
Select the
Future
Progeno
identifies genetic potential

Progeno
identifies genetic potential

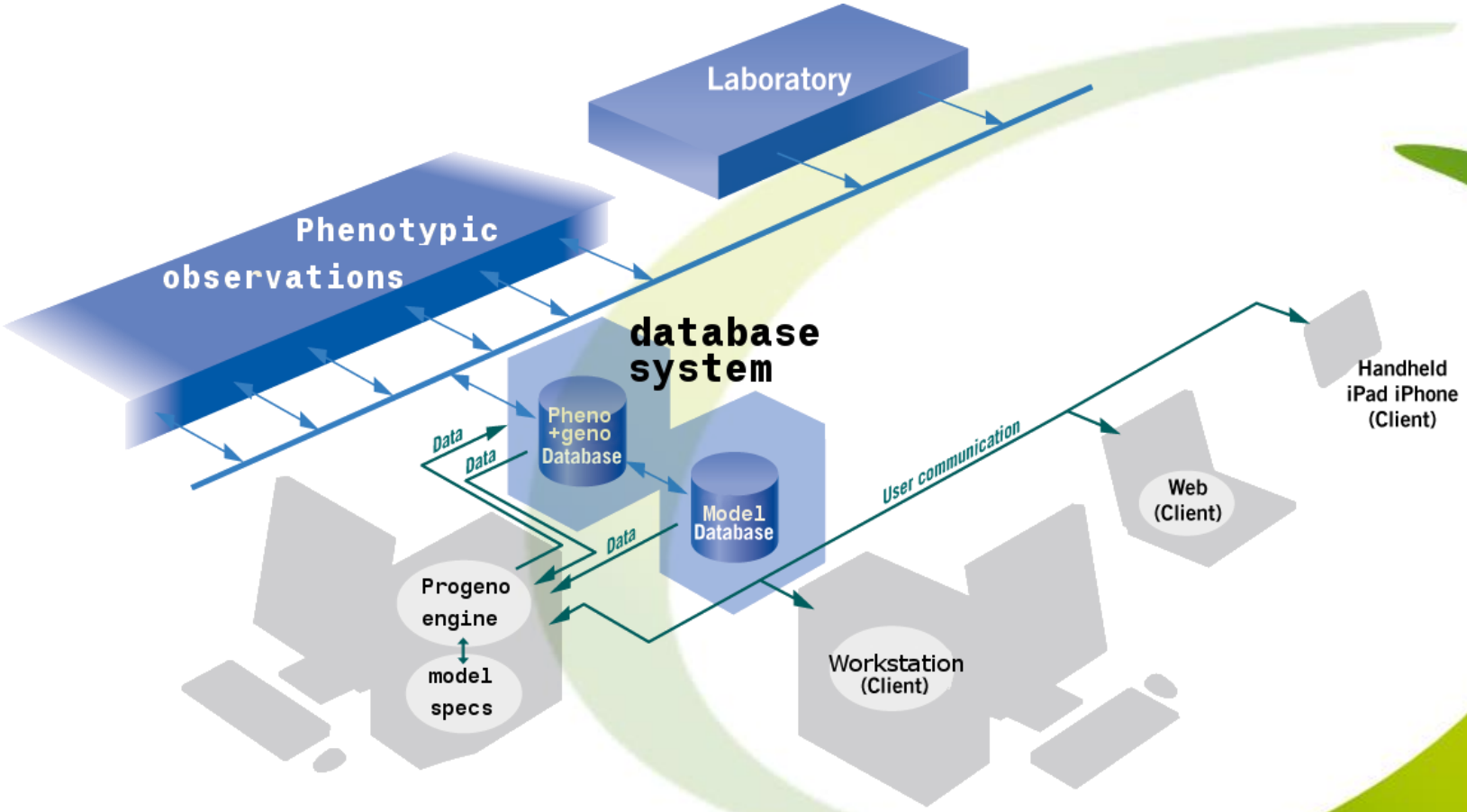
Progeno Analytics Training, May 20 / 26

Steven Maenhout

e-mail: Steven.Maenhout@progeno.net
site: https://www.progeno.net

Progeno
identifies genetic potential

Progeno architecture



Progeno
identifies genetic potential

Progeno workflow

Breeding activities

- germplasm, trait, protocol specifications
- trial design, observations, metadata

Analytics

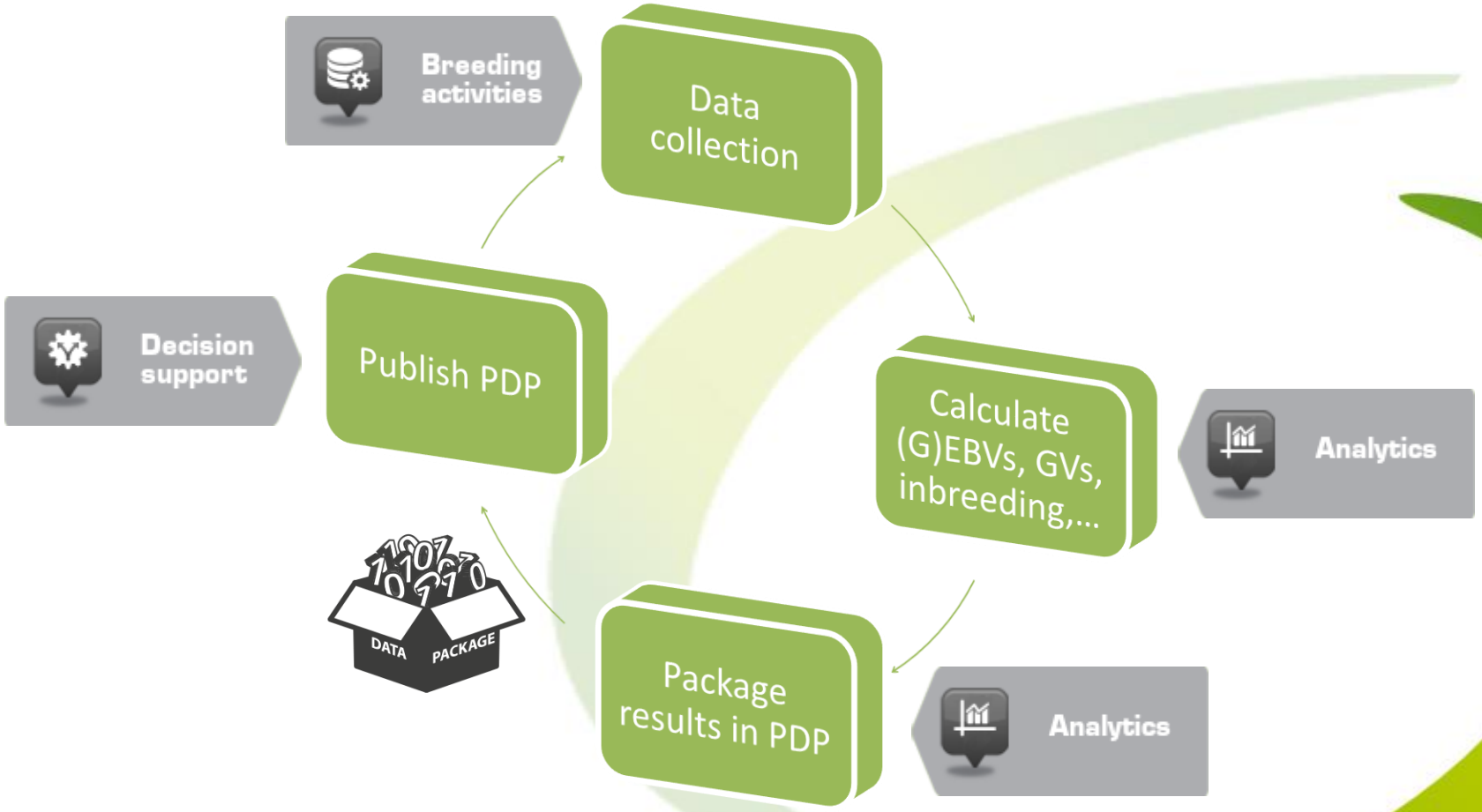
- phenotypic analysis, breeding values, …
- GWAS, genomic prediction, …

Decision support

- cross predictions, genomic prediction, marker-assisted selection
- pedigree browser, dendrogram, breeding pool visualisations

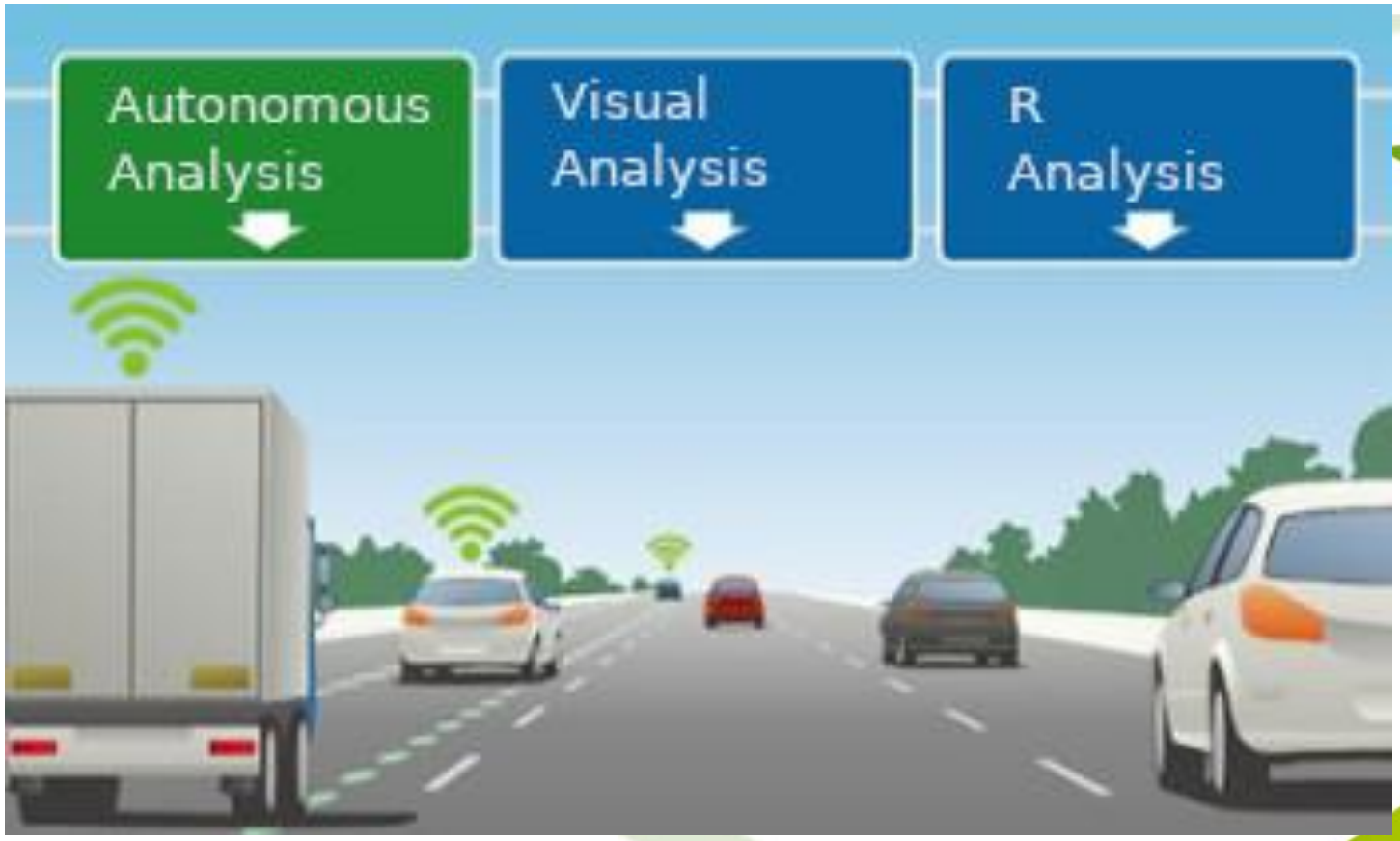
Progeno
identifies genetic potential

PDP = Processed Data Package



Progeno
identifies genetic potential

Progeno Analytics



Progeno
identifies genetic potential

Lane 1: Autonomous data analysis

- **Autonomous**: automatically processes incoming trial and marker data to breeding values
- **Self-learning**: uses all available data for each analysis, (genomic) breeding values become more reliable as new data is added.
- **Correcting**: detects and removes outliers in the observations
- **Adapting**: analysis models adjust automatically to the properties of the breeding pool

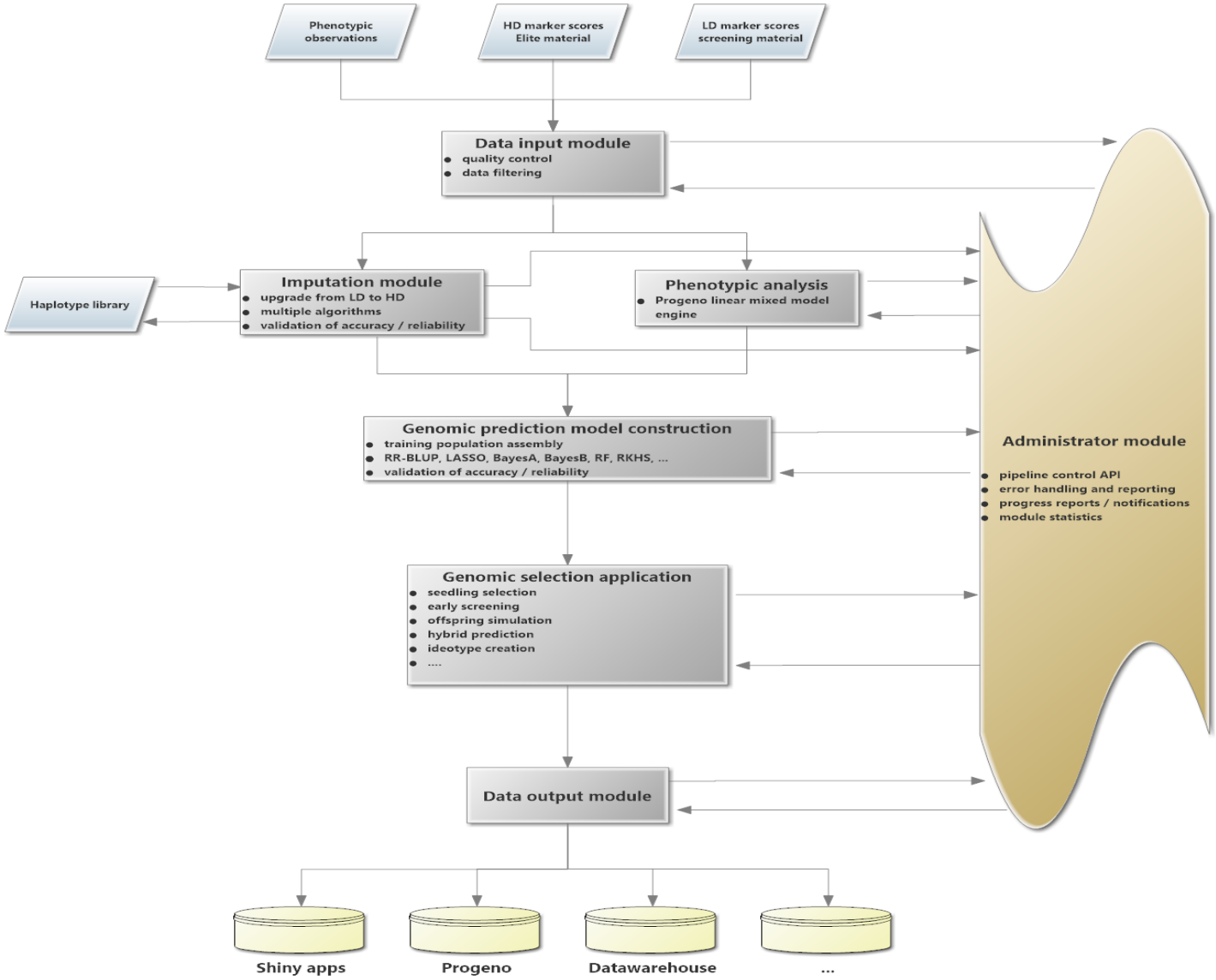
Progeno
identifies genetic potential

Lane 2: Progeno Analytics module

- provides a range of **statistical analysis** and **visualisation** techniques:
  - histograms, density plots, scatter plots, blox plots, GGE biplots, 3D scatter or surface plots, correlations, …
  - linear models (ANOVA, t-test, regression, …), mixed models, genomic predictions, …
- direct connection to the **Progeno Datawarehouse**
- direct connection to **R** via **Rstudio Server**
- save datasets, analysis results and reports to personal or shared workspace
- **package and publish breeding values** as custom PDPs

Progeno
identifies genetic potential

Lane 3: Progeno-R:



Progeno
identifies genetic potential

Database query

- each phenotypic value has at least 3 coordinates:
  1. **who**: the physical **position** of the observed accession / plot: row, column, trial, location, …
  2. **when**: the date/**time** of the observation
  3. **what**: the observed **trait**
- multiple ways to represent 3-dimensional observations in a 2-dimensional sheet

Progeno
identifies genetic potential

3D => 2D

EuclegID	Row	Column	2019_11_28_Plantlength	2019_11_28_Seednumber
EUC_TU_001	3	10	104.8	85.4
EUC_TU_001	3	3	98.8	69.2
EUC_TU_001	1	1	76	40.6
EUC_TU_001	2	6	81.8	55.6
EUC_TU_002	3	8	115.4	49.4
EUC_TU_002	3	1	82.2	35.4
EUC_TU_002	2	5	80.4	38.2
EUC_TU_002	3	6	86.4	34
EUC_TU_003	3	7	97.2	39.8
EUC_TU_003	1	3	87.6	41.2
EUC_TU_004	1	9	99.4	52
EUC_TU_004	2	3	86.2	49.6
EUC_TU_005	1	7	81.8	99
EUC_TU_005	2			
EUC_TU_006	1			

EuclegID	Row	Column	Year	Month	Day	Plantlength	Seednumber
EUC_TU_001	3	10	2019	11	28	104.8	85.4
EUC_TU_001	3	3	2019	11	28	98.8	69.2
EUC_TU_001	1	1	2019	11	28	76	40.6
EUC_TU_001	2	6	2019	11	28	81.8	55.6
EUC_TU_002	3	8	2019	11	28	115.4	49.4
EUC_TU_002	3	1	2019	11	28	82.2	35.4
EUC_TU_002	2	5	2019	11	28	80.4	38.2
EUC_TU_002	3	6	2019	11	28	86.4	34
EUC_TU_003	1	3	2019	11	28	87.6	41.2
EUC_TU_003	3	7	2019	11	28	97.2	39.8
EUC_TU_004	2	3	2019	11	28	86.2	49.6
EUC_TU_004	1	9	2019	11	28	99.4	52
EUC_TU_005	1	7	2019	11	28	81.8	99
EUC_TU_005	2	1	2019	11	28	80.8	99.4
EUC_TU_006	2	8	2019	11	28	104.2	-

Progeno
identifies genetic potential

Exercise Database query

1. What is the largest observed value for the trait Plantlength? Which accession in which trial?
2. Calculate the average observed “Thousandseedweight” for accessions “Tommy” and “William” per trial. Put the results side-by-side (Name and Year in rows, accessions in columns).
3. Which trial shows the lowest average Seedyield scores when only considering German accessions?
4. List the median “Rstage” for accession Munro for all available months and trials. In which trial / month was the highest median observed?

Exercise Box-plot

- Create dataset that contains all “SeedYield” observations from trial “ART trial 2018”. Add the “ProvenanceCountry” of each accession to this dataset and make a boxplot showing the SeedYield distribution for each ProvenanceCountry.

Exercise scatter plot

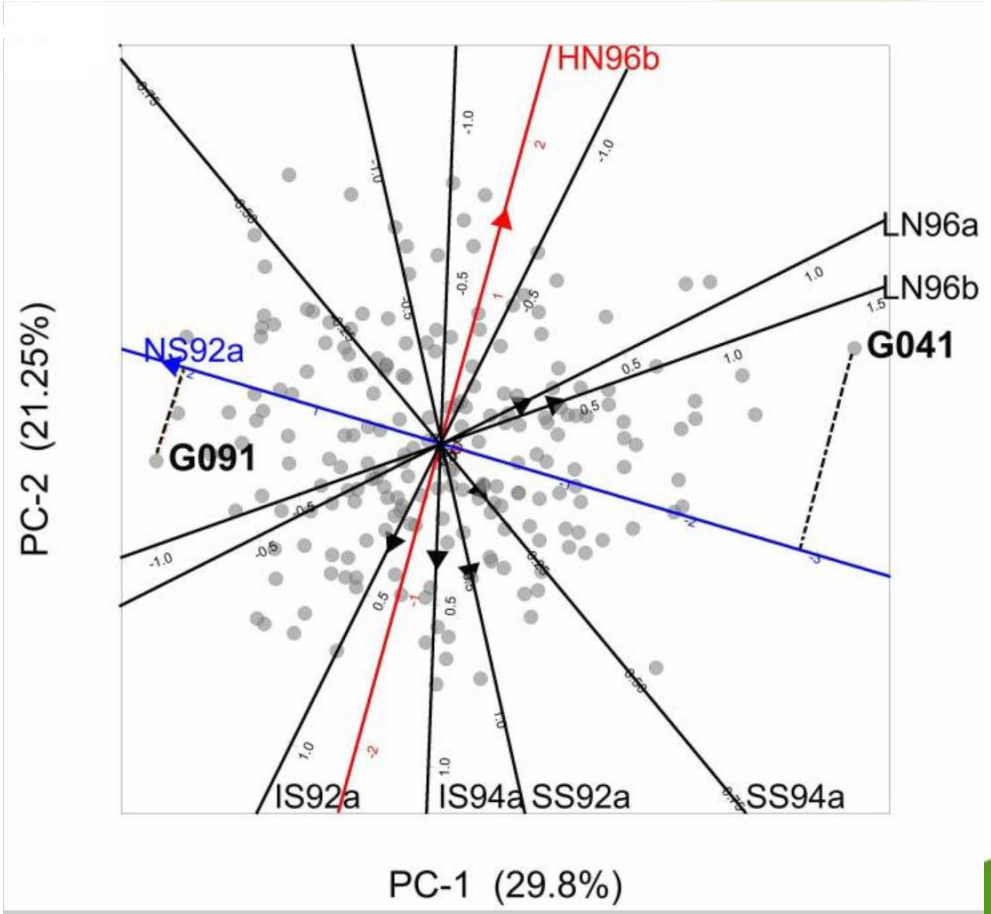
- Which trial seems to indicate a positive correlation between the traits Seedyield and Proteincontent?

Exercise 3D + heatmap

- Extract the “Plantlength” observations from trial “Vuk Djordjevic” including the row and column coordinates. Examine the spatial distribution with three different plot types:
  - 3D scatter
  - 3D surface
  - Heatmap
- Which plot type is the most informative in your opinion?

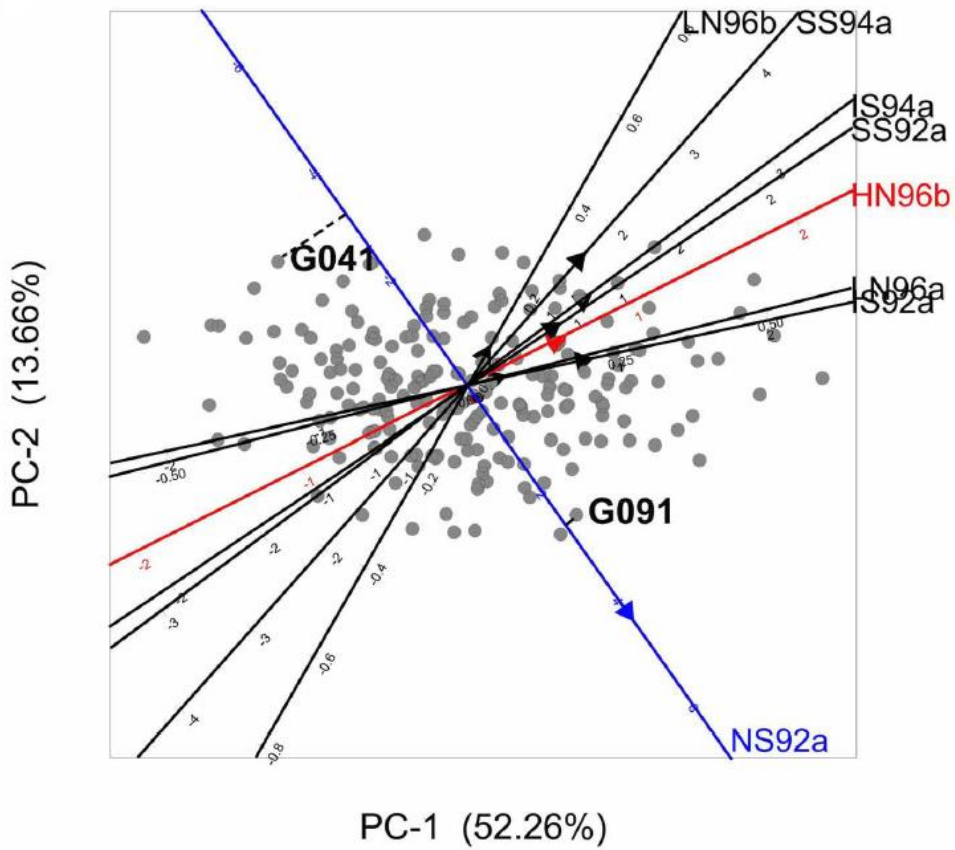
AMMI: Additive Main effects and Multiplicative Interactions model

$$\underline{\mu}_{ij} = \mu + \mathrm{G}_i + \mathrm{E}_j + \sum_{k=1}^{\mathrm{K}} \mathrm{b}_{ik}\mathrm{z}_{jk} + \underline{\varepsilon}_{ij}$$



GGE: Genotype main effects and GEI model

$$\underline{\mu}_{ij} = \mu + E_j + \sum_{k=1}^{K} b_{ik} z_{jk} + \underline{\varepsilon}_{ij}$$

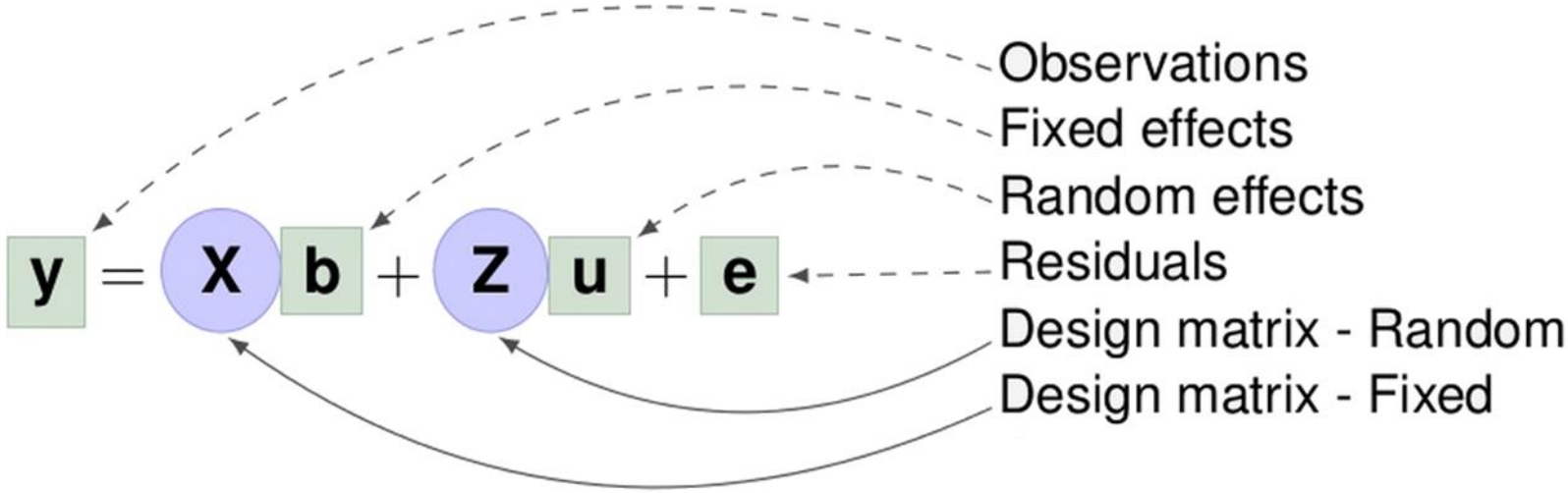


Exercise correlation matrix

- Calculate the correlation matrix for the traits Plantlength, Proteincontent, Seednumber, SeedYield and the row/column coordinates of trial ART2018.

Single trial analysis Mixed Model

- Incomplete block design, partial replication:
  - mixed model approach
  - focus on getting a good model fit
  - various options for modelling spatial trends



with $\mathrm{Var}(\mathbf{e}) = \mathbf{R}$ & $\mathrm{Var}(\mathbf{u}) = \boldsymbol{\Sigma}$

Candidate model fitting

- Base model:
  - fixed effects:
    - *trial mean*
    - *full replication*
  - random effects:
    - *treatment (genotype)*
- candidate random effects => use AIC criterion
  - incomplete block
  - row
  - column
- candidate spatial models => use AIC criterion
  - none (identity residual variance)
  - AR1 x AR1

$$y = mu + rep + genotype + e$$

Spatial correlations

- First-order autoregressive in rows

$$\mathbf{R}_{\text{row}} = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{21} \\ \rho & 1 & \rho & \dots & \rho^{20} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{21} & \rho^{20} & \rho^{19} & \dots & 1 \end{bmatrix}$$

- If we assume no correlations between columns

$$\mathbf{R}_{\text{col}} = \mathbf{I}_5$$

- then

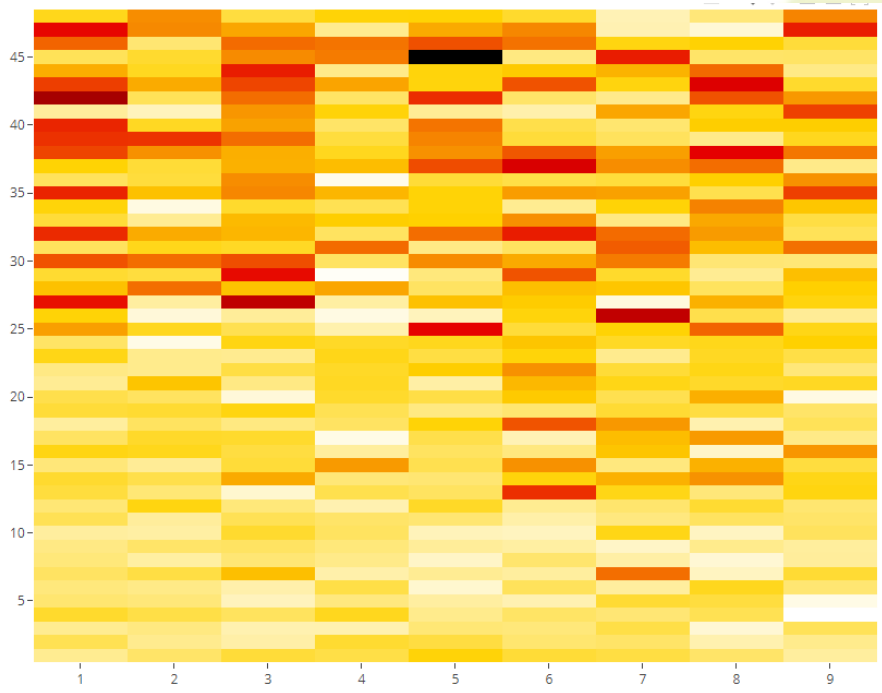
$$\mathbf{R} = \sigma_e^2(\mathbf{R}_{\text{col}} \otimes \mathbf{R}_{\text{row}})$$

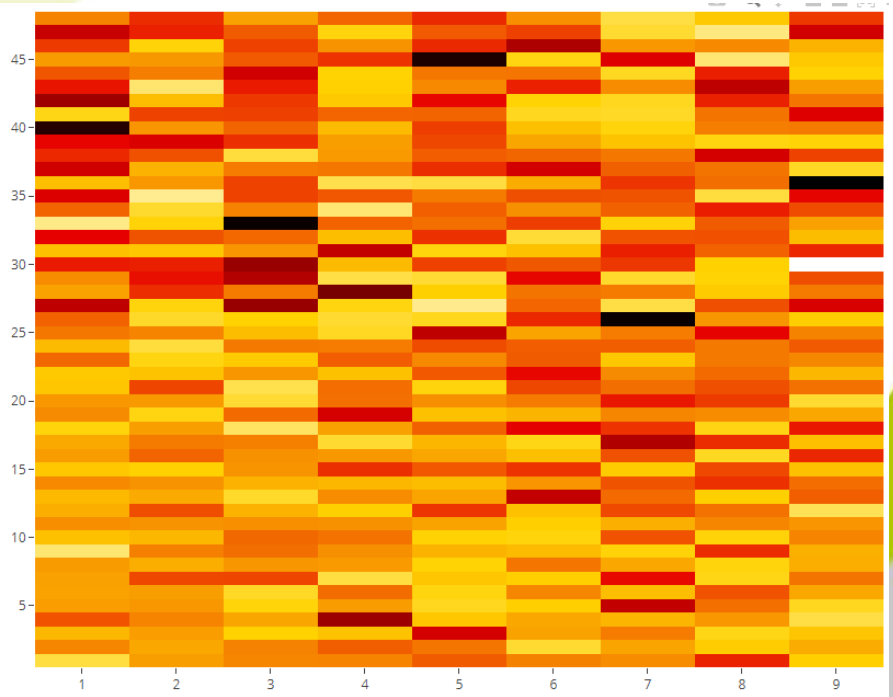
Model Comparison

- information criteria like (AIC, BIC) generally preferred over likelihood ratio tests
- AIC = $-2L + 2t$
- BIC = $-2L + t \log n$
- t = # of variance parameters
- n = # of residual degrees of freedom =
- (nobservations - nfixed_parameters).
- best model has smallest AIC/BIC
- if using REML estimates AIC/BIC only allow to compare models that have the same fixed effects structure

Trait ‘Plantlength’ of trial ‘Vuk Djordjevic’

Model	AIC
Base model: y = mu + genotype	2957.08
+ random row	2949.05
+ random column	2948.93
+ AR1 x AR1 residual	3544.55





Progeno
identifies genetic potential

Automated trial / trait model search

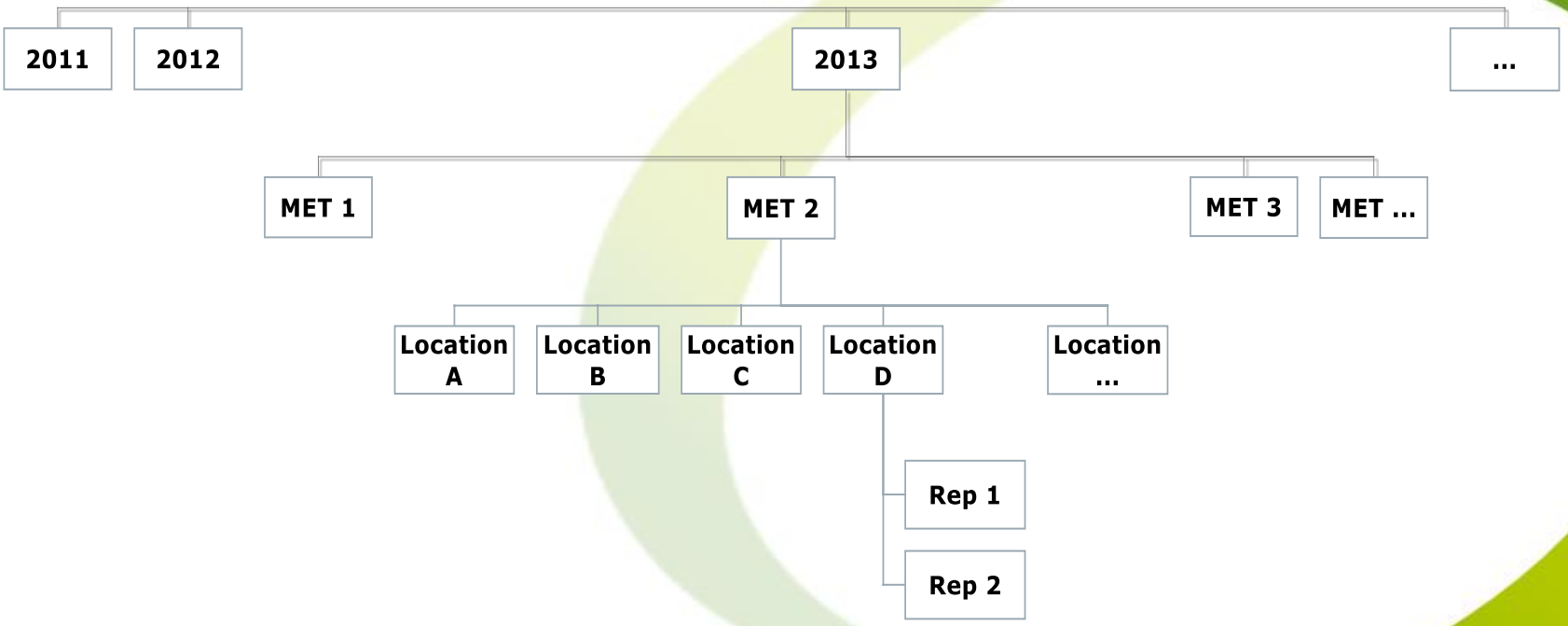
- fits all combinations of model term candidates
- chooses candidate model with lowest AIC
- allows to extract genotypic effects (BLUPs) and residuals of best-fitting model

Exercise single trial model

- Find the best-fitting linear mixed model to analyse the trait “Seedyield” of trial “ART trial 2019”
- Make a scatter plot having the BLUPs of the accession on the X-axis and their standard error on the Y-axis.
- Can you explain the different levels of standard errors of the BLUPs?

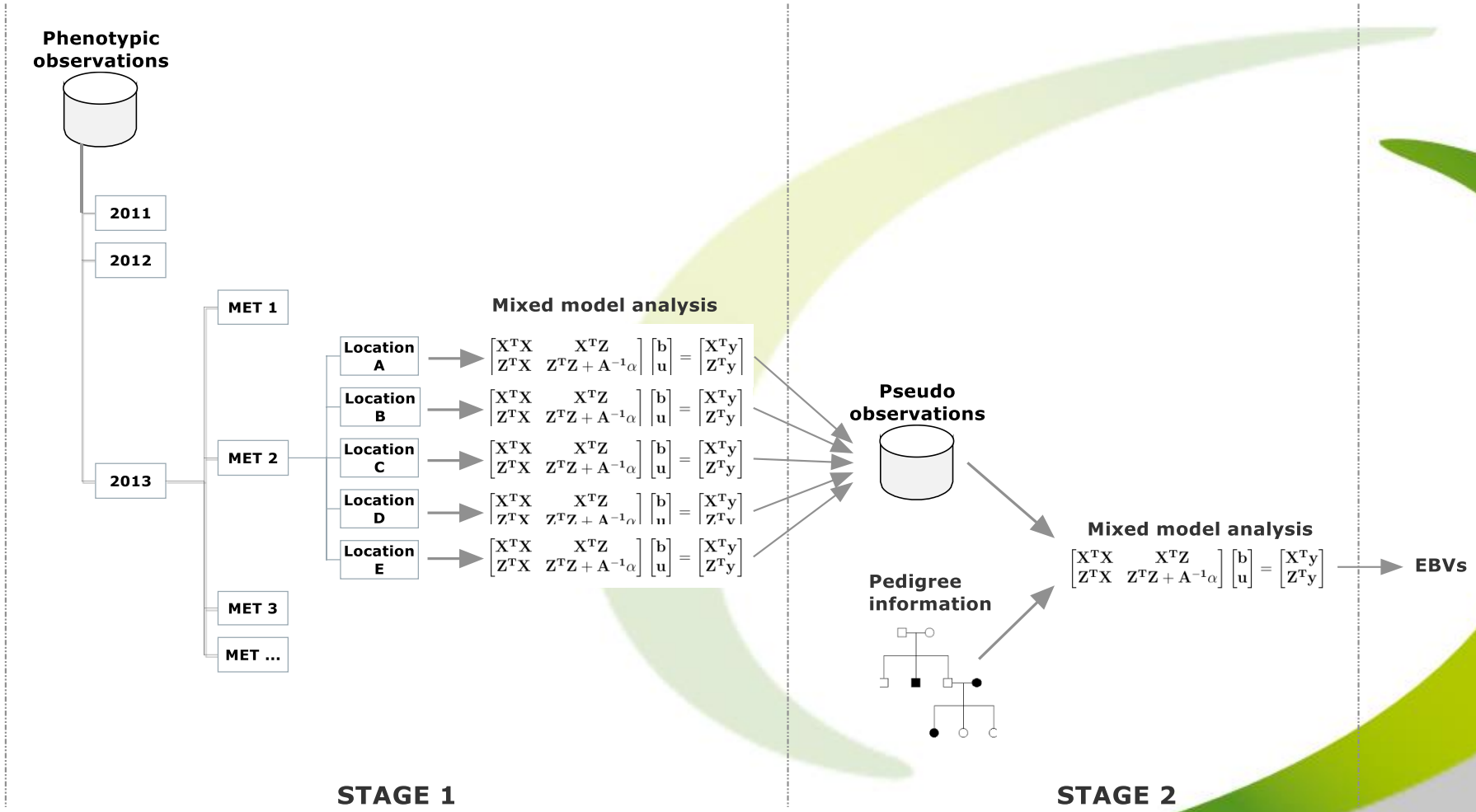
Multi-environment trial analysis

- plant breeding trials generally organised in multi-year / multi-environment trials (MET)



Mixed model approach: Two-stage MET analysis

2-stage procedure

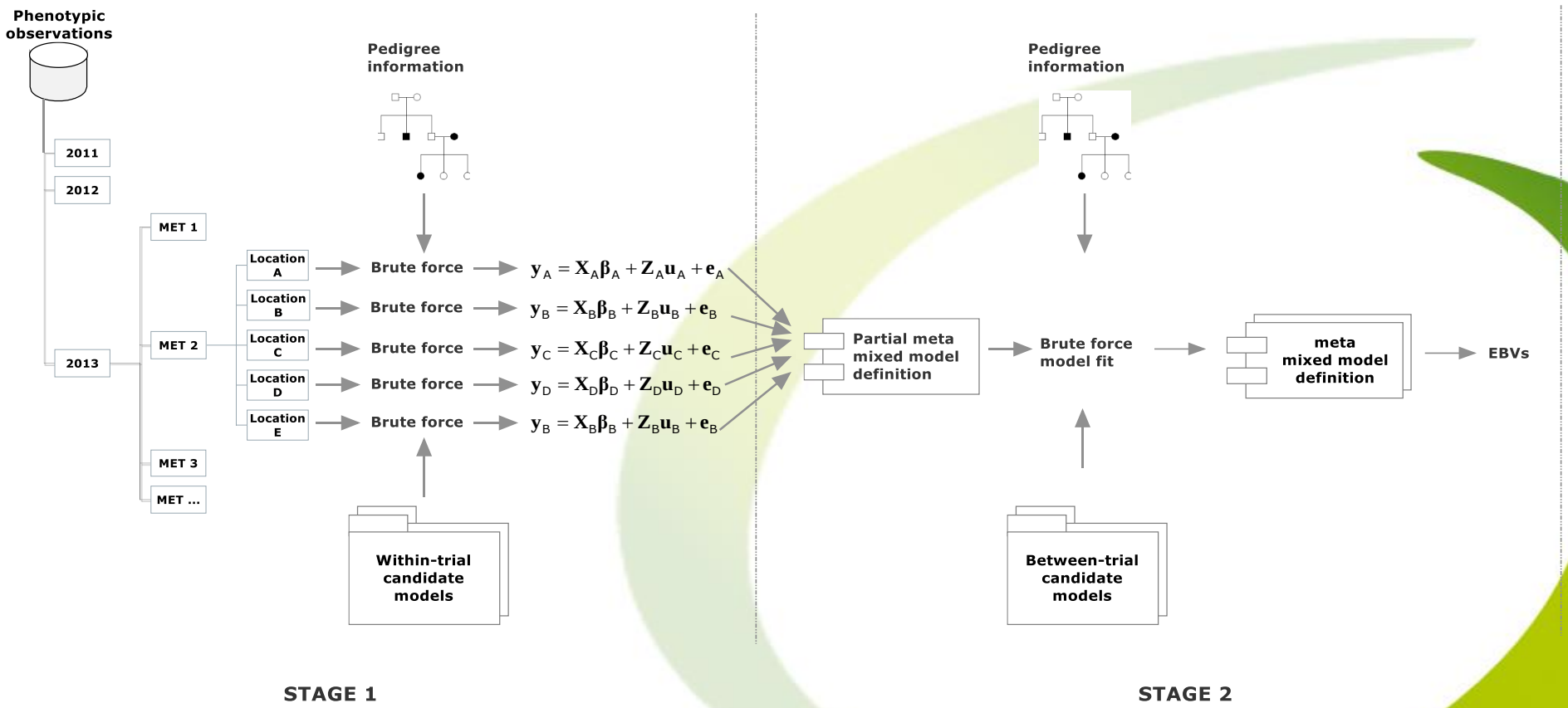


Progeno
identifies genetic potential

Two-stage MET analysis

- genotypes should be fitted as fixed effects in stage 1 to prevent “double shrinkage”
- various weighing mechanisms are used in stage 2 to correct for the difference in variance (and covariance) between adjusted means from stage 1
- simplified weighing models => adverse affects on estimates
- realistic weighing => prohibitive computational burden
- single-stage mixed model analysis is generally considered to be the **golden standard**

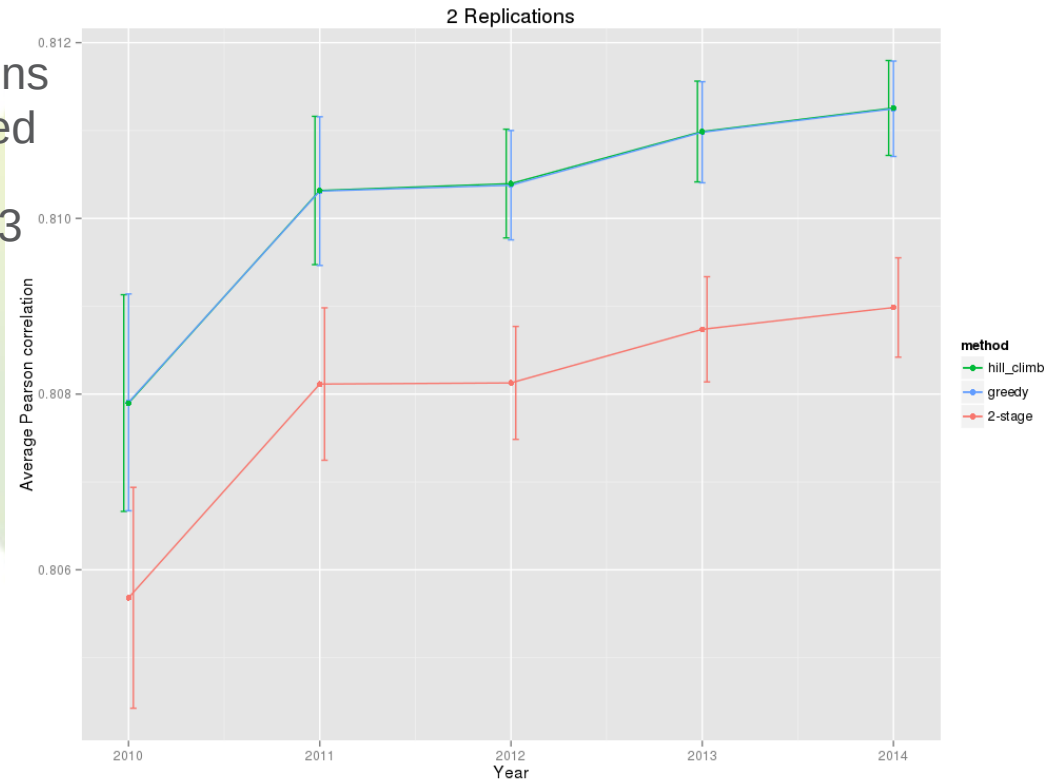
Multi-trial, single-stage model search



Efficiency gain

- Simulated MET phenotypes:
  - 5 trial years
  - 20 METs / year
  - 5 trials / MET
  - 50 parcels / replication / trial
  - random, heterogeneous spatial model for residuals
  - randomized complete block designs
  - connectivity by check and repeated varieties
  - entry mean heritability: 0.67 – 0.73
  - 100 iterations

- 3 model fit strategies:
  1. unweighted two-stage
  2. greedy search
  3. hill climbing search



Trial connectivity

- two trials are directly connected if they have at least one common accession
- two trials can be indirectly connected by having common accessions with a third trial
- fitting a linear mixed model to disconnected data is perfectly possible but the results are bogus
- Progeno Multi trial model search verifies trial connectivity before analysis

	h_1	h_2	h_3	h_4
e_1	3	6	0	0
e_2	3	4	0	0
e_3	0	0	7	5

Progeno
identifies genetic potential

Exercise multi-trial

- Query the database for all observations of the trait “Thousandseedweight”. Export a dataset with the columns EuclegID, Name, Row, Column and Thousandseedweight.
- Make a boxplot per trial
- Find the best fitting multi-trial linear mixed model formulation, export the BLUPs
- Calculate the average Thousandseedweight (first by trial and then by accession) and compare these averages with the BLUPs in a scatter plot. What is the correlation between BLUPs and averages?

Exercise 2 multi-trial

- Query the database for all observations of the trait “Seedyield”. Export a dataset with the columns EuclegID, Name, Block1, Row, Column and Seedyield.
- Try to find the best-fitting multi-trial linear mixed model
- Scale the SeedYield variable and retry
- Save the resulting BLUPs in your Personal storage

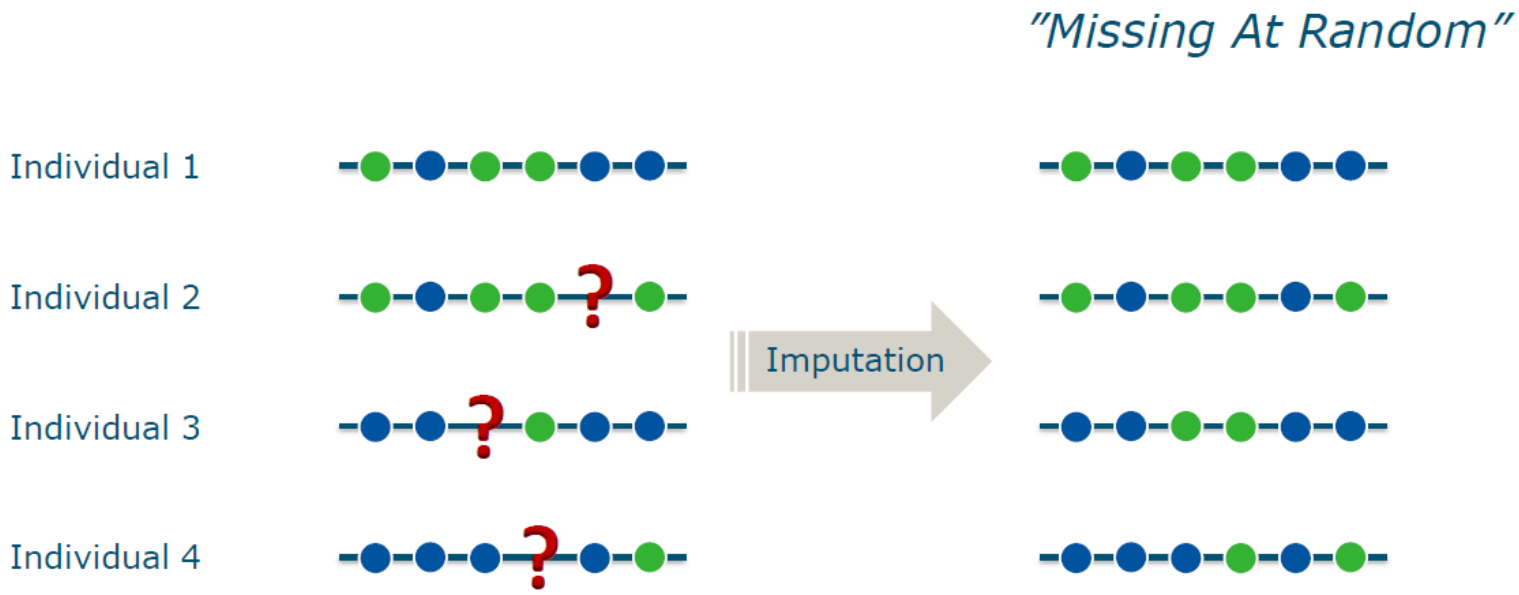
Progeno
identifies genetic potential

Genotype imputation

- the statistical inference of unobserved genotypes:
  - impute randomly missing genotypic data
  - impute genotypic data for alignment of different SNP arrays
  - impute genotypic data from low-density SNP array to high-density SNP array
  - impute genotypic data from low coverage sequencing data

Randomly missing genotypic data

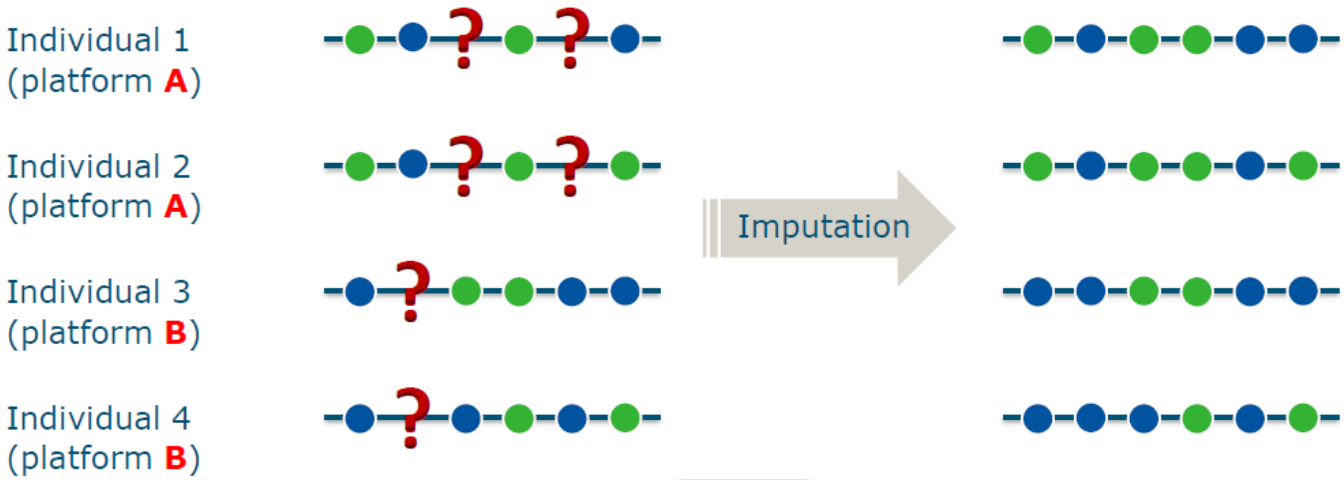
- low percentage of genotypes have not been called, e.g., <5%



Alignment of different SNP arrays

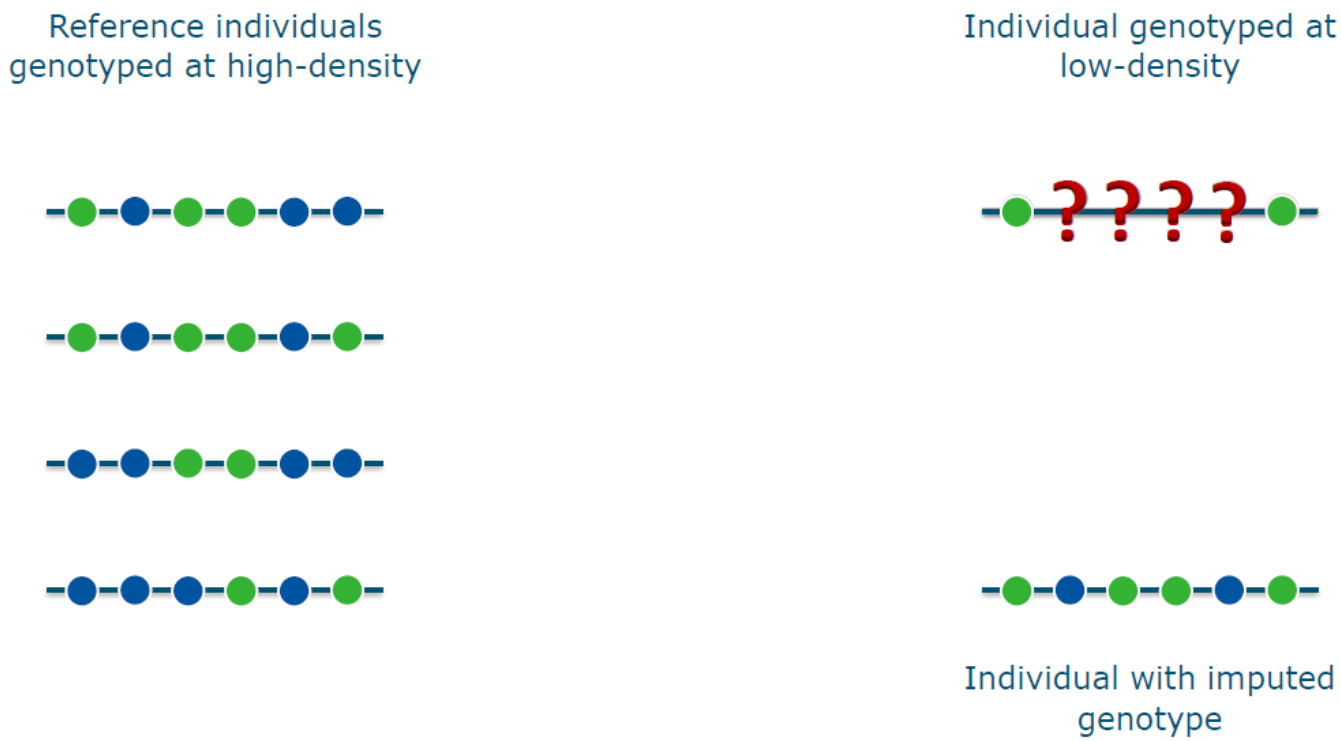
- different genotyping platforms are used with only a certain percentage of SNP common across platforms

"Missing Not At Random"



Impute from low-density to high-density

- makes genomic prediction economically feasible:
  - elite material: genotype with high-density array
  - test / new material: genotype with low-density and impute missing SNP scores



Imputation methods

- naïve approaches
- basic statistical approaches
- population-based approaches
- family-based approaches

Naïve imputation

- often used when GS software does not handle missing data:
  - marker mean value (i.e. related to allele frequency)
  - heterozygous value (inbred lines)

Progeno
identifies genetic potential

Population-based approaches

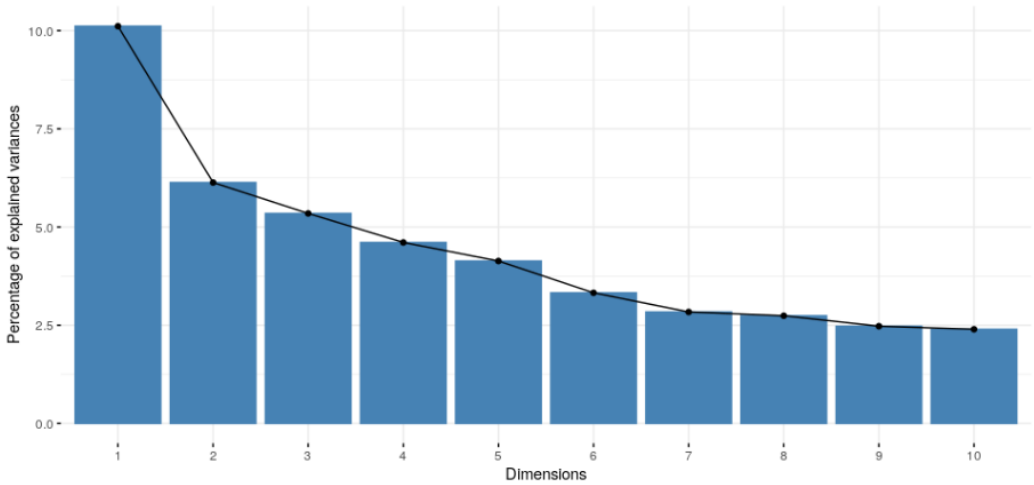
- commonly used in human studies
- based on short-range LD information in a population
- often based on Hidden Markov models (HMM)
- several methods / implementations available that differ in accuracy, computational requirements, speed, ….
  - Beagle (Browning and Browning. 2009. AJHG 84: 210-223)
  - Impute v2 (Howie, et al. 2009. PLoS Genet 5: e1000529)
  - fastPHASE (Scheet and Stephens. 2006. AJHG 78: 629-644)

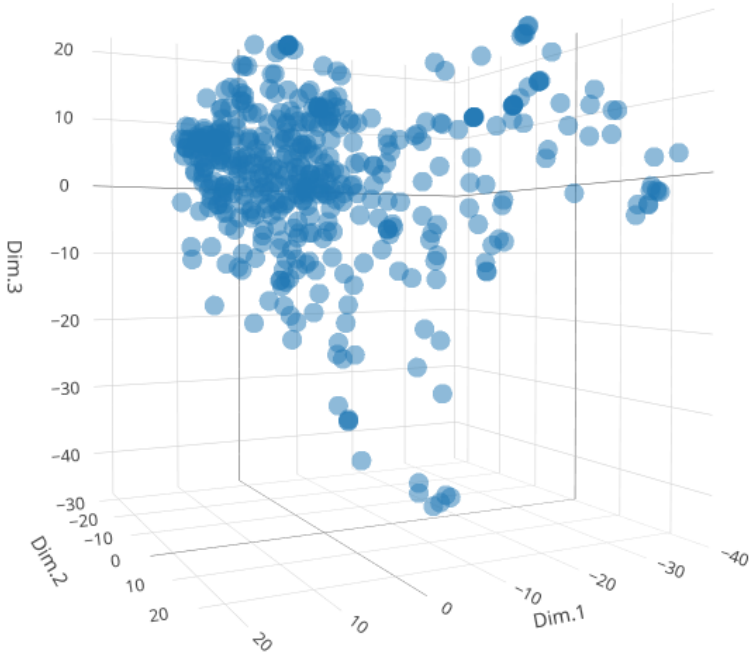
Marker redundancy

- remove markers that are either
  - monomorphic (var(RAF) < 0.01)
  - strongly correlated (R >= 0.99) with other marker(s)
  - parallel computation => might produce different outcome on consecutive runs !

Population analysis

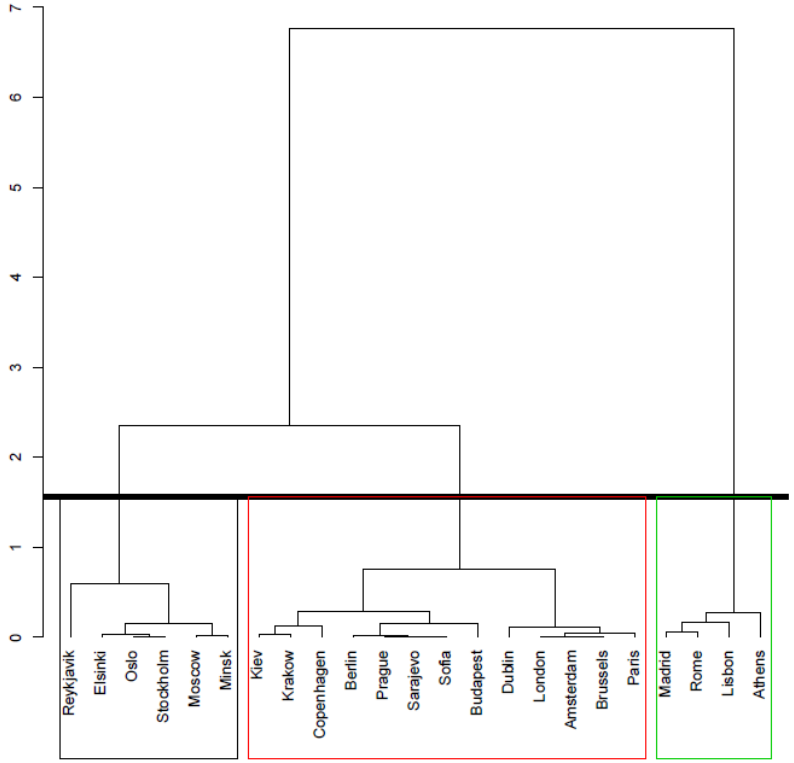
- Dimensionality of the marker matrix is reduced by means of principal component analysis (PCA):
- finds linear combinations of markers that explain the most variance
- Scree plot helps to identify the number of loadings to maintain

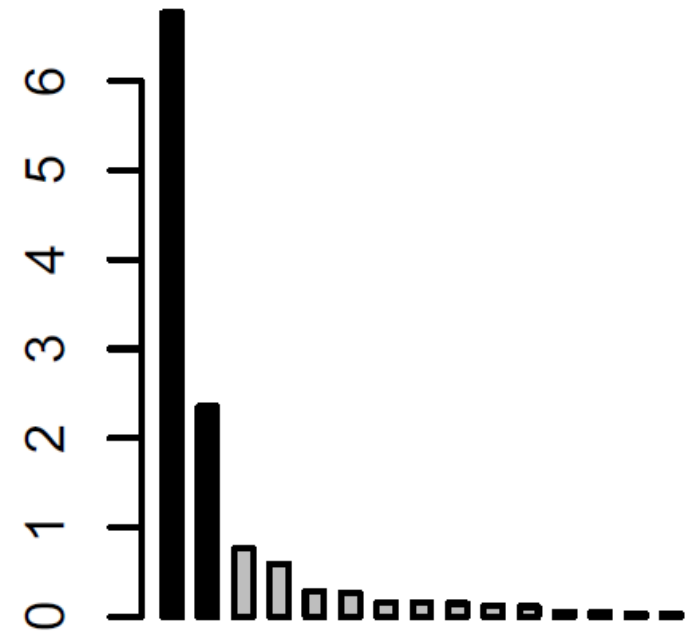




HCPC: Hierarchical Clustering on Principal Components

- clustering method that builds a hierarchical tree (i.e. dendrogram) from accession PCA scores
- Agglomerative clustering: consecutively merging (groups of) accessions until all belong to the same cluster (i.e. the root of the tree)
- Identify the position in the three where the “between inertia” ratio is minimized





Progeno
identifies genetic potential

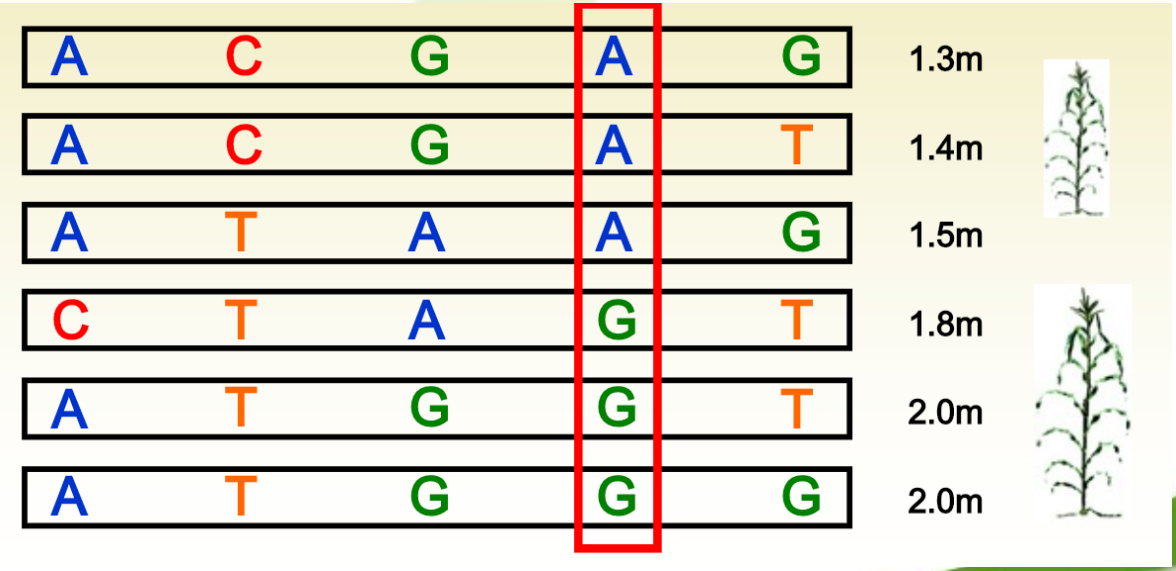
Exercise Population analysis

- Perform a PCA and HCPC analysis on the RAFs of markers located on chromosome 4
- Retrieve the supplier information of the accessions and merge it with the HCPC cluster results. Can you link suppliers to HCPC clusters?

Progeno
identifies genetic potential

Genome-wide Association Study

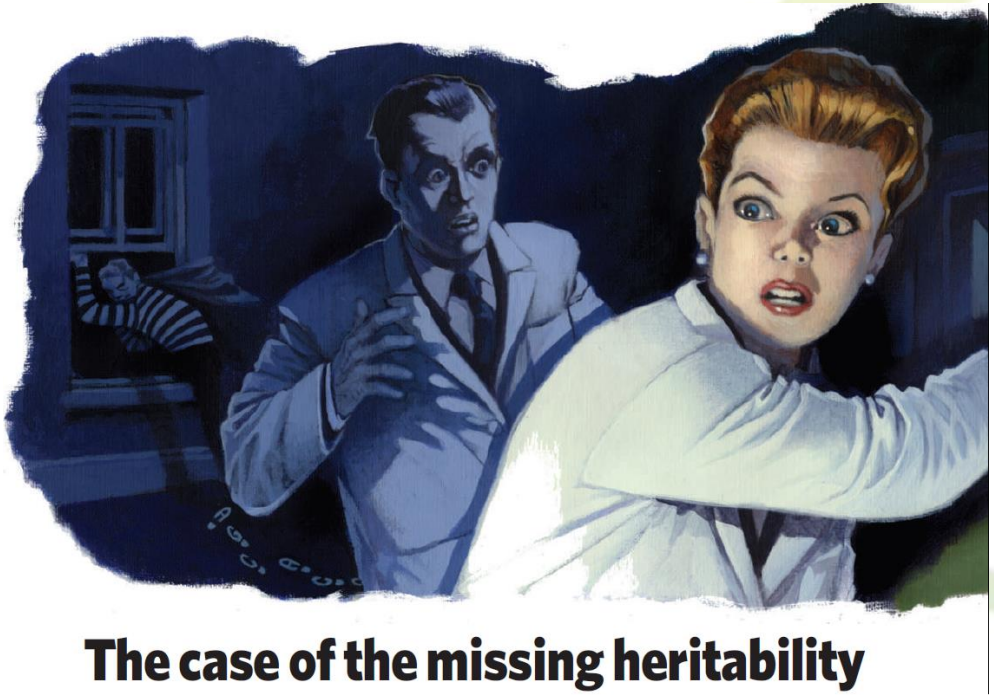
- marker alleles are correlated with a trait on a population level
- can detect association by looking at unrelated individuals from a population
- does not necessarily imply that markers are linked to (are close to) genes influencing the trait.
- requires high-density molecular marker panel

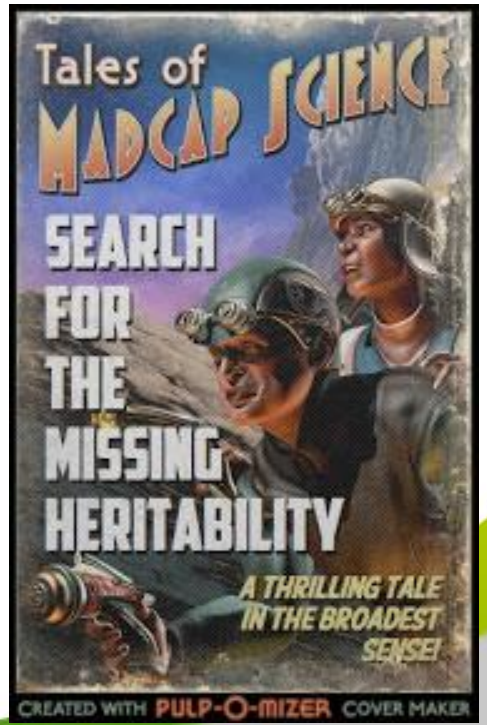


Progeno
identifies genetic potential

Limitations of GWAS

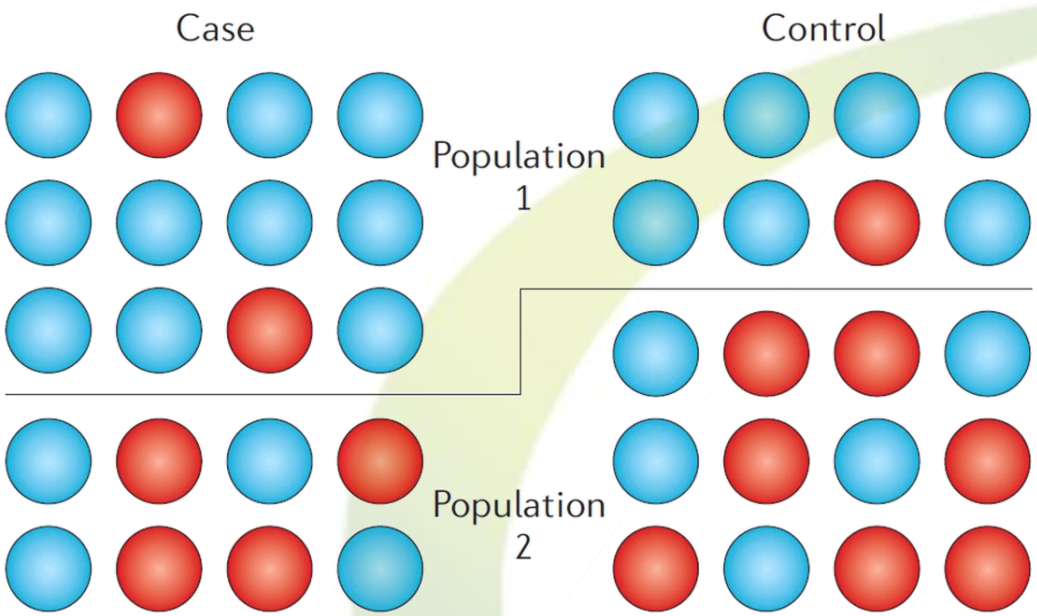
- not very predictive for complex traits
- generally explains little heritability
- focus on common variation
- many associated variants are not causal





QK model

- Both kinship and population stratification can bias GWA studies



- yij = mean + m(i) + population(i) + Individual(i) + residual
- Var(Individual) = scaled Kinship matrix

Progeno
identifies genetic potential

Correcting for multiple testing

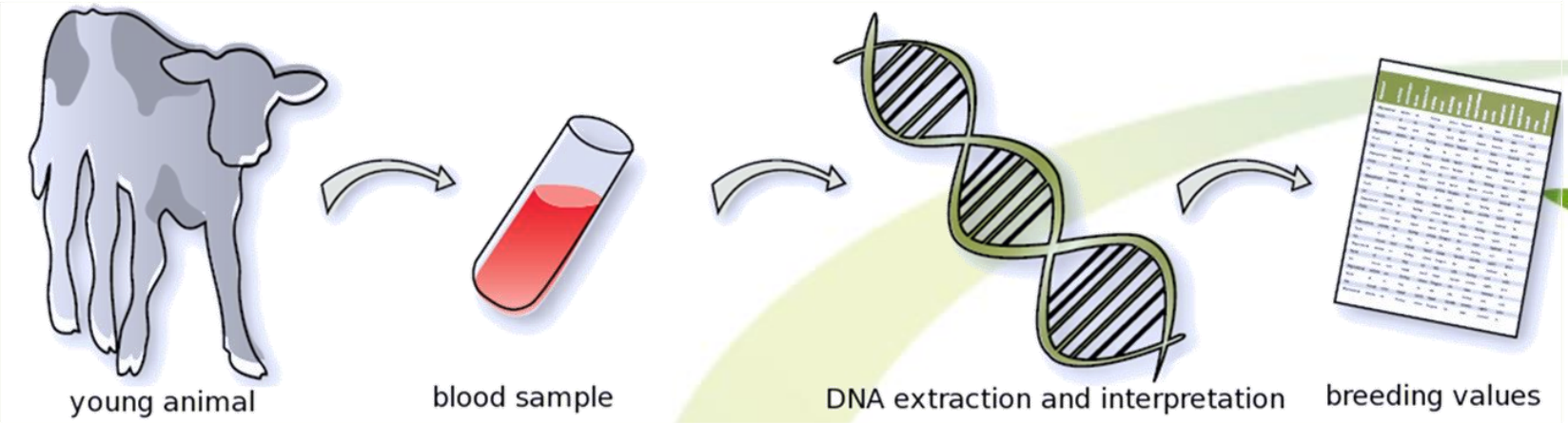
- Performing more statistical test will always yield more hits
- **Bonferroni**: Threshold = alpha / NumTrials
- **False Discovery Rate** (Benjamini-Hochberg):
  - Threshold1 = (1 * alpha) / NumTrials
  - Threshold2 = (2 * alpha) / NumTrials
  - Threshold3 = (3 * alpha) / NumTrials
  - …
- If you are willing to accept a “fraction of false discoveries”, FDR correction is acceptable

Progeno
identifies genetic potential

Exercise GWAS

- Extract the marker RAFs for SNPs located on chromosome 13 at a physical position greater than 35.000.000
- Impute missing values and remove redundant columns
- Perform population analysis and export cluster information
- Perform GWAS using the SeedyieldBLUP dataset from the shared storage space.
- Check “Reuse variance estimates” to speed up computations
- Report significant SNPs after FDR correction

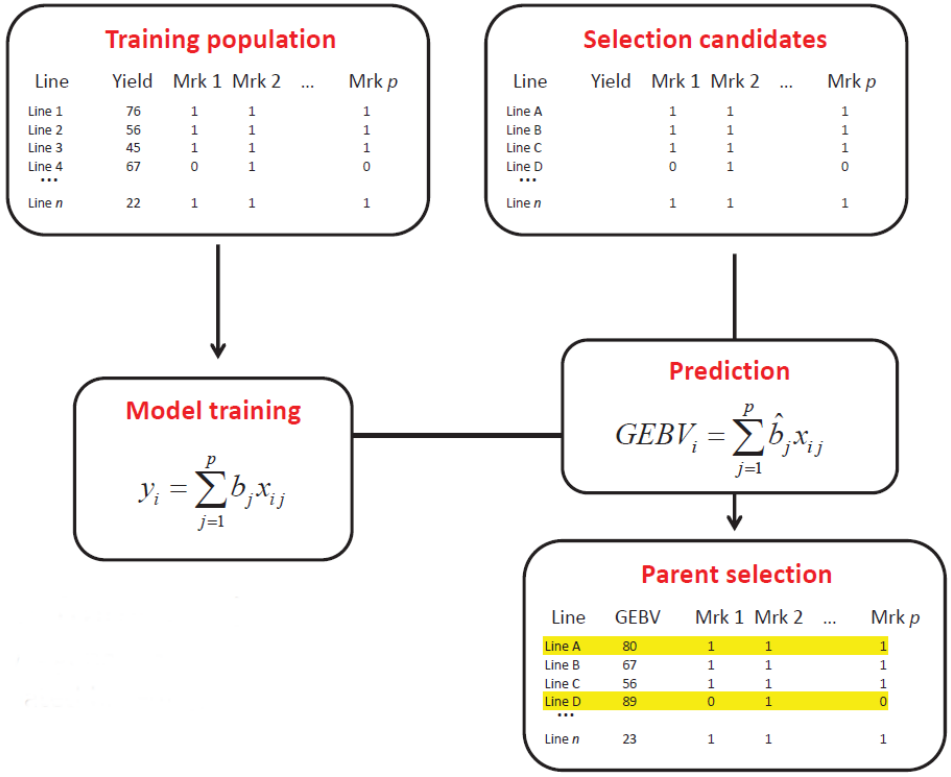
Genomic selection



- GS aims to trace all the QTLs in the genome using large number of markers. - all markers can have some effect
- First introduced by Meuwissen et al. in 2001
- Very successful! It is routinely applied in breeding of many plant and animal species

Progeno
identifies genetic potential

GS Concept



1. the effects of chromosome segments (markers, haplotypes, ..) are estimated in a training population

2. breeding values of individuals without phenotypes are predicted

Progeno
identifies genetic potential

Genomic Prediction methods

- RRBLUP or Ridge Regression BLUP: linear mixed model approach fitting a random effect for each marker

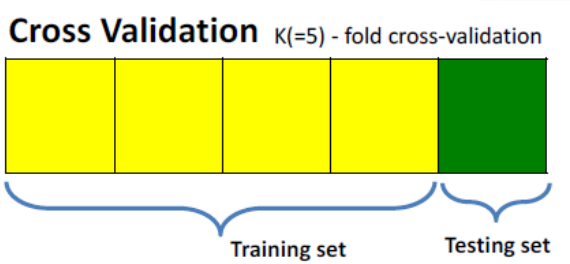
  => fast computations
- GBLUP or Genomic BLUP: linear mixed model approach fitting a random effect for each accession => slower than RRBLUP

  => provides reliabilities
- Random Forests: machine learning technique using resampled decision trees

  => requires parameter optimization

Genomic prediction accuracy

- **predictive ability**: the expected Pearson correlation between genomic prediction and the actual phenotype
- **prediction accuracy**: the expected Pearson correlation between genomic prediction and the breeding value
- estimated through k-fold cross-validation:
  - randomly split the training data in k subsets
  - for each subset i:
    - *remove the phenotypic data for all the members of subset I*
    - *construct a genomic prediction model from the remaining data*
    - *predict the phenotypes of the members of subset I from their genotypic data*
  - correlate the predictions with the actual phenotypic values



Progeno
identifies genetic potential

Genomic Prediction exercise

- Analyze the available multi-trial data for the trait “ThousandSeedWeight”. Export the BLUPs to a separate dataset.
- Train a genomic prediction model using the BLUPs as phenotypes and the “allMarkers.imputed.informative” matrix. What is the average cross-validation accuracy.
- Use the trained model to make genomic predictions for the matrix of selection candidates named “selectionMatrix”
- Which selection candidate has the highest predicted ThousandSeedWeight?

Progeno
identifies genetic potential

PDP = Processed Data Package