A Hands-on Introduction to GWAS

Danièle Filiault
Český Krumlov
30 Jan 2020

Part of Vienna BioCenter

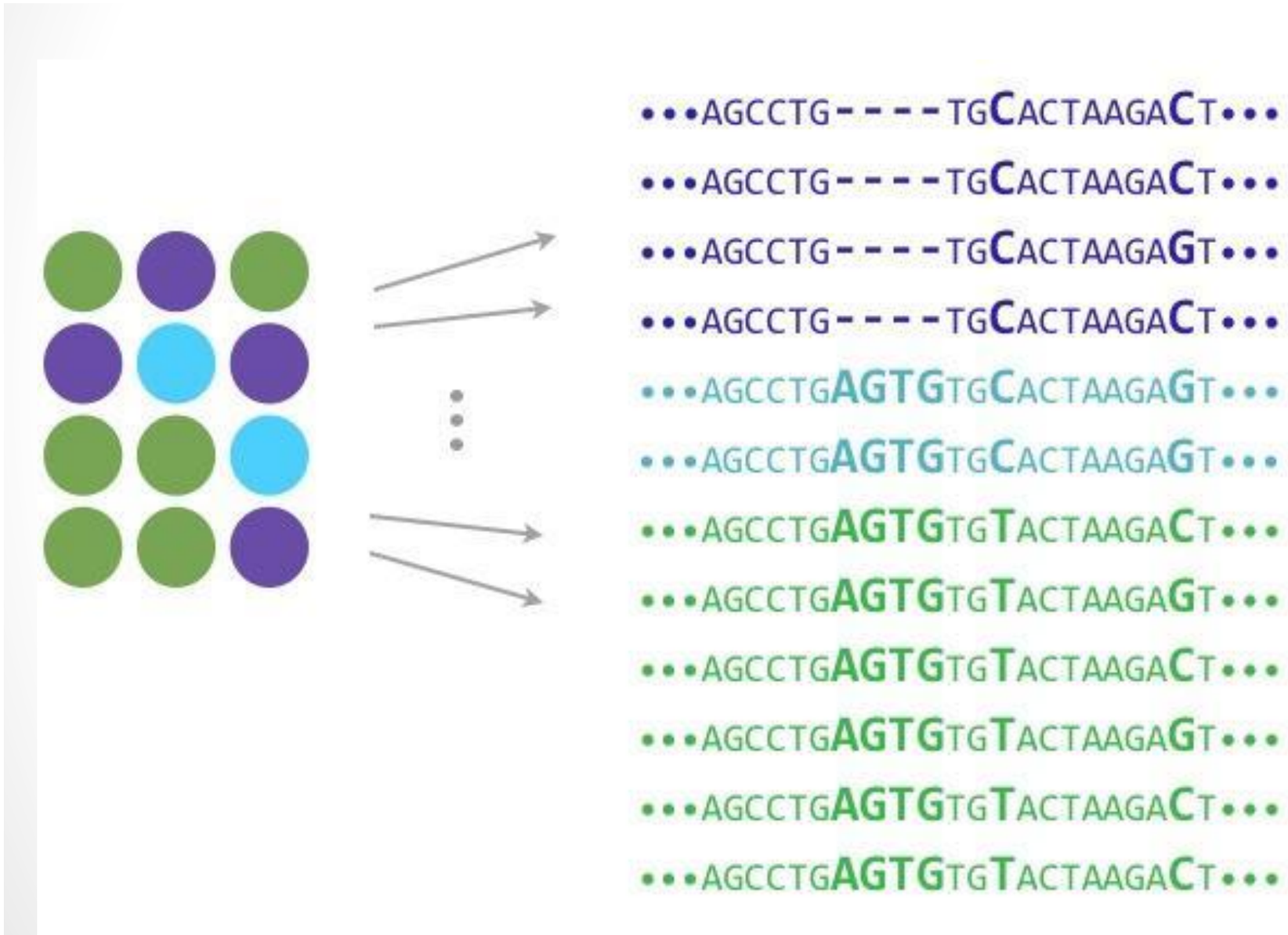
The goals for today

- Give a practical, non-technical introduction of how to do a “standard” linear mixed model GWAS
- Address some FAQs
- Identify possible sticking points
- 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?

Use genome-wide SNPs/variants
Associate phenotype and genotype



LD is our frenemy!

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

Why? proximity, population structure, selection

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

What do we need to run a GWAS?

Fit a linear model for each SNP

Linear Mixed Model

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

Phenotype = **Genotype** + **K** + error

Genotype is a fixed effect, K is a random effect

P-value is determined by comparing models with and without the genotype term

Model also estimates β (the effect of the SNP)

Phenotypes and Accessions

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

Phenotype and accession 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 chips

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 accessions
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 (especially in humans)
- output from the Structure program

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. Walk through a GWAS together (3 jupyter notebooks) with a flowering time phenotype

2. You will do some GWAS by yourself with two additional phenotypes
   - work through the 3 notebooks in order
   - make sure you change input and output file names as appropriate (in section 1b of notebooks).

3. We will do a final conclusion after everyone has finished at least one phenotype independently

I did my 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!
- Sanger 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)?
- The reference allele is not necessarily the ancestral allele.

Moving beyond simple GWAS

1. General 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?

2 phenotypes

Cadmium concentrations in leaves (growth chamber)
https://doi.org/10.1371/journal.pgen.1002923

Leaf color (field-grown plants)
my unpublished data

If you're fast (or motivated), try to do a GWAS with different minor allele frequencies. Or do full flowering time data (1135 accessions - how long does this take to run?) Or if you like coding exercises, try to understand the code line by line. ;)