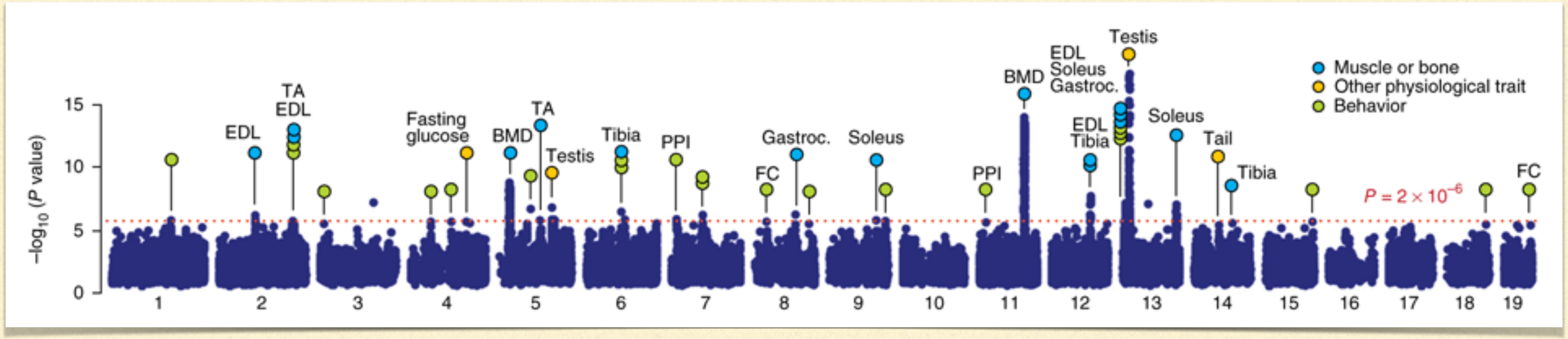


GENOME WIDE ASSOCIATION STUDY

Using PLINK and GEMMA

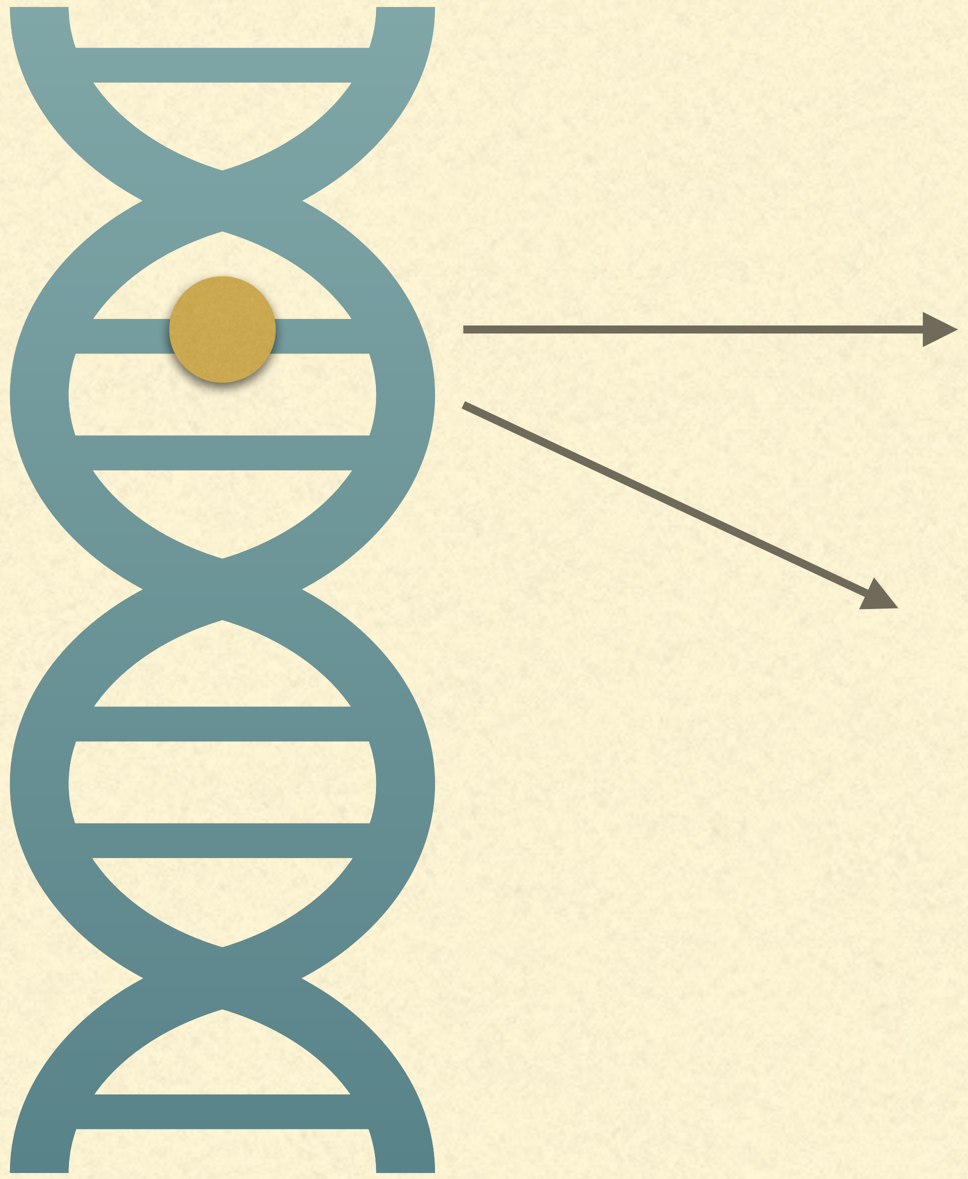
Quantitative Biology Class
14th December 2023

Nhu L.T. Tran

Short theory:
What is Genome Wide Association Study?

WHAT IS **G**ENOME **W**IDE **A**SSOCIATION **S**TUDY

- *GWAS* identifies genomic **variants** that are statistically associated with a particular trait



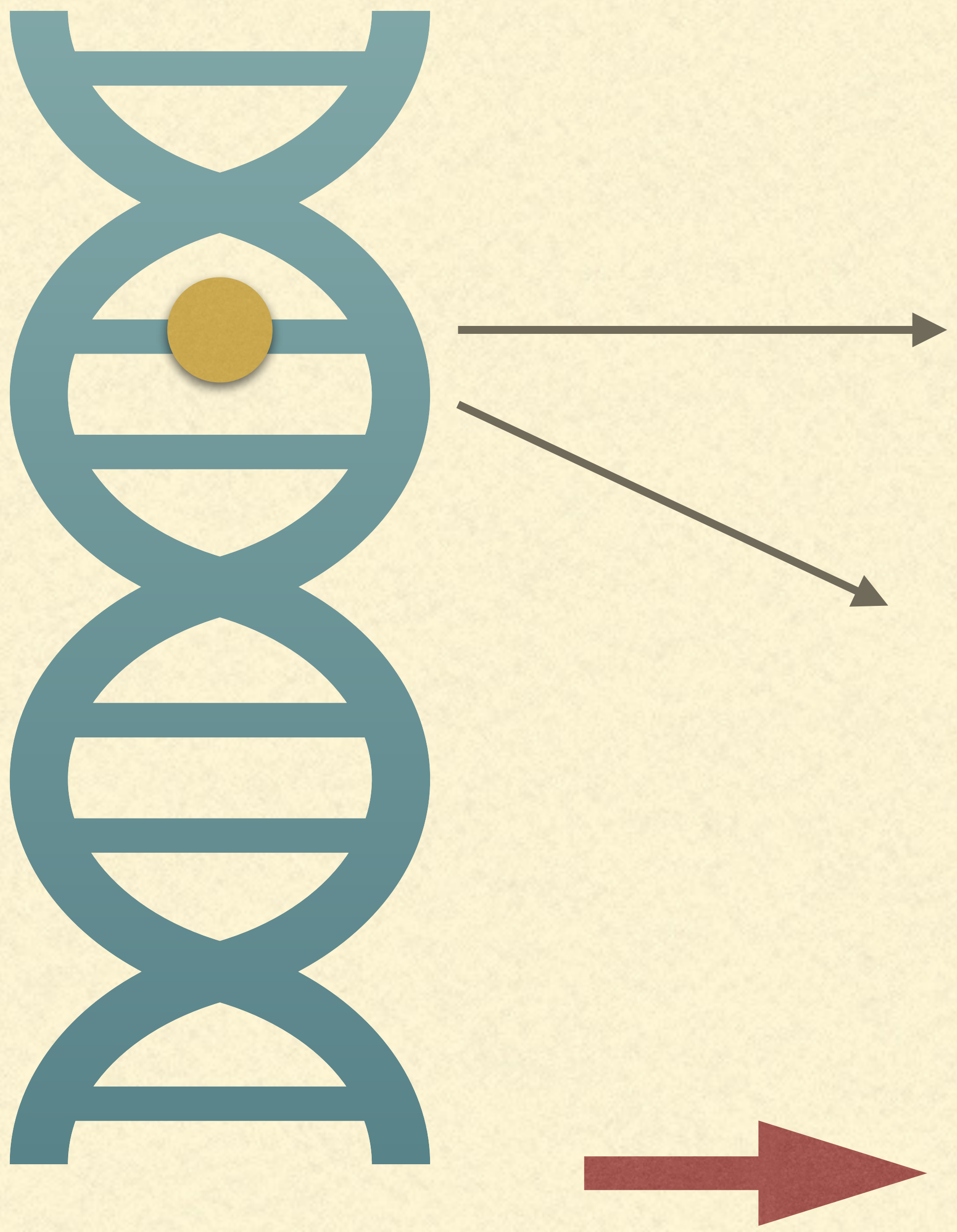
10% have A allele

90% have T allele

Variants due to single nucleotide polymorphism (**SNP**)

WHAT IS **G**ENOME **W**IDE **A**SSOCIATION **S**TUDY

- *GWAS* identifies genomic **variants** that are statistically associated with a particular trait.

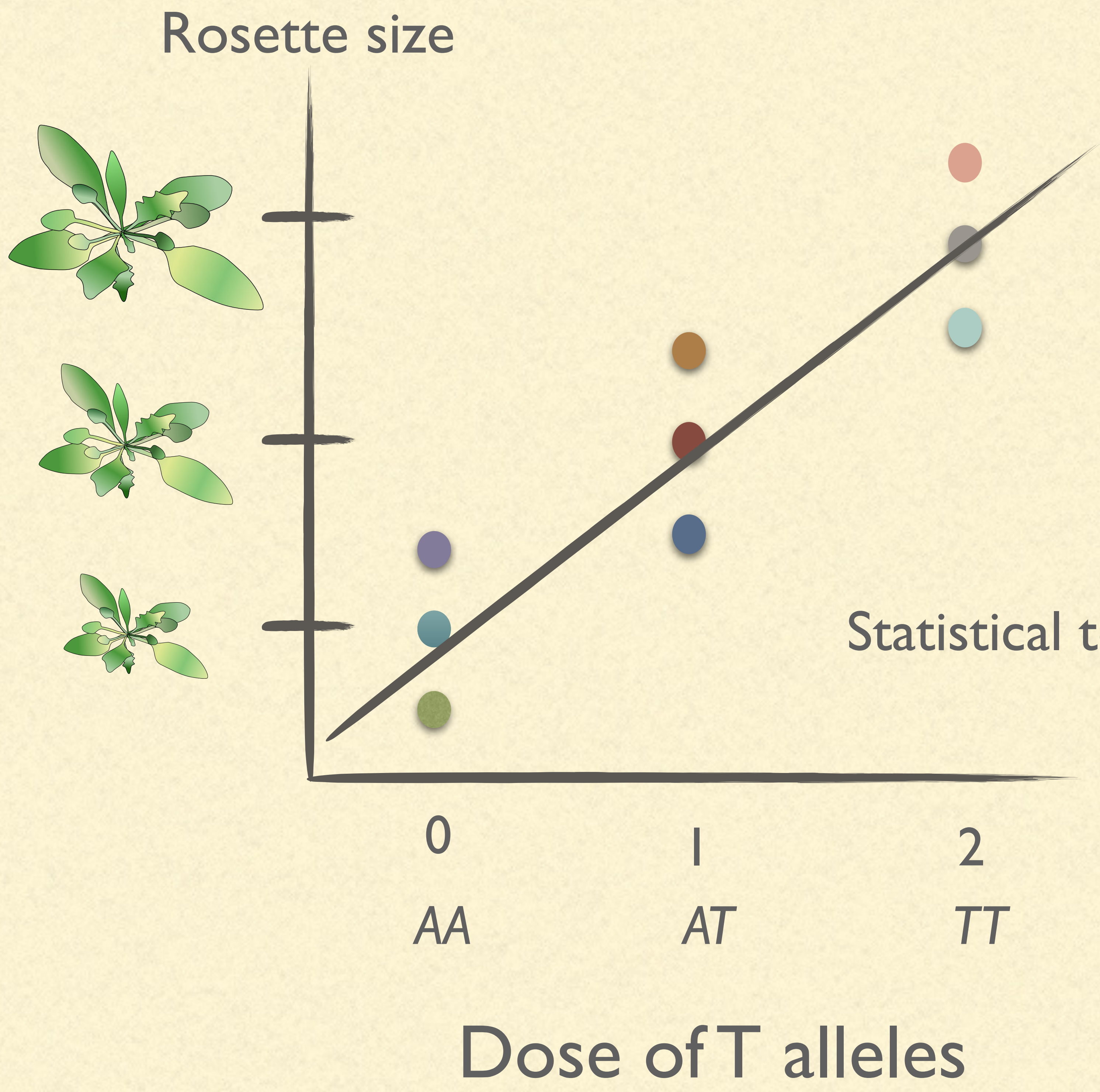


At SNP #1

10% have A allele

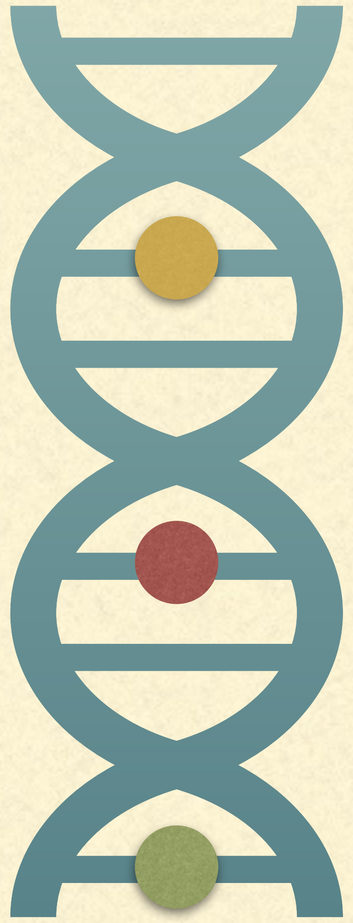
90% have T allele

SNP#1 is associated with rosette size



Statistical testing: p value = 0.05

WHICH DATA ARE NEEDED FOR GWAS?



SNP data

	SNP #1	SNP #2	SNP #3	…	SNP # 1M
Plant 1	A	T	G	…	T
Plant 2	A	A	C	…	C
Plant 3	T		G	…	C
…	…	A	…	…	C
Plant 1000	T	T	C	…	C

Phenotypes of interest

	HEIGHT	ROSETTE SIZE	ETC.
Plant 1	15	10	…
Plant 2	18	9	…
Plant 3	23	25	…
…	…	…	…
Plant 1000	12	20	…

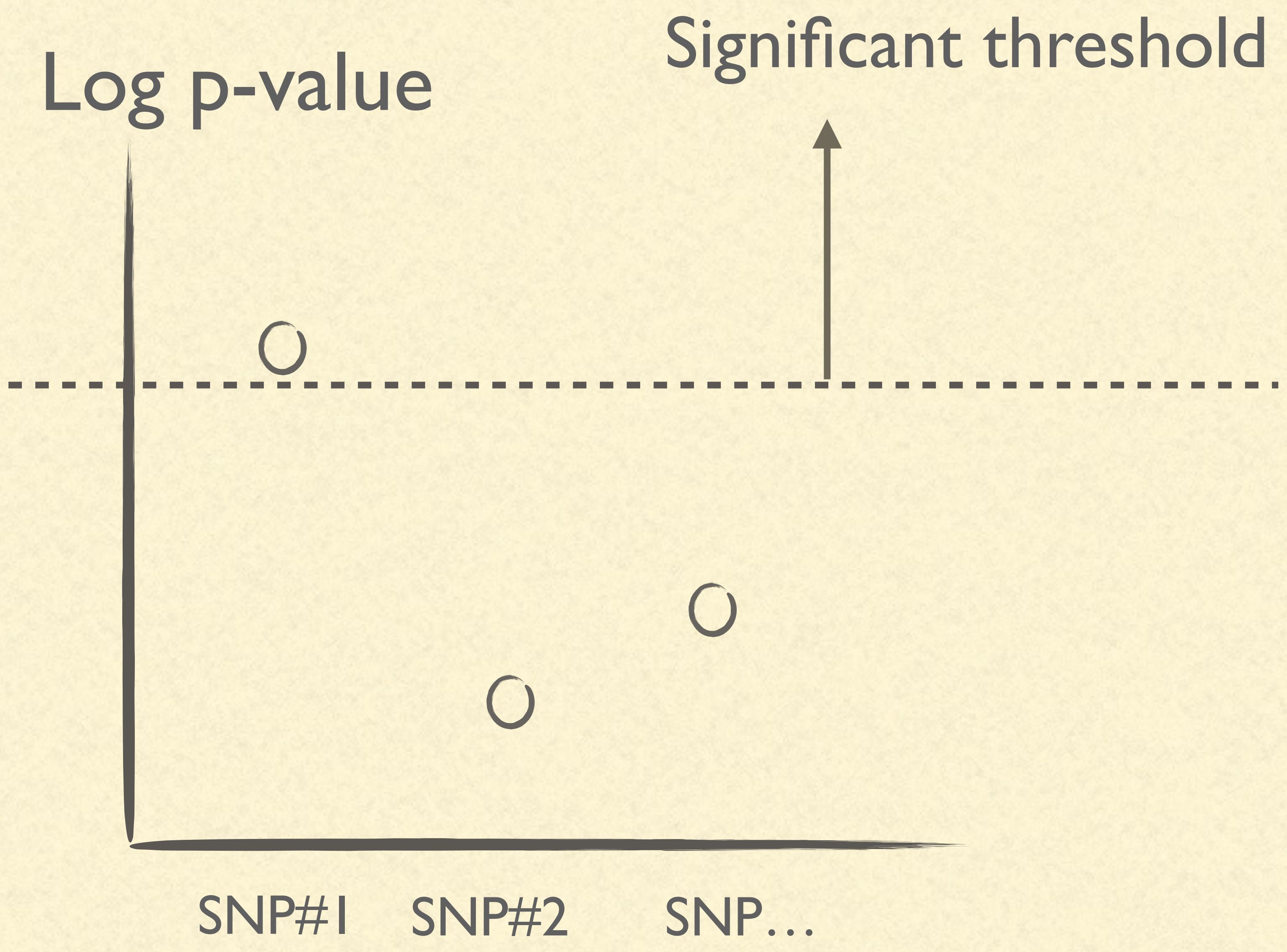
HOW DOES GWAS RESULT LOOK?

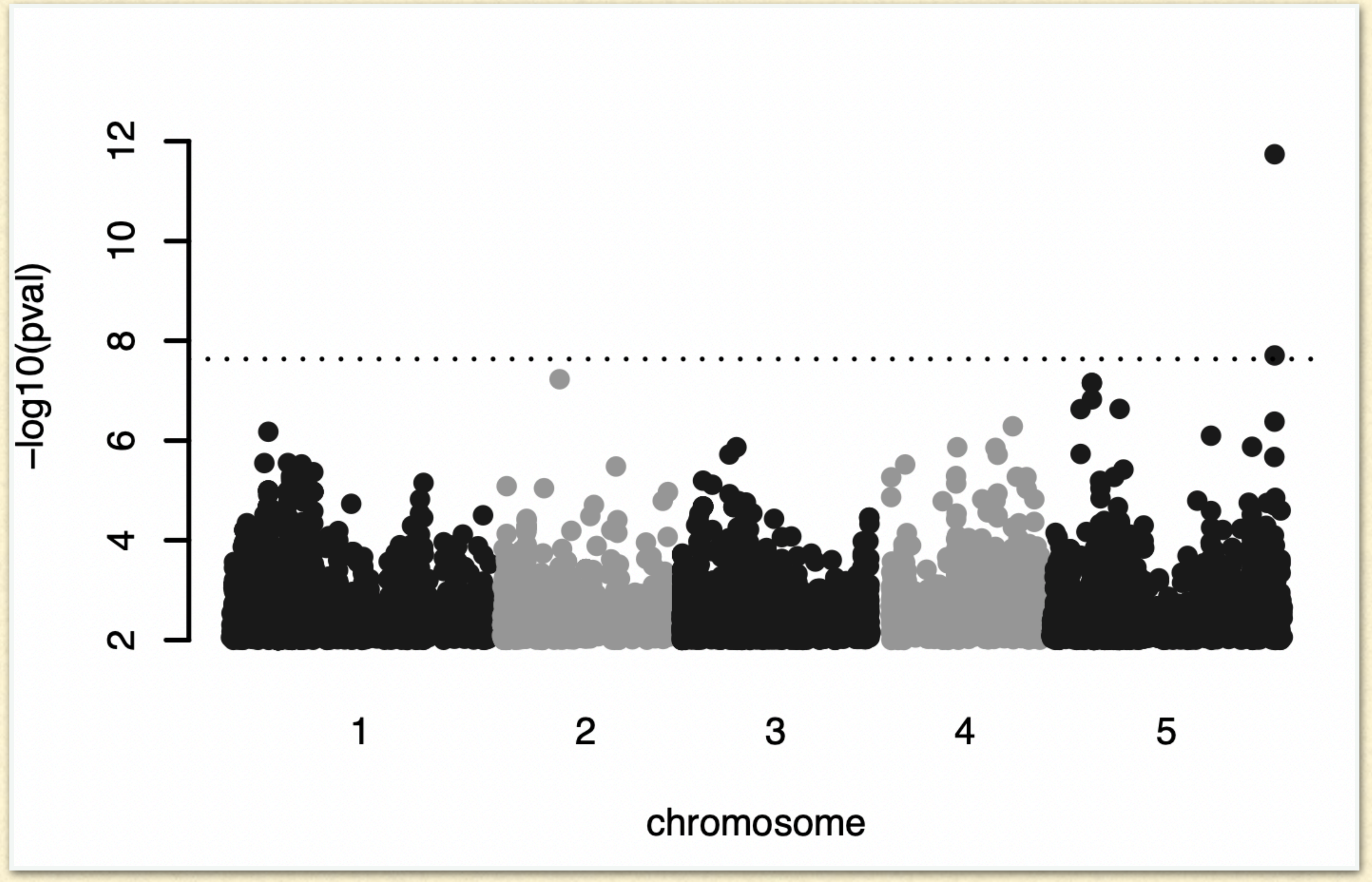
SNP data +

Phenotype data

=> Association test

- All the SNPs are laid out on x axis
- p-value on the y axis



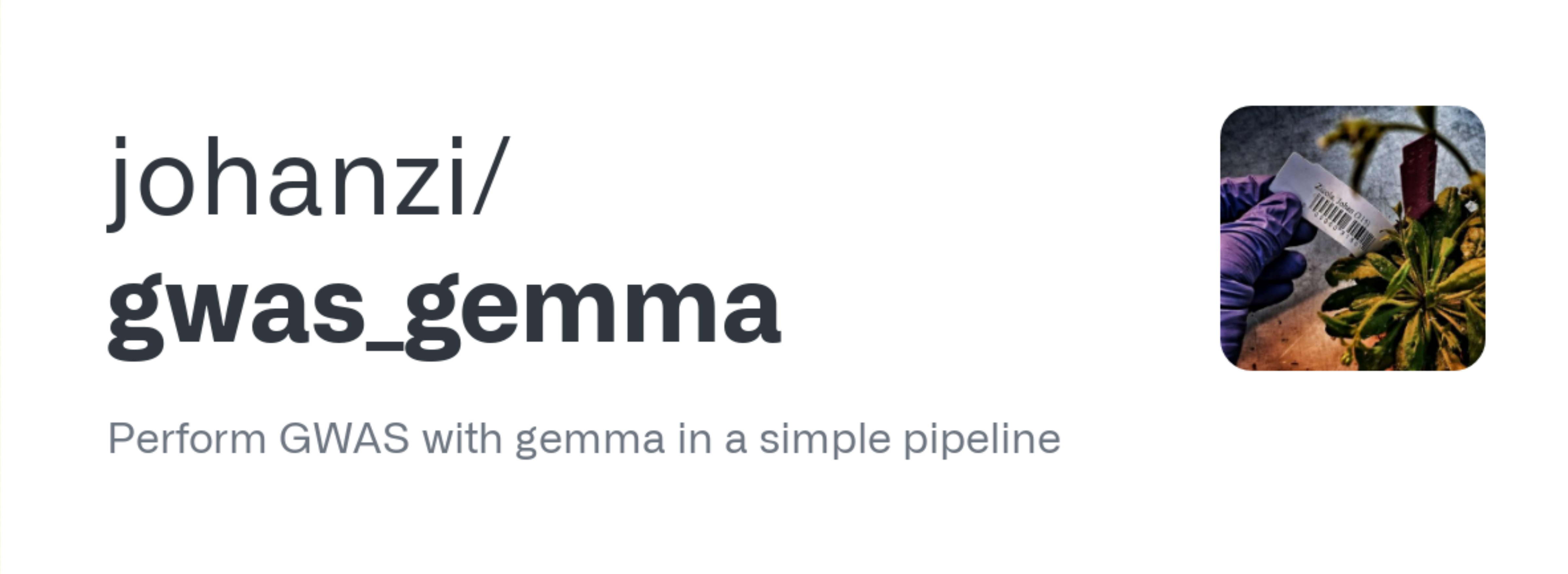


Variant at chromosome 5, position […], is associated with [phenotype]

Practical:
A general and simple workflow to run GWAS

TO BE INSTALLED

- **Gemma GWAS**



johanzi/
gwas_gemma

Perform GWAS with gemma in a simple pipeline

https://github.com/genetics-statistics/GEMMA#installation

- PLINK cog-genomics

1. BEFORE GWAS

A) Convert VCF to Plink binary files

Binary version of SNP info, not any readable format for human

```
plink --vcf input.vcf.gz --make-bed --out prefix_of_plink_bfiles
```

Your file name

To make bed file

Name of your output

This command creates three files: **.bed**, **.bim**, and **.fam** files to be used for GWAS

.bim map file contained all **variants** (SNPs)

.fam map file of all **individuals**

**.bim file example*

Chr	SNP	Position	Allele1	Allele 2
1	-	286	C	T
1	-	291	A	T
1	-	303	T	C

1. BEFORE GWAS

B) Perform PCA for population structure

```
plink --vcf input.vcf.gz --make-bed --pca --out output_pca_prefix
```

Your file name

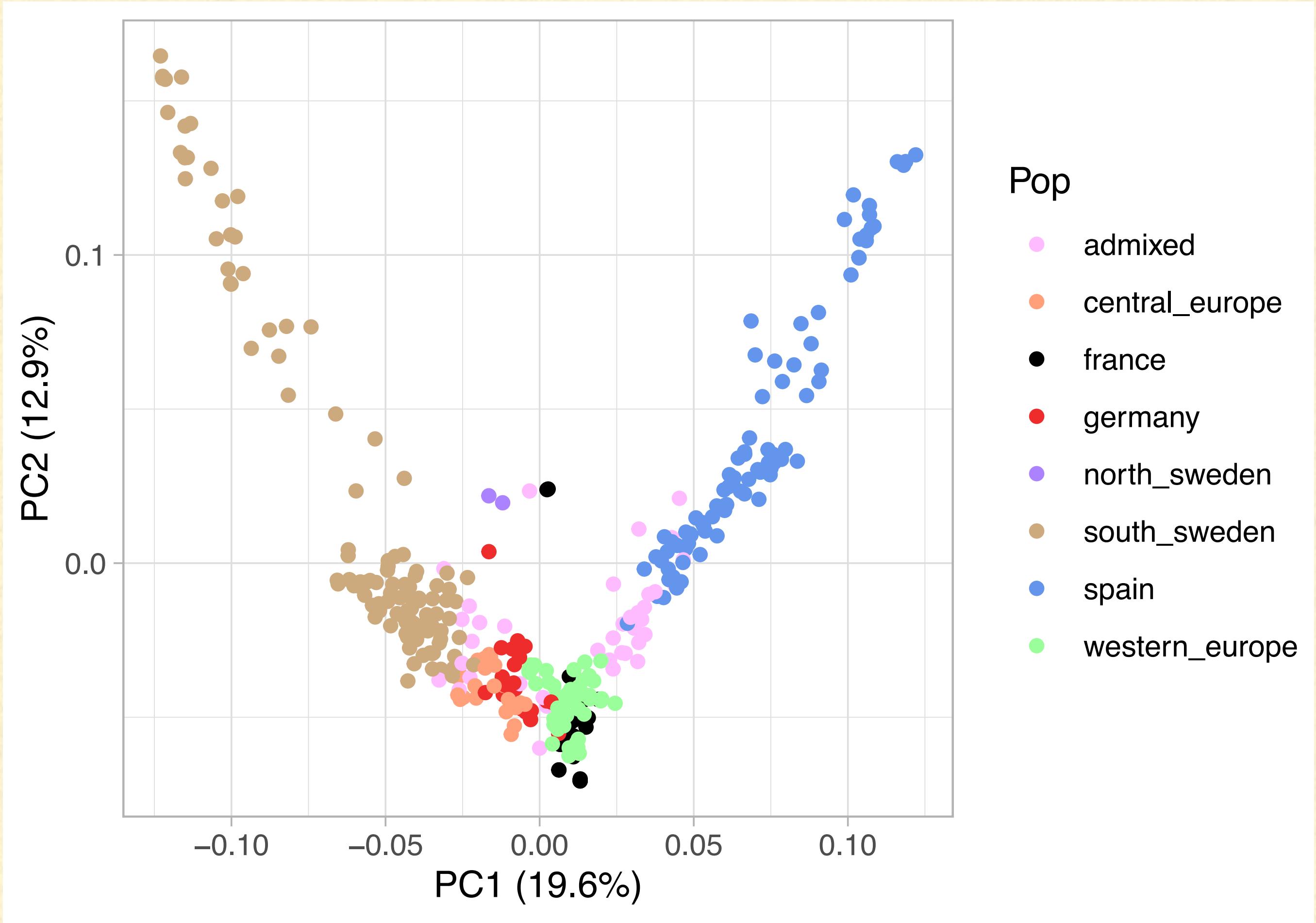
To make bed file and perform PCA

Name of your output

This step and previous steps can be combined, because

this outputs *.**eigenval** and *.**eigenvec** files to plot PCA in R;

and *.bed, *.bim, *.fam files at the same time for GWAS



Example of PCA plots

I. BEFORE

C) Making kinship matrix

```
gemma --bfile prefix_of_plink_bfiles -p phenotype_file.txt -gk -o your_kinship_output_name
```

Prefix of your three plink files

Your phenotype file

In same order as vcf sample order!

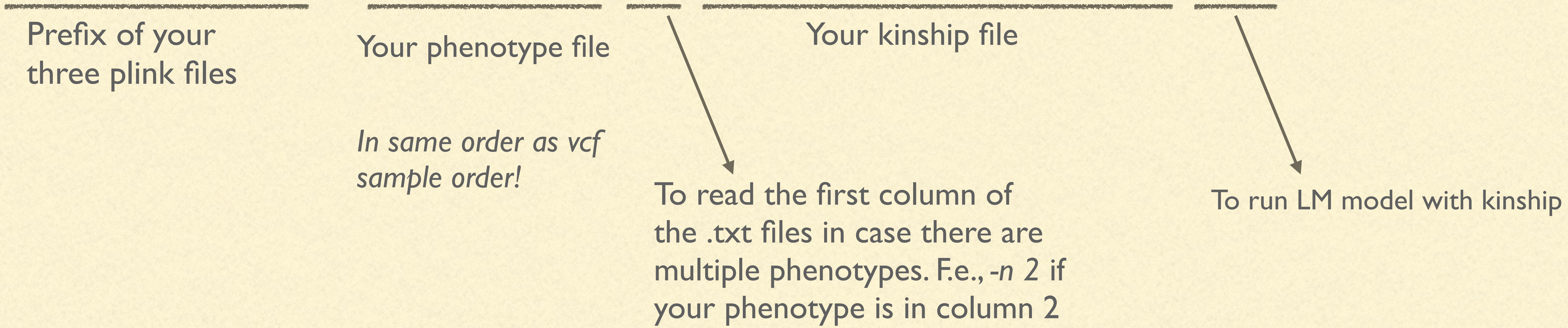
Your kinship name

This command creates three files: ***.cXX.txt** file - this is your kinship

2. RUNNING GWAS

RUN GWAS USING LINEAR MIXED MODEL

```
gemma --bfile prefix_of_plink_bfiles -p phenotype_file.txt -n 1 -k your_kinship_output_name.cXX.txt -lmm 4 -o your_gwas_output_name
```



Prefix of your three plink files

Your phenotype file

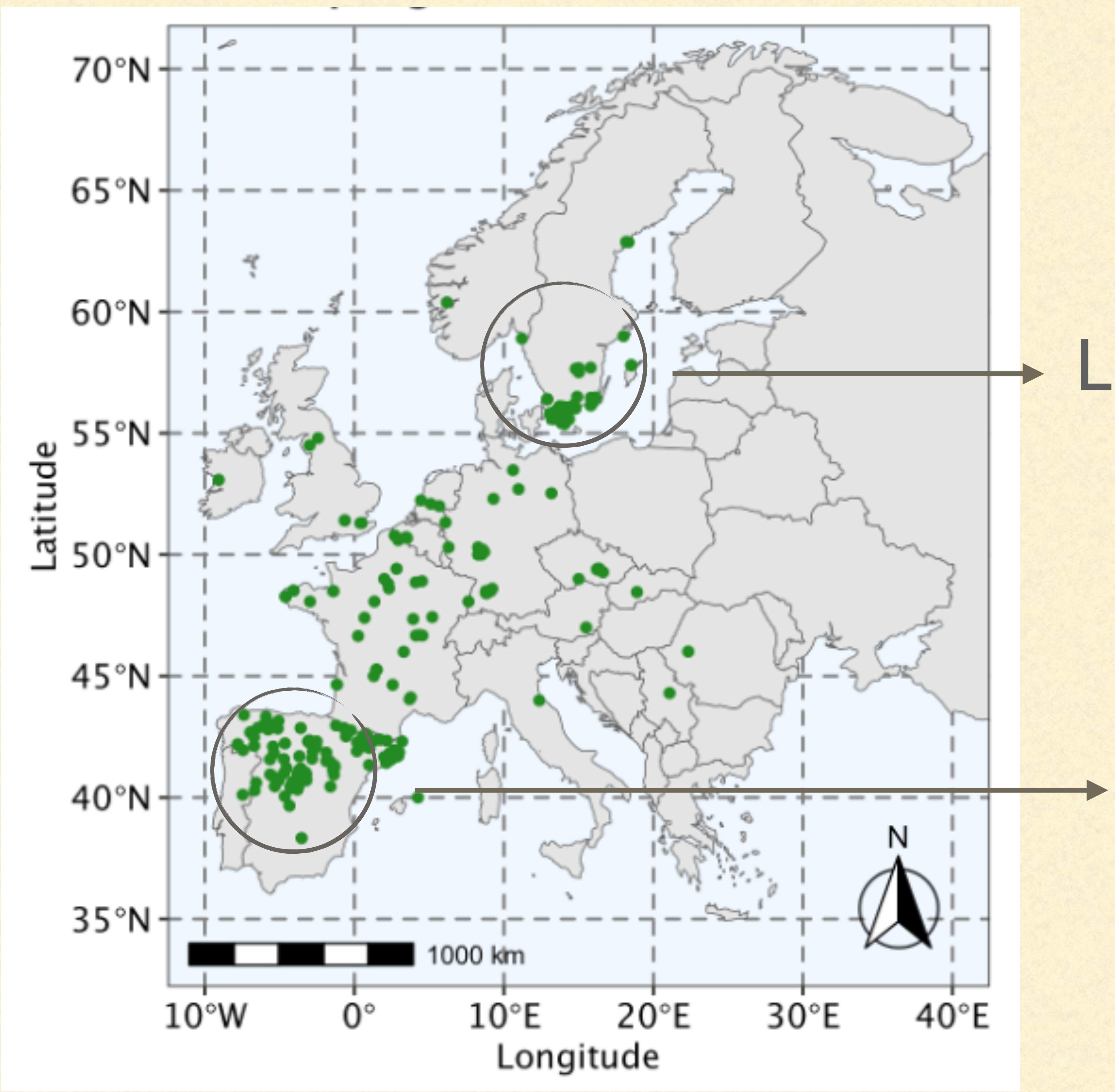
In same order as vcf sample order!

Your kinship file

To read the first column of the .txt files in case there are multiple phenotypes. F.e., *-n* 2 if your phenotype is in column 2

To run LM model with kinship

GWAS VARIATION - GENE ENVIRONMENT ASSOCIATION (GEA)



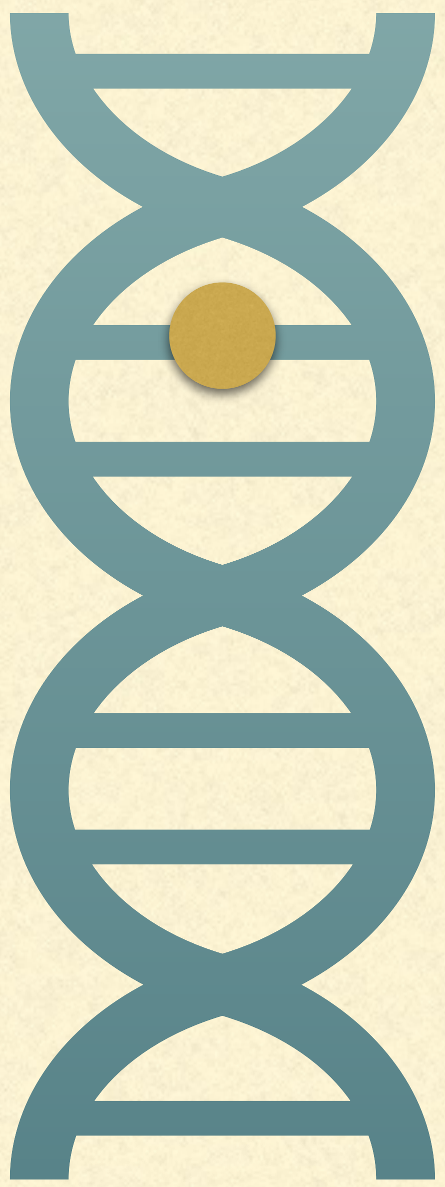
Low temperature

High temperature

“Environment” can be anything - temperature of different seasons, precipitation, light intensity, etc.

GWAS VARIATION - GENE ENVIRONMENT ASSOCIATION (GEA)

Same principle as GWAS, instead of phenotype, we use the bioclimatic factor

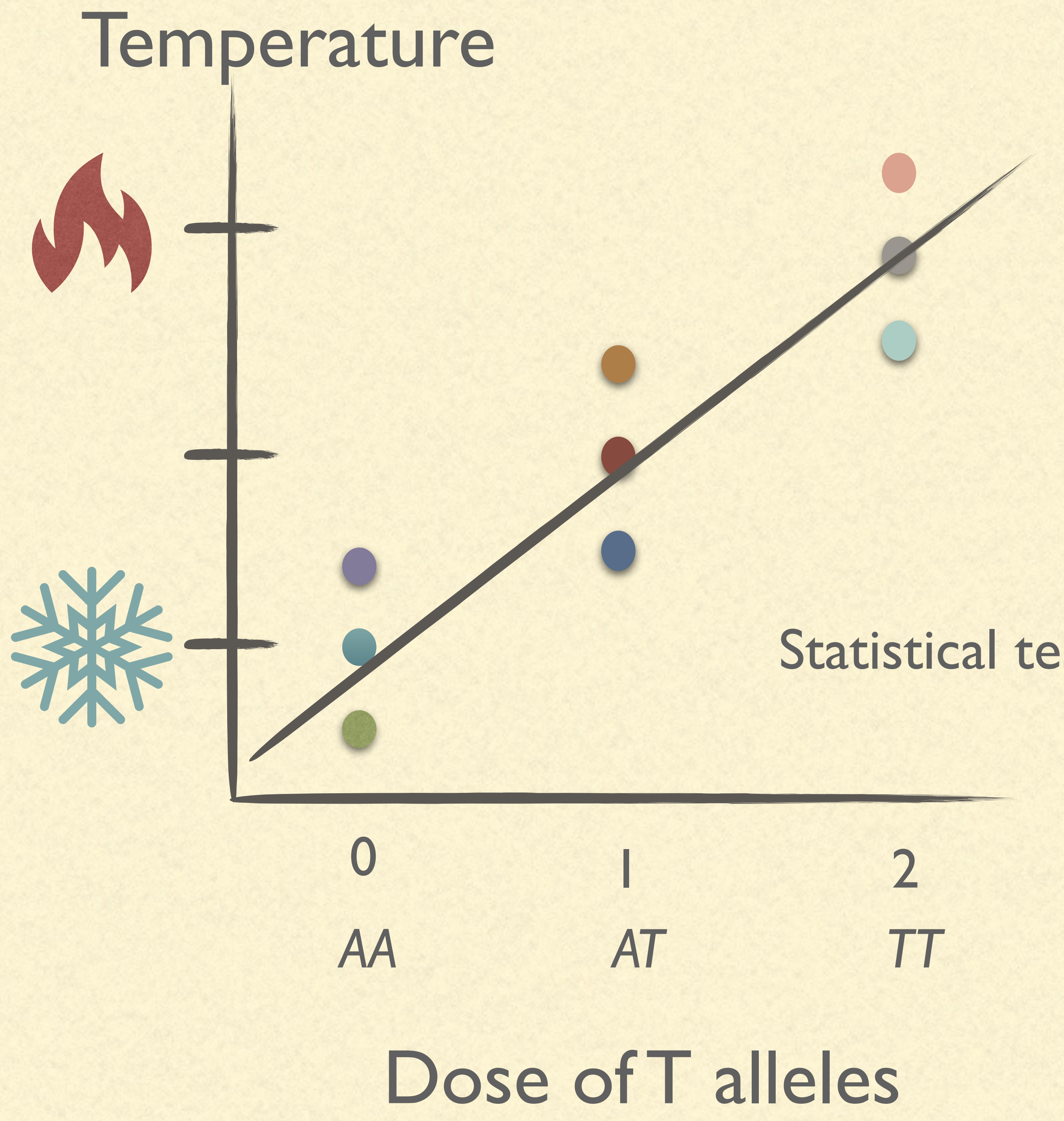


At SNP #1

10% have A allele

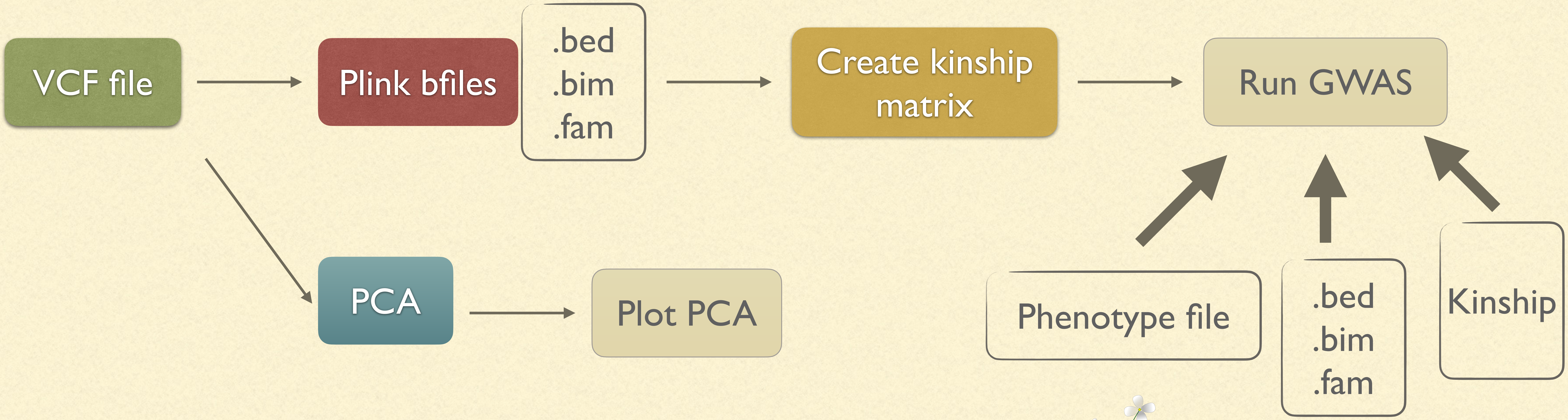
90% have T allele

SNP#1 is associated with temperature



Statistical testing: p value = 0.05

SUMMARY



VCF file

Plink bfiles

.bed
.bim
.fam

Create kinship matrix

Run GWAS

PCA

Plot PCA

Phenotype file

.bed
.bim
.fam

Kinship

——Good luck——-