Title: The Connection Between GWAS and Linear Regression

The Connection Between GWAS and Linear Regression

1. Introduction

Genome-Wide Association Studies (GWAS) have revolutionised the understanding of the genetic architecture of complex traits and diseases. By systematically scanning the genome for single-nucleotide polymorphisms (SNPs) associated with phenotypic variation, GWAS enables researchers to identify genetic loci that influence traits ranging from height to susceptibility to common diseases such as type 2 diabetes.

A critical component underlying GWAS is the application of statistical models to assess the association between genotype and phenotype. Among these models, linear regression serves as the foundational approach for analysing continuous traits. This essay explores the intimate relationship between GWAS and linear regression, demonstrating how regression models facilitate SNP discovery and interpretation.

2. Overview of GWAS

GWAS involves scanning hundreds of thousands to millions of SNPs across the genome in large cohorts of individuals. The primary objective is to detect statistical associations between genetic variants and phenotypic traits.

Typical GWAS workflow includes:

1. **Data collection and preprocessing**: Collecting genotype data (SNP arrays or sequencing) and corresponding phenotypic measurements.

2. **Quality control (QC)**: Filtering SNPs and individuals based on missingness, minor allele frequency (MAF), and genotype call rates.

3. **Statistical association testing**: Evaluating the relationship between each SNP and the phenotype using appropriate models.

4. **Multiple testing correction and visualisation**: Adjusting for the large number of tests and visualising results through Manhattan and QQ plots.

Challenges in GWAS include controlling for population stratification, multiple testing burden, and potential confounding variables.

3. Linear Regression in GWAS

Linear regression models the relationship between a dependent variable (phenotype) and independent variables (SNPs). For a single SNP:

$$Y = \beta 0 + \beta 1X + \varepsilon$$

Where:

- Y: continuous phenotype
- X: genotype coded as 0,1,2
- β1: effect size of SNP on phenotype
- ε: residual error

In GWAS, linear regression is applied SNP-by-SNP, yielding effect sizes and p-values to identify significant SNPs.

4. Connecting GWAS and Linear Regression

Each GWAS result for a SNP is derived from a linear regression model. Key points:

- Genotype coding: 0, 1, or 2 minor alleles.
- Regression coefficient (β1): magnitude of SNP effect.
- P-value: statistical significance of association.

Example: For 100 individuals and height as a phenotype, linear regression for each SNP identifies SNPs significantly affecting height.

5. Practical Considerations

- **Multiple testing**: Millions of SNPs require stringent significance thresholds (e.g., Bonferroni: $p < 5e\text{-}8$).
- **Covariates**: Age, sex, population structure.
- **Assumptions**: Linearity, homoscedasticity, normal residuals.
- **Alternative models**: Logistic regression for binary traits.

6. Visualisation

- **Manhattan plot**: -log10(p-values) across the genome.
- **QQ plot**: compare observed vs expected p-values.
- Linear regression coefficients can also be visualised.

7. Example Mini GWAS

7.1 Simulate mini genotype and phenotype

```
plink2 --simulate-geno 50 100 0.1 --make-bed --out mini_geno
awk '{if(NR>1) print $1, $2, rand()*10}' mini_geno.fam > mini_pheno.txt
```

7.2 Run linear regression GWAS

```
plink2 --bfile mini_geno --pheno mini_pheno.txt --glm allow-no-covars --out mini_assoc
```

7.3 Analyse with R

```
library(qqman)
gwas <- read.table("mini_assoc.PHENO1.glm.linear", header=TRUE)
if("#CHROM" %in% colnames(gwas)) colnames(gwas)[colnames(gwas) == "#CHROM"] <-
"CHR"
gwas$BP <- as.numeric(gwas$BP)
gwas$PVAL <- as.numeric(gwas$P)
png("mini_manhattan.png", width=1000, height=600)
manhattan(gwas, chr="CHR", bp="BP", snp="ID", p="PVAL", main="Mini GWAS Manhattan
Plot", suggestiveline=-log10(1e-5), genomewideline=-log10(5e-8))
dev.off()
top_snps <- gwas[order(gwas$PVAL), ][1:10, ]
print(top_snps)
```

- The Manhattan plot illustrates SNPs with significant p-values.
- The top SNPs table shows the strongest associations.

8. Conclusion

GWAS fundamentally relies on linear regression to detect associations between SNPs and continuous traits. Regression coefficients provide effect size estimates, and p-values determine significance. Understanding this relationship is crucial for designing, analysing, and interpreting GWAS, and translating findings into biological insights.

9. References

1. Visscher PM et al., 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet.

2. Balding DJ, 2006. A tutorial on statistical methods for population association studies. Nat Rev Genet.

3. Purcell S et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet.

4. R Core Team, 2024. R: A Language and Environment for Statistical Computing.