Introduction to genome- and epigenome-wide association studies

Nuno Sepúlveda, 27.11.2025

Introduction

I will tell you about myself.

Introduction

You tell me about yourself.

Course content

1. Review of basic concepts in genetics
    A. Genotype/phenotype
    B. Genotype-phenotype mapping
    C. Mendelian genetics
    D. Non-mendelian genetics
2. Introduction to genome-wide association studies (GWAS)
    A. Genetic variation, genetic linkage, linkage disequilibrium
    B. Fisher’s infinitesimal/additive models for quantitative and binary traits
    C. Basic concepts in GWAS - data quality checks, statistical methodology, and reporting (Manhattan plots, qq-plots)

Course content

3. Introduction to epigenome-wide association studies (EWAS)
    A. Basics of epigenetic data - proportion of methylation per probe
    B. Quality controls
    C. Data analysis methods - simple statistical tests, linear regression, beta regression, multiple testing corrections
    D. Main outputs - Manhattan plots, qq-plots

Course materials

https://github.com/immune-stats/gwas-ewas-course-2025/

Recommendation:

Take notes, the slides are not sufficient.

1. Introduction to Genetics

What is genetics?

Science that studies the laws of heredity.

Which the hereditary factors play a role in these laws?
Which (biological) mechanisms are associated with these factors?
How these mechanisms evolve over time and in different populations?

Medical genetics

Population genetics and evolution

Experimental genetics

The main objective of genetics

Understand the causality between genetic variation and phenotypic variation

Genetic variation → Phenotypic variation

Medical genetics

Genetic variation → Non-disease versus Disease

Gene, genotype and phenotype

Gene
Heredity factor or unit
Stretch of DNA that encodes a protein

Allele (or alellemorph)
A variation of this heredity factor

Genotype
The allelic composition of a given gene in an individual

Phenotype
Observable biological characteristic

Mendelian genetics

Single gene, two alleles, single binary trait/phenotype

Complete penetrance

Rules of dominance/recessiveness

Penetrance - probability of expressing the phenotype given the genotype

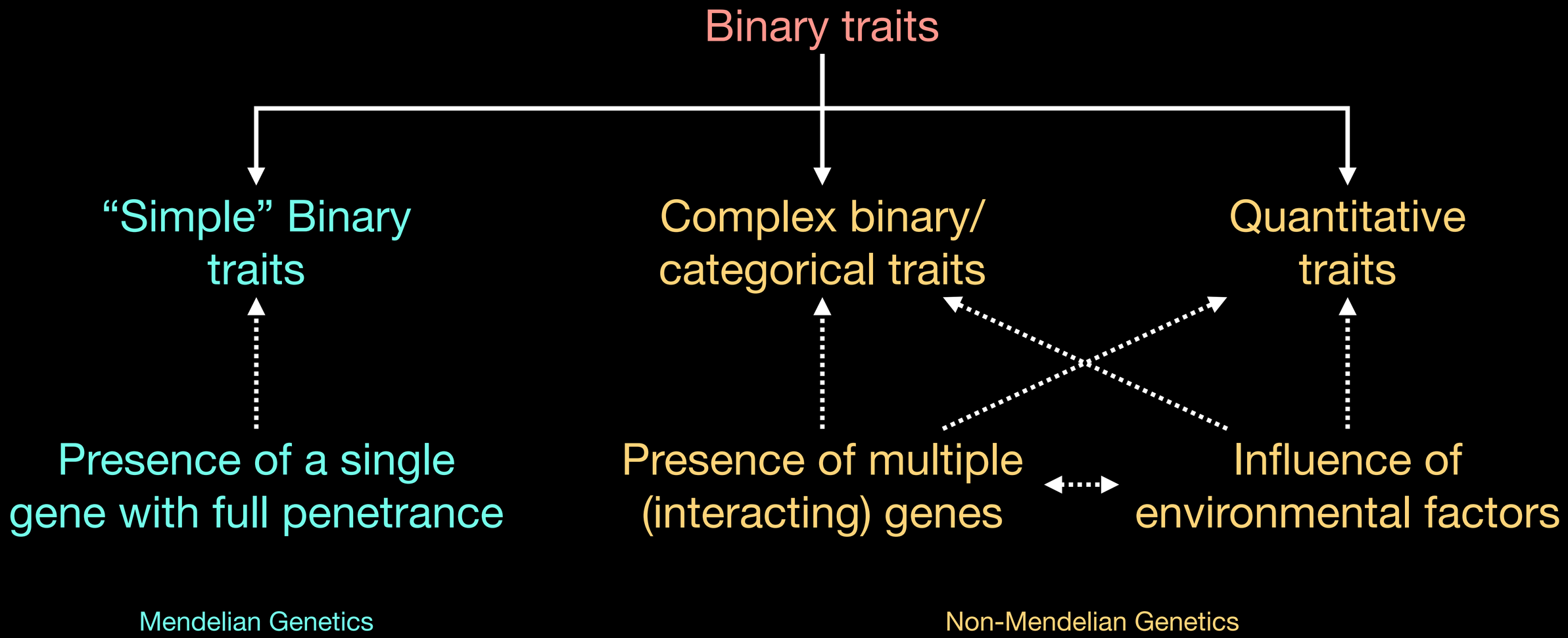
Mendelian genetics

Genotype	Phenotype
AA	***A***
Aa	***A***
aa	***a***

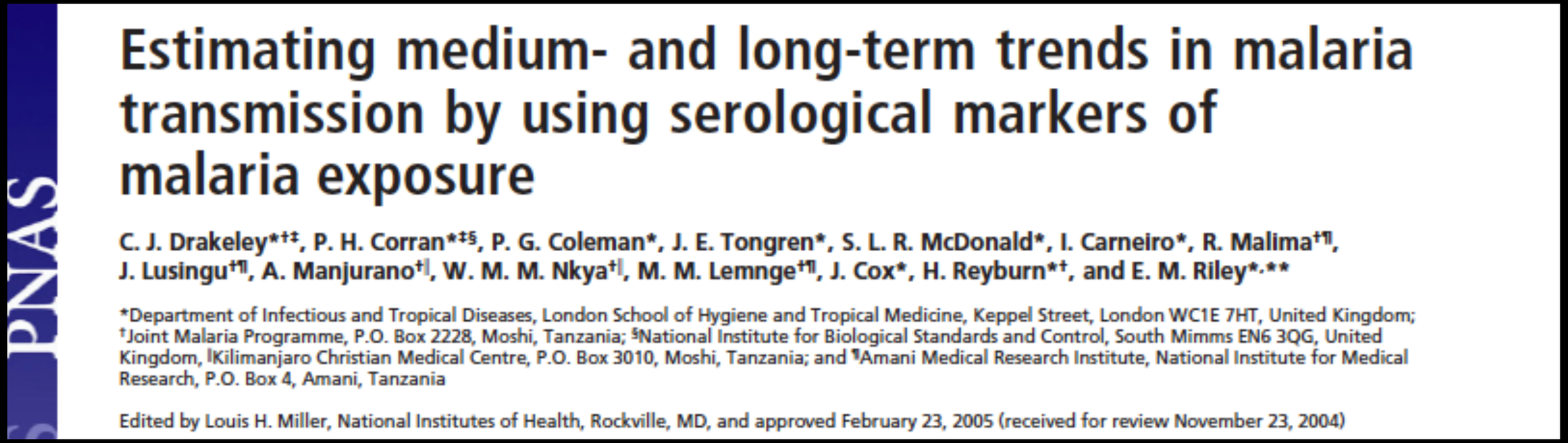
Allele A is dominant over allele a

Allele a is recessive over allele A

Mendelian genetics versus Non-mendelian genetics



Working exercise: data_tanzania.csv



PNAS

Estimating medium- and long-term trends in malaria transmission by using serological markers of malaria exposure

C. J. Drakeley*†‡, P. H. Corran*‡§, P. G. Coleman*, J. E. Tongren*, S. L. R. McDonald*, I. Carneiro*, R. Malima†¶, J. Lusingu†¶, A. Manjurano†‖, W. M. M. Nkya†‖, M. M. Lemnge†¶, J. Cox*, H. Reyburn*†, and E. M. Riley*,**

*Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom; †Joint Malaria Programme, P.O. Box 2228, Moshi, Tanzania; §National Institute for Biological Standards and Control, South Mimms EN6 3QG, United Kingdom, ‖Kilimanjaro Christian Medical Centre, P.O. Box 3010, Moshi, Tanzania; and ¶Amani Medical Research Institute, National Institute for Medical Research, P.O. Box 4, Amani, Tanzania

Edited by Louis H. Miller, National Institutes of Health, Rockville, MD, and approved February 23, 2005 (received for review November 23, 2004)

Working exercise: Tanzania dataset

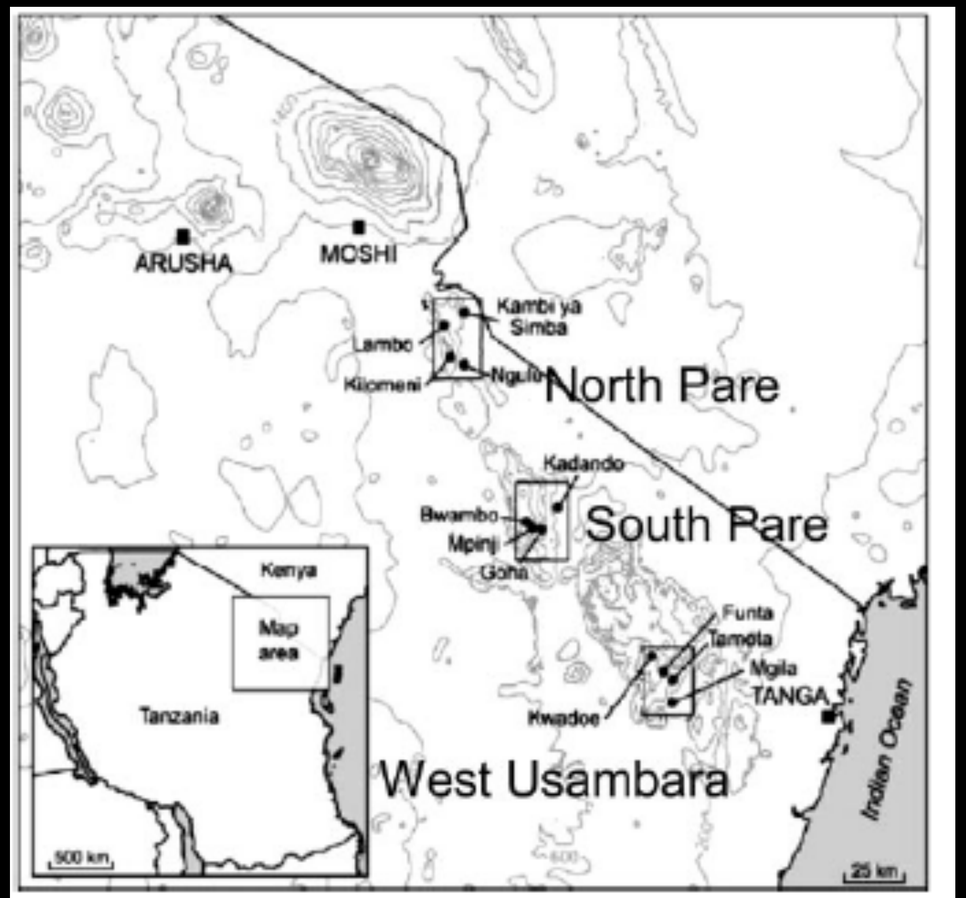


Fig. 1. Map of the study area showing the three altitude transects and 12 study villages.

Cross-sectional study

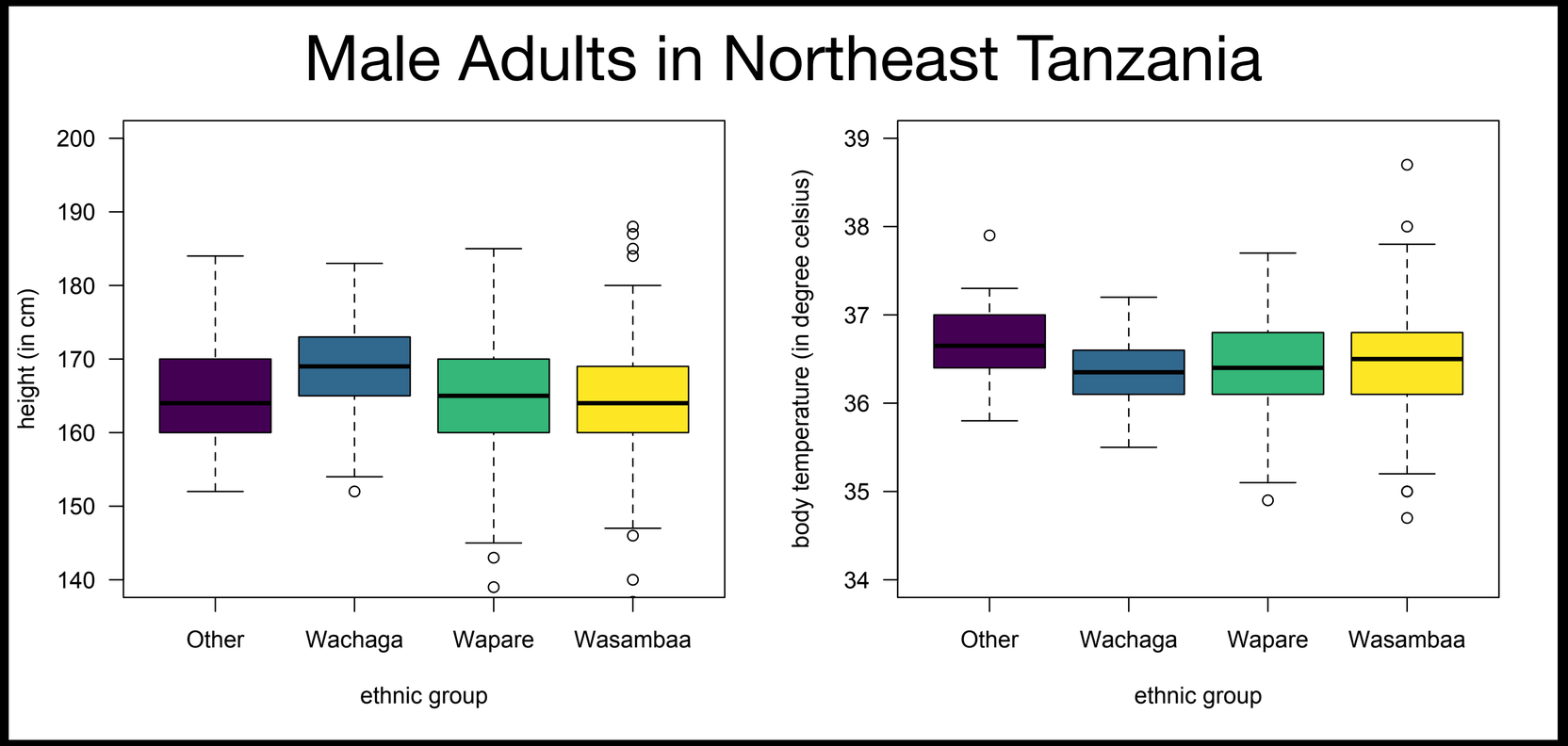
Stratified sampling (three age groups: 0-4, 5-14, 15-45

24 villages in 6 altitude transects

~8146 individuals (6 months-45 years old)

Gender and age distributions matched across villages

Working exercise: Tanzania dataset



Can you reproduce the plots in R?

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

Basic question

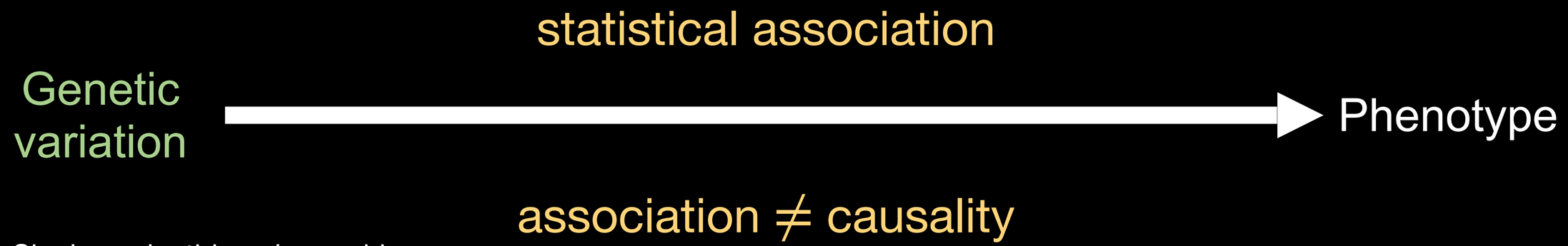
What are the genes involved and what is their action on the phenotypic expression?

Genetic association studies

- Candidate-gene association studies
- Genome-wide association studies
- Pedigree/family/Twin/Nuclear family studies

2. Introduction to GWAS

Genetic association studies



Single nucleotide polymorphisms

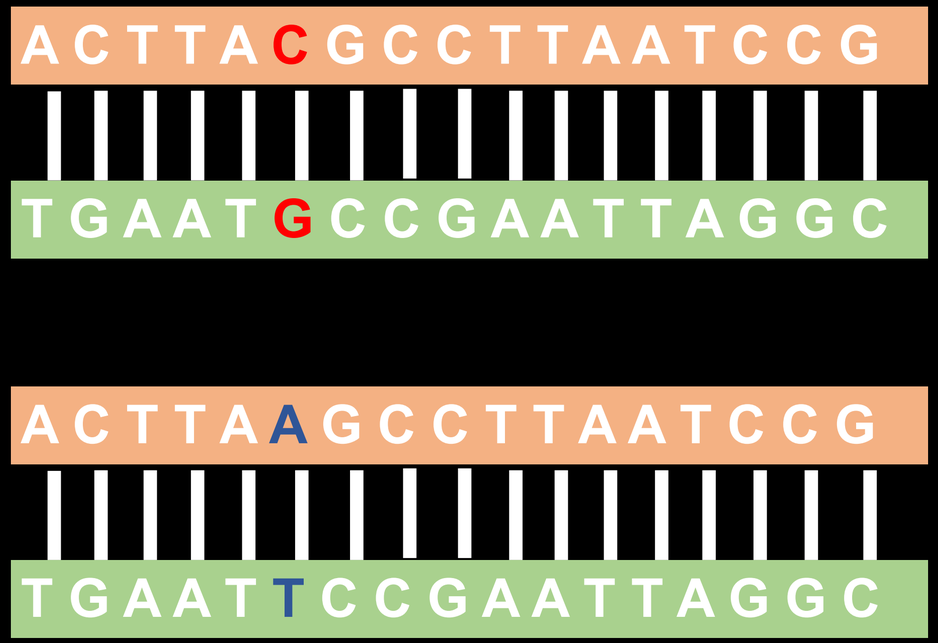
Structural variants (>1 kbp)

- Copy number variation
- Indels
- Inversions
- Etc

Genetic variation - Single nucleotide polymorphisms

Known by the acronym SNP

Rs_number (e.g., rs334)



SNP catalogue (NCBI)

Input and Output Counts for ALFA Release 4

Input	Count
Studies	105
Subjects	408,709
Genotypes	5,897,518,457,092

Output	Count
Total RefSNPs	904,623,795
Exist in dbSNP [157]	904,097,097
Novel (ALFA R4)	526,698

Release May 15, 2025

Reference Human Genome (very important!)

GRCh38.p14

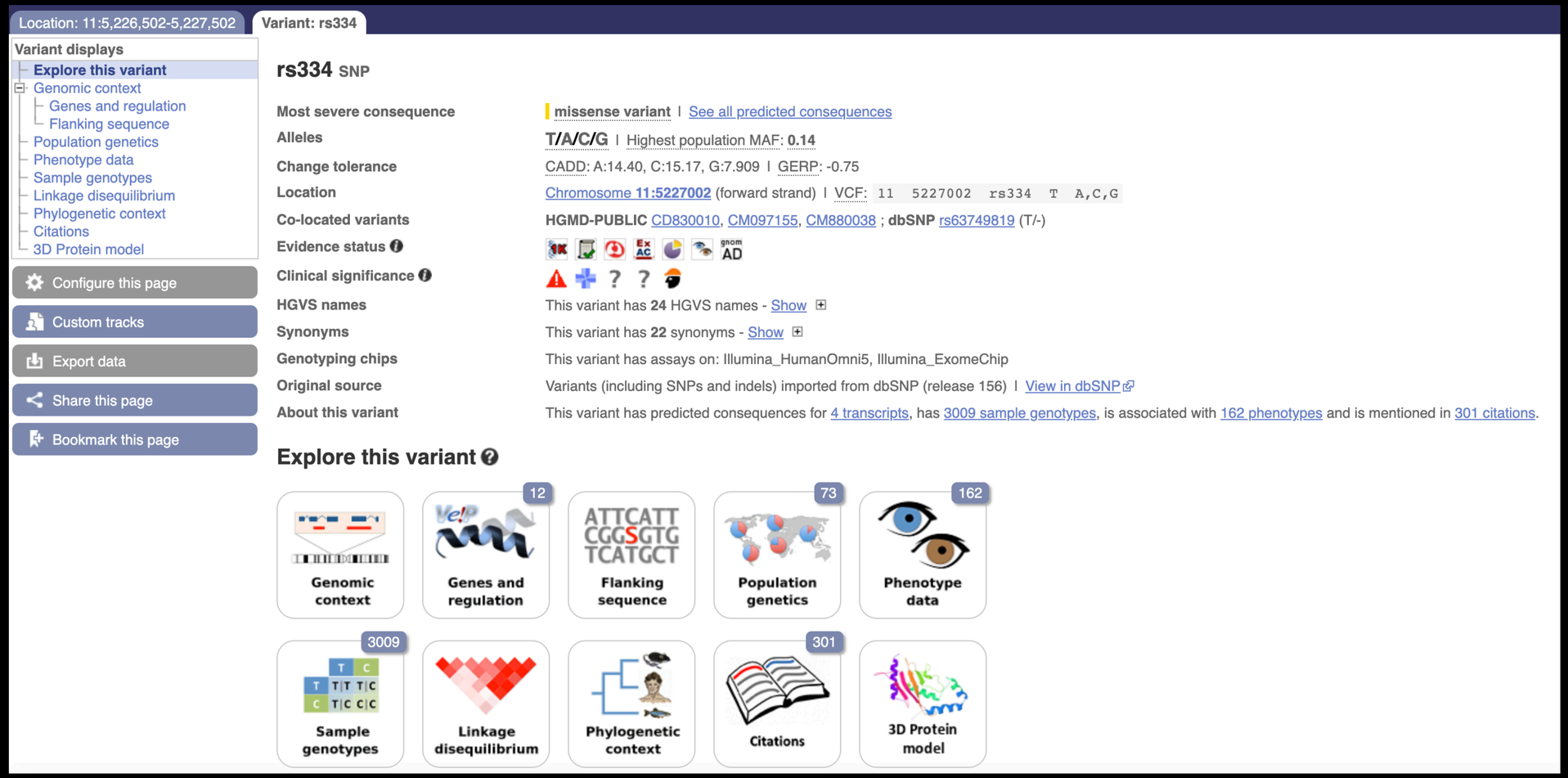
https://www.ncbi.nlm.nih.gov/snp/

Finding information about a specific SNP

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

https://useast.ensembl.org/index.html

Example: information about rs334

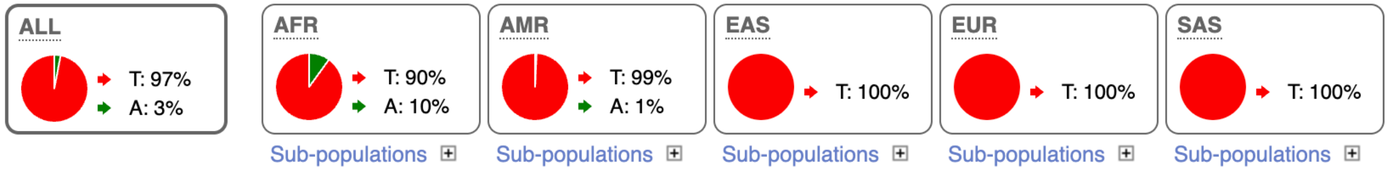


https://useast.ensembl.org/index.html

Example: information about rs334

Population genetics

1000 Genomes Project Phase 3 allele frequencies



Jump to: 1000 Genomes Project Phase 3 (32) | gnomAD exomes v4.1 (10) | gnomAD genomes v4.1 (11) | NCBI ALFA (12) | TOPMed (1) | NHLBI Exome Sequencing Project (2) | Gambian Genome Variation Project (5)

1000 Genomes Project Phase 3 (32)

Show All entries | Show/hide columns | Filter

Population	Allele: frequency (count)		Genotype: frequency (count)		Genotypes
ALL	**T: 0.973 (4871)**	**A: 0.027 (137)**	**TIT: 0.945 (2367)**	**AIT: 0.055 (137)**	**Show**
AFR	**T: 0.900 (1190)**	**A: 0.100 (132)**	**TIT: 0.800 (529)**	**AIT: 0.200 (132)**	**Show**
ACB	T: 0.953 (183)	A: 0.047 (9)	TIT: 0.906 (87)	AIT: 0.094 (9)	Show
ASW	T: 0.984 (120)	A: 0.016 (2)	TIT: 0.967 (59)	AIT: 0.033 (2)	Show
ESN	T: 0.879 (174)	A: 0.121 (24)	TIT: 0.758 (75)	AIT: 0.242 (24)	Show
GWD	T: 0.885 (200)	A: 0.115 (26)	TIT: 0.770 (87)	AIT: 0.230 (26)	Show
LWK	T: 0.899 (178)	A: 0.101 (20)	TIT: 0.798 (79)	AIT: 0.202 (20)	Show
MSL	T: 0.876 (149)	A: 0.124 (21)	TIT: 0.753 (64)	AIT: 0.247 (21)	Show
YRI	T: 0.861 (186)	A: 0.139 (30)	TIT: 0.722 (78)	AIT: 0.278 (30)	Show

https://useast.ensembl.org/index.html

Genetic variation - Structural variants

nstd186

Study Page: nstd186 (NCBI Curated Common Structural Variants).

Region Type	Region count	Call Type	Call count	Example Variant Call	Query
copy number variation	63566	copy number gain	430	nssv16206451	"nstd186"[study] AND "copy number gain"[variant_type]
		copy number loss	335	nssv16201745	"nstd186"[study] AND "copy number loss"[variant_type]
		copy number variation	5730	nssv17968517	"nstd186"[study] AND "copy number variation"[variant_type]
		deletion	60280	nssv17968567	"nstd186"[study] AND "deletion"[variant_type]
		duplication	12326	nssv17968562	"nstd186"[study] AND "duplication"[variant_type]
insertion	12292	insertion	12806	nssv17968565	"nstd186"[study] AND "insertion"[variant_type]
mobile element deletion	3749	alu deletion	4874	nssv16888762	"nstd186"[study] AND "alu deletion"[variant_type]
		herv deletion	1	nssv16193961	"nstd186"[study] AND "herv deletion"[variant_type]
		line1 deletion	585	nssv16888760	"nstd186"[study] AND "line1 deletion"[variant_type]
		mobile element deletion	43	nssv17968248	"nstd186"[study] AND "mobile element deletion"[variant_type]
		sva deletion	173	nssv16888721	"nstd186"[study] AND "sva deletion"[variant_type]
mobile element insertion	13327	alu insertion	11380	nssv17968566	"nstd186"[study] AND "alu insertion"[variant_type]
		line1 insertion	1441	nssv17968250	"nstd186"[study] AND "line1 insertion"[variant_type]
		mobile element insertion	57	nssv17968337	"nstd186"[study] AND "mobile element insertion"[variant_type]
		sva insertion	758	nssv17968334	"nstd186"[study] AND "sva insertion"[variant_type]
Total	92934		111219		

https://www.ncbi.nlm.nih.gov/dbvar/

Genetic association studies

Single nucleotide polymorphisms

Why are they chosen to perform this type of studies?

Source of statistical association

Genetic linkage

Association arising at the level of the individual

Linkage disequilibrium

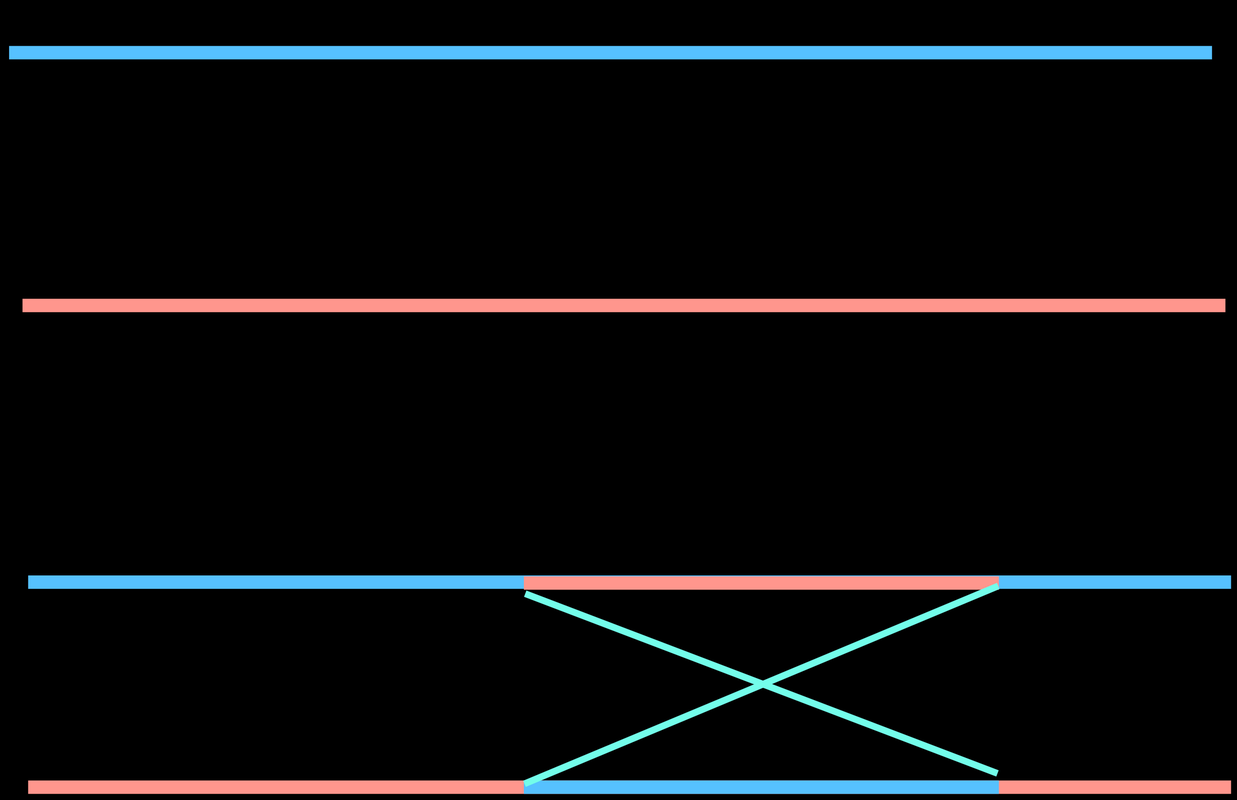
Association arising at the level of the population (where mutation, random drift, migration, and selection occurs)

Genetic linkage and recombination

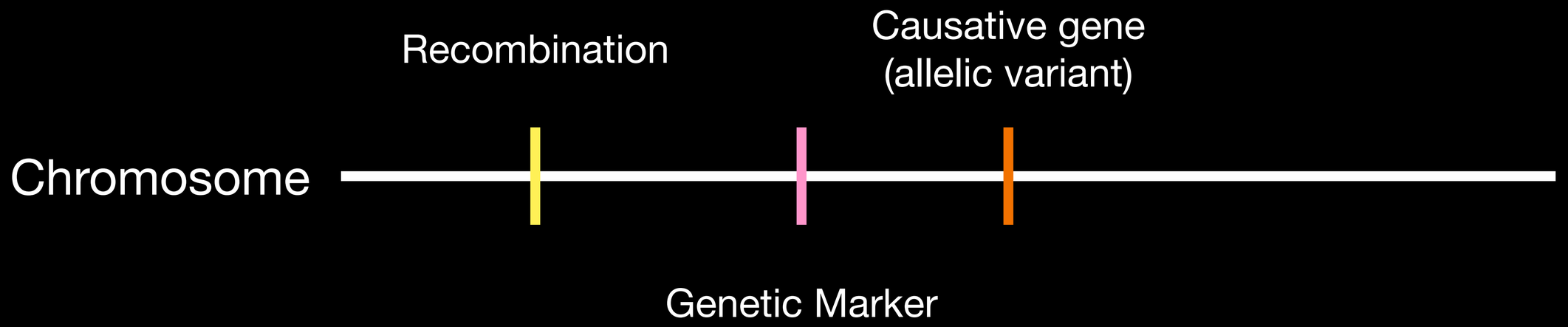
Paternal Chromosome

Maternal Chromosome

During formation of gametes (sperm/egg cells)



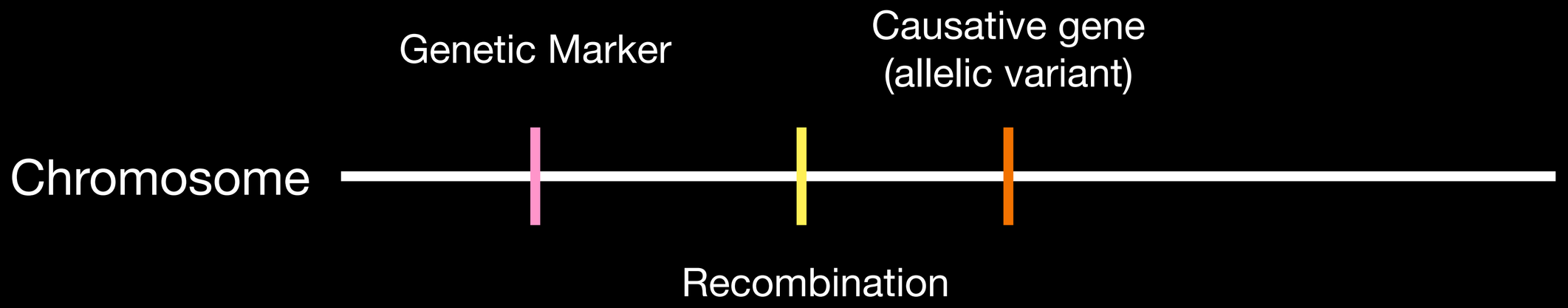
Complete linkage



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 data fully captures the true statistical association between the causative gene and the phenotype.

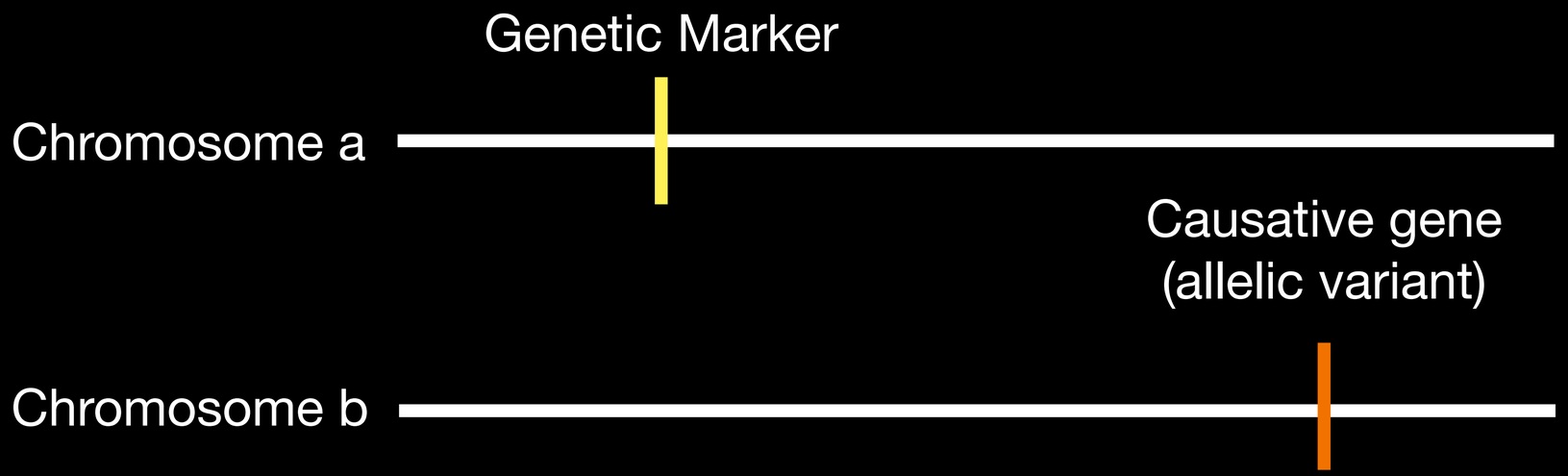
Incomplete linkage



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 data partially captures the true statistical association between the causative gene and the phenotype.

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 causative gene/phenotype.

Linkage disequilibrium

It is the degree of statistical association between two loci when collecting a sample from a given population.

Locus A/Locus B	**Allele B**	**Allele b**
Allele A	n_{AB}	n_{Ab}
Allele a	n_{aB}	n_{ab}

It reflects not only the existent genetic linkage between the loci but also their joint selection at the population.

Linkage disequilibrium

Statistical definition

$$D = p_{AB} - p_A p_B$$

$$r = \frac{D}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$

Estimation

$$D = \hat{p}_{AB} - \hat{p}_A \hat{p}_B = \frac{n_{AB}}{n} - \frac{n_{AB} + n_{Ab}}{n} \times \frac{n_{AB} + n_{aB}}{n}$$

Linkage disequilibrium

It is the degree of statistical association between the genetic markers when collecting a sample from a given population.

M1/M2	BB	Bb	Bb
AA			
Aa			
Aa			

Practical implications of genetic linkage/linkage disequilibrium

It allows the mapping of the causal genetic variation responsible for the expression of the phenotype using statistical techniques of association.

The existence of markers in complete linkage allows:
- to reduce the number of genetic markers under analysis (important for reducing the problem of multiple testing):
- to perform accurate data imputation using the existing (strong correlation structure).

But how to perform the statistical analysis of a given set of genetic markers?

Fisher's infinitesimal (polygenic) model

XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

CONTENTS.

Fisher's infinitesimal (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.

$$Y_i = \beta_0 + \sum_{j=1}^{\infty} \beta_j X_{ij} \rightsquigarrow ?$$

where

$X_{ij} =$ number of a given allele in the genotype of gene j in individual i

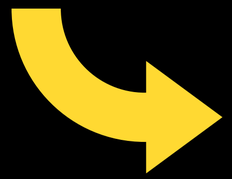
$\beta_0 =$ overall average of environmental factors and alleles located at other genes

$\beta_j =$ phenotype effect of adding an allele to the genotype at gene j

Theoretical implications of the Fisher's model

Common variant common disease hypothesis

Common diseases are caused by many variants with small effects towards the disease (e.g., cardiovascular diseases, type II diabetes, schizophrenia)



Rare variant rare disease (Mendelian diseases)

Rare diseases are caused by a single variant with a very strong (biological) effect (e.g., IPEX - Immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome)

Statistical implications of the Fisher's 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 can be 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
    - What about the sample size?

Additive model for a single genetic marker for diploid organisms

Y_i = random variable for the quantitative trait in individual i

$$Y_i \mid \mu_i, \sigma \rightsquigarrow \text{Normal}(\mu_i, \sigma^2)$$

X_{ij} = number of a given allele in the genotype of individual i for the genetic marker j

$$X_{ij} \in \{'aa', 'aA', 'AA'\} \longrightarrow X_{ij} \in \{0,1,2\}$$

$$\mu_i = E[Y_i] = \beta_0 + \beta_j X_{ij}$$

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

Additive model as simple linear regression

$$\mu_i = E[Y_i] = \beta_0 + \beta_j X_i$$

What are the assumptions?

What if there are two individuals from the same family? Is this model applicable?

Testing the association between a given marker and the phenotype

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

Wald's Score test (for large samples)

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

P-value$< \alpha$, evidence for association between the genetic marker and the phenotype.

$\alpha =$ significance level of the test.

Testing the association between a given marker and the phenotype

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

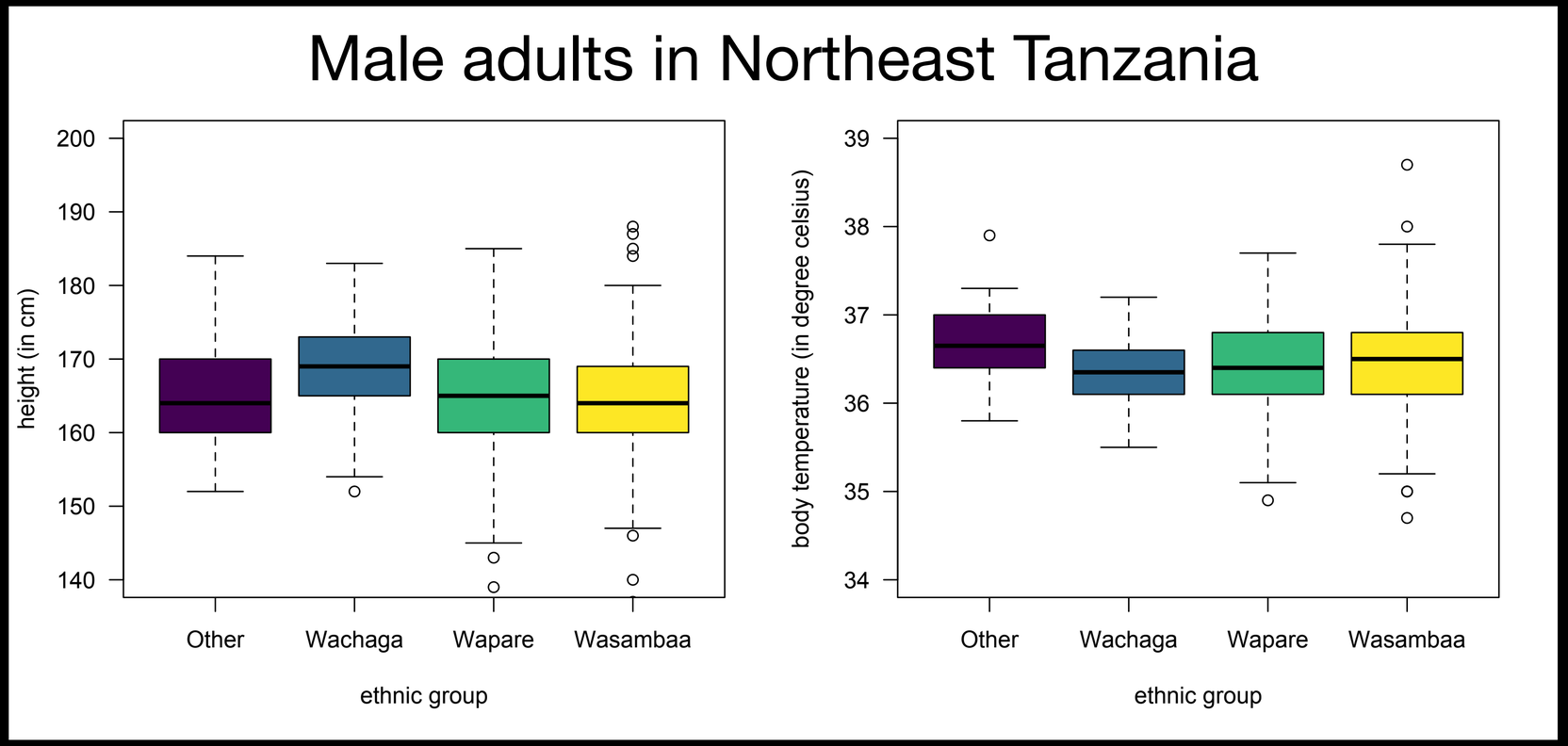
Wilks' likelihood ratio test

$$\Lambda = (-2) \times \left(\log L(\hat{\beta}_0 \,|\, M_0) - \log L(\hat{\beta}_0, \hat{\beta}_1 \,|\, M_1) \right) | \, H_0 \rightsquigarrow \chi^2_{(1)}$$

$\log L(\hat{\beta}_0 \,|\, M_0) =$ maximised log-likelihood of the regression model without the covariate

$\log L(\hat{\beta}_0, \hat{\beta}_j \,|\, M_1) =$ maximised log-likelihood of the regression model M_1 with the covariate

Working exercise: Tanzania dataset



Check online the information about the SNP rs1801033. Test the association of this genetic marker with height in the adults from Tanzania using the additive model. Draw your conclusions.

Extending the additive model

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

Under the assumption of sampling unrelated individuals

$$\mu_i = \beta_0 + \beta_1 X_{ij} + \underbrace{\beta_1^* X_{i1}^* + \cdots + \beta_p^* X_{ip}^*}_{\text{Non-genetic covariates}} + \epsilon_i$$

$\epsilon_i \mid \sigma^2 \rightsquigarrow N(0, \sigma^2)$

$$\mathbf{X}_i = (X_{i1}^*, \ldots, X_{ip}^*)$$

Which non-genetic covariates to include in the model?

Final model (?)

$Y_i \,|\, \mu_i, \sigma \rightsquigarrow N(\mu_i, \sigma^2)$ Under the assumption of sampling unrelated individuals

$$\mu_i = \beta + \underbrace{\beta_1 X_{i1} + \cdots + \beta_m X_{im}}_{\text{Associated genetic markers}} + \underbrace{\beta_1^* X_{i1}^* + \cdots + \beta_p^* X_{ip}^*}_{\text{Non-genetic covariates}} + \epsilon_i$$

Can you extend further this model?

Statistical validation of the model - Residuals analysis

$e_i = y_i - \hat{y}_i$ Residuals

Residuals should follow a normal distribution

Which plots/summary statistics can you used?

Do you know a test for testing this assumption?

Statistical validation of the model - Residuals analysis

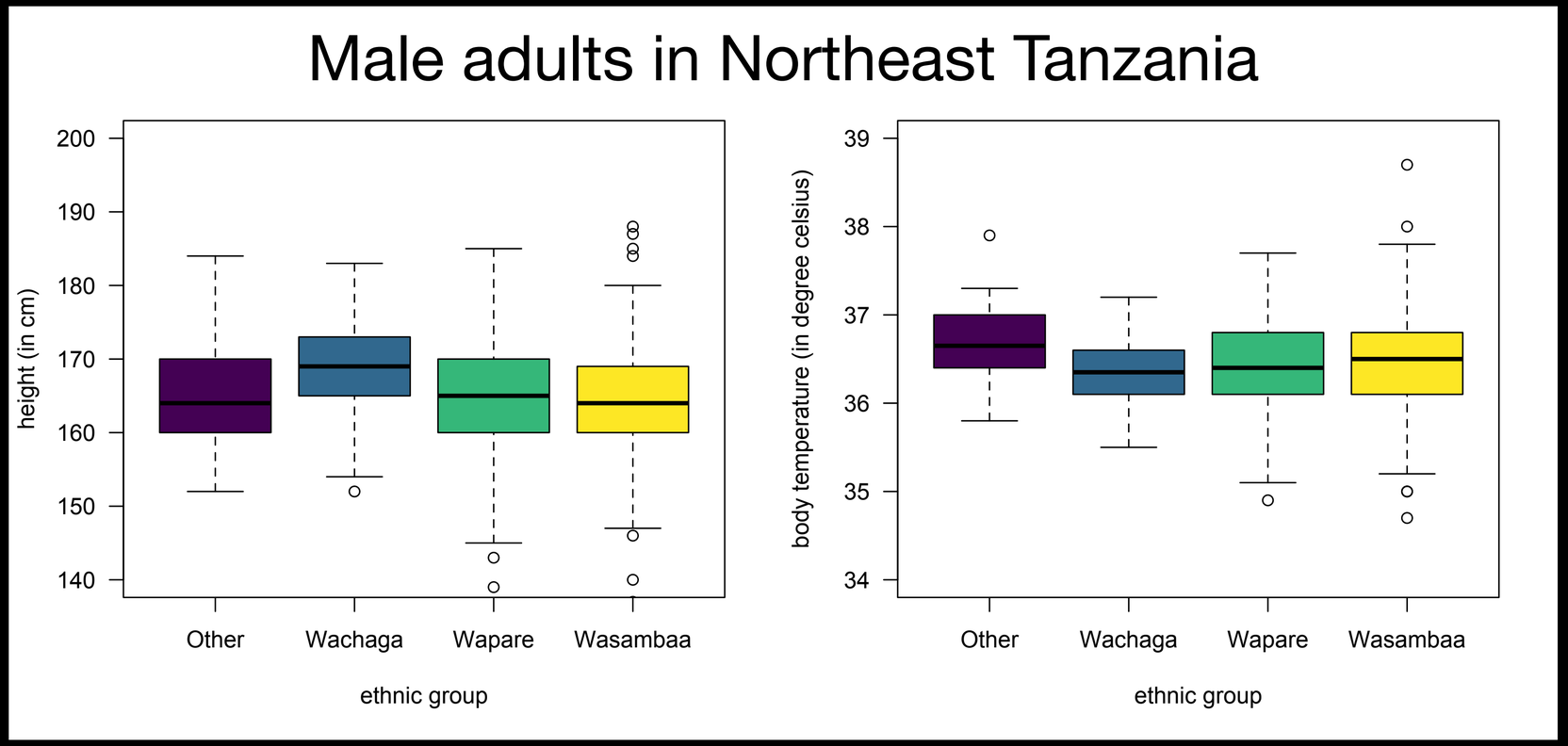
$e_i = y_i - \hat{y}_i$ Residuals

Residuals should be homoscedastic (same variance)

Which plots/summary statistics can you used?

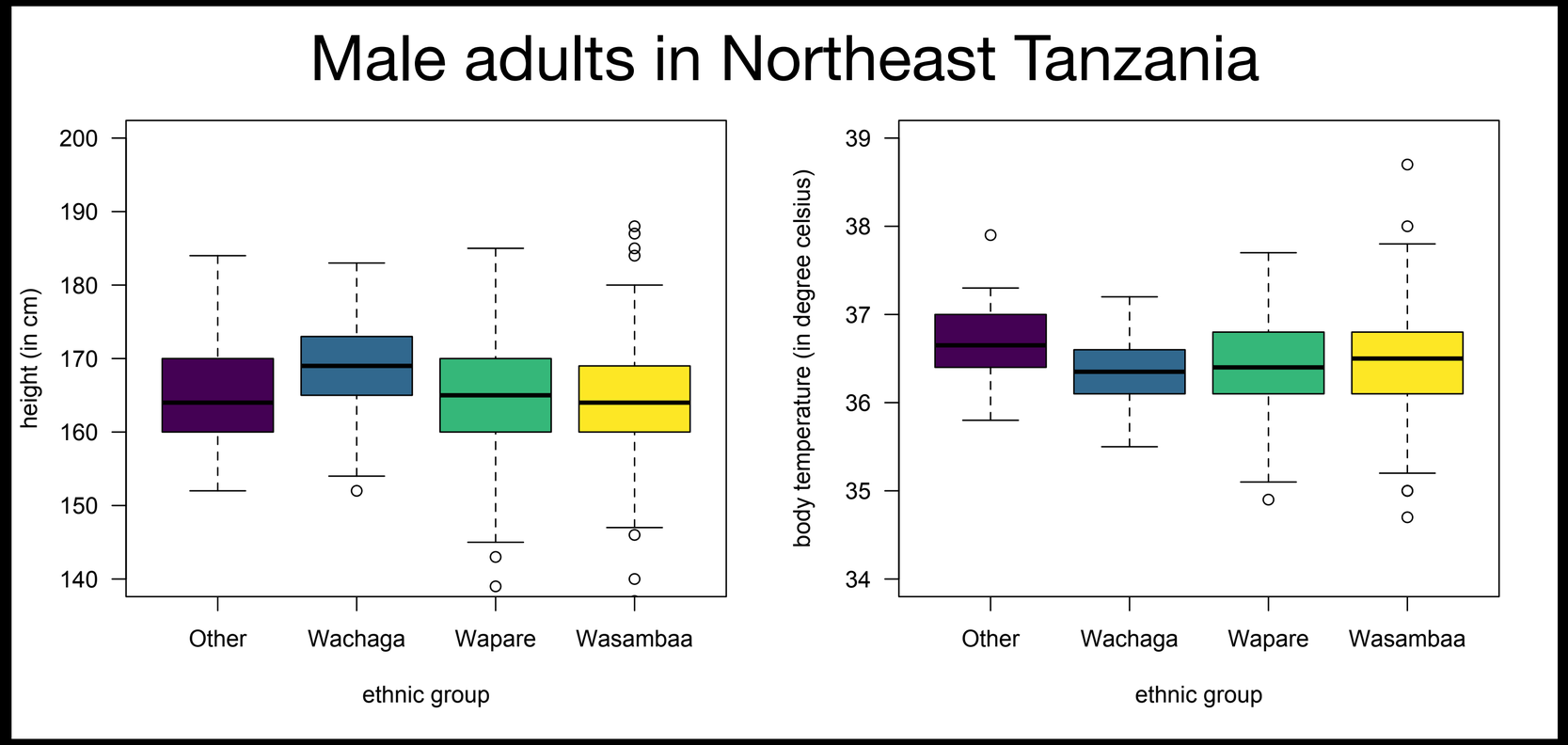
Do you know a test for testing this assumption?

Working exercise: Tanzania dataset



Check online the information about the SNP rs1801033. Test the association of this genetic marker with height but including “Ethnic.Group” as an additional covariate in the additive model. (Perform a residual analysis of the respective model.) Draw your conclusions.

Homework: Tanzania dataset

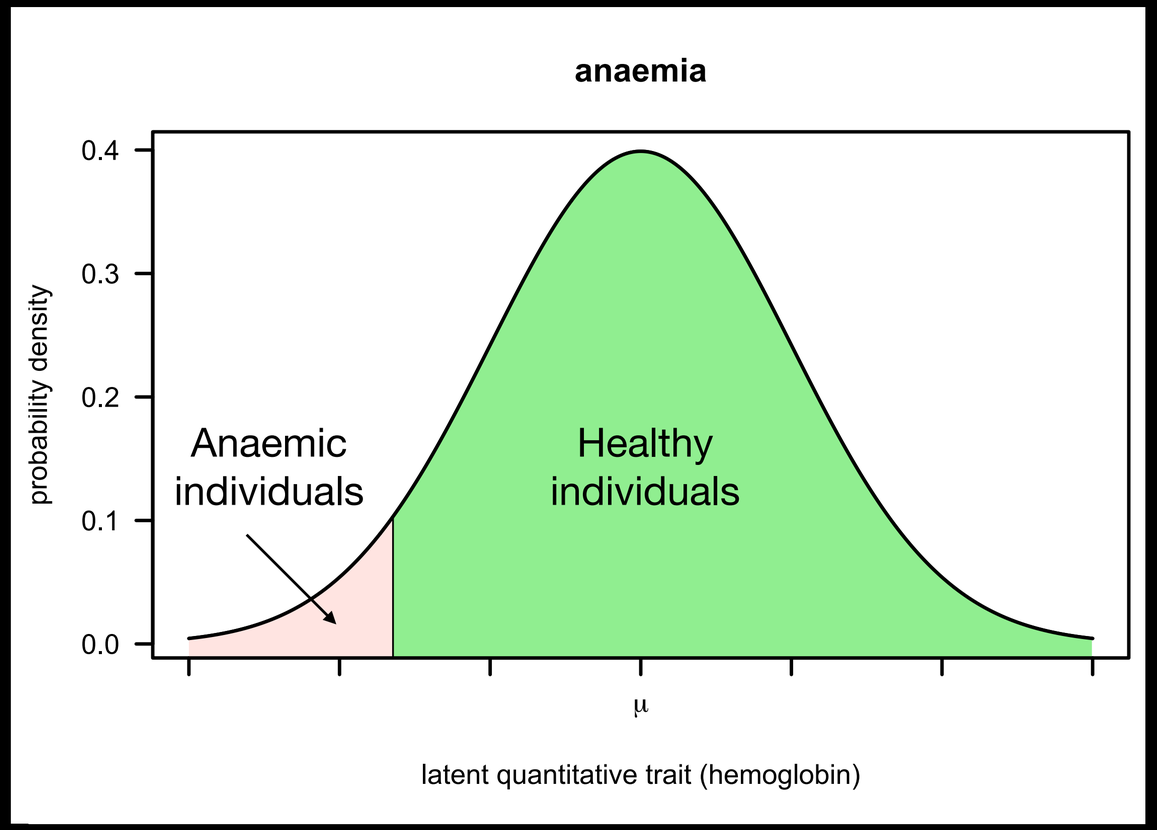


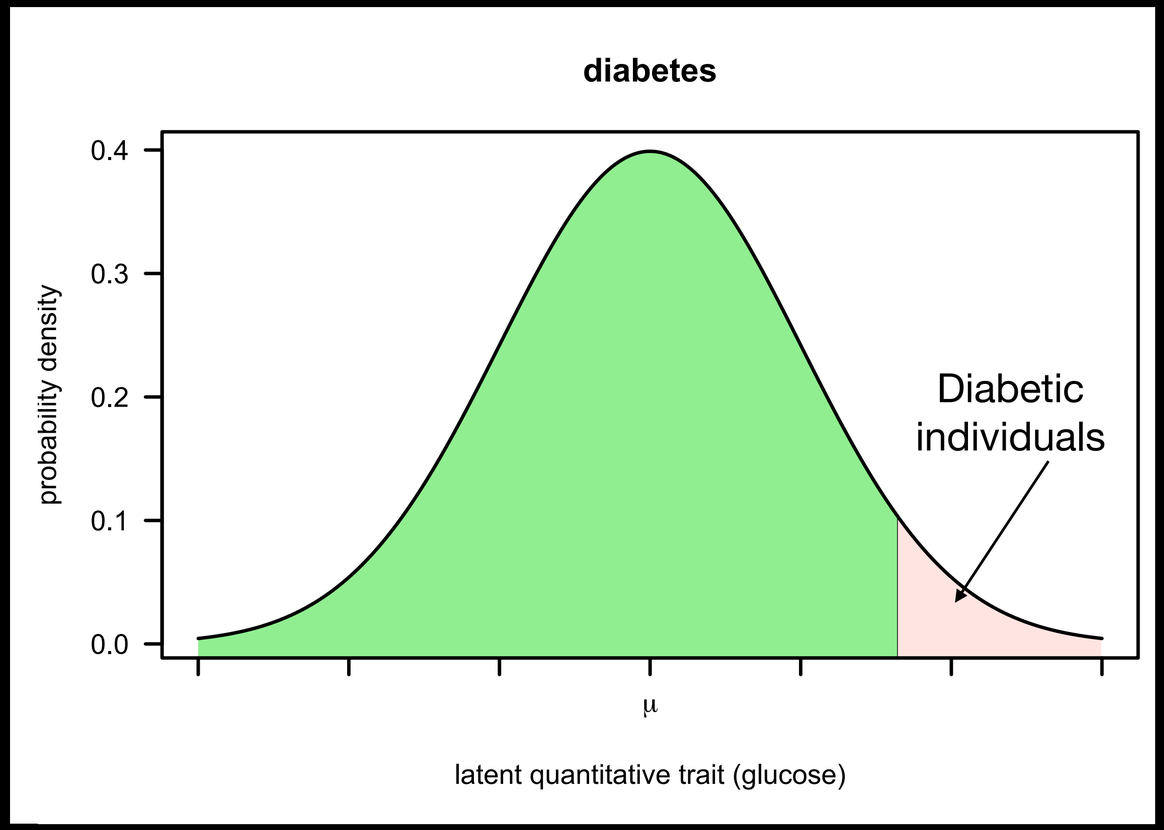
Test the association of rs1799964 with body temperature first alone and then including additionally malaria infection and ethnic group as additional covariates. Perform a residual analysis of the respective model. Draw your conclusions.

Application of Fisher's infinitesimal 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

Y_i = random variable for the binary trait in individual i

$$Y_i \mid \pi_i \rightsquigarrow \text{Bernoulli}(\pi_i)$$

π_i = Probability of the *i*-th individual expressing the phenotype of interest (e.g., presence of disease

Probit regression

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_j X_{ij} \qquad X_{ij} \in \{0,1,2\}$$ (single marker)

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_j X_{ij} + \beta_1^* X_{i1}^* + \cdots + \beta_p^* X_{ip}^*$$ (including other non-generic covariates)

Testing the association between a given marker and the phenotype

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

Wald's Score test (for large samples)

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

P-value$< \alpha$, evidence for association between the genetic marker and the phenotype.

$\alpha =$ significance level of the test.

Testing the association between a given marker and the phenotype

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

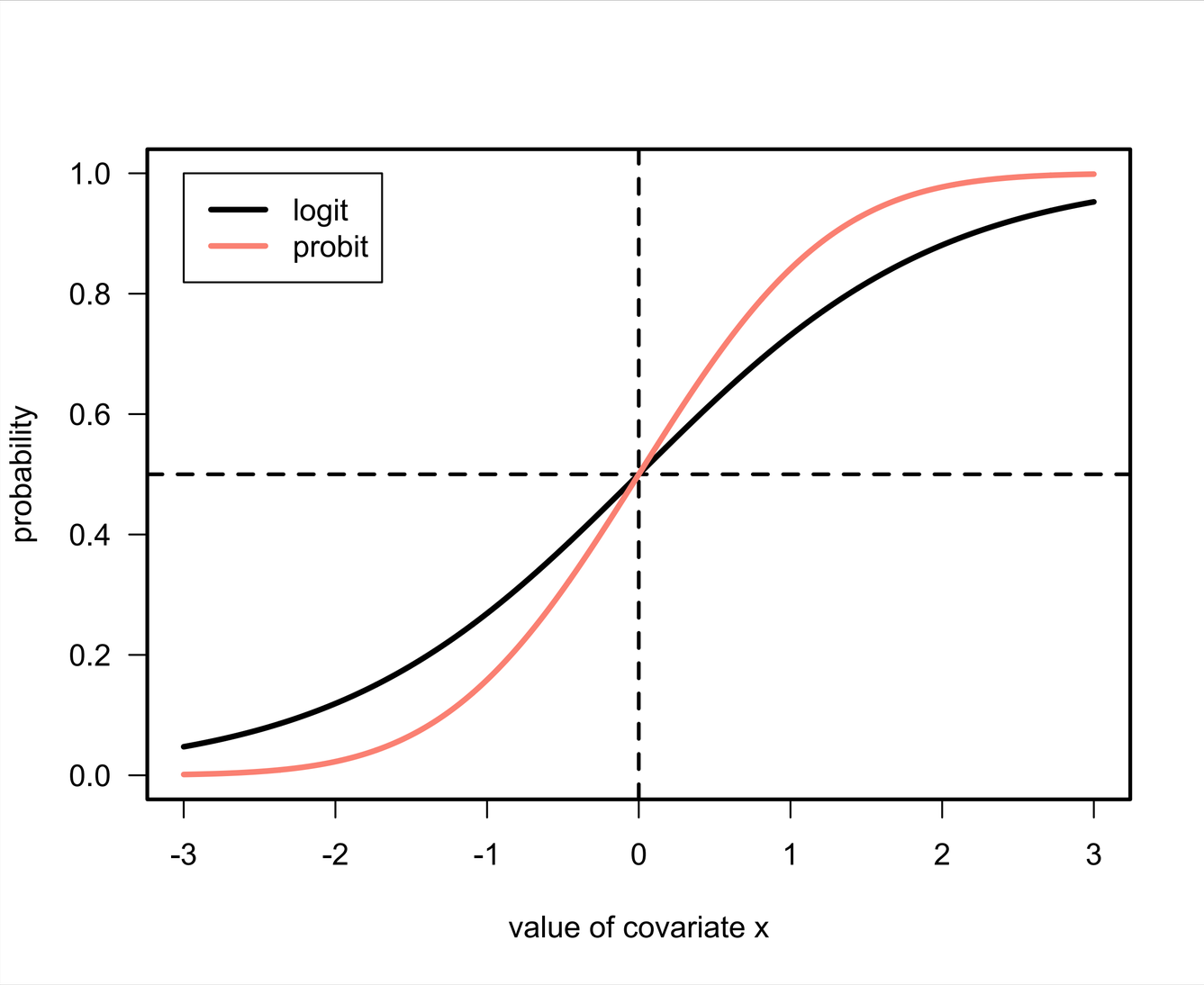
Wilks' likelihood ratio test

$$\Lambda = (-2) \times \left(\log L(\hat{\beta}_0 \,|\, M_0) - \log L(\hat{\beta}_0, \hat{\beta}_1 \,|\, M_1) \right) | \, H_0 \rightsquigarrow \chi^2_{(1)}$$

$\log L(\hat{\beta}_0 \,|\, M_0) =$ maximised log-likelihood of the regression model without the covariate

$\log L(\hat{\beta}_0, \hat{\beta}_j \,|\, M_1) =$ maximised log-likelihood of the regression model *M1* with the covariate

Practical note on probit regression



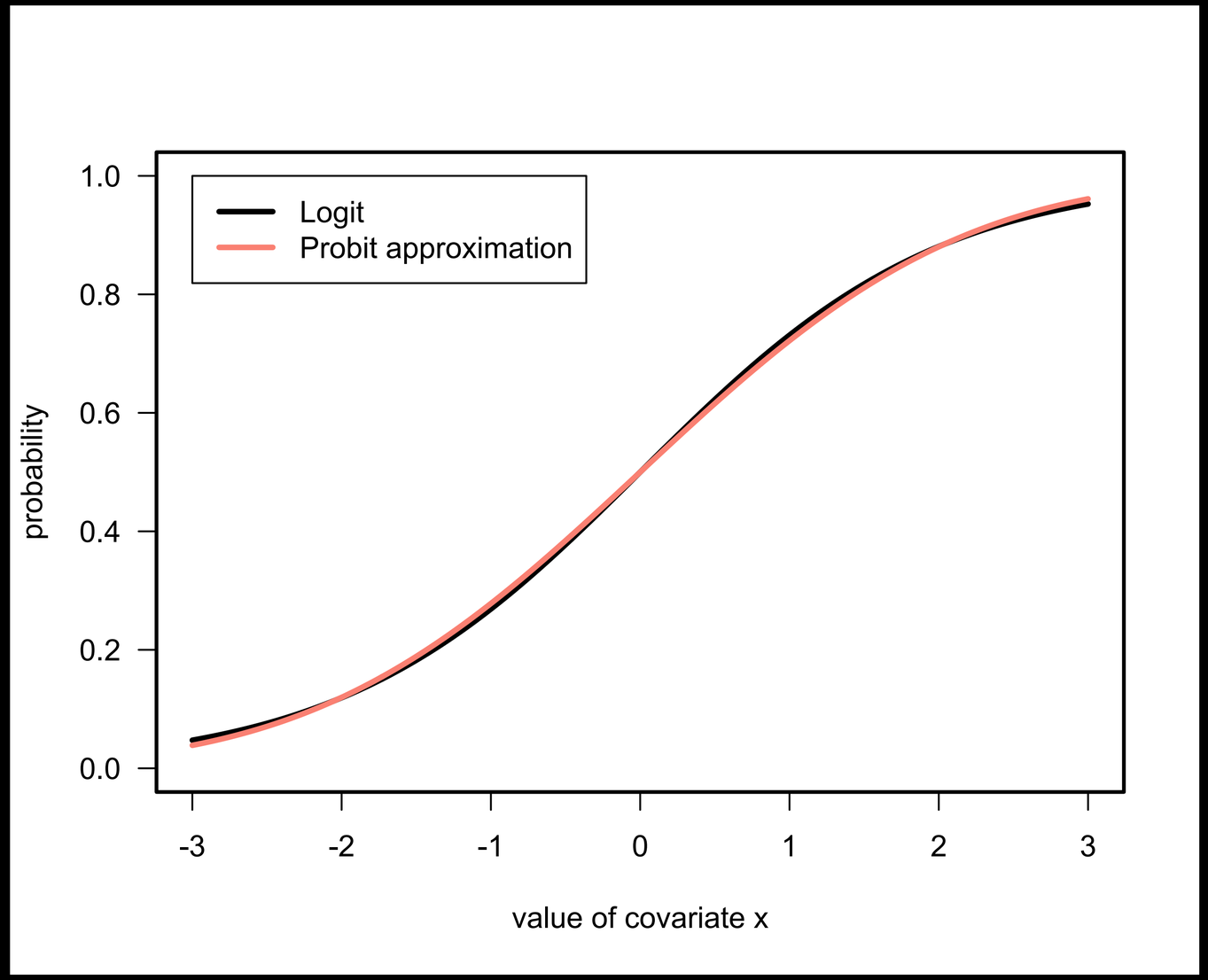
Logistic regression

$$\pi = \frac{e^{x_i}}{1 + e^{x_i}} \Leftrightarrow \log \frac{\pi_i}{1 - \pi_i} = x$$

Probit regression

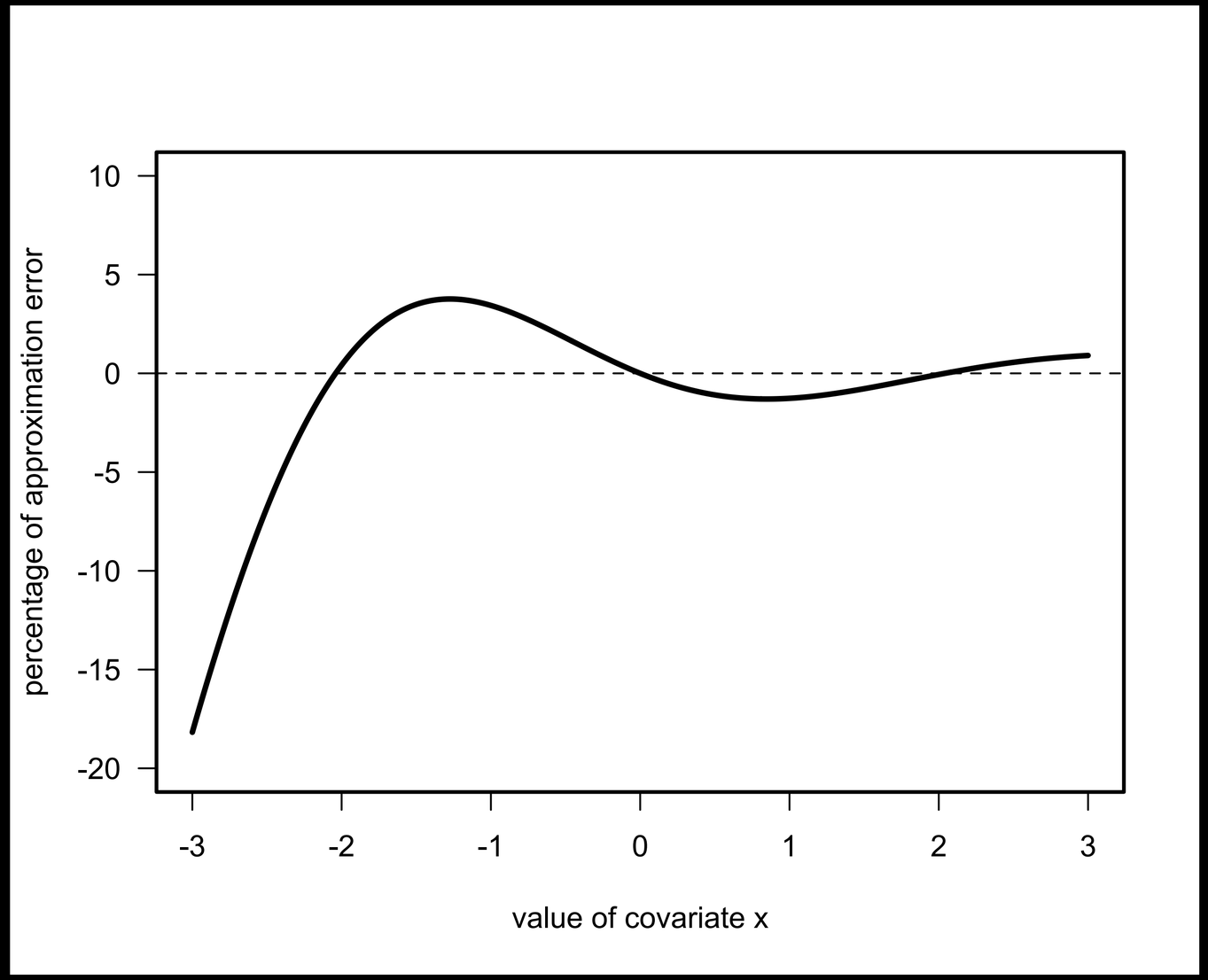
$$\pi_j = \Phi\left(x_i\right) \Leftrightarrow \Phi^{-1}\left(\pi_i\right) = x_i$$

Practical note on probit regression



$$\frac{e^{x_i}}{1+e^{x_i}} \approx \Phi\left(\frac{x_i}{1.70}\right)$$

Practical note on probit regression



Probit regression

You can fit a logistic regression model and then divide the corresponding estimated coefficients by 1.70

Testing the association between a given marker and the phenotype

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

Wilks' likelihood ratio test

$$\Lambda = (-2) \times \left(\log L(\hat{\beta}_0 \,|\, M_0) - \log L(\hat{\beta}_0, \hat{\beta}_1 \,|\, M_1) \right) | \, H_0 \rightsquigarrow \chi^2_{(1)}$$

$\log L(\hat{\beta}_0 \,|\, M_0) =$ maximised log-likelihood of the regression model without the covariate

$\log L(\hat{\beta}_0, \hat{\beta}_j \,|\, M_1) =$ maximised log-likelihood of the regression model *M1* with the covariate

Working exercise: Tanzania dataset

Assume sampling unrelated individuals

Check information online about rs6874639 and rs3024500. Test the association of these genetic markers with anaemia using the probit and logistic additive models.

Do the same test including age and malaria infection as covariates. Draw your conclusions.

Homework: Tanzania dataset

Repeat previous analysis but now for low haemoglobin as the binary phenotype.

Draw your conclusions.

Statistical validation of the probit/logistic additive model

Hosmer-Lemeshow test
It is like Pearson’s goodness-of-fit test.

le Cessie-Houwelingen test

$$e_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

Common genetic association studies

Candidate gene association studies (Obsolete)

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

10-250 SNPs under analysis

Genome-wide association studies (GWAS)

“Fishing expedition”

Hundred thousands/millions of SNPs under analysis

What are the practical problems of these studies?

Basic characteristics of the Illumina™ chips (human)

Array	Number of SNP evaluated	Number of samples per array	Number of samples scanned per week
HumanOmni2.5-Quad	~2.5 Million	4	672
Infinium Global Screening Array - 48 Kit	650,321 markers + 50,000 custom marker capacity	48	~11,520
Infinium Global Screening Array - 24 Kit	654,027 markers + 100,000 custom marker capacity	24	~5760
Infinium Global Diversity Array-8 Kit	1,825,277 markers	8	~1728
Infinium CytoSNP-850K	848,902 markers	8	~960

Check Affymetrix Arrays

Illumina chips (human)

Infinium Global Screening Array - 48 Kit

Table 3: Marker information

Marker categories			No. of markers
Exonic markers[a]			81,168
Intronic markers[a]			257,722
Nonsense markers[b]			5269
Missense markers[b]			45,829
Synonymous markers[b]			8476
Mitochondrial markers[b]			1089
Indels[c]			8471
Sex chromosomes[c]	X	Y	PAR/homologous
	27,036	3887	823

a. RefSeq-NCBI Reference Sequence Database.[19]
b. Compared against the UCSC Genome Browser.[6]
c. NCBI Genome Reference Consortium, Version GRCh37.[21]

Infinium Global Screening Array - 24 Kit

Table 3: Marker information

Marker categories			No. of markers
Exonic markers[a]			85,342
Intronic markers[a]			262,173
Nonsense markers[b]			5904
Missense markers[b]			51,188
Synonymous markers[b]			9273
MItochondrial markers[b]			1138
Indels[c]			10,118
Sex chromosomes[c]	X	Y	PAR/homologous
	27,176	4138	879

a. RefSeq - NCBI Reference Sequence Database.[20] Accessed May 2020.
b. Compared against the UCSC Genome Broswer.[4] Accessed May 2020.
c. NCBI Genome Reference Consortium, Version GRCh37.[21] Accessed May 2020.
Abbreviations: indel, insertion/deletion; PAR, pseudoautosomal region.

Illumina chips (human)

Infinium Global Diversity Array-8 Kit

Table 3: Marker information

Marker category			No. of markers
Exonic markers[a]			482,865
Intronic markers[a]			654,467
Nonsense markers[b]			26,140
Missense markers[b]			308,978
Synonymous markers[b]			35,518
Mitochondrial markers[b]			1354
Indels[c]			38,694
Sex chromosomes	X	Y	PAR/homologous
	62,353	6456	5490

a. RefSeq NCBI Reference Sequence Database.[20] Accessed April 2021

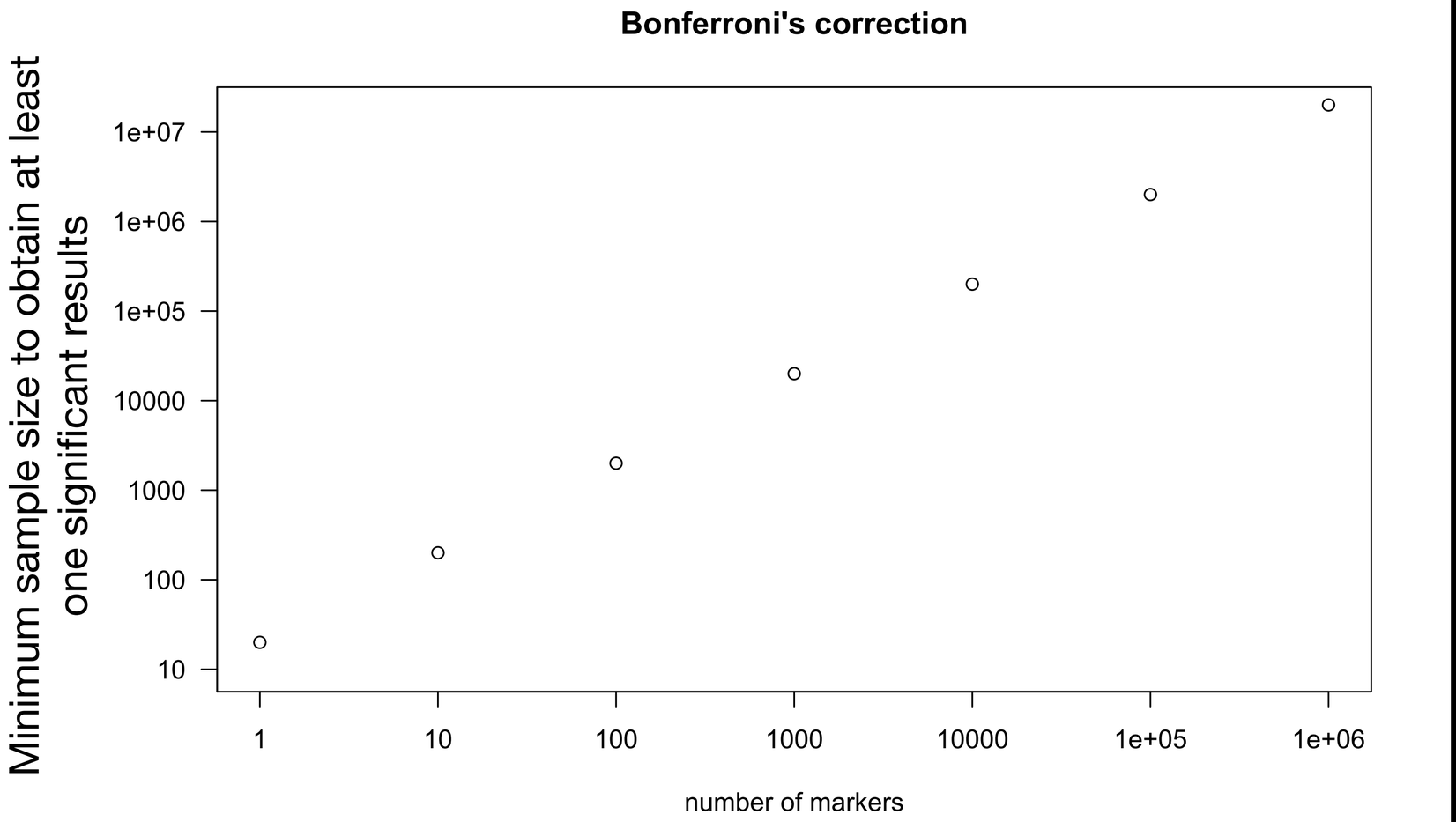
b. Compared against the UCSC Genome Browser.[4] Accessed April 2021

c. NCBI Genome Reference Consortium, Version GRCh37.[21] Accessed April 2021

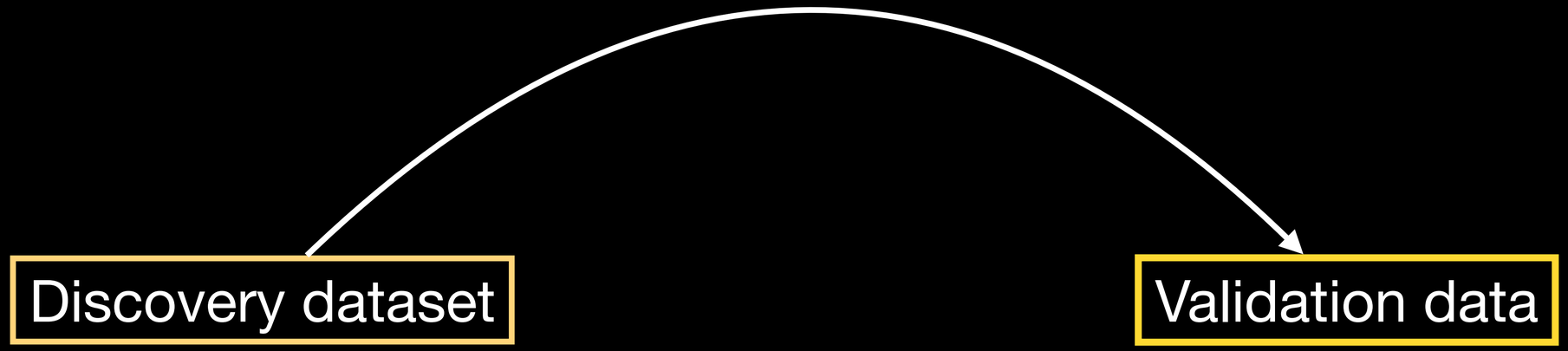
Abbreviations: indel, insertion/deletion; PAR, pseudoautosomal region

What are the (technical) problems of GWAS?

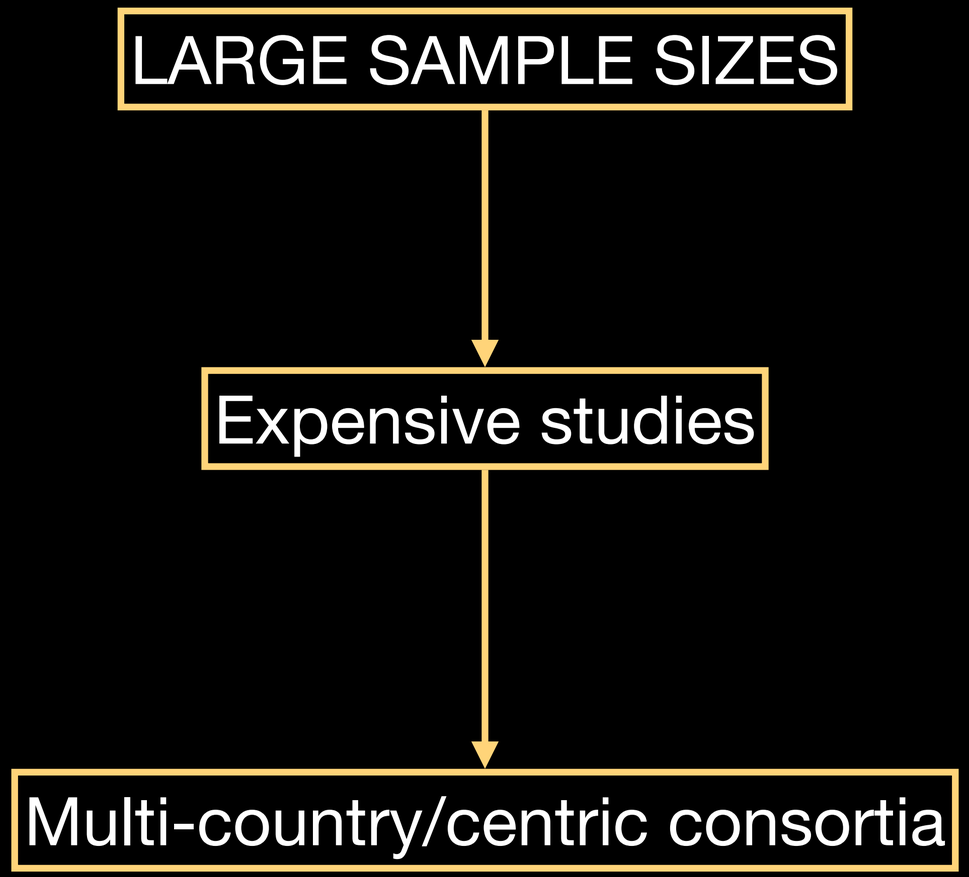
The curse of multiple testing



The curse of reproducibility/validation



Solution


LARGE SAMPLE SIZES
Expensive studies
Multi-country/centric consortia

Examples of consortia

nature genetics ARTICLES
https://doi.org/10.1038/s41588-018-0303-9

Trans-ethnic association study of blood pressure determinants in over 750,000 individuals

Ayush Giri[1,2,98], Jacklyn N. Hellwege[2,3,98], Jacob M. Keaton[2,3,98], Jihwan Park[4,98], Chengxiang Qiu[4], Helen R. Warren[5,6], Eric S. Torstenson[2,3], Csaba P. Kovesdy[7], Yan V. Sun[8,9], Otis D. Wilson[2,10], Cassianne Robinson-Cohen[10], Christianne L. Roumie[11,12], Cecilia P. Chung[13], Kelly A. Birdwell[10,14], Scott M. Damrauer[15,16], Scott L. DuVall[17,18], Derek Klarin[19,20,21,22], Kelly Cho[23,24,25], Yu Wang[26], Evangelos Evangelou[27,28], Claudia P. Cabrera[5,6], Louise V. Wain[29,30], Rojesh Shrestha[4], Brian S. Mautz[3], Elvis A. Akwo[10], Muralidharan Sargurupremraj[31], Stéphanie Debette[31,32], Michael Boehnke[33], Laura J. Scott[33], Jian'an Luan[34], Jing-Hua Zhao[34], Sara M. Willems[34], Sébastien Thériault[35,36], Nabi Shah[37,38], Christopher Oldmeadow[39], Peter Almgren[40], Ruifang Li-Gao[41], Niek Verweij[42], Thibaud S. Boutin[43], Massimo Mangino[44,45], Ioanna Ntalla[5], Elena Feofanova[46], Praveen Surendran[47], James P. Cook[48], Savita Karthikeyan[47], Najim Lahrouchi[21,49,50], Chunyu Liu[51], Nuno Sepúlveda[52], Tom G. Richardson[53], Aldi Kraja[54,55,56], Philippe Amouyel[57], Martin Farrall[58], Neil R. Poulter[59], Understanding Society Scientific Group[60], International Consortium for Blood Pressure[60], Blood Pressure-International Consortium of Exome Chip Studies[60], Markku Laakso[61], Eleftheria Zeggini[62], Peter Sever[63], Robert A. Scott[34], Claudia Langenberg[34], Nicholas J. Wareham[34], David Conen[64], Colin Neil Alexander Palmer[37], John Attia[39,65], Daniel I. Chasman[66], Paul M. Ridker[66], Olle Melander[40], Dennis Owen Mook-Kanamori[41], Pim van der Harst[42], Francesco Cucca[67,68], David Schlessinger[69], Caroline Hayward[43], Tim D. Spector[44], Marjo-Riitta Jarvelin[70,71,72,73,74], Branwen J. Hennig[75,76,77], Nicholas J. Timpson[53], Wei-Qi Wei[78], Joshua C. Smith[78], Yaomin Xu[26,78], Michael E. Matheny[11,12,26,78], Edward E. Siew[10,12], Cecilia Lindgren[21,79,80], Karl-Heinz Herzig[81,82], George Dedoussis[83], Joshua C. Denny[78], Bruce M. Psaty[84,85,86,87], Joanna M. M. Howson[47], Patricia B. Munroe[5,6], Christopher Newton-Cheh[49], Mark J. Caulfield[5,6], Paul Elliott[70,88,89], J. Michael Gaziano[23,66], John Concato[90,91], Peter W. F. Wilson[92,93], Philip S. Tsao[94,95], Digna R. Velez Edwards[1,2,78], Katalin Susztak[4,96,99], Million Veteran Program[60], Christopher J. O'Donnell[66,97,99], Adriana M. Hung[2,10,99*] and Todd L. Edwards[2,3,99*]

nature genetics

Reappraisal of known malaria resistance loci in a large multicenter study

Malaria Genomic Epidemiology Network*

Many human genetic associations with resistance to malaria have been reported, but few have been reliably replicated. We collected data on 11,890 cases of severe malaria due to *Plasmodium falciparum* and 17,441 controls from 12 locations in Africa, Asia and Oceania. We tested 55 SNPs in 27 loci previously reported to associate with severe malaria. There was evidence of association at $P < 1 \times 10^{-4}$ with the *HBB*, *ABO*, *ATP2B4*, *G6PD* and *CD40LG* loci, but previously reported associations at 22 other loci did not replicate in the multicenter analysis. The large sample size made it possible to identify authentic genetic effects that are heterogeneous across populations or phenotypes, with a striking example being the main African form of G6PD deficiency, which reduced the risk of cerebral malaria but increased the risk of severe malarial anemia. The finding that G6PD deficiency has opposing effects on different fatal complications of *P. falciparum* infection indicates that the evolutionary origins of this common human genetic disorder are more complex than previously supposed.

DecodeME

Add languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

This article is about the ongoing study whether ME/CFS has a genetic risk factor. For the defunct genome testing service, see deCODE genetics.

Not to be confused with Decode Entertainment.

DecodeME is a genome-wide association study searching for genetic risk factors for Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). With a planned recruitment of 25,000 patients, it is expected to be the largest such study to date.[4][5] Recruitment closed on 15 November 2023 and preliminary results were published in a preprint on 7 August 2025.[6]

DecodeME

Decode ME
The ME/CFS Study

The study's official logo

Purpose	Research (DNA)
ICD-10-PCS	G93.32.[1][2]

Background [edit]

ME/CFS is a chronic medical condition that often causes significant disability, and

Do you know any GWAS done in Poland?

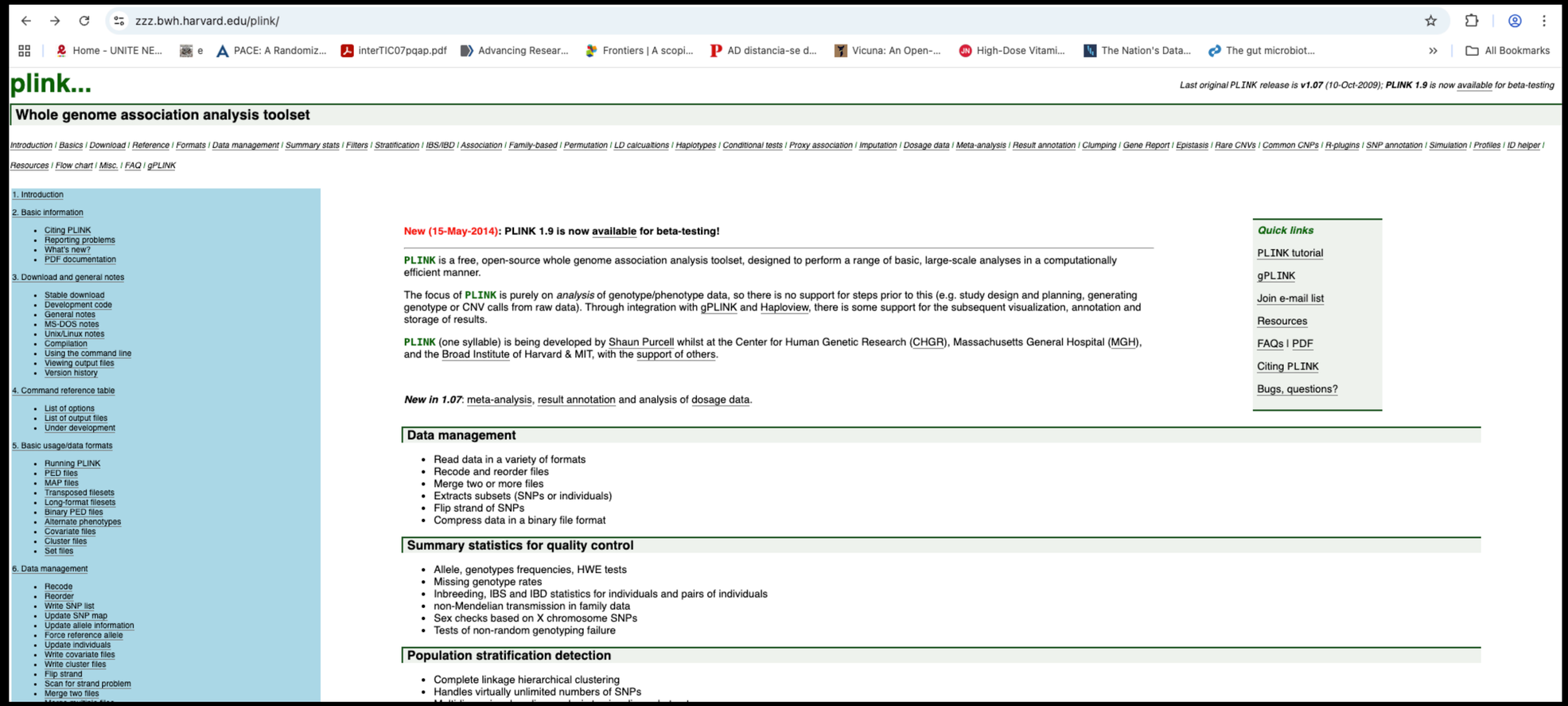
Steps of a GWAS

Data quality checks
- Minor allele frequency
- Missing data (SNP/individual)
- Hardy-Weinberg equilibrium test
- Heterozygosity
- Biological sex

Data analysis
- Data imputation (optional - outside the scope of the course)
- Association tests
- Manhattan plot
- Multiple testing correction (Bonferroni, false discovery rate)
- Haplotype analysis

Software

PLINK



How to use the software (outside the scope of this course)

Steps of a GWAS

Data quality checks
- Minor allele frequency
- Missing data (SNP/individual)
- Hardy-Weinberg equilibrium test
- Heterozygosity
- Biological sex

Data analysis
- Data imputation (optional - outside the scope of the course)
- Association tests
- Manhattan plot
- Multiple testing correction (Bonferroni, false discovery rate)
- Haplotype analysis

Data quality checks

frontiers in Pediatrics

OPINION
published: 12 June 2020
doi: 10.3389/fped.2020.00293

Check for updates

Review of the Quality Control Checks Performed by Current Genome-Wide and Targeted-Genome Association Studies on Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

Anna D. Grabowska[1], *Eliana M. Lacerda*[2], *Luís Nacul*[2,3] *and Nuno Sepúlveda*[4,5]*

[1] *Department of Biophysics and Human Physiology, Medical University of Warsaw, Warsaw, Poland,* [2] *Department of Clinical Research, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom,* [3] *Complex Chronic Diseases Program, British Columbia Women's Hospital and Health Centre, Vancouver, BC, Canada,* [4] *Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom,* [5] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal*

OPEN ACCESS

Edited by:
Marco Carotenuto,
University of Campania Luigi

Data quality checks: minor allele frequency

Minor allele frequency (MAF - 0-50%)
Frequency of the allele with less frequency in the population.

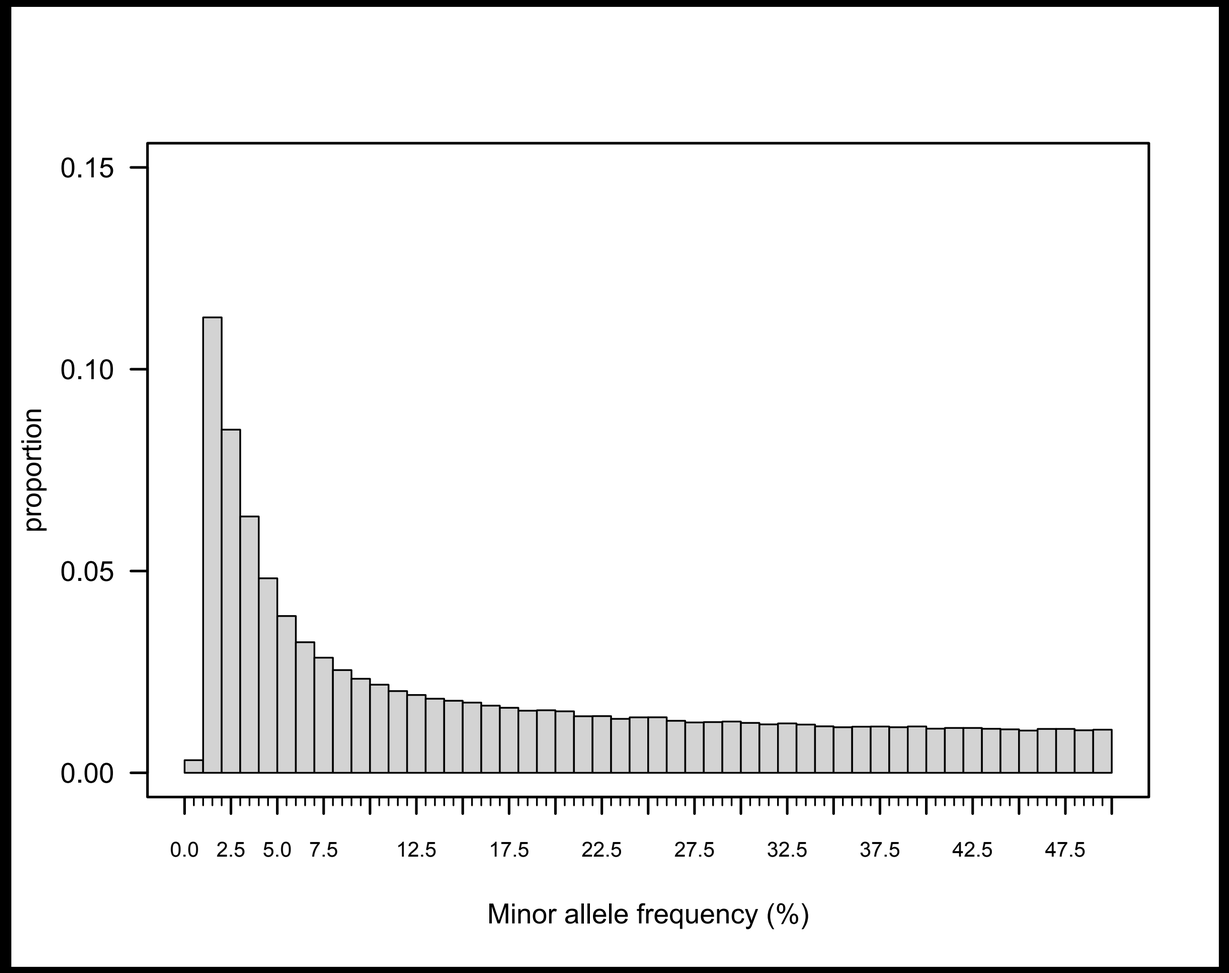
$$\text{MAF} = \min\left\{\frac{2n_{aa}+n_{ab}}{2n}, \frac{2n_{bb}+n_{ab}}{2n}\right\}$$

Action:

1. Calculate the MAF frequency distribution
2. Remove all SNP with a MAF < cutoff (say 0.025 or 0.01)

Why to remove these SNPs?

Data quality checks: minor allele frequency

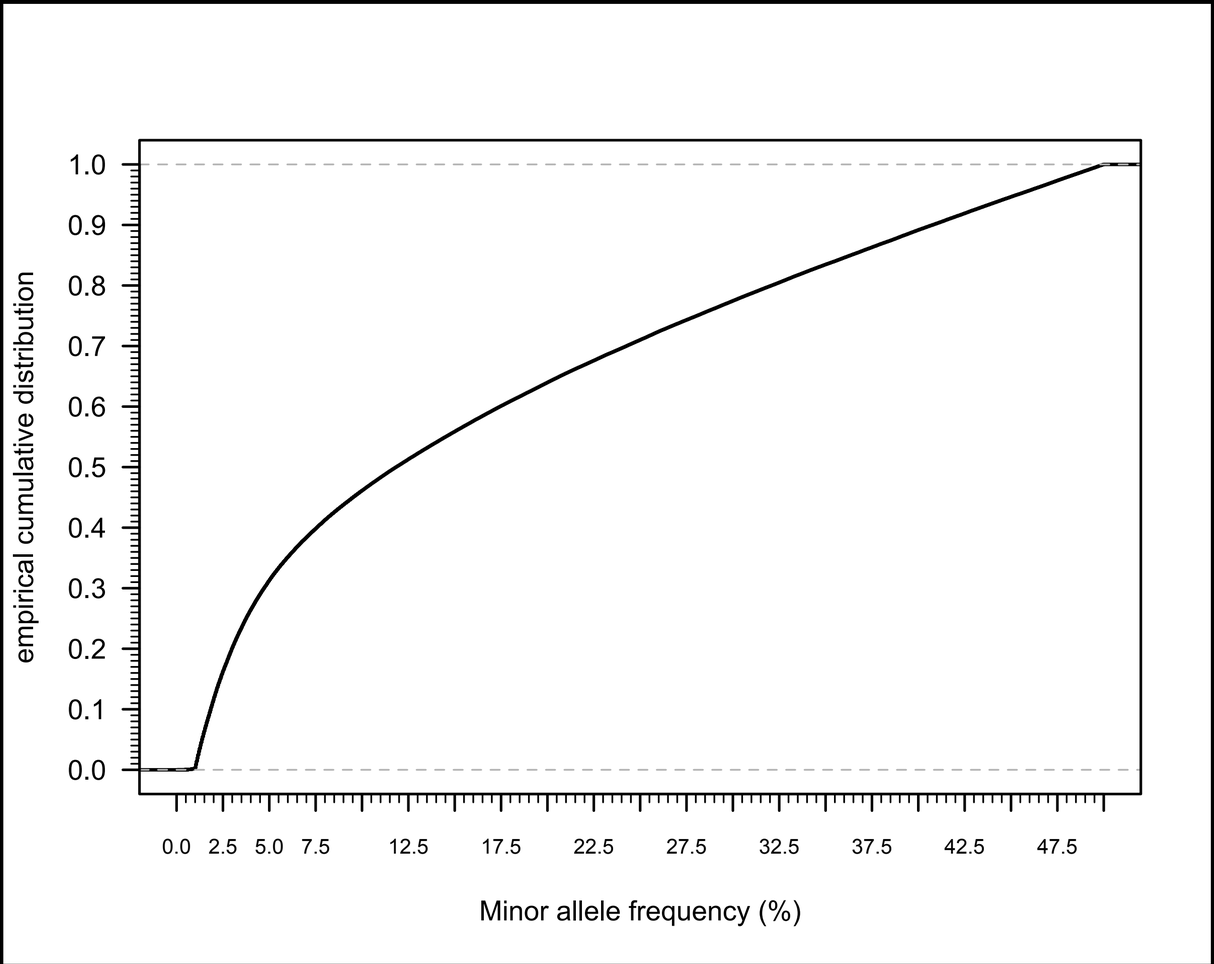


COVID-19 study in Portugal
n=1219

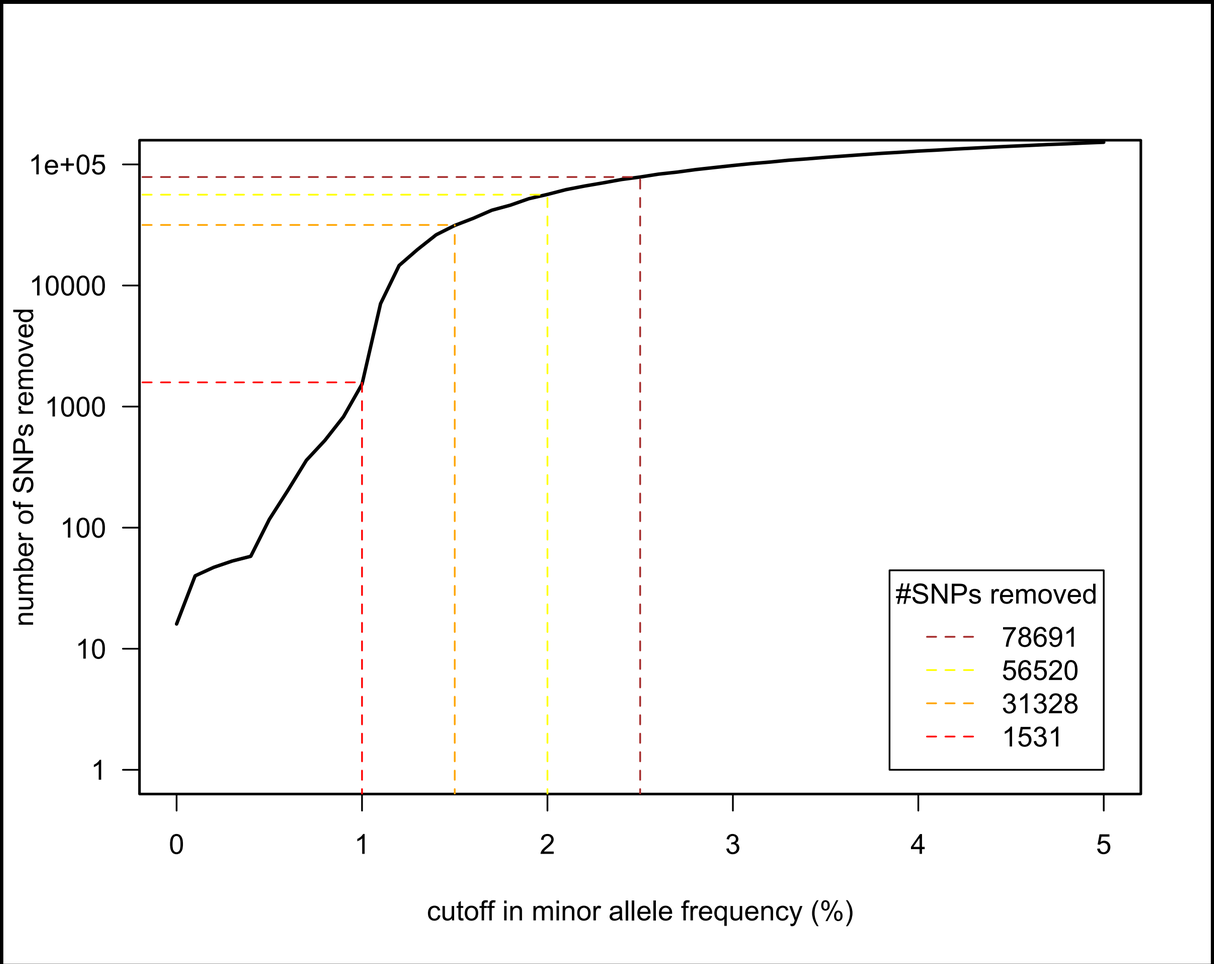
Affymetrix Axiom® 480 K SNP

Total SNPs in the array = 487,314

Data quality checks: minor allele frequency



Data quality checks: minor allele frequency



Choose the cutoff:

1. Number of SNPs removed
2. Expected frequency per genotype

(more on Hardy-Weinberg equilibrium)

Data quality checks: missing data (SNP)

Action:

1. Calculate the proportion of missing genotypes per SNP

$$\pi_{missing,k} = \frac{n_{missing,k}}{n}$$

2. Remove all SNP with a proportion > cutoff (say 0.05 or 0.1)

Data quality checks: missing data (individual)

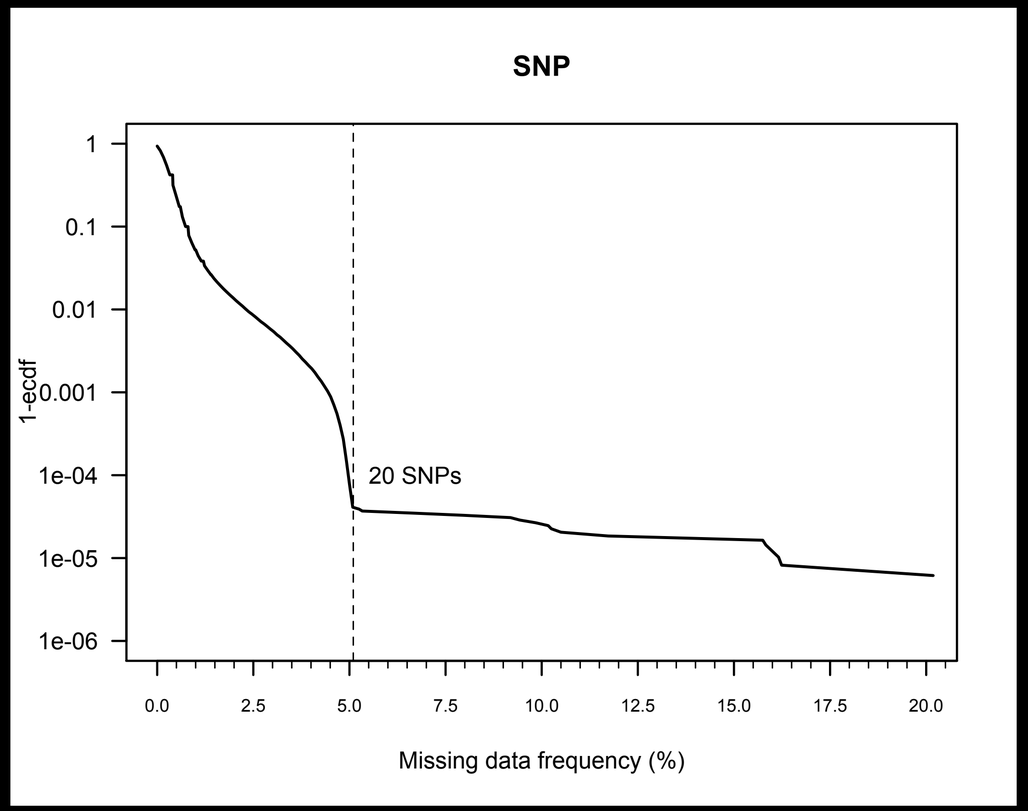
Action:

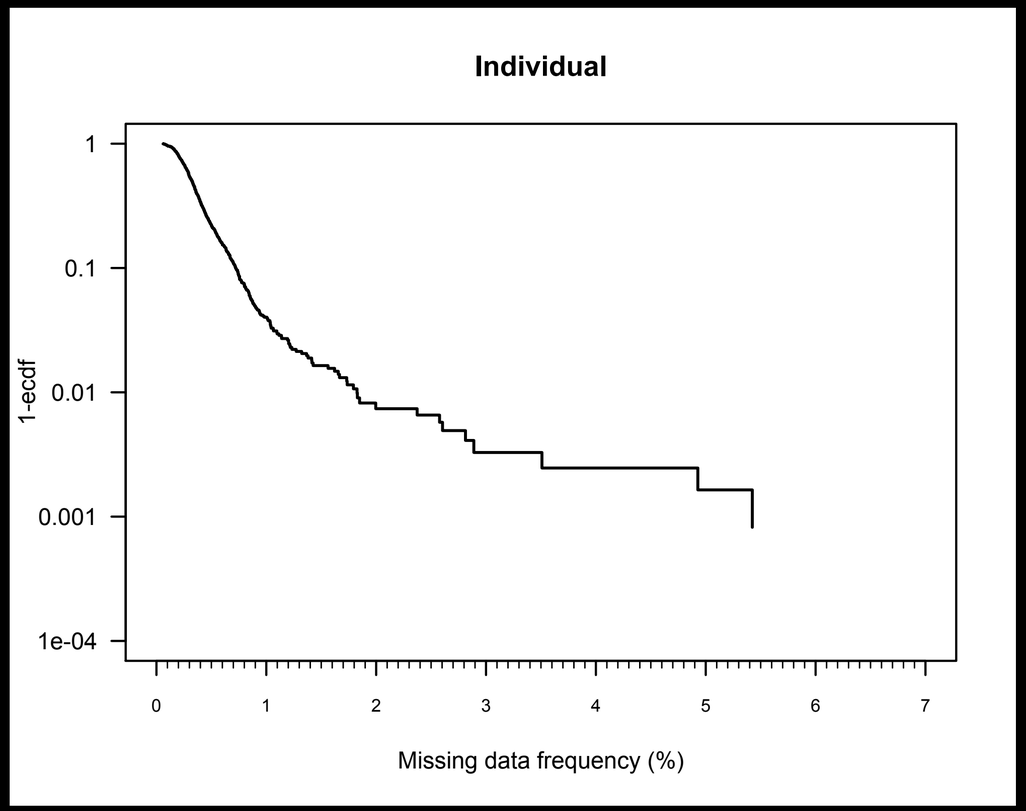
1. Calculate the proportion of missing genotypes per individual

$$\pi_{missing,i} = \frac{n_{missing,i}}{n}$$

2. Remove all SNP with a proportion > cutoff (say 0.05 or 0.1)

Data quality checks: missing data (individual)





Data quality checks: Hardy-Weinberg equilibrium test

Genotype	Frequency	Probability
aa	n_{aa}	π_a^2
ab	n_{ab}	$2\pi_a(1-\pi_a)$
bb	n_{bb}	$(1-\pi_a)^2$

(Healthy/Cases data only)

Assumptions

No genotype errors

No selection/migration/mixture

Random mating

Under Multinomial sampling

$$\hat{\pi}_a = \frac{2n_{aa} + n_{ab}}{2(n_{aa} + n_{ab} + n_{bb})}$$ (MLE)

Data quality checks: Hardy-Weinberg equilibrium test

$$X^2 = \sum_{i=aa,ab,bb} \frac{(n_i - e_i)^2}{e} \mid H_0 \rightsquigarrow \chi_1^2$$

1. P-value < α^*, reject the null hypothesis (remove SNP)

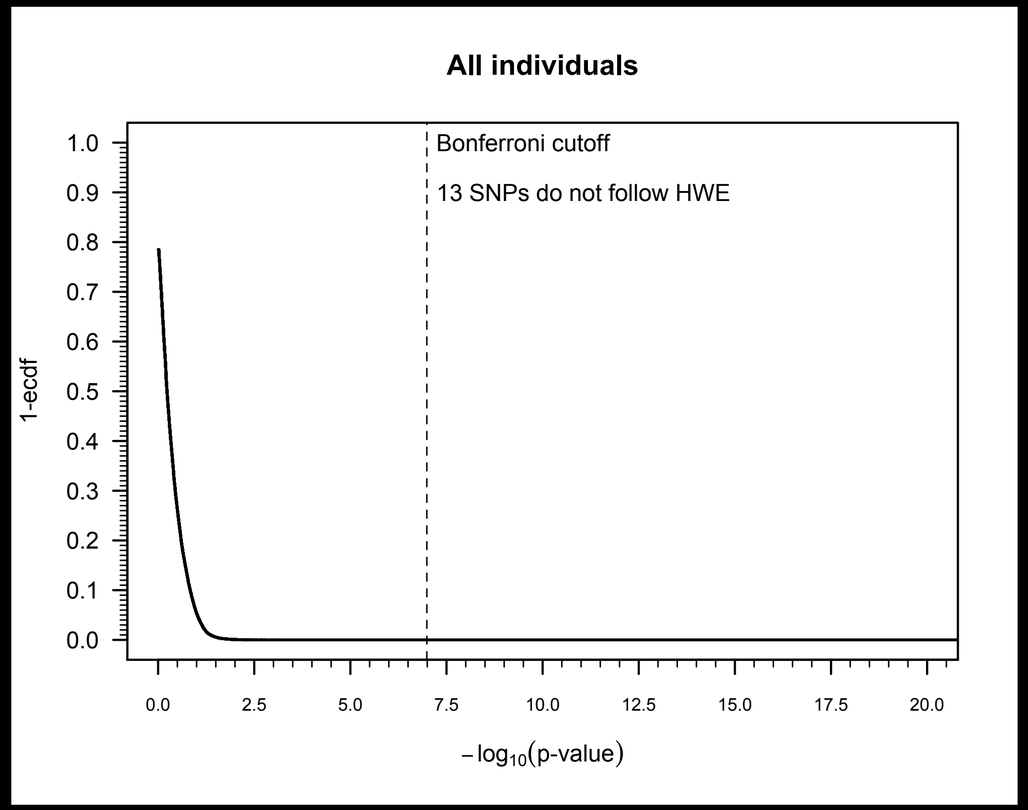
$$\alpha^* = \frac{\alpha}{p}$$ (Bonferroni correction)

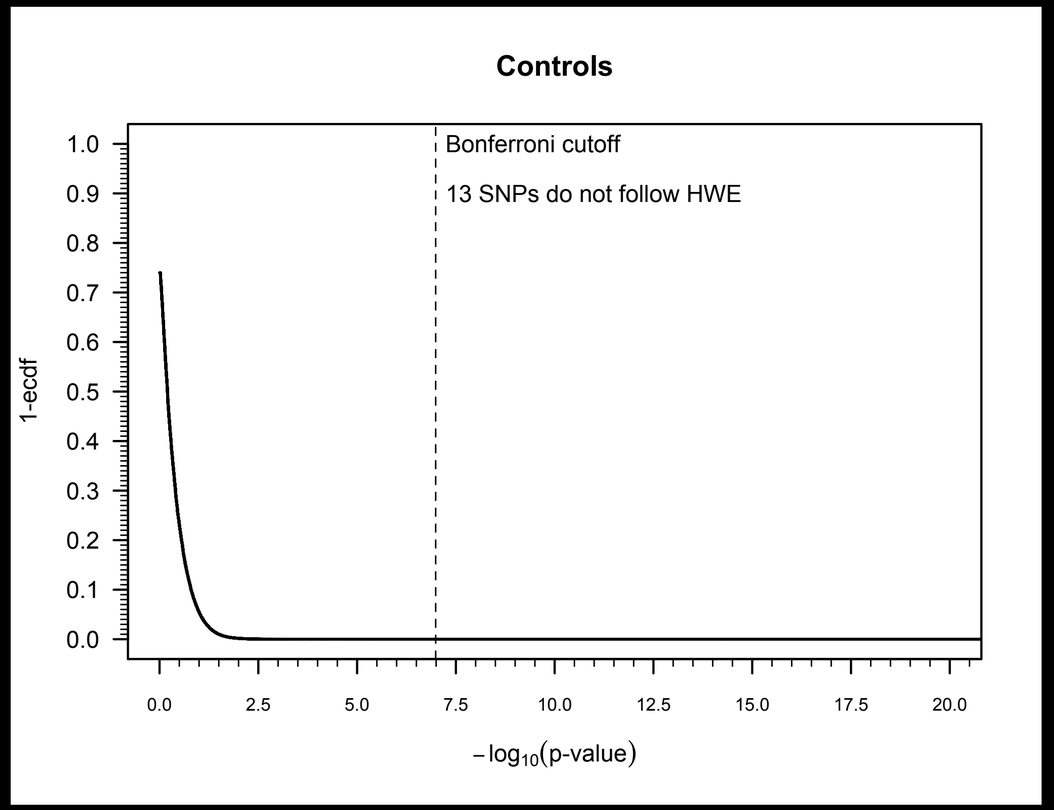
where α is the significance level of the overall analysis

2. P-values $\mid H_0 \rightsquigarrow$ Uniform$(0,1)$

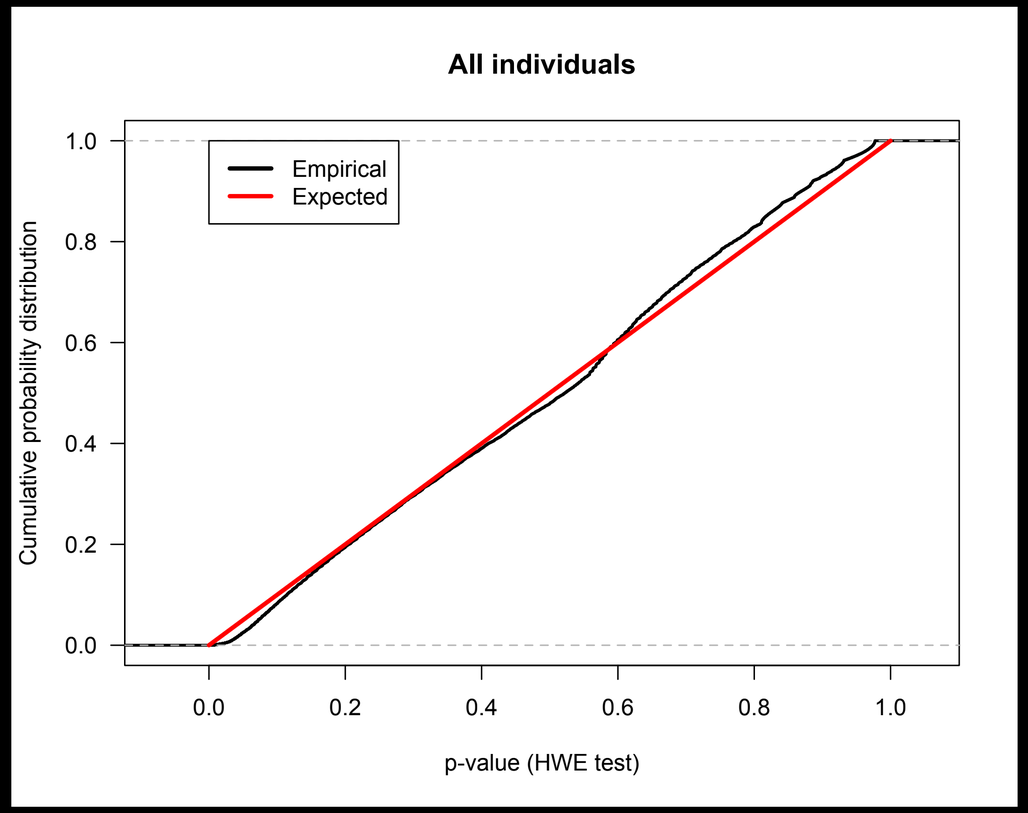
Assess the empirical distribution visually (e.g., PP-plot or QQ-plot) and check whether there are some strong deviation from the uniform distribution.

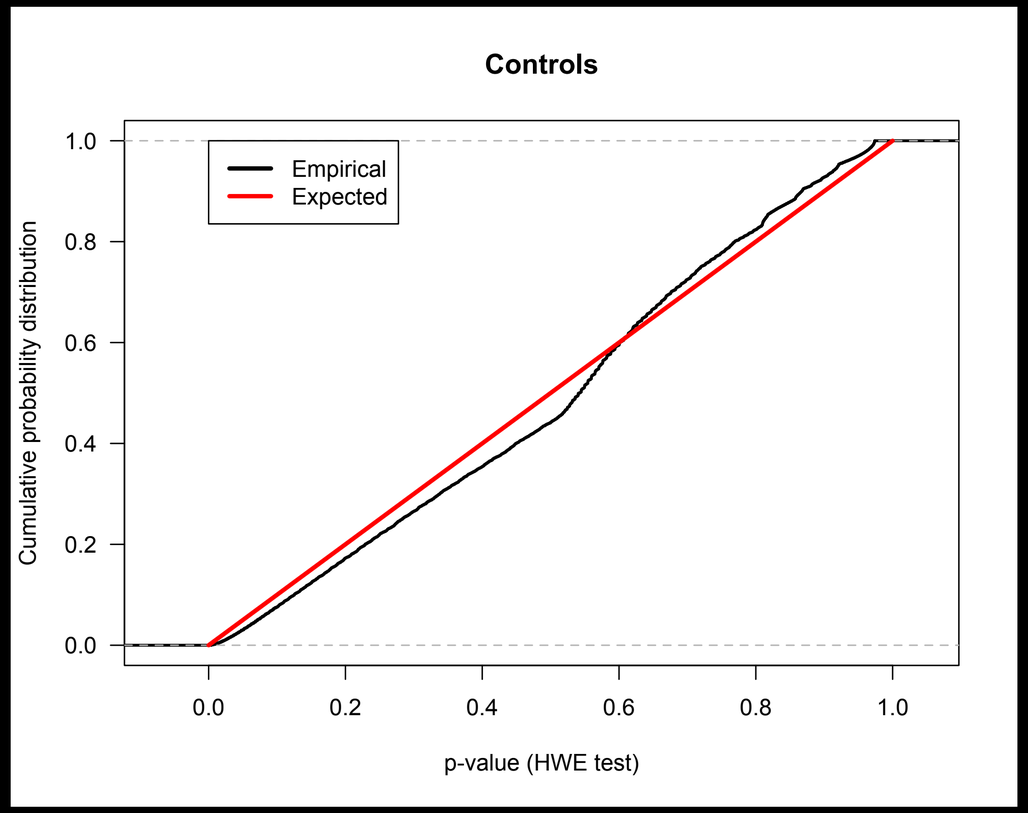
Data quality checks: Hardy-Weinberg equilibrium test





Data quality checks: Hardy-Weinberg equilibrium test





Working exercise: Tanzania dataset

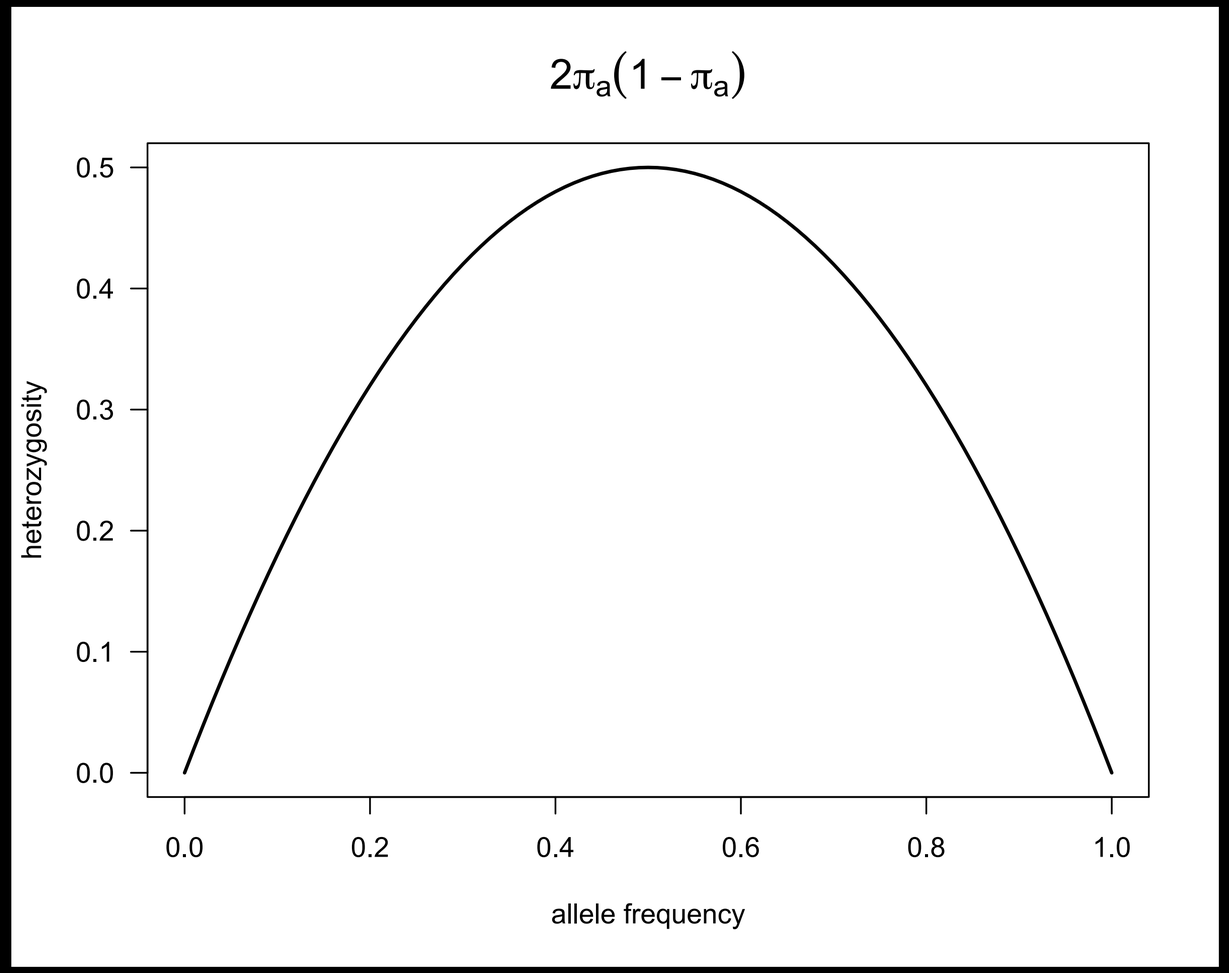
Test the Hardy-Weinberg equilibrium of the genotype distribution of rs1801033, rs6874639, and rs3024500 using the Pearson’s chi-square goodness-of-fit test. Draw your conclusions.

Note on the use of Hardy-Weinberg equilibrium test

COVID-19 study in Portugal, n=1219

Genotype	**Probability**	**Expected (MAF=0.01)**	**Expected (MAF=0.02)**	**Expected (MAF=0.025)**
aa	π_a^2	1194.7	1170.7	1158.8
ab	$2\pi_a(1-\pi_a)$	24.1	47.8	59.4
bb	$(1-\pi_a)^2$	0.1	0.5	0.8

Data quality checks: heterozygosity



Data quality checks: heterozygosity

Action:

1. Calculate the proportion of heterozygous genotypes per SNP

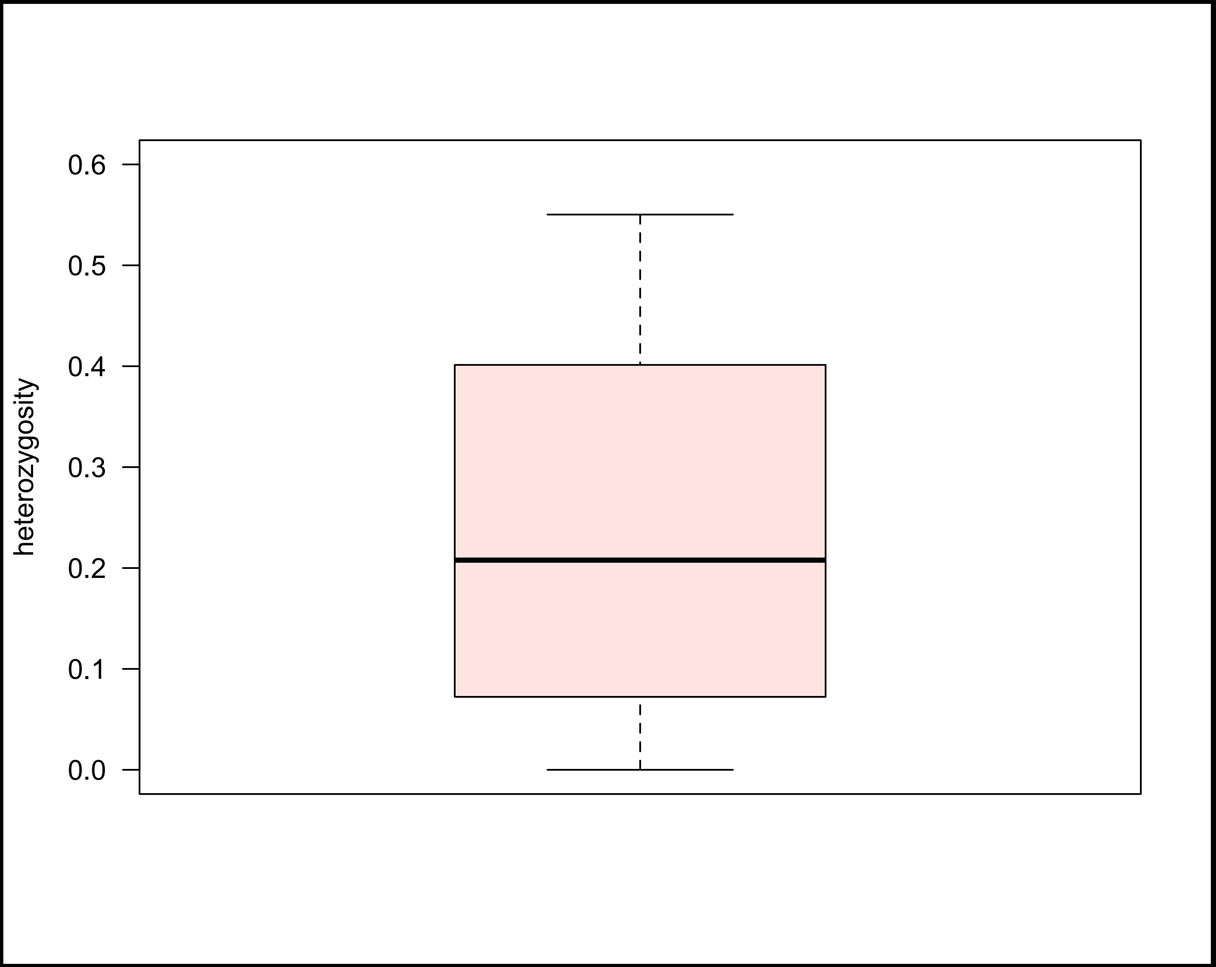
$$\pi_{ab,k} = \frac{n_{n_{ab},i}}{n} \quad \text{(heterozygosity)}$$

2. Remove all SNP with a proportion > 0.5 (or $0.5 + 1.96 \times \frac{0.5}{\sqrt{n}}$ or

$$0.5 + \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) \times \frac{0.5}{\sqrt{n}})$$

Note: SNP with a too-high level of heterozygosity are likely to fail HWE test.

Data quality checks: heterozygosity



$$\text{cutoff} = 0.5 + 1.96 \times \frac{0.5}{\sqrt{1219}} = 0.5288$$

$$\text{cutoff} = 0.5 + \Phi^{-1}(1 - \frac{0.05}{2 \times 487314}) \times \frac{0.5}{\sqrt{1219}}$$

$$= 0.5762$$

Max heterozygosity =0.5503

No SNP removed

Data quality checks: heterozygosity (X-linked SNP)

Action:

1. Calculate the heterozygosity per X-linked SNP in males only
2. Remove all X-linked SNP with a proportion > cutoff (0.01 or 0.025)

Data quality checks: biological sex

Action:

1. Calculate the X-linked homozygosity per individual
2. Predict the sex of individual
3. Check whether the predicted sex is similar to the coded/pedigree sex
4. Remove individual from the sex-mismatched individuals or use genomic-predicted sex as covariate in the analysis

Note:

Genomic data can predict sex from individuals who have this information missing.

Data quality checks: biological sex

Coded sex/Predicted sex	Male	Female	Indeterminate
Male (1)	492	2	0
Female (2)	0	706	9
Missing (0)	4	1	0

Let’s do all the data quality checks for COVID19 GWAS

Data quality checks

frontiers in Pediatrics

OPINION
published: 12 June 2020
doi: 10.3389/fped.2020.00293

Check for updates

Review of the Quality Control Checks Performed by Current Genome-Wide and Targeted-Genome Association Studies on Myalgic Encephalomyelitis/Chronic Fatigue Syndrome

Anna D. Grabowska[1], *Eliana M. Lacerda*[2], *Luís Nacul*[2,3] *and Nuno Sepúlveda*[4,5]*

[1] *Department of Biophysics and Human Physiology, Medical University of Warsaw, Warsaw, Poland,* [2] *Department of Clinical Research, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom,* [3] *Complex Chronic Diseases Program, British Columbia Women's Hospital and Health Centre, Vancouver, BC, Canada,* [4] *Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom,* [5] *CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal*

OPEN ACCESS

Edited by:
Marco Carotenuto,
University of Campania Luigi

Data quality checks

TABLE 1 | Summary of the QC checks performed in published GWAS and TGAS on ME/CFS.

Reference, type of study	Monomorphic SNPs or SNPs with low MAF	HWE	Heterozygosity	Genotyping rate
Smith et al. (5), GWAS	• The total number of monomorphic SNPs was reported • SNPs were not excluded according to MAF	• The HWE was tested using data from healthy controls alone • A significance level of 0.05 was used in the statistical tests	• Heterozygosity of SNPs in the X chromosome was used for confirming gender of the samples	• SNPs with genotyping rates <80% were excluded • Individual samples with genotyping rates <92% were repeated
Schlauch et al. (6), GWAS	• The total number of SNPs with too-low MAF was reported • SNPs with MAF<0.05 were excluded	• The HWE was tested using data from both healthy controls and patients • A significance level of 0.0008 was used in the statistical tests	• Heterozygosity of SNPs in the X chromosome was only used for confirming gender	• SNPs with genotyping rates < 95% were excluded • Individual samples with genotyping rates <95% were excluded
Herrera et al. (7), GWAS	• SNPs with MAF < 0.01 were excluded	• The HWE was tested using data from both healthy controls and patients • A significance level of 0.00001 was used in the statistical tests	• Samples with heterozygosity rate higher or lower than two standard deviations of the average heterozygosity for all samples were excluded from the analysis • Heterozygosity of SNPs in X chromosome was also used for confirming gender	• SNPs with genotyping rates <97% were excluded. • Individual samples with genotyping rates <90% were excluded
Perez et al. (8), GWAS	• SNPs with MAF <0.10 in either patients or reported in the Kaviar database were excluded.	• Not reported	• Not reported	• Not reported
Rajeevan et al. (9), TGAS	• SNPs with MAF<0.05 were excluded	• The HWE was tested using data from both healthy controls and patients • A significance level of 0.01 was used in the statistical tests	• Not performed	• SNPs with genotyping rates <80% were excluded • Genotyping rates were performed in each individual sample
Johnston et al. (10), TGAS	• SNP with MAF <0.01 were excluded	• Not reported	• Heterozygosity was reported as a QC check but there was no information about the criterium used	• Not reported

Steps of a GWAS

Data quality checks
- Minor allele frequency
- Missing data (SNP/individual)
- Hardy-Weinberg equilibrium test
- Heterozygosity
- Biological sex

Data analysis
- Data imputation (outside the scope of the course)
- Association tests (controlling for population structure outside the scope of the course)
- Manhattan plot
- Multiple testing correction (Bonferroni, false discovery rate)
- Haplotype analysis (outside the scope of the course)

Data analysis: association tests

Test association between each marker and the phenotype

Additive model (with additional covariates)

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

Report the p-value for marker tested

Check the significance of the associations adjusting for multiple testing

Check the distribution of the p-values (is it similar to a Uniform distribution?)

Great deal of computational efficiency

Data analysis: Manhattan plot

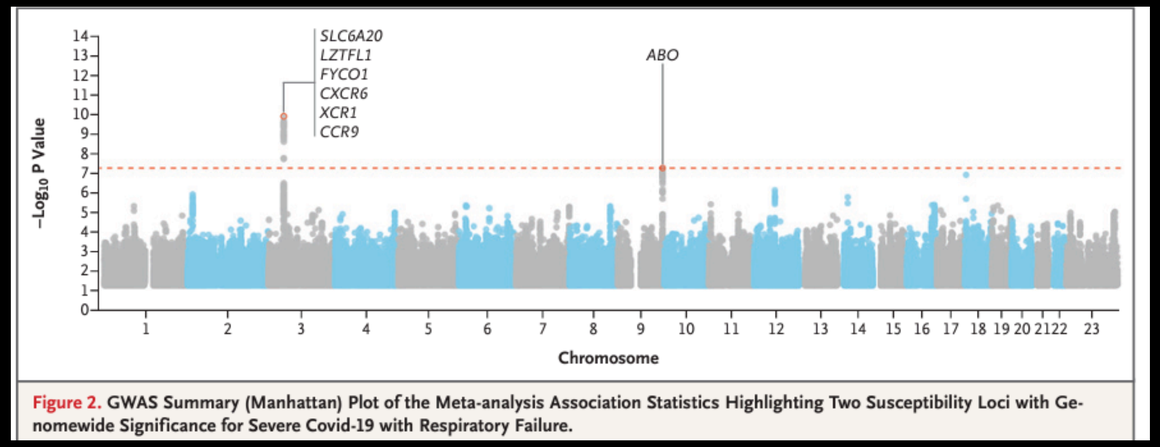


Figure 2. GWAS Summary (Manhattan) Plot of the Meta-analysis Association Statistics Highlighting Two Susceptibility Loci with Genomewide Significance for Severe Covid-19 with Respiratory Failure.

Data analysis: Manhattan plot

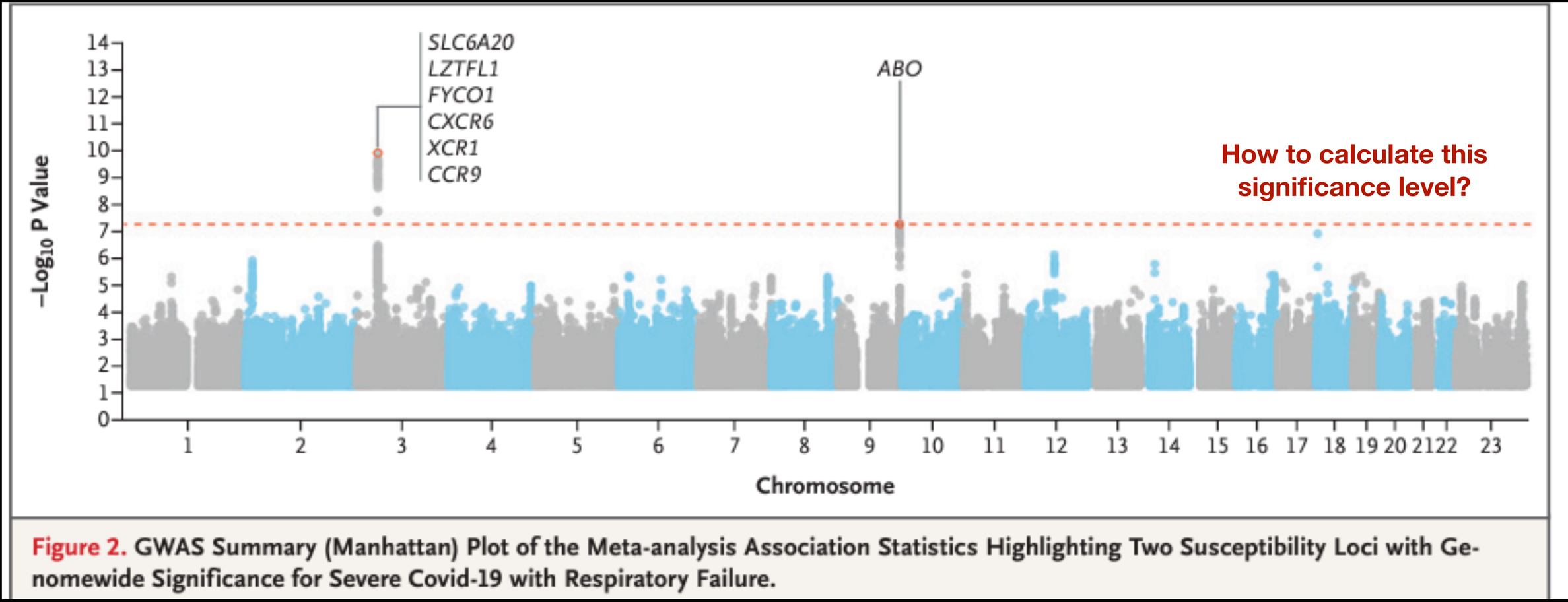


Figure 2. GWAS Summary (Manhattan) Plot of the Meta-analysis Association Statistics Highlighting Two Susceptibility Loci with Genomewide Significance for Severe Covid-19 with Respiratory Failure.

Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F, et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med. 2020;383(16):1522-1534.

Exercise: COVID19_association_tests.csv

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

Data analysis: multiple testing

In absence of association

$\alpha = 0.05$ $m =$ number of genetic markers (statistical tests)

$Y =$ Number of significant tests $| H_0, \alpha = 0.05 \rightsquigarrow$ Binomial$(m, p = \alpha)$

Expected number of false positive associations

$E[Y | H_0, \alpha] = m \times \alpha$

Data analysis: adjusting multiple testing (classical methods)

Redefine the type I error for the overall analysis

$$P[Y \geq 1 \,|\, H_0, \alpha^*] = \alpha$$

$$1 - (\alpha^*)^m = \alpha \Leftrightarrow \alpha^* = 1 - (1 - \alpha)^{1/m}$$

Sidak-Dunn correction

$$E[Y \,|\, H_0, \alpha^*] = \alpha$$

$$m \times \alpha^* = \alpha \Leftrightarrow \alpha^* = \frac{\alpha}{p}$$

Bonferroni correction

Data analysis: adjusting multiple testing (false discovery rate)

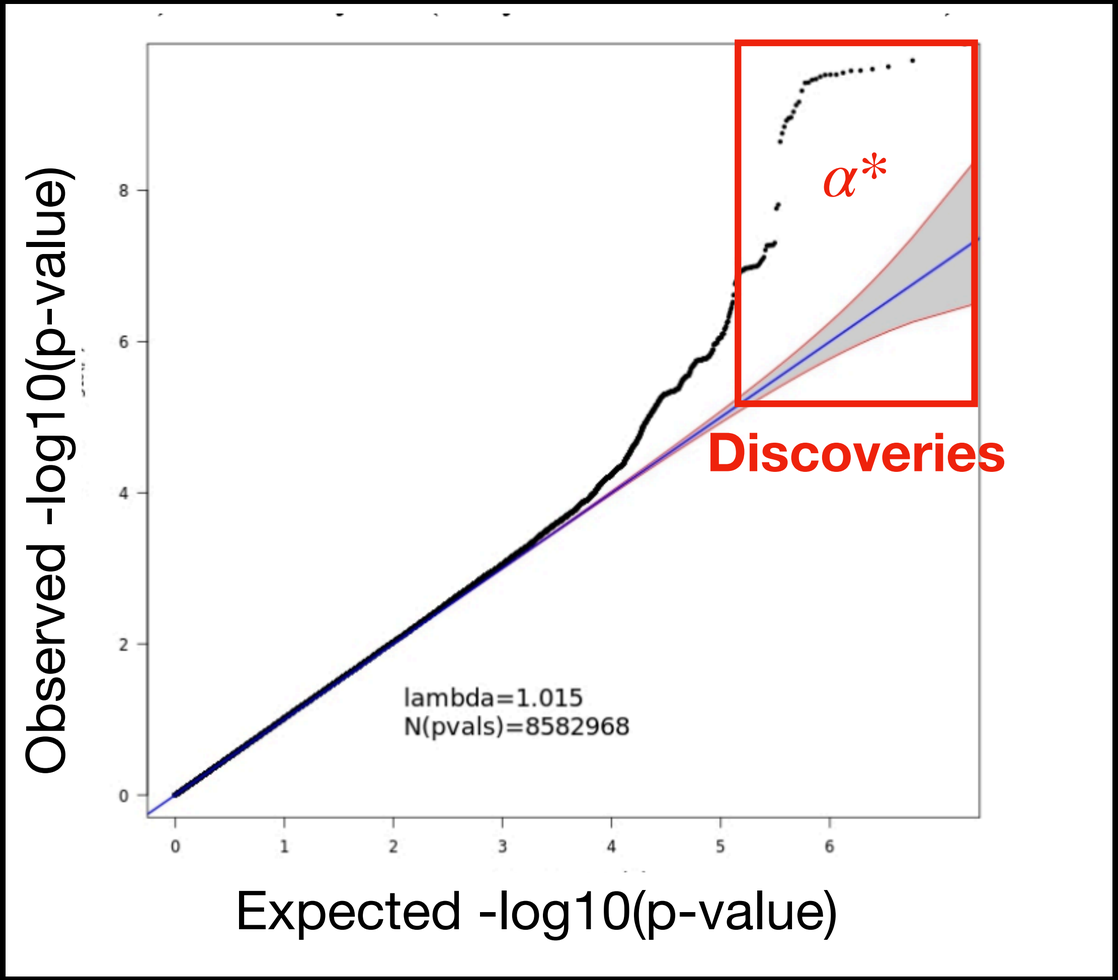
	True Positive	False Negative	Total
Significant results	m_{11}	m_{12}	m_{1+}
Non-significant results	m_{21}	m_{22}	m_{2+}
Total	m_{+1}	m_{+2}	m

Data analysis: adjusting multiple testing (false discovery rate)

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

$$E\left[\frac{m_{11}}{m_{11} + m_{21}}\right] = \frac{q}{\alpha} = \alpha^* \qquad \alpha^* = 0.05$$

Data analysis: visual idea



Data analysis: Benjamini-Hochberg 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)

Data analysis: 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 $p_{(i)}^{adj} = \min \left\{ 1, \min_{j \geq i} \frac{m p_{(j)}}{j} \right\}$

   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

Data analysis: other variants of Benjamini-Hochberg procedure

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 MASS - BH and BY procedures

Package mutoss - other procedures for controlling FDR

Exercise: COVID19_association_tests.csv

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

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

Data analysis: data imputation

Impute missing genotypes from a given SNP using information of genetic linkage on the flanking SNP

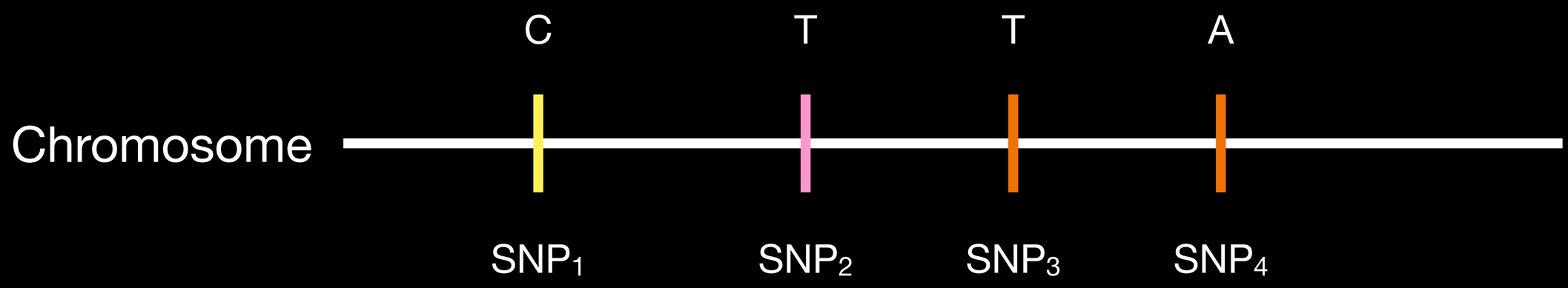
Extend SNPs under analysis by predicting the genotypes of untyped SNP using information

Development of statistical data imputation methods specific for SNP data

Softwares
Beagle, Impute, EagleImp

Data analysis: haplotype analysis

Combine allele composition from flanking SNP genetically linked to the associated SNP. This forms haplotypes

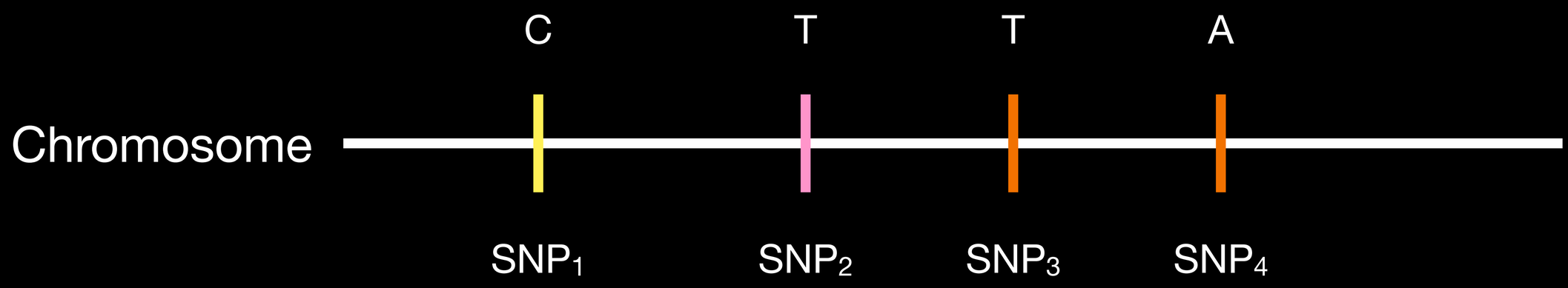


Haplotype formed by four genetically-linked SNP

Haplotype could be the causal factor of the phenotype

Data analysis: haplotype analysis

Combine allele composition from flanking SNP genetically linked to the associated SNP. This forms haplotypes



Haplotype formed by four genetically-linked SNP

Haplotype could be the causal factor of the phenotype

Final word on reproducibility