24th International Master in Plant Genetics, Genomics and Breeding

Unit 7: Marker enabled prediction and selection

IBD, IBS, Genetic Distance, Population Structure

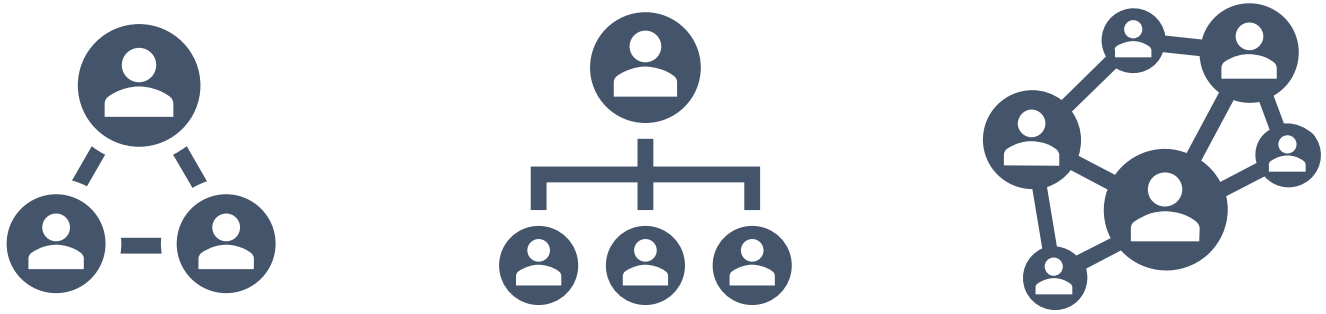
Feb 8 – 9, 2024

CJ Yang

Scotland's Rural College

cyang@sruc.ac.uk

cjyang-work.github.io



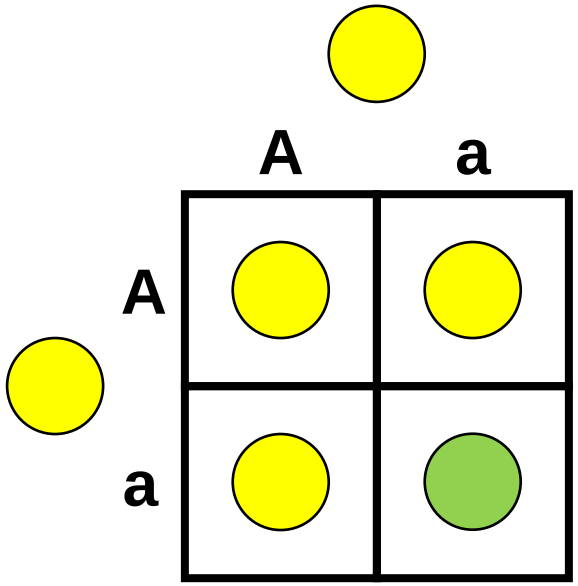
Lecture outline

1. Background & introduction
2. Identity-by-descent (IBD)
3. Identity-by-state (IBS)
4. Genetic distance
5. Population structure
6. Summary

- Concepts
- Examples
- Applications
- Challenges

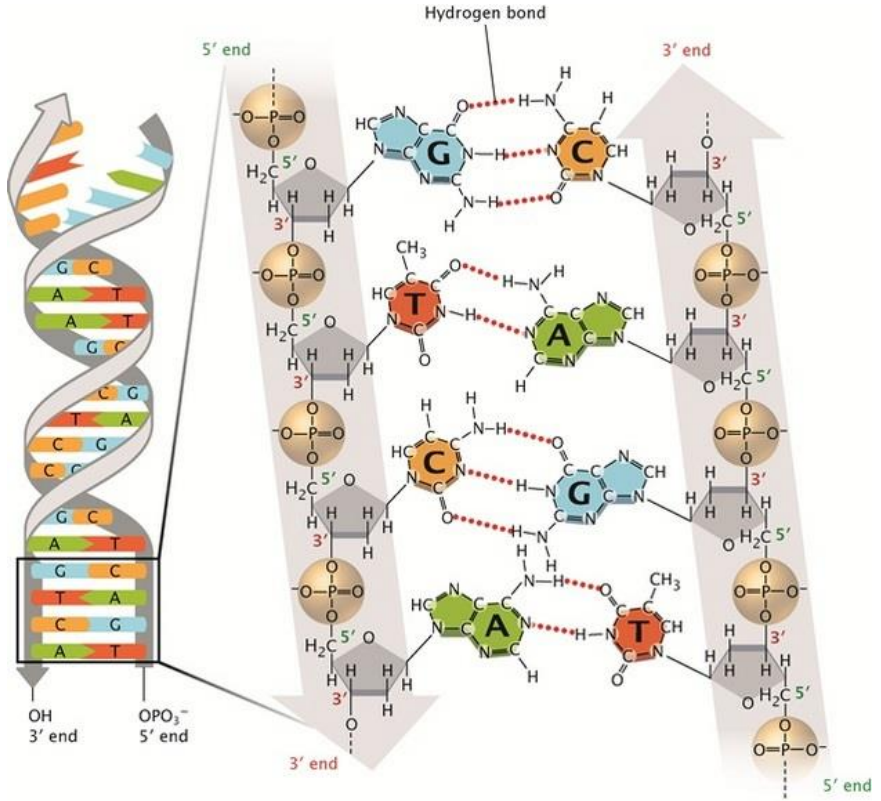
Background

Mendelian genetics



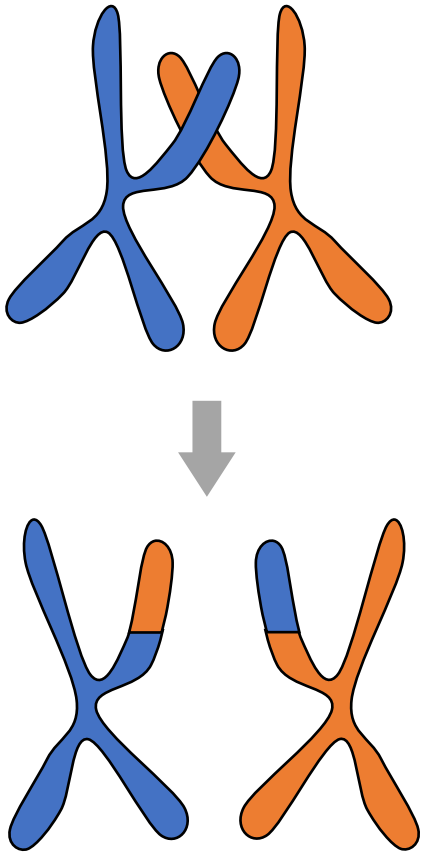
Dominant vs recessive alleles

DNA



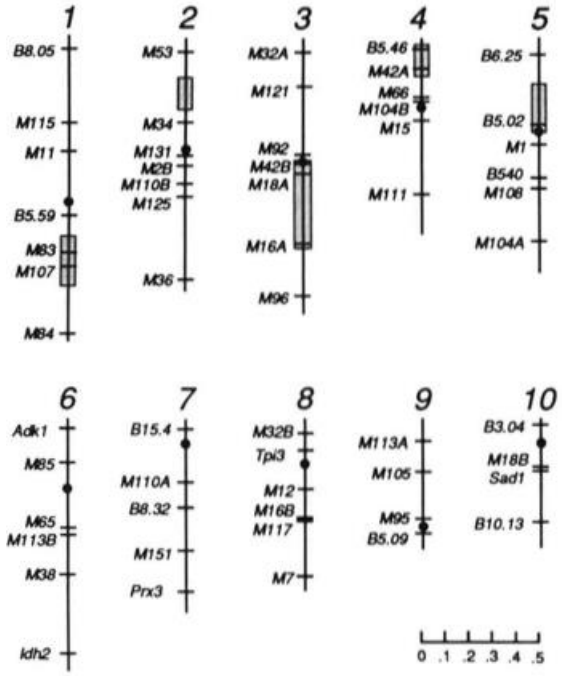
Pray L (2008)

Recombination



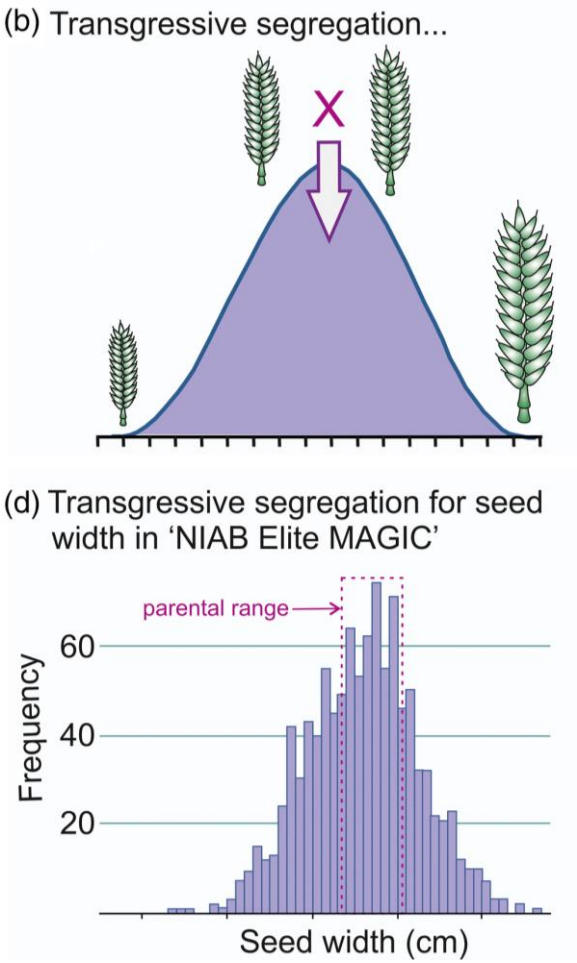
Background

Genetic linkage



Doebley and Stec (1991)

Quantitative traits



Mackay et al (2021)

Evolution

https://www.pinterest.com/pin/570620215266518401/

Motivation for understanding genetic relationships

Academic research perspective

- Understanding of selection, adaptation, ecology and evolution.
- Gene mapping.
- Quantitative genetics and complex traits.
- Method development.

Breeding perspective

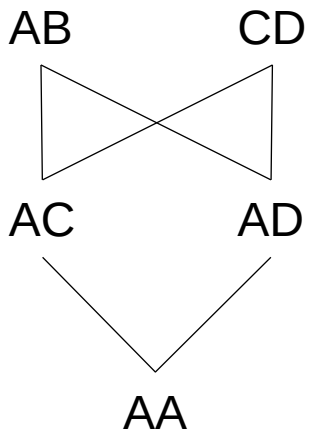
- Genetic diversity management, optimal parental selection.
- Target environment breeding.
- Marker-based mapping and selection.

Identity-by-descent (IBD)

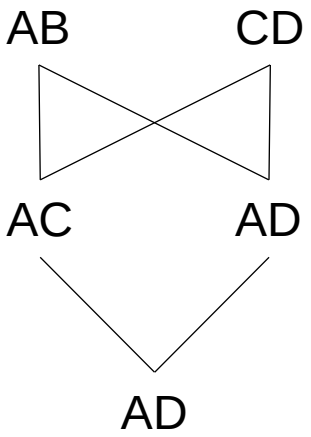
"Two alleles that have originated from the replication of one single allele in a previous generation may be called identical by descent"

Falconer and Mackay (1996) Introduction to Quantitative Genetics

Example with 4 alleles



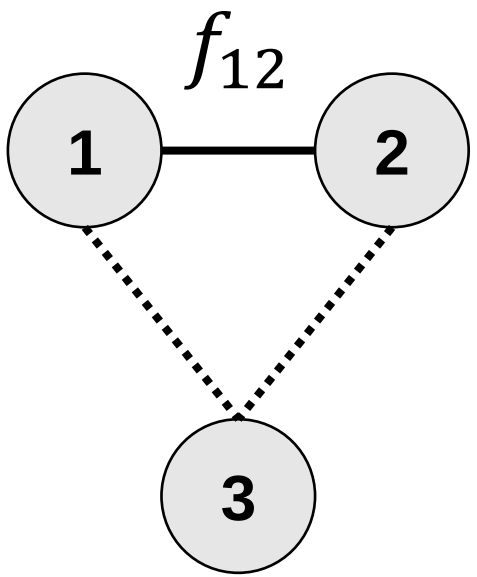
IBD



not IBD

Coefficient of coancestry/consanguinity/kinship (f)

f = Probability of two alleles are IBD.



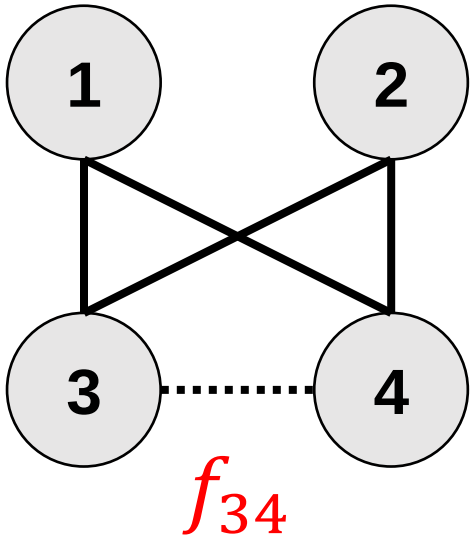
F_3

Coefficient of inbreeding

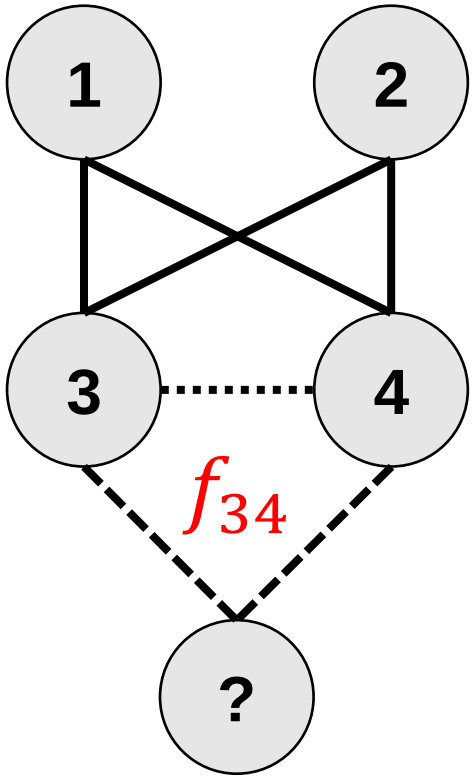
$$0 \leq f_{12} \leq 1$$

If 1 and 2 do not share a recent common ancestor, $f_{12} \to 0$

Example 1

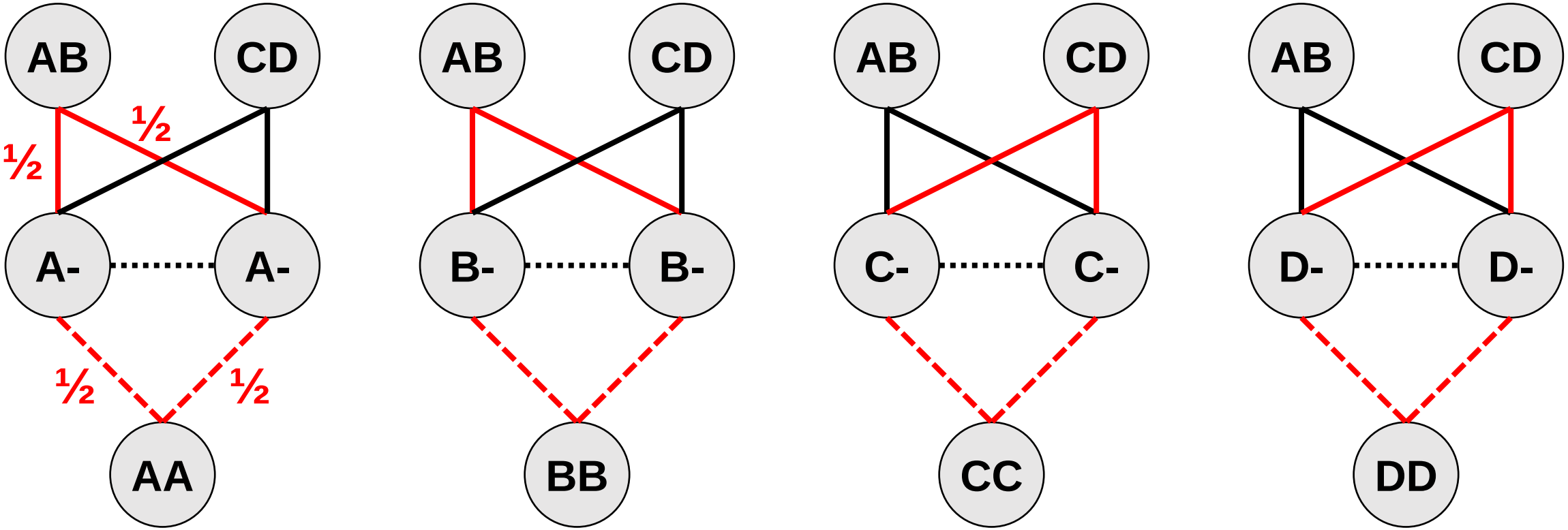


Example 1



Hypothetical progeny

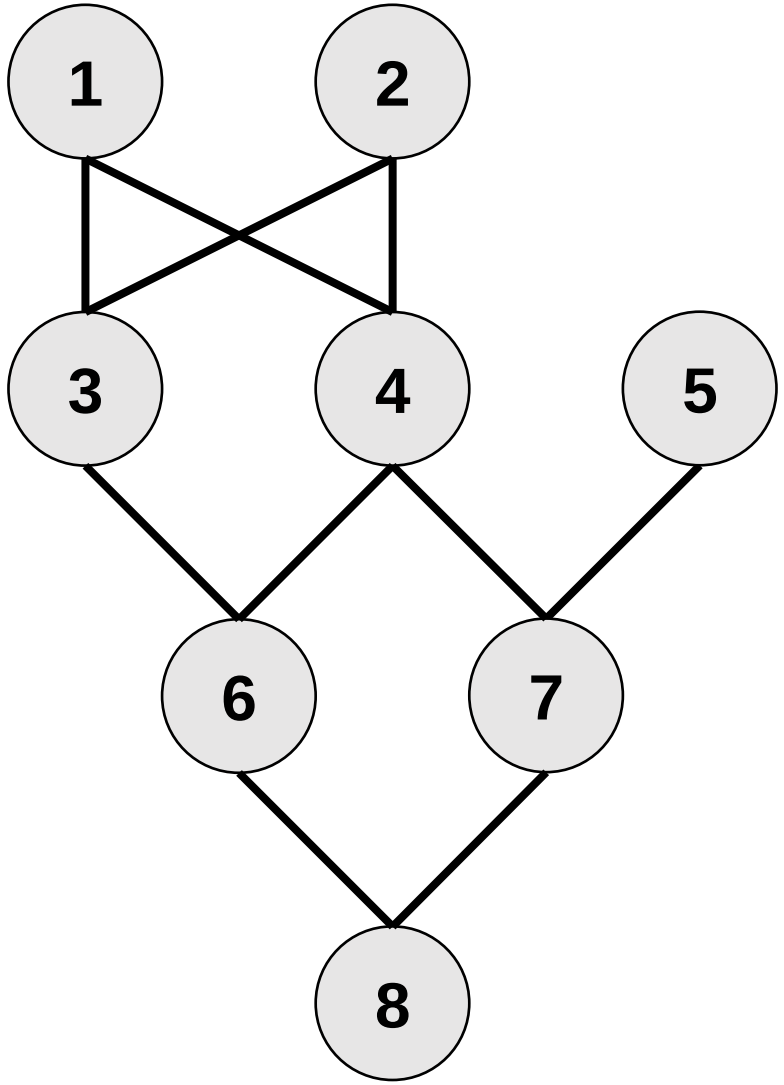
Example 1



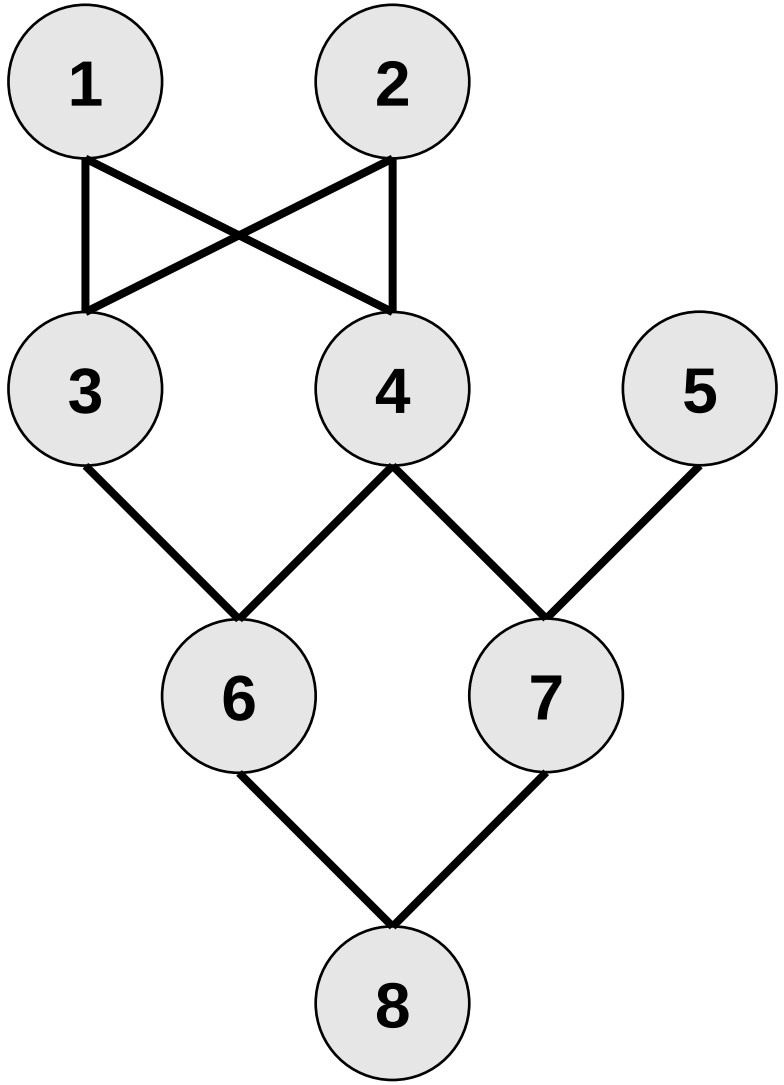
$$f_{34} = \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^4 = \frac{1}{4}$$

Dirty method

Example 2



Example 2

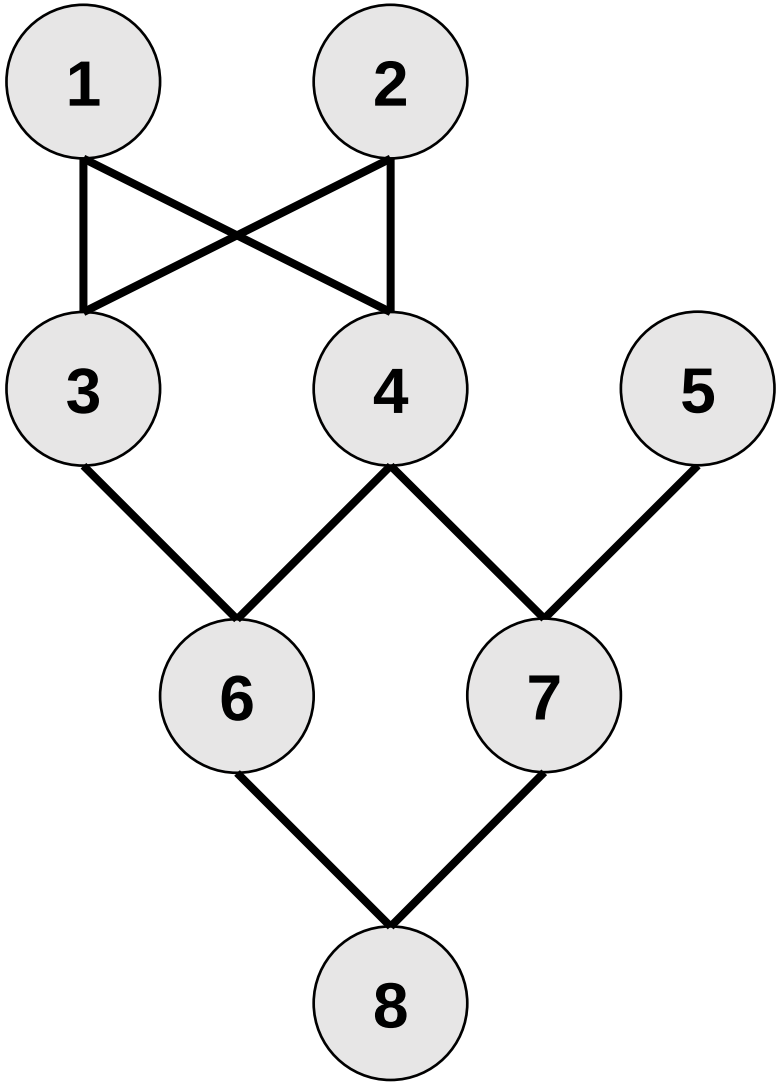


1. Create a table and arrange it from oldest to youngest generations.

i	j	f_{ij}
1	1	
2	1	
2	2	
3	1	
3	2	
3	3	
4	1	
4	2	
4	3	
4	4	
5	1	
5	2	
5	3	
5	4	
5	5	
6	1	
6	2	
6	3	

i	j	f_{ij}
6	4	
6	5	
6	6	
7	1	
7	2	
7	3	
7	4	
7	5	
7	6	
7	7	
8	1	
8	2	
8	3	
8	4	
8	5	
8	6	
8	7	
8	8	

Example 2



2. Fill in the known and easy ones.

Without additional information, we assume $F_1 = F_2 = F_5 = 0$ and $f_{21} = f_{51} \cdots f_{54} = f_{65} = 0$.

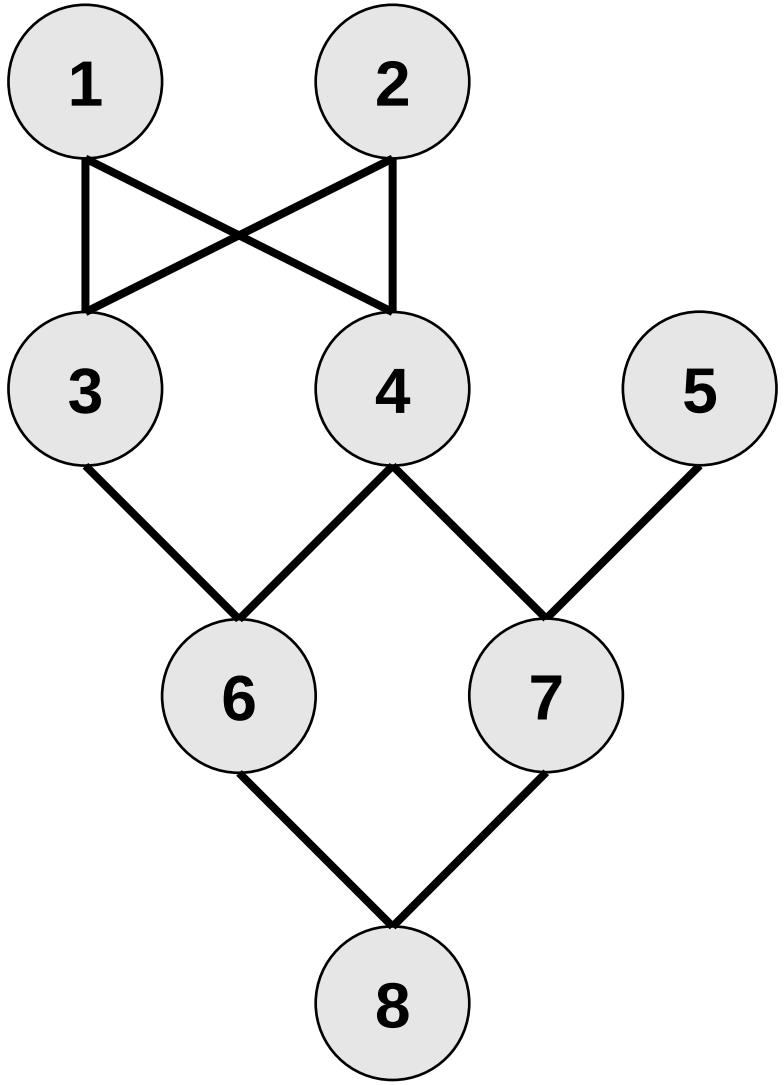
$$f_{ii} = \frac{1}{2}(1 + F_i)$$

$$f_{ij} = f_{ji}$$

i	j	f_{ij}
1	1	0.5
2	1	0
2	2	0.5
3	1	
3	2	
3	3	
4	1	
4	2	
4	3	
4	4	
5	1	0
5	2	0
5	3	0
5	4	0
5	5	0.5
6	1	
6	2	
6	3	

i	j	f_{ij}
6	4	
6	5	0
6	6	
7	1	
7	2	
7	3	
7	4	
7	5	
7	6	
7	7	
8	1	
8	2	
8	3	
8	4	
8	5	
8	6	
8	7	
8	8	

Example 2



3. Fill in the rest.

f_{ij} is the average of f's between j and parents of i.

$$f_{ij} = \frac{1}{2}(f_{pj} + f_{qj})$$

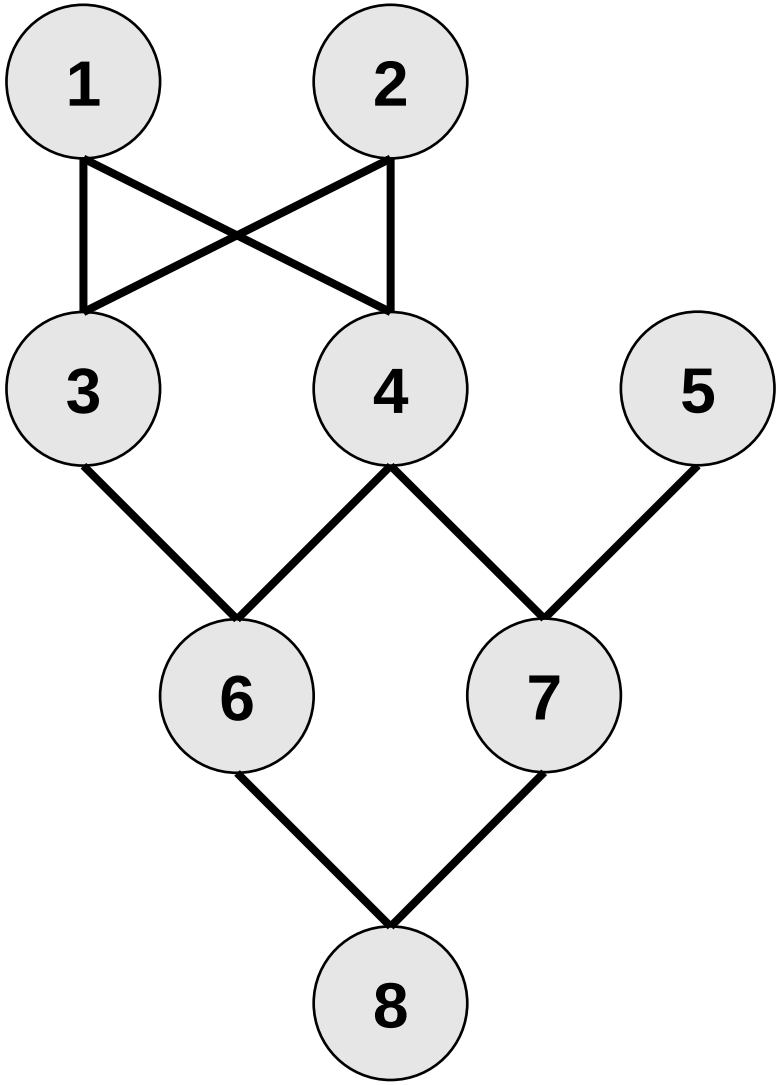
Note: p and q are the parents of i.

e.g. $f_{31} = \frac{f_{11}+f_{21}}{2}$.

i	j	f_{ij}
1	1	0.5
2	1	0
2	2	0.5
3	1	0.25
3	2	0.25
3	3	
4	1	0.25
4	2	0.25
4	3	
4	4	
5	1	0
5	2	0
5	3	0
5	4	0
5	5	0.5
6	1	
6	2	
6	3	

i	j	f_{ij}
6	4	
6	5	0
6	6	
7	1	
7	2	
7	3	
7	4	
7	5	
7	6	
7	7	
8	1	
8	2	
8	3	
8	4	
8	5	
8	6	
8	7	
8	8	

Example 2



If i and j are in the same generation, we can also calculate f_{ij} as the average of f's between parents of i and j.

$$f_{ij} = \frac{1}{4}(f_{pr} + f_{ps} + f_{qr} + f_{qs})$$

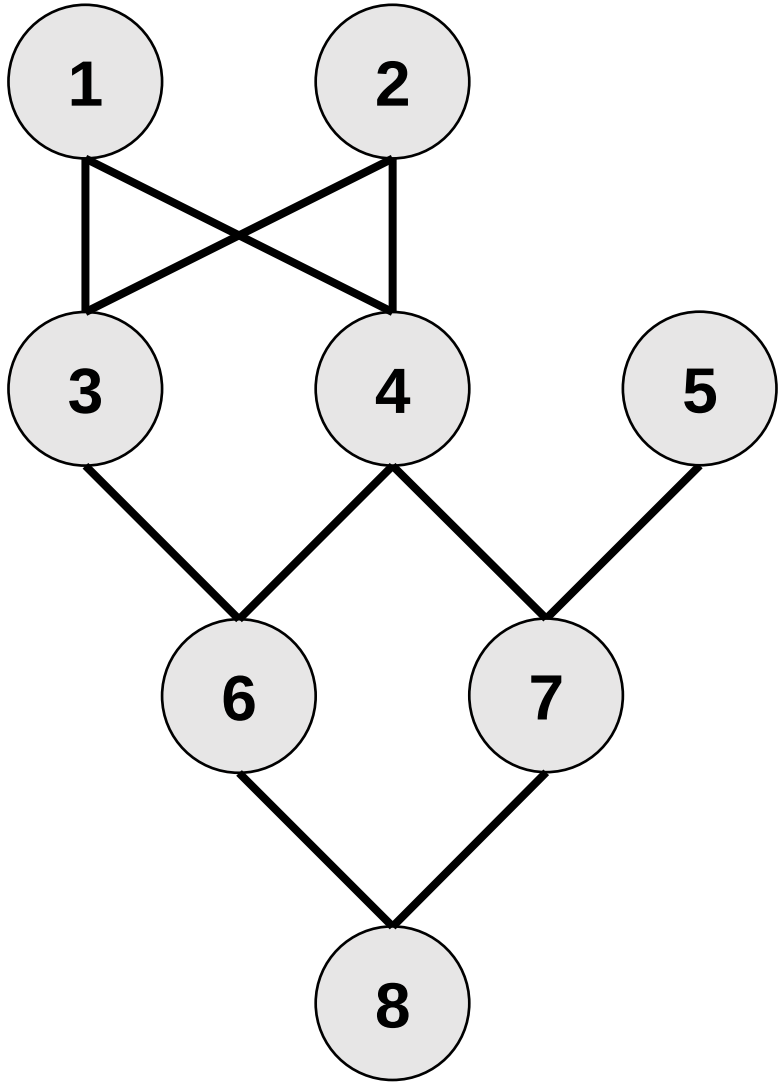
Note: r and s are the parents of j.

e.g. $f_{43} = \frac{f_{11}+f_{12}+f_{21}+f_{22}}{4}$.

i	j	f_{ij}
1	1	0.5
2	1	0
2	2	0.5
3	1	0.25
3	2	0.25
3	3	
4	1	0.25
4	2	0.25
4	3	0.25
4	4	
5	1	0
5	2	0
5	3	0
5	4	0
5	5	0.5
6	1	
6	2	
6	3	

i	j	f_{ij}
6	4	
6	5	0
6	6	
7	1	
7	2	
7	3	
7	4	
7	5	
7	6	
7	7	
8	1	
8	2	
8	3	
8	4	
8	5	
8	6	
8	7	
8	8	

Example 2



Recall these.

$$f_{ii} = \frac{1}{2}(1 + F_i)$$

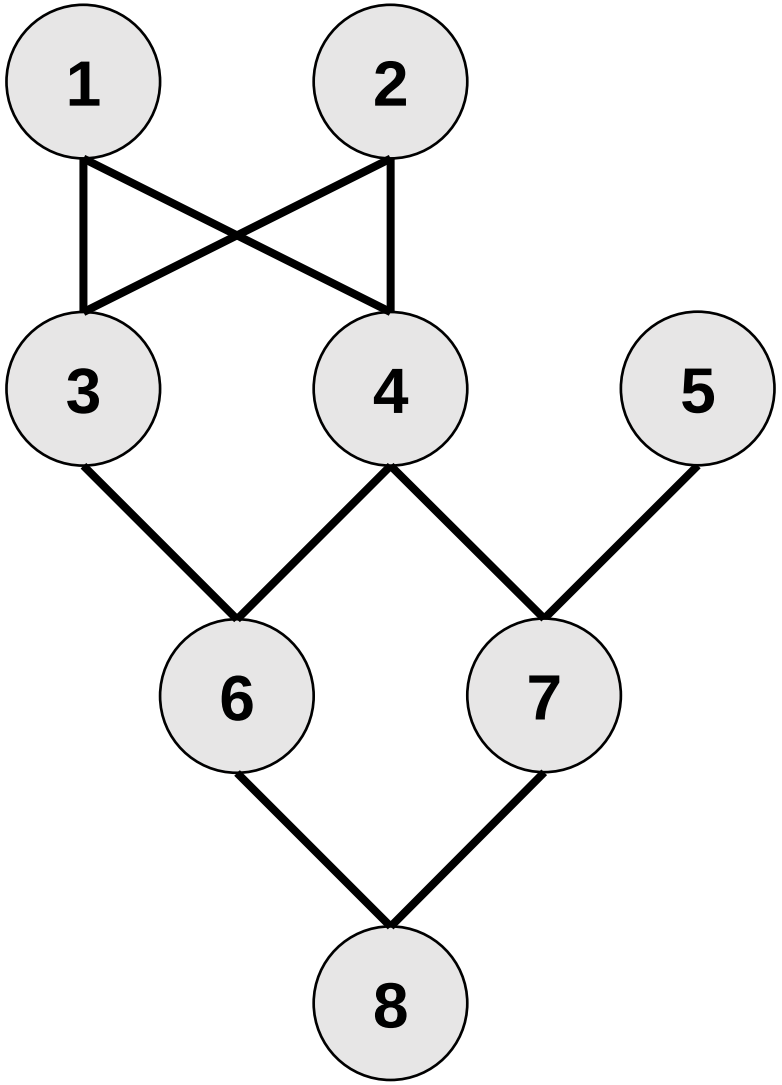
$$F_i = f_{pq}$$

e.g. $f_{33} = \frac{1}{2}(1 + f_{12})$.

i	j	f_{ij}
1	1	0.5
2	1	0
2	2	0.5
3	1	0.25
3	2	0.25
3	3	0.5
4	1	0.25
4	2	0.25
4	3	0.25
4	4	0.5
5	1	0
5	2	0
5	3	0
5	4	0
5	5	0.5
6	1	
6	2	
6	3	

i	j	f_{ij}
6	4	
6	5	0
6	6	
7	1	
7	2	
7	3	
7	4	
7	5	
7	6	
7	7	
8	1	
8	2	
8	3	
8	4	
8	5	
8	6	
8	7	
8	8	

Example 2



Use these to complete the rest.

$$f_{ii} = \frac{1}{2}(1 + F_i)$$

$$F_i = f_{pq}$$

$$f_{ij} = \frac{1}{2}(f_{pj} + f_{qj})$$

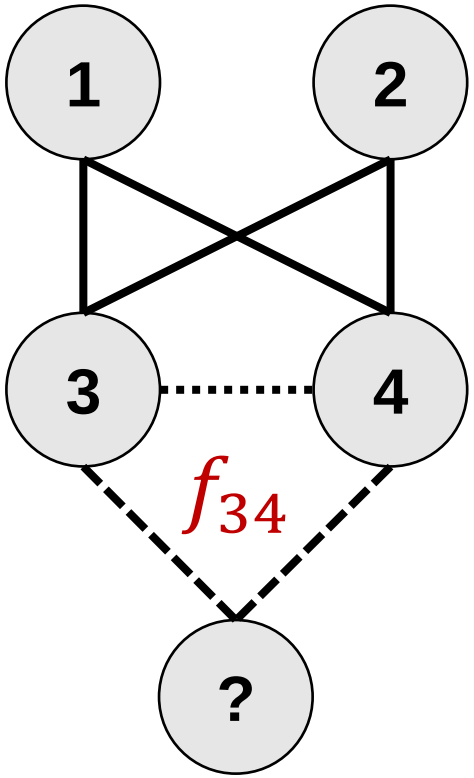
e.g. $f_{76} = \frac{1}{2}(f_{46} + f_{56})$.

i	j	f_{ij}
1	1	0.5
2	1	0
2	2	0.5
3	1	0.25
3	2	0.25
3	3	0.5
4	1	0.25
4	2	0.25
4	3	0.25
4	4	0.5
5	1	0
5	2	0
5	3	0
5	4	0
5	5	0.5
6	1	0.25
6	2	0.25
6	3	0.375

i	j	f_{ij}
6	4	0.375
6	5	0
6	6	0.625
7	1	0.125
7	2	0.125
7	3	0.125
7	4	0.25
7	5	0.25
7	6	0.1875
7	7	0.5
8	1	0.1875
8	2	0.1875
8	3	0.25
8	4	0.3125
8	5	0.125
8	6	0.40625
8	7	0.34375
8	8	0.59375

Example 1.1

What if 1 and 2 have inbreeding history?



$$f_{34} = \frac{1}{4}(f_{11} + f_{12} + f_{21} + f_{22})$$

$$f_{11} = \frac{1}{2}(1 + F_1)$$

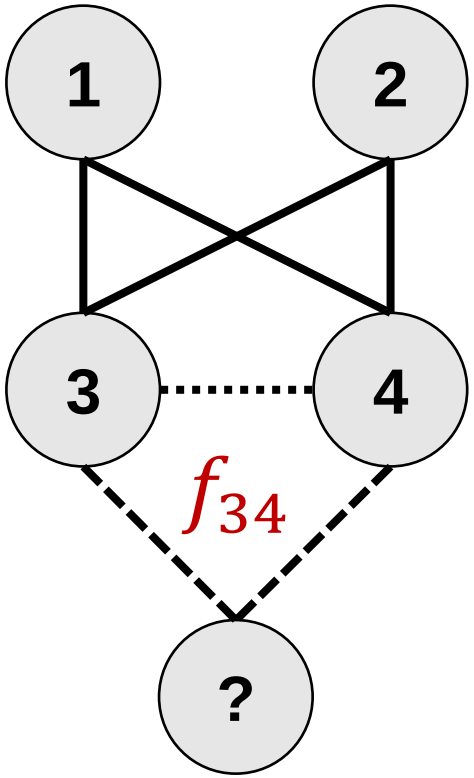
$$f_{12} = f_{21} = 0$$

$$f_{22} = \frac{1}{2}(1 + F_2)$$

$$f_{34} = \frac{1}{4}\left(\frac{1}{2}(1 + F_1) + f_{12} + f_{12} + \frac{1}{2}(1 + F_2)\right) = \frac{1}{4} + \frac{1}{8}F_1 + \frac{1}{8}F_2$$

Example 1.2

What if 1 and 2 have inbreeding history and they are also related?



$$f_{34} = \frac{1}{4}(f_{11} + f_{12} + f_{21} + f_{22})$$

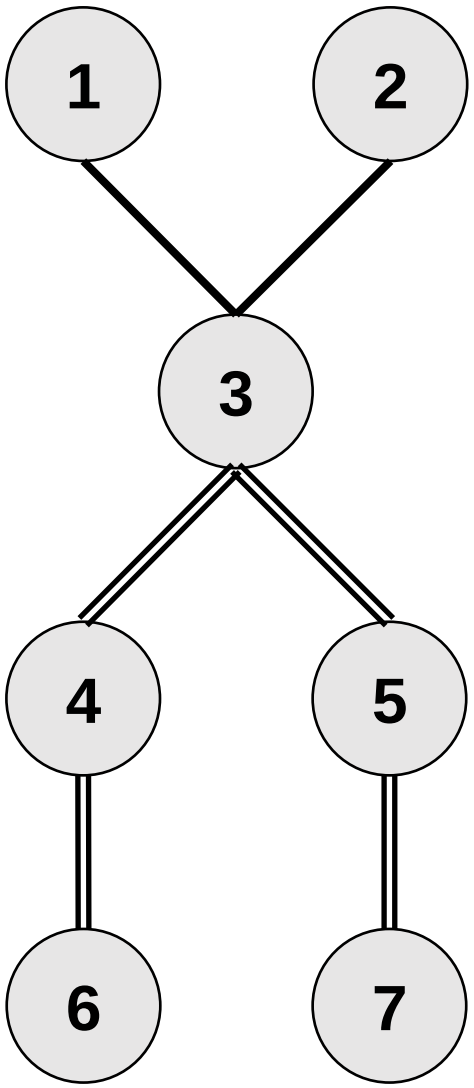
$$f_{11} = \frac{1}{2}(1 + F_1)$$

$$f_{12} = f_{21} \neq 0$$

$$f_{22} = \frac{1}{2}(1 + F_2)$$

$$f_{34} = \frac{1}{4}\left(\frac{1}{2}(1 + F_1) + f_{12} + f_{12} + \frac{1}{2}(1 + F_2)\right) = \frac{1}{4} + \frac{1}{8}F_1 + \frac{1}{8}F_2 + \frac{1}{2}f_{12}$$

Example 3



Bi-parental crosses, $f_{12} \neq 0$

$$f_{33} = \frac{1}{2}(1 + F_3) = \frac{1}{2}(1 + f_{12})$$

$$f_{34} = \frac{1}{2}(f_{33} + f_{33}) = f_{33}$$

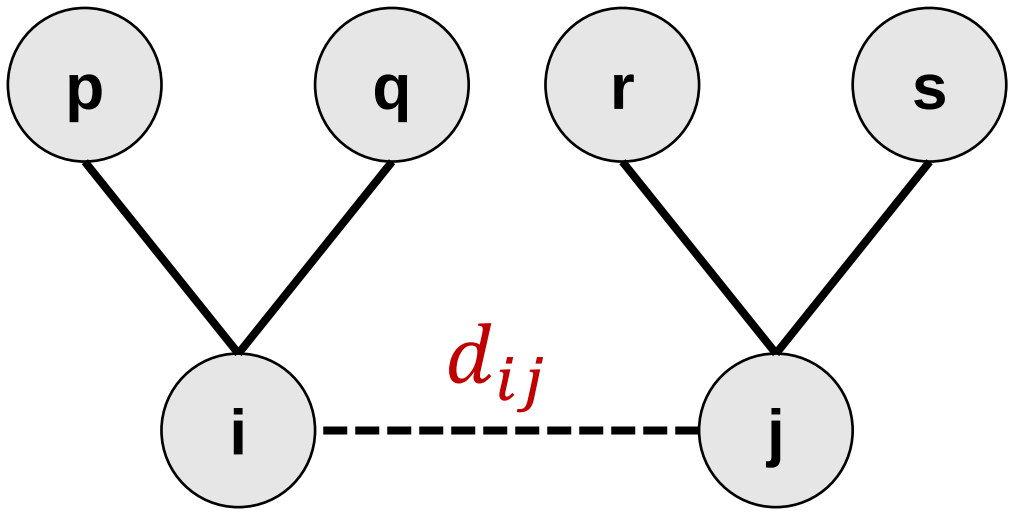
$$f_{45} = \frac{1}{2}(f_{34} + f_{34}) = f_{34}$$

$$f_{56} = \frac{1}{2}(f_{45} + f_{45}) = f_{45}$$

$$f_{67} = \frac{1}{2}(f_{56} + f_{56}) = f_{56} = \frac{1}{2}(1 + f_{12})$$

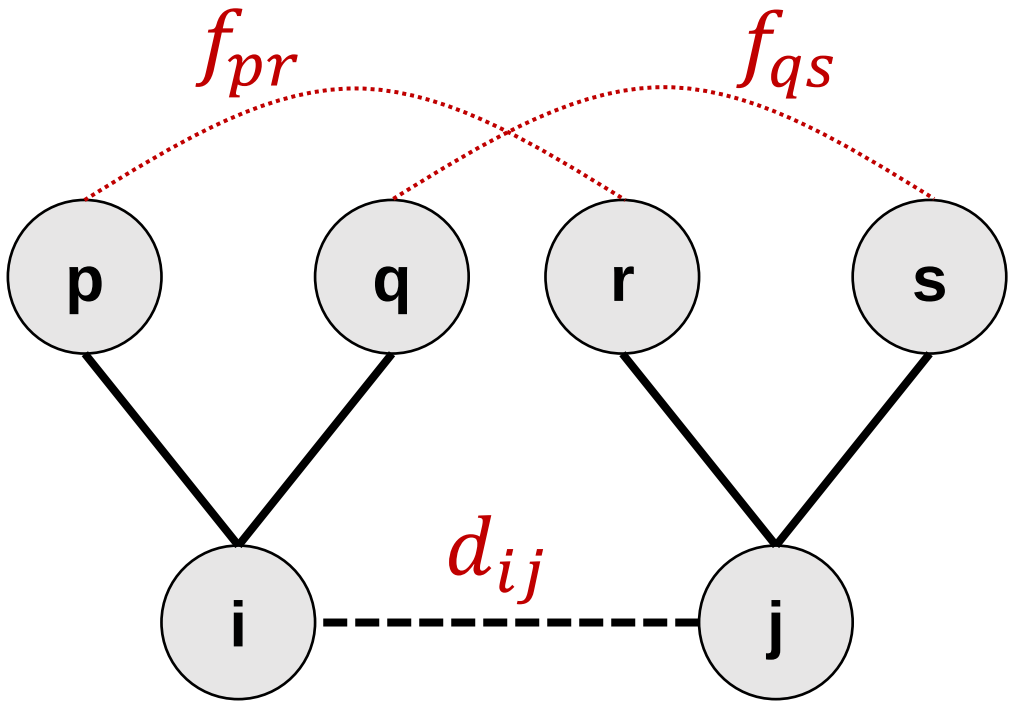
Coefficient of fraternity (d)

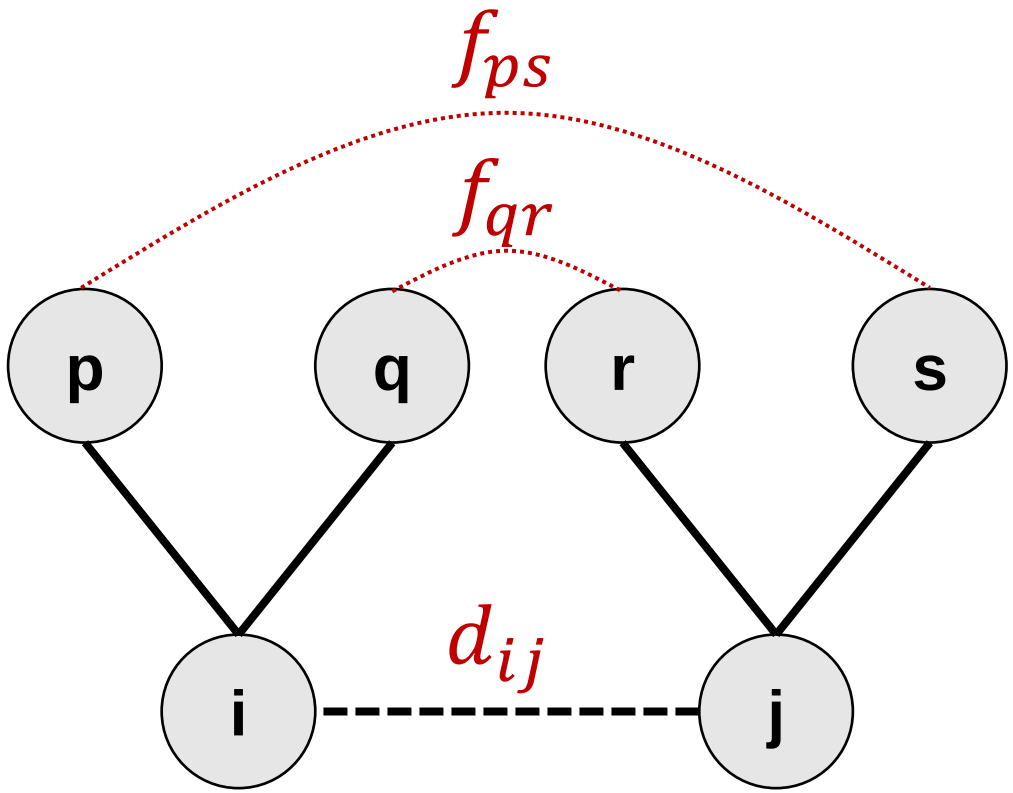
d = Probability of two genotypes are IBD.



Coefficient of fraternity (d)

d = Probability of two IBD genotypes.

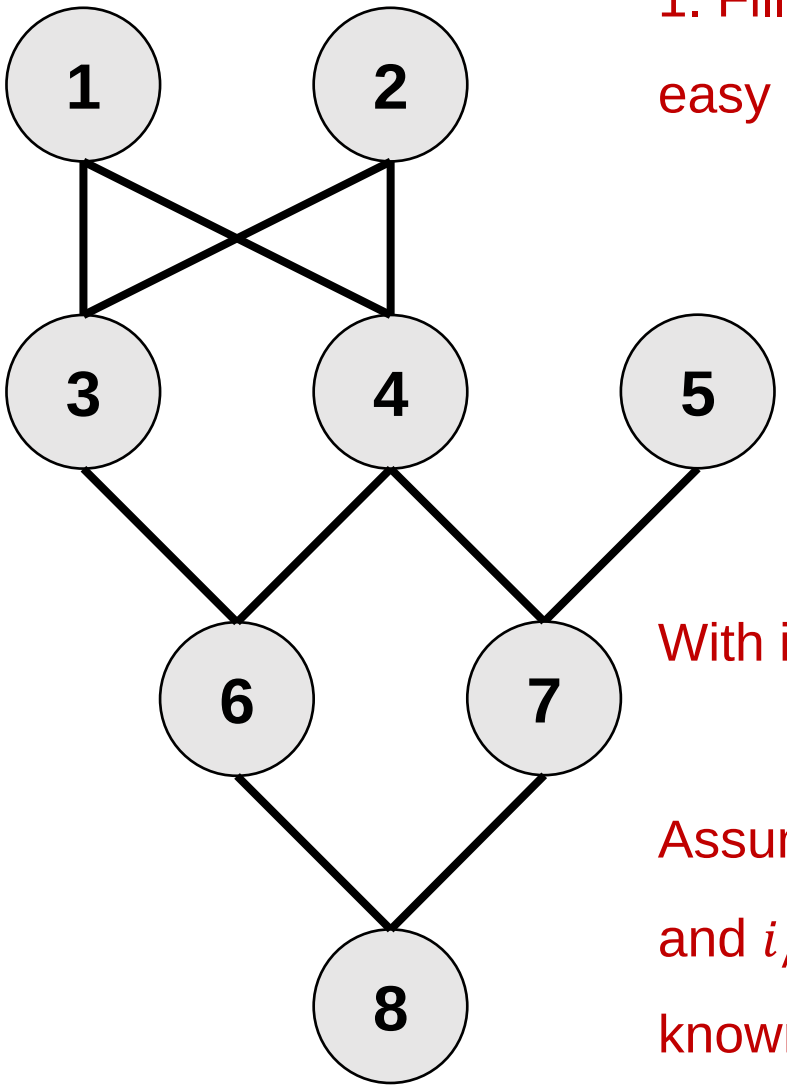




Only possible if the parents are related.

$$d_{ij} = f_{pr}f_{qs} + f_{ps}f_{qr}$$

Example 2



1. Fill in the known and easy ones.

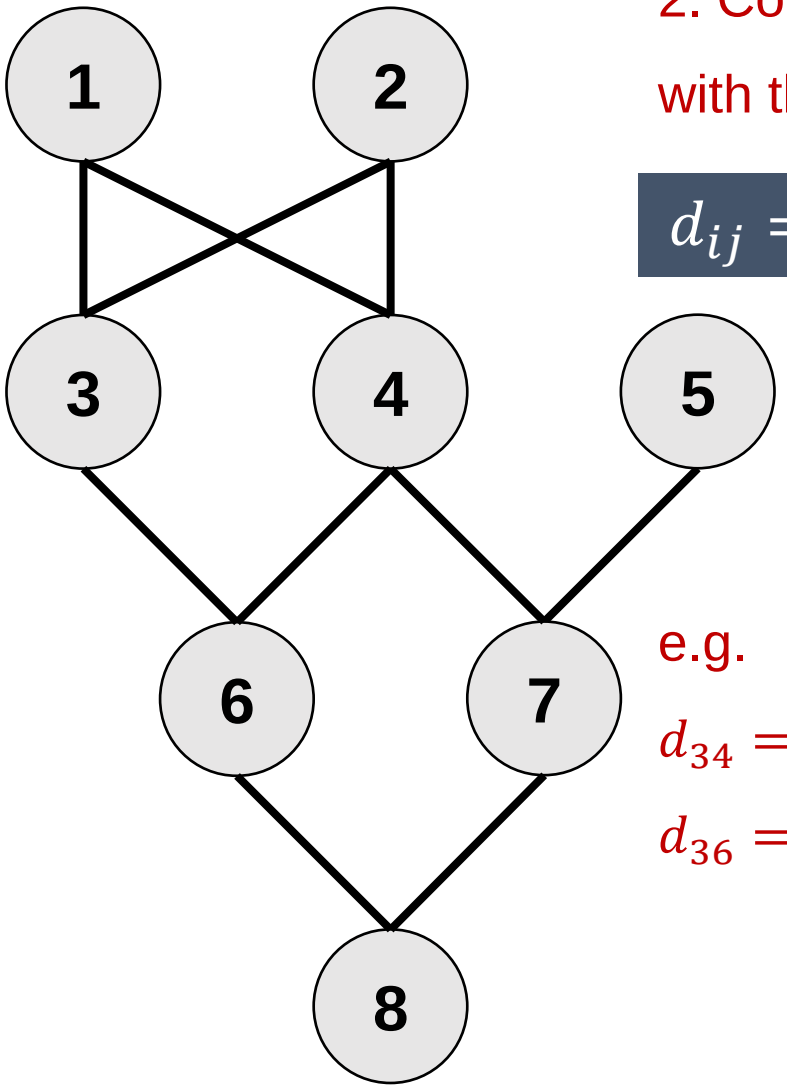
With itself, $d_{ii} = 1$.

Assume $d_{ij} = 0$ if $i \neq j$ and i/j does not have known parents.

i	j	f_{ij}	d_{ij}
1	1	0.5	1
2	1	0	0
2	2	0.5	1
3	1	0.25	0
3	2	0.25	0
3	3	0.5	1
4	1	0.25	0
4	2	0.25	0
4	3	0.25	
4	4	0.5	1
5	1	0	0
5	2	0	0
5	3	0	0
5	4	0	0
5	5	0.5	1
6	1	0.25	0
6	2	0.25	0
6	3	0.375	

i	j	f_{ij}	d_{ij}
6	4	0.375	
6	5	0	0
6	6	0.625	1
7	1	0.125	0
7	2	0.125	0
7	3	0.125	
7	4	0.25	
7	5	0.25	0
7	6	0.1875	
7	7	0.5	1
8	1	0.1875	0
8	2	0.1875	0
8	3	0.25	
8	4	0.3125	
8	5	0.125	0
8	6	0.40625	
8	7	0.34375	
8	8	0.59375	1

Example 2



2. Compute the rest with this formula.

$$d_{ij} = f_{pr}f_{qs} + f_{ps}f_{qr}$$

e.g.

$$d_{34} = f_{11}f_{22} + f_{12}f_{21}$$

$$d_{36} = f_{13}f_{24} + f_{14}f_{23}$$

i	j	f_{ij}	d_{ij}
1	1	0.5	1
2	1	0	0
2	2	0.5	1
3	1	0.25	0
3	2	0.25	0
3	3	0.5	1
4	1	0.25	0
4	2	0.25	0
4	3	0.25	0.25
4	4	0.5	1
5	1	0	0
5	2	0	0
5	3	0	0
5	4	0	0
5	5	0.5	1
6	1	0.25	0
6	2	0.25	0
6	3	0.375	0.125

i	j	f_{ij}	d_{ij}
6	4	0.375	0.125
6	5	0	0
6	6	0.625	1
7	1	0.125	0
7	2	0.125	0
7	3	0.125	0
7	4	0.25	0
7	5	0.25	0
7	6	0.1875	0
7	7	0.5	1
8	1	0.1875	0
8	2	0.1875	0
8	3	0.25	0.0625
8	4	0.3125	0.0625
8	5	0.125	0
8	6	0.40625	0.140625
8	7	0.34375	0.09375
8	8	0.59375	1

Some known values for f and d

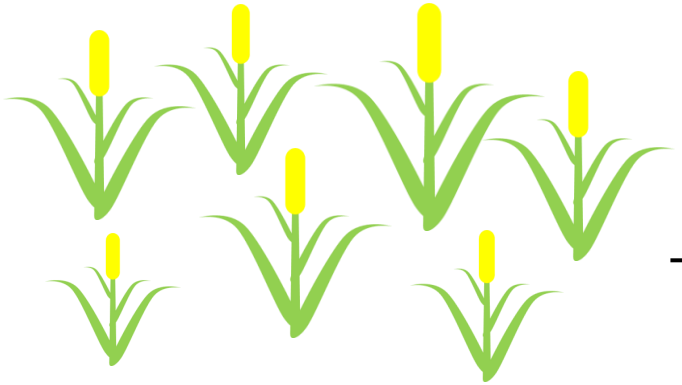
Relationship	f	d
Parent - progeny	1/4	0
Half siblings	1/8	0
Full siblings	1/4	1/4
First cousins	1/16	0
Monozygotic twins	1/2	1

These values assume no inbreeding history.

Applications of IBD

1. QTL linkage mapping
2. Marker genotype imputation and phasing
3. Genetic relationship matrix (GRM) calculation
4. Breeding and evolutionary history understanding
5. Population demographic inference
6. Genealogy and ancestry identification
7. Forensics

IBD in QTL linkage mapping



Genotype a bi-parental population →

```
ID1  A  G  C  A
ID2  A  C  C  T
ID3  G  C  C  T
```

Compute P(IBD) using hidden Markov model →

P(IBD) with founder 1

```
ID1  1  1  .5  1
ID2  1  1  .5  0
ID3  0  0  .5  0
```

What is a hidden Markov model (HMM)?

Dall-E (2024)

As you leave your home every morning, you look up to the sky and wonder if it will be rainy or sunny later in the day.

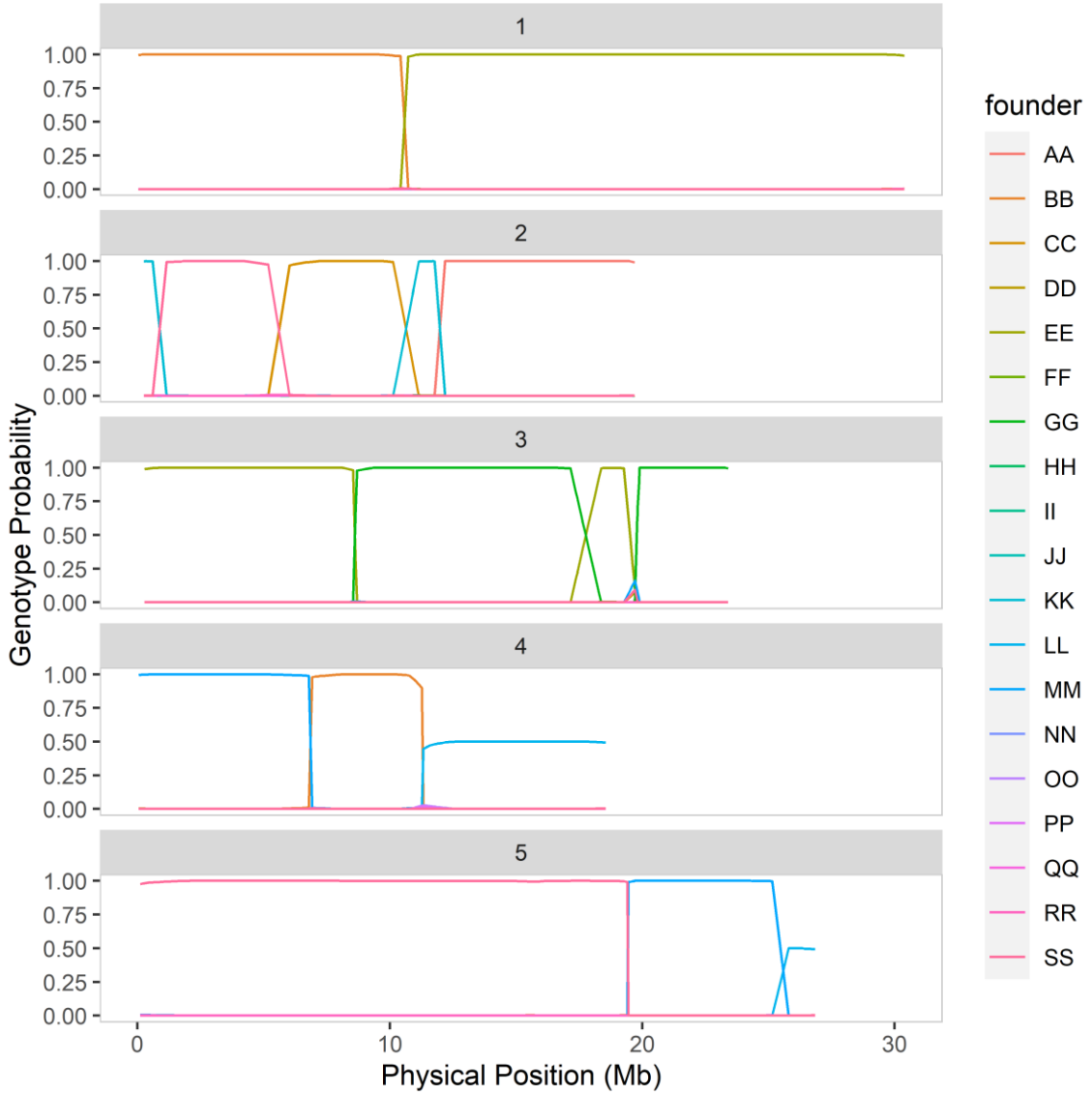
Assume that the weather forms a Markov chain, you pause at your front door for a second. You check the cloud cover, feel the breeze, look around and try to remember what happened in the last few days. Then you decide if you should pack an umbrella.

Cloud, breeze = marker genotype
Weather = founder IBD

IBD in QTL linkage mapping (multiparental population)

Similarly, P(IBD) can also be calculated in multi-parental populations.

Here is an example showing P(IBD) in a RIL from the Arabidopsis 19-founder MAGIC population (Kover et al 2009).

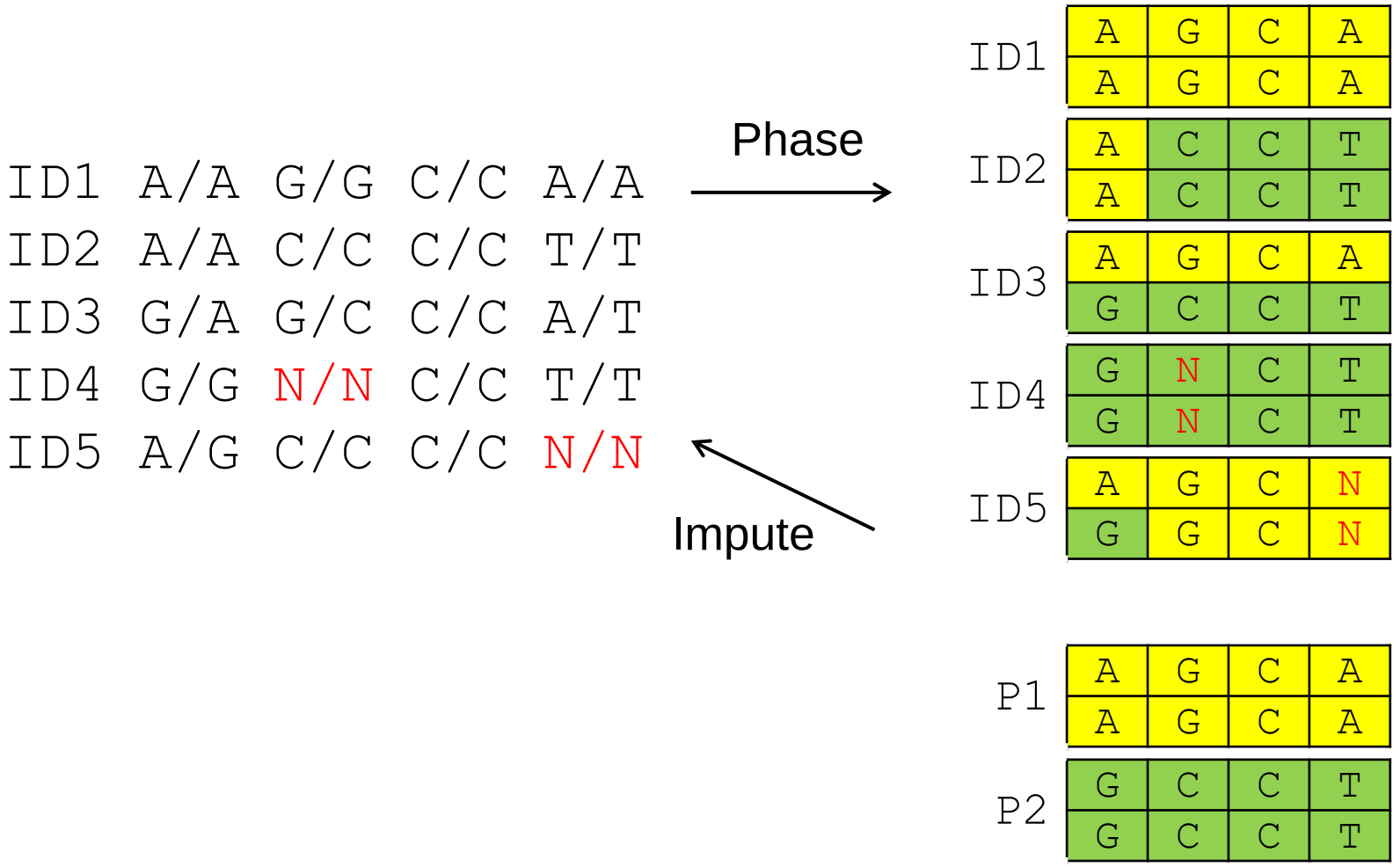


IBD in marker genotype imputation and phasing

This process is similar to the previous application in QTL linkage mapping.

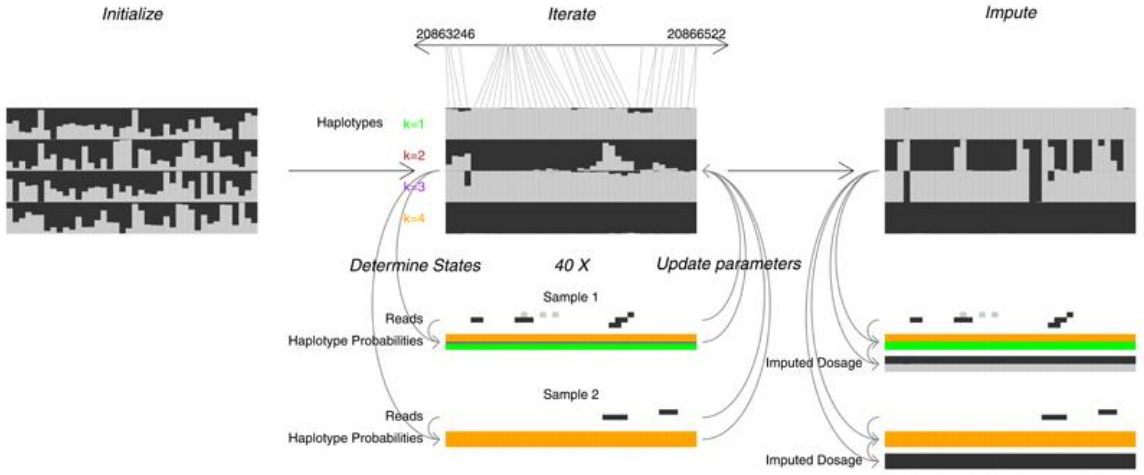
It starts with using HMM to identify (phase) the parent origin of each allele.

Then, it uses the phase information to fill in (impute) the missing data.

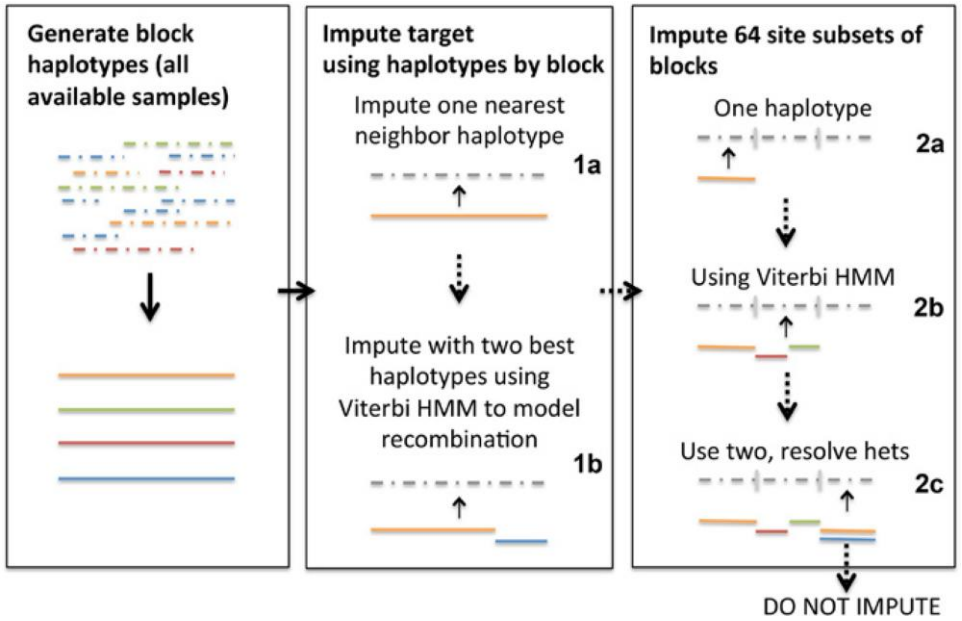


IBD in marker genotype imputation and phasing – common software

STITCH (Davies et al 2016)



FILLIN/FSFHap (Swarts et al 2014)



AlphaImpute (Hickey et al 2012)

Beagle (Browning et al 2021)

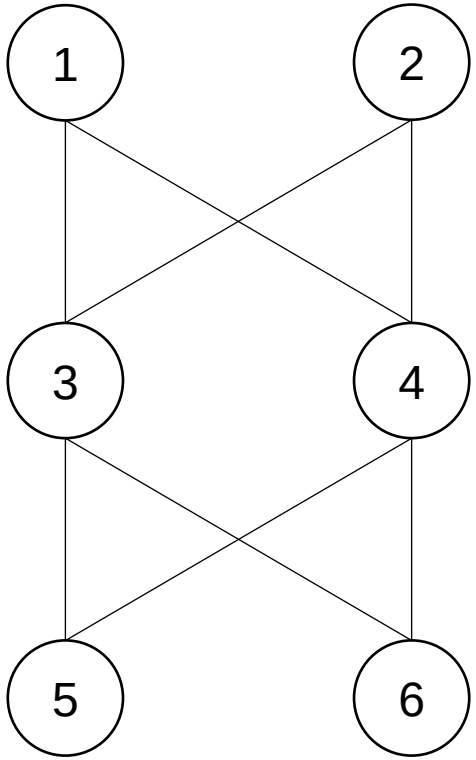
Many options, but they are fundamentally similar – identify founder haplotype,

fill in missing data from others that shared the same haplotype.

IBD in genetic relationship matrix (GRM) calculation

Pedigree-based additive genetic relationship matrix (A)

	1	2	3	4	5	6
1	1.00	0.00	0.50	0.50	0.50	0.50
2	0.00	1.00	0.50	0.50	0.50	0.50
3	0.50	0.50	1.00	0.50	0.75	0.75
4	0.50	0.50	0.50	1.00	0.75	0.75
5	0.50	0.50	0.75	0.75	1.25	0.75
6	0.50	0.50	0.75	0.75	0.75	1.25



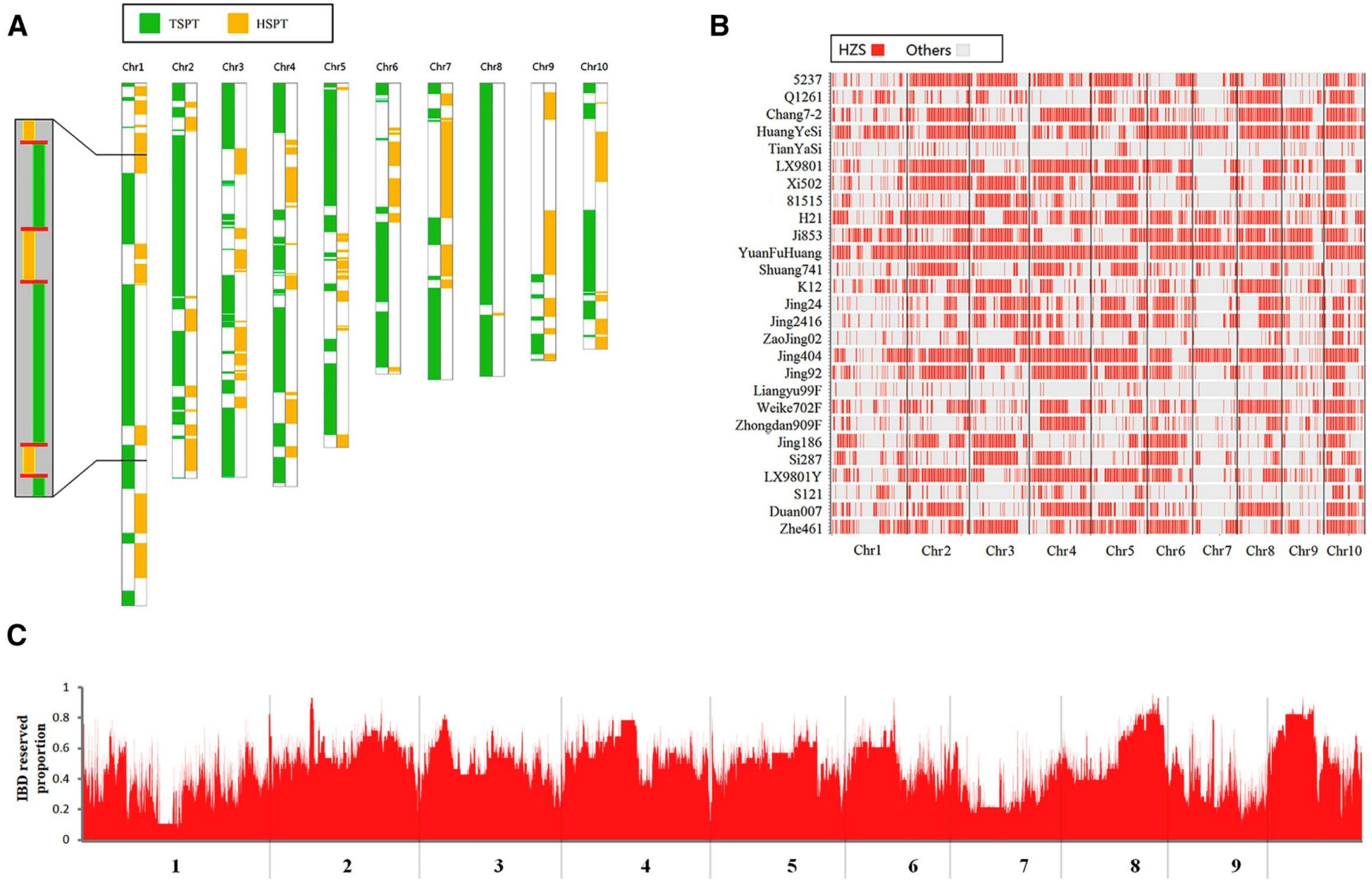
Coefficient of additive genetic covariance between individuals, $A_{ij} = 2f_{ij}$.

Recall that $f_{ii} = \frac{1}{2}(1 + F_i)$, so the diagonals are $1 + F_i$.

Note: A is a symmetric matrix, $A_{ij} = A_{ji}$.

Similar can be done for the dominance GRM.

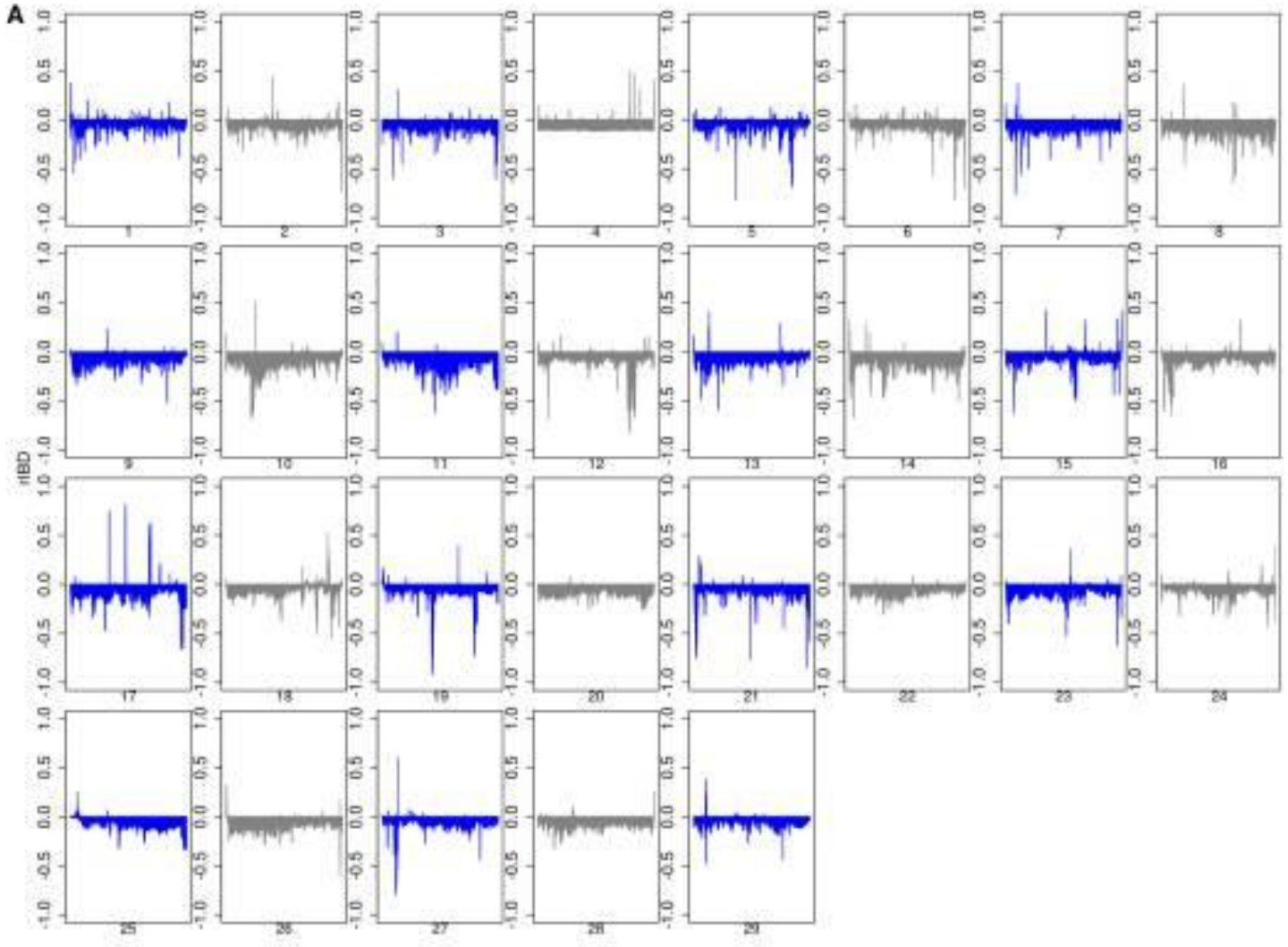
IBD in breeding and evolutionary history understanding



This highlights the prevalence of a single variety in the Chinese maize breeding history.

Li et al (2019) The HuangZaoSi maize genome provides insights into genomic variation and improvement history of maize.

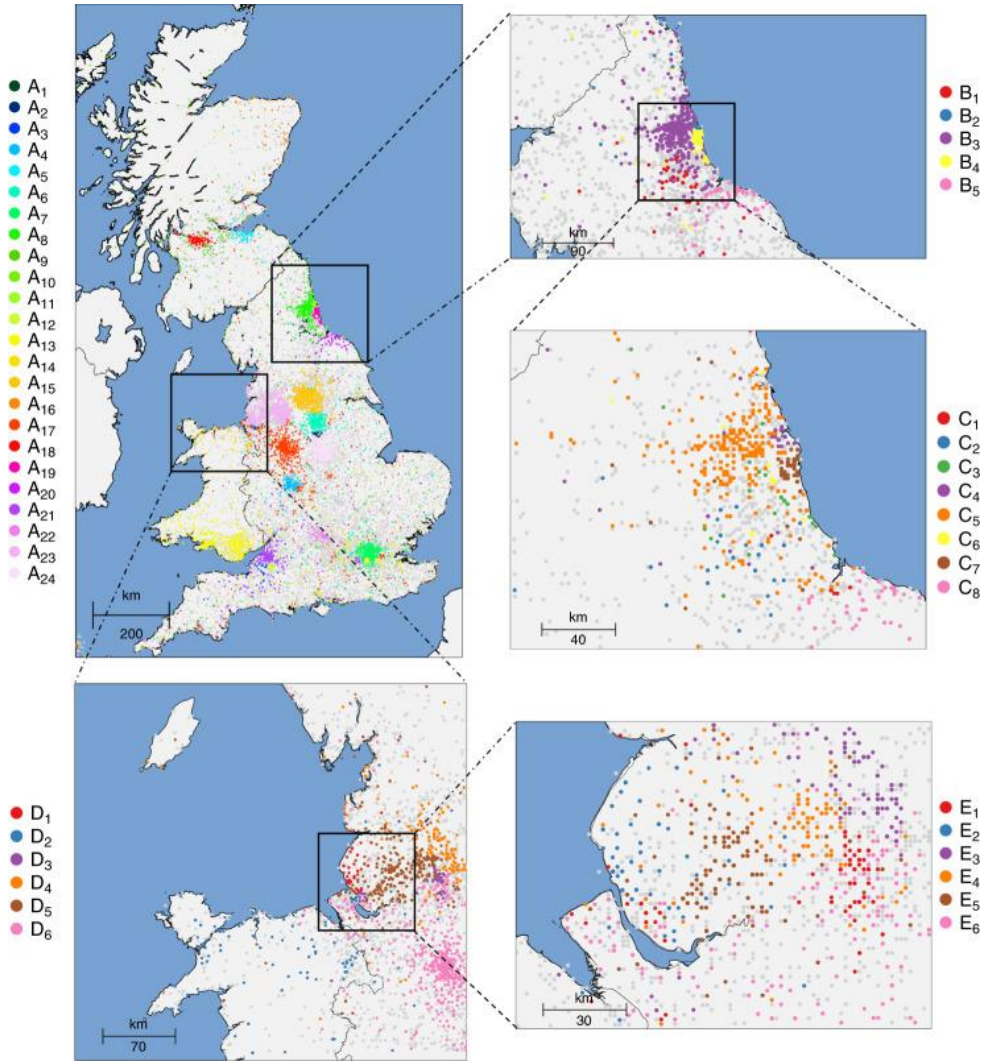
IBD in breeding and evolutionary history understanding



Similar work in cattle, this shows the genome-wide Holstein introgression in Danish cattle.

Zhang et al. (2018) Human-mediated introgression of haplotypes in a modern dairy cattle breed

IBD in population demographic inference



Saada et al. (2020) Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations

- Historical migration patterns
- Population admixture
- Genetic diseases

IBD in genealogy and ancestry identification

IBD haplotype inference as a commercial product.

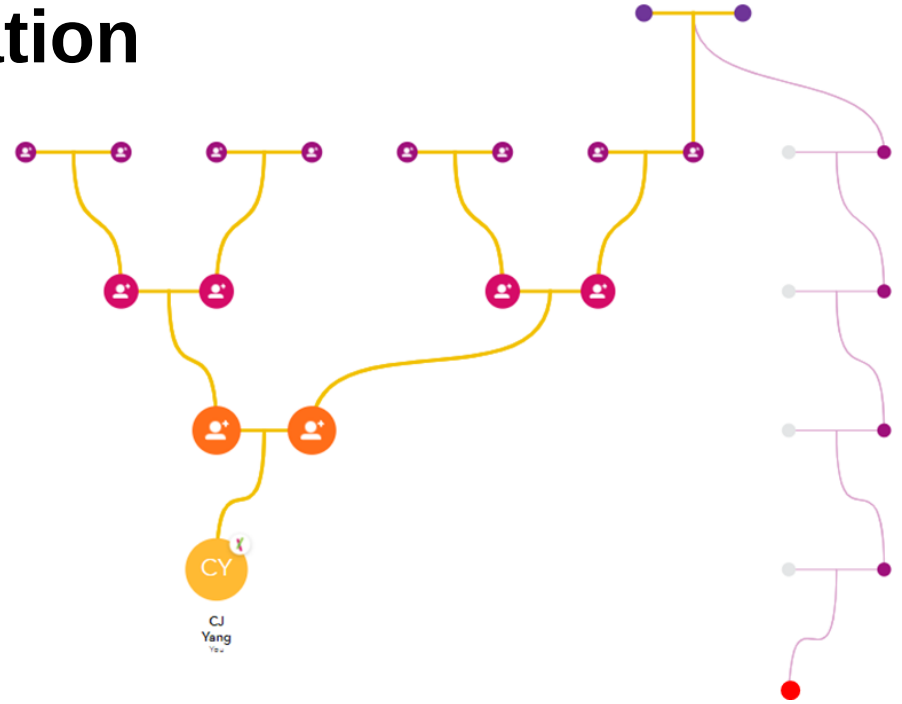
23andMe

You have new DNA Relatives

Dear CJ,

17 people who share DNA with you have joined DNA Relatives over the past 40 days.

Visit DNA Relatives →



Your genetic relationship

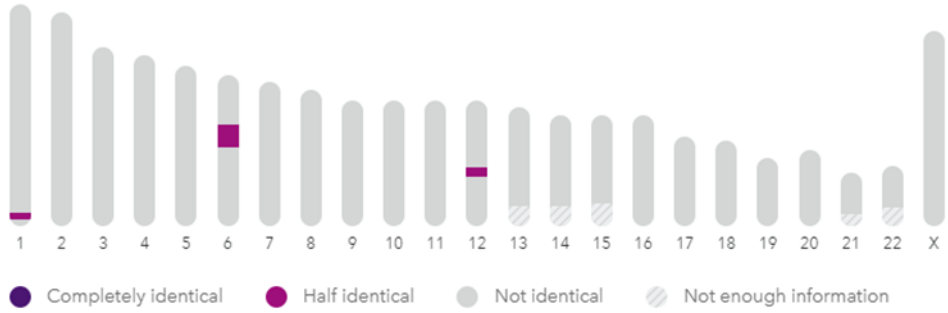
Predicted relationship

Third Cousin Once Removed

Shared DNA

0.68%

51cM



IBD in forensics

Dall-E (2024)

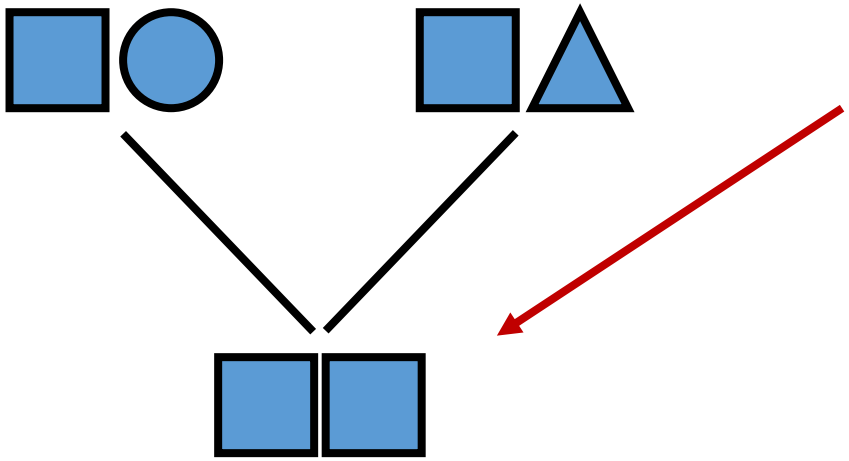
- Matching IBD segments between crime scene samples and database/relatives.
- Cooler in TV than it actually is.

Challenges in working with IBD

- Genotypic data quality (errors, missing data)
- Complicated population demography
- Assumptions in estimating coefficients
- Small sample size
- Recombination
- Limitation in computational power
- Ethical concerns in data usage
- It is average, at best

Identity-by-state (IBS)

Two of the same alleles regardless of their ancestral origins are considered identical by state.



If the parents are “unrelated”, then they are IBS.

If the parents are related, then they are IBD.

If two alleles are IBD, they are also IBS.

But, if two alleles are IBS, they are not necessarily IBD.

Example 1

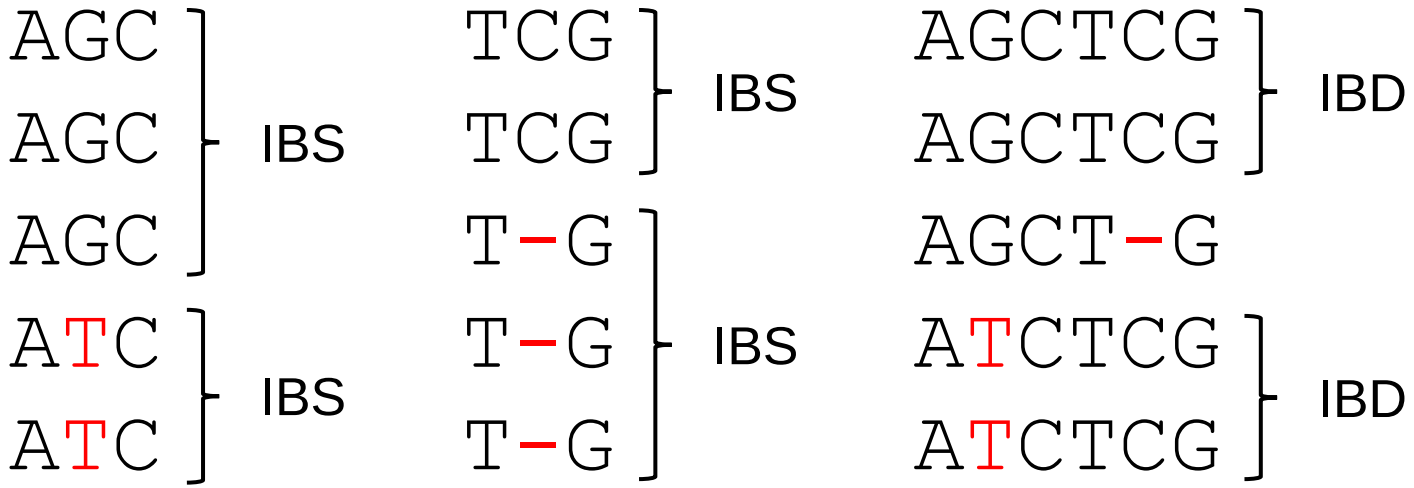
IBS is often described in the context of molecular markers.

Biallelic single nucleotide polymorphisms (SNP) vs multiallelic gene.

For example, there are 4 alleles in the founders.

```
AGCTCG
AGCT-G
ATCTCG
ATCT-G
```

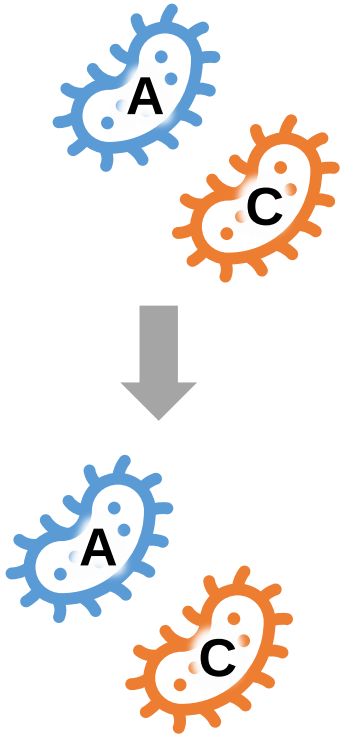
Some generations later, we genotype the descendants and find these polymorphisms.



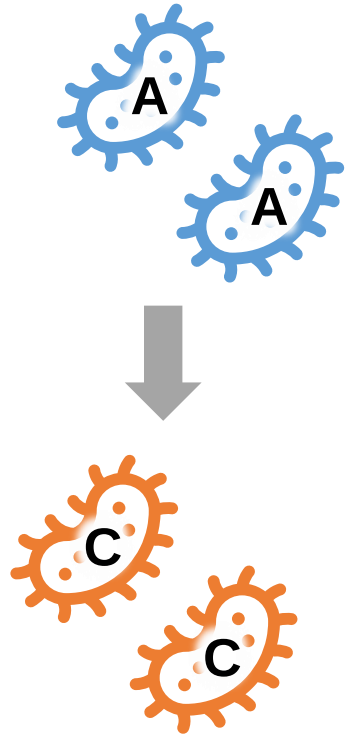
Example 2

IBS can happen due to convergent evolution.

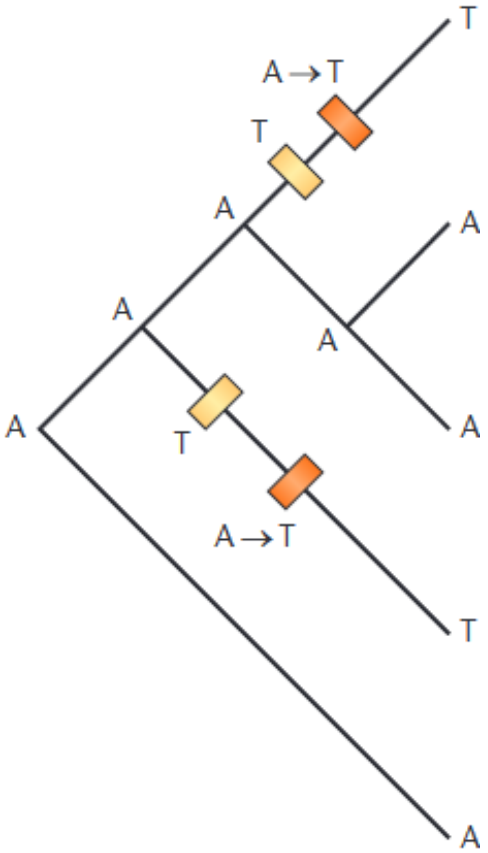
Environment 1



Environment 2



Ability to feed on poisonous milkweed



Stern (2013)

IBS, in a nutshell

https://knowyourmeme.com/memes/spider-man-pointing-at-spider-man

Applications of IBS

1. Genetic relationship matrix (GRM) calculation
2. Linkage disequilibrium identification
3. Association mapping

IBS in genetic relationship matrix (GRM) calculation

Genomic-based (vs pedigree-based)

Suppose we have a genotype matrix (X) with 6 individuals and 10 markers,

	m01	m02	m03	m04	m05	m06	m07	m08	m09	m10
id1	2	2	0	0	0	2	0	0	2	2
id2	2	2	0	0	0	2	2	2	0	2
id3	0	2	0	2	0	0	0	2	0	0
id4	0	0	0	0	2	0	0	0	0	0
id5	2	0	2	2	0	0	2	2	0	0
id6	0	0	2	0	2	0	0	0	0	0

we can calculate the K matrix as $K = \frac{WW'}{\sum 2p_j(1-p_j)}$

where $W_{ij} = X_{ij} - 2p_j$

and p_j is the allele frequency for j^{th} marker.

	id1	id2	id3	id4	id5	id6
id1	2.50	1.00	-0.65	-0.65	-1.25	-0.95
id2	1.00	2.20	-0.35	-1.25	-0.05	-1.55
id3	-0.65	-0.35	1.60	-0.20	0.10	-0.50
id4	-0.65	-1.25	-0.20	1.60	-0.80	1.30
id5	-1.25	-0.05	0.10	-0.80	2.20	-0.20
id6	-0.95	-1.55	-0.50	1.30	-0.20	1.90

Additive GRM

Additive GRM is often denoted as $\boldsymbol{K_A}$.

Given a genotype matrix $\boldsymbol{X}$ with n rows of individuals and m columns of markers coded as 0/1/2, there are several variants of how $\boldsymbol{K_A}$ can be calculated.

VanRaden (2008) method 1

Endelman and Jannink (2012) rrBLUP

$$\boldsymbol{K_A} = \frac{\boldsymbol{WW'}}{\sum 2p_j(1-p_j)} \; where \; W_{ij} = X_{ij} - 2p_j$$

VanRaden (2008) method 2

Yang et al. (2011) GCTA

$$\boldsymbol{K_A} = \frac{\boldsymbol{WW'}}{m} \; where \; W_{ij} = \frac{X_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)}}$$

Speed et al. (2012) LDAK

$$\boldsymbol{K_A} = \frac{\boldsymbol{WW'}}{m} \; where \; W_{ij} = \frac{X_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)}} \cdot \sqrt{\frac{\omega_j m}{\sum \omega_j}}$$

Note: ω_j is the marker weight adjusted by LD.

Dominance GRM

Dominance GRM is often denoted as $\boldsymbol{K_D}$.

Su et al. (2012)

$$\boldsymbol{K_D} = \frac{\boldsymbol{WW'}}{\sum\left(2p_j(1-p_j)\right)\left(1-2p_j(1-p_j)\right)} \; where \; W_{ij} = H_{ij} - 2p_j\left(1-p_j\right)$$

$$H_{ij} = \begin{cases} 0 \; if \; X_{ij} = 0 \\ 1 \; if \; X_{ij} = 1 \\ 0 \; if \; X_{ij} = 2 \end{cases}$$

Zhu et al. (2015) GCTA

$$\boldsymbol{K_D} = \frac{\boldsymbol{WW'}}{m} \; where \; W_{ij} = \frac{H_{ij} - 2{p_j}^2}{2p_j\left(1-p_j\right)}$$

$$H_{ij} = \begin{cases} 0 \; if \; X_{ij} = 0 \\ 2p_j \; if \; X_{ij} = 1 \\ 4p_j - 2 \; if \; X_{ij} = 2 \end{cases}$$

Uses of GRM

Typically, any quantitative trait can be partitioned into mean (μ), genetic (g) and residual (e) effects.

$$Y = \mu + g + e$$

To do that, we need to know the relationships of g and e among the individuals in a population.

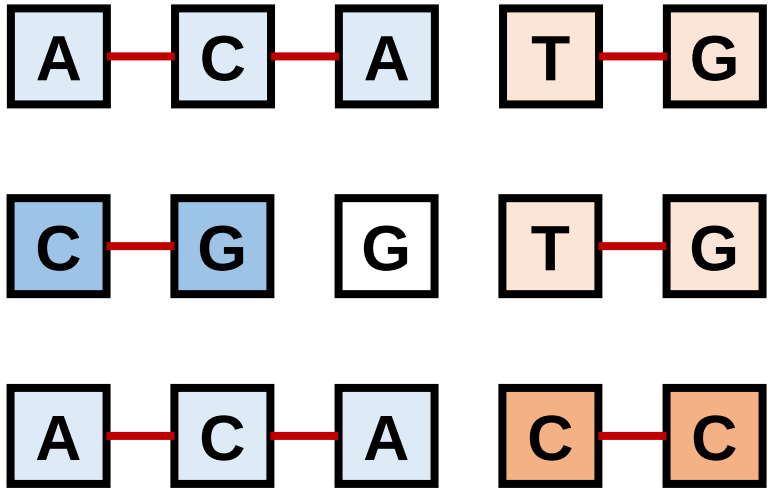
e is assumed uncorrelated among the individuals, so the relationship is just an identity matrix.

g is assumed correlated among the individuals, and the relationship is the GRM.

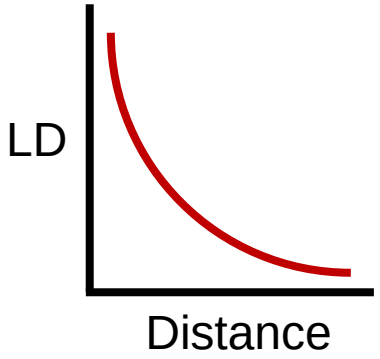
So, GRM is important in modern quantitative genetics – principal components, GWAS, genomic prediction, variance components, etc.

IBS in linkage disequilibrium (LD) identification

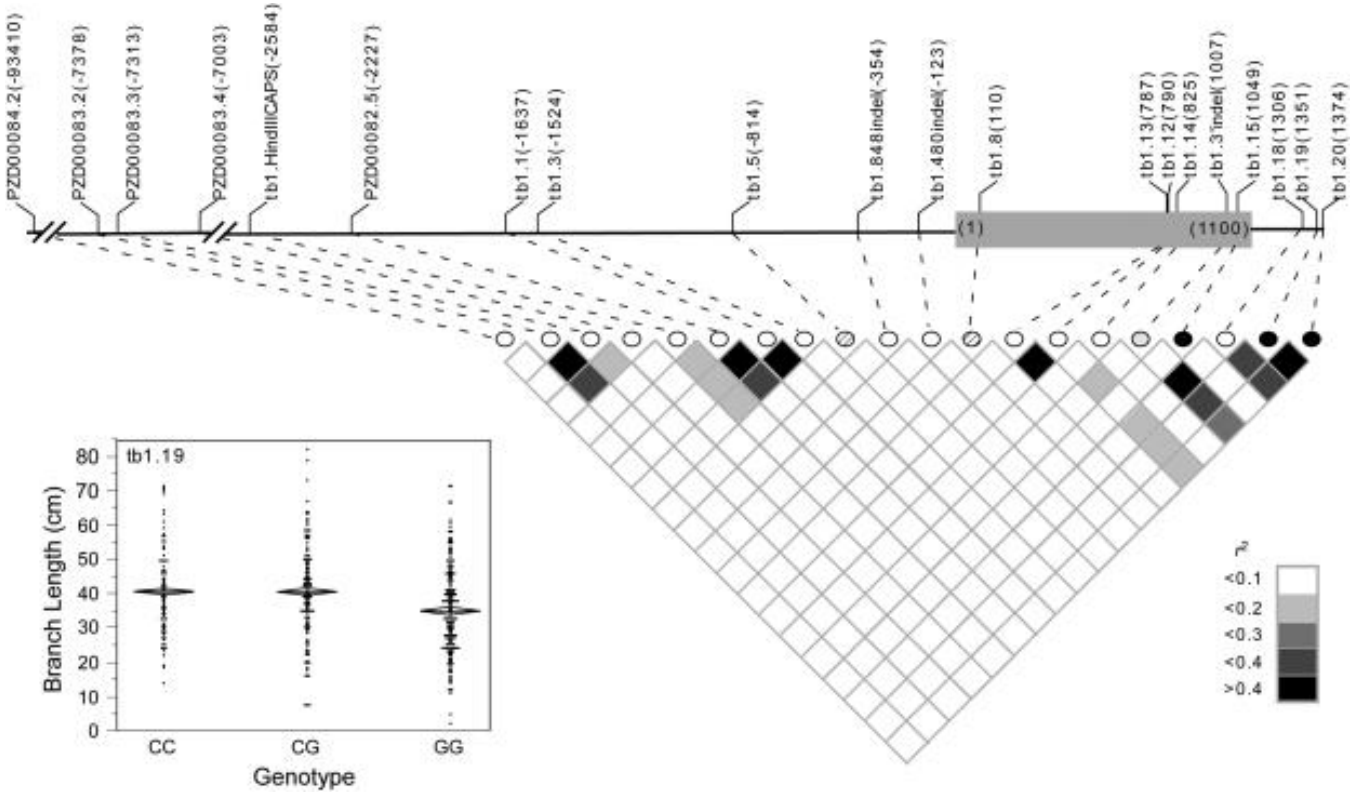
LD blocks



LD decay



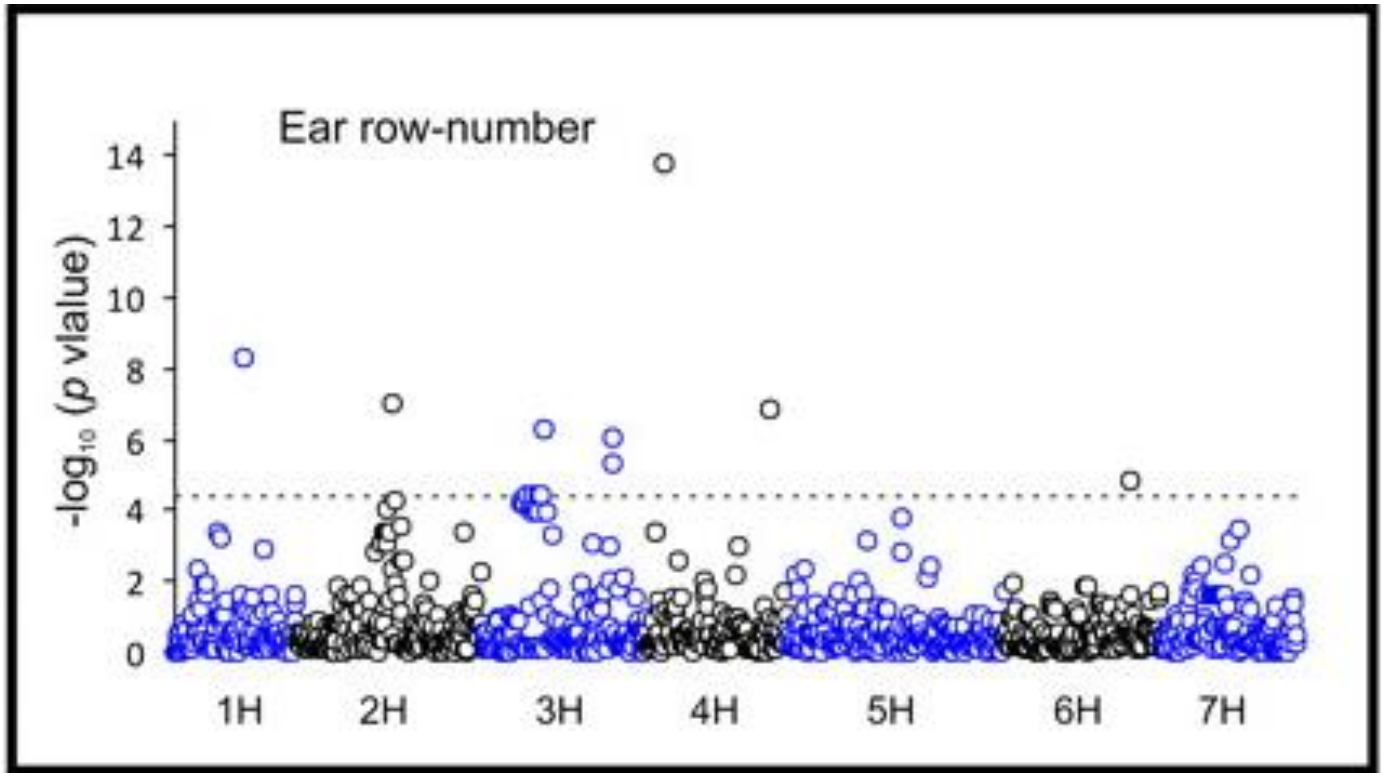
LD in a teosinte population



Weber et al (2007)

Association mapping

GWAS in barley varieties



Cockram et al (2010)

- GRM to correct for genetic background effect.
- Biallelic SNP testing.
- LD between significant SNP and candidate genes.

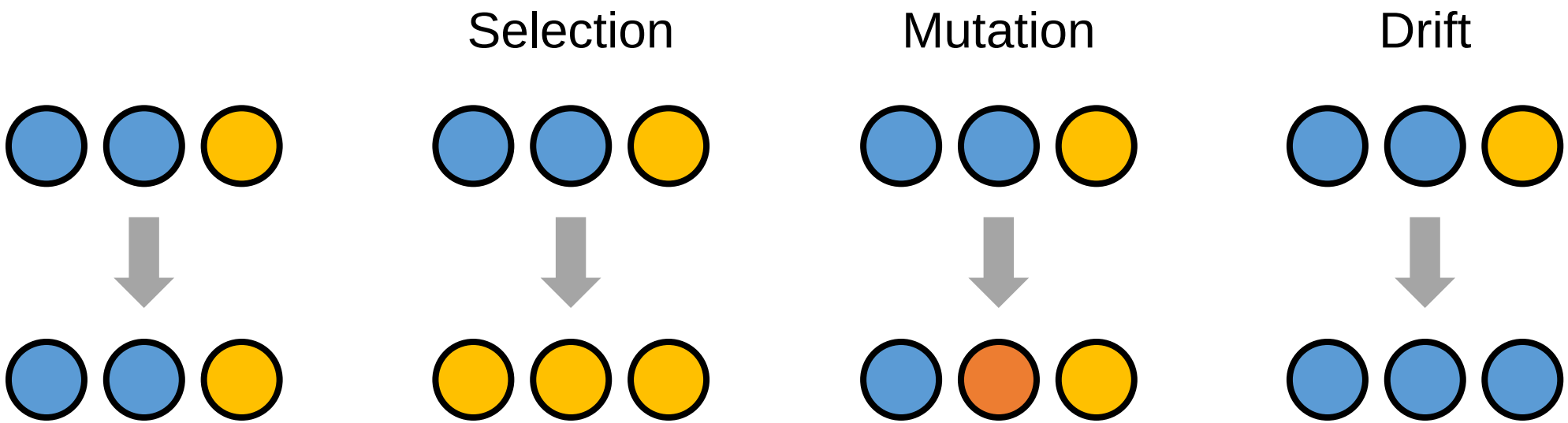
Challenges in working with IBS

- Genotypic data quality (errors, missing data)
- Complicated population demography
- Small sample size
- Limitation in computational power
- Lack of genealogical support for biological significance

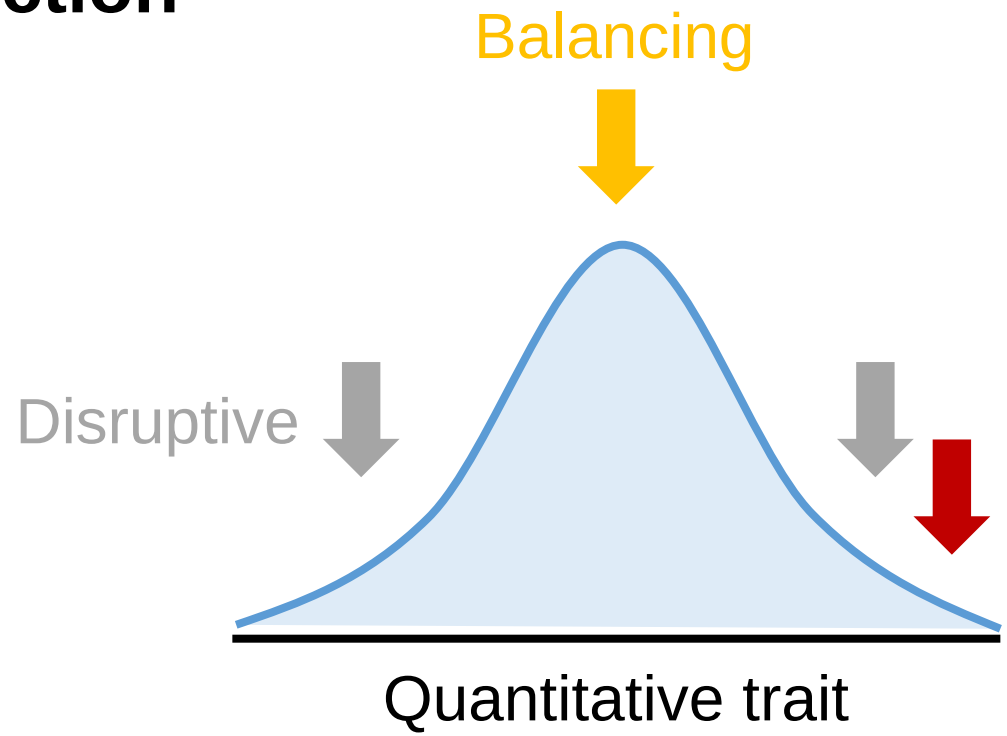
Genetic distance

A measure of genetic divergence between populations or sub-populations.

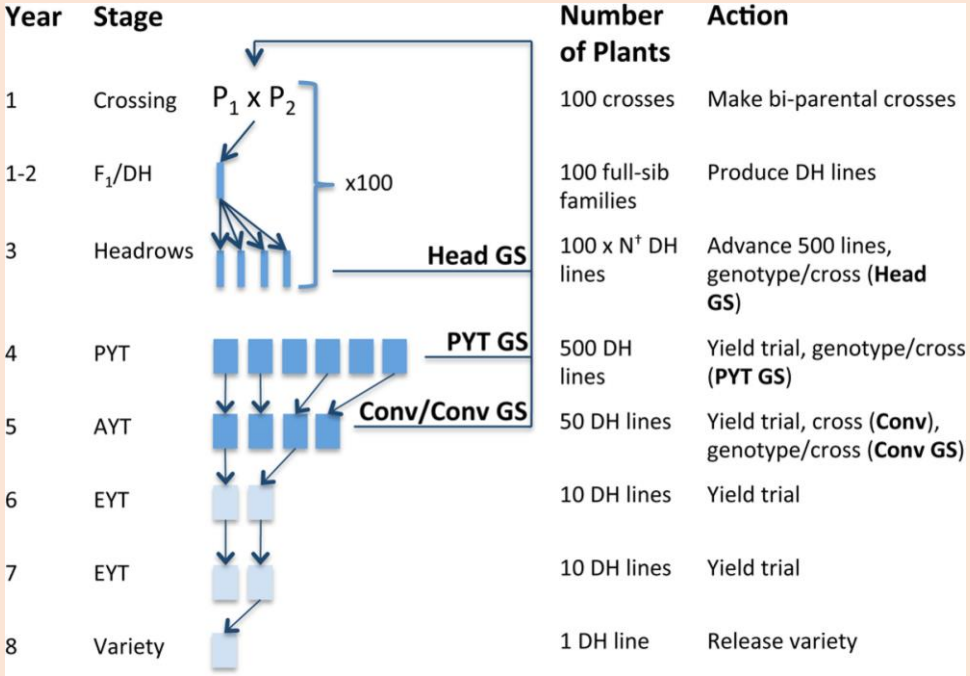
Common causes:



Selection



Directional/truncation

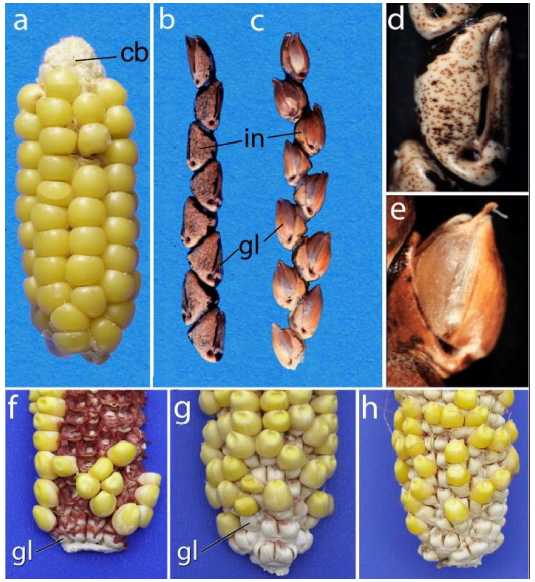


Gaynor et al (2017)

Mutation → Novel variation by interfering with the DNA replication and repair system

Natural mutation

- Inherent mutation rate
- Long term divergence
- Can be important for crop evolution



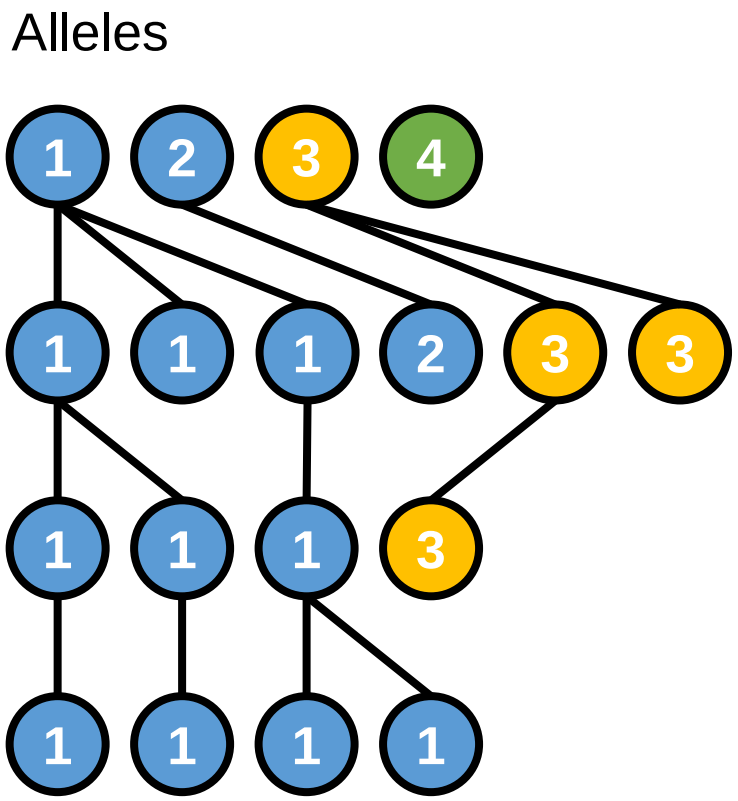
Wang et al (2005)

E.g. *tga1* in maize

Induced mutation

- Random: mutagen like EMS, γ-ray
- Targeted: transgenic, gene editing
- Common approach for breeding vegetable and ornamental

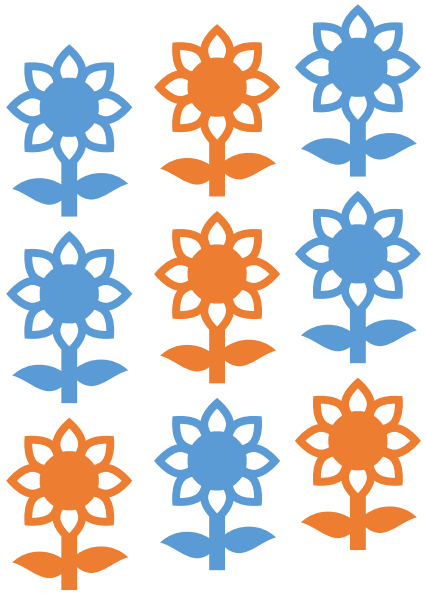
Drift



- Unpredictable.
- Follows a multinomial distribution.
- Binomial distribution in the case of two alleles.
- Mean = p_i.
- Variance = $\frac{p_i\,(1-p_i)}{2N_{i+1}}$.
- In the absence of mutation, $F_t = 1 - \left(1 - \frac{1}{2N}\right)^t$.
- Heterozygosity, $H_t = 1 - F_t = \left(1 - \frac{1}{2N}\right)^t$.

Wahlund effect

Under Hardy-Weinberg equilibrium (HWE), presence of subpopulations can reduce overall heterozygosity even without inbreeding.



Flowers in the morning

Flowers in the afternoon

$$H_{1+2} < H \; if \; p_1 \neq p_2$$

H_{1+2} = combined heterozygosity

H = overall heterozygosity

Proof for Wahlund effect

Under HWE, $P(AA) + P(Aa) + P(aa) = p^2 + 2p(1-p) + (1-p)^2 = 1$

Heterozygosity is $H = 2p(1-p)$

Heterozygosities in subpopulation 1 and 2 are $H_1 = 2p_1(1-p_1)$

$H_2 = 2p_2(1-p_2)$

Combined heterozygosity is the weighted average of H_1 and H_2

$$H_{1+2} = \omega_1 H_1 + \omega_2 H_2 \qquad \omega_1 + \omega_2 = \frac{n_1}{n_1+n_2} + \frac{n_2}{n_1+n_2} = 1$$

$$H_{1+2} = 2\omega_1 p_1(1-p_1) + 2\omega_2 p_2(1-p_2)$$

$$H_{1+2} = 2\omega_1 p_1 - 2\omega_1 p_1^2 + 2\omega_2 p_2 - 2\omega_2 p_2^2$$

Note:

n is the population size

p is the allele frequency

Proof for Wahlund effect

Recall $H_{1+2} = 2\omega_1 p_1 - 2\omega_1 p_1^2 + 2\omega_2 p_2 - 2\omega_2 p_2^2$

Overall heterozygosity is $H = 2p(1-p)$ and because $p = \omega_1 p_1 + \omega_2 p_2$

$$H = 2(\omega_1 p_1 + \omega_2 p_2)\big(1 - (\omega_1 p_1 + \omega_2 p_2)\big)$$

$$H = 2\omega_1 p_1 - 2\omega_1^2 p_1^2 - 2\omega_1\omega_2 p_1 p_2 + 2\omega_2 p_2 - 2\omega_1\omega_2 p_1 p_2 - 2\omega_2^2 p_2^2$$ Expand the equation

$$H = 2\omega_1 p_1 - 2\omega_1^2 p_1^2 + 2\omega_2 p_2 - 2\omega_2^2 p_2^2 - 4\omega_1\omega_2 p_1 p_2$$ Rearrange the terms

$$H = 2\omega_1 p_1 - 2\omega_1 p_1^2 + 2\omega_1 p_1^2 - 2\omega_1^2 p_1^2 + 2\omega_2 p_2 - 2\omega_2 p_2^2 + 2\omega_2 p_2^2 - 2\omega_2^2 p_2^2 - 4\omega_1\omega_2 p_1 p_2$$ Tricks

$$H = H_{1+2} + 2\omega_1 p_1^2 - 2\omega_1^2 p_1^2 + 2\omega_2 p_2^2 - 2\omega_2^2 p_2^2 - 4\omega_1\omega_2 p_1 p_2$$ Replace the terms

$$H = H_{1+2} + 2\omega_1 p_1^2 - 2\omega_1(1-\omega_2) p_1^2 + 2\omega_2 p_2^2 - 2\omega_2(1-\omega_1) p_2^2 - 4\omega_1\omega_2 p_1 p_2$$ Tricks

$$H = H_{1+2} + 2\omega_1 p_1^2 - 2\omega_1 p_1^2 + 2\omega_1\omega_2 p_1^2 + 2\omega_2 p_2^2 - 2\omega_2 p_2^2 + 2\omega_1\omega_2 p_2^2 - 4\omega_1\omega_2 p_1 p_2$$ Expand the equation

$$H = H_{1+2} + 2\omega_1\omega_2 p_1^2 + 2\omega_1\omega_2 p_2^2 - 4\omega_1\omega_2 p_1 p_2$$ Simplify the equation

$$H = H_{1+2} + 2\omega_1\omega_2(p_1^2 + p_2^2 - 2p_1 p_2)$$ Rearrange the terms

$$H = H_{1+2} + 2\omega_1\omega_2(p_1 - p_2)^2$$

$H_{1+2} < H \; if \; p_1 \neq p_2$

How to compute genetic distance?

- $p_{x,ij}$ and $p_{y,ij}$ are the frequencies of allele j at locus i in population x and y.
- m is the number of markers.

Roger's distance (~ Euclidean)

$$D = \frac{1}{m}\sum_i \sqrt{\frac{1}{2}\sum_j \left(p_{x,ij} - p_{y,ij}\right)^2}$$

Special case: biallelic $D = \frac{1}{m}\sum_i \left|p_{x,ij} - p_{y,ij}\right|$

Pop	A1	A2	A3
X	0.2	0.5	0.3
Y	0.3	0.1	0.6

Pop	B1	B2	B3
X	0.4	0.2	0.4
Y	0.4	0.5	0.1

$$D = \frac{1}{2}\left\{\sqrt{\frac{1}{2}\left[(0.2-0.3)^2 + (0.5-0.1)^2 + (0.3-0.6)^2\right]} + \sqrt{\frac{1}{2}\left[(0.4-0.4)^2 + (0.2-0.5)^2 + (0.4-0.1)^2\right]}\right\}$$

$$D = 0.3303$$

How to compute genetic distance?

- $p_{x,ij}$ and $p_{y,ij}$ are the frequencies of allele j at locus i in population x and y.
- m is the number of markers.

Nei's distance (Nei 1972)

$$D = -ln\left(\frac{\sum_i \sum_j p_{x,ij} p_{y,ij}}{\sqrt{\sum_i \sum_j p_{x,ij}^2 \cdot \sum_i \sum_j p_{y,ij}^2}}\right)$$

Pop	A1	A2	A3
X	0.2	0.5	0.3
Y	0.3	0.1	0.6

Pop	B1	B2	B3
X	0.4	0.2	0.4
Y	0.4	0.5	0.1

$$D = -ln\left[\frac{0.2 \cdot 0.3 + 0.5 \cdot 0.1 + 0.3 \cdot 0.6 + 0.4 \cdot 0.4 + 0.2 \cdot 0.5 + 0.4 \cdot 0.1}{\sqrt{(0.2^2 + 0.5^2 + 0.3^2 + 0.4^2 + 0.2^2 + 0.4^2) \cdot (0.3^2 + 0.1^2 + 0.6^2 + 0.4^2 + 0.5^2 + 0.1^2)}}\right]$$

$$D = 0.3132$$

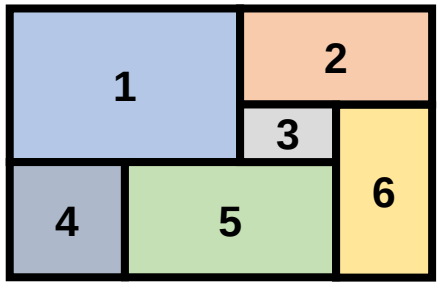
There are many other ways to compute genetic distance that we will not cover here.

Wright's F statistics, F_{ST}

In the presence of inbreeding, the genotype frequencies in a population are:

$$P(AA) = p^2 + Fp(1-p)$$
$$P(Aa) = 2p(1-p)(1-F)$$
$$P(aa) = (1-p)^2 + Fp(1-p)$$

In the presence of sub-populations, the genotype frequencies are:



$$P(AA) = p_i^2 + F_i p_i(1-p_i)$$
$$P(Aa) = 2p_i(1-p_i)(1-F_i)$$
$$P(aa) = (1-p_i)^2 + F_i p_i(1-p_i)$$

Wright's F statistics, F_{ST}

Let the heterozygosity of sub-population i be the same as the population, we have

$$2p_i(1-p_i)(1-F_i) = 2p(1-p)(1-F)$$

$$\frac{(1-F)}{(1-F_i)} = \frac{2p_i(1-p_i)}{2p(1-p)}$$ ← Rearrange the terms

$$\frac{(1-F)}{(1-F_i)} = 1 - F_{ST,i}$$ ← Tricks $F_{ST,i} = 1 - \frac{2p_i(1-p_i)}{2p(1-p)}$

$$(1-F) = \left(1-F_{ST,i}\right)(1-F_i)$$ ← Rearrange the terms

$$(1-F_{IT}) = (1-F_{ST})(1-F_{IS})$$

0?

F_{IT} is the population (overall) inbreeding coefficient.

F_{IS} is the sub-population inbreeding coefficient.

F_{ST} is the fixation index, i.e. population differentiation.

Alternative ways to compute F_{ST}

$$F_{ST} = \frac{1}{n}\sum F_{ST,i}$$

Overall F_{ST} while ignoring sub-population size differences.

$$F_{ST} = \frac{Var(p')}{p(1-p)}$$

Can be derived from the above.
Note that $Var(p')$ is the variance of sub-population allele frequencies.

$$F_{ST,i} = 1 - \left(1 - \frac{1}{2N_i}\right)^t$$

F_{ST} for sub-population under drift alone.
Note that N_i is the sub-population size and t is the number of generation.

$$F_{ST,i} = \frac{\pi_b - \pi_{w,i}}{\pi_b}$$

π_b is the nucleotide diversity (average number of pairwise difference) between sub-pops. $\pi_{w,i}$ is the nucleotide diversity within sub-pop i.

$$F_{ST,i} = \frac{M_{w,i} - M_b}{1 - M_b}$$

M is the average allele matches (Weir and Goudet 2017).

And many more methods not described here.

How to derive $F_{ST} = \frac{Var(p')}{p(1-p)}$

Using $n = 2$ as example,

$$F_{ST} = \frac{1}{2}\sum F_{ST,i}$$

$$= \frac{1}{2}\left(1 - \frac{p_1(1-p_1)}{p(1-p)} + 1 - \frac{p_2(1-p_2)}{p(1-p)}\right)$$

$$= \frac{1}{2}\left(\frac{p(1-p) - p_1(1-p_1) + p(1-p) - p_2(1-p_2)}{p(1-p)}\right)$$

$$= \frac{1}{2}\left(\frac{2p + p_1^2 + p_2^2 - 2p^2 - p_1 - p_2}{p(1-p)}\right)$$

$$= \frac{1}{2}\left(\frac{p_1^2 + p_2^2 - 2p^2}{p(1-p)}\right)$$

Note that $p_1 + p_2 = 2p$

$$= \frac{1}{2}\left(\frac{p_1^2 + p_2^2 + 2p^2 - 4p^2}{p(1-p)}\right)$$

$$= \frac{1}{2}\left(\frac{p_1^2 + p_2^2 + 2p^2 - 4p\left(\frac{p_1 + p_2}{2}\right)}{p(1-p)}\right)$$

$$= \frac{1}{2}\left(\frac{p_1^2 + p_2^2 + 2p^2 - 2pp_1 - 2pp_2}{p(1-p)}\right)$$

$$= \frac{1}{2}\left(\frac{p_1^2 - 2pp_1 + p^2 + p_2^2 - 2pp_2 + p^2}{p(1-p)}\right)$$

$$= \frac{1}{2}\left(\frac{(p_1 - p)^2 + (p_2 - p)^2}{p(1-p)}\right)$$

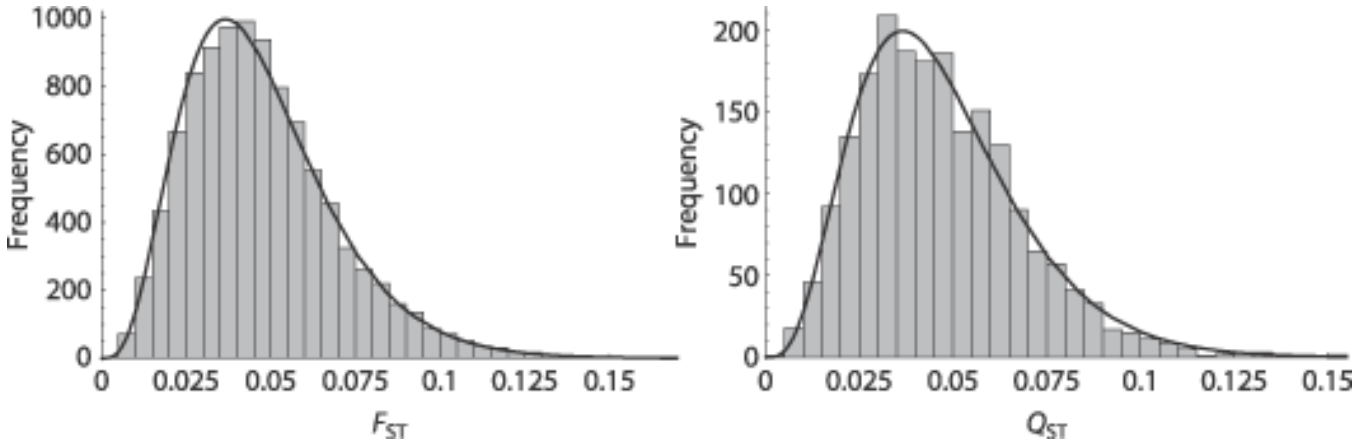
$$= \frac{1}{p(1-p)}\frac{1}{2}\sum (p_i - p)^2$$

$$= \frac{Var(p')}{p(1-p)}$$

Q_{ST}

Q_{ST} is a measure of population differentiation in quantitative traits.

$$Q_{ST} = \frac{V_{G,between}}{V_{G,between} + 2V_{G,within}}$$

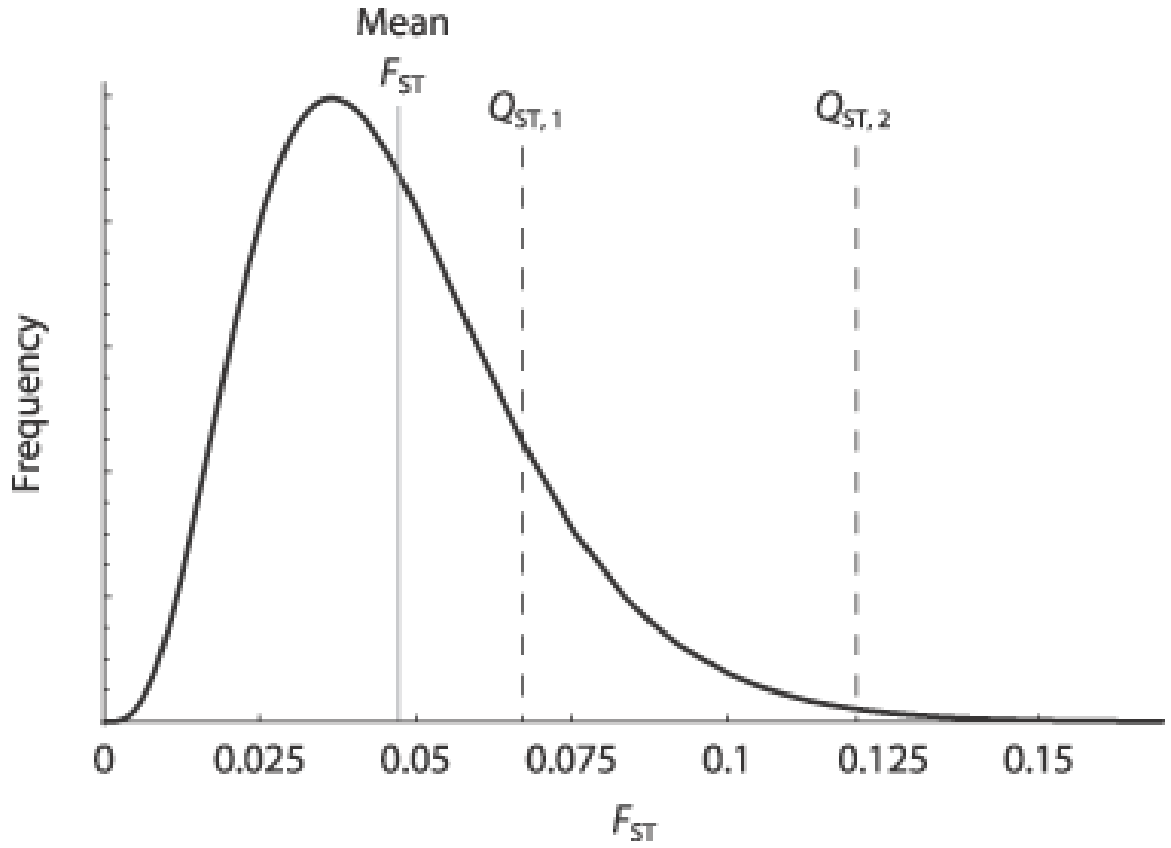


Whitlock (2008)

Figure 1. The distribution of neutral F_{ST} and Q_{ST}. Each value in these histograms is derived from a single neutral locus or a single neutral trait in an island model with 10 demes sampled. The solid lines are the χ^2 distribution predicted by Lewontin & Krakauer (1973) and are the same in both figures. For these simulations, the local population size was N = 100, and the migration rate was m = 0.05. The traits were controlled by five unlinked loci with mutational effects chosen from an exponential distribution.

Q_{ST} vs F_{ST}

Q_{ST} vs F_{ST} comparison for identifying possible selection.



Whitlock (2008)

Figure 6. Comparing a single Q_{ST} value to the distribution of F_{ST}. The Q_{ST} value for Trait 1 is greater than the mean F_{ST} (= 0.0475), but it is not unrepresentative of the distribution of F_{ST} values. It would be quite likely to generate a Q_{ST} value like $Q_{ST,1}$ from a neutral trait. Trait 2 has a Q_{ST} value that is in the tail of the distribution of F_{ST}; this trait has a Q_{ST} value that would be very unusual for a neutral trait.

Example 1a

Ind	Sub-pop	Marker
1	A	2
2	A	1
3	A	2
4	A	1
5	B	0
6	B	0
7	B	1
8	B	1

Δ_{ij}

$$F_{ST,i} = \frac{\pi_b - \pi_{w,i}}{\pi_b}$$

1. Make all possible pairwise comparisons within sub-pop A.

$$\pi_{w,1} = \frac{\Delta_{12} + \Delta_{13} + \Delta_{14} + \Delta_{23} + \Delta_{24} + \Delta_{34}}{6} = \frac{1 + 0 + 1 + 1 + 0 + 1}{6} = \frac{2}{3}$$

2. Make all possible pairwise comparisons between sub-pop A and B.

$$\pi_b = \frac{\Delta_{15} + \Delta_{16} + \Delta_{17} + \Delta_{18} + \Delta_{25} + \Delta_{26} + \Delta_{27} + \Delta_{28} + \Delta_{35} + \Delta_{36} + \Delta_{37} + \Delta_{38} + \Delta_{45} + \Delta_{46} + \Delta_{47} + \Delta_{48}}{16}$$

$$\pi_b = \frac{2 + 2 + 1 + 1 + 1 + 1 + 0 + 0 + 2 + 2 + 1 + 1 + 1 + 1 + 0 + 0}{16} = 1$$

3. Apply the formula.

$$F_{ST,1} = \frac{1 - \frac{2}{3}}{1} = \frac{1}{3}$$

What are $\pi_{w,2}$ and $F_{ST,2}$?

This method is biased when the sub-population sizes are different.

Example 1b

Ind	Sub-pop	Marker
1	A	2
2	A	1
3	A	2
4	A	1
5	B	0
6	B	0
7	B	1
8	B	1

m_{ij}

$$F_{ST,i} = \frac{M_{w,i} - M_b}{1 - M_b}$$

1. Compute all pairwise matches within sub-pop A.

$$M_{w,1} = \frac{m_{12} + m_{13} + m_{14} + m_{23} + m_{24} + m_{34}}{6} = \frac{\frac{1}{2} + 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}}{6} = \frac{7}{12}$$

2. Make all possible pairwise comparisons between sub-pop A and B.

$$\pi_b = \frac{m_{15} + m_{16} + m_{17} + m_{18} + m_{25} + m_{26} + m_{27} + m_{28} + m_{35} + m_{36} + m_{37} + m_{38} + m_{45} + m_{46} + m_{47} + m_{48}}{16}$$

$$M_b = \frac{0 + 0 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 0 + 0 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}}{16} = \frac{3}{8}$$

3. Apply the formula.

$$F_{ST,1} = \frac{\frac{7}{12} - \frac{3}{8}}{1 - \frac{3}{8}} = \frac{1}{3}$$

What are $M_{w,2}$ and $F_{ST,2}$?

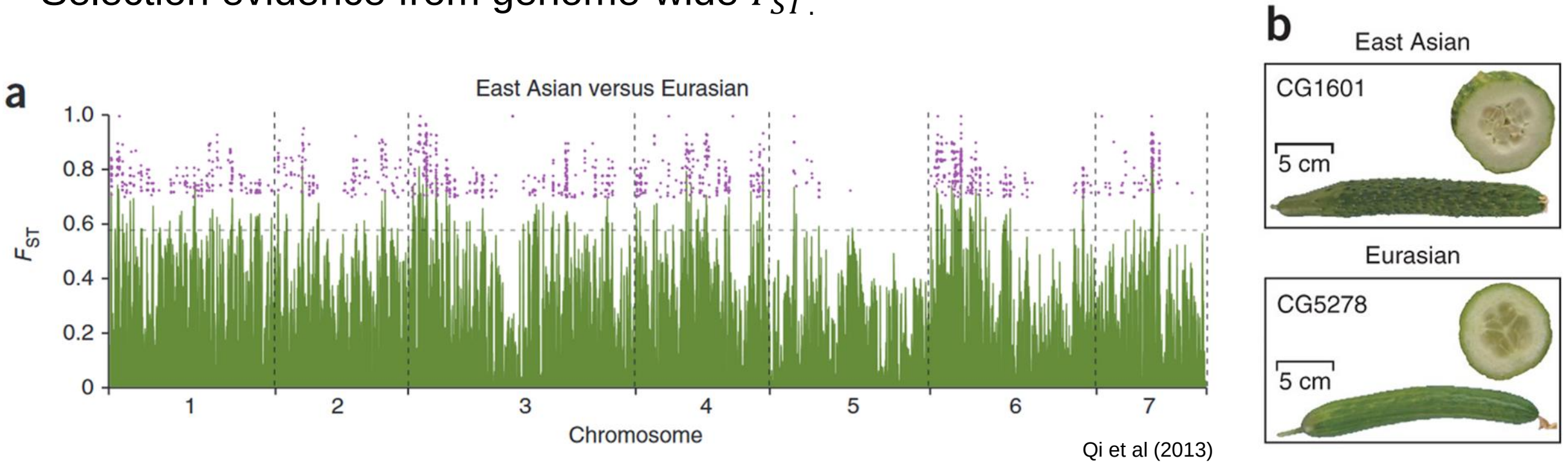
If any individual is taken out, this F_{ST} value will be different from the previous slide.

Applications of genetic distance

1. Identifying the impact of selection and breeding.
2. Understanding evolution from a population genetics angle.
3. Phylogenetics.
4. Data encryption.

Identifying the impact of selection and breeding

Selection evidence from genome-wide F_{ST}.



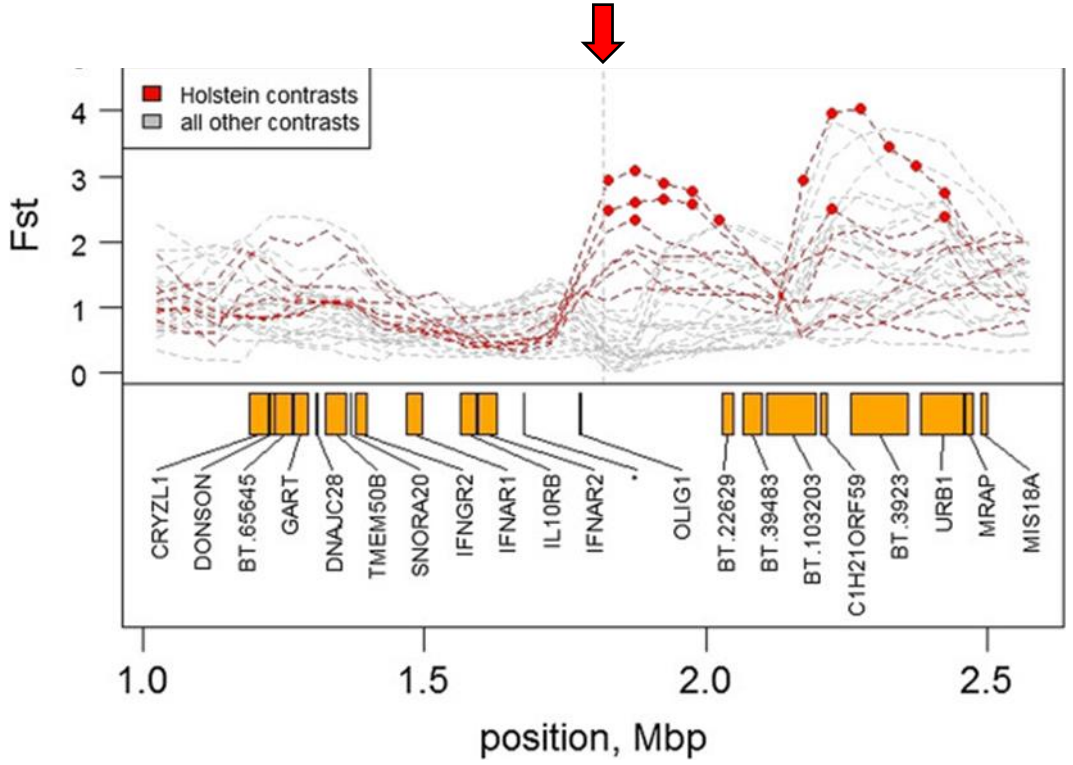
Qi et al (2013)

Figure 3. Highly divergent regions (top 5%; $F_{ST} \geq 0.57$) and nonsynonymous SNPs (top 5%; $F_{ST} \geq 0.70$) between the East Asian and Eurasian groups. Green vertical bars higher than the dashed line ($F_{ST} = 0.70$) indicate highly divergent regions; purple dots indicate highly divergent nonsynonymous SNPs.

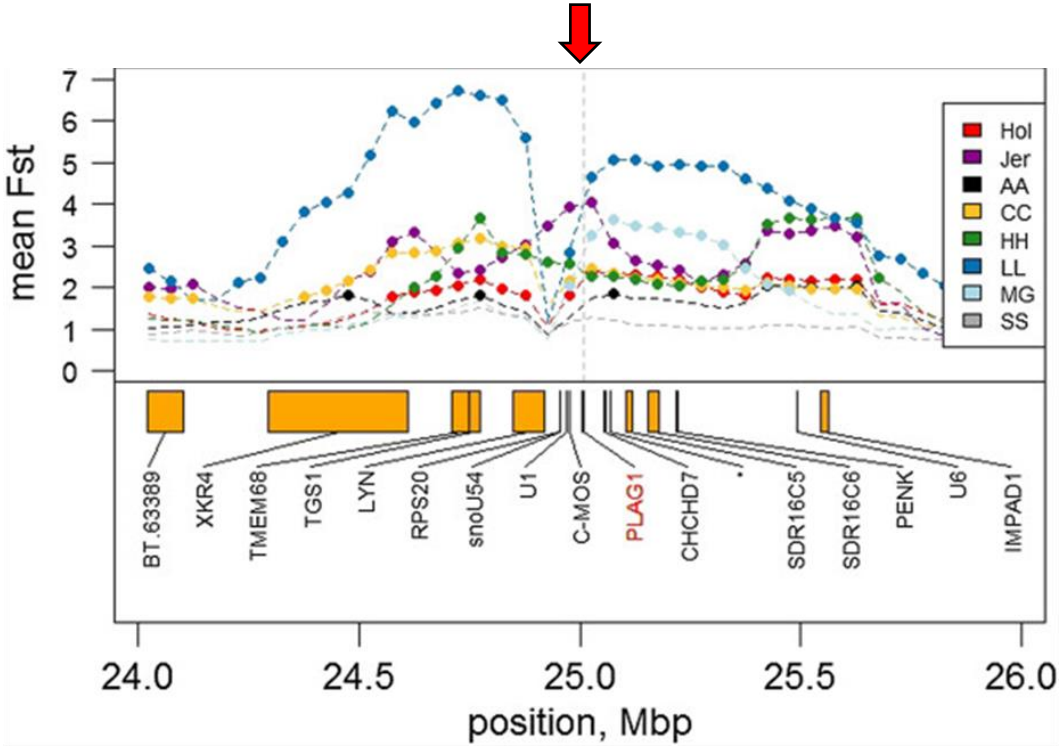
Identifying the impact of selection and breeding

F_{ST} may not always produce selection evidence.

POLLED locus – simple trait

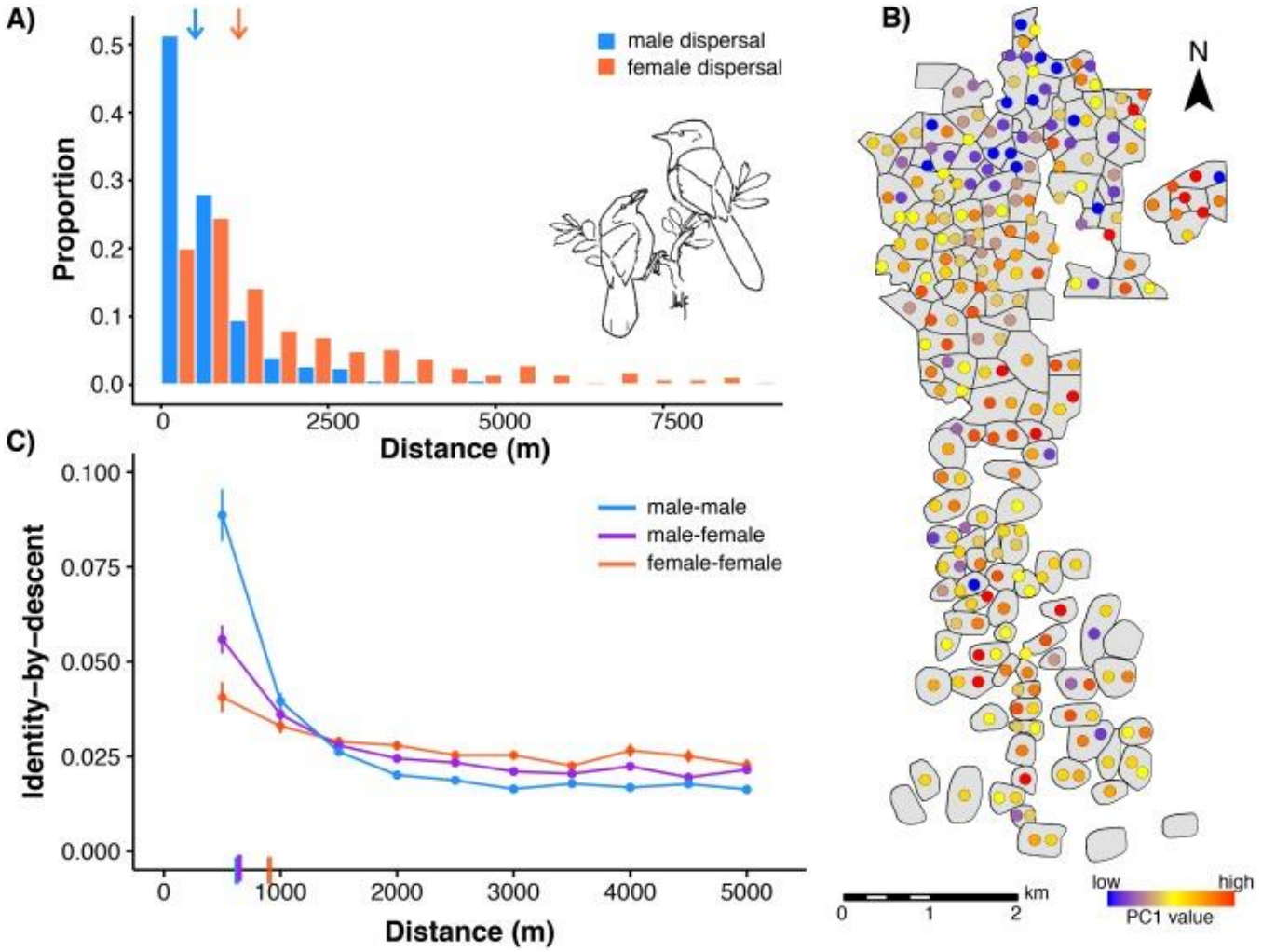


PLAG1 locus – complex trait



Kemper et al (2014)

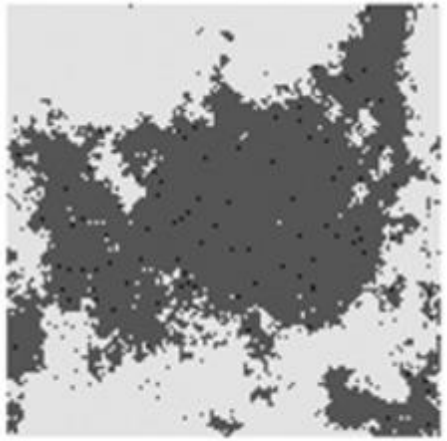
Understanding evolution from a population genetics angle



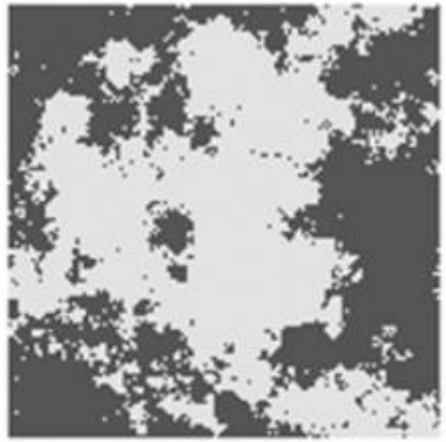
- Example of isolation-by-distance.
- Genetic distance vs geographical distance.
- Higher dispersal distance in Florida Scrub-Jay females than males.
- At short distance, the males are more genetically similar to each other than the females.

Aquillon et al (2017)

Landscape genetics



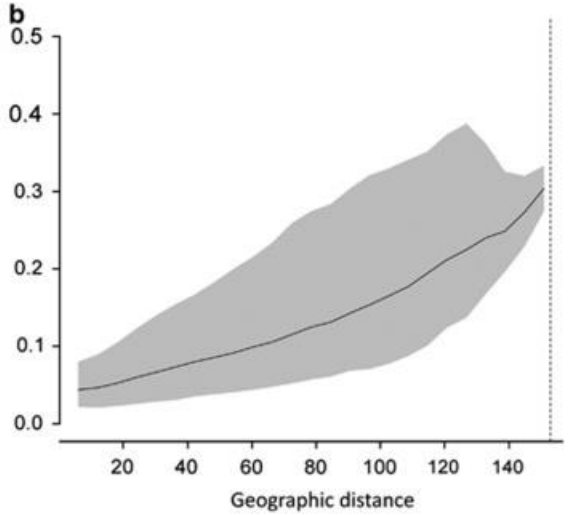
Landscape A

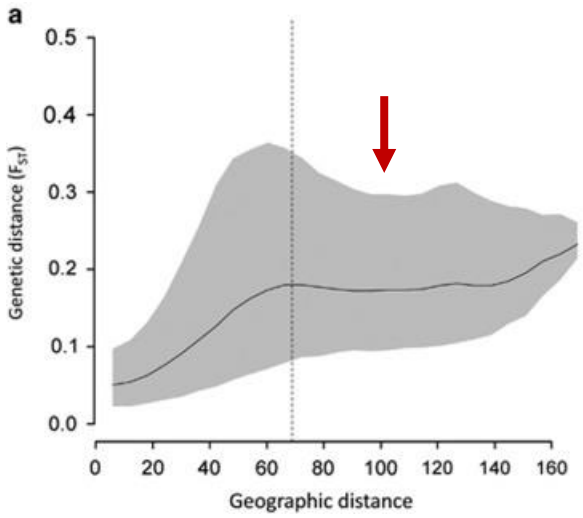


Landscape E

= Habitat

= Matrix



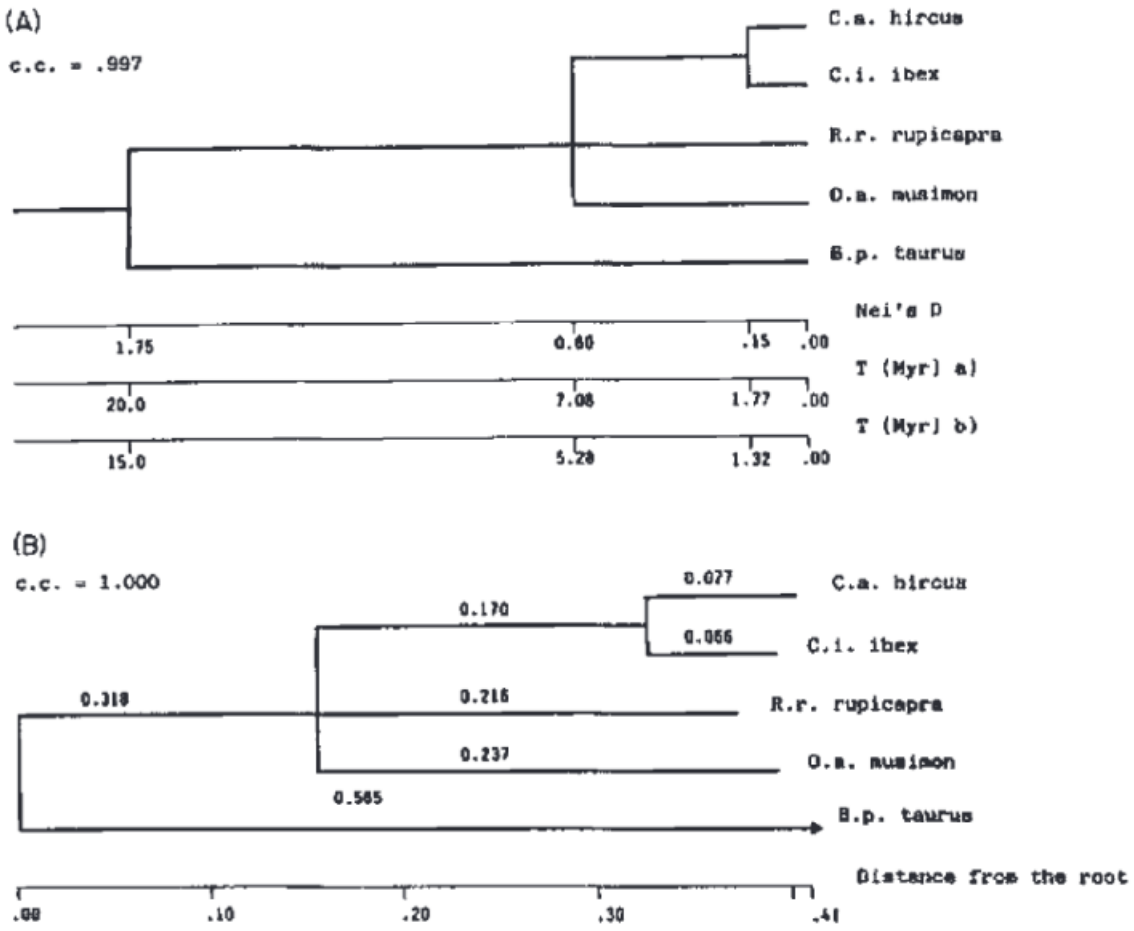


van Strien et al (2015)

The relationship between genetic distance and geographical distance is not that straightforward.

It also depends on the landscape, which determines how much gene flow vs drift between sub populations.

Phylogenetics

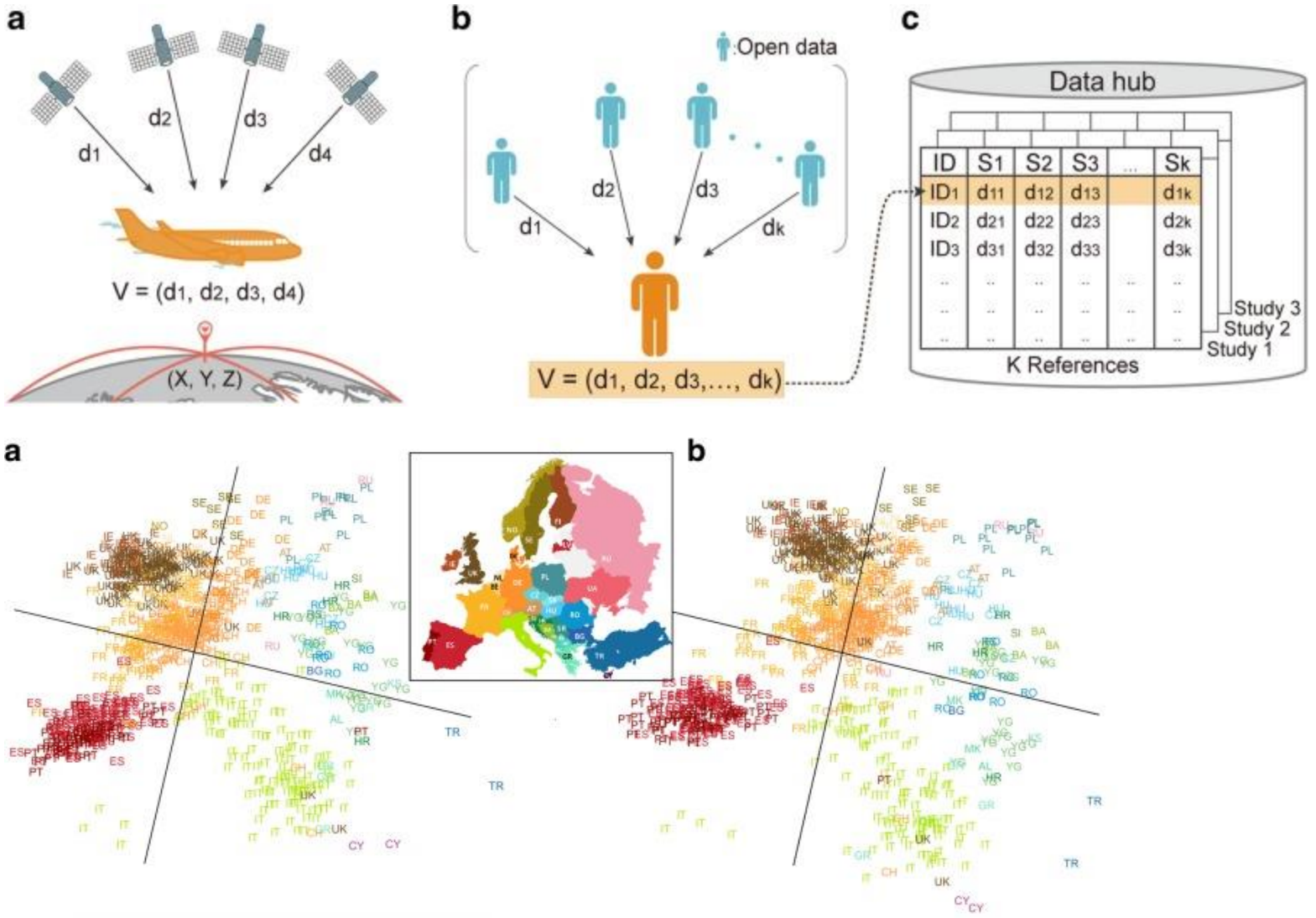


Randi et al (1991)

Phylogenetic trees are built from distance matrix.

Fig. 1 (A) UPGMA dendrogram obtained with Nei's (1978) standard unbiased genetic distances. Time scales according to (a) the lower and (b) the upper divergence time. (B) WAGNER tree computed with Rogers' (1972) genetic distances and rooted using *Bos* as outgroup. c.c. = cophenetic correlation.

Data encryption



- Conversion of genomic markers into distances from reference individuals as a form of data encryption.
- May have potential for use in crop and animal breeding in terms of data sharing.

Kim et al (2019)

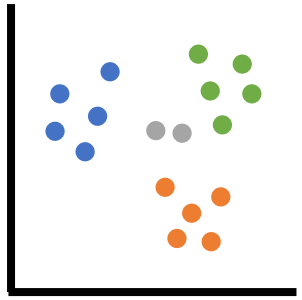
Challenges in working with genetic distance

- Distance choice
- Genotypic data quality (errors, missing data)
- Assumptions in evolution rate, molecular clock, etc
- Genetic marker type: SNP, InDel, SSR, etc
- Statistical significance threshold (e.g. F_{ST})

Population structure

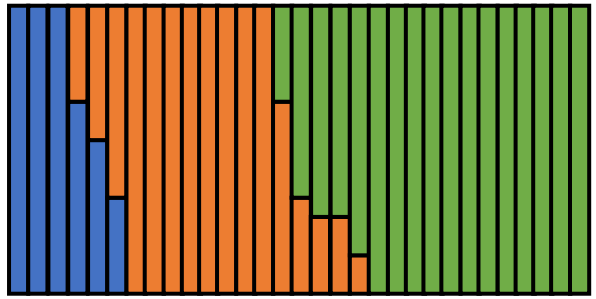
A consequence of genetic divergence between sub-populations.

Principal component analysis (PCA)

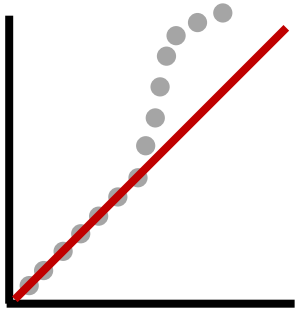


https://knowyourmeme.com/photos/1630278-why-is-it-when-something-happens-its-always-you-three

STRUCTURE



Population structure control



Principal component analysis (PCA)

Reduce dimensionality in data to a form that is easier to understand.

Given a set of genetic marker data, how can we quickly infer the population structure?

	m01	m02	m03	m04	m05	m06	m07	m08	m09	m10
id1	2	0	0	0	0	0	2	0	0	2
id2	2	0	0	0	0	0	0	0	0	2
id3	2	0	0	2	0	0	2	0	0	2
id4	0	0	2	0	2	2	0	2	2	0
id5	0	2	2	0	2	0	0	0	0	0
id6	0	0	2	0	2	0	0	0	0	0

1. Calculate the correlations (e.g. K-matrix/GRM: genetic relationship matrix).

Principal component analysis (PCA)

1a. Calculate the allele frequencies and means.

	m01	m02	m03	m04	m05	m06	m07	m08	m09	m10
p	0.50	0.17	0.50	0.17	0.50	0.17	0.33	0.17	0.17	0.50

$\mu = 2p$

	m01	m02	m03	m04	m05	m06	m07	m08	m09	m10
id1	2	0	0	0	0	0	2	0	0	2
id2	2	0	0	0	0	0	0	0	0	2
id3	2	0	0	2	0	0	2	0	0	2
id4	0	0	2	0	2	2	0	2	2	0
id5	0	2	2	0	2	0	0	0	0	0
id6	0	0	2	0	2	0	0	0	0	0

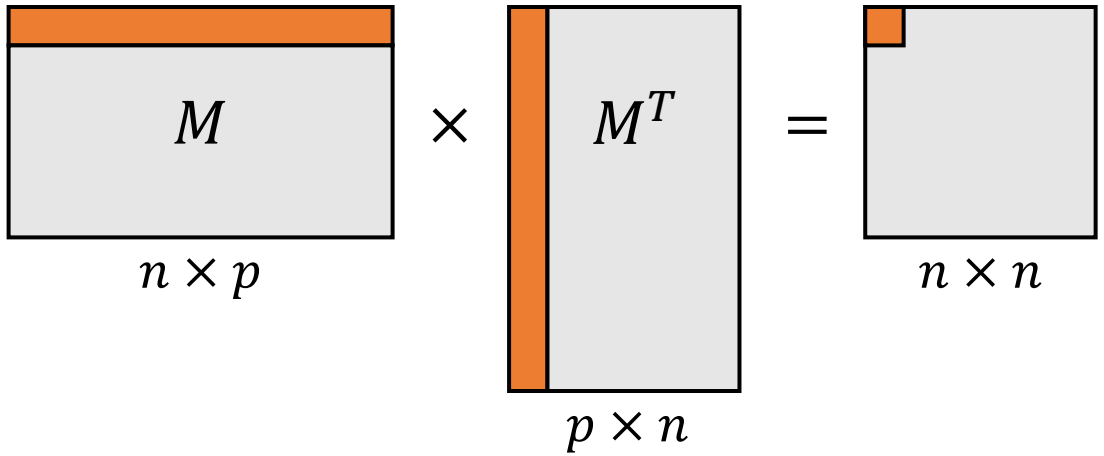
1b. Center the marker data by the means.

	m01	m02	m03	m04	m05	m06	m07	m08	m09	m10
id1	1	-0.33	-1	-0.33	-1	-0.33	1.33	-0.33	-0.33	1
id2	1	-0.33	-1	-0.33	-1	-0.33	-0.67	-0.33	-0.33	1
id3	1	-0.33	-1	1.67	-1	-0.33	1.33	-0.33	-0.33	1
id4	-1	-0.33	1	-0.33	1	1.67	-0.67	1.67	1.67	-1
id5	-1	1.67	1	-0.33	1	-0.33	-0.67	-0.67	-0.33	-1
id6	-1	-0.33	1	-0.33	1	-0.33	-0.67	-0.67	-0.33	-1

Principal component analysis (PCA)

1c. Multiply the matrix by its transpose.

	id1	id2	id3	id4	id5	id6
id1	6.33	3.67	5.67	-6.33	-5.00	-4.33
id2	3.67	5.00	3.00	-5.00	-3.67	-3.00
id3	5.67	3.00	9.00	-7.00	-5.67	-5.00
id4	-6.33	-5.00	-7.00	13.00	2.33	3.00
id5	-5.00	-3.67	-5.67	2.33	7.67	4.33
id6	-4.33	-3.00	-5.00	3.00	4.33	5.00



1d. Divide the matrix by $\sum 2p(1-p)$ [not actually needed for PCA].

	id1	id2	id3	id4	id5	id6
id1	1.65	0.96	1.48	-1.65	-1.30	-1.13
id2	0.96	1.30	0.78	-1.30	-0.96	-0.78
id3	1.48	0.78	2.35	-1.83	-1.48	-1.30
id4	-1.65	-1.30	-1.83	3.39	0.61	0.78
id5	-1.30	-0.96	-1.48	0.61	2.00	1.13
id6	-1.13	-0.78	-1.30	0.78	1.13	1.30

Principal component analysis (PCA)

2. Perform eigen-decomposition ($\boldsymbol{K}v_i = \lambda_i v_i$).

$$v_i = \begin{matrix} -0.42 & -0.06 & -0.08 & -0.14 & 0.79 & -0.41 \\ -0.30 & 0.02 & -0.73 & -0.05 & -0.45 & -0.41 \\ -0.48 & -0.11 & 0.66 & 0.10 & -0.38 & -0.41 \\ 0.52 & -0.74 & 0.01 & -0.13 & -0.01 & -0.41 \\ 0.36 & 0.58 & 0.17 & -0.57 & -0.06 & -0.41 \\ 0.32 & 0.31 & -0.02 & 0.79 & 0.12 & -0.41 \end{matrix}$$

eigenvectors

$$\lambda_i = \begin{matrix} 8.08 \\ 2.21 \\ 0.93 \\ 0.44 \\ 0.34 \\ 0.00 \end{matrix}$$

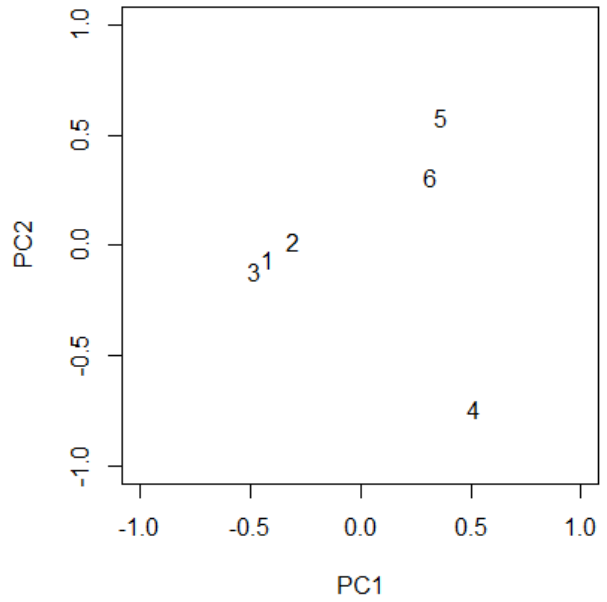
eigenvalues

3. Plot the eigenvectors (principal components).

% variation explained ($PVE = \frac{\lambda_i}{\sum \lambda_i} \cdot 100$).

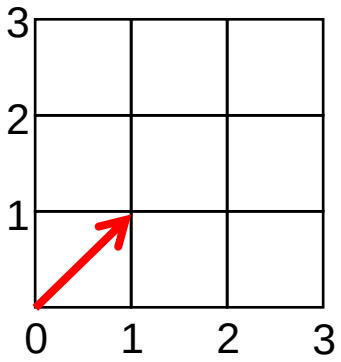
PC1 explains 67% variation;

PC2 explains 18% variation.



Eigenvectors and eigenvalues

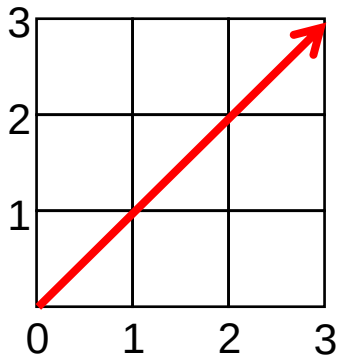
Remember from physics classes, a vector has direction and magnitude.



Vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is represented by the red arrow.

Let's try to transform the vector by $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$.

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$



Notice that $\begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, so the transformation changed the magnitude by 3 but not the direction.

In that case, we can call $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ as the eigenvector and 3 as the corresponding eigenvalue.

Eigenvectors and eigenvalues

Dall E (2024)

Think of v_i as a slice of $\boldsymbol{K}$,

and λ_i is the measure of

how big the slice is.

Interpreting PCA plot

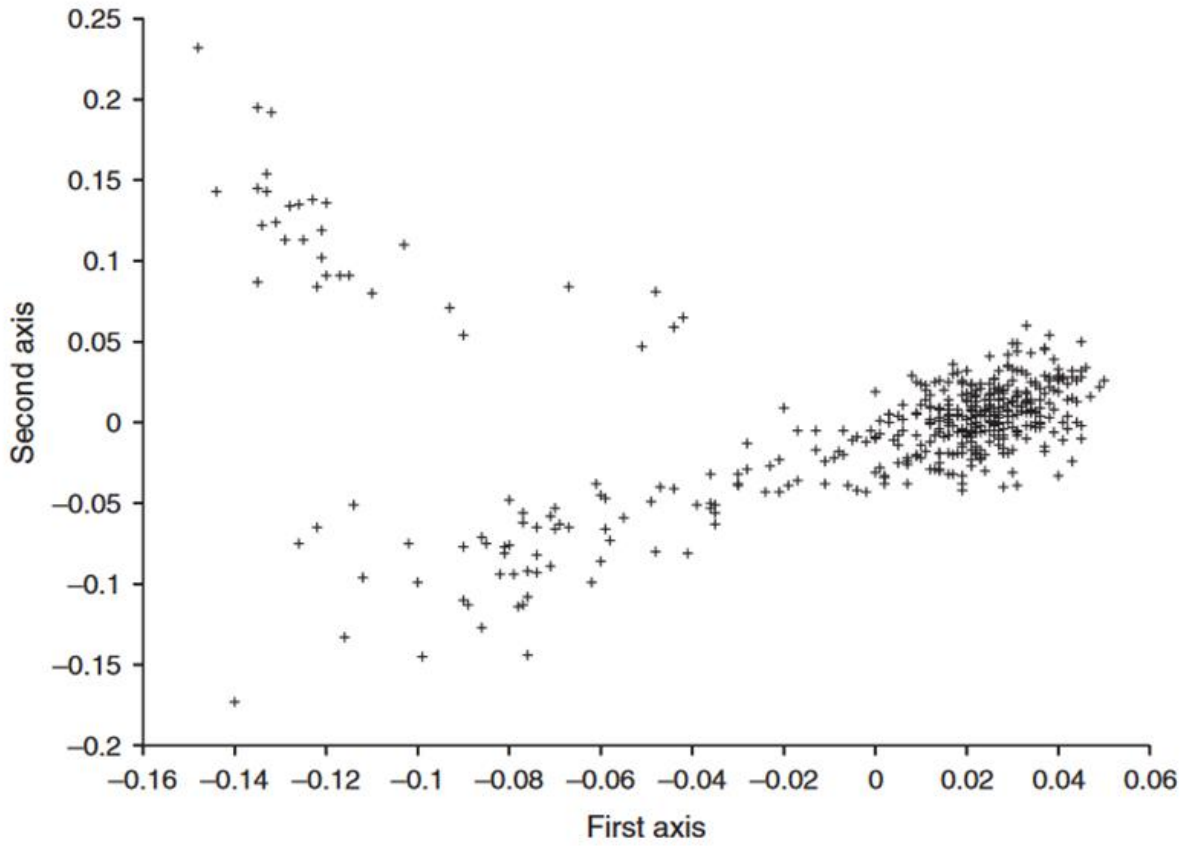
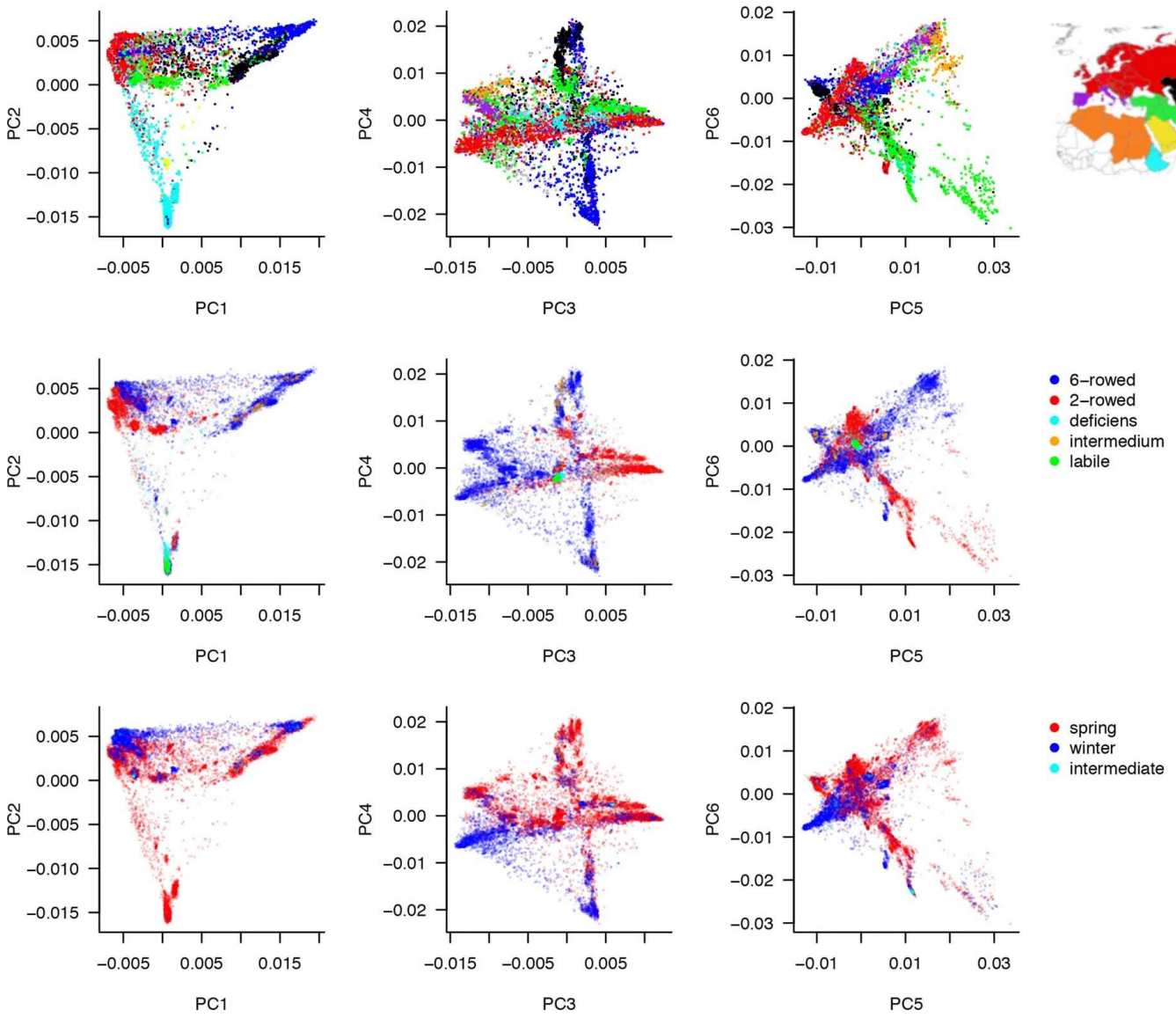


Figure 2 The top two axes of variation of European American samples. We hypothesize that the first axis reflects genetic variation between northwest and southeast Europe, with a fraction of the samples showing southeast European ancestry (first axis < 0; see text). It follows that the second axis separates two southeast European subpopulations.

Price et al (2006)

- Clear separation on the first axis (PC1).
- Finer-scale separation on PC2.

Interpreting PCA plots



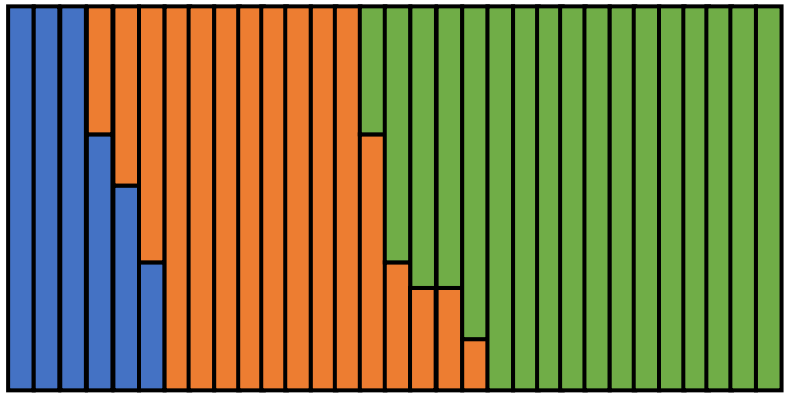
Jayakodi et al (2020)

Extended Data Figure 1. PCA with genotyping-by-sequencing data of 19,778 varieties of domesticated barley sampled from the gene bank of the IPK9. The first six principal components are shown. Samples are colored according to geographic origin, row type or annual growth habit.

STRUCTURE et al

- These are software to model population admixture.
- Typically uses Bayesian clustering methods ($Posterior = Prior \times Observation$).
- The model takes a prior (K clusters of populations), adjusts it according to observation (allele frequencies) over many iterations, and produces the posterior (ancestry admixture estimate).

$K = 3$

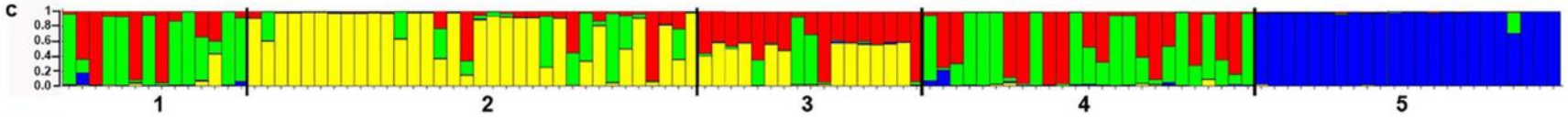


STRUCTURE in action: Welwitschia

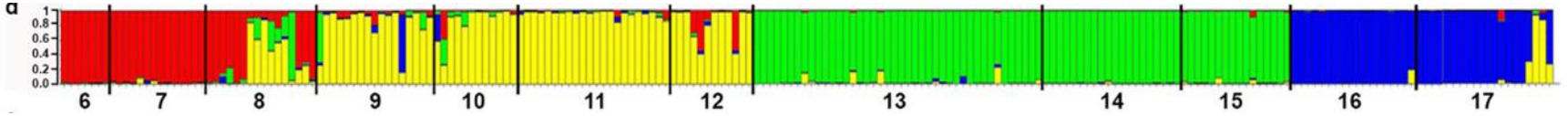
powo.science.kew.org/taxon/urn%3Alsid%3Aipni.org%3 Anames%3A383591-1

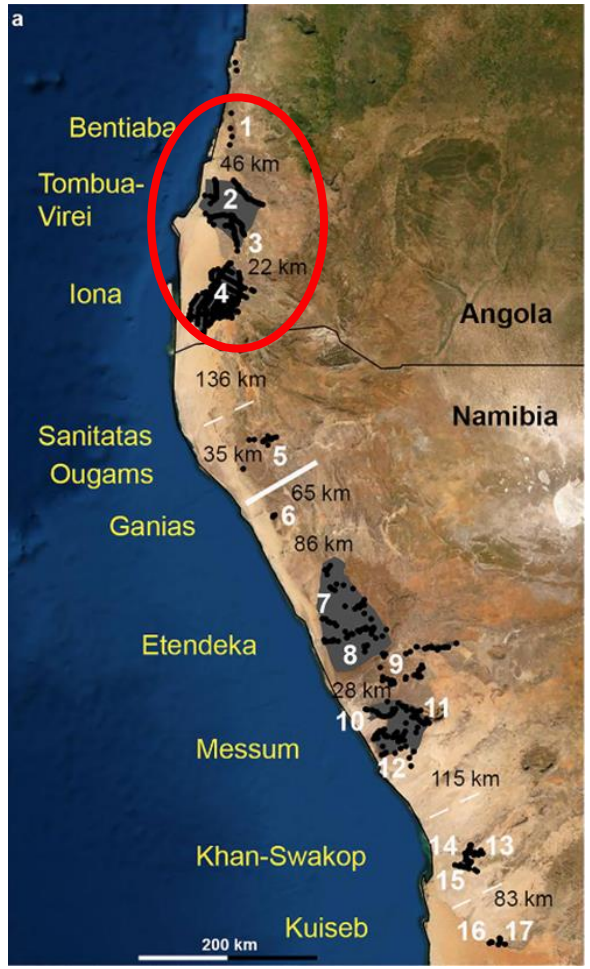
Both ssp, $K = 2$

ssp mirabilis, $K = 4$



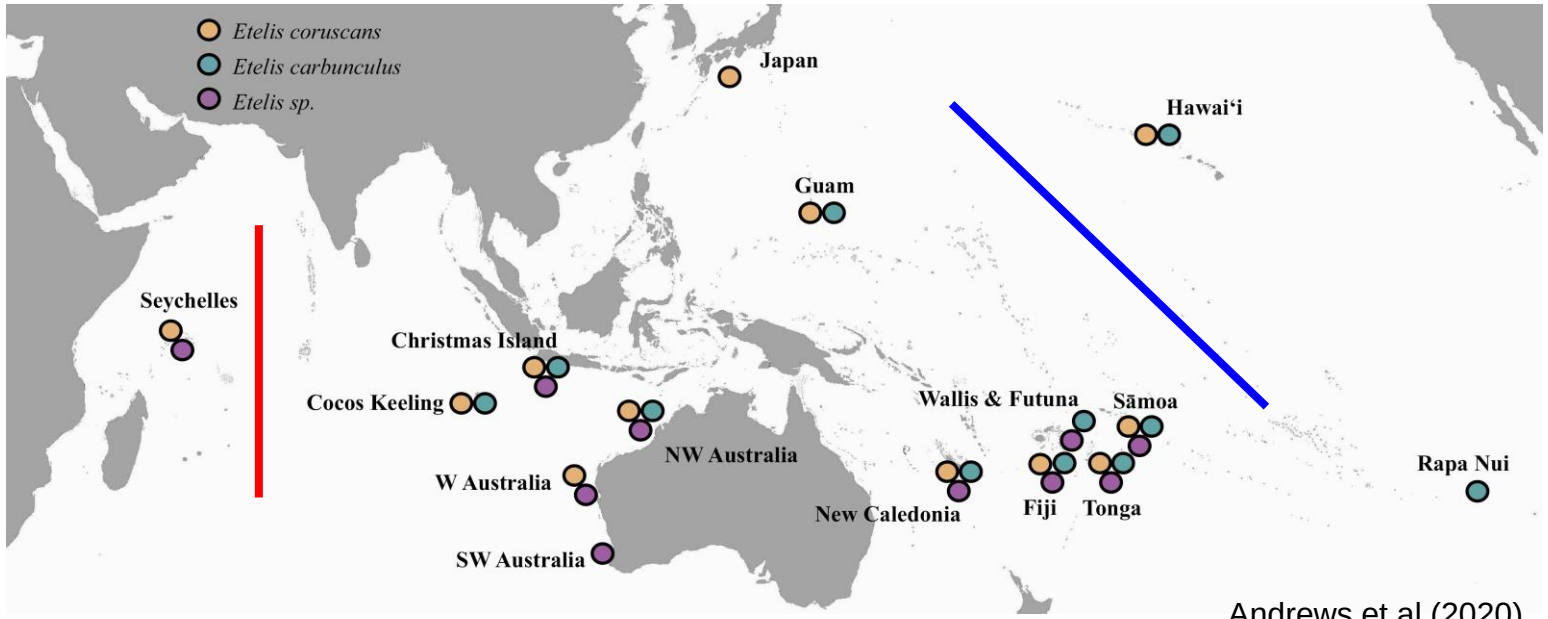
ssp namibiana, $K = 4$



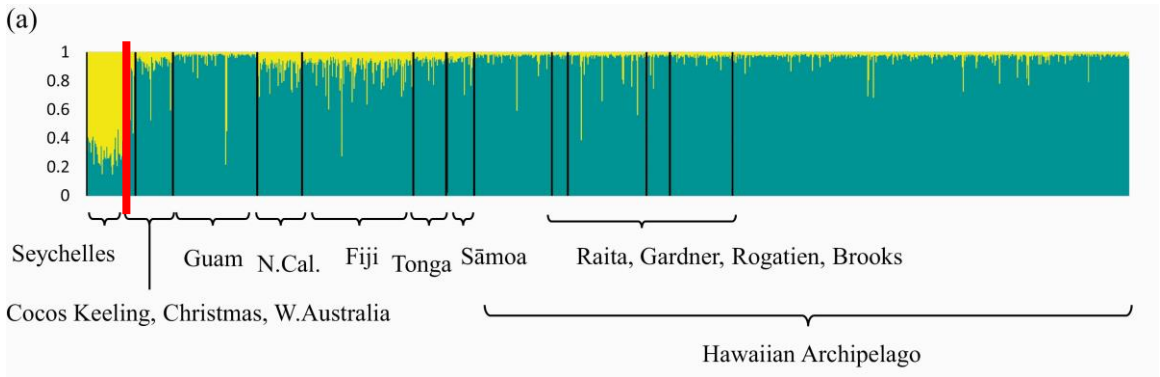


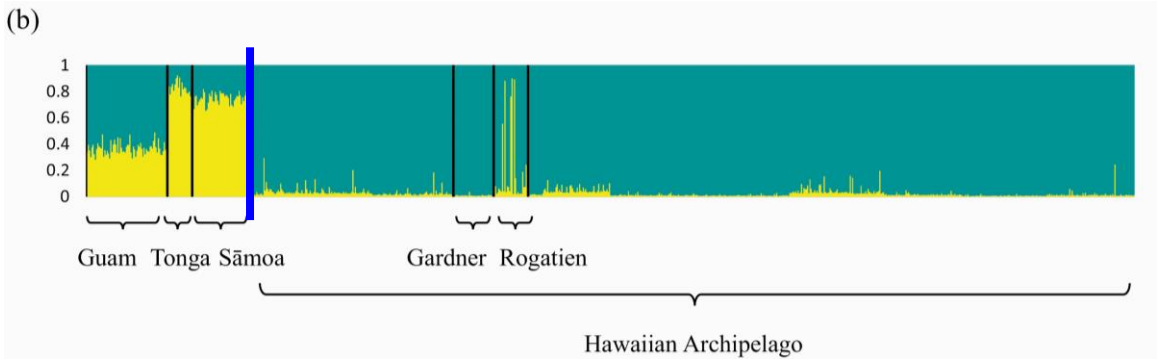
Jurgens et al (2021)

STRUCTURE in action: Eteline snapper



Andrews et al (2020)





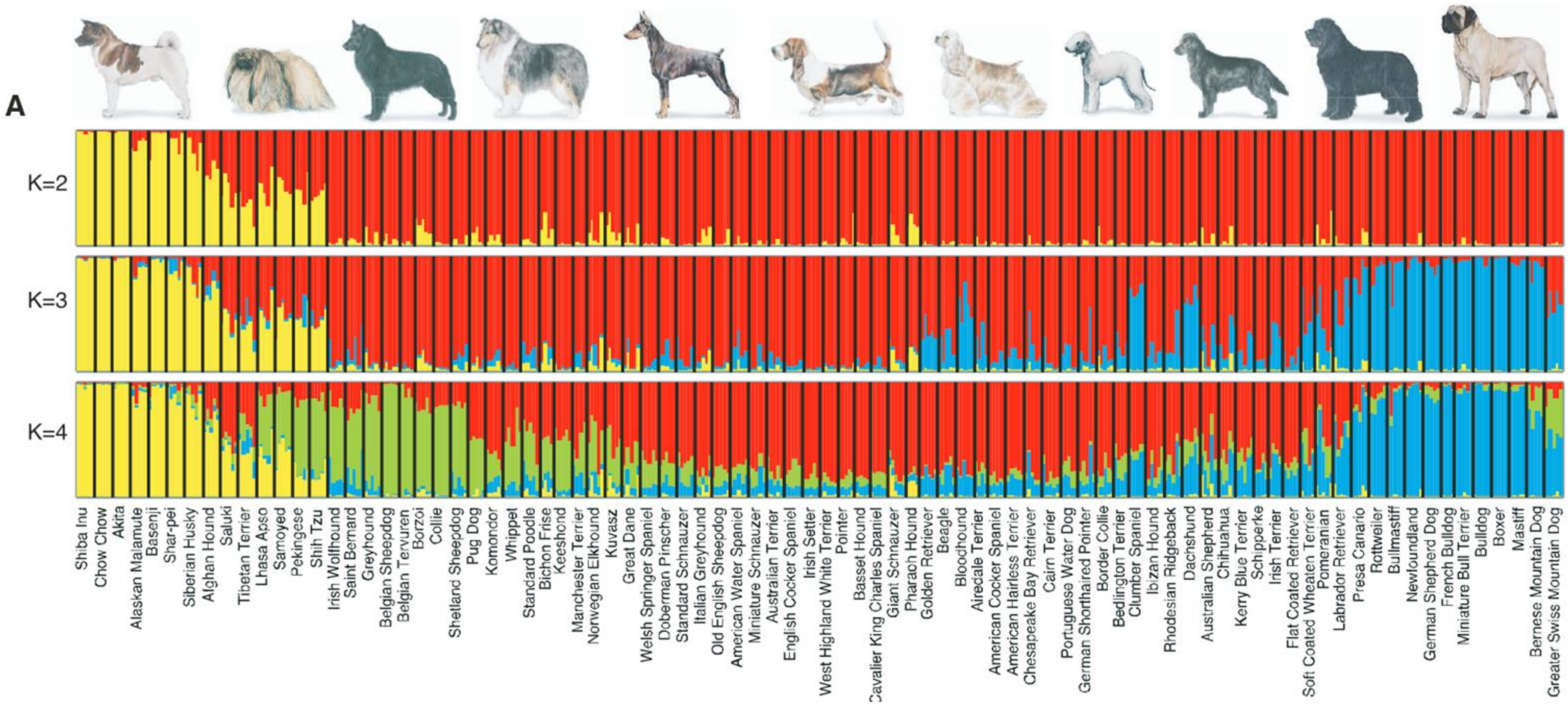
E coruscans, $K = 2$

www.fishbase.se/summary/1385

E carbunculus, $K = 2$

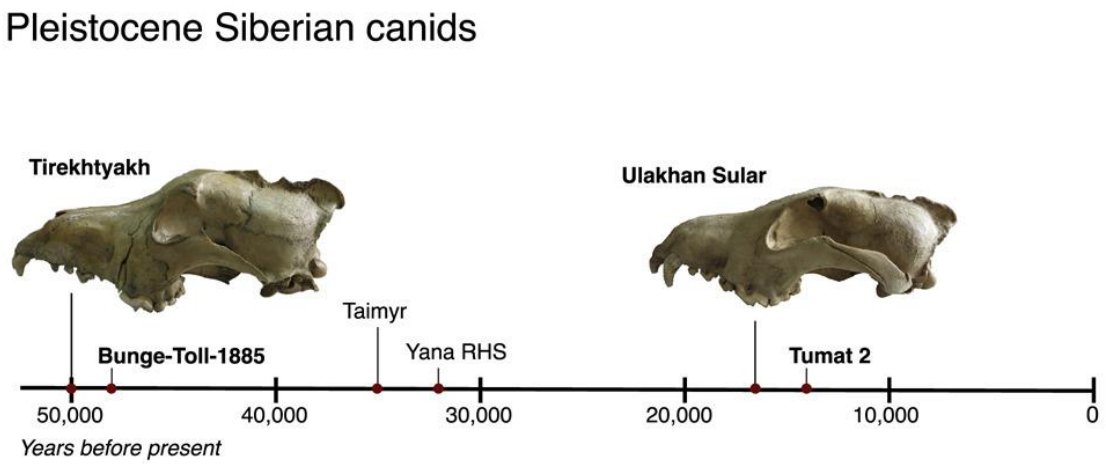
www.fishbase.se/summary/Etelis-carbunculus

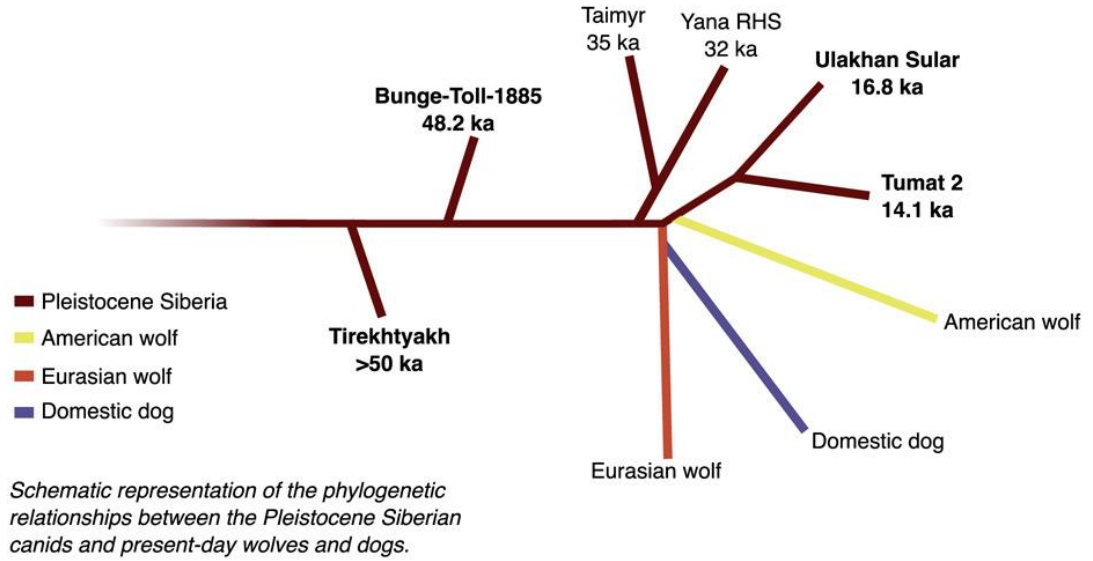
STRUCTURE in action: canine



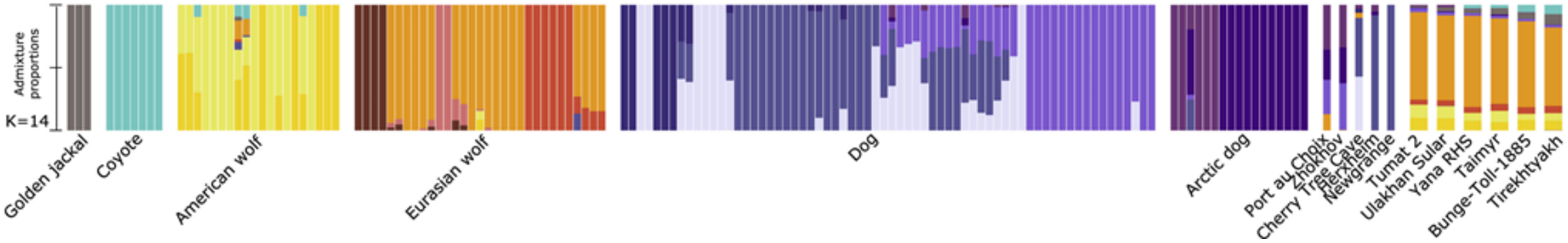
Parker et al (2004)

ADMIXTURE in action: canine





Schematic representation of the phylogenetic relationships between the Pleistocene Siberian canids and present-day wolves and dogs.



Ramos-Madrigal et al (2021)

PCA + STRUCTURE + Phylogenetics in action: Wheat

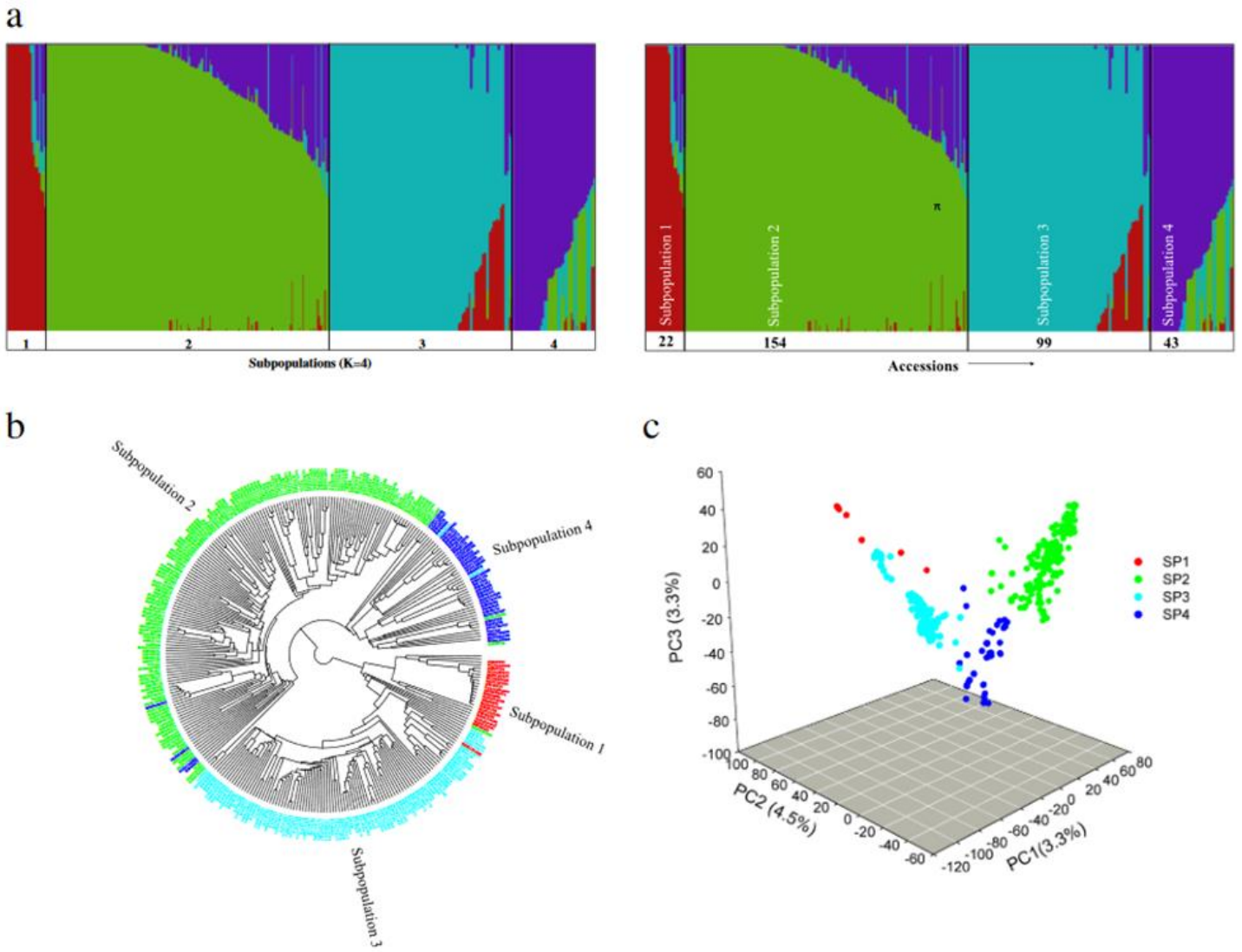
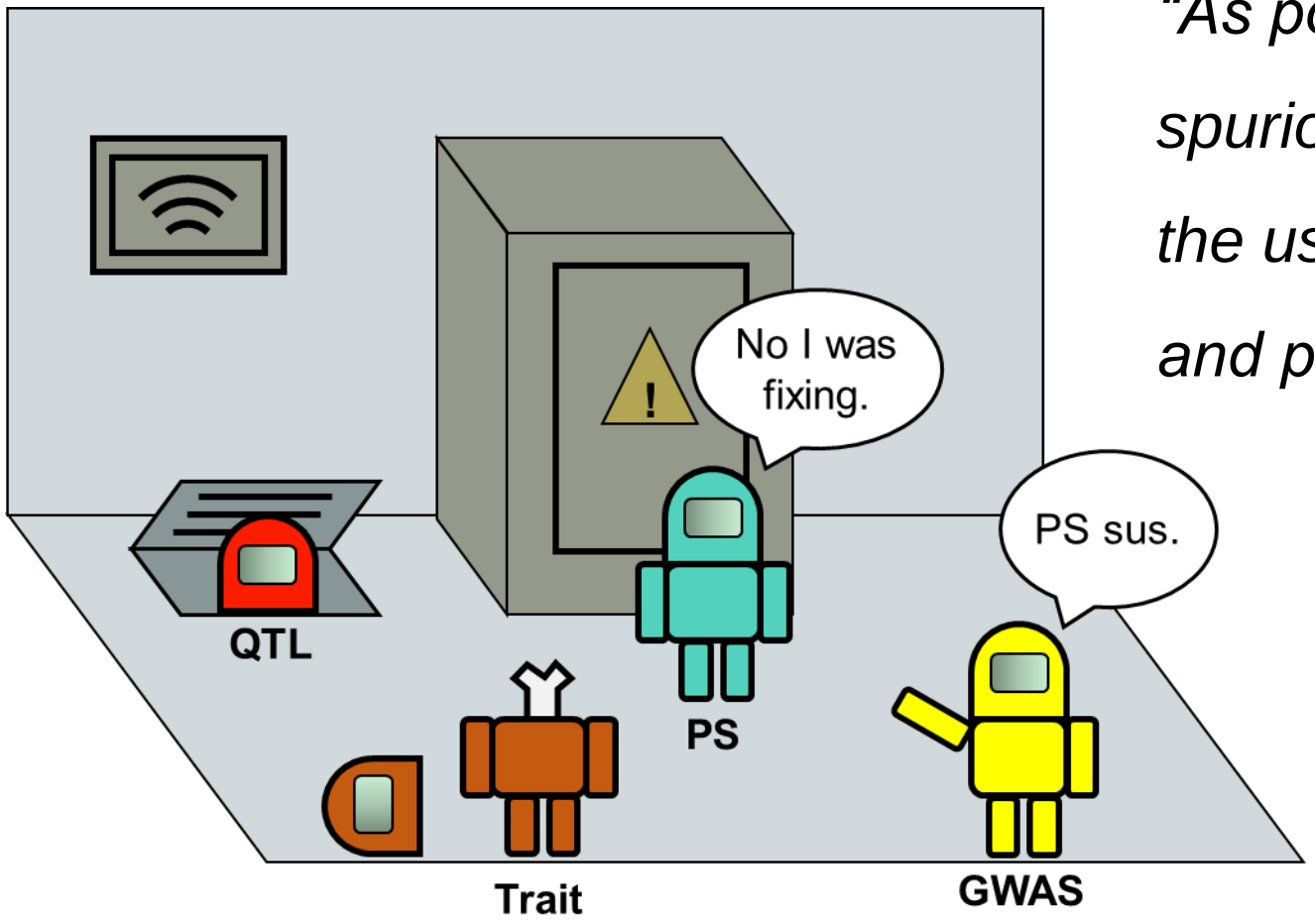


Fig. 2 Population structure analysis of the NWDP. **a** Estimated population structure of 318 spring wheat genotypes from Nepal on K = 4. Columns represent individual wheat accessions, while the length represents the proportion of each subpopulation (indicated by the colour) belonging to that accession. **b** Dendrogram based on cluster analysis using pairwise genetic distances. **c** Principal component analysis (PCA) using 95 K GBS markers. The labels SP1, SP1, SP3 and SP4 correspond to the subpopulations 1, 2, 3 and 4

Khadka et al (2020)

www.teagasc.ie/crops/soil--soil-fertility/crop-n-p-k-advice/spring-cereals/spring-wheat/

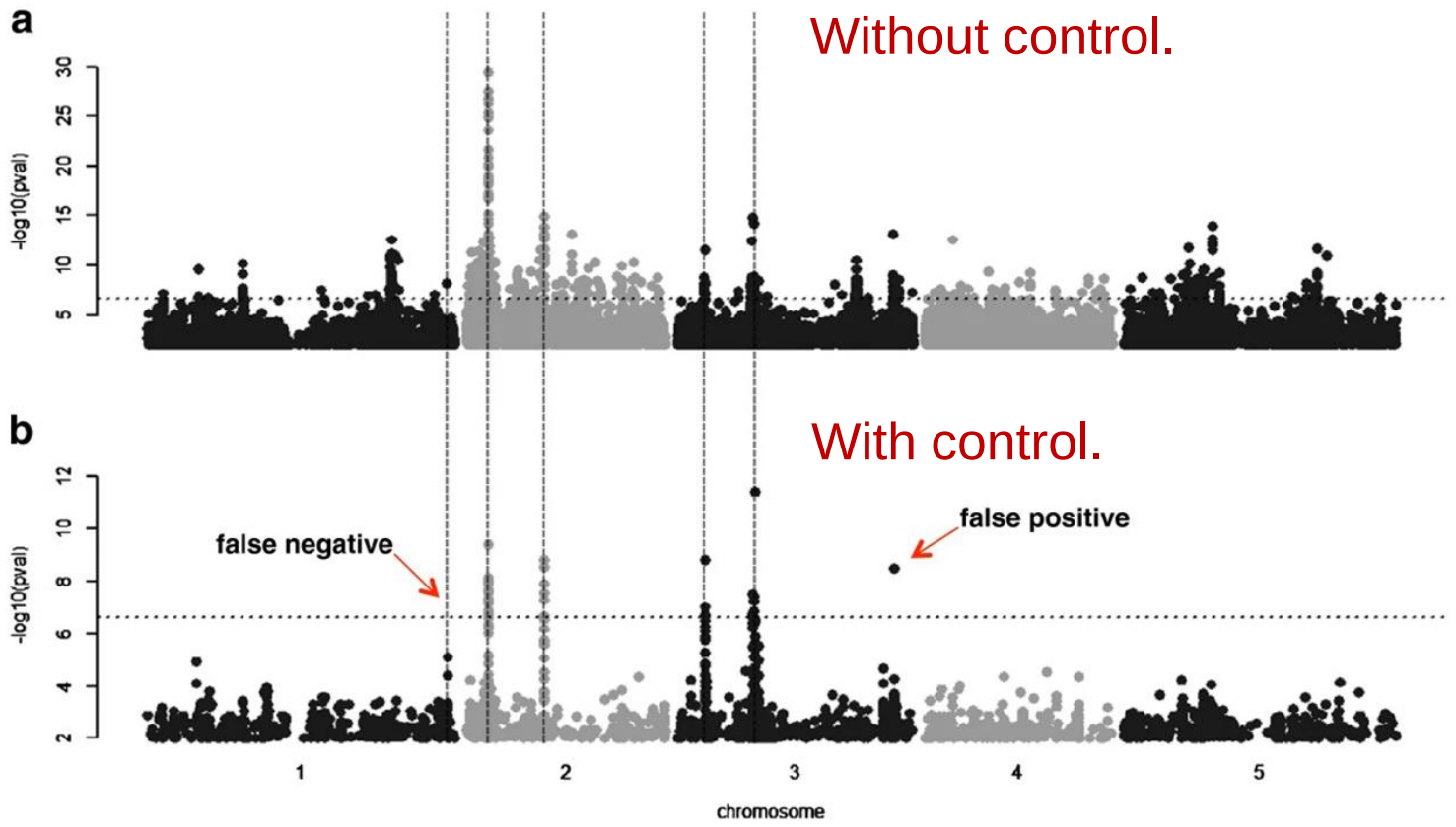
Population structure and spurious associations



“As population structure can result in spurious associations, it has constrained the use of association studies in human and plant genetics”.

Yu et al (2006)

Population structure control



Taking genetic background into account improves the performance of GWAS. Manhattan plots for a simulated trait, in which each data point represents a genotyped SNP, ordered across the five chromosomes of *Arabidopsis*. Five SNPs (indicated by vertical dashed lines) were randomly chosen to be 'causative' and account for up to 10% of the phenotypic variance each. GWAS using **a)** a linear model, and **b)** a mixed model that accounts for population structure and other background genomic factors. The simple linear model leads to heavily inflated p-values and the five causative markers are not the strongest associations. The mixed model is superior, but still leads to one false negative and one false positive. A dashed horizontal line denotes the 5% Bonferroni threshold.

Korte and Farlow (2013)

- Commonly used in GWAS.
- Many spurious results without control.
- Control is not perfect.

How to control for population structure?

Fix it

1. Structured association, e.g. MLM Q + K.
2. Genomic/delta control (GC/DC).
3. Stepwise regression.
4. Machine learning.
5. Functional validation.
6. Replicated studies.

Avoid it

1. Family-based, e.g. transmission disequilibrium test (TDT).
2. Mapping population design.

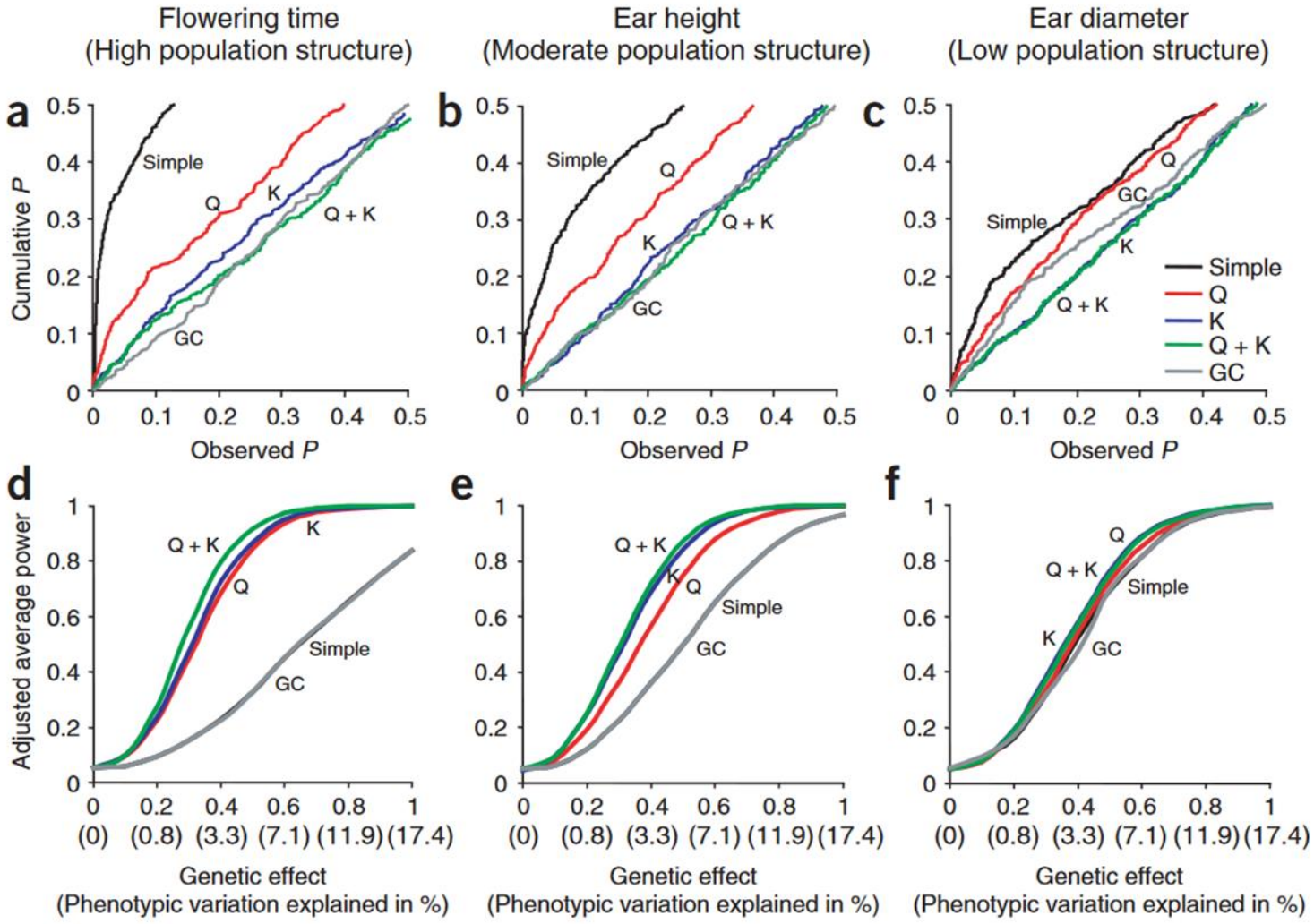
Mixed linear model, Q + K

Here is an “incorrect” GWAS model to illustrate the point.

$$y = \mu + marker + Q + K + e$$

Q = principal components

K = GRM.



Yu et al (2006)

Genomic control (GC)

$$\chi^2_{GC} = \frac{\chi^2_{marker}}{\lambda}$$

Note: $\lambda = \frac{median\left(\chi^2_{bg}\right)}{0.456}$ and for large n, $F_{df=1,n} \approx \chi^2_{df=1}$

Marker	n	Effect	SE	T-stat	F-stat	p
m01	100	1.5	0.5	3	9	0.003
m02	100	1.0	0.5	2	4	0.048
m03	100	0.8	0.4	2	4	0.048
m04	100	1.0	1.0	1	1	0.320
m05	100	0.5	0.5	1	1	0.320
m06	100	0.6	0.3	2	4	0.048

Using m03 to m06 as the background markers, $\lambda = \frac{2.5}{0.456} = 5.482.$

With GC, then:

$$F_{GC,m01} = \frac{9}{5.482} = 1.642 \quad p_{GC,m01} = 0.203$$

$$F_{GC,m02} = \frac{4}{5.482} = 0.730 \quad p_{GC,m02} = 0.395$$

Delta control (DC)

$$\chi^2_{GC} = \chi^2_{marker} - \delta^2$$

Note: $\delta^2 \approx mean(\chi^2_{bg})$ and for large n, $F_{df=1,n} \approx \chi^2_{df=1}$

Marker	n	Effect	SE	T-stat	F-stat	p
m01	100	1.5	0.5	3	9	0.003
m02	100	1.0	0.5	2	4	0.048
m03	100	0.8	0.4	2	4	0.048
m04	100	1.0	1.0	1	1	0.320
m05	100	0.5	0.5	1	1	0.320
m06	100	0.6	0.3	2	4	0.048

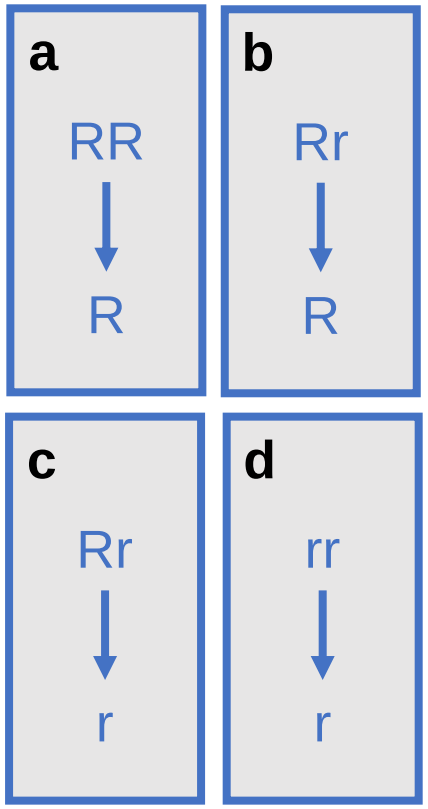
Using m03 to m06 as the background markers, $\delta^2 = \frac{4+1+1+4}{4} = 2.5$.

With DC, then:

$F_{DC,m01} = 9 - 2.5 = 6.5 \quad p_{DC,m01} = 0.012$

$F_{DC,m02} = 4 - 2.5 = 1.5 \quad p_{DC,m02} = 0.224$

Transmission disequilibrium test (TDT)



- Identify n individuals (progeny) with disease (simple trait).
- Genotype n progeny and $2n$ parents.
- For each parent-progeny pair, classify the allele transmission.

	parent _R	parent _r
parent R_, transmits R	**a**	**b**
parent r_, transmits r	**c**	**d**

$$\chi^2_{df=1} = \frac{(b-c)^2}{b+c}$$

$$a + b + c + d = 2n$$

- Homozygous parents are uninformative, guaranteed to transmit same allele.
- Heterozygous parents can transmit either allele, $b \approx c$ if the marker is not linked to the causative locus.

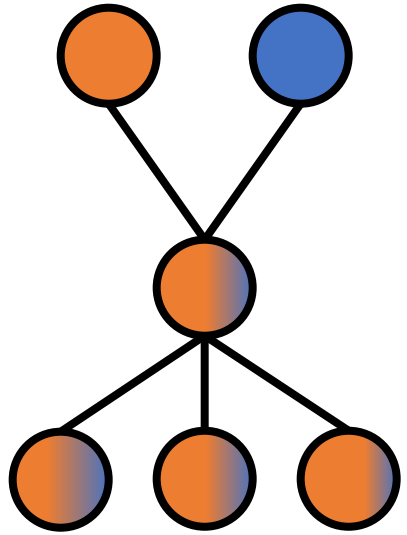
Mapping population design

Populations under random mating

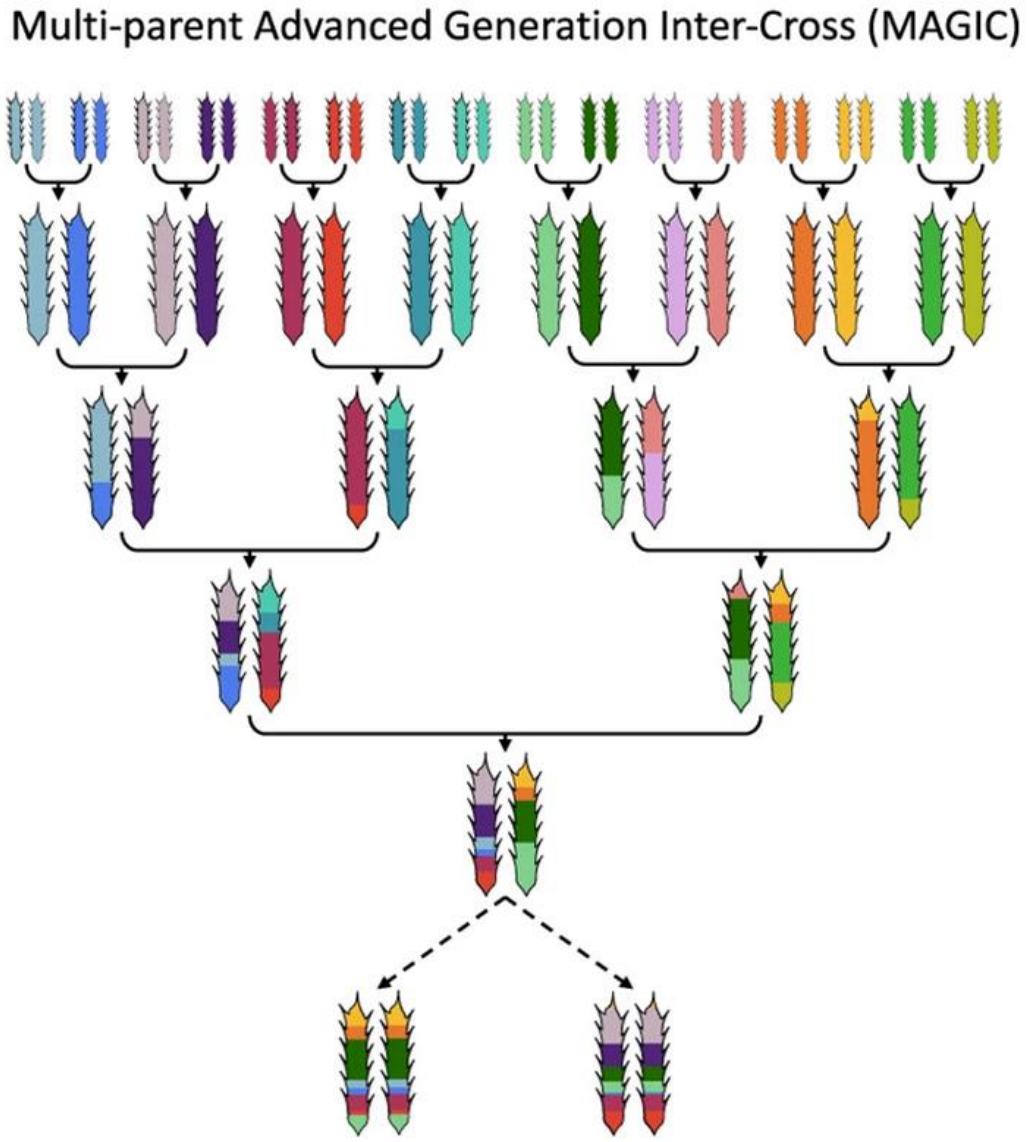
- Uncommon.
- Require us to have already studied the population structure.

Bi-parental populations

- No population structure to worry about.
- Recombinations between the two parental genomes.
- QTL linkage mapping.
- Limited allelic diversity.



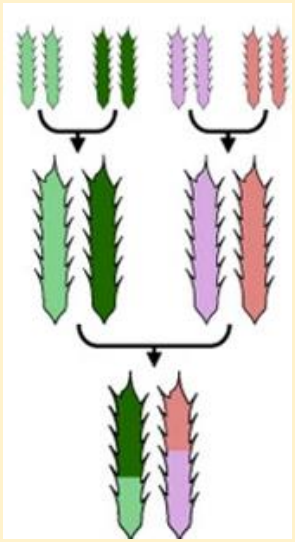
Mapping population design



Scott et al (2020)

- Typically involves $n = 2^x$ founders, e.g. 4, 8, 16.
- The founders are crossed at equal probabilities.
- Minimal population structure.

An example with 4 founders.



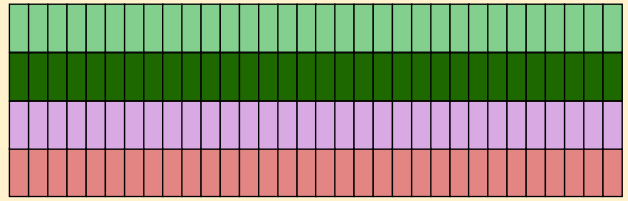
All possible combinations:

(1 x 2) x (3 x 4)

(1 x 3) x (2 x 4)

(1 x 4) x (2 x 3)

Admixture analysis example:



MAGIC population in wheat with 16 founders

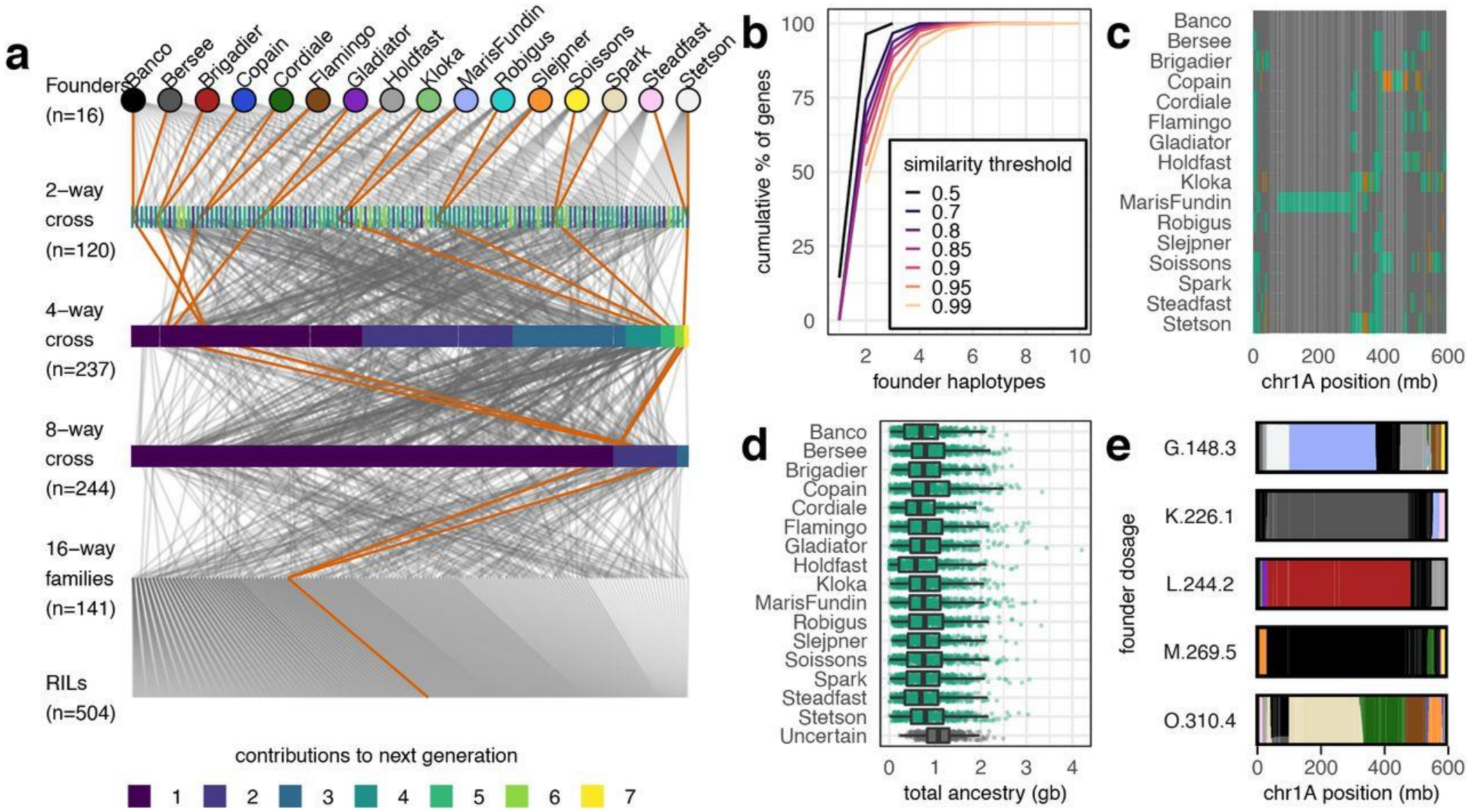


Figure 2. NDM population design and haplotypic diversity. (a) Pedigree showing the construction of 504 Recombinant Inbred Lines (RILs). One exemplar pedigree is highlighted to show how all 16 founders are intercrossed into each RIL. (b) Founder haplotype groups at 73,982 promoter-gene loci with SNP variation, where founders with the same haplotype have genotypic similarity fractions that exceed the corresponding threshold. (c) Pairwise similarity/dissimilarity between founders on chromosome 1A, determined using a dynamic programming algorithm to infer founder similarity and breakpoint position. Founders that are inferred to have similar haplotypes for each region are the same colour. (d) The total length of genomic blocks in NDM lines inferred to come from each founder; uncertain ancestry blocks have a maximum founder dosage of <90%. (e) Inferred founder dosage and ancestry mosaics across chromosome 1A for five example RILs, with founders colored as in (a).

Scott et al (2021)

Applications related to population structure

1. Association mapping (discussed previously).
2. Pan-genome assembly.
3. Selection mapping.
4. Precision medicine.

Population structure: need for a pan-genome

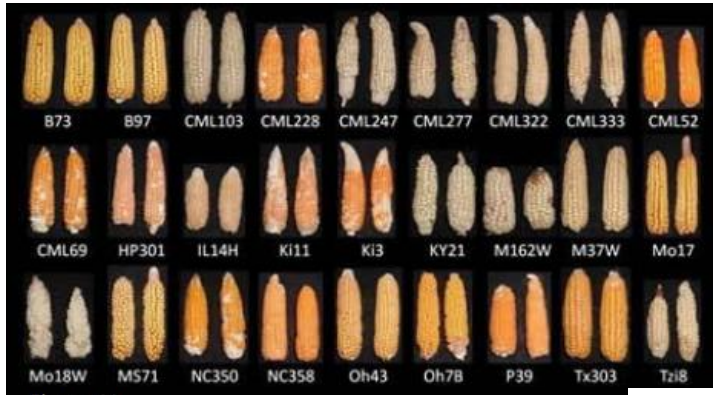
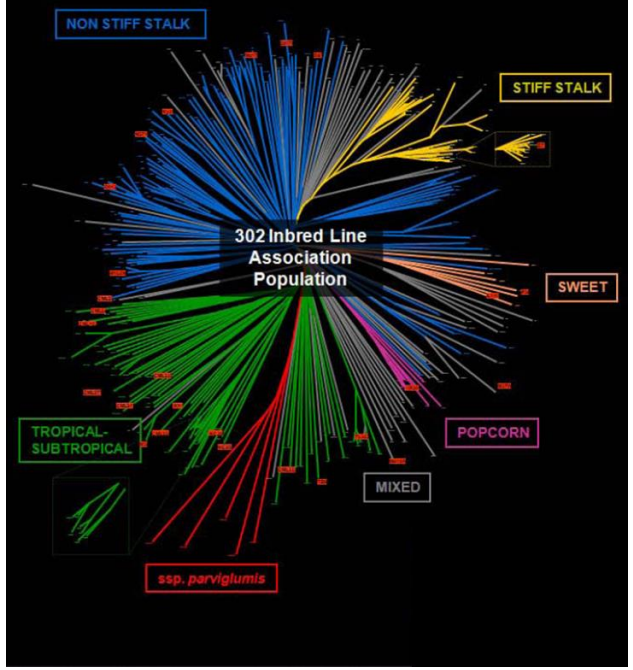
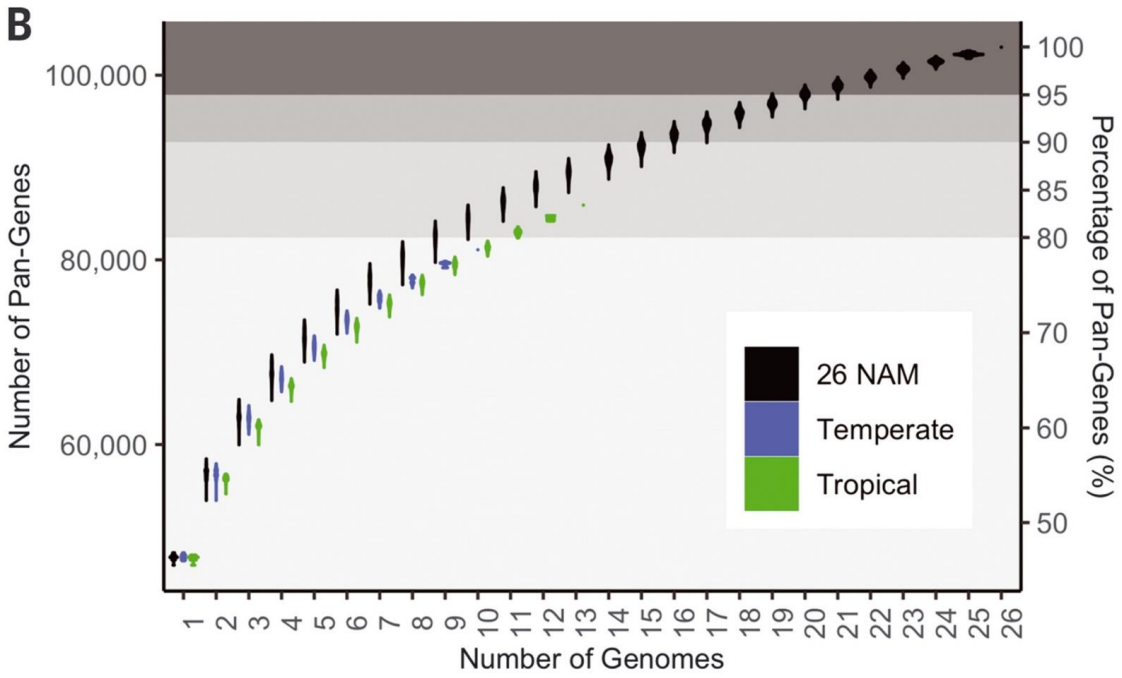


Figure S2.



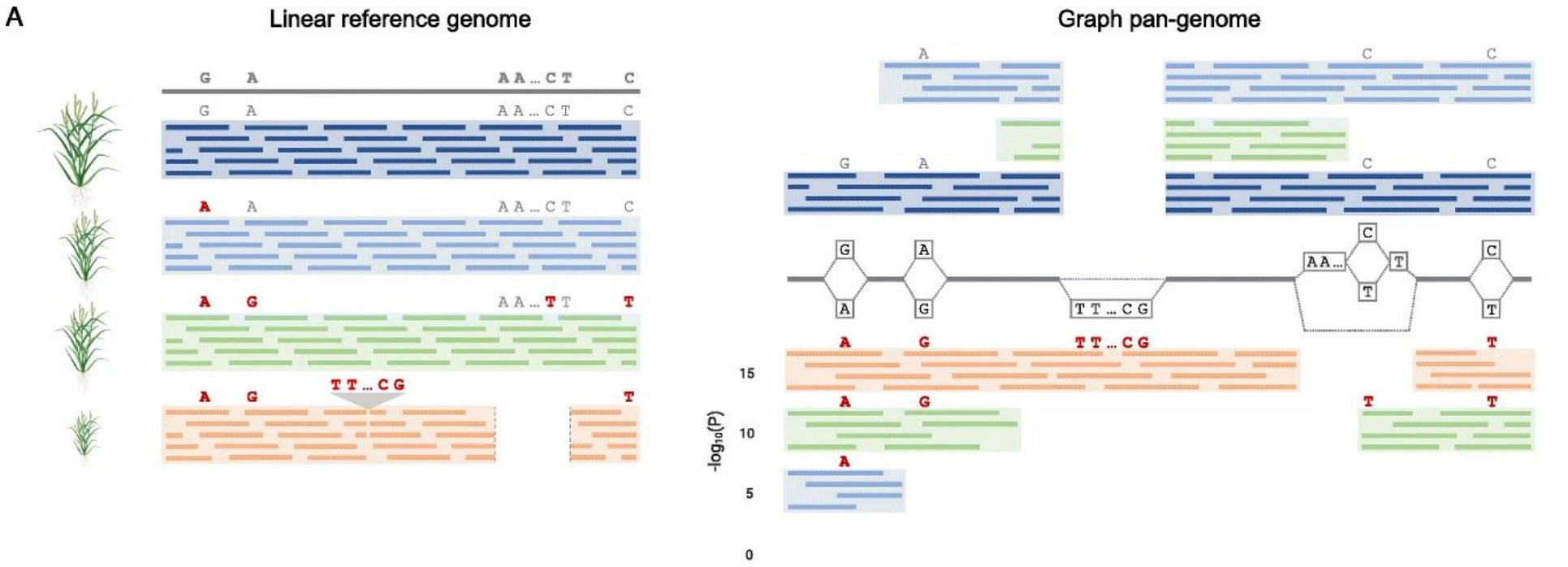
McMullen et al (2009)

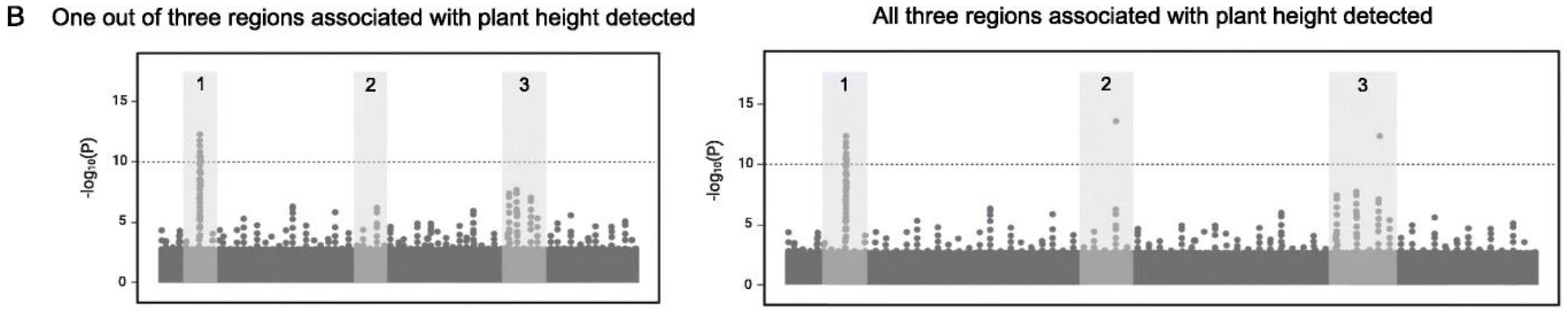


Hufford et al (2021)

- High level of presence/absence variants.
- Bias from alignment to a single reference genome.

One of many benefits of pan-genome





Improvement to mapping power.

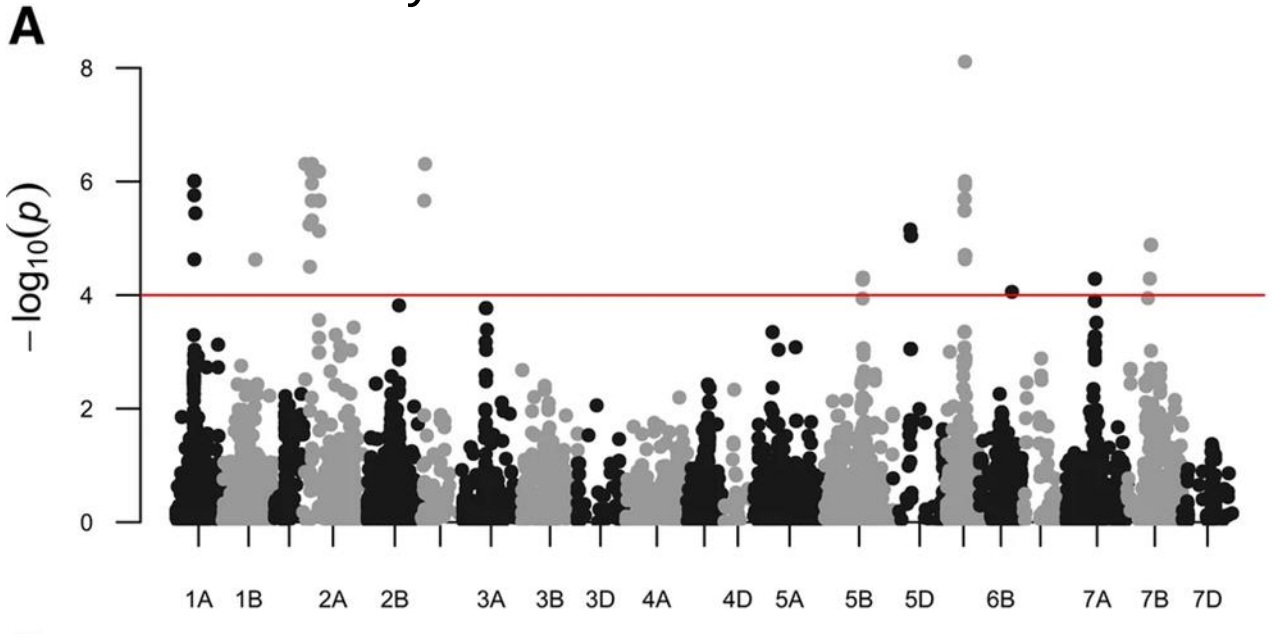
Coletta et al (2021)

Selection mapping

Selection/breeding results in population structure.

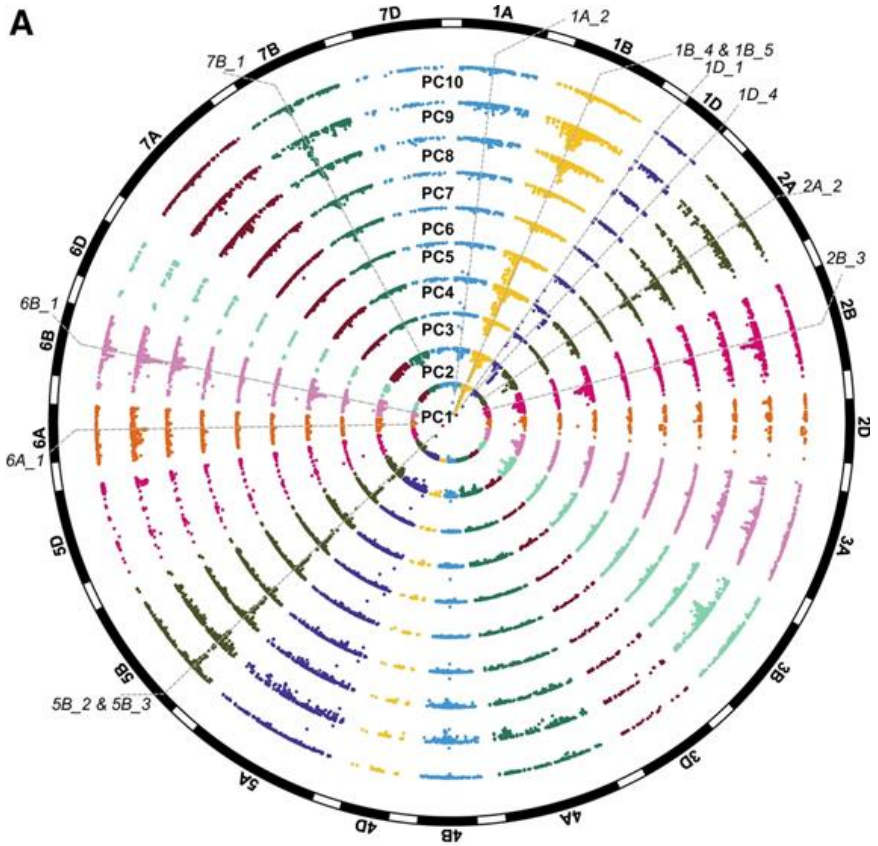
Selection evidence in winter wheat can be identified using env/eigen-GWAS approaches .

Year of variety release as trait



Sharma et al (2021)

PC as trait



Precision medicine (as explained in plants)

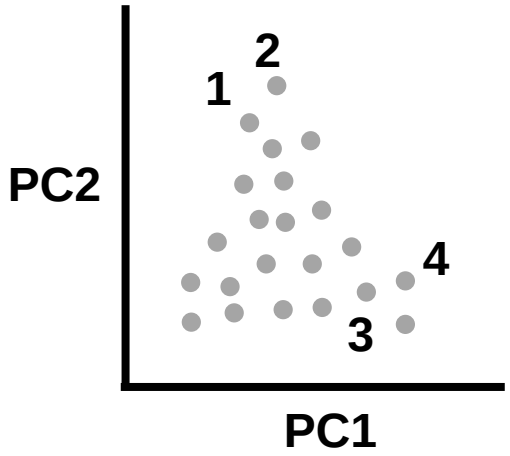
Pop 1

Pop 2

Pop 3

Pop 4

- Assume the causative locus has not been identified.
- Case 1: find an association at the left marker in Pop 1 > this marker gets it wrong in Pop 3 and 4.
- Case 2: find an association at the right marker in Pop 3 > this marker gets it wrong in Pop 1 and 2.



If we know the population structure, then we know which marker is more appropriate for the populations.

Challenges in working with population structure

- Population structure vs true positive
- What is the right K?
- Genotyping data quality (e.g. partially solved by pan-genome).
- Small sample size
- Statistical methods

Summary

Background. Why do we to study genetic relationship?

IBD. Definition: two alleles that have originated from the replication of one single allele in a previous generation ♦ coefficient of coancestry and fraternity – methods for calculating them.

IBS. Definition: two of the same alleles regardless of their ancestral origins ♦ IBS-based coefficients.

Genetic distance. Definition: a measure of genetic divergence between populations or sub-populations ♦ how do sub-populations diverge? ♦ methods for computing genetic distance.

Population structure. Definition: a consequence of genetic divergence between sub-populations ♦ common topics like PCA, admixture analysis, control approaches.

Examples

Applications

Challenges

Resources

Methods for calculating coefficient of coancestry

https://doi.org/10.1007/978-3-030-83940-6_11

https://www.genetic-genealogy.co.uk/supp/calc_inbreeding_coan.html

Relevant references

GRM: vanRaden (2008) Efficient methods to compute genomic predictions.

MLM Q + K: Yu et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.

GC: Devlin and Roeder (1999) Genomic control for association studies.

DC: Gorroochurn et al (2011) An improved delta-centralization method for population stratification.

TDT: Spielman et al (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).

MAGIC: Cavanagh et al (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants.

Free online resources

https://github.com/cooplab/popgen-notes

https://felsenst.github.io/pgbook/pgbook.pdf

https://excellenceinbreeding.org/toolbox

Paid resources

Walsh and Lynch (2018) Evolution and selection of quantitative traits

Falconer and Mackay (1995) Introduction to quantitative genetics

Contact me

cyang@sruc.ac.uk

cjyang-work.github.io

@hataraku_cj