Biostatistics

Applications in Genetic and Epigenetic Data

Nuno Sepúlveda, 08.12.2025

Syllabus

1. General review
    a. Population/Sample/Sample size
    b. Type of Data – quantitative and qualitative variables
    c. Common probability distributions/popular tests
2. Applications in Medicine
    a. Construction and analysis of diagnostic tools – Binomial distribution, ROC curve, sensitivity, specificity, Rogal-Gladen estimator
    b. Estimation of treatment effects - generalized linear models
    c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model
3. Applications in Genetic and Epigenetic Data
    a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
    b. Methylation association studies – M versus beta values, estimation of biological age
4. Applications in Serological Data Analysis
    a. Determination of seropositivity using Gaussian mixture models
    b. Reversible catalytic models for estimating seroconversion rate
    c. Sample size calculation for estimating seroconversion rate

Mendelian Genetics

Single gene, single binary trait

Complete penetrance

Rules of dominance/recessiveness

Non-Mendelian Genetics

Complex binary traits

Presence of common diseases (diabetes, multiple sclerosis, COVID-19 lethality)

Categorical traits

Eye color

Multiple interacting gene involved

Quantitative traits

Height, Haemogloblin levels, blood pressure

Basic question

What are the genes involved and what is their action on the phenotype?

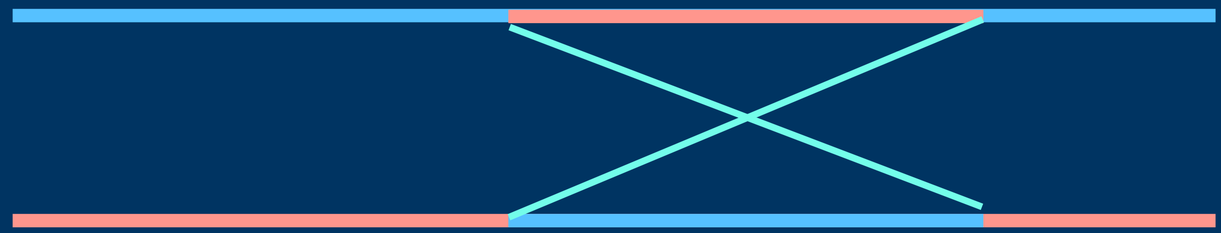
The identification of the genes involved is expected to improve human health by targeting the causative genes

Recombination

Paternal Chromosome a

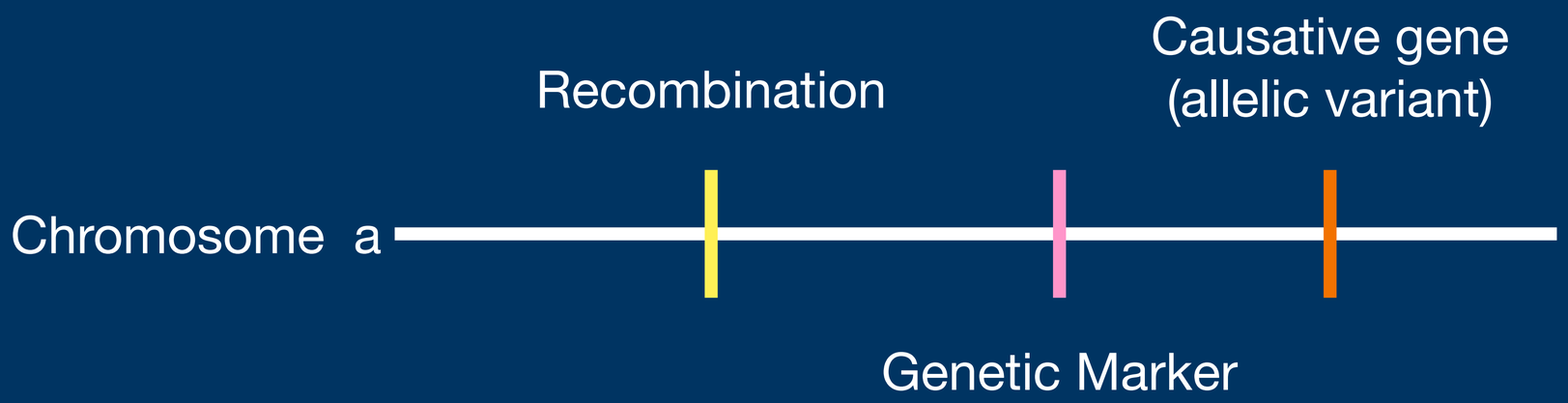
Maternal Chromosome a

During formation of gametes (sperm/egg cells)



Linkage and recombination

Complete linkage

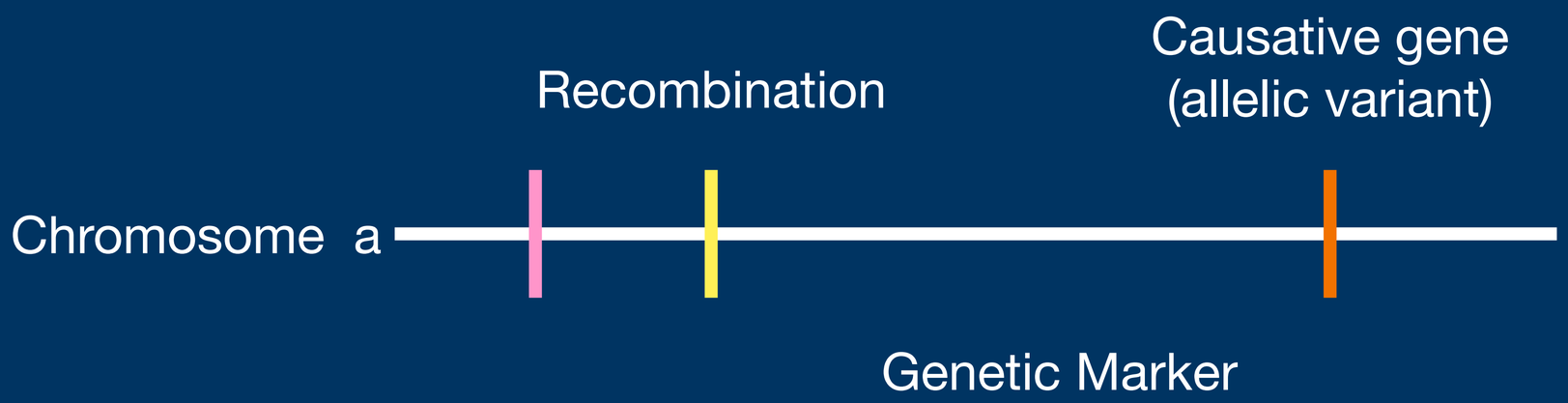


If a genetic marker and the causative gene are physically close to each other, then the genetic marker and the gene are inherited together.

The genetic marker fully represents the true statistical association between the causative gene and the phenotype.

Linkage and recombination

Incomplete linkage



If a genetic marker and the causative gene are not physically close to each other, then a recombination might occur during gametogenesis. This recombination is passed onto the offsprings

The genetic marker partially represents the true statistical association between the causative gene and the phenotype.

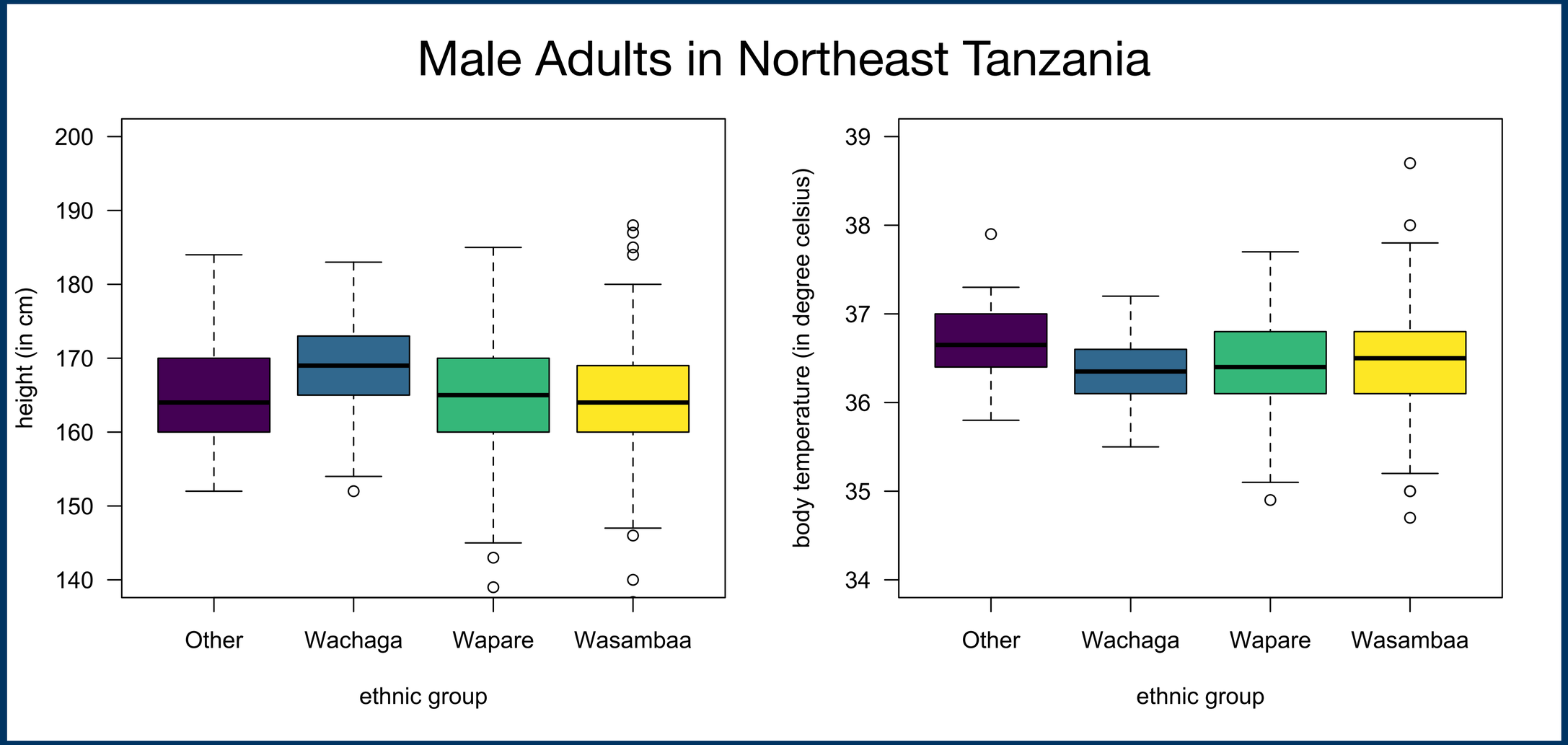
Linkage and recombination

No linkage

If a genetic marker and the causative gene are in different chromosomes, there is no association between the genetic marker and the causative gene due to independent segregation of chromosomes.

The genetic marker is not associated with the phenotype.

A bit about quantitative traits



What are the genes controlling the height and body temperature of these individuals?

Fisher's infinitesimal (or polygenic) model

A quantitative trait is affected by a large number of alleles located at different genes. The effect of these alleles is additive on the quantitative.

$$Y_i = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j X_i \rightsquigarrow ?$$

where

X_{ij} is the number of alleles in genotype at gene j in individual i

α_0 is the overall average of environmental factors and alleles located at other genes

α_j is the phenotype effect ofadding an allele to the genotype at gene j

Practical implications of Fisher's infinite allele model

The genotype of an individual is converted in the number of a given allele (no rules of dominance and recessiveness as proposed by Mendel)

Interaction among different causative genes is discarded

↓

We can simplify the analysis of multiple genetic markers by analysis each genetic marker separately

↓

Additive model for the analysis of single genetic marker

Additive model for a single genetic marker for diploid organisms (humans!)

Y_i = random variable for the quantitative trait in individual i

$$Y_i \mid \mu_i, \sigma \rightsquigarrow N(\mu_i, \sigma^2)$$

X_i = the number of a given allele in the genotype of individual i for the genetic marker

$$X_i^* \in \{'aa', 'aA', 'AA'\} \longrightarrow X_i \in \{0,1,2\}$$

$$\mu_i = \alpha + \alpha_1 X_i$$

Assume sampling of unrelated individuals from the population

Additive model is a simple linear regression using a covariate with three numeric levels

Note: if sampling includes individuals from the same family, we need to include a random effect to contemplate the correlation among individuals due to genetic relatedness (similar to repeated measurement models - linear mixed models)

Testing the effect of a marker on the phenotype

$H_0 : \alpha_1 = 0$ versus $H_1 : \alpha_1 \neq 0$

Wald's Score test

$$S = \frac{\hat{\alpha}_1}{se(\hat{\alpha}_1)} \mid H_0 \rightsquigarrow Normal(\mu = 0, \sigma^2 = 1)$$

Wilks' likelihood ratio test

$$\Lambda = (-2)\frac{L(\hat{\alpha}_0^*)}{L(\hat{\alpha}_0, \hat{\alpha}_1)} \mid H_0 \rightsquigarrow \chi^2_{(1)}$$

$L(\hat{\alpha}_0^*) =$ maximised log-likelihood of the regression model without the covariate

$L(\hat{\alpha}_0, \hat{\alpha}_1) =$ maximised log-likelihood of the regression model with the covariate

Extending the additive model

$$Y_i \,|\, \mu_i, \sigma \rightsquigarrow N(\mu_i, \sigma^2)$$

Under the assumption of sampling unrelated individuals

$$\mu_i = \alpha + \alpha_1 X_i + \underbrace{\beta_1 Z_{1i} + \cdots + \beta_p Z_{pi}}$$

Non genetic covariates

Under the assumption of sampling unrelated individuals

$$\mu_i = \alpha + \underbrace{\alpha_1 X_{1i} + \cdots + \alpha_1 X_{mi}} + \underbrace{\beta_1 Z_{1i} + \cdots + \beta_p X_{pi}}$$

Associated genetic markers

Non-genetic covariates

What are the most common genetic markers in human genetics?

Single nucleotide polymorphisms

CTCTCTTCTGAGTC

CTCTCTGCTGAGTC

Located in a specific region of the genome

84.7 million SNPs are already catalogued (the 1000 genomes project consortium)

The SNPs are denoted by their rs_number (e.g., rs334)

Detaled information about SNP can be found in:

https://www.ncbi.nlm.nih.gov/snp/ or https://www.ensembl.org/

The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526, 68–74 (2015).

Go to https://www.ncbi.nlm.nih.gov/snp/

and

Check information about the following SNPs:

rs334 rs1050829 rs1050828

What is your interpretation of the information provided?

Minor allele frequency (MAF)

Frequency of the allele with less frequency in the population

In practice, SNPs with alleles with a MAF < 0.025 are not included in the analysis

Why?

Exercise (additive model): data_Tanzania_males.csv

Assume sampling unrelated individuals

Calculate the genotype and allele distributions of rs334 in the male adults from Tanzania. Compare with a statistical test with the data provided in the NCBI website. Draw your conclusions.

Check information online about rs1801033. Now test the association of this genetic marker with height in the male adults from Tanzania using the additive model. Do the same test now also including age as a covariate. Draw your conclusions.

Check information online about rs1799964. Now the association of this genetic marker with body temperature first alone and then including additionally malaria infection (environmental factor) and ethnic group (proxy of distinct genetic populations) using the additive model. Draw your conclusions.

Hardy-Weinberg equilibrium

Genotype	Frequency	Probability
AA	n_{AA}	π_A^2
Aa	n_{Aa}	$2\pi_A(1-\pi_A)$
aa	n_{aa}	$(1-\pi_A)^2$

Assumptions

No genotype errors

No selection/migration/mixture

Random mating

Under Multinomal sampling

$$\hat{\pi}_A = \frac{2n_{AA} + n_{Aa}}{2(n_{AA} + n_{Aa} + n_{aa})} \quad \text{(MLE)}$$

Can you prove this estimator?

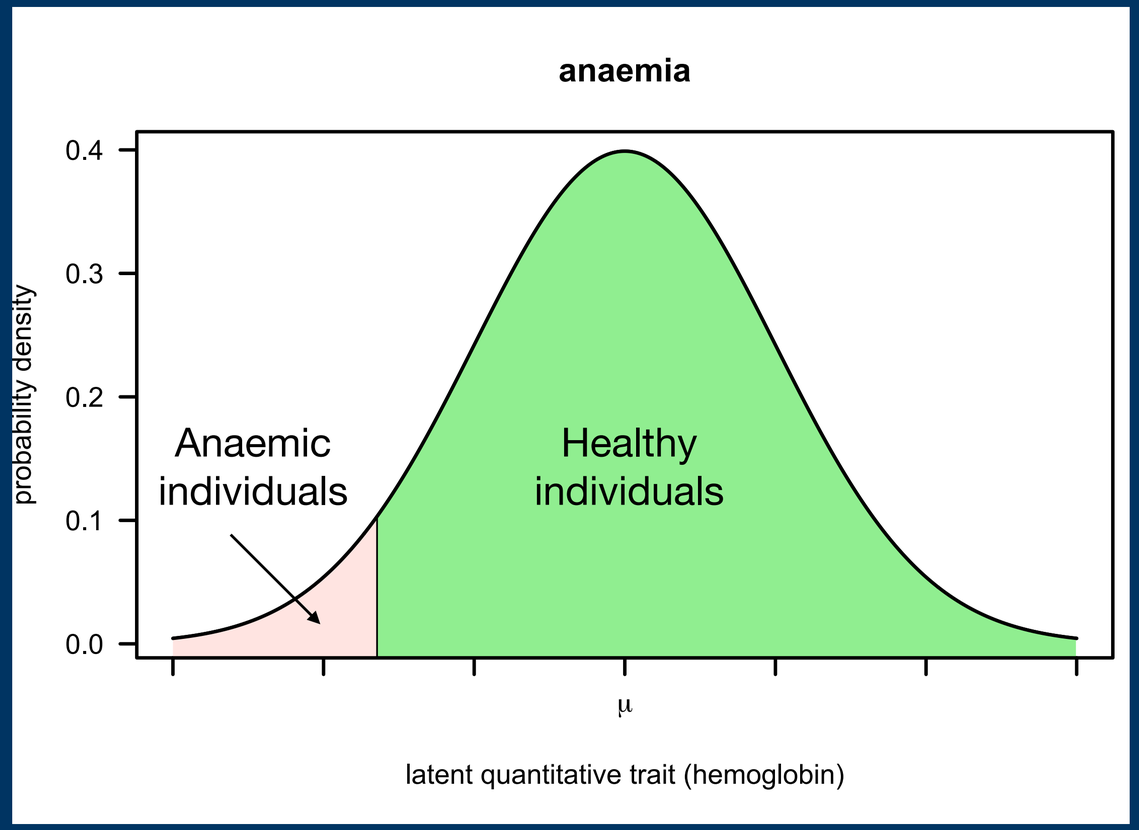
Exercise: data_Tanzania_males.csv

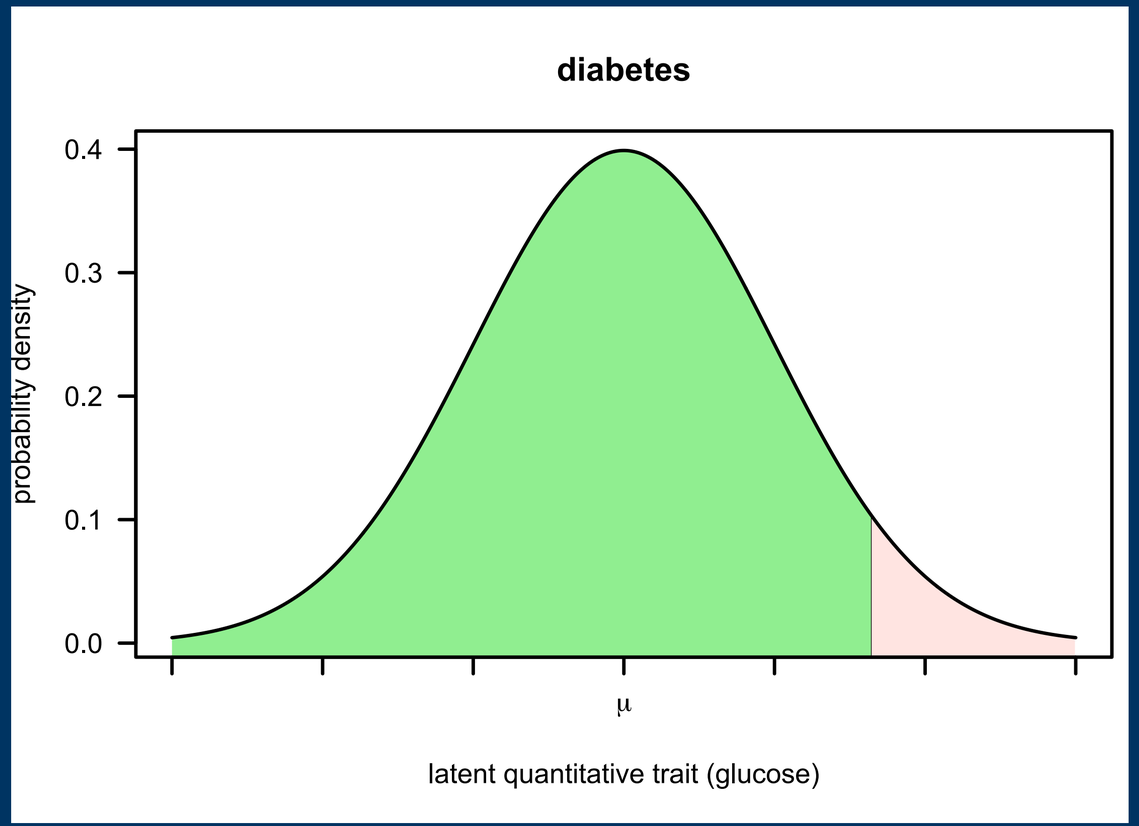
Test the Hardy-Weinberg equilibrium of the genotype distribution of rs334, rs1801033, rs1799964, rs6874639, and rs3024500 using the Pearson’s chi-square goodness-of-fit test. Draw your conclusions.

Application of Fisher's infinite models to binary traits

Anaemia	Dwarfism (?)	Diabetes
Haemoglobin level (Hb)	Height (cm)	Fasting glucose
< 130 g/L in men <120 g/L in women	< 147cm	>7.0 mmol/l >126 mg/dl
2273 SNPs possible associated with Hb	21954 SNPs possibly associated with height	405 SNPs possibly associated with fasting glucose

Liability models





Additive probit regression is a liability model

Probit regression

$$\Phi^{-1}(p_i) = \alpha_0 + \alpha_1 X_i \qquad X_i \in \{0,1,2\}$$

(single marker)

$$\Phi^{-1}(p_i) = \alpha_0 + \alpha_1 X_i + \beta_1 X^*_{1i} + \cdots + \beta_p X^*_{pi}$$

(including other non-generic covariates)

In practice, logistic regression works well (see lecture on GLM)

Probit and logit link functions are only different at the extremes

Again, testing the effect of a marker on the phenotype

$H_0 : \alpha_1 = 0$ versus $H_1 : \alpha_1 \neq 0$

Wald's Score test

$$S = \frac{\hat{\alpha}_1}{se(\hat{\alpha}_1)} \mid H_0 \rightsquigarrow Normal(\mu = 0, \sigma^2 = 1)$$

Similarly to the additive model for quantitative traits

Wilks' likelihood ratio test

$$\Lambda = (-2)\frac{L(\hat{\alpha}_0^*)}{L(\hat{\alpha}_0, \hat{\alpha}_1)} \mid H_0 \rightsquigarrow \chi^2_{(1)}$$

$L(\hat{\alpha}_0^*) =$ maximised log-likelihood of the regression model without the covariate

$L(\hat{\alpha}_0, \hat{\alpha}_1) =$ maximised log-likelihood of the regression model with the covariate

Exercise (probit additive model): data_Tanzania_males_lecture_10.csv

Assume sampling unrelated individuals

Check information online about rs6874639 and rs3024500. Test the association of this genetic marker with anaemia using the probit additive model. Do the same test including age and malaria infection as covariates. Draw your conclusions.

Repeat the above analysis but now for low haemoglobin as the binary phenotype. Draw your conclusions.

Two main types of studies

Candidate gene association studies

SNPs located in genes known to be the biological pathway leading to the trait under analysis

10-250 SNPs under analysis

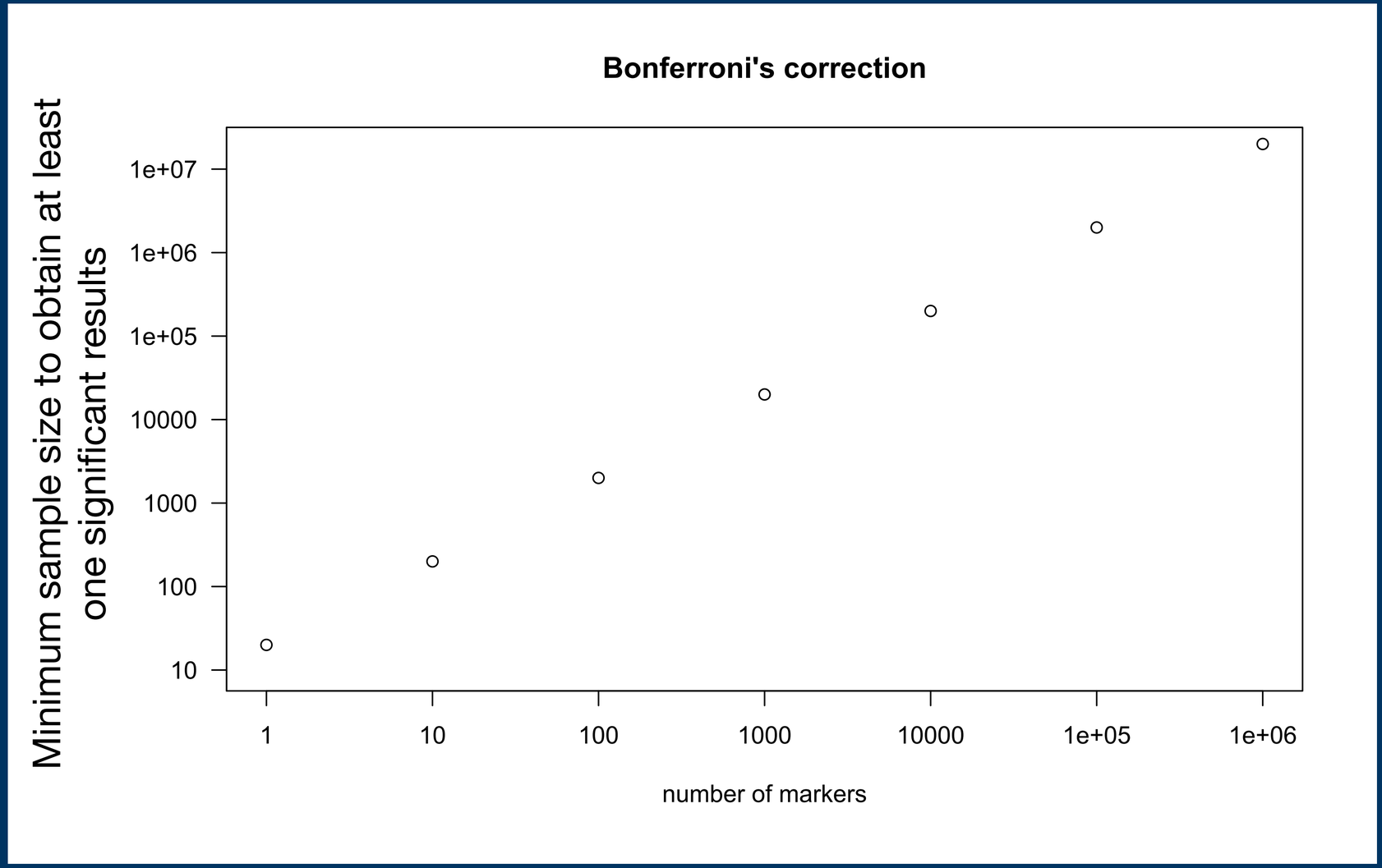
Genome-wide association studies (GWAS)

“Fishing expedition”

Millions of SNPs under analysis

What are the practical problems of these studies?

Practical problems of GWAS



Global strategy for the analysis
Candidate gene association studies

Test association between each marker and the phenotype

Additive model $\mu_{AA} = \mu + 2\mu_A,\ \mu_{Aa} = \mu + \mu_A,$ and $\mu_{aa} = \mu$

Dominance/Recessiveness model $\mu_{AA} = \mu_{Aa} = \mu + \mu_A,$ and $\mu_{aa} = \mu$

Heterosis model $\mu_{AA} = \mu_{aa} = \mu,\ \mu_{Aa} = \mu + \mu_{AA}$

General model $\mu_{AA}, \mu_{Aa},\ \mu_{aa}$

Report the lowest p-value among all the models tested

Correct significance level for multiple testing (Bonferroni/Sidak-Dunn)

Check the distribution of p-values (deviations from the uniform distribution are evidence for true associations

Global strategy for the analysis
GWAS

Test association between each marker and the phenotype

Additive model $\mu_{AA} = \mu + 2\mu_A,\ \mu_{Aa} = \mu + \mu_A,$ and $\mu_{aa} = \mu$

Report the p-value for marker tested

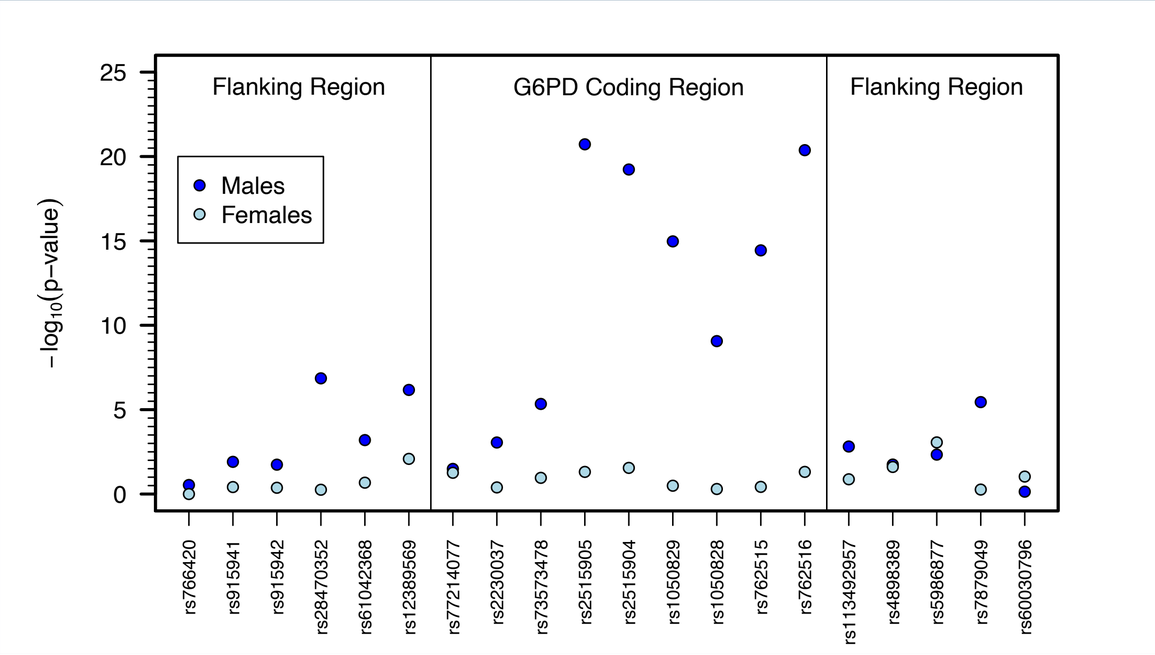
Adjust the p-values for multiple testing

Check the distribution of the p-values as for

Great deal of computational efficiency

Note: GWAS is usually analysed in the standalone PLINK software (not in the software R).

Main outputs (Candidate gene association study)



Main outputs - GWAS

Do you know how this plot is called?

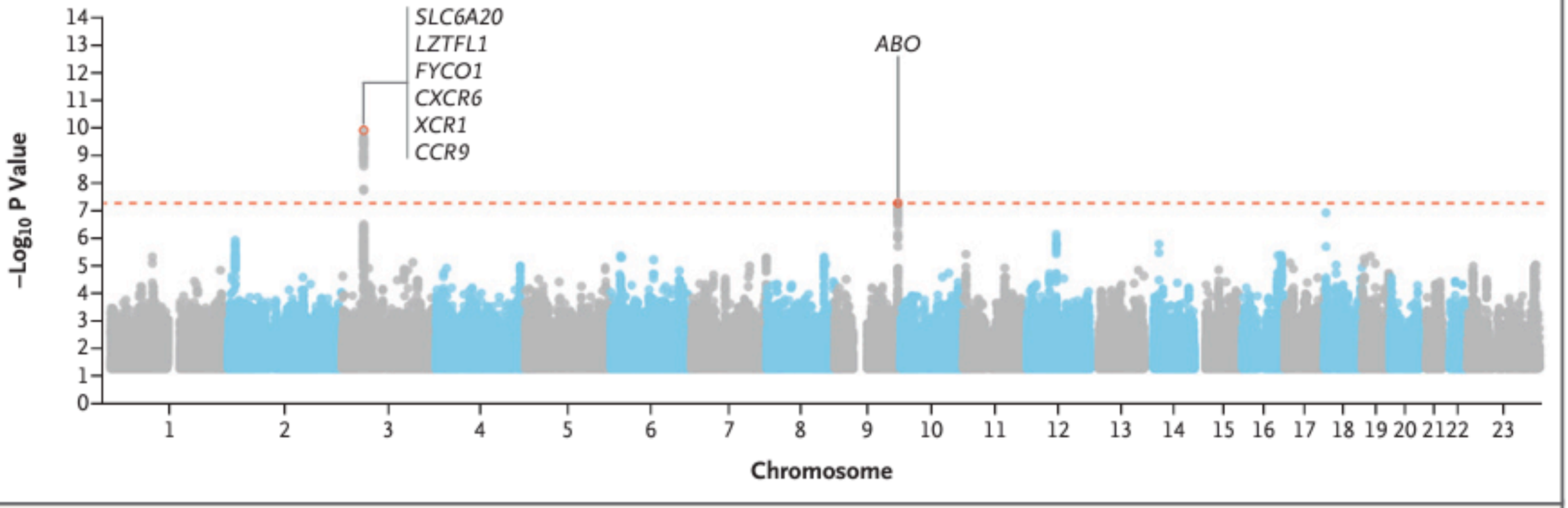


Figure 2. GWAS Summary (Manhattan) Plot of the Meta-analysis Association Statistics Highlighting Two Susceptibility Loci with Genomewide Significance for Severe Covid-19 with Respiratory Failure.

Manhattan plot

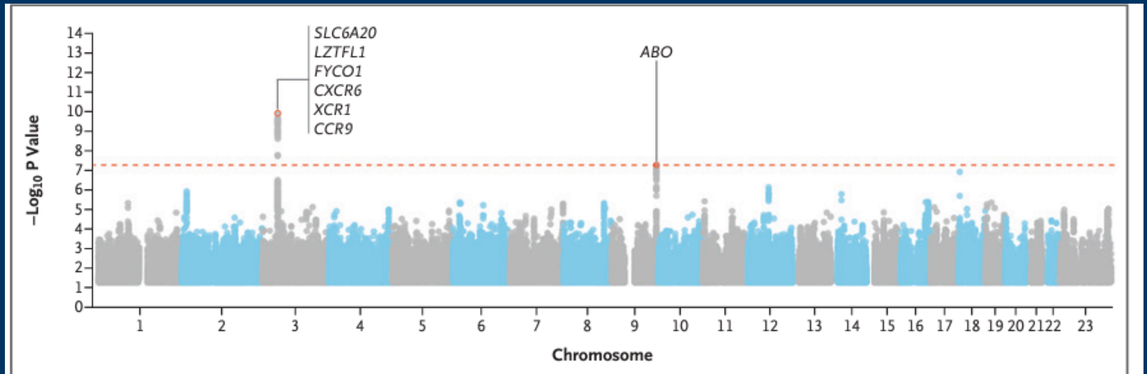
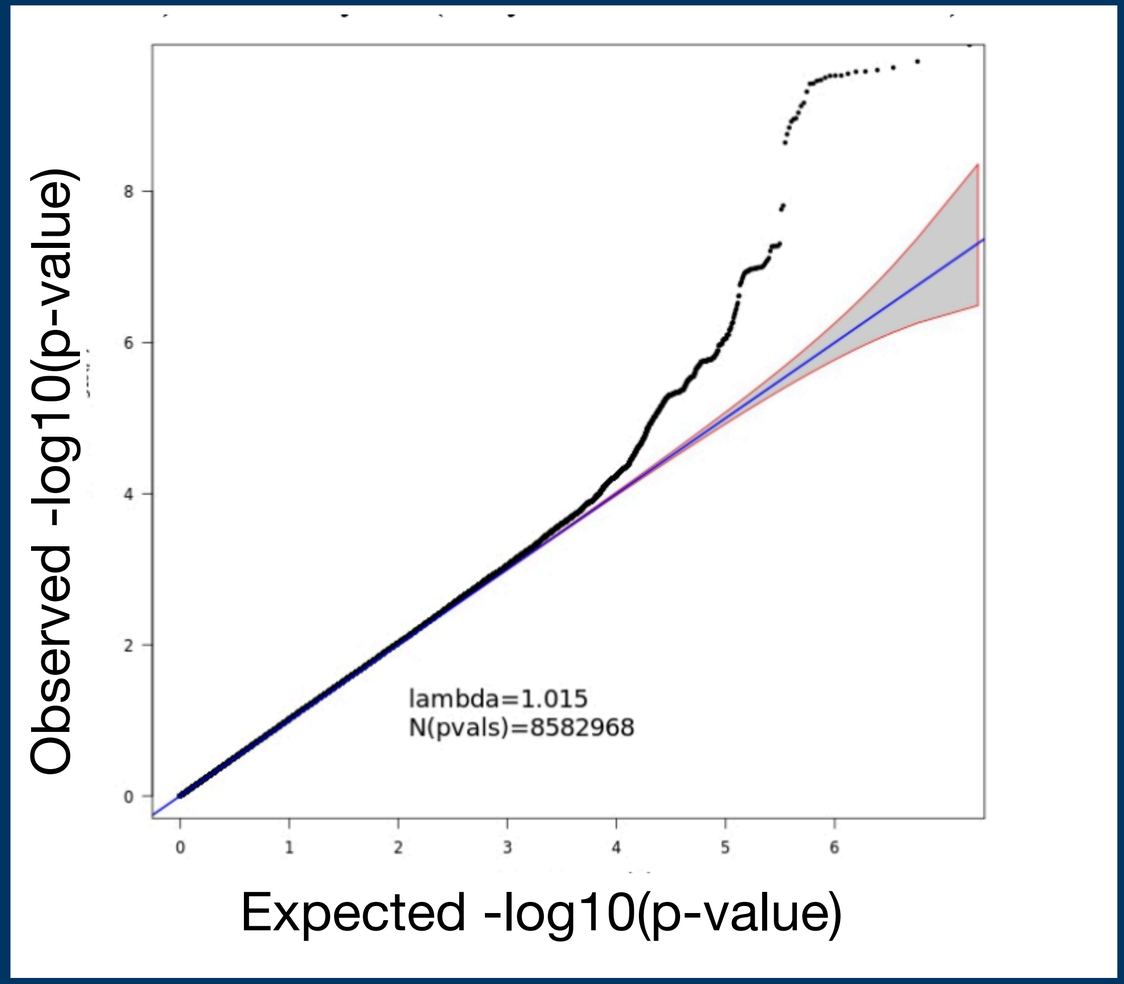


Figure 2. GWAS Summary (Manhattan) Plot of the Meta-analysis Association Statistics Highlighting Two Susceptibility Loci with Genomewide Significance for Severe Covid-19 with Respiratory Failure.

Main outputs - GWAS



Practical - data_gwas.csv

Let’s recreate the typical visual outputs from a GWAS on COVID-19 in the Portuguese population