Ethan O’Keefe
Gen 711
5/22/2022
Gen 711 Final Project Update

Identifying Genetic Variance Linkage to Increased Alcohol Consumption in Wistar Rats - Updated Analysis

Continuing the Project

Since the conclusion of the project on its initial due date (5/13/2022), work on obtaining proper variant calls and analysis continued in close collaboration with Professor Miller. The most consequential roadblock in the initial project pipeline was the size of the data and my computer's inability to process it correctly or entirely. This was solved by using RON, one of the University of New Hampshire's remote computer clusters to process the large scale data. Once the data was processed, further analysis could continue. All citing conventions introduced in the initial project report will continue to be used (command[ln#], process[tr#])

Variant Calling

This process was backed up a few steps from its progress in the initial project submission to include more samples. Bcftools mpileup was used to assemble the sorted .bam files which were then piped into bcftools call command to create a .bcf file containing SNP calls for 4 assembled sequences[ln242-243].

```
Bcftools mpileup -Ou -f ref.fasta input1.bam input2.bam… | bcftools call -mv -Ob -o output.bcf
```

The output .bcf calls were then converted[ln232] to vcf format so they could be read by vcftools for further analysis. The bctools convert command was utilized for this step in tandem with the -O v option to specify a .vcf output.

```
Bcftools convert -O v -o output.vcf input.bcf
```

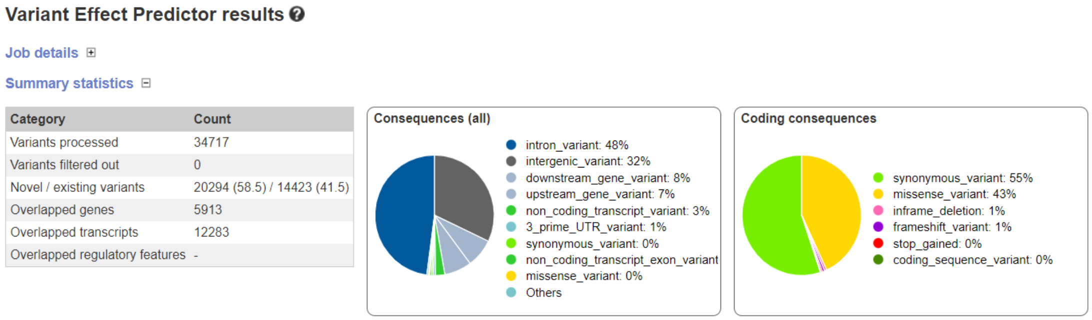
The resulting file was a text based .vcf file containing all SNP calls for the 4 aligned Wistar sequences.

Variant Effect Prediction

In any given sequence, when compared in reference to a different strain, there are bound to be an unusably large number of variants between the two sequences. This was no different in my own analysis. The .vcf calls output from the previous step contained far too many variants for any selective set to be attributed to a phenotypical propensity to consume. As such a further specification was needed. Vcftools[12] was installed with conda and was used to restrict present calls to only those which were consistent amongst all 4 samples and were different from the reference[ln236]. This greatly reduced the number of calls by reporting only those conserved enough to occur in all examined samples.

```
Vcftools --gzvcf input.vcf.gz  --non-ref-af 0.99 --max-missing 1.0 --remove-filtered-all --recode
--stdout | gzip -c > output.vcf.gz
```

The output file was compressed into a .vcf.gz format to limit excessive file sizes that might be created during tinkering of command options. The filtered file was finally uploaded to Ensembl's 'Variant Effect Predictor'[13]. The VEP program analyzes the impact any given variant is likely to have on a gene within the organism. The program reported variants mostly occurred synonymously (55%) and as missense variants (43%, Figure 5). The program outputs a text file containing all variants and they're estimated impact which can be found in this github under the VEP folder. The overwhelming majority of calls were low or 'modifying' impact, which occur in non-coding areas of the sequence. There were several high impact variants due to framshifts or splice acceptor variants, high impact calls can be accessed in the same VEP folder on github.



Category	Count
Variants processed	34717
Variants filtered out	0
Novel / existing variants	20294 (58.5) / 14423 (41.5)
Overlapped genes	5913
Overlapped transcripts	12283
Overlapped regulatory features	-

Figure 5. VEP analysis on filtered variant calls

Analysis

The goal of this project was to identify genetic variation between alcohol preferring and non-preferring rats strains that could be attributed to an increased propensity to consume alcohol. The final generated data, reports conserved and significant variation between the alcohol preferring Wistar rat, and non-alcohol preferring Brown rat (Norway rat, Bn7.2), and the degree of its consequence on the genome. Unfortunately, without an association study, there is no way to identify which variants are directly responsible for the increased propensity, although the performed analysis narrows the list down to a select few variants. The phenotypical increase in consumption propensity is not caused by a single gene but likely a combination of genes and subsequent traits, as such its difficult to determine which variants were more directly responsible for the change. If one assumes consumption propensity was the only trait that was bred for when developing the Wistar rat strain, it's an easy conclusion to suggest the summative effect of all identified variants are responsible for the increase. This is unlikely however as the breeding could not possibly have occured on a large enough scale to isolate such a complex trait, and so additional, unintentional traits were likely bred into the Wistar rats as well. This reiterates the problem with correlating gene variants to physical traits without a further association study. Due to this, I will report the genes most closely associated with alcohol abuse and the significant variation within them, as this is as specific a correlation I can derive without making excessive assumptions.

Figures 6,7,8 and 9 detail an IGV visualization of the ALDH2, FTO, SLC39A8 and VRK2 genes respectively and their sequence variance. It's quite noticeable that none of these variations occur within an exon of their corresponding gene. This could be used to argue for the incomplete nature of the analysis, or opposingly, the lack of involvement in said genes to the increased consumption propensity. Of the 7 genes of interest detailed in the initial report, 5 had variants within the gene, and none had variants within an exon. This not to say these genes are not involved in alcohol consumption, as there is a case to be made for the small sample size used in this project, but that the phenotypical variation in Wistar rats specifically, is likely derived from variants in a different set of genes. This is an interesting conclusion to arrive at, as heritable vulnerability to alcohol abuse is a well known phenomenon, and the genes of interest selected for this project were chosen based on their presence in association studies. These findings could suggest selectively bred alcohol propensity, and naturally bred alcohol propensity are affected by different sets of genes. This is a conclusion that is certainly worth further investigation. Should the opportunity arise, I would like the chance to run an association study on these linkages, any further updates will be uploaded to this github.

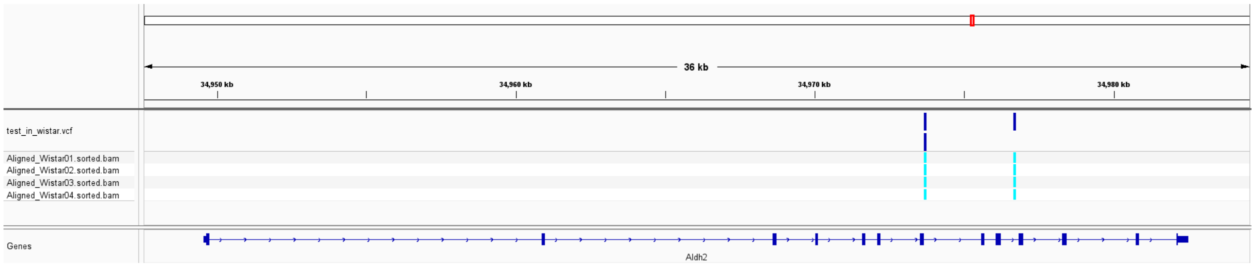

36 kb
34,950 kb
34,960 kb
34,970 kb
34,980 kb
test_in_wistar.vcf
Aligned_Wistar01.sorted.bam
Aligned_Wistar02.sorted.bam
Aligned_Wistar03.sorted.bam
Aligned_Wistar04.sorted.bam
Genes
Aldh2

Figure 6. IGV visualization of ALDH2 variants

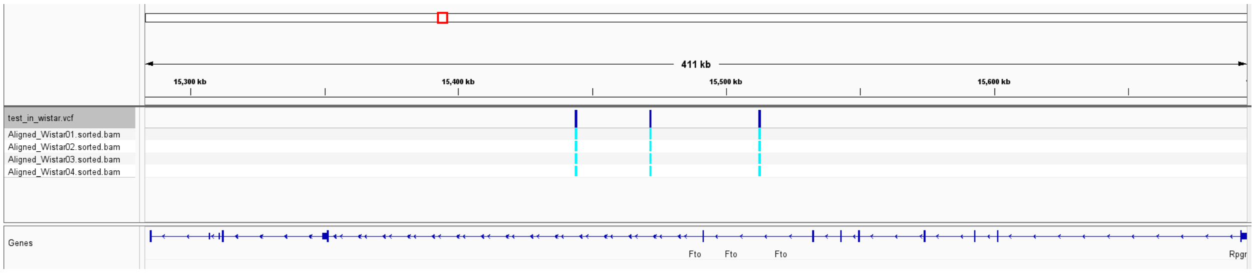

411 kb
15,300 kb
15,400 kb
15,500 kb
15,600 kb
test_in_wistar.vcf
Aligned_Wistar01.sorted.bam
Aligned_Wistar02.sorted.bam
Aligned_Wistar03.sorted.bam
Aligned_Wistar04.sorted.bam
Genes
Fto
Fto
Fto
Rpgr

Figure 7. IGV visualization of FTO variants

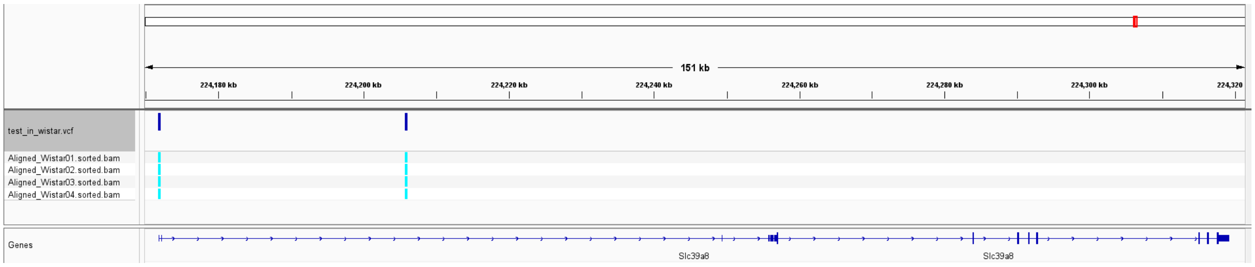


Figure 8. IGV visualization of SLC39A8 variants

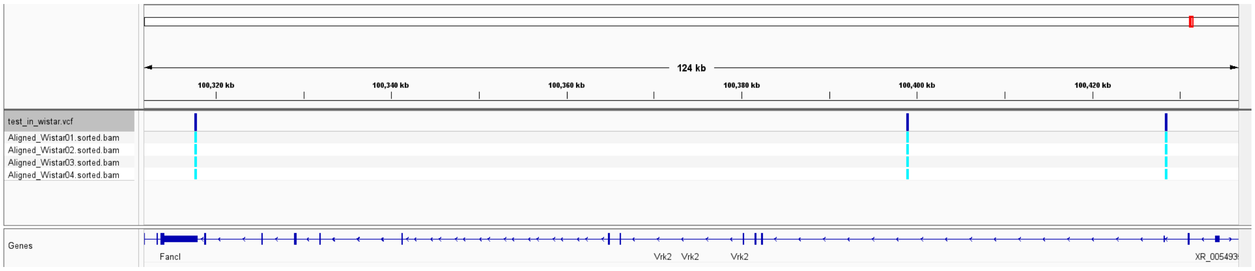


Figure 9. IGV visualization of VRK2 variants

Citations continued

12. “VCFtools Manual.” *VCF Manual*, http://vcftools.sourceforge.net/man_latest.html.

13. *Ensembl Variant Effect Predictor (VEP)*, https://useast.ensembl.org/info/docs/tools/vep/index.html.

Appendix

- All referenced images can be found in the ‘Appendix images” folder on this project's Github.