ETH zürich

Data science and statistics in molecular plant breeding

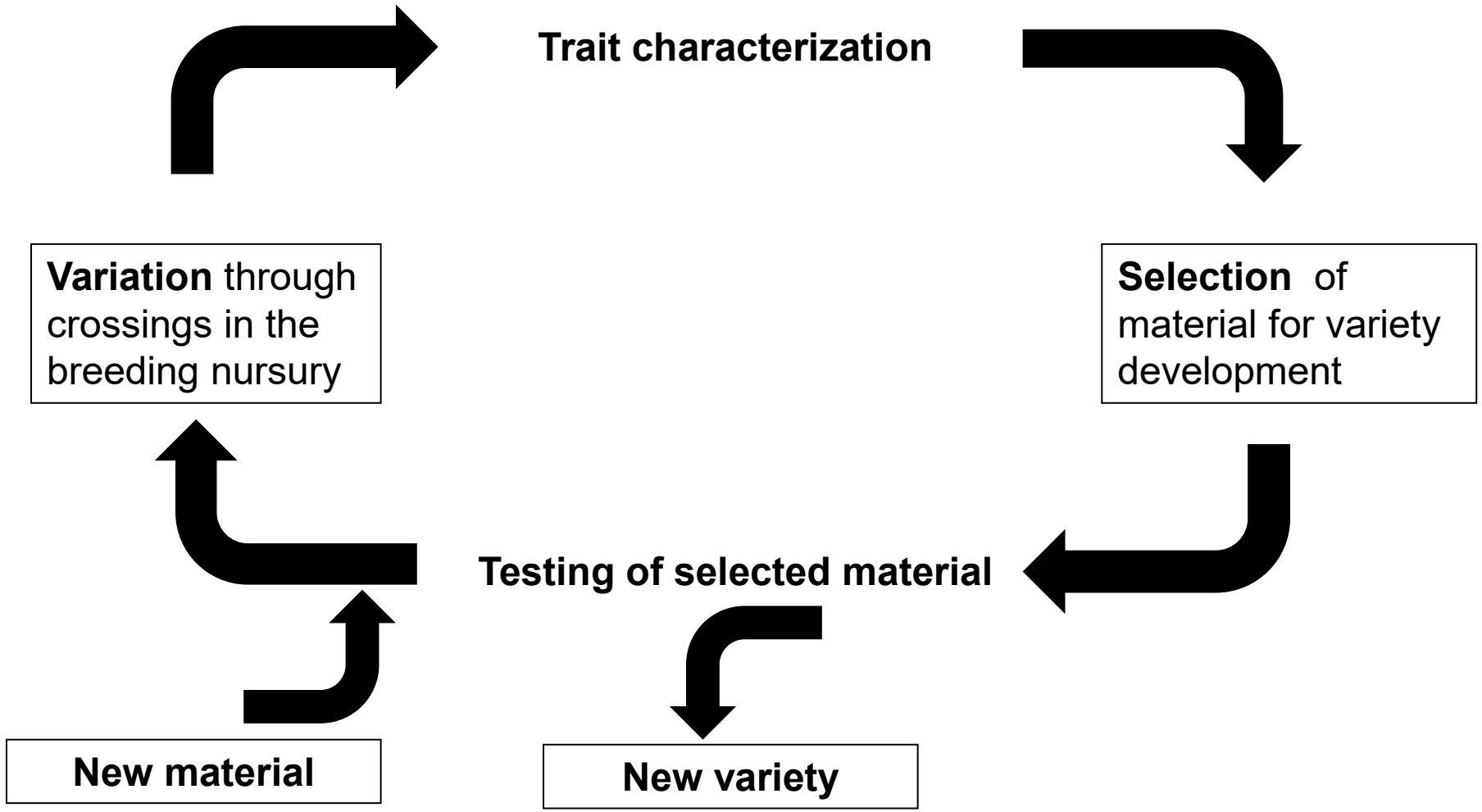
Roland Kölliker

16.04.24 LegumeLegacy Event 3, Zurich

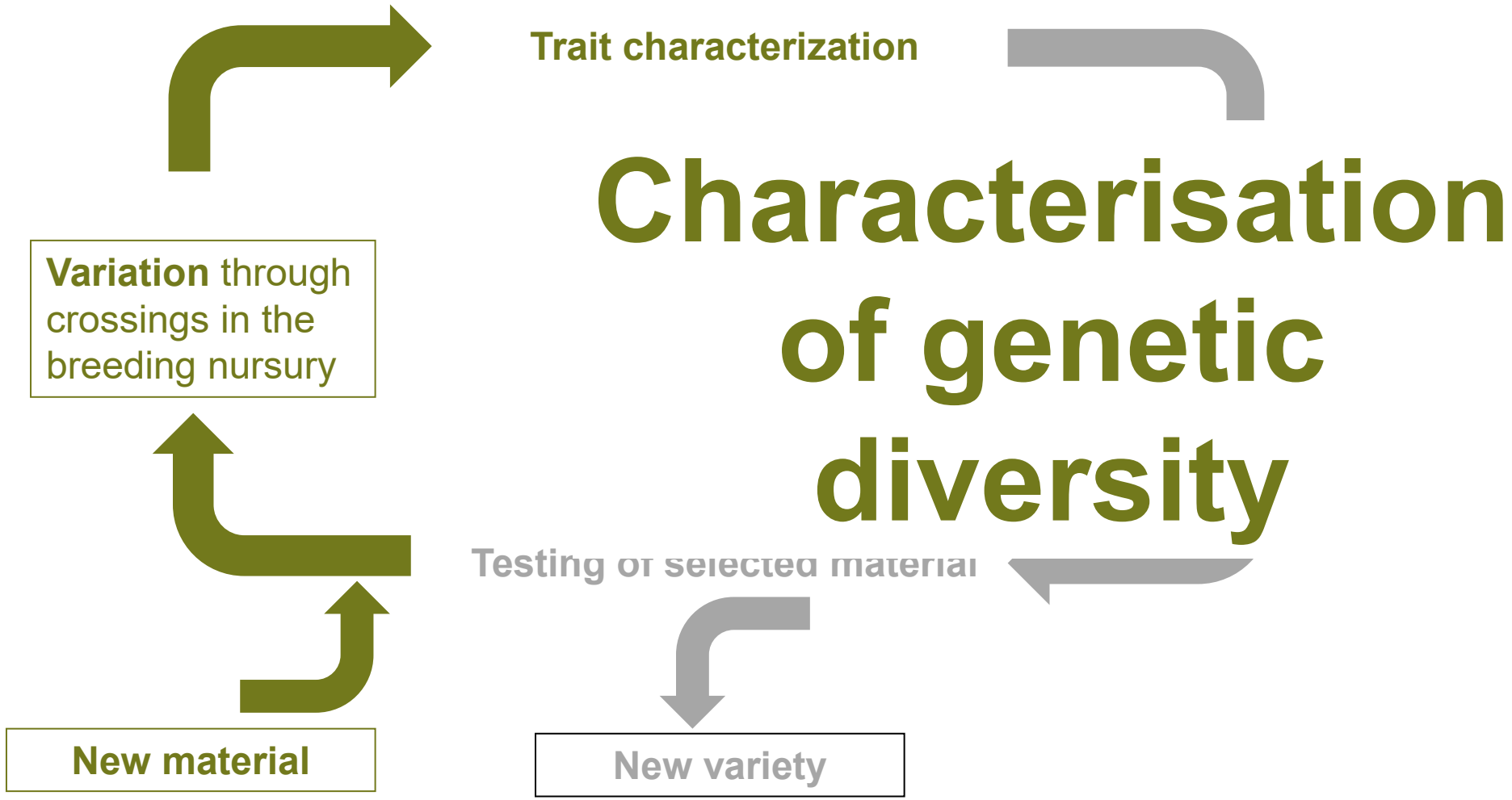
LegumeLegacy

Training event Switzerland, Zurich, 15.4. - 19.4.2024

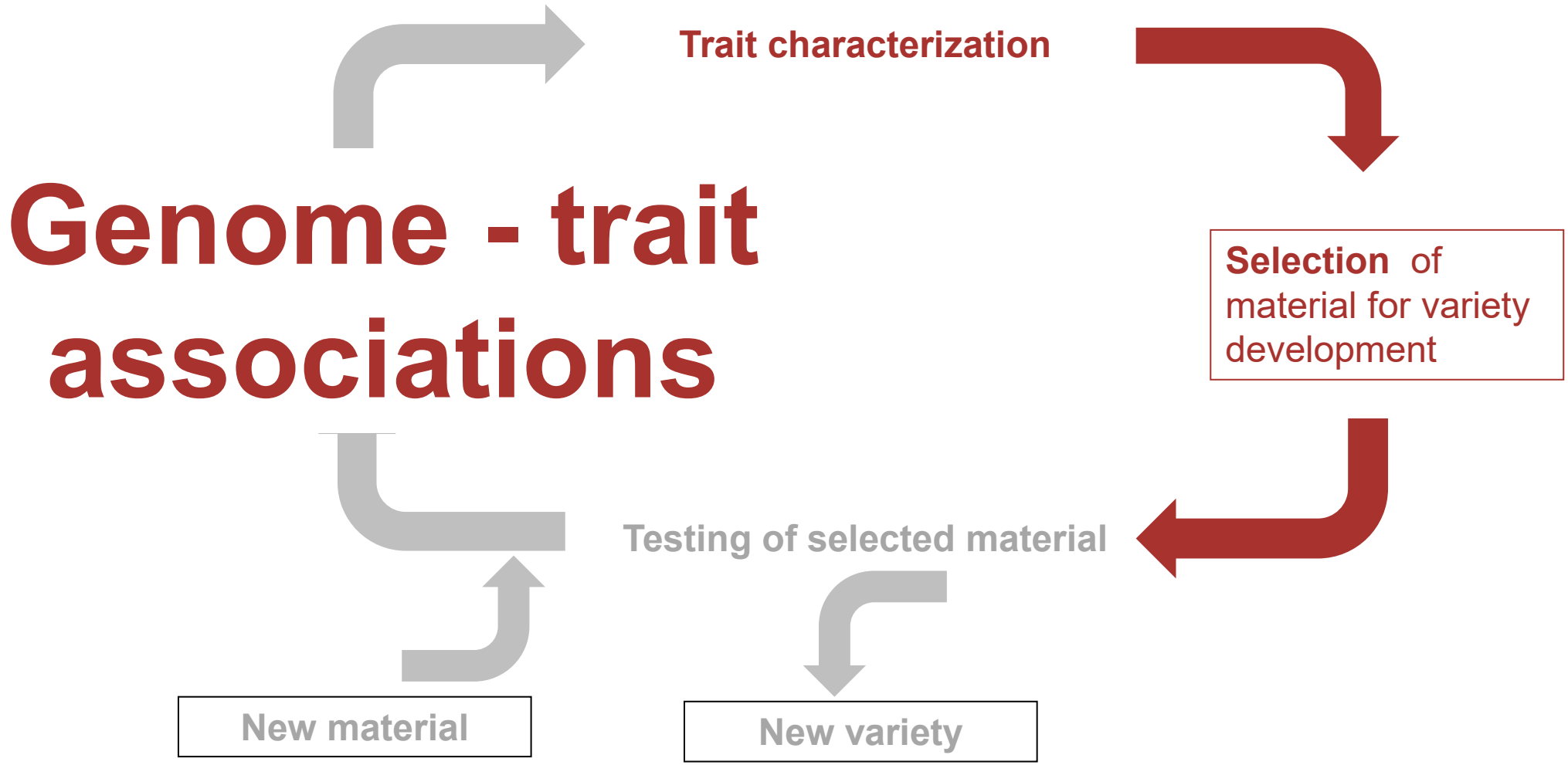
Data science and statistics in molecular plant breeding



Data science and statistics in molecular plant breeding



Data science and statistics in molecular plant breeding



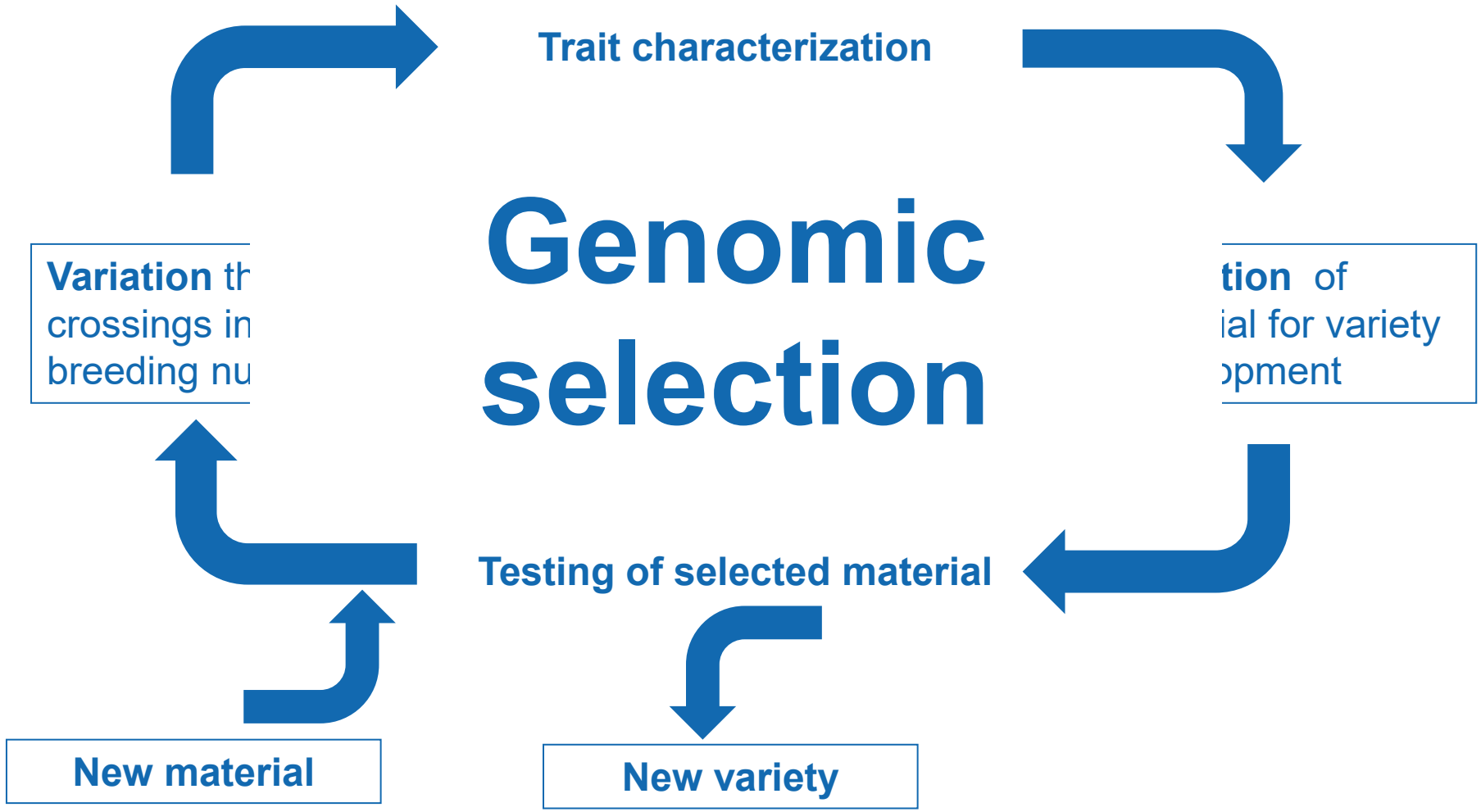
Genome – trait associations

LegumeLegacy

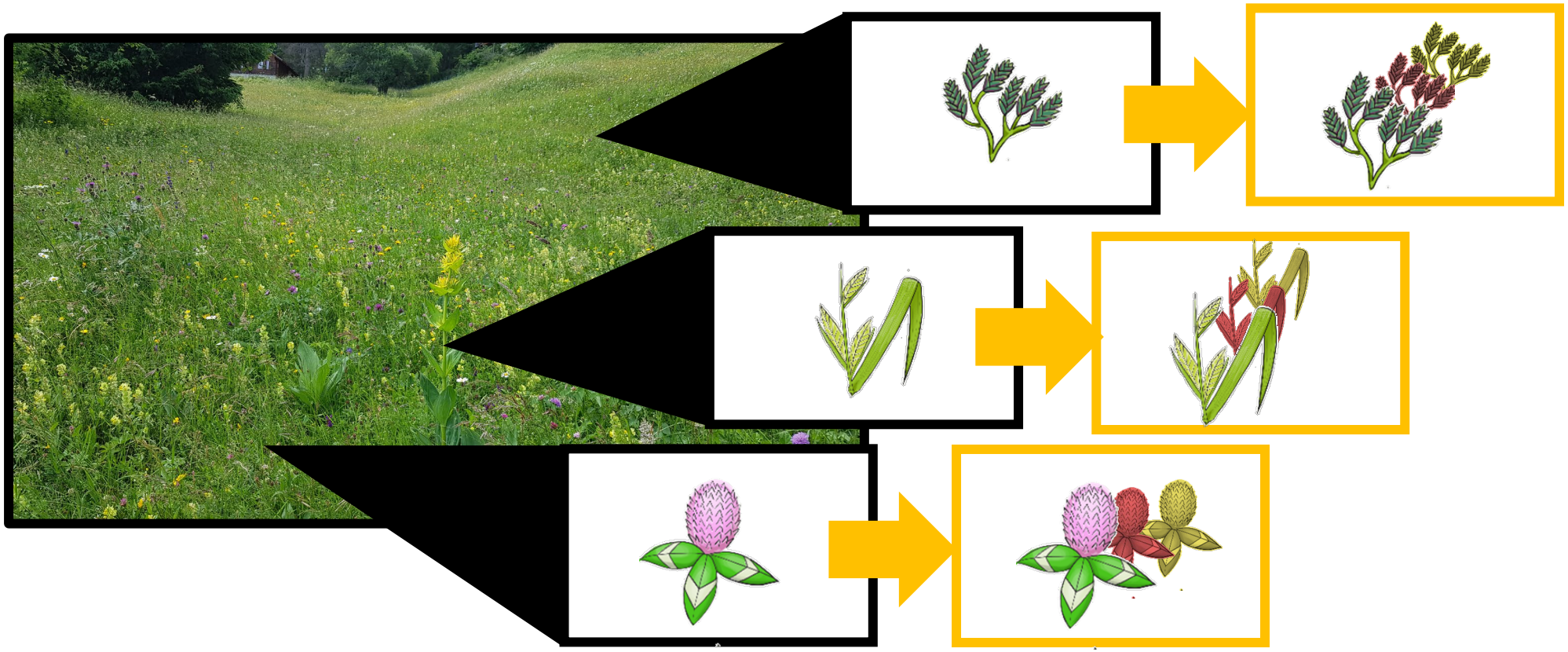
Identify genomic regions (DNA sequences) that control / influence target phenotypic traits

- Marker – trait associations
  - Linkage mapping and quantitative trait locus (QTL) analysis
    - Single marker regression, interval mapping, multiple regression, …
- Genome – wide association mapping
  - Analyse the whole genome in a large number of diverse populations
  - Generalised linear models, linear mixed models, Bayesian approaches, …
- Identification of candidate genes (i.e. genes controlling the trait)
  - Sequence comparisons to model species (BLAST analysis in large databases)
  - Transcriptomics (analysis of gene expression)
- Validation of candidate genes

Data science and statistics in molecular plant breeding



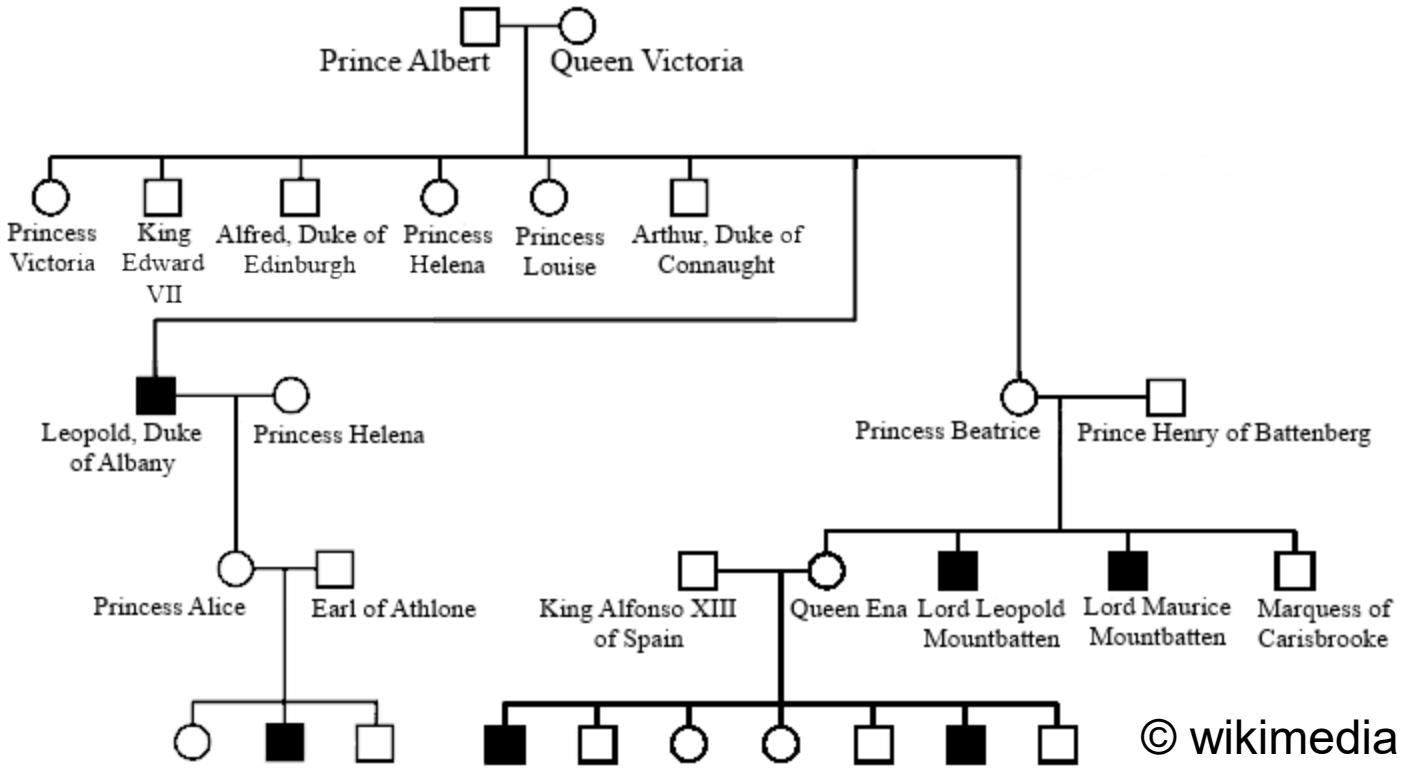
What is genetic diversity?



- Genetic diversity can be defined as the genetic differences between individuals within a species or a population

How can genetic diversity be measured?

- Calculation of genetic diversity (or similarity) based on **pedigree information**
  → **Identity By Descent (IBD)**



© wikimedia

Coefficient of coancestry

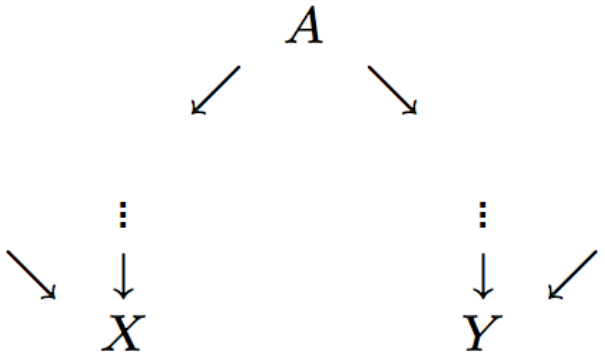
- The similarity (or diversity) between two individuals can be expressed using the **coefficient of coancestry Θ_{ij}**

- The concestry coefficient is defined as **the probabilty that two alleles at a locus, drawn at random from two individuals are identitical by descent**

- Examples:
  - parent – offspring $\Theta = (1/2)^2 = 1/4$
  - half-sibs $\Theta = (1/2)^3 = 1/8$
  - full-sibs $\Theta = (1/2)^3 + (1/2)^3 = 1/4$

Calculating the coancestry coefficient

- Algorithms to obtain kinship coefficients often use a technique called "path counting«
- To get the coancestry coefficient for X and Y, we would identify the path linking them through their common ancestor(s)

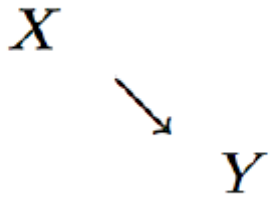


- If the X and Y have ancestor A in common, and if there are n individuals (including X, Y) in the path linking them through A, then the coancestry of X and Y , is

$$\theta_{XY} = \left(\frac{1}{2}\right)^n$$

- If there are several ancestors, this expression is summed over all the ancestors

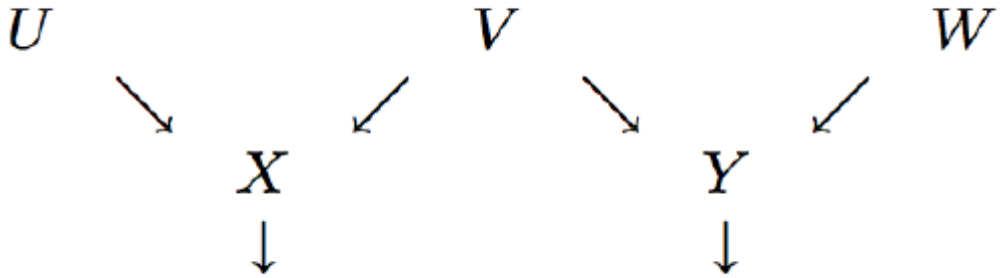
Path counting: parent-offspring



- The common ancestor of parent X and child Y is X. The path linking X; Y to their common ancestor is YX and this has n = 2 individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

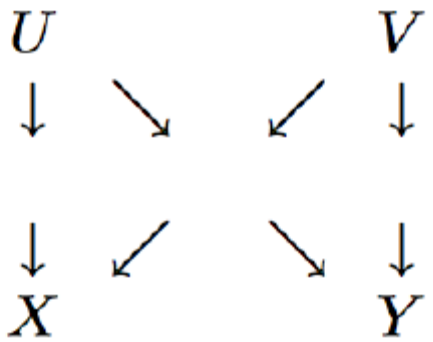
Path counting: half sibs



- The common ancestor of half sibs X and Y is V. The path linking X, Y to their common ancestor is XVY and this has n = 3 individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

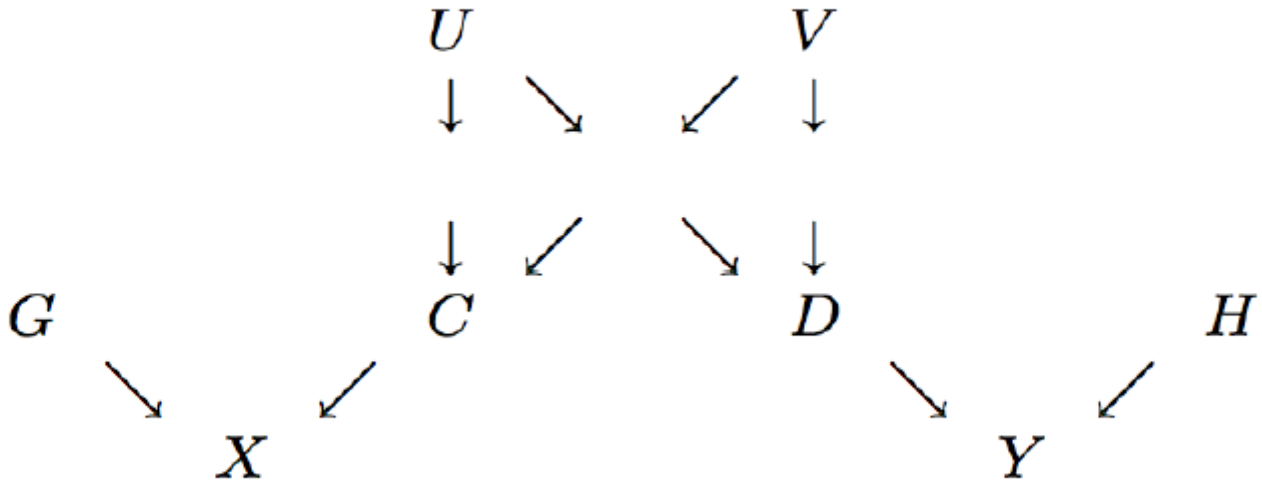
Path counting: full sibs



- The common ancestors of full sibs X and Y are U and V. The paths linking X and Y to their common ancestors are XUY and XVY and these each have n = 3 individuals, therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \frac{1}{4}$$

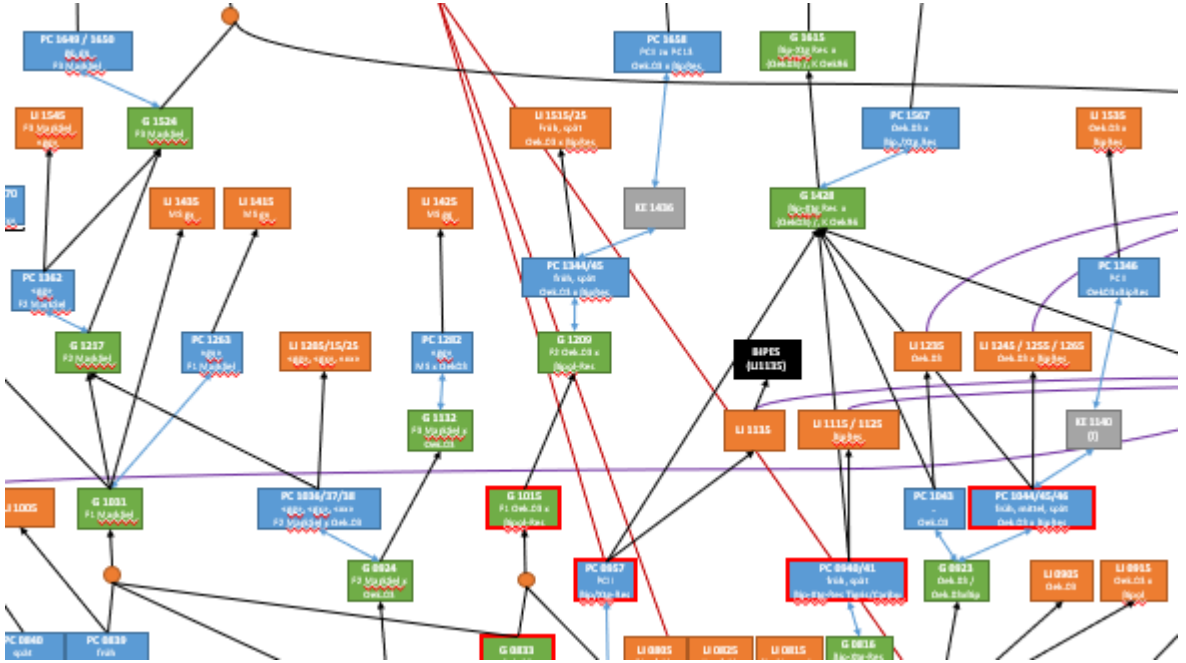
Path counting: first cousins



- Calculate the kinship coefficient for first cousins X, Y using path counting

Limitations of pedigree-based estimates of genetic diversity

- Often very complex pedigrees in breeding schemes, particularly for population-based cultivars in outbreeding crops
- Pedigree information not or only partially available
- No comparison possible to unrelated populations, wild ancestors etc.



Estimating genetic diversity

Compare heritable properties of individuals and calculate genetic diversity

- **Phenotypic markers**
  - binary traits
    - leaf marks, awns
  - quantitative traits
    - leaf width, spike length, plant height, flower morphology,…
  - need to eliminate effects caused by environmental factors

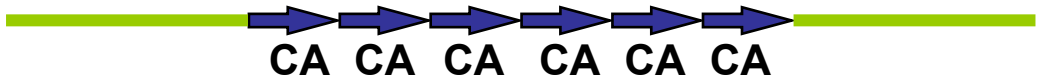
    → replicated field trials

Estimating genetic diversity

- **Molecular genetic markers**
  - Differences in DNA sequences
    - Large number of markers available
    - Not influenced by the environment
  - Various marker systems
    - **S**imple **S**equence **R**epeats
    - **A**mplified **F**ragment **L**ength **P**olymorphism
    - **S**ingle **N**ucleotide **P**olymorphism
    - **G**enotyping **B**y **S**equencing
    - …

Simple Sequence Repeat (SSR) markers

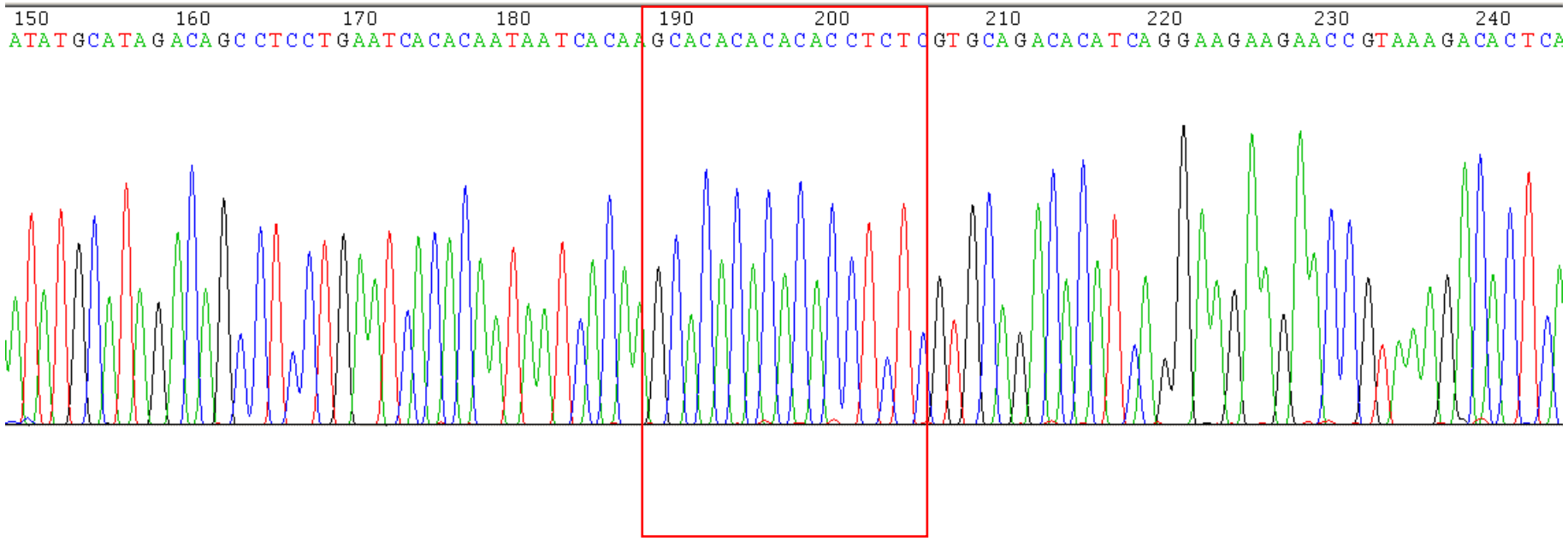
- Repetitive DNA motifs (2-4bp)



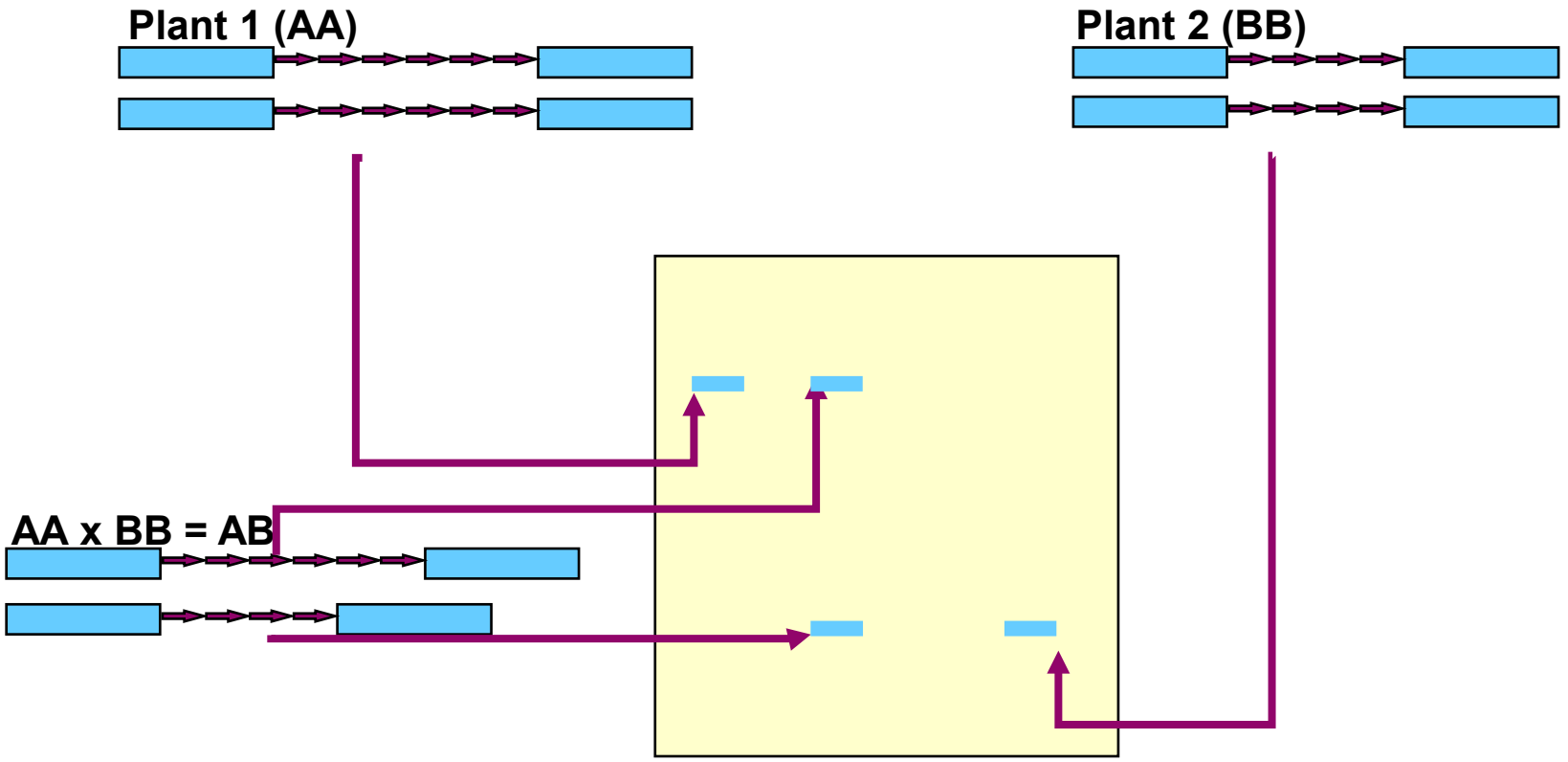
- Polymorphisms: variable number of repeated elements

- Flanking regions are often conserved → PCR amplification

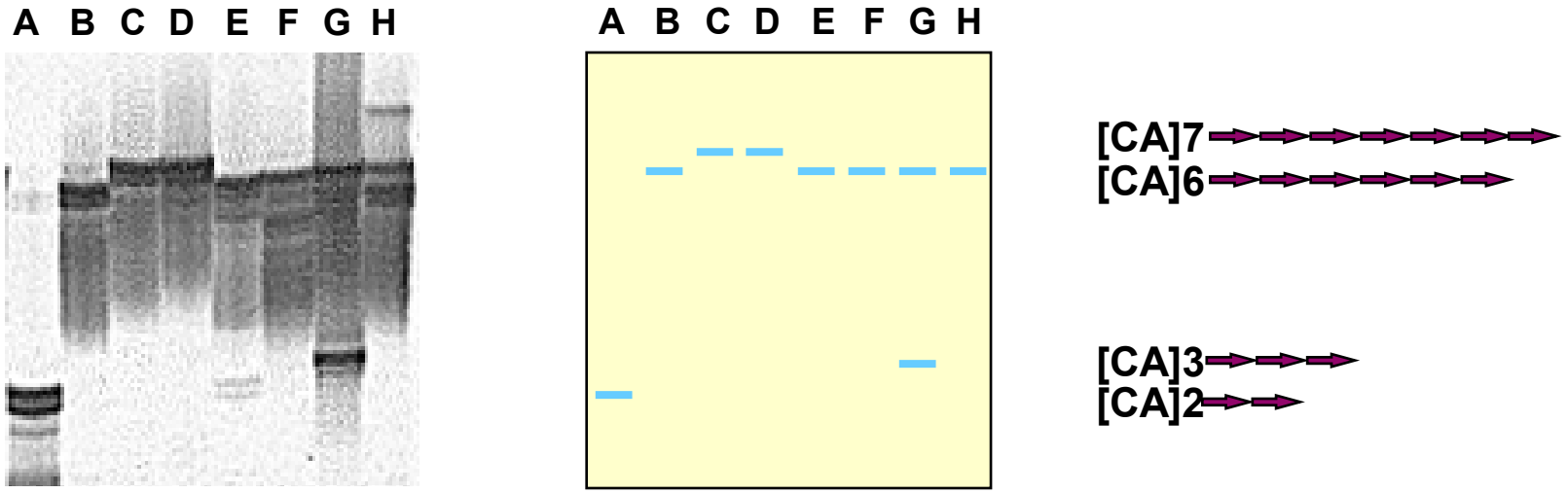
Simple Sequence Repeat (SSR) markers



Simple Sequence Repeat (SSR) markers



Simple Sequence Repeat (SSR) markers



Estimating genetic diversity

- Describe individuals under investigation with as many markers (phenotypic or genotypic) as available
- Compute pairwise differences between individuals using all marker information
- Different measures available depending on marker data
  - **Euclidean Squared Distance**

$$E_{ij}^2 = \sum_k (x_{ki} - x_{kj})^2$$

$_{i,j}$ = individual plants, $_k$ = marker locus

  - phenotypic data (qualitative and quantitative)
  - dominant marker data (e.g. AFLP): equals the number of marker differences between two individuals

Euclidean squared distance

- Example
  - Two individual plants (A, B)
  - Two traits (plant height, number of flowers)
  - Data:

Plant	Plant height	Number of flowers
A	150 cm	15
B	120 cm	35

$$E_{ij}^2 = \sum_k (x_{ki} - x_{kj})^2$$

  - $E^2 = (150\text{-}120)^2 + (15 - 35)^2 = 130$
    - Different scales → scale (and center) data

Measures of genetic diversity

- Simple Matching Coefficient
  - (a + d)/(a + b + c + d)
  - for dominant marker data (AFLP)
  - considers that absence corresponds to homozygous loci
- Jaccard Coefficient
  - a/(a + b + c)
  - for co-dominant data (SSR)
  - only counts bands present in either individual
- Nei-Li Coefficients
  - 2a/(2a + b + c)
  - for co-dominant data (SSR)
  - percentage of shared bands

		Indiv.$_j$		
		1	0	
Indiv.$_i$	1	a	b	a + b
	0	c	d	c + d
		a + c	b + d	n

Measures of genetic diversity

- **Identity By State (IBS)**
  - proportion of loci at which two individuals share the same alleles (1 for complete identity)
  - note the difference to IBD (independent mutations

- **Rogers Distance** $RD = 1 - IBS$

- **Modified Rogers Distance** $MRD = \sqrt{1 - IBS}$
  - for co-dominant data (SNP)

Data analysis / interpretation

- Analysis of genetic diversity usually involves a large number of individuals characterised at a large number of loci

	AccID	LG1_1001490	LG2_26125099	LG5_6958144	LG6_10423959	LG7_27042118	scaf156_111229	scaf358_82618	scaf999_42394	scaf999_42400
1	TP001	0.95	0.88	0.82	0.37	0.74	0.75	0.52	0.77	0.78
2	TP002	0.98	0.96	0.78	0.48	0.22	0.91	0.49	0.74	0.98
81	TP081	0.81	1.00	0.59	0.44	0.57	0.82	0.44	0.55	0.99
101	TP102	0.85	0.94	0.59	0.39	0.57	0.91	0.82	0.80	0.77
180	TP183	0.90	0.97	0.70	0.37	0.54	0.97	0.40	0.92	0.89
188	TP191	0.81	0.98	0.58	0.42	0.63	0.86	0.42	0.83	0.96
241	TP253	0.88	0.95	0.62	0.37	0.55	0.70	0.50	0.86	0.90
255	TP267	0.90	1.00	0.69	0.39	0.42	0.63	0.81	0.89	0.97
396	TP661	0.97	1.00	0.26	0.43	0.45	0.95	0.68	1.00	0.99
397	TP662	0.90	0.93	0.47	0.38	0.68	0.65	0.31	0.94	0.97

>

- Multivariate descriptive analyses facilitate the identification of groups of individuals
  - cluster analysis
  - principle component analysis
  - multidimensional scaling

Cluster analysis

- group observations using objective criteria
- calculate similarity or difference between individual observations (e.g.using Euclidean distance)
- draw graphical representation starting with most similar observation
- and ’let the tree grow’
- various clustering algorithms depending on research question

	1	2	3	4
2	0.20			
3	0.25	0.40		
4	0.45	0.35	0.30	
5	0.80	0.50	0.60	0.70

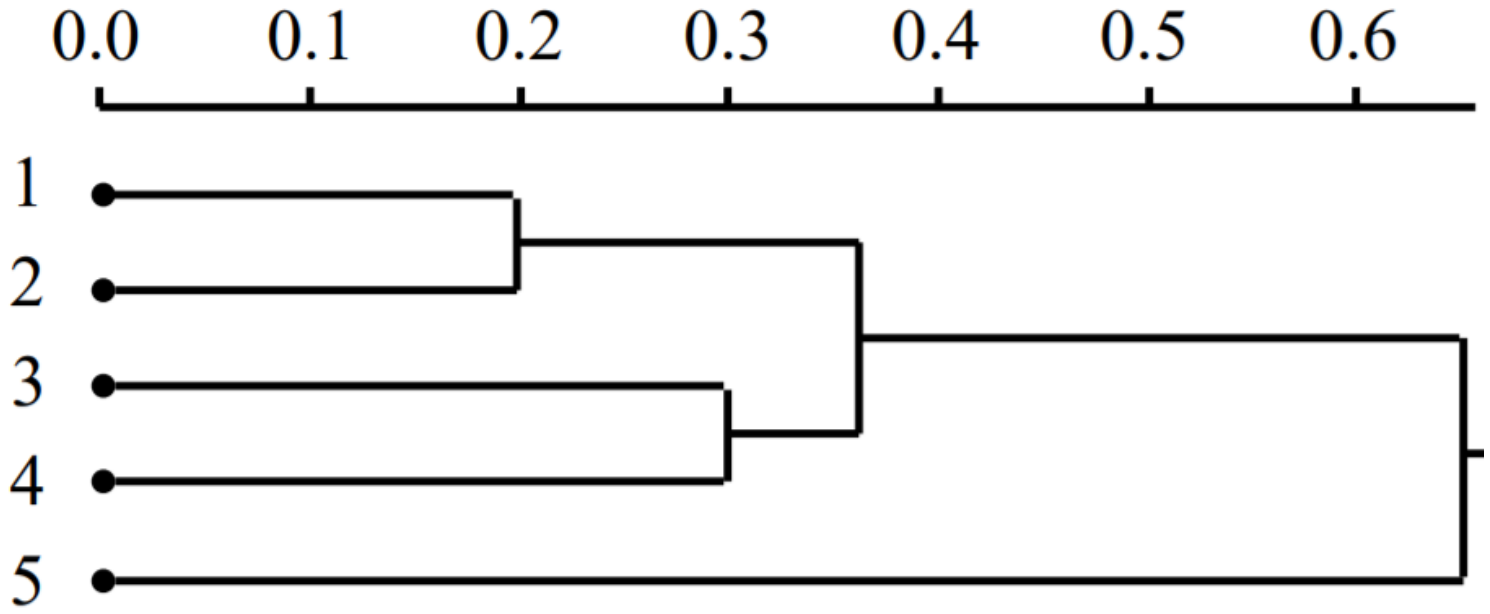
D1	Pairs of objects
0.20	1 - 2
0.25	1 - 3
0.30	3 - 4
0.35	2 - 4
0.40	2 - 3
0.45	1 - 4
0.50	2 - 5
0.60	3 - 5
0.70	4 - 5
0.80	1 - 5

Clustering algorithms

- UPGMA: unweighted pair group method with arithmetic mean
  - Proximity between two clusters is the arithmetic mean of all the proximities between the objects of one, on one side, and the objects of the other, on the other side.

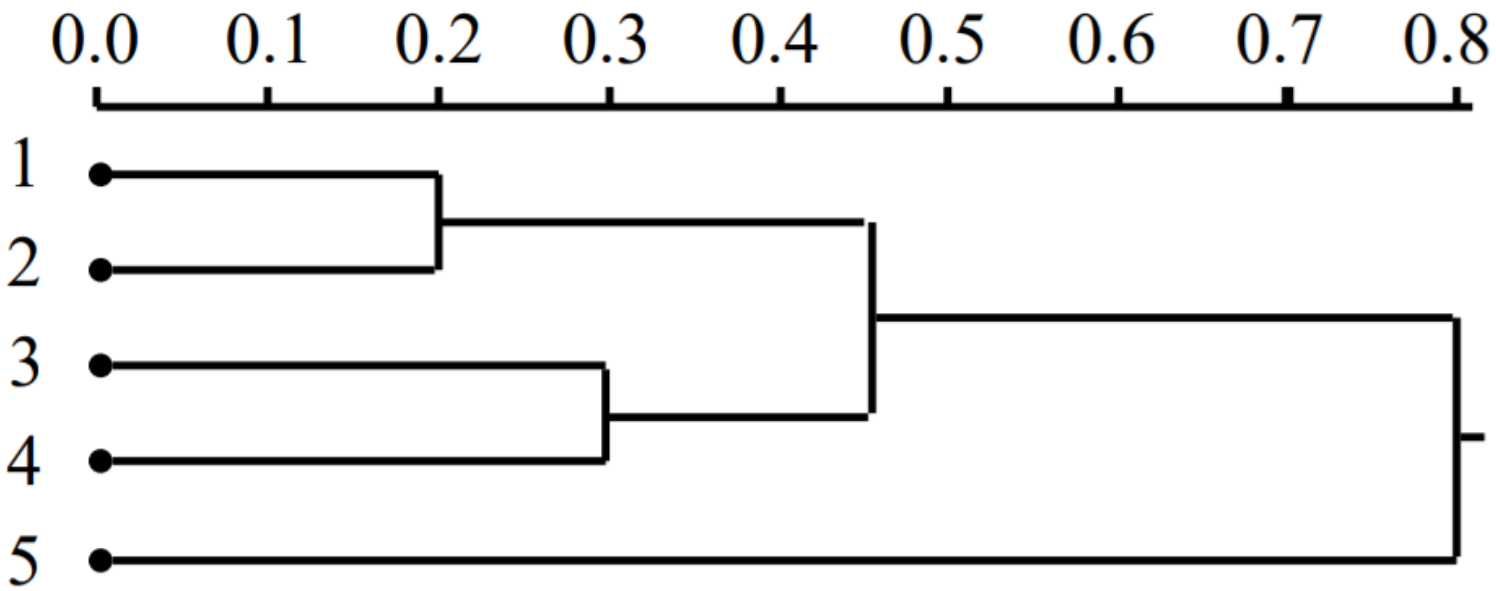
- Ward's method of minimal increase of sum of squares
  - Proximity between two clusters is the magnitude by which the summed square in their joint cluster will be greater than the combined summed square in these two clusters
- Roland

UPGMA clustering



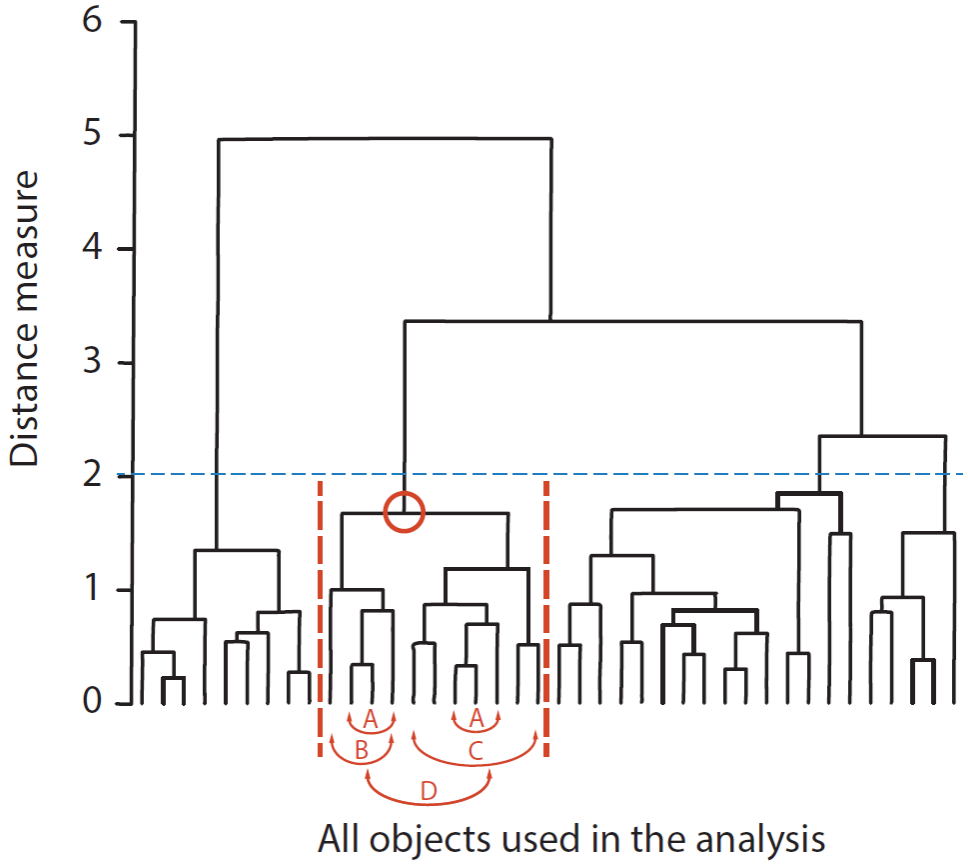
D1	Pairs of objects
0.20	1 - 2
0.25	1 - 3
0.30	3 - 4
0.35	2 - 4
0.40	2 - 3
0.45	1 - 4
0.50	2 - 5
0.60	3 - 5
0.70	4 - 5
0.80	1 - 5

Ward clustering



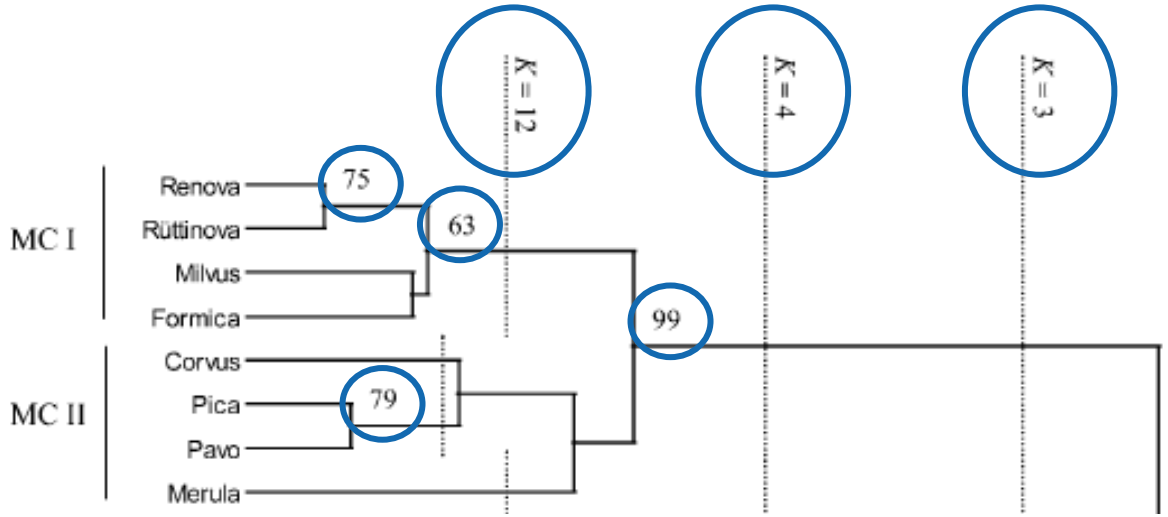
D1	Pairs of objects
0.20	1 - 2
0.25	1 - 3
0.30	3 - 4
0.35	2 - 4
0.40	2 - 3
0.45	1 - 4
0.50	2 - 5
0.60	3 - 5
0.70	4 - 5
0.80	1 - 5

Interpreting dendrograms



Interpreting dendrograms

- Define numer of relevant clusters
- Bootstrap analysis for cluster support



Principal Component Analysis

basic idea: visualise multidimensional data

- reduce the dimension of the data set
- retain most important variation of the data
- transform original data into new uncorrelated variables in a way that variation on new variables is maximised

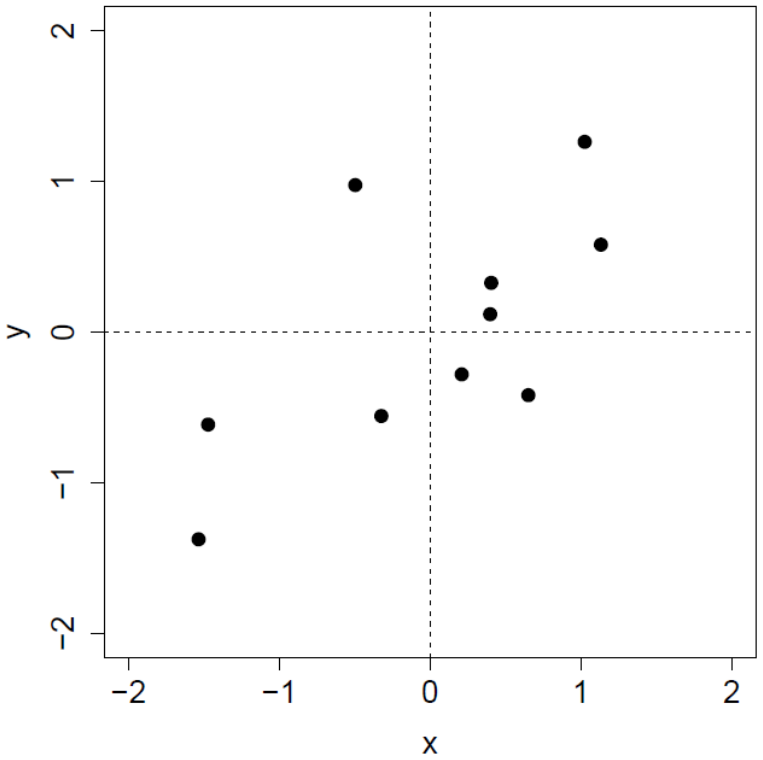
history

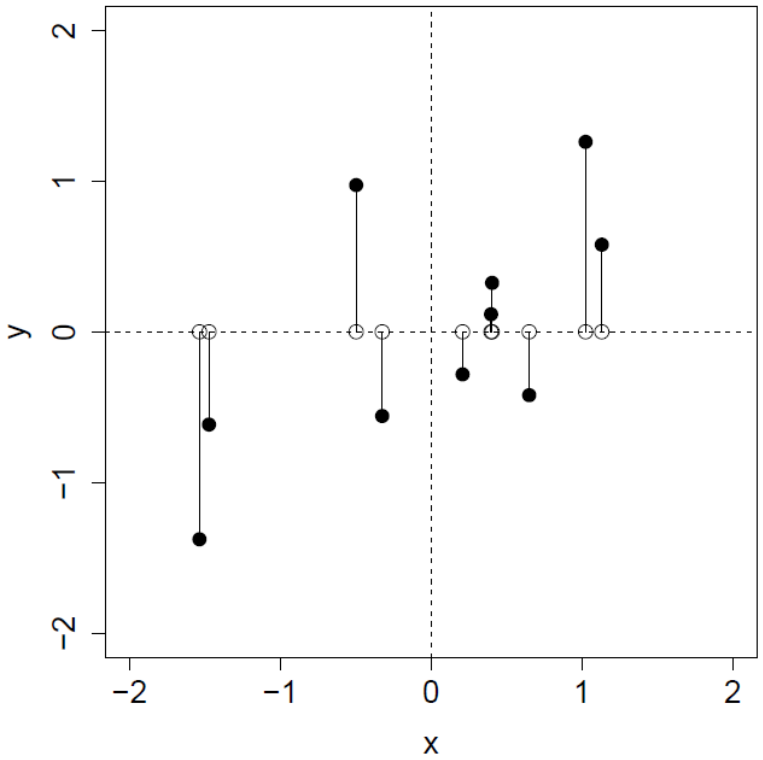
- 1901: first proposed (Karl Pearson)
- 1933: general procedures established (Harold Hotelling)
- 1970's: widely adopted

The principles of principal component analysis

Principle Components (PC)

- represent the underlying structure in the data
- give the directions where there is the most variance

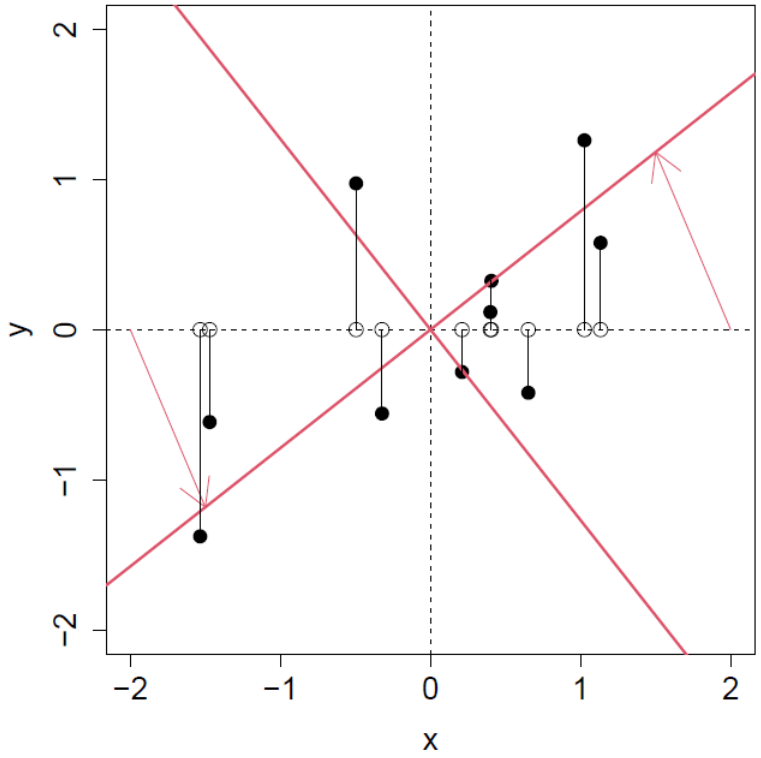


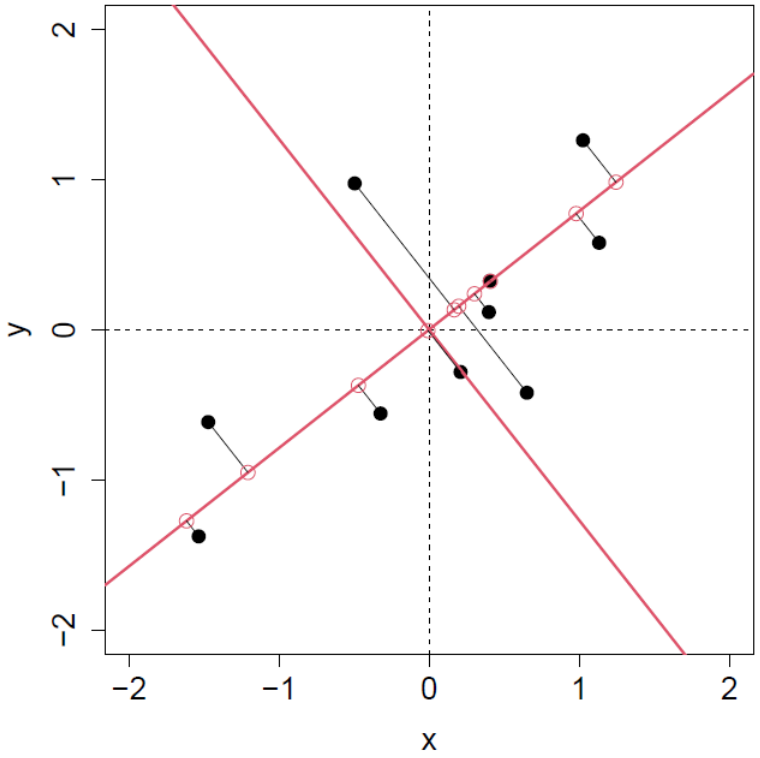


The principles of principal component analysis

Principle Components (PC)

- give the directions where there is the most variance
- are linear combinations of the original variables

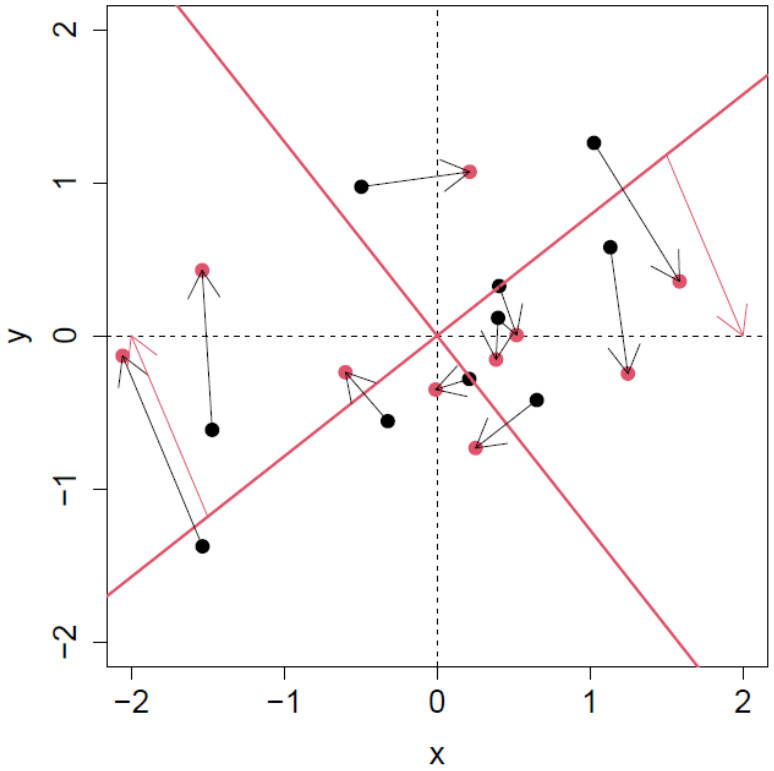


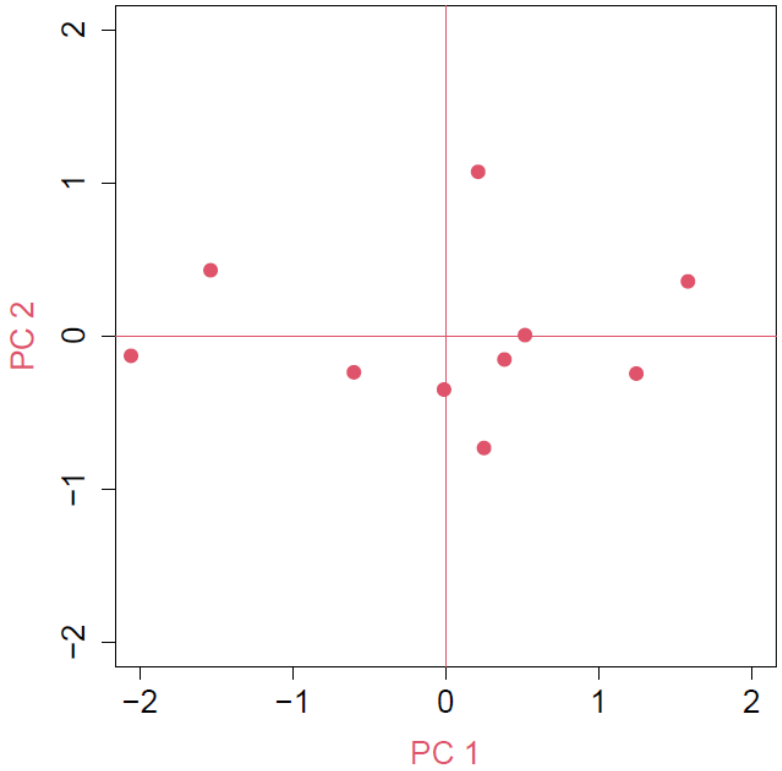


The principles of principal component analysis

Principle Components (PC)

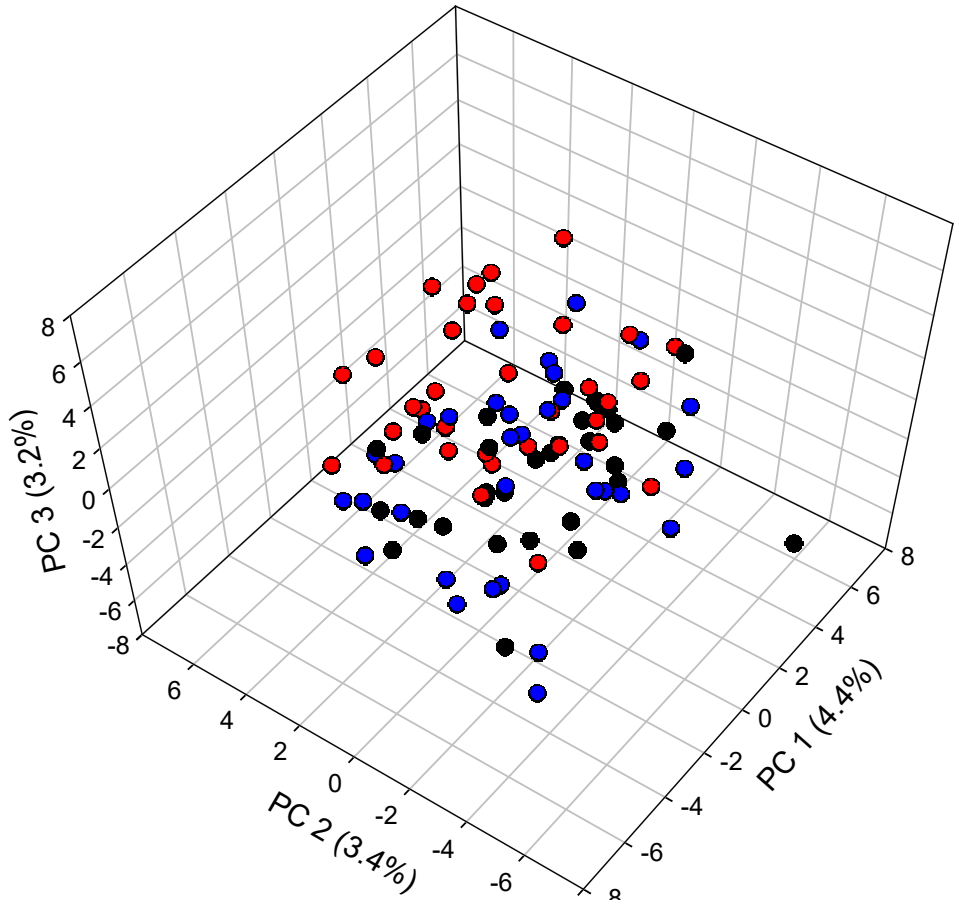
- are linear combinations of the original variables (PC 1 $= a_1 x + b_1 y$)
- are orthogonal (uncorrelated) to each other





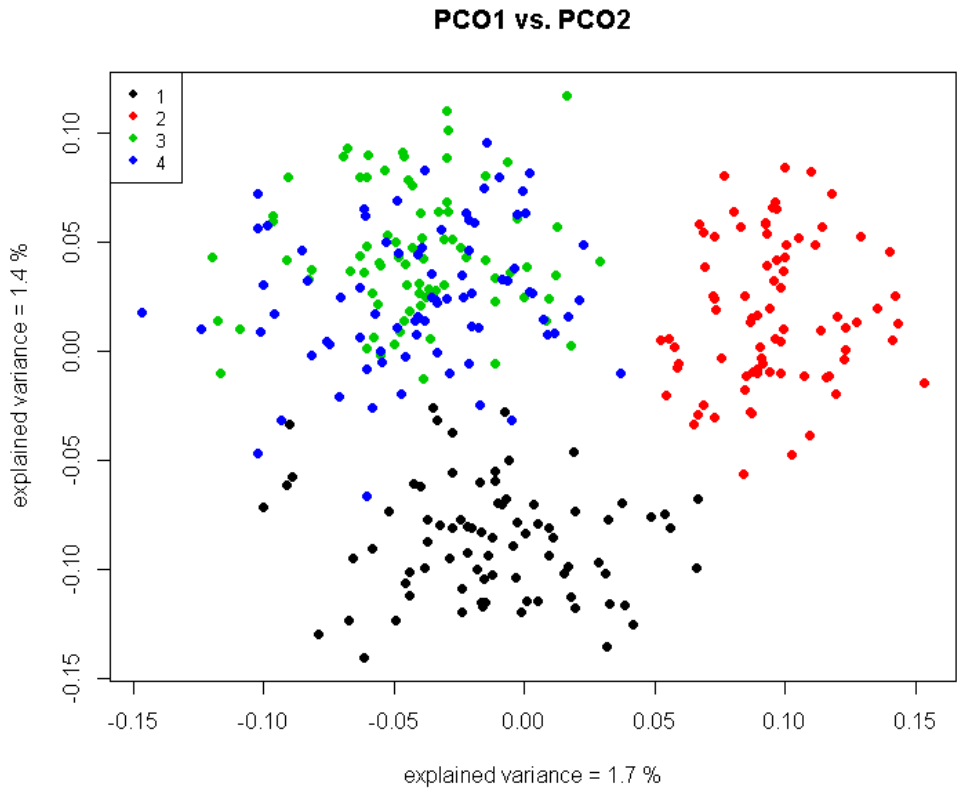
Principal component analysis

- Visualisation of similarity or relatedness of samples
- The closer the point, the more similar the samples



Multidimensional scaling

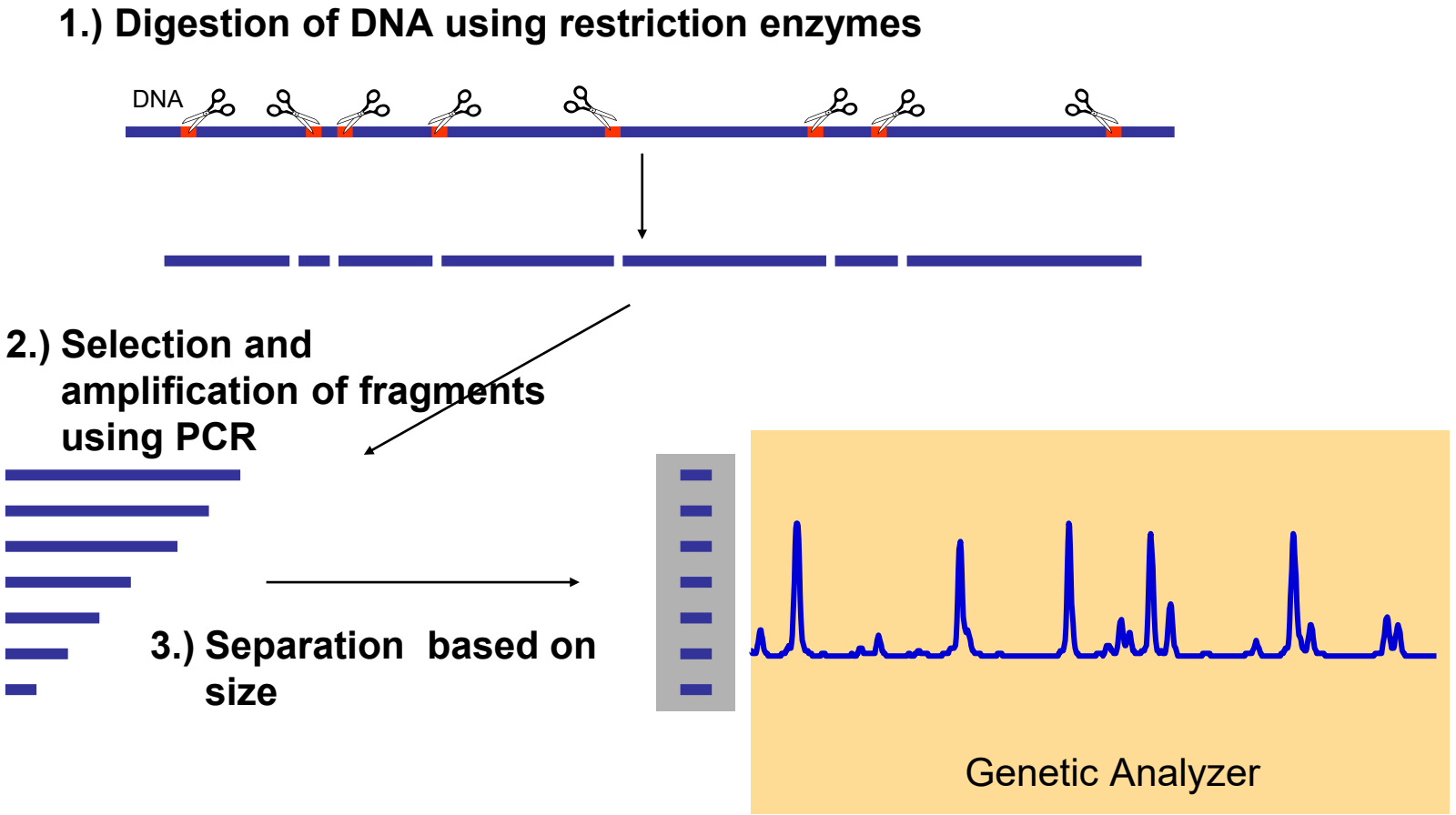
- visualizing between-object distances in a multi-dimensional space by minimizing a loss function
- based on distance matrix
- when based on Euclidean distance == Principle component analysis!



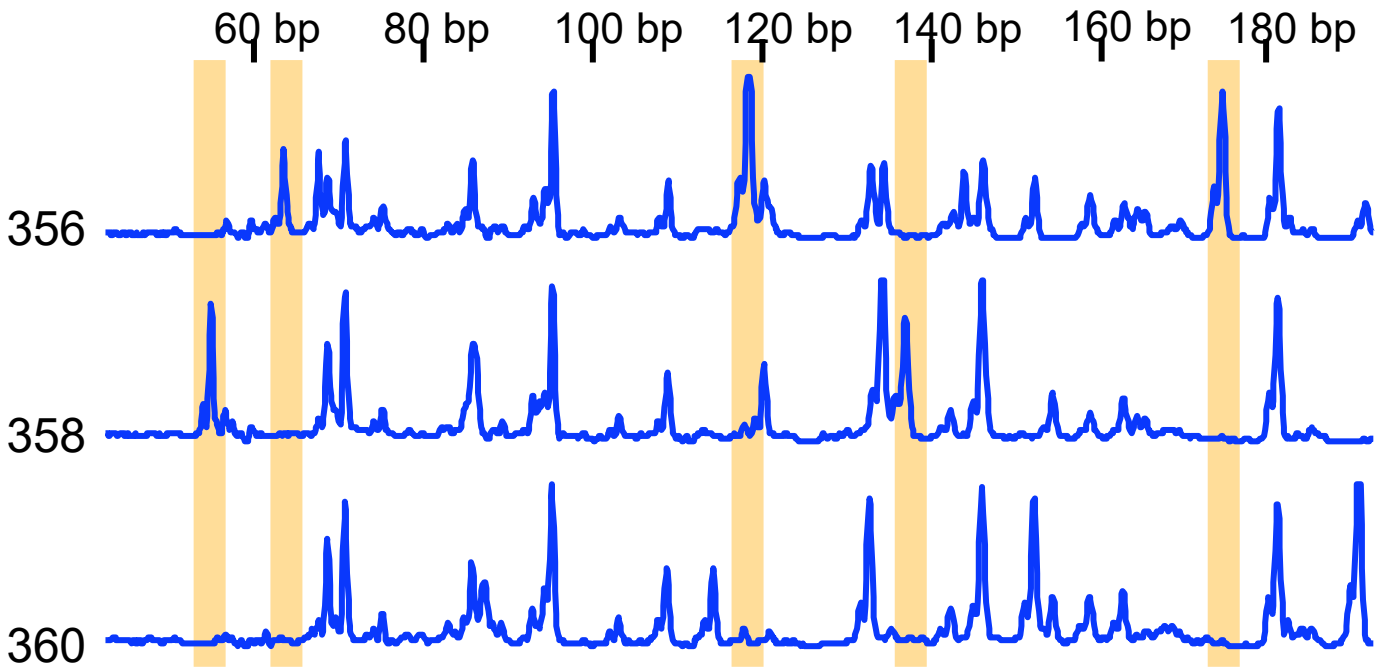
LegumeLegacy

FROM MARKER DATA TO POPULATION STRUCTURE

AFLP (Amplified Fragment Length Polymorphism)



AFLP peaks



AFLP patterns of three red clover cultivars

AFLP data

Pflanze	AFLP1	AFLP2	AFLP3	AFLP4	AFLP5	AFLP6
351	0	0	0	1	0	0
352	0	0	0	1	0	0
353	0	0	0	1	0	0
354	0	0	1	0	0	0
355	0	0	0	1	0	0
356	0	0	0	1	0	0
357	0	0	0	1	0	0
358	0	0	0	1	0	0
359	0	0	0	1	0	0
360	0	0	0	1	0	0
361	0	0	0	1	0	0

	351	352	353	354	355	356	357	358	359	360	361
351	0										
352	7	0									
353	9	4	0								
354	17	20	22	0							
355	12	9	13	19	0						
356	10	9	9	21	10	0					
357	9	8	8	20	9	3	0				
358	16	11	11	23	14	10	11	0			
359	11	8	12	18	9	7	8	11	0		
360	9	10	10	20	11	9	8	13	8	0	
361	7	10	10	16	11	9	6	11	10	4	0

Data matrix >>
principle component analysis

Distance matrix >>
cluster analysis,

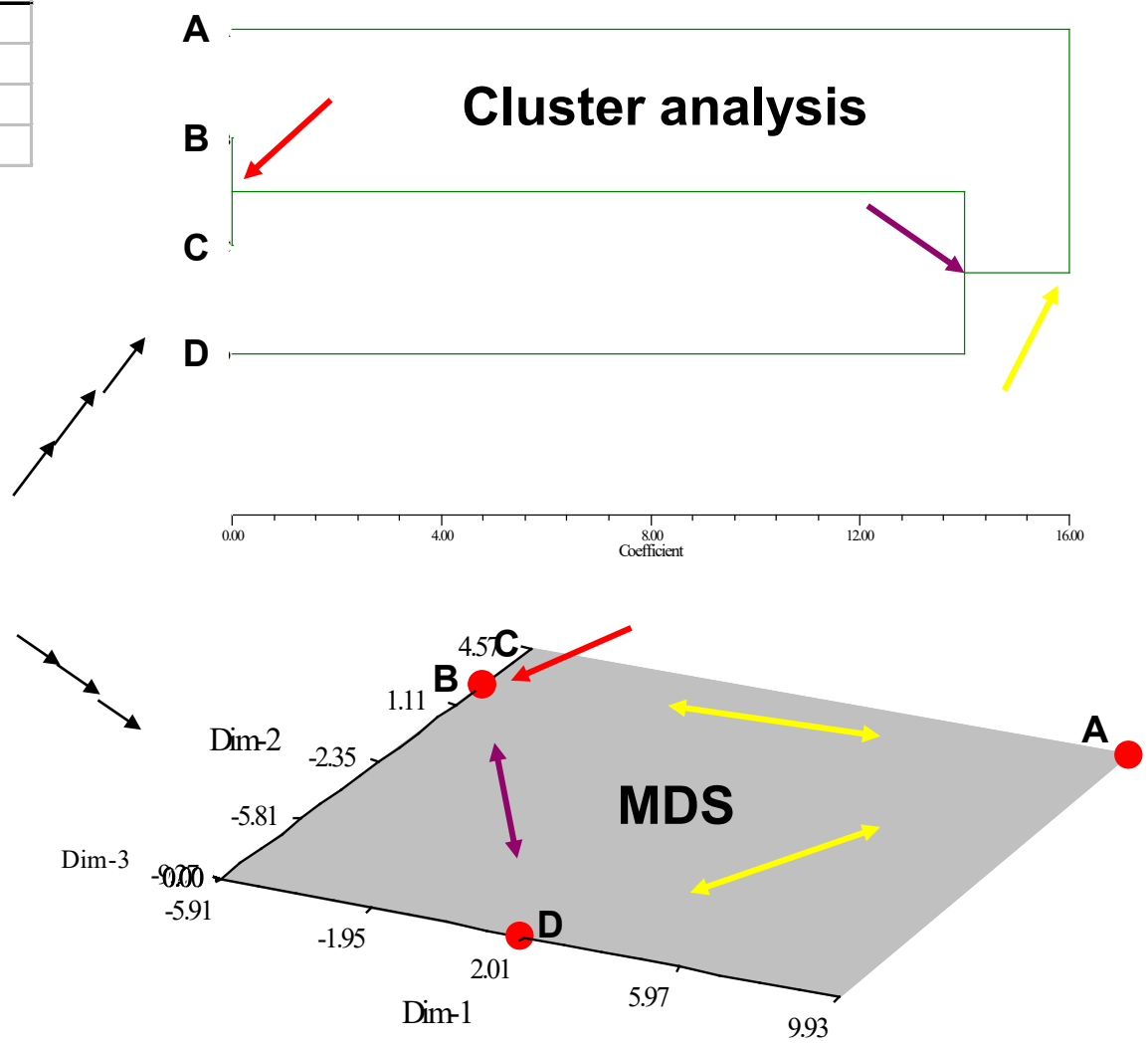
multidimensional scaling

Data analysis

	1	2	3	4	5	6	u.s.w.
A	0	1	1	0	1	0	.....
B	1	1	0	1	1	1	.....
C	1	1	0	1	1	1	.....
D	0	1	1	1	1	0	.....

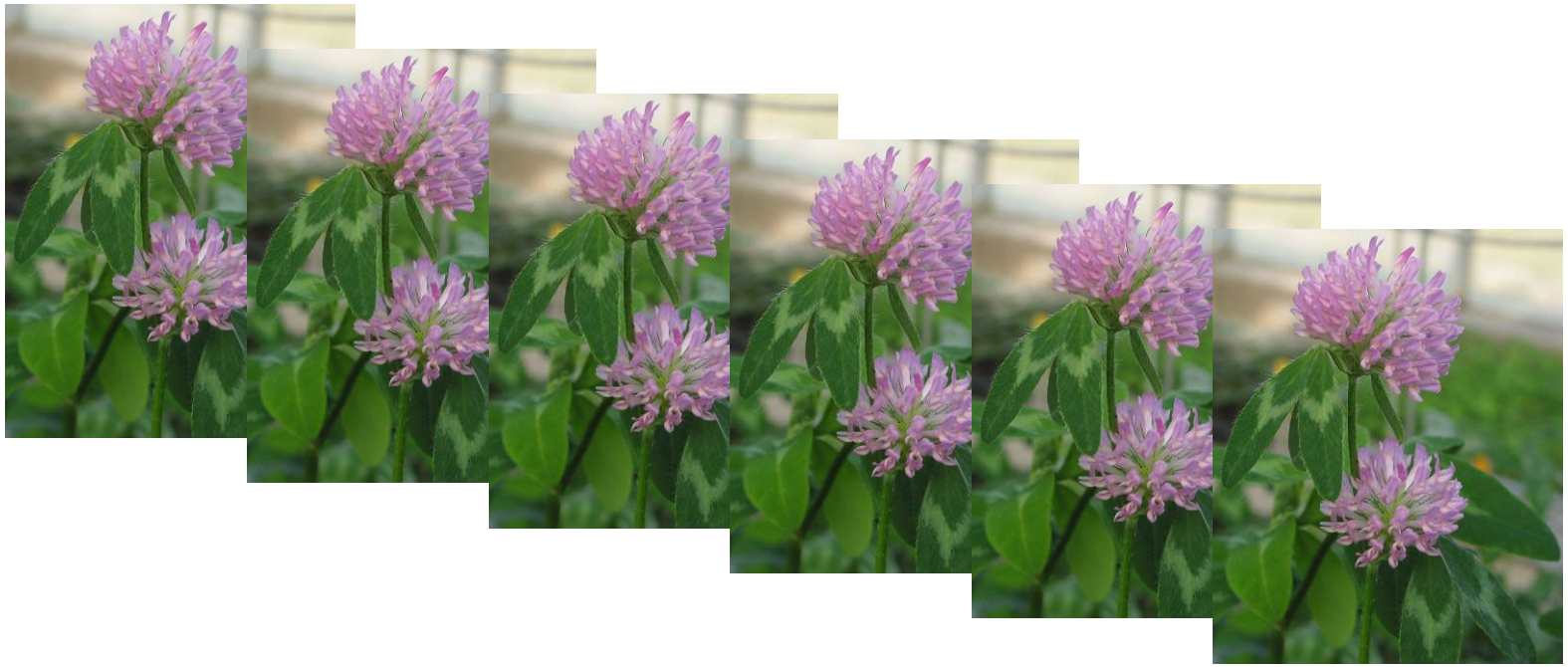
Distance matrix

	A	B	C	D
A	0			
B	16	0		
C	16	0	0	
D	16	14	14	0

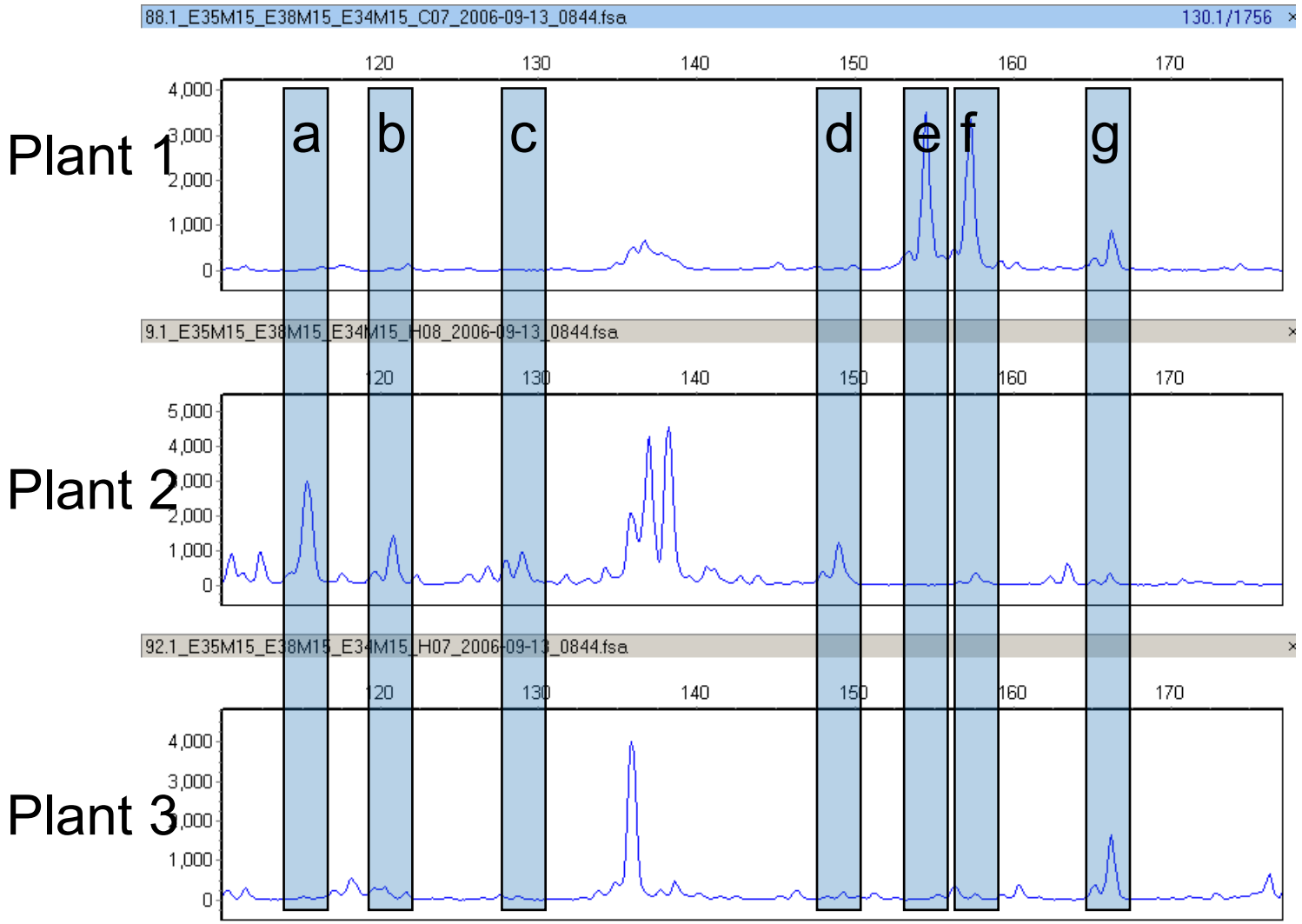


Exercise

- Genetic diversity among individuals



Exercise - AFLP data



Exercise - Genetic distance

- Calculate the genetic distance between plants 1, 2 and 3 using the following formula:

$$E_{ij}^2 = \sum_k (x_{ki} - x_{kj})^2$$

i,j = individual plants, k = marker locus

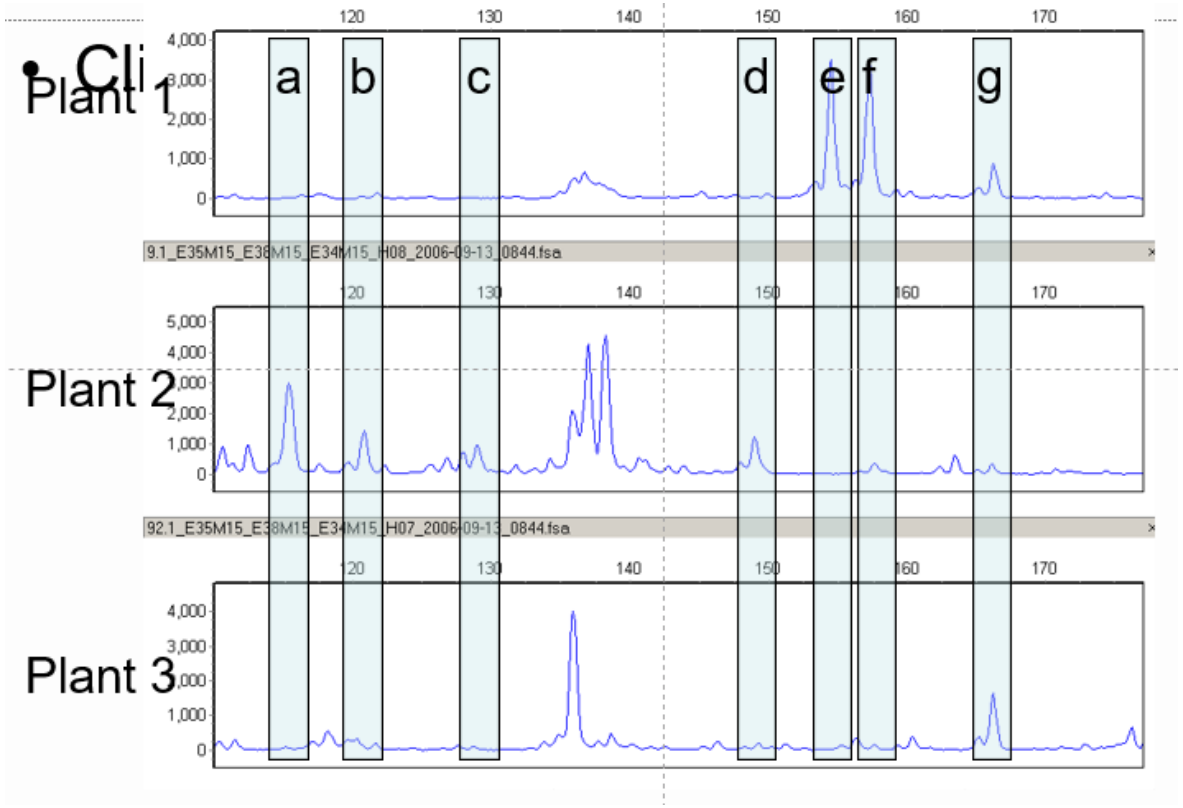
- Draw a dendrogram illustrating the relationships among the plants

- Calculate the genetic distance between plants 1, 2 and 3 using the following formula:

$$E_{ij}^2 = \sum_k (x_{ki} - x_{kj})^2$$

i,j = individual plants, k = marker locus

- Draw a dendrogram illustrating the relationships among the plants



Exercise - solution

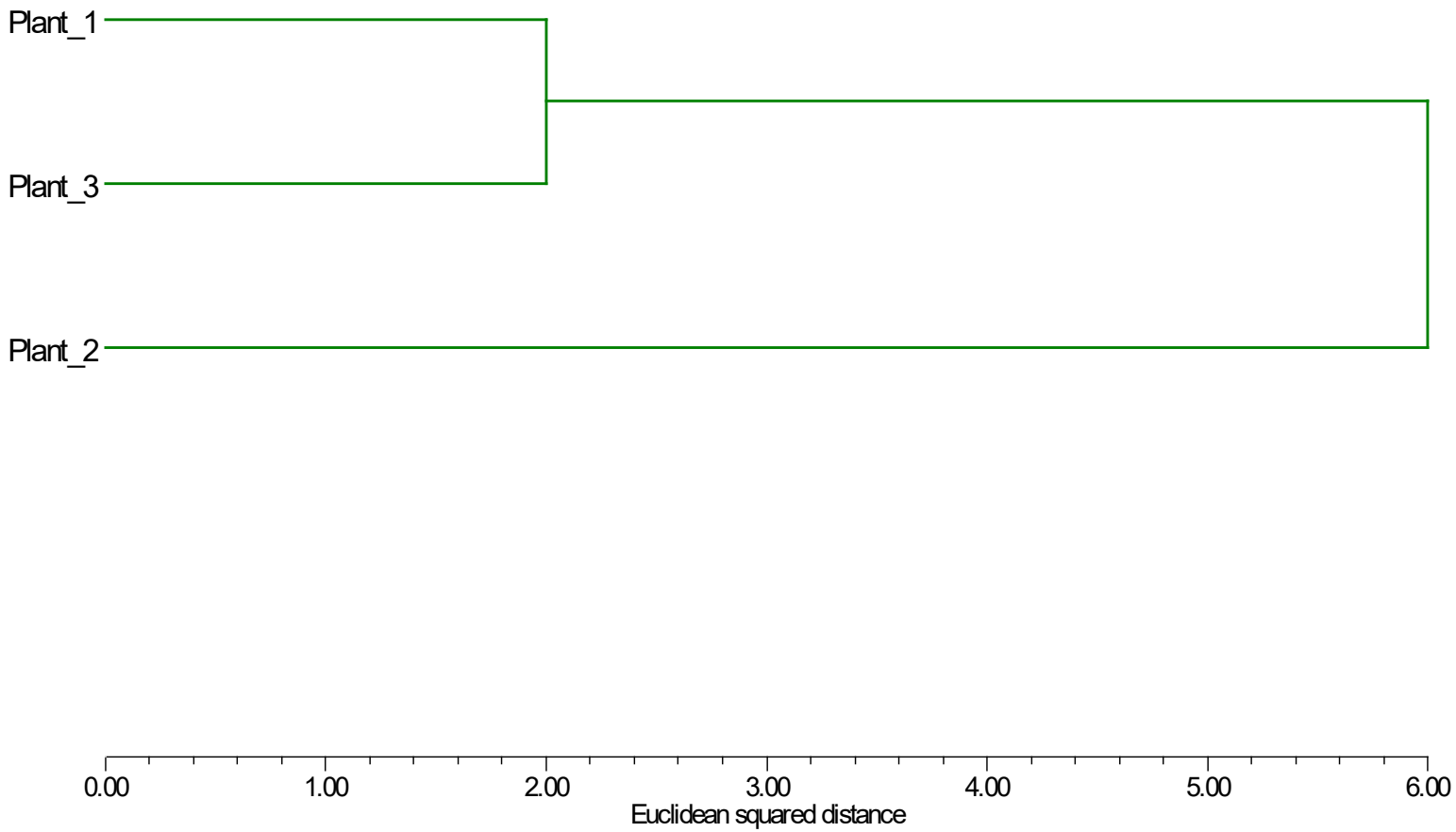
- Binary matrix

Rows\Col	a	b	c	d	e	f	g
Plant_1	0	0.000	0.000	0.000	1.000	1.000	1.000
Plant_2	1.000	1.000	1.000	1.000	0.000	0.000	0.000
Plant_3	0.000	0.000	0.000	0.000	0.000	0.000	1.000

- Distance matrix

Rows\Col	Plant_1	Plant_2	Plant_3
Plant_1	0.000		
Plant_2	7.000	0.000	
Plant_3	2.000	5.000	0.000

Exercise - dendrogram



Exercise

Five populations of *Centaurea jacea* have been sampled from five countries (**Switzerland; CH, Hungary; HU, Italy; IT, Norway; NO and Slovenia; SL**). Populations consisted of **19 individual plants each** and were analysed using amplified fragment length polymorphism (AFLP) markers. The file "centData.txt“ contains the data of **268 markers** for the 95 individual plants. Use **multivariate analyses such as cluster analysis, principle components analysis (PCA) and analysis of molecular variance (AMOVA).** You profit the most if you try to find the solutions yourself. However, don’t hesitate to ask for assistance during the lecture. Also, a possible solution is given in "https://n.ethz.ch/~rolandko/download/cent_fancy.R".

Centaurea jacea

Example

- **MARKER ASSISTED POLYCROSS BREEDING IN ITALIAN RYEGRASS**

Breeding forage crops

- Mostly cross-pollinating species
  - Wind or insect – pollination
  - High degree of self-incompatibility

- Breeding mainly focused on population cultivars
  - Open pollinated cultivars
    - Population improvement through recurrent selection

  - Synthetic cultivars
    - Intercrossing of a limited number of selected parents
    - Multiplication by random open pollination in isolation

Ryegrass breeding

- Perennial ryegrass
  - Important forage grass of temperate regions
  - Outbreeding species
  - Poly cross breeding; cultivar = heterogeneous population
- Genetic diversity
  - Heterosis, combining ability
  - Inbreeding depression, self-incompatibility
  - Uniformity >> cultivar registration
  - Variability >> performance, adaptability

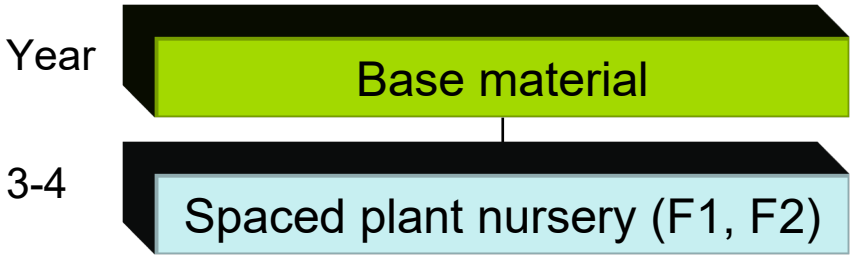
Task: Finding optimal diversity

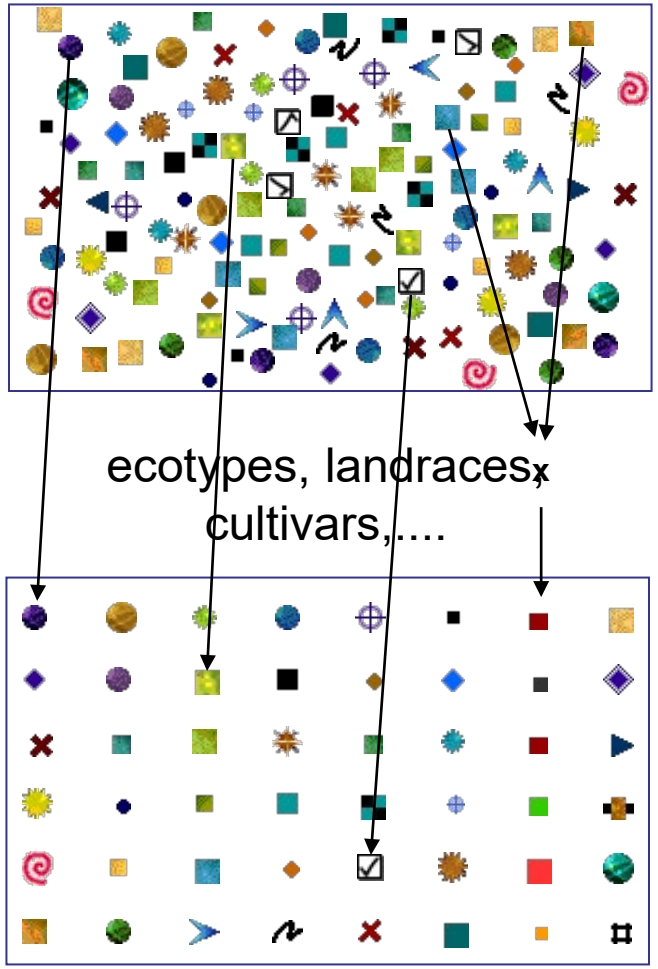
Aim of the study

LegumeLegacy

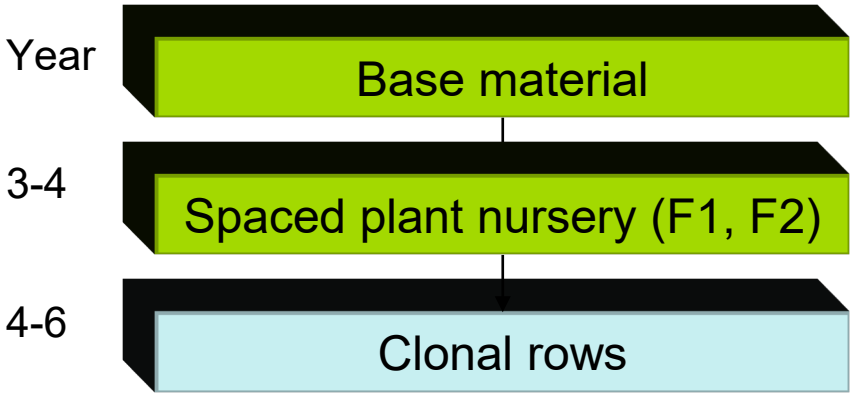
- To assess the effect of genetic diversity among parental plants on agronomic performance and diversity of polycross progenies

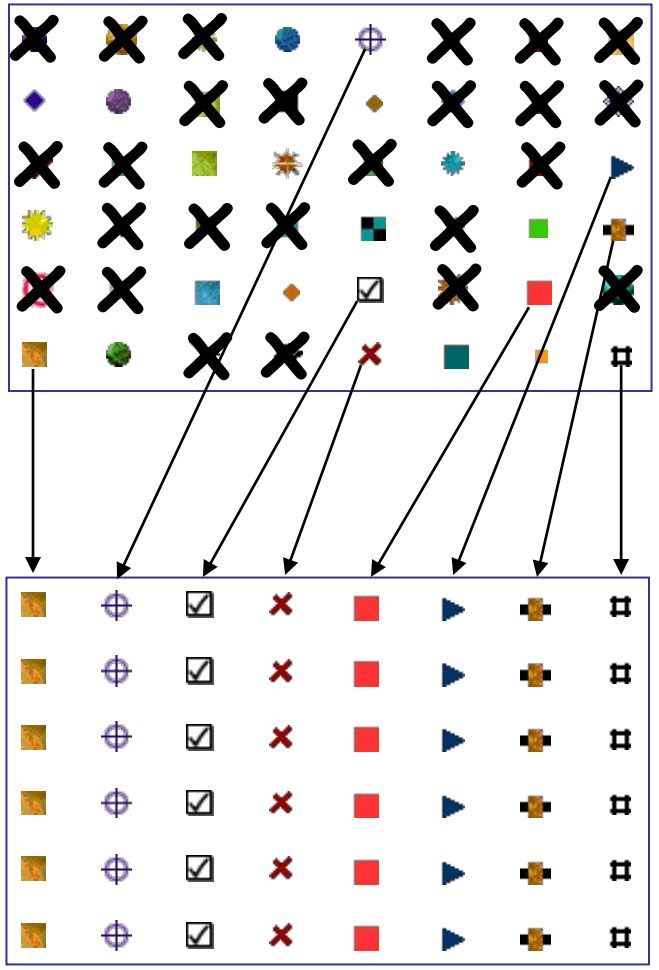
Polycross breeding



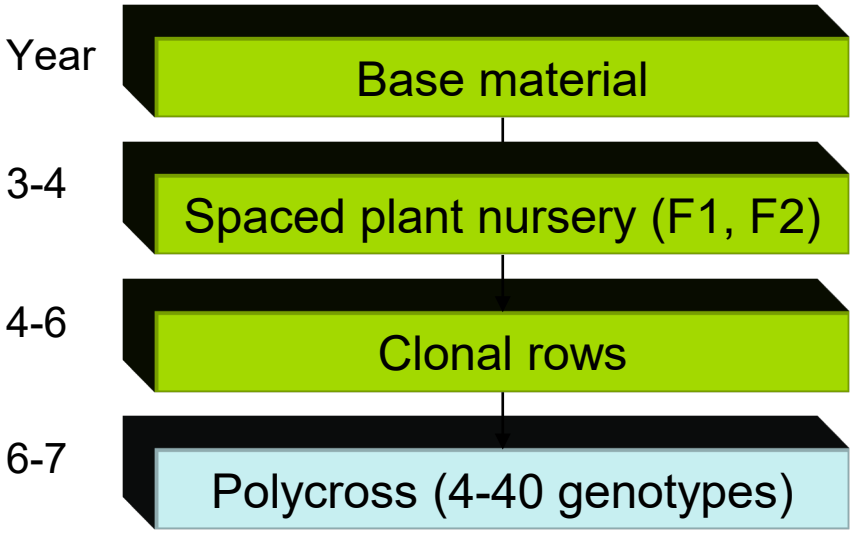


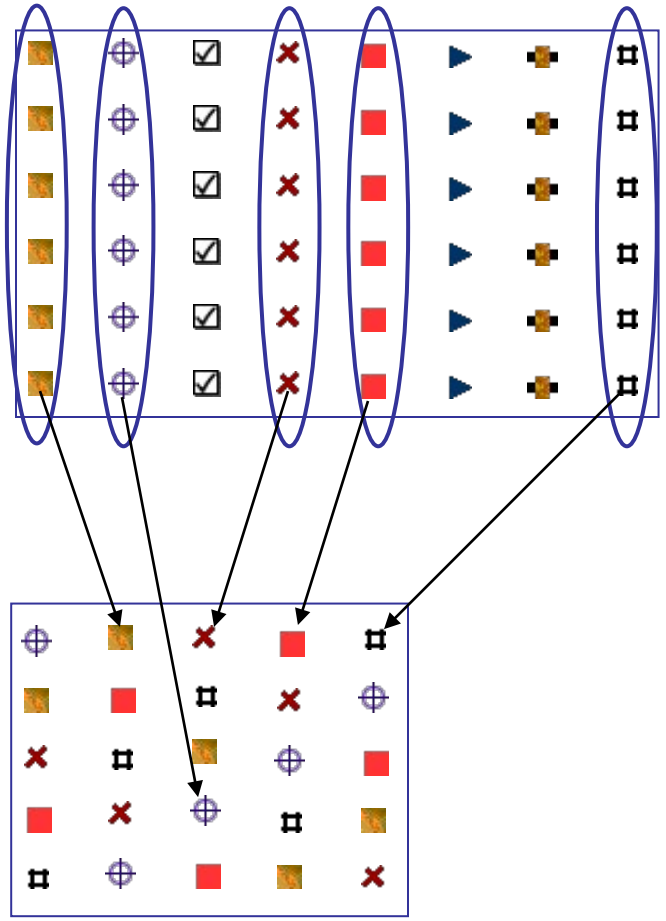
Polycross breeding



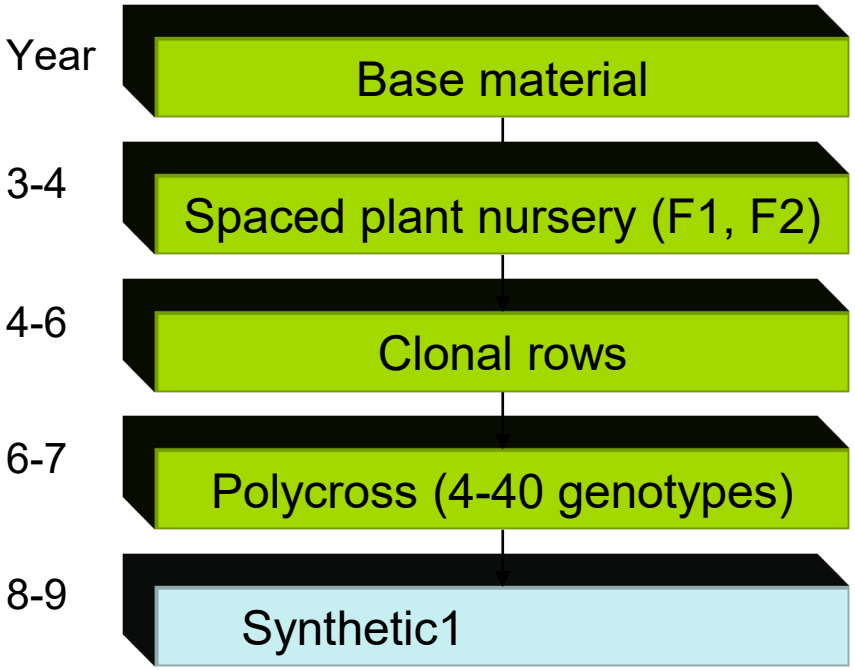


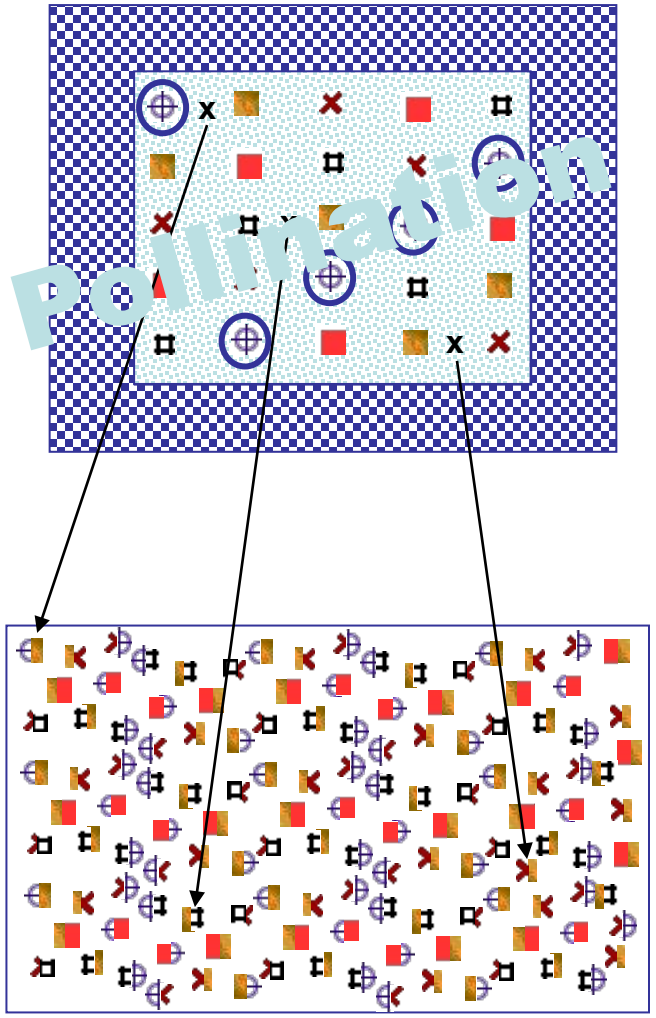
Polycross breeding



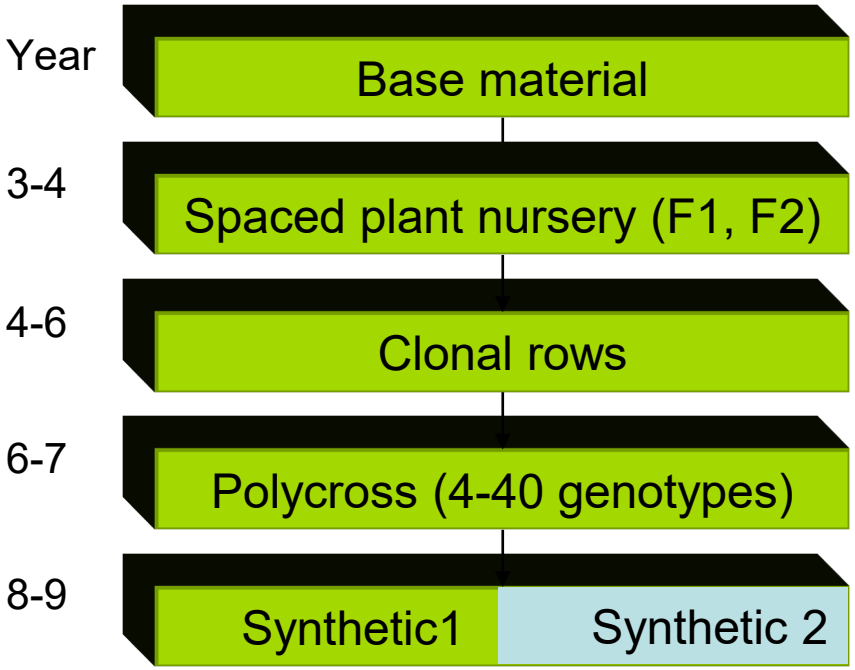


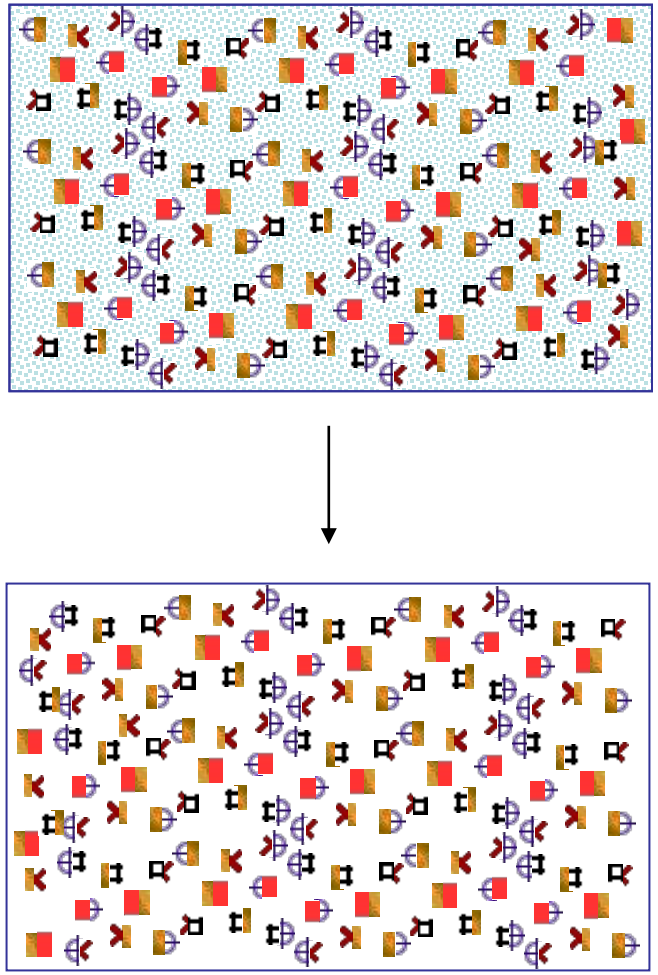
Polycross breeding



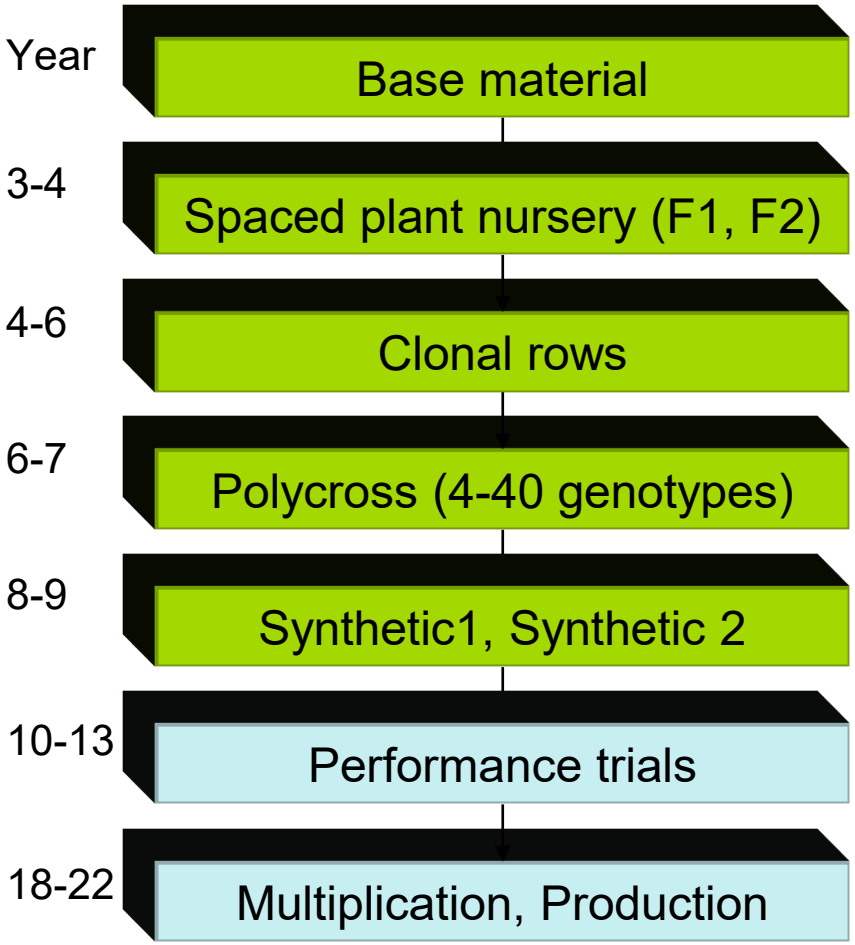


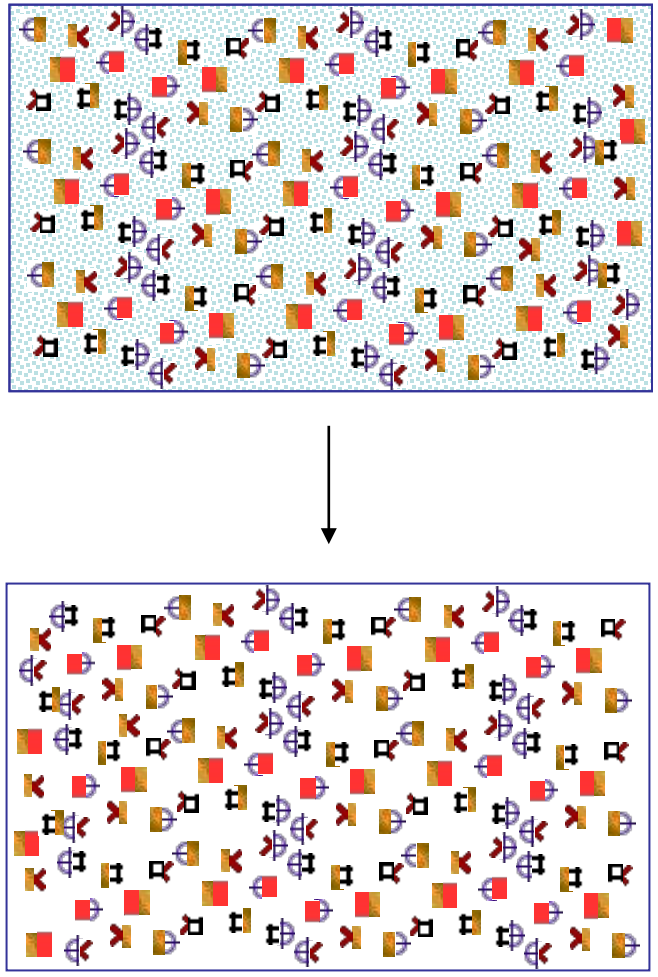
Polycross breeding



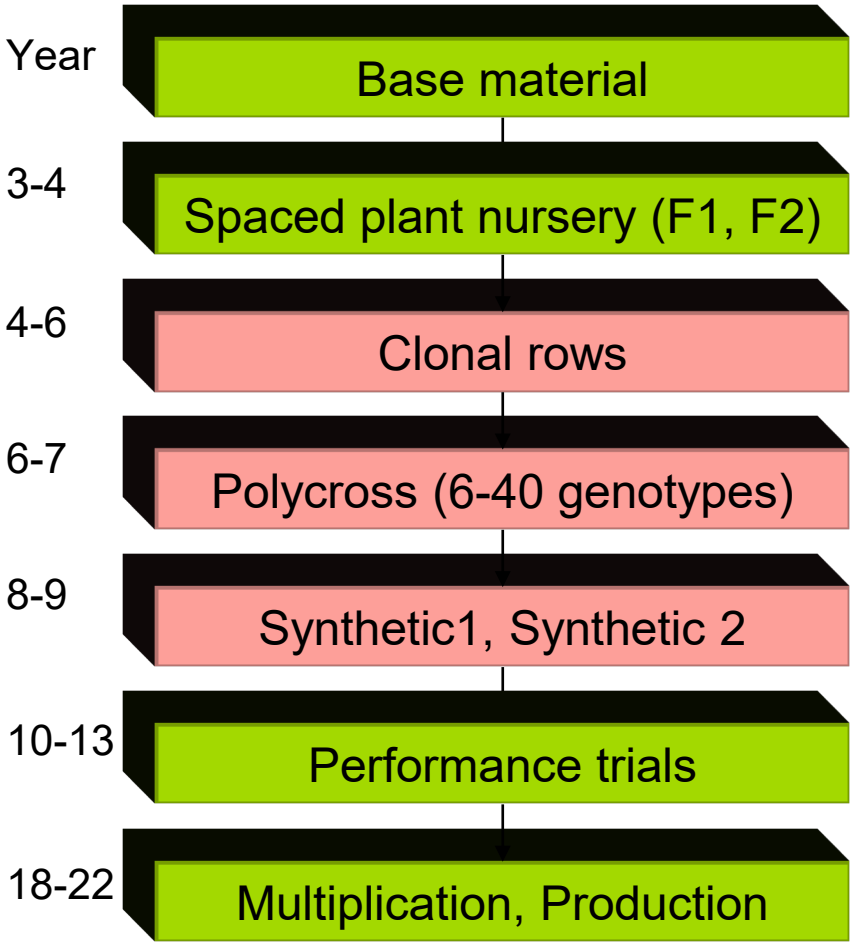


Polycross breeding





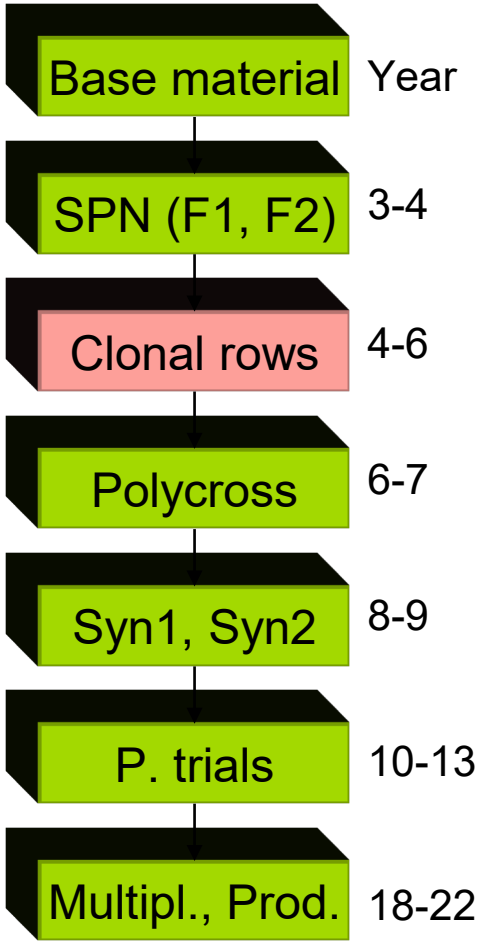
Experimental setup



- Molecular characterisation of potential parental plants
- Selection of parents based on AFLP marker diversity
- Genetic and phenotypic assessment of progenies

Selection of parental plants

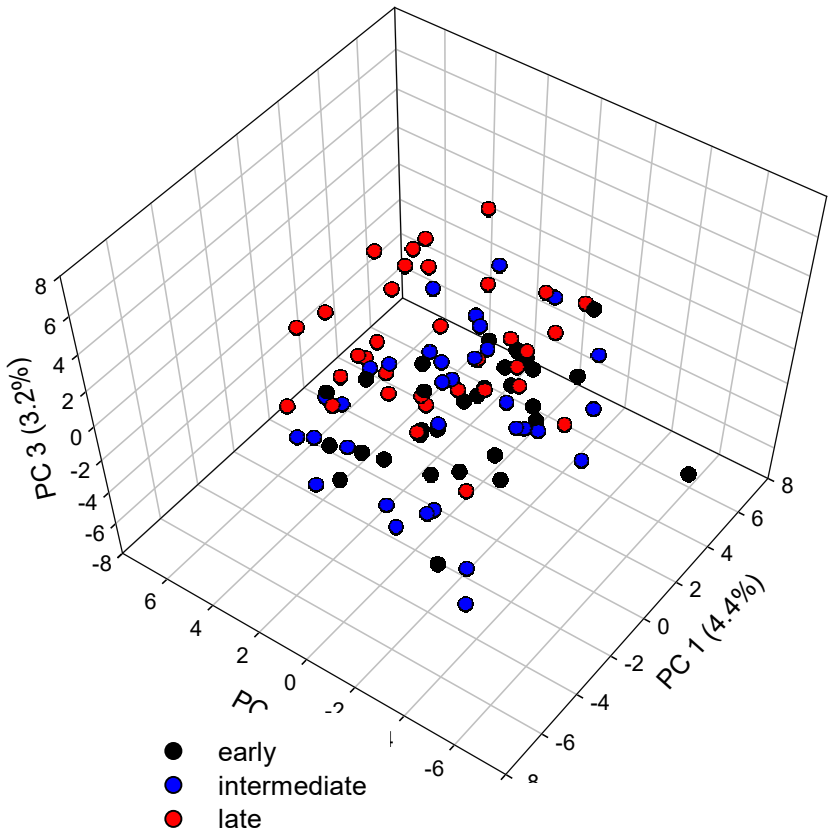
- **Plant material**
  - 98 perennial ryegrass plants
  - Advanced breeding germplasm
  - 3 groups (date of heading)
    - Early, intermediate, late
      - ~4 days difference between groups
      - All plants early flowering
- **Genetic diversity**
  - 184 AFLP polymorphic markers
  - Pairwise comparison of plants
  - Euclidean squared distance (E^2)
    - Marker diversity (E^2/No. of markers)
  - Multivariate analyses
  - Selection of parental plants based on genetic diversity



Diversity among parents

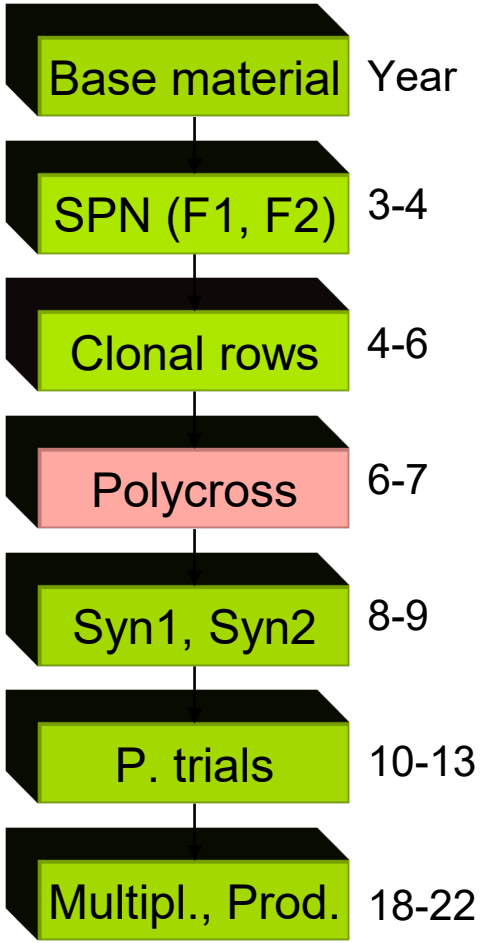
- No grouping of parental germplasm according to date of heading
- Considerable genetic diversity (E^2 = 51.7)
- AFLP results reflected pedigree information

Principle Component Analysis

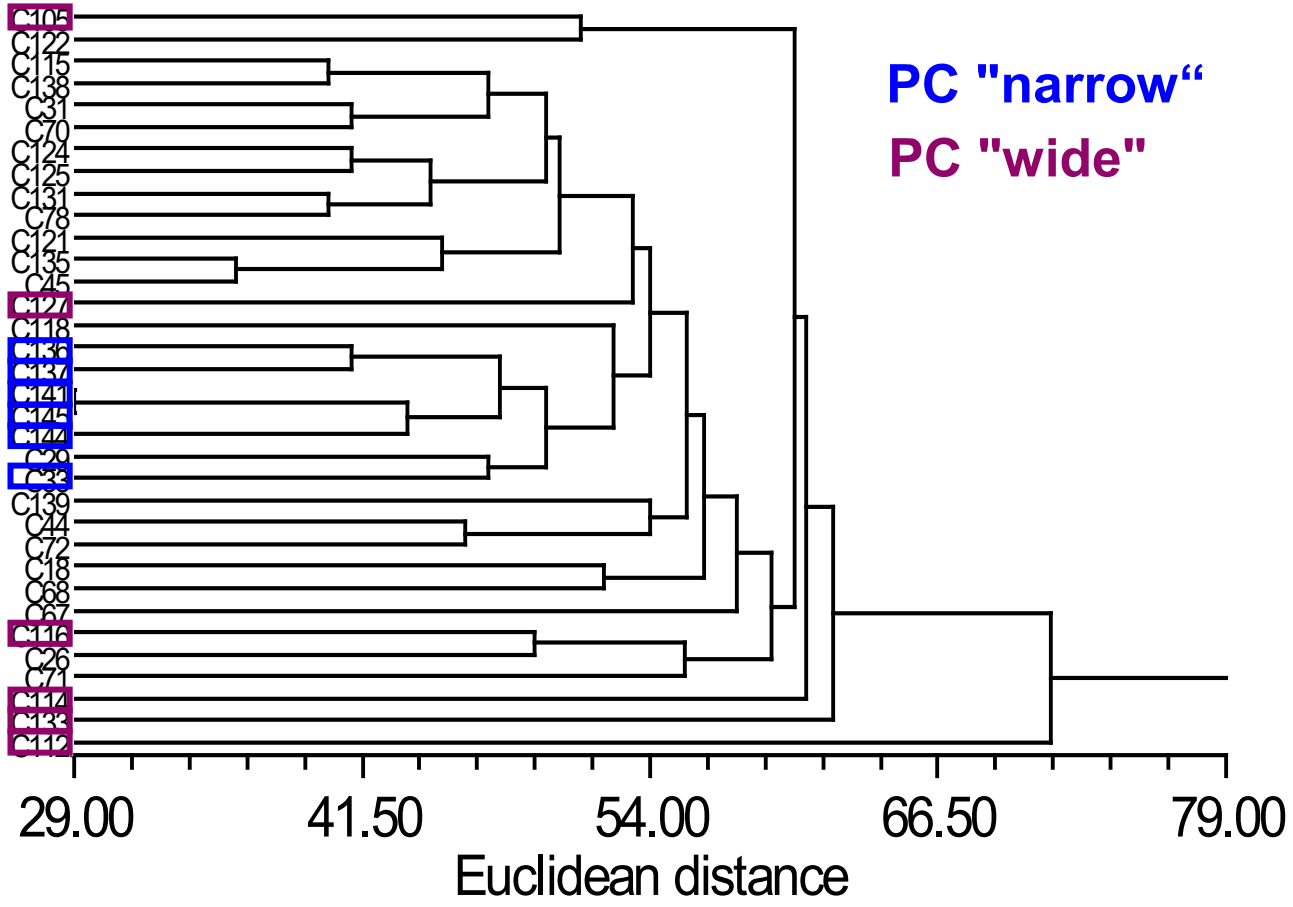


Selection of parental plants

- Polycrosses (PC) with different levels of genetic diversity
  - Selection of parental plants based on molecular markers
    - 6 closely related plants >> PC narrow
    - 6 more distantly related plants >> PC wide
    - 2 PC per group

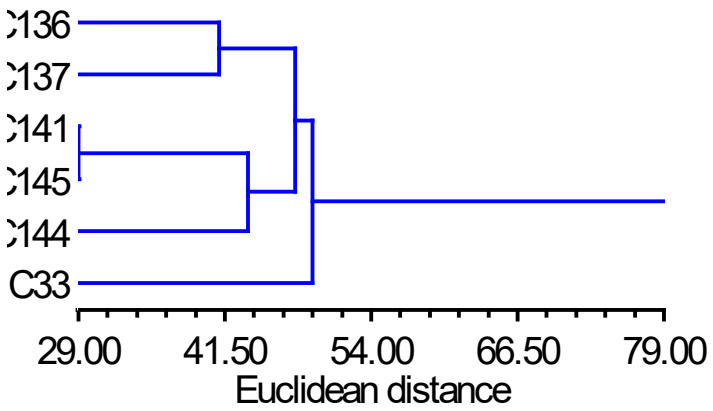


Cluster analysis group “late”

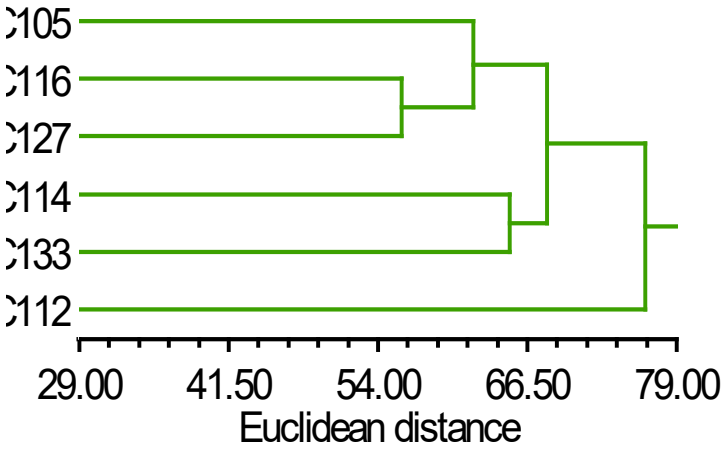


Cluster analysis of individual PC

PC "narrow"



PC "wide"



Cluster analysis based on 184 AFLP markers

Diversity among selected parental plants

	PC "narrow"		PC "wide"		Within group	
Groups	E^2	%	E^2	%	E^2	%
Early	34.1	18	60.4	33	50.7	28
Intermediate	42.2	23	59.3	32	52.5	29
Late	42.4	23	64.5	35	52.0	28
Average	39.6	22	61.4	33	51.7	28

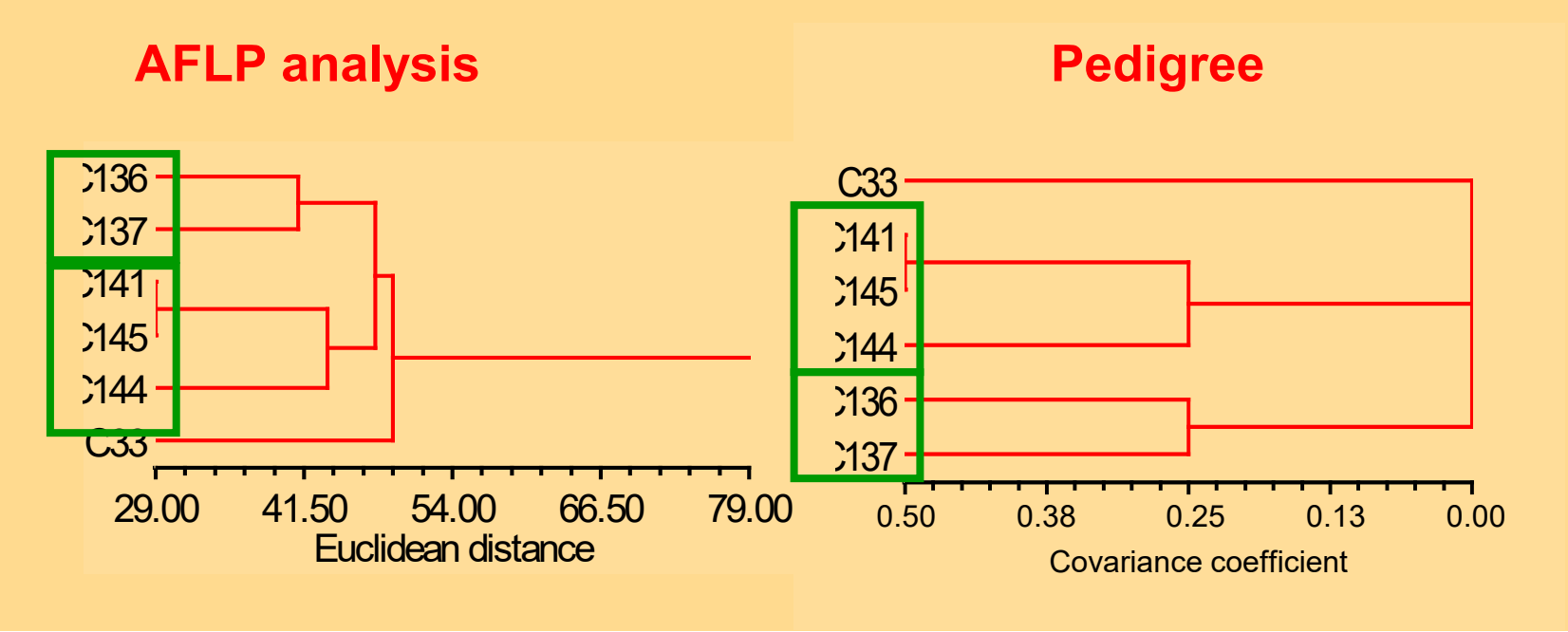
E^2=Euclidean distance, %=Marker diversity (E^2/no. of markers)

Diversity is considerably lower in "narrow" polycrosses

Validation of marker data

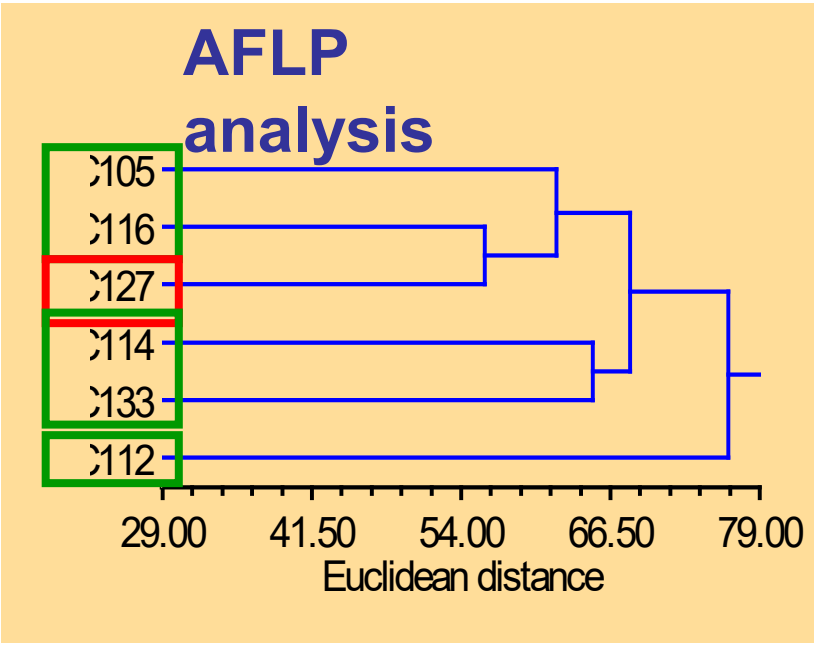
- Does AFLP diversity reflect the breeding history of the plants?
- Pedigree Information
  - Pair-crosses
  - Mutual pollination (only 1 parent known)
  - Self pollination
- Covariance coefficient
  - 2 x probability of a shared allele at a locus

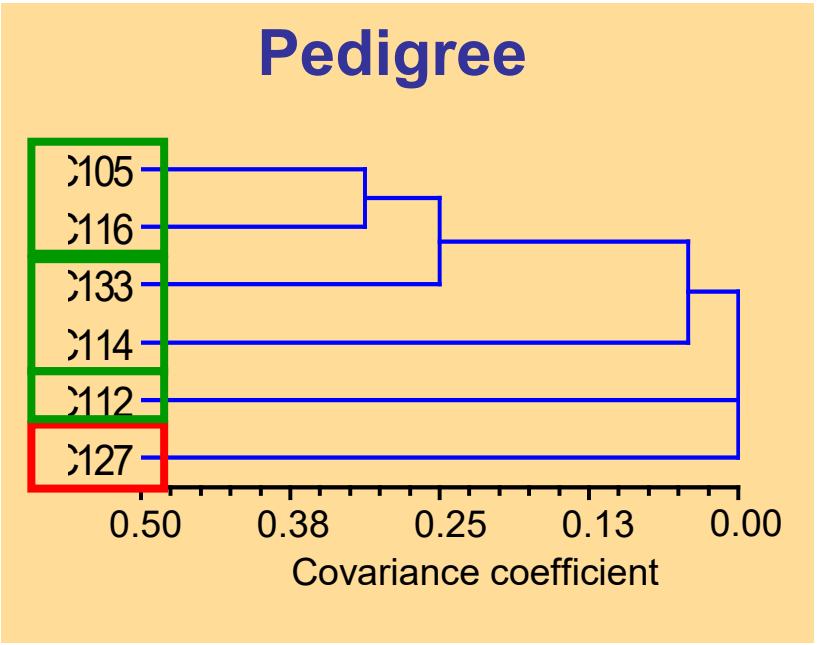
AFLP – pedigree: PC “narrow”



AFLP and pedigree data are consistent

AFLP – pedigree: PC “wide”





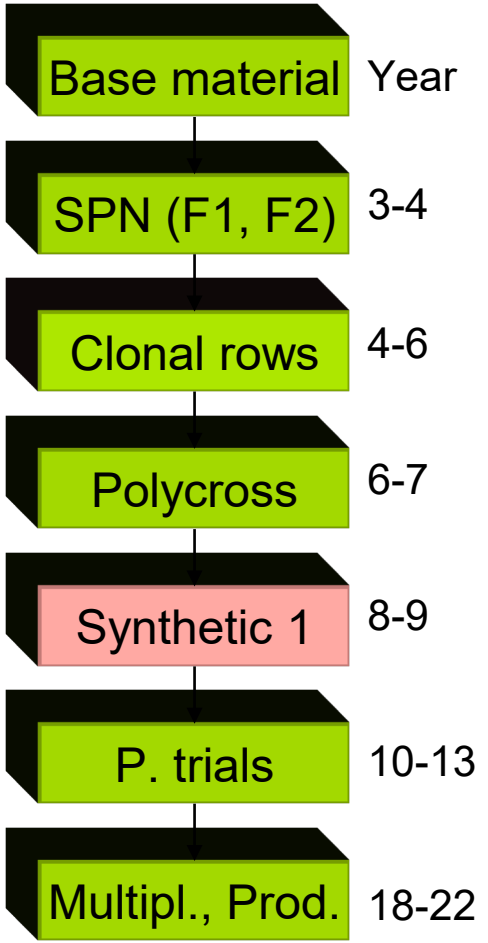
Group C

Partial consistence

Genetic diversity of progenies

- **216 Syn1 plants**
  - 36 progeny per PC
  - 6 progeny per motherplant

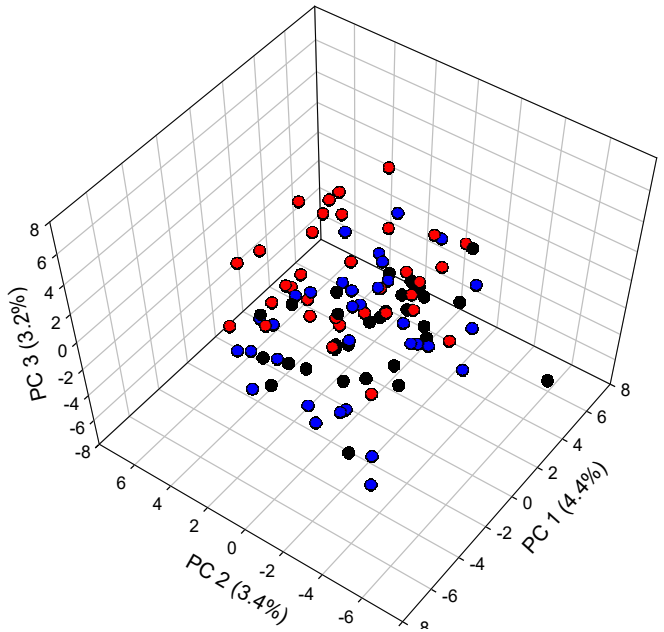
- **Genetic diversity**
  - 184 AFLP markers previously scored in parental plants
  - Euclidean squared distance (E^2)



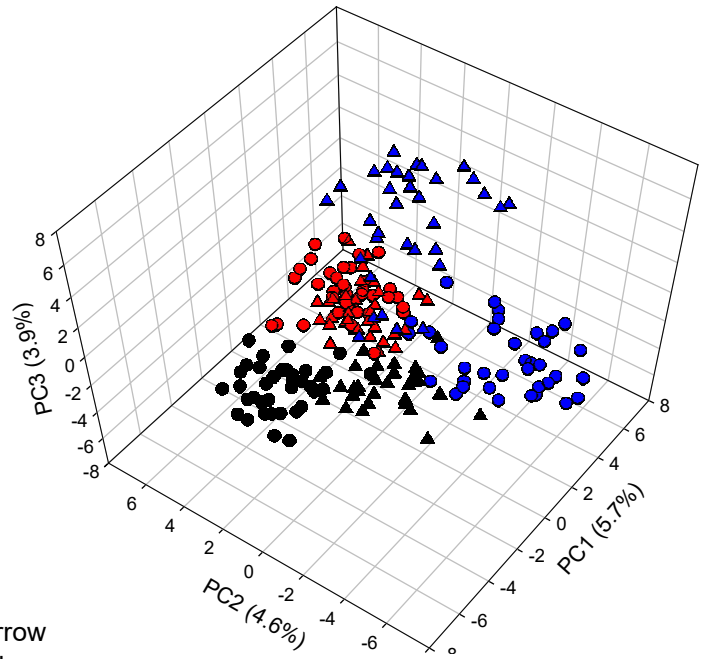
Separation of populations

Principle Component Analysis

Parental germplasm (98 plants)

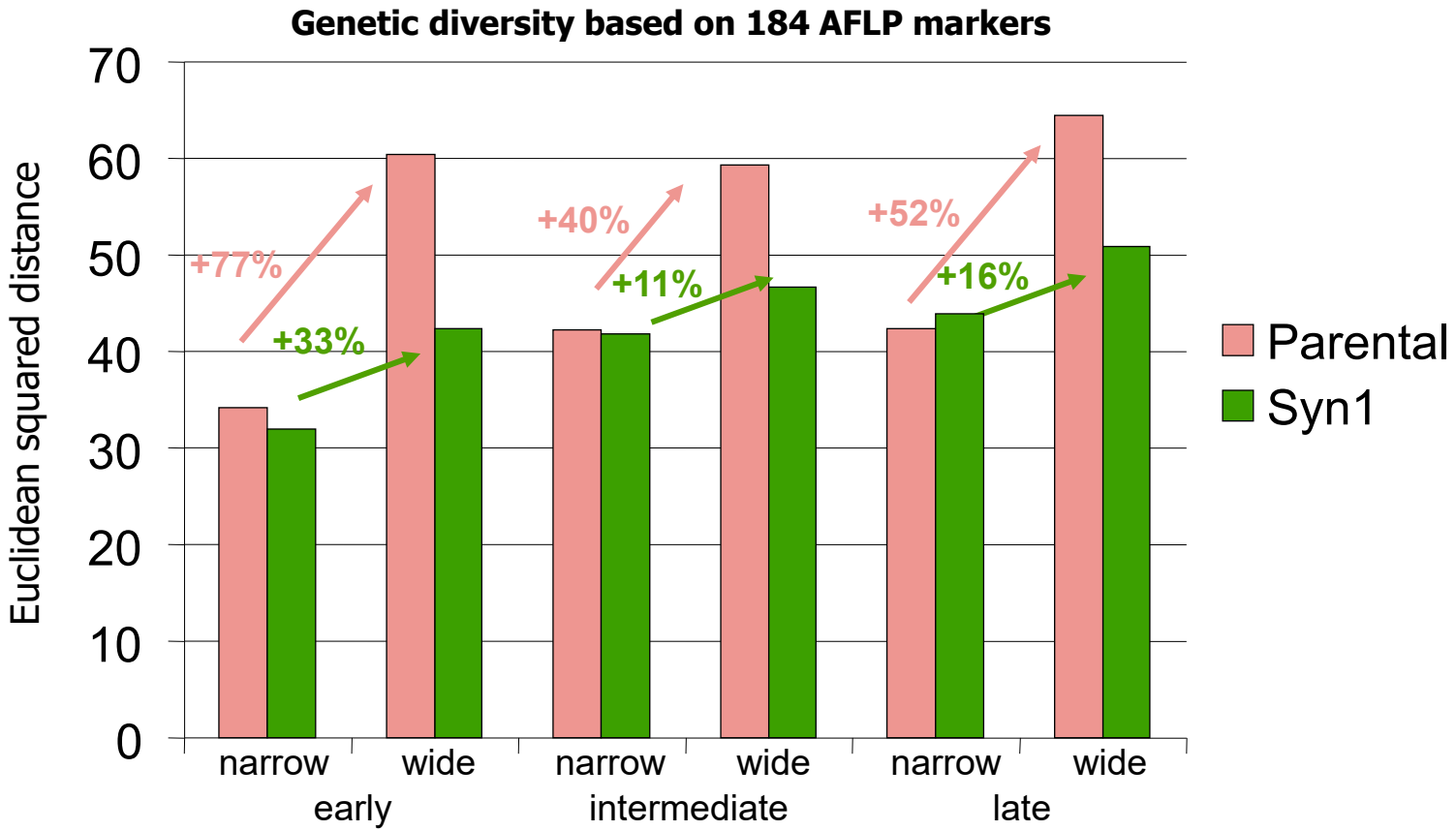


Syn1 progeny (216 plants)



- Early narrow
- Early wide
- Intermed narrow
- Intermed wide
- Late narrow
- Late wide

Diversity in parents and Syn 1



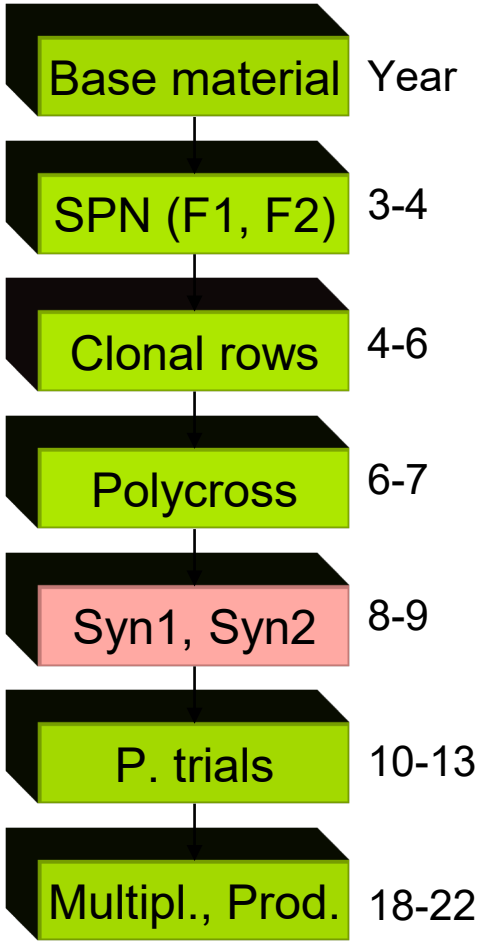
Clear differences between narrow and wide populations

Phenotypic characterisation

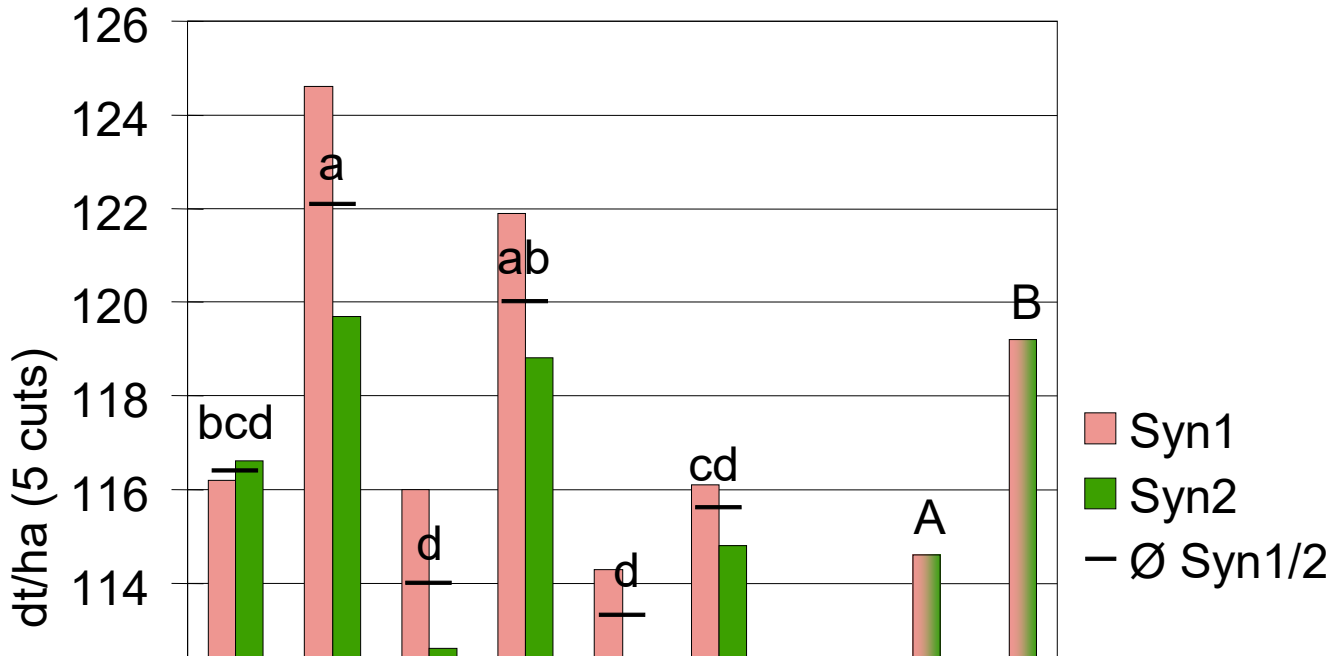
- Field trials with Syn1 and Syn2 populations in 2004

- **Agronomic performance**
  - Dry matter yield
  - Plot trials

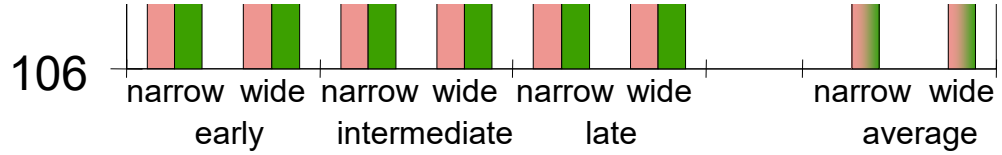
- **Uniformity**
  - Heading date (UPOV)
  - Spaced plant measurements



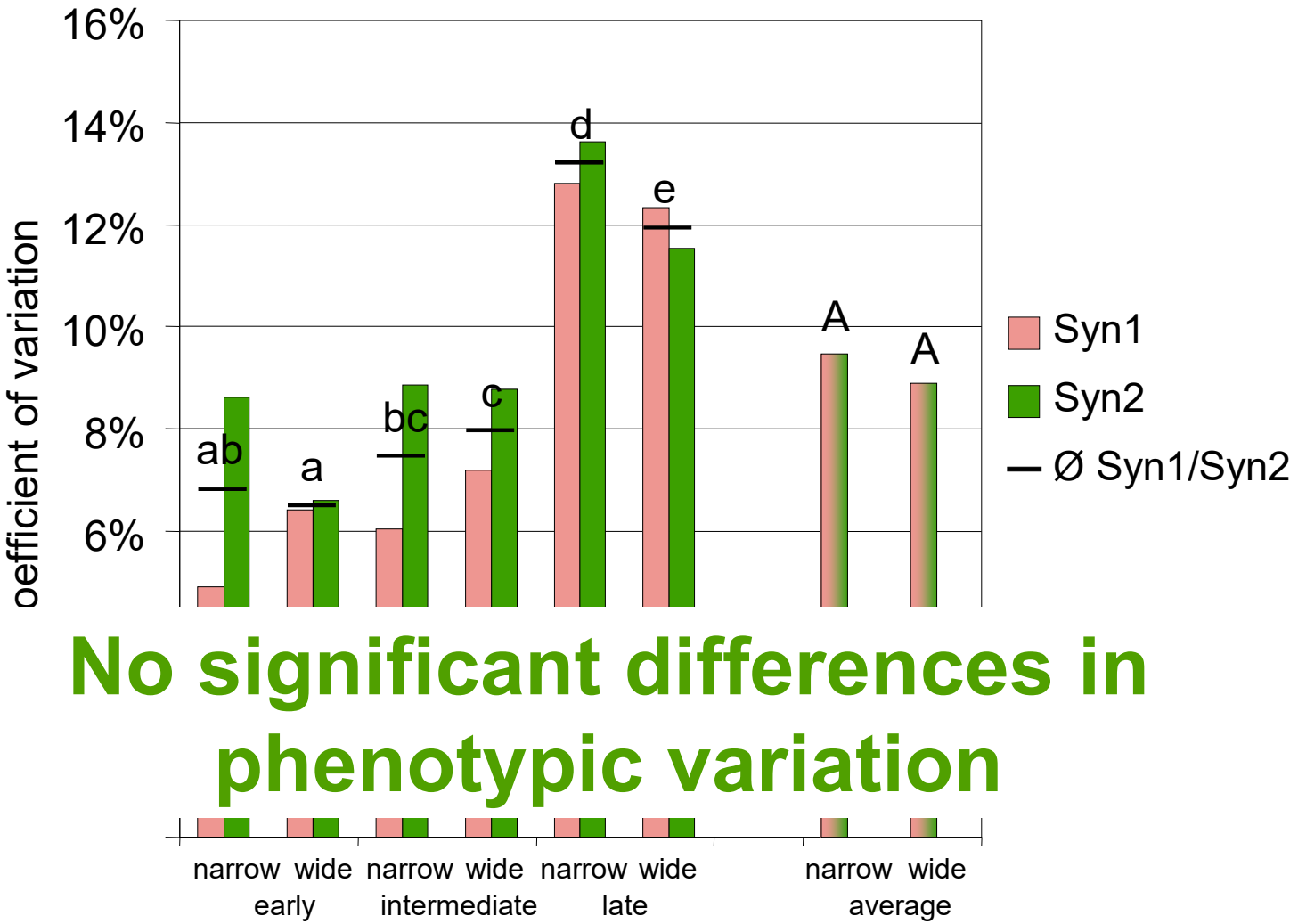
Dry matter yield



Average yield increase of 4% in wide polycross progenies



Phenotypic variation (heading date)



No significant differences in phenotypic variation

Conclusions

- AFLP markers allow for the selection of parental plants with different levels of genetic diversity
- Differences in genetic diversity are partially transmitted to Syn1 progenies
- High genetic diversity among polycross parents can have a positive effect on agronomic performance of progenies