(Effective) population size and coalescent theory

Brendan Reid

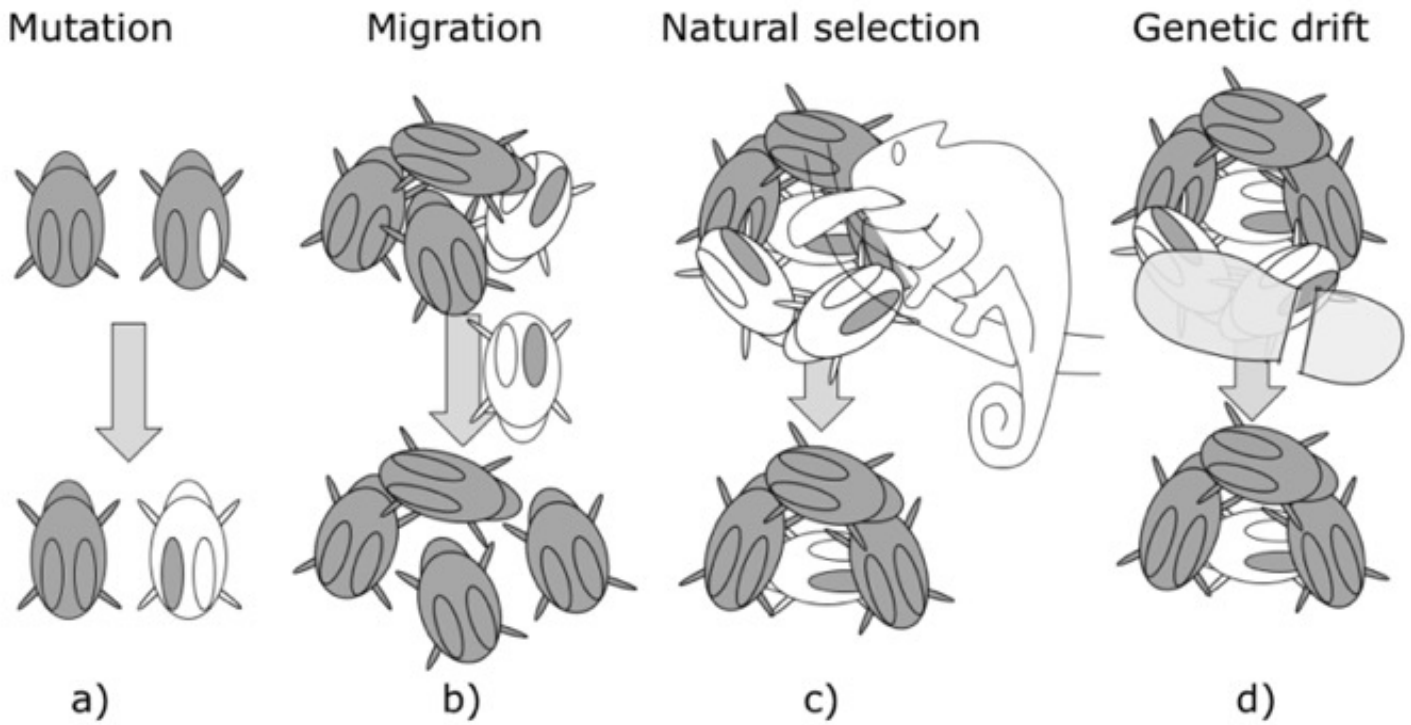
2022 Philippines PIRE ‘Omics Workshop

Dumaguete, Philippines

photo: earth.com

Four fundamental forces that drive evolution

- Mutation
- Migration
- Natural selection
- **Genetic drift**



Population size affects all of these forces!

Figure from Schlieter et al. 2019

Genetic drift

- Random change in frequency of genes over time
- Basically a random sampling effect

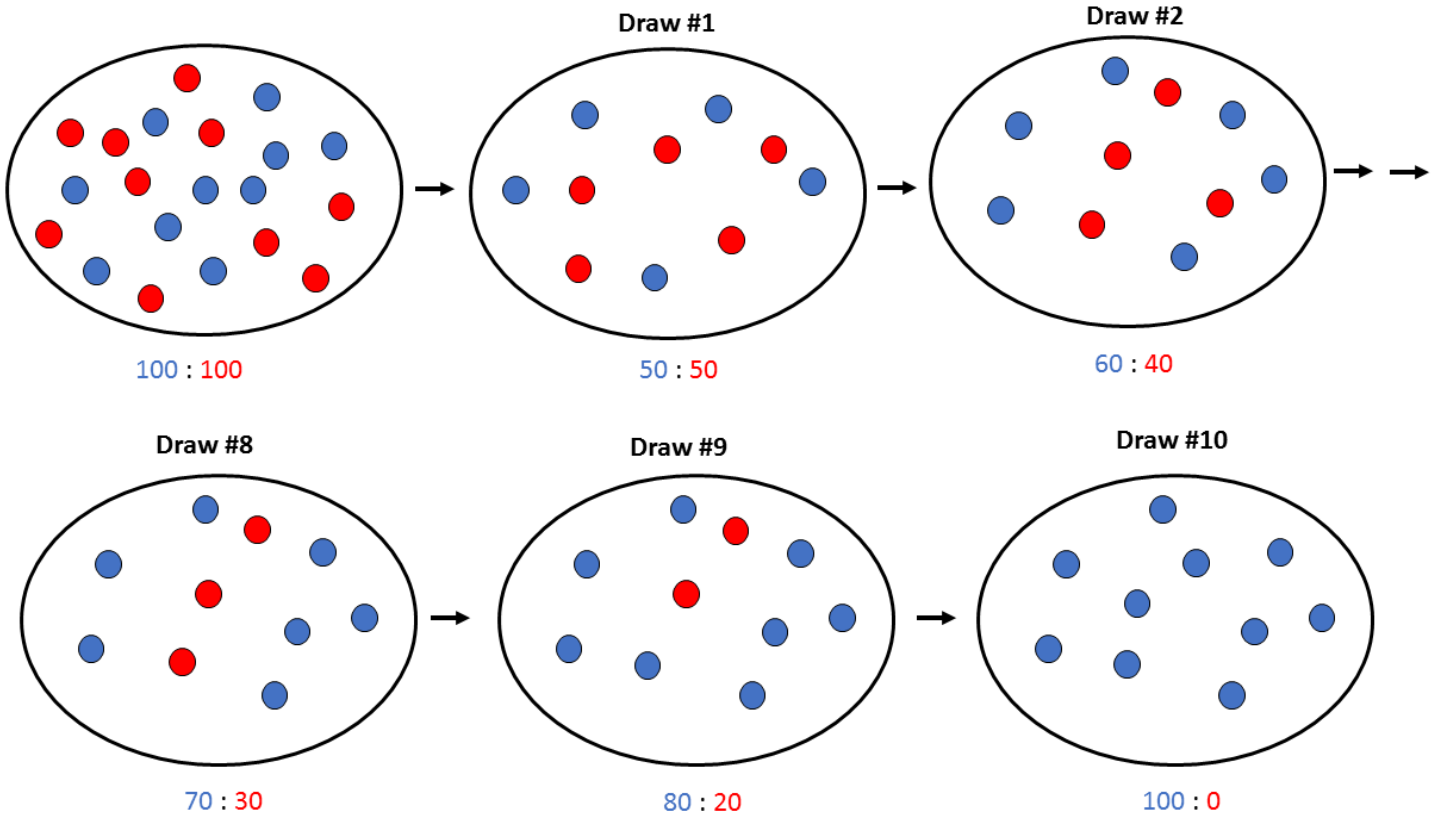


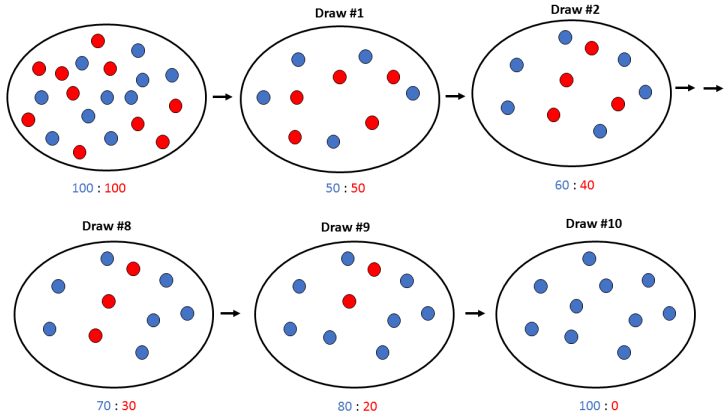
Figure from steemit.com

Genetic drift and population size

- Would variance in allele frequency over time be larger in a population of 10,000 individuals or 100 individuals?

Smaller population size =
higher variance from generation to generation =
more + faster genetic drift!

Infinite population size = no drift!

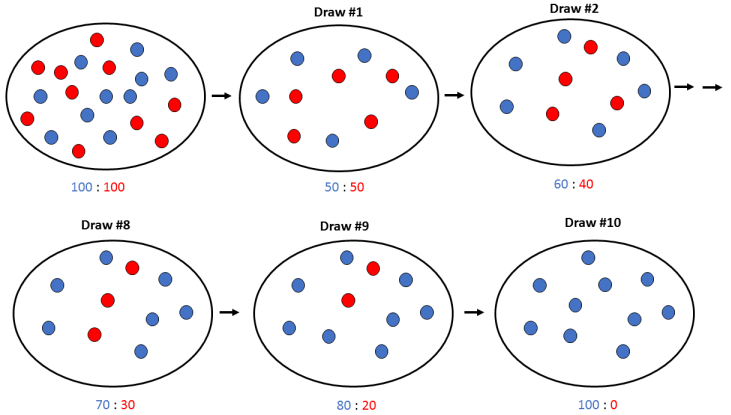


The pace of genetic drift

Time to fixation for an allele with frequency *p*:
$E(T) = -4N(p \ln p + (1-p) \ln(1-p))$ generations

What's the expected fixation time
(in generations) when:

N = 1000, *p* = 0.5?	~2770 generations
N = 1000, *p* = 0.1?	~1300 generations
N = 100, *p* = 0.5?	~277 generations
N = 100, *p* = 0.1?	~130 generations



Change in heterozygosity over time: $H_t = (1 - 1/2N)^t H_o$

How fast is heterozygosity lost?

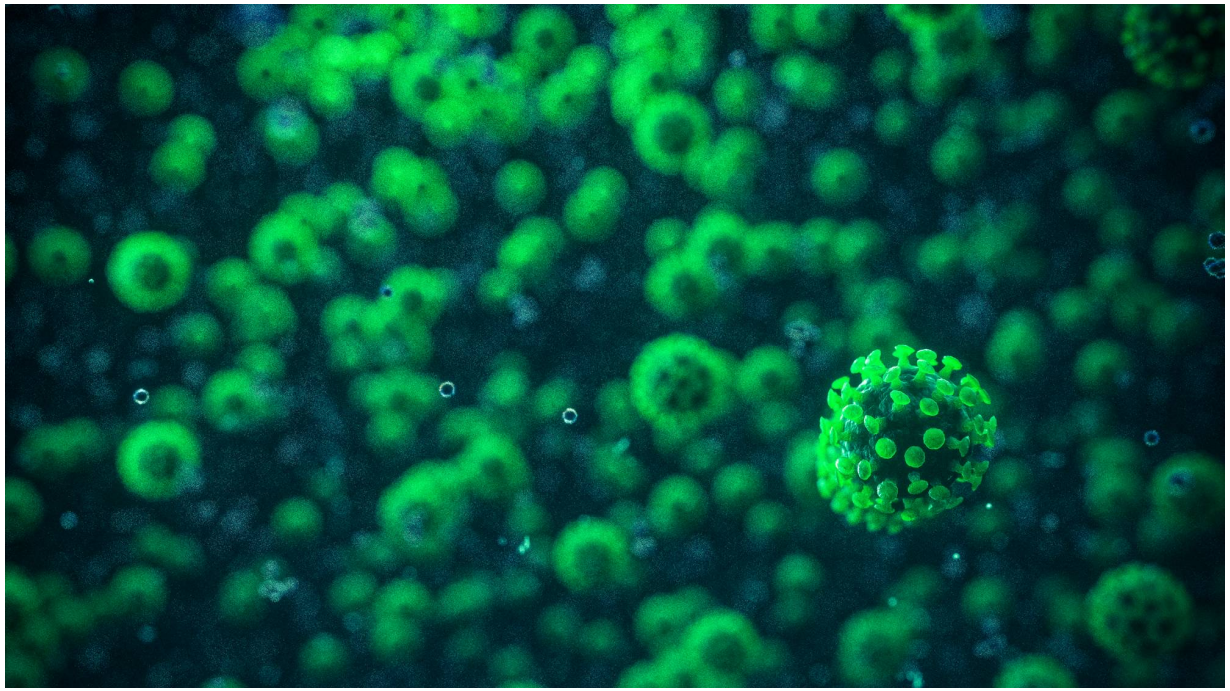
Why does any population have genetic diversity if it's constantly being lost to drift??

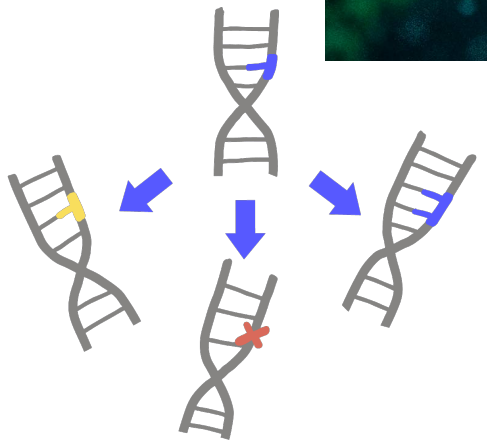
Mutation and population size

Random changes in the genome that generate diversity

Unicellular organisms/viruses: mutations per replication

Eukaryotes: mutations per meiosis





Figures: Johns Hopkins, Wikipedia

Population size is extremely important in conservation biology too!

- Probability of extirpation/extinction
- Genetic diversity and adaptive potential
- Identifying trends in population size is important for setting management priorities

Photos from Princeton, Wikipedia

“Ideal” populations vs real populations

- “Ideal” (Wright-Fisher) populations have:
  - Equal sex ratio
  - Equal reproductive success among individuals
  - Discrete generations
  - All individuals equally likely to mate with every other individual (“panmixia”)
  - No migration from outside
  - Neutral evolution (no selection)
- Real populations have:
  - Variance in sex ratios and reproductive success
  - Assortative mating
  - Overlapping generations
  - Population structure
  - Selection

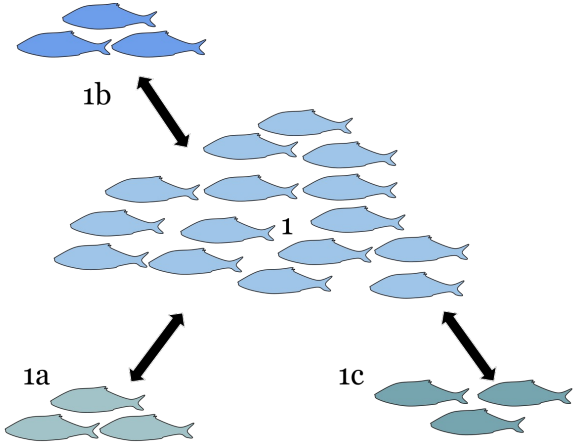
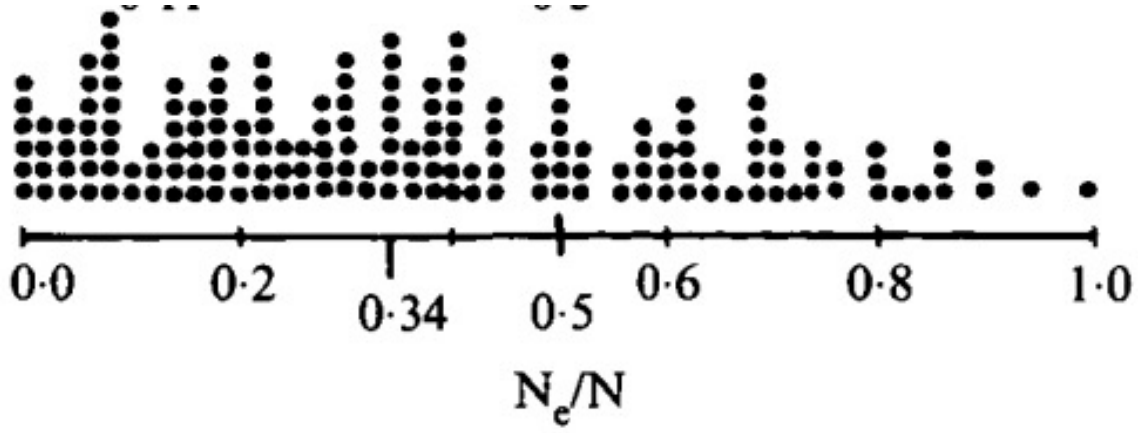


Figure from Wikipedia

Effective population size (N_e)

- Size of an “ideal” population that would experience genetic drift at the same rate as a real population of interest
- Usually smaller than census population size (N_c)
  - Frankham 1995
    - Review of Ne/N across 100 species of animals and plants
    - Average Ne/N = 10%
  - Waples 2002
    - Ne/N = 20%
  - Palstra & Ruzzante 2008
    - Ne/N = 14%



Marine species… an exception?

- Very high fecundity, high mortality early in life
  - Type III survivorship
- Very high variance in reproductive success?
- Estimates of Ne from 10-10,000
  - but estimates of N often in the millions
- Ne/N from 10^{-3} to 10^{-5}
- Remain controversial

Relationship between N_e and genetic diversity

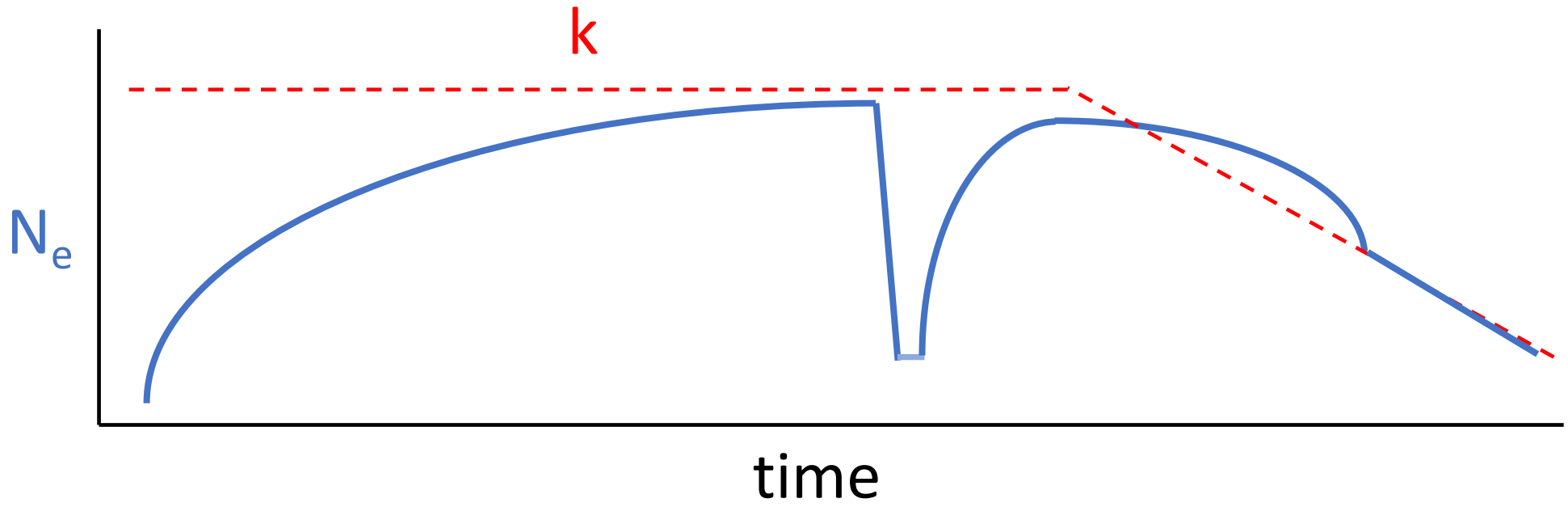
- At a long-term stable state, genetic diversity has a predictable relationship to genetic diversity
- Represents equilibrium between loss of diversity (through drift) and gain of diversity (through mutation)
- Watterson (+ Wu) 1975
- Expected number of polymorphic sites (Θ) related to “population mutation rate”

$\Theta = 4N_e\mu$

ya I know dude

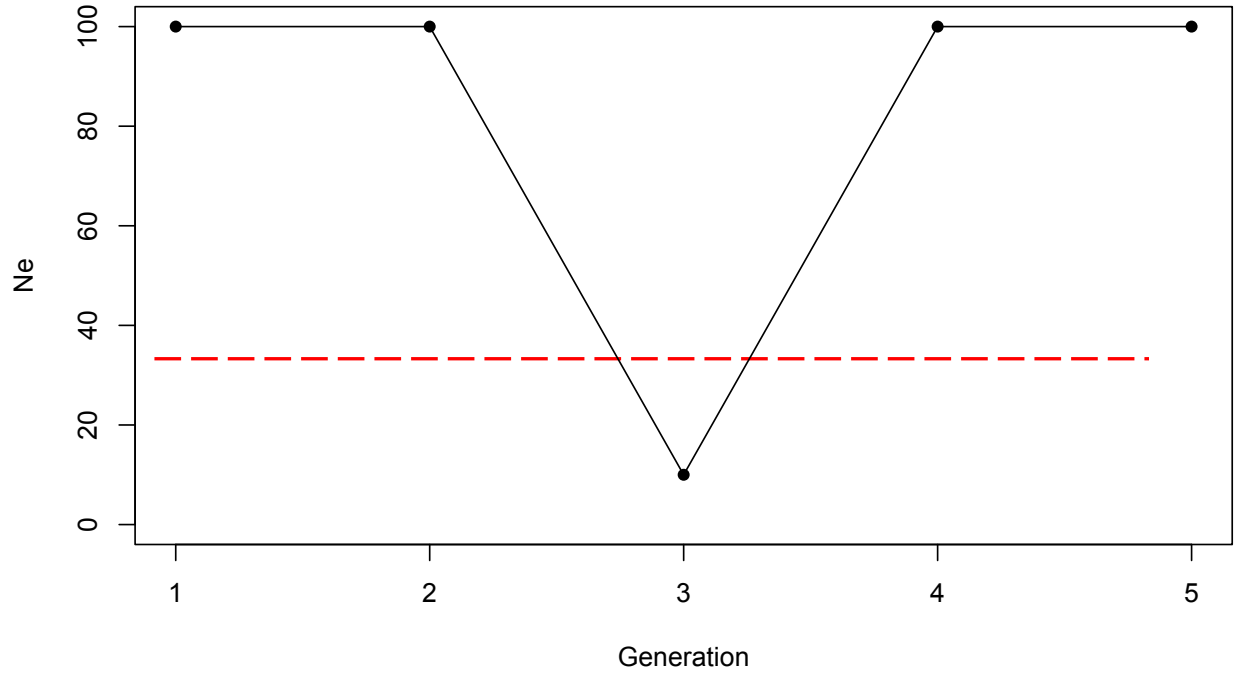
Populations can be stable or dynamic

- Equilibrium / at carrying capacity (k)
- Expansions
- Bottlenecks
- Declines



Effects of size change on N_e

- Long-term average of N_e can be calculated as the harmonic mean (reciprocal of average of reciprocals) of N_e in each generation
  - For this bottleneck, N_e = 1/(((1/100)+(1/100)+(1/10)+(1/100)+1/100))/5)
- Periods of low N_e affect long-term average more than periods of high N_e



Genetic consequences of size change

- Alterations to the site frequency spectrum (SFS)
- Bottlenecks cause reductions in diversity
  - Loss of rare alleles
- Expansions can cause increases in diversity
  - Excess of rare alleles

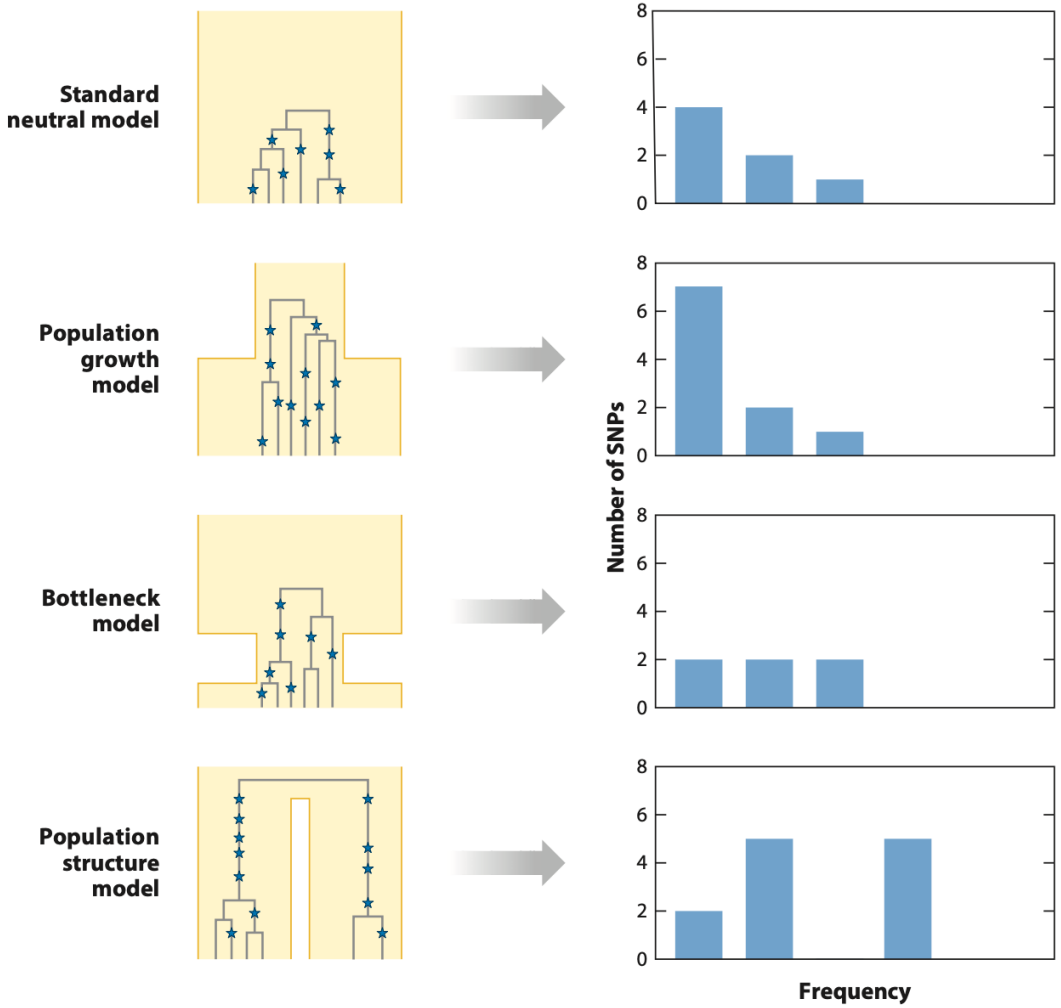


Figure from Beichman et al. 2018

Many methods used to estimate (changes in) N_e from genetic data

- SFS-based inference
- Linkage disequilibrium
- Approximate Bayesian computation (ABC)
- **Coalescence-based methods**
- Differing data requirements
- Differing resolution at various time scales and different ability to detect change

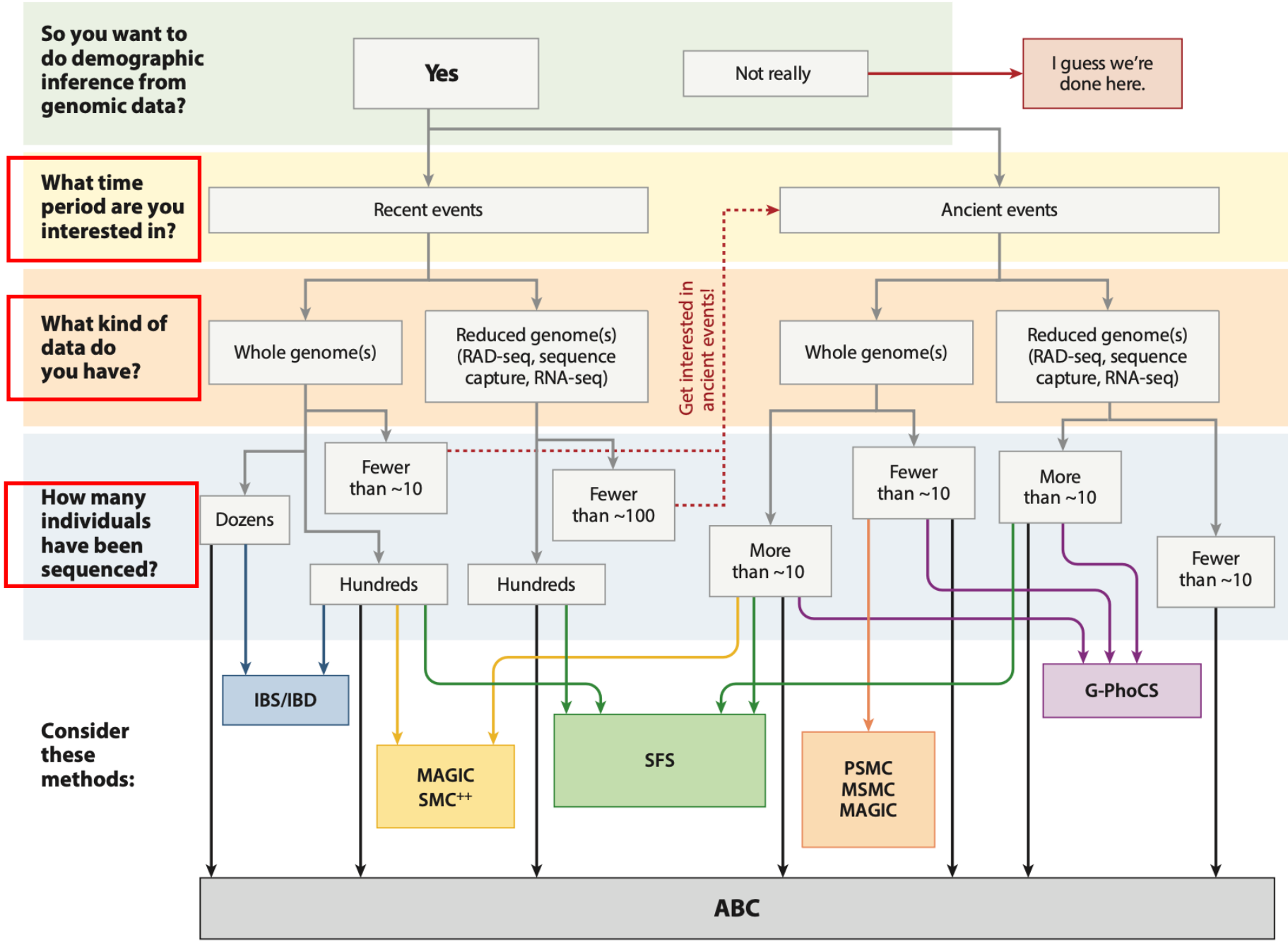


Figure from Beichman et al. 2018

Site frequency spectrum (e.g. momi2)

- Generate by simulation or use approximations
- Compare observed SFS to expected SFS under different demographic scenarios
- Very flexible, can incorporate multiple populations and migration
- Need many individuals to detect recent declines

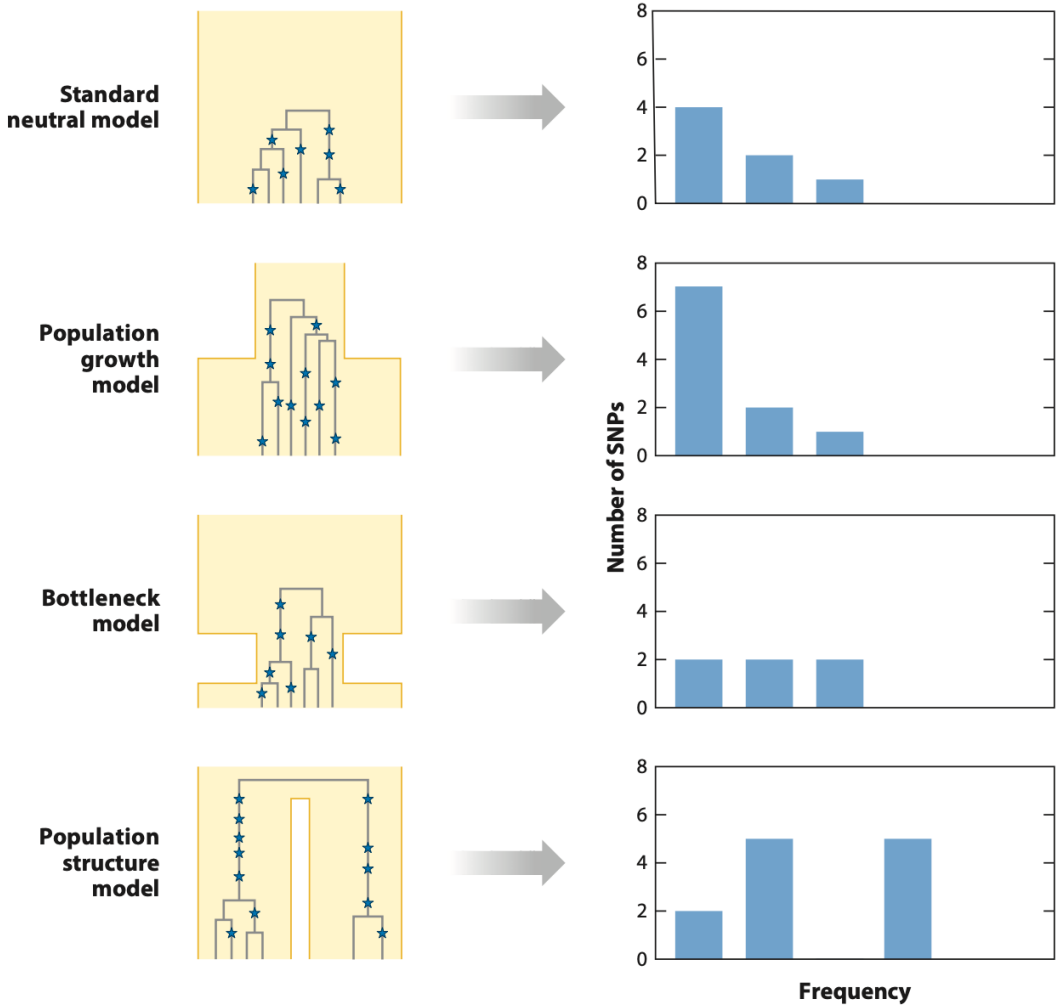
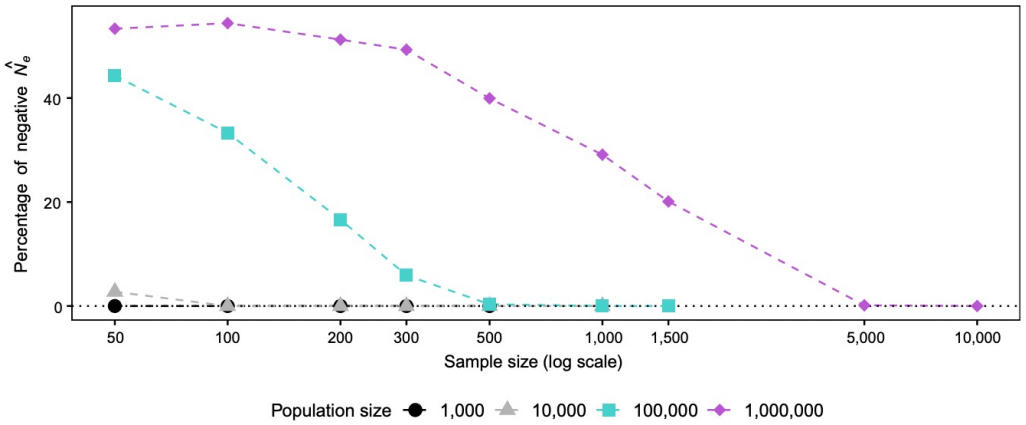


Figure from Beichman et al. 2018

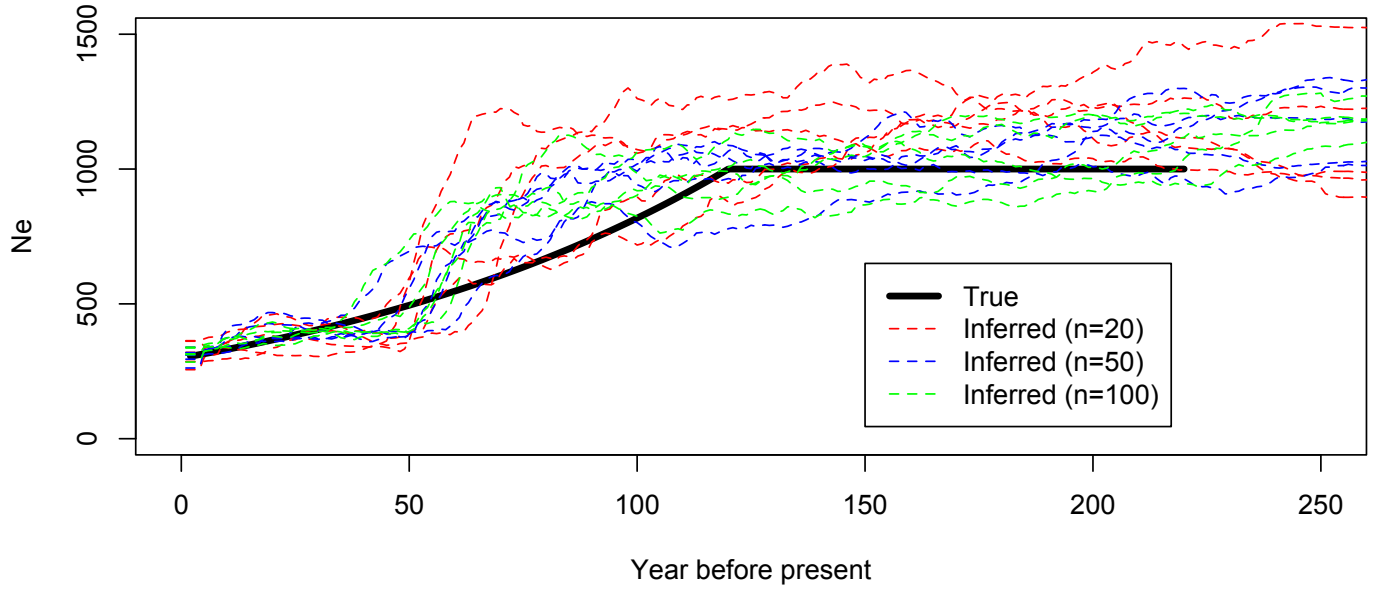
Linkage disequilibrium (e.g. NeEstimator)

- Linkage *equilibrium* = alleles on different chromosomes are randomly associated / at Hardy-Weinberg Equilibrium
- r^2 should = 0, *but* only when the population size is infinite
- Smaller populations will start to show nonrandom associations between alleles ($r^2 > 0$)
- Point estimate of population size at time of sampling
- Requires large sample sizes (10% of N_e / 1% of N_c)



Linkage disequilibrium (e.g. GONE/SNeP)

- Use pattern of linkage *within* chromosomes to infer changes in population size
- Need a good reference genome
- Smaller sample sizes, accurate for recent time points



Approximate Bayesian Computation

- Compare simulated summary statistics to observed statistics
- Can customize the summary statistics and models used
- Sensitive to model mis-specification

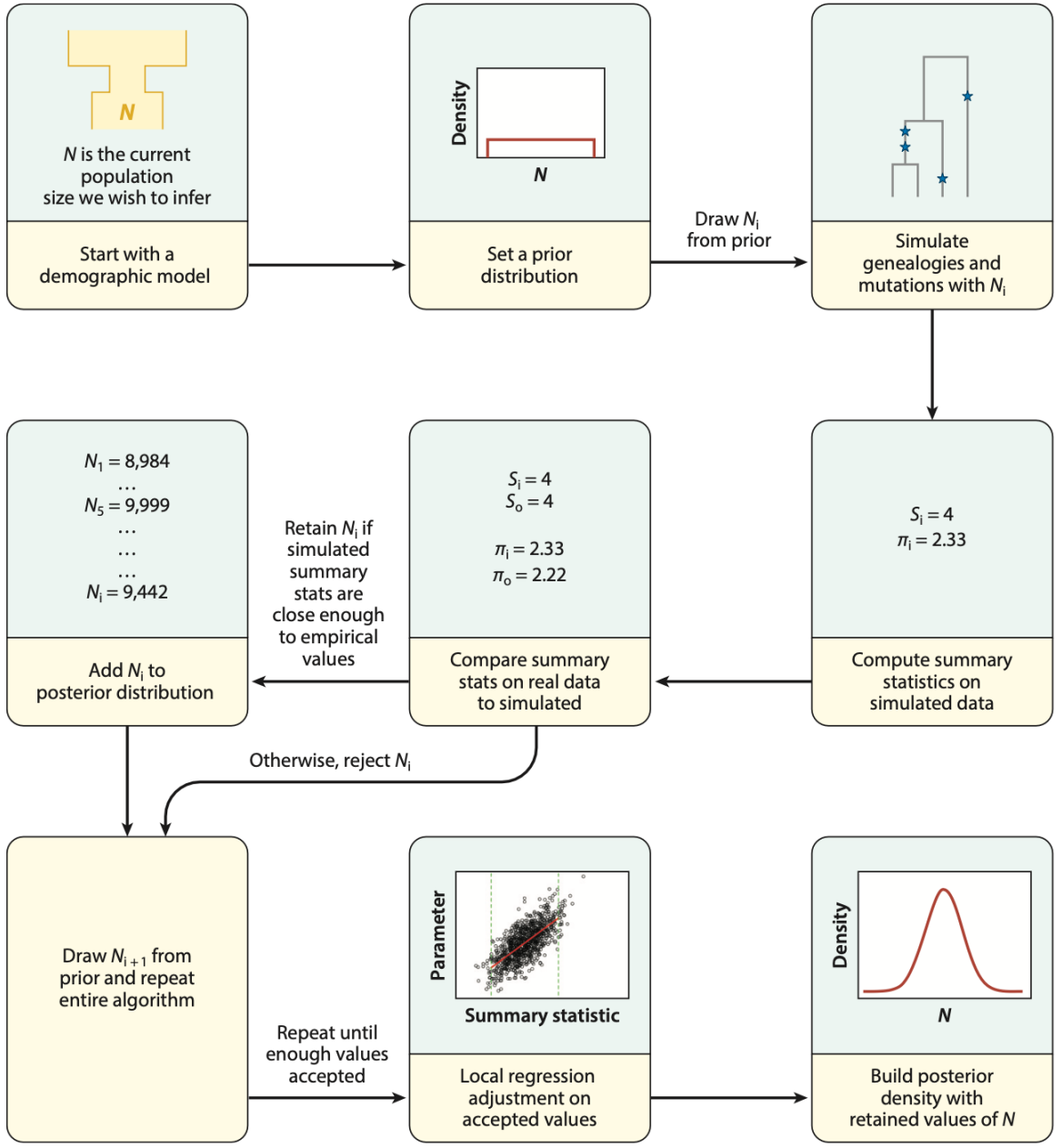


Figure from Beichman et al. 2018

Coalescent theory

- Relates population size and genealogy
- Probability that two gene copies sampled in the present shared a common ancestor in the previous generation = 1/N
- Average time for two gene copies t coalesce (TMRCA) =$2N_e$ generation
  - However, different genes and different sets of individuals can have different TMRCAs!

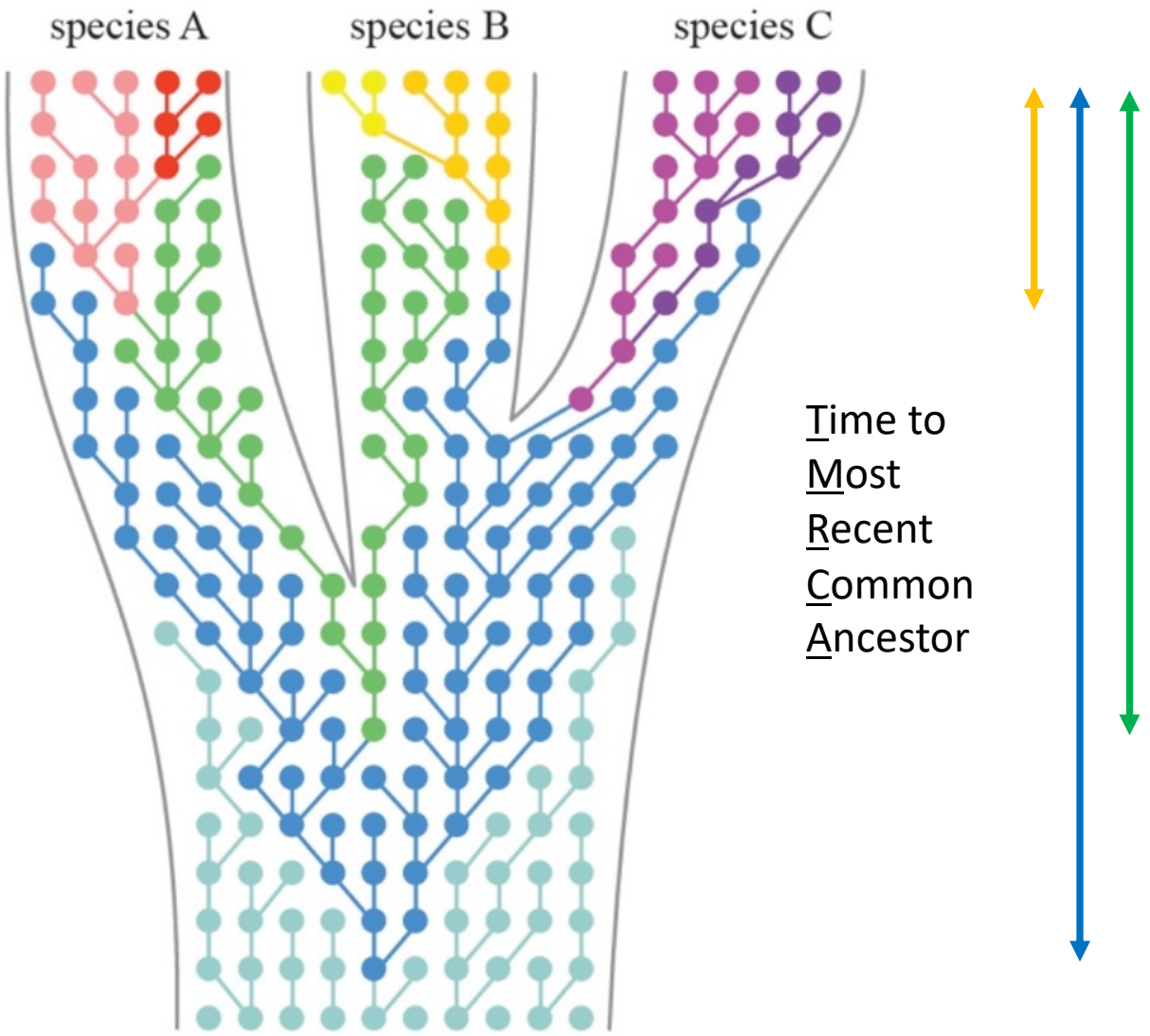


Figure from Leliaert et al. 2014

Advantages of coalescent theory

- Well-developed theoretical expectations
- Thinking backwards (from our sampled individuals) is often easier and less computationally intensive than thinking forwards (from a common ancestor)

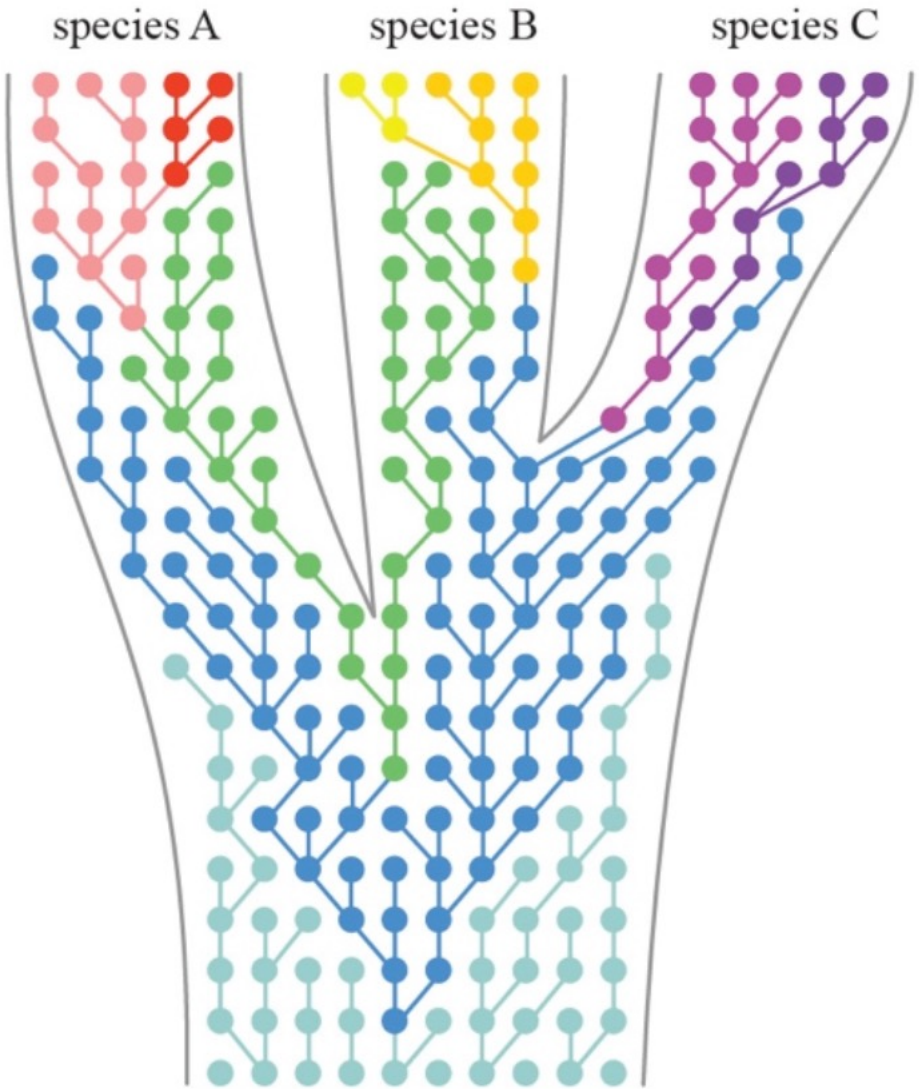


Figure from Leliaert et al. 2014

Coalescent methods

- Make inferences about changes in population size over time based on genealogical patterns in the data
- Can use many individuals to accurately reconstruct the tree(s)
  - Bayesian skyline plot
  - IMA3
- Can use many different genes for a few individuals
  - **Sequentially Markovian methods**

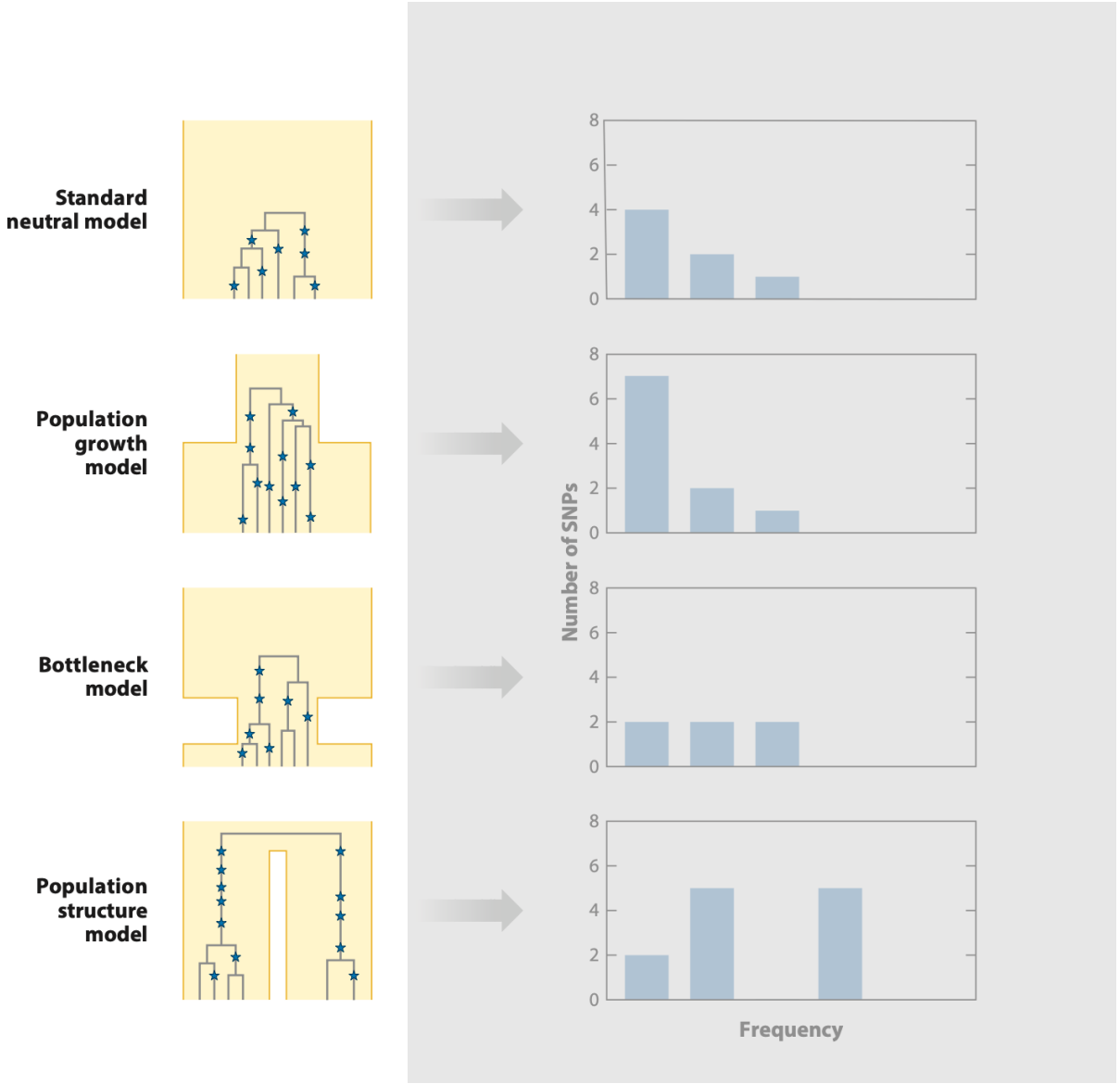


Figure from Beichman et al. 2018

LETTER

doi:10.1038/nature10231

Inference of human population history from individual whole-genome sequences

Heng Li[1,2] & Richard Durbin[1]

Received: 28 May 2019 | Revised: 11 October 2019 | Accepted: 12 November 2019

DOI: 10.1002/ece3.5888

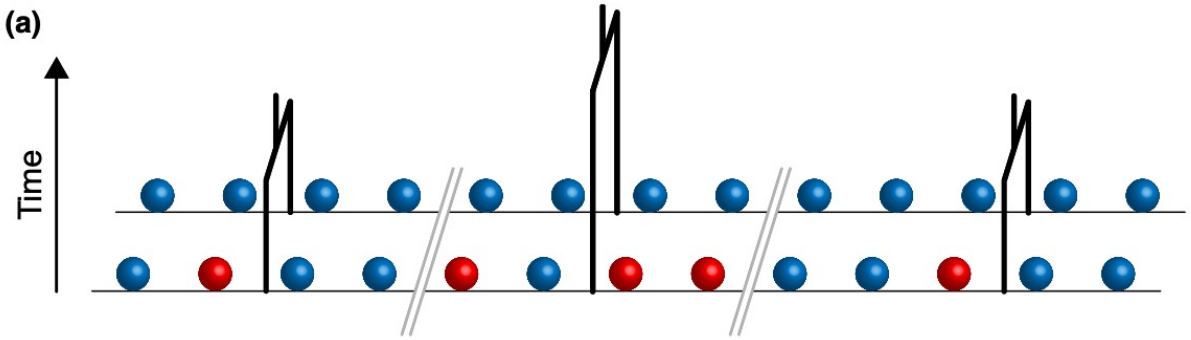
REVIEW ARTICLE

Ecology and Evolution Open Access WILEY

A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data

Niklas Mather | Samuel M. Traves | Simon Y. W. Ho

PSMC input data: genome sequence from a single diploid individual

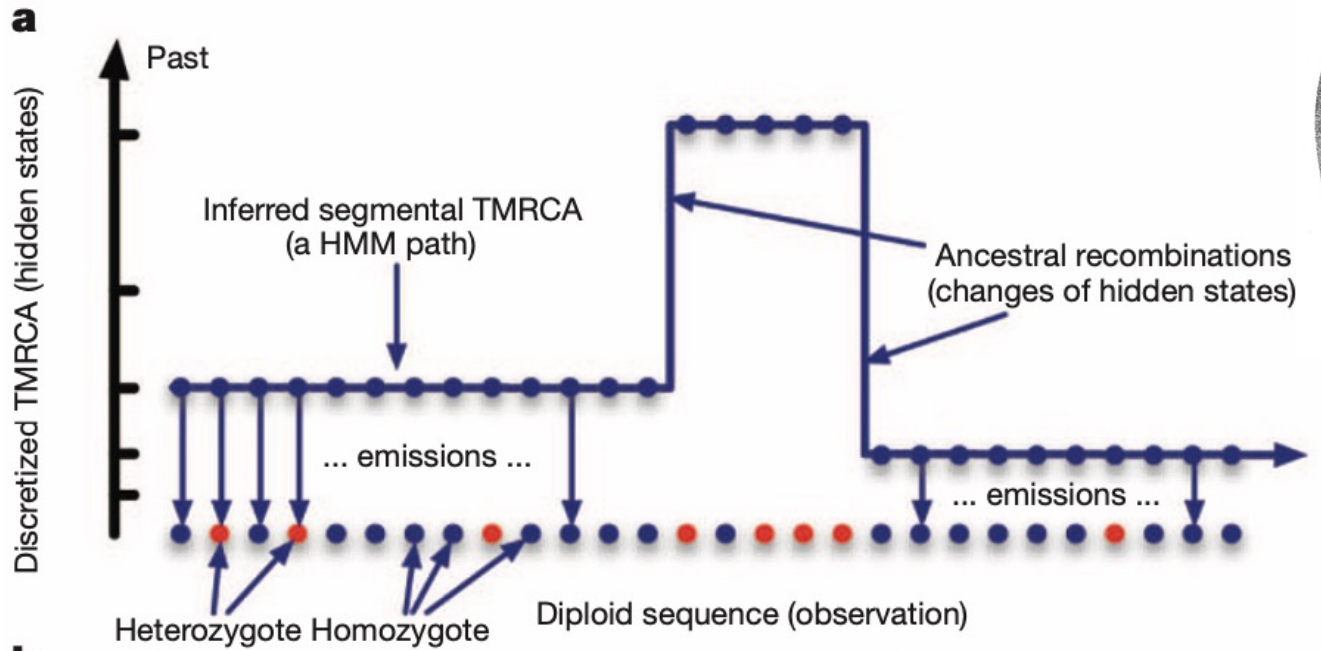


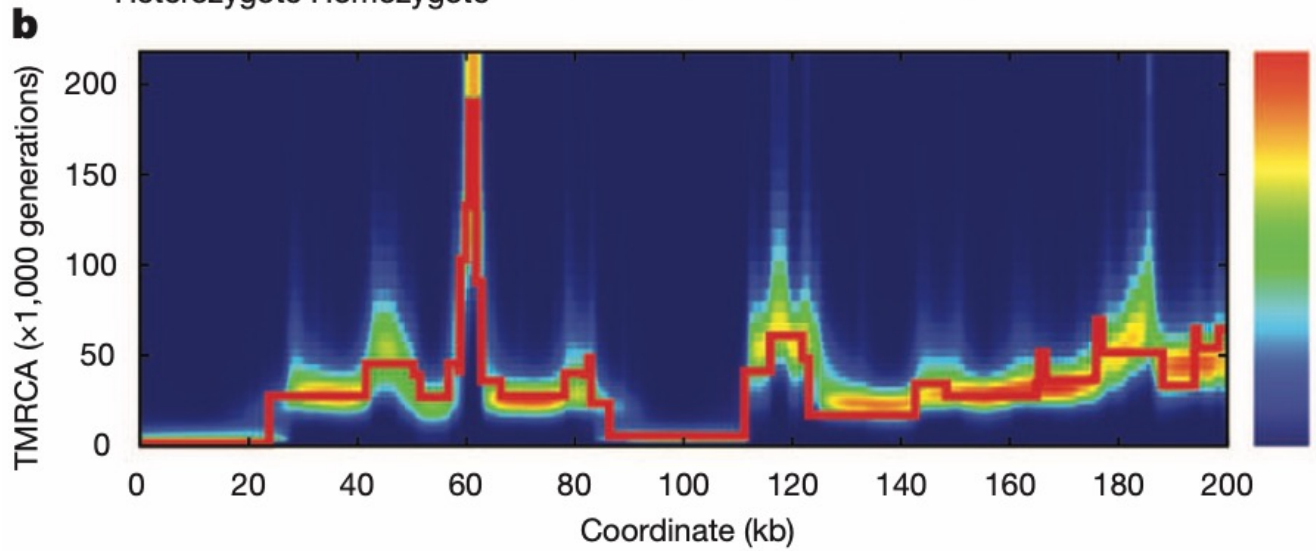
Walking a “path” through the genome

The “landscape” of that path is the pattern of heterozygosity we see within the individual and was influenced by the pattern of coalescence

The landscape is determined by a few parameters:

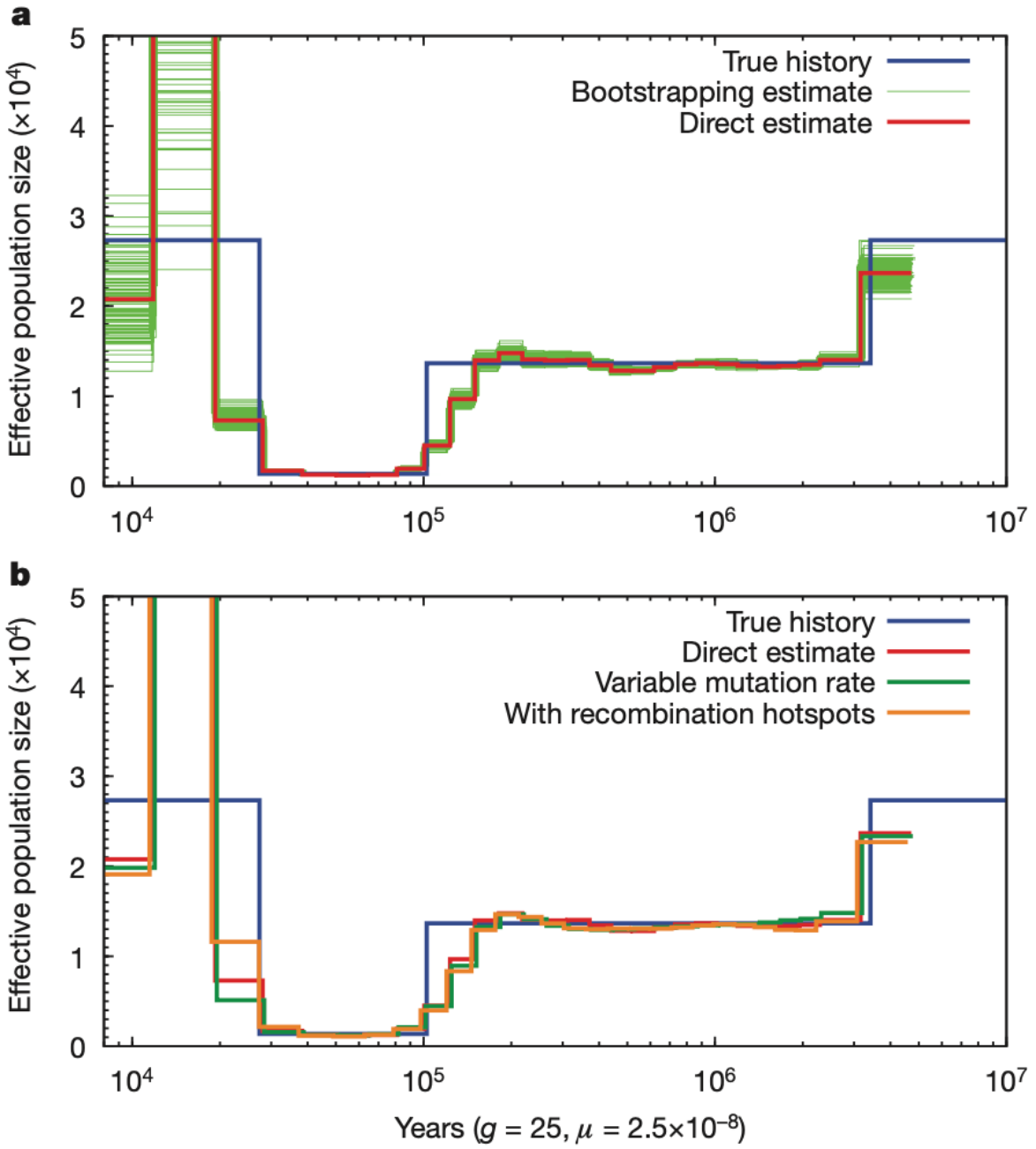
1) Θ_0 (scaled mutation rate)
2) ρ_0 (scaled recombination rate)
3) $\lambda(t)$s (relative population sizes)





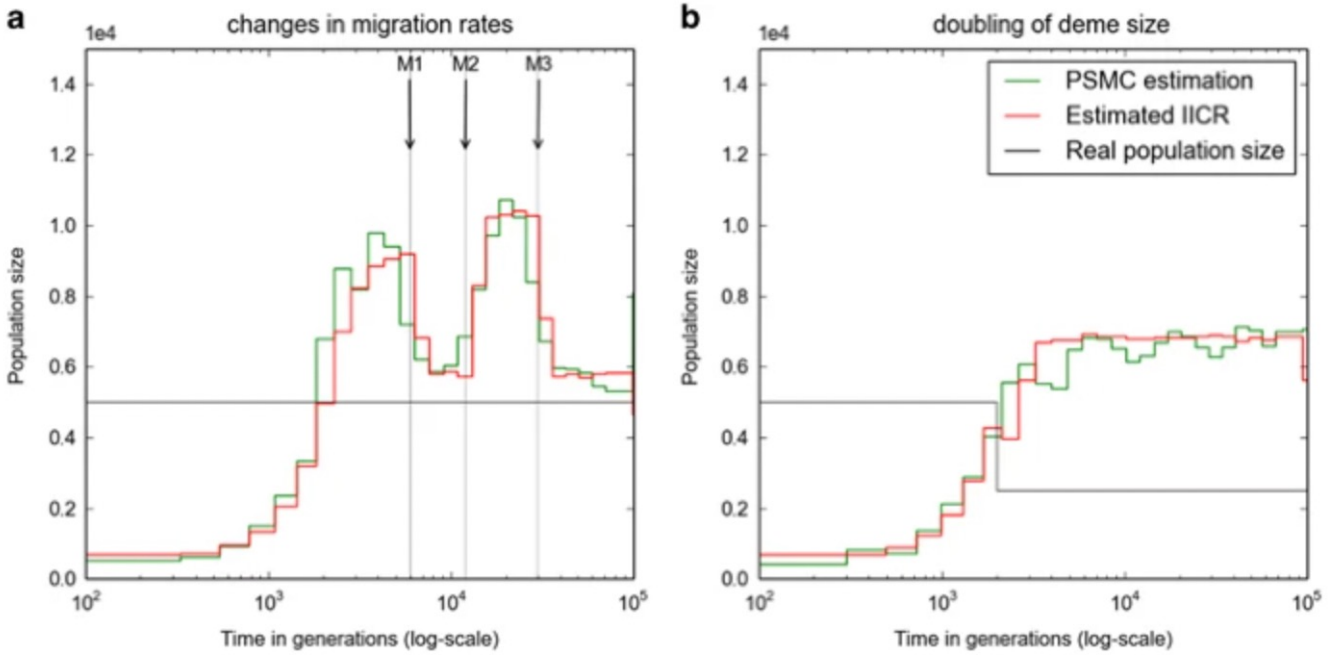
Advantages of PSMC

- Works well with small sample sizes (one individual!)
- Can work even without full reference genomes or whole-genome data
- Can infer complex population histories
- Robust to variation in mutation/recombination rate



Limitations of PSMC

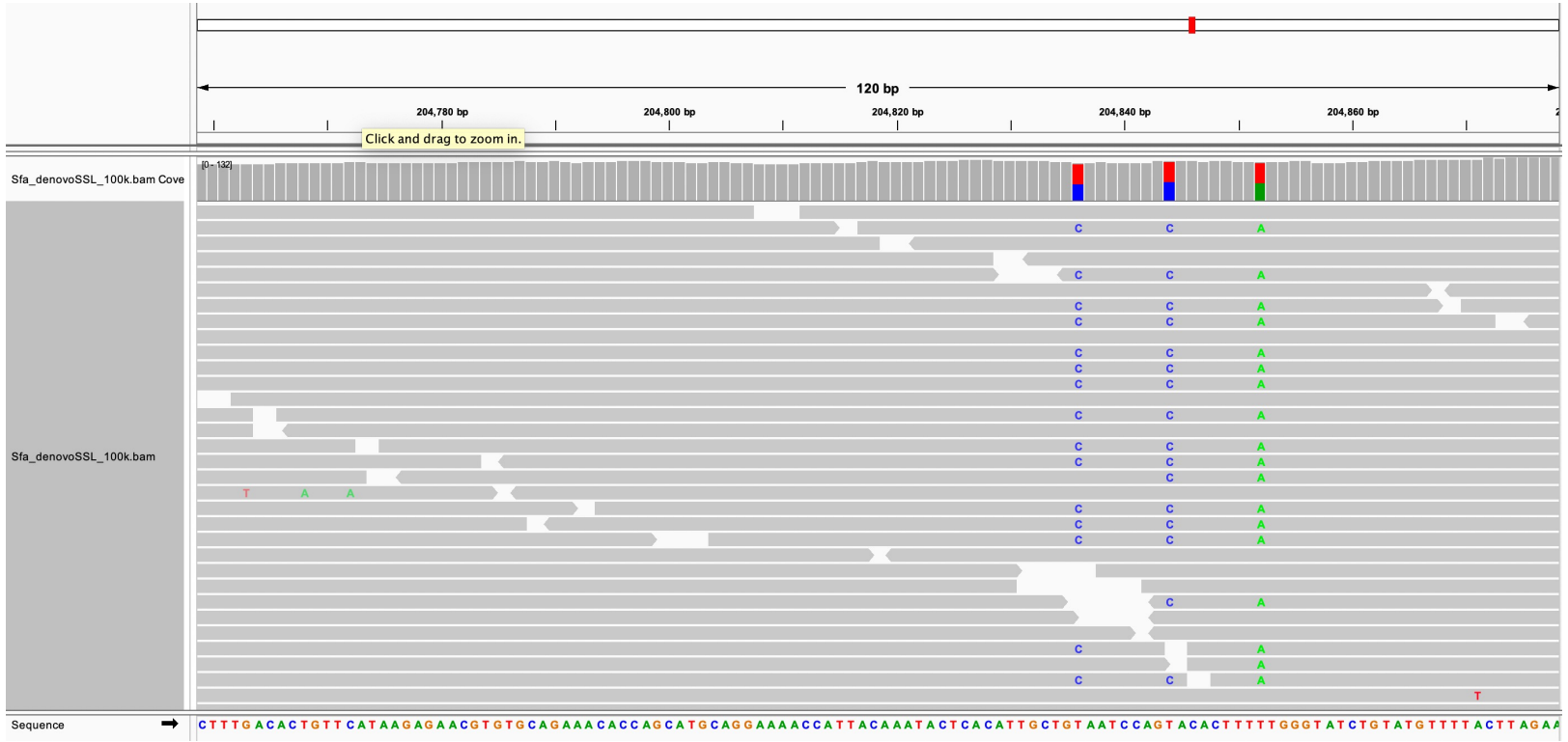
- Limited accuracy for the recent past
  - How recent?
- Need mutation rate/generation time to accurately scale estimates
- Actually measures distribution of coalescence rates
  - For an ideal population this will track size closely
  - Population structure/migration and selection also influence the distribution of coalescent events!



Mazet et al. 2016

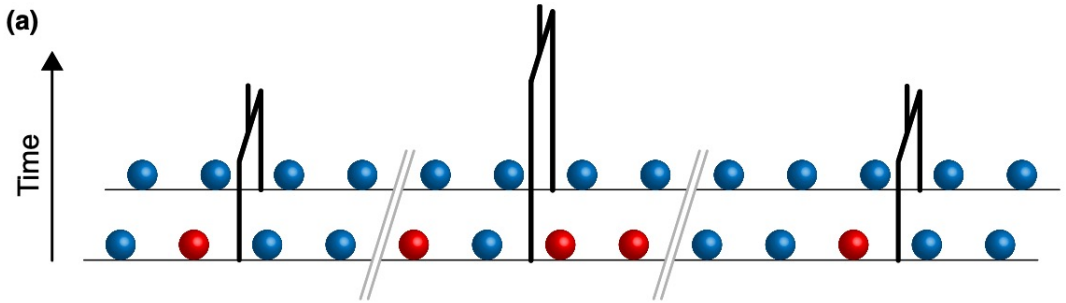
Workshop exercise

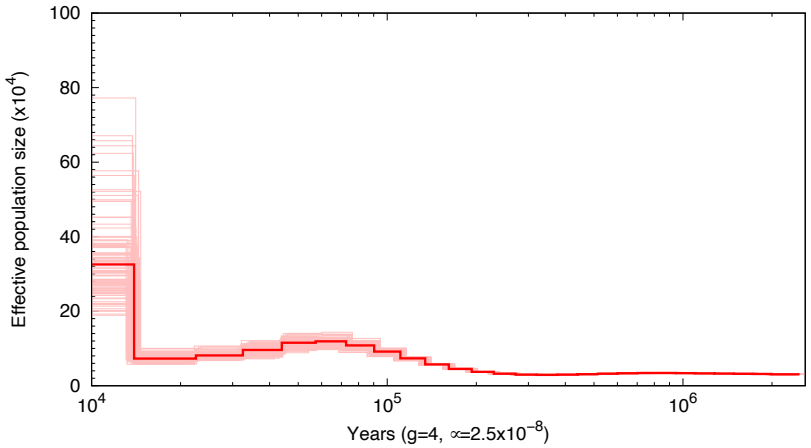
- Use shotgun sequencing data from *S. fasciatus*, combined with our *de novo* genome and a reference genome from Genbank to conduct a PSMC analysis!
  - Prepare reference genomes
  - Map reads to genomes
  - Estimate coverage
  - Call consensus sequences
  - Convert to PSMC format
  - Run PSMC
  - Estimate confidence intervals



Workshop exercise

- Use shotgun sequencing data from *S. fasciatus*, combined with our *de novo* genome and a reference genome from Genbank to conduct a PSMC analysis!
  - Prepare reference genomes
  - Map reads to genomes
  - Estimate coverage
  - Call consensus sequences
  - Convert to PSMC format
  - Run PSMC
  - Estimate confidence intervals





Questions?

Acknowledgments

- Malin Pinsky
- Liz Wallace
- Philippines PIRE Project
- Silliman University
- YOU!