Introduction to **GWAS:**
Linkage disequilibrium and Linear regression

Christian Werner
(Computational and quantitative geneticist) **EiB CIMMYT**, Texcoco (Mexico)

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

Oscar González-Recio
OscarGenomics
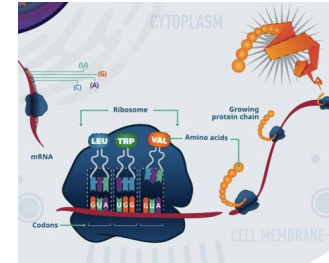*(Computational biologist and quantitative geneticist)* **INIA-UPM**, Madrid (Spain)

A quick overview of GWAS

Genetic basis of GWAS-GWP

Phenotypic variation

Genetic variants

- Allele substitution effect

- Infinitesimal model

Genetic basis of GWAS-GWP

Phenotypic variation

Genetic variants

$$P = G + E$$

- Continuous trait

- Categorical trait (threshold model)

Genetic basis of GWAS-GWP

Phenotypic variation

Genetic variants

Linkage disequilibrium

$$P = G + E$$

Genotype-Phenotype association → causal or **functional** link

Genetic basis of GWAS-GWP

- A study that agnostically tests hundreds of thousands of single nucleotide polymorphisms (SNPs) densely spaced across the genome for association with a given disease or trait.

- Rationale:
  - Not limited by a priori knowledge of disease process or the results of linkage studies
  - Preferable approach (so far) for diseases of complex etiology
  - Evidence for functional SNPs outside of coding regions

"*First, our studies provide convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including non-protein-coding transcripts, and those that extensively overlap one another*." ENCODE, Nature June 2007

Genetic basis of GWAS-GWP

- Exposure:
  - SNPs
  - Indels
  - *Might* - copy number variants (CNV), haplotypes, other type of markers
- Outcome: easily adaptable to either dichotomous or quantitative outcomes
- Study Populations: Case-control, cross-sectional, or cohort. Family or population based.
  - Need to have large sample sizes!
- Analytic methods: logistic/linear regression
  - Flexibility for different genetic models - additive, dominant, recessive.

Genetic basis of GWAS-GWP

- Genotyping – Since early 2000 (25-100€)
- Genome sequencing - (approx. 1000 €)

- Quantitative genetics – many genes – need many markers
- Use LD between gene and marker
  - We assume that all genes are in LD with at least one marker (we need a high coverage throughout the genome)
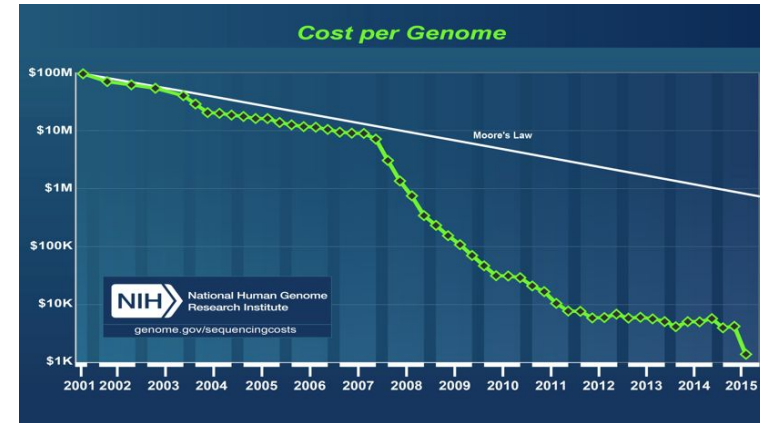
Costs matter - approximate current **SNP array prices**

SNP chip	n. SNP	approximate price
Cow BovineLD	7,900	**$32**
Cow Bovine SNP50	53,000	**$50**
Cow BovineHD	777,000	**$150**
Dog CanineHD	172,000	**$125**

Commercial SNP chips

MaizeSNP50 BeadChip, **GeneChip**® **Rice 44K SNP** $\rightarrow \sim$ **100$**
per sample

Genome-wide association studies (GWAS)

1. Genotyping costs have decreased dramatically over the last 10 years, in most species
2. Large number of genotyping experiments

   → explosion of GWAS experiments!

Powerful (*supposedly*) approach to the identification of genes/**genomic regions involved in plant, animal and human phenotypes**

Genome-wide association studies (GWAS)

association is one type of statistical problem

- discovery of interesting relationships among variables in large data sets (i.e., **association**);

- division of data sets into several discrete groups (i.e., **clustering**);

- assignment of observations to groups (i.e., **classification**);

- Extrapolate quantitative outputs based on attributes of observational units (i.e., **prediction**);

- etc.

Inference vs Prediction

Inference

- Determine the effect of a covariate on the response

- Determine the <u>causal</u> relationship between a covariate and the response

- More difficult (in general)

Prediction

- Educated <u>guess</u> of the outcome

- Expected behaviour in the future

- Based on proxies/markers

Inference vs Prediction
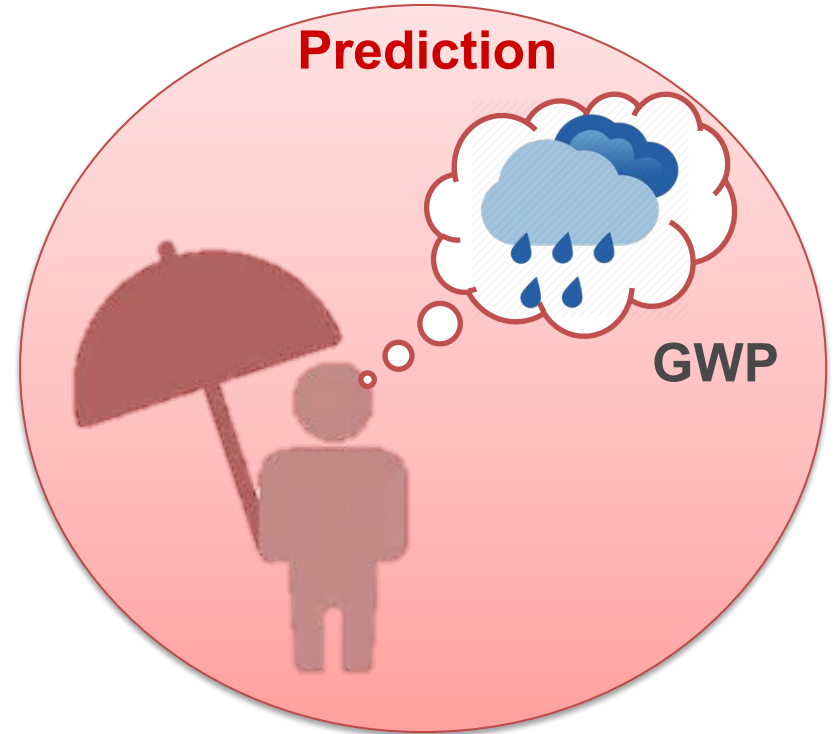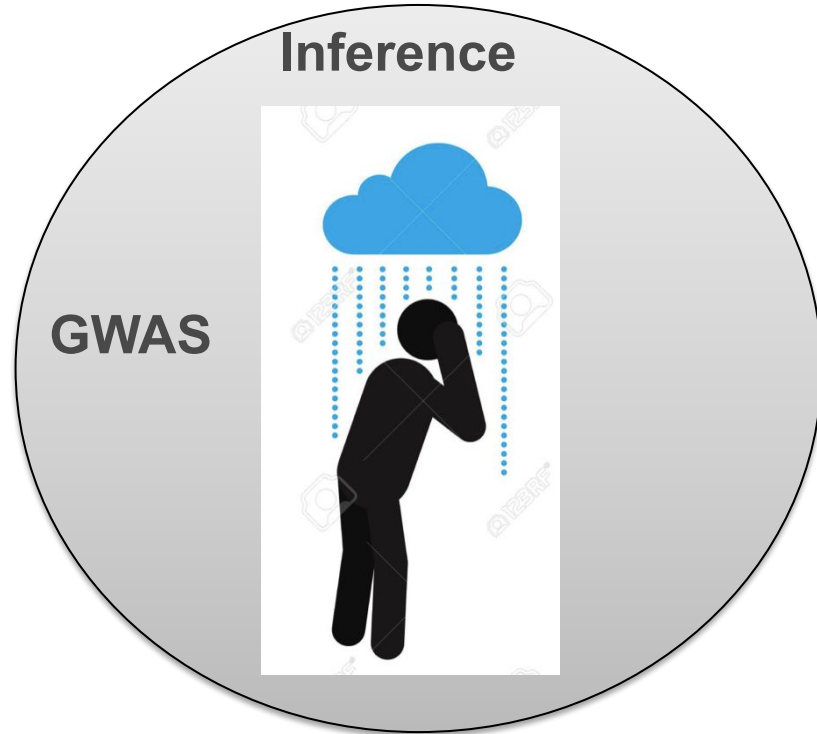
Inference vs Prediction

- Know the past
- Predict the future
- Act consequently

Genetic basis of GWAS-GWP

GWAS goal

- Detect genomic markers/regions associated to phenotypes (traits) of interest

- Find biological pathways of interest

- Interaction between treatments/drugs and genes

GWP goal

- Calculate the genomic risk score /predisposition of a disease or trait

- Calculate the genomic merit of individuals

- Predict future performance

- Sire/dam selection in animal breeding

Some examples
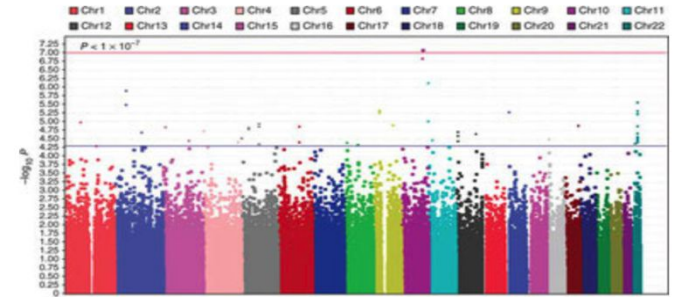
GWAS – Rheumatoid arthritis

BMC Proceedings

Home About Articles Submission Guidelines

Volume 3 Supplement 7

Genetic Analysis Workshop 16

Proceedings | Open Access | Published: 15 December 2009

Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model

Oscar González-Recio, Evangelina López de Maturana, Andrés T Vega, Corinne D Engelman ✉ & Karl W Broman

BMC Proceedings 3, Article number: S63 (2009) | Cite this article

504 Accesses | 4 Citations

GWAS – **Milk yield dairy cattle**
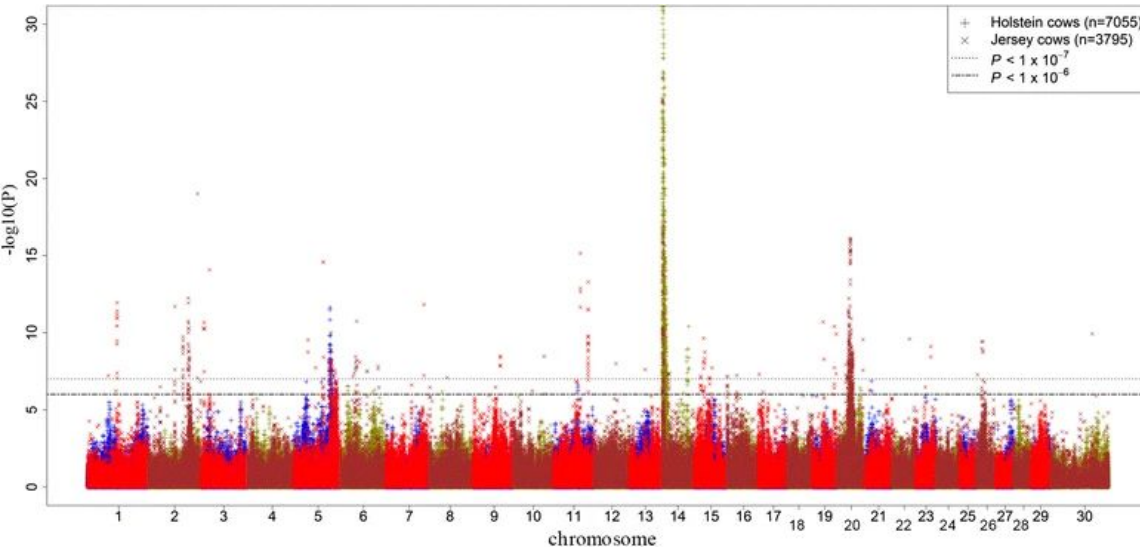
- **DGAT1** gene on BTA 14 in dairy cows (HOL and JER)
- milk fat content

GWAS – Fertility dairy cattle

A number of
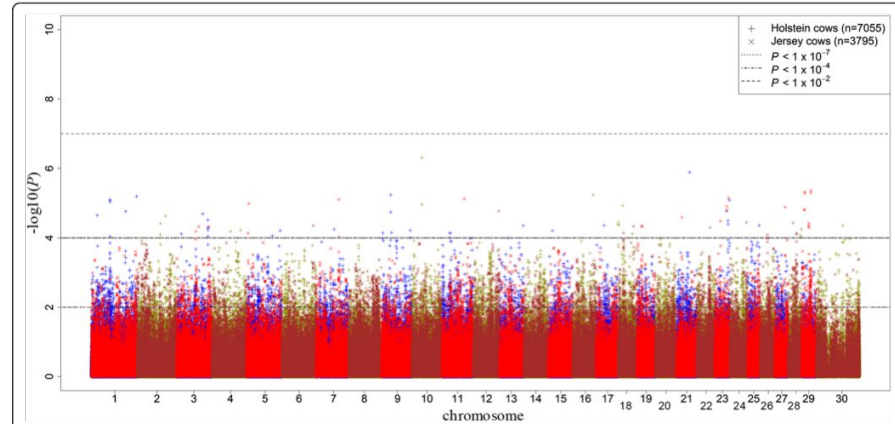**fertility-related** traits
(additive and
dominance)

Fig. 4 Distribution of dominance SNP effects for fertility. Manhattan plot of all dominance SNP effects for calving interval in discovery and validation populations with chromosome number on horizontal axis and $-\log_{10}(P\text{-value})$ on vertical axis
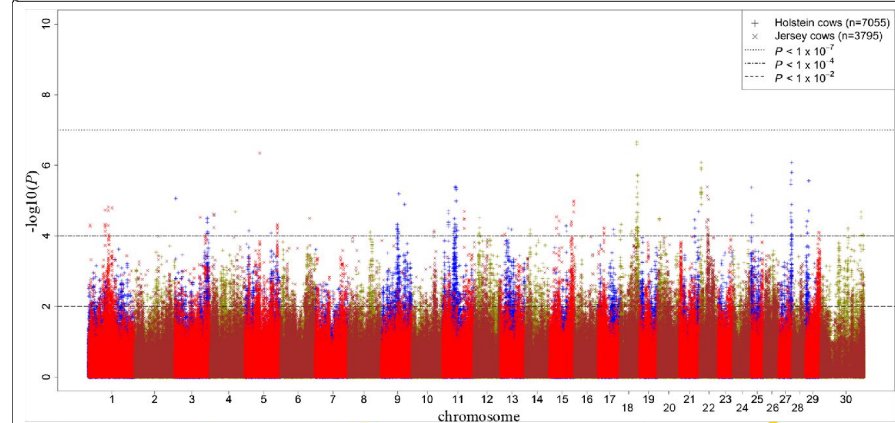

Fig. 2 Distribution of additive SNP effects for fertility. Manhattan plot of all additive SNP effects for calving interval in discovery and validation populations with chromosome number on horizontal axis and $-\log_{10}(P\text{-value})$ on vertical axis
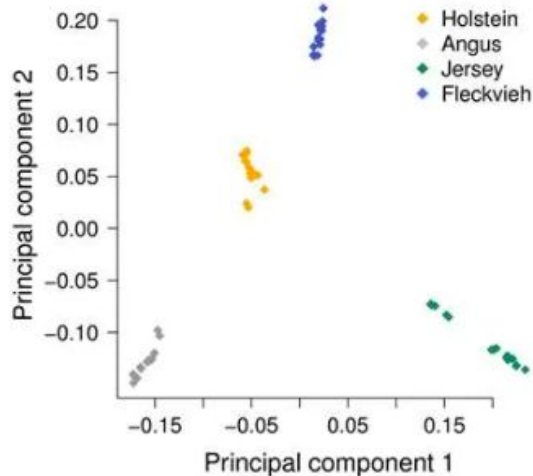
GWAS –curly hair in cattle
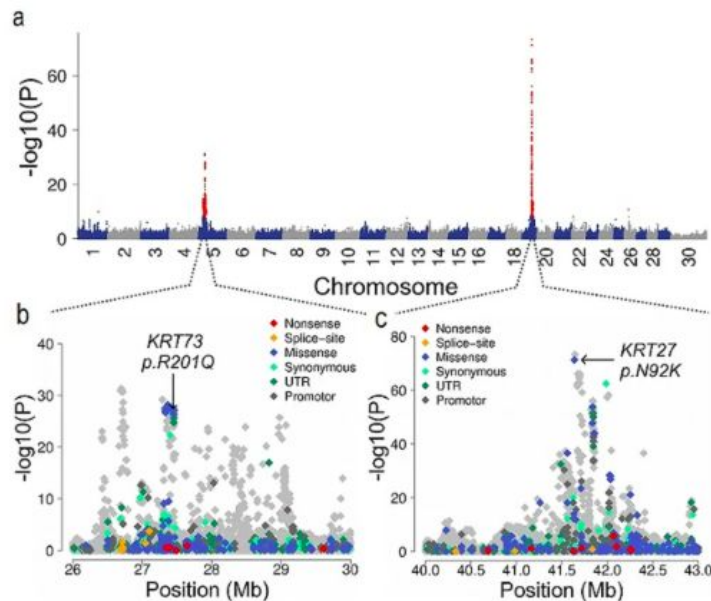
Manhattan plot showing the association of 17,640,970 imputed variants with the proportion of daughters with curly hair in 3222 Fleckvieh bulls (**a**). Red dots represent variants with P < 10⁻⁹. Detailed overview of the associated regions on chromosomes 5 (**b**) and 19 (**c**). Variants in the promoter (defined to encompass 1,000 bp upstream of the transcription start), in the untranslated regions (UTR) and in the amino acid coding region are highlighted with different color. The associated region on BTA5 encompasses Krt71, which underlies curly hair in various species. Variant calling yielded four missense mutations in Krt71 (p.R133W, p.F143I, p.N177I, p.P452H); however, none of them was polymorphic in the 43 resequenced Fleckvieh animals. Functional annotation of the variants within the QTL region on BTA5 revealed that 12 closely linked missense mutations in Krt73, Krt2 and Krt76 are highly significantly associated with curly hair in Fleckvieh cattle. Among those, only the p.R201Q mutation in Krt73 (c.G602A, chr. 5: 27,445,800 bp, ss682156288) was predicted to be damaging by PolyPhen-2 and SIFT analysis.

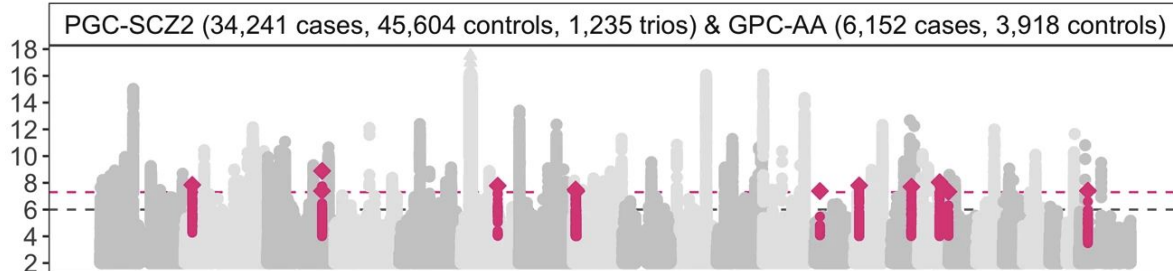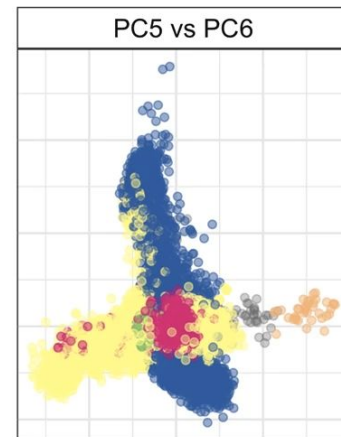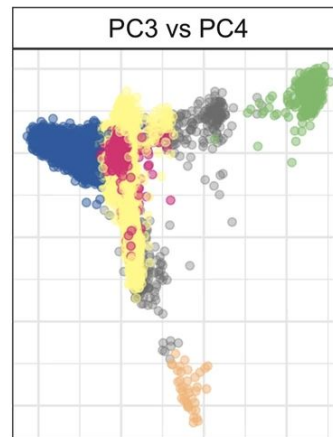# GWAS - Schizophrenia

Molecular Psychiatry
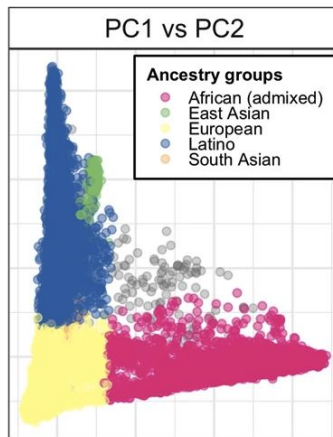
Article | Open Access | Published: 07 October 2019

Contributions of common genetic variants to risk of schizophrenia among individuals of African and Latino ancestry

Tim B. Bigdeli ✉, Giulio Genovese, [...] Carlos N. Pato

Molecular Psychiatry (2019) | Cite this article

2537 Accesses | 1 Citations | 63 Altmetric | Metrics

GWAS - Covid-19

nature

Explore content ∨   Journal information ∨   Publish with us ∨
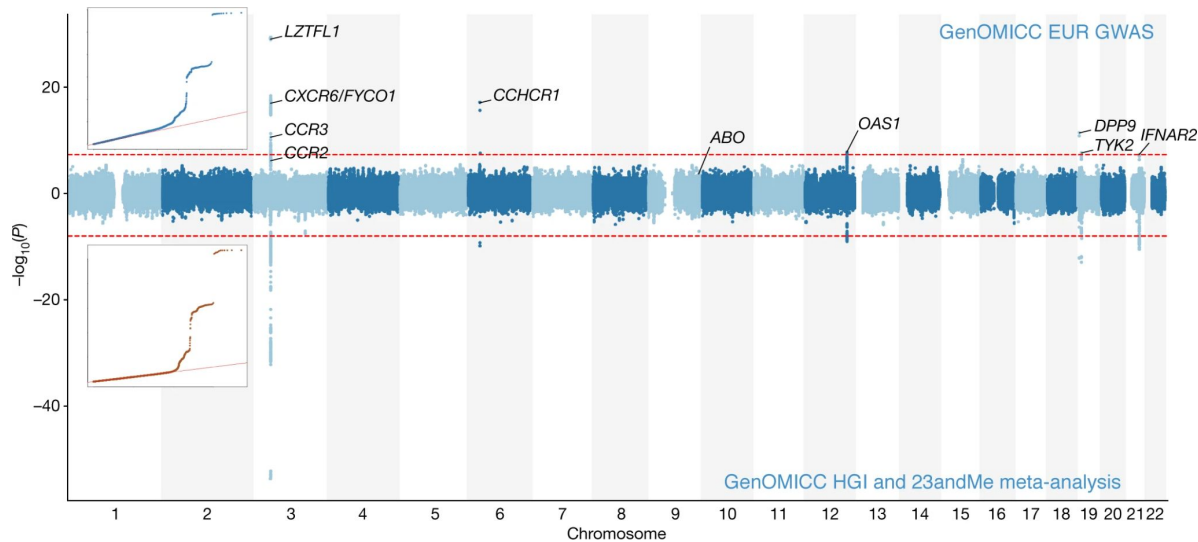
nature › articles › article

Article | Published: 11 December 2020

Genetic mechanisms of critical illness in COVID-19

Erola Pairo-Castineira, Sara Clohisey, [...] J. Kenneth Baillie ✉

Nature **591**, 92–98 (2021) | Cite this article

240k Accesses | 55 Citations | 2731 Altmetric | Metrics

Doing a GWAS is "easy"

Doing a GWAS is "easy"

Doing a GWAS CORRECTLY is not

What will you learn

- How GWAS work
- Use the right type of analyses
- Identify and understand the individual steps involved in a GWAS project
- Understand the limitations of GWAS
- Visualize results of GWAS
- Assemble the different steps into a reproducible pipeline

Regarding GWAS - what we will do

0. getting the data
1. data preprocessing **(EDA)**
2. data preprocessing **(filtering)**
3. **imputation** of missing genotype data
4. **GWAS** basic models
5. single SNP vs many SNP
6. continuous/binary traits
7. population structure
8. Manhattan plots/qq-plots (post-hoc analysis)
9. build the **pipeline**

NEXT LECTURE

Introduction to GWAS: Linkage disequilibrium and Linear Regression