A Hands-on Introduction to GWAS

Pieter Clauw
pieter.clauw@gmi.oeaw.ac.at

PopGen intro course

30 September 2022

The goals for today

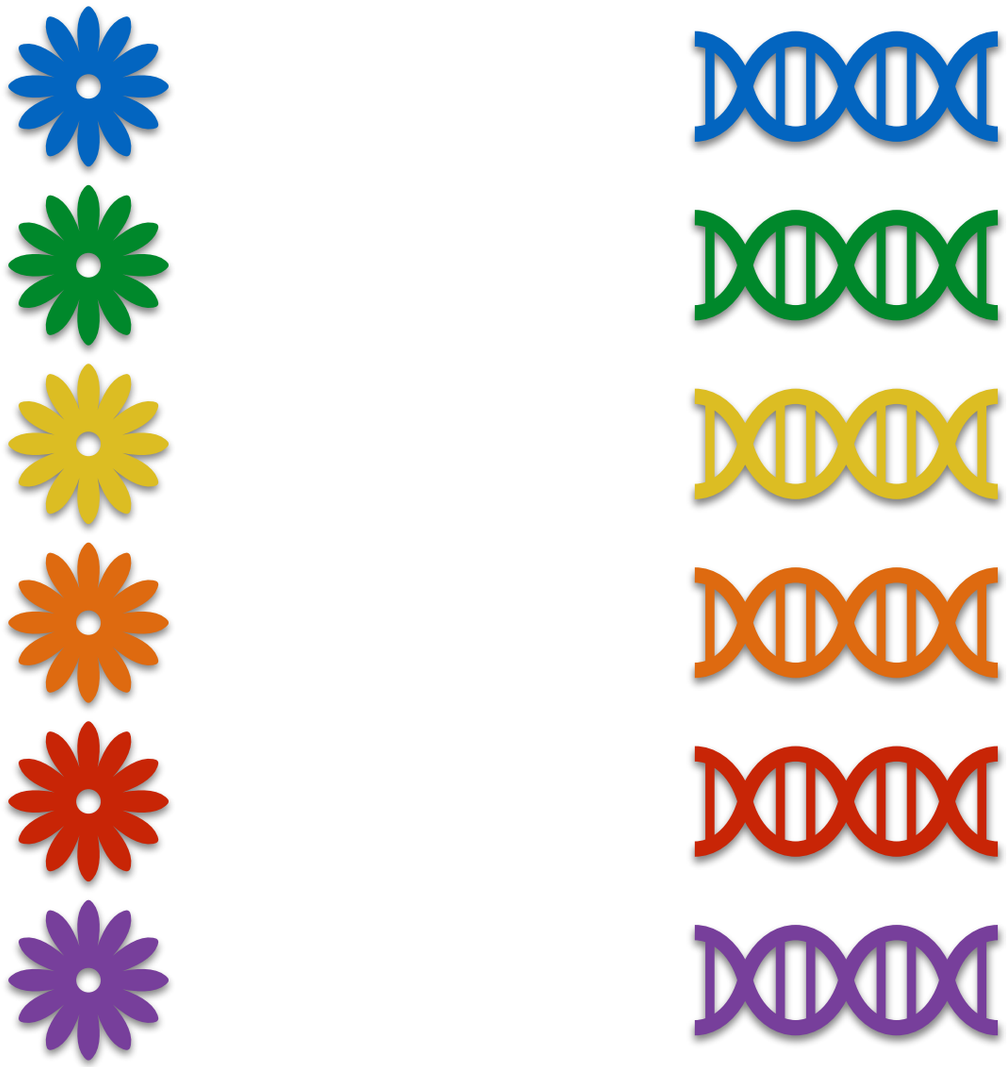
- What is a genome-wide association study?
- When to do GWAS?
- Give a practical, non-technical introduction of how to do a “standard” linear mixed model GWAS
- Address some FAQs
- Identify possible challenges and pitfalls
- Understand the process enough to begin exploring GWAS in your organism of choice

Build a foundation for you to start learning more about GWAS on your own!

What is a GWAS?

Genome **W**ide **A**ssociation **S**tudy

Link genetic differences to phenotypic differences



What is a GWAS?

Genome **W**ide **A**ssociation **S**tudy

Link differences at specific genetic loci to phenotypic differences

…ATGTTTAGCGTAGCGA…

…ATGTTTAGCGTAGCGA…

…ATGTTTAGCGTAGCGA…

…ATGTTTATCGTAGCGA…

…ATGTTTATCGTAGCGA…

…ATGTTTATCGTAGCGA…

All possible polymorphisms: SNPs, INDELs, SVs, LOF …

Test each locus independently

When to do a GWAS?

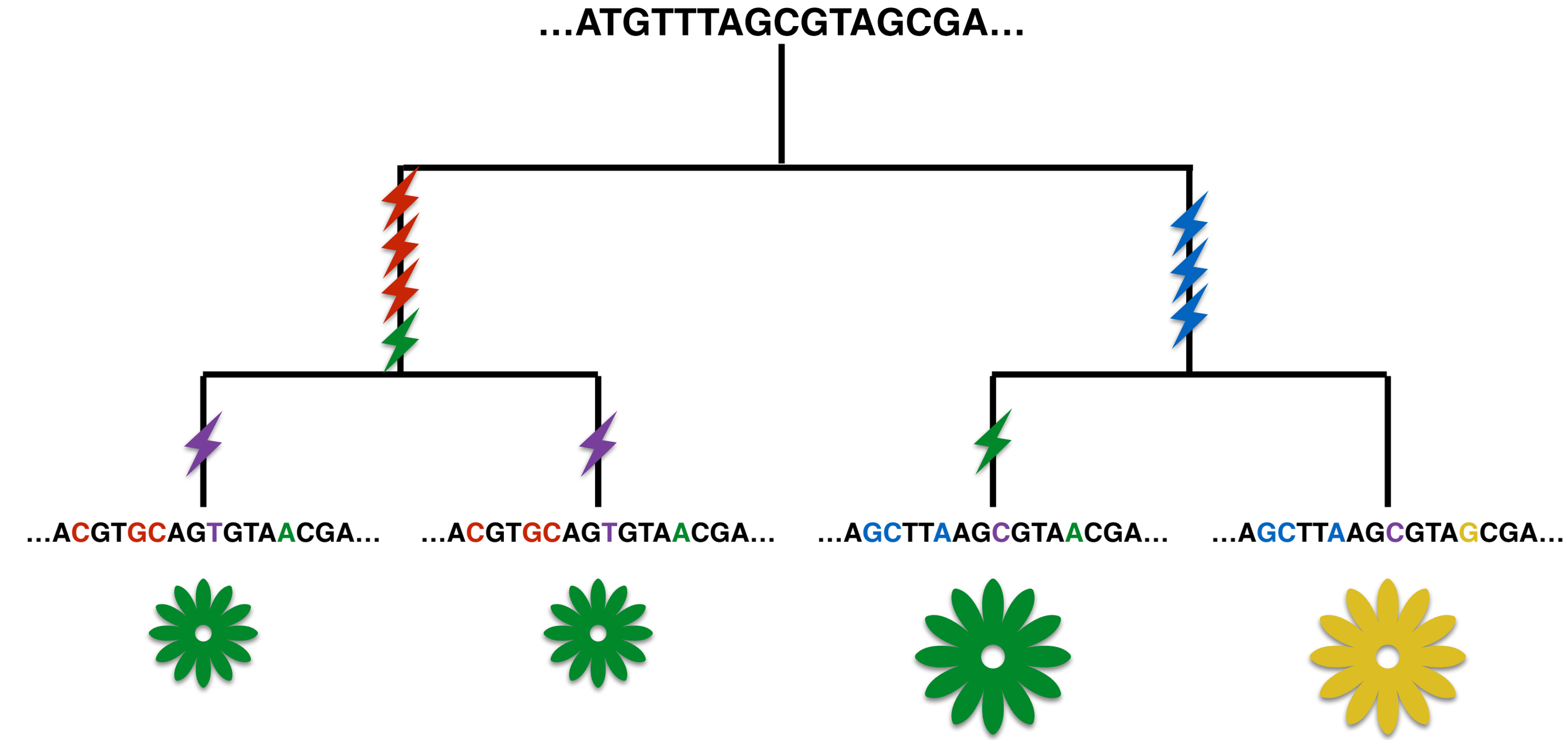
Genotype to phenotype map

- which loci/genes underly a specific phenotype
- Breeding, human/medical genetics, molecular biology
- Generates mechanistic hypotheses
- Genomic prediction/ polygenic scores
- Genetic architecture

Evolutionary genetics

- which loci are involved in selection
- Presumes selection on the phenotype under study
- Identifies loci for further investigation

Origins of genetic variation



Structure of genetic variation: LD

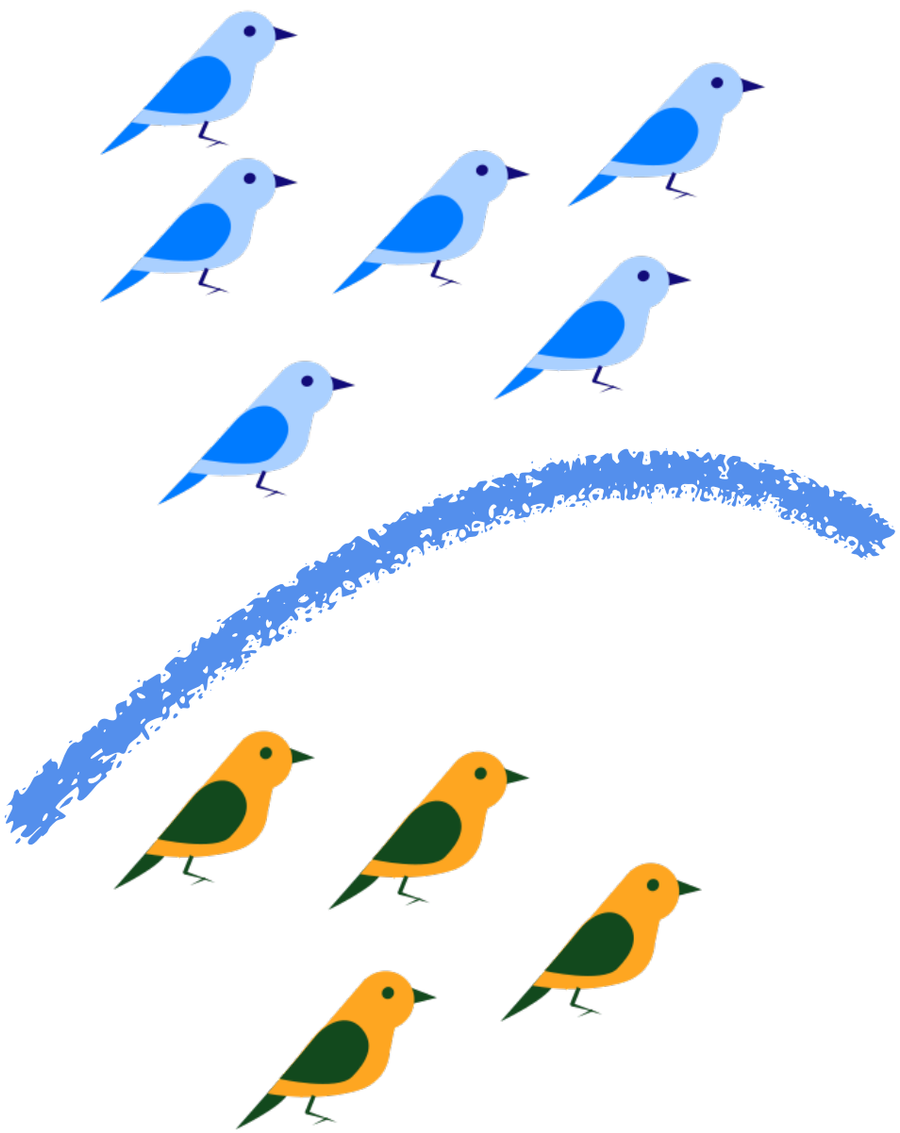
Linkage disequilibrium (LD) = tendency for SNP variants to be inherited together

Why? proximity in genome, population structure, selection
Broken down by recombination - linkage blocks

Friend
We don’t need to test every SNP (one SNP can tag many variants).

Enemy
LD can confound our results!
Complicated confounding patterns (especially if many alleles or loci underlie the trait)
Population structure causes confounding genome-wide
If we don’t try to correct for population structure, we will end up with too many false positives.

How to correct for population structure?



Individuals within a population are more related than those between populations.

Here, birds in a population share not only ***causative*** variants, but also ***non-causative*** variants that are more common in the population (genetic background).

We take this background into account in GWAS to try to reduce the significance of non-causative variants. The **K-matrix** represents this background relatedness.

What do we need to run a GWAS?

What the GWAS does:

Fit a linear mixed model for each SNP

$$Y = X\beta + K + \varepsilon$$

Phenotype (Y) = Genotype (X) + Background relatedness (K) + error

(Genotype is a fixed effect, K is a random effect)

What we get as results:

1. P-value for each SNP (from comparing the full model to a model without the genotype term)

$$Y = K + \varepsilon$$

2. The effect of the SNP on the phenotype (β)

Phenotypes and Genotypes

GWAS is testing an underlying hypothesis!
Genetic variation underlies the phenotypic variation we observe among these individuals.

Phenotype and individual/genotype choice shape this hypothesis (and interpretation of results).

Phenotypes need to be heritable traits.

Simple vs. complex phenotypes

Quantitative (can do binary or categorical with different models)

Remember, we are doing linear modeling, so there is the assumption that residuals will be normal.

Genotypes

Common genotyping methods
Whole-genome resequencing
RADseq/Genotyping by Sequencing
mRNA-Seq
Exome sequencing
SNP arrays

- Understanding LD is key in selecting/interpreting results given a particular dataset.
- How much of the genome are you actually assaying?
- Coding versus non-coding variation?
- SNPs aren’t the only type of polymorphism.

K matrix

Genetic relationship between pairs of individuals
Pairwise relatedness matrix / IBS matrix

Used for population structure correction

Many programs for estimation: limix, plink, R packages, etc.
Today we will use a pre-calculated matrix - calculation is slow.

Sometimes population structure correction will take other forms:

- PCA
- output from the Structure program

Main point is that this represents the genetic background relatedness between individuals

Options for running GWAS

Online options

EasyGWAS (https://easygwas.ethz.ch/)
GWA-Portal (https://gwas.gmi.oeaw.ac.at/)

- easy interface
- limited data availability
- limited types of analysis
- not flexible

Do-it-yourself

limix (https://github.com/limix/limix)
gemma (https://github.com/genetics-statistics/GEMMA)
and MANY others…

- requires coding skills
- maximum flexibility and options
- allows for more complicated GWAS analyses

The plan for hands-on work today

1. We will walk through GWAS together using a flowering time phenotype (3 Jupyter notebooks).

2. You will do some GWAS by yourself with two additional phenotypes. This independent exercise is outlined in:

GWAS_course_independent_work_instructions.pdf

Think about the different questions as you work.
Please just ask questions if you aren't sure you can answer them!

3. Conclusions, post-GWAS, advanced GWAS

The GWAS jupyter notebooks

0_running_Jupyter_notebooks.ipynb
Learn how to navigate a Jupyter notebook

1_phenotype_exploration.ipynb
Explore the phenotype we will use

2_GWAS.ipynb
Prepare input variables, run GWAS, and output results

3_GWAS_interpretation.ipynb
Visualize and understand GWAS results

https://github.com/picla/GWAS_workshop_CK

I did a GWAS, now what?

- **Association is not causality!**
- **Most significant SNP is not necessarily causative** (confounding, incomplete knowledge of variants)**.**
- **The "peakiness" of a peak is not necessarily a sign of its importance**

Interested in candidate genes?

- Use biology, annotation, expression, etc. Be creative!
- Sequence candidate genes/regions
- Test causality by QTL and/or transgenics with different alleles in common background

Interested in evolutionary or ecological inferences?

- Consider the consequences of your choice of accessions and phenotypes
- Associations with selective pressures (environment/climate)?
- Other signs of selection?
- The reference allele is not necessarily the ancestral allele.

Moving beyond simple GWAS

1. Generalized linear models for different data distributions
2. Adding covariants
3. Incorporating genotype-by-environment interactions
4. Analyzing correlated traits together (increases power)
5. Bayesian approaches
6. Polygenic scores
7. Genomic prediction

There are endless packages/programs available to do GWAS, with more published all the time. Limix and gemma are great places to start.

Human GWAS is a bit of a different animal (no experiments, huge sample sizes, environmental confounding).

Final Questions/Comments?