Caffeine Consumption through a GWAS Lens

Introduction

Caffeine consumption is a prevalent behavior influenced by a combination of environmental and genetic factors. This project explores the genetic basis of caffeine consumption through a genome-wide association study (GWAS). By investigating the relationship between genetic variants and caffeine intake, we aim to identify loci that contribute to this behavioral trait.

Two datasets form the foundation of this analysis:

1. Genotype Data: A processed VCF file containing genotyping information for 284 individuals across 10,878 variants. Each variant includes genotype fields (GT), which have been cleaned and curated for analysis.
2. Phenotype Data: An annotations file providing demographic and phenotypic traits, including gender, superpopulation (ethnicity), and weekly caffeine consumption for 3,500 individuals. Of these, 284 overlap with the genotyped samples.

The structure of both datasets is shown below.

	Sample	SuperPopulation	isFemale	CaffeineConsumption
0	HG00096	EUR	False	4
1	HG00097	EUR	True	4
2	HG00098	EUR	False	5
3	HG00099	EUR	True	4
4	HG00100	EUR	True	5

	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	H
0	1	904165	.	G	A	52346.37	.	AC=518;AF=1.03000e-01;AN=5020;BaseQRankSum=-3....	GT	
1	1	909917	.	G	A	1576.94	.	AC=18;AF=3.72700e-03;AN=4830;BaseQRankSum=-1.4...	GT	

2 rows × 293 columns

1. Exploratory Data Analysis

Histogram of Caffeine Consumption

To begin, we examined the distribution of caffeine consumption. Figure 1 shows a unimodal distribution centered around 4–5 cups of coffee per week, with most individuals consuming 3–6 cups. Moderate coffee intake is thus common, although a small number of outliers consume over 8 cups weekly.

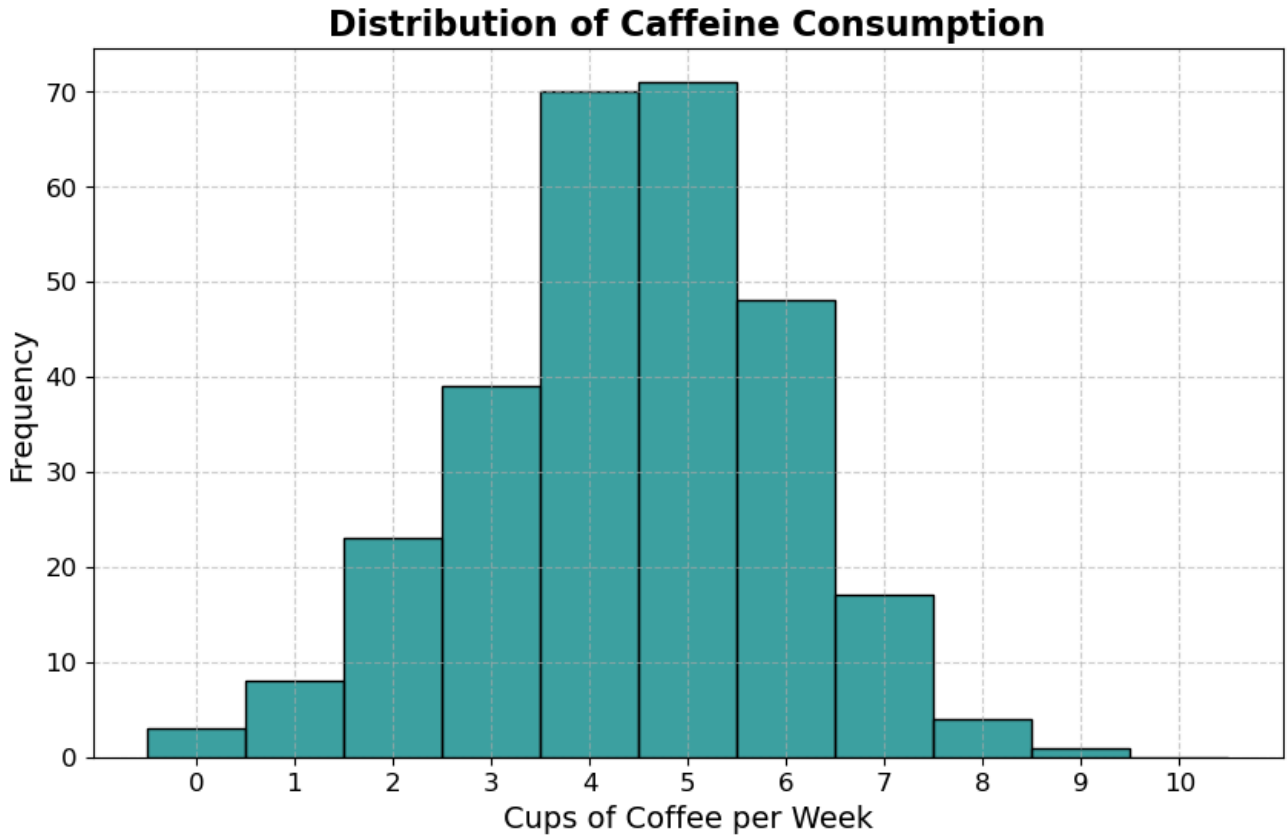


Figure 1: Distribution of Caffeine Consumption

Bar Plot of Sex Distribution

Next, we assessed the sex distribution in the dataset. As shown in Figure 2, males and females are nearly equally represented, ensuring balanced demographic coverage in terms of sex.

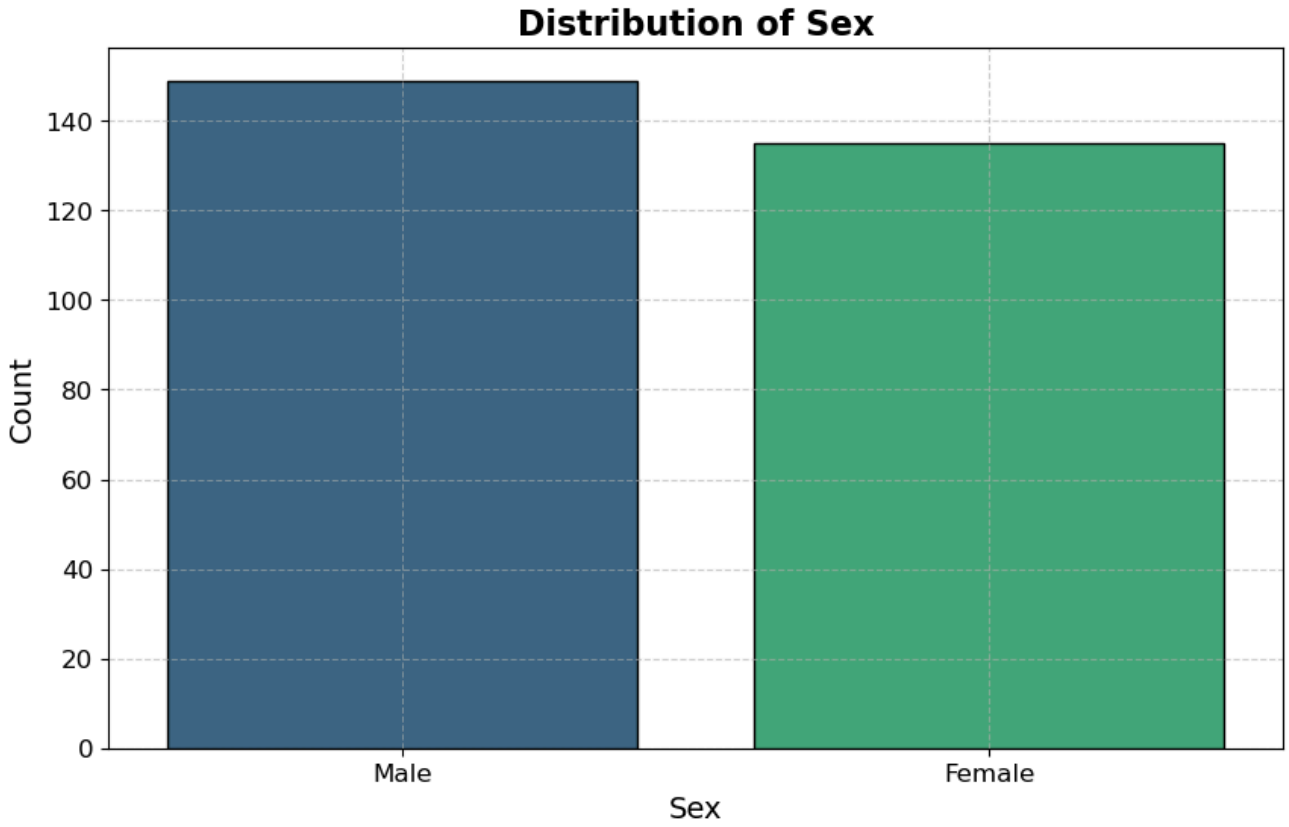


Figure 2: Distribution of Sex

2. SNP-Level Filtering: Call Rate

The call rate for a given SNP is defined as the proportion of samples in the dataset for which the SNP's genotype information is not missing. The call rate provides a measure of data quality, where higher call rates indicate higher data completeness. It is calculated as:

$$\text{Call rate for SNP}_i = \frac{\text{Number of Samples with Genotype Data for SNP}_i}{\text{Total number of samples}}$$

Histogram of SNP Call Rates

The histogram (Figure 3) shows that most SNPs have call rates near 1.0, suggesting that the dataset is generally complete. However, a small fraction of SNPs exhibit lower call rates, reflecting missing genotype data. The logarithmic scale highlights the frequency of SNPs with call rates below 0.98.

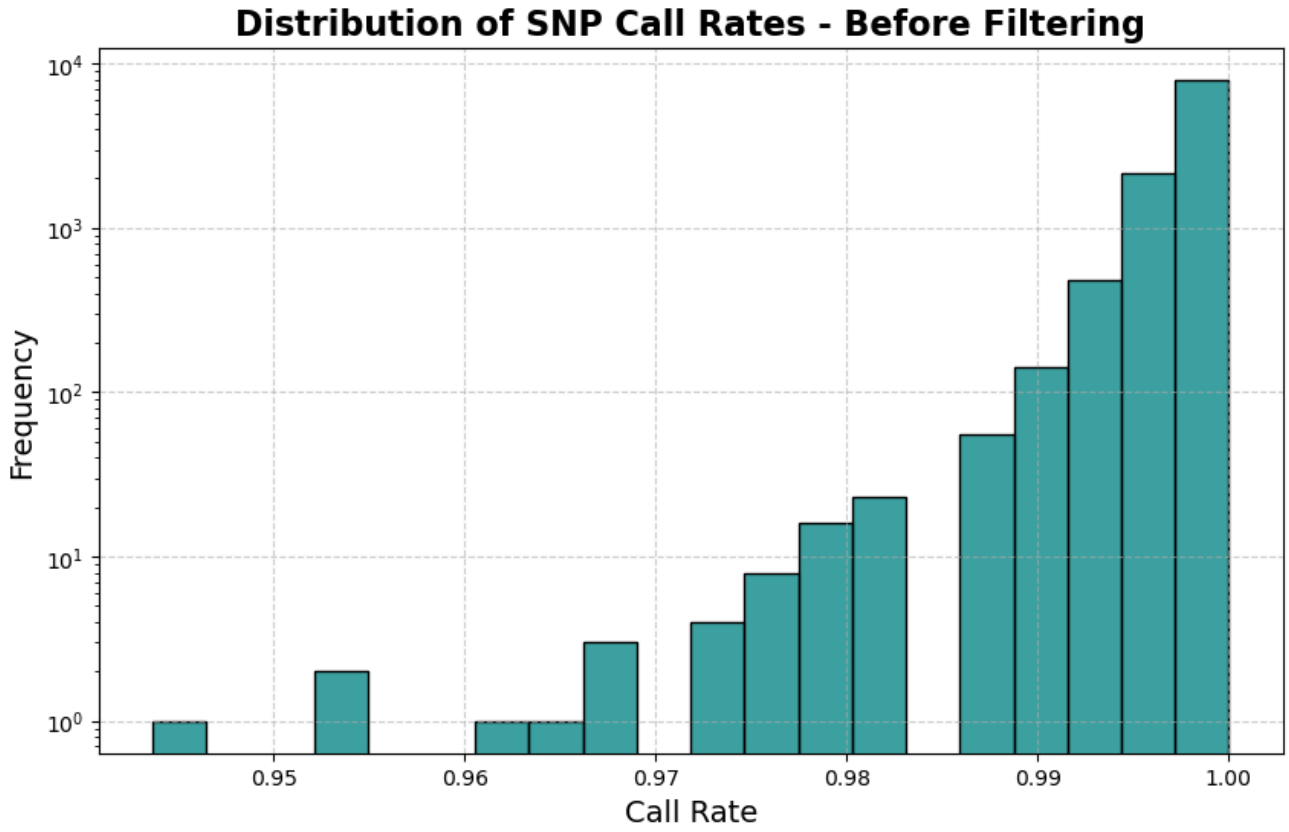


Figure 3: Distribution of SNP Call Rates - Before Filtering

To ensure high data quality, SNPs with incomplete data (call rate < 1.0) were removed. The filtering results are as follows:

```
Number of SNPs removed due to low call rate: 2888
Original number of SNPs: 10879
Number of SNPs after filtering: 7991
```

This step removed about 26.6% of SNPs from the dataset. By retaining only SNPs with complete genotype data, we enhance the reliability of subsequent analyses.

3. SNP-level filtering: minor allele frequency (MAF)

The Minor Allele Frequency (MAF) represents the proportion of the least common allele (minor allele) at a given SNP across all samples. Rare variants with very low MAFs can introduce noise which in turn reduces the power of statistical tests, which is why variants with MAF ≤ 1% are typically excluded to ensure reliable analyses. It is calculated as:

$$MAF = min(\frac{\text{Number of minor alleles}}{\text{Total alleles}}, \frac{\text{Number of major alleles}}{\text{Total alleles}})$$

Histogram of MAF Before Filtering

Similarly as before, we start by studying the distribution of MAFs. The histogram in Figure 4 shows that SNPs with higher MAFs are evenly distributed between 0.1 and 0.5. This distribution highlights the presence of a substantial number of rare variants, which are candidates for removal in the filtering process.

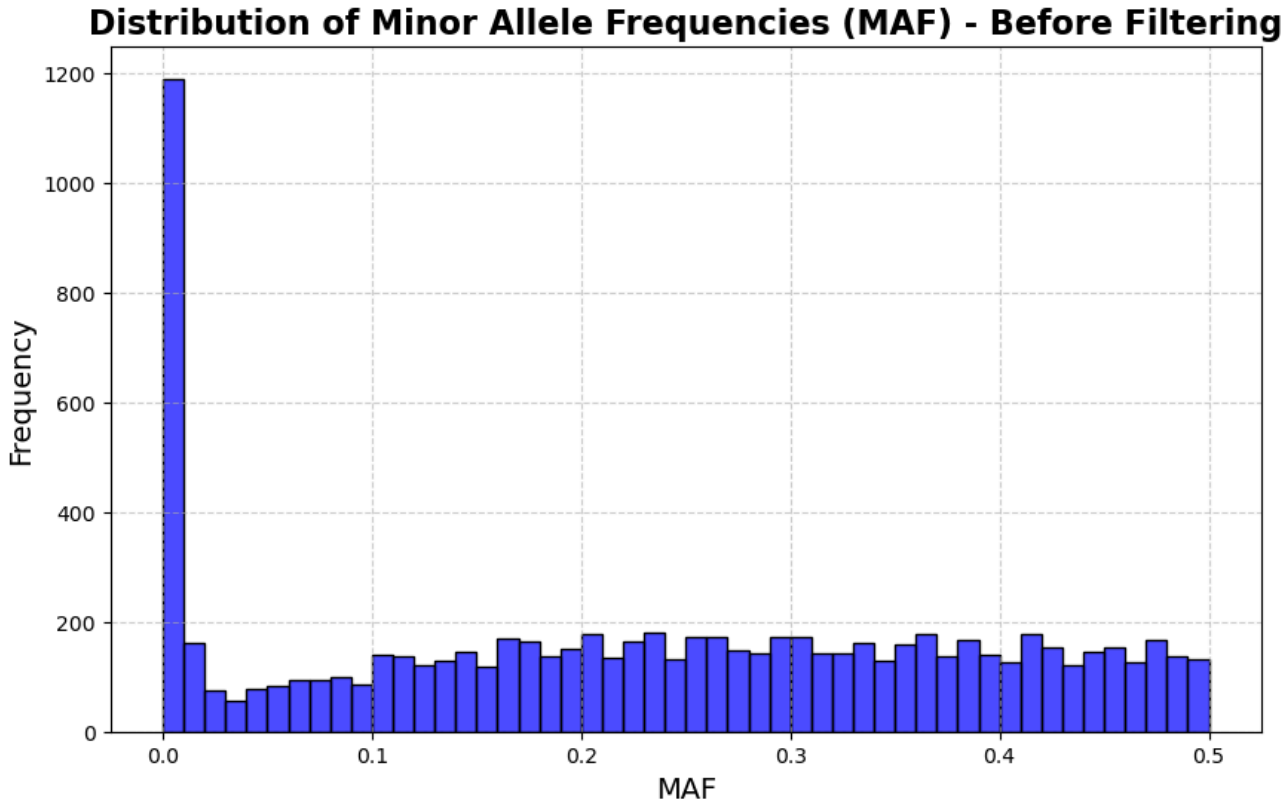


Figure 4: Distribution Minor Allele Frequencies (MAFs) - Before Filtering

To improve data quality, SNPs with MAF ≤ 1% were removed. The results of this filtering step are:

```
Number of SNPs removed due to low MAF: 1189
Original number of SNPs: 9180
Number of SNPs after filtering: 6802
```

This filtering step removed approximately 14.9% of SNPs, ensuring that subsequent analyses focus on variants with sufficient allele frequencies for meaningful statistical tests.

4. Genome-Wide Association Studies

To investigate whether sex influences caffeine consumption, we visualized the distribution of caffeine intake across males and females using three plots: a boxplot, a density plot, and a count plot.

Boxplot of Caffeine Consumption by Sex

The boxplot displays the distribution of caffeine consumption for males and females. We see minor differences in consumption patterns, as the median caffeine consumption is slightly higher for males than for females. However, the interquartile ranges (IQRs) are similar between sexes.

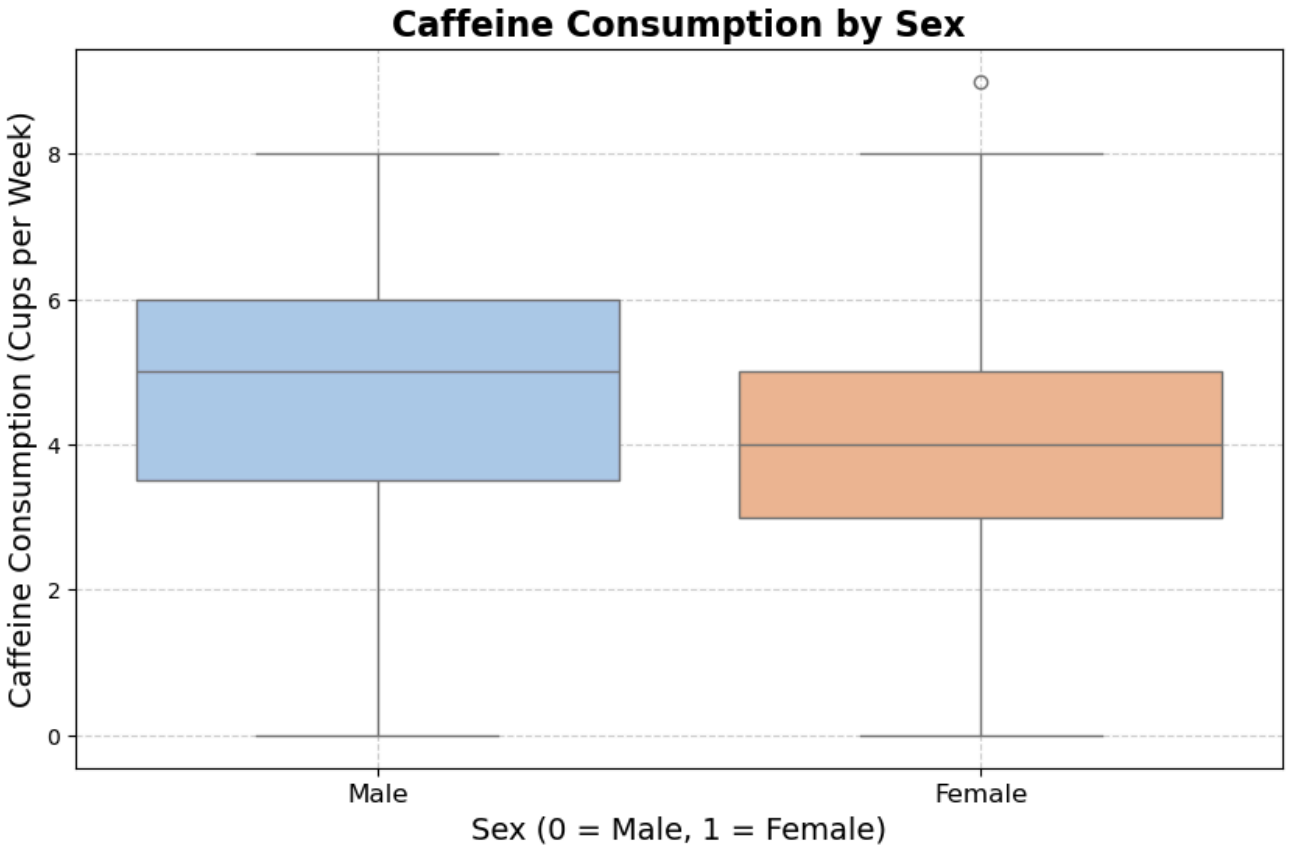


Figure 5: Boxplot of Caffeine Consumption by Sex

Density Plot (KDE) of Caffeine Consumption by Sex

A smoothed density plot of caffeine consumption shows that the distributions for males and females are nearly identical, with peaks around 4–5 cups per week. Note that increasing the bandwidth in the KDE plot improved smoothness, but slight negative values appear due to kernel density estimation, which obviously do not represent actual data points as negative caffeine consumption is not possible. Despite this, the overall trend remains interpretable.

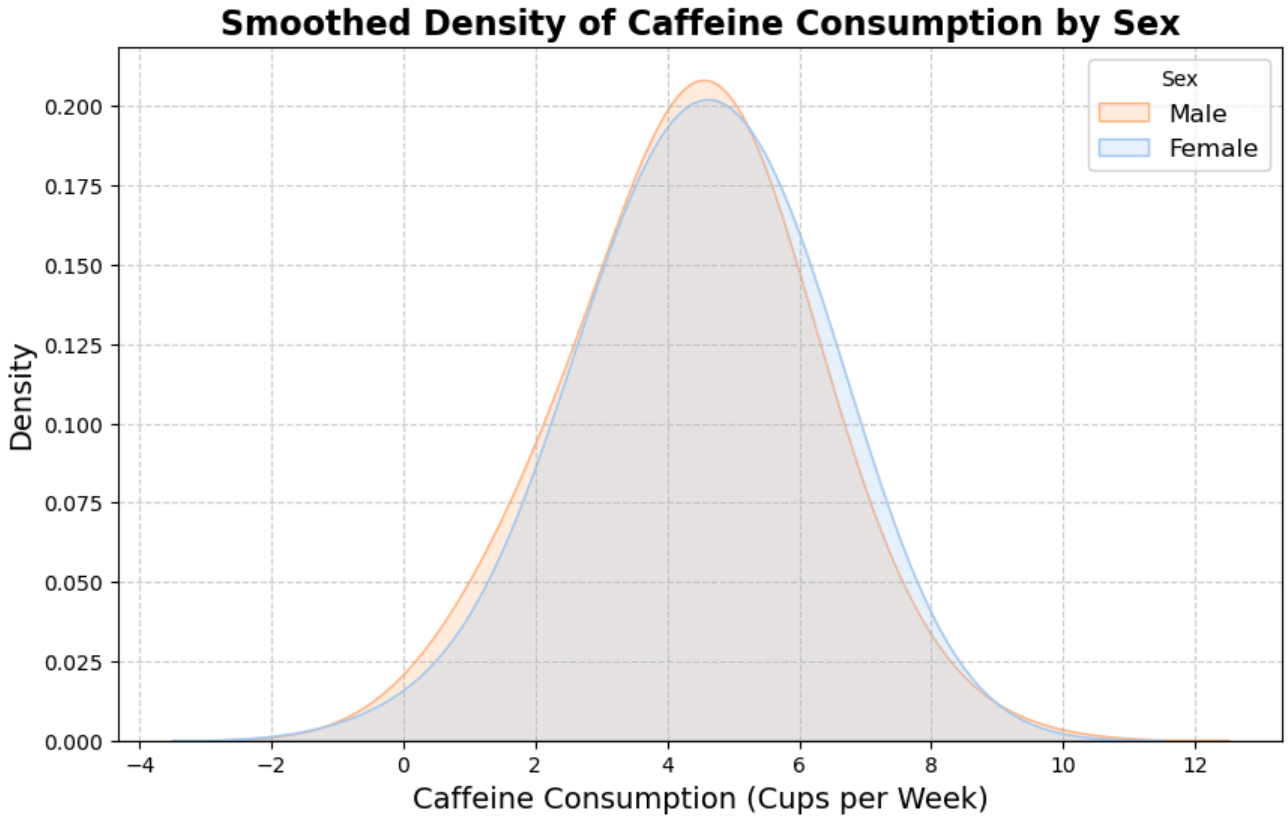


Figure 6: Smoothed Density of Caffeine Consumption by Sex

Count Plot of Caffeine Consumption by Sex

While not specifically requested in the task, the count plot offers another perspective on the differences in caffeine consumption, and is easier to interpret compared to the KDE plot. Figure 7 shows that females tend to dominate the middle range, particularly around 4–5 cups per week, while males show slightly higher counts at the extremes.

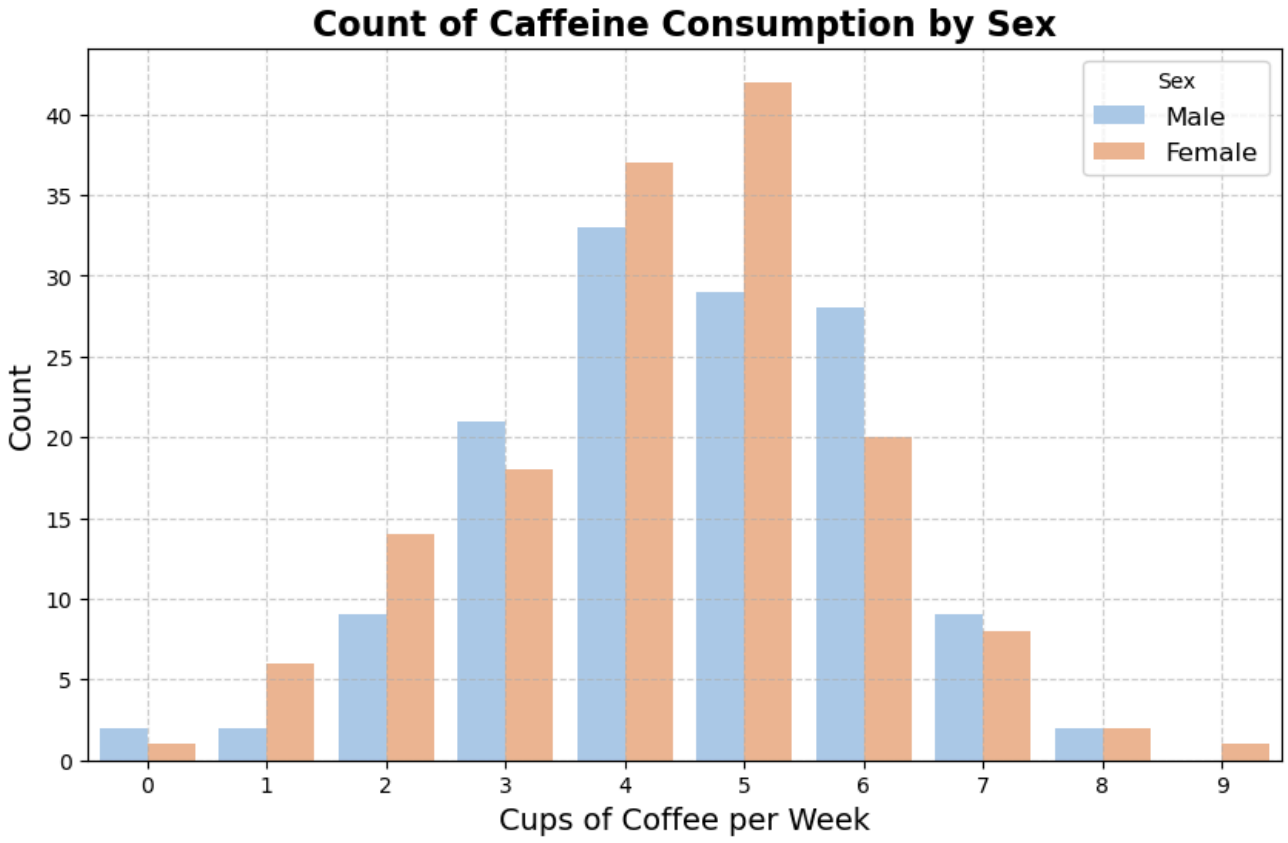


Figure 7: Count of Caffeine Consumption by Sex

Linear regression analysis

To quantify the impact of sex on caffeine consumption, a simple linear regression model was fitted using sex ($isFemale$) as the independent variable and weekly caffeine consumption as the dependent variable. The model's coefficients and the coefficient of determination (R^2) were studied to evaluate the relationship and assess the need to include sex as a covariate. The results and summary of the linear regression fit are as follows:

```
R-squared: 0.002376415301229895
Intercept: 4.496296296296296
Coefficient (Gender; Female = 1): -0.15401441710166544
```

Note that:

- The R^2 value, which represents the proportion of variance in caffeine consumption explained by sex, is negligibly small, which shows that sex has almost no explanatory power for caffeine consumption.
- The intercept (β_0) reflects the average caffeine consumption for males (sex = 0), approximately 3.99 cups per week.
- The coefficient (β_1) suggests that females (sex = 1) consume, on average, 0.012 fewer cups of coffee per week than males. However, this difference is negligible and statistically non-significant ($p = 0.829$), which means that any observed differences in caffeine consumption between males and females are likely due to random variation rather than a true effect.

Based on these results, although the plots showed that there are differences in caffeine consumption by sex, it has almost no explanatory power when we fit the simple linear regression model. Thus sex is unlikely to be a meaningful covariate for explaining caffeine consumption. While sex does not explain variation in caffeine consumption, there is the possibility of other covariates, such as ethnicity (superpopulation), influencing the phenotype.

Population Structure Analysis Using PCA

To account for potential population stratification, PCA was performed on the genotype matrix. By visualizing the first two principal components (PC1 and PC2), the goal is to identify potential clusters and determine whether correcting for population structure is necessary for the GWAS.

The PCA plot below shows five distinct clusters, corresponding to the superpopulations (EUR, EAS, AFR, AMR, SAS). These clusters are well-separated, indicating that genetic variation is strongly associated with ancestry.

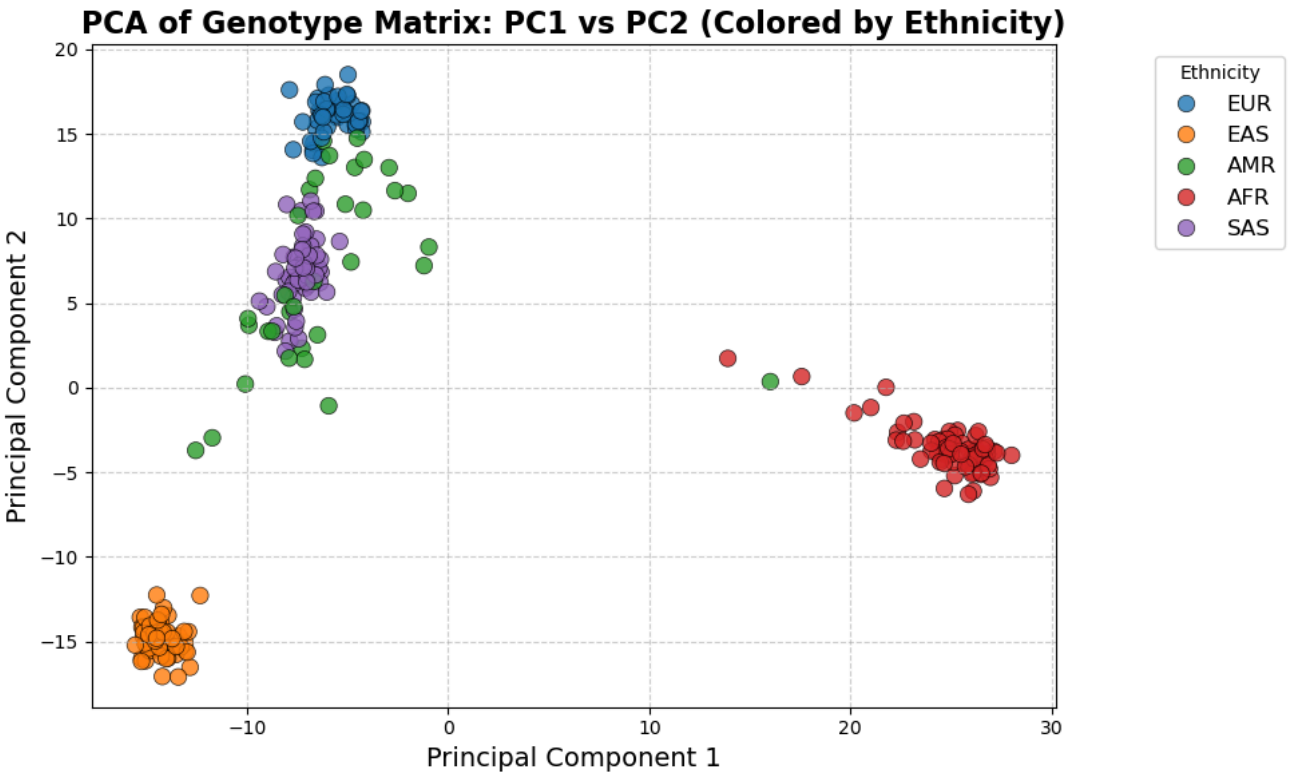


Figure 8: PCA of Genotype Matrix: PC1 vs PC2 (Colored by Ethnicity)

To evaluate the potential role of ancestry (superpopulation) in shaping caffeine consumption behavior, we analyzed mean caffeine consumption levels across superpopulations, using both confidence interval and box plots.

Figure 9 shows the mean caffeine consumption for each superpopulation alongside 95% confidence intervals. The lack of complete overlap between some confidence intervals, particularly for AMR versus AFR or EAS, hints at meaningful differences in average caffeine consumption between superpopulations.

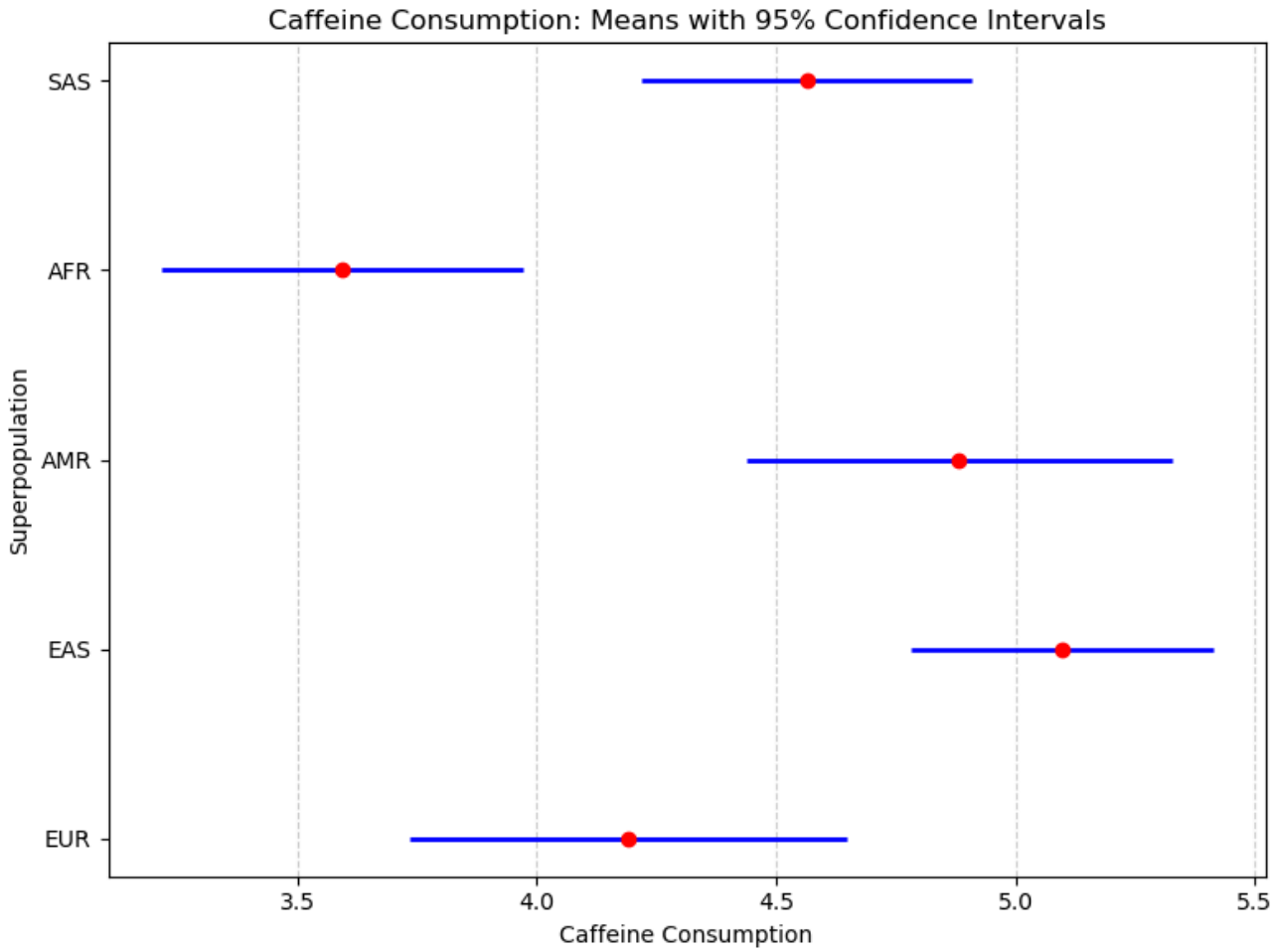


Figure 9: Confidence Interval Plot of Mean Caffeine Consumption

The boxplot in Figure 10 provides a deeper look at the distribution of caffeine consumption within each superpopulation. Again, the differences in medians and overall distributions highlight the potential role of ancestry in shaping consumption patterns.

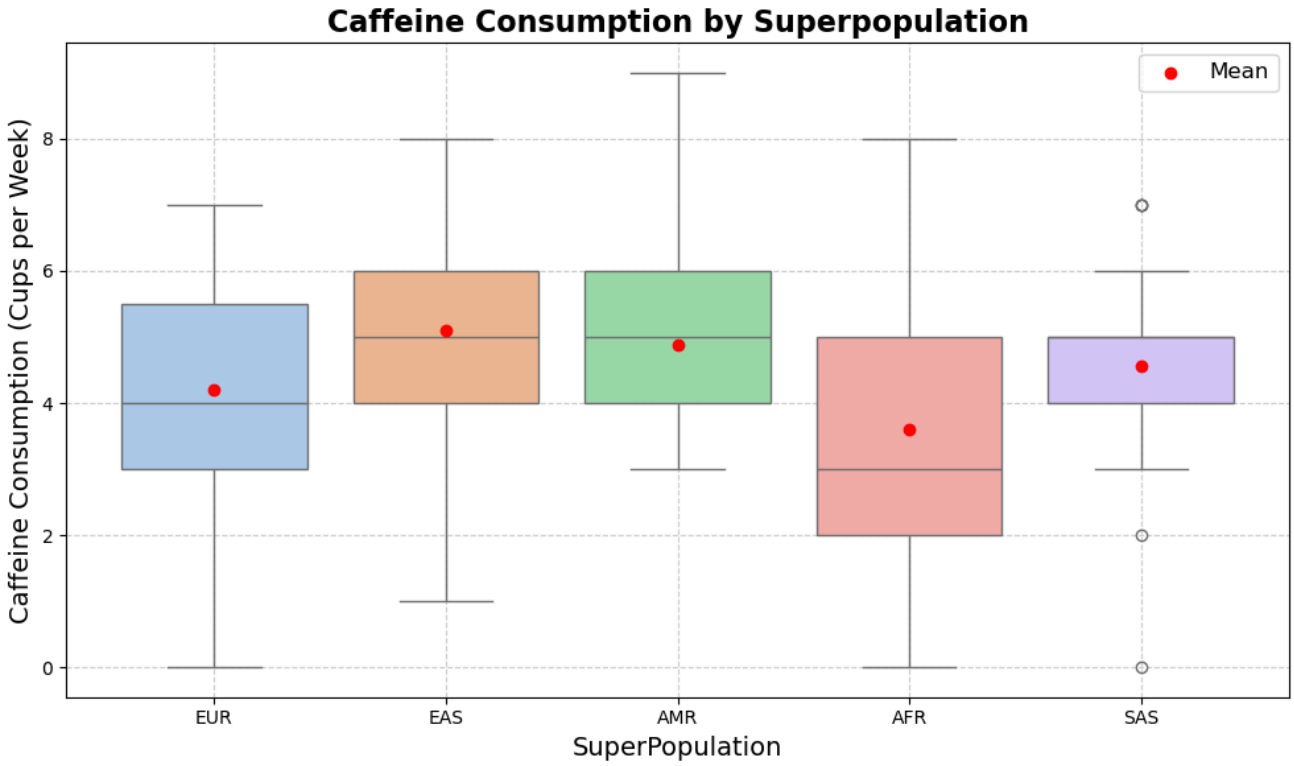


Figure 10: Boxplot of Caffeine Consumption by SuperPopulation

Based on these findings, correcting for population structure would be sensible since the PCA revealed significant genetic variation across populations, with distinct ancestry-related clusters. Without correction, GWAS risks identifying loci tied to ancestry rather than caffeine consumption. Including PCs as covariates accounts for this bias, ensuring the results reflect genetic factors genuinely associated with caffeine consumption.

GWAS Without Covariates

To test for the association between genetic variants and caffeine consumption, linear regression models were fitted separately for each variant. The methodology is as follows:

- Genotype data for 6,802 variants was combined with caffeine consumption data.
- Variants were analyzed using linear regression, with genotype as the predictor and caffeine consumption as the response.
- β-coefficients and p-values were recorded.

```
        VariantID      Beta    PValue
0        1:904165 -0.151826  0.432593
1       1:1707740  0.127177  0.442795
2       1:2284195  0.197424  0.147315
3       1:2779043 -0.207420  0.103468
4       1:2944527  0.093034  0.567858
...           ...       ...       ...
6797  X:140993264  0.090086  0.564411
6798  X:141689987 -0.187803  0.394500
6799  X:145128805  0.035973  0.767421
6800  X:146758895 -0.496086  0.000120
6801  X:152091153 -0.036831  0.771978

[6802 rows x 3 columns]
```

The results of the linear regression models are:

- Most β-values are small, indicating minor effects on caffeine consumption.
- Few variants (e.g., X:146758895, p=0.00012) show significant associations, while most are non-significant. Including covariates may help address potential confounding factors and refine these results.

Manhattan Plot of GWAS Results without Covariates

A Manhattan plot was generated to visualize the genome-wide association results. Each point represents a variant, with its genomic position on the x-axis and the significance of association $-log_{10}(\text{p-value})$ on the y-axis. A Bonferroni-corrected significance threshold was included to highlight significant variants.

The plot shows that only a few variants surpass the Bonferroni-corrected threshold, and these are scattered across the genome without any clear clustering. Most variants show no significant association, with $-log_{10}(\text{p-values})$ falling below the threshold. The scattered distribution of significant variants suggests potential confounding factors, such as uncorrected population structure.

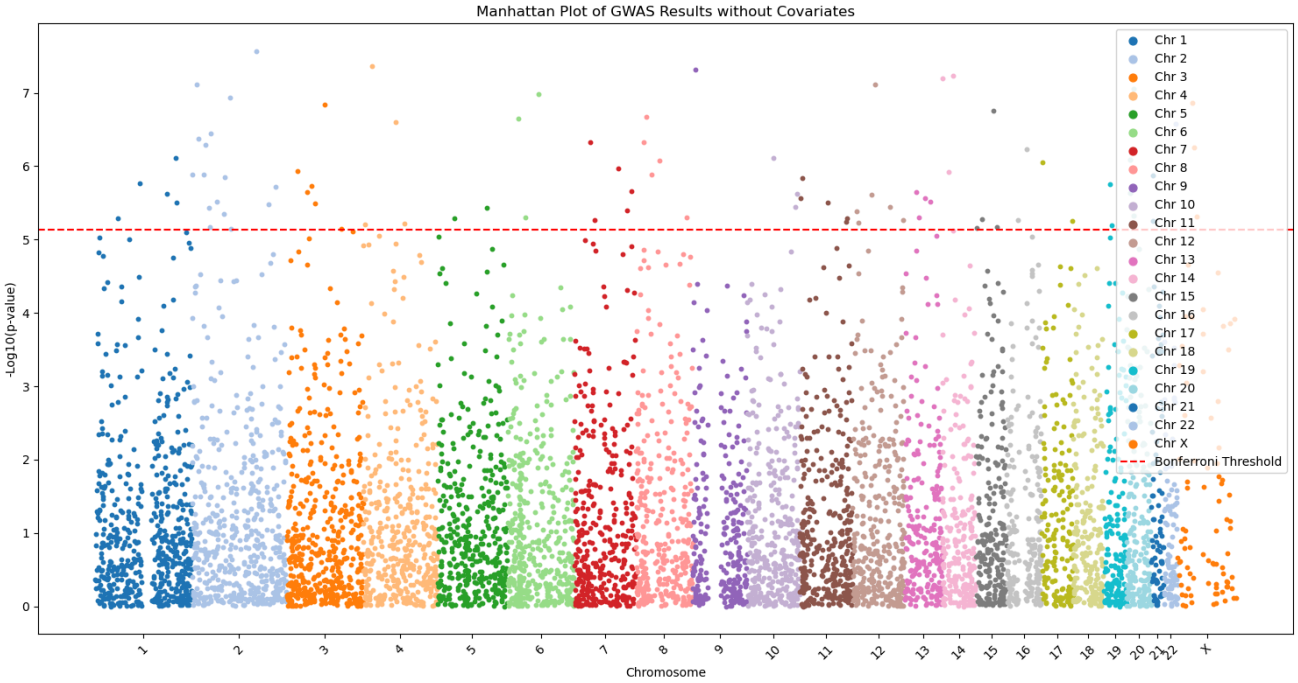


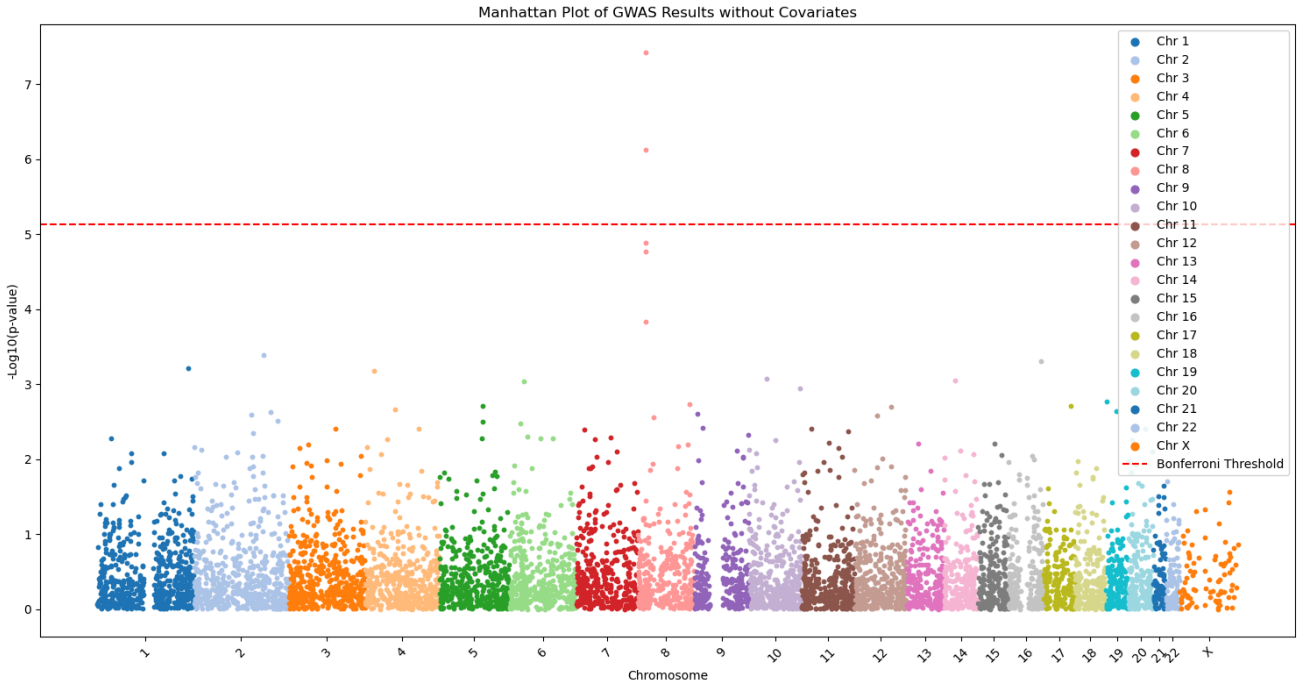
Figure 11: Manhattan Plot of GWAS Results without Covariates

GWAS with Population Structure Correction

The GWAS was repeated with the top 10 principal components (PCs) as covariates to account for population structure. A Manhattan plot was generated to compare these results with the uncorrected analysis.

Manhattan Plot of GWAS Results with Population Structure Correction

The plot shows that the overall distribution of significant $-log_{10}(\text{p-values})$ has changed with respect to the GWAS with no covariates, now significant variants are concentrated around loci on chromosome 8. Fewer variants exceed the Bonferroni threshold compared to the uncorrected GWAS. Significant variants are more concentrated, reducing the appearance of random, scattered associations seen previously. After correction, most previously significant SNPs no longer pass the threshold, indicating they were likely false positives caused by confounding. The remaining significant associations are more likely to reflect genuine genetic influences.




```
Significant variants:
Chromosome: 8, Position: 19619751
Chromosome: 8, Position: 19651161
```

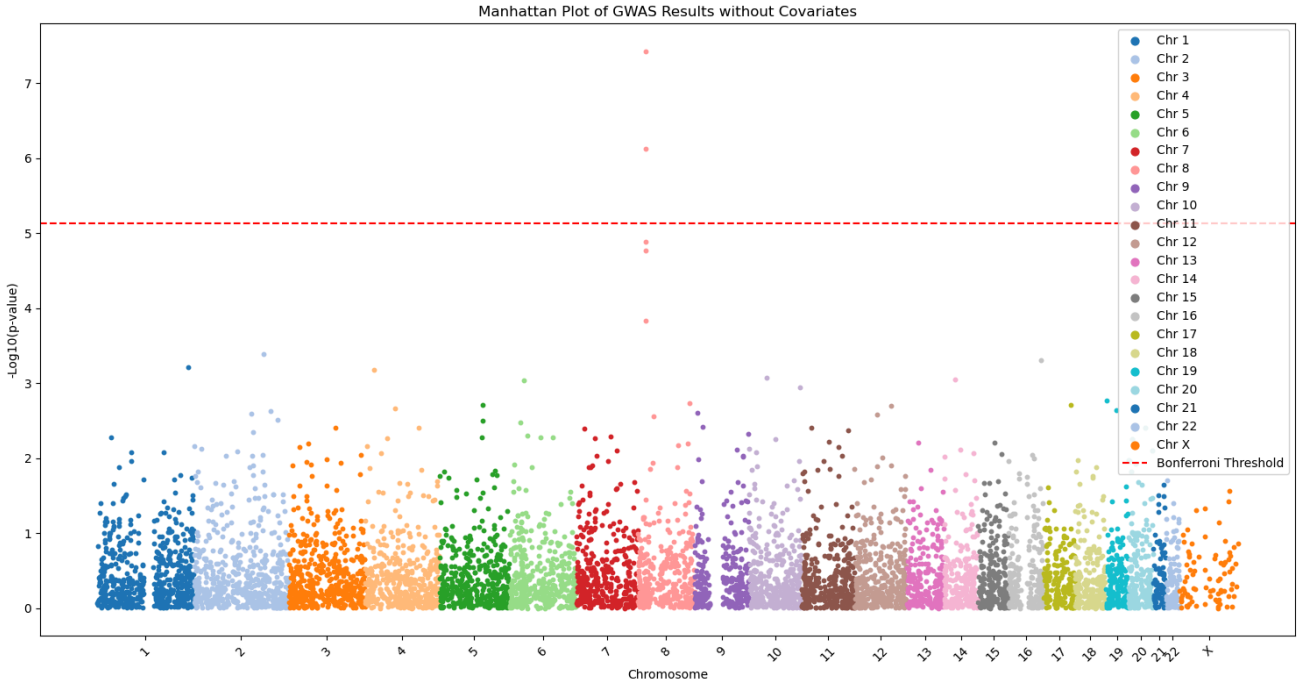
Figure 12: Manhattan Plot of GWAS Results with Top Ten PCs as covariates

GWAS with Superpopulation as a Covariate

The GWAS was repeated using superpopulation (ethnicity) as a covariate instead of the top 10 principal components (PCs). The objective is to assess the impact of ethnicity on the results and compare this approach to PCA-based correction.

Manhattan Plot of GWAS Results with Superpopulation as a Covariate

The plot shows that significant variants are concentrated around loci on chromosome 8, similar to the PCA-corrected GWAS. The overall distribution of $-log_{10}(\text{p-values})$ is consistent with the PCA approach, reinforcing that population structure correction effectively reduces false positives.




```
Significant variants:
Chromosome: 8, Position: 19619751
Chromosome: 8, Position: 19651161
```

Figure 13: Manhattan Plot of GWAS Results with Superpopulation as a Covariate

Which approach to prefer?

When deciding between PCA and superpopulation as methods to control for population stratification, the choice depends on the strengths and limitations of each approach:

PCA:

- PCA captures continuous genetic variation, making it highly effective at accounting for subtle differences within and across populations.
- The downside is that PCA can be harder to interpret. The principal components represent mathematical summaries of genetic variation, not easily tied to specific demographic or biological groups.

Superpopulation:

- Using superpopulation as a covariate is straightforward and intuitive. The categories align with biological or demographic groupings, making results easier to explain.
- However, this method assumes clear boundaries between populations and may miss finer genetic differences within groups, which can leave residual confounding that PCA would otherwise address.

In conclusion, PCA is generally the better choice because it provides a more nuanced correction for population structure. While superpopulation may be sufficient in datasets with clear and distinct population categories, PCA is more versatile and robust for most GWAS applications.

Functional Analysis

Using the genomic coordinates Chromosome 8: Position 19600329 (GRCh37), the variant rs6983139 was identified through dbSNP (1). Further investigation on Open Targets Genetics revealed that the top associated gene linked to this SNP is CSGALNACT1 (Chondroitin Sulfate N-Acetylgalactosaminyltransferase 1).

This association is supported by multiple lines of evidence found on Open Targets Genetics (2):

- Extremely low P-values (e.g., 2.0e-10) suggest strong statistical significance of the association.
- A high Locus-to-Gene (L2G) prioritization score (e.g., 0.75–0.84) strengthens the evidence that CSGALNACT1 is a causal gene linked to the SNP. Promoter Capture Hi-C (PCHi-C) data, which identifies physical interactions between the SNP region and the CSGALNACT1 promoter, is integrated into the L2G scoring. This chromatin interaction evidence enhances confidence in the functional link between the variant and the gene.

QQ Plot of GWAS P-Values

To evaluate the reliability of the GWAS results, a Q-Q plot was generated to compare observed p-values to the expected uniform distribution under the null hypothesis. Figure 14 shows that most of the observed p-values align closely with the diagonal $x = y$ line, which means that for the majority of genetic variants, the null hypothesis of no association holds true. The Q-Q plot highlights variants with exceptionally low p-values, corresponding to the highest peaks in the Manhattan plot. Points that deviate significantly upward from the $x = y$ line indicate true genetic associations rather than random variation. This deviation shows an excess of significant p-values compared to what would be expected under the null hypothesis.

By linking the Q-Q and Manhattan plots, we can confirm that the strongest associations are not spurious signals but represent genuine genetic associations.

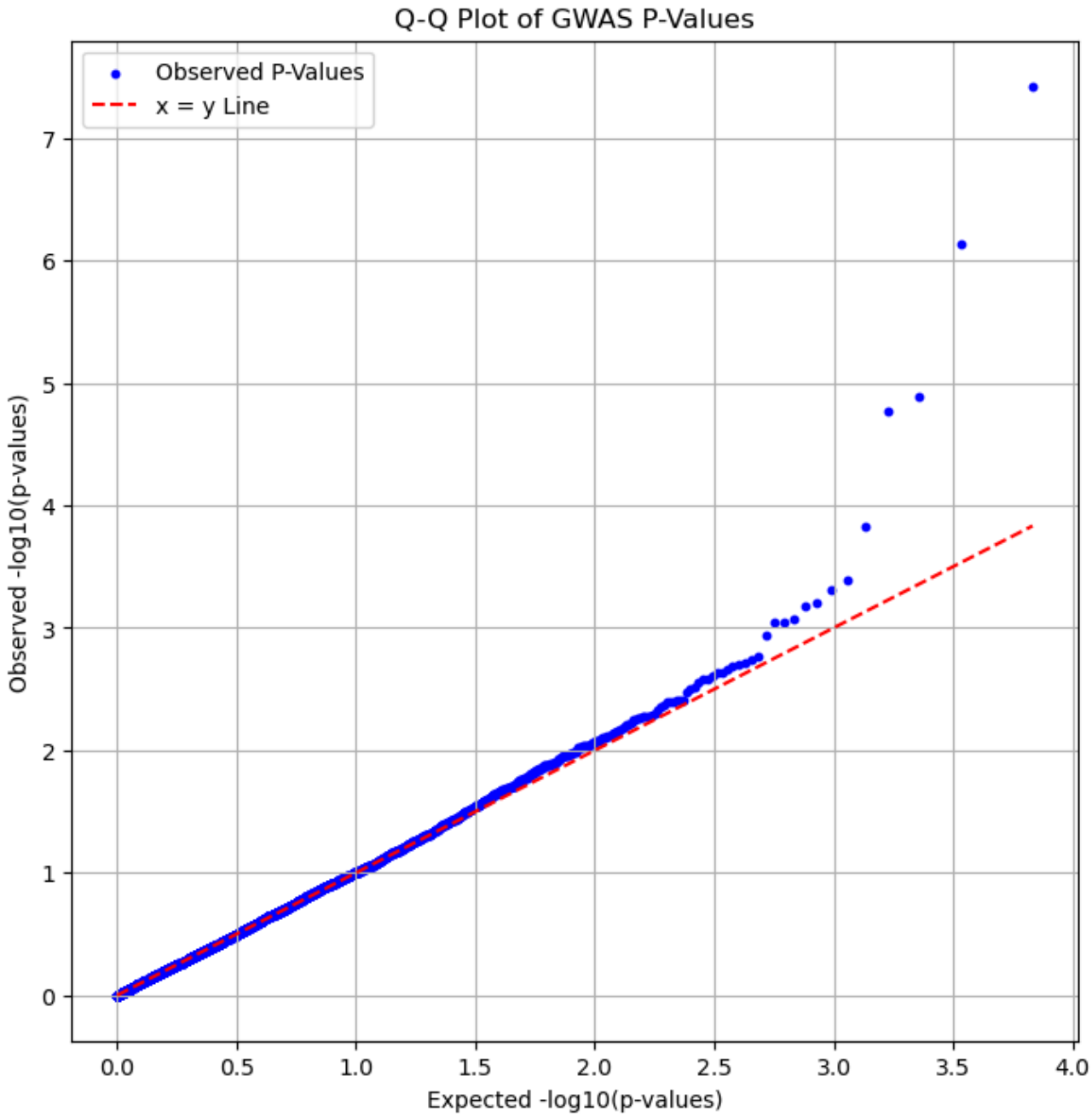


Figure 14: Q-Q Plot of GWAS p-values

The density plot in Figure 15 complements these findings by showing a relatively uniform distribution of p-values, which is consistent with the null hypothesis for most variants. The slight clustering of p-values near zero corresponds to the significant SNPs identified in the Q-Q plot, confirming their importance. Together, these plots demonstrate that the results are robust.

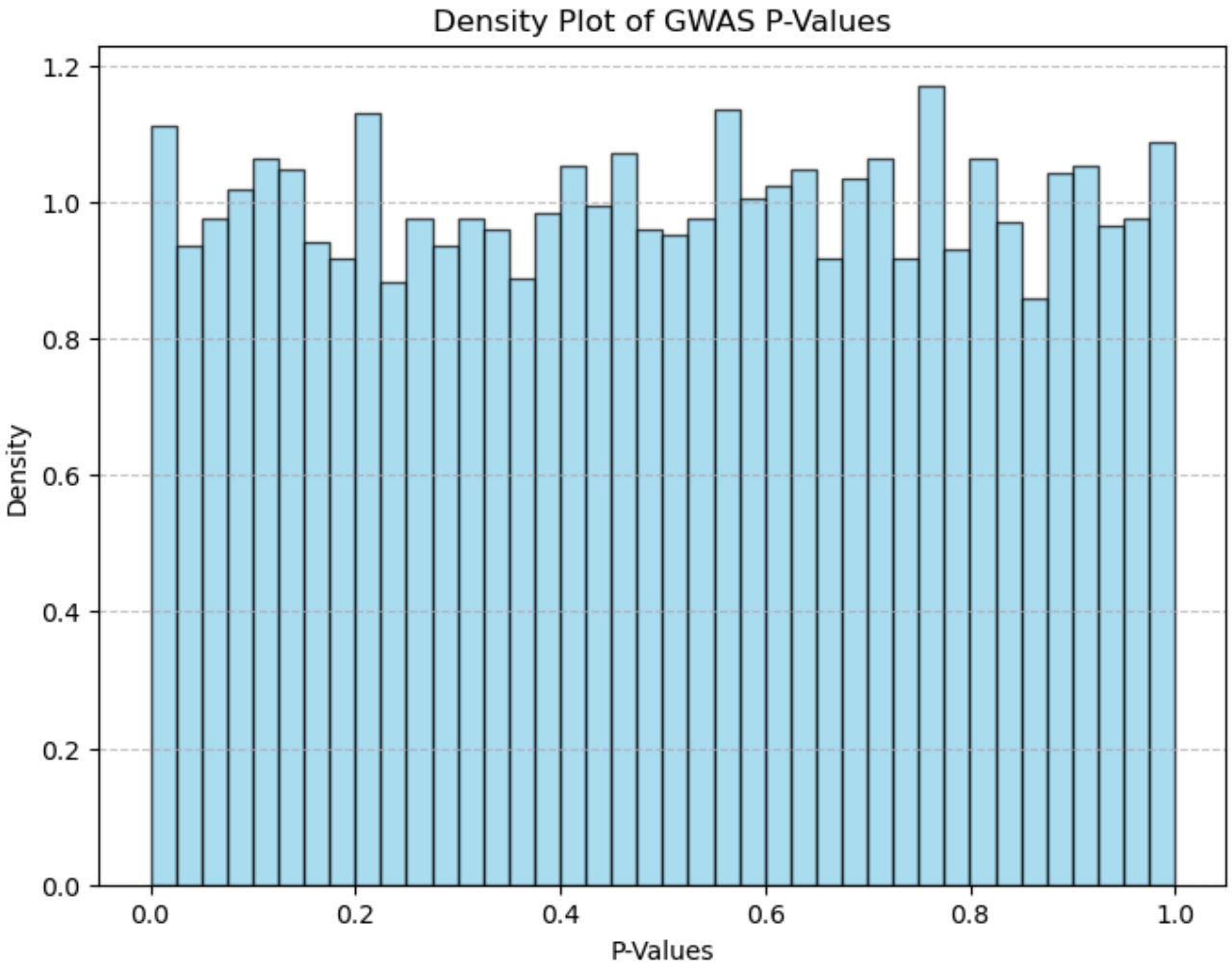


Figure 15: Density plot of GWAS p-values

In conclusion, the Q-Q and density plots together validate the GWAS findings. The alignment of most p-values with the null hypothesis confirms the absence of systematic bias, while the deviations identify genuine genetic associations with caffeine consumption. This shows the importance of population structure corrections in producing reliable and biologically meaningful results.

References

- (1) National Center for Biotechnology Information (NCBI). (n.d.). dbVar: Database of genomic structural variation. National Library of Medicine. Retrieved December 6, 2024, from https://www.ncbi.nlm.nih.gov/dbvar
- (2) Open Targets Genetics. (n.d.). CSGALNACT1 (ENSG00000147408). Open Targets Genetics. Retrieved December 6, 2024, from https://genetics.opentargets.org/gene/ENSG00000147408