Day 1
Basics: Best practices for breeding program simulation

Jon Bancic, Chris Gaynor, Daniel Tolhurst, Gregor Gorjanc

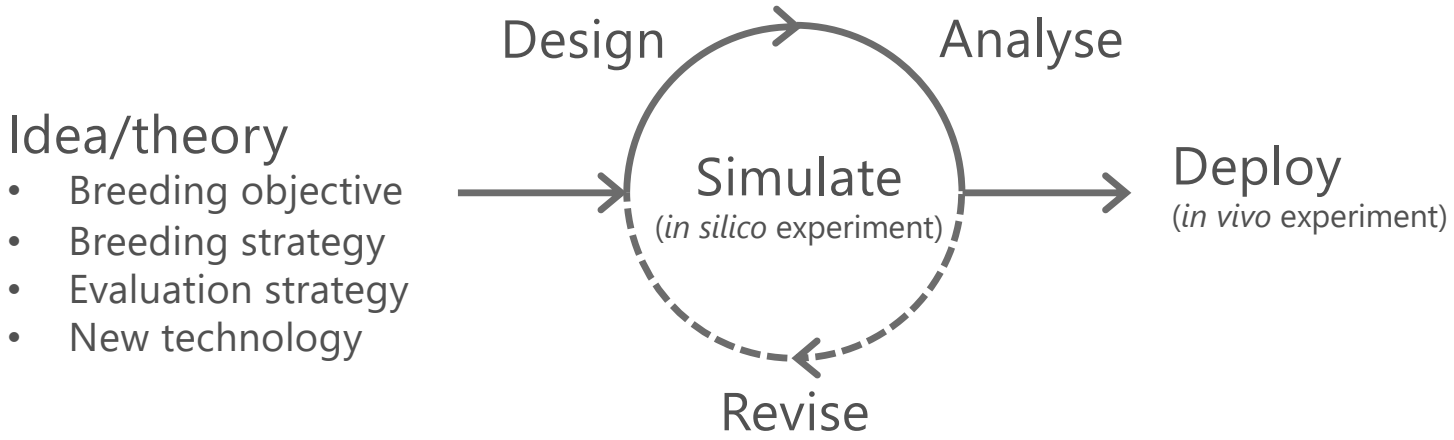
UNE, Armidale
2024-02-05

Learning objectives

- Understand why we simulate breeding programs
- Understand the steps for simulating a breeding program
- Learn to simulate a plant breeding program
- Learn to simulate an animal breeding program

Why simulate breeding programs?

- Breeding programs are complex
  (genetics, reproduction, production, disease, data, statistics,…)
- Empirical testing is ineffective, expensive and time-consuming
- Great tool for testing ideas and theory before practical implementation



Why simulate breeding programs?

- Tool to gain insight and understanding
  - "*If you can simulate it, you can understand it*"
  - Educational at all stages of career and skill set

- From breeder's perspective
  - Forces you to understand the breeding program
  - Forces communication and justification of breeding actions

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. Filling the breeding pipeline
6. Running the burn-in phase
7. Running the future phase with competing scenarios
8. Replication and statistical comparison

1. Defining questions of interest

- What is the research question?
- Determine whether simulation is necessary
- What level of complexity? → Start simple!

2. Outlining the breeding program

Crop details

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

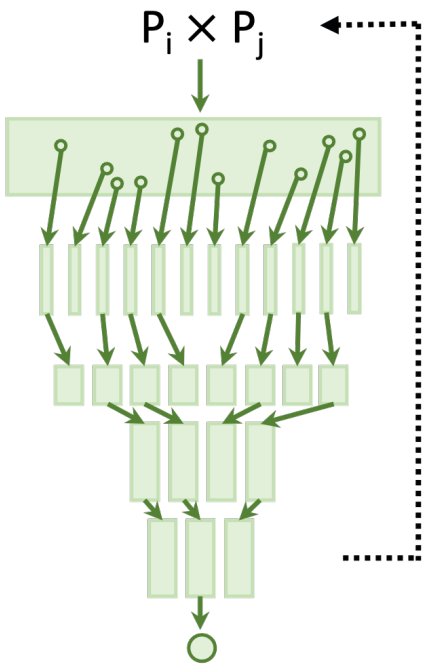
Breeding program

- 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!

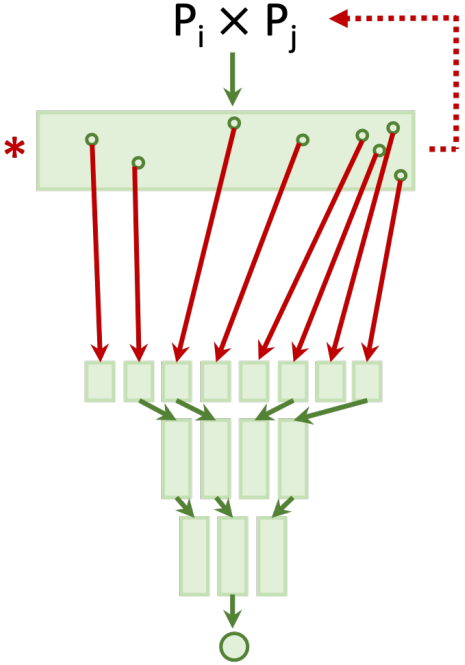
Year	Stage	Population	Env	h^2	Action
1	Crossing	100 crosses			Crosses with 100 parents
2	F_1/DH	100 families			Produce 100 DHs per cross
3	HDRW	10,000	1	0.10	Grow DHs
4	PYT	500	2	0.20	Field trial
5	AYT	50	5	0.50	Field trial
6	EYT	10	20	0.70	Field trial
7	Variety		1		Release inbred variety



Alternative breeding program

Sketch it out!

Year	Stage		Population	Env	h^2	Action
1	Crossing	$P_i \times P_j$	100 crosses			Crosses with 100 parents
2	F_1/DH	*	100 families			Produce *80* DHs per cross
3	PYT		500	2	0.20	Field trial
4	AYT		50	5	0.50	Field trial
5	EYT		10	20	0.70	Field trial
6	Variety			1		Release inbred variety



2. Outlining the breeding program

Obtain approximate costs of key actions for fair comparison

			Phenotypic		Genomic	
Action	Cost ($)	Env	# Units	Cost ($)	# Units	Cost ($)
Cross	30/cross	/	100	3,000	100	3,000
Grow F_1s	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

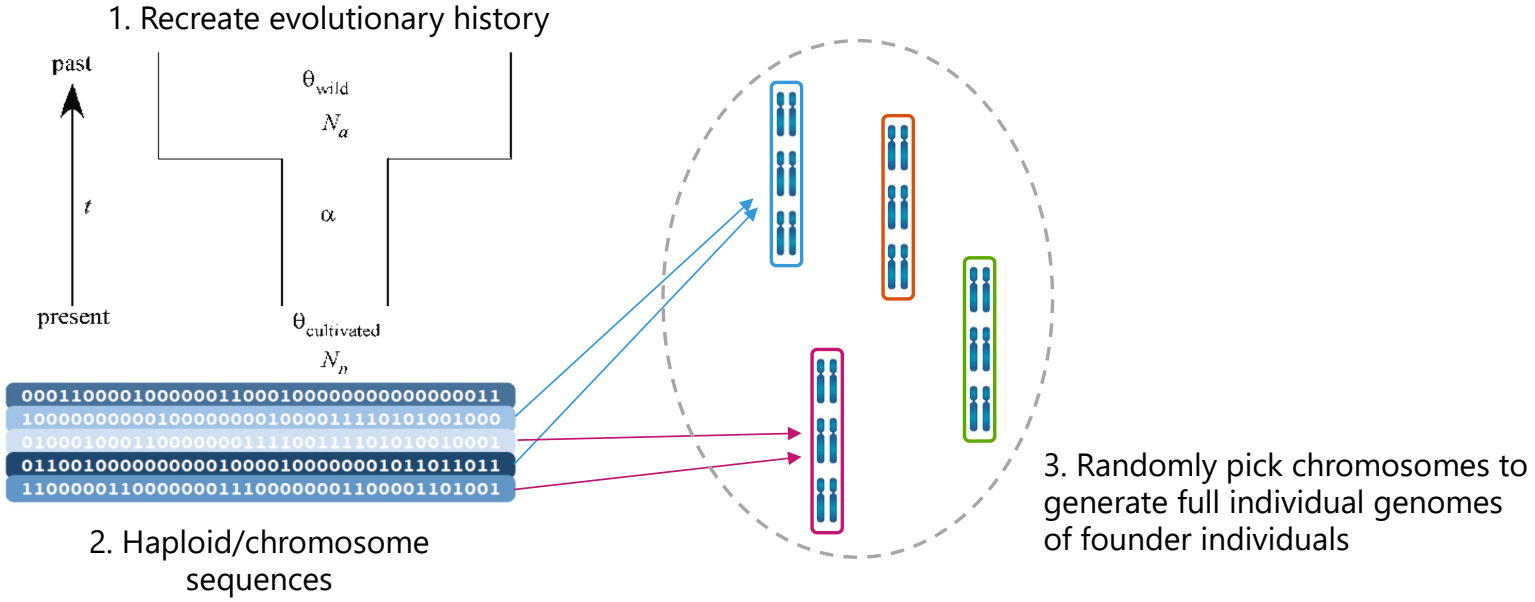
Pick simulation parameters that mimic a real breeding program

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	Trait parameters
`nSnp`	Number of SNPs per chromosome	400*	
`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	
`nParents`	Number of parents to start a breeding cycle	50	Program parameters
`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

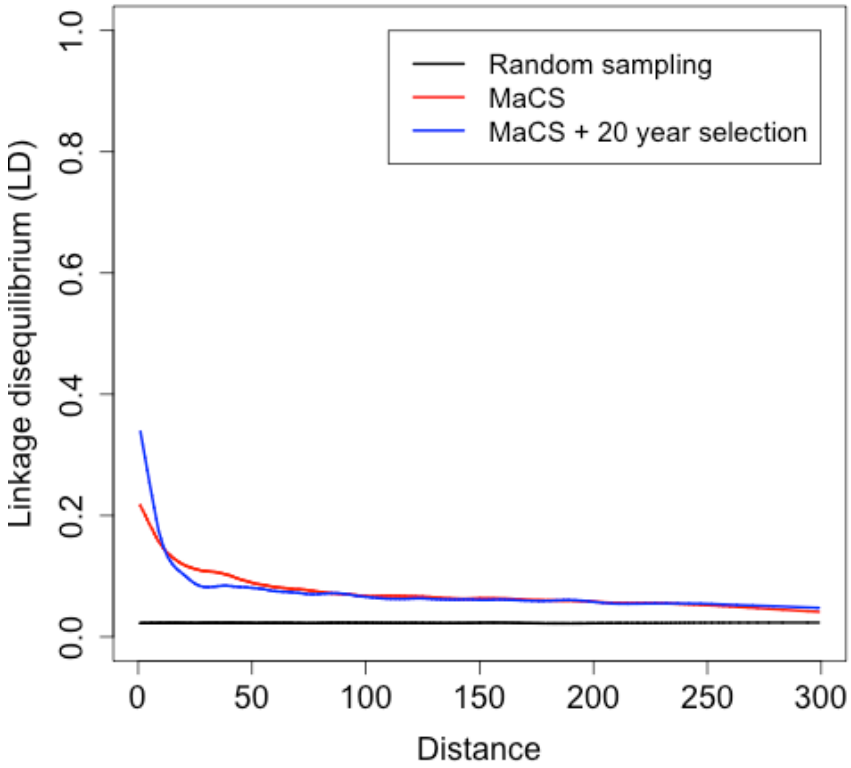
Backward-in-time coalescent simulation (MaCS, Chen et al. 2009)

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
		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

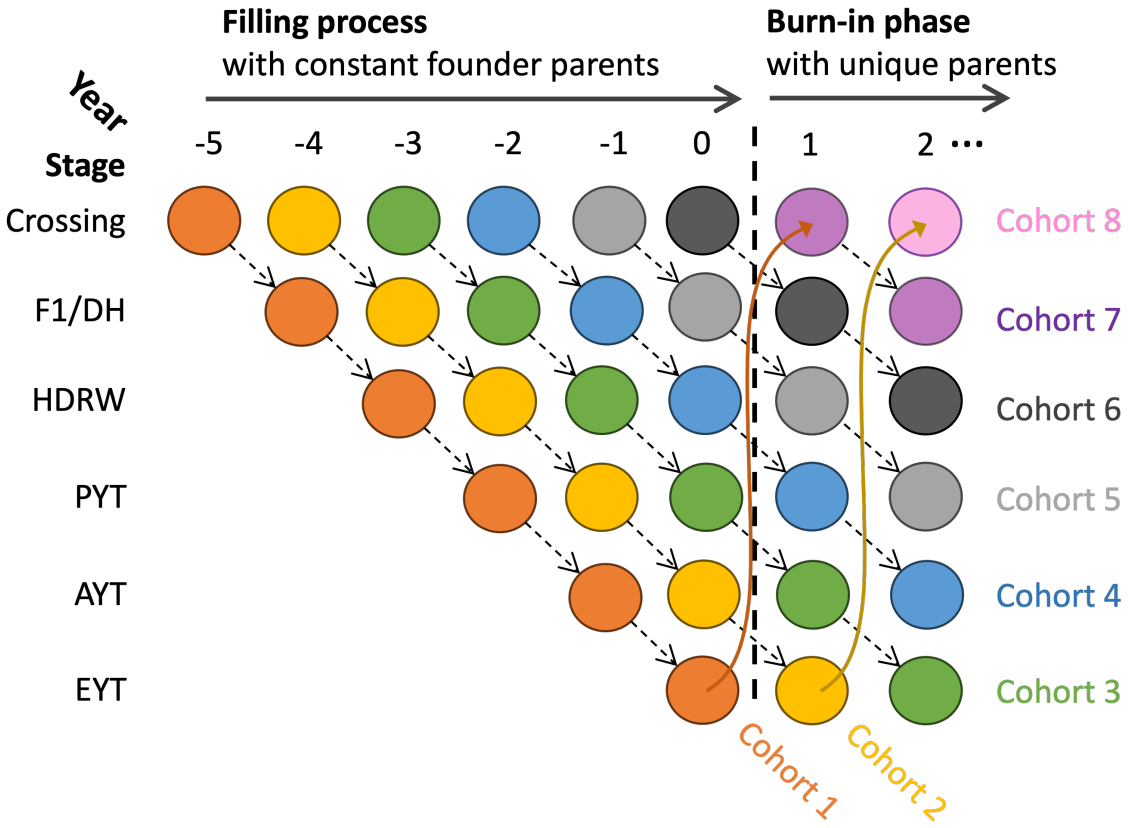
Trait parameters

5. Filling the breeding pipeline

- Start of *forward-in-time* simulation in AlphaSimR
  - Model traits and recombination
  - Model breeding programs

- Fill stages with distinct populations (or cohorts) to mimic overlapping generations
  - Use constant founder parents to create cohorts
  - Unique cohorts arise due to randomness in crosses, selection, genetic drift and environmental noise

5. Filling the breeding pipeline



6. Running the burn-in phase

- Before formal evaluation of alternative 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 with competing scenarios

- Evaluation of alternative 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. Capture and understand the key sources of variation
  2. Calculate summary statistics (mean and variance) of tracked simulation parameters and test for significance

- 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

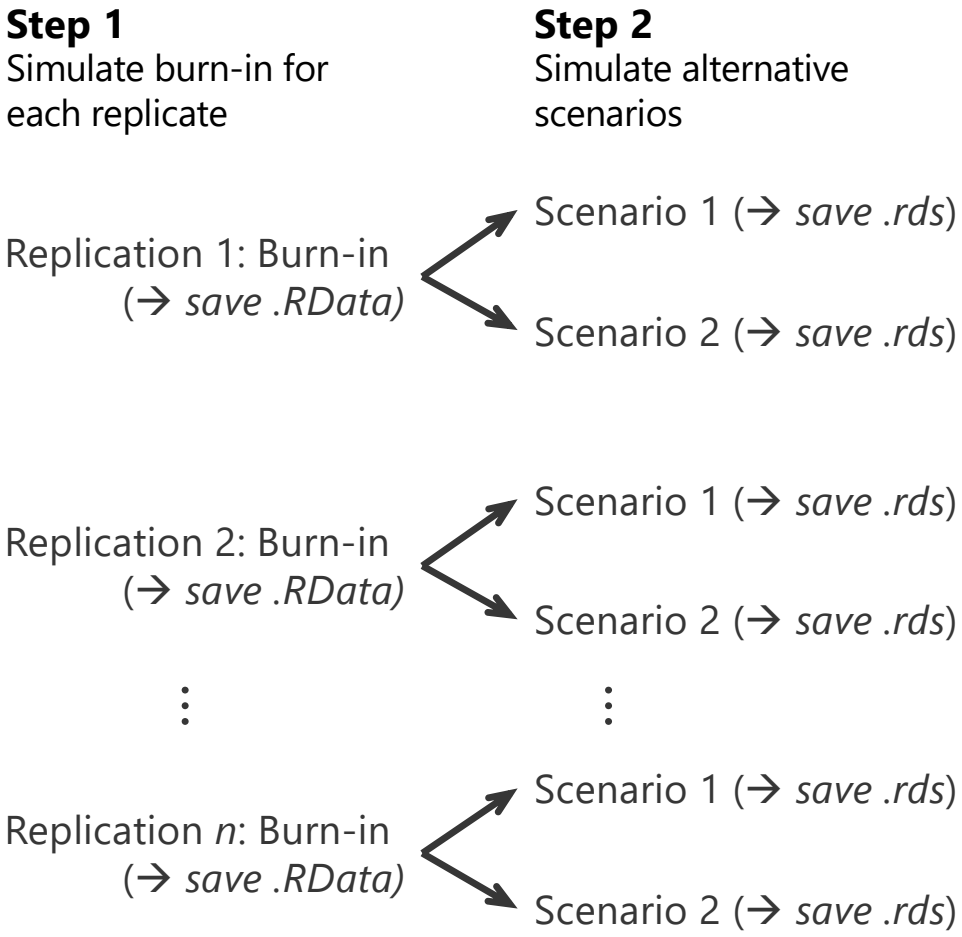
Replication 1: Burn-in
(→ *save .RData)*

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

⋮

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

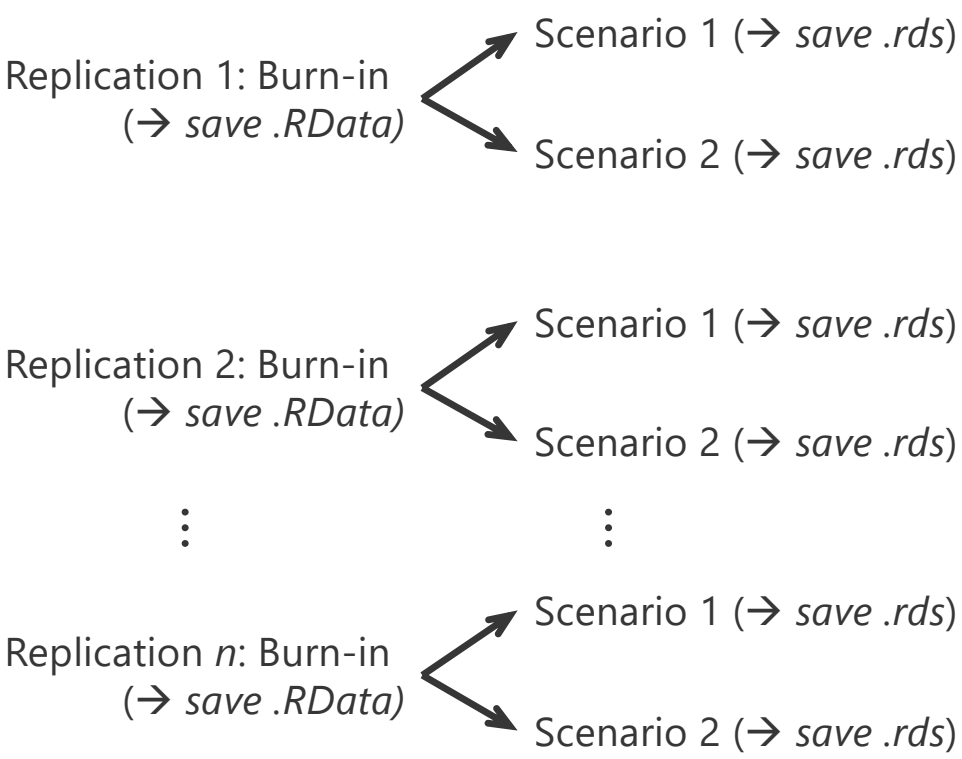
Parallelizing replication



Parallelizing replication

Step 1
Simulate burn-in for each replicate

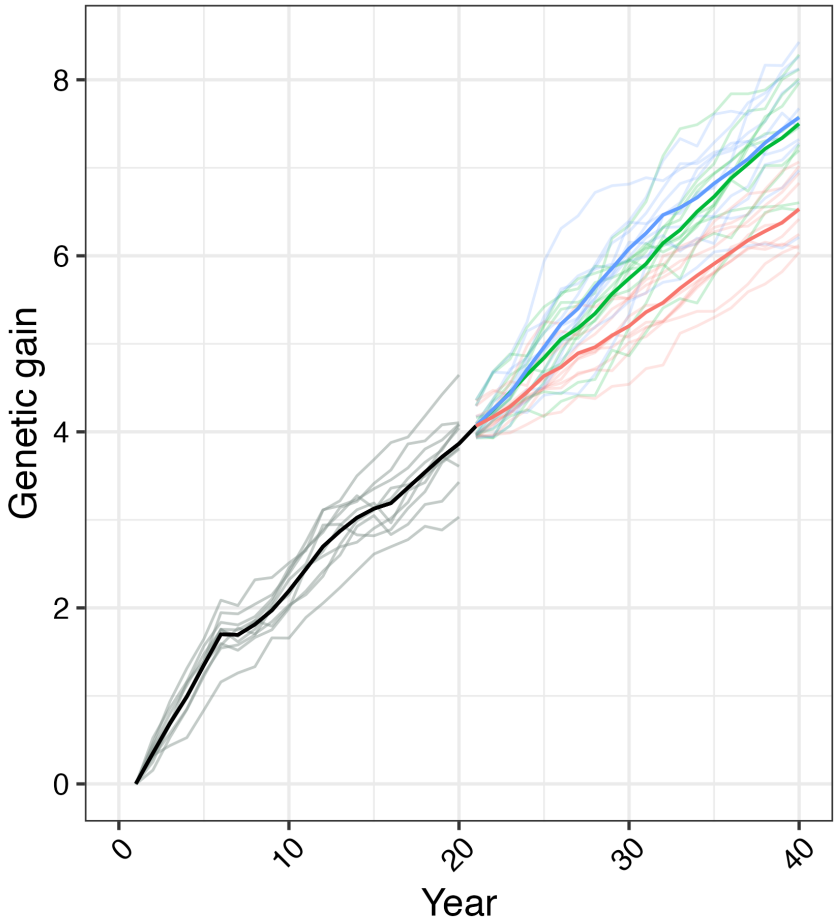
Step 2
Simulate alternative scenarios

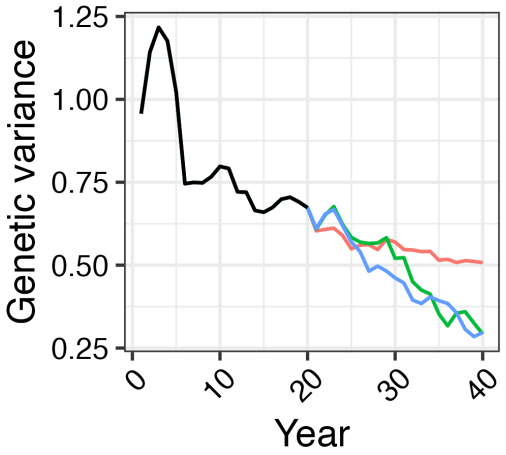


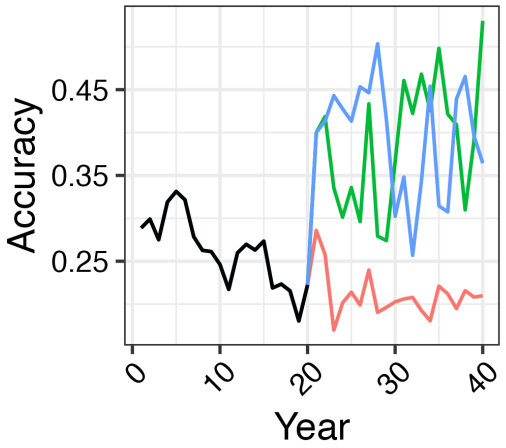
Step 3
Collate results, summarize across replicates and apply statistical tests

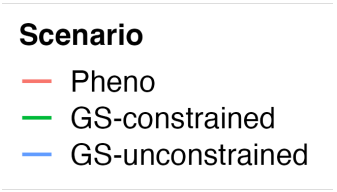
	Year	Stage	Program	GEI	variable	value_Mean	value_SE
	<dbl>	<fct>	<fct>	<fct>	<fct>	<dbl>	<dbl>
1	20	AYT	Genomic Selection	High	GeneticMean	1.72	0.101
2	20	AYT	Genomic Selection	Low	GeneticMean	6.93	0.321
3	20	AYT	Genomic Selection	Moderate	GeneticMean	4.75	0.198
4	20	AYT	Genomic Selection	No	GeneticMean	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

Take away messages

- Every breeding program should have a digital twin
- Clearly define the research problem
- Gather as much information
- Start simple and gradually build up
- Track as many parameters as you want and look for bugs
- Use existing templates

Plant breeding simulations with AlphaSimR

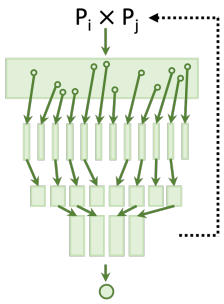
Jon Bančič[1*†], Philip Greenspoon[1†], Chris R. Gaynor[1,2] and Gregor Gorjanc[1]

[1*]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh, EH25 9RG, Midlothian, United Kingdom.
[2]Bayer Crop Science, 700 Chesterfield Pkwy W., Chesterfield, MO 63017, Missouri, USA.

https://www.biorxiv.org/content/10.1101/2023.12.30.573724v1.article-metrics

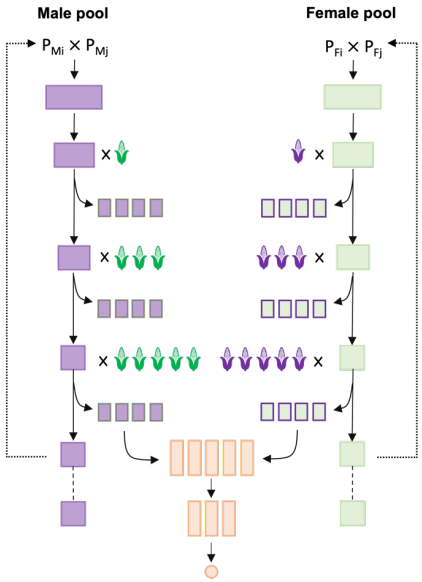
Clonal breeding program

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	Population	Env	h^2	Action
1	Parents	80 crosses			Crosses with 50 parents
2	F_1	80 families			Grow F1s and produce DHs
2	TC1	4,000 inbreds			Testcross with 1 tester
2	TC1-YT	4,000 entries	1	0.06	Testcross yield trial 1
3	TC2	400 inbreds			Testcross with 3 testers
3	TC2-YT	1,200 entries	2	0.10	Testcross yield trial 2
4	TC3	40			Testcross with 5 testers
4	TC3-YT	200 entries	4	0.20	Testcross yield trial 3
5	Hybrid YT1	20 hybrids	8	0.40	Hybrid yield trial 1
6	Hybrid YT2	4 hybrids	100	0.90	Hybrid yield trial 2
7	Release		1		Release hybrid variety



Know simulation limitations

- Simplified versions of reality
- Lengthy tunning process

AlphaSimR demonstration

10 minute Tutorial (Jon)**:** Plant breeding program simulation

30 min Exercise: Plant breeding program

10 minute Tutorial (Gregor)**:** Animal breeding program simulation

30 min Exercise: Animal breeding program

Remaining time: Other AlphaSimR features