and stored them in a dictionary in my python.

```
status, output = subprocess.getstatusoutput(f'samtools faidx
human/hg38.fa {chrom}:{begin}-{end}')
```

I created a dictionary whose keys were seq concatenated with the index number I was on for each associated sequence and whose associated values were the actual sequences I was able to obtain with the line of code above. My program is highly customizable so it can be applied to other species and adjusted easily to work with different ranges of deletion length if need be. Below, I will show the histograms for each of the five species in order to verify I was working with the correct set of data.

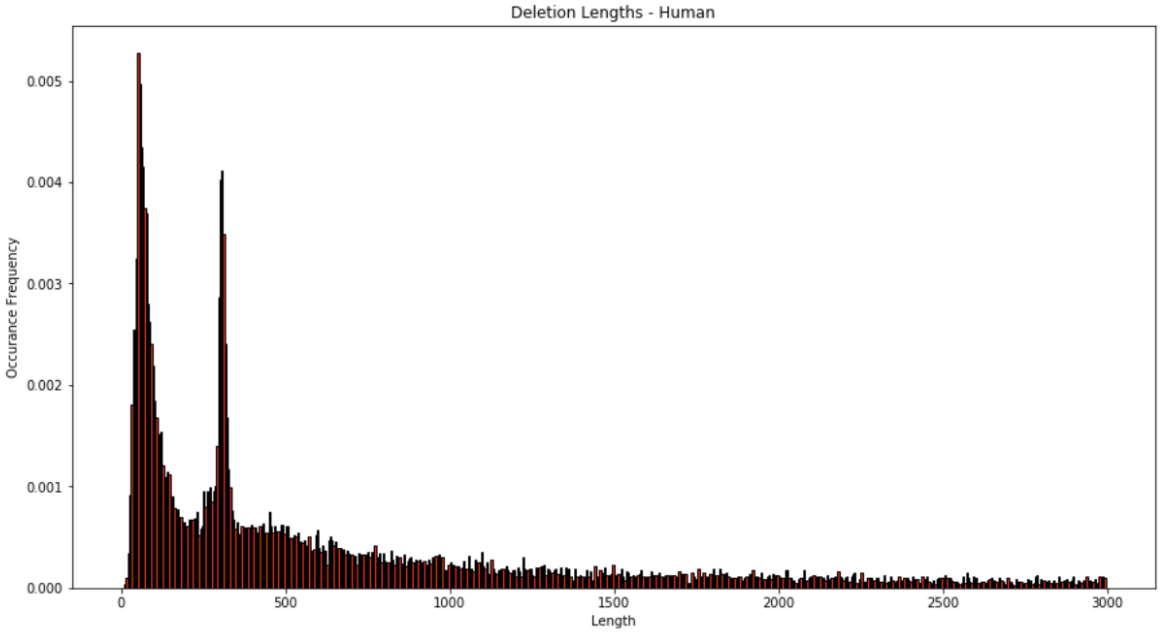

Figure 1: Human Deletion Lengths

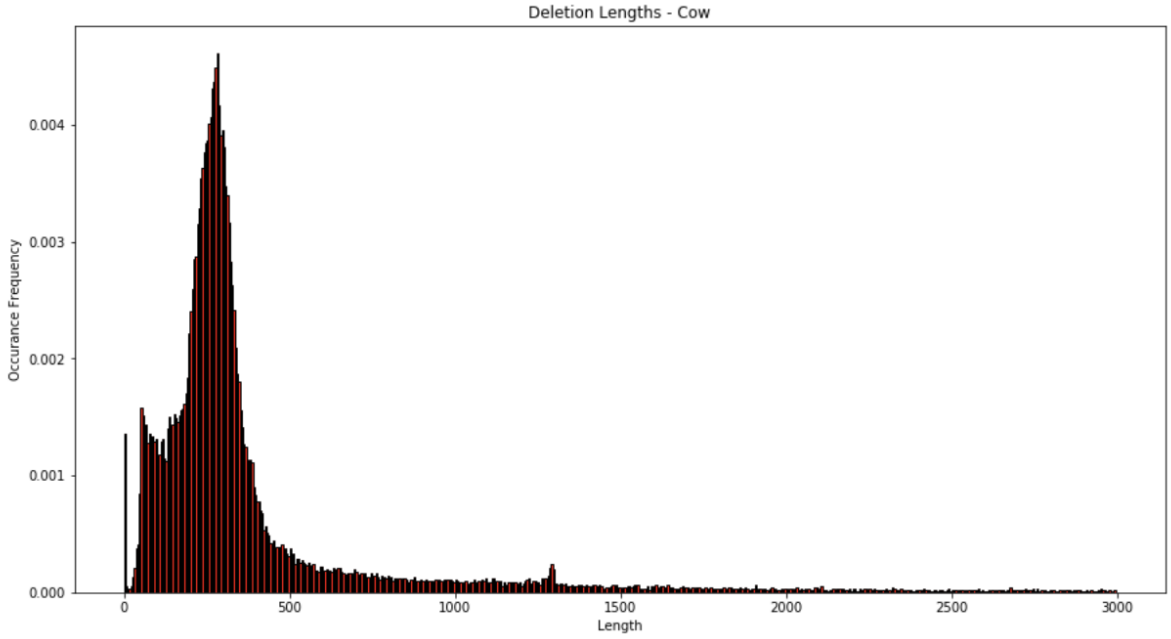

Figure 2: Cow Deletion Lengths

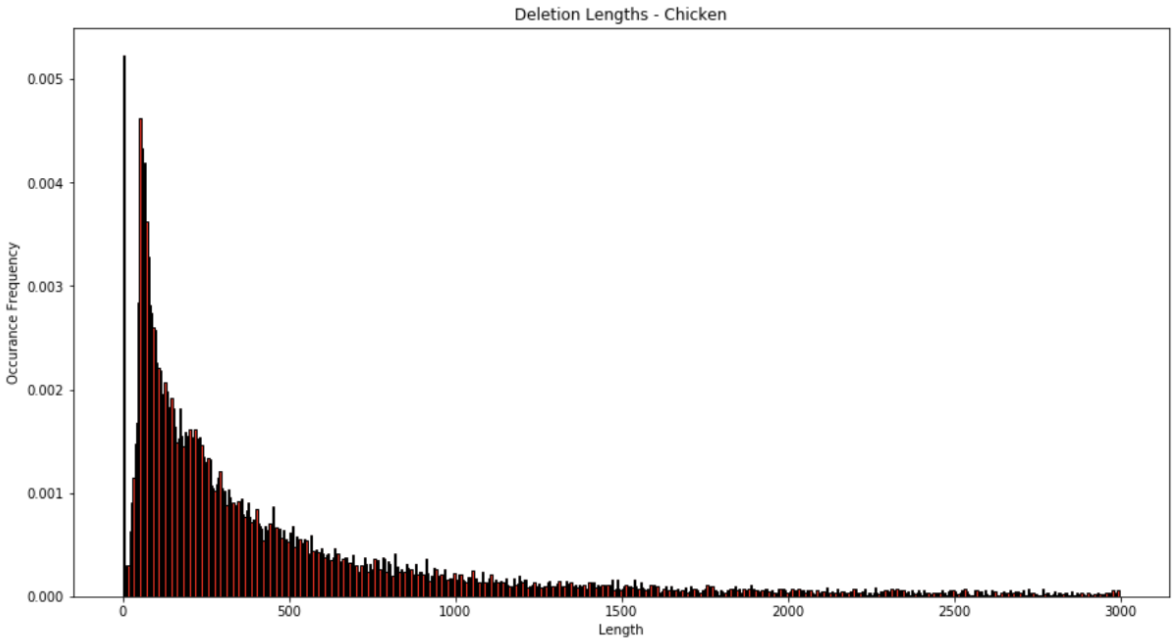

Figure 3: Chicken Deletion Lengths

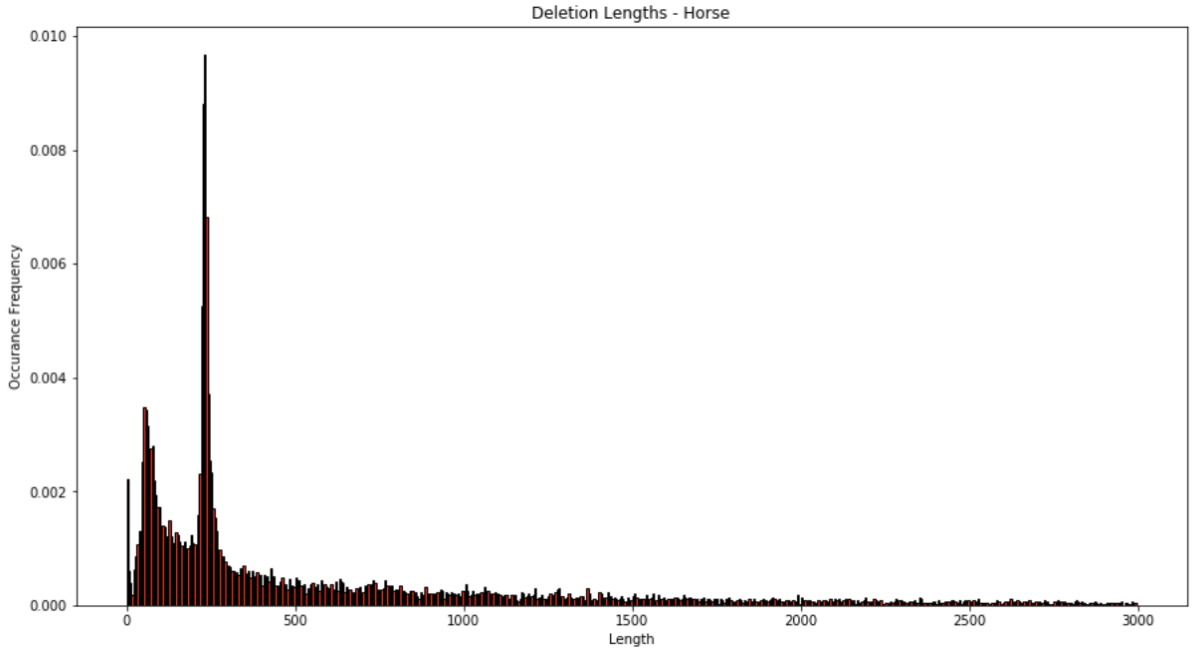

Figure 4: Horse Deletion Lengths

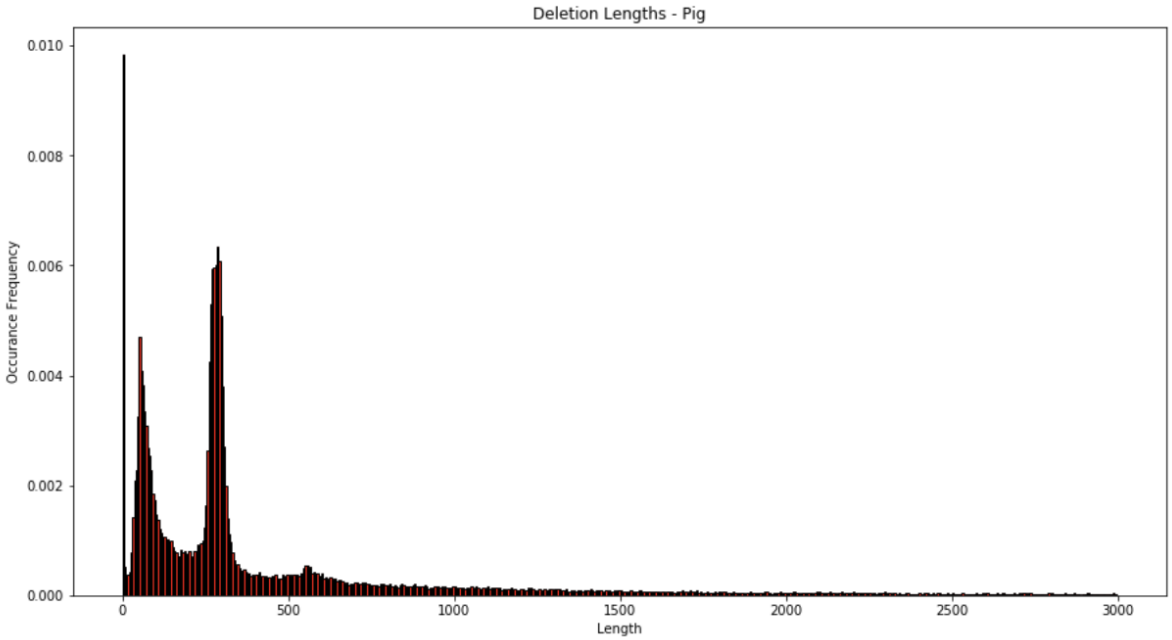

Figure 5: Pig Deletion Lengths

I wrote the appropriate sequences to fasta files for each species. After this, I was then able to begin analysis on sequence alignment.

The next necessary step was to create local blast databases for each fasta file that was saved as described above. This step is crucial in order to provide quick analysis when comparing sequences as the local databases will be indexed appropriately. The command below is what is used in order to create these local databases. Trying to compare sequence similarity without first creating local (or remote if desired) databases reduces efficiency as there are no indices to search on. More on creating local blast databases can be found here: https://www.ncbi.nlm.nih.gov/books/NBK569841/

```
makeblastdb -in seq_of_interest.fsa -parse_seqids -blastdb_version 5
-dbtype nucl -out sequences -title -sequences
```

In order to create these databases, we must assign a unique identifier to each sequence as it allows you to retrieve the sequence by identifier and associate every sequence with a taxonomic node (through the taxid of the sequence). The unique identifiers I used were *seq#* and I incremented the number as I assigned each sequence. According to the Blast documentation, being able to associate a database sequence with a taxonomic node is especially powerful when using the version 5 databases that BLAST uses to limit the search by taxonomy.

My goal is to discover if there exist significant alignments within sub populations of different species. To begin this analysis, I follow two different methods: One for humans and another for other animal species. For humans, I used the thousand genomes data on samples in order to assign their heritage. The data has 9 features as shown in the figure below.

	Sample name	Sex	Biosample ID	Population code	Population name	Superpopulation code	Superpopulation name	Population elastic ID	Data collections
0	HG00271	male	SAME123417	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes 30x on GRC...
1	HG00276	female	SAME123424	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes 30x on GRC...
2	HG00288	female	SAME1839246	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes 30x on GRC...
3	HG00290	male	SAME1839057	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes 30x on GRC...
4	HG00303	male	SAME1840115	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38
...	...	...	...	...	...	...	...	...	...
4973	HGDP00773	female	SAMEA3302906	NaN	Japanese	NaN	East Asia (SGDP)	JapaneseSGDP	Simons Genome Diversity Project
4974	R3	male	SAMEA3302714	NaN	Relli	NaN	South Asia (SGDP)	RelliSGDP	Simons Genome Diversity Project
4975	NA12236	female	SAMEA6604124	CEU	CEPH	EUR	European Ancestry	CEU	1000 Genomes 30x on GRCh38
4976	HGDP00456	male	NaN	NaN	Mbuti	NaN	Africa (HGDP)	MbutiHGDP	HGDP Transcriptome
4977	GM19129	female	NaN	YRI	Yoruba	AFR	African Ancestry	YRI	Human Genome Structural Variation Consortium, ...

Figure 6: Thousand Genomes Data in Pandas Dataframe

I found there to be 6 unique population groups: 'African Ancestry', 'American Ancestry', 'East Asian Ancestry', 'European Ancestry', 'South Asia (SGDP),South Asian Ancestry', and 'South Asian Ancestry'. I then went through all the human variants and if I found it to belong to a population I added it to the list associated with the proper population dictionary key. I could then write each of these sequence groups to fasta files, create local databases, and run the Blastn process to find if there were any significant alignments within populations.

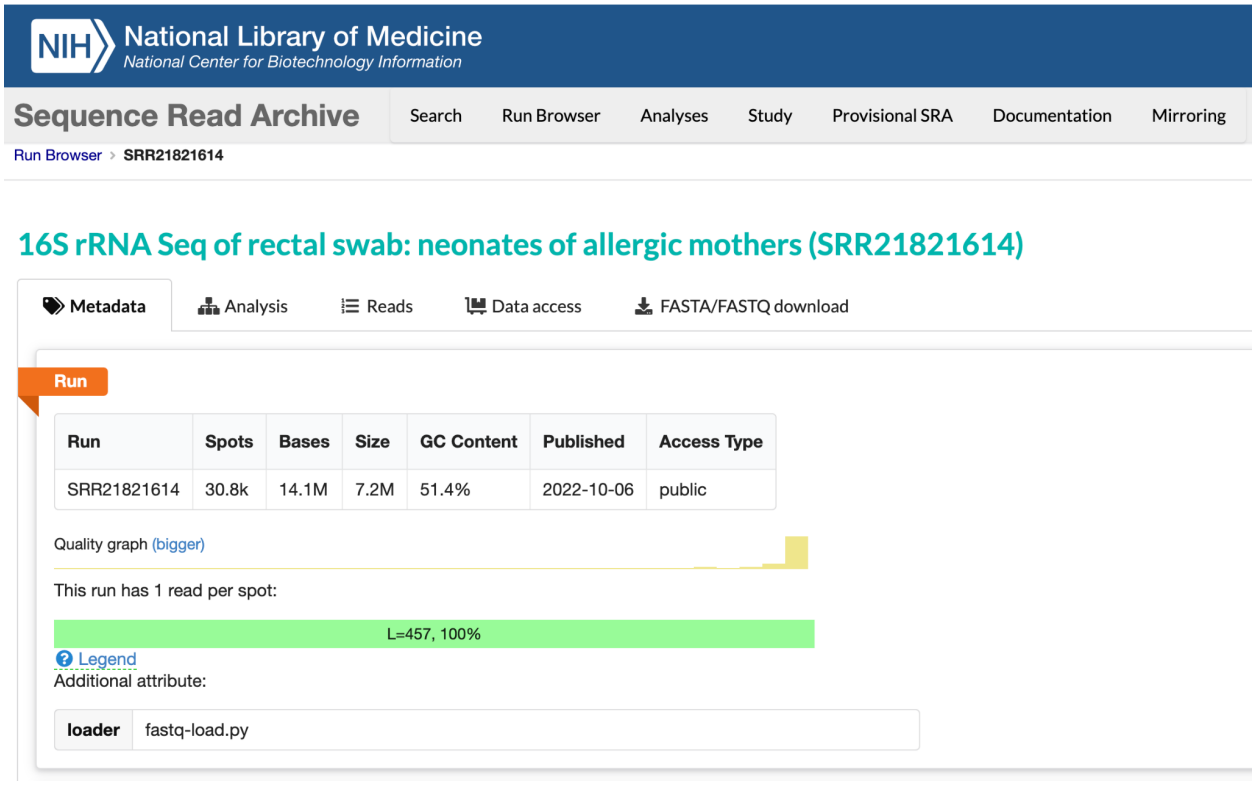

Figure 7: NIH Sequence Read Archive Sample Run

For the other species, I went through the same process of finding all the unique sample names and then used an api connection to query the National Library of Medicine's sequence read archive. The data found here gave me information about the different sub genuses of the upper level species. From there I followed the same process where I went through all the species

variants, and if I found it to belong to a population I added it to the list associated with the proper population dictionary key. I could then write each of these sequence groups to fasta files, create local databases, and run the Blastn process in order to find if there were any significant alignments within populations.

FINDINGS

After all of this work was done, I then proceeded to run the actual analysis on significant alignments within subspecies. After creating the fasta files and subsequent databases as described above, the following command can be used to find significant regions of similarity between sets of sequences.

```
blastn -query seq_of_interest.fsa -subject seq_of_interest.fsa -out
results2.txt
```

If you prefer to analyze the output of the blast command in a different format, such as xml for example, then you can add a *-outfmt* option and specify the output type as defined in the blast documentation.

```
blastn -query seq_of_interest.fsa -subject seq_of_interest.fsa -out
results2.xml -outfmt 5
```

I ran the blast command twice, first saving the output as a txt file, and secondly saving the output as a xml file so I could open, visualize, and parse the output in python. The Blastn application searches a nucleotide query against nucleotide subject sequences or a nucleotide database. The output from the command is written to whatever txt file is specified. In the case above we write to *results2.txt.* It is then important to analyze the contents of the output file. Within the file we are able to see the sequences that produce significant alignments and the E values associated with them. The E values are the expected value, which is the number of

BLAST hits one would expect to see by chance, with the observed score or higher. If we had an E value of 1x10^-44 then you would expect to see that alignment 1x10^-44 times by chance. In other words, it's not random. So, the tinier the E value the more likely the alignment is not random and may be due to a biologically meaningful relationship between the two sequences. We can infer some degree of homology from alignments like this. Another good question is what is a good cutoff in terms of E values in order to say confidently that two sequences are related biologically. There isn't a single answer since E values depend on sequence and database lengths but at NCBI they use a cutoff of 1x10^-6 for some internal processes. A good resource for blast analysis can be found here: https://www.ncbi.nlm.nih.gov/books/NBK569856/ (Bethesda (MD), 2002).

For all the human populations there are a great deal of significant alignments found at good (small) E values. Below are figures containing the E value distributions for each human population. The actual alignment outputs can be found in the github repository which is linked at

the top of the paper.

```
The min e value is 0.0
The max e value is 2.707e-08
```

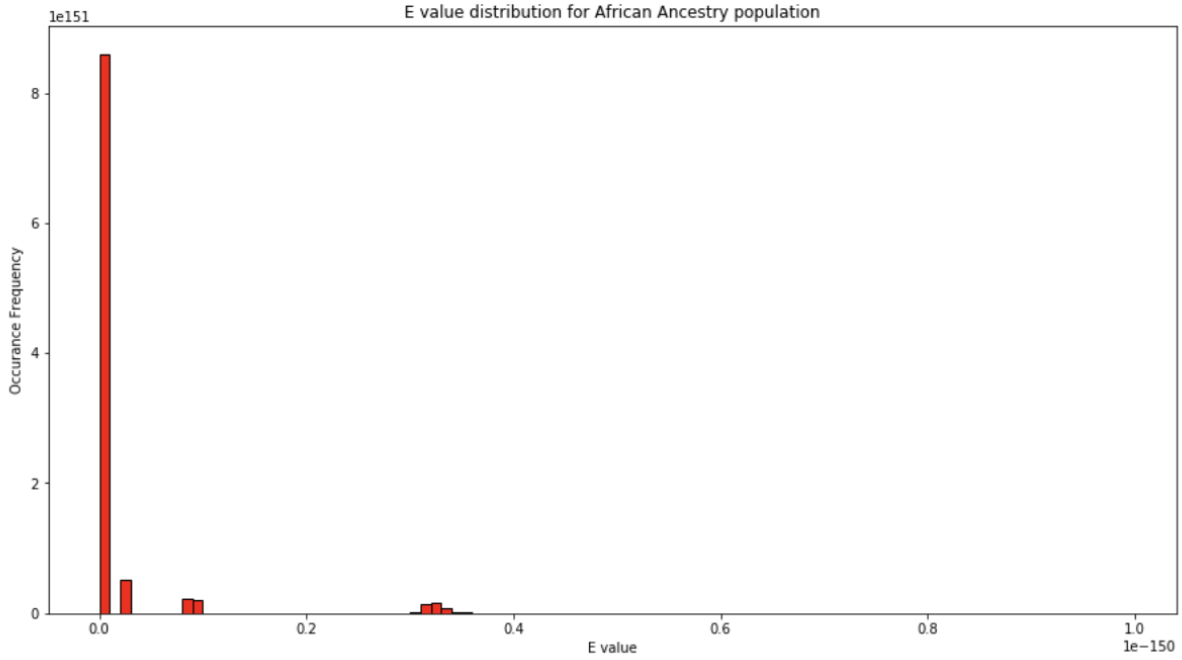

Figure 8: E value distribution for African Ancestry population

```
The min e value is 0.0
The max e value is 2.21024e-08
```

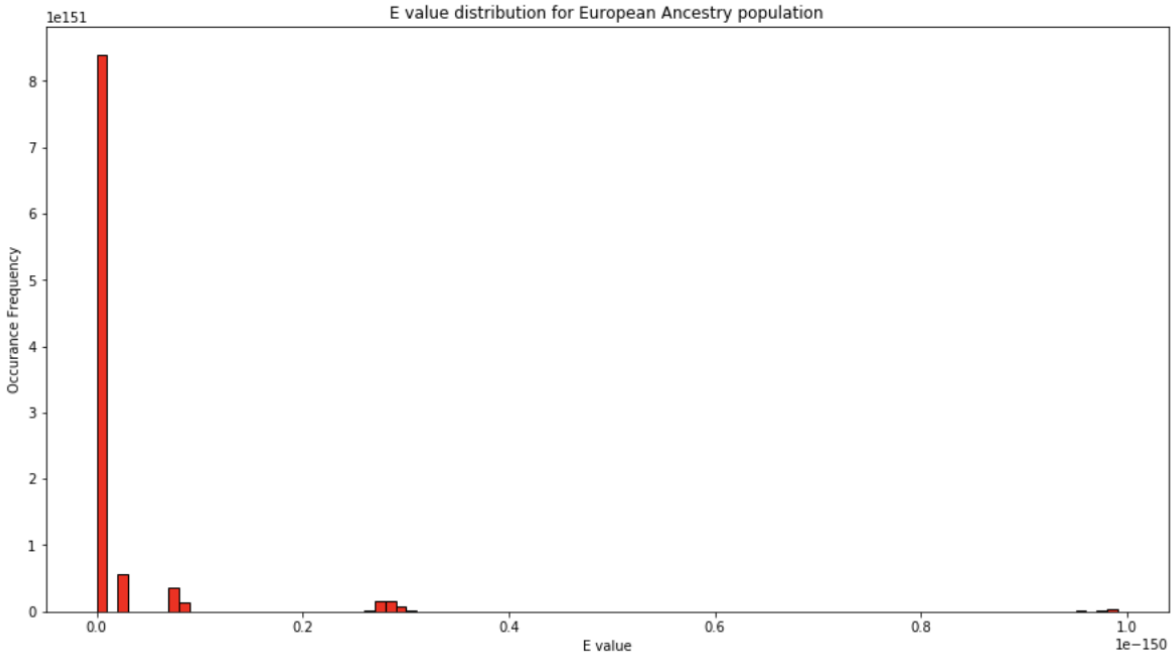

Figure 9: E value distribution for European Ancestry population

```
The min e value is 0.0
The max e value is 2.26406e-08
```

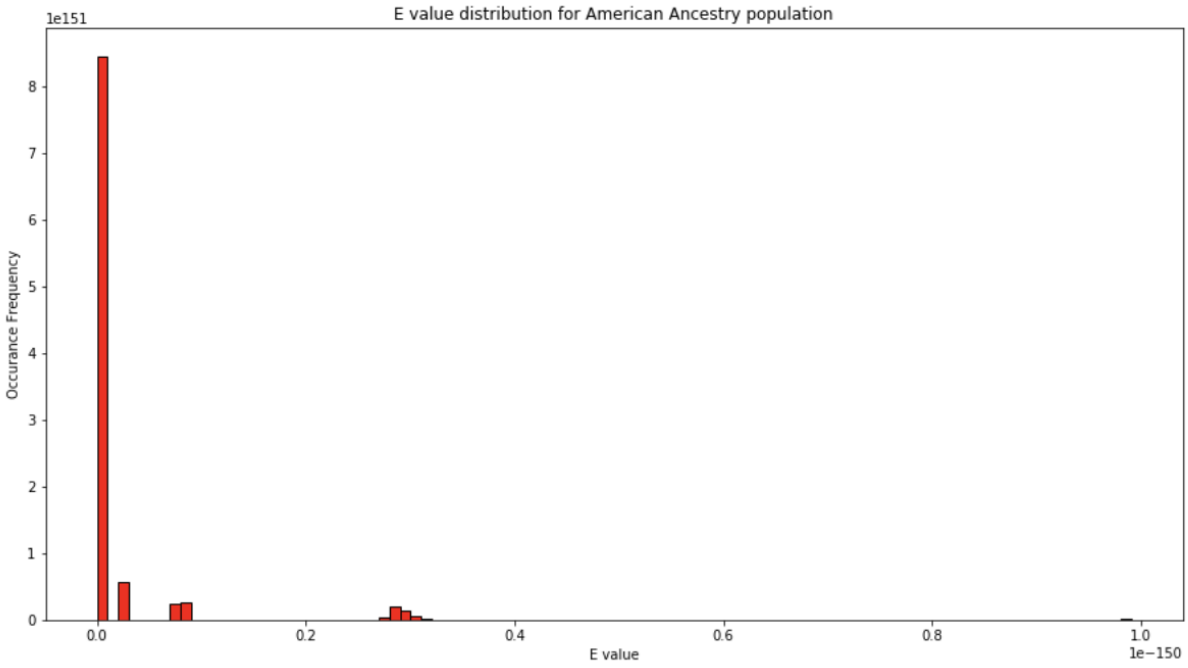

Figure 10: E value distribution for American Ancestry population

```
The min e value is 0.0
The max e value is 2.10949e-08
```

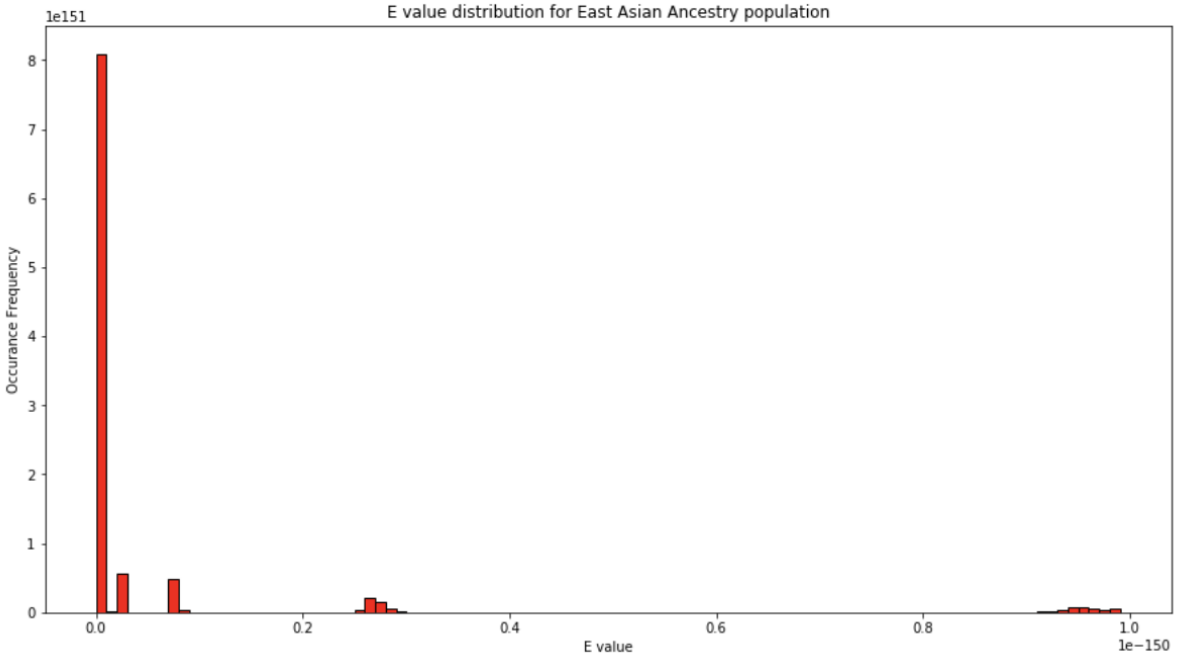

Figure 11: E value distribution for East Asian Ancestry population

```
The min e value is 0.0
The max e value is 2.13438e-08
```

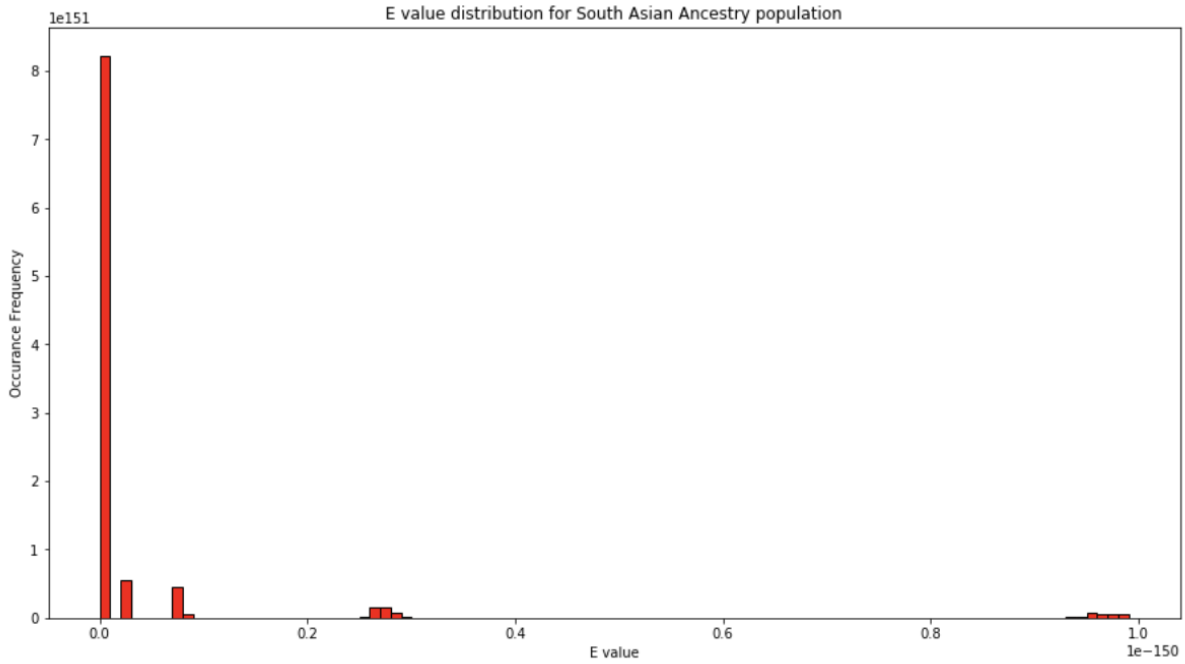

Figure 12: E value distribution for South Asian Ancestry population

```
The min e value is 0.0
The max e value is 1.15784e-08
```

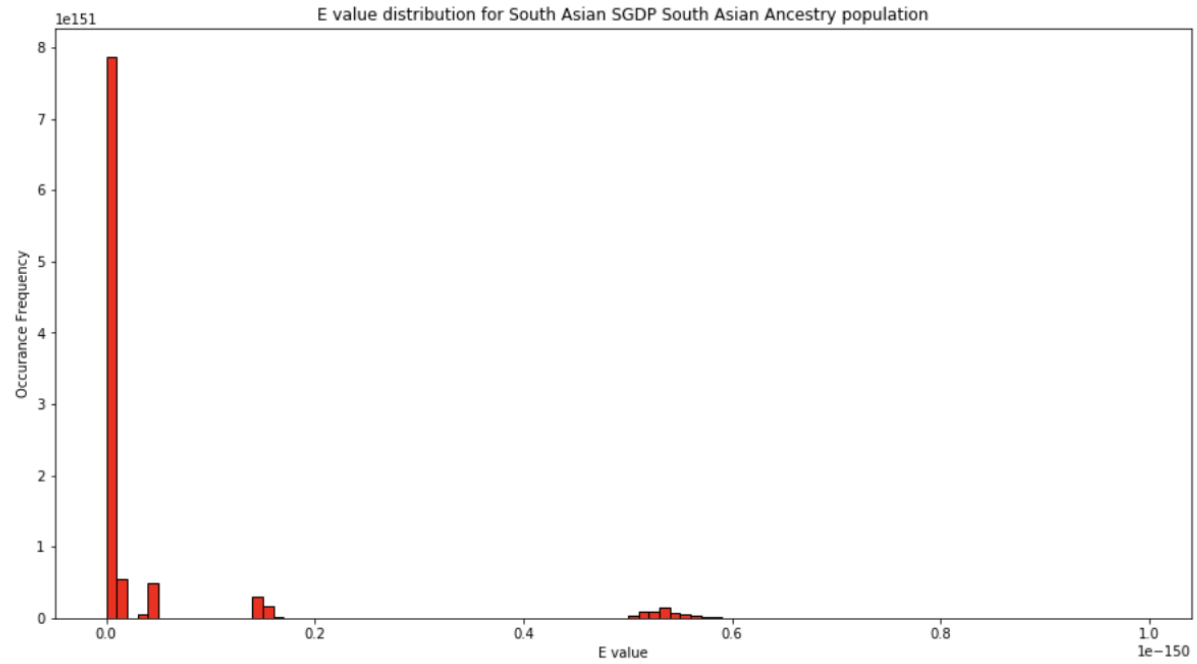

Figure 13: E value distribution for South Asian SGDP South Asian Ancestry population

We can see all the E values are very tiny so it is very likely the alignments are not random and are likely due to a biologically meaningful relationship between the sequences. Additionally these populations share traits which leads me to believe that these variants are affecting the traits displayed in each population to some extent. There is certainly more work to be done here but this is a solid starting point for further exploration.

Additionally, for each sub genus of animal species I did analysis there is a great deal of knowledge to be extracted from this analysis of the significant sequence alignments. For example, I will look at the different horse subspecies. After initial filtering and analysis I found the three unique sub genuses to be “thoroughbred”, “Przewalski horse”, and “P06074M1”. This last group is representative of the Jeju poney which are native horses on Jeju island in the Republic of Korea.

```
: relevent_sub_genuses

: {'SRR515208': 'thoroughbred',
   'SRR515214': 'thoroughbred',
   'SRR516118': 'P06074M1',
   'SRR12719743': 'Przewalski_horse_4'}
```

Figure 14: Relevant horse subgenuses

I then ran Blastn on each of the sub population groups of variants. All of this code is shared on the github for further analysis and understanding. I found similar E values and alignments for

each sub genus with E values averaging to the -155th making it very unlikely that the alignments are random at all. These E value graphs are shown below.

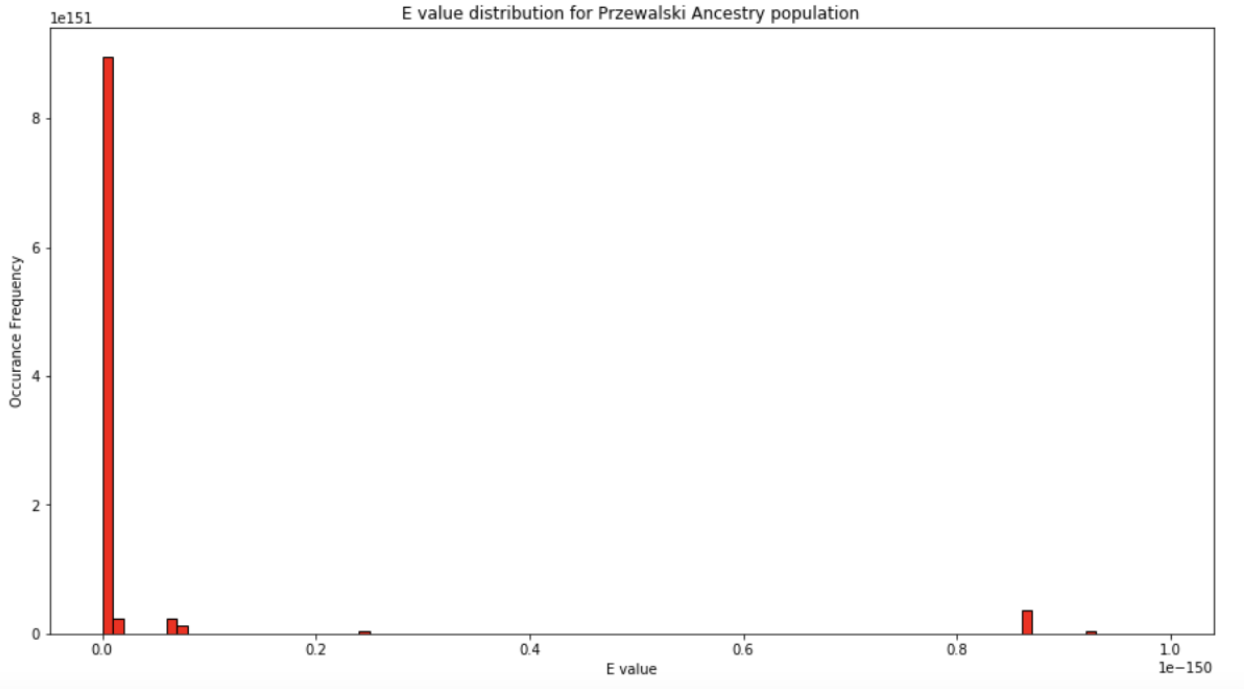

Figure 15: E value distribution for Przewalski Horse Population

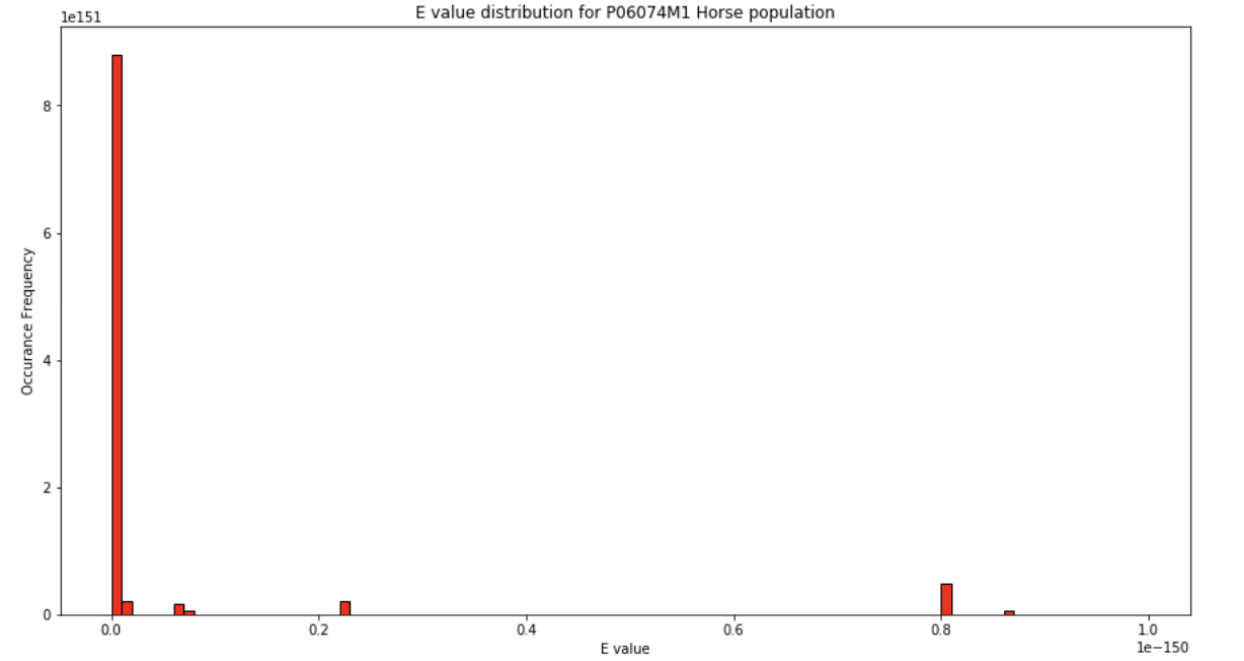

Figure 15: E value distribution for P06074M1 Horse Population

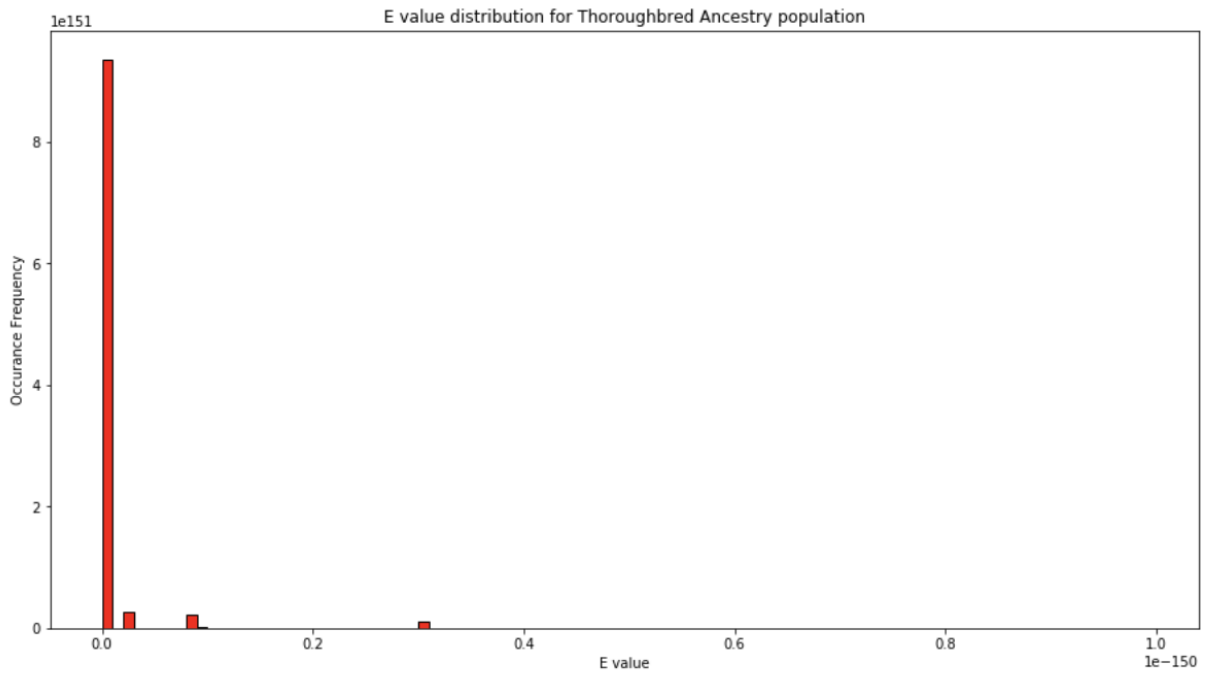

Figure 16: E value distribution for Thoroughbred Horse Population

Each variant that contained either the SRR515208 or SRR515214 SNAME was in the thoroughbred population. Variants with the sample names SRR515208 or SRR515214 have been identified to contribute to enhancing Actin-based mortality in racing horses by the artificial selection of the Ral-GTPases (https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR515208&display=metadata) . So, we see variants that are prevalent in populations of horse species, such as in the thoroughbred sub-gensuses, that cause certain traits to appear across the population.

Additionally, we can look at the Przewalski horse which is a rare, endangered species and the last existing wild horse. There has been a great deal of research that has gone into its genetic characterization as it is important to help facilitate the preservation of its gene pool. There is a very close genetic relationship between Przewalski and Mongolian horses as there is a gene

flow between these two lineages following the split from their common ancestor. There has been debate about the admixture of Przewalskis and domestic horses, and discussion around if they are truly wild horses, or domestic ones. After some studies and analysis on the genomic variants of Przewalski and Mongolian horses it has been found that there are differing abundances of heterozygous and homozygous SNPs in the two types of horse, and the SNPs that were shared between the Mongolian horse and Przewalski horse were lower in number than those that are typically shared within species (Kyoung-Tag et al., 2014). Finally, we can look at the Jeju horse which has been hypothesized to have originated from Mongolian horses similarly to the Przewalski horse as discussed above. These horses have many adaptations due to their harsh local environment. They exhibit traits such as small bodies, stocky heads, and shorter limbs. There has been work to sequence and compare the genomes of Mongolian breeds and the Jeju horse, especially their SNPs and variations/mutations. This analysis has shown that although the Jeju horse is genetically most similar to Mongolian breeds it has genetic ancestry that is independent of Mongolian breeds (Srikanth et al., 2019). We can see that variants are able to help identify biological similarity and that they can also help us identify if some sub populations have common ancestry. In the cases described above, there are significant alignments between variants that occur within subpopulations which follows naturally as this alignment points to biological relationships between variants.

For each subspecies I analyzed, I found significant alignment similarities in each group's variants. I found a great deal of information on many variants. For example, when I looked at the SRR10303810 mutation in chickens, I found that it is involved in comb development and the molecular mechanism of comb formation in chickens. Additionally I did some in depth analysis on structural variants that occur in pigs. The data I was able to work with for pigs contains a lot of samples from Chinese pig species. This is because there has been previous work done in order to improve the pig genome of Chinese local pigs for breeding and genetic resource

conservation reasons. This work has been focused on revealing the full range of structural variations between local Chinese and European pigs. This helps image the existence, or lack thereof, of common ancestry between these breeds of pigs. In the variant data I worked with, there are a great number of different Chinese pig populations that have been bred from the Ningxiang pig, a Chinese indigouse pig breed. Comparisons between this Ningxiang pig and other European pig genomes have since revealed the existence of many structural variants in genes that are involved in the immune system, nervous system, lipid metabolism, and environmental adaptation. More specifically, the genetic variants include 47 Chinese domestic pig-specific structural variants and the associated 74 genes that are likely to contribute to many of the differences that are displayed in domestic traits of Chinese pigs compared to European pigs (Ma et al., 2022). When I ran an analysis on the pig variants I had access to, I found significant alignments between all the variants that occurred within the Ningxiang pigs and the other Chinese native pig breeds of interest. These breeds and their associated variants are found below in Figure 17. Additionally, in Figure 18, I show a sample of significant alignments between variants for Ningxiang and Hechuan pigs.

{'Hechuan': ['SRR13786979', 'SRR13786980', 'SRR13786981', 'SRR13786982', 'SRR13786976', 'SRR13786978', 'SRR13786977'], 'Tiegu': ['SRR13786971', 'SRR13786968', 'SRR13786969', 'SRR13786970', 'SRR13786974', 'SRR13786973', 'SRR13786975'], 'Qinshaohua': ['SRR13786991', 'SRR13786993', 'SRR13786942', 'SRR13786941', 'SRR13786940', 'SRR13786992', 'SRR13786990'], 'Ningxiang': ['SRR13786948', 'SRR13786949', 'SRR13786945', 'SRR13786946', 'SRR13786947', 'SRR13786952', 'SRR13786951'], 'Xiaoer': ['SRR13786984', 'SRR13786986', 'SRR13786987', 'SRR13786985', 'SRR13786989', 'SRR13786988', 'SRR13786944'], 'Shaziling': ['SRR13786958', 'SRR13786959', 'SRR13786954', 'SRR13786957', 'SRR13786953', 'SRR13786956', 'SRR13786955'], 'Daweizi': ['SRR13786964', 'SRR13786965', 'SRR13786963', 'SRR13786967', 'SRR13786962', 'SRR13786966', 'SRR13786960'], 'Debao': ['SRR13786972', 'SRR13786943', 'SRR13786995', 'SRR13786983', 'SRR13786961', 'SRR13786950', 'SRR13786994']}

Figure 17: Chinese native pig breeds and their associated variant names

```
> seq0
Length=287

 Score = 531 bits (287),  Expect = 2e-151
 Identities = 287/287 (100%), Gaps = 0/287 (0%)
 Strand=Plus/Plus

Query  1    CCCCCACAGGTTTCGGGGGAGGAGAACCATCCCCCAGAGTTACTGGCGGGGGAGAACCAT  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    CCCCCACAGGTTTCGGGGGAGGAGAACCATCCCCCAGAGTTACTGGCGGGGGAGAACCAT  60

Query  61   CCCCCAGAGTTACTGGCAGGGGAGAGCCACCCCCACAGGTACTGGCGGGGAGAACCACCC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   CCCCCAGAGTTACTGGCAGGGGAGAGCCACCCCCACAGGTACTGGCGGGGAGAACCACCC  120

Query  121  CCCACAGGCACTGGGTGGGGAGAACCATCCCCCACAGGTACTgggggggggAGAACCACCC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  CCCACAGGCACTGGGTGGGGAGAACCATCCCCCACAGGTACTGGGGGGGGGAGAACCACCC  180

Query  181  CCCACAGGTACTgggggggggAGAACCACCCCCCAGAGTTACTGGCGGGGGAGAGCCACCC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  CCCACAGGTACTGGGGGGGGGAGAACCACCCCCCAGAGTTACTGGCGGGGGAGAGCCACCC  240

Query  241  CCCACAGGCACTgggggggggAGAACCATCCCCTACAGGTACTGGGGG  287
            |||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  CCCACAGGCACTGGGGGGGGGAGAACCATCCCCTACAGGTACTGGGGG  287
```

Figure 18: Sequence Alignments between Ningxiang and Hechuan Pigs

The next step to take would be to do analysis across European pig sequences to identify and verify any common ancestry that exists between these two groups. Since I do not yet have these samples on hand this has been filed as a piece of future work. However, it has already been uncovered through a structural variant analysis study that Chinese fatty pig breeds, such as the Ningxiang, have a fat percentage that is up to 50% while the European Duroc pig has less than 35% body fat. It was found that there exists a 281-bp deletion in the first intron of the MYL4 gene which was found in 0/6 Duroc pigs in the study and 6/8 Ningxiang pigs in the study. This is one of the variants of interest found in the data set that I am working with. On a large scale, it was found in 0/25 European pigs and 18/74 Chinese pigs in the study. The gene is apparently located in the core genome region of a trait locus for back fat at the last rib (Ma et al.,

2022). The knowledge gained from structural variants analysis is very applicable and useful in the real world. In this case, it gives great insight into how to optimize modern pig breeding.

The analysis I did on the populations above can be done for each species sub genus populations. My main focus is building out a python program so individuals can perform their own analysis on species of their choice. As a result, in depth analysis on any one species population exceeds the scope of this paper. Additionally, the number of sub populations varies greatly for each species so it is not reasonable nor practical to do in depth analysis on all of these spices unless they are of interest.

CONCLUSION

Overall this project was a very interesting learning experience and has the potential to grow over time. There is a lot to be learned about how different species are affected by variants and how these variants affect traits that appear in different species. The python program found in the associated project github can be used for in depth analysis and research on structural variants for any species. Depending on the format of the data that one is working with, there may need to be small changes in the functions for proper processing.

FURTHER WORK

There is still further work to be done. The first piece of additional work to do would be to look more in depth at each species sub population and understand more about the variants. Additionally, it would be beneficial to categorize what traits are associated with what variants. Secondly, I would like to look at more species and do similar analysis on a greater variety of species and sub-gensuses within these species. I would also like to continue to develop the generalized analysis python notebook I created so it continues to become more widely applicable to different species and more usable with different file formats. Finally, it would be

helpful to work on answering the question of whether there exist some traits that are affected more severely by mobile elements than others.

RESOURCES

Bertolotti, A.C., Layer, R.M., Gundappa, M.K. et al. The structural variation landscape in 492 Atlantic salmon genomes. Nat Commun 11, 5176 (2020). https://doi.org/10.1038/s41467-020-18972-x

BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Building a BLAST database with your (local) sequences. 2008 Jun 23 [Updated 2021 Jan 7]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK569841/

Kyoung-Tag Do, Hong-Sik Kong, Joon-Ho Lee, Hak-Kyo Lee, Byung-Wook Cho, Heui-Soo Kim, Kung Ahn, Kyung-Do Park, Genomic characterization of the Przewalski's horse inhabiting Mongolian steppe by whole genome re-sequencing, Livestock Science, Volume 167, 2014, Pages 86-91, ISSN 1871-1413, https://doi.org/10.1016/j.livsci.2014.06.020.

Ma, H., Jiang, J., He, J. et al. Long-read assembly of the Chinese indigenous Ningxiang pig genome and identification of genetic variations in fat metabolism among different breeds. Molecular Ecology Resources, Volume 22, Issue 4, Mary 2022, Pages 1508-1520, https://doi.org/10.1111/1755-0998.13550

National Library of Medicine. (2022). https://www.ncbi.nlm.nih.gov/

Srikanth, Krishnamoorthy & Kim, Nam-Young & Park, WonCheoul & Kim, Jae-Min & Kim, Kwondo & Lee, Kyung-Tai & Son, Ju-Hwan & Chai, Han-Ha & Choi, Jung-Woo & Jang, Gul-Won & Kim, Heebal & Ryu, Youn-Chul & Nam, Jin-Wu & Park, Jong-eun & Kim, Jun-Mo & Lim, Dajeong. (2019). Comprehensive genome and transcriptome analyses reveal genetic relationship, selection signature, and

transcriptome landscape of small-sized Korean native Jeju horse. Scientific Reports. 9. 10.1038/s41598-019-53102-8.

Yiwei Niu, Xueyi Teng, Honghong Zhou, Yirong Shi, Yanyan Li, Yiheng Tang, Peng Zhang, Huaxia Luo, Quan Kang, Tao Xu, Shunmin He, Characterizing mobile element insertions in 5675 genomes, Nucleic Acids Research, Volume 50, Issue 5, 21 March 2022, Pages 2493–2508, https://doi.org/10.1093/nar/gkac128