



THE UNIVERSITY
of EDINBURGH



Biotechnology and
Biological Sciences
Research Council



THE ROYAL
SOCIETY

Best practices for breeding program simulation

Jon Bancic & Gregor Gorjanc

Athens, Greece

2025-01-30



Learning objectives

Lecture

- Understand why we simulate breeding programs
- Review the steps for simulating a breeding program

Practical (enrol into the online course)

- Simulate a plant breeding program
- Simulate an animal breeding program

Steps for simulating a breeding program

1. Defining questions of interest
2. Outlining the breeding program
3. Specifying global parameters
4. Simulating genomes and founders
5. Populating the breeding pipeline
6. Running the burn-in phase
7. Running the future phase
8. Replication and statistical comparison

1. Defining questions of interest

- What is the question of interest?
- Determine whether simulation is necessary
- What level of complexity?

2. Outlining the breeding program

Species details

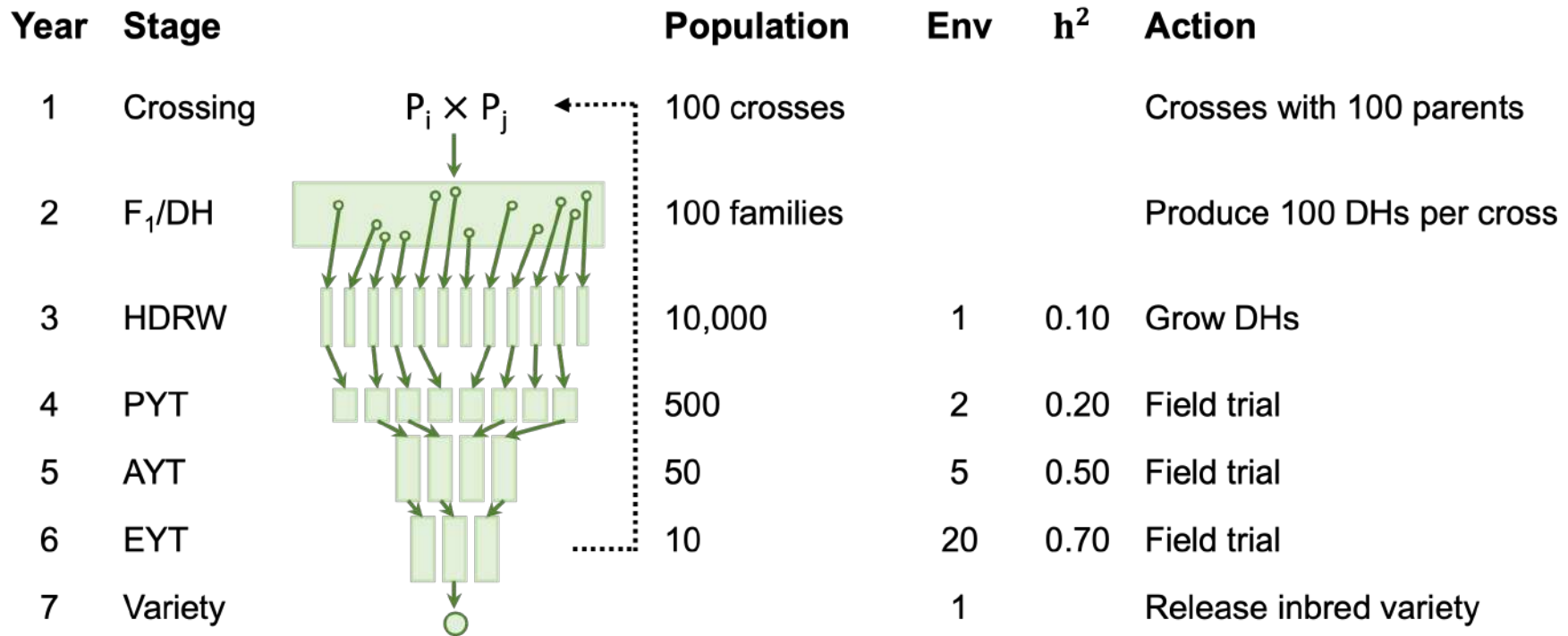
- Biology (e.g., type of mating, reproduction rate)
- Genome and evolution (genome size, mutation and recombination rates, demography)

Breeding program

- Objectives (e.g., yield, protein content)
- Numbers per breeding stage (e.g., genotypes, trials, heritabilities)
- Type of selection (e.g., individual-, family-, testcross-based)
- Breeding population (e.g. mean, variance, inbreeding, trait correlations)
- Program specificities (e.g. target growing region, statistical model)
- Logistical constraints (e.g. nursery space, number of growing environments)

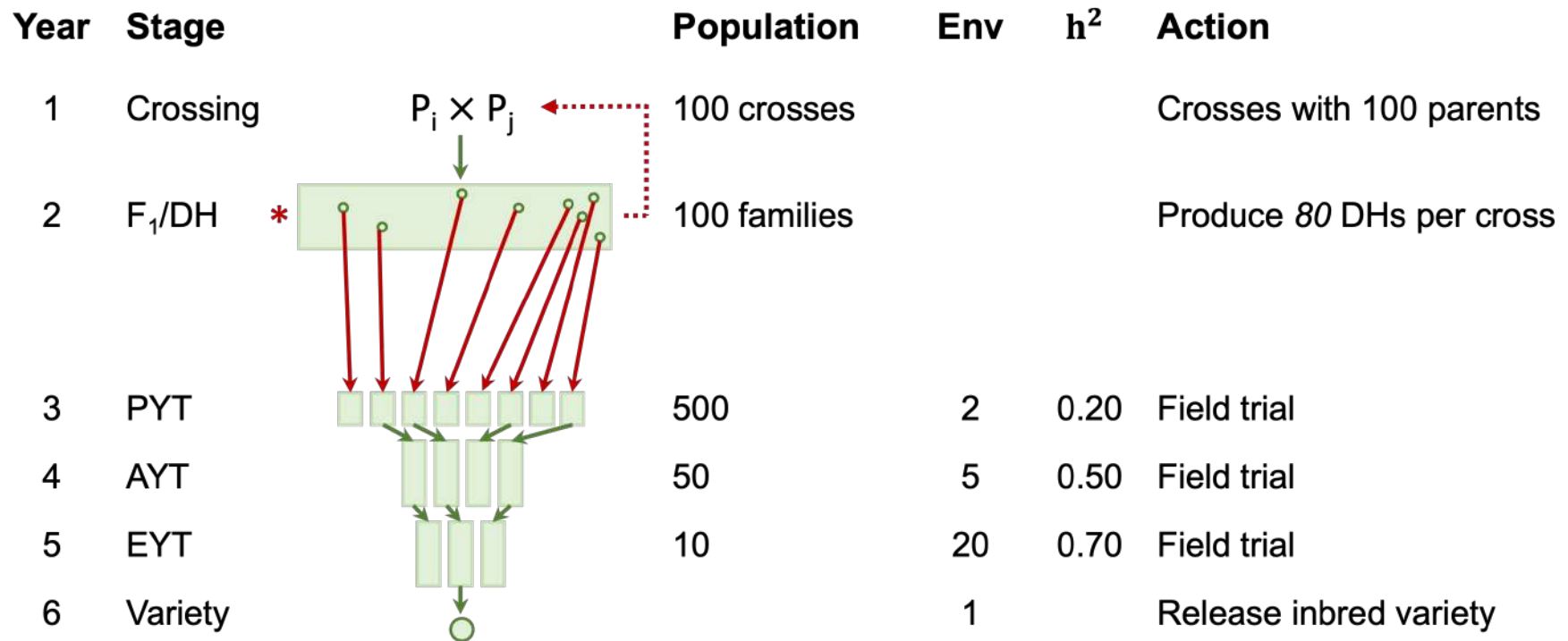
Base breeding program

Sketch it out!



Alternative breeding program

Sketch it out!



2. Outlining the breeding program

Obtain approximate costs of key actions for fair comparison

| Action | Cost (\$) | Env | Phenotypic | | Genomic | |
|-----------------------|-----------|-----|--------------|-----------|--------------|-----------|
| | | | # Units | Cost (\$) | # Units | Cost (\$) |
| Cross | 30/cross | / | 100 | 3,000 | 100 | 3,000 |
| Grow F ₁ s | 30/plant | / | 100 | 3,000 | 100 | 3,000 |
| Make DHs | 30/plant | / | 10,000 | 300,000 | 8,900 | 267,000 |
| Genotype | 15/plant | / | / | / | 8,900 | 133,400 |
| HDRW | 10/plot | 1 | 10,000 | 100,000 | / | / |
| PYT | 20/plot | 5 | 500 | 50,000 | 500 | 50,000 |
| AYT | 50/plot | 15 | 50 | 37,500 | 50 | 37,500 |
| EYT | 50/plot | 20 | 10 | 10,000 | 10 | 10,000 |
| | | | Total | 503,500 | Total | 504,500 |

3. Specifying global parameters

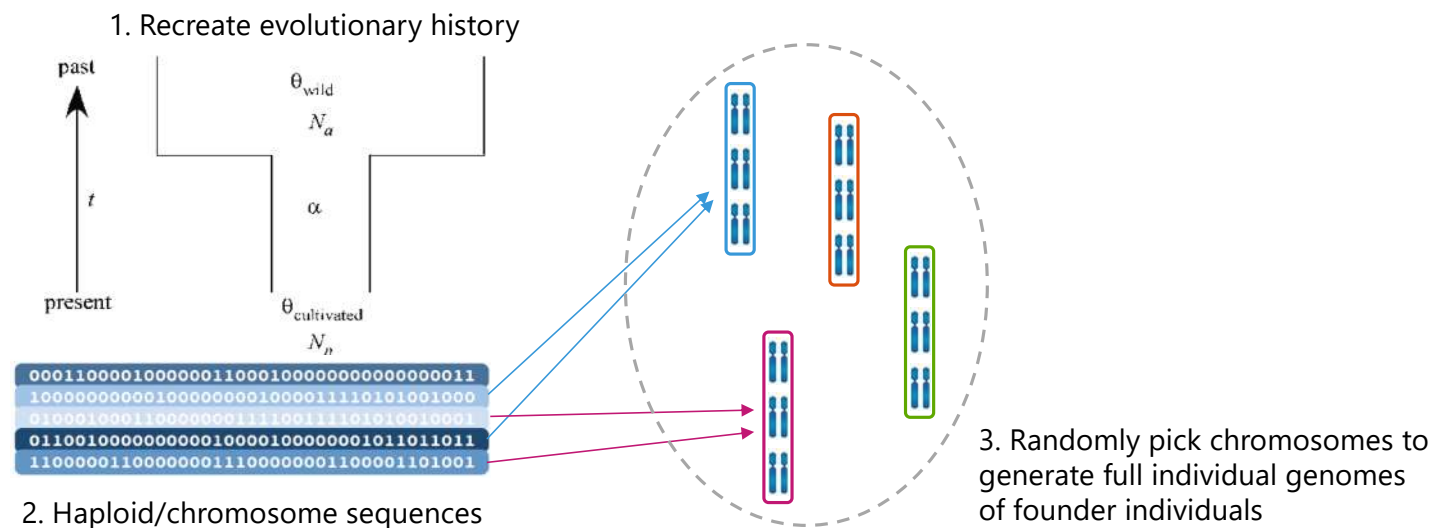
Pick simulation parameters that mimic a real breeding program

| Parameter | Definition | Value | |
|------------|---|--------|--------------------|
| nReps | Number of simulation replications | 10 | General parameters |
| nBurnin | Number of years in the burn-in phase | 20 | |
| nFuture | Number of years in future phase | 20 | |
| nQTL | Number of QTL per chromosome | 20 | |
| nSnp | Number of SNPs per chromosome | 400* | Trait parameters |
| initMeanG | Initial population mean genetic value for yield trait | 0 | |
| initVarG | Initial population genetic variance for yield trait | 1 | |
| initVarGE | Initial GxE interaction variance for yield trait | 2 | |
| varE | Yield trial error variance for yield trait | 4 | Program parameters |
| nParents | Number of parents to start a breeding cycle | 50 | |
| newParents | Number of new parents each breeding cycle | 50 | |
| nCrosses | Number of crosses among parents to start a breeding cycle | 100 | |
| nDH | Number of DH individuals produced per cross | 100/89 | |
| famMax | Maximum number of DH individuals per cross to enter PYT | 10 | |
| nPYT | Number of entries in PYT | 500 | |
| nAYT | Number of entries in AYT | 50 | |
| nEYT | Number of entries in EYT | 10 | |
| repHDRW | Effective replication in HDRW | 4/9 | |
| repPYT | Effective replication in PYT | 1 | |
| repAYT | Effective replication in AYT | 4 | |
| repEYT | Effective replication in EYT | 8 | |
| startTP | Year to start collecting training records for GS | 18* | |

4. Simulating genomes and founders

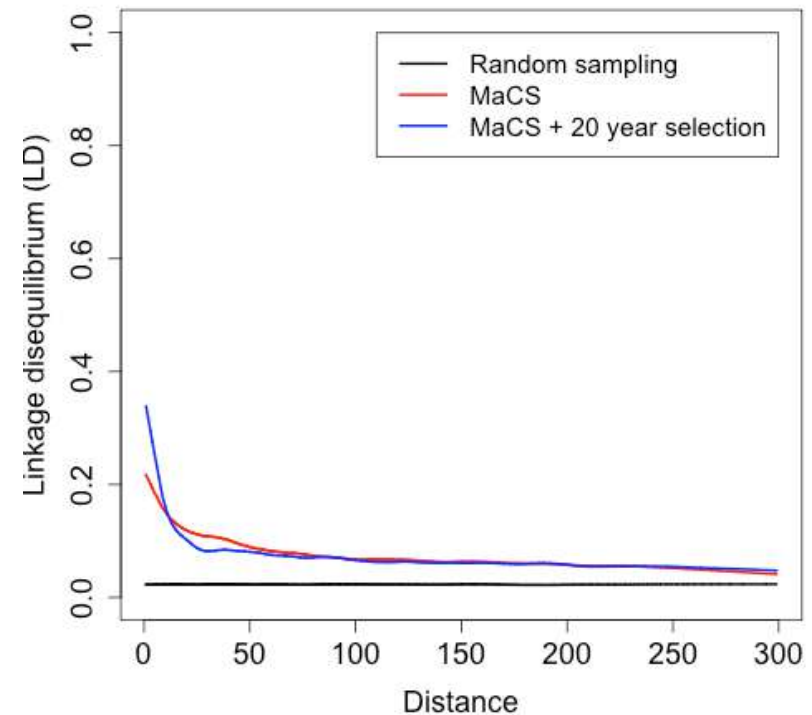
Backward-in-time coalescent simulation (MaCS, Chen et al. 2009)

1. Recreate the evolutionary history of the species
(chromosome size, mutation and recombination rate, effective population size)
2. Produce genome and haplotypes
3. Generate founder individuals



Creating founder haplotypes in AlphaSimR

- Random sampling of haplotypes
- MaCS coalescent simulation
 - Select from pre-defined species
 - Specify own evolution history
- Externally obtained haplotypes
 - SNP data
 - Other simulators (e.g., msprime)



4. Simulating genomes and founders

| Phase | Action | Feature |
|---------|----------------------------|---|
| Burn-in | Specifying Founder Genomes | 100,000 Generations of Evolution 50 inbred founders 10 chromosome pairs 1.43 Morgans per chromosome 8×10^8 base pairs per chromosome 2×10^{-9} mutation rate |
| | Specifying Trait Features | Grain yield 1,000 QTL per chromosome Normally distributed QTL effects For other values, see Table 3 |
| | Simulating Recent Breeding | 20 years of breeding Doubled haploid lines Phenotypic selection Track mean, variance, and selection accuracy |
| Future | Simulating Future Breeding | 20 years of breeding Test genomic selection Constrained and unconstrained costs 4K SNP array 5 years of training records for genomic selection Ridge regression BLUP for genomic selection |

Genome parameters

4. Simulating genomes and founders

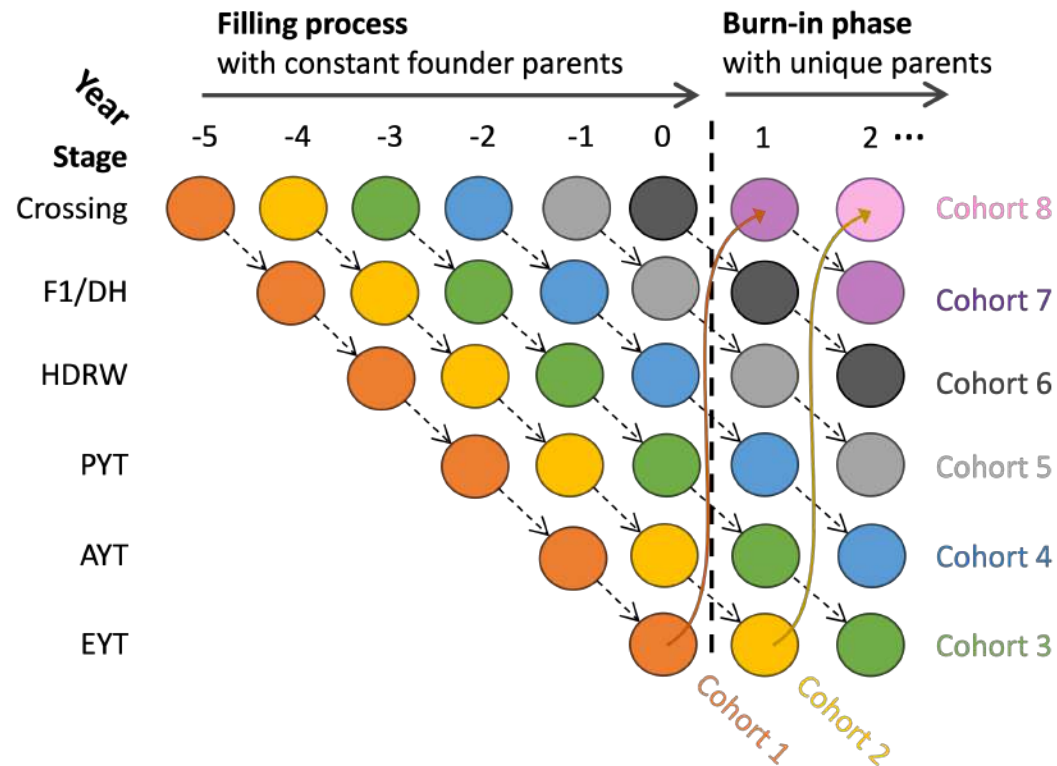
| Phase | Action | Feature |
|---------|----------------------------|---|
| Burn-in | Specifying Founder Genomes | 100,000 Generations of Evolution |
| | | 50 inbred founders |
| | | 10 chromosome pairs |
| | | 1.43 Morgans per chromosome |
| | | 8×10^8 base pairs per chromosome |
| | | 2×10^{-9} mutation rate |
| | Specifying Trait Features | Grain yield |
| | | 1,000 QTL per chromosome |
| | | Normally distributed QTL effects |
| | Simulating Recent Breeding | For other values, see Table 3 |
| | | 20 years of breeding |
| | | Doubled haploid lines |
| | | Phenotypic selection |
| | | Track mean, variance, and selection accuracy |
| | | 20 years of breeding |
| Future | Simulating Future Breeding | Test genomic selection |
| | | Constrained and unconstrained costs |
| | | 4K SNP array |
| | | 5 years of training records for genomic selection |
| | | Ridge regression BLUP for genomic selection |

Trait parameters

5. Populate the breeding pipeline

- Start of *forward-in-time* simulation in AlphaSimR
 - Model traits and recombination
 - Model breeding programs
- Populate stages with distinct cohorts to mimic generations overlap
 - Use constant founder parents to create cohorts
 - Unique cohorts arise due to randomness in crosses, selection, genetic drift and environmental noise

5. Populate the breeding pipeline



6. Running the burn-in phase

- Before formal evaluation of competing scenarios commences
 1. Represent historical breeding and create realistic starting point
 2. Generate a population structure that reflects a real population
 3. Removes burn-in oddities
- Uses the simplest program as the template

Start collecting simulation parameters

```
# A tibble: 4 × 7
  ScenarioName Rep Year Stage GeneticMean GeneticVariance SelectionAccuracy
  <chr>      <dbl> <dbl> <chr>      <dbl>      <dbl>      <dbl>
1 Base         1   21 HDRW         2.4         0.5         0.3
2 Alternative   1   21 HDRW         2.8        0.45         0.4
3 Base         1   21 EYT         3.1         0.2         0.4
4 Alternative   1   21 EYT         3.3        0.23         0.4
```

Store lists

```
[[1]]
An object of class "Pop"
Ploidy: 2
Individuals: 100
Chromosomes: 10
Loci: 1000
Traits: 1

[[2]]
An object of class "Pop"
Ploidy: 2
Individuals: 100
Chromosomes: 10
Loci: 1000
Traits: 1
```

7. Running the future phase

- Evaluation of competing scenarios commences
- Approach depends on the purpose of the study
 - **Sensitivity analysis** (e.g., number of parents)
 - Vary a single simulation parameter at the time to avoid confounding
 - Pick extreme and reasonable values from parameter space → find the breaking point
 - **Method development** (e.g., breeding program restructuring, parent selection strategy)
 - Consider unconstrained and constrained costs
- Continue collecting simulation parameters
- Use external software for specific analysis (e.g., pedigree model)

8. Replication and statistical comparison

- Replication is necessary to account for stochasticity
 1. Calculate summary statistics (mean and variance) of tracked simulation parameters and test for significance
 2. Capture and understand the key sources of variation
- No rule of thumb (at least 10 replications)
 - Number of replications depends on the desired precision, complexity and computing resources

Parallelizing replication

Step 1

Simulate burn-in for
each replicate

Replication 1: Burn-in
(\rightarrow *save .RData*)

Replication 2: Burn-in
(\rightarrow *save .RData*)

\vdots

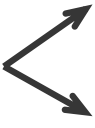
Replication n : Burn-in
(\rightarrow *save .RData*)

Parallelizing replication

Step 1

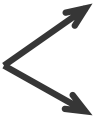
Simulate burn-in for
each replicate

Replication 1: Burn-in
(\rightarrow *save .RData*)



Scenario 1 (\rightarrow *save .rds*)
Scenario 2 (\rightarrow *save .rds*)

Replication 2: Burn-in
(\rightarrow *save .RData*)

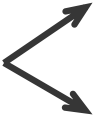


Scenario 1 (\rightarrow *save .rds*)
Scenario 2 (\rightarrow *save .rds*)

⋮

⋮

Replication n : Burn-in
(\rightarrow *save .RData*)

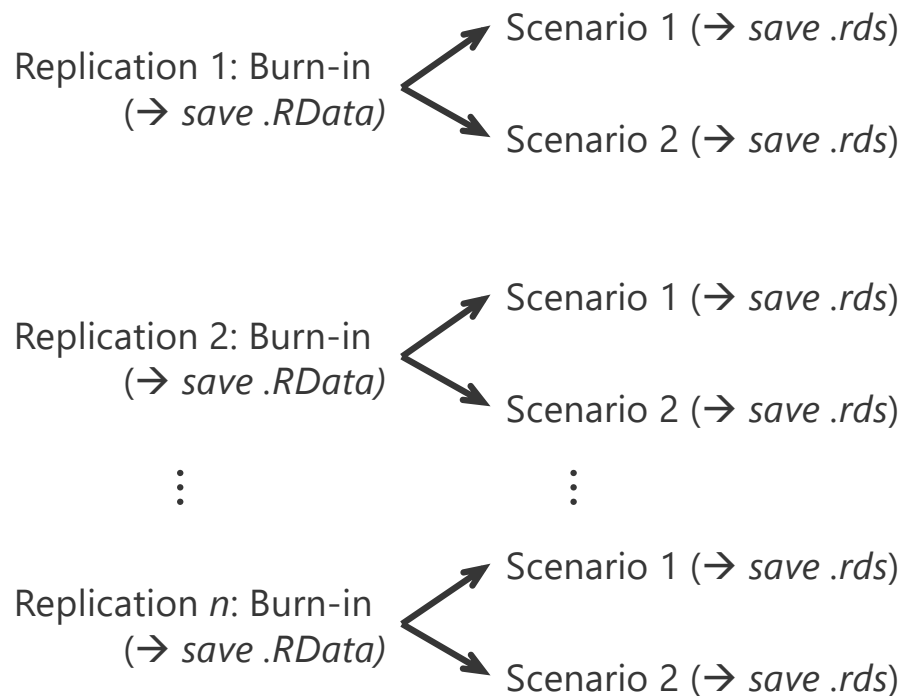


Scenario 1 (\rightarrow *save .rds*)
Scenario 2 (\rightarrow *save .rds*)

Parallelizing replication

Step 1

Simulate burn-in for each replicate



Step 2

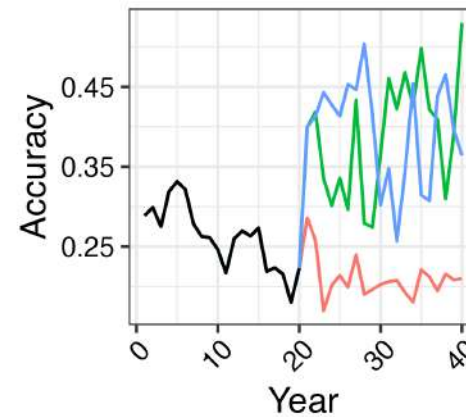
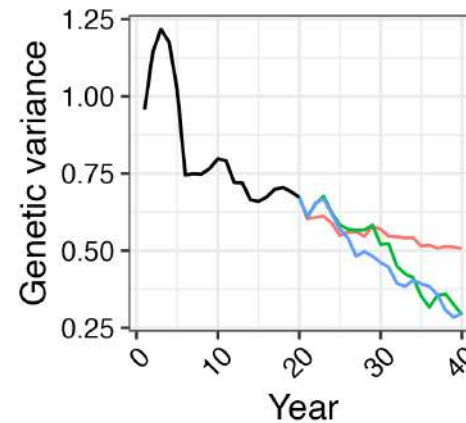
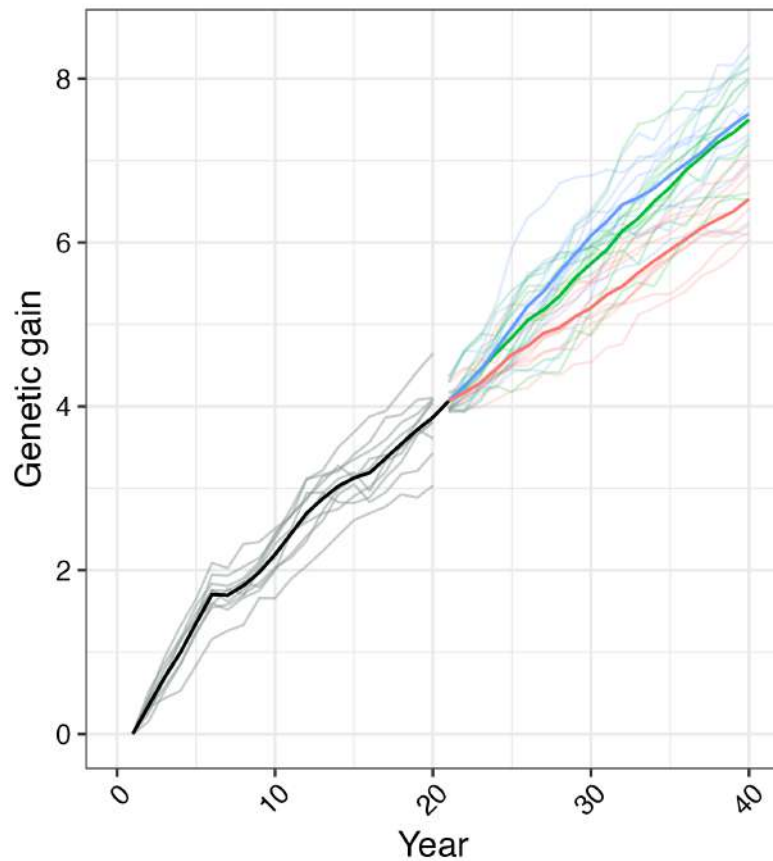
Simulate alternative scenarios

Step 3

Collate results, summarize across replicates and apply statistical tests

| | Year | Stage | Program | GEI | variable | value_Mean | value_SE |
|---|-------|-------|----------------------|----------|-------------|------------|----------|
| | <dbl> | <fct> | <fct> | <fct> | <fct> | <dbl> | <dbl> |
| 1 | 20 | AYT | Genomic Selection | High | GeneticMean | 1.72 | 0.101 |
| 2 | 20 | AYT | Genomic Selection | Low | GeneticMean | 6.93 | 0.321 |
| 3 | 20 | AYT | Genomic Selection | Moderate | GeneticMean | 4.75 | 0.198 |
| 4 | 20 | AYT | Genomic Selection | No | GeneticMean | 11.3 | 0.442 |
| 5 | 20 | AYT | Phenotypic Selection | High | GeneticMean | 0.885 | 0.0911 |
| 6 | 20 | AYT | Phenotypic Selection | Low | GeneticMean | 4.74 | 0.175 |
| 7 | 20 | AYT | Phenotypic Selection | Moderate | GeneticMean | 3.21 | 0.135 |
| 8 | 20 | AYT | Phenotypic Selection | No | GeneticMean | 8.34 | 0.344 |

Summarising and examining results



1. Expect the outcomes
2. Examine trends
3. Discuss the results
4. Look out for bugs

Scenario

- Pheno
- GS-constrained
- GS-unconstrained

Take away messages

- Every breeding program should have a digital twin
- Clearly define the question of interest
- Gather as much information
- Start simple and gradually build up
- Track as many parameters as you want and look for bugs
- Use existing templates
- Assess averages and their variation (both are important)

Received: 6 January 2024

Accepted: 16 June 2024

DOI: 10.1002/csc.21312

Crop Science

ORIGINAL ARTICLE

Special Section: Computational Design of Changing Cropping Systems

Plant breeding simulations with AlphaSimR

Jon Bančić^{1,†}  | Philip Greenspoon^{1,†}  | R. Chris Gaynor^{1,2}  | Gregor Gorjanc¹ 

HighlanderLab / jbancic_alphasimr_plants

<> Code

Issues 2

Pull requests

Actions

Projects

Wiki

Security

jbancic_alphasimr_plants

Public

main

1 Branch

0 Tags

Go to file

gregorgorjanc

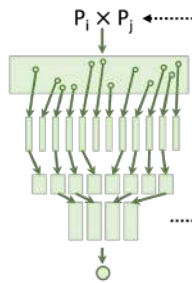
Update README.md - CS paper

8ed5cb0 · 3

| | |
|-------------------|---|
| 01_LineBreeding | Two-part program corrections, figures added |
| 02_ClonalBreeding | Two-part program corrections, figures added |
| 03_HybridBreeding | Two-part program corrections, figures added |
| 04_Features | Update miscellaneousSlot.R |
| .gitignore | Polish of all the scripts |
| LICENSE | Update LICENSE |
| LineBreeding.Rmd | Some updates |
| LineBreeding.html | Some updates |
| README.md | Update README.md - CS paper |

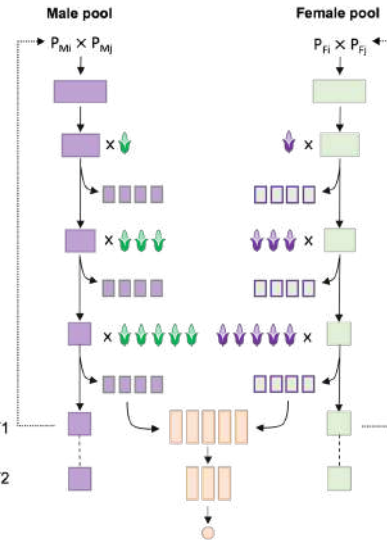
Clonal breeding program

| Year | Stage | Population | Env | h^2 | Action |
|-------|-----------|--------------|-----|-------|---|
| 1 | Crossing | 100 crosses | | | Crosses with 20 parents |
| 2 | Seedlings | 100 families | | <0.1 | Germinate 100 seeds per family in nursery |
| 3-5 | PCT | 2,000 | 1 | <0.1 | Grow propagated clones |
| 6-10 | ACT | 500 | 2 | 0.50 | Multi-year yield trials |
| 11-16 | ECT | 40 | 7 | 0.70 | Multi-year yield trials |
| 17 | Variety | | 1 | | Release clonal variety |

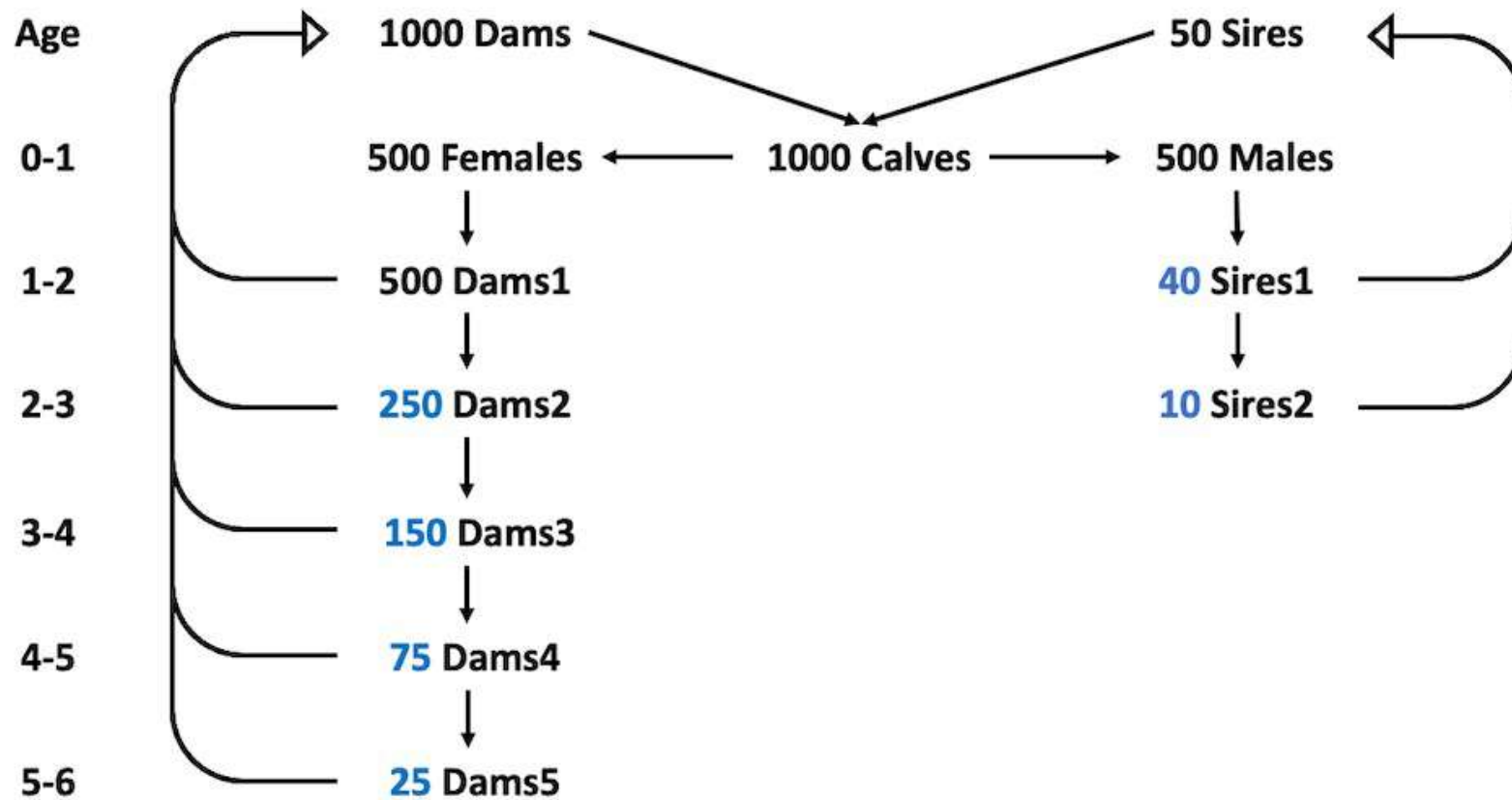


Hybrid breeding program

| Year | Stage | Male pool | Female pool | Population | Env | h^2 | Action |
|------|------------|---------------------------|----------------------------|---------------|-----|-------|------------------------------|
| 1 | Parents | $P_{Mi} \times P_{Mj}$ | $P_{Fi} \times P_{Fj}$ | 80 crosses | | | Crosses with 50 parents |
| 2 | F_1 | | | 80 families | | | Grow F_1 s and produce DHs |
| 2 | TC1 | \times (green plant) | \times (purple plant) | 4,000 inbreds | | | Testcross with 1 tester |
| 2 | TC1-YT | 4 purple bars | 4 purple bars | 4,000 entries | 1 | 0.06 | Testcross yield trial 1 |
| 3 | TC2 | \times (4 green plants) | \times (4 purple plants) | 400 inbreds | | | Testcross with 3 testers |
| 3 | TC2-YT | 4 purple bars | 4 purple bars | 1,200 entries | 2 | 0.10 | Testcross yield trial 2 |
| 4 | TC3 | \times (8 green plants) | \times (8 purple plants) | 40 | | | Testcross with 5 testers |
| 4 | TC3-YT | 4 purple bars | 4 purple bars | 200 entries | 4 | 0.20 | Testcross yield trial 3 |
| 5 | Hybrid YT1 | 4 orange bars | 4 orange bars | 20 hybrids | 8 | 0.40 | Hybrid yield trial 1 |
| 6 | Hybrid YT2 | 4 orange bars | 4 orange bars | 4 hybrids | 100 | 0.90 | Hybrid yield trial 2 |
| 7 | Release | | | | 1 | | Release hybrid variety |



Simple animal breeding simulation (beef, 2 sex paths)



Phenotypically best animals

Practical

Work through `10_Animal_breeding_programme.Rmd`

Study 2 papers

- 11_Slagboom_..

Slagboom et al. *Genetics Selection Evolution* (2024) 56:71
<https://doi.org/10.1186/s12711-024-00938-y>

Genetics Selection Evolution

RESEARCH ARTICLE

Open Access

The effect of phenotyping, adult selection, and mating strategies on genetic gain and rate of inbreeding in black soldier fly breeding programs



Margot Slagboom^{1*} , Hanne Marie Nielsen¹, Morten Kargo^{1,2}, Mark Henryon³ and Laura Skrubbeltrang Hansen^{1,4}

- 12_Obsteter_..

Obšteter et al. *Genetics Selection Evolution* (2023) 55:31
<https://doi.org/10.1186/s12711-023-00798-y>

Genetics Selection Evolution

SOFTWARE

Open Access

SIMplyBee: an R package to simulate honeybee populations and breeding programs



Jana Obšteter^{1*} , Laura K. Strachan², Jernej Bubnič¹, Janez Prešern¹ and Gregor Gorjanc^{2,3}



THE UNIVERSITY
of EDINBURGH



Biotechnology and
Biological Sciences
Research Council



THE ROYAL
SOCIETY

Best practices for breeding program simulation

Jon Bancic & Gregor Gorjanc

Athens, Greece

2025-01-30

