Forward-time Individual-based Simulations in Ecology and Evolution

Module 2: Basic forward-time population simulations

Olivier Cotto
Frédéric Guillaume
Max Schmid

Tvärminne Zoological Station – University of Helsinki

March 2023

Module 2

Basic forward-time population simulations

OUTLINE

- Simulations in Population genetics: historical perspective
- Backward *vs.* Forward in time IBS
- Model assumptions of forward IBS: SLiM vs. Nemo
- The Wright-Fisher model without selection
- Simulating population demographic history with SLiM and Nemo

Simulations in population genetics

- Long history, starting in the 60's (Kimura, Nei, Felsenstein, Turelli, etc.)
- Originally locus-based, with (few) neutral loci or loci under selection, forward-in-time.
- Developed to verify assumptions and expectation of theoretical models about allele frequency dynamics, maintenance of genetic variation, etc.
- Later developed into sequence-based simulation with backward-in-time simulations (coalescent) (Hudson, 2002, MS software).
- Now fully developed forward-time IBS software for simulations in a large variety of contexts.

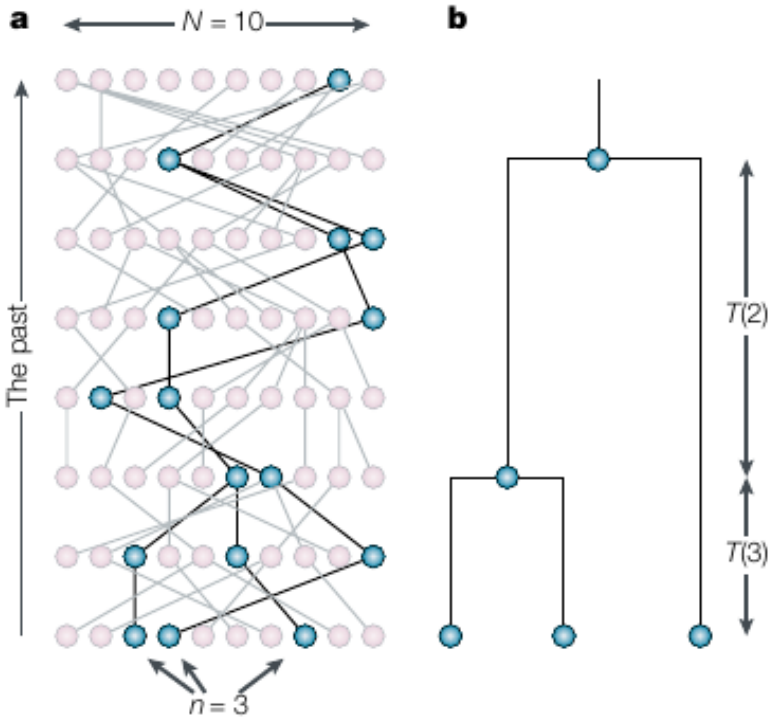
Simulations in population genetics

- Long history, starting in the 60's (Kimura, Nei, Felsenstein, Turelli, etc.)
- Originally locus-based, with (few) neutral loci or loci under selection, forward-in-time.
- Developed to verify assumptions and expectation of theoretical models about allele frequency dynamics, maintenance of genetic variation, etc.
- Later developed into sequence-based simulation with backward-in-time simulations (coalescent) (Hudson, 2002, MS software).
- Now fully developed forward-time IBS software for simulations in a large variety of contexts.

Backward *vs.* Forward in Time IBS

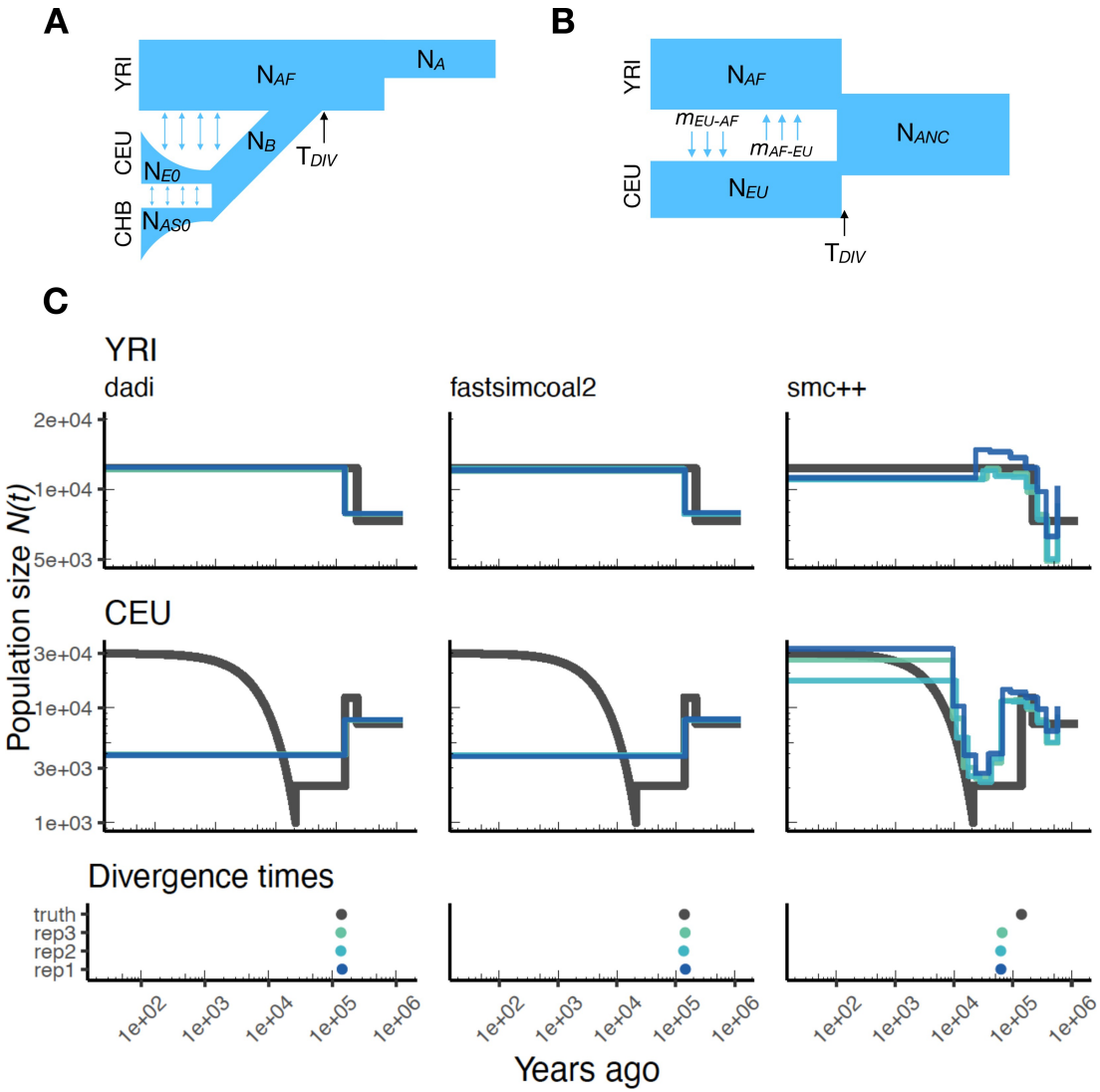
Backward-in-time

- *coalescent* simulations
- generates genealogy for current sample of $n \ll N_e$ individuals
- population is *Wright-Fisher* (constant, random mating)
- neutral evolution at *di-allelic loci*, with migration (1 selected locus)
- remains an *approximation*, works best when $N_e \geq 1000$
- low flexibility
- extremely fast



Rosenberg & Nordborg, Nature Genetics, 2002

Coalescent simulations – What for?

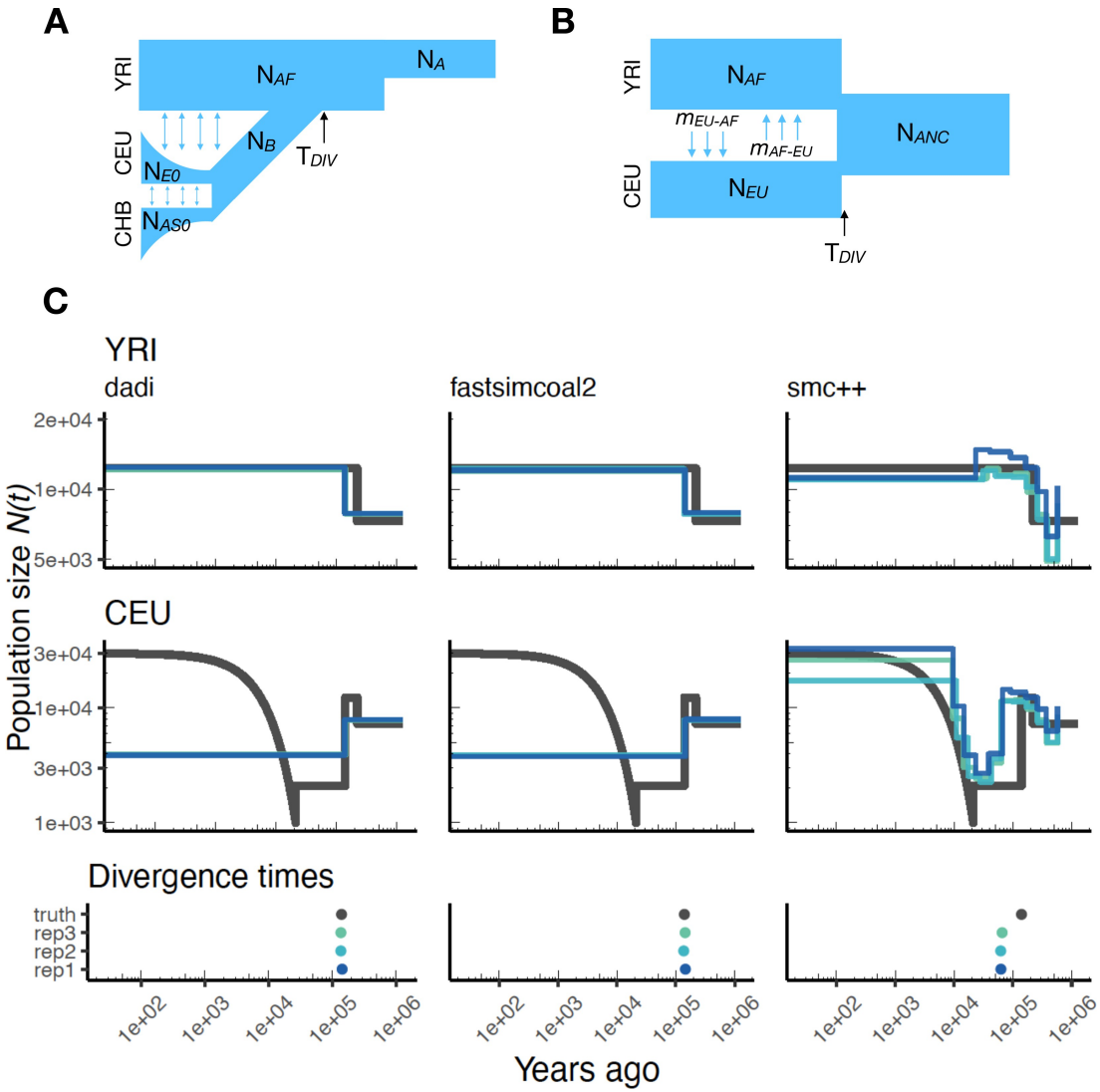


Adrion et al., eLife, 2020

Demographic inference

- reconstruct demographic history from current pattern of genome-wide variation
- use simulations to test stat methods against the “ground truth”
- assumes neutral variation

Coalescent simulations – What for?



Adrion et al., eLife, 2020

Demographic inference

- reconstruct demographic history from current pattern of genome-wide variation
- use simulations to test stat methods against the "ground truth"
- assumes neutral variation

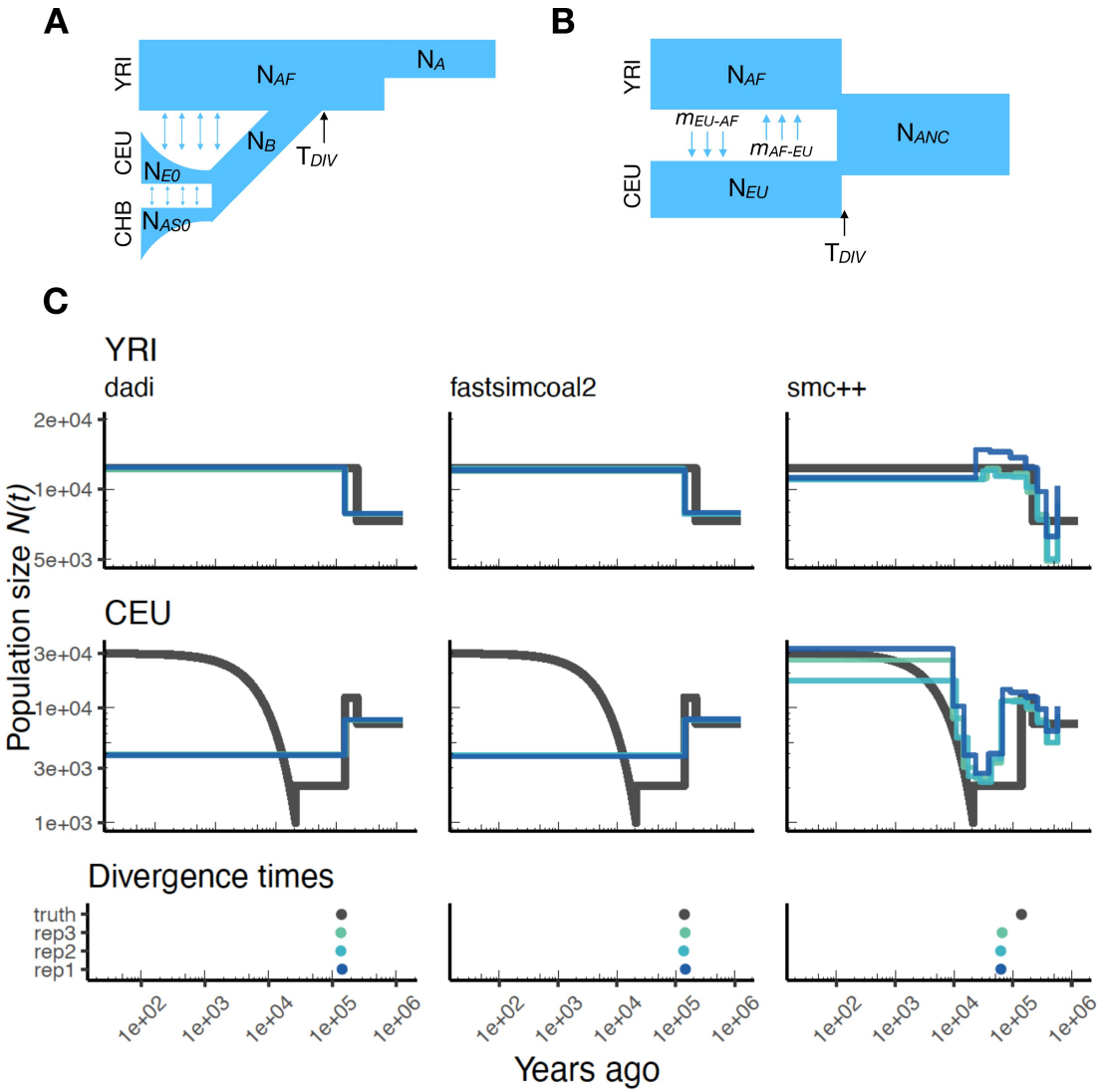
Out of Africa

Estimate past **population sizes**, **migration rates**, and **divergence time** of human populations during expansion from Africa into Europe

Figure:

black line: simulation; blue lines: inferences

Coalescent simulations – What for?



Adrion et al., eLife, 2020

Demographic inference

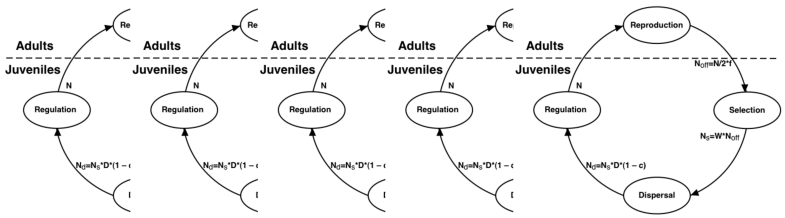
- reconstruct demographic history from current pattern of genome-wide variation
- use simulations to test stat methods against the “ground truth”
- assumes neutral variation

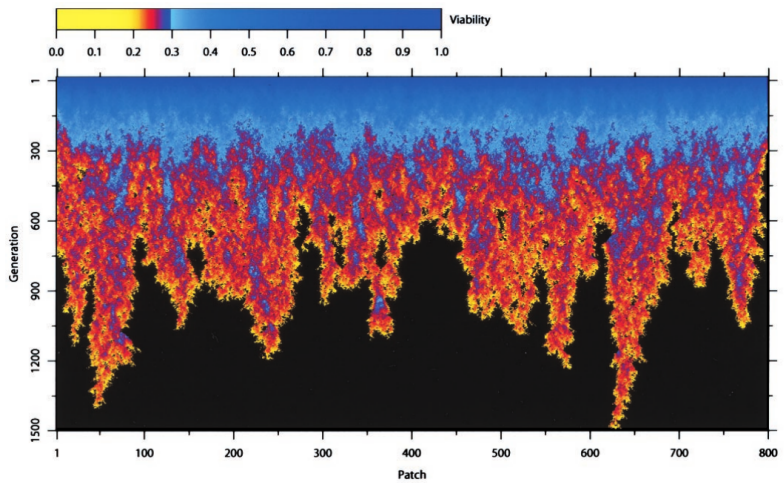
But

- how to assess effects of selection?
- e.g.: genome-wide background selection, partial/soft sweeps at loci underlying polygenic traits, etc.
- forward-time simulations are necessary here

Backward *vs.* Forward in Time IBS

Iteration of a life cycle forward in time ($T_0 \rightarrow T_{end}$)





Higgins & Lynch, PNAS, 2001

Forward-in-time

- simulates *the whole population*
- incorporates demographic stochasticity, selection, non-random mating, etc.
- in non-Wright-Fisher populations
- high flexibility (complex eco-evo scenarios)
- slower
- harder to parameterize

Forward-time simulations
A plethora of simulation software

forward-in-time

1990's	FPG (Jody Hey), easypop (Francois Balloux)
2000's	(Metapop, 2003), Nemo (2006), SimuPop (2006), SFSCode (2008), quantiNEMO (2008)
2010's	SLiM 1 (2013), fwdpp (2014), SLiM 2,3,4 (2017, 2019, 2022), and many more

Forward-time simulations
A plethora of simulation software

forward-in-time - backward-in-time

1990's	FPG (Jody Hey), easypop (Francois Balloux)
2000's	(Metapop, 2003), Nemo (2006), SimuPop (2006), SFSCode (2008), quantiNEMO (2008)
	Hudson's MS (2002)
2010's	SLiM 1 (2013), fwdpp (2014), SLiM 2,3,4 (2017, 2019, 2022), and many more
	MSprime (2016)

Forward-time simulations
How to choose the right tool?

Key issues to consider

- know the model assumptions
- understand the impact of assumptions on performances
- validation of simulation results
- parameterization of complex models
- usability: ease of parameter input, scripting, replication
- interoperability: type and format of output

Forward-time simulations – Model assumptions

Forward-time IBS are genetically (and spatially) explicit, stochastic simulations (sometimes also called Monte Carlo simulations)

- Each individuals is explicitly represented
- Genetic elements within each individual are explicitly represented
- Individuals have a spatial location (live in a deme)
- Individuals live through one or more iteration of a life cycle

Forward time IBS – Model assumptions

Genetic model:

- infinite site model (ISM) – SLiM
- finite site model (FSM) – Nemo

Infinite *vs* finite site models

ISM: implicit sequence representation

- models segregating sites only, within chromosome blocks (e.g., 1Mb)
- mapping: 1 site (locus) = 1 nucleotide = di-allelic loci (SNP)
- each mutation is unique (position drawn at random on chromosome)
- fixed mutations are removed
- individuals carry no mutation at the start

Infinite *vs* finite site models

ISM: implicit sequence representation

- models segregating sites only, within chromosome blocks (e.g., 1Mb)
- mapping: 1 site (locus) = 1 nucleotide = di-allelic loci (SNP)
- each mutation is unique (position drawn at random on chromosome)
- fixed mutations are removed
- individuals carry no mutation at the start

FSM: sparse genome representation

- models a predefined, fixed number of loci with their map position
- mapping: 1 locus = k nucleotides, $k \in [1, many]$
- models multi-allelic loci
- individuals carry all loci (starting variation possible)
- more flexibility in kind of locus (SNP to QTL)

Infinite *vs* finite site models

Applicability

ISM best for:

- sequence-based variation (SNP)
- simulate large genomic regions with low sequence variation ($\mu \sim 10^{-8}$)
- selection on di-allelic loci (e.g., deleterious mutations)

FSM best for:

- variation at large, multi-allelic loci (μ-satellites, QTL)
- gene or haplotype level variation
- selection at QTL, phenotypes, complex traits etc.

Forward-time simulations – Population

Population models:

- Wright-Fisher models – populations of constant size
- non-Wright-Fisher models – stochastic demography

Forward-time simulations – Population

Population models:

- Wright-Fisher models – populations of constant size
- non-Wright-Fisher models – stochastic demography

The Wright-Fisher model

A population of constant size N with random mating, in which N offspring are generated by random sampling of N zygotes from 2N gametes from N parents, with replacement.

The Wright-Fisher model assumptions

- constant finite size N
- random-mating
- non-overlapping generations
- one locus – or free recombination
- no selection
- no mutation

The Wright-Fisher model assumptions

- constant finite size N
- random-mating
- non-overlapping generations
- one locus – or free recombination
- no selection
- no mutation

Under Wright-Fisher model assumptions, we precisely know the **sampling variance** of the allele frequencies in the *next generation* **under drift** only:

$$\sigma_p^2 = \frac{pq}{2N}.$$

This is used to define the **effective population size** N_e of a population not following the WF assumptions whose properties are those of an equivalent WF population of size N_e.

The Wright-Fisher model assumptions

- constant finite size N
- random-mating
- non-overlapping generations
- one locus – or free recombination
- no selection
- no mutation

Under Wright-Fisher model assumptions, we precisely know the **sampling variance** of the allele frequencies in the *next generation* **under drift** only:

$$\sigma_p^2 = \frac{pq}{2N}.$$

This is used to define the **effective population size** N_e of a population not following the WF assumptions whose properties are those of an equivalent WF population of size N_e.

$\Rightarrow$ We can model a population as a WF population as long as we know its N_e.

Population models

Wright-Fisher *vs.* non-Wright-Fisher populations

- Wright-Fisher populations are good for simple population-genetics simulations with soft-selection; match the coalescent for neutral variation
- Same for simple Island-model of migration in spatially structured population
- Simple models are good for benchmarking and validation
- Non Wright-Fisher populations necessary for more stochasticity and demographic "complexity" (age/stage-structure, density dependence, etc.)

Population models

Wright-Fisher *vs.* non-Wright-Fisher populations

- Wright-Fisher populations are good for simple population-genetics simulations with soft-selection; match the coalescent for neutral variation
- Same for simple Island-model of migration in spatially structured population
- Simple models are good for benchmarking and validation
- Non Wright-Fisher populations necessary for more stochasticity and demographic "complexity" (age/stage-structure, density dependence, etc.)

How to model WF populations

- SLiM: populations are WF by default
- Nemo: populations are non-WF by default

Practice: drift and mutation in one WF pop

Model:

- N: explore a range of population sizes $\{100, 10000\}$
- genetic elements: neutral mutations (SNP)
- mutation: μ defaults to $10^{-7} - 10^{-6}$ per nucleotide
- recombination: r defaults to 10^{-8} per base pair
- chromosome length: 10^6 bases = 1cM; results in $10^6 \times 10^{-8}$ = 1% chance of x-over per individual per generation on a 1cM chromosome block
- Time: 10N generations

What to monitor:

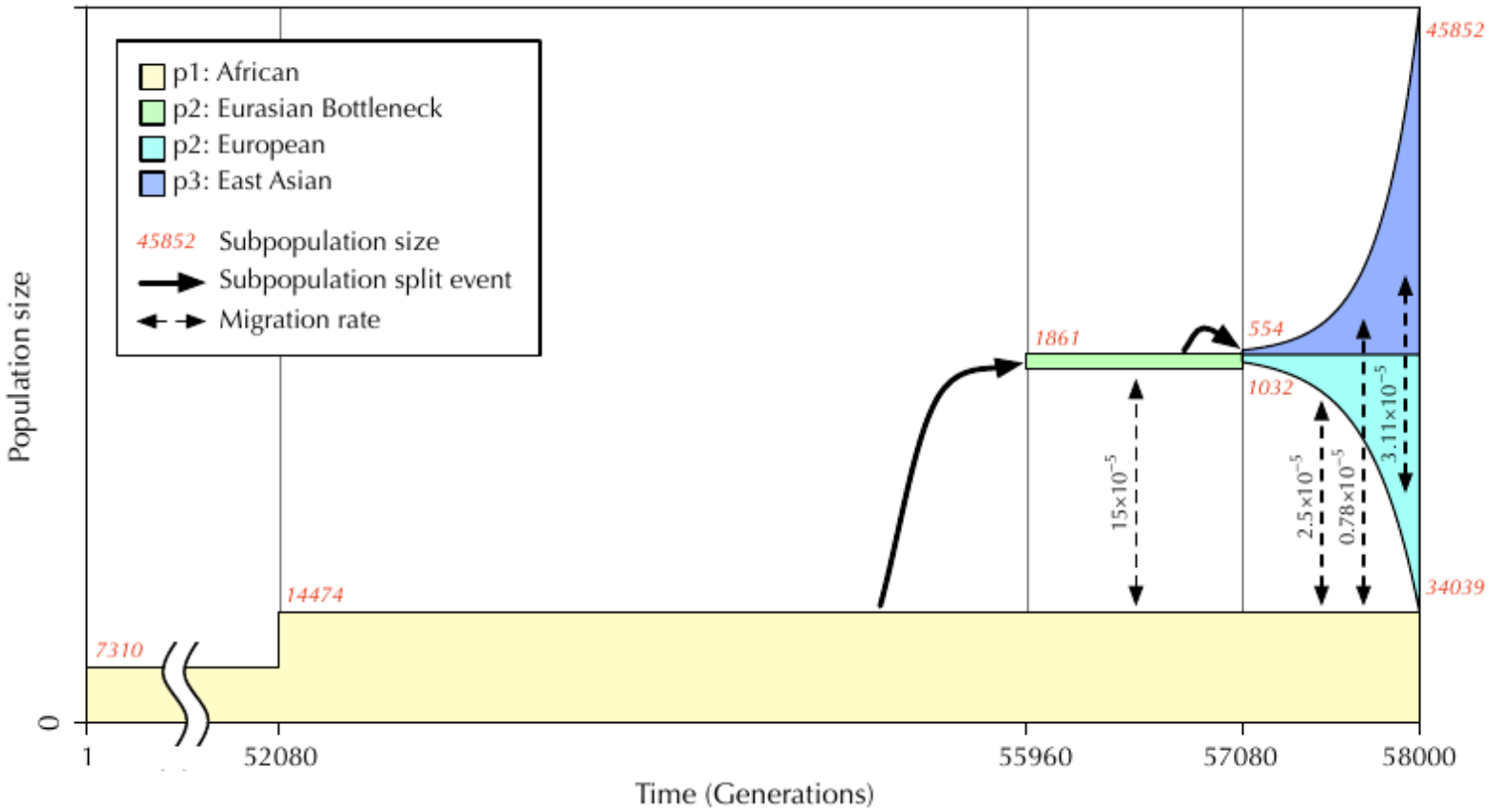
- genetic diversity measure (heterozygosity or θ)
- expectation: $\theta = 4N_e\mu$ (diversity, heterozygosity)
- expectation: $F = \frac{1}{4N_e\mu+1}$ (homozygosity)

Exercises

1. Run the `oneWFpop-ntrl` simulation with SLiM and Nemo
2. Graph the evolution of heterozygosity H_o or θ over time in R
3. Compare effects of population size and mutation rates on H, keeping $N_e\mu$ constant. Discuss the effect of scaling population size to mutation rate to lower the generation number.
4. Nemo: compare results for SNPs (2 alleles) and μ-satellites ($>$10 alleles/locus) when decreasing the number of multi-allelic loci to the same number of segregating SNPs.
5. SLiM: (optional) try and run multiple replicates of the same simulation. For this, you will need to use a `for` loop in `bash`, either in a script or directly in the terminal (or in `python`/`R` with `system()` calls).

Practice: demographic history with multiple WF pops

Out of Africa (Gravel model, 2011)



©Ben Haller

Exercise: Out of Africa

- Calculate and graph F_{ST} over time (use the Weir&Cockerham estimator in Nemo). Compare SLiM's and Nemo's results.
- Verify that population size is as expected after exponential growth, at the end of the simulation.
- Verify that migration has been modeled correctly (with the `migrants.patch` stat recorder).
- Evaluate the effect of population scaling on the genetic diversity after population bottleneck and growth by running un-scaled simulations.

Exercise: sub-divided WF populations with dispersal

1. Explore the functionalities of the `disperse` LCE, inherited by `breed_disperse` for WF populations.
2. Split `breed_disperse` into its two base components `breed` and `disperse` and try to match your previous simulation(s), this time wit a non-WF population. A key parameter now is `mean_fecundity` which influences how many offspring are produced.