Simulation of breeding programs with the Modular Breeding Program Simulator (MoBPS)

Torsten Pook

28/04/2025

WAGENINGEN
UNIVERSITY & RESEARCH

@tpook92

Torsten.pook@wur.nl

- You do not learn how to use a software by listening
  - Most of this course will be practical sessions
- Form small groups (~1 – 3)
  - In particular when you have limited programming background
- Joint discussion & sample solution
- All sample solutions and slides will be shared

Agenda

- Monday & Tuesday
  - Program: 9:00 – 17:00
  - Lunch break: 12:30 – 13.30
  - Coffee breaks at ~10.30, 15.00

Agenda

- General introduction of the MoBPS framework
- Basic functionality of the web-interface
- More advanced features of the web-interface
- Scenario comparison in the web-interface
- Basic functionality of the R-package
- Trait generation in the R-package
- Import of real genotype data (vcf) in the R-package

Agenda

- Breeding value estimation in the R-package
- Breeding value estimation with commercial software (MiXBLUP, blupf90)
- Implementing own methodology within R
- Computational efficiency within R
- Inbreeding management in the R-package

Agenda

- Addressing specific individuals in the R-package
- Discrete phenotypes & litter sizes
- Simulating multiple scenarios
- Evaluating different scenarios

Optional! When we need more time for the other Tasks that has priority!

- Topics of choice / Open questions

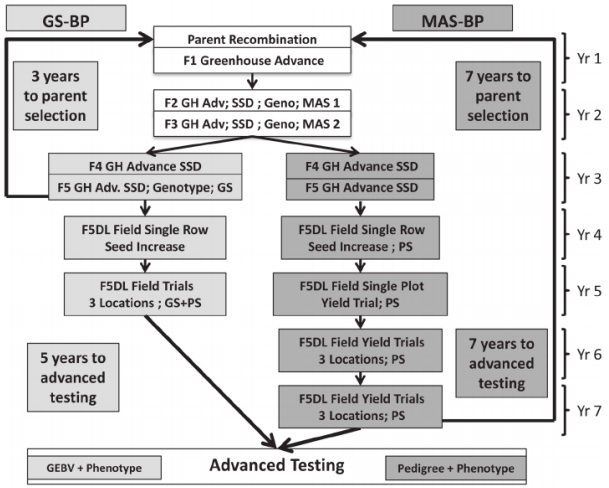
- If questions arise long-term: Torsten.pook@wur.nl

General introduction

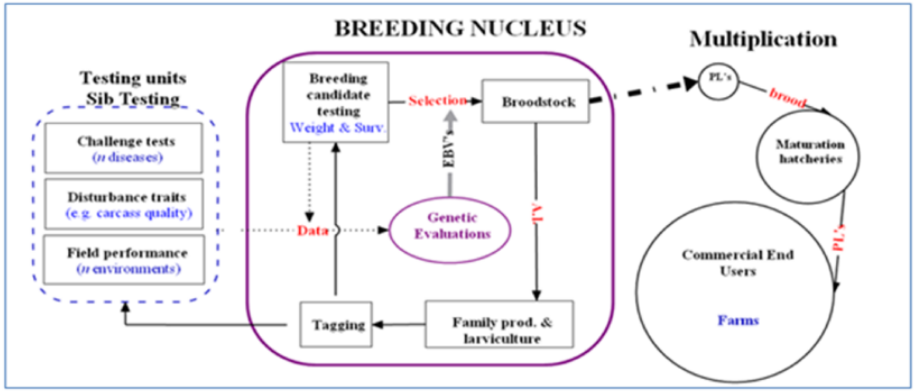
Interrupt me whenever you want!

What is a breeding program?

- Planned breeding of a group of animals or plants over several generations
- Goal: Change characteristics of animals/plant through careful selection of breeding partners
- General language to describe breeding programs (Simianer et al. 2021)



(Wheat breeding: Heffner et al., 2010)



(Fish breeding: Rye 2012)

What are we interested in?

- What is our breeding objective?
  - Genetic progress
  - Maintenance of genetic diversity
  - Risk (variability of the outcome)
  - Economic efficiency

How to control it?

- How many animals/plants to use
- Generate genotype / phenotype data
  - How many? Which?
- Mating scheme
- Selection technique
- Use of biotechnology
- And much more...

→ Complex optimization problem!

Possible ways to answer this?

- Experience of the breeder
- Simulation study
- Cohort-based deterministic (ZPLAN+, Täubert et al. 2010)

$$R = i \cdot h \cdot \sigma_a$$

- Good approximation but formulas are limited to specific applications and are constructed to handle „easy“ scenarios

→ Stochastic simulation

What is a stochastic simulation?

- Every breeding action in a breeding scheme is simulated
  - Low number of assumptions
  - Flexibility
  - High level of detail
  - High computational demands & work!
- Results of a simulation will not be an expected gain but the realization of a stochastic process

What is the MoBPS?

- Environment for the simulation
- The software takes care of backend-"stuff" that you have to account for but is not main part of analysis
  - Meiosis
  - Trait simulation
  - Efficient data storage
- Pre-implemented functions common breeding actions
  - Breeding value estimation
  - Phenotyping
  - Selection
- Function to help with down-stream analysis

About the R-package

- Mainly distributed via GitHub
- Design philosophy:
  - Generate a framework that is able to simulate all breeding programs
  - When something is not yet possible and we see a general value in it, we are going to add it

Version 1.11.50 (17.06.24)

Improved speed of meiosis simulation (~40%) + improved parallelization

"GBLUP" is not an accepted input where previously "vanRaden" was expected to perform a genomic BVE ("vanRaden" still works as well)

Added inbreeding control features (max.selection.fullsib/halfsib, selection.index.kindship, avoid.mating)

Added option to phenotype/genotype selected individuals (genotyped.selected, phenotype.selected)

Added link to blupf90/renumf90 (bve.blupf90)

Added advanced mixblup features (restart/nopeek/calcinbr.s)

Number of recombination points can be linked to a heritable trait (recombination.rate.trait)

Fixed bug in OGC implementation for optiSel where contributions where not always correctly assigned to individuals

Version 1.11.03 (29.12.23)

QTL-effects can now be sampled from gaussian and gamma distribution (default: effect.distribution = "gauss")

Added features to trait time.point of phenotyping, culling, genotyping

Extended max.offspring feature to limit each individual to an individual threshold value

Added link to mixblup (bve.mixblup)

Added option to cull all non-selected individuals

Added snapshot function to get an overview of population (get.snapshot / get.snapshot.single)

Version 1.10.48 (29.03.23)

Documentation overhaul (?breeding.diploid, ?creating.diploid + Guidelines)

Added function to add additional genetic diversity to existing population (add.diversity)

- New MoBPS release (v.1.12) end of this week!

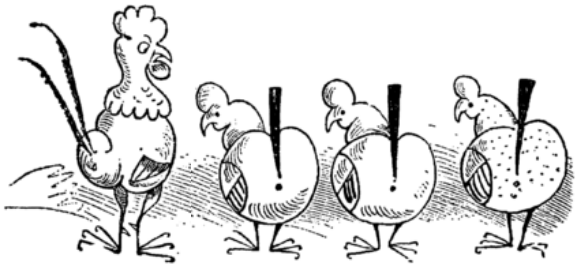
- We will use v1.12 this workshop – so today please not use Github but the Dropbox-version!

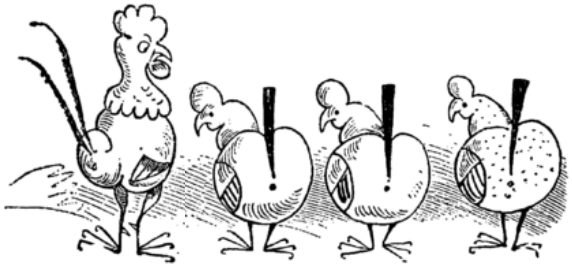
About the R-package

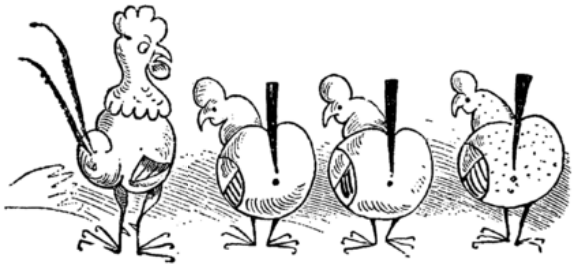
- The parameters in MoBPS are used to perform specific breeding actions
- You can not tell the tool to simulate a population for you with given effective population size / levels of LD / a population specific trait
    - You yourself have to set up a mating scheme to obtain those target values
        - Simulate 100 generations of random mating in a population
        - Analyze the final population
        - If levels of LD are too low
            - simulate more generations / reduce the number of individuals

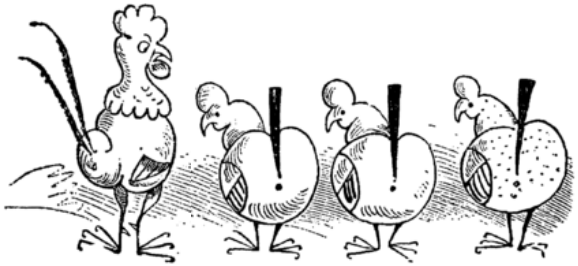
Some simulations to inspire

Cock rotation





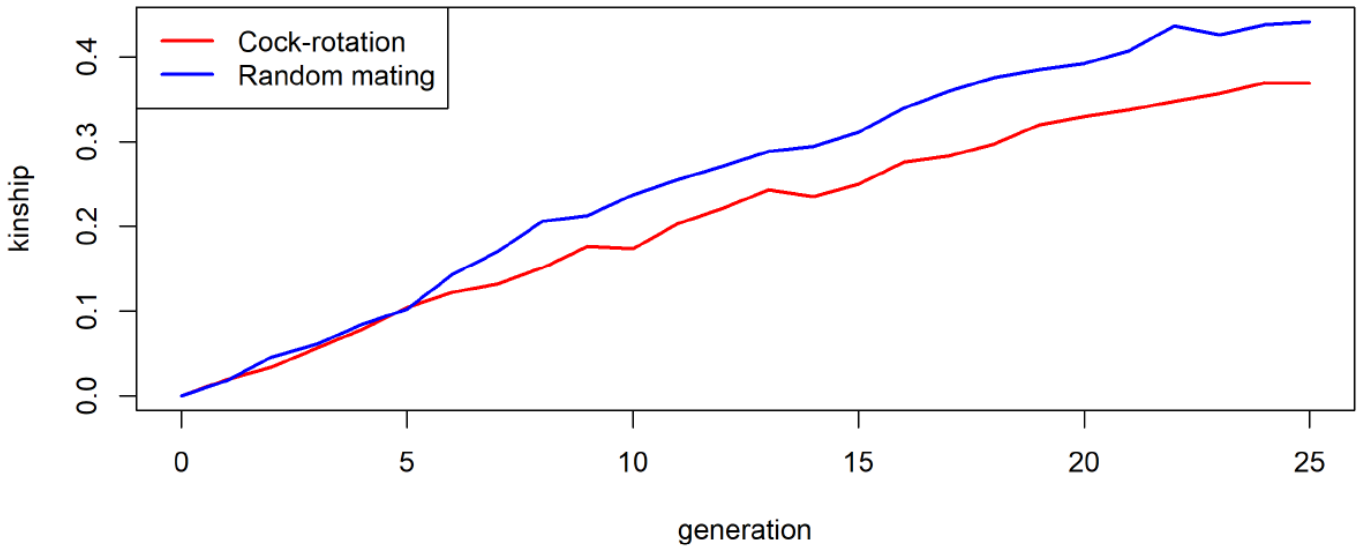




WAGENINGEN
UNIVERSITY & RESEARCH

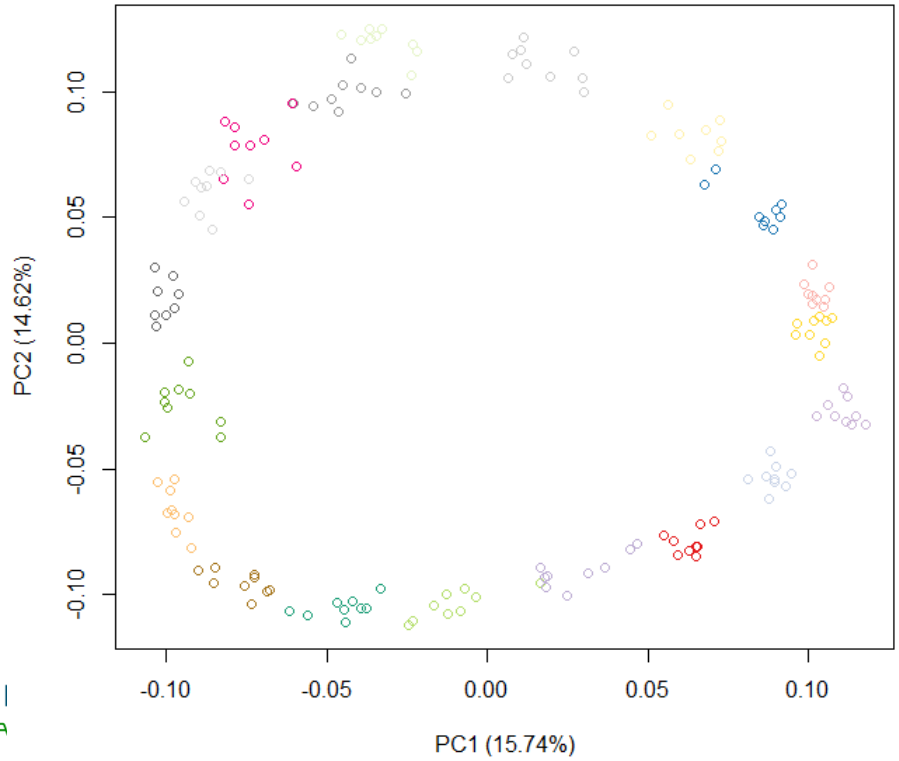
Cock rotation

- Breeding scheme to reduce loss of genetic diversity



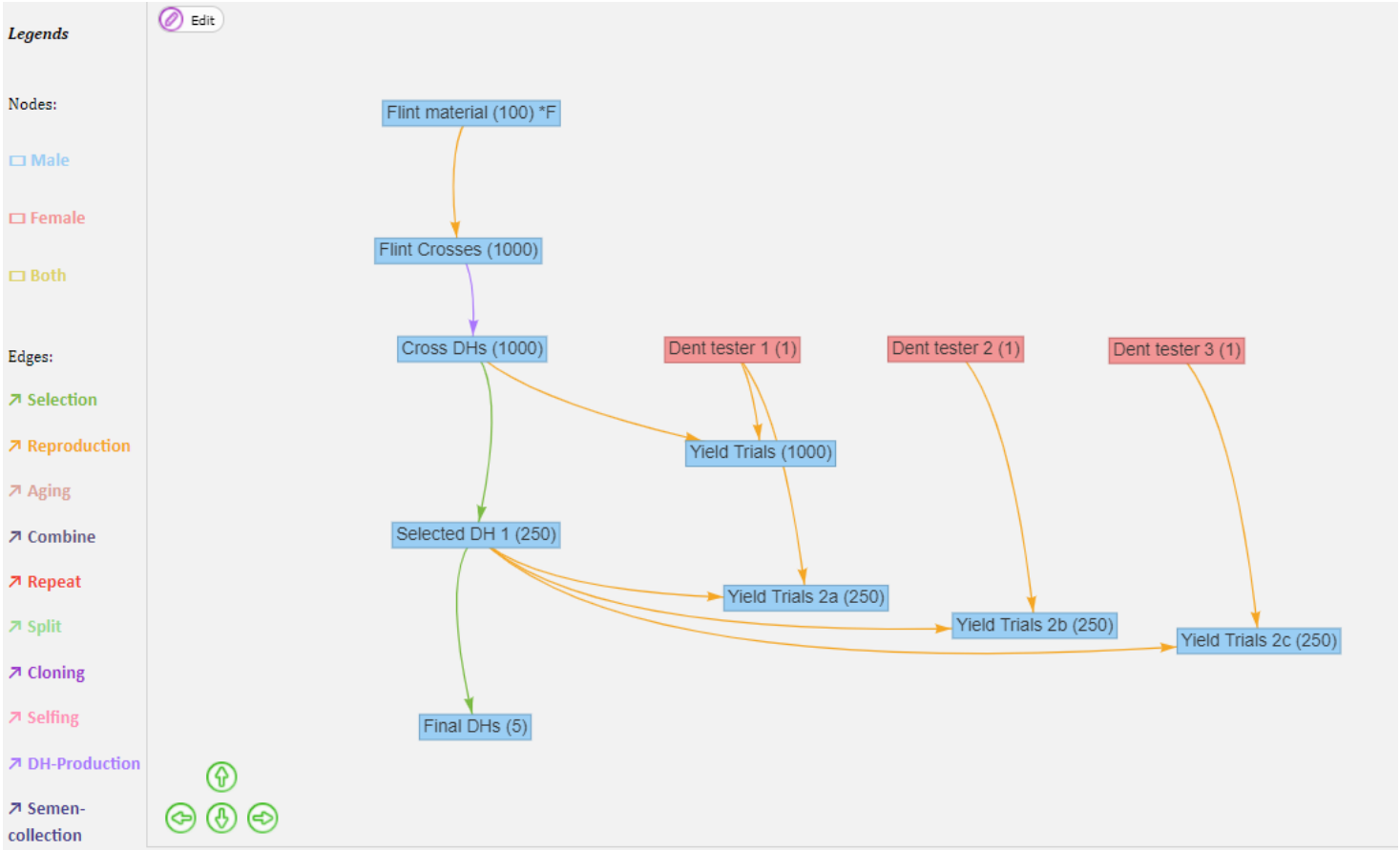
Cock rotation

- Genetic distances between animals according to a principle component analysis



Each color represents one box

Test crossing scheme in maize



Test crossing scheme in maize

- Phenotypes are collected in multiple different environments
- Different number of repetitions / plots

Genetic Correlation

	Grain Yield Loc1	Grain Yield Loc2	Grain Yield Loc3
Grain Yield Loc1	1	0.8	0.8
Grain Yield Loc2	0.8	1	0.8
Grain Yield Loc3	0.8	0.8	1

Phenotype Information

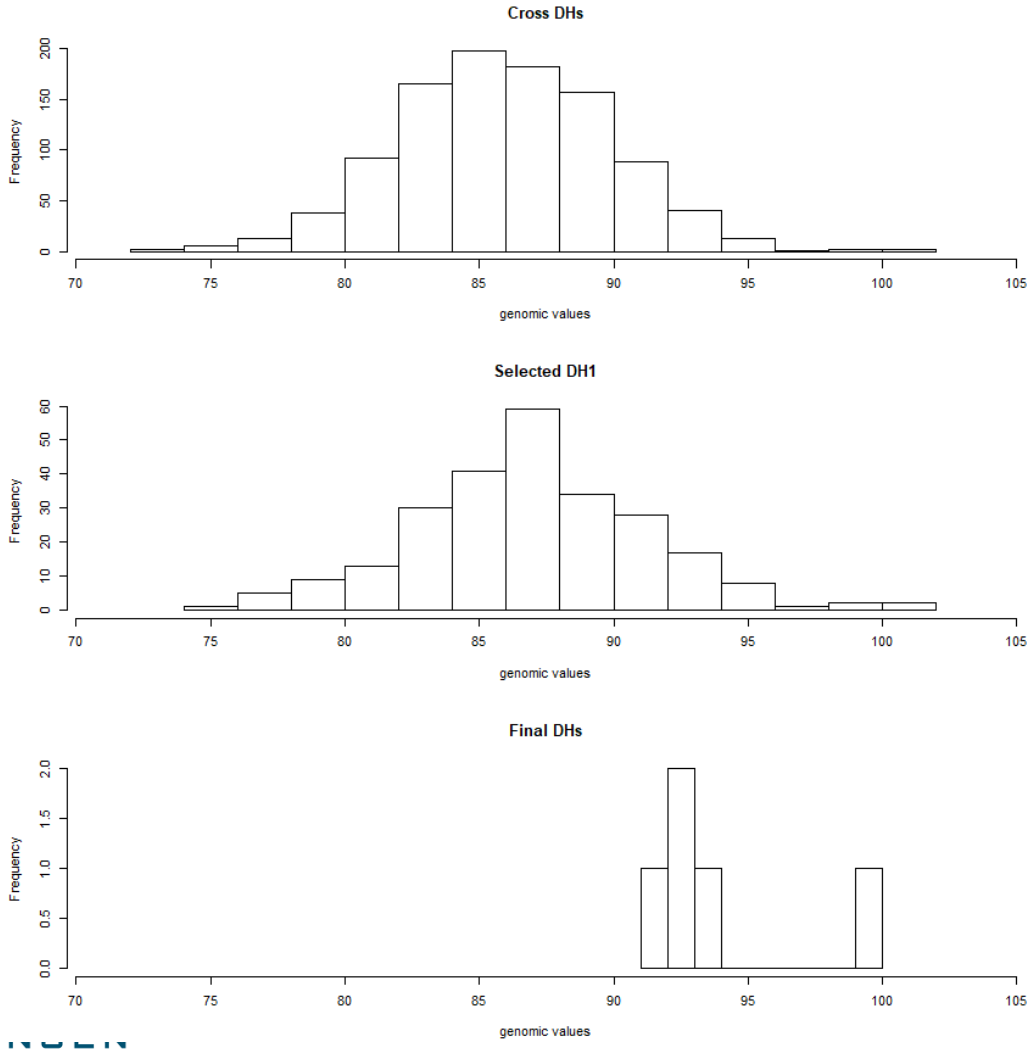
Press here to get info on how this module works!

Add new phenotype | Show/Hide 3 phenotypes | Show/Hide QTLs | Show/Hide residual correlation | Show/Hide genetic correlation

Phenotype	Unit	Pheno. Mean	Pheno. SD	Heritability	Repeatability	# additive QTL	# dominant QTL	# qualitative epistatic QTL	# quantitative epistatic QTL	Major QTL	Value per unit (€)	Show Cor		
Grain Yield Lc		100	10	0.3	0.5	500	500	0	0	0	0	☑	X
Grain Yield Lc		100	10	0.3	0.5	500	500	0	0	0	0	☑	X
Grain Yield Lc		100	10	0.3	0.5	500	500	0	0	0	0	☑	X

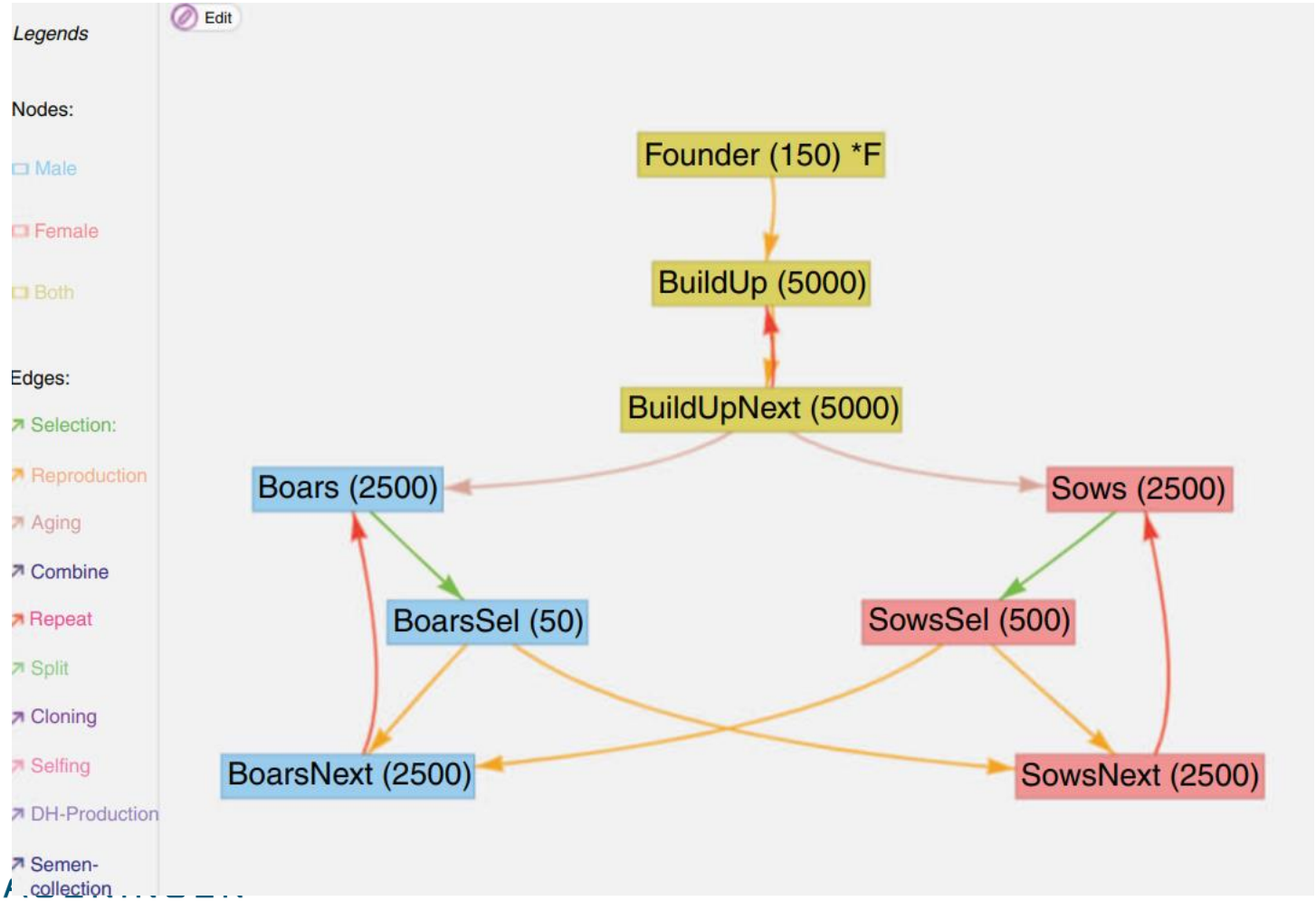
WAGENINGEN
UNIVERSITY & RESEARCH

Test crossing scheme in maize

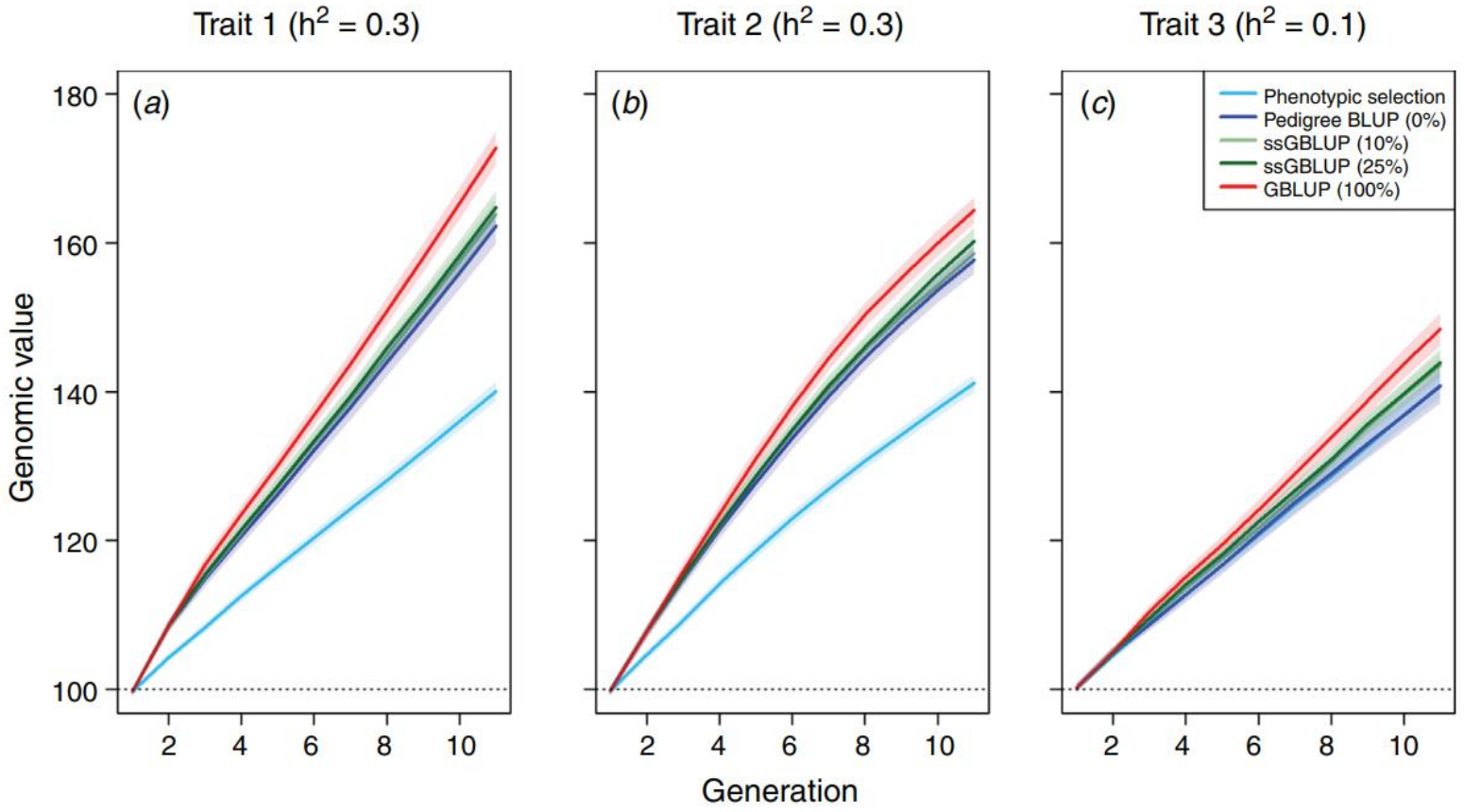


WAGENINGEN
UNIVERSITY & RESEARCH

Pig breeding (Pook et al. 2021)

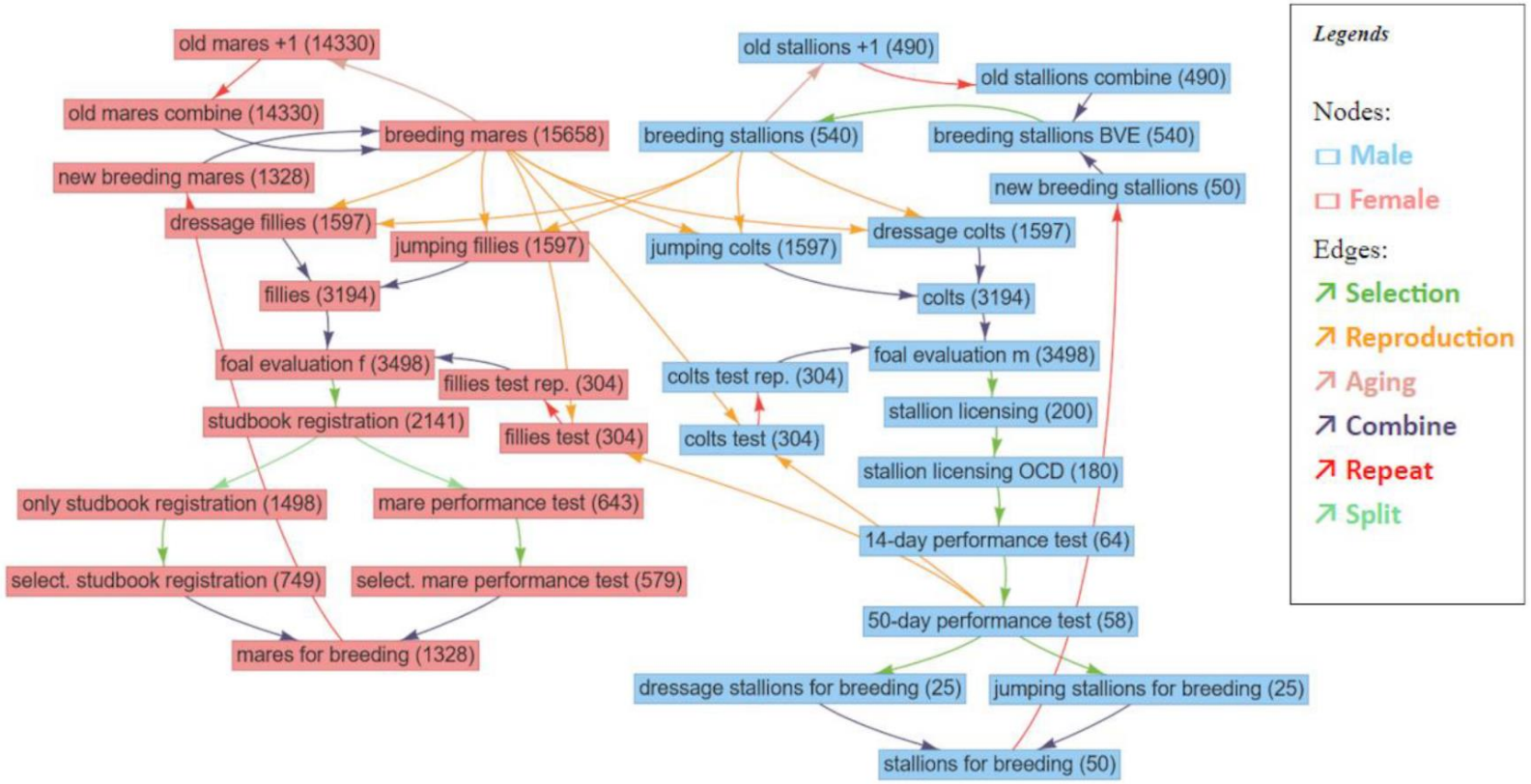


Pig breeding (Pook et al. 2021)



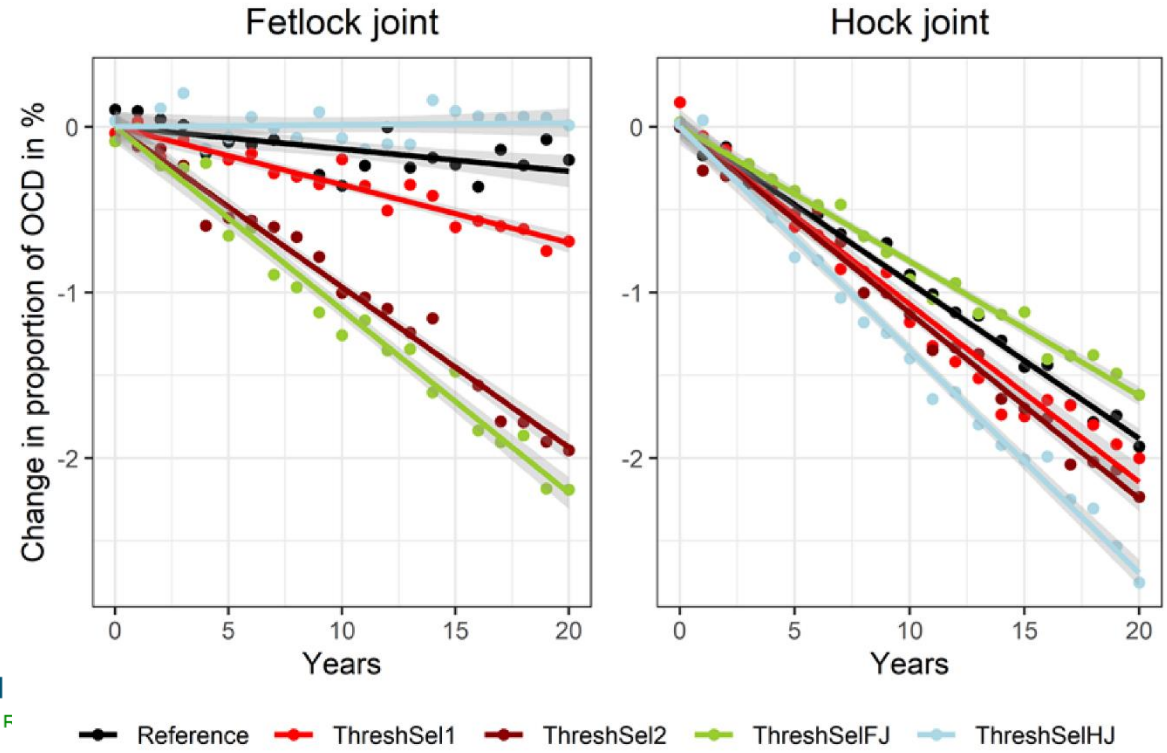
WAGENINGEN
UNIVERSITY & RESEARCH

Horse breeding scheme (Buettgen et al. 2020)

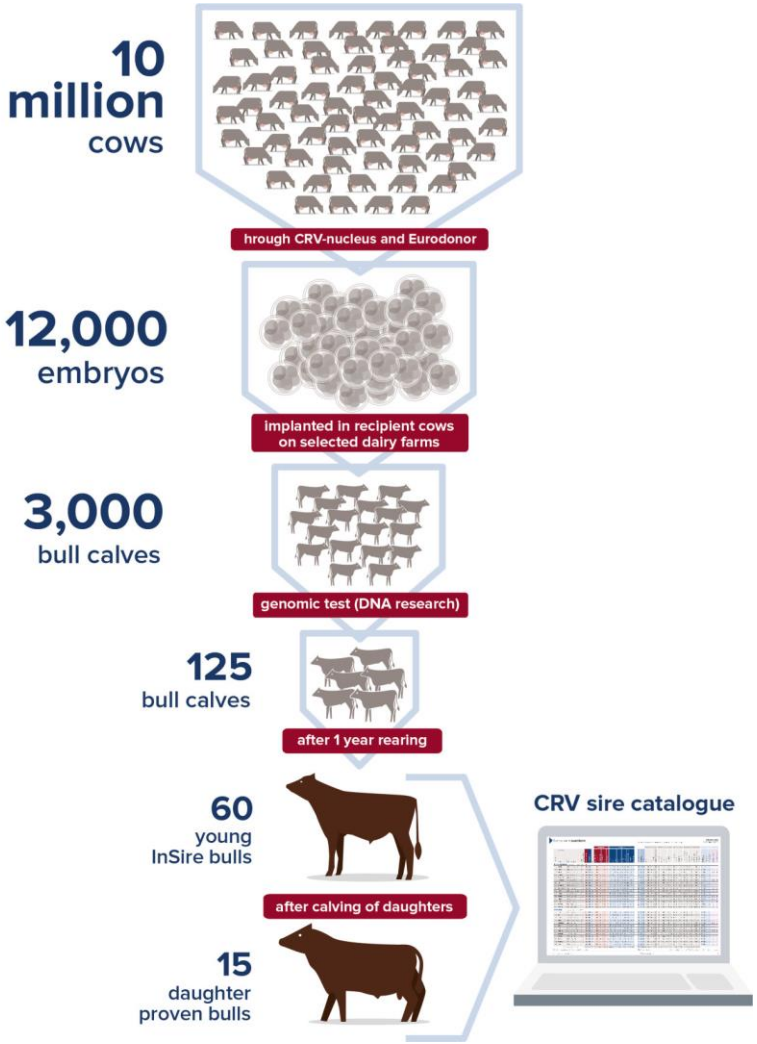


Horse breeding scheme (Buettgen et al. 2020)

- Impact of changes to the breeding program
- Exclude horses with osteochondritis dissecans from selection



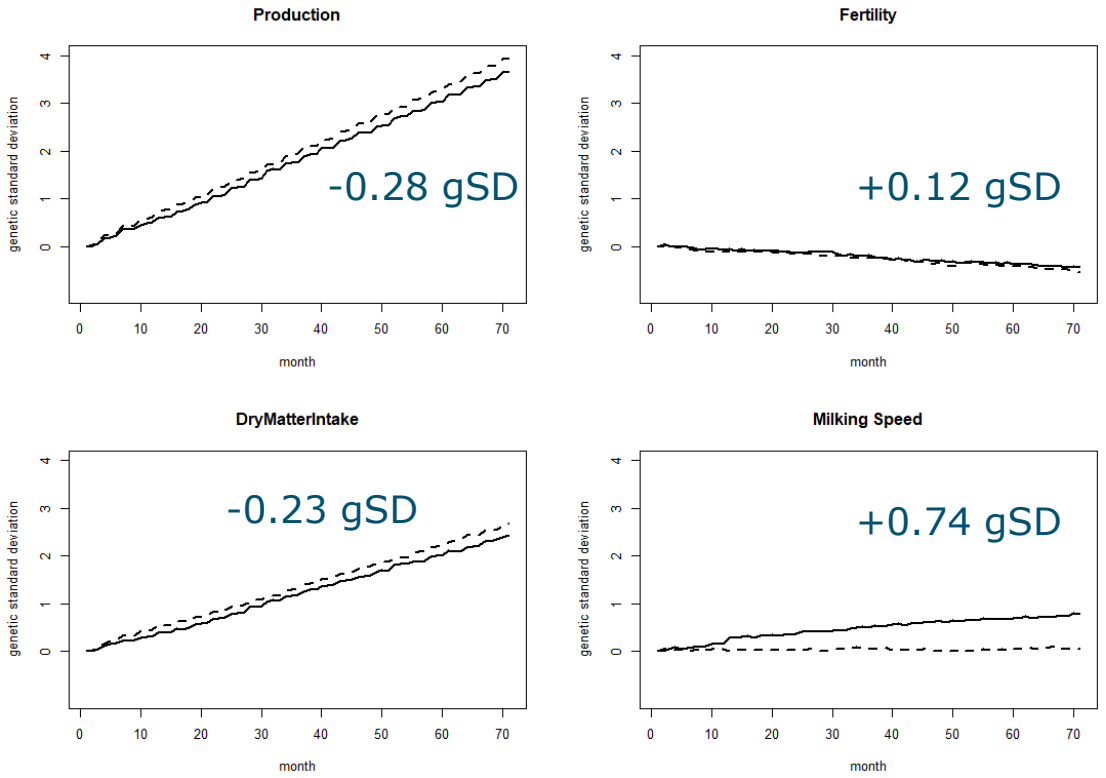
Dairy cattle breeding (CRV)



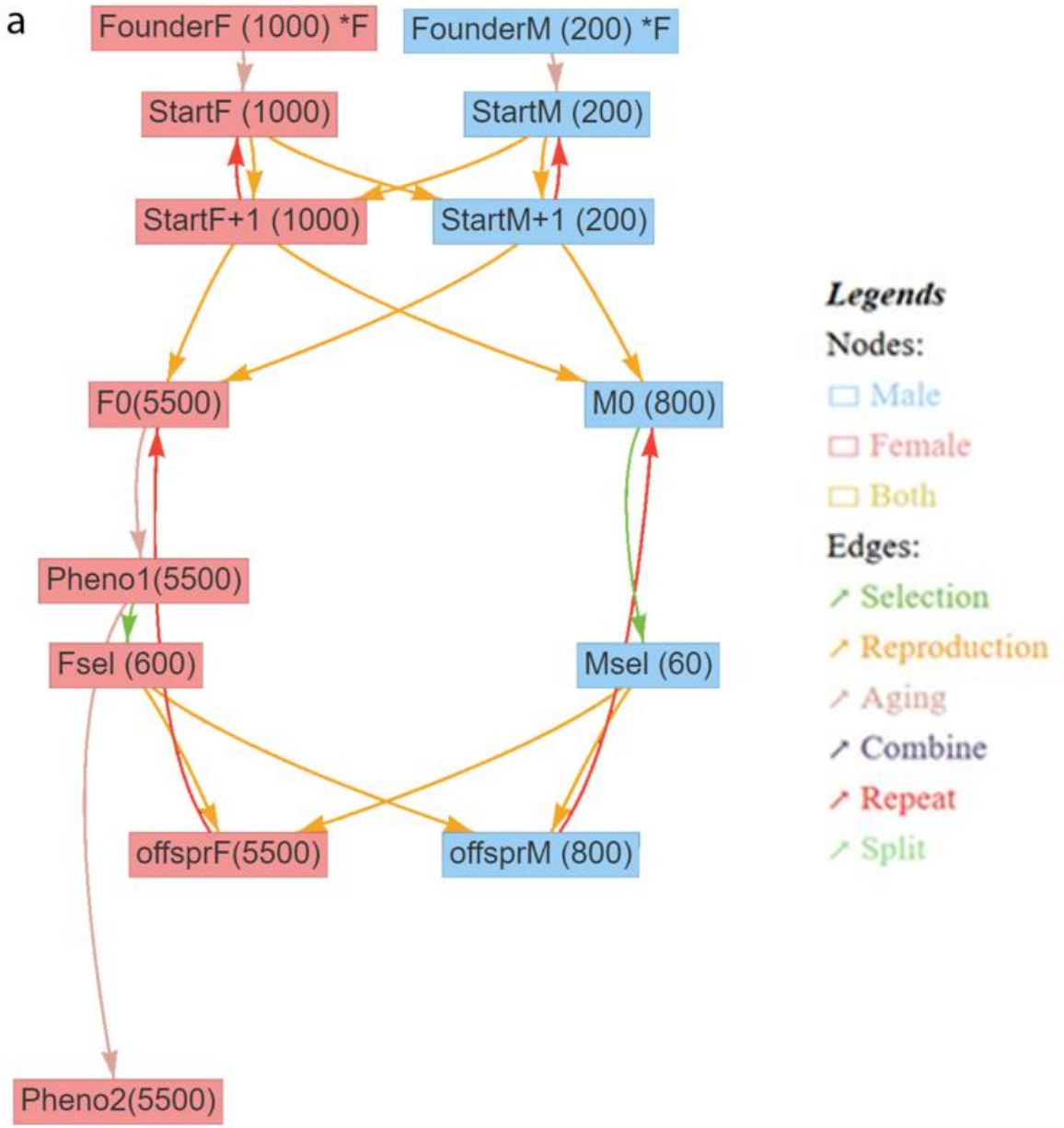
WAGENINGEN
UNIVERSITY & RESEARCH

Dairy cattle breeding (CRV)

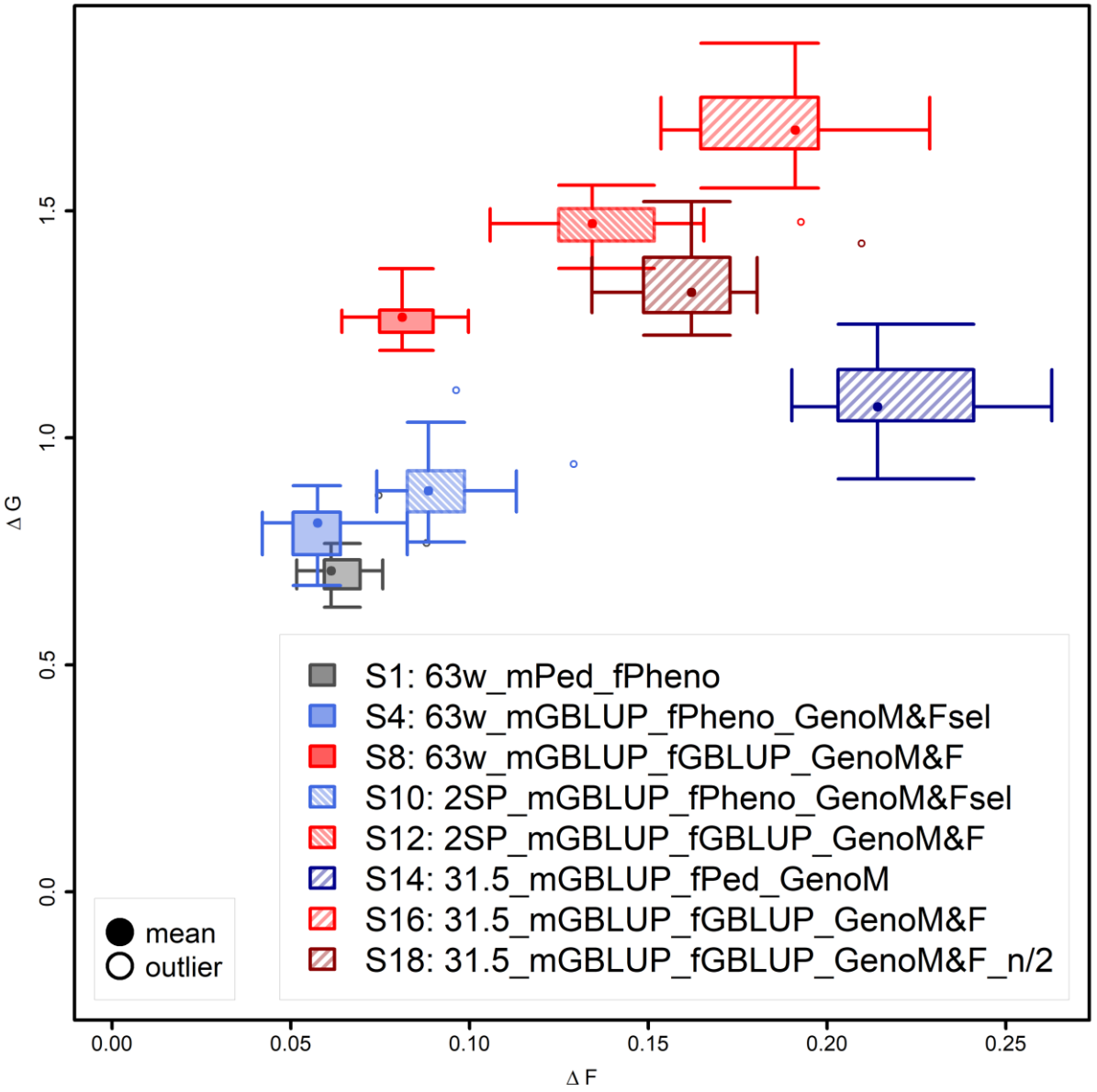
- Threshold selection for fertility & milking speed
- Fertility negatively correlated with production traits



Shortening the generation interval in chicken (Büttgen et al., 2025)

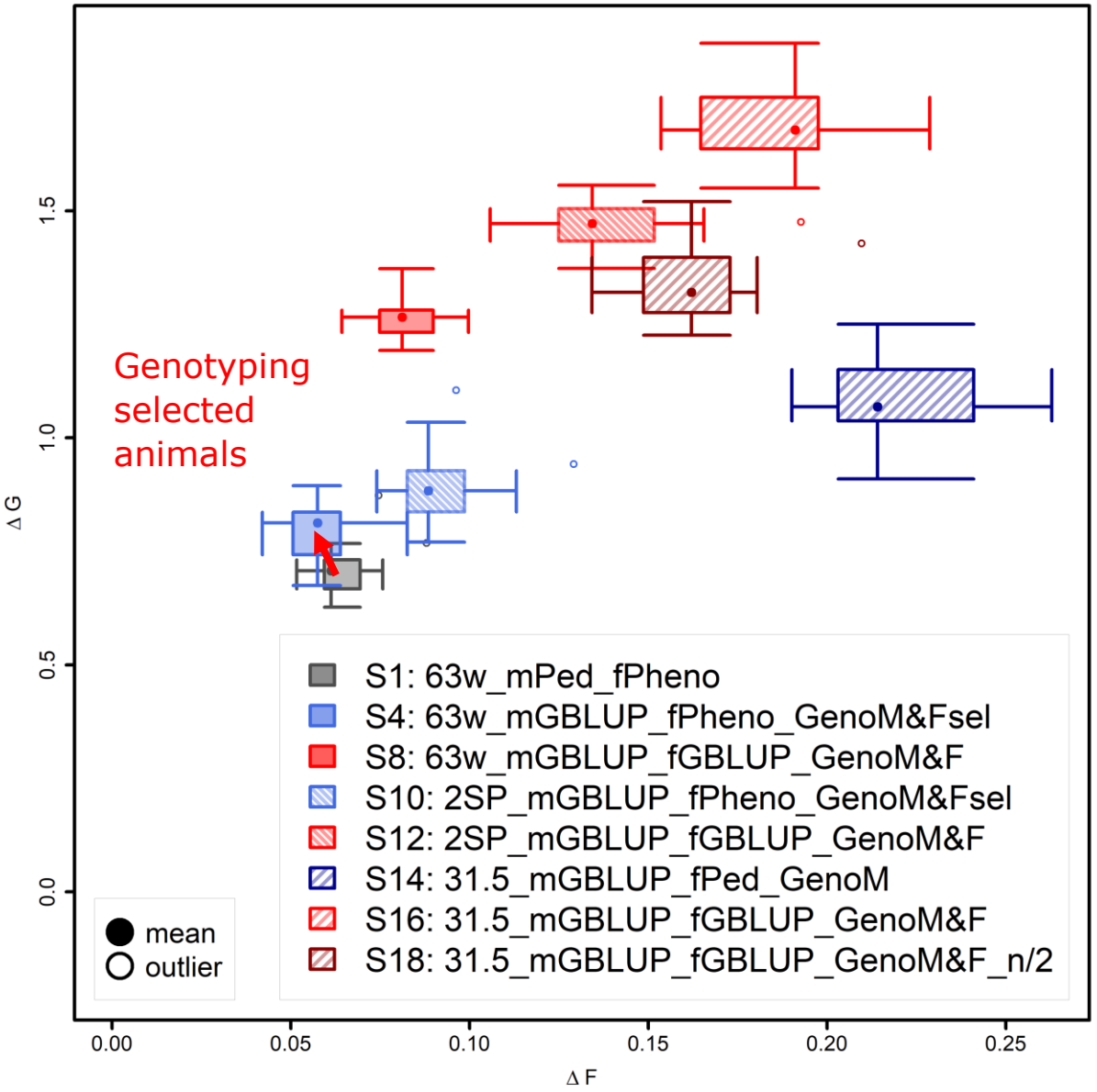


Genetic gain



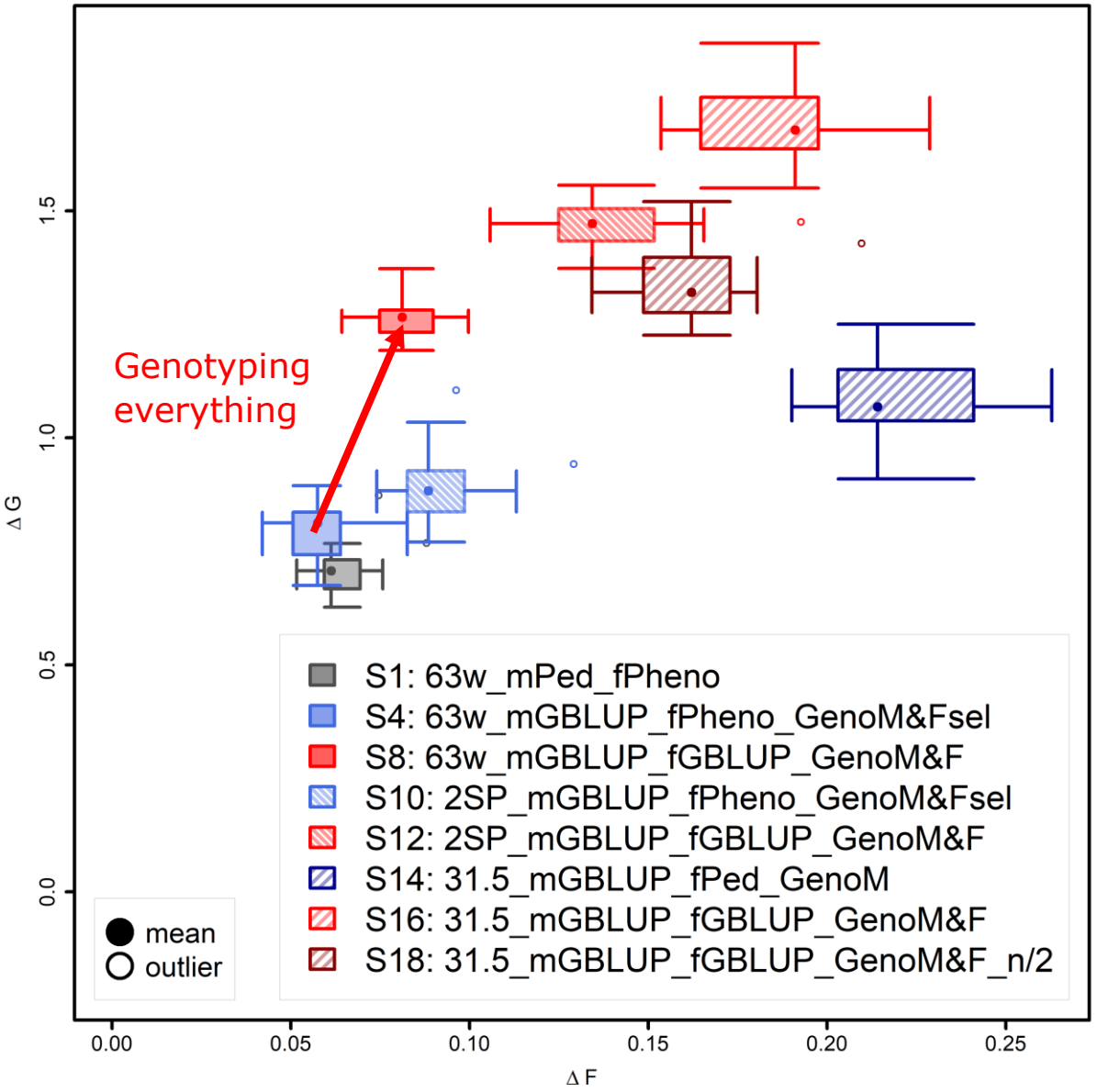
inbreeding

Genetic gain



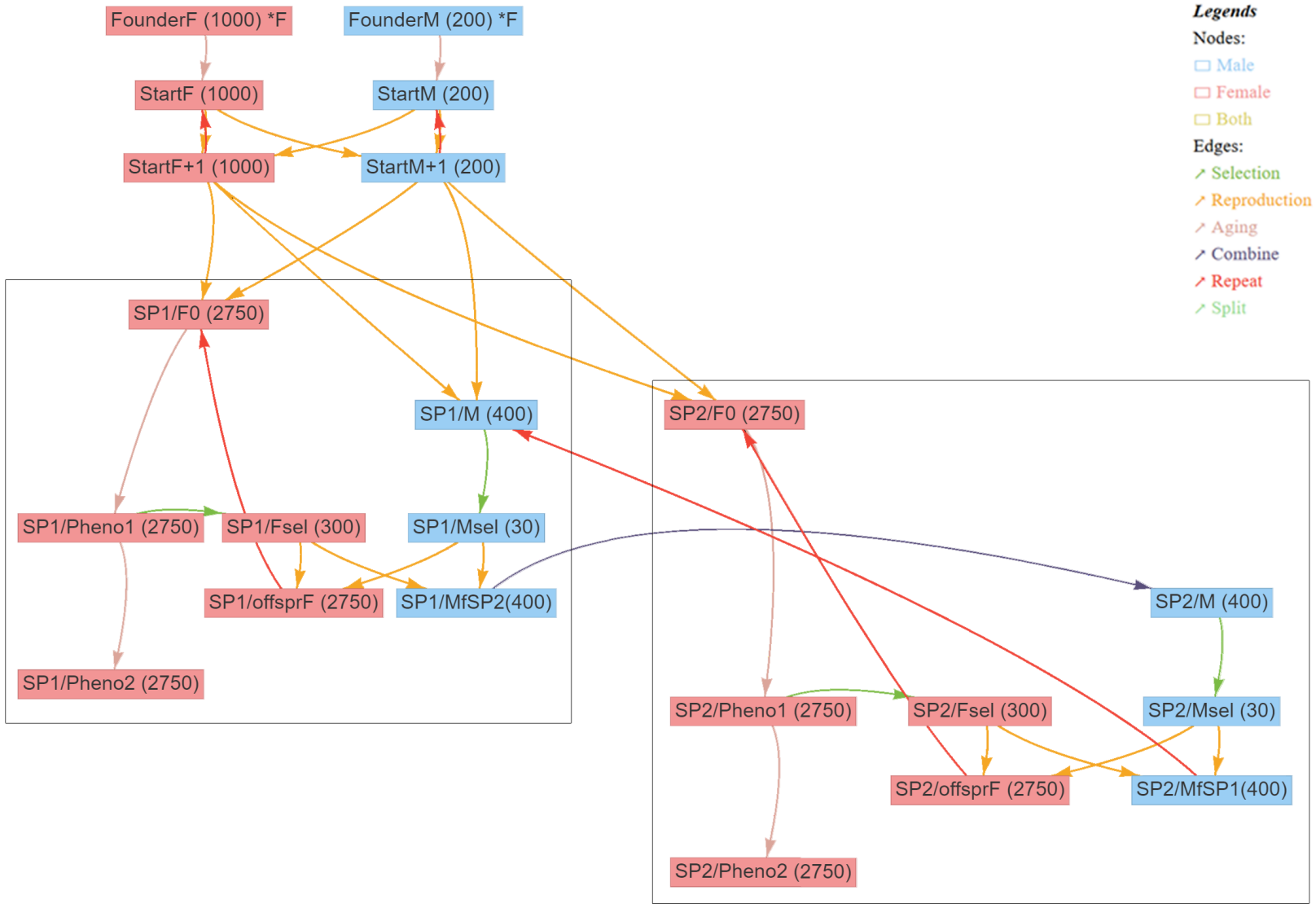
inbreeding

Genetic gain

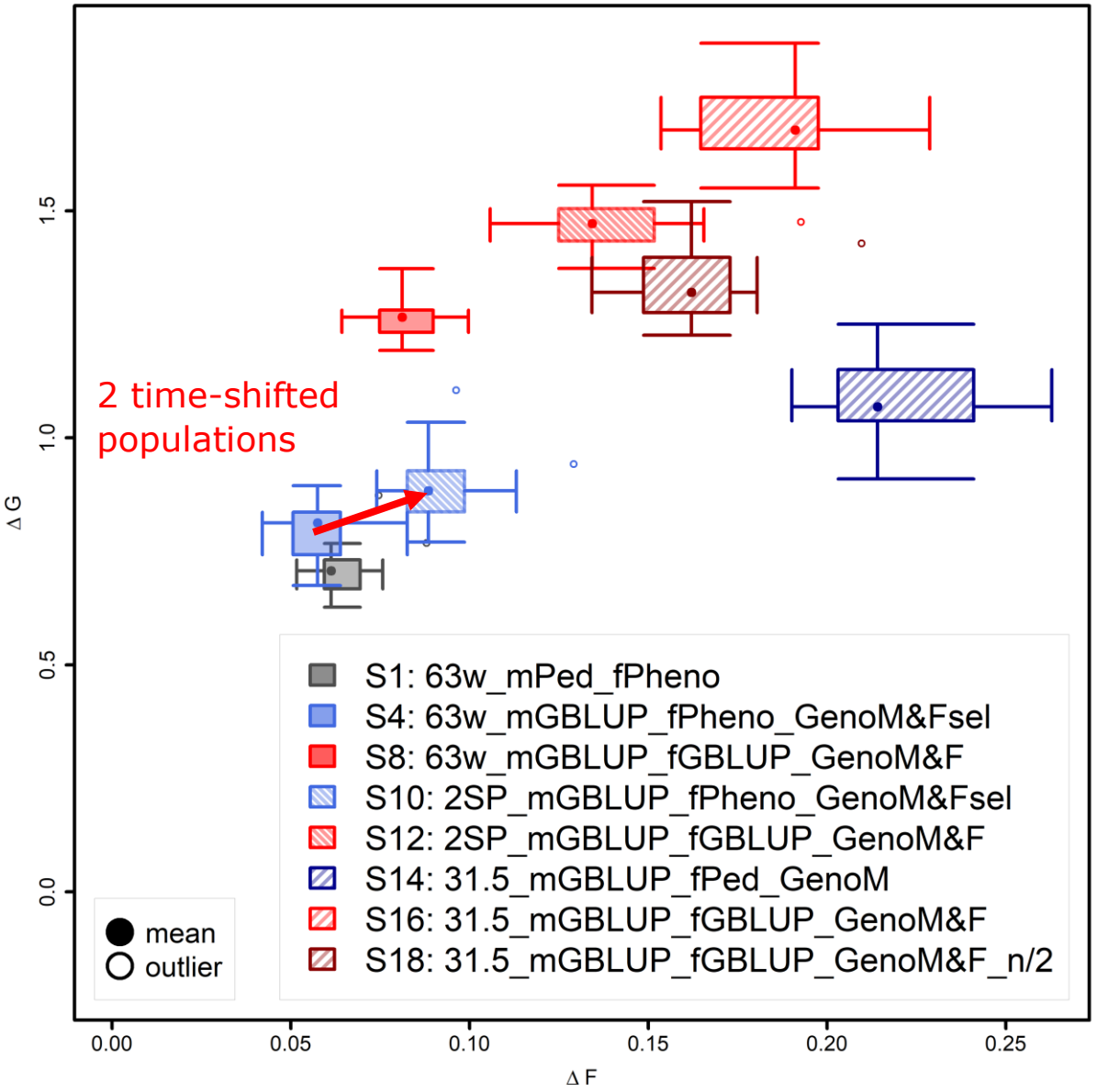


inbreeding

Shortening the generation interval in chicken (Büttgen et al., 2025)

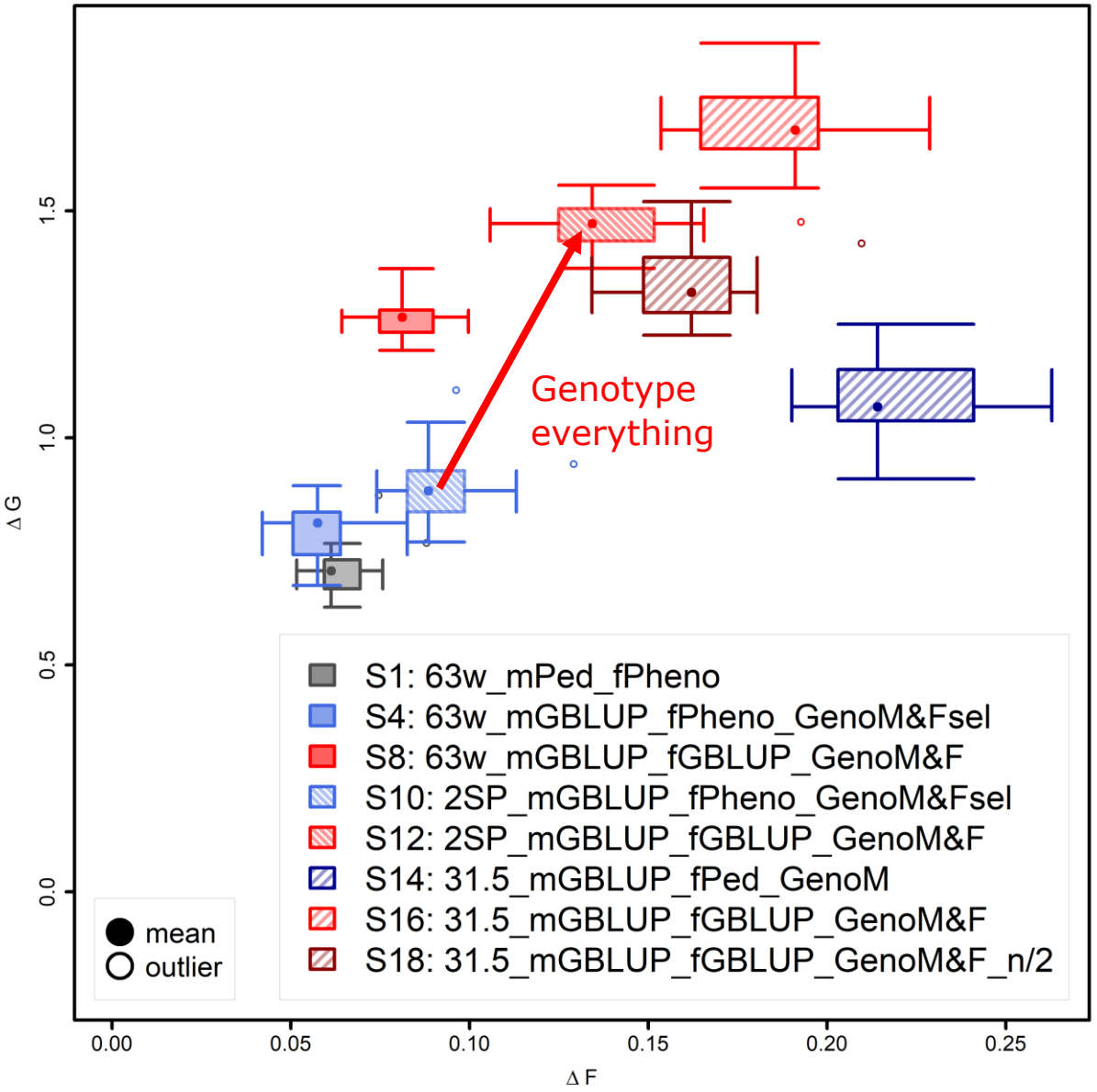


Genetic gain



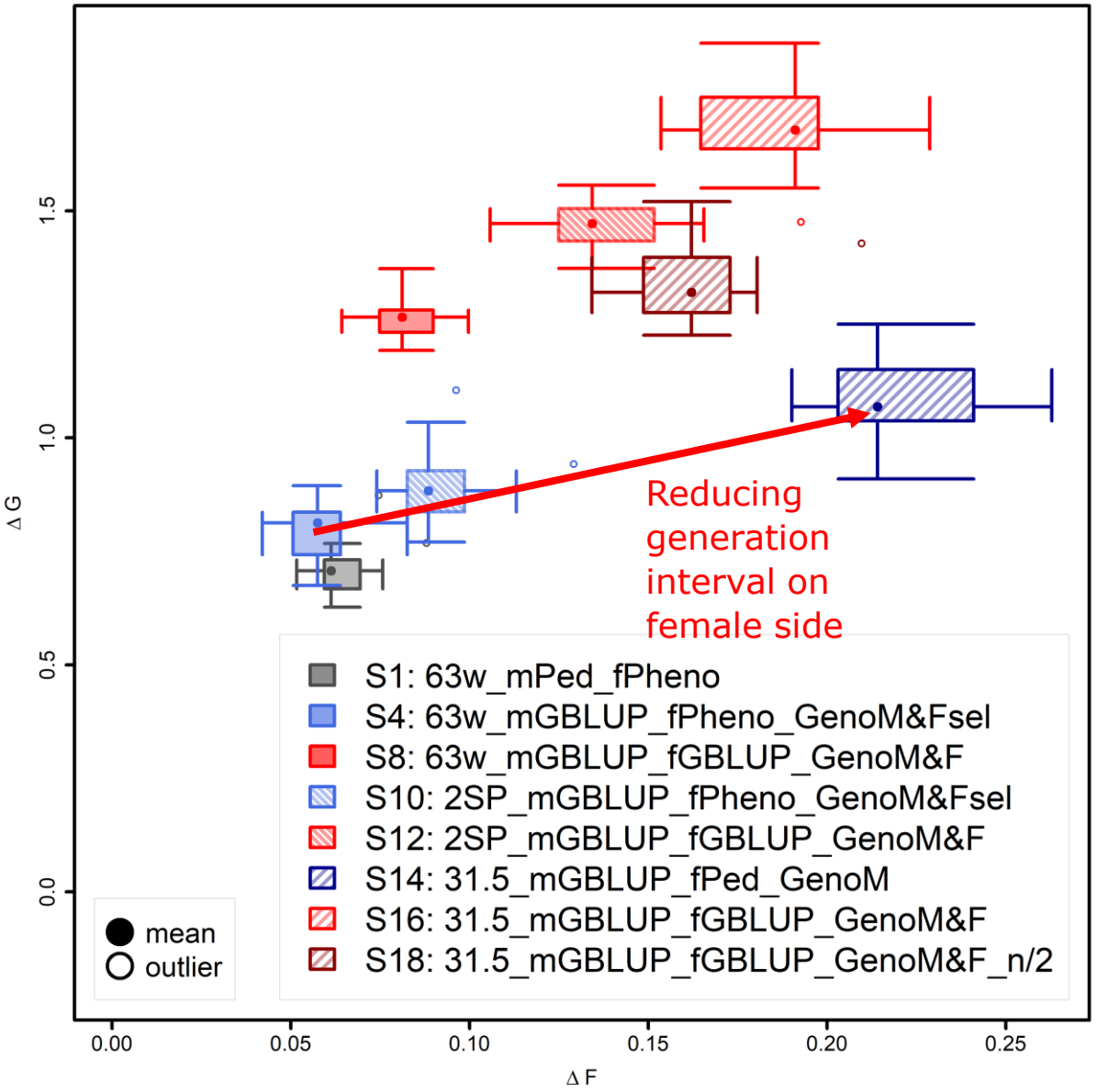
inbreeding

Genetic gain



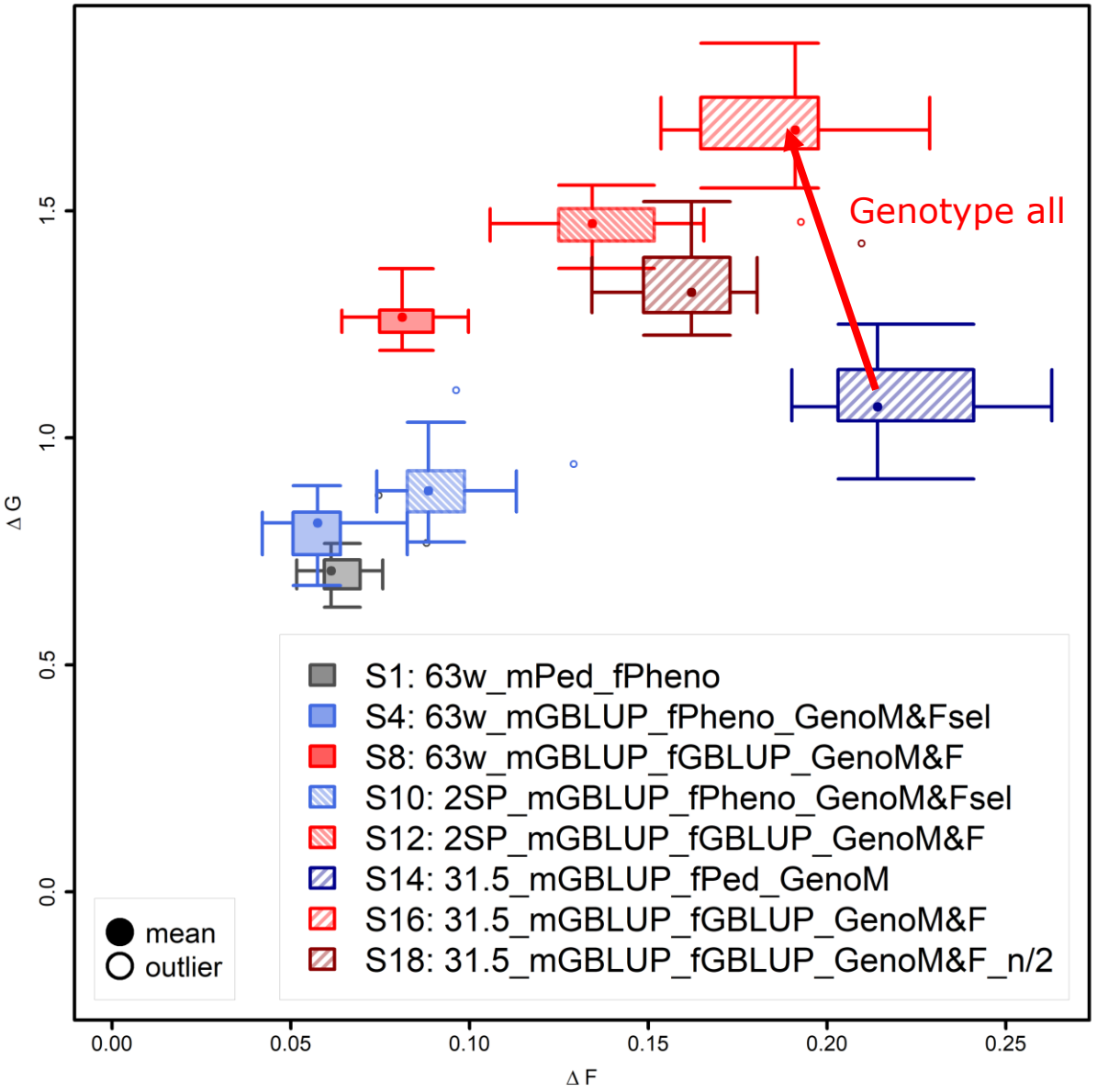
inbreeding

Genetic gain



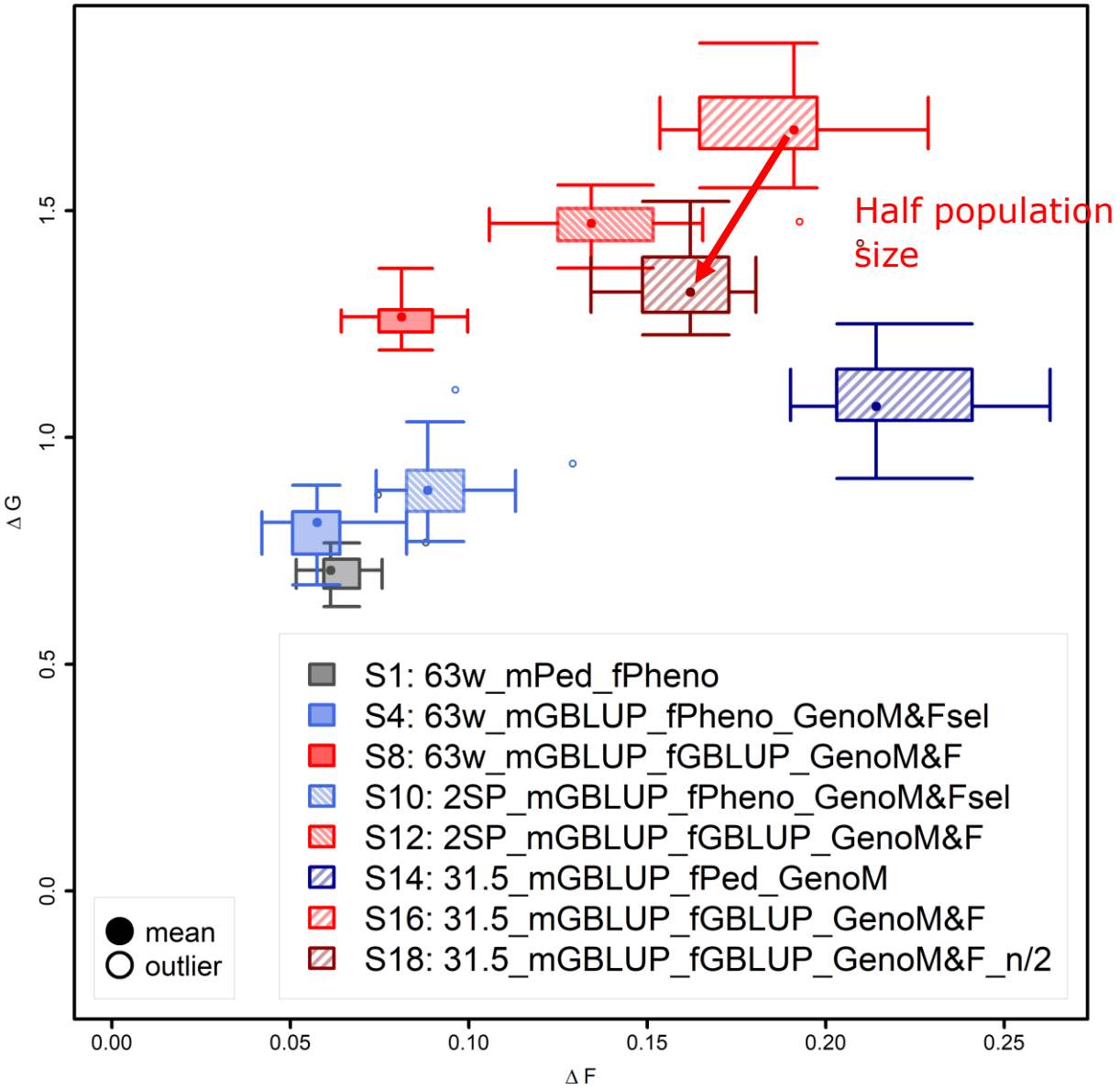
inbreeding

Genetic gain



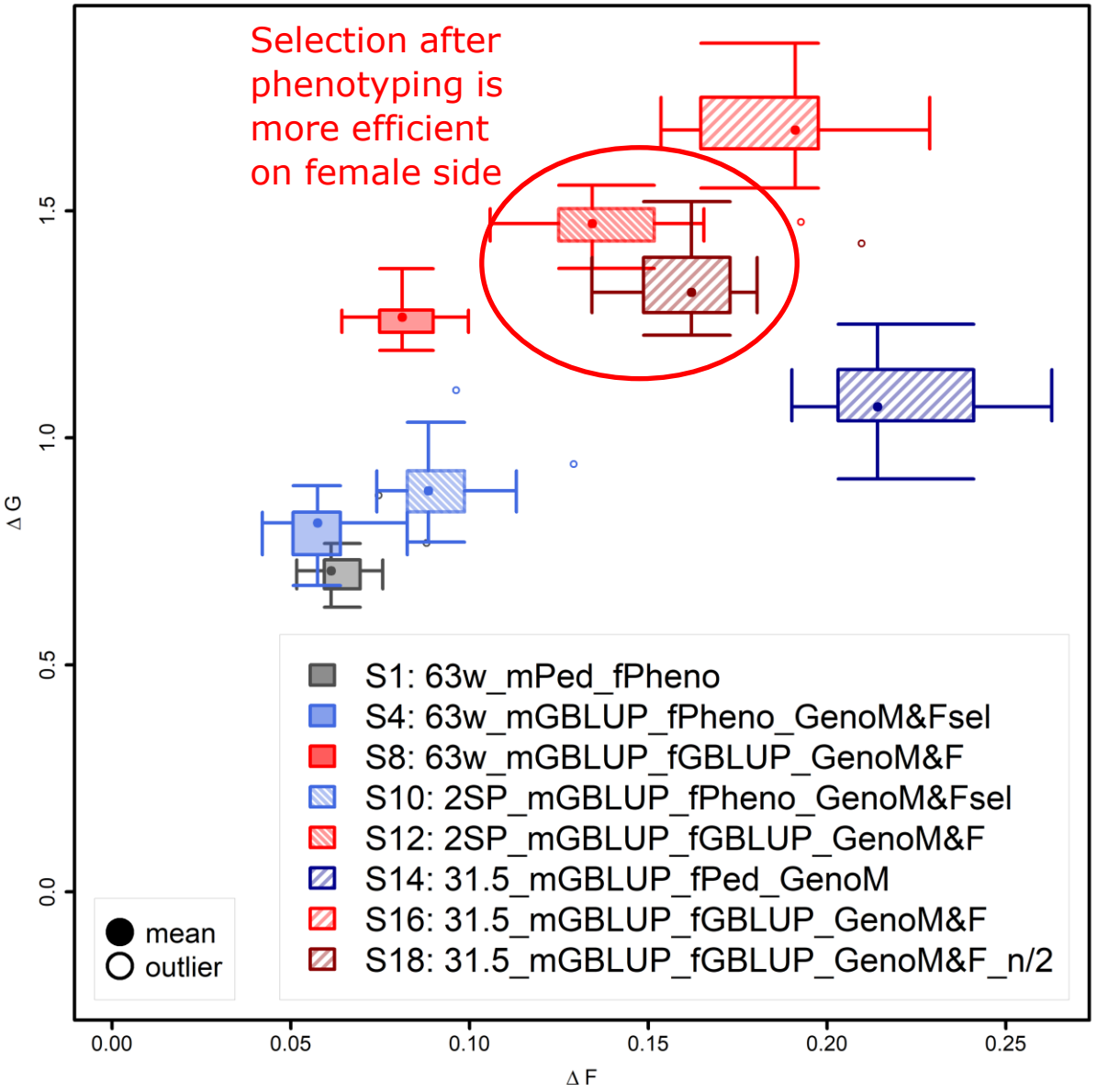
inbreeding

Genetic gain



inbreeding

Genetic gain



inbreeding

- Various small exemplary script in Guidelines (section 6)
- Exemplary script on my GitHub:
  - https://github.com/tpook92/MoBPS/tree/master/Exemplary_scripts
- Tobias Github:
  - https://github.com/tobiasniehoff/exploring_MoBPS

Questions?

Getting started with MoBPS

- MoBPS is mostly used as an R-package
- Most flexible & efficient
- Not the easiest to get started!
- User-Manual: https://github.com/tpook92/MoBPS/blob/master/Guidelines%20to%20MoBPS.pdf

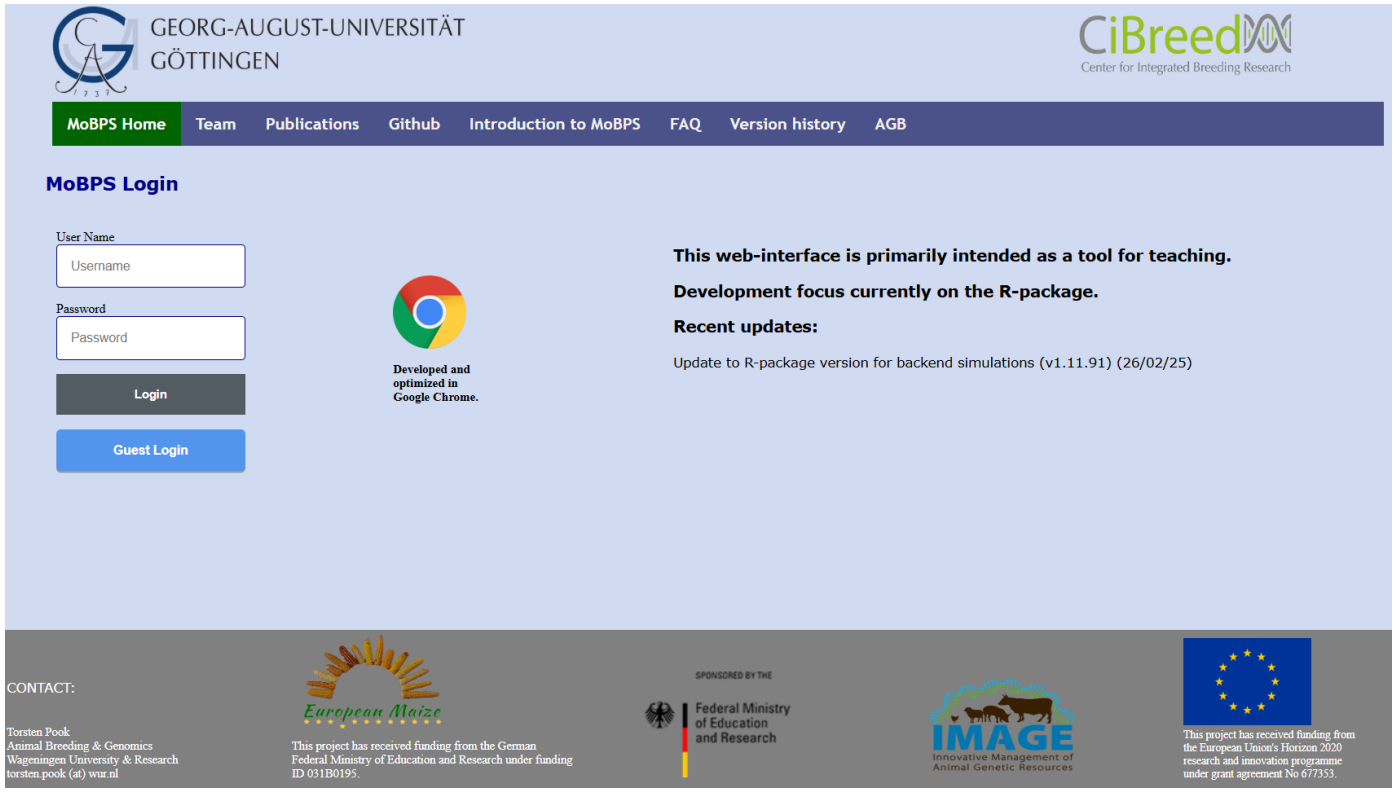
```
breeding.diploid <- function(population, mutation.rate = 10^-5, remutation.rate = 10^-5, recombination.rate = 1,
          selection.m = "random", selection.f = NULL, new.selection.calculation = TRUE, selection.function.matrix = NULL,
          selection.size = 0, ignore.best = 0, breeding.size = 0, breeding.sex = NULL, breeding.sex.random = FALSE,
          used.generations.m = 1, used.generations.f = NULL, relative.selection = FALSE, class.m = 0, class.f = 0,
          add.gen = 0, recom.f.indicator = NULL, recom.f.polynom = NULL, duplication.rate = 0,
          duplication.length = 0.01, duplication.recombination = 1, new.class = 0L, bve = FALSE, sigma.e = NULL, sigma.g = 100,
          new.bv.child = "mean", computation.A = "vanRaden", delete.haplotypes = NULL, delete.individuals = NULL,
          fixed.breeding = NULL, fixed.breeding.best = NULL, max.offspring = Inf, store.breeding.totals = FALSE, forecast.sigma.g = TRUE,
          multiple.bve = "add", multiple.bve.weights = 1, store.bve.data = FALSE, fixed.assignment = FALSE,
          reduce.group = NULL, reduce.group.selection = "random", selection.critera = c(TRUE,TRUE), selection.criteria.type = c("bve", "bve"),
          same.sex.activ = FALSE, same.sex.sex = 0.5, same.sex.selfing = TRUE, selfing.mating = FALSE, selfing.sex = 0.5,
          praeimplantation = NULL, heritability = NULL, multiple.bve.scale = FALSE, use.last.sigma.e = FALSE,
          save.recombination.history = FALSE, martini.selection = FALSE, BGLR.bve = FALSE, BGLR.burnin = 500,
          BGLR.iteration = 5000, copy.individual = FALSE, dh.mating = FALSE, dh.sex = 0.5, n.observation = 1,
          bve.0isNA = TRUE, phenotype.bv = FALSE, standardize.bv = FALSE, standardize.bv.level = 100,
          standardize.bv.gen = 1, delete.same.origin = FALSE, remove.effect.position = FALSE, estimate.u = FALSE,
          BGLR.print = FALSE, new.phenotype.correlation = NULL, new.breeding.correlation = NULL, estimate.add.gen.var = FALSE,
          estimate.pheno.var = FALSE, best1.from.group = NULL, best2.from.group = NULL, best1.from.cohort = NULL,
          best2.from.cohort = NULL, add.class.cohorts = TRUE, store.comp.times = TRUE, store.comp.times.bve = TRUE,
          store.comp.times.generation = TRUE, special.comb = FALSE, max.auswahl = Inf, predict.effects = FALSE,
          SNP.density = 10, use.effect.markers = FALSE, use.effect.combination = FALSE, import.position.calculation = NULL,
          special.comb.add = FALSE, BGLR.save = "RKHS", BGLR.save.random = FALSE, ogc = FALSE, ogc_cAc = NA, emmreml.bve = FALSE,
          sommer.bve = FALSE, breedR.bve = FALSE, breedR.groups = NULL, nr.edits = 0, gene.editing.offspring = FALSE,
          gene.editing.best = FALSE, gene.editing.offspring.sex = c(TRUE,TRUE), gene.editing.best.sex = c(TRUE,TRUE),
          gwas.u = FALSE, approx.residuals = TRUE, sequenceZ = FALSE, maxZ = 5000, maxZtotal = 0, delete.sex = 1:2,
          gwas.group.standard = FALSE, y.gwas.used = "pheno", gen.architecture.m = 0, gen.architecture.f = NULL,
          add.architecture = NULL, ncore = 1, ncore.generation = 1, Z.integer = FALSE, store.effect.freq = FALSE,
          backend = "doParallel", randomSeed = NULL, randomSeed.generation = NULL, Rprof = FALSE, miraculix = FALSE,
          miraculix.mult = NULL, fast.compiler = 0, miraculix.cores = 1, store.bve.parameter = FALSE, miraculix.chol = TRUE,
          best.selection.ratio.m = 1, best.selection.ratio.f = NULL, best.selection.criteria.m = "bv", best.selection.criteria.f = NULL,
          best.selection.manual.ratio.m = NULL, best.selection.manual.ratio.f = NULL, bve.class = NULL, parallel.generation = FALSE,
          name.cohort = NULL, display.progress = TRUE, max.ticks = Inf, combine = FALSE, repeat.mating = 1, time.point = 0,
          creating.type = 0, multiple.observation = FALSE, new.bv.observation = NULL, new.bv.observation.gen = NULL,
          new.bv.observation.cohorts = NULL, new.bv.observation.database = NULL, bve.gen = NULL, bve.cohorts = NULL,
          bve.database = NULL, sigma.e.gen = NULL, sigma.e.cohorts = NULL, sigma.e.database = NULL, sigma.g.gen = NULL,
          sigma.g.cohorts = NULL, sigma.g.database = NULL, gwas.gen = NULL, gwas.cohorts = NULL, gwas.database = NULL,
          bve.insert.gen = NULL, bve.insert.cohorts = NULL, bve.insert.database = NULL, reduced.selection.panel.m = NULL,
          reduced.selection.panel.f = NULL, breeding.all.combination = FALSE, depth.pedigree = Inf, copy.individual.keep.bve = TRUE,
          bve.avoid.duplicates = TRUE, report.accuracy = TRUE, share.genotyped = 1, singlestep.active = FALSE,
          remove.non.genotyped = TRUE, added.genotyped = 0, fast.uhat = FALSE, offspring.bve.parents.gen = NULL,
          offspring.bve.parents.database = NULL, offspring.bve.parents.cohort = NULL, offspring.bve.offspring.gen = NULL,
          offspring.bve.offspring.database = NULL, offspring.bve.offspring.cohort = NULL){
```

Table of Contents

WA UNIVE

MoBPSweb

- Web-based application to enter a breeding program in a more intuitive way (www.mobps.de)



Some general tips

- Change your password
- Frequently press “Save”
- Exemplary templates can provide inspiration
- Do not ignore these buttons:

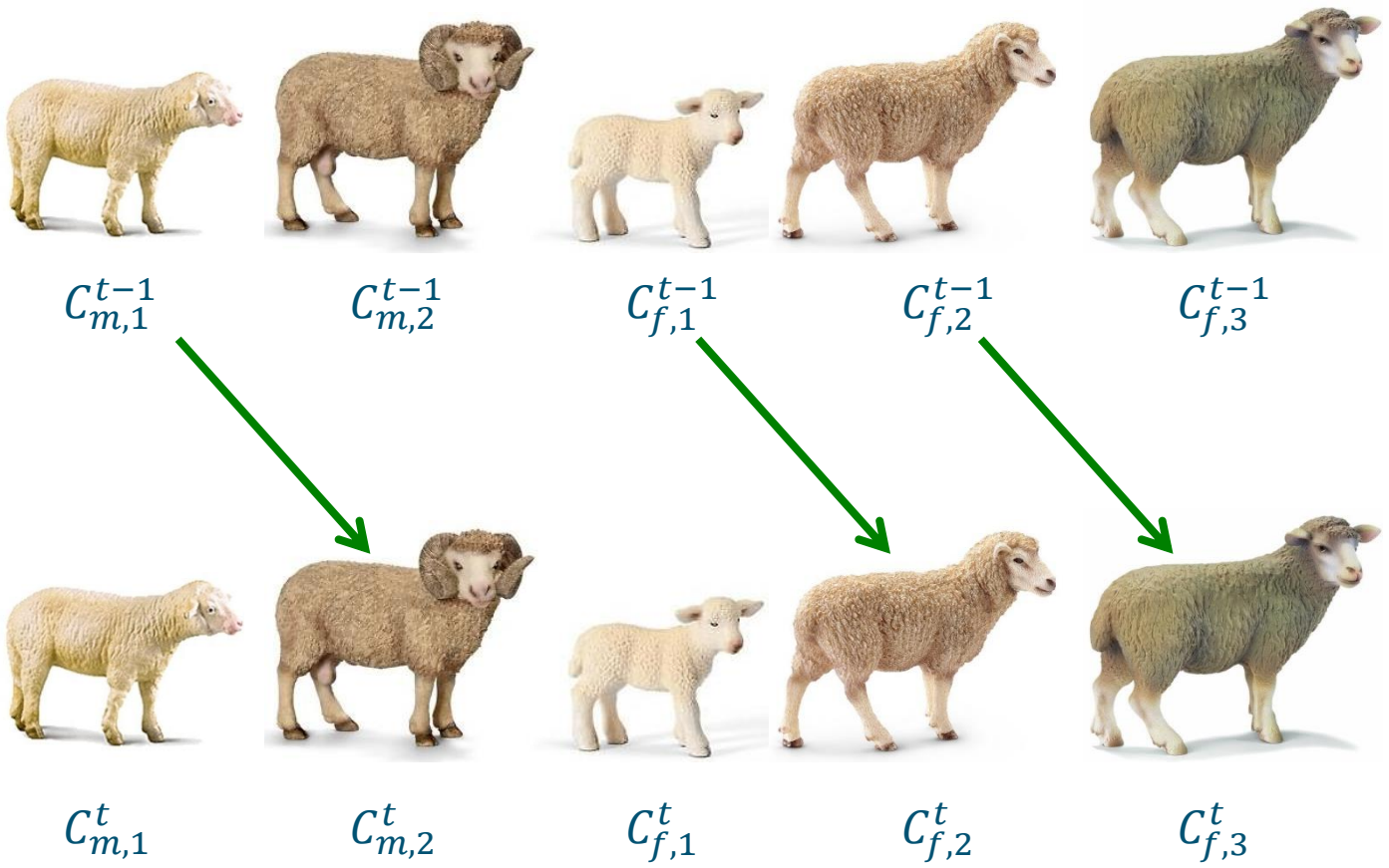
- Do not ignore warnings:

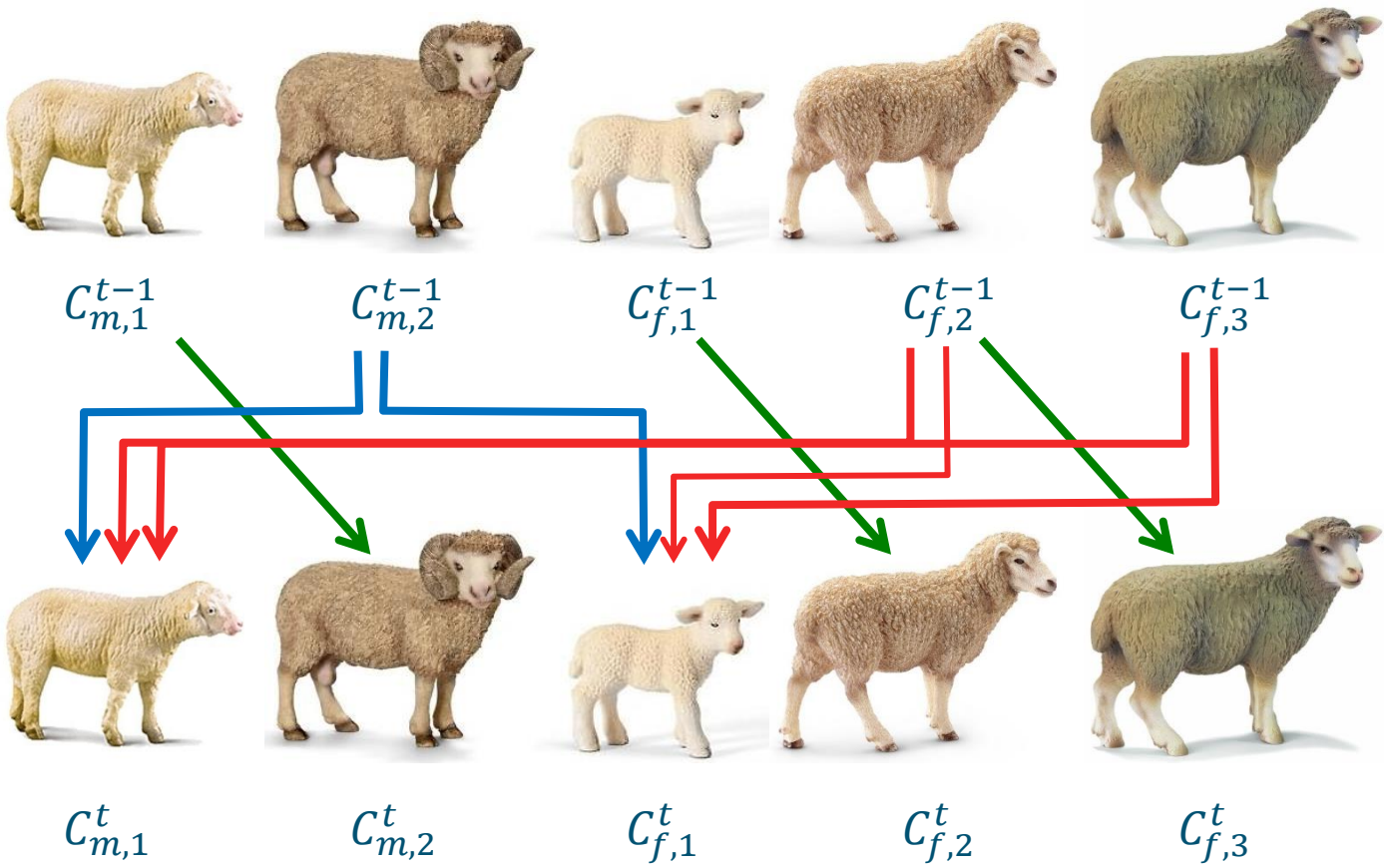
Attention, there are 2 warnings! R Simulation most likely cannot be run unless they are fixed.

Warnings

Warning 0 : Heritability for phenotype1 must be between 0 and 1

Warning 1 : Different sex between nodes SelectedCows and Bulls


$C_{m,1}^{t-1}$
$C_{m,2}^{t-1}$
$C_{f,1}^{t-1}$
$C_{f,2}^{t-1}$
$C_{f,3}^{t-1}$
$C_{m,1}^{t}$
$C_{m,2}^{t}$
$C_{f,1}^{t}$
$C_{f,2}^{t}$
$C_{f,3}^{t}$


$C_{m,1}^{t-1}$
$C_{m,2}^{t-1}$
$C_{f,1}^{t-1}$
$C_{f,2}^{t-1}$
$C_{f,3}^{t-1}$
$C_{m,1}^{t}$
$C_{m,2}^{t}$
$C_{f,1}^{t}$
$C_{f,2}^{t}$
$C_{f,3}^{t}$

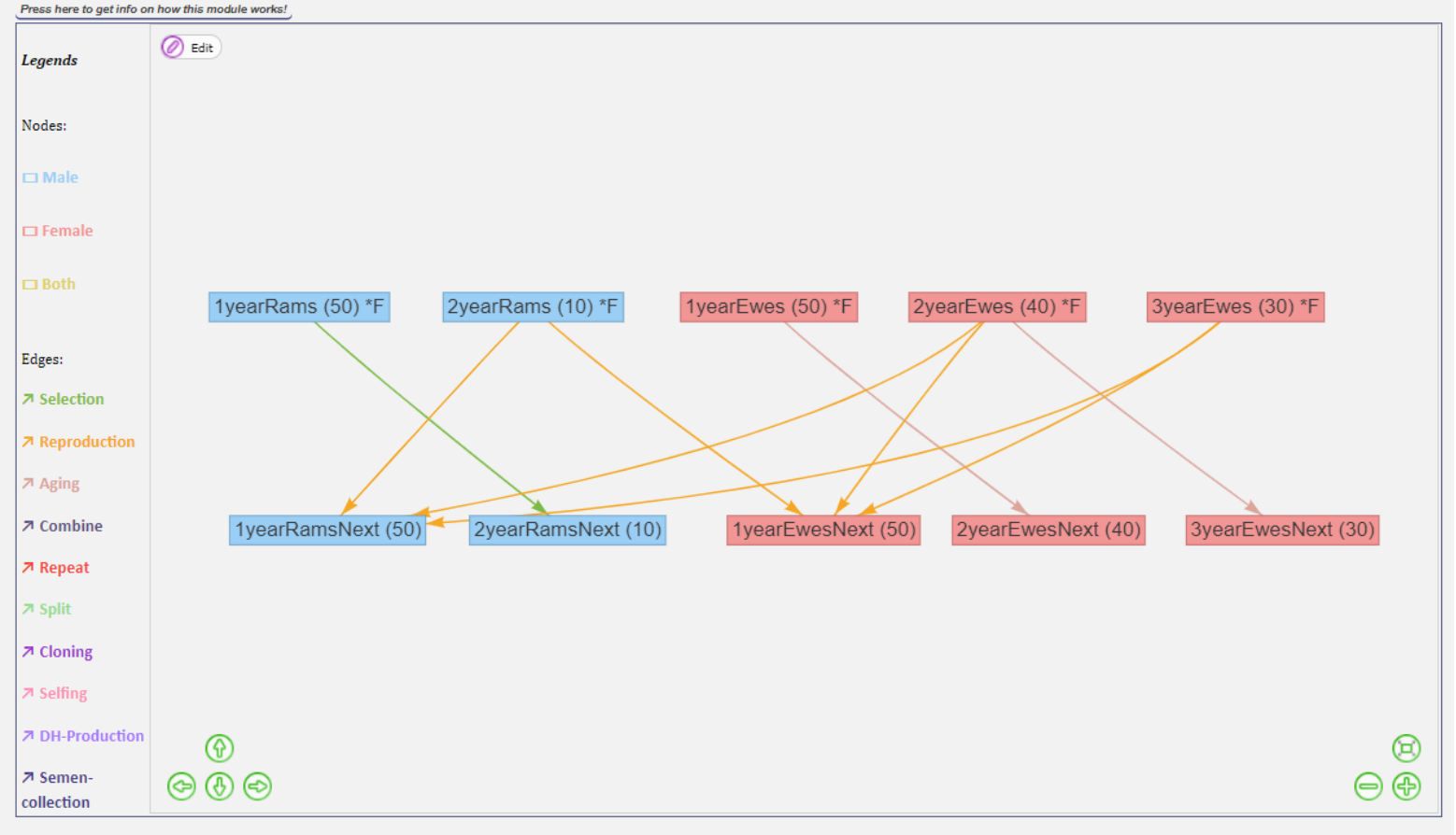
Task 1: Sheep Breeding Program

- Simulate one cycle of the sheep breeding program with:
  - 50 1-year rams & 10 2-year rams
  - 50 1-year ewes & 40 2-year ewes & 30 3-year ewes
- One trait (Meat)
  - Phenotypic mean: 100
  - Heritability $h^2 = 0.3$, 1000 QTLs
- Selection on the male side is based on phenotypes
- Selection on the female side is done at random
- 2-year old rams and 2 & 3 –year old ewes are used for reproduction
- Simulate a genome with 5 chromosomes with 1000 SNPs / 1 Morgan

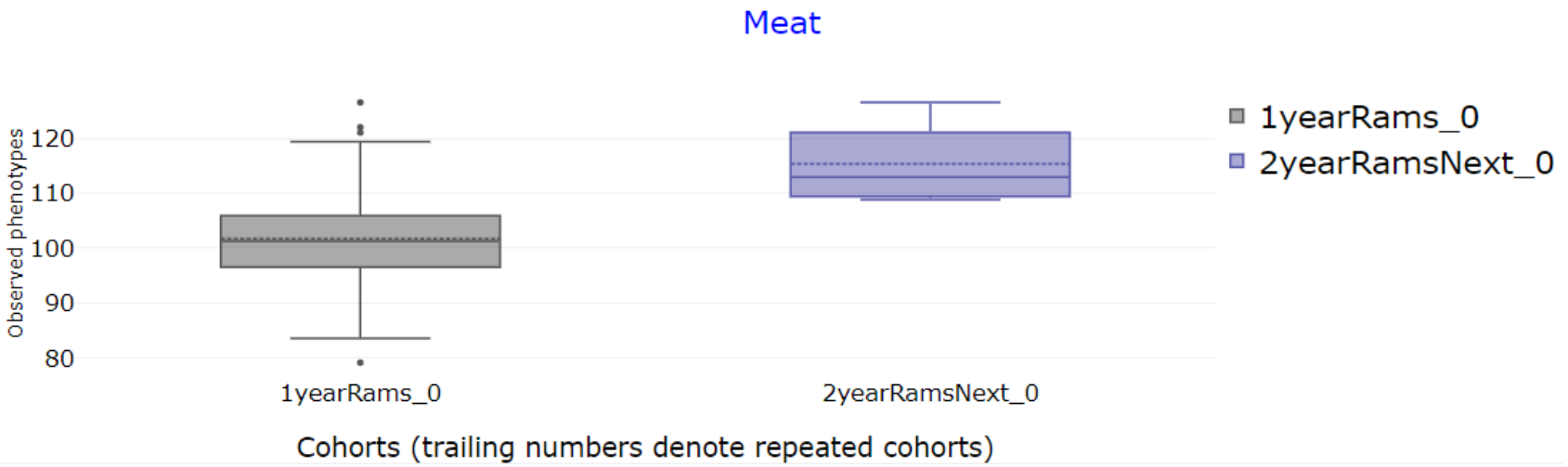
WAGENINGEN
UNIVERSITY & RESEARCH

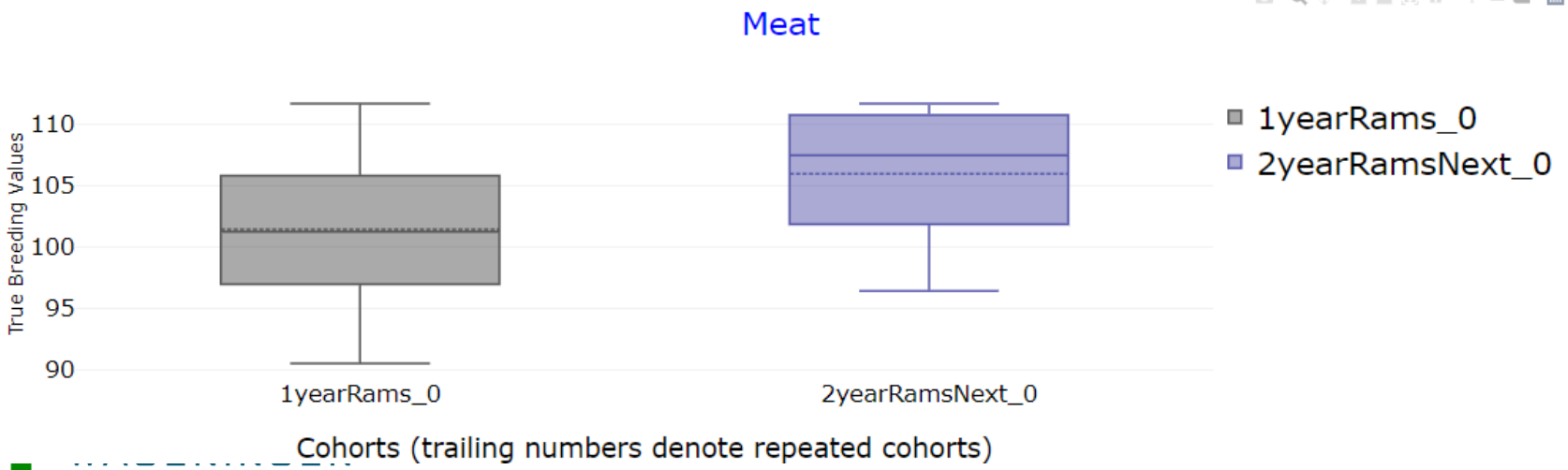
Solution Task 1

- All details are given in the template: „Simple Sheep“



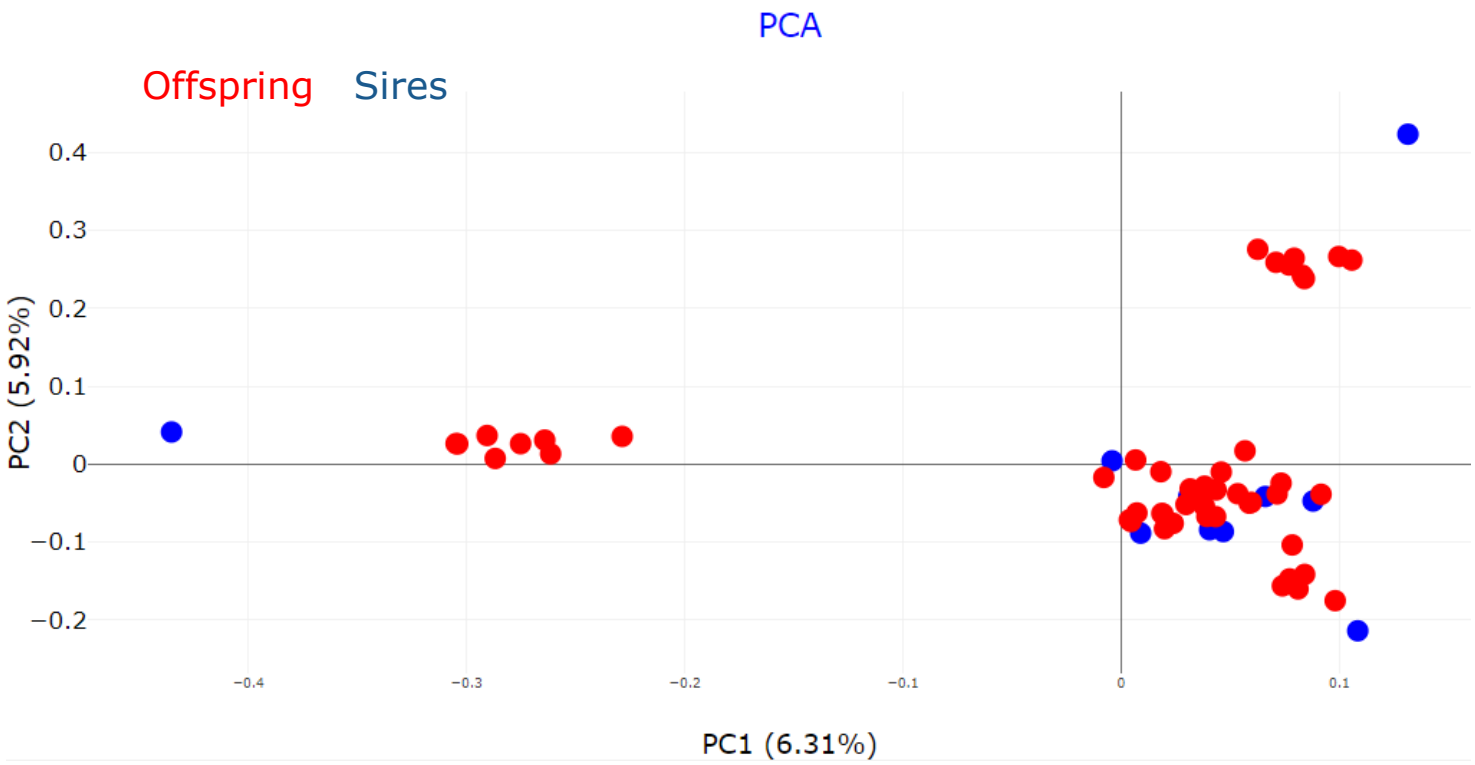
Solution Task 1



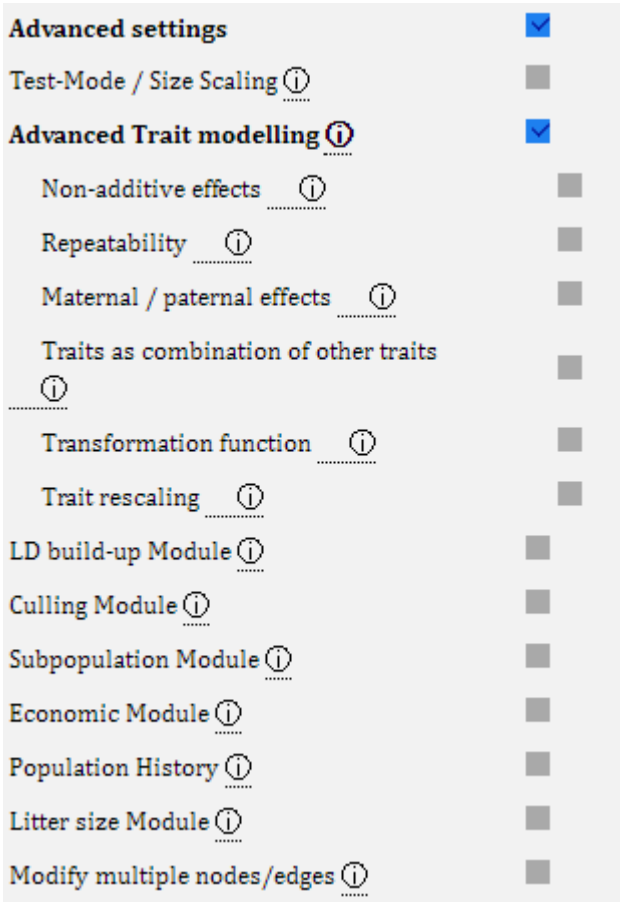


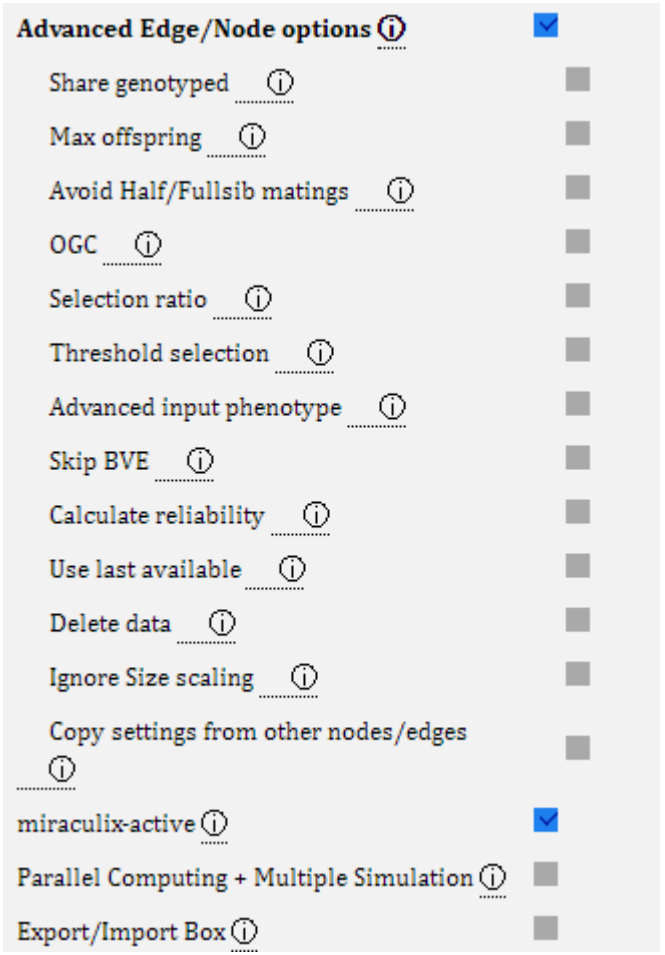
UNIVERSITY & RESEARCH

Solution Task 1



Advanced options in MoBPSweb

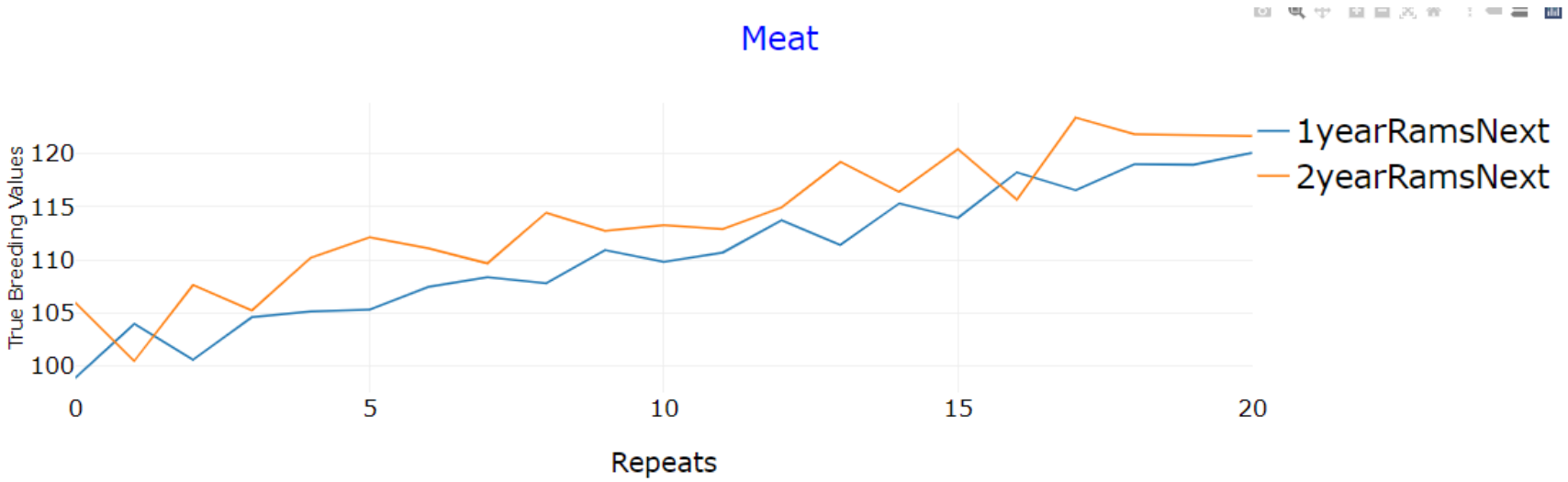


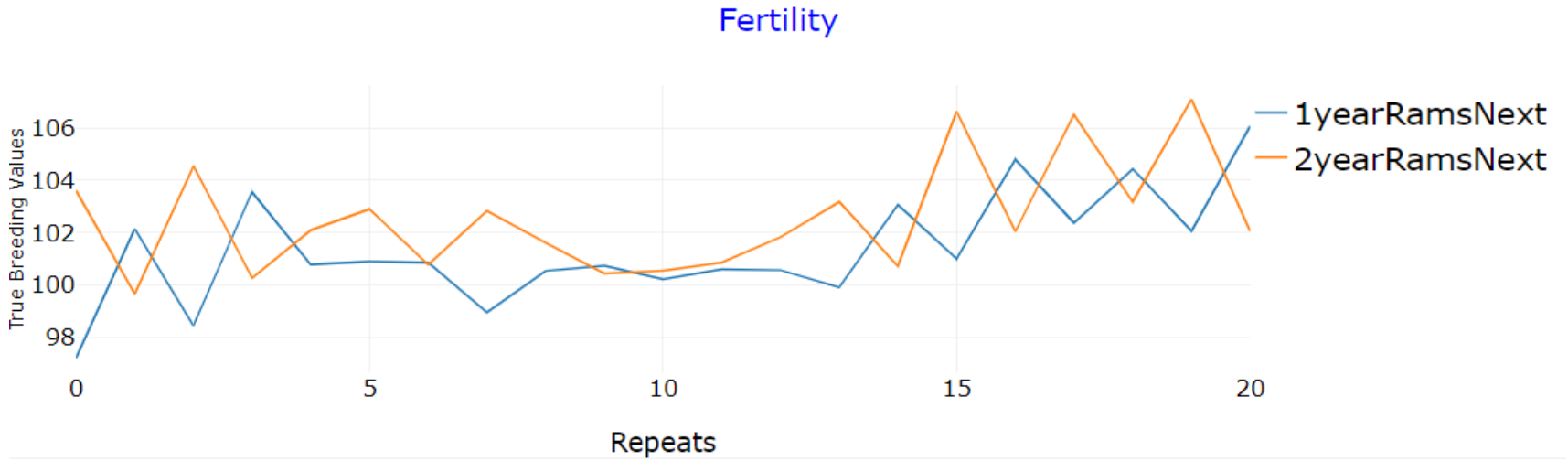


Task 2: Make it more realistic

- Perform an LD build up to ensure that founders are related
  - Use 5 generations with 100 individuals
- Add a second trait (Fertility)
  - Phenotypic Mean: 100 & Heritability $h^2 = 0.2$
  - Only two and three-year-old animals should be phenotyped
  - Traits have a genetic correlation of 0.2
  - Residual effects are not correlated
- Simulate 20 cycles of the breeding program
- The number of offspring from a given mating (litter) should be 2 with a probability of 50%, 3 with a probability of 30% and 4 with a probability of 20%
- Use genomic selection for the selection of rams
  - Use all animals from the current cycle in the breeding value estimation
  - Put equal weight on both traits (use the scaling: „Per Genomic Value SD“)
  - Assume all individuals to be genotyped
- Use the IlluminaOvineSNP50 array as an underlying genomic map

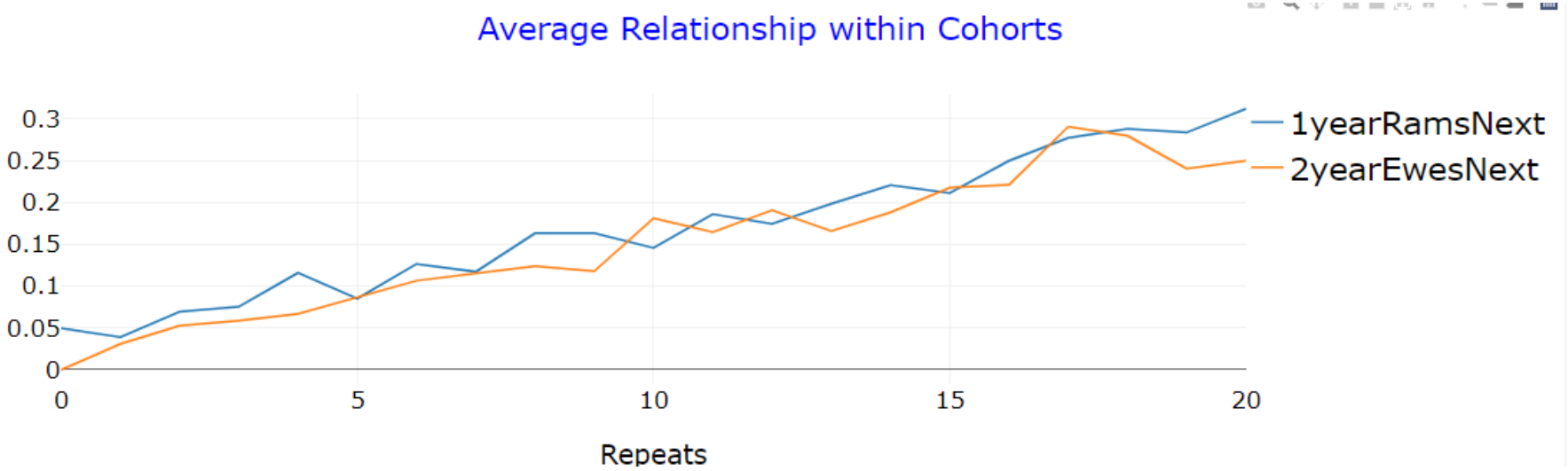
Solution Task 2

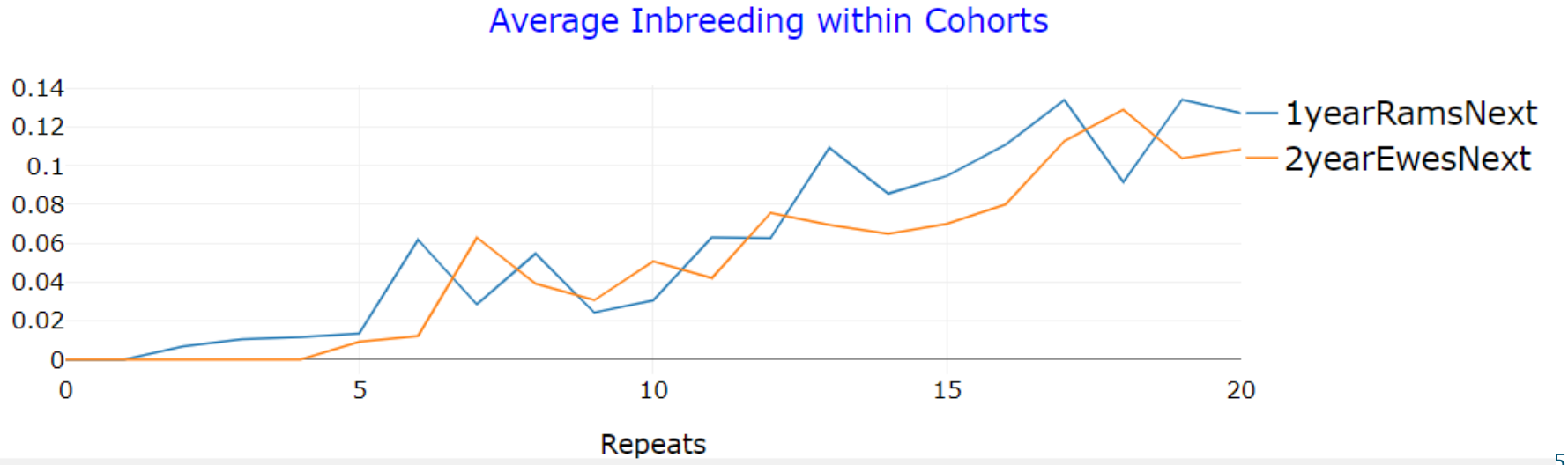




WAGENINGEN
UNIVERSITY & RESEARCH

Solution Task 2

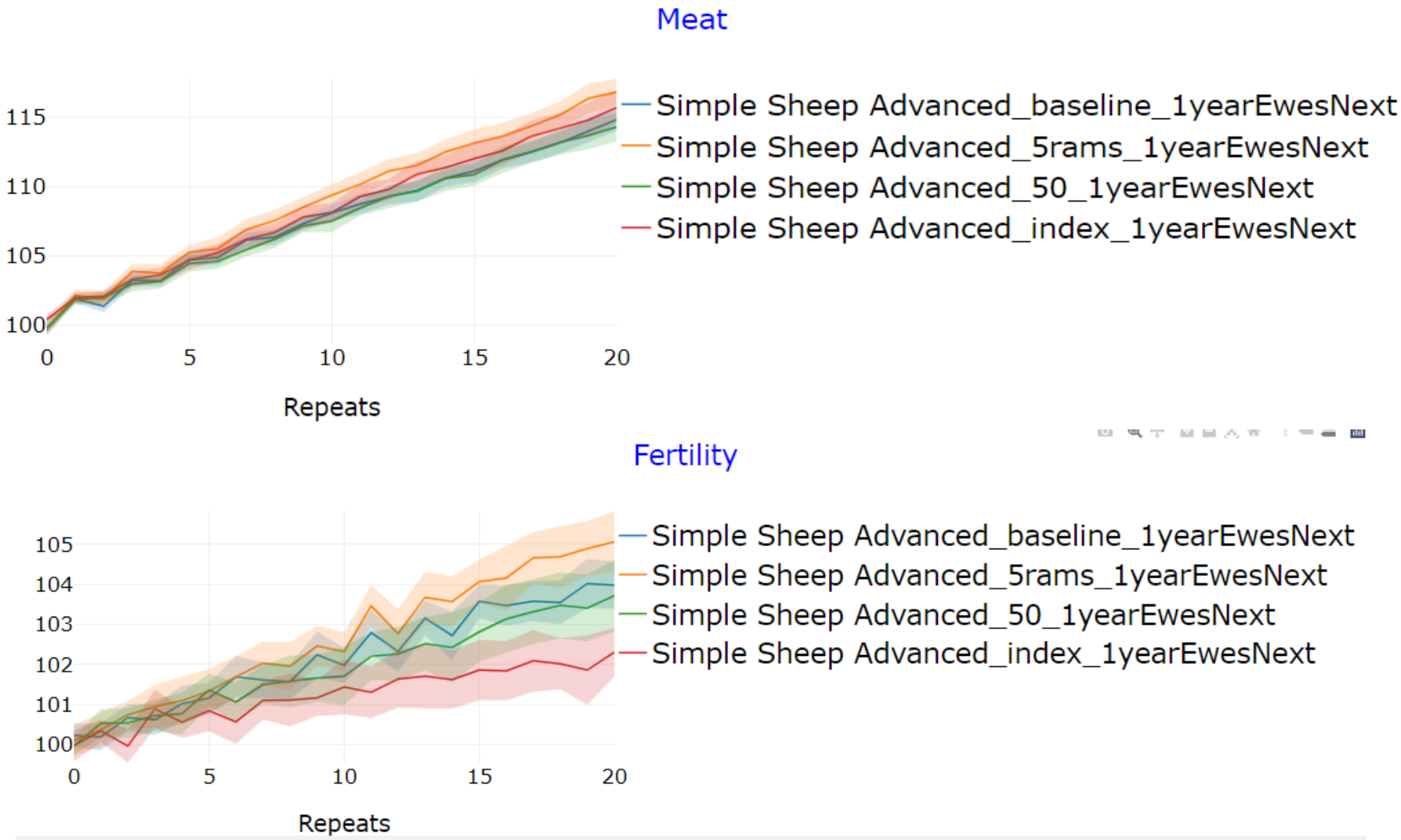




Task 3: Comparison of scenarios

- Consider the following different variations of the breeding scheme as separate projects:
  1. Only 5 rams are selected for reproduction
  2. Only 50% of all ewes are genotyped
  3. Use a selection index with three times as much weight on the meat trait
- Simulate each breeding scheme at least 3 times
- Use the „Compare Project“ module to compare the different scenarios in regard to genetic gain and inbreeding
- Hint: Confidence intervals could be benefical to decide which differences are significant and which are not!

Solution Task 3



Solution Task 3

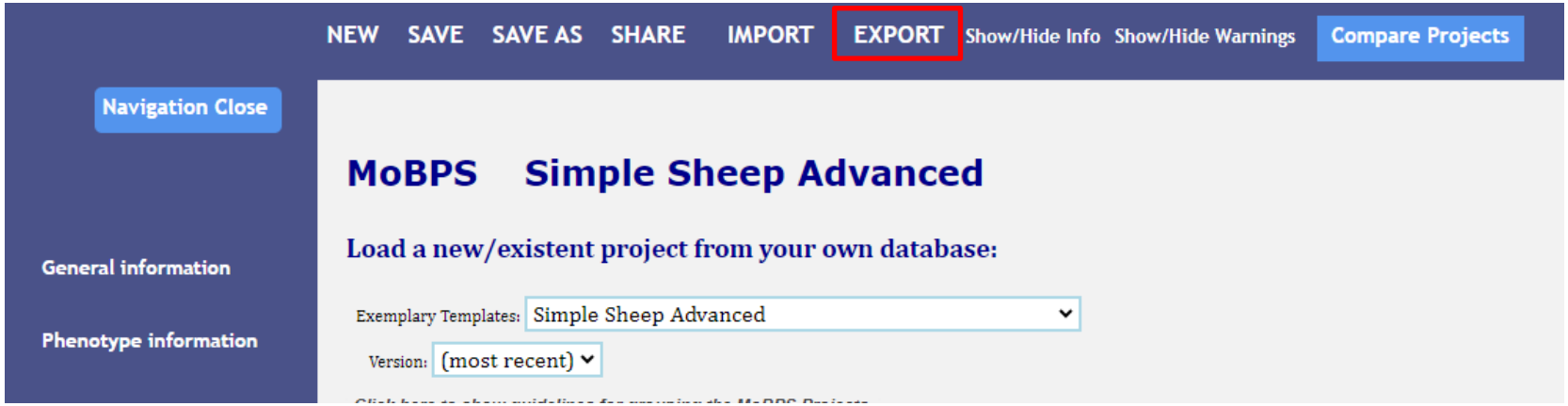
- Even 25 simulations over 20 generations are barely enough to distinguish between scenarios
  - High variance in the outcome of simulations
  - Small differences between scenarios
- Breeding in practice also has random factors!

Task 4

- Install our packages from source / local file (not CRAN)
- Files are available at https://www.dropbox.com/scl/fo/n0h8fa92gubv0mnj8qylj/ANeUixV1SN1Mvf6B0onhPU8?rlkey=b6sq2uztthzvkq67582g5c1m5&st=gzatg6rk&dl=0
- MoBPS (V1.12.000)
    - For Windows this requires Rtools (https://cran.r-project.org/bin/windows/Rtools/rtools40.html) on some systems
- Usually you will find the most recent version at https://github.com/tpook92/MoBPS

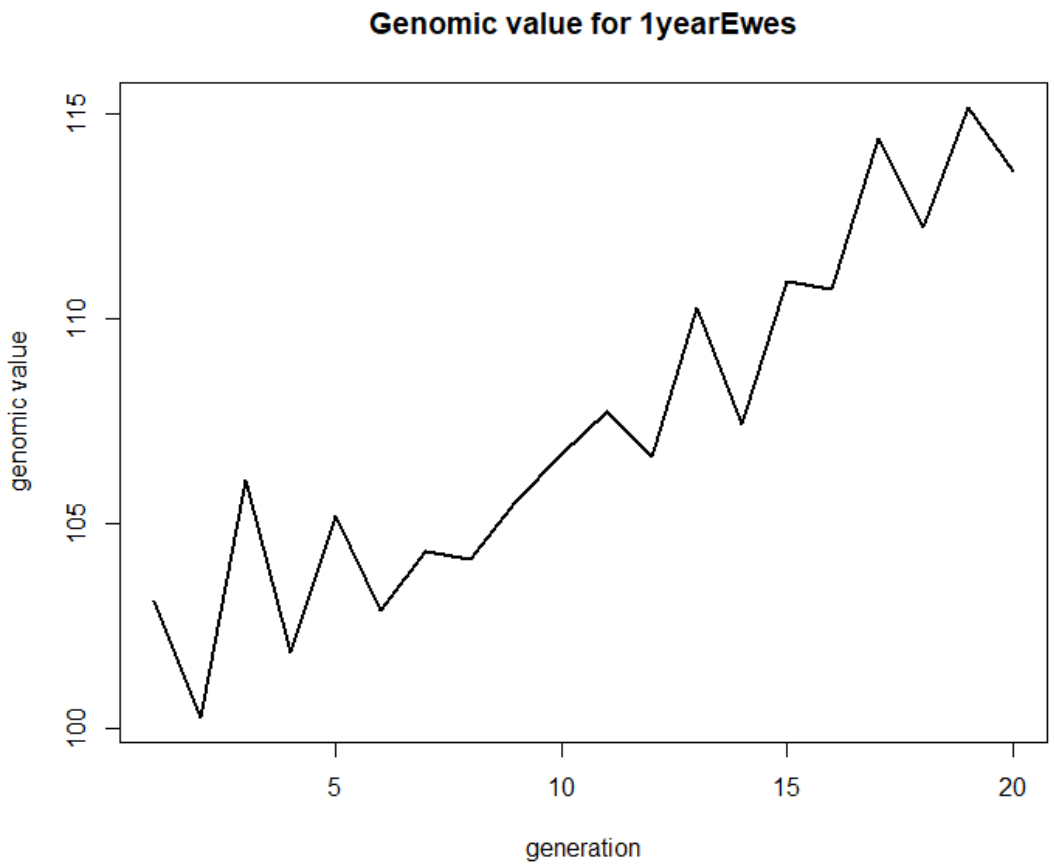
- For computational efficiency:
  - RandomFieldsUtils & Miraculix
  - Examples in this workshop will be so small that this is not needed!
  - Only specific versions of these package work together
  - Don’t expect frequent updates on these packages
  - Current recommendation: v.1.5.1.1 for RandomFieldsUtils & miraculix on all operating systems
- To use real genetic maps: MoBPSmaps V0.1.14

To check if you are properly set up



- Simulate the scheme by use of *json.simulation()*
  - To write a log-file set the parameter log to the path you want the file to be written to

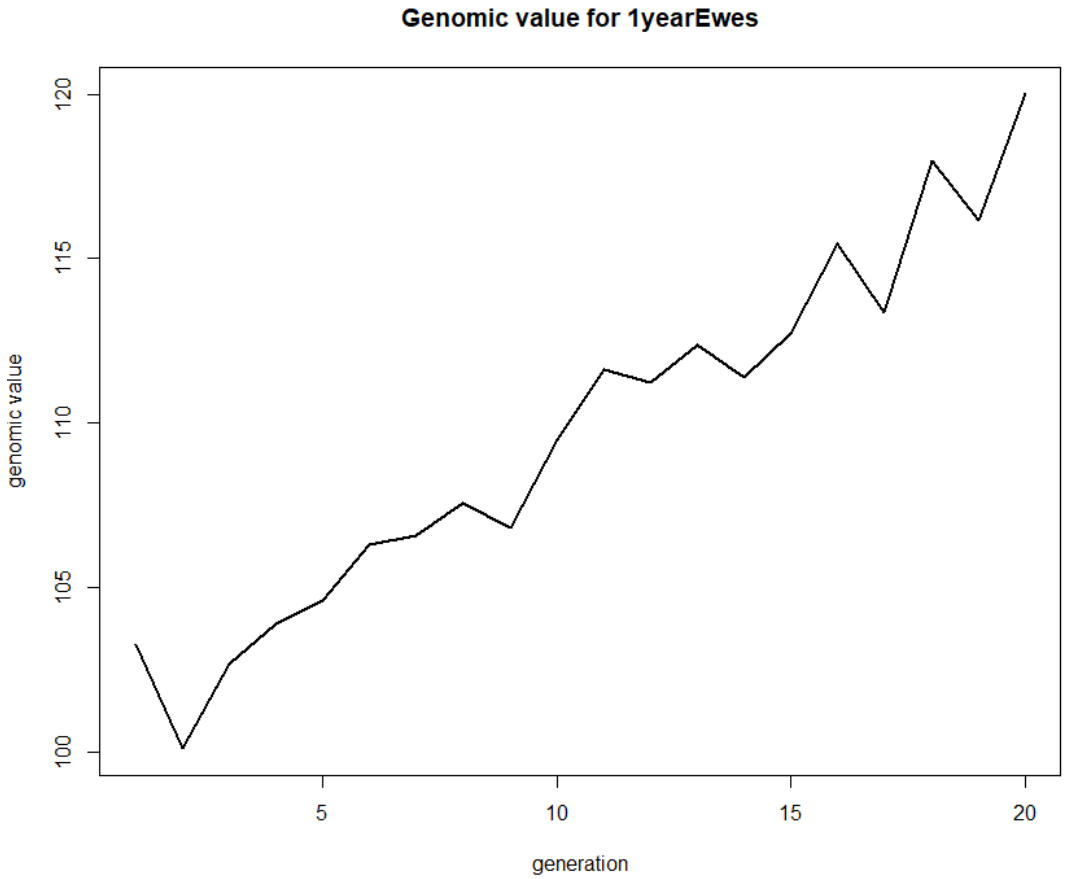
- Extract the underlying true Breeding Values for the 1 year old ewes for the different cycles
- Exemplary code: Task4/Simulation_from_JSON.R



Solution

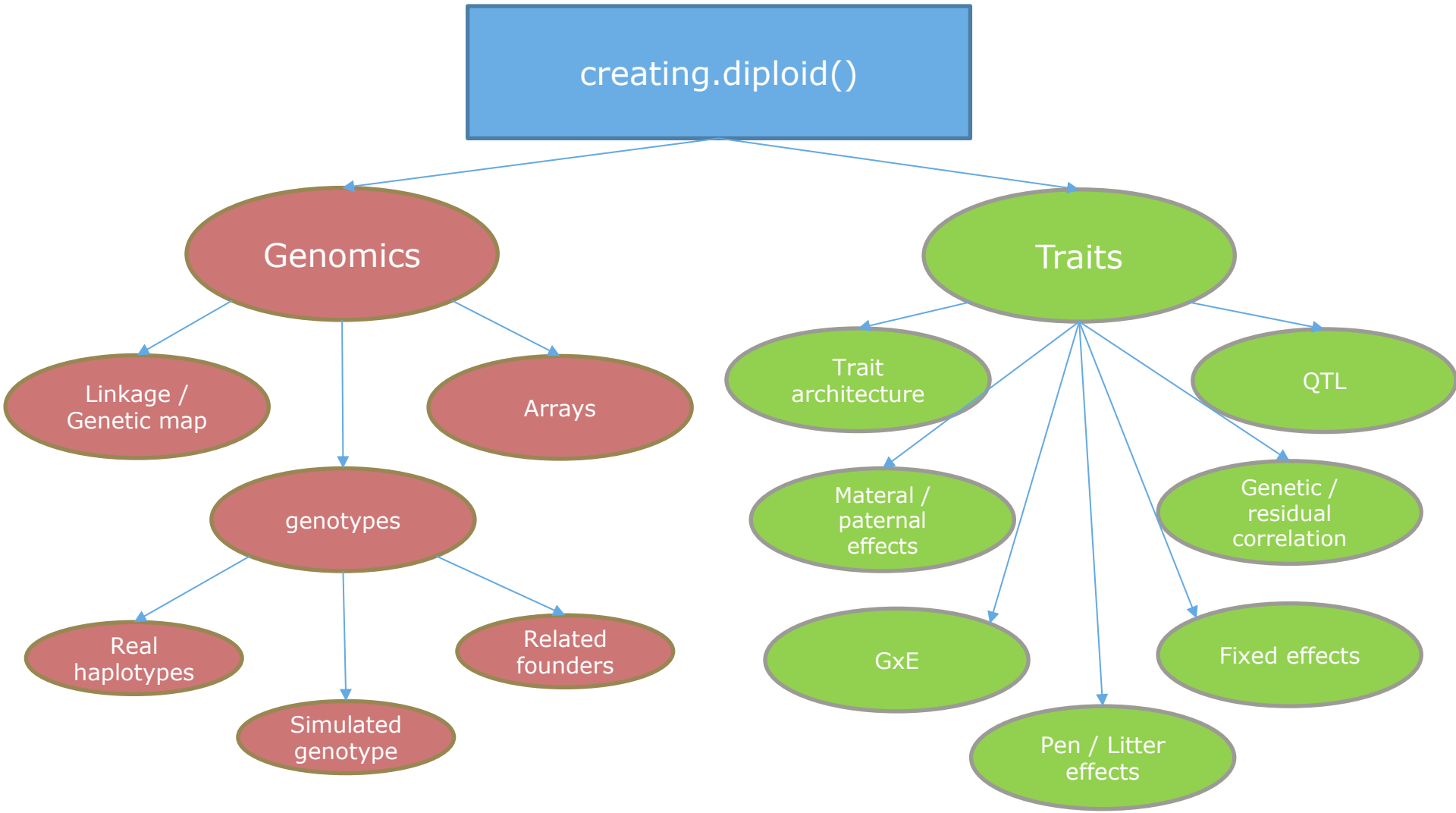
```
> summary(population)
Population size:
Total: 6060 Individuals
Of which 2370 are male and 3690 are female.
There are 22 generations
and 152 unique cohorts.
3780 individuals are copies of previously generated individuals.

Genome Info:
There are 26 unique chromosomes.
In total there are 5000 SNPs.
The genome has a total length of 24.30970424 Morgan.
The genome has a physical size of about: 2.431 GB

Trait Info:
There are 2 modelled traits.
Of which 2 have underlying QTL.
Trait names are:Meat Fertility
Highest correlation between genetics of traits is 0.2 (absolut value).
There are no interactions between residual effects.
Total time spent for generation: 16 seconds.

Time spent per step:
0.8 seconds for creation of founder population.
0.3 seconds for calculation of true genomic values.
0.1 seconds for phenotyping.
0.7 second for breeding value estimation.
0.8 seconds for selection.
13.3 seconds for generation of new individuals.
```

Solution


Genomic value for 1yearEwes
genomic value
100
105
110
115
120
5
10
15
20
generation

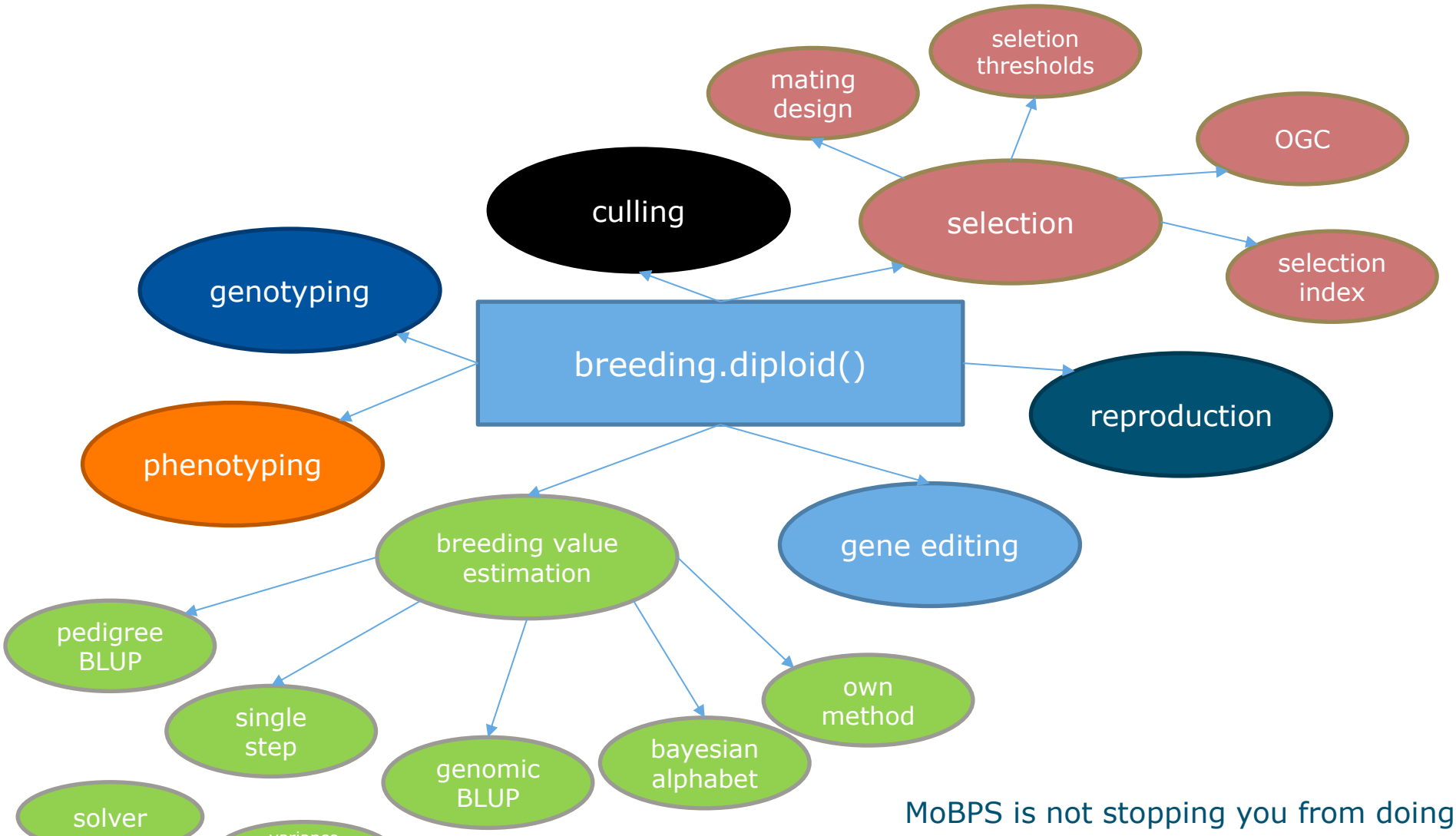
Steps of your simulation

Initialization of the founder population



→ The closer your design is to reality the more reliable your later results will be

Simulation of breeding actions



MoBPS is not stopping you from doing practically impossible things!

Output of a simulation

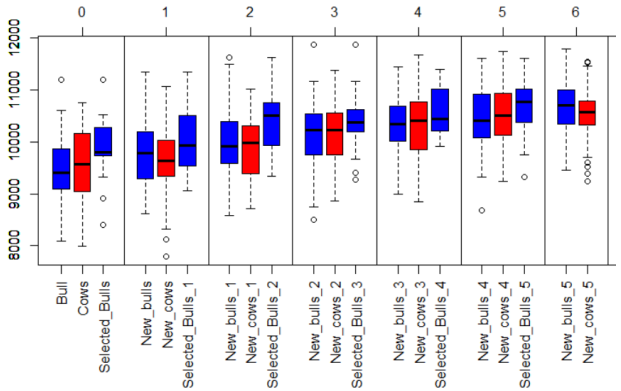
- Highly compressed object (population list)
- This contains information on all individuals ever generated in our simulation and is an expected input for breeding.diploid()
- We know the exact truth!
  - Genotypes / haplotypes
  - Where are the QTLs in the genome
  - What are the underlying genomic values of individuals
  - When/Where were recombination events happen
  - Pedigree without errors

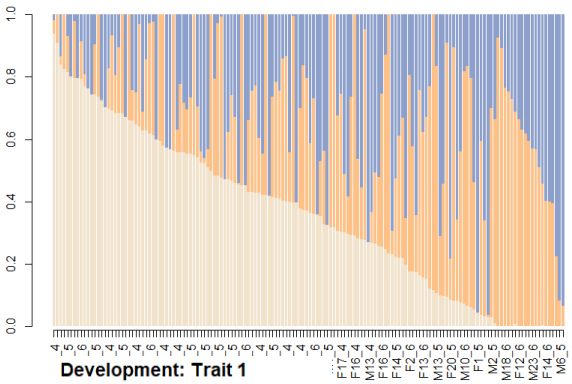
Position of individuals

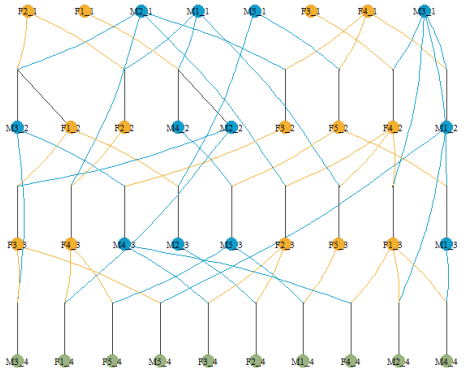
- „gen“: Each new group of individuals is assigned to a generation
- „database“: Specification within a generation
  - database = cbind(5,1,20,30)
  - generation: 5
  - sex: 1 (Male)
  - individual number: 20 to 30
- „cohorts“: The names you assigned them to have

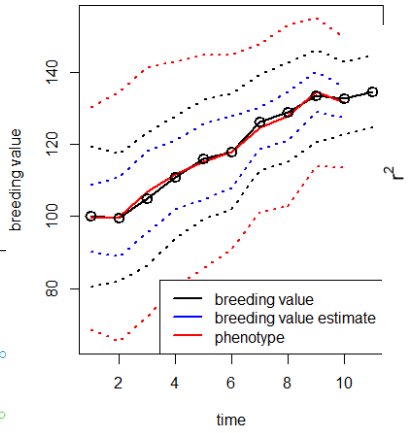
```
Start generation of new individuals (cohort: Offspring).
  |============================================================| 100%Successfully
 generated 50 individuals.

Added _M, _F to cohort names!
Successfully generated cohort: Offspring_M
Database position: 3 (gen), 1 (sex), 1 (first), 20 (last).
Successfully generated cohort: Offspring_F
Database position: 3 (gen), 2 (sex), 1 (first), 30 (last).
Generation of new individuals took 0.103 seconds
```

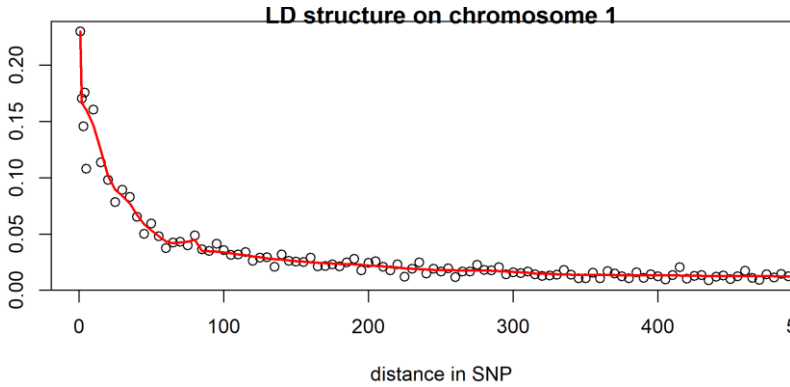
Analysis functior

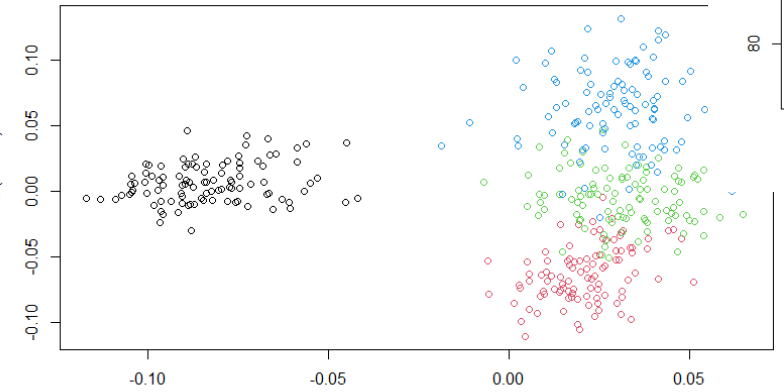




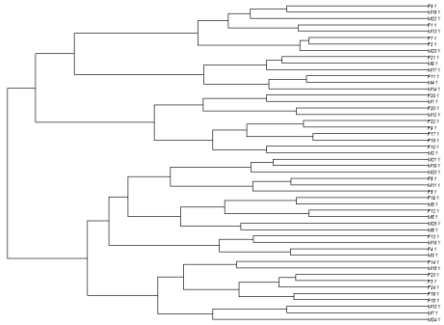


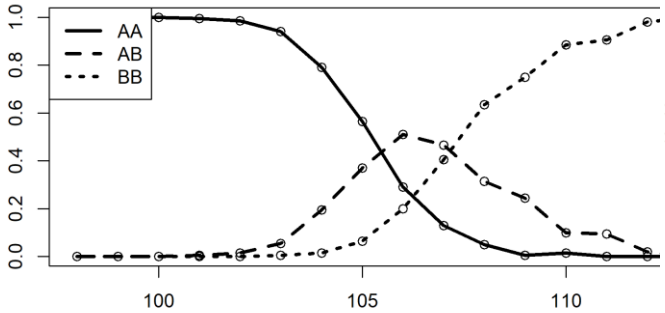


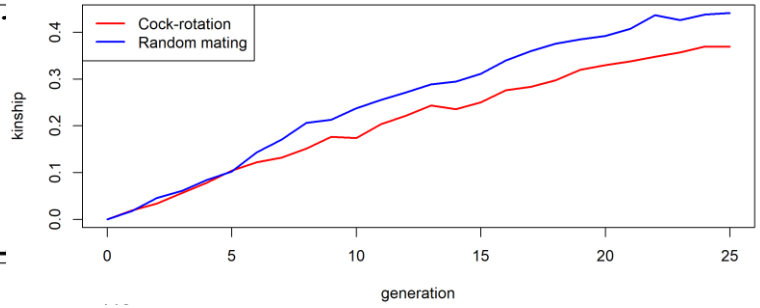




	time point	cohort	n_indi	n_geno	n_pheno
[1,]	"8"	"E-line_piglets_8_M"	"75"	"0"	"0"
[2,]	"9"	"E-line_CT_9"	"21"	"21"	"0"
[3,]	"9"	"E-line_FT_9_M"	"21"	"21"	"0"
[4,]	"13"	"E-line_CT_sel_13"	"1"	"1"	"1"
[5,]	"13"	"E-line_FT_sel_13_M"	"4"	"4"	"4"
[6,]	"16"	"E-line_Elite_16"	"1"	"1"	"1"
[7,]	"17"	"E-line_Elite_17"	"1"	"1"	"1"
[8,]	"18"	"E-line_Elite_18"	"1"	"1"	"1"
[9,]	"19"	"E-line_Elite_19"	"1"	"1"	"1"







Basic example

```
# Generation of a founder population
# with 100 individuals and 10,000 SNPs
# The genome contains 5 chromosomes with a length of 2 Morgan each
# Generation of one trait with 60 underlying QTL
# of which 50 are purely additive and 10 have a dominate effect
# The genomic variance of the trait simulated to be 1
# The generated cohort is named "Founder"

pop <- creating.diploid(nsnp=10000, nindi=100,
                        chr.nr=5, chromosome.length=2,
                        n.additive=50, n.dominant=10,
                        name.cohort="Founder",
                        var.target = 1)

# Generate phenotypic observations for all individuals
# Residual variance is set to result in a heritablity of 0.5
pop <- breeding.diploid(pop, heritability=0.5,
                        phenotyping="all")
```

- First input of breeding.diploid is your current population and the output is an updated object that includes all previous information as well
- Unless you want to store intermediate steps you can overwrite the population

Basic example

```
# Generate 100 offspring
# Use the top 20 male and top 20 female based on their BVE
# All male individuals from the "Founder" cohort are used
# as potential sires.
# All female individuals from the "Founder cohort are used
# as potential dams.
# The resulting cohort in named "Offspring".
pop1 <- breeding.diploid(pop, breeding.size=100,
                         selection.size=c(20,20),
                         selection.criteria = "bve",
                         selection.m.cohorts="Founder_M",
                         selection.f.cohorts="Founder_F",
                         name.cohort="Offspring")

# Same procedure, just with a higher selection intensity
# on the male side.
pop2 <- breeding.diploid(pop, breeding.size=100,
                         selection.size=c(5,20),
                         selection.criteria = "bve",
                         selection.m.cohorts="Founder_M",
                         selection.f.cohorts="Founder_F",
                         name.cohort="Offspring")
```

How to figure out what does what?

- Enter ?breeding.diploid & ?creating.diploid in your R console
- Guidelines_to_MoBPS.pdf
  - This is a 127 page Manual!
  - Do not read it from front to back in one piece
    - Ctrl + F / Search
  - Depending on how you learn / what you need different sections might be more helpful
  - Asking me questions gives me feedback on which sections need updates / are used

Prints & warnings are your friend!

```
> population = breeding.diploid(population, selection.size = c(10,10))
No individuals for selection provided (male side). Use last available.
No individuals for selection provided (female side). Use last available.
Start selection procedure.
Selection male size:
Select 10 individuals out of  2.
Selection female size:
Select 10 individuals out of  3.
Warning messages:
1: In breeding.diploid(population, selection.size = c(10, 10)) :
  Less individuals available for selection than given in selection.size.
 Automatically reduce the number of selected individuals to 2
2: In breeding.diploid(population, selection.size = c(10, 10)) :
  Less individuals available for selection than given in selection.size.
 Automatically reduce the number of selected individuals to 3
```

- Just as a warning when going through the sample solutions:
  - The sample solutions files .R contain more than just the results to the Tasks at hand but also include inserts arising from questions etc.

Task 5: Switching to R

- Perform the simulation from „Simple Sheep“ directly in R

Solution Task 5

```
> get.cohorts(population, extended = TRUE)[,1:4]
                name             generation male individuals female individuals
1yearRams       "1yearRams"      "1"        "50"              "0"
2yearRams       "2yearRams"      "1"        "10"              "0"
1yearEwes       "1yearEwes"      "1"        "0"               "50"
2yearEwes       "2yearEwes"      "1"        "0"               "40"
3yearEwes       "3yearEwes"      "1"        "0"               "30"
1yearRamsNext   "1yearRamsNext"  "2"        "50"              "0"
1yearEwesNext   "1yearEwesNext"  "2"        "0"               "50"
2yearRamsNext   "2yearRamsNext"  "2"        "10"              "0"
2yearEwesNext   "2yearEwesNext"  "2"        "0"               "40"
3yearEwesNext   "3yearEwesNext"  "2"        "0"               "30"
```

```
> summary(population)
Population size:
Total: 360 Individuals
Of which 120 are male and 240 are female.
There are 2 generations
and 10 unique cohorts.
80 individuals are copies of previously generated individuals.

Genome Info:
There are 5 unique chromosomes.
In total there are 5000 SNPs.
The genome has a total length of 5 Morgan.
The genome has a physical size of about: 0.4998 GB

Trait Info:
There is 1 modelled trait.
The trait has underlying QTL
The trait is named: Meat
Total time spent for generation: 1.4 seconds.

Time spent per step:
0.6 seconds for creation of founder population.
0.1 seconds for calculation of true genomic values.
0.1 seconds for phenotyping.
0.6 seconds for generation of new individuals.
```

Extension Task 5 (snapshots and genotyping)

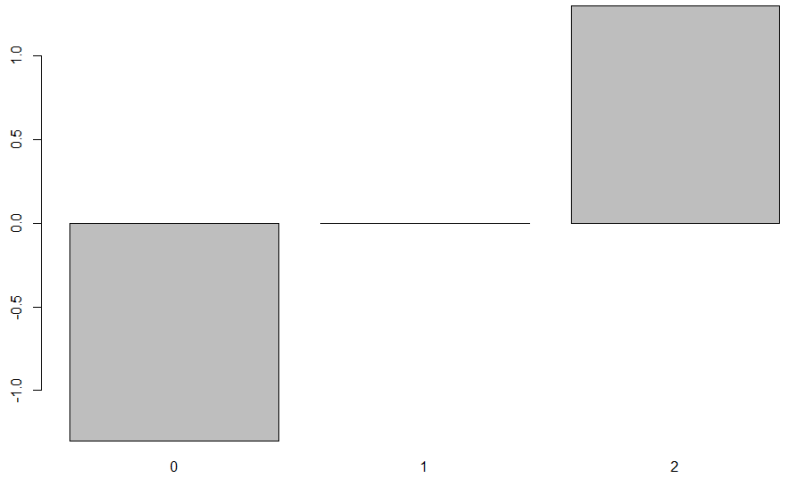
```
> get.geno(population, gen=3, non.genotyped.as.missing = TRUE)
           M1_3 M2_3 M3_3 M4_3 M5_3 M6_3 M7_3 M8_3 M9_3 M10_3
Chr1SNP2      2   NA   NA   NA   NA    2   NA   NA   NA     2
Chr1SNP4      0   NA   NA   NA   NA    1   NA   NA   NA     1
Chr1SNP6      0   NA   NA   NA   NA    1   NA   NA   NA     0
Chr1SNP8      2   NA   NA   NA   NA    0   NA   NA   NA     1
Chr1SNP10     2   NA   NA   NA   NA    0   NA   NA   NA     1
Chr1SNP12     0   NA   NA   NA   NA    0   NA   NA   NA     0
Chr1SNP14     2   NA   NA   NA   NA    1   NA   NA   NA     1
```

```
> # Overview on time points whenever an individual is switching cohorts:
> get.snapshot(population, cohorts = "1yearRams")
     time point cohort          n_indi n_geno n_pheno
[1,] "0"        "1yearRams"     "50"   "0"    "50"
[2,] "0"        "2yearRamsNext" "10"   "0"    "10"
> # Overview on time points whenever the status of some individuals changes:
> get.snapshot.single(population, cohorts = "1yearRams")
Generate analysis starting from cohort: 1yearRams
     time point cohort      n_indi n_geno n_pheno
[1,] "0"        "1yearRams" "50"   "0"    "50"
[2,] "1"        "1yearRams" "50"   "10"   "50"
```

On the generation of traits

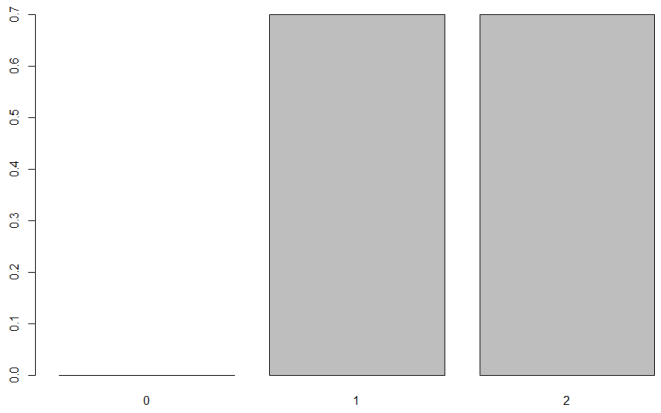
On the generation of traits

- Predefined architectures:
  - Purely additive QTLs with effect size drawn from $N(0,1)$ (n.additive)
  - Purely additive QTLs with equal effect size (n.equal.additive)



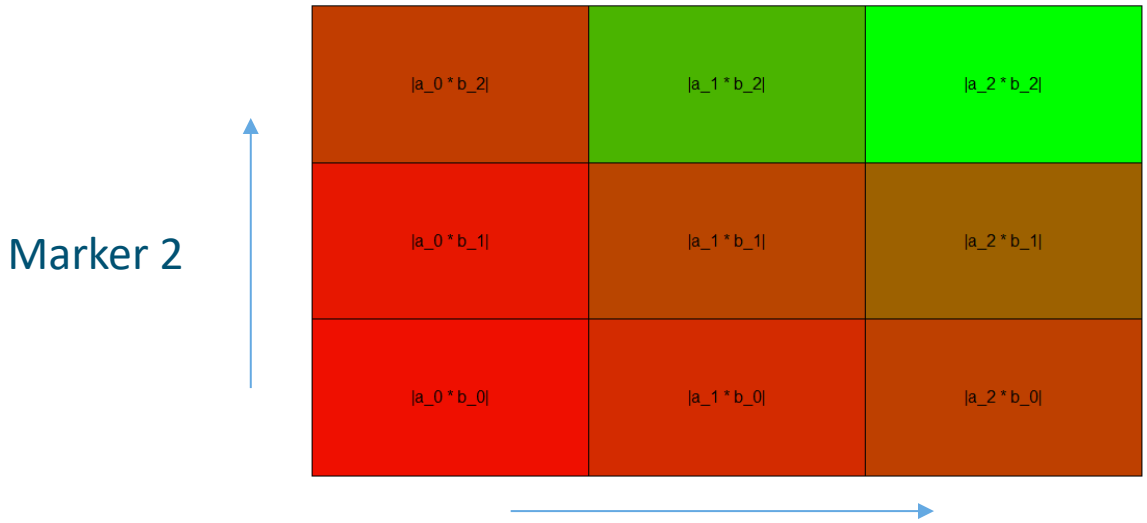
On the generation of traits

- Predefined architectures:
  - Dominant QTLs with effect size drawn from $N(0,1)$ (n.dominant)
  - default: a = d → dominant.only.positive = TRUE: a = |d|
  - Dominant QTLs with equal effect size (n.equal.dominant)
- QTLs are part of your pool of SNPs



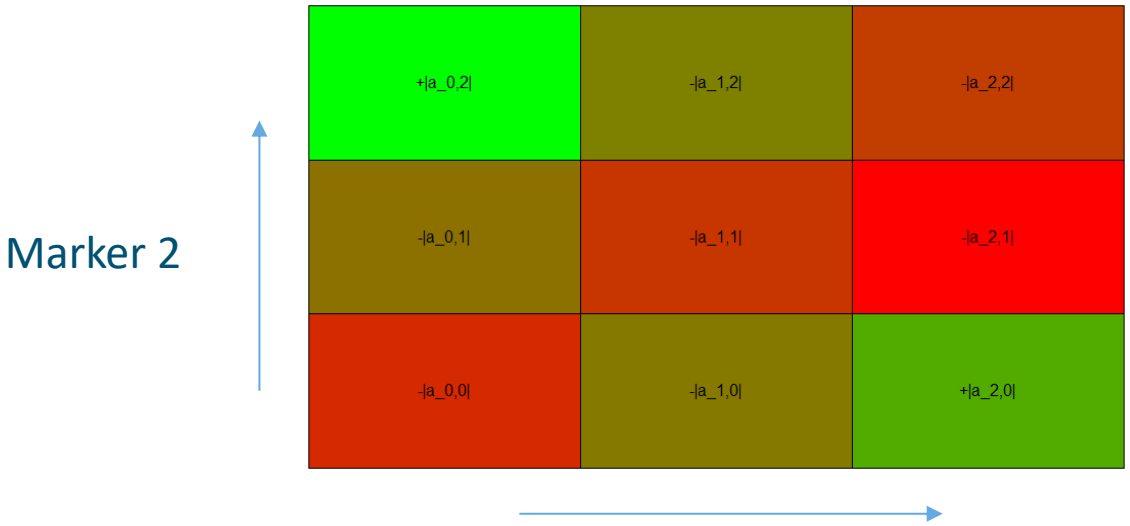
On the generation of traits

- Quantitative epistatic effects (n.quantitative)
  - Generate three $N(0,1)$ random variables for each of the two markers and sort them according to absolute size $(a_0, a_1, a_2)\ and\ (b_0, b_1, b_2)$.
  - Effect of (X,Y) is $|a_X \cdot b_Y|$



On the generation of traits

- Qualitative epistatic effects (n.qualitative)
  - Different effects for each marker combination ($N(0,1)$)
  - Effects for (0,2) and (2,0) are positive, rest negative



What if this is not enough?

- Manual general of QTLs via:
  - Single marker QTL: real.bv.add

	SNP	Chromo	Effect AA	Effect AB	Effect BB
[1,]	1	1	0.4294301864	0	-0.4294301864
[2,]	2	1	-0.0756397721	0	0.0756397721
[3,]	8	1	0.2381887064	0	-0.2381887064

- There is a QTL on marker 8 on chromosome 1
  - Homozygosity in the first allele has an effect of -0.238...
  - Heterozygosity has an effect of 0
  - Homozygosity in the second allele has an effect of 0.238...

More advanced features

	SNP	Chromo	Effect AA	Effect AB	Effect BB	Overall position	Pool	Pool-based Effect
[1,]	1	1	0.4294301864	0	-0.4294301864	1	0	0
[2,]	2	1	-0.0756397721	0	0.0756397721	2	0	0
[3,]	8	1	0.2381887064	0	-0.2381887064	8	0	0

- Traits can only be evaluated based on:
  - Maternal / paternal genotype
  - Different effects for alleles from different founder pools
  - Not based on realized allele but originating founder pool

Epistatic effects

```
> real.bv.mult
     First SNP First chromosome Second SNP Second chromosome effect 00 effect 01 effect 02 effect 10
[1,]       144                1        145                 1      1.00      0.00      0.00      0.00
[2,]         6                3        188                 5      0.37      0.16      1.33      1.49
[3,]         5               17          1                10      1.18      2.60      0.18      1.74
     effect 11 effect 12 effect 20 effect 21 effect 22
[1,]      0.00      0.00      0.00      0.00      0.00
[2,]      1.58      2.51      0.38      2.12      0.98
[3,]      0.69      1.39     -1.21      0.96      1.94
```

```
> real.bv.dice
$location
$location[[1]]
     SNP chromosome
[1,]  11          1
[2,]  12          1
[3,]  16          4

$location[[2]]
     SNP chromosome
[1,]  14          2
[2,]  77          6
[3,]  15          9


$effects
$effects[[1]]
 [1]  1.8212212  1.5939013  1.9189774  1.7821363  1.0745650 -0.9893517  1.6198257  0.9438713  0.8442045 -0.4707524  0.5218499  1.4179416  2.3586796
[14]  0.8972123  1.3876716  0.9461950 -0.3770596  0.5850054  0.6057100  0.9406866  2.1000254  1.7631757  0.8354764  0.7466383  1.6969634  1.5566632
[27]  0.3112443

$effects[[2]]
 [1]  0.29250484  1.36458196  1.76853292  0.88765379  1.88110773  1.39810588  0.38797361  1.34111969 -0.12936310  2.43302370  2.98039990  0.63277852
[13] -0.04413463  1.56971963  0.86494540  3.40161776  0.96076000  1.68973936  1.02800216  0.25672679  1.18879230 -0.80495863  2.46555486  1.15325334
[25]  3.17261167  1.47550953  0.29005357
```

- In case SNP/chromosome positions set to NA, they are automatically filled with reasonable values
- If effects are placed on SNPs that are not there, effects will be removed

Simulation of correlated traits

- Correlated traits are a combination of original traits
- Cholesky decomposition to calculate weights

Target correlation: $\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$

Cholesky decomposition: $\begin{pmatrix} 1 & 0.2 \\ 0 & 0.9797 \end{pmatrix}$

- Trait 1 QTL effects: old trait 1 / trait1_sd
- Trait 2 QTL effects:
  old trait 1* 0.2 / trait1_sd
  + old trait 2 * 0.9797 / trait2_sd
- Number of QTLs will be higher for some traits (or QTL positions shared: shared qtl.position.shared = TRUE)

Definition of heritability

- In line with definitions from animal breeding
  - Heritability for a single observation / plot
  - Higher number of observations / plots will reduce residual variance
  - The residual variance is split into a permanent effect made for all observations and a temporary part for a single observation

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

$$\sigma_e^2 = \sigma_{PU}^2 + \sigma_{TU}^2$$

$$w^2 = \frac{\sigma_a^2 + \sigma_{PU}^2}{\sigma_a^2 + \sigma_{PU}^2 + \sigma_{TU}^2}$$

- The repeatability gives information on the share of the permanent effect on the total residual effect

Use of genomic data

- For all individuals in the population, underlying haplotypes are stored
- This is also the case when they are not used directly
- You can control if individuals are genotyped / partially genotyped (low-density array) and the tool will automatically take care of the use in a breeding value estimation
- You can also just manually selected which markers to include in the breeding value estimation directly

Founder genotypes

- Default: randomly generated
  - Very fast but no linkage structure
- Simulated LD build-up
  - External tools (MaCS, Chen et al. 2009)
  - founder.simulation() within MoBPS
  - Manual simulation of the population history in MoBPS
  - Burn-in cycles (e.g. start analysis in cycle 10)
- Import of real data
  - Haplotype matrix (nSNP x nHaplotypes – matrix) + map in creating.diploid() in dataset / map parameter
  - VCF / PedMap format

Make sure data is phased! (e.g. BEAGLE, Browning et al. 2018)

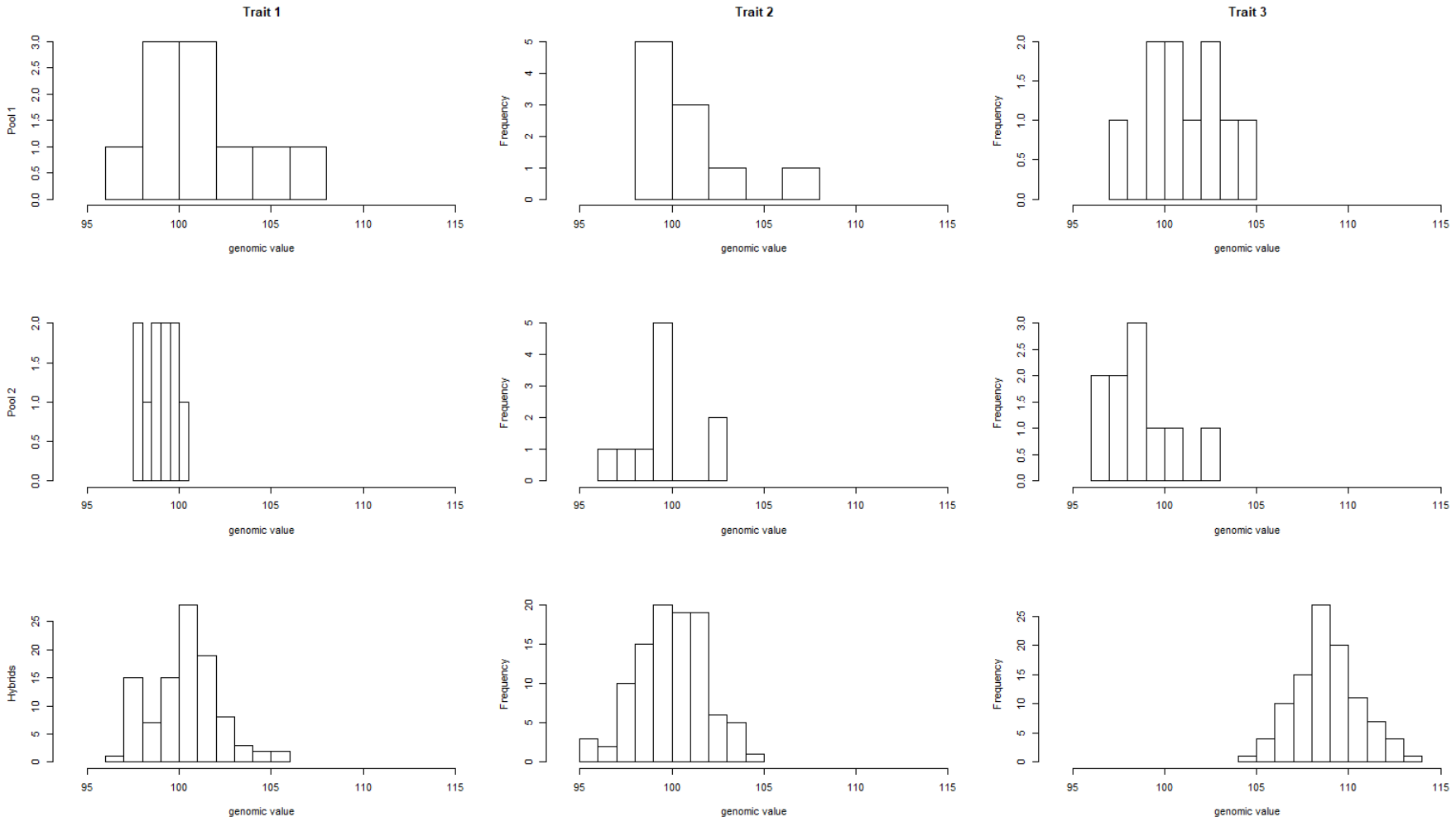
Task 6: Import of real data and generation of traits

- Generate a founder population with 10 individuals from two different gene pools
- Genotypic data for the individuals is given via Pool1.vcf and Pool2.vcf
- Add three traits:
  - One with 10 purely additive QTLs
  - One with 1000 purely additive QTLs
  - One with 500 purely additive QTLs and 500 dominant QTLs (with positive effect)
  - Traits should be uncorrelated
  - For all traits phenotypic mean for the founders should be 100 with a genetic variance of 5
- Generate 100 offspring by random mating between individuals from the two gene pools
- Compare the genomic values of the parents and offspring
- How often was each individual used for reproduction?
- Generate a PCA for all simulated individuals

Solution Task 6

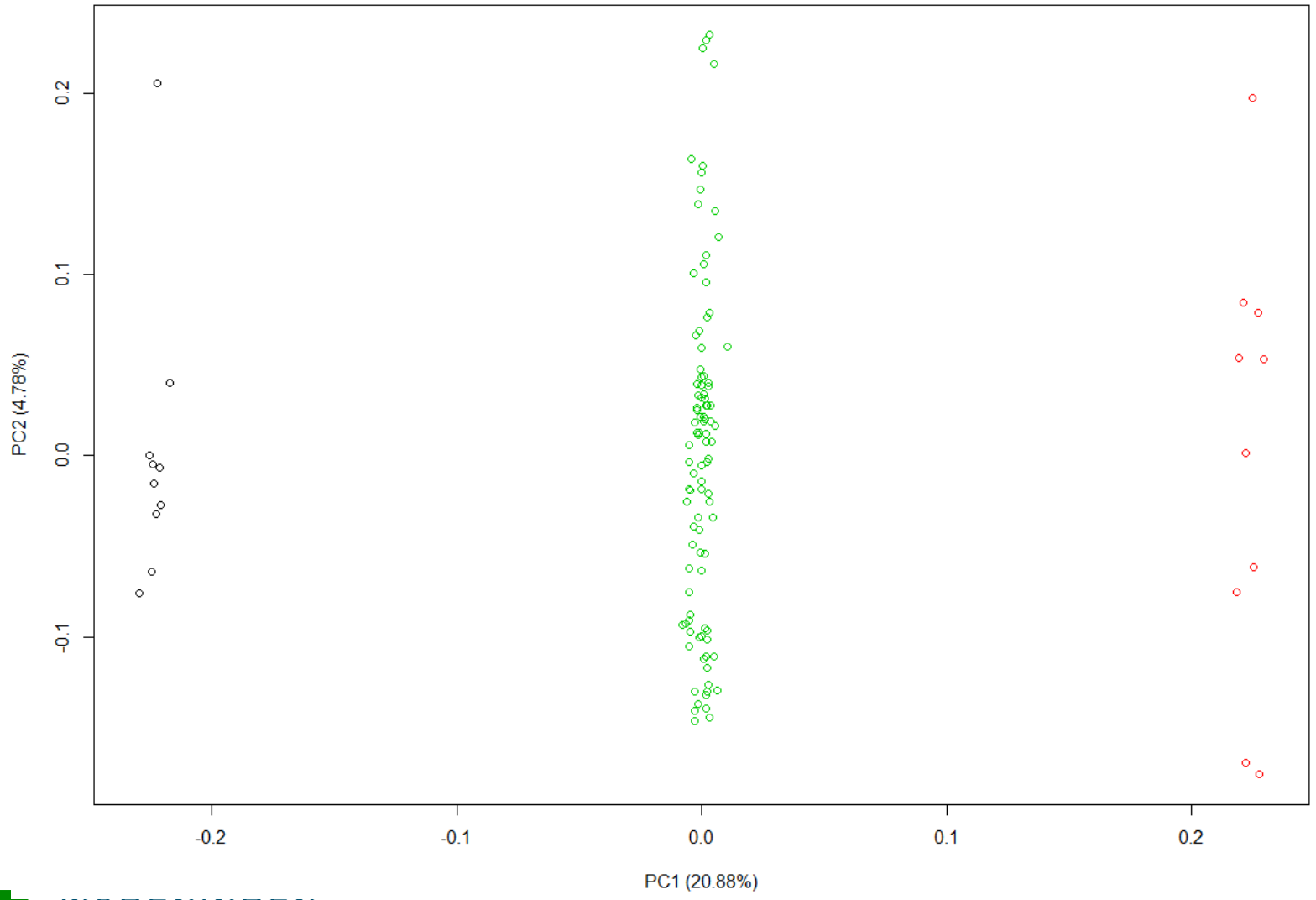
```
> summary(population)
Population size:
Total: 120 Individuals
Of which 110 are male and 10 are female.
There are 2 generations
and 3 unique cohorts.

Genome Info:
There are 5 unique chromosomes.
In total there are 5000 SNPs.
The genome has a total length of 25 Morgan.
The genome has a physical size of about: 0.5 GB

Trait Info:
There are 3 modelled traits.
Of which 3 have underlying QTL.
Trait names are:Trait 1 Trait 2 Trait 3
Genetics of traits are uncorrelated.
There are no interactions between residual effects.
Total time spent for generation: 0.8 seconds.

Time spent per step:
0.4 seconds for creation of founder population.
0.4 seconds for generation of new individuals.
```

Solution Task 6



WAGENINGEN
UNIVERSITY & RESEARCH

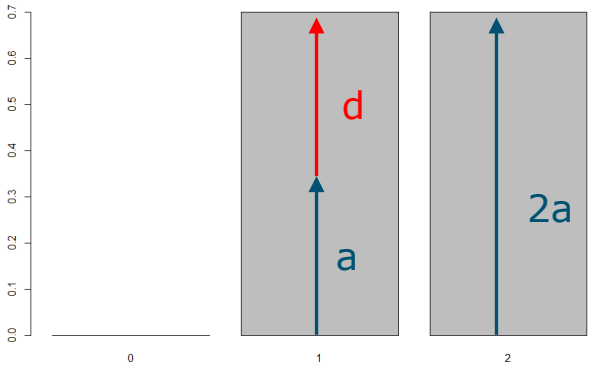
Solution Task 6



WAGENINGEN
UNIVERSITY & RESEARCH

Extension Task 6 (genetic variance)

```
> population = breeding.diploid(population, heritability = rep(0.3,3), phenotyping.cohorts = "Crossbred")
Start deriving enviromental variance (according to given heritability).
Estimated residual variances: 2.898 11.406 3.9537
Start simulating phenotypes.
> get.variance(population, cohorts = "Crossbred")
        h2_narrow  h2_broad sigma_g2  sigma_e2 sigma_a2  sigma_d2
Trait 1 0.3689922 0.3689922 1.242016  2.123952 1.242016 0.0000000
Trait 2 0.2772849 0.2772849 4.888271 12.740788 4.888271 0.0000000
Trait 3 0.2424259 0.2573124 1.694463  4.890772 1.565061 0.2645772
```



- A dominant QTL is partly additive!
- Share of dominance will be dependent on genotypes etc.
- There is no MoBPS parameter to set a given share of dominance variance

Extension Task 6 (limiting offspring per parent)

```
# To make sure all individuals are used at most 10 times ((max.offspring = 10))
population <- breeding.diploid(population, breeding.size = c(100,0),
                               selection.m.cohorts = "Pool1", selection.f.cohorts = "Pool2",
                               name.cohort = "Crossbred2", max.offspring = 10)
```

```
# To make sure each mating combination is done once (( breeding.all.combination = TRUE))

population <- breeding.diploid(population, breeding.size = c(100,0),
                               selection.m.cohorts = "Pool1", selection.f.cohorts = "Pool2",
                               name.cohort = "Crossbred3", breeding.all.combination = TRUE)
```

```
# To manually select mating pairs fixed.breeding

# Generate 3 new individuals:
# First: mate individual from generation 1, sex 1, nr 1 with individual from generation 1, sex 2, nr 1
# Second: mate individual from generation 1, sex 1, nr 1 with individual from generation 1, sex 2, nr 2
# Third: mate individual from generation 1, sex 1, nr 4 with individual from generation 1, sex 2, nr 6
fixed <- matrix(c(1,1,1,1,2,1,
                  1,1,1,1,2,2,
                  1,1,4,1,2,6), byrow=TRUE, ncol=6)

population <- breeding.diploid(population, fixed.breeding = fixed)
```

Handling of phenotypes, breeding values and genomic values

- For each individuals and each trait, we are storing:
  - The observed phenotype („pheno“)
  - The estimated breeding value („bve“ / “ebv”)
  - The underlying true genomic value („bv“ / “gv”)
- Selection can be based on these criteria
- The underlying true genomic value can usually not be observed in practice
  - Powerful tool to evaluate methods!

Simulation of fixed effects

```
4 population = creating.trait(population, n.additive = 100, fixed.effects = cbind(10,5),
5                             var.target = 100, mean.target = 100,
6                             trait.name = "Trait")
7
```

- Two fixed effects with effect size 10 and 5

```
18 population = breeding.diploid(population, phenotyping.database = cbind(12,2), genotyped.gen = 12,
19                               heritability = 0.3,
20                               fixed.effects.p = cbind(c(0,0,1,1),c(0,1,0,1)),
21                               fixed.effects.freq = c(0.1,0.2, 0.3, 0.4))
22
```

Fixed effect 1	Fixed effect 2	Probability
0	0	0.1
0	1	0.2
1	0	0.3
1	1	0.4

```
> population = breeding.diploid(population, bve = TRUE, bve.gen = 12, rrblup.bve = TRUE)
Start genomic BVE.
Use Genomic BLUP (All individuals are genotyped)!
500 phenotyped individuals in BVE (Trait: Trait).
1000 individuals considered in BVE.
Variance components in BVE: sigma_g^2 = 83.0279; sigma_e^2 = 251.4819 ; h^2 = 0.248
Estimated fixed effects:
[1] 100.437756   6.284264   3.985076
1.36 seconds for BVE.
Correlation between genetic values and BVE:
0.626921
```

- My blupf90 implementation does **not** include fixed effects yet!

Recap Day 1

Agenda

- General introduction of the MoBPS framework
- Basic functionality of the web-interface
- More advanced features of the web-interface
- Scenario comparison in the web-interface
- Basic functionality of the R-package
- Trait generation in the R-package
- Import of real genotype data (vcf) in the R-package

Agenda

- Breeding value estimation in the R-package (Task 7)
- Breeding value estimation with commercial software (MiXBLUP, blupf90)
- Computational efficiency within R
- ~~Inbreeding management in the R-package (Task 8)~~
- Offspring phenotypes / hybrid yield trial in the R-package (Task 8b)
- Spotting errors in code (Task 13)

Agenda

- ~~Addressing specific individuals in the R-package (Task 9)~~
- ~~Discrete phenotypes & litter sizes~~
- Simulating multiple scenarios & evaluating different scenarios (Task 10 – 12)
    - Primary as a frontal-presentation
- Remaining open questions

Task 7: Breeding value estimation

- This task is split into two subparts – if you are struggling with the first part you can load in the .Rdata object with an already generated population
- Generate a population list with 12 generations:
    - Each generation contains 50 males, 50 female with parents of the previous generation
    - Use a genetic map with 25.000 SNPs, 5 chromosomes with a length of 3 Morgan each
    - Generate a trait with heritability of 0.3 and 1‘000 purely additive QTLs
    - Make sure that:
        - In generation 10 only males are phenotyped
        - In generation 11 & 12 all individuals are phenotyped
        - In generation 10 & 11 all individuals are genotyped
        - In generation 12 only 20% of all individuals are genotyped
- Perform a breeding value estimations for individuals in generation 10, 11, 12 or combinations of the cohorts

  Use:
    - Genomic breeding value estimation (GBLUP)
    - Pedigree-based breeding value estimation (PBLUP)
    - Single-step breeding value estimation (ssGBLUP)
    - Assume individuals are only genotyped for 10‘000 / 2‘000 / 100 randomly selected markers
    - Generate a plot to showcase real genomic values and breeding values for generation 10.

Solution Task 7

- Having a look at prints / logs can be very helpful to check if the tool is doing what you expect it to do

```
> population <- breeding.diploid(population, bve=TRUE, bve.gen=12)
Start genomic BVE.
Use Single Step GBLUP
Start derive Single-step relationship matrix
Construct pedigree matrix for 100 individuals.
Derive pedigree-matrix based for  100  individuals based on  793  individuals.
Derived pedigree matrix in  0.09 seconds.
Start deriving of H matrix for 16 genotyped and 84 non-genotyped individuals.
Derived H matrix in  0 seconds.
100 phenotyped individuals in BVE (Trait: Trait 1).
100 individuals considered in BVE.
Variance components in BVE: sigma_g^2 = 456.1813; sigma_e^2 = 738.2976 ; h^2 = 0.382
chol (own), 5 cores
0 seconds for BVE.
Correlation between genetic values and BVE:
0.6376494
```

Solution Task 7

```
> population <- breeding.diploid(population, bve=TRUE, bve.gen=11)
Start genomic BVE.
100 phenotyped individuals in BVE (Trait: Trait 1).
100 individuals considered in BVE.
Variance components in BVE: sigma_g^2 = 337.1231; sigma_e^2 = 1111.8964 ; h^2 = 0.233
0 seconds for BVE.
Correlation between genetic values and BVE:
0.6032743
> # BVE with various different arrays
> population <- breeding.diploid(population, bve=TRUE, bve.gen=11, bve.array = 2)
Start genomic BVE.
10000 markers survived filtering for BVE.
100 phenotyped individuals in BVE (Trait: Trait 1).
100 individuals considered in BVE.
Variance components in BVE: sigma_g^2 = 337.1231; sigma_e^2 = 1111.8964 ; h^2 = 0.233
0 seconds for BVE.
Correlation between genetic values and BVE:
0.5975752
> population <- breeding.diploid(population, bve=TRUE, bve.gen=11, bve.array = 3)
Start genomic BVE.
2000 markers survived filtering for BVE.
100 phenotyped individuals in BVE (Trait: Trait 1).
100 individuals considered in BVE.
Variance components in BVE: sigma_g^2 = 337.1231; sigma_e^2 = 1111.8964 ; h^2 = 0.233
0 seconds for BVE.
Correlation between genetic values and BVE:
0.6137002
> population <- breeding.diploid(population, bve=TRUE, bve.gen=11, bve.array = 4)
Start genomic BVE.
100 markers survived filtering for BVE.
100 phenotyped individuals in BVE (Trait: Trait 1).
100 individuals considered in BVE.
Variance components in BVE: sigma_g^2 = 337.1231; sigma_e^2 = 1111.8964 ; h^2 = 0.233
0 seconds for BVE.
Correlation between genetic values and BVE:
0.4980535
```

25k SNPs

10k SNPs

2k SNPs

0.1k SNPs

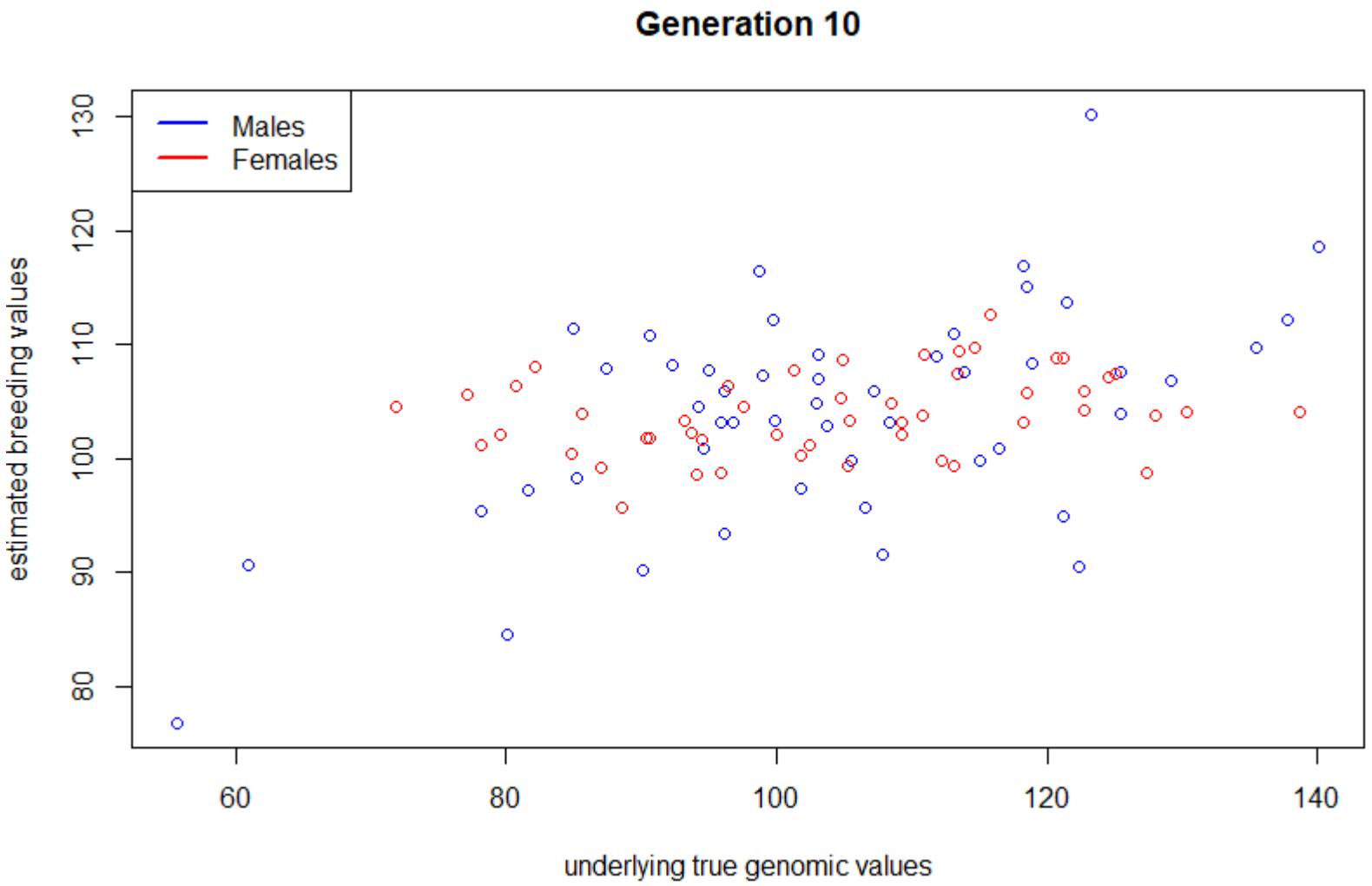
With 100 markers prediction accuracies are going down substantially

Solution Task 7

- Prediction accuracies for males are much higher than for females
- Only males are phenotyped in generation 10!

```
> analyze.bv(population, database = cbind(10,1))
[[1]]
              Trait 1
BV / BVE    0.5439709
BV / Pheno  0.5372799
BVE / Pheno 0.9727460

[[2]]
 Trait 1
316.4133

> analyze.bv(population, database = cbind(10,2))
[[1]]
              Trait 1
BV / BVE    0.2907731
BV / Pheno         NA
BVE / Pheno        NA

[[2]]
 Trait 1
255.9466
```

Solution Task 7



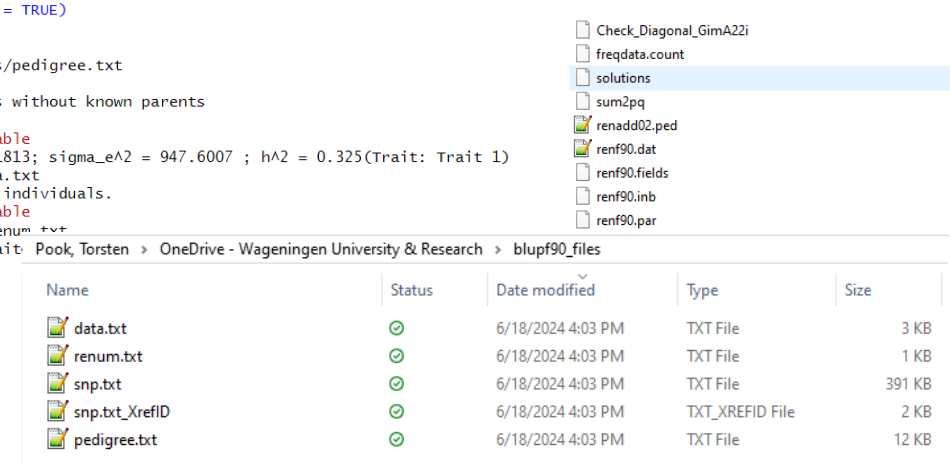
Solution Task 7

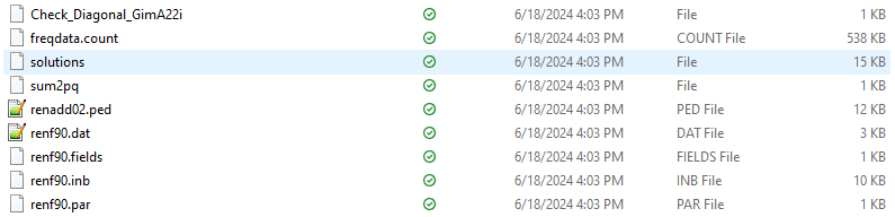
- MoBPS „only“ provides standard approaches for selection & breeding value estimation
- Direct link to commercial software MiXBLUP & blupf90
  - Internally write blupf90/renumf90-input files
  - Still relatively basic / reflective on what has been used yet
  - Contact me if you feel something is missing for your purposes (I am not a blupf90-expert)
- Use of own methodology is also possible
- MoBPS takes care of phenotype simulation, meiosis, computational efficiency etc.

Extension Task 7 (blupf90 / rrblup / MiXBLUP)

- Base-R, MiXBLUP, blupf90 implementation will assume heritability and trait-correlations to be known
- rrblup & sommer will estimate variance components (slow!!!)
- Base-R & rrblup will apply multiple single-trait models
- Base-R & rrblup handle pedigree and genomic BLUP in the same way ((PBLUP will be much slower than it could

```
> # make use software is available in your working directory ((or provide a path to the software))
> population <- breeding.diploid(population, bve=TRUE, bve.gen=12,
+                                blupf90.bve = TRUE)
Start genomic BVE.
Use Single Step GBLUP
Start writting pedigree file at blupf90_files/pedigree.txt
Start collecting pedigree
Pedigree contains 881 animals with 88 animals without known parents
Start writting: blupf90_files/pedigree.txt
x being coerced from class: matrix to data.table
Variance components in BVE: sigma_g^2 = 456.1813; sigma_e^2 = 947.6007 ; h^2 = 0.325(Trait: Trait 1)
Start writting datafile at blupf90_files/data.txt
Start writting blupf90 genotype file with 16 individuals.
x being coerced from class: matrix to data.table
Start writting input file at blupf90_files/renum.txt
100 phenotyped individuals in BVE (Trait: Trait 1)
100 individuals considered in BVE.
Remove previous solutions file!
Remove previous renf file!
Remove previous renadd file!
Start RENUM
Finished RENUM BVE
Estimate took 0.47 seconds.
Start BLUPF90
Finished BLUPF90 BVE
Estimate took 2.26 seconds.
Correlation between genetic values and BVE:
0.7362108
```





Automatic generation of renum/blupf90 input files in working directory / blupf90_files

Computational efficiency of MoBPS

On the computational efficiency of MoBPS

- Smallest Unit of saving things in R: Integer
- 4 Bytes (32 bits)
  - Can contain any integer number from -2^31 to 2^31
  - Adding 1 + 1:

```
        0000000000 0000000000 0000000000 01
+       0000000000 0000000000 0000000000 01
-------------------------------------------
=       0000000000 0000000000 0000000000 10
=                                         2
```

miraculix

- Genotypes only take values 0, 1, 2
- 30 of 32 bits are wasted
- This is implemented in the R-package miraculix
  - ~ 15 times less memory than use of integers
  - Matrix operations are performed bit-wise
    - Between 14 – 34 times faster than regular R
    - internal screening if things like avx2, sse4 or a graphics card are available
  - This is more effective than the previous MiXBLUP implementations and has now been adapted there
- Genotypes of non-founders are not stored
  - On-the-fly calculation based on points of recombination

BLAS / LAPACK

- Basic Linear Algebra Subprograms are internally used operations from linear algebra (e.g. matrix multiplications on non- 0/1/2 data)
- To check if your system is good, running the following R code:

```
n <- 5000
T <- matrix(0, nrow=n, ncol=n)
A <- T %*% T
```

- This matrix multiplication on a very good system takes ~ 2 seconds
- I am using OpenBlas // Intel MKL from a conda container
- Check if your HPC supports blas – e.g. I start R this way: LD_PRELOAD=/shared/apps/openblas/gcc-6.1.0/sandy-bridge/0.3.3/lib/libopenblas.so R --vanilla

Planning of computing time and memory

- Identify computational heavy steps of your simulation:
  - Run simulation with lower individual numbers (size.scaling)
  - How much time does it take to generate one individual
  - Breeding value estimation using GBLUP (in R)
    - Inversion of a matrix: o(n^3)
    - Calculation of the G matrix: o(p*n^2)
    - The G matrix for 11.000 individuals requires about 1 GB of memory (scales quadratically)
  - get.comp.times() // sacct
  - There is usually tricks to reduce computing times

Simplifications of reality

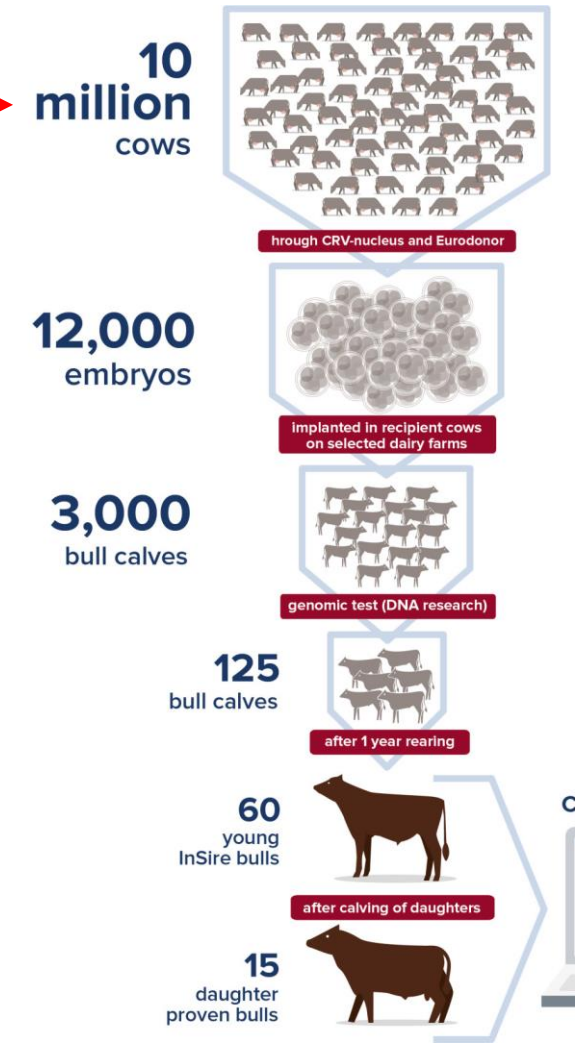
"All models are wrong, but some are useful." (George E.P. Box)

- Dairy cattle simulations:
  - We want to simulate multiple years of breeding
  - Different scenarios
  - Each scenario has to be simulated multiple times

  → Extremely high computational burden

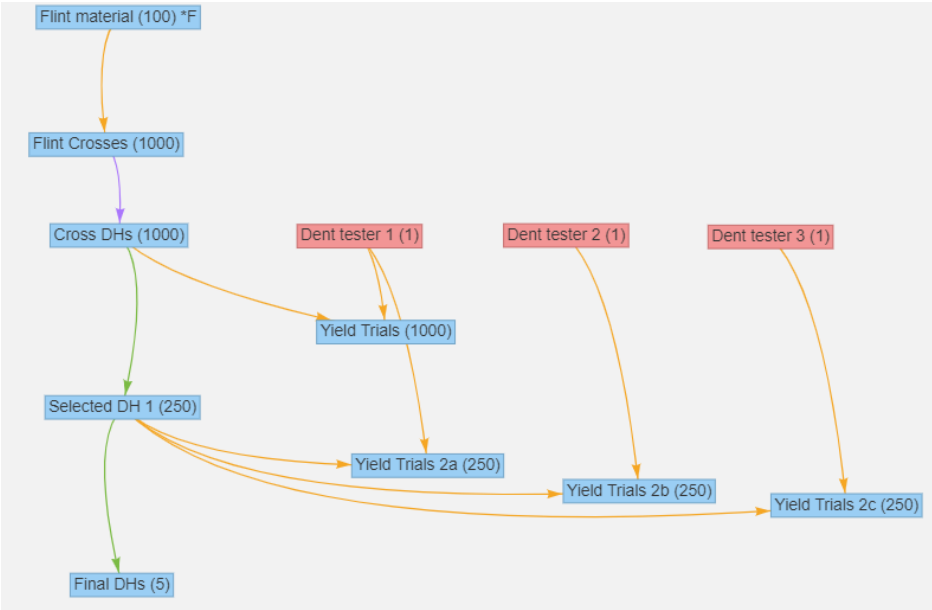
Simplifications of reality

- Highest computational burden:
  - Breeding value estimation with millions of animals
- Reduce number of EuroDonor animals but be aware:
  - Genetic diversity might go down stronger
  - Accuracy of breeding values might be lower
  - Best EuroDonors are relatively worse to DeltaNucleus
- Separate prediction models for different trait complexes



Task 8: Offspring phenotypes / yield trials

- Simulate the following breeding scheme:
- Founders:
    - Simulate a founder population with 100 lines from one pool (flint lines) and 3 lines from a second pool (dent lines)
    - Make sure that allele frequencies in the two pools are different
    - Simulate a single trait with 1‘000 QTLs
    - Use 10‘000 SNPs. You can use a subset of the maize 600k array from MoBPSmaps
- Generate 1000 crosses within the Flint gene pool
- Generate 1000 DH lines from the crosses
- Mate the DHs to one of the three dent lines
- Phenotype the offspring (h^2 = 0.3)
- Select the top 250 DHs lines based on the performance in the yield trial
    - Hint: make sure that each DH is cross to the tester exactly once!
- Mate the selected DHs to all three dent lines
- Phenotype the offspring (h^2 = 0.3)
- Select the top 5 DH lines based on the performance in the yield trail
    - Hint: You can use breeding.all.combination or fixed.breeding (for a challenge) to generate your second yield trial

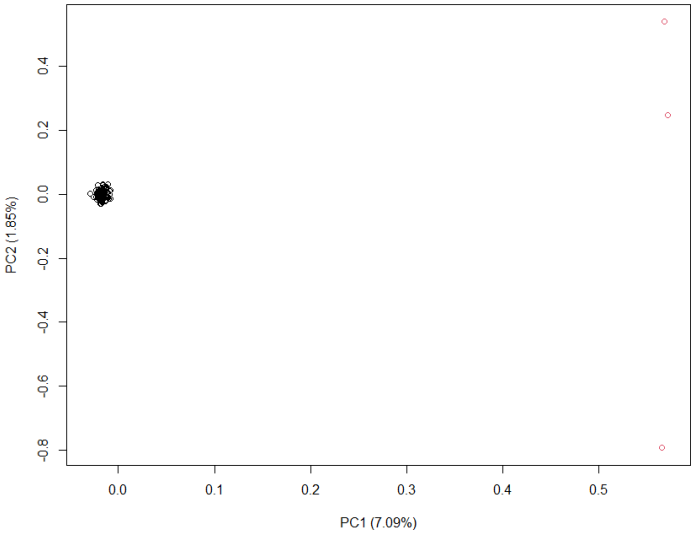


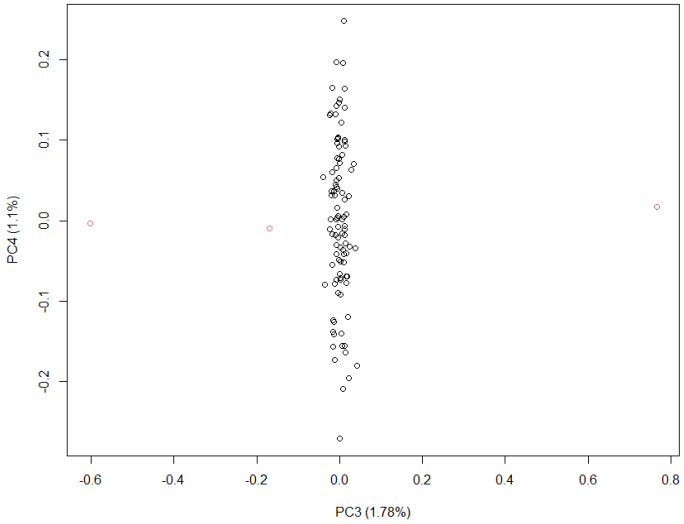
Solution Task 8

```
> summary(population)
Population size:
Total: 4108 Individuals
Of which 4105 are male and 3 are female.
There are 7 generations
and 10 unique cohorts.
255 individuals are copies of previously generated individuals.

Genome Info:
There are 10 unique chromosomes.
In total there are 10000 SNPs.
The genome has a total length of 21.05182721 Morgan.
The genome has a physical size of about: 2.102 GB

Trait Info:
There is 1 modelled trait.
The trait has underlying QTL
The trait is named: Trait 1
Total time spent for generation: 18.3 seconds.

Time spent per step:
0.7 seconds for creation of founder population.
0.8 seconds for phenotyping.
7.3 second for breeding value estimation.
0.1 seconds for selection.
9.3 seconds for generation of new individuals.
```

Solution Task 8

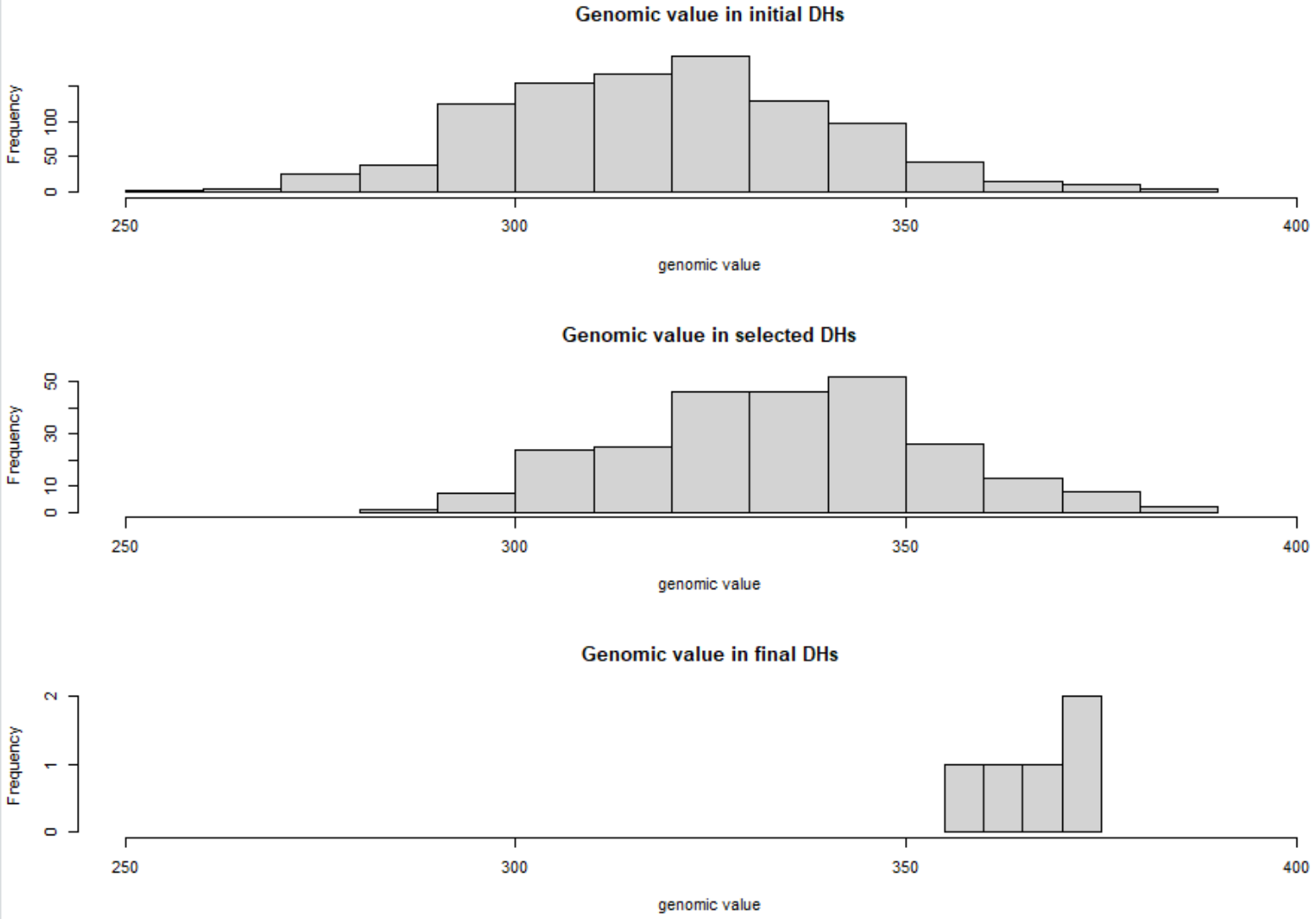
- PC 1 splits between the two pools
- PC 2,3 seem to mainly differentiates between the lines in pool 2
- PC 4 differentiates between lines in pool 1





WAGENINGEN
UNIVERSITY & RESEARCH

Solution Task 8

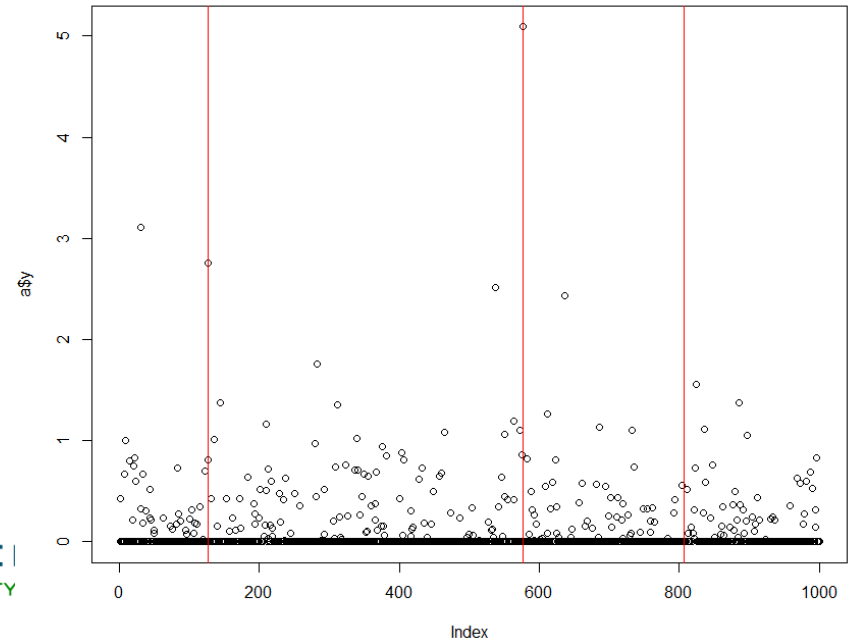


Use of own methods for GWAS/BVE

```
library(MoBPS)

# Fixate random seed for a uniform result
set.seed(1)
dataset <- founder.simulation(nsnp=1000)
set.seed(2)

population <- creating.diploid(dataset = dataset)

geno <- get.geno(population, gen=1)

hist(rowMeans(geno))

# Place QTL effects on some markers with
QTLs <- matrix(c(126,1, 0,1,2,
                 577,1,0,1,2,
                 806,1,0,1,2), byrow=TRUE, ncol=5)

population <- creating.trait(population, real.bv.add = QTLs)


population <- breeding.diploid(population, heritability = 0.5, phenotyping = "all")

pheno <- get.pheno(population, gen=1)
```

Genome-wide association study

```
# Use any software solution for GWAS analysis
ge <- data.frame(marker=1:1000, chrom=rep(1,1000), pos = 1:1000, geno, check.names = FALSE)
ph <- data.frame(line = colnames(geno), y = pheno[1,])

a <- rrBLUP::GWAS(pheno=ph, geno=ge, plot=FALSE)
plot(a$y)
abline(v=c(126,577,806), col="red")
```



The QTL at marker 806 is not found

Depending on the used p-value markers 126 & 577 will be found but there will also be false positives then!

Marker assisted selection

```
## Marker assistent selection

# use 10 randomly selected markers

geno_mas <- geno[sample(1:1000,10),]

model <- lm(pheno[1,] ~t(geno_mas))

y_hat <- model$fitted.values

cor(y_hat, pheno[1,])

new.bve <- cbind(colnames(pheno), y_hat)
population <- insert.bve(population, bves = new.bve)

get.bve(population, gen=1)
```

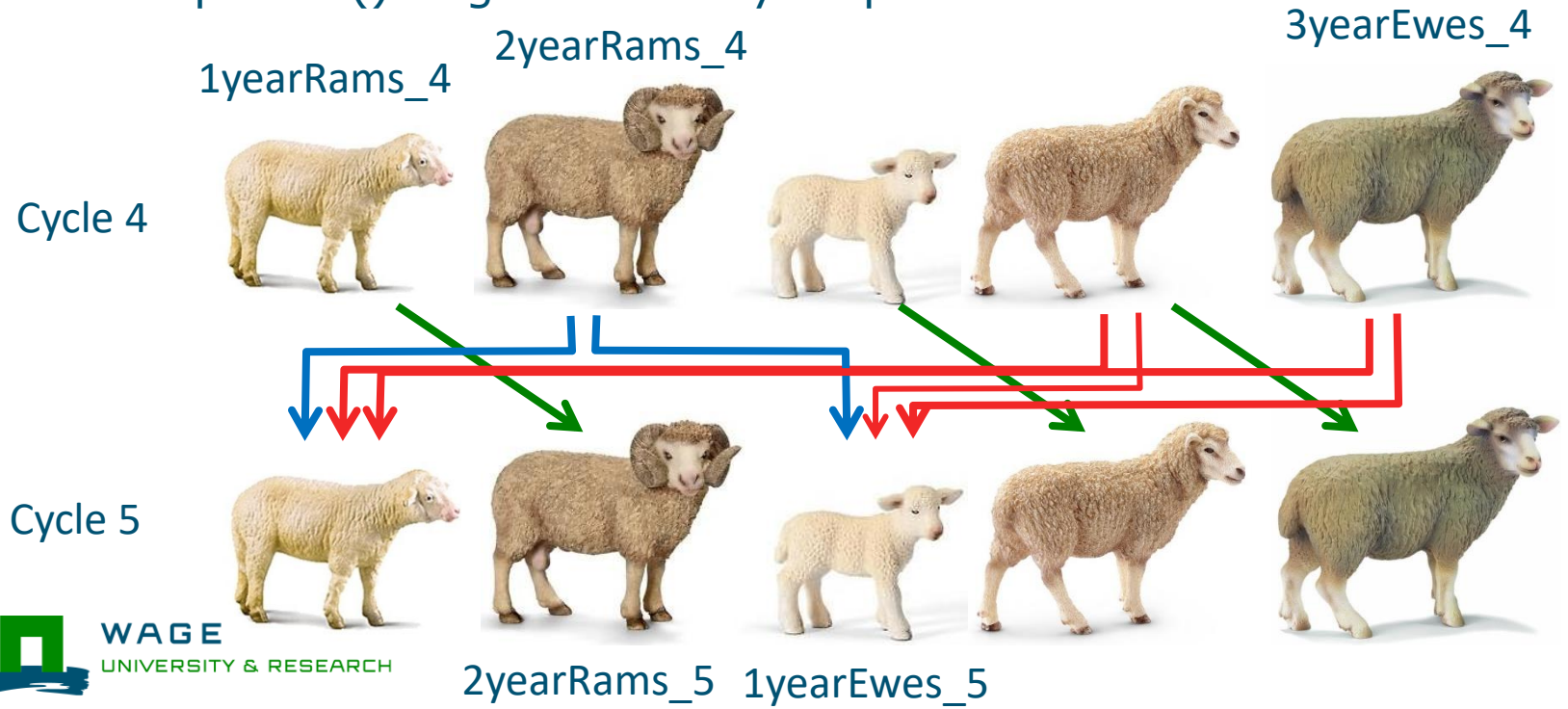
- Setting up a simulation script takes a lot of effort & time
- Most programming mistakes will lead to errors
  - These are actually the good issue
  - You have to fix them
- Minor issue are more difficult to stop
  - Check log-files for issues
  - Make sure that warnings are investigated
  - Question results when they are not in-line with your expectation

Task 13

- Download the script & log-file
- Check the simulation for potential issues & fix them

Task 10: Simulation of multiple breeding cycles

- Open the breeding scheme „Simple_Sheep_Advanced_single“
- Extend the code to simulate 20 generations of your breeding scheme
- Hints:
  - Try to use a loop instead of writting the same code multiple times
  - Use cohort names that include the cycle each cohort is in
  - paste0() might be a very helpful function!



Task 10: Simulation of multiple breeding cycles

```
> summary(population)
Population size:
Total: 5780 Individuals
Of which 2260 are male and 3520 are female.
There are 21 generations
and 145 unique cohorts.
3600 individuals are copies of previously generated individuals.

Genome Info:
There are 26 unique chromosomes.
In total there are 46545 SNPs.
The genome has a total length of 24.57031306 Morgan.
The genome has a physical size of about: 2.4488 GB

Trait Info:
There are 2 modelled traits.
Of which 2 have underlying QTL.
Trait names are:Trait 1 Trait 2
Highest correlation between genetics of traits is 0.2 (absolut value).
There are no interactions between residual effects.
Total time spent for generation: 21.3 seconds.

Time spent per step:
1.1 seconds for creation of founder population.
0.2 seconds for calculation of true genomic values.
2.3 seconds for phenotyping.
4.4 second for breeding value estimation.
0.9 seconds for selection.
12.5 seconds for generation of new individuals.
```

Task 11: Simulating different scenarios in R

- Extend the script from Task 10 to allow for the simulation of multiple runs and multiple scenarios
- Think of an interesting alternative scenario to compare the baseline against:
  1. Baseline
  2. Only 5 rams are selected for reproduction
  3. Only 50% of all ewes are genotyped
  4. Use a selection index with three times as much weight on the meat trait
- Use set.seed() to make sure that the genetic architecture of the trait is the same between the four simulations
- Store results of an individual simulation in an .RData object

Solutions Task 11

- Evaluate underlying true genomic values, kinships and prediction accuracies for all cohorts:

```
cohorts <- get.cohorts(population)

genomic_values <- accuracies <- kinships <- matrix(0, nrow=2, ncol=length(cohorts))

for(index in 1:length(cohorts)){
  # This extracts the genomic values for animals from the cohort and calculates the mean
  genomic_values[,index] <- rowMeans(get.bv(population, cohorts=cohorts[[index]]))
  # This extracts accuracies of the breeding value estimation for selected cohorts
  accuracies[,index] <- analyze.bv(population, cohorts=cohorts[index])[[1]][1,]
  # This approximates avg. kinship between animals and within ((kinship -0.5) *2 is inbreeding)
  kinships[,index] <- kinship.emp.fast(population=population, cohorts = cohorts[index])

}

save(file=paste0("Sheep_simulation_scenario", scenario, "run", run,".RData"), list=c("cohorts", "genomic_values", "accuracies", "kinships"))
```

Evaluating different scenarios in R

- Main recommendations:
  - Avoid using different scripts for different scenarios!

  →Error prone when you have to do changes in different script

  →Initialize all parameters subject to change in the beginning of your script

  - Documentation / comments on what you are doing

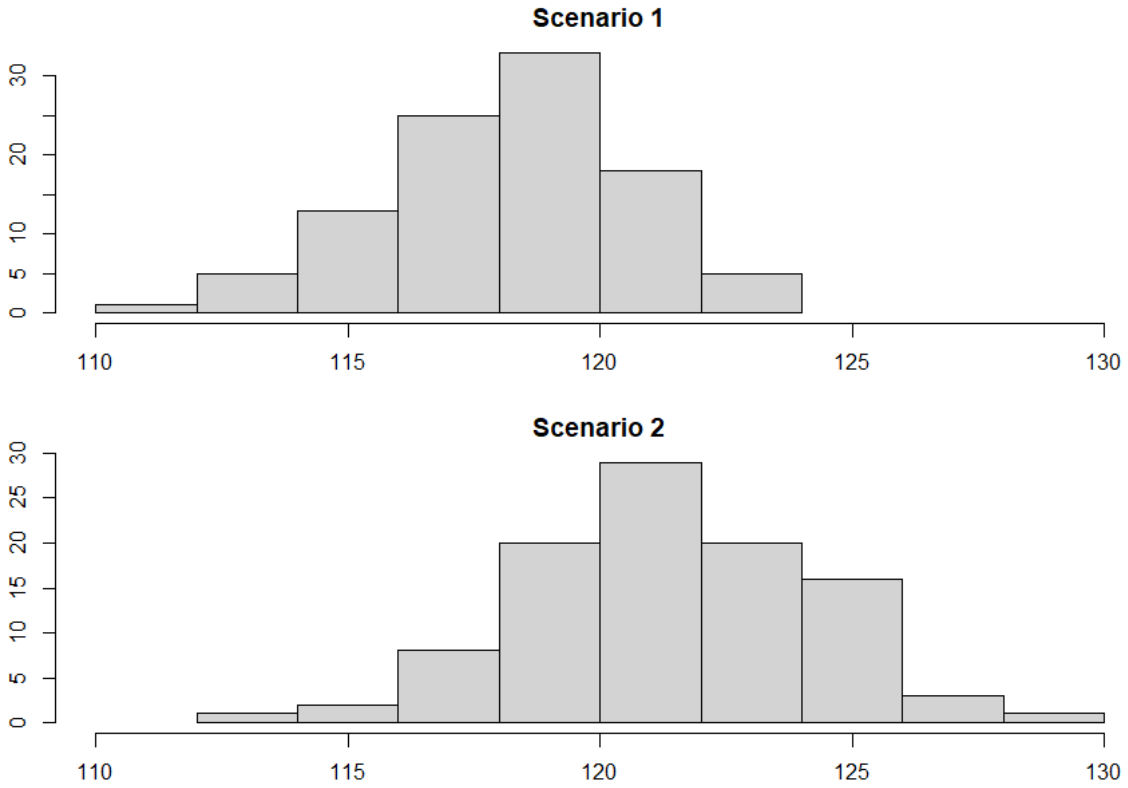
  →Good readability of code leads to less errors

  →Setting parameters not needed can help both you and other understand your code when returning to it

Task 12: Evaluating different scenarios in R

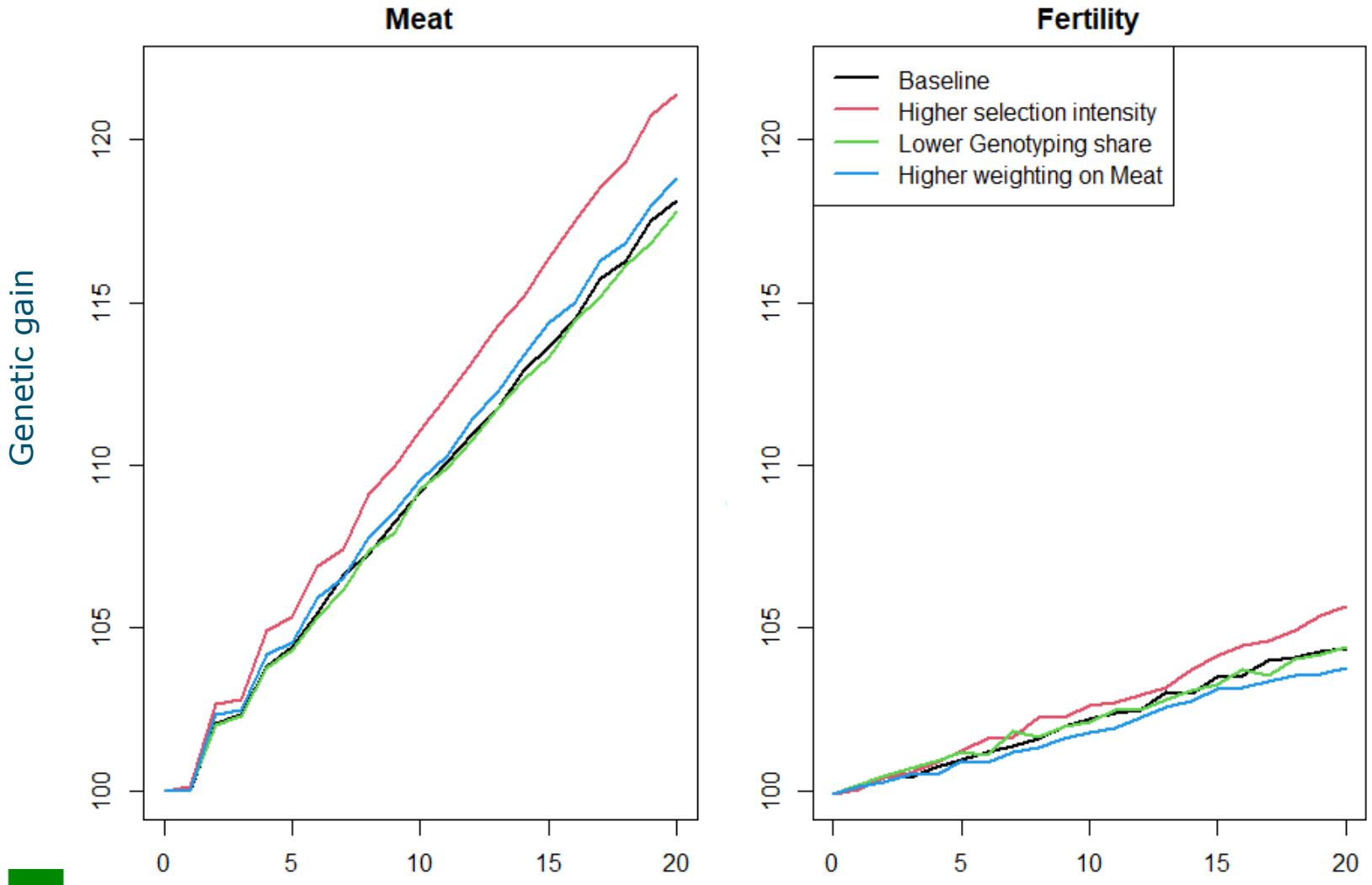
- Each of the four scenarios has been simulated 100 times
- Analyze the results of the simulation:
  - Are genetic gains for the meat trait different between scenario 1 & 2 after 20 cycles – use a t-test
  - Are the obtained prediction accuracies different in scenario 1 & 3
  - Generate a plot showing the avg. genomic values for the 1-year rams for both traits in all scenarios in the different cycles

Solution Task 12:

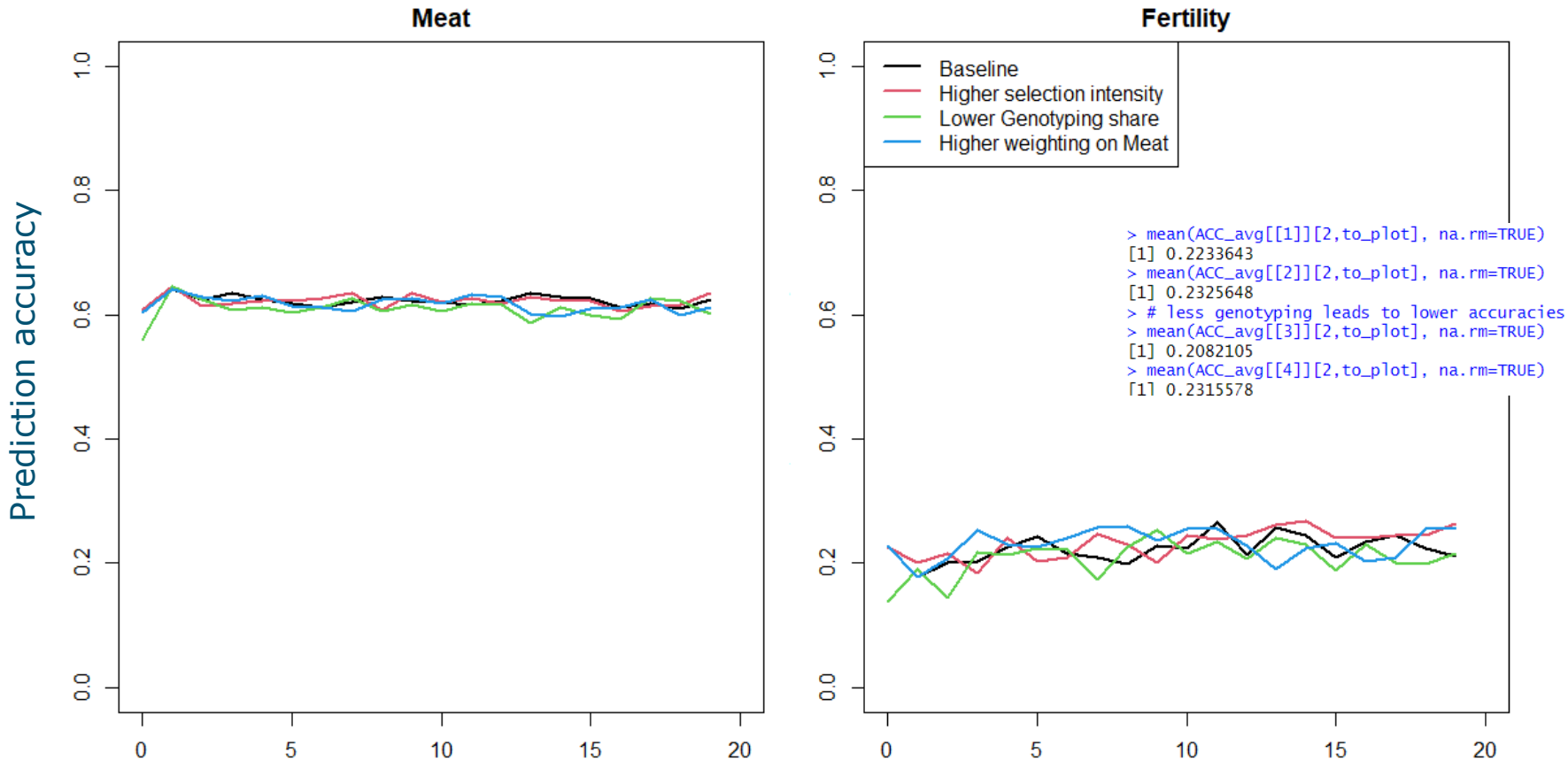



```
        Welch Two Sample t-test

data:  bvs[[1]] and bvs[[2]]
t = -8.67, df = 193.84, p-value = 1.718e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.015541 -2.527181
sample estimates:
mean of x mean of y
 118.1367  121.4080
```

Solution Task 12:



Solution Task 12:



Discussion of open questions

How to simulate XYZ?

Auxiliary material (not used in class)

Task 9 Addressing specific individuals

- Simulate a breeding program with:
  - 11 generations containing 20 males, 80 female each
  - No selection
  - Use a genetic map with 1.000 SNPs, 5 chromosomes with a length of 3 Morgan each
  - Generate a trait with heritability of 0.3 and 100 purely additive QTLs
- Phenotype female in generation 11
- Genotype all males and the best 20 females based on their phenotype
- Generate 5 offspring using the third-best male (underlying true genomic value) with the 3, 7, 17, 42, 79 best-female of generation 11 (based on phenotype)
- Check if any sires of generation 10 do not have genotyped offspring
- If yes additionally genotype one offspring for each of these sires

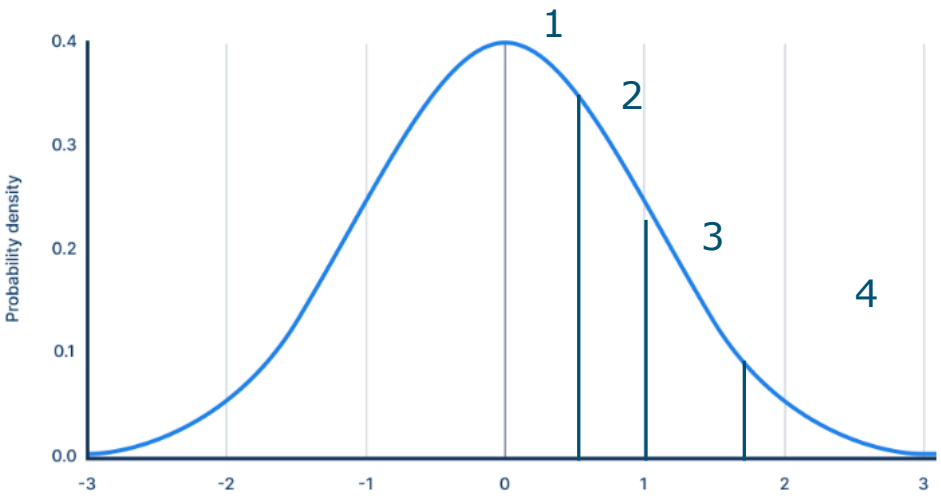
```
> is_genotyped = get.genotyped(population, gen = current_gen)
> sum(is_genotyped)
[1] 43
>
> # number of genotyped offspring
> table(pedigree[is_genotyped,6])

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 1  4  2  2  1  3  2  2  1  5  2  1  3  4  1  2  2  1  3  1
>
> # number of non-genotyped offspring
> table(pedigree[!is_genotyped,6])

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 2  4  3  2  1  3  1  3  3  5  3  3  5  3  1  7  1  1  2  4
```

Discrete phenotypes & litter sizes
Guidelines section 6.19

Transformation from normal distribution to any distribution

- Target distribution:

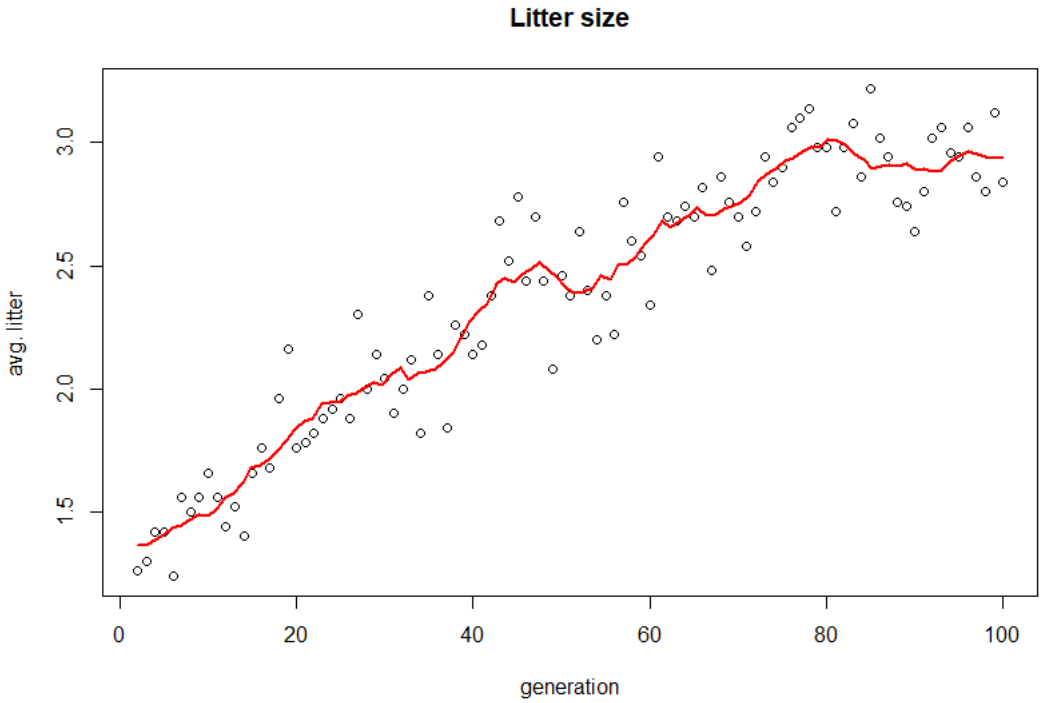
Probability	Litter size
0.70	1
0.15	2
0.10	3
0.05	4



70% / 85%/95% quantile

- Lowest values are projected to lowest discrete realizations
- Variance of the normal distribution is known (use mean/var.target)

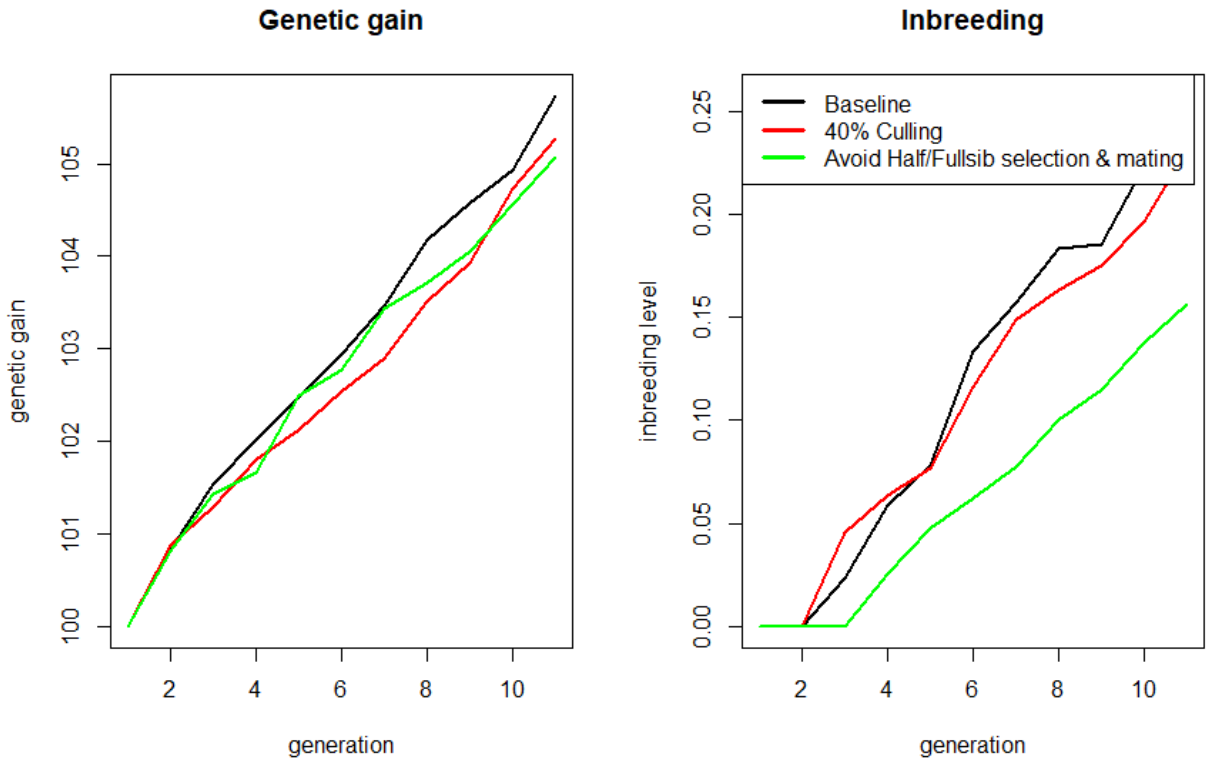
WAGENINGEN
UNIVERSITY & RESEARCH


Litter size
avg. litter
1.5
2.0
2.5
3.0
0
20
40
60
80
100
generation

Task 8a (analysis function & inbreeding management & culling)

- Simulate a breeding program with:
  - 11 generations containing 50 males, 50 female each
  - Parents for the next generation are selected as the best 10 males and 10 females based on phenotype (both males and females are phenotyped)
  - Use a genetic map with 1.000 SNPs, 5 chromosomes with a length of 3 Morgan each
  - Generate a trait with heritability of 0.3 and 100 purely additive QTLs
- Calculate genetic gain and inbreeding levels per generation
- Addition 1: Modify the selection process to reduce inbreeding (while not compromising genetic gain too much)
- Addition 2: Simulate that 40% of your animals are culled before selection (at random).

Solution Task 8a



- This is just a single replicate!
- See Task 12 for a more general solution