Biostatistics

Applications in Genetics, Genomics, and other ‘omics data

Nuno Sepúlveda, 18.12.2024

Syllabus

1. General review

    a. What is Biostatistics?
    b. Population/Sample/Sample size
    c. Type of Data – quantitative and qualitative variables
    d. Common probability distributions
    e. Work example – Malaria in Tanzania

2. Applications in Medicine

    a. Construction and analysis of diagnostic tools – Binomial distribution, sensitivity, specificity, ROC curve,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 Genetics, Genomics, and other 'omics 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
    c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial

4. Other Topics

    a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
    b. Serological analysis – Gaussian (skew-normal) mixture models
    c. Advanced sample size and power calculations

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

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

Can you prove this estimator?

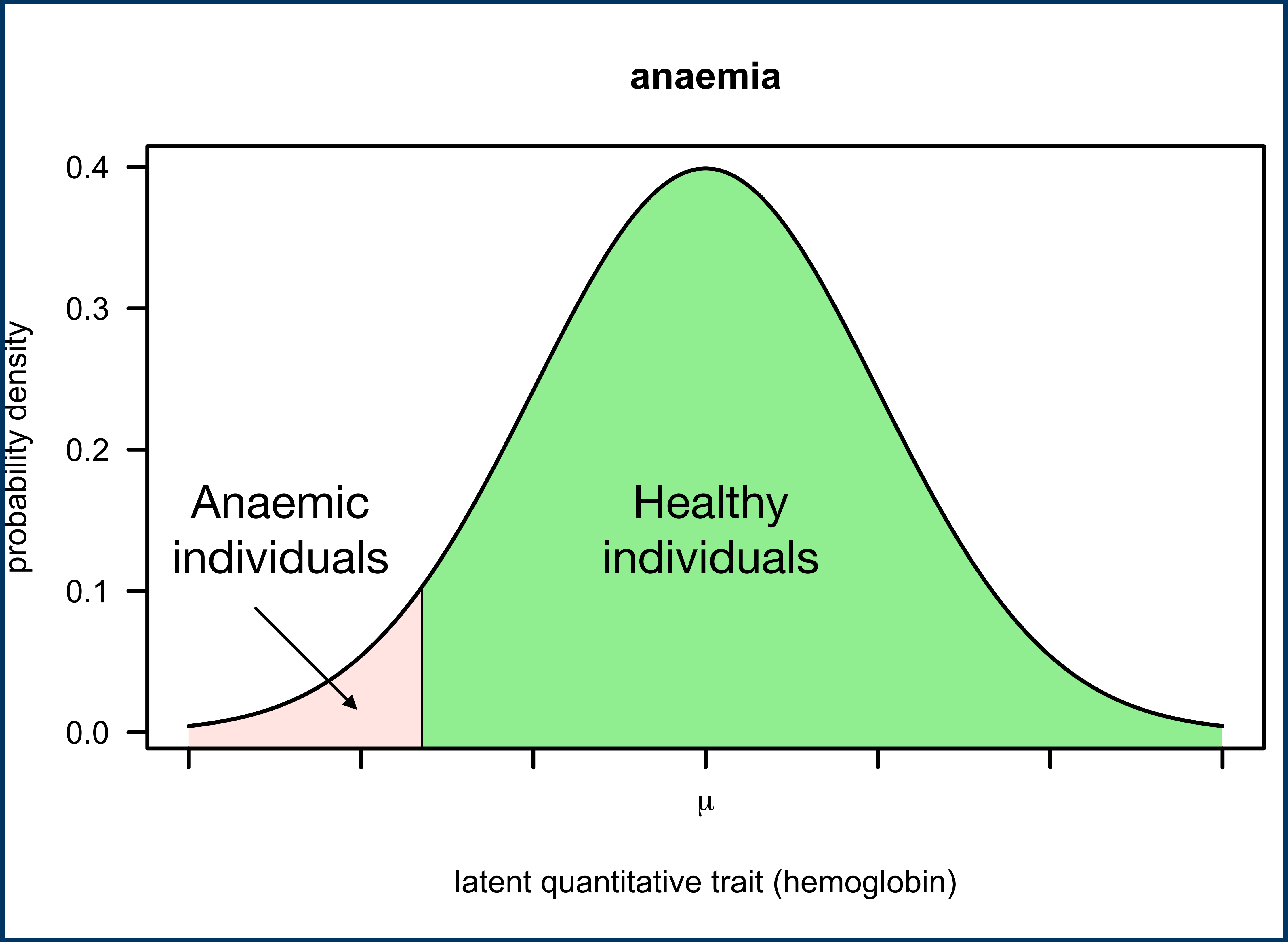
Exercise: data_Tanzania_males_lecture_10.csv

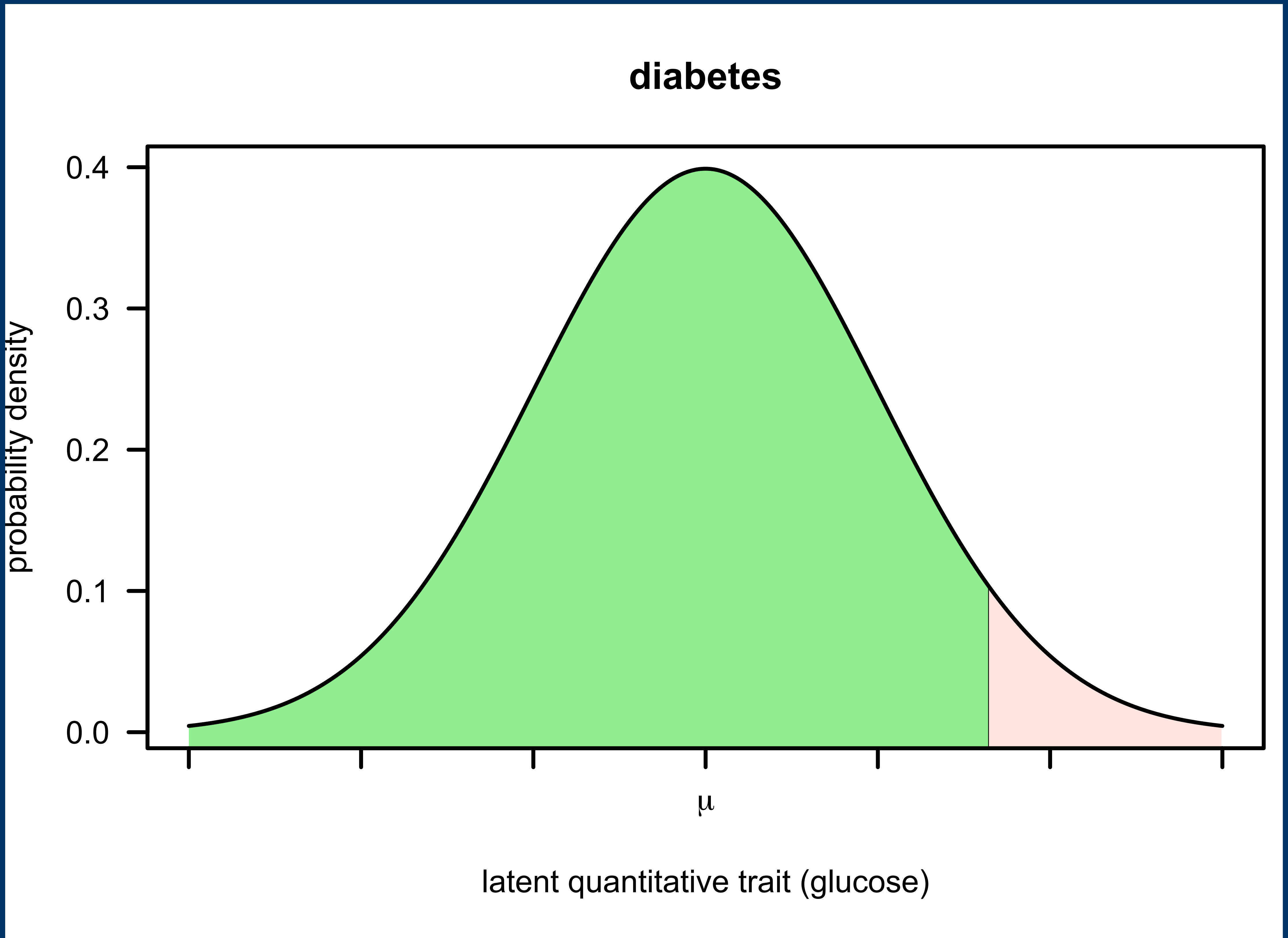
Test the Hardy-Weinberg equilibrium of the genotype distribution of rs334, rs1801033, rs1799964, rs6874639, and rs3024500 using the Pearson's chi-square goodness-of-fit test. Draw your conclusions.

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

Similarly to the additive model for quantitative traits

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

Wilks' likelihood ratio test

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

$L(\hat{\alpha}_0^*) =$ 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.

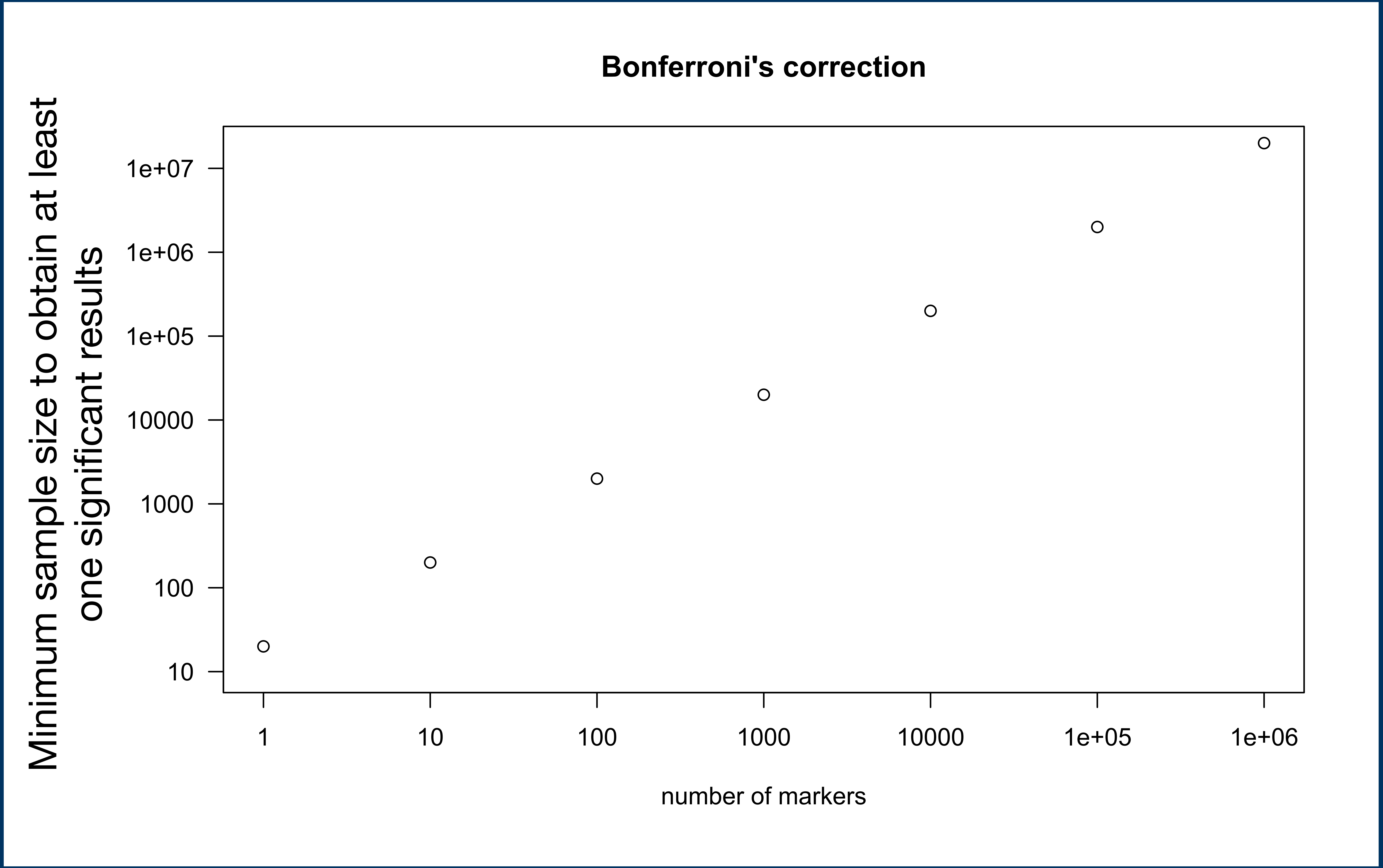
Repeat the above analysis but now for low haemoglobin as the binary phenotype. Draw your conclusions.

Two main types of studies

Candidate gene association studies	Genome-wide association studies (GWAS)
SNPs located in genes known to be the biological pathway leading to the trait under analysis	“Fishing expedition”
10-250 SNPs under analysis	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

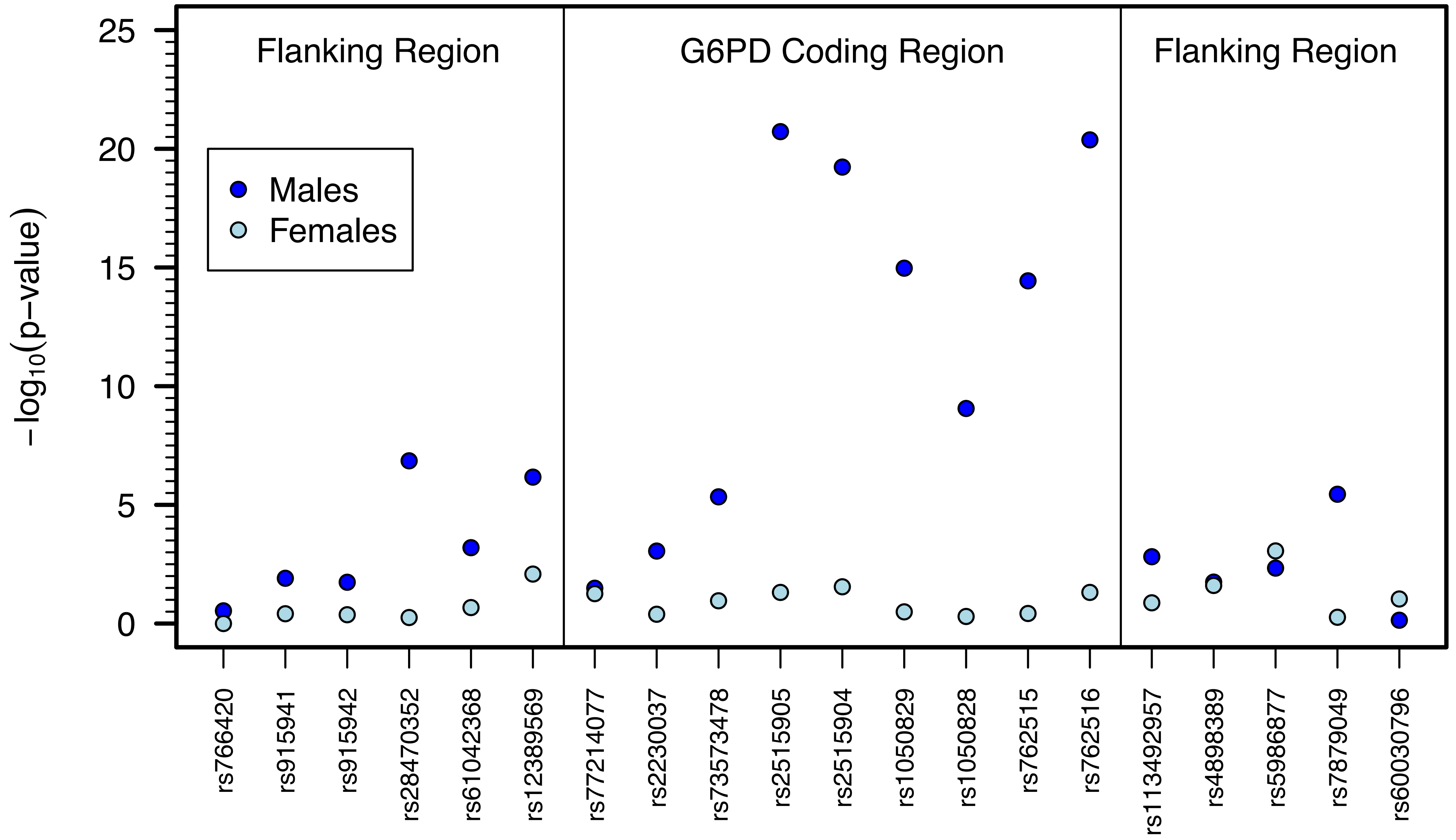
Great deal of computational efficiency

Adjust the p-values for multiple testing

Check the distribution of the p-values as for

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?

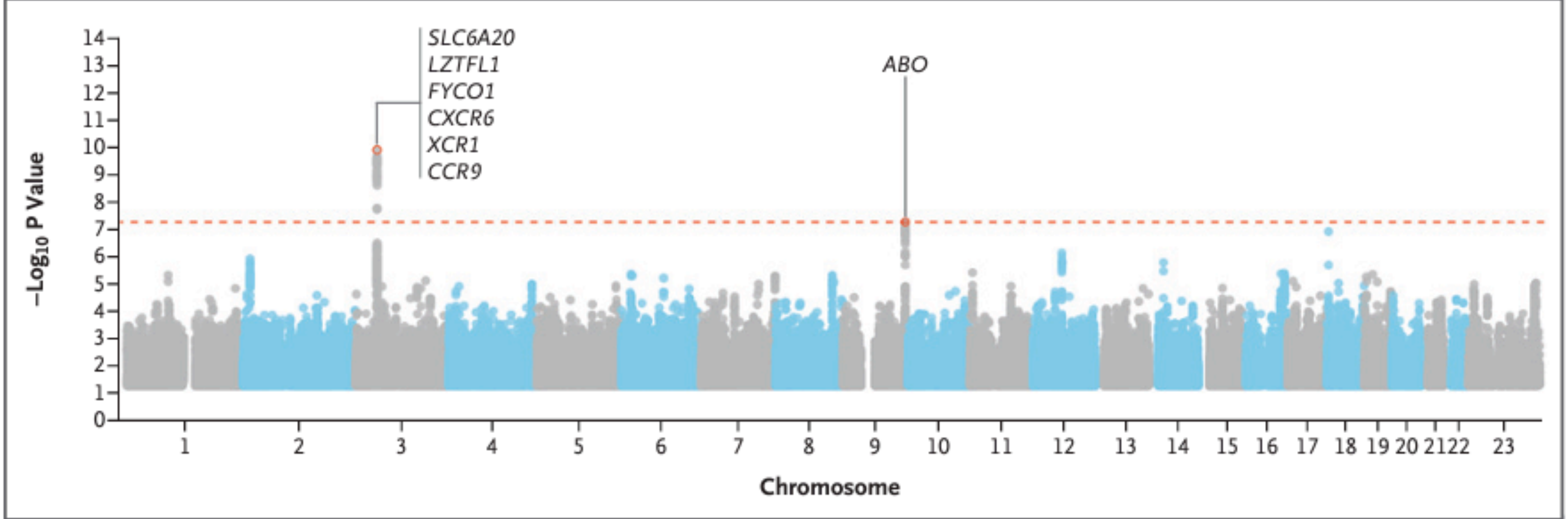


Figure 2. GWAS Summary (Manhattan) Plot of the Meta-analysis Association Statistics Highlighting Two Susceptibility Loci with Genomewide Significance for Severe Covid-19 with Respiratory Failure.

Manhattan plot

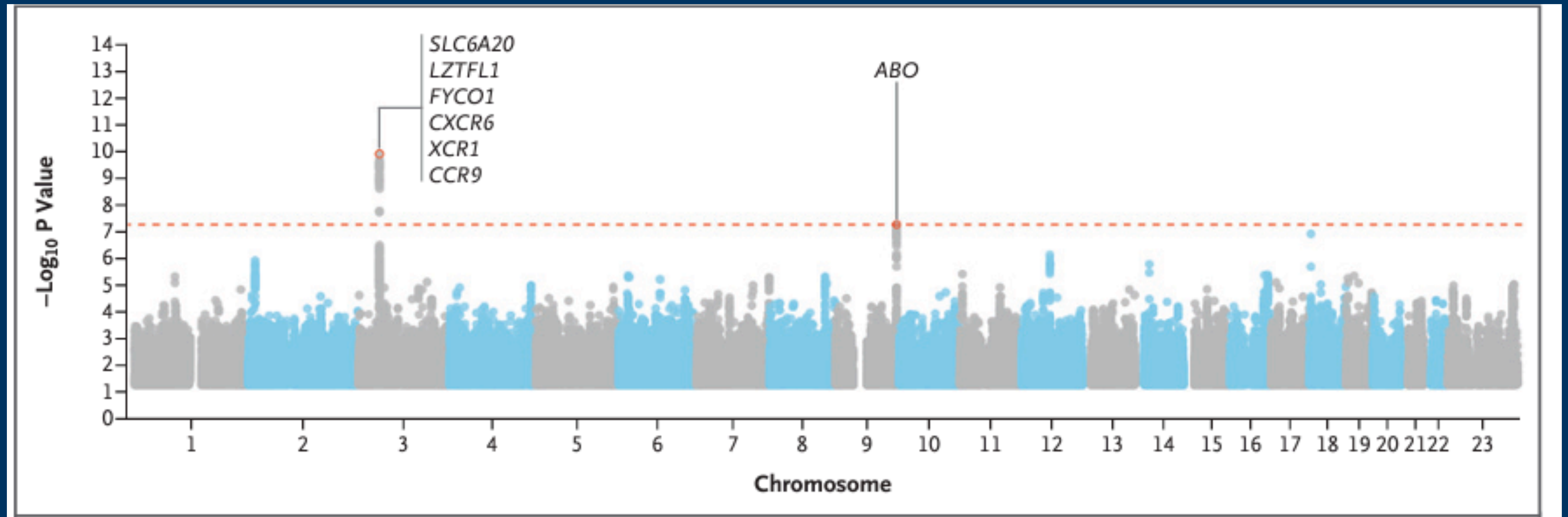
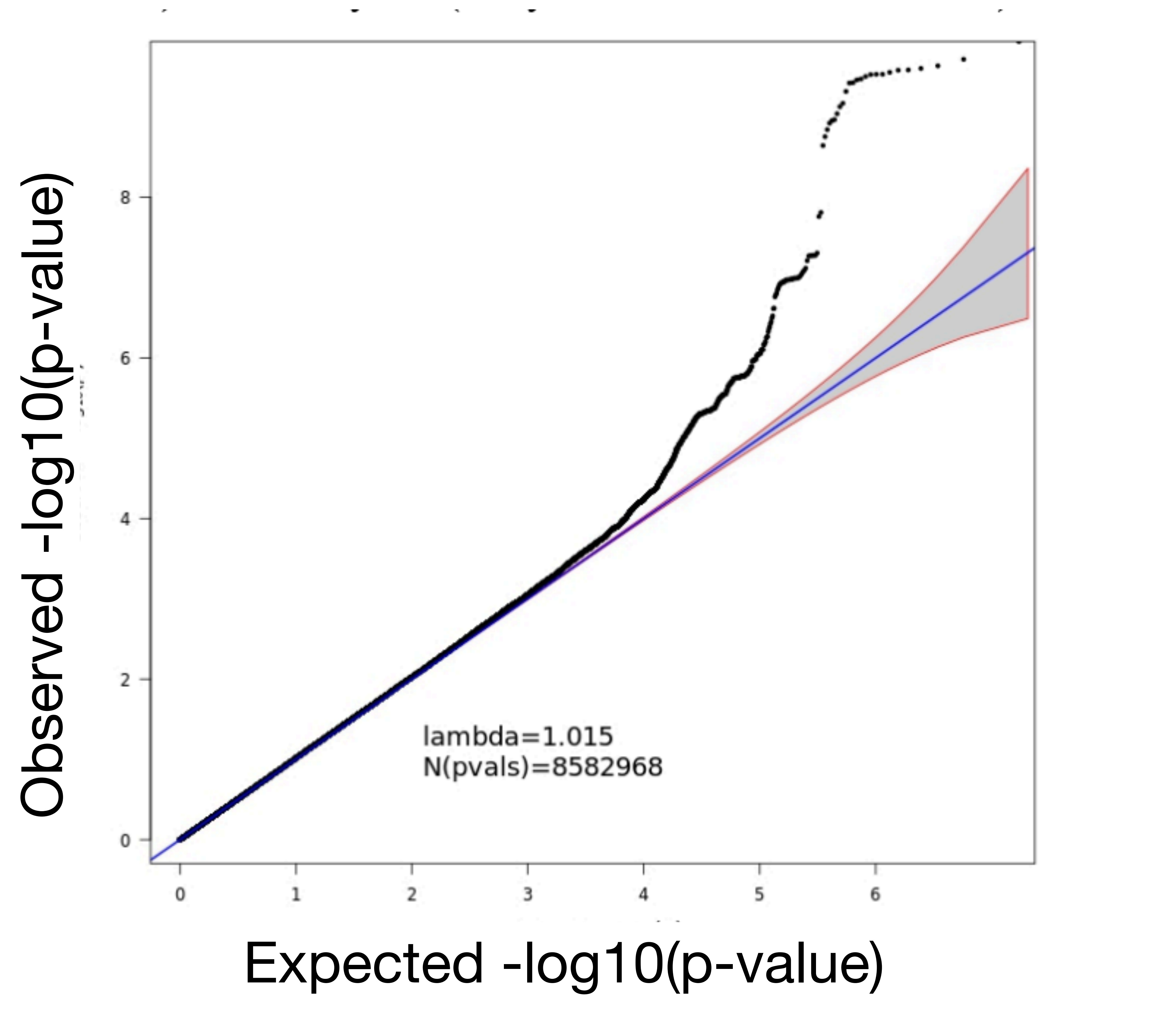


Figure 2. GWAS Summary (Manhattan) Plot of the Meta-analysis Association Statistics Highlighting Two Susceptibility Loci with Genomewide Significance for Severe Covid-19 with Respiratory Failure.

Main outputs - GWAS



Practical - data_lecture_11_gwas.csv

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

Controlling the proportion of true positive among significant results

	True Positive	**False Negative**	**Total**
Significant results	m11	m12	m1
Non-significant results	m21	m22	m2
Total	m_1	m_2	m

False discovery rate
(Benjamini-Hochberg (BH) procedure)

$$E\left[\text{True positives} \,|\, H_0 \text{ rejected}\right] = \alpha*$$

$$E\left[\frac{m11}{m11 + m21}\right] = \frac{q}{\alpha} = \alpha*$$

$$\alpha* = 0.05$$

False discovery rate
(Benjamini-Hochberg (BH) procedure)

Algorithm (under the assumption of independent tests)

1. Order all the p-values by increasing order, $p_{(1)}, \ldots, p_{(n)}$
2. For α^*, find k such as $p_{(k)} \leq \frac{k}{m}\alpha^*$
3. Reject the null hypothesis (i.e., declare discoveries) for all the genetic markers associated with p-values less $p_{(k)}$

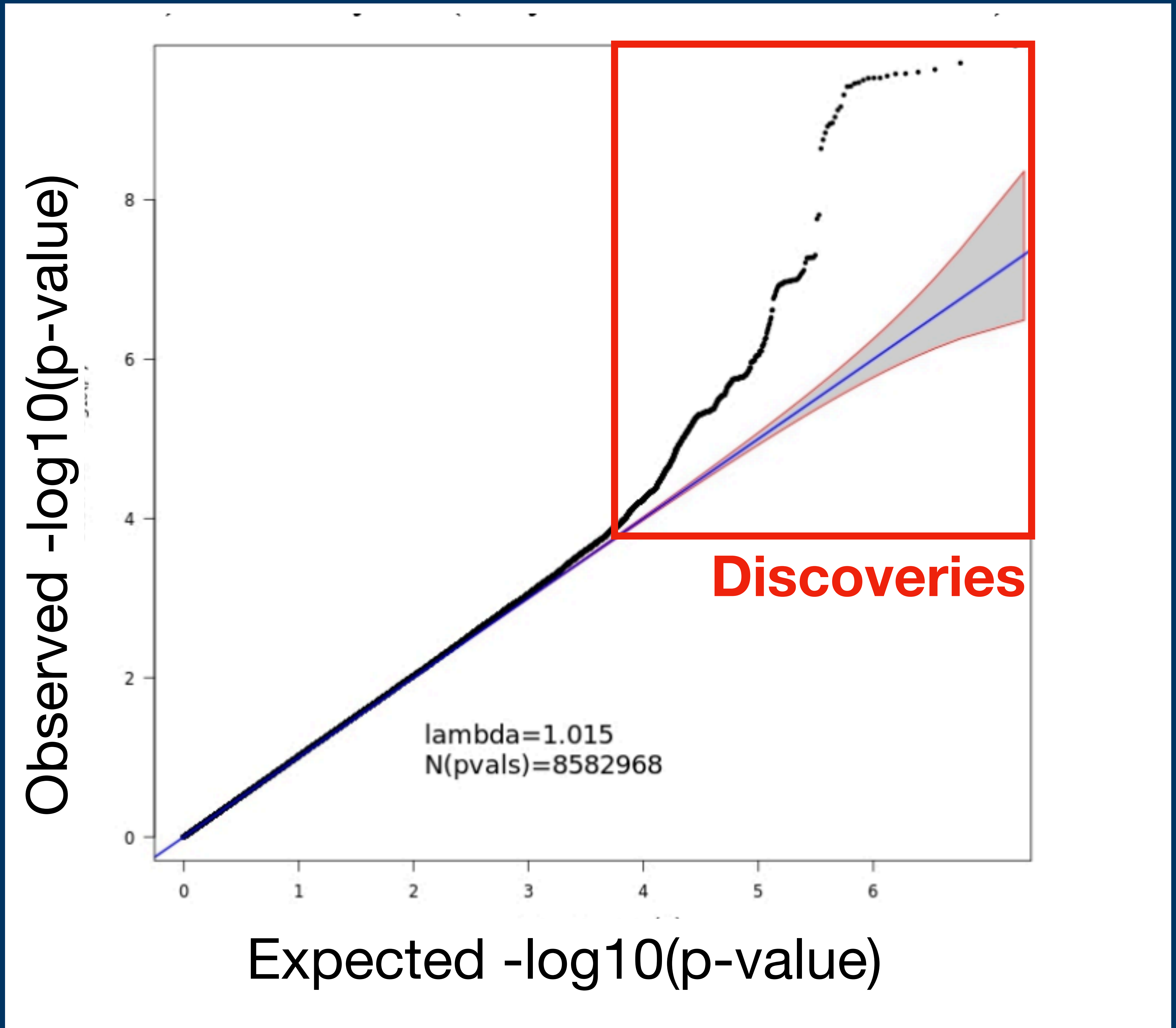
where

$p_{(1)}$ is the minimum p-value (with rank 1)

$p_{(k)}$ is the p-value with rank k

$p_{(n)}$ is the maximum p-value (with rank n)

Visual interpretation



Adjusting p-values

1. Order all the p-values by increasing order
2. Assign the ranking or position to each p-value
3. Calculate the adjusted by $p_{(i)}^{adj} = \min\left\{1, \min_{j \geq i} \frac{mp_{(j)}}{j}\right\}$

   where

   $p_{(i)}^{adj}$ is the adjusted p-value with rank i

   $p_{(j)}$ is the p-value with rank j

   m is the number of tests

Reject null hypothesis of tests whose the adjusted p-values are below the FDR

Useful variants of Benjamini-Hochberg

Benjamini-Yekutieli (BY) procedure for dependent tests

Adequate when analysing data from genetic markers in the same genetic locus

Benjamini-Krieger-Yekutieli (BKY) (improved) procedure for independent tests

Package mutoss has implementations of these and other procedures

Package MASS has implementations of the BH and BY procedures

Exercise - data_lecture_11_gwas.csv

Apply Benjamini-Hochberg and Benjamini-Krieger-Yekutieli procedures to p-values from genetic markers in chromosome 10

How many genetic markers are statistically significant according to these procedures?