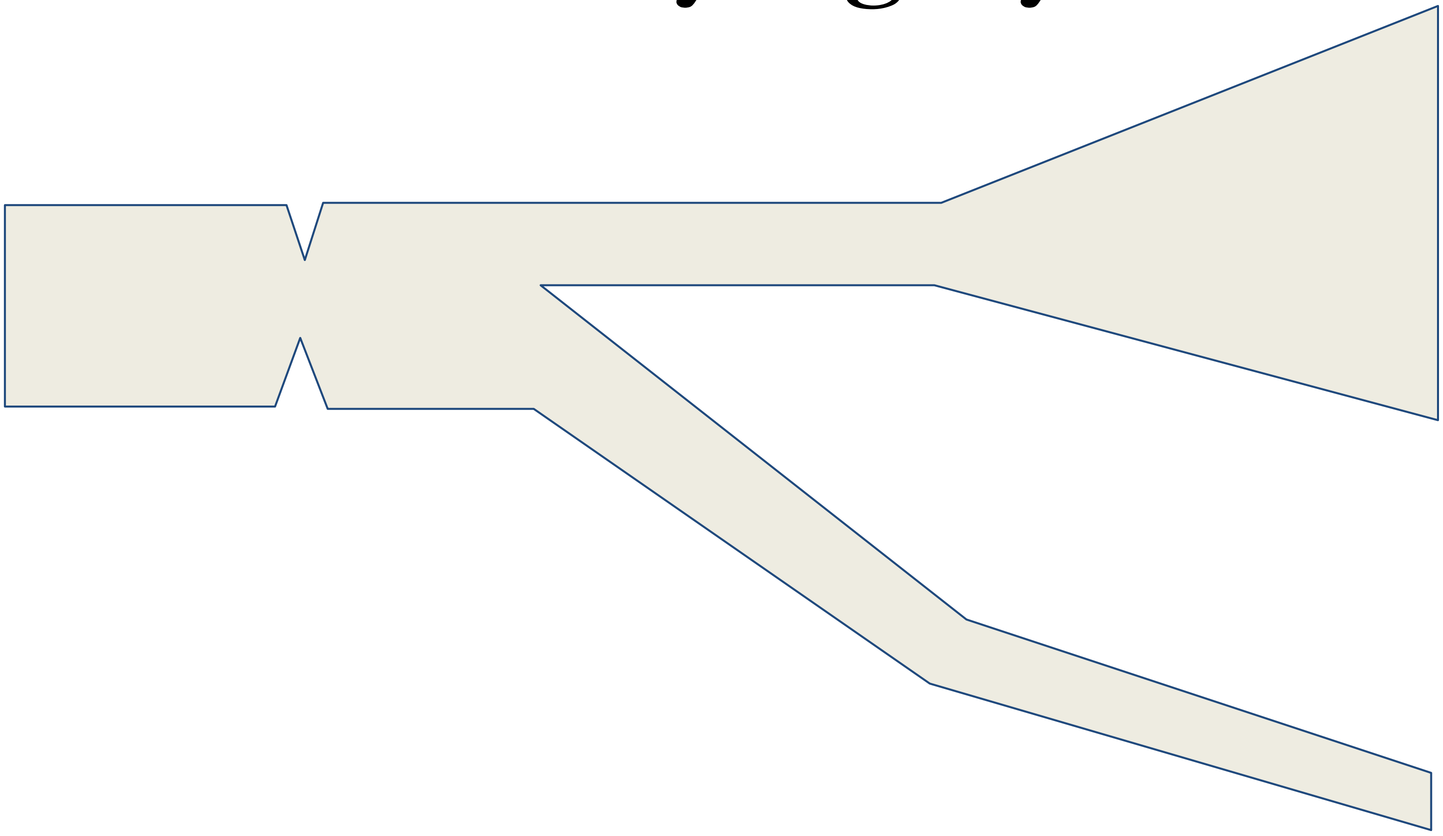
Analysing SNP data

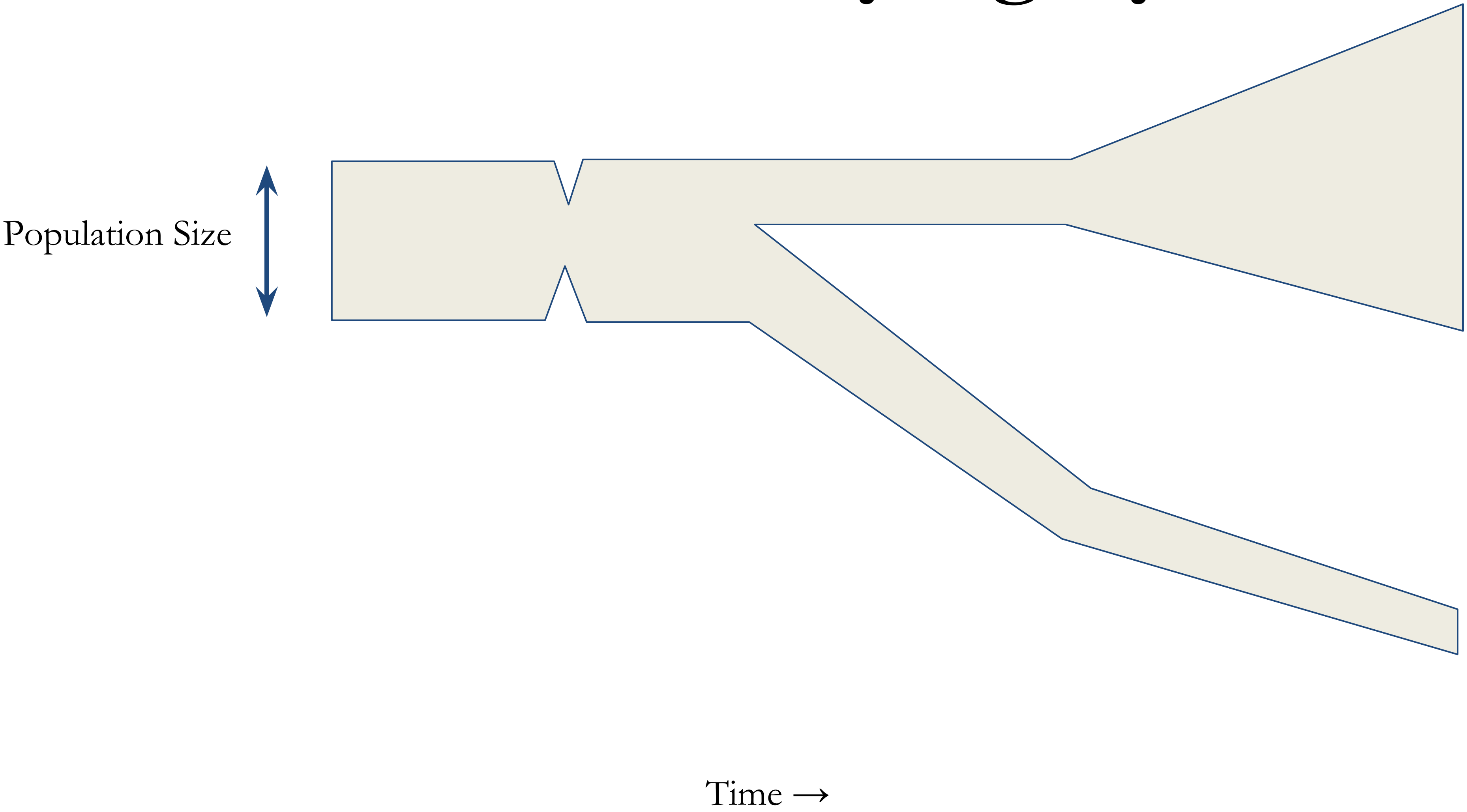
PCA and GWAS

Population Genetic Diversity: An evolutionary legacy

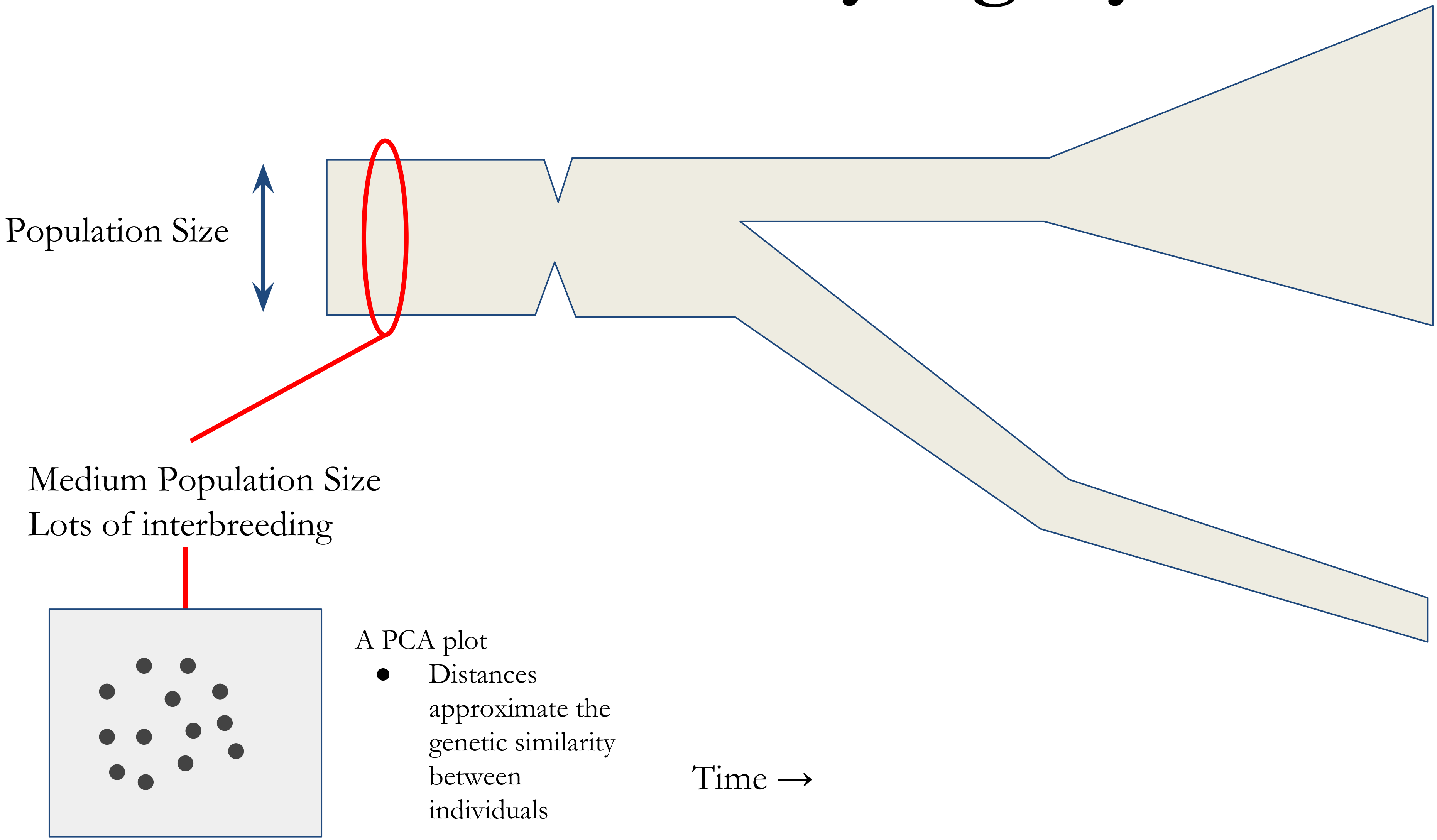


Time →

Population Genetic Diversity: An evolutionary legacy



Population Genetic Diversity: An evolutionary legacy



Population Size

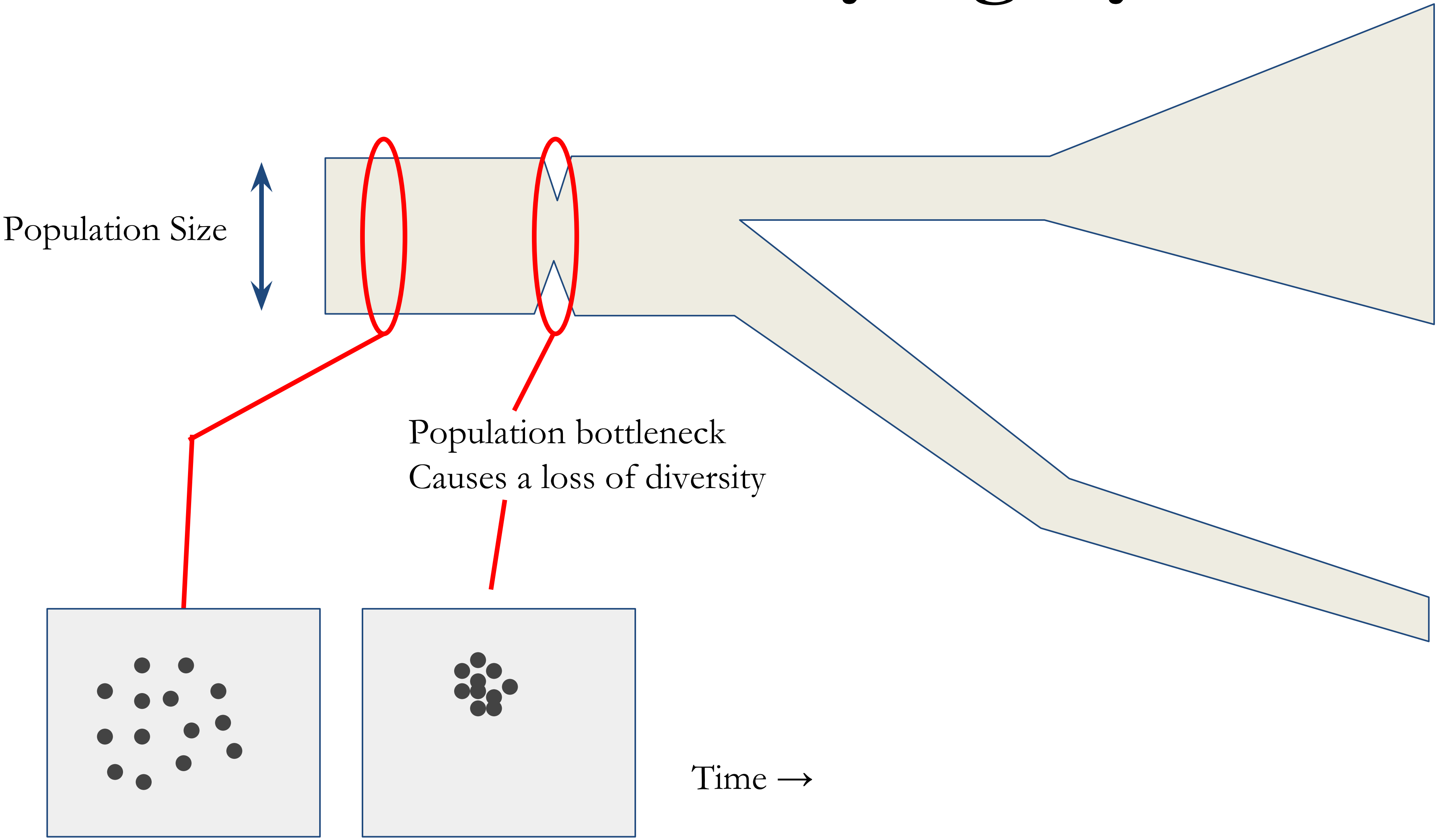
Medium Population Size
Lots of interbreeding

A PCA plot

- Distances approximate the genetic similarity between individuals

Time →

Population Genetic Diversity: An evolutionary legacy

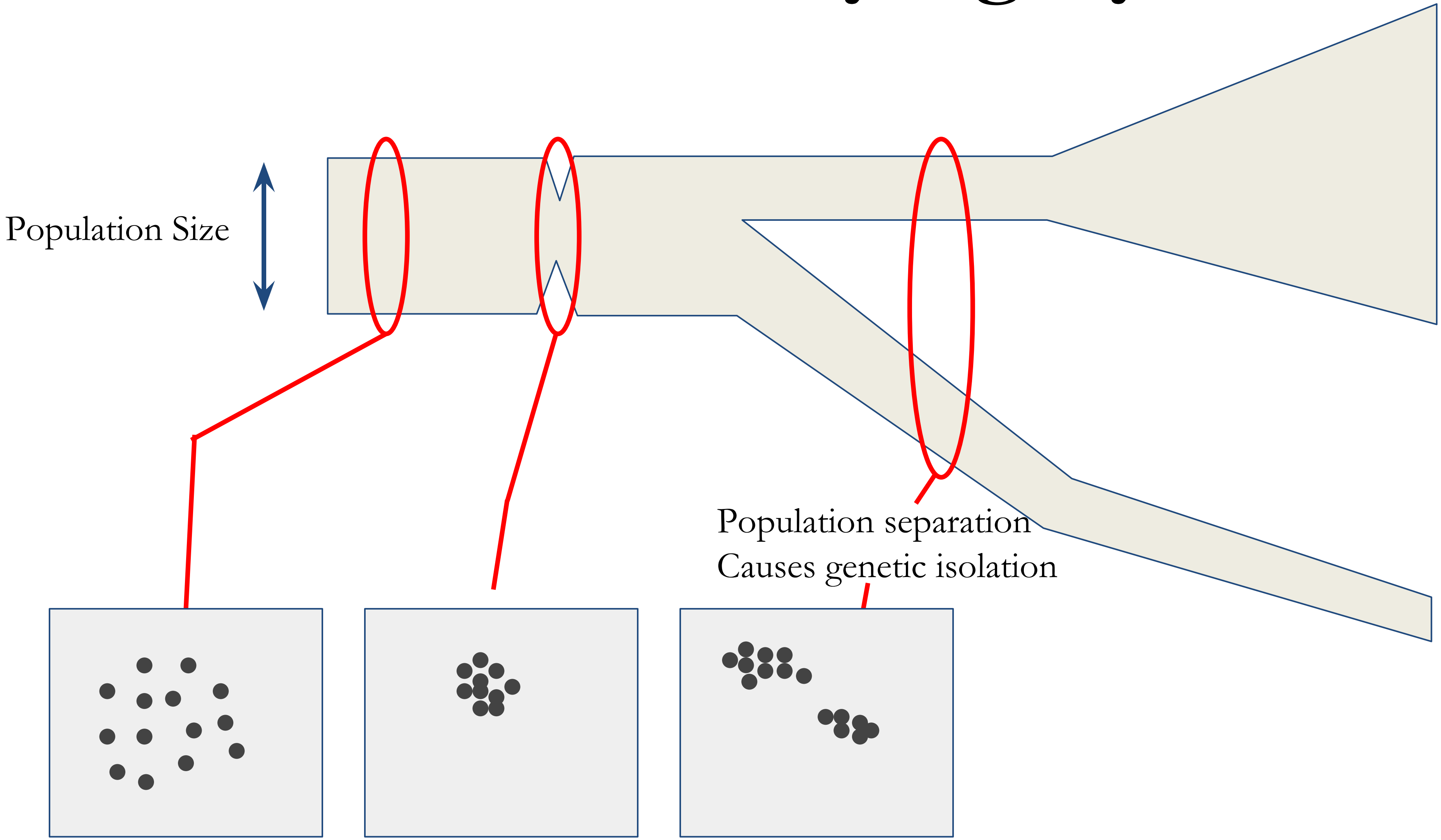


Population Size

Population bottleneck
Causes a loss of diversity

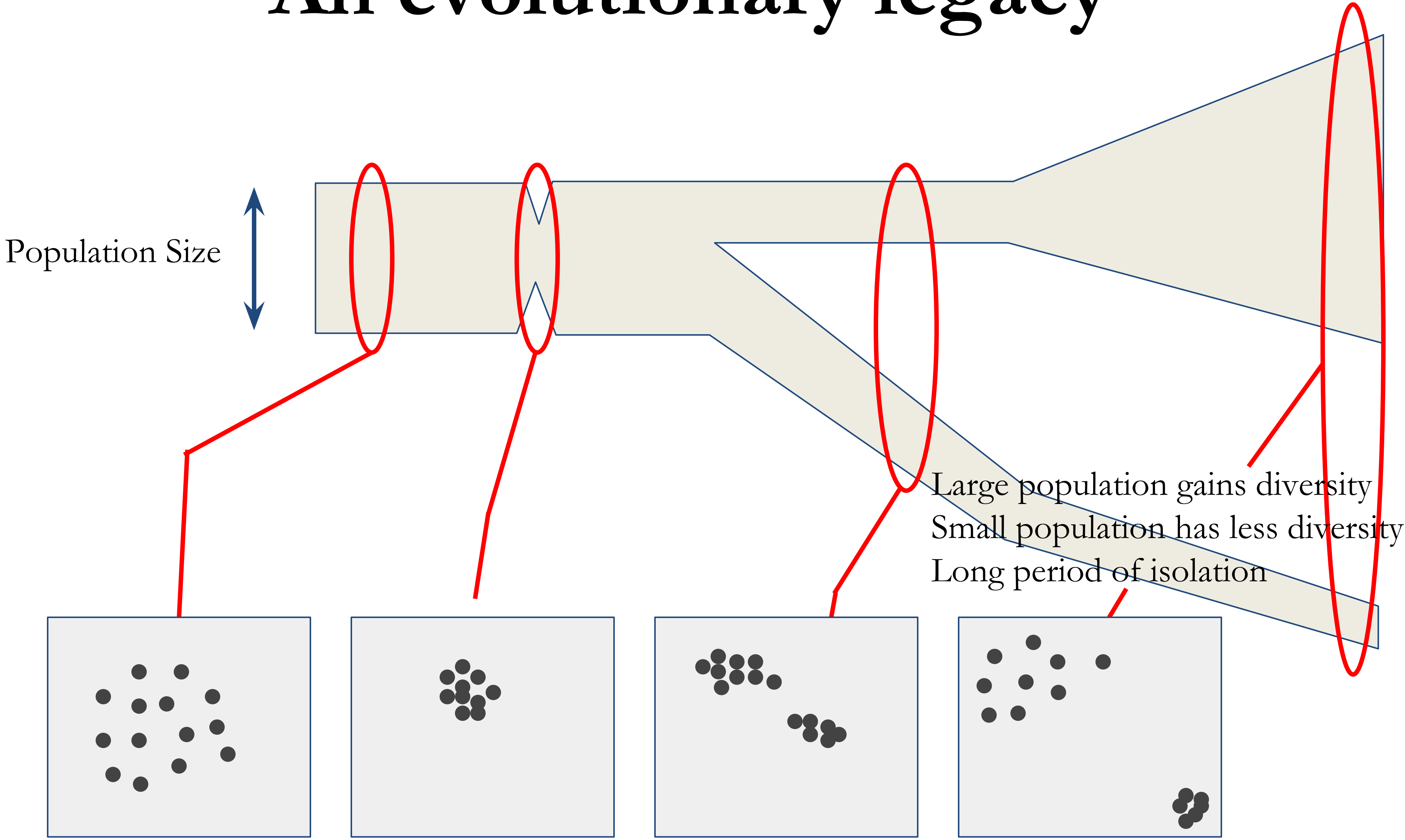
Time →

Population Genetic Diversity: An evolutionary legacy



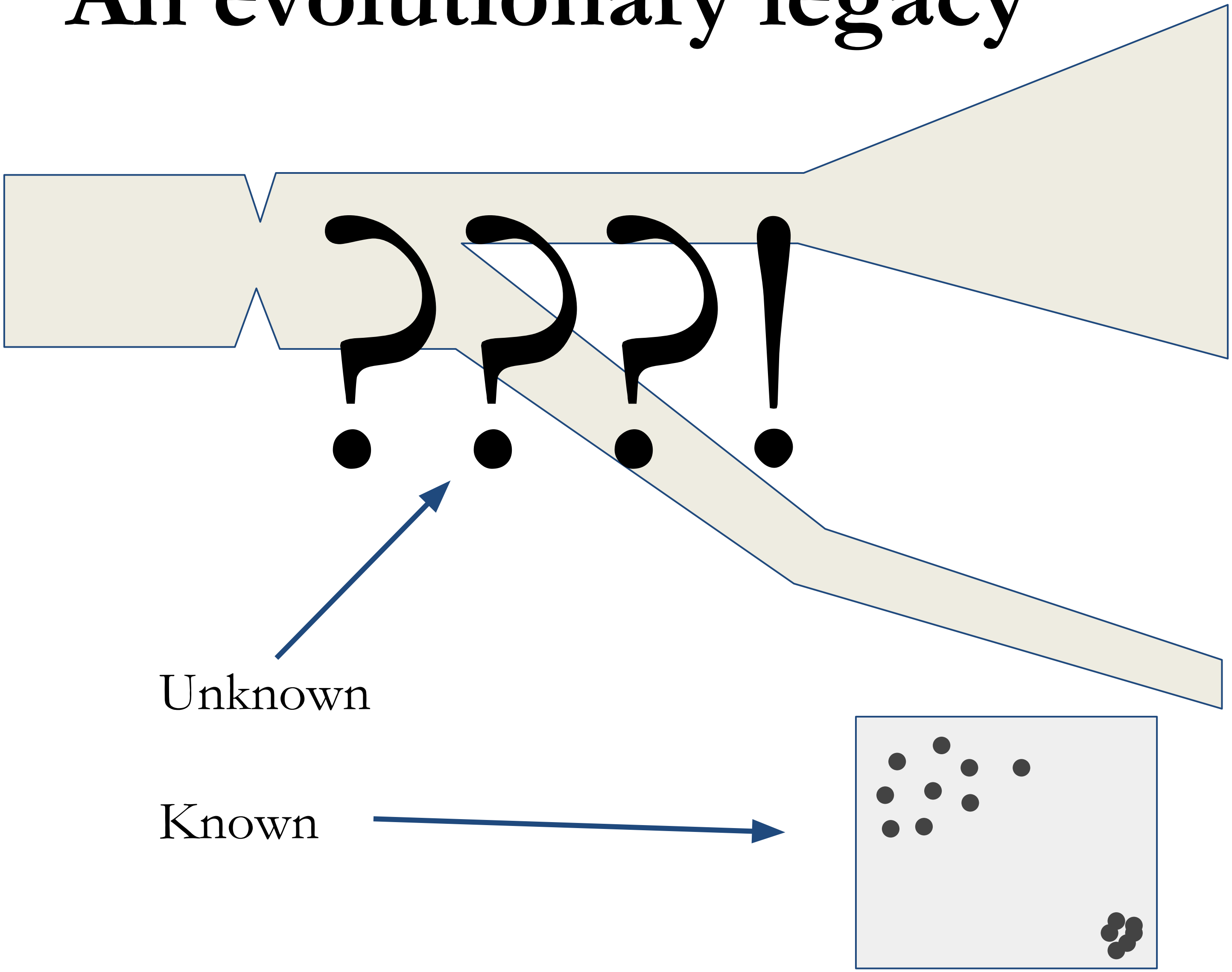
Population separation
Causes genetic isolation

Population Genetic Diversity: An evolutionary legacy



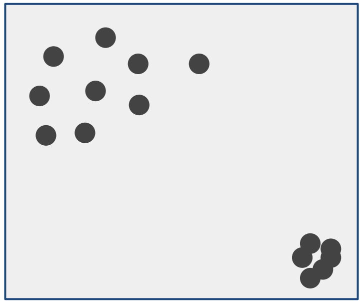
Large population gains diversity
Small population has less diversity
Long period of isolation

Population Genetic Diversity:
An evolutionary legacy



Population Genetic Diversity: A valuable resource

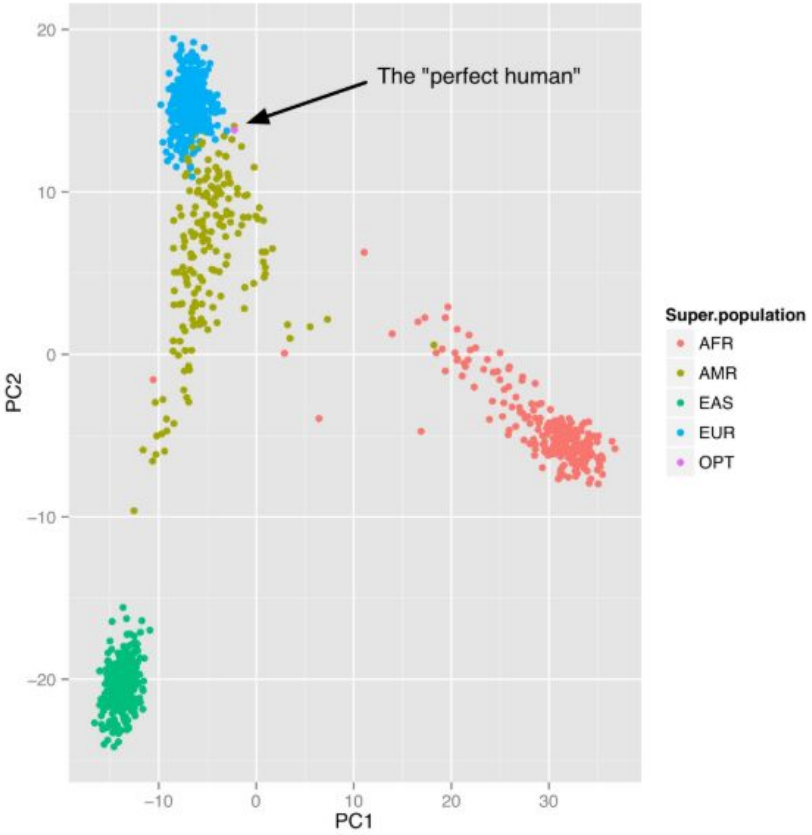
Each individual's genetic material is the product of millions of years of evolution to to tolerate different ...



- Temperatures
- Moisture levels
- Viruses
- Soil conditions
- Day lengths
- Cold periods or frost
- Wind conditions
- Bacteria
- Nematodes, insects
- Fungal pathogens
- Seasonal extremes
- Droughts
- Nutrient deficiencies
- Nutrient excesses
- Salt levels
- Herbivores
- Symbioses
  - … etc etc etc

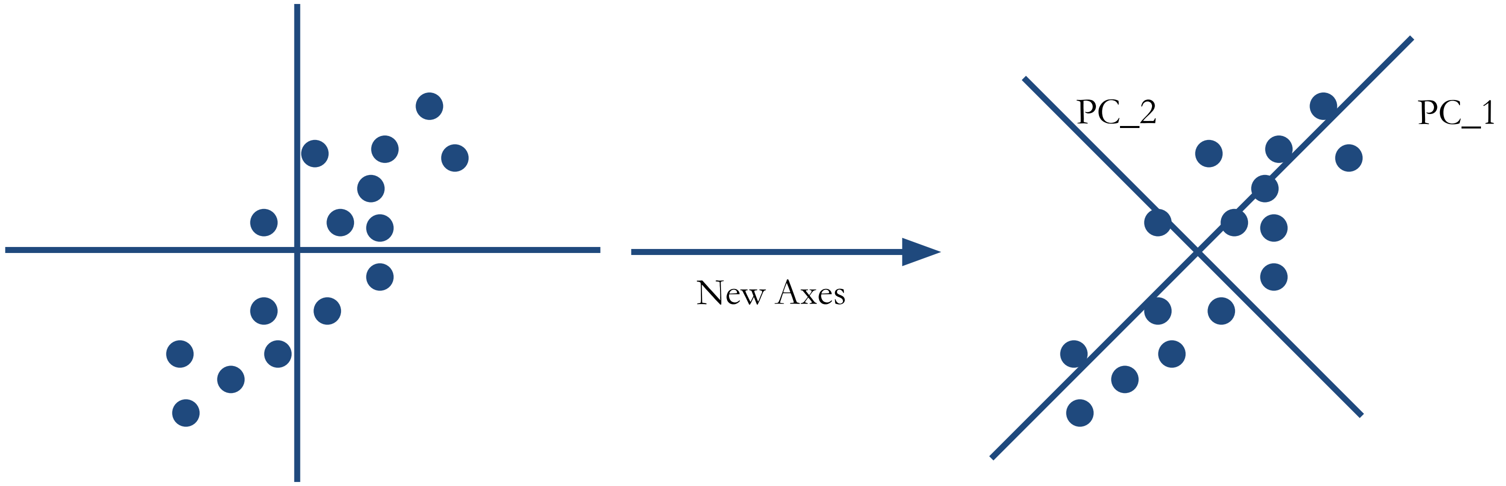
PCA with SNP data:

- An excellent way to visualise diversity, and
- An efficient mathematical way to specify population structure



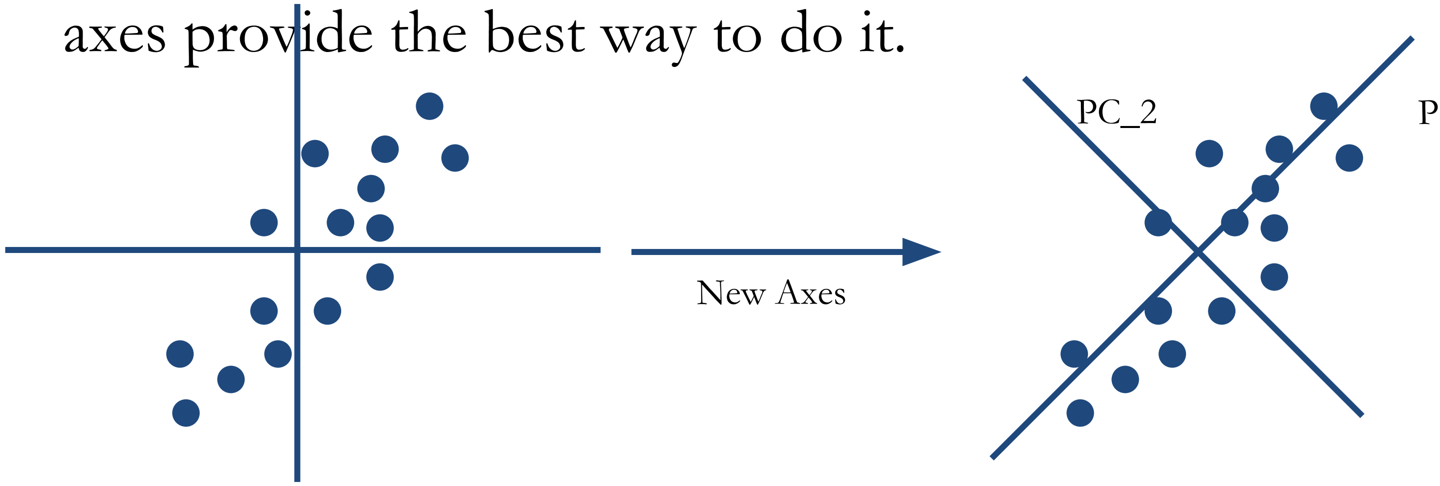
PCA

- A way of specifying new “axes” to the data, so that the new axes express the variation in the data best.



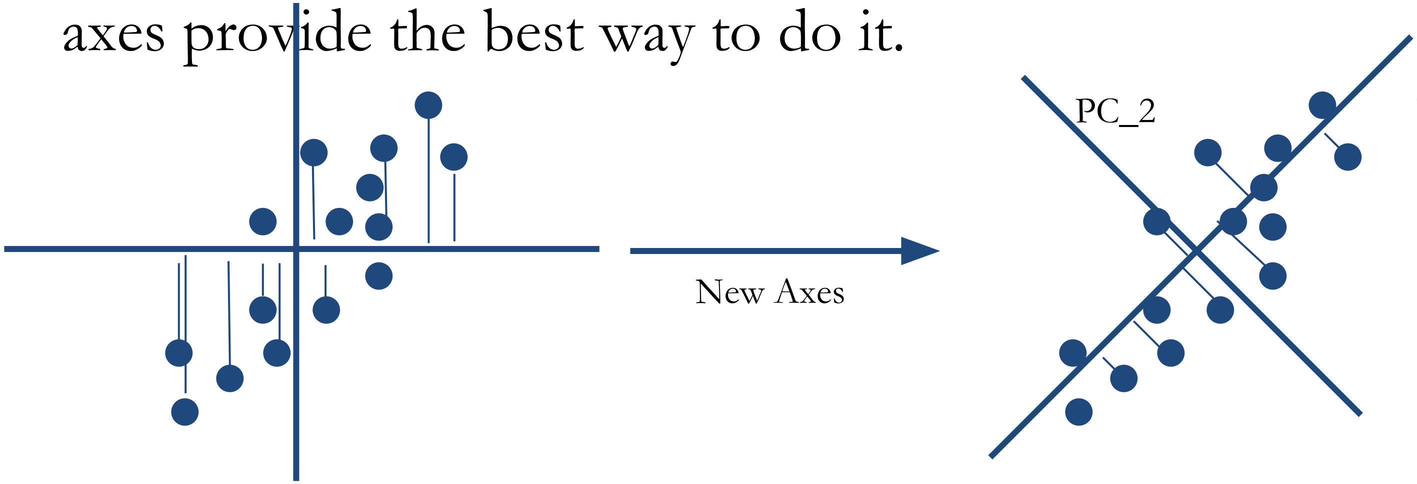
PCA

- A way of specifying “new axes” for the data, so that the new axes (or Principal Components) capture the highest possible variation in the data.
- If we need to describe the data using fewer dimensions, the new axes provide the best way to do it.



PCA

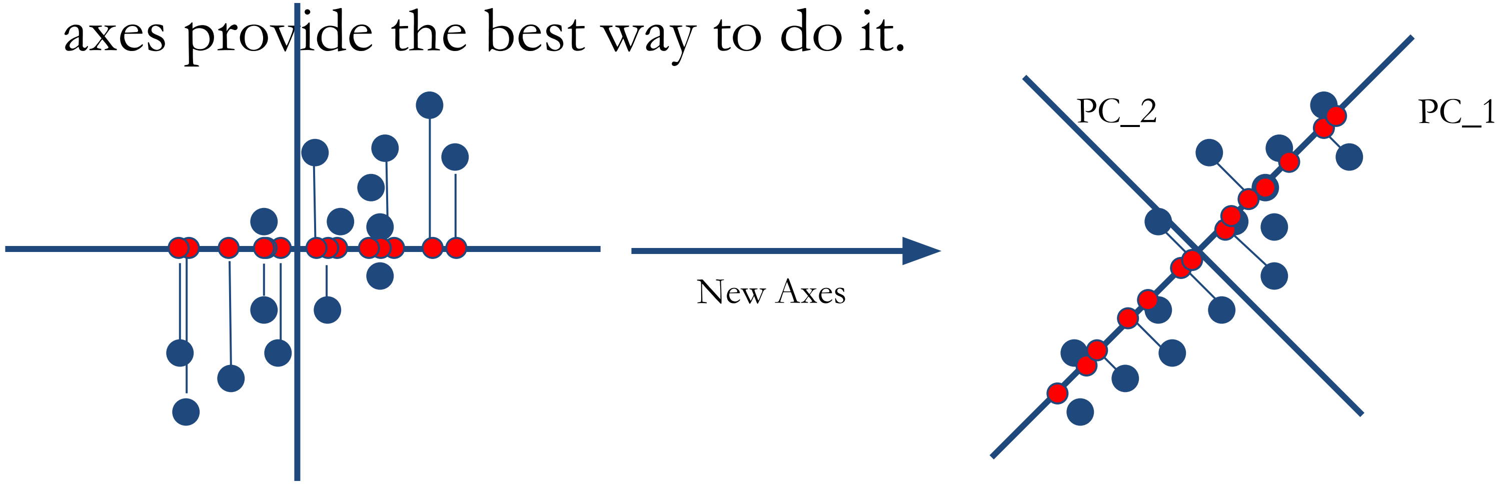
- A way of specifying “new axes” for the data, so that the new axes (or Principal Components) capture the highest possible variation in the data.
- If we need to describe the data using fewer dimensions, the new axes provide the best way to do it.



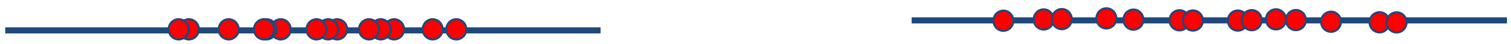
From 2D to 1D:

PCA

- A way of specifying “new axes” for the data, so that the new axes (or Principal Components) capture the highest possible variation in the data.
- If we need to describe the data using fewer dimensions, the new axes provide the best way to do it.

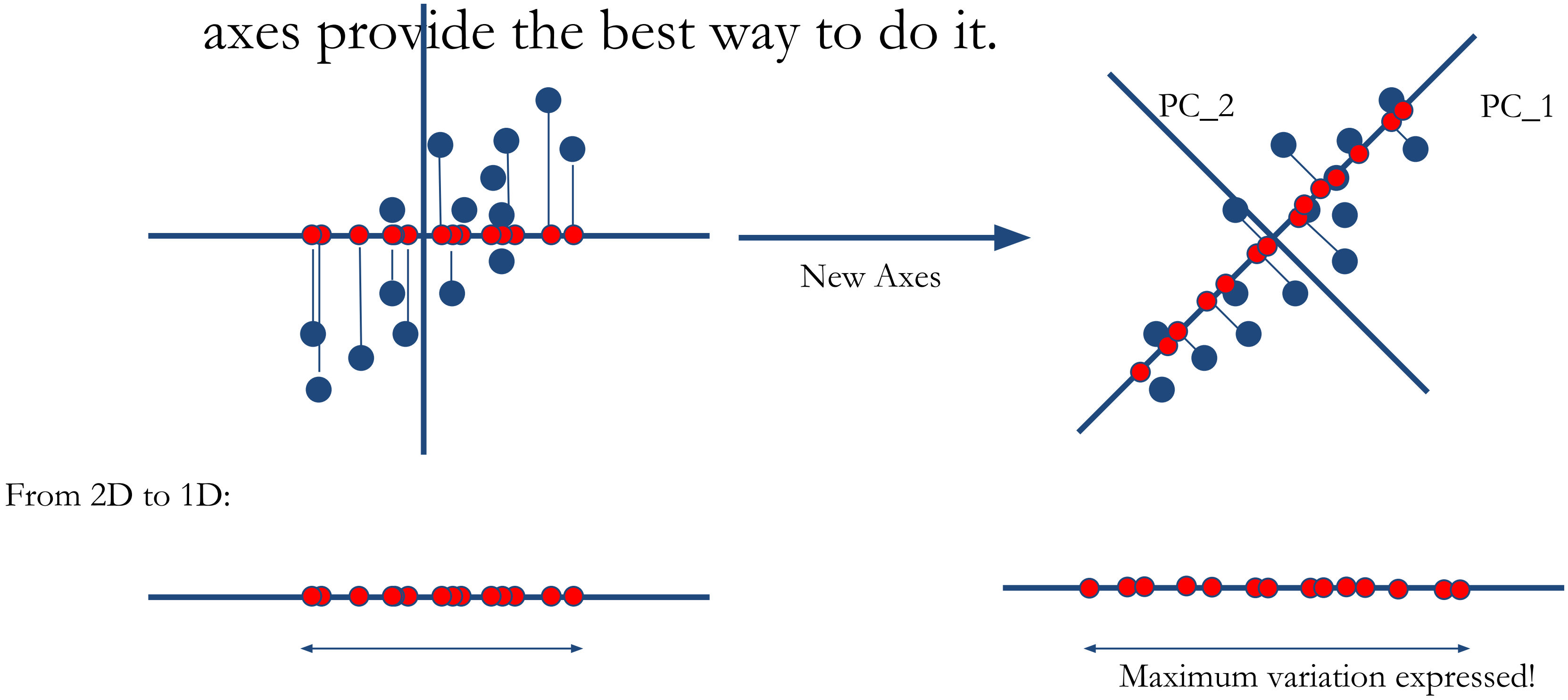


From 2D to 1D:



PCA

- A way of specifying “new axes” for the data, so that the new axes (or Principal Components) capture the highest possible variation in the data.
- If we need to describe the data using fewer dimensions, the new axes provide the best way to do it.



PCA

- A way of specifying “new axes” for the data, so that the new axes (or Principal Components) capture the highest possible variation in the data.
- If we need to describe the data using fewer dimensions, the new axes provide the best way to do it.
- From 3D to 2D:

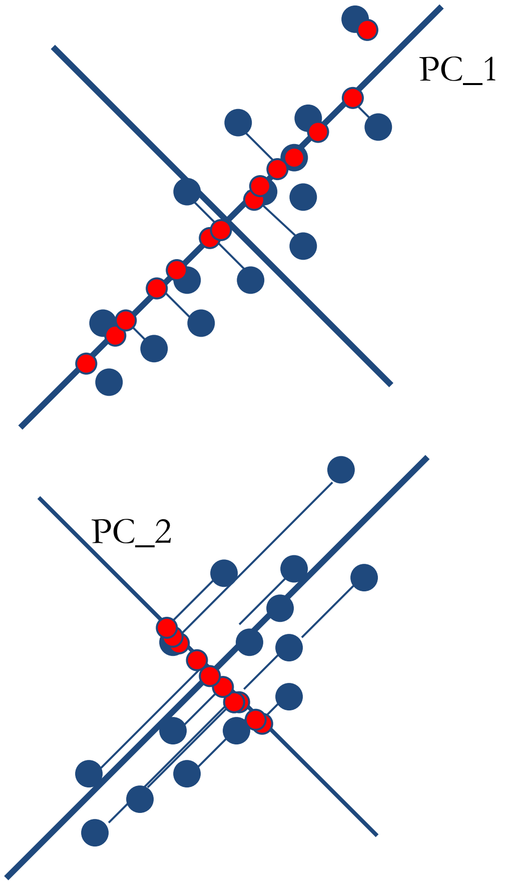
https://www.google.com/search?biw=1440&bih=767&tbm=isch&sa=1&ei=cw2cXLOzAZGblwT-o6TQCQ&q=eigenvectors+3d+gif&oq=eigenvectors+3d+gif&gs_l=img.3...551553.556272..556385...0.0..0.84.386.6......1....1..gws-wiz-img.2TdLFyVmTjY#imgrc=30k1xua4SqZHkM:

PCA

- A way of specifying "new axes" for the data, so that the new axes (or Principal Components) capture the highest possible variation in the data.
- If we need to describe the data using fewer dimensions, the new axes provide the best way to do it.
- From 3D to 2D
- A SNP dataset has as many "dimensions" as there are SNPs.
    - And the values in it are discrete (e.g., 0, 1, and 2)

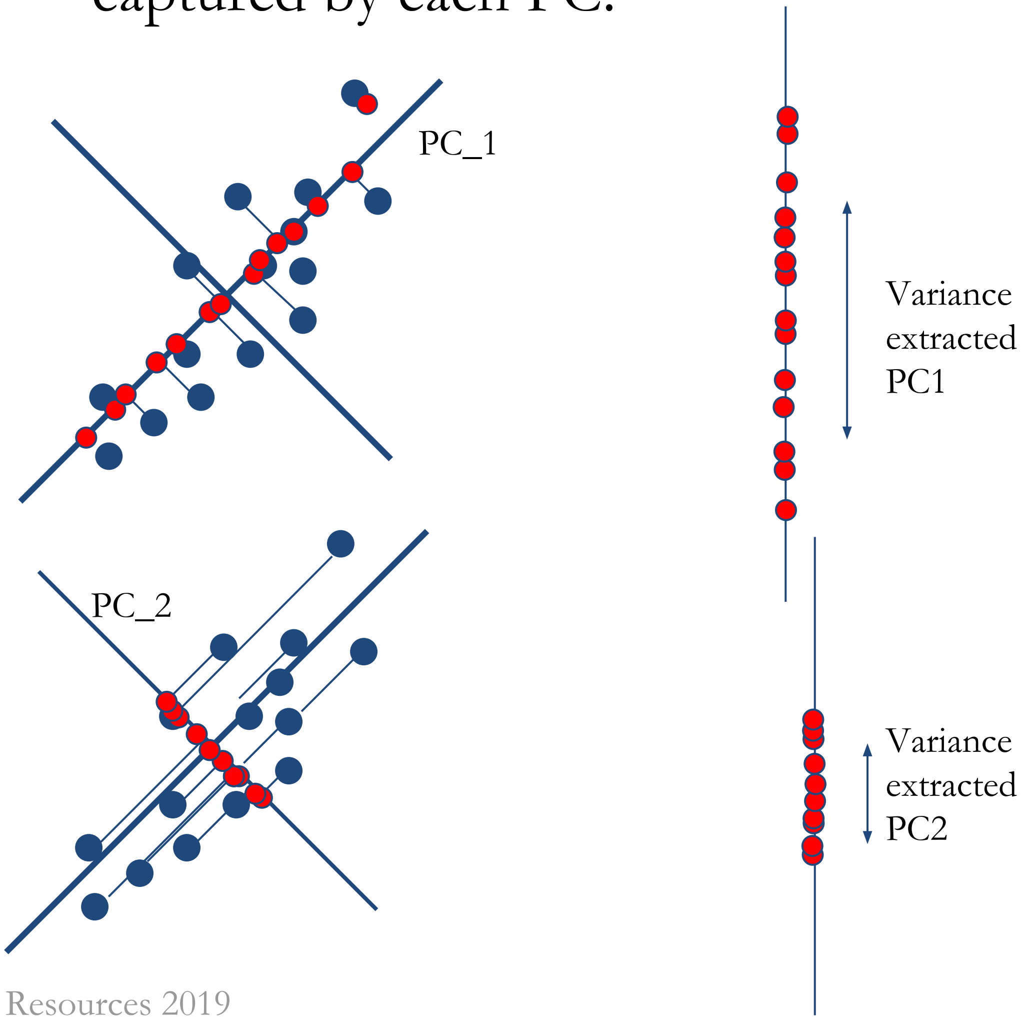
PCA

- A “scree plot” shows how much of the variation in the data is captured by each PC.



PCA

- A “scree plot” shows how much of the variation in the data is captured by each PC.



PC_1

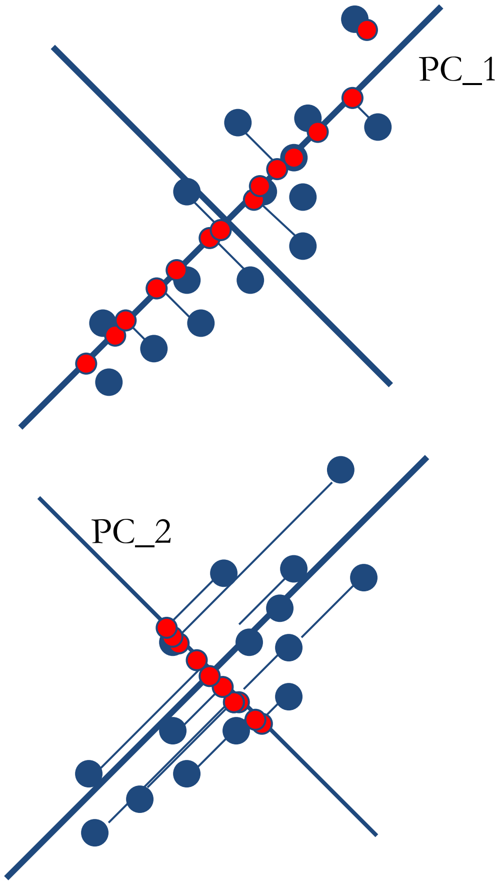
PC_2

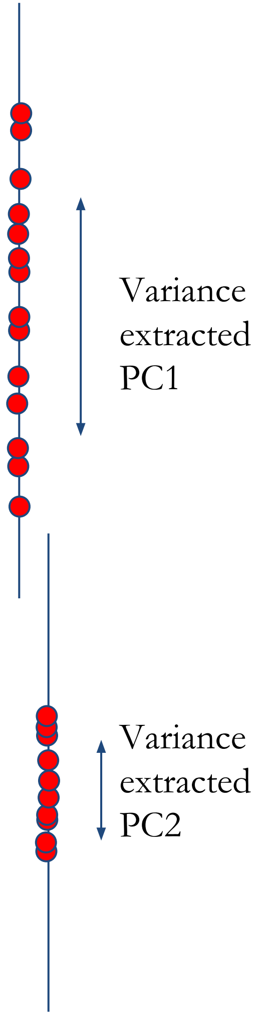
Variance
extracted
PC1

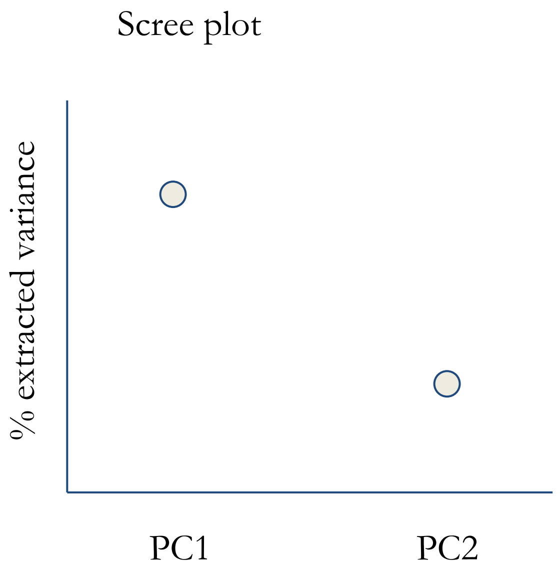
Variance
extracted
PC2

PCA

- A “scree plot” shows how much of the variation in the data is captured by each PC.

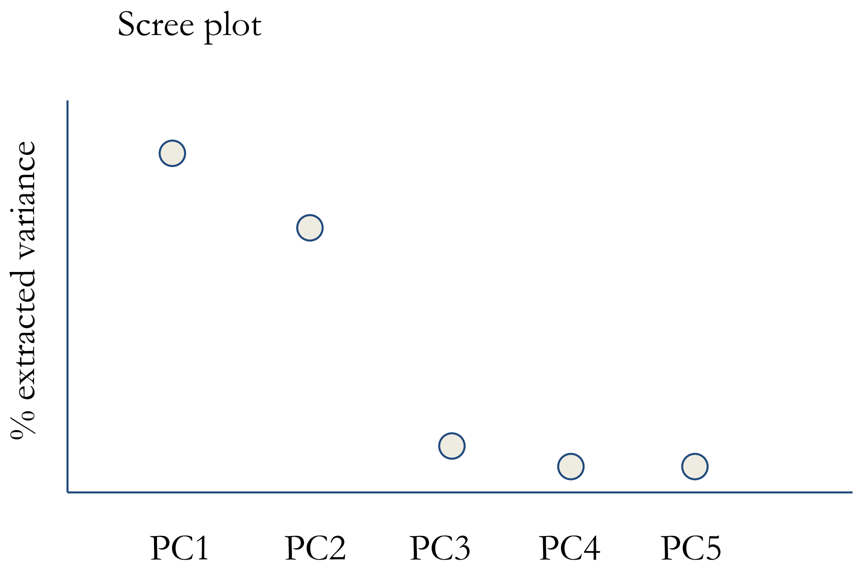




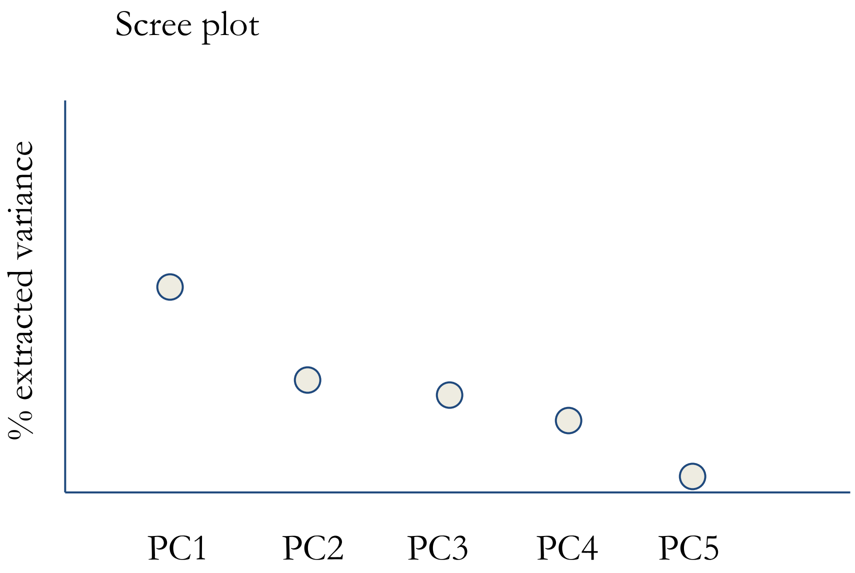


PCA

- A “scree plot” shows how much of the variation in the data is captured by each PC.
    - Interpreting scree plots



PCs 1 and 2 can be used to describe the dataset very well, because they explain most of the variance!

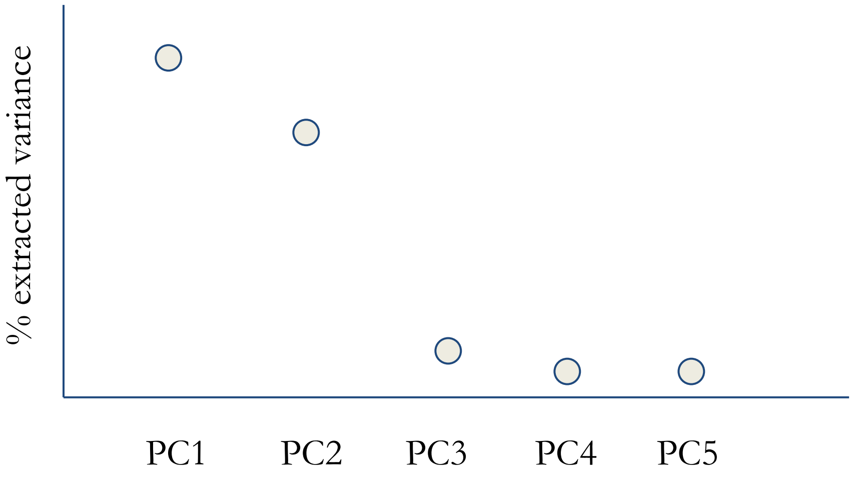


Need more PCs to describe most of the variation in the data. Use PCs 1--4?

PCA

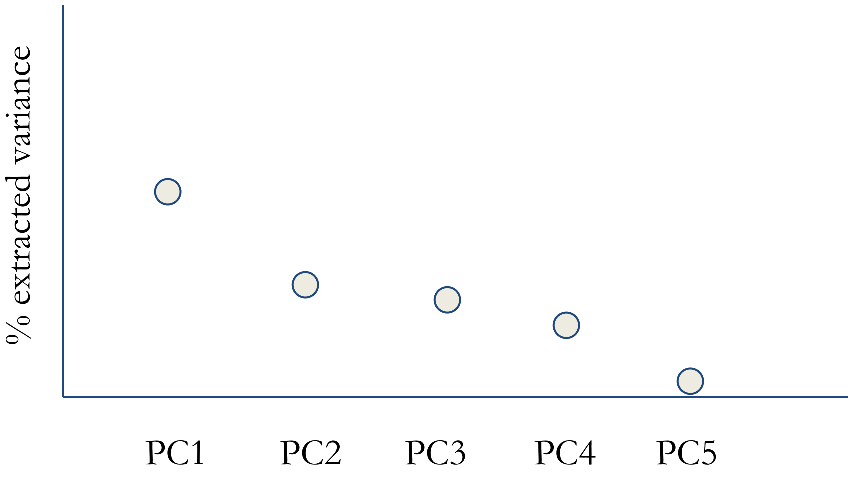
- A “scree plot” shows how much of the variation in the data is captured by each PC.
  - Interpreting scree plots

Scree plot



PCs 1 and 2 can be used to describe the dataset very well, because they explain most of the variance!

Scree plot



Need more PCs to describe most of the variation in the data. Use PCs 1--4?

A geneticist might use a rule like “use as many PCs as needed to describe 95% of the variation”.

Genetic Associations

- Testing whether genetic markers (SNPs in our case) are associated with a phenotype.
    - That's 'associated with' … not 'cause'!

Genetic Associations

- Testing whether genetic markers (SNPs in our case) are associated with a phenotype.
  - That's 'associated with' … not 'cause'!
  - Association suggests *linkage to* the cause of the phenotype

Genetic Associations

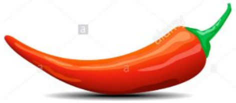
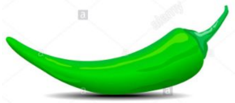
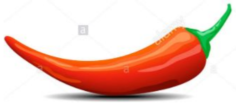
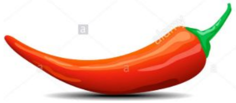
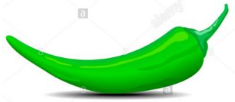
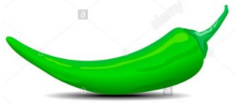
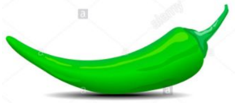
- Testing whether genetic markers (SNPs in our case) are associated with a phenotype.
    - That's 'associated with' … not 'cause'!
    - Association suggests *linkage to* the cause of the phenotype
    - Means the SNP *may* be physically near a gene that causes the phenotype

Genetic Associations

- Testing whether genetic markers (SNPs in our case) are associated with a phenotype.
    - That’s ‘associated with’ … not ‘cause’!
    - Association suggests *linkage to* the cause of the phenotype
    - Means the SNP **may** be genetically near a gene that causes the phenotype
    - Means the SNP allele can be used as a marker to screen for the (probable) presence of the causal gene(s)

Genetic Associations

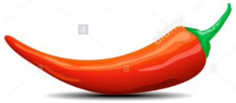
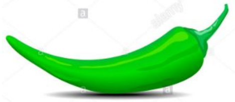
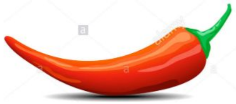
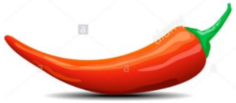
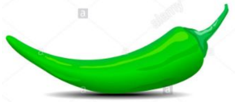
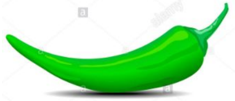
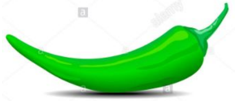
Intuition

Phenotype	Genotype
	2
	0
	2
	2
	0
	0
	0

Strong association!

Genetic Associations

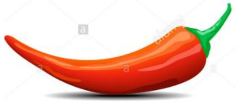
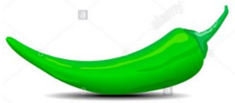
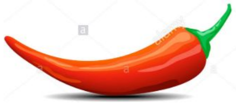
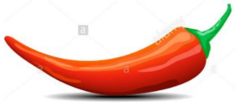
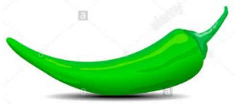
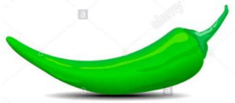
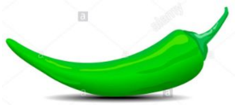
Intuition

Phenotype	Genotype
	0
	0
	2
	0
	0
	2
	0

Weak / no association

Genetic Associations

Intuition

Phenotype	Genotype
	2
	0
	2
	2
	0
	2
	0

Maybe associated .... ?

But how do we judge
what is significant?

Genetic Associations

Any appropriate statistical test can be used to judge whether a SNP is likely associated with a phenotype, e.g.:

Chi squared

Phenotype

		2	0
Phenotype		20	52
		32	32

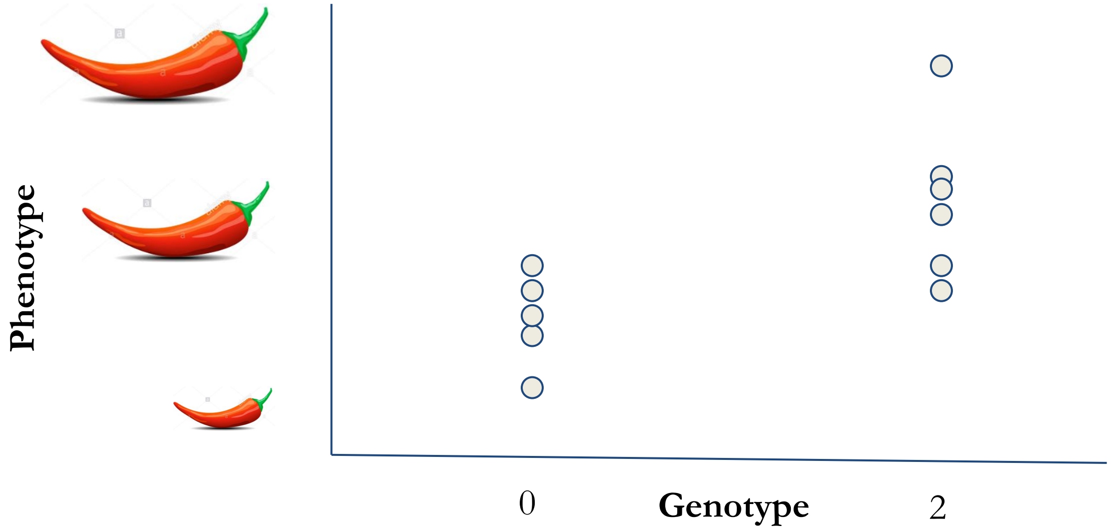
Null hypothesis: The genotype and the phenotype are independent

… generate a p-value ...

Genetic Associations

Any appropriate statistical test can be used to judge whether a SNP is likely associated with a phenotype, e.g.:

t-test



Null Hypothesis: The mean sizes of plants of the two genotypes are equal

… generate a p-value ….

Genetic Associations

Any appropriate statistical test can be used to judge whether a SNP is likely associated with a phenotype, e.g.:

We will use a linear regression

- Fast
- Helps account for population structure
- *You are not required to know details*, just:
    - ○ Null hypothesis: “The SNP allele is required to explain the phenotype.”

Genetic Associations

Any appropriate statistical test can be used to judge whether a SNP is likely associated with a phenotype, e.g.:

We will use a linear regression

- Fast
- Helps account for population structure
- Not required to know details, just:
    - Null hypothesis: “The SNP allele is required to explain the phenotype.”
- The test is done for every SNP we chose in the genome … hence

Genome Wide Association Study (GWAS)

Genetic Associations

Any appropriate statistical test can be used to judge whether a SNP is likely associated with a phenotype, e.g.:

We will use a linear regression

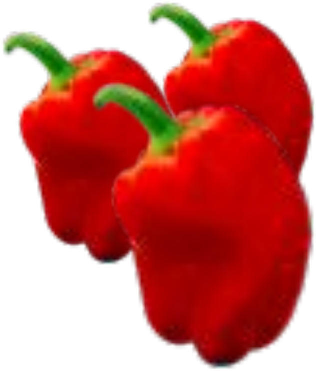
- Fast
- Helps account for population structure
- Not required to know details, just:
  - Null hypothesis: "The SNP allele is required to explain the phenotype."
- The test is done for every SNP we chose in the genome … hence

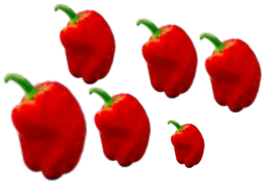
Genome Wide Association Study (GWAS)

(we just use chromosome 1H)

Associations and population structure

```
CGATTCGTGCGGGGCTCCTCTCAGGATGCTTAAA
CGATTCGTGCGGGGCTCCTCTCAGGATGCTTAAA
CGATTCGTGCGGGGCTCCTCTCAGGATGCTTAAA
CGATTCGTGCGGGGCTCCTCTCAGGATGCTTGAA
CGATTCGTGCGGGGCTCCTCTCAGGATGCTTGAA
CGATTCGCGCGGGGCTCCTCTCAGGATGCTTGAA
CGATTCGCGCGGGGCTCCTCTCAGGATGCTTGAA
CGATTCGCGCGGGGCTCCTCTCAGGATGCTTGAA
CGATTCGCGCGGGGCTCCTCTCAGGATGCTTGAA
CGATTCGCGCGGGGCTCCTCTCAGGATGCTTGAA
```





Does either SNP associate with fruit size?

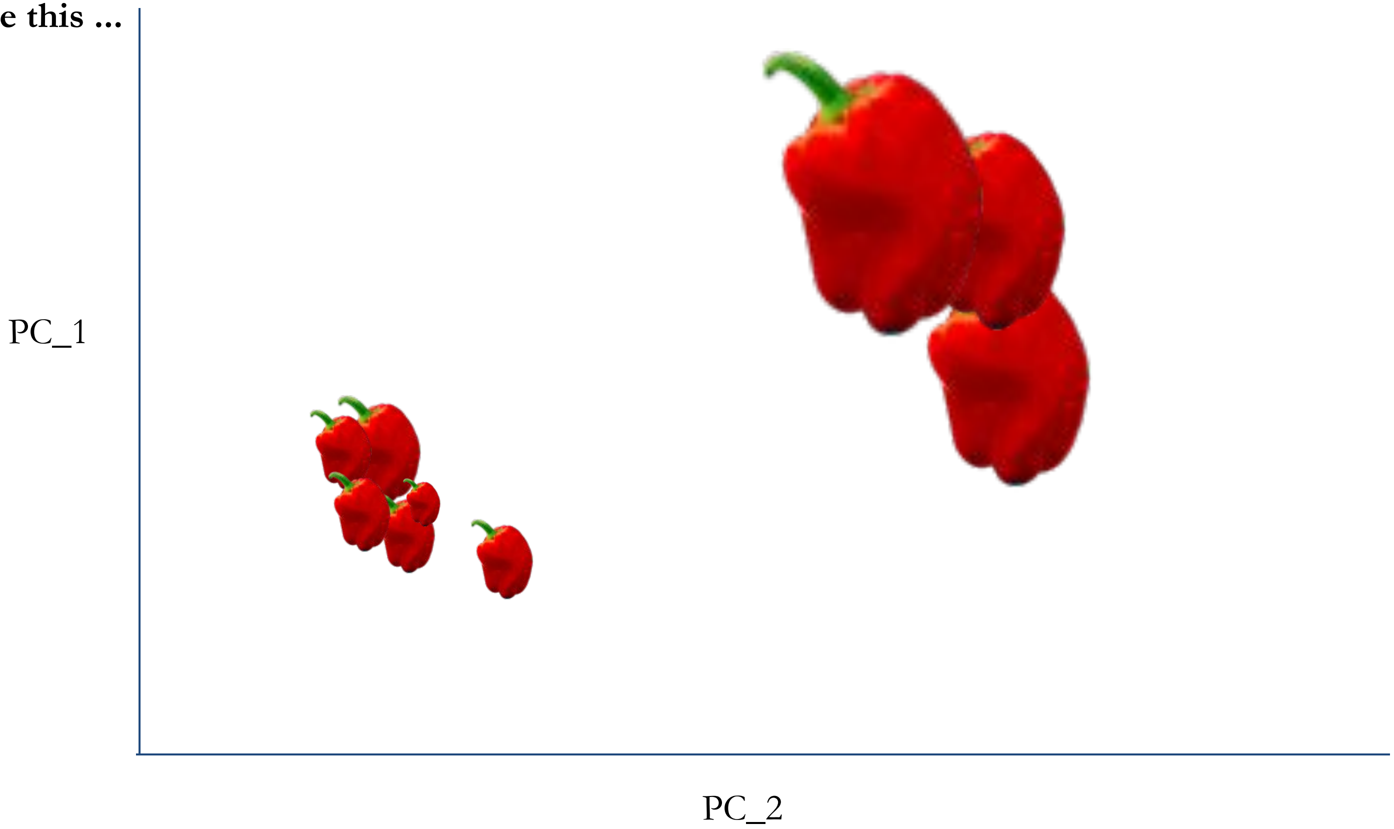
Associations and population structure

AGCCCTGTCTCTAGGCTCTCGATTCGTGCGGGGCTCCTCTCAGGATGCTTAAA
AGCCCTGTCTCTAGGCTCTCGATTCGTGCGGGGCTCCTCTCAGGATGCTTAAA
AGCCCTGTCTCTAGGCTCTCGATTCGTGCGGGGCTCCTCTCAGGATGCTTAAA
AGCCCCGTCTCTGGGCTCTCGATTCGTGCGGGGCTCCTCTCAGGATGCTTGAA
AGCCCCGTCTCTGGGCTCTCGATTCGTGCGGGGCTCCTCTCAGGATGCTTGAA
AGCCCCGTCTCTGGGCTCTCGATTCGCGCGGGGCTCCTCTCAGGATGCTTGAA
AGCCCCGTCTCTGGGCTCTCGATTCGCGCGGGGCTCCTCTCAGGATGCTTGAA
AGCCCCGTCTCTGGGCTCTCGATTCGCGCGGGGCTCCTCTCAGGATGCTTGAA
AGCCCCGTCTCTGGGCTCTCGATTCGCGCGGGGCTCCTCTCAGGATGCTTGAA
AGCCCCGTCTCTGGGCTCTCGATTCGCGCGGGGCTCCTCTCAGGATGCTTGAA

What about now?

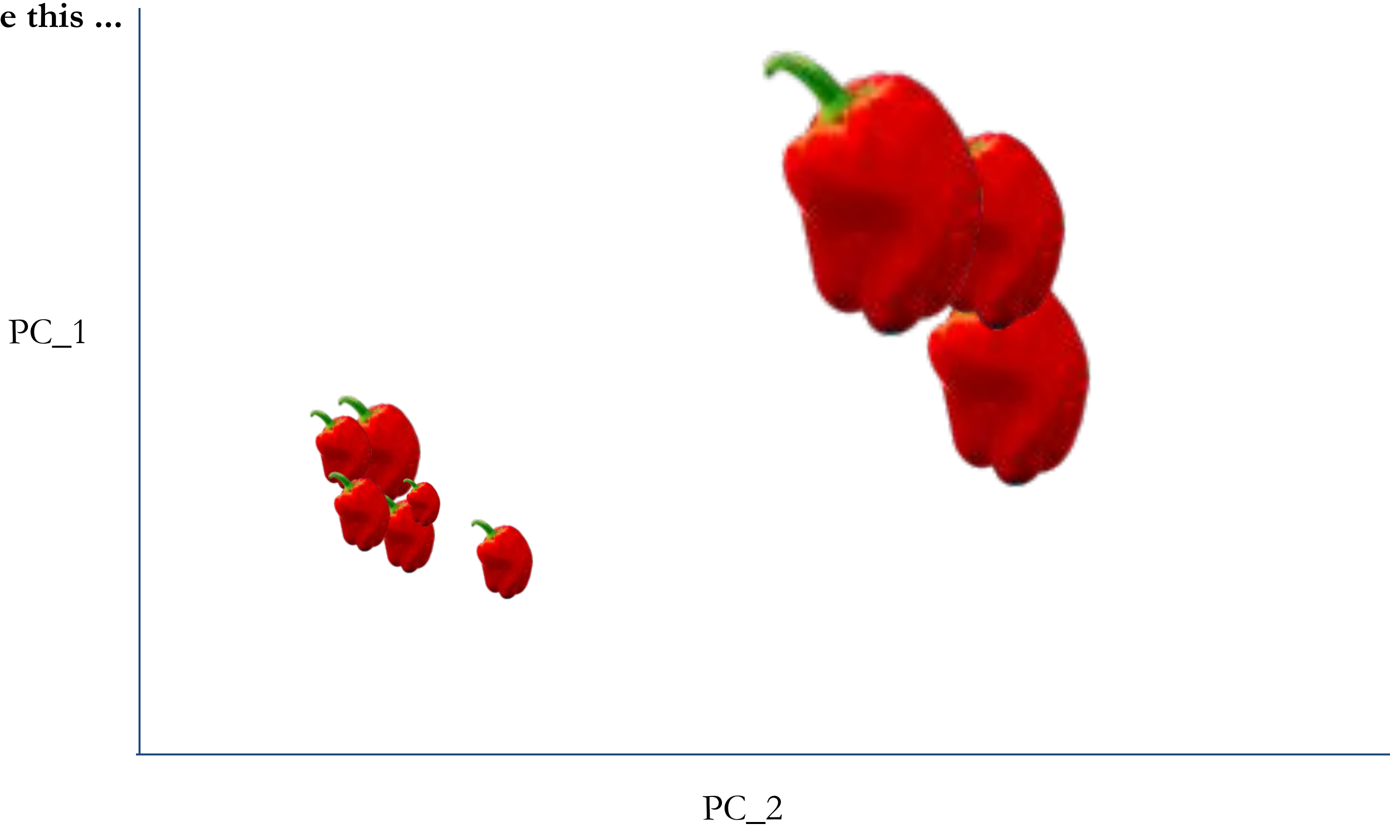
Associations and population structure

A PCA plot would look a bit like this …



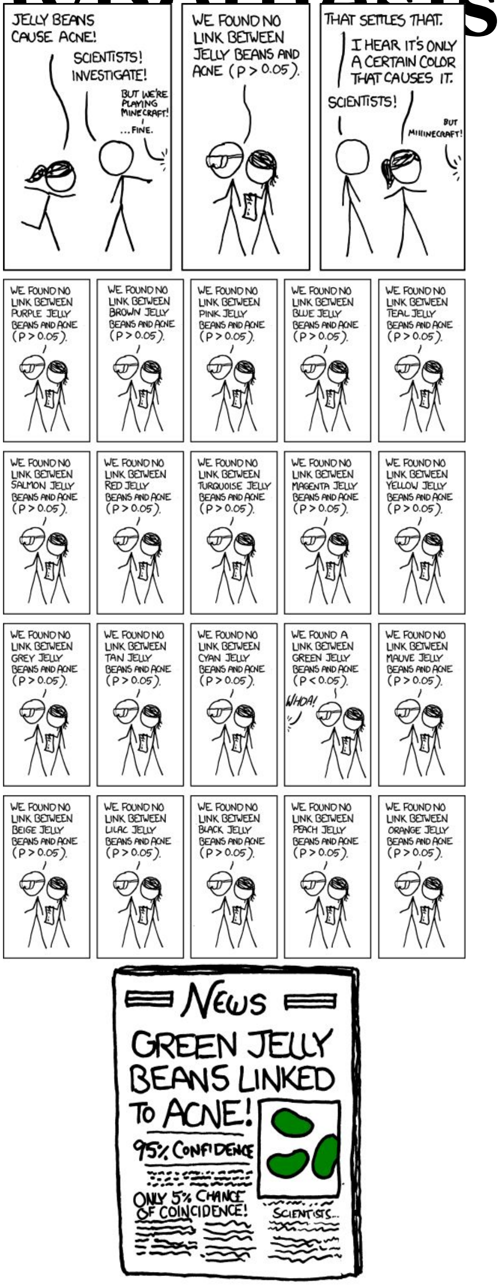
Associations and population structure

A PCA plot would look a bit like this ...

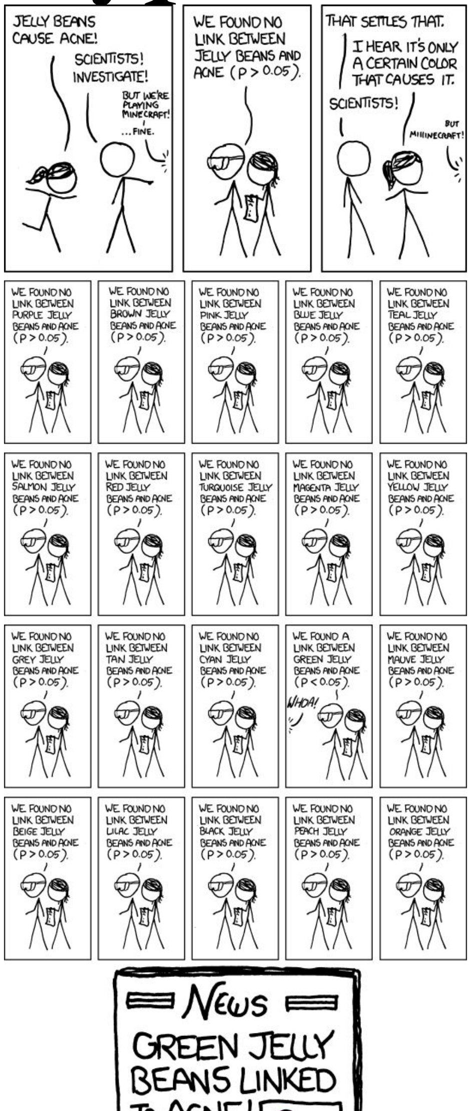


- The principal components are able to predict the phenotype!
  - i.e., population structure is a strong influence
- We can control for this in the linear model
  - The model will test how much extra predictive power the SNP allele gives us, when the population structure (summarised by the PCs) are also used to predict the phenotype.

Multiple hypothesis testing ...

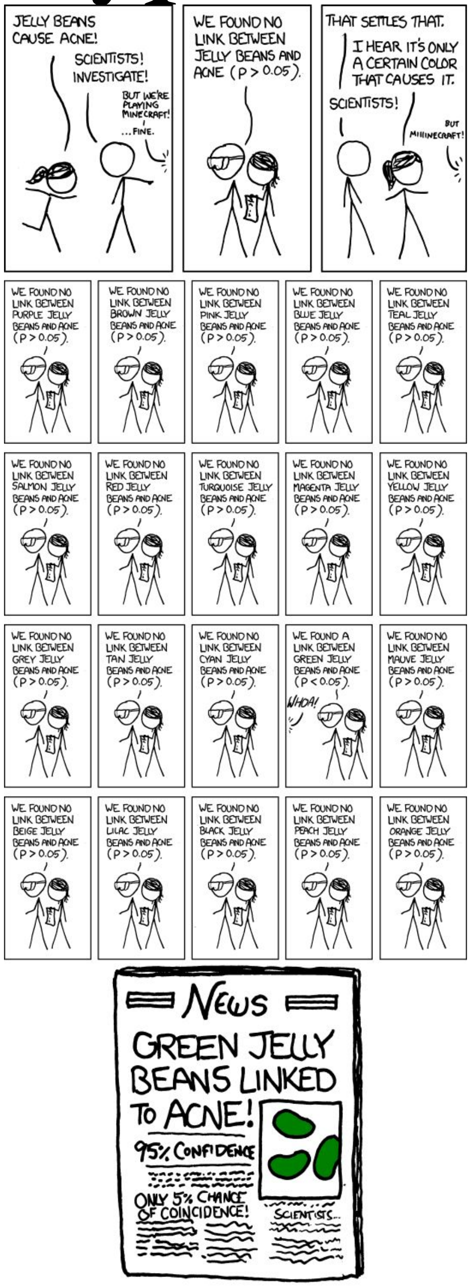

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Multiple hypothesis testing …


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

- The normal “significance threshold” of p-value = 0.05 causes false positives when we test many hypotheses

Multiple hypothesis testing …


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

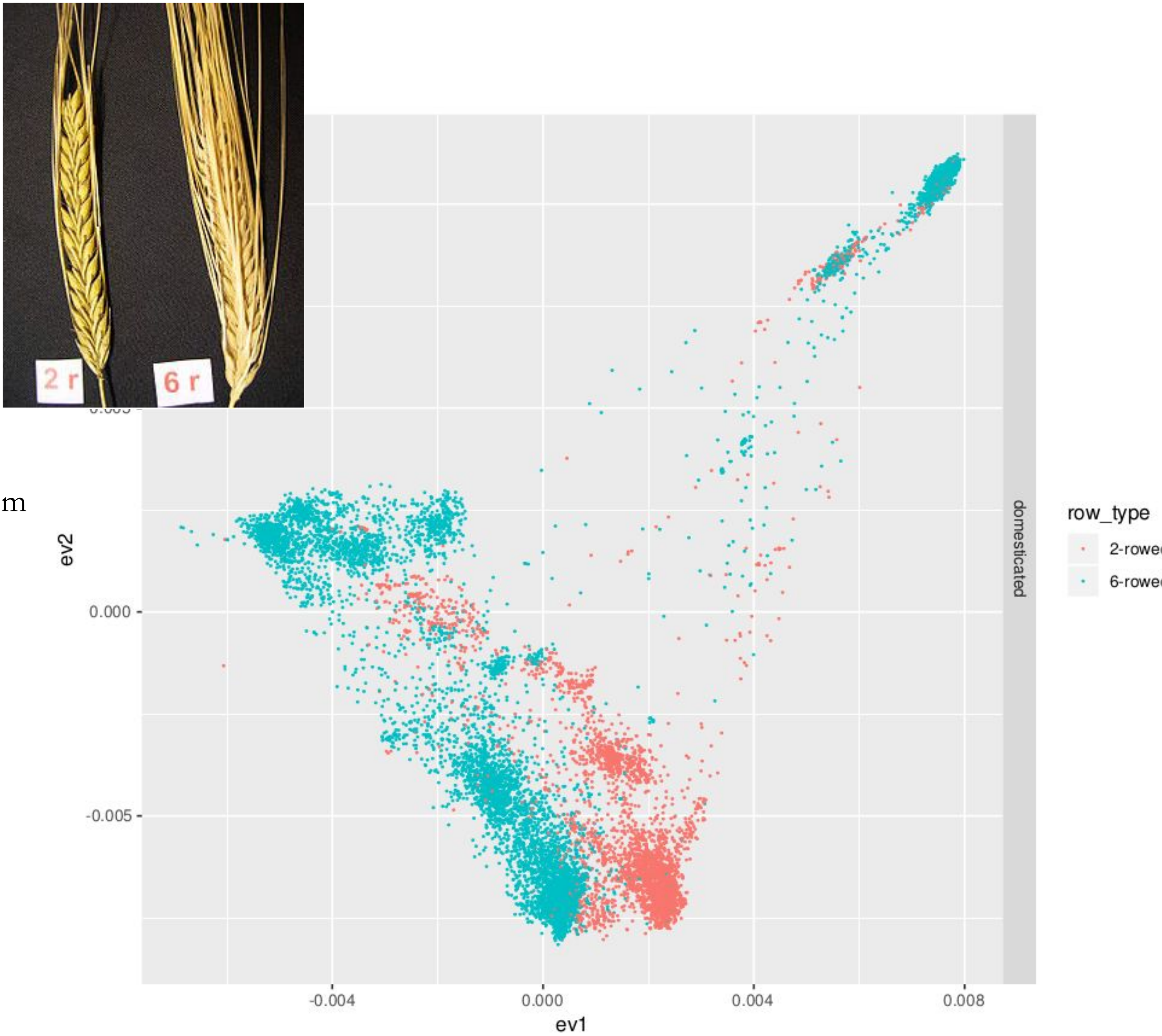
- The normal “significance threshold” of p-value < 0.05 causes false positives when we test many hypotheses.
- We correct for this by setting the threshold much more stringently.
- The “Bonferroni correction” involves simply dividing the p-value significance threshold (0.05) by the number of tests.
- The number of tests is the number of SNPs.

Association testing in this session

Test for differences in row type (6-rowed vs 2-rowed)

Association testing in this session

Test for differences in row type
(6-rowed vs 2-rowed)

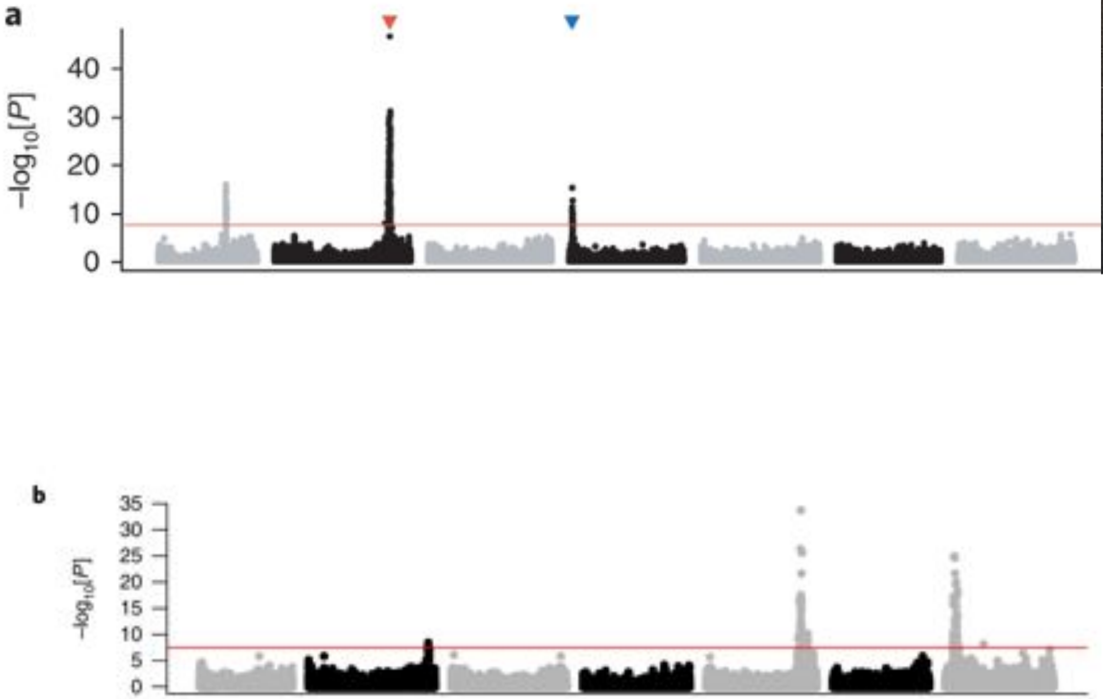


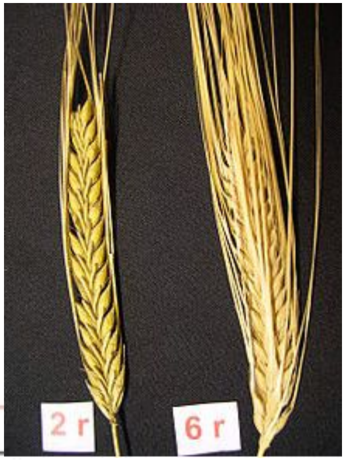
The collection our sample came from shows some relationship between row-type and population structure!

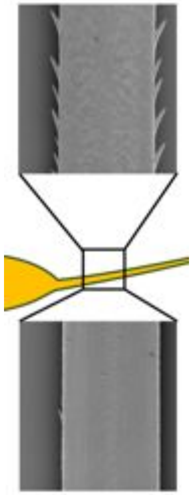
Association testing in this session

Previous GWAS results show genetic associations with awn barbs and row type

Fig. 3: Genome-wide association scans for morphological characters.





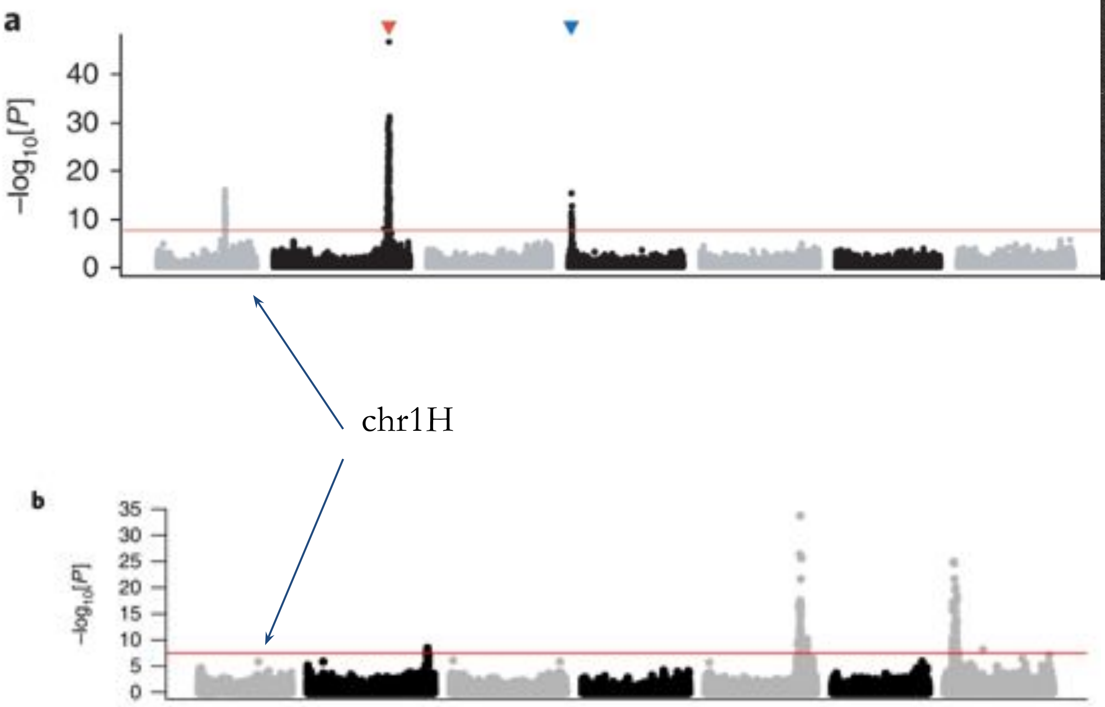


https://www.nature.com/articles/s41588-018-0266-x

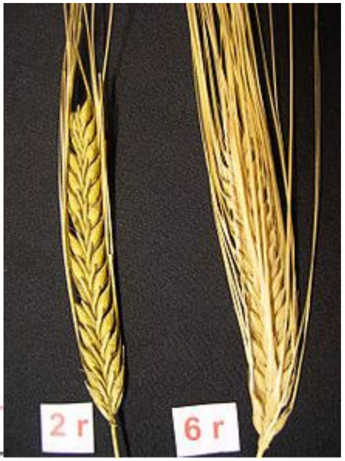
Association testing in this session

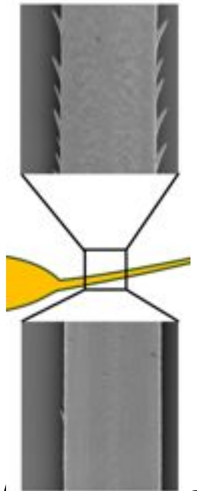
Previous GWAS results show genetic associations with row type (and also awn barbs, below)

Fig. 3: Genome-wide association scans for morphological characters.



chr1H



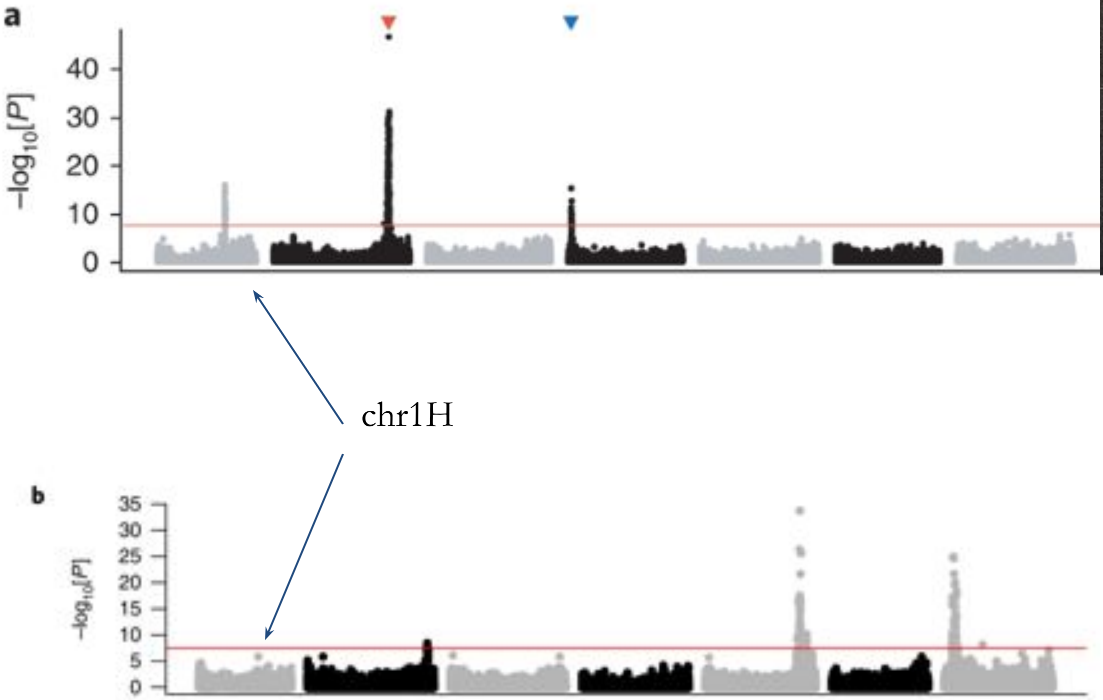


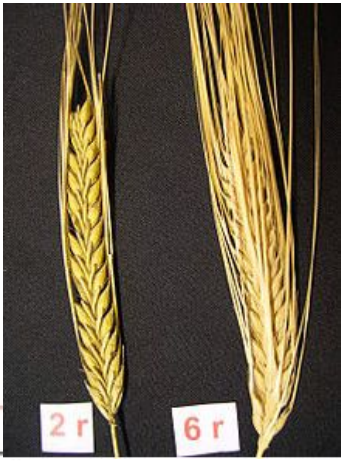
https://www.nature.com/articles/s41588-018-0266-x

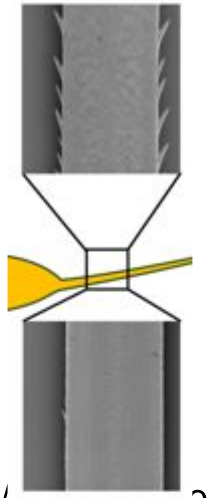
Association testing in this session

Previous GWAS results show genetic associations with row type (and also awn barbs, below)

Fig. 3: Genome-wide association scans for morphological characters.







We are unlikely to find any strong correlations owing to small sample size …

https://www.nature.com/articles/s41588-018-0266-x

Association testing in this session

So … primary aims:

- Load our cleaned SNP data into R
- Create a PCA to summarise the population structure between samples
  - Plot a PCA plot!
- Link it to phenotype data from a database
  - Using a new data.table trick ...
  - More PCA plots …
- Link three datasets: **The PCs** that summarise the population structure, t**he SNPs**, and **the phenotype data**
- Run a GWAS to test associations at each SNP
- Plot the GWAS results as a Manhattan Plot with the significance threshold shown on the plot.