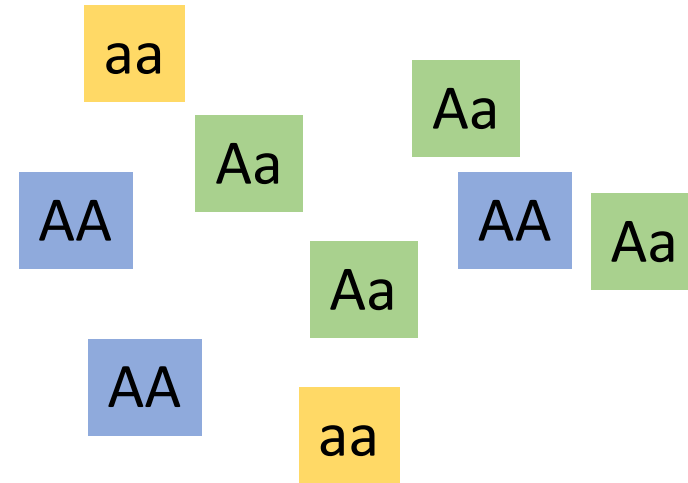# Drift, mutation, the coalescent

Week 2

- How and why do allele frequencies change through time?
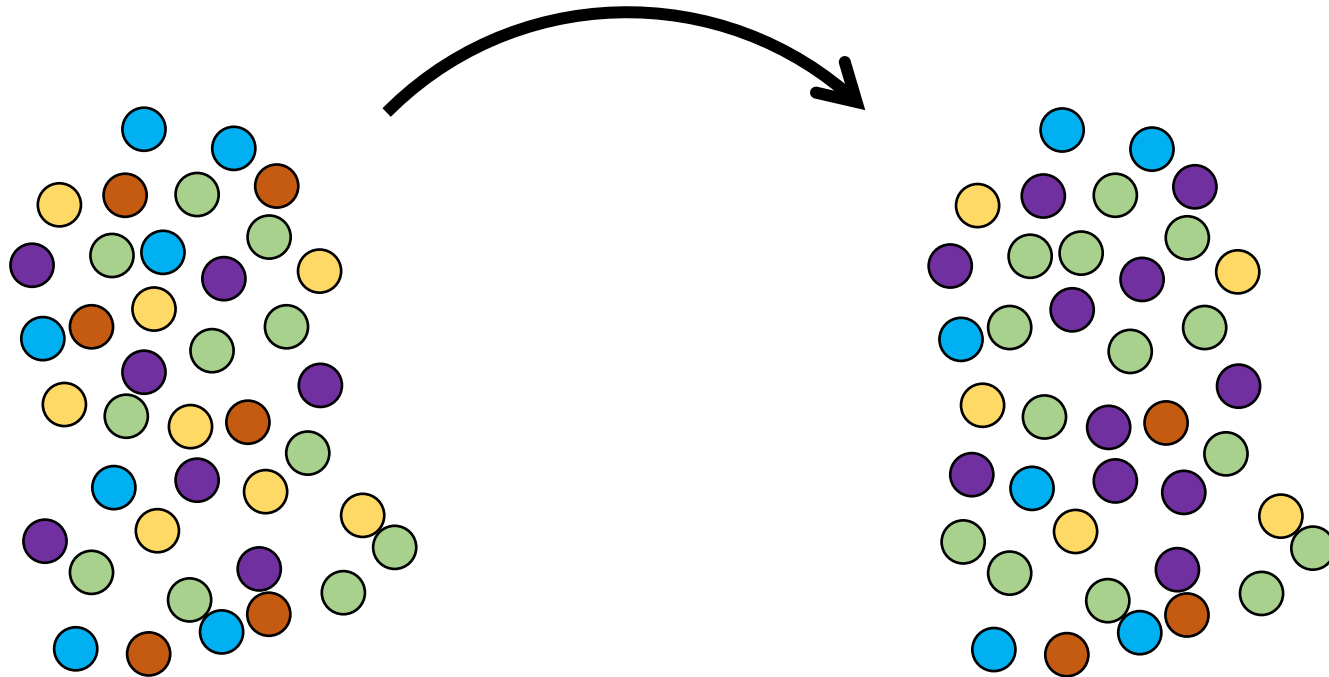
- Two main factors:
  - Drift
  - Selection

Wright-Fisher Model

- Gene copies are randomly sampled from 1 generation to the next

- Think of a bag of marbles

Wright-Fisher Model

- Gene copies are randomly sampled from 1 generation to the next

- What is the expected allele frequency in the next generation?
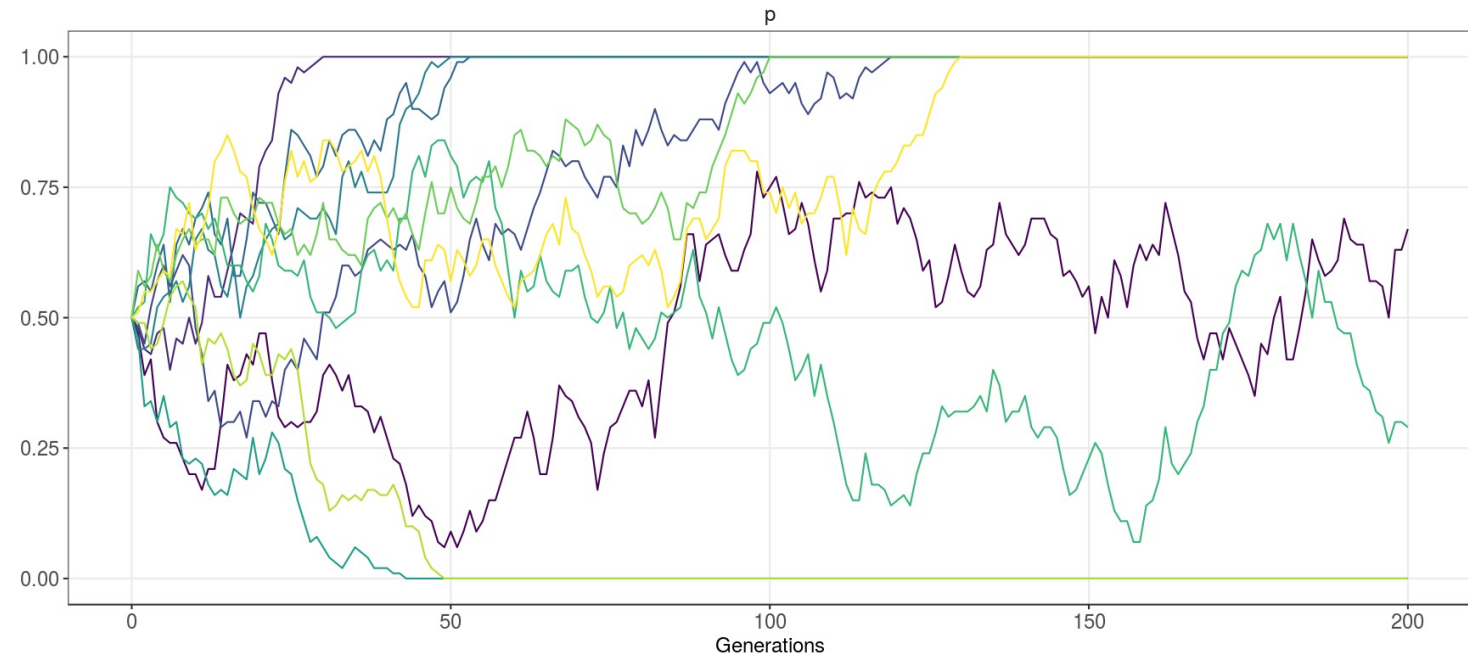
$$E[f_{A(t+1)}] = \frac{(2N)(f_A)}{2N} = fA$$

Expected allele frequencies are equal across generations

Wright-Fisher Model

- Gene copies are randomly sampled from 1 generation to the next
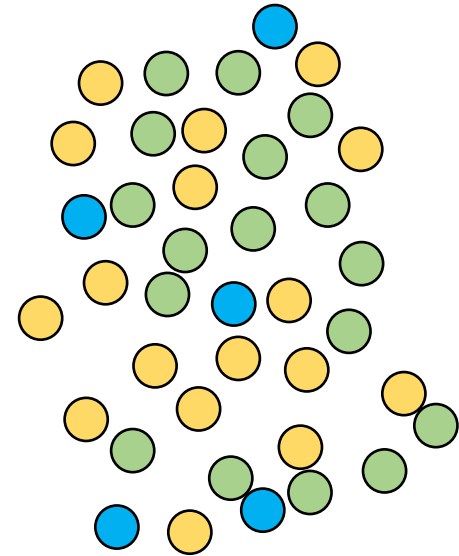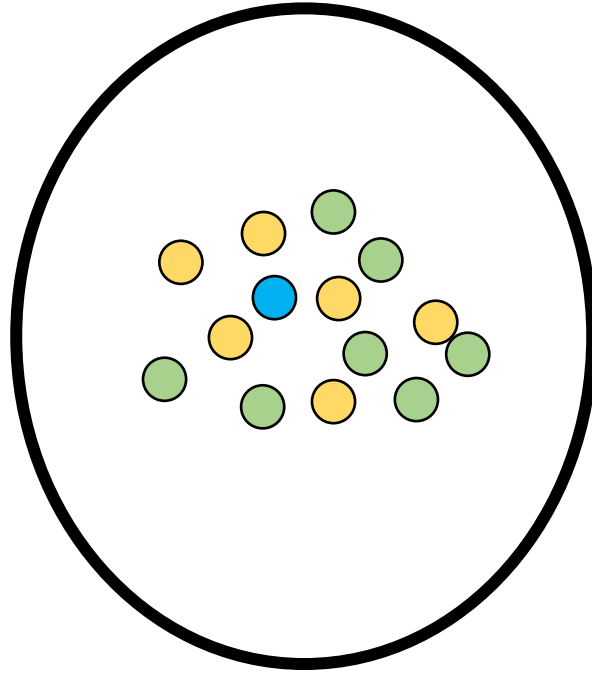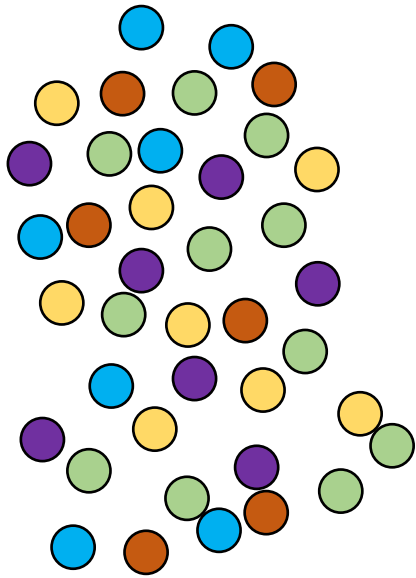
DRIFT

- Think about flipping a coin
- We have finite populations, so alleles are randomly fixed or lost.
- All alleles will eventually be fixed or lost

Effective population size (N$_e$)

- The number of individuals that would produce the same amount of drift as the observed population
- In contrast to the census size

Northern elephant seal

Michael L. Baird

~20 individuals

2000

1990s {
A-A-G (1) : $N = 107$ seals
G-T-A (2) : $N = 42$ seals

1980 {
A-A-G (1) : BN-6
G-T-A (2) : BN-10

1965 — A-A-G (1) : BN-5

1914 {
A-A-G (1) : BN-23
A-A-G (1) : BN-24
G-T-A (2) : BN-25

1892
Known demographic
bottleneck {
G-T-A (2) : TS-1
G-T-A (2) : TS-2

ca. 1800 {
A-A-G (1) : BN-20
G-A-A (3) : BN-26
G-A-A (3) : BN-7
A-A-A (4) : BN-22

ca. 1000 — G-A-G (5) : BN-29

Current Biology

An empirical genetic assessment of the severity of the northern elephant seal population bottleneck

Diana S. Weber*, Brent S. Stewart[†‡], J. Carlos Garza[¶] and Niles Lehman*

Abadía-Cardoso et al 2017

Effective population size (N_e)

- The number of individuals that would produce the same amount of drift as the observed population
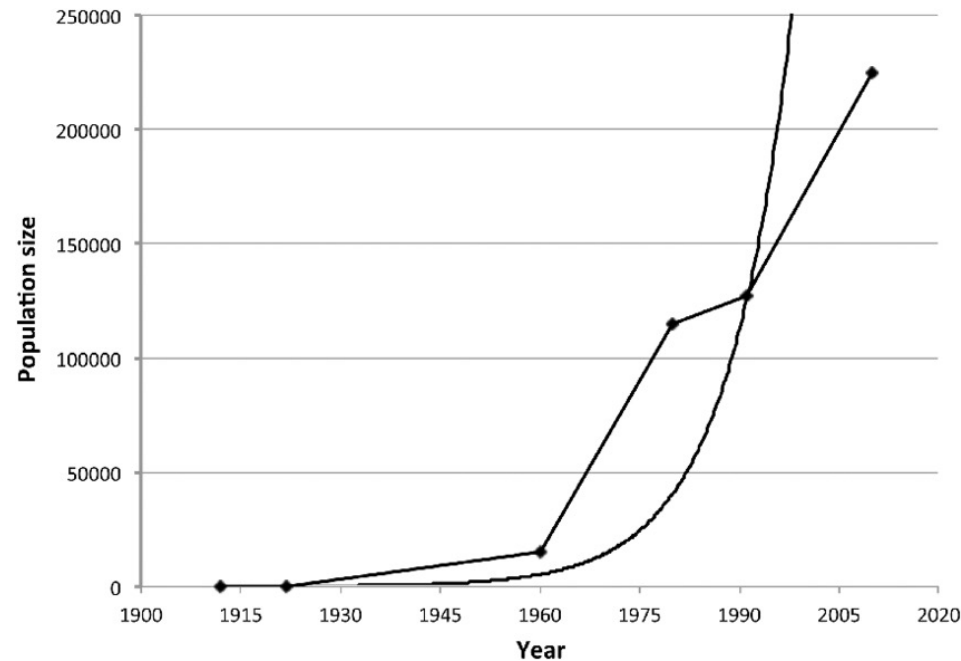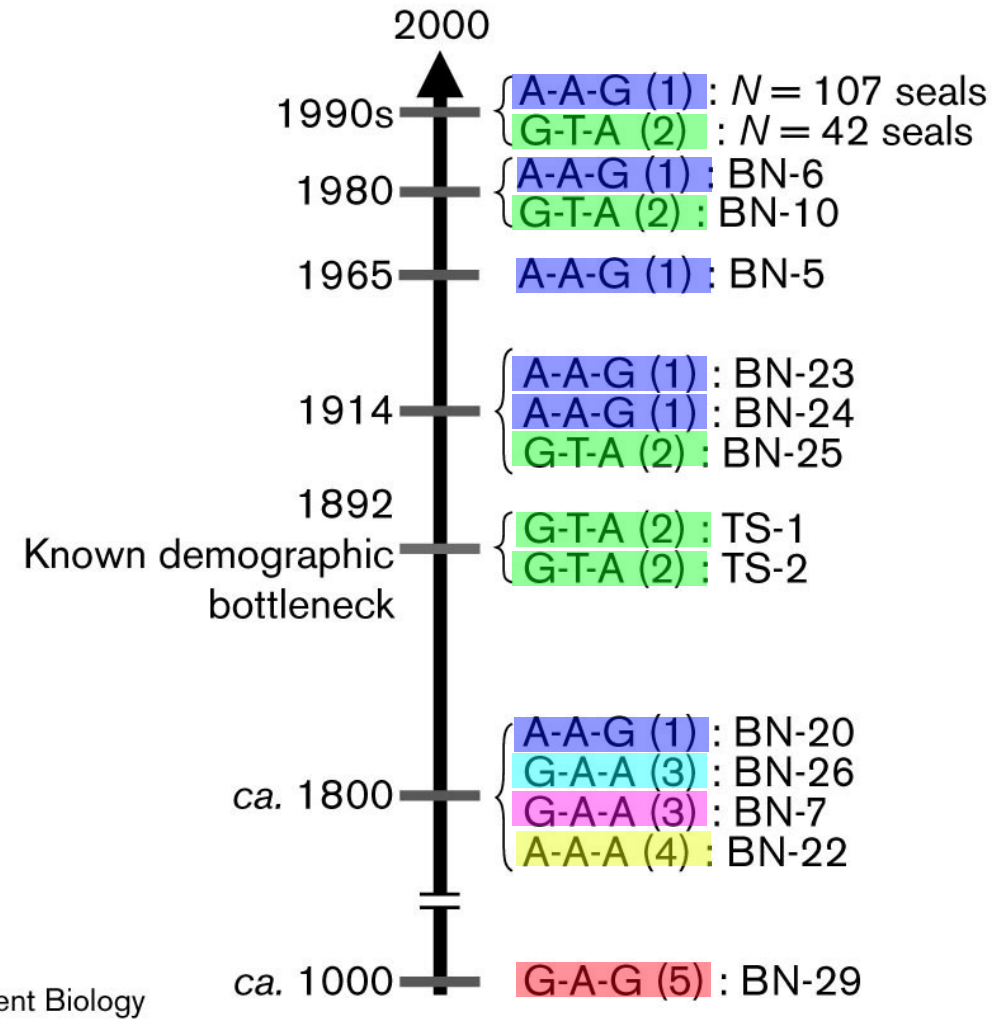- In contrast to the census size

Across 3 generations, population size of: 100, 10, 100

What is N_e?

Harmonic mean $\quad \dfrac{n}{\sum_{i=1}^{t} \dfrac{1}{N_i}} = \dfrac{3}{\left[\dfrac{1}{100} + \dfrac{1}{10} + \dfrac{1}{100}\right]}$

$N_e = 25$

Arithmetic mean = 70

Effective population size (N$_e$)

- What can influence N$_e$?
  - Change in population size; variance in reproductive success; unequal sex ratio…

What else can affect N_e?

Uneven breeding sex ratios (the less frequent sex can be seen as an allelic bottleneck)

$$N_e = 4 \frac{N_m N_f}{N_m + N_f}$$

N_m: number of breeding males
N_f: number of breeding females
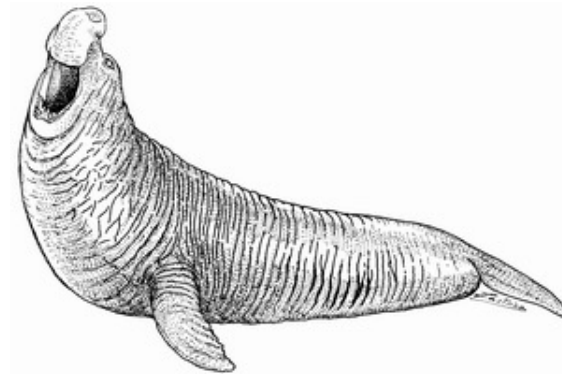All other assumptions of the Wright-Fisher model are met

Example: southern elephant seals (*Mirounga leonina*),
Falkland Islands (Fabiani *et al*. 2004)

N = 625
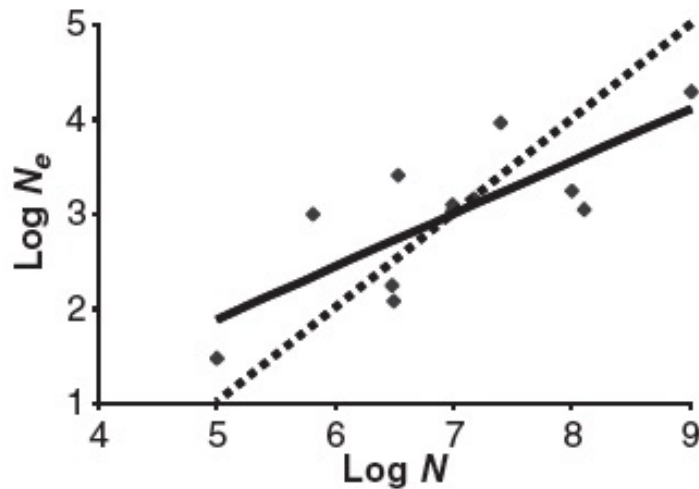N_m = 21 (estimated using microsatellite markers)
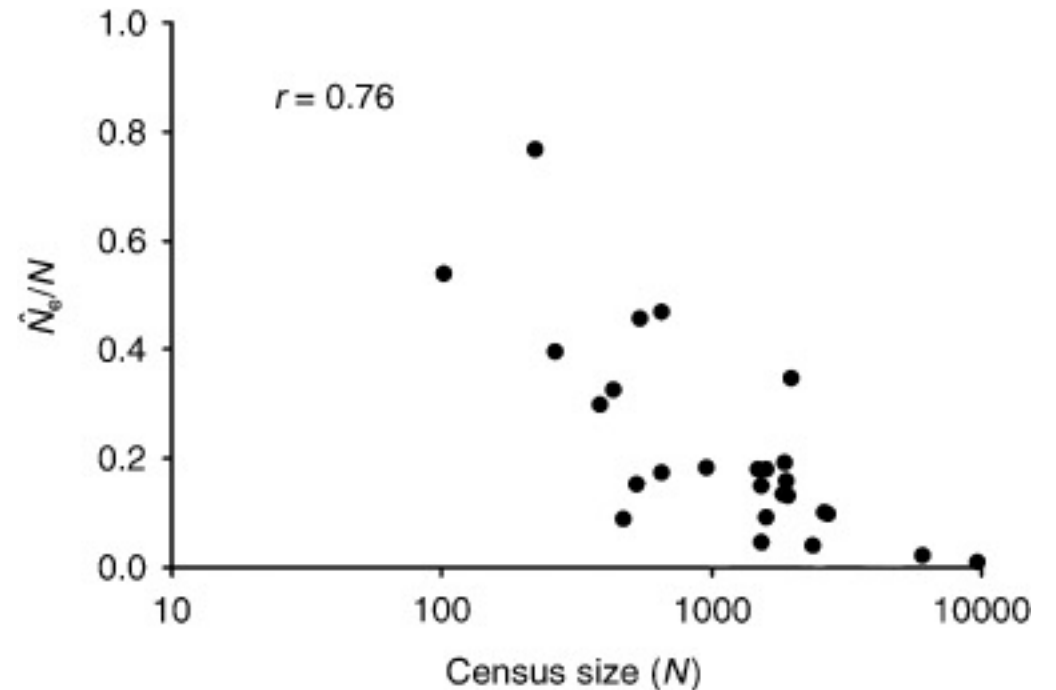N_f = 550
N_e = 80.9

N_e/N estimates in wild populations:

- In fish, a 100's to 10,000's typical for N_e

- N_e/N estimates: ~0.1 (Frankham 1995)

- N_e/N estimates in *marine* populations

  - 0.01 – 0.000001 (Hauser & Carvalho 2008)
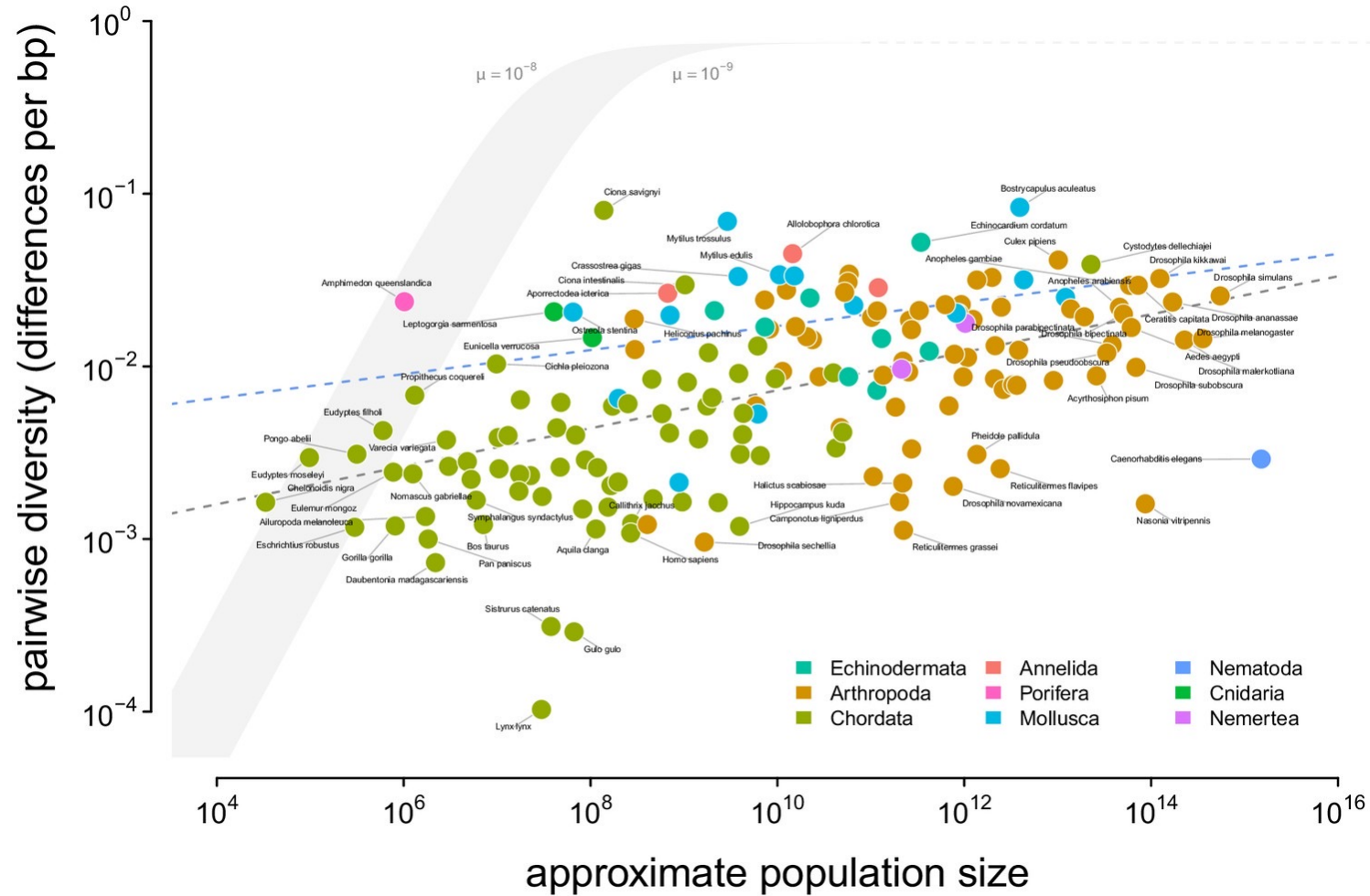
Decreases with increasing population size

Hauser & Carvalho 2008

Palstra and Ruzzante 2008

N$_e$/diversity does not scale with population size

Bufflao 2021, eLife

Why are N_e/N estimates so low in marine populations?

- Publication bias?
  - Difficult to obtain finite N_e estimates for large populations
- Fluctuations in population sizes?
  - Yes, but some of the methods used apply to the short-term N_e
- Uneven breeding sex ratios?
  - Yes, but extreme biases in sex ratios would be required to explain the N_e/N ratios,
    - e.g. one of the sexes 97.4% of the spawning population for $N_e/N = 0.1$
- Selection
- "sweepstakes" reproductive systems
  - High fecundity and mortality during the pelagic larval phase where entire families either survive or die (Hedgecock 1994)
  - Lower diversity within and higher differentiation among larval cohorts
    - Larvae or juveniles in recruiting batches tend to be related
  - N_e/N primarily affected by the small proportion of winners,
    - e.g. 0.004 for a pop of 100,000 where 100 produce 1000 offspring, 50,000 two and 49900 zero

Estimating effective population size

    Many methods (not reviewed here!)

    Three commonly used approaches (Hare *et al.* 2011)

        Single time point (contemporary)

            Linkage disequilibrium

            Excess heterozygosity

            Frequency of half and full siblings

        Temporal (contemporary)

            Variance N_e

        Coalescent (long-term)

Wright-Fisher simulation

- Exercise today:
  - You're going to look at the effect of population size on drift.
  - Hint- it matters… a lot.

If drift causes all alleles to eventually be fixed or lost, why do we have diversity?

Mutation

- Drift ↓ variation
- Mutation ↑ variation

- Deletions, insertions, inversions, point mutations….

- Mutation rate = μ
  - Probability of A → a
  - Typically 10^{-7} to 10^{-9} in eukaryotes

Species divergence and molecular clocks

- This is when things get cool…

- The rate at which mutations (fixed) accumulate between species

- If:
    - Mutation rate = μ
    - 2N gene copies
    - How many new mutation each generation?
        - $2N\mu$
    - What is frequency of each new mutation?
        - 1/2N
    - Probability that any mutation goes to fixation?
        - 1/2N
    - How many mutations each generation will go to fixation?
        - $2N\mu * 1/(2N) = \mu$
        - **Rate of subsitution = mutation rate!**

Molecular clock

- If the substitution rate is constant, we can use this as a "clock"
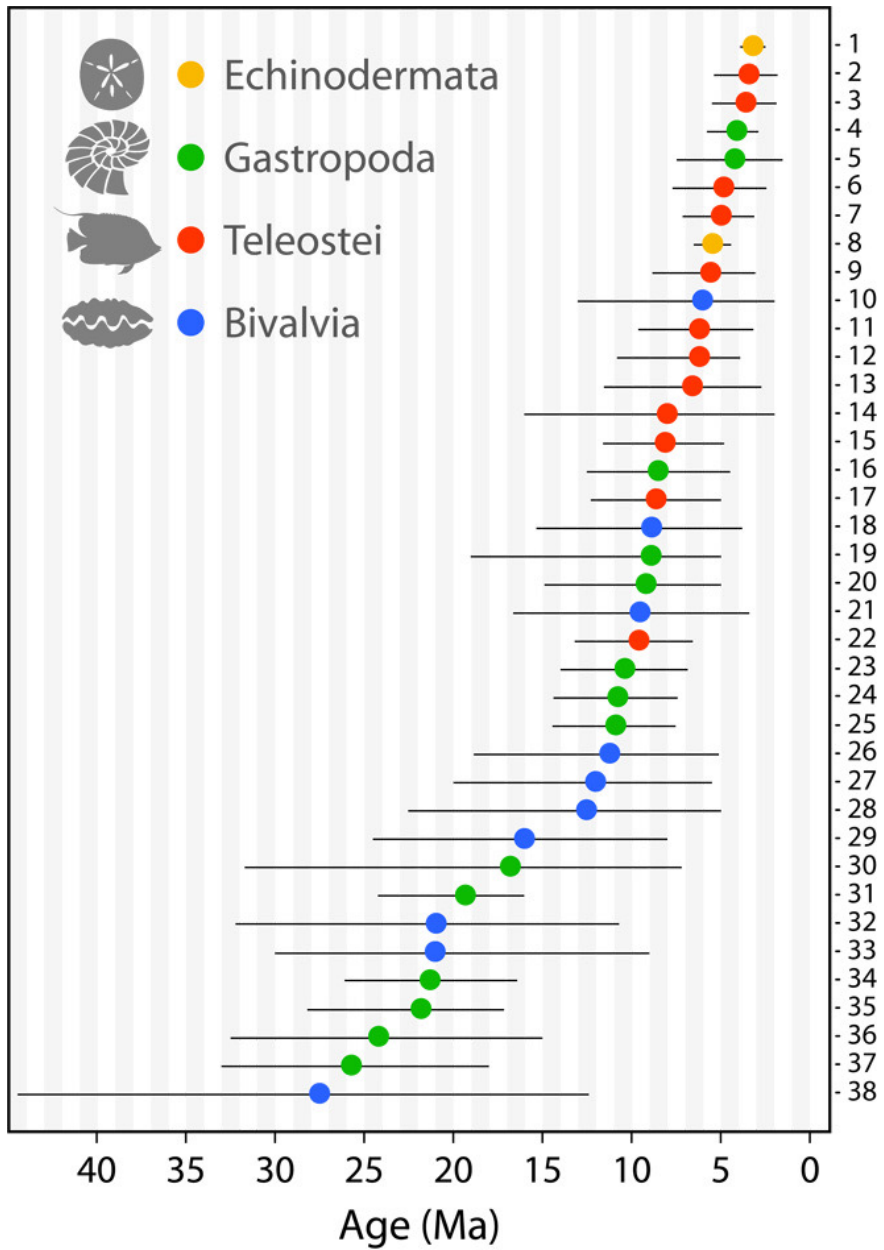
Ancestor

Species A

Species B

t
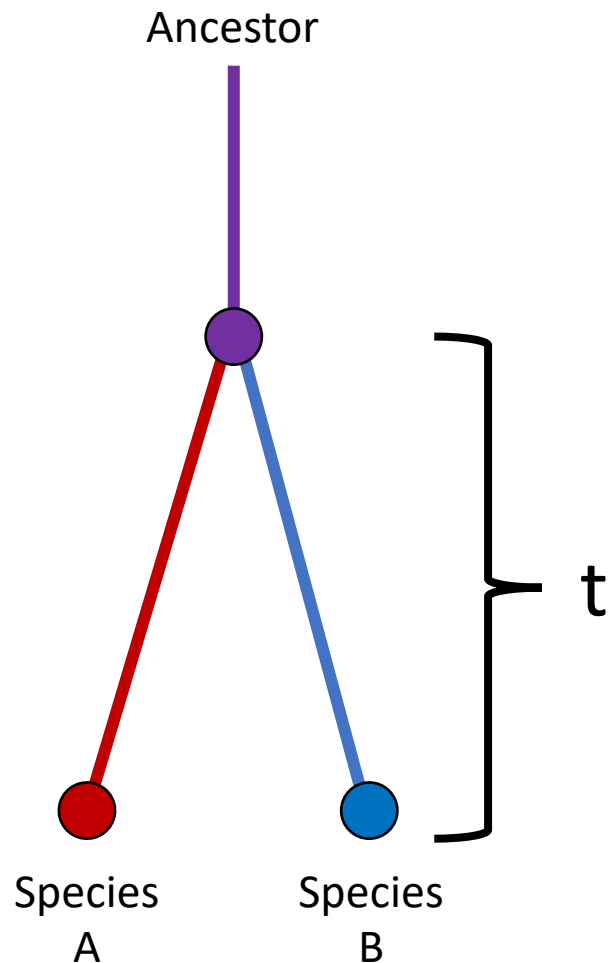
Expected number of fixed differences:

$$d = 2t\mu$$
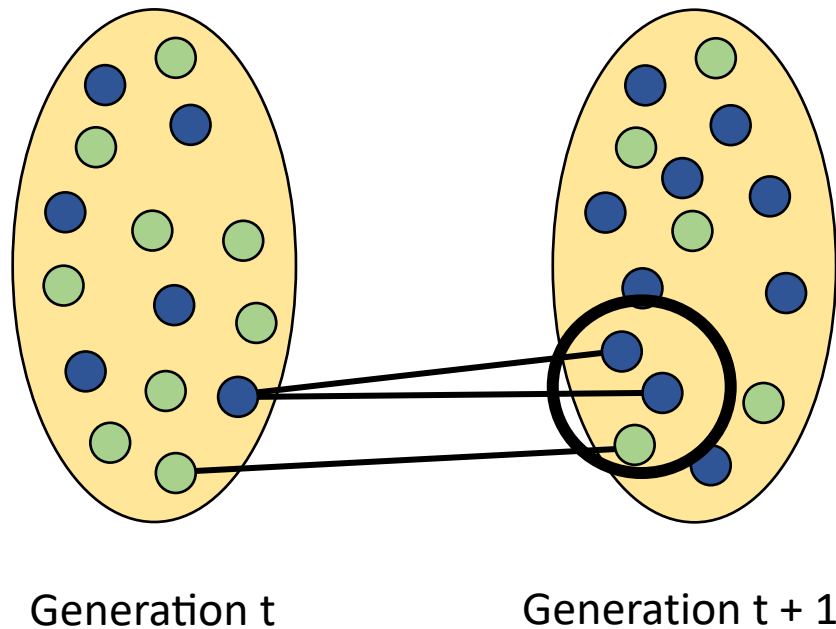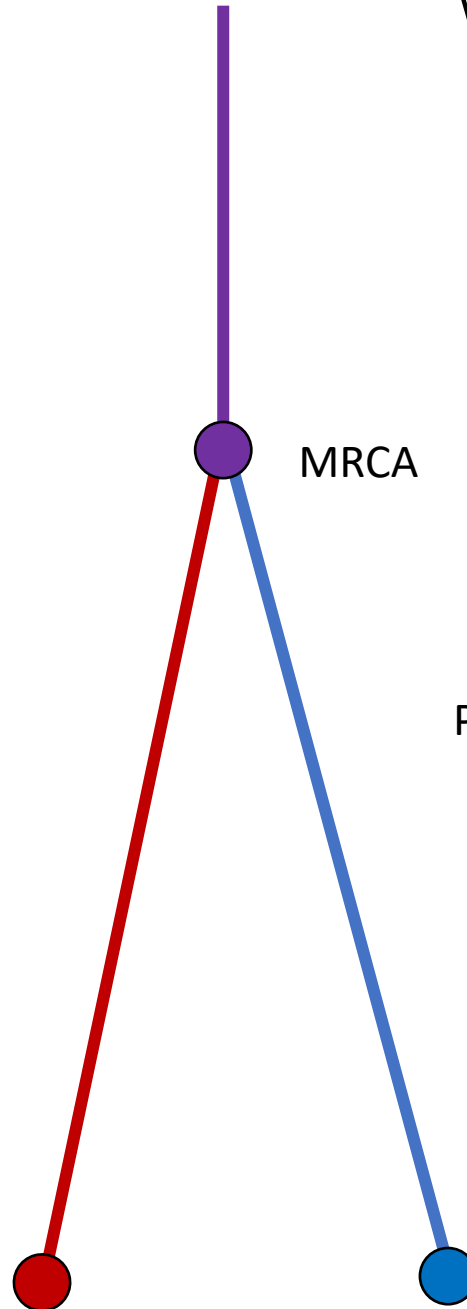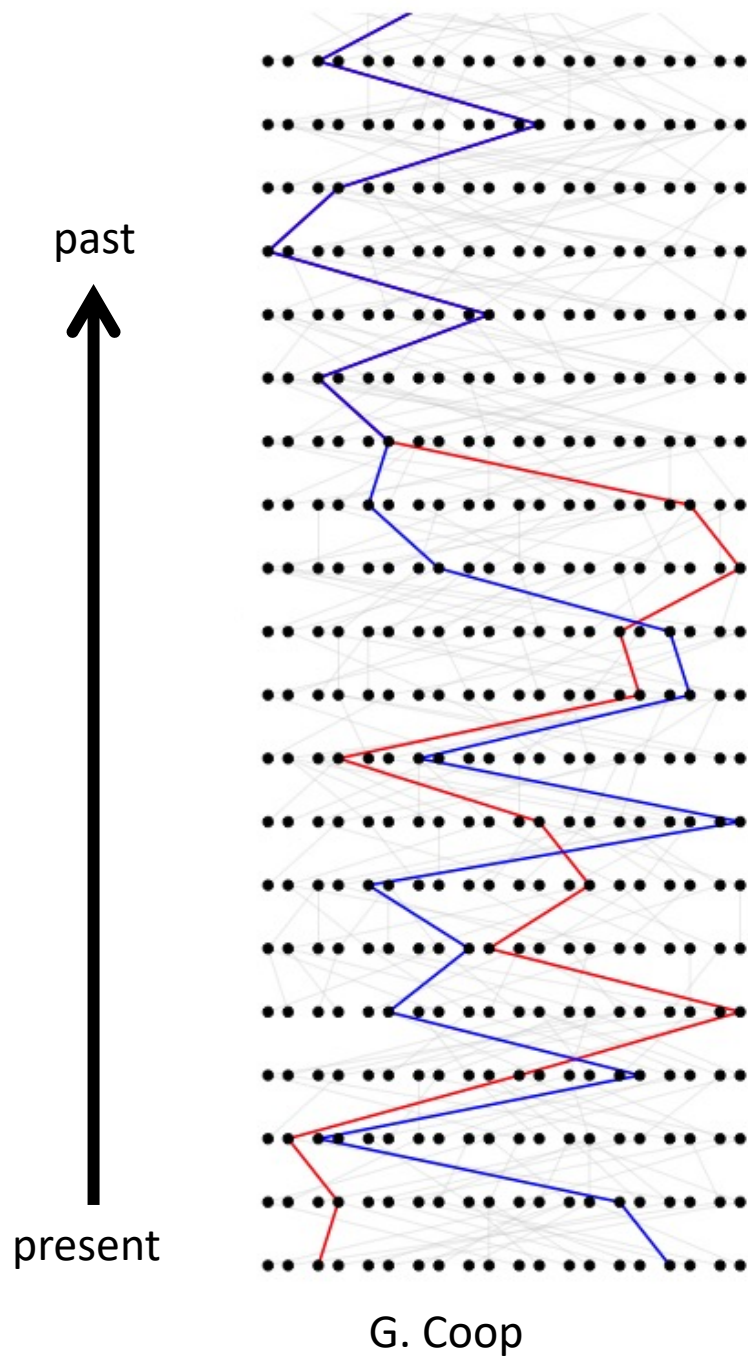
Why 2t?

How do you get t?

O'Dea et al., 2006, Science Advances

Coalescent

- Thinking backwards in time.
- When do two gene copies have the same parent?
  - When do they **coalesce**?

Generation t   Generation t + 1

What is the probability that two individuals have the same parent in the previous generation?

Pr(2 gene copies have same parent in previous gen) =

$1/(2N)$

Pr(2 gene copies don't have same parent) =

$1 - 1/(2N)$

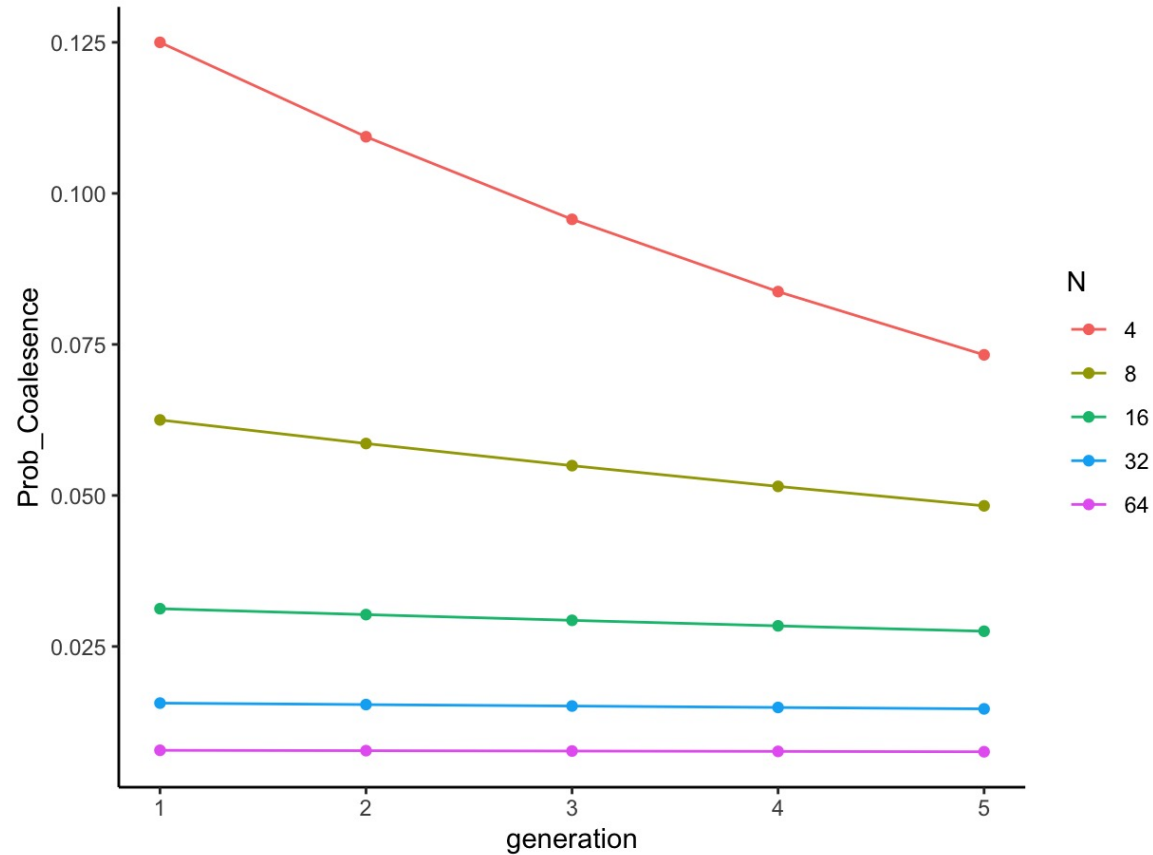Pr(2 gene copies don't have same parent in t generations) =

$[1 - 1/(2N)]^t$

Pr(2 gene copies coalesce in generation t + 1) =

$[1 - 1/(2N)]^t [1/2N]$

past

present

G. Coop

MRCA

Larger population = longer coalescent time

Mean time to coalescence is 2N generations

Large populations have more variation than small ones

Genetic variation

- Expected number of mutations separating any two alleles in a population

$$4N\mu = \theta$$

- Links amount of variation to population size
- Large populations have long coalescent times and high variation.

Genetic variation

- Tajima's estimator: π
  - Average number of pairwise differences
    - $\frac{Sum\ of\ differences}{total\ comparisons}$
    - Total compairsons: n(n-1)/2

  - 2+2+2
  - 0+1
  - 1
  - Total= 8
  - Comparisons = 4(3)/2 = 6
  - π= 8/6 = 1.33

A: 0 0 0
B: 0 1 1
C: 0 1 1
D: 1 0 1

Genetic variation

Watterson's estimator: θ_w

- The number of segregating sites

$$\theta_w = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

- $S = 3$
- $n = 4$
- $1/1 + \frac{1}{2} + 1/3 = 1.83$
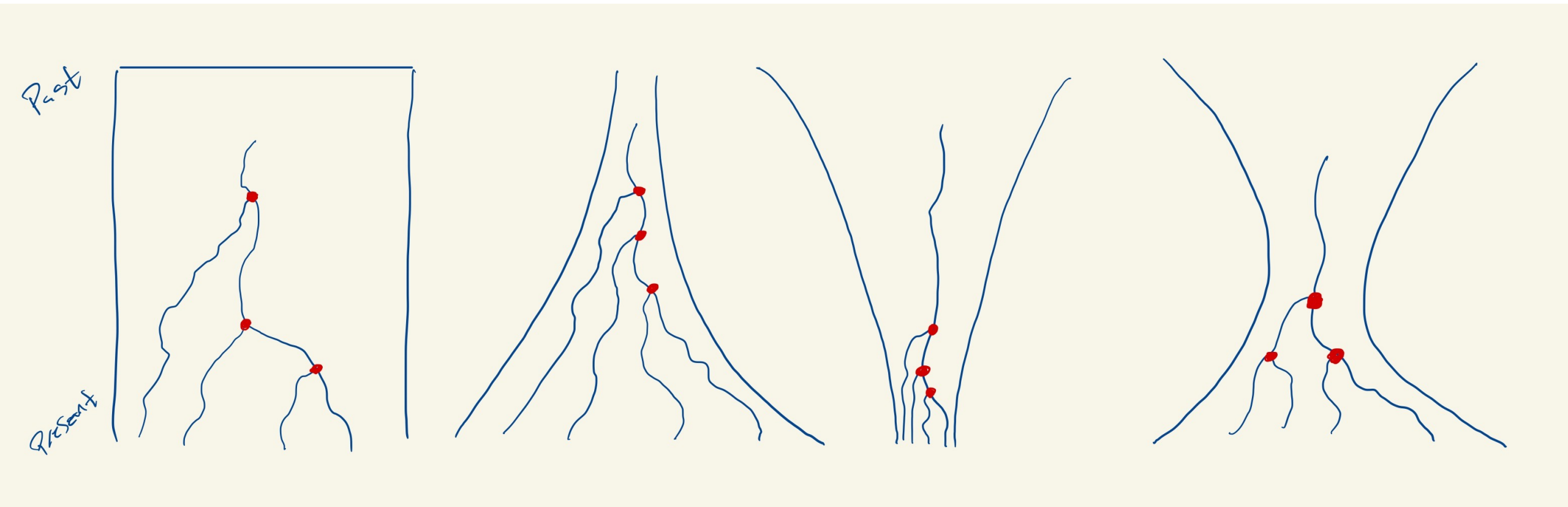- $\theta_w = 3/1.83 = 1.64$

A: 0 0 0
B: 0 1 1
C: 0 1 1
D: 1 0 1

$\pi = 1.33$

Changes in population size have large effect

Coalesence patterns can tell us about demographic history

Paper for discussion

Demographic histories and genetic diversity across pinnipeds are shaped by human exploitation, ecology and life-history

M.A. Stoffel[1,2], E. Humble[1,3], A.J. Paijmans [1], K. Acevedo-Whitehouse[4], B.L. Chilvers[5], B. Dickerson[6], F. Galimberti [7], N.J. Gemmell [8], S.D. Goldsworthy[9], H.J. Nichols[2,18,19], O. Krüger[1], S. Negro[10,20], A. Osborne[11], T. Pastor[12], B.C. Robertson[13], S. Sanvito [7], J.K. Schultz[14], A.B.A. Shafer[15], J.B.W. Wolf[16,17] & J.I. Hoffman[1,3]