Todd Gavin
Dr. Peter Calabrese
QBIO401 – Quantitative Biology
12 December 2022

Final Project Report

Analyzing the statistical differences and frequency of the genotypes of cancer-risk SNPs between populations.

Introduction

Cancer rates, prevalence, and outcomes vary among different population groups in the United States. Researchers have identified genetic variations that are specific to certain ancestries and may contribute to these disparities. In the human genome, there are SNPs that increase the likelihood of developing risk to certain types of cancers. In this project I have studied three of them: rs72699833 found on chromosome 1, rs4713266 found on chromosome 6, and rs6983267 found on chromosome 8.

1. Research Questions
    a. Are the genotypic frequencies of certain cancer-risk SNPs statistically different in certain ethnic groups compared to others?
    b. Is there a linear correlation between the chi-square values and ALT allele frequency of many different SNPs on a single chromosome?
    c. Do the chi-square values of many different SNPs on a single chromosome follow a chi-square distribution?
2. Hypotheses'
    a. I hypothesize that the genotypic frequencies of certain cancer-risk SNPs statistically different in certain ethnic groups compared to others. As for what these ratios will be, we will have to refer to the analysis.

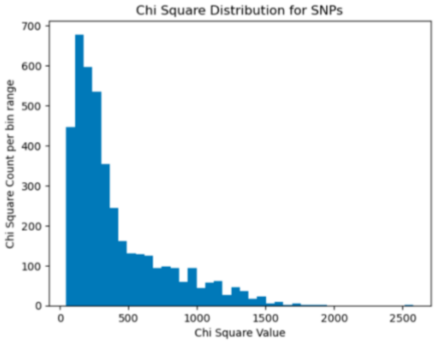
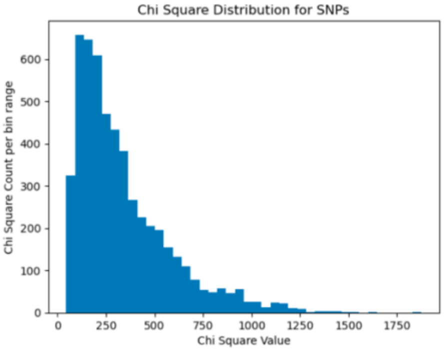
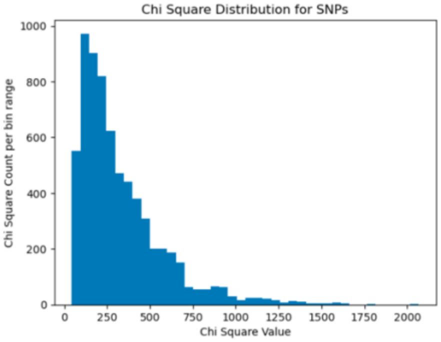
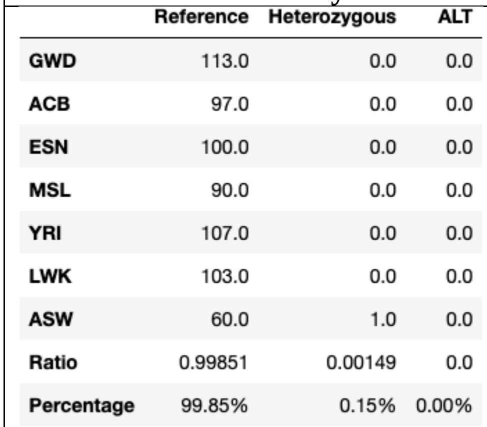
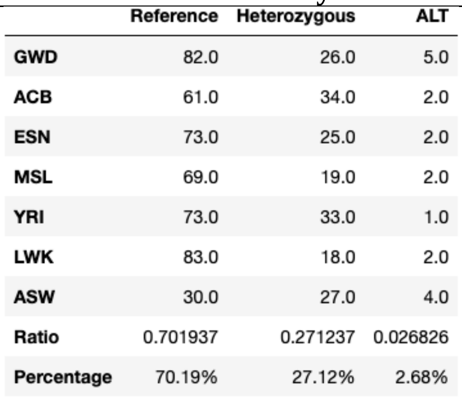
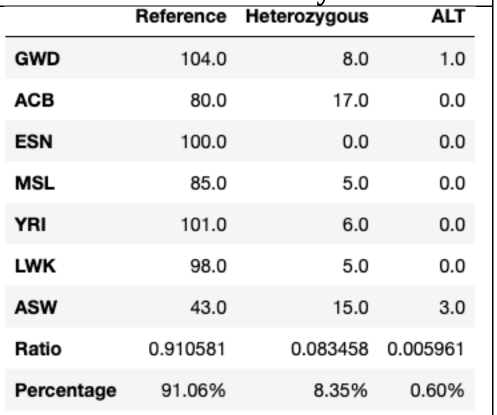
b. For the other two research questions, I do not have a hypothesis, I will let the analysis show the results.

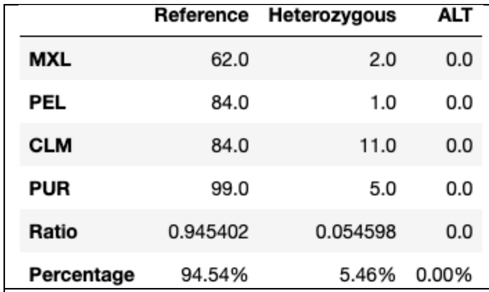
As we have seen from the data, there seems to be a statistically significant association between an individual's population (their origin of ethnicity) and the probability that they have a certain genotype that increases their likelihood for developing a certain type of cancer. SNP rs72699833 is found on chromosome 1 and is linked to PHGDH in cis. PHGDH is a gene involved in the metabolism of serine, and its overexpression has been observed in certain subtypes of breast, cervical, colorectal, and non-small-cell lung cancer. In these diseases, overexpression of PHGDH is generally associated with a worse outcome (Fagny et al., 2019). In a recent issue of Cancer Research, Han and colleagues discovered that SNP rs4713266 is associated with an increased risk for developing prostate cancer. The study also found that this SNP alters the activity of a NEDD9 enhancer, leading to increased NEDD9 expression. This research provides both epidemiological and mechanistic insight into the factors that may cause disparities in prostate cancer (Mavura et al., 2021). The inherited variant on chromosome 8q24, rs6983267, is linked to the development of colorectal cancer. Evidence from the study Pomerantz et al. states that this region acts as a transcriptional enhancer and physically interacts with the MYC proto-oncogene. The rs6983267 alleles also bind to transcription factor 7-like 2 (TCF7L2) differently. Their findings provide strong support for a biological mechanism behind this non-protein coding risk variant.

Methods

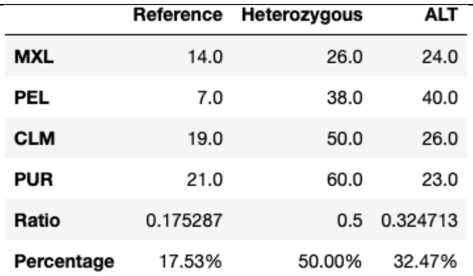
For my final project I used python notebooks to collect, organize, aggregate, and analyze SNPs data from the 1000 Genome Project from the UC-Santa Cruz open research library.

Data

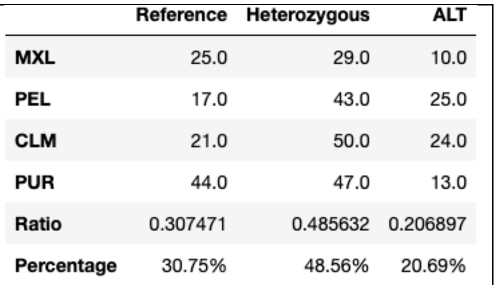
<table>
<tr><th>SNP #1: rs72699833-chr1</th><th>SNP #2: rs4713266-chr6</th><th>SNP #3: rs6983267-chr8</th></tr>
<tr><td><b>Percentage of noncomputable SNPs</b>: 80.90%</td><td><b>Percentage of noncomputable SNPs</b>: 81.50%</td><td><b>Percentage of noncomputable SNPs</b>: 80.11%</td></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
<tr><td><b>99th Percentile:</b> 1461.24
<b>95th Percentile:</b> 1151.31
<b>Median</b>: 277.83
<b>Average</b>: 403.88</td><td><b>99th Percentile:</b> 1147.69
<b>95th Percentile</b>: 846.54
<b>Median</b>: 266.81
<b>Average</b>: 336.86</td><td><b>99th Percentile</b>: 1239.06
<b>95th Percentile</b>: 838.71
<b>Median</b>: 254.19
<b>Average</b>: 328.18</td></tr>
<tr><td><b>SNP File Path:</b> TargetSNPsData/rs72699833-chr1.csv
<b>Chi2</b>: 302.88
<b>P value:</b> 6.87e-38</td><td><b>SNP File Path</b>: TargetSNPsData/rs4713266-chr6.csv
<b>Chi2</b>: 1016.58
<b>P value</b>: 2.66e-180</td><td><b>SNP File Path</b>: TargetSNPsData/rs6983267-chr8.csv
<b>Chi2</b>: 1091.82
<b>P value</b>: 6.77e-196</td></tr>
<tr><td><b>AFR - African Ancestry</b></td><td><b>AFR - African Ancestry</b></td><td><b>AFR - African Ancestry</b></td></tr>
<tr><td>
<table>
<tr><th></th><th>Reference</th><th>Heterozygous</th><th>ALT</th></tr>
<tr><td>GWD</td><td>113.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td>ACB</td><td>97.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td>ESN</td><td>100.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td>MSL</td><td>90.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td>YRI</td><td>107.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td>LWK</td><td>103.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td>ASW</td><td>60.0</td><td>1.0</td><td>0.0</td></tr>
<tr><td>Ratio</td><td>0.99851</td><td>0.00149</td><td>0.0</td></tr>
<tr><td>Percentage</td><td>99.85%</td><td>0.15%</td><td>0.00%</td></tr>
</table></td><td>
<table>
<tr><th></th><th>Reference</th><th>Heterozygous</th><th>ALT</th></tr>
<tr><td>GWD</td><td>82.0</td><td>26.0</td><td>5.0</td></tr>
<tr><td>ACB</td><td>61.0</td><td>34.0</td><td>2.0</td></tr>
<tr><td>ESN</td><td>73.0</td><td>25.0</td><td>2.0</td></tr>
<tr><td>MSL</td><td>69.0</td><td>19.0</td><td>2.0</td></tr>
<tr><td>YRI</td><td>73.0</td><td>33.0</td><td>1.0</td></tr>
<tr><td>LWK</td><td>83.0</td><td>18.0</td><td>2.0</td></tr>
<tr><td>ASW</td><td>30.0</td><td>27.0</td><td>4.0</td></tr>
<tr><td>Ratio</td><td>0.701937</td><td>0.271237</td><td>0.026826</td></tr>
<tr><td>Percentage</td><td>70.19%</td><td>27.12%</td><td>2.68%</td></tr>
</table></td><td>
<table>
<tr><th></th><th>Reference</th><th>Heterozygous</th><th>ALT</th></tr>
<tr><td>GWD</td><td>104.0</td><td>8.0</td><td>1.0</td></tr>
<tr><td>ACB</td><td>80.0</td><td>17.0</td><td>0.0</td></tr>
<tr><td>ESN</td><td>100.0</td><td>0.0</td><td>0.0</td></tr>
<tr><td>MSL</td><td>85.0</td><td>5.0</td><td>0.0</td></tr>
<tr><td>YRI</td><td>101.0</td><td>6.0</td><td>0.0</td></tr>
<tr><td>LWK</td><td>98.0</td><td>5.0</td><td>0.0</td></tr>
<tr><td>ASW</td><td>43.0</td><td>15.0</td><td>3.0</td></tr>
<tr><td>Ratio</td><td>0.910581</td><td>0.083458</td><td>0.005961</td></tr>
<tr><td>Percentage</td><td>91.06%</td><td>8.35%</td><td>0.60%</td></tr>
</table></td></tr>
<tr><td><b>AMR - American Ancestry</b></td><td><b>AMR - American Ancestry</b></td><td><b>AMR - American Ancestry</b></td></tr>
</table>



	Reference	Heterozygous	ALT
MXL	62.0	2.0	0.0
PEL	84.0	1.0	0.0
CLM	84.0	11.0	0.0
PUR	99.0	5.0	0.0
Ratio	0.945402	0.054598	0.0
Percentage	94.54%	5.46%	0.00%

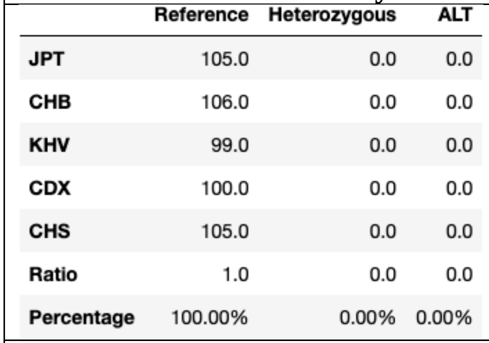


	Reference	Heterozygous	ALT
MXL	14.0	26.0	24.0
PEL	7.0	38.0	40.0
CLM	19.0	50.0	26.0
PUR	21.0	60.0	23.0
Ratio	0.175287	0.5	0.324713
Percentage	17.53%	50.00%	32.47%



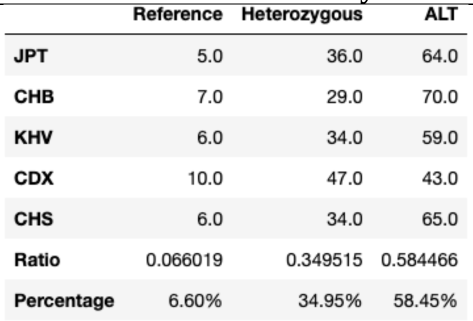
	Reference	Heterozygous	ALT
MXL	25.0	29.0	10.0
PEL	17.0	43.0	25.0
CLM	21.0	50.0	24.0
PUR	44.0	47.0	13.0
Ratio	0.307471	0.485632	0.206897
Percentage	30.75%	48.56%	20.69%

EAS – East Asian Ancestry



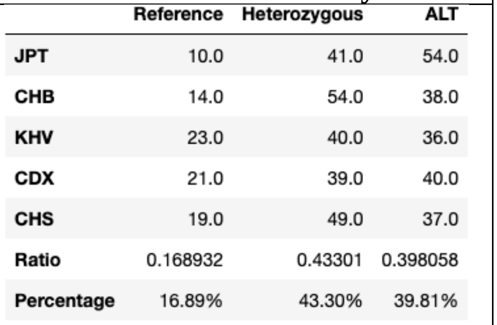
	Reference	Heterozygous	ALT
JPT	105.0	0.0	0.0
CHB	106.0	0.0	0.0
KHV	99.0	0.0	0.0
CDX	100.0	0.0	0.0
CHS	105.0	0.0	0.0
Ratio	1.0	0.0	0.0
Percentage	100.00%	0.00%	0.00%

EAS – East Asian Ancestry



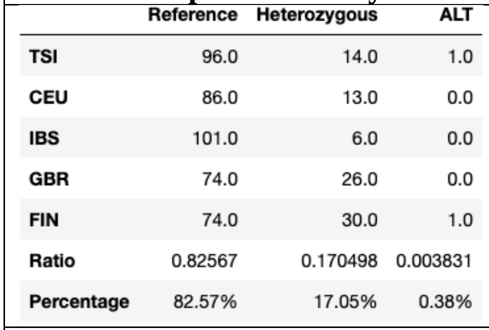
	Reference	Heterozygous	ALT
JPT	5.0	36.0	64.0
CHB	7.0	29.0	70.0
KHV	6.0	34.0	59.0
CDX	10.0	47.0	43.0
CHS	6.0	34.0	65.0
Ratio	0.066019	0.349515	0.584466
Percentage	6.60%	34.95%	58.45%

EAS – East Asian Ancestry



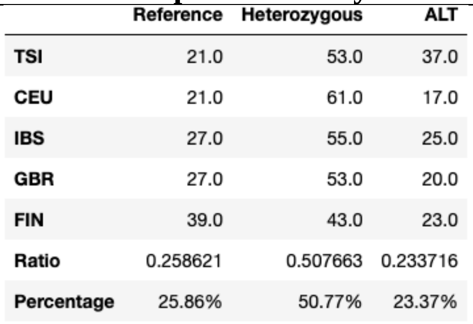
	Reference	Heterozygous	ALT
JPT	10.0	41.0	54.0
CHB	14.0	54.0	38.0
KHV	23.0	40.0	36.0
CDX	21.0	39.0	40.0
CHS	19.0	49.0	37.0
Ratio	0.168932	0.43301	0.398058
Percentage	16.89%	43.30%	39.81%

EUR – European Ancestry



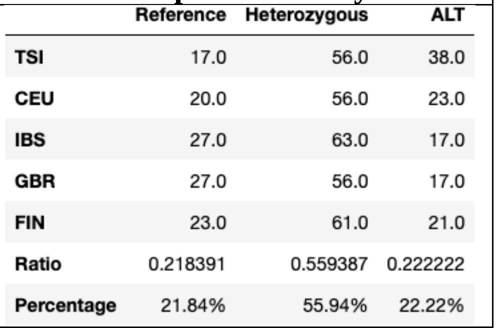
	Reference	Heterozygous	ALT
TSI	96.0	14.0	1.0
CEU	86.0	13.0	0.0
IBS	101.0	6.0	0.0
GBR	74.0	26.0	0.0
FIN	74.0	30.0	1.0
Ratio	0.82567	0.170498	0.003831
Percentage	82.57%	17.05%	0.38%

EUR – European Ancestry



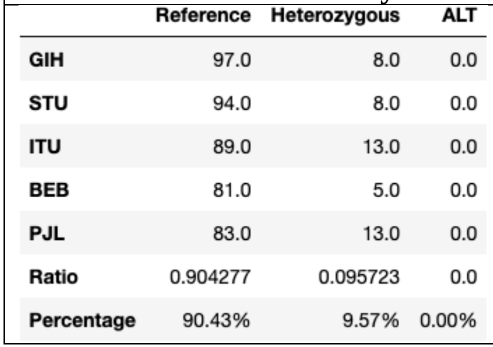
	Reference	Heterozygous	ALT
TSI	21.0	53.0	37.0
CEU	21.0	61.0	17.0
IBS	27.0	55.0	25.0
GBR	27.0	53.0	20.0
FIN	39.0	43.0	23.0
Ratio	0.258621	0.507663	0.233716
Percentage	25.86%	50.77%	23.37%

EUR – European Ancestry



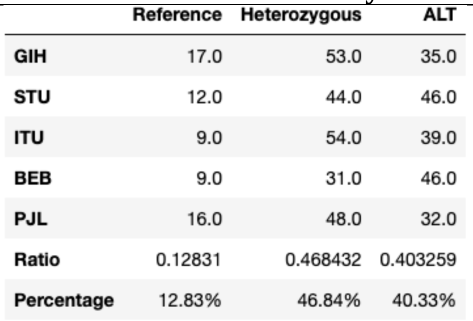
	Reference	Heterozygous	ALT
TSI	17.0	56.0	38.0
CEU	20.0	56.0	23.0
IBS	27.0	63.0	17.0
GBR	27.0	56.0	17.0
FIN	23.0	61.0	21.0
Ratio	0.218391	0.559387	0.222222
Percentage	21.84%	55.94%	22.22%

SAS – South Asian Ancesry



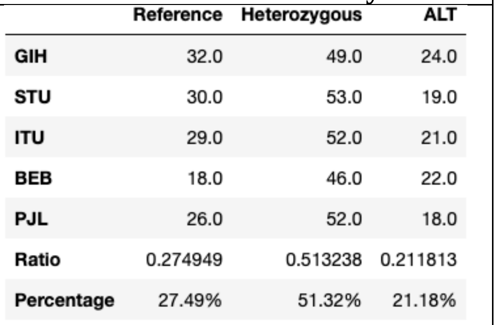
	Reference	Heterozygous	ALT
GIH	97.0	8.0	0.0
STU	94.0	8.0	0.0
ITU	89.0	13.0	0.0
BEB	81.0	5.0	0.0
PJL	83.0	13.0	0.0
Ratio	0.904277	0.095723	0.0
Percentage	90.43%	9.57%	0.00%

SAS – South Asian Ancesry

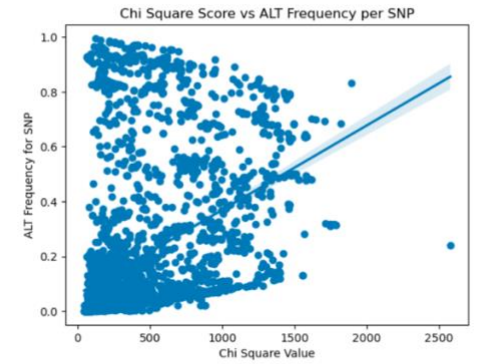
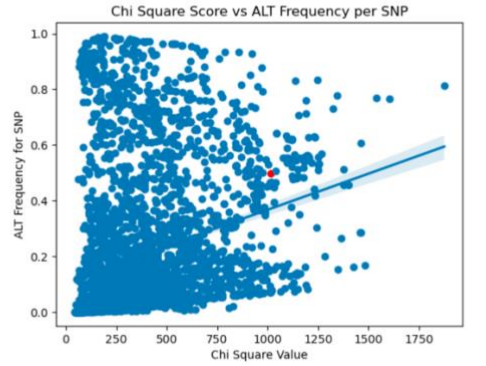
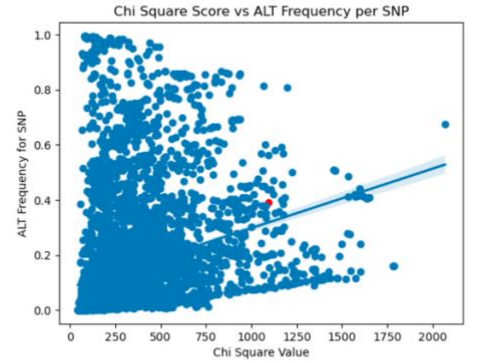


	Reference	Heterozygous	ALT
GIH	17.0	53.0	35.0
STU	12.0	44.0	46.0
ITU	9.0	54.0	39.0
BEB	9.0	31.0	46.0
PJL	16.0	48.0	32.0
Ratio	0.12831	0.468432	0.403259
Percentage	12.83%	46.84%	40.33%

SAS – South Asian Ancesry



	Reference	Heterozygous	ALT
GIH	32.0	49.0	24.0
STU	30.0	53.0	19.0
ITU	29.0	52.0	21.0
BEB	18.0	46.0	22.0
PJL	26.0	52.0	18.0
Ratio	0.274949	0.513238	0.211813
Percentage	27.49%	51.32%	21.18%

		
Target SNP: **Chi2:** 302.88 **ALT Frequency:** 0.0314 (ALT Frequency so small, you cannot view it on the plot.)	**Target SNP**: **Chi2:** 1016.58 **ALT Frequency:** 0.497	**Target SNP**: **Chi2:** 1091.82 **ALT Frequency:** 0.391
R_value: 0.147 **P_value:** 1.07e-146	**R_value**: 0.056 **P_value**: 9.45e-69	**R_value**: 0.063 **P_value**: 4.11e-98

Discussion and Analysis

- A range of about 1,000,000 SNPs were selected for which about 20,000-30,000 SNPs were actually collected (1-3%) for the control data. Of those SNPs collected, about 80% of them are incomputable, meaning that they did not generate a chi-square value because there was no variance in the genotypes.
- As we look at the histograms of chi-square values vs. chi-square value count, we can see that, for the most part, the distribution follows a chi-square distribution.
- Looking across the super populations' genotypic frequencies for all three-target cancer-risk SNPs, there does not seem to be any pattern that associates the super populations with one another. However, the populations inside each respective super population seems to associate closely when having the same genotypic counts.
- When viewing the scatter plots of the chi-square score vs. the ALT frequency per SNP, there is no linear correlation between the two variables, however, there does seem to be a

very slight association when chi-square value increases, ALT allele frequency does as well.

- The chi-square scores for target SNPS 2 and 3 both fall into the 95th percentile for the total control amount of chi-square values for its respective chromosomes. Additionally, their chi-square values are statistically significant as their p-values are less than 1%.
- However, for SNP 1, it has a very low chi-square value and does not follow into the 95th percentile of its respective control data. When taking a closer look at the genotypic frequencies, there does not seem to be a significant variation, meaning that most of the populations all have the reference alleles.
- As we can see from the data, SNP rs6983267-chr8 has the highest ALT allele expression in East Asian Ancestry.
- SNP rs4713266-chr6 has the highest ALT allele expression in East Asian and South Asian Ancestry.

Limitations

There are some limitations to this study. For example, we did not analyze the SNPs of individuals who did end up developing the cancer we are studying from our three target SNPs. Additionally, just because an individual had the DNA combination for a particular cancer, does not mean they are guaranteed to develop it. There is a correlation, however, there is not a causation. There may be tertiary factors that influence an individual's likelihood of developing the cancers studied in this project such as methylation and epigenetics. However, there is some reason to believe that there is an association between the SNP-type frequency and the likelihood of developing certain cancers.

Conclusion

This study showed that there is a statistically significant association between the genomic population for which and individual originates from and the frequency for which they have certain genotypes that increases their risk of developing certain types of cancers.

Works Cited

Fagny, M., Platig, J., Kuijjer, M. L., Lin, X., & Quackenbush, J. (2019). Nongenic cancer-risk snps affect oncogenes, tumour-suppressor genes, and immune function. *British Journal of Cancer*, *122*(4), 569–577. https://doi.org/10.1038/s41416-019-0614-3

Mavura, M. Y., & Huang, F. W. (2021). How cancer risk snps may contribute to prostate cancer disparities. *Cancer Research*, *81*(14), 3764–3765. https://doi.org/10.1158/0008-5472.can-21-1146

Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H., Beckwith, C. A., Chan, J. A., Hills, A., Davis, M., Yao, K., Kehoe, S. M., Lenz, H. J., Haiman, C. A., Yan, C., Henderson, B. E., Frenkel, B., Barretina, J., Bass, A., Tabernero, J., … Freedman, M. L. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nature genetics*, *41*(8), 882–884. https://doi.org/10.1038/ng.403