Introduction to GWAS

Data Pre-Processing: Initial & Exploratory Data Analysis

Christian Werner

(Quantitative geneticist and biostatistician) **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini

HerrFalloppio

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)

Oscar González-Recio

OscarGenomics

(Computational biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)

Initial data analysis mainly focuses on data cleaning, a first screening, and transformation (if necessary) to ensure data quality and confirm that our data set meets the relevant distributional and model assumptions. Data cleaning may include steps such as elimination of duplicate records, handling of missing values, identification of systematic errors, or correction of coding inconsistencies.

Exploratory data analysis is used to examine data sets and summarize their main characteristics. EDA helps to discover patterns in the data, spot anomalies and outliers, test a hypothesis and check our assumptions. EDA tells us what data can reveal beyond the formal modeling or hypothesis testing and provides a better understanding of data set variables and their interactions. It can also help to determine if the statistical techniques you are considering for data analysis are appropriate.

Initial and Exploratory Data Analysis

(“crap in, crap out”)

Initial and Exploratory Data Analysis – IDA & EDA

Before we conduct a GWAS, there are two types of data to explore

- **Genotypic data**
- **Phenotypic data**

Genotypic data

IDA & EDA - genotypic data

Various metrics – statistics are performed either across SNP or across samples

Key concept: detect SNP and samples that should be removed prior to GWAS

Use of some metrics and thresholds is species- and population-specific

Still some level of subjectivity in the thresholds

IDA & EDA - genotypic data

Also referred to as Quality Control (QC)

Some parameters to look at…

- Genotype calling and signal intensities (not covered here…)
- Marker allele frequencies
- Missing rate per marker and per individual
- Hardy-Weinberg equilibrium
- Heterozygosity

IDA & EDA - genotypic data

Missing rate per marker & per individual

SNPs might be of poor quality if their genotyping failed in many individuals

- Should be investigated separately for all study groups (if known)
- Common SNP array thresholds are 2 – 10% (based on sample size & SNP number)

Sample DNA might be of poor quality if there are many missing SNPs in an individual

- Too many missing SNPs per individual can be an indication of poor DNA quality
- (or true deletions…)
- Common SNP array thresholds are between 2 – 10% (based on sample size & SNP number)
- Includes monomorphic SNP !!

IDA & EDA - genotypic data

Marker allele frequencies

- Allele counts & genotype counts
- Minor allele frequency (MAF)
  - Some SNPs will be monomorphic
  - One of the alleles may be at very low frequency
  - Might be due to genotyping errors
  - Power to detect the association is very low

Common MAF thresholds are between 1 – 3% (but think about what's reasonable in your data set!)

In samples with known group structure, MAF should be checked within groups

IDA & EDA - genotypic data

Minor allele count - an alternative to MAF

- Genotyping platforms have become extremely accurate and may not justify a filtering based on a “relatively high” expected genotyping error rate anymore.
- Minor allele count is based on the idea that you need a certain number of samples (alleles) to reliably estimate the mean effect of the allele.
- “Think about it in terms of a *t*-test” - you want to check if two alleles have a different effect on a trait.
- ***Rule of thumb***: 30-40 counts of the minor allele

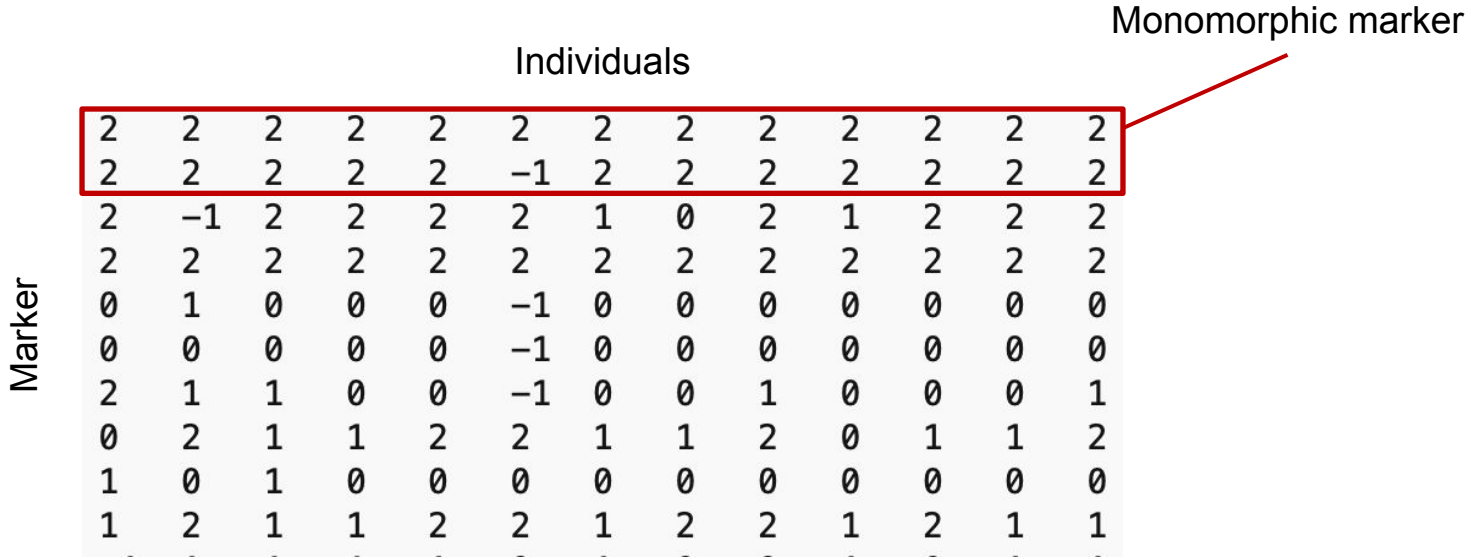
Exploratory data – genotypes

Individuals

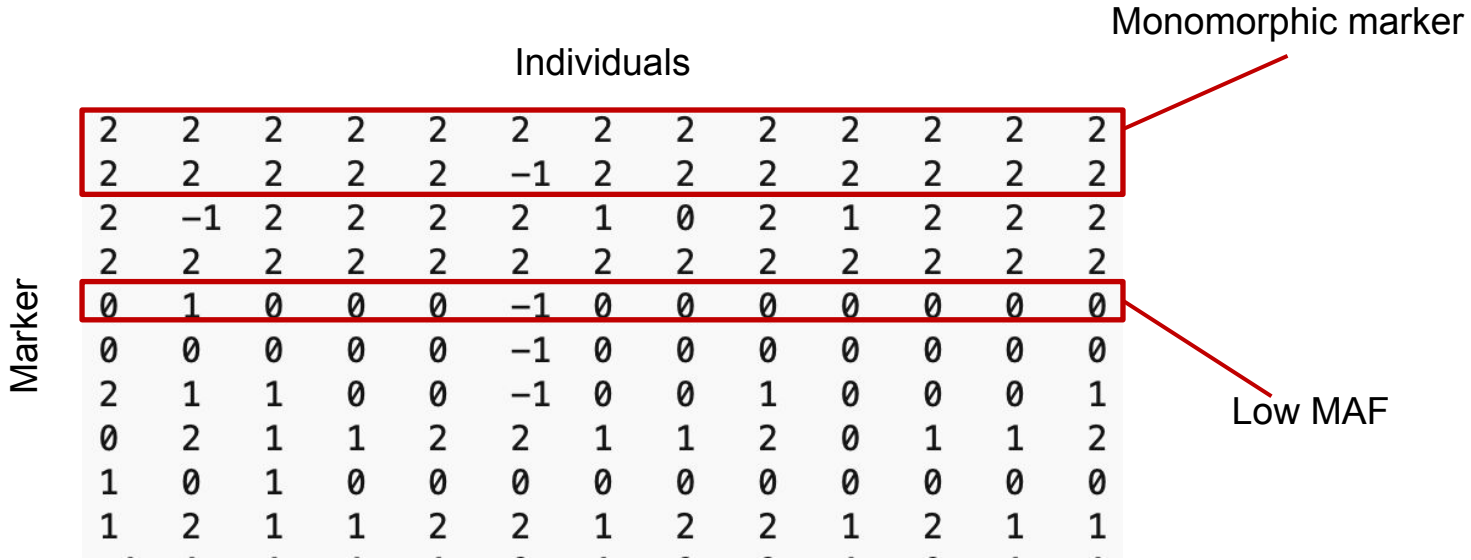
Marker

2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	−1	2	2	2	2	2	2	2
2	−1	2	2	2	2	1	0	2	1	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2
0	1	0	0	0	−1	0	0	0	0	0	0	0
0	0	0	0	0	−1	0	0	0	0	0	0	0
2	1	1	0	0	−1	0	0	1	0	0	0	1
0	2	1	1	2	2	1	1	2	0	1	1	2
1	0	1	0	0	0	0	0	0	0	0	0	0
1	2	1	1	2	2	1	2	2	1	2	1	1

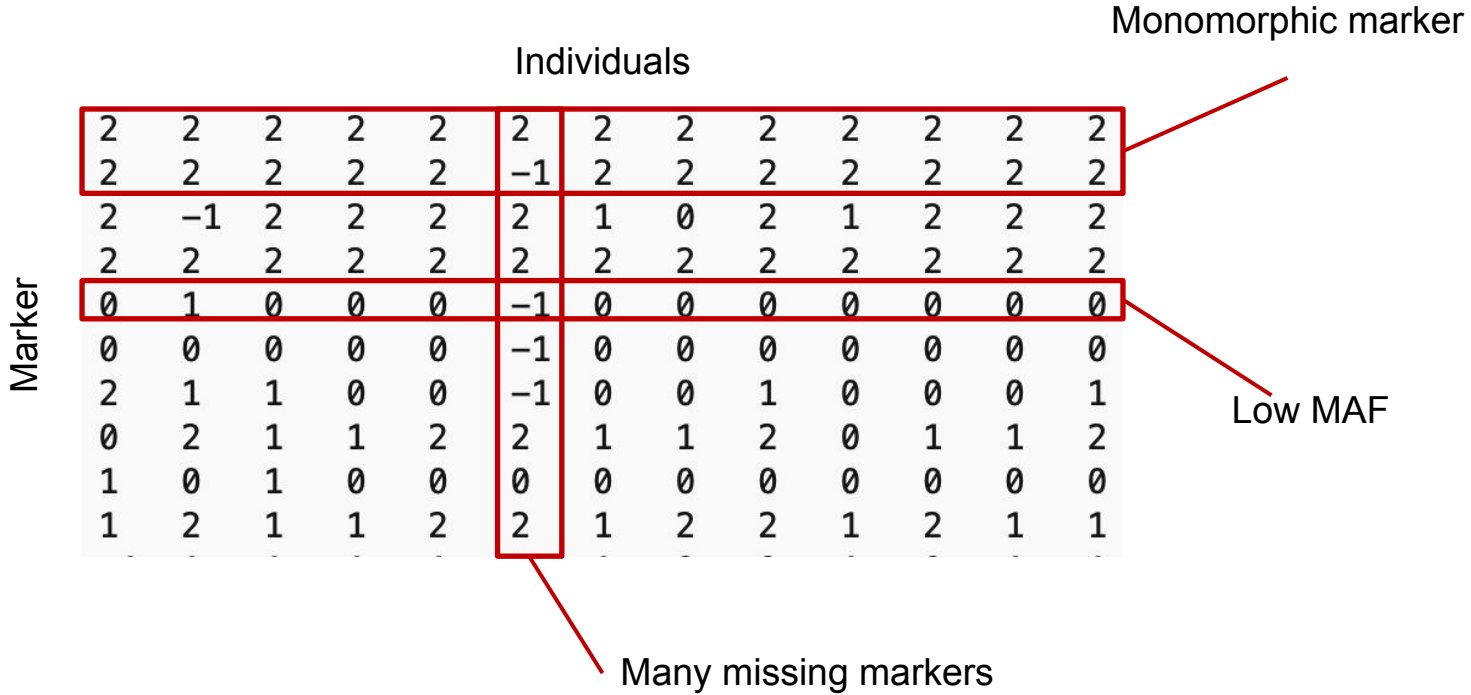
Exploratory data – genotypes



Exploratory data – genotypes

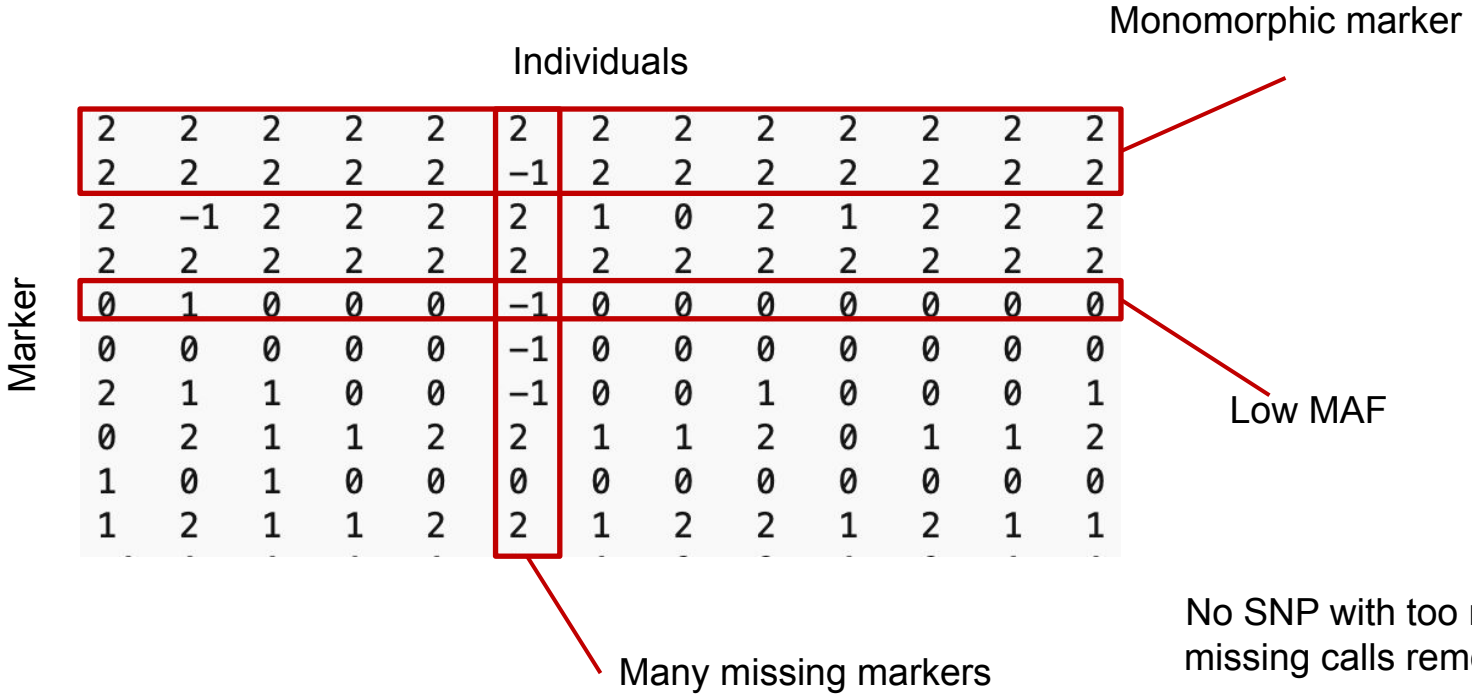


Exploratory data – genotypes



2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	−1	2	2	2	2	2	2	2
2	−1	2	2	2	2	1	0	2	1	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2
0	1	0	0	0	−1	0	0	0	0	0	0	0
0	0	0	0	0	−1	0	0	0	0	0	0	0
2	1	1	0	0	−1	0	0	1	0	0	0	1
0	2	1	1	2	2	1	1	2	0	1	1	2
1	0	1	0	0	0	0	0	0	0	0	0	0
1	2	1	1	2	2	1	2	2	1	2	1	1

Exploratory data – genotypes

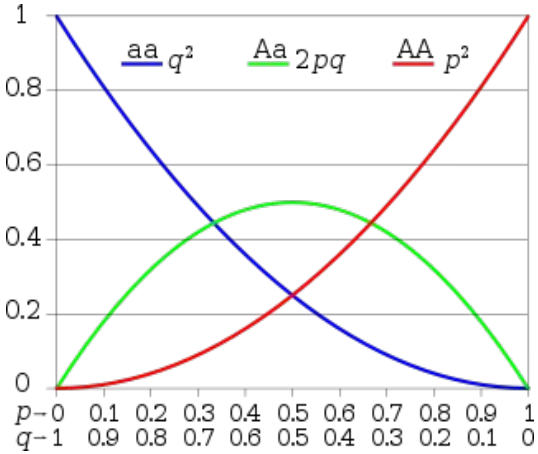


IDA & EDA - genotypic data

Hardy-Weinberg equilibrium: relationship between allele and genotype frequencies

$$(p + q) = p^2 + 2pq + q^2$$

- diploid genomes
- autosomal loci
- large population
- random mating
- equal frequencies in both sexes
- no selection
- no migration
- no mutations



IDA & EDA - genotypic data

Deviations from HW equilibrium

- systematic genotyping errors
- violation of assumptions

Test for HW deviation

- chi -squared (χ^2) test
- Fisher's (exact) test
- many, many more ...

But also selection, assortative mating, population structure and inbreeding cause deviations from HWE!

HWE is, in most cases, NOT a reasonable assumption...

IDA & EDA - genotypic data

Heterozygosity

- Proportion of heterozygotes
- Heterozygosity can be checked per locus & per marker

Very high sample heterozygosity can be an indication of DNA contamination

- But also could be that a small proportion of samples are truly very different from the rest…
- Removal of samples that depart ± 3 SD from the mean

Very high heterozygosity per marker could also indicate poor DNA quality, but also be due to...

- the breeding scheme (e.g. hybrid breeding in plants, or very low heterozigosity in lines)
- Genome duplications

Phenotypic data

IDA & EDA - phenotypic data

Data type

- Continuous (e.g. height)
- Binary (e.g. case/control)
- Categorial (e.g. scores (ordered), eye colour (ordered))

Measure of centrality: mean, mode, median

Measures of dispersion: range, variance, standard deviation

Distribution of the data

- Distribution of values as expected? Outliers?
- representative sample of the population?
- Other explanatory covariables

IDA & EDA - phenotypic data

Covariables

Are there any variables which may have a relationship with the phenotype?

- E.g. sex, breed, age, treatments, year effects, … (population structure)

The data needs to be corrected for these effects. Otherwise they can be confounded with allelic variants with an effect on the phenotype which we try to identify.

Covariables with significant effects on the phenotype can be identified using ANOVA (requires balanced datasets - ANOVA is outdated) or **linear mixed models**. However, a comprehensive preparation of phenotypic data including model comparison is not covered here…

IDA & EDA - phenotypic data

Assumptions for continuous variables (different for binary traits…)

- Normally distributed **residuals** (prerequisite of GWAS model assumptions)
- Homogeneity of variance (differences in variance might indicate a factor that has not been included in the phenotype processing)

IDA & EDA - phenotypic data

Outliers

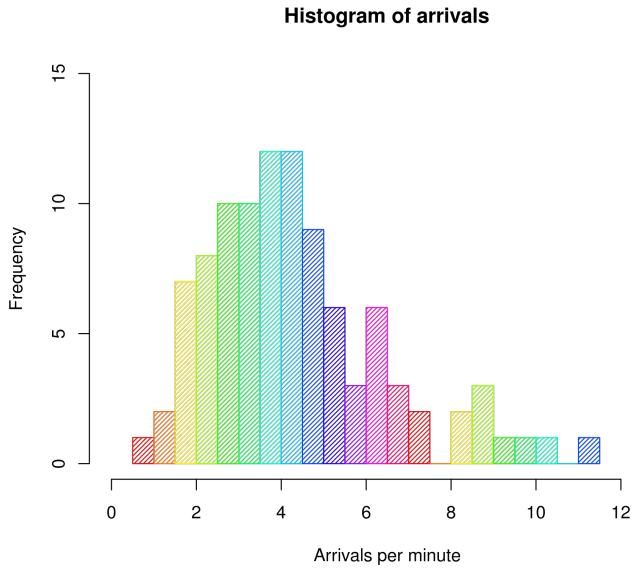
- Apparently rare phenotypes are often a result of errors or poor models rather than true outliers
- However, the values might be real - outliers should be investigated thoroughly rather than relying on statistical tests

Data transformation

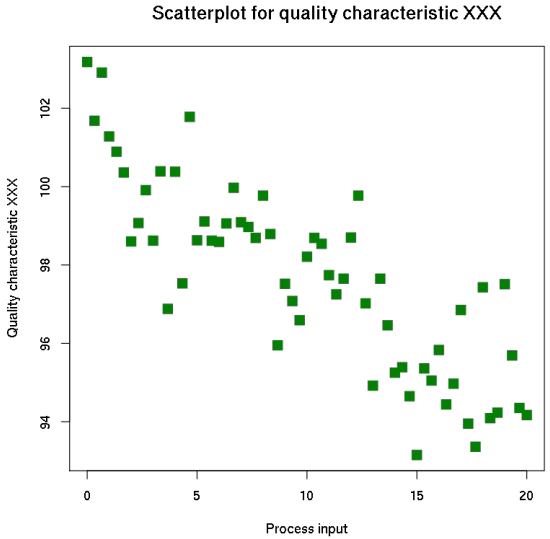
- **Positively skewed distributions of the residual** with the long tail into the positive direction can be corrected with a logarithmic or square root transformation.
- **Negatively skewed distributions of the residual** that have a long tail in the negative direction can be corrected with cubing or squaring
- Transformations only if really necessary…

IDA & EDA - phenotypic data

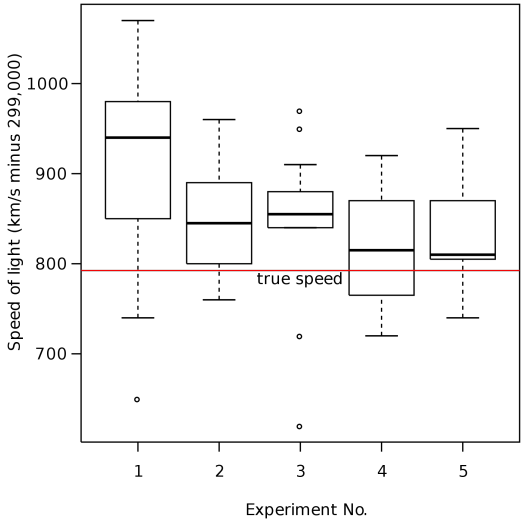
Histograms,



Scatter plots,



Box plots,



IDA & EDA - phenotypic data

Normality of residuals

- Look at the data (Histogram, QQ-Plot)
- Don’t rely on Shapiro-Wilk test

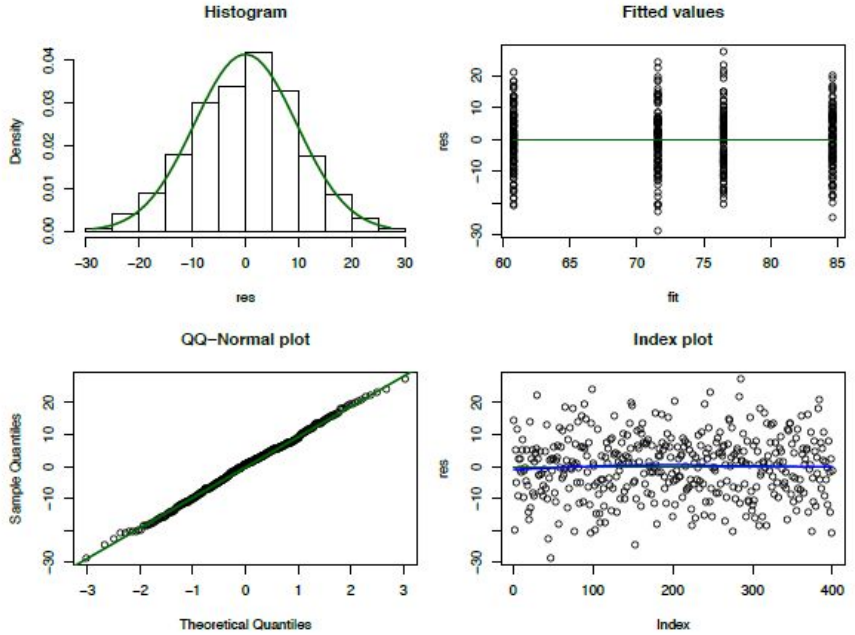
Variance homogeneity

- Look at the data (Scatterplot)
- Don’t rely on Leven’s test

Tests are conservative and might indicate a violation of the assumptions of normality and variance homogeneity in a suitable “real-world” dataset.

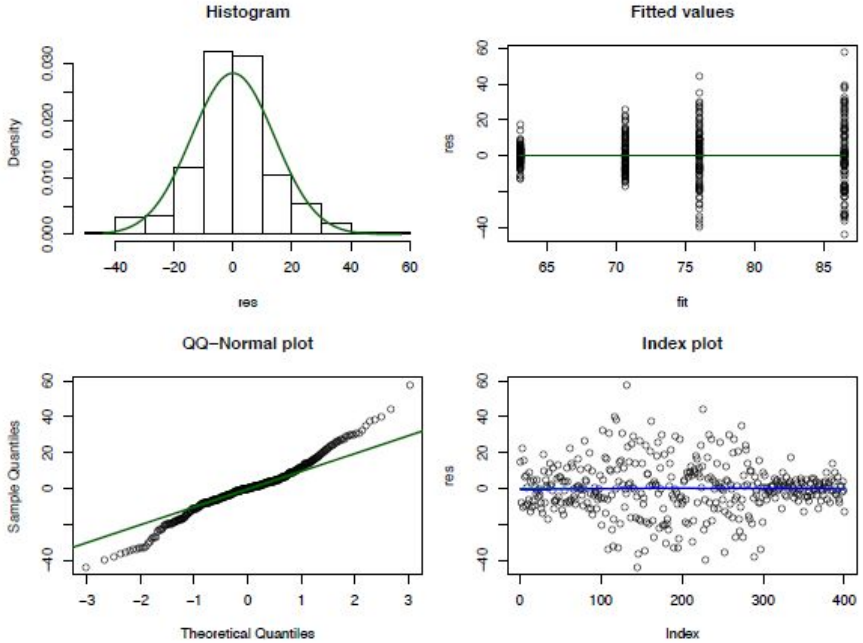
IDA & EDA - phenotypic data

Almost perfectly distributed residuals



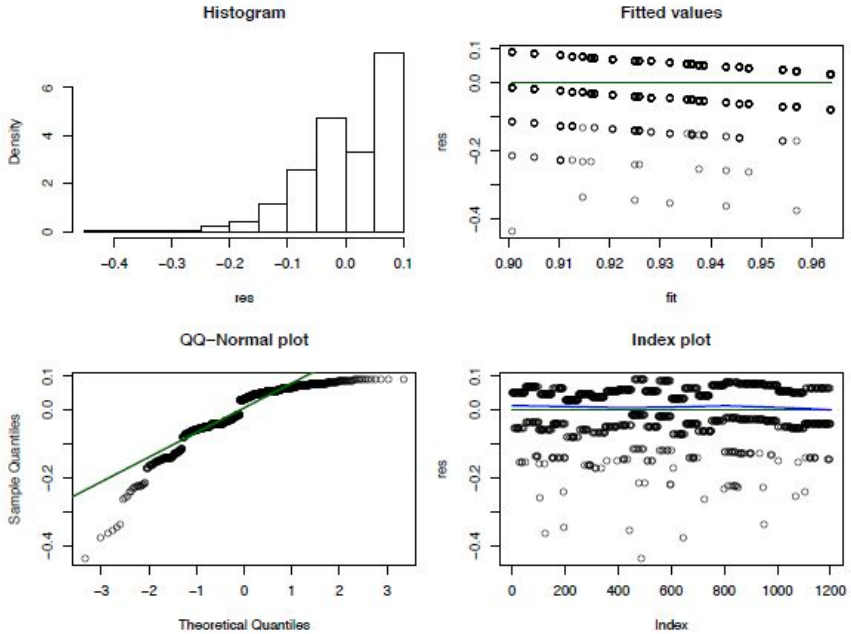
IDA & EDA - phenotypic data

Still good...



IDA & EDA - phenotypic data

There might be something wrong here....



Summary of pre-processing for genotype data

Pre-processing genotypic data – Standard procedures

missing rate

- per-sample (e.g. max 10% missing SNP genotypes per sample)
- per-site (e.g. max 5% missing genotype per variant/SNP)
- stricter/looser thresholds depending on data/experiment (e.g. SNP array, GBS, quality of reference sequence, ...)

MAF

Hardy Weinberg equilibrium

others

Pre-processing genotypic data – Standard procedures

missing rate

MAF

- remove monomorphic variants → non-informative
- remove variants at low frequency (“rare”) → spurious associations
- threshold may depend on sample size
- usually (re)done after imputation

Hardy Weinberg equilibrium

others

Pre-processing genotypic data – Standard procedures

missing rate

MAF

Hardy Weinberg equilibrium

- set **low threshold** for p-value (e.g. exp(-10))
- **questionable**: some of the forces driving HW disequilibrium are characteristic for breeding (selection, migration, mutation, adaptation etc.)

others

Pre-processing genotypic data – Standard procedures

missing rate

MAF

Hardy Weinberg equilibrium

others

- sex chromosomes (might need to be removed / analyzed separately)
- Mendelian errors
- quality scores (vcf files)
- relatedness (between samples - check for duplicates)

Imputation of missing genotypes

Imputation of missing genotypes – why?

Imputation - the process of replacing missing data with substituted values

Preliminary step for a wide range of genetic analyses

Most models and software for population genetics, genomic selection (GS) and genome-wide association studies (GWAS) do not handle missing data by default and require complete datasets

1. **Genotyping techniques generate a proportion of missing data (uncalled genotypes)**
   - SNP arrays ~**5%**
   - RAD-Seq (e.g. GBS) ~**50%**

2. **Optimization/efficiency of genotyping strategies (low → high density data)**
   scaling-up: **low → high density** (mixed genotyping strategies)
   whole-genome sequence imputation

Imputation of missing genotypes – methods

1. **General methods for the imputation of any type of data**
   - mean substitution (replacing missing values with the mean of the SNP across the population), median imputation
   - K-Nearest Neighbour Imputation (KNNI)
   - many more ...

2. **Methods specific for the imputation of missing genotypes**

   Two groups (and combinations of them):

   - based on LD and allele frequency
   - based on pedigree information