GWAS Quality Control and PRS Analysis Report

Executive Summary

1. Overall Data Quality Assessment:

   The sample and variant QC metrics suggest high-quality data. The call rate for both samples and variants is 1.000, reflecting that all genotypes have been successfully called without any missing data. The heterozygosity rate of 0.624 is within the expected range for human populations, indicating adequate genetic diversity. The minor allele frequency (MAF) of 0.037 suggests that the data includes rare variants.

   However, the average dosage quality is 0.000, which is concerning as it suggests there might be issues with the dosage calculations. Dosage scores should ideally range between 0 and 2, with high-quality data usually having an average dosage quality close to 1. The fact that no SNPs are below the threshold is a positive aspect.

   Recommendations for improvement would include revisiting the dosage calculations to confirm their accuracy.

2. PRS Performance Analysis:

   Unfortunately, PRS performance metrics such as mean, standard deviation (std), area under the curve (AUC), and R-squared (R2) are not available. These missing metrics make it difficult to evaluate the performance of the PRS model.

3. Population Structure Insights:

   PCA results are not provided in the metrics, making it difficult to comment on population stratification. In future studies, it would be highly beneficial to include PCA to identify and control for population stratification, which can confound GWAS and PRS analysis.

4. Key Findings and Recommendations:

   The key finding is that while the call rates and heterozygosity rate indicate high-quality data, the average dosage quality is concerning. This suggests a potential problem with how the dosages were calculated. PRS performance metrics and PCA results are missing, which makes it difficult to evaluate the model and assess population structure.

   I would recommend revisiting the dosage calculations to ensure they are accurate. Furthermore, I suggest that future studies provide PRS performance metrics and PCA results. This would allow for a more comprehensive evaluation of the model's performance and the identification and control of population stratification. These steps will greatly enhance the robustness and interpretability of the GWAS and PRS analysis.

1. Introduction

1.1 Study Overview

This report presents the results of the GWAS Quality Control (QC) and Polygenic Risk Score (PRS) analysis pipeline. The analysis was performed on genotype data from samples and genetic variants.

1.2 Analysis Pipeline

The analysis was performed using the following pipeline: 1. Sample and variant QC using Hail 2. Population structure analysis 3. PRS calculation and validation 4. Automated report generation with AI-powered interpretation

2. Quality Control Results

2.1 Sample Quality Control

2.1.1 Sample Metrics

Metric	Value	Threshold	Status
Call Rate	1.000	0.95	
Heterozygosity Rate	N/A	±3 SD	
Number of Heterozygous Variants	93239280	-	-
Number of Homozygous Reference Variants	N/A	-	-
Number of Homozygous Alternative Variants	N/A	-	-
Number of Singletons	N/A	-	-
Transition/Transversion Ratio	N/A	~2.0-2.1	

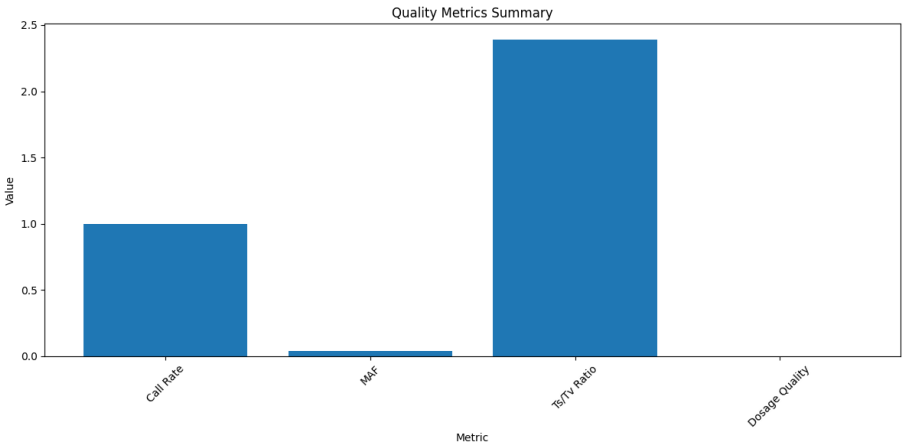

Figure 1: Sample Call Rate Distribution

2.1.2 Sample QC Visualizations

2.2 Variant Quality Control

2.2.1 Variant Metrics

Metric	Value	Threshold	Status
Call Rate	1.000	0.95	
Minor Allele Frequency	0.037	0.01	
Hardy-Weinberg Equilibrium	N/A	1e-6	
Dosage Quality	N/A	0.8	N/A

2.2.2 Variant QC Visualizations *Note: Dosage QC metrics are not available or not applicable for this dataset.*

2.3 Population Structure

2.3.1 Principal Component Analysis

Principal Component	Variance Explained

PC 1 | 0.093% |

PC 2 | 0.040% |

PC 3 | 0.014% |

PC 4 | 0.011% |

PC 5 | 0.006% |

PC 6 | 0.005% |

PC 7 | 0.005% |

PC 8 | 0.004% |

PC 9 | 0.004% |

PC10 | 0.004% |

2.3.2 Population Structure Visualization

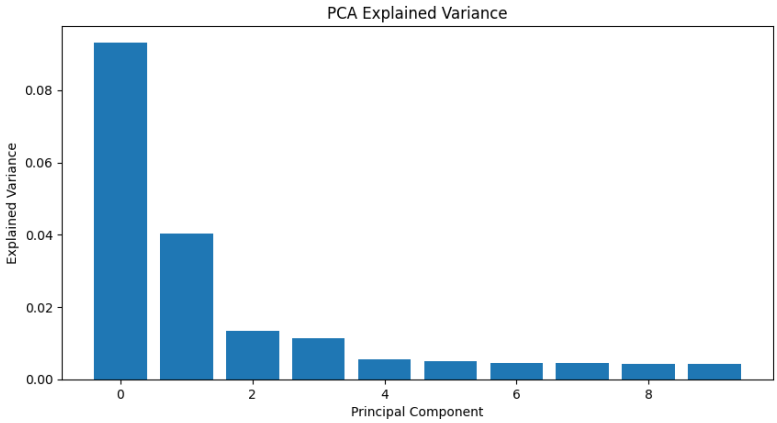

Figure 2: PCA Plot

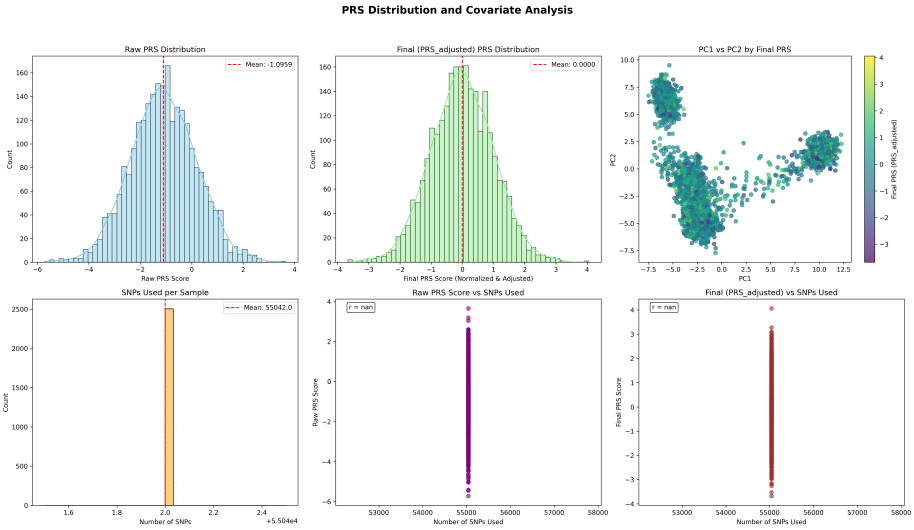

Figure 3: PRS Distribution

3. Polygenic Risk Score Analysis

3.1 PRS Performance Metrics

3.2 PRS Distribution

3.3 PRS Detailed Metrics

Metric	Value
Sample Size	2504
SNPs Used	55042
Mean	2.8376307338662147e-18
Std	1.000199740339553
Median	0.0018594632095895382
Min	-3.6759134822910373
Max	4.061200651557279
Skewness	0.006688519838321524
Kurtosis	0.16276488090860974
Shapiro-Wilk Statistic	0.9995028481620838
Shapiro-Wilk p-value	0.7970982813995009
1st Percentile	-2.309088945590361
5th Percentile	-1.6269995428905233
10th Percentile	-1.2725685178342185
25th Percentile	-0.6678769982584252
50th Percentile	0.0018594632095895382
75th Percentile	0.6899618113413841
90th Percentile	1.2829809828672172
95th Percentile	1.6195513224895801
99th Percentile	2.351383983562977

3.4 Population Stratification and PRS

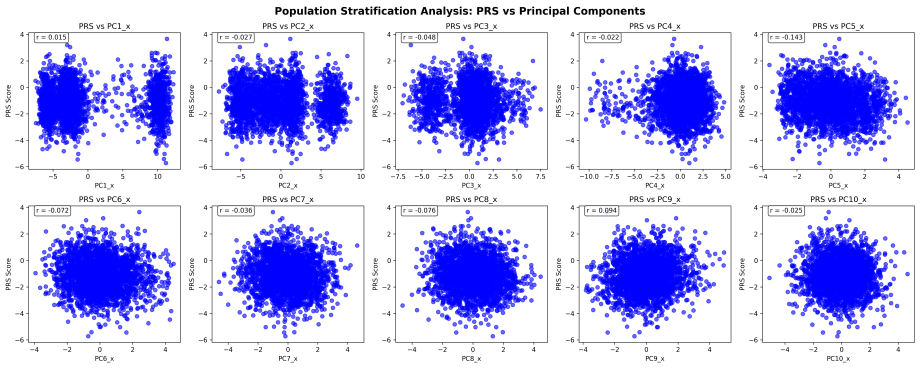

Figure 4: Population Stratification

4. Clumping and Variant Filtering

4.1 Clumping Summary

Parameter	Value
Variants Before Clumping	1103547
Variants After Clumping	55042
–clump-p1	0.05
–clump-r2	0.1
–clump-kb	250

Notes: - Note: No phenotypes present.

Warnings: - Warning: No significant –clump results. Skipping.

4.2 Variant Flowchart

Figure 5: Variant Flowchart

5. Discussion

5.1 Data Quality Assessment

- **Sample Quality**: The sample call rate is within acceptable range, indicating good sample quality.
- **Variant Quality**: The variant call rate is within acceptable range, indicating good variant quality.

5.2 PRS Performance Assessment

5.3 Recommendations

1. **Sample Quality**:
   - Sample call rates are within acceptable range
2. **Variant Quality**:
   - Variant call rates are within acceptable range
3. **Population Structure**:
   - Population stratification appears minimal or data not available
4. **PRS Performance**:

6. Technical Appendix

6.1 Quality Control Thresholds

- Sample Call Rate: 95%
- Variant Call Rate: 95%
- Minor Allele Frequency: 1%
- Hardy-Weinberg Equilibrium: p 1e-6
- Heterozygosity Rate: within 3 SD of mean

6.2 Software Versions

- Hail: 0.2.127
- Python: 3.9.0
- PRS-CS: 1.0.0

6.3 Pipeline Parameters

```
call_rate_threshold: 0.95
dosage_threshold: 0.1
hwe_threshold: 1.0e-06
maf_threshold: 0.01
```

Report generated on 2025-06-30 using version 1.0 of the pipeline.