Introduction to GWAS:
A quick overview

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)

Something you need to carefully look at, or that may impair your GWAS

Something to do, or that optimizes your GWAS

Don’t. Discourage to use this.

Smart tip. Something that makes the trick.

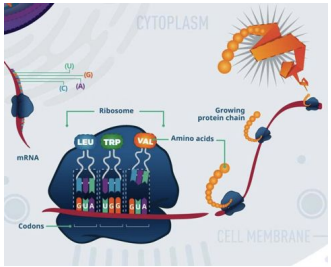
Advanced. Something to dive in.

Genetic basis of GWAS-GWP

Phenotypic variation

Genetic variants

- Allele substitution effect



- Infinitesimal model

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
  - If large SNP density, each gene is expected to be in LD with at least 1 SNP
  - 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

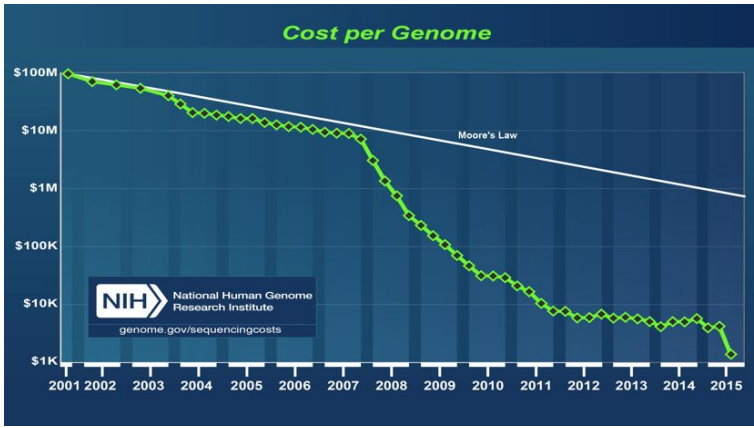
Genetic basis of GWAS-GWP

- 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)



conduction a GWAS

Study population

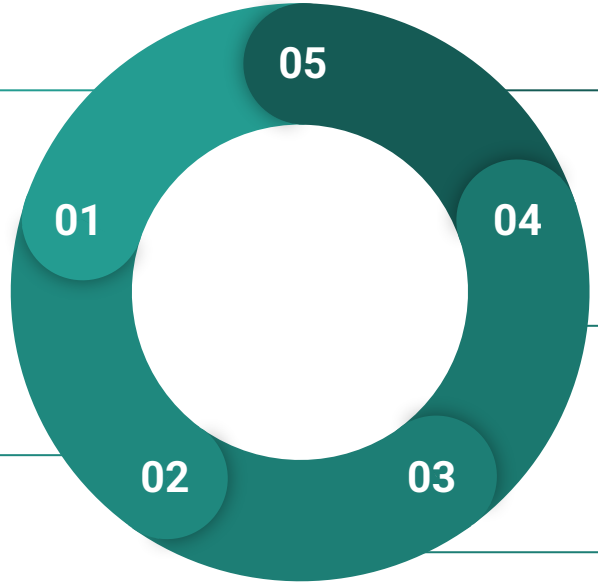
Large sample size

Statistical power

Recruitment/Enrolment

Genotyping

-



Accounting for false discovery

-Validation, metaGWAS, Replication, postGWAS

Testing for association

Statistical models

Data Processing

Filtering / Quality Control

Imputation

Costs matter - approximate current SNP array prices

- Bacteria
- Bovine
- Canine
- Chicken
- Drosophila
- Equine
- Hamster
- Human
- Maize
- Mammalian
- Mouse
- Nematode
- Ovine
- Plant
- Porcine
- Rat
- Rice
- Shrimp
- Soybean
- Tilapia
- Tomato
- Virus
- Yeast
- Zebrafish

Microarray
For Research Use Only
DNA

BovineHD DNA Analysis Kit

This comprehensive genome-wide bovine genotyping array kit features over 777,000 SNPs, and is compatible with any breed of beef or dairy cattle.

Microarray
For Research Use Only
DNA

BovineSNP50 v3 DNA Analysis BeadChip

This BeadChip microarray provides high density multi-sample bovine genotyping for genome characterization of major dairy and beef cattle breed types.

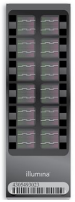
EquineSNP50 Genotyping BeadChip

The EquineSNP50 Genotyping BeadChip features more than 54,000 evenly spaced and validated SNPs derived from the EquCab2 assembly. This 12-sample Infinium® BeadChip presents a cost-effective and high-quality genotyping solution for equine research.

Canine HD (12-sample)

170k

Featuring highly polymorphic SNP content and providing uniform genomic coverage, the CanineHD BeadChip enables the interrogation of genetic variation in any domestic dog breed. Importantly, this BeadChip presents an average of greater than 70 markers per megabase (Mb), providing ample SNP density for robust within-breed association and copy number variation (CNV) studies. This BeadChip contains more than 170,000 markers placed on the CanFam2.0 reference sequence. Illumina developed the BeadChip in collaboration with the LUPA Consortium, which includes 22 European universities and other partners such as the Broad Institute.



OvineSNP50 DNA Analysis Kit

This sheep microarray features over 54,241 evenly spaced SNP probes for genome-wide association studies, genome-wide selection, and genetic merit determination.Read More...

The BeadChip was developed in collaboration with leading ovine researchers from AgResearch, Baylor UCSC, CSIRO, and the USDA as part of the International Sheep Genomics Consortium. It features over 54,241 evenly spaced probes that target single nucleotide polymorphisms (SNPs).

GGP Porcine HD Array

This genome-wide porcine genotyping array is ideal for marker-assisted selection and prediction applications. Array content includes:

- 70,000 SNPs for all major porcine breeds
- Average marker spacing of ~42 kb
- 20 key causative mutations

GGP Porcine LD Array

This array is designed for marker-assisted selection, Illumina PorcineSNP60 imputation, GGP Porcine HD imputation, and prediction applications. Array content includes:

- More than 10,000 SNPs for all major porcine breeds
- Average marker spacing of ~250 kb
- ~20 important causative mutations

Costs matter - approximate current SNP array prices

Microarray
For Research Use Only
DNA

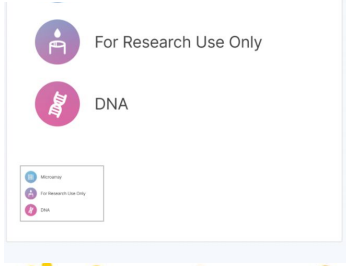
MaizeLD BeadChip Kit

Microarray kit for maize breeding applications and assessment of essentially derived varieties. Samples used include the Plant Variety Protection Act panel.

Microarray
For Research Use Only
DNA

MaizeSNP50 DNA Analysis Kit

This array enables genetic variation analysis across maize lines. It includes over 50,000 validated markers derived from the B73 corn reference sequence.



GGP Potato Arrays 12k

Identify resistance regions with maximum coverage using the GeneSeek Genomic Profiler (GGP) potato microarray.Read More...

Select Product(s)
What size kit do I need?

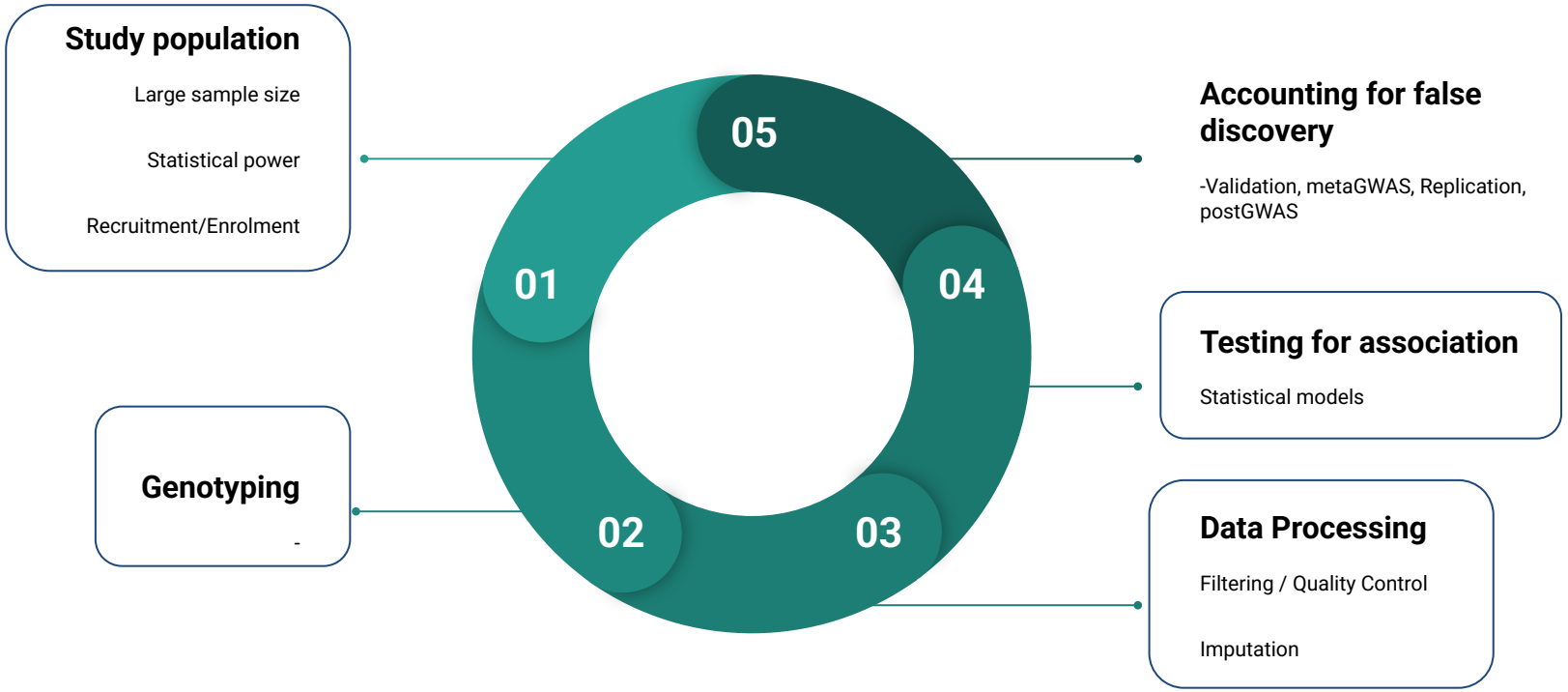
0	GGP Potato-24 v4.0 (48 samples) 20044459	Sign in to see pricing and favorite products.
0	GGP Potato-24 v4.0 (288 samples) 20044820	Sign in to see pricing and favorite products.
0	GGP Potato-24 v4.0 (1152 samples) 20044821	Sign in to see pricing and favorite products.

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**

conduction a GWAS



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

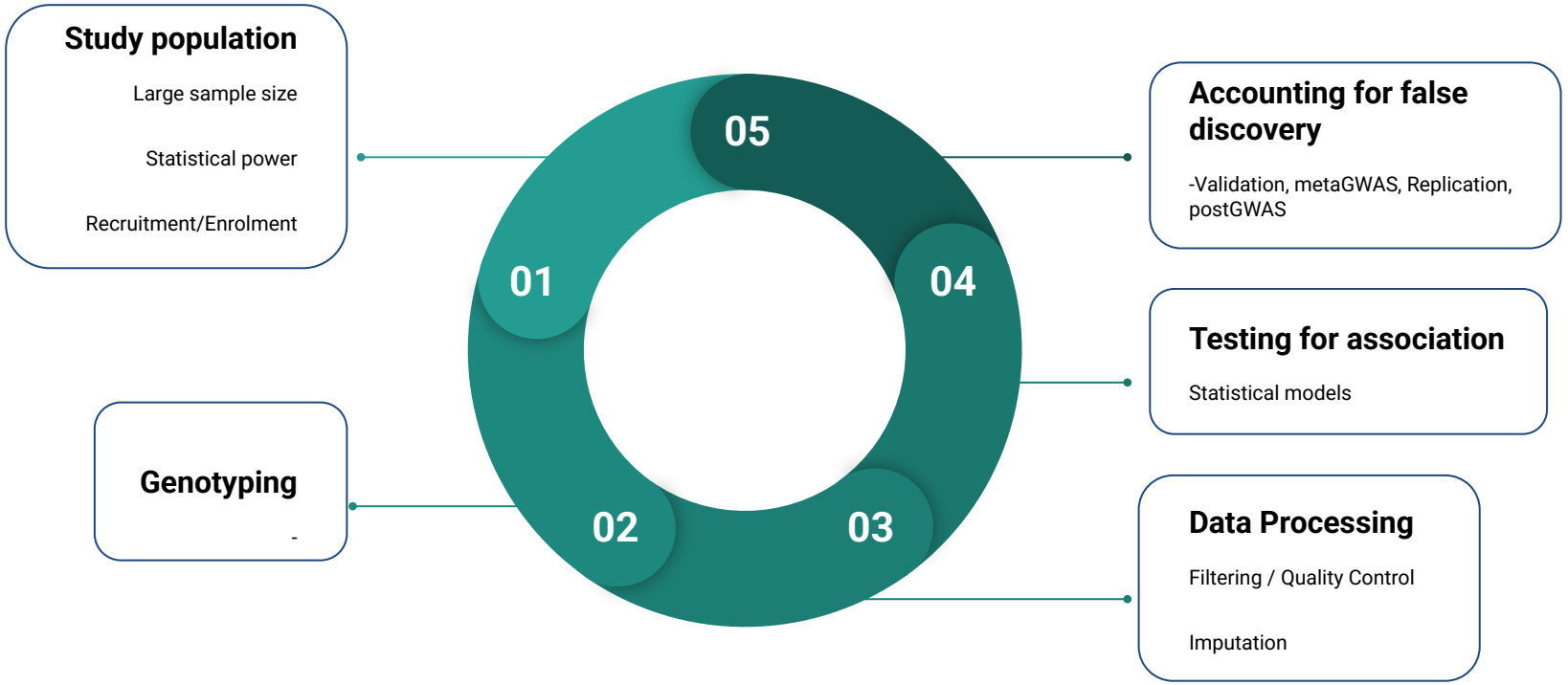
Linkage disequilibrium

$$P = G + E$$

Baby model $y=\mu+x\beta+\epsilon$

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

conduction a GWAS



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

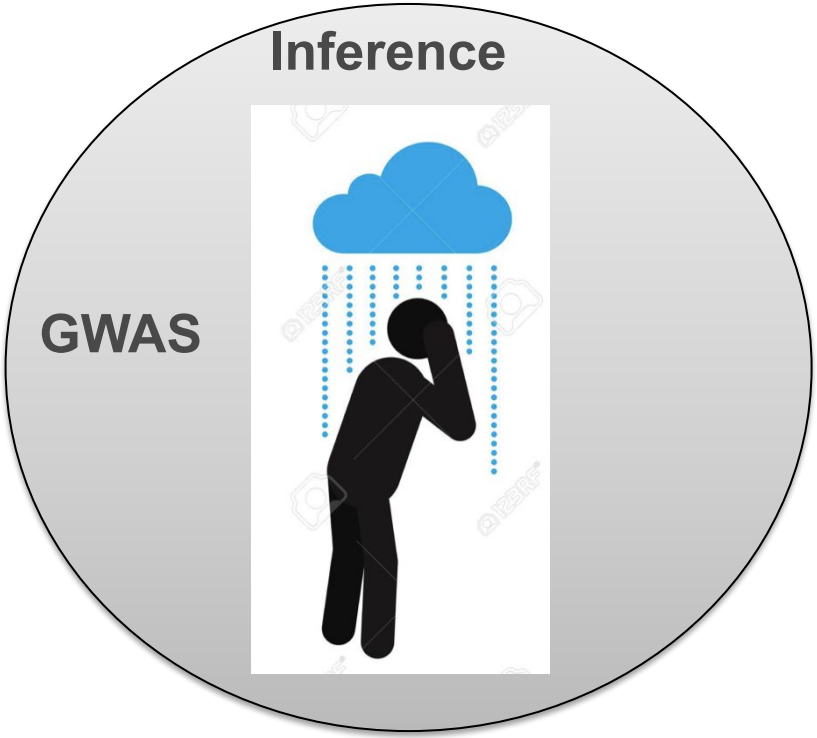
Inference

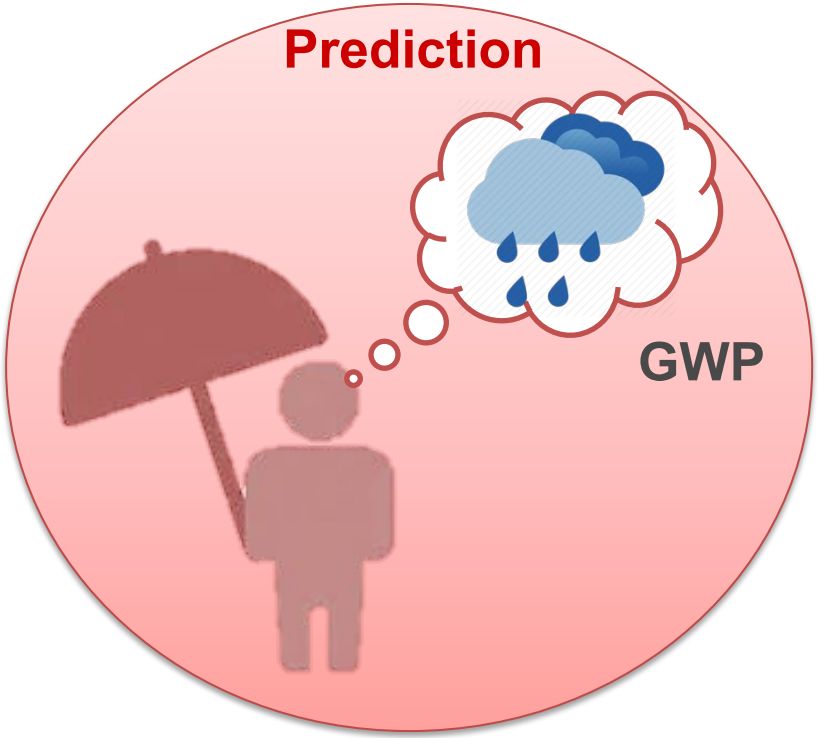
- Determine the effect of a covariate on the response
- Determine the causal relationship between a covariate and the response (Need specific experimental designs)

Prediction

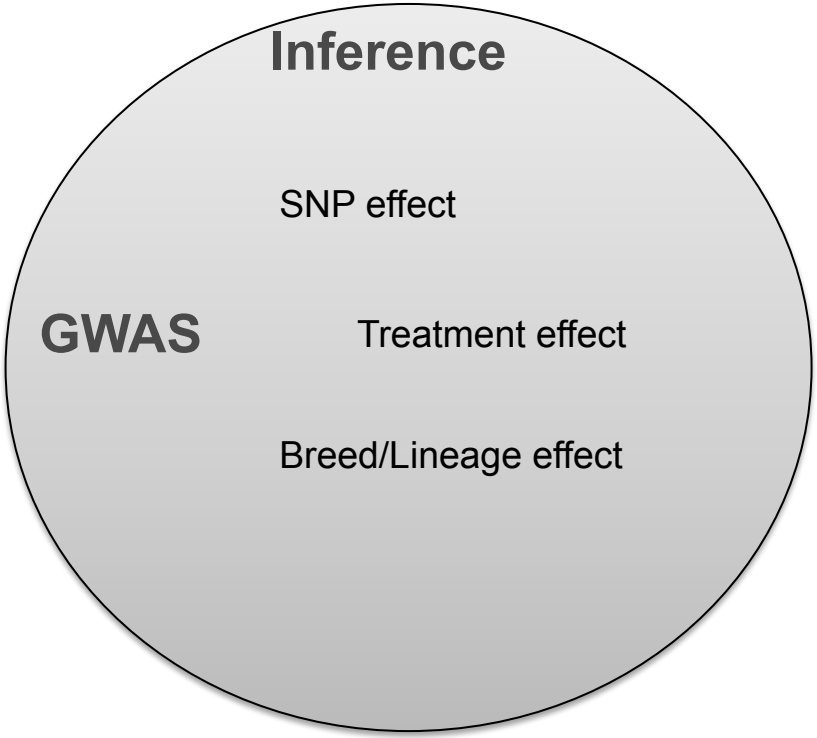
- Educated guess of the outcome
- Expected behaviour in the future
- Based on proxies/markers

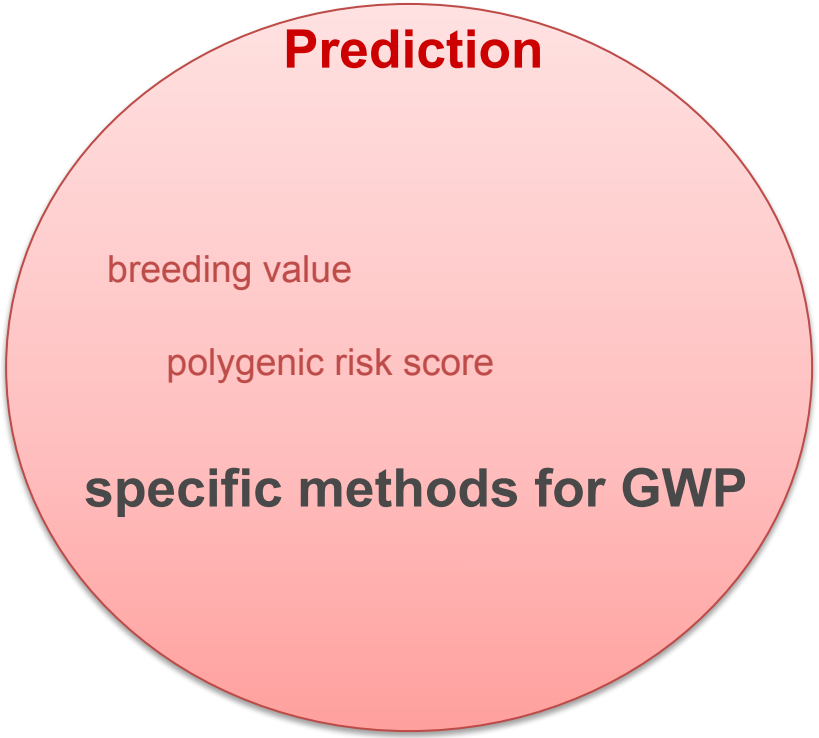
Inference vs Prediction





Inference vs Prediction





Genetic basis of GWAS-GWP

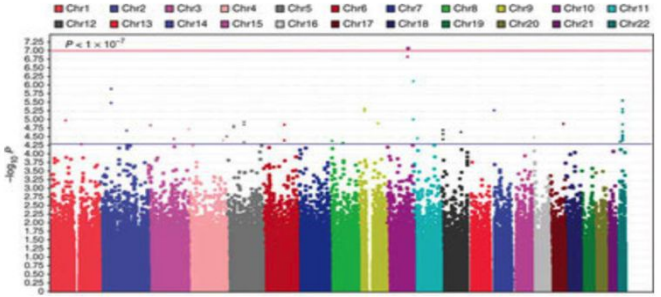
GWAS goal

- 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
- Use methods that apply shrinkage on inference

Some examples



GWAS –Rheumatoid arthritis

BMC Proceedings

Home About Articles Submission Guidelines

Volume 3 Supplement 7

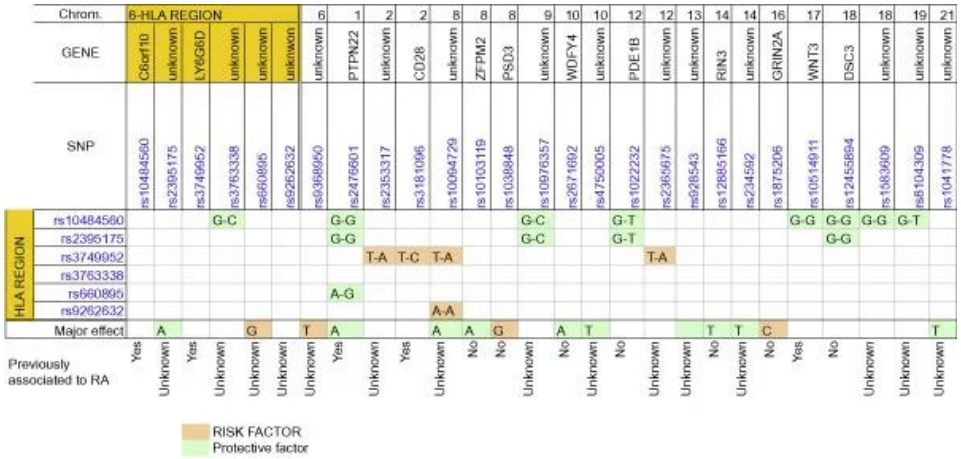
Genetic Analysis Workshop 16

Proceedings | Open Access | Published: 15 December 2009

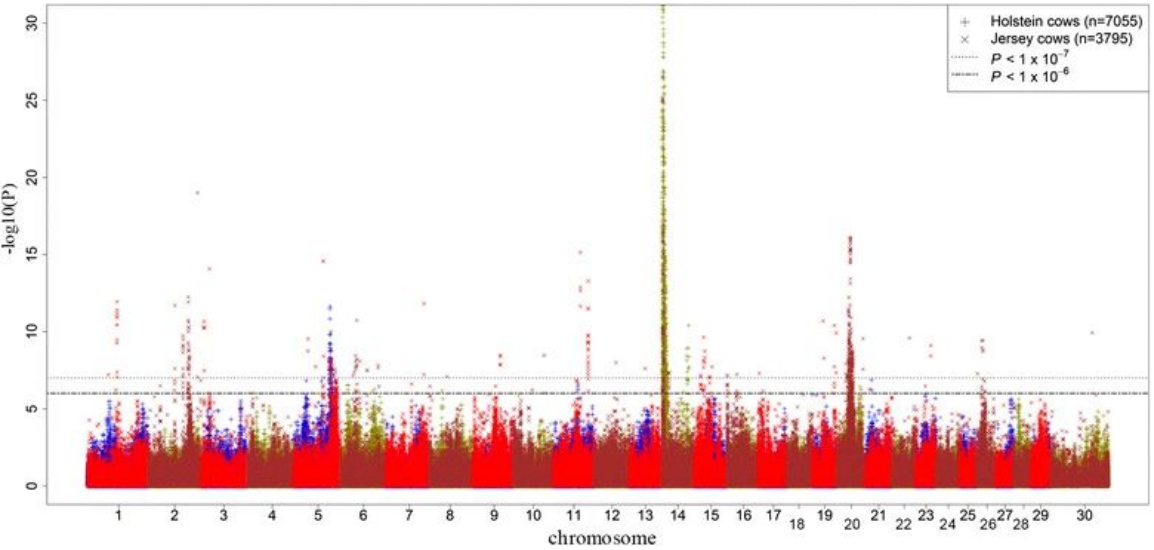
Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model

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



GWAS – Milk yield dairy cattle



- **DGAT1** gene on BTA 14 in dairy cows (HOL and JER)
- milk fat content

BMC Genetics

Home About Articles Submission Guidelines

Research article | Open Access | Published: 22 July 2015

Validation of markers with non-additive effects on milk yield and fertility in Holstein and Jersey cows

Hassan Aliloo, Jennie E. Pryce, Oscar González-Recio, Benjamin G. Cocks & Ben J. Hayes

BMC Genetics 16, Article number: 89 (2015) | Cite this article

4154 Accesses | 8 Citations | 3 Altmetric | Metrics

GWAS – Fertility dairy cattle

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

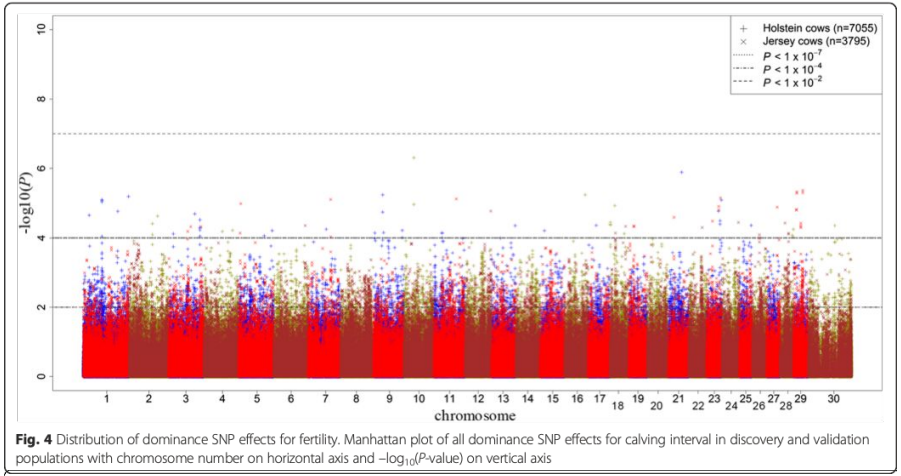


Fig. 4 Distribution of dominance SNP effects for fertility. Manhattan plot of all dominance SNP effects for calving interval in discovery and validation populations with chromosome number on horizontal axis and $-\log_{10}(P\text{-value})$ on vertical axis

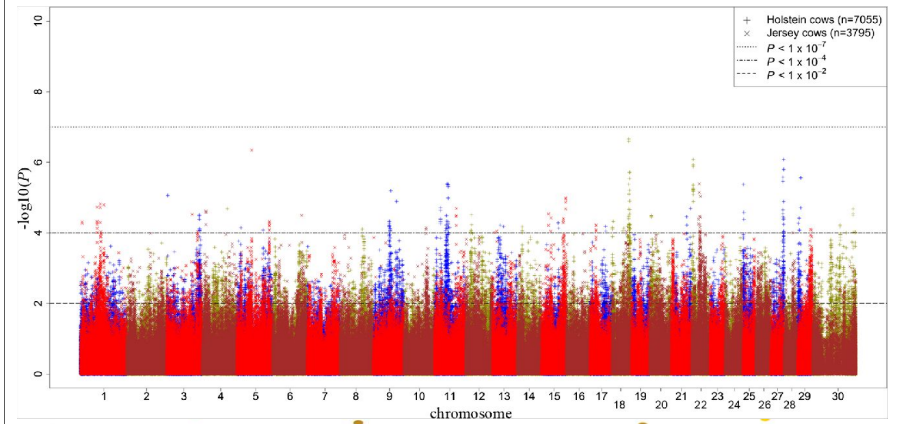


Fig. 2 Distribution of additive SNP effects for fertility. Manhattan plot of all additive SNP effects for calving interval in discovery and validation populations with chromosome number on horizontal axis and $-\log_{10}(P\text{-value})$ on vertical axis

GWAS –curly hair in cattle

nature genetics

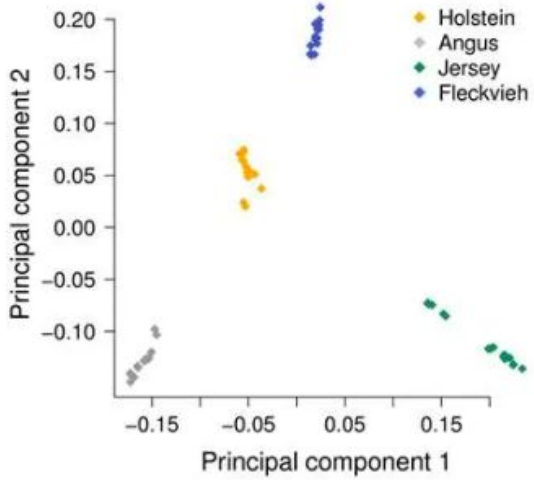
Article | Published: 13 July 2014

Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle

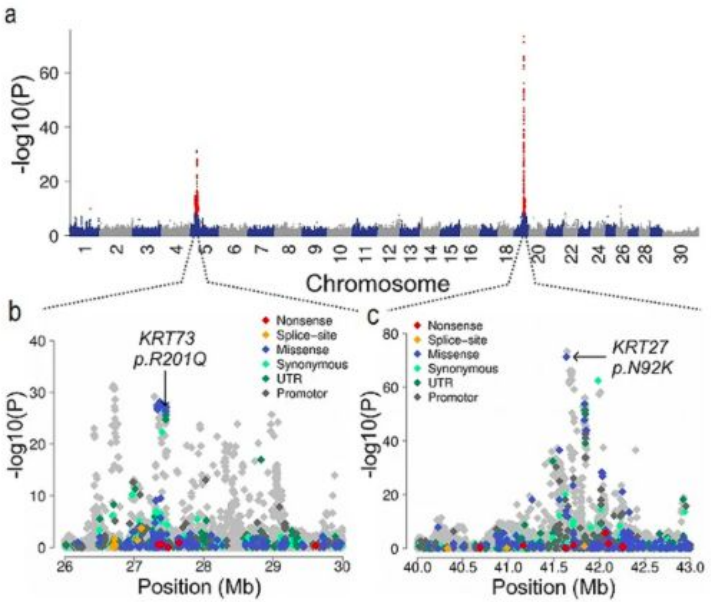
Hans D Daetwyler, Aurélien Capitan, [...] Ben J Hayes

Nature Genetics **46**, 858–865(2014) | Cite this article

1351 Accesses | **351** Citations | **113** Altmetric | Metrics



From: Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits 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

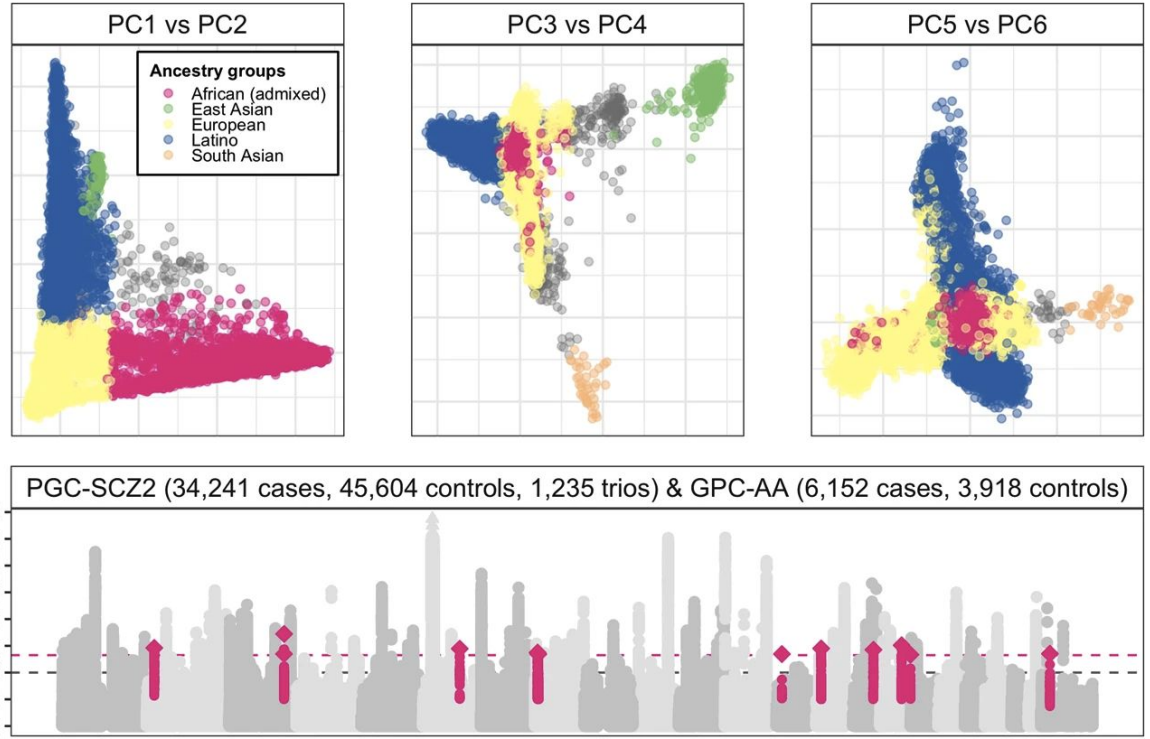
Molecular Psychiatry

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



GWAS - Covid-19

nature

Explore content | Journal information | Publish with us

nature > articles > article

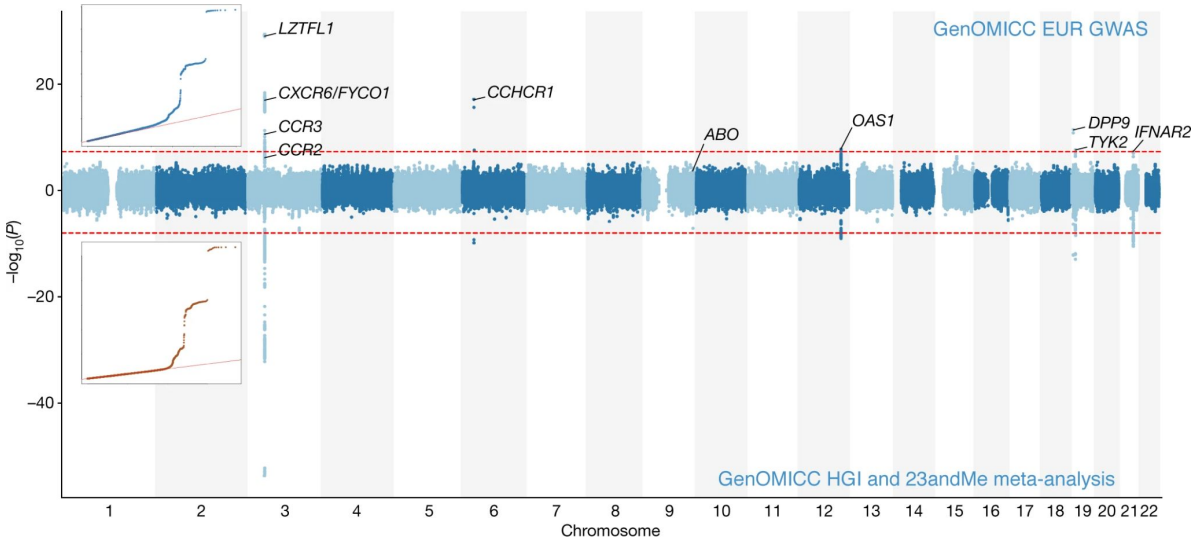
Article | Published: 11 December 2020

Genetic mechanisms of critical illness in COVID-19

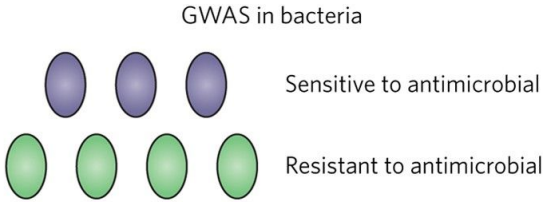
Erola Pairo-Castineira, Sara Clohisey, [...]J. Kenneth Baillie

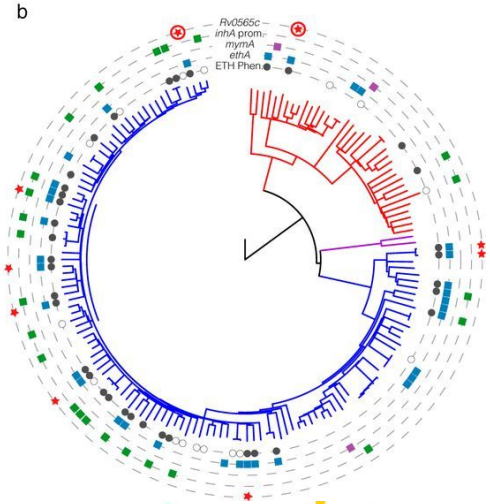
Nature **591**, 92–98(2021) | Cite this article

240k Accesses | **55** Citations | **2731** Altmetric | Metrics

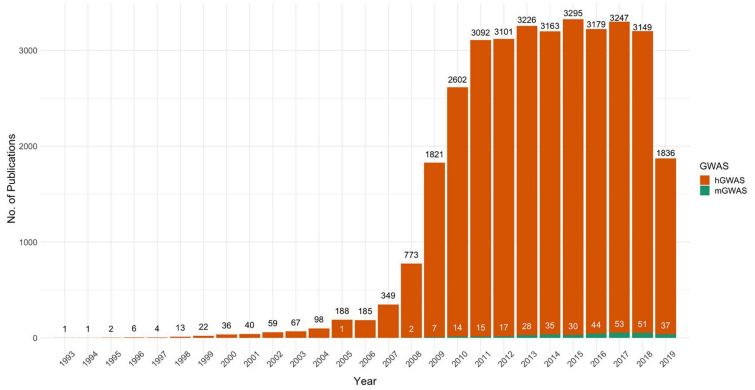


GWAS - bacterial





Not as popular as human GWAS (hGWAS)



Boosted by WholeGenomeSequencing of bacterial genomes along with phenotypes (*e.g.* antimicrobial resistance).

GWAS - bacterial

Particularities

- Haploid organisms
- Strong population structure (Difficult to find Causal mutation and epistatic effects)
- Great variation of the pan-genome
- Low sample sizes
- Unavailability of SNP-chips (necessary whole-genome sequencing)

Relevant in

- Antibiotic susceptibility
- Susceptibility to disinfectants
- Transmissibility
- Disease presentation/severity
- Carriage duration
- Invasiveness

Doing a GWAS is “easy”

Doing a GWAS is “easy”

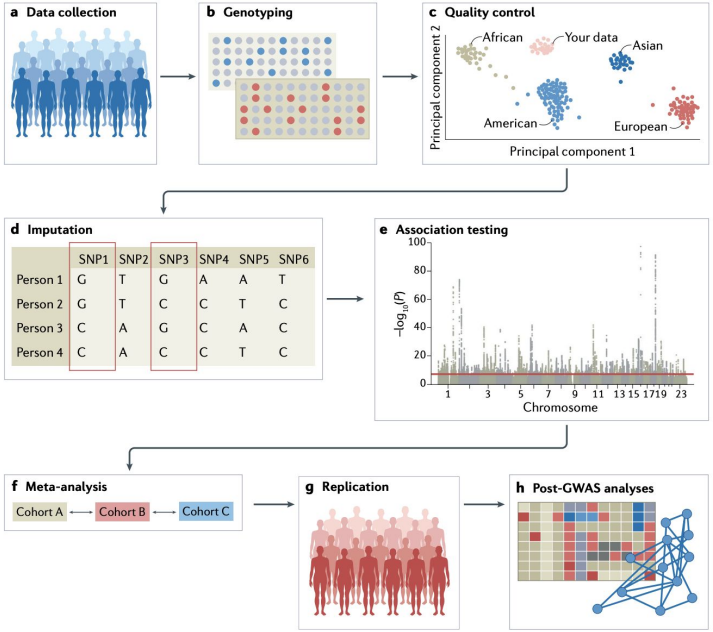
Doing a GWAS CORRECTLY is not

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