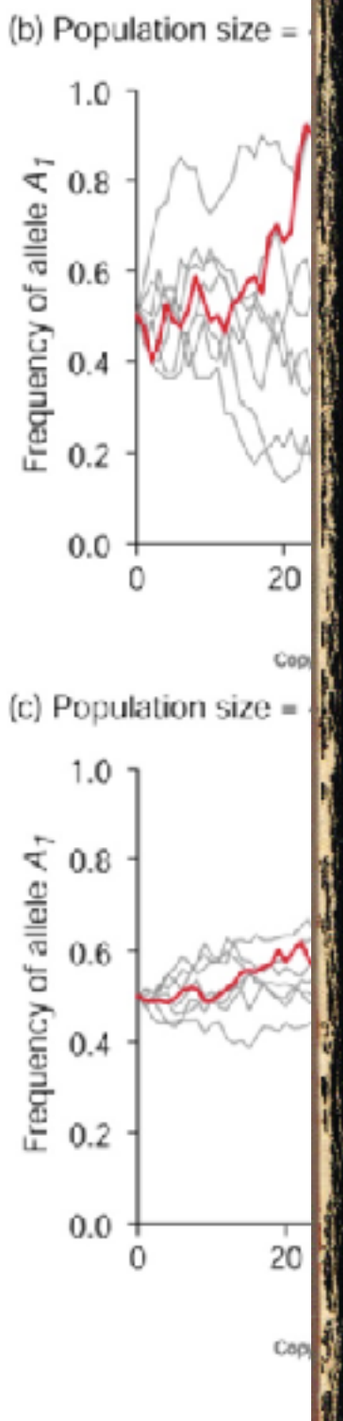
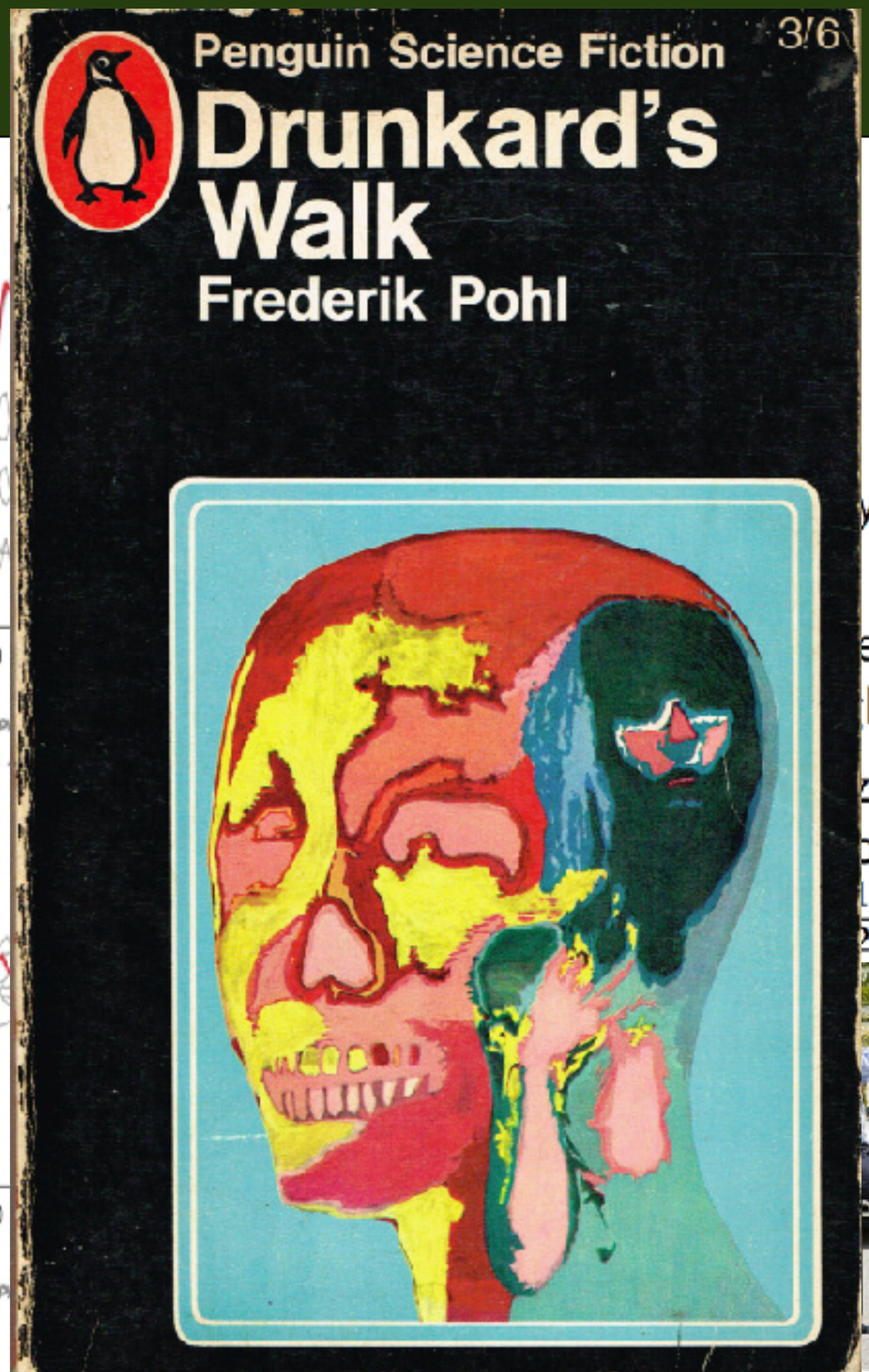
EEB Population Genetic Simulation and Inference

Lecture 1b (continued):
Models of Evolution, intro, etc

Professor Mike Hickerson
Department of Biology - City College
CUNY GC subprogram in EEB

Example allele frequency trajectories due to genetic drift





y to a "drunkard's

epping left and
ht at random

ze of the steps
portional here to:
$\frac{p(1-p)}{2N}$

genetic drift experiment in *Drosophila melanogaster*

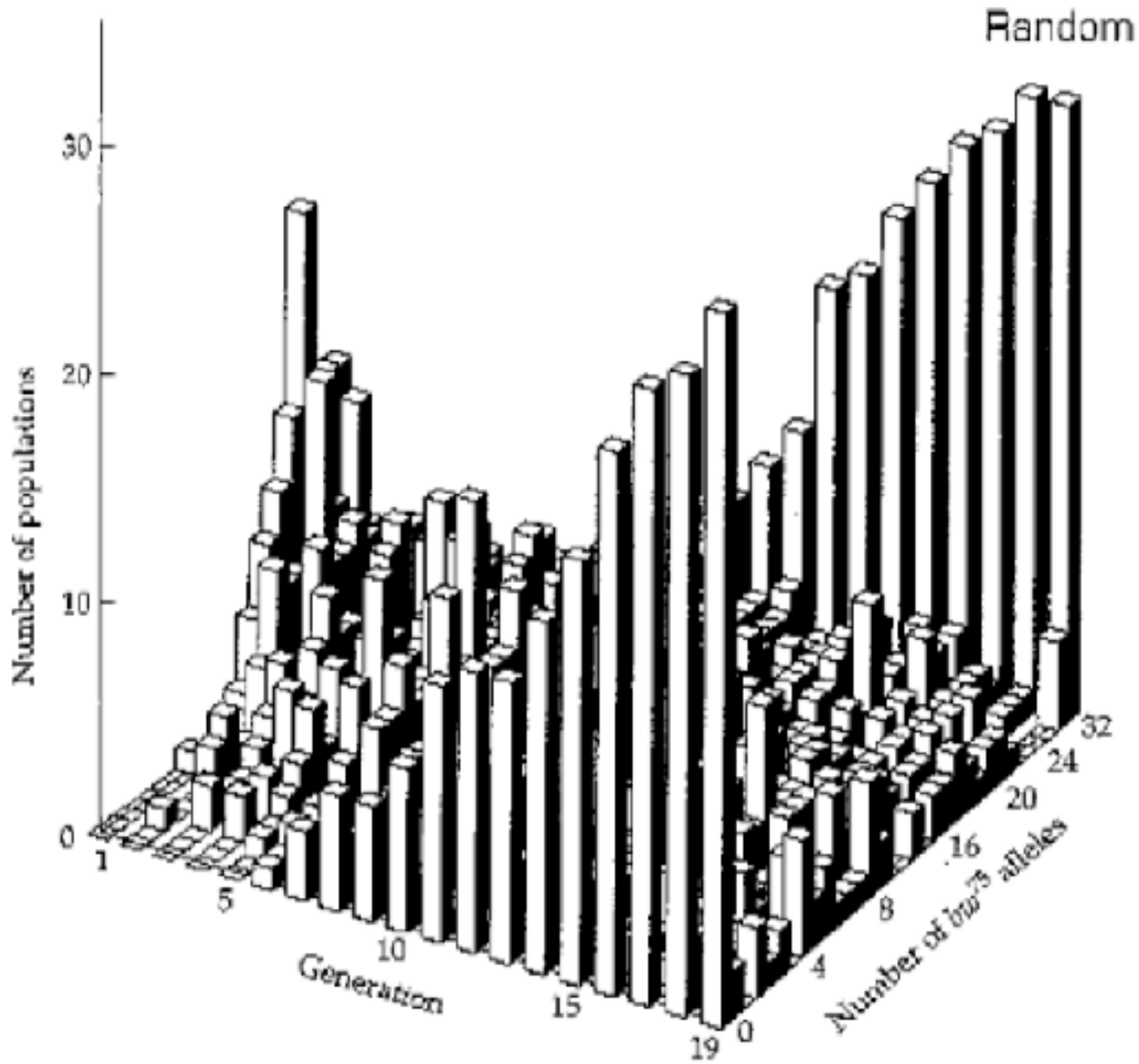


FIGURE 3.4 Random genetic drift in 107 actual populations of *Drosophila melanogaster*. Each of the initial 107 populations consisted of 16 bw^{75}/bw heterozygotes ($N = 16$; bw = brown eyes). From among the progeny in each generation, eight males and eight females were chosen at random to be the parents of the next generation. The horizontal axis of each curve gives the number of bw^{75} alleles in the population, and the vertical axis gives the corresponding number of populations. (Data from Buri 1956.)

theory meets data

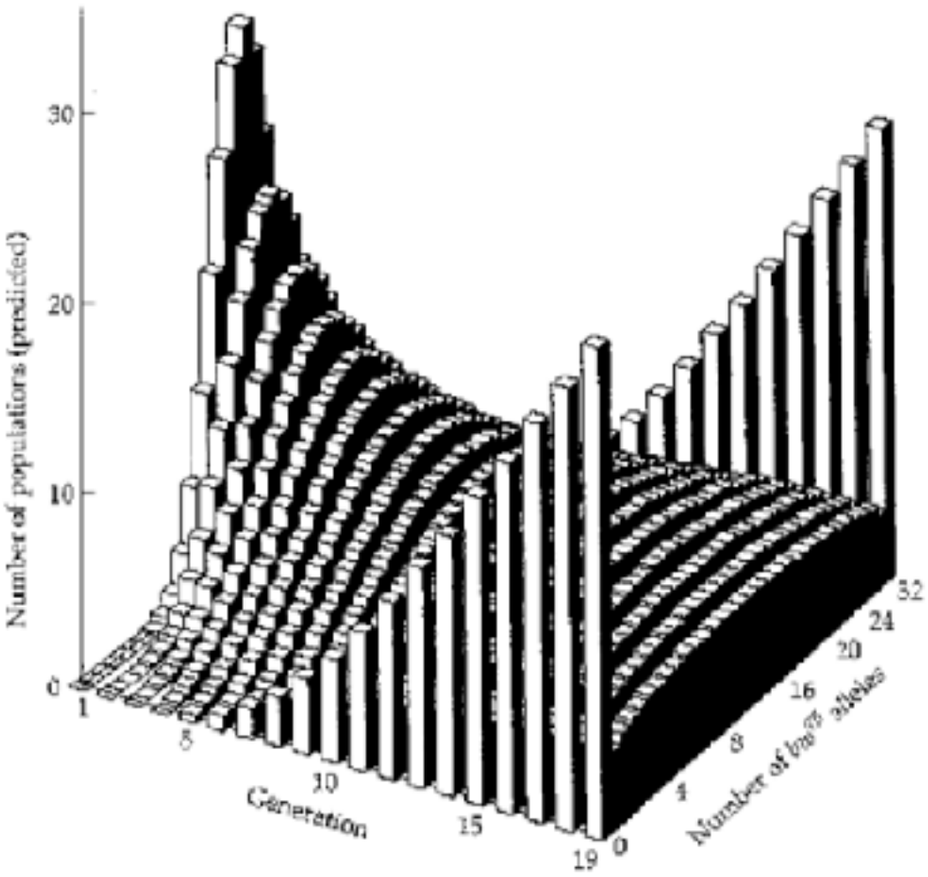


FIGURE 3.5 Prediction of the Wright-Fisher model for the distribution of allele frequencies $\phi(p, x; t)$ in subpopulations of size $N = 16$, where x represents the allele frequency in generation t. Time runs for 19 generations, and all subpopulations start with an initial allele frequency of $p = 0.5$. The values of $\phi(p, x; t)$ were generated by successive multiplication of the Markov transition probability matrix, whose entries are given by the binomial distribution in Equation 3.2. The model with $2N = 32$ predicts that fewer populations have fixed by generation 19 than actually did go to fixation in the experiment in Figure 3.4. This is because the variance in offspring number is about 70% greater than that assumed in the Wright-Fisher model.

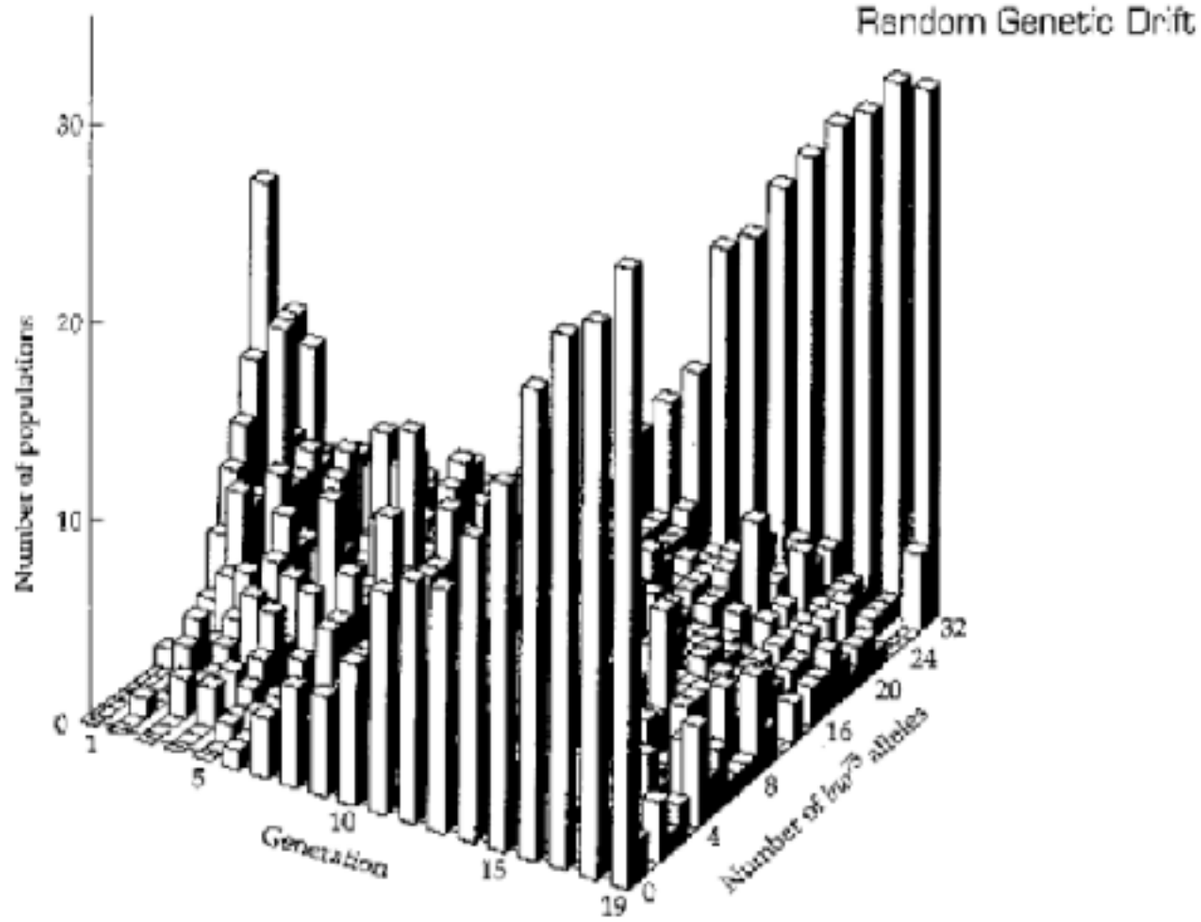
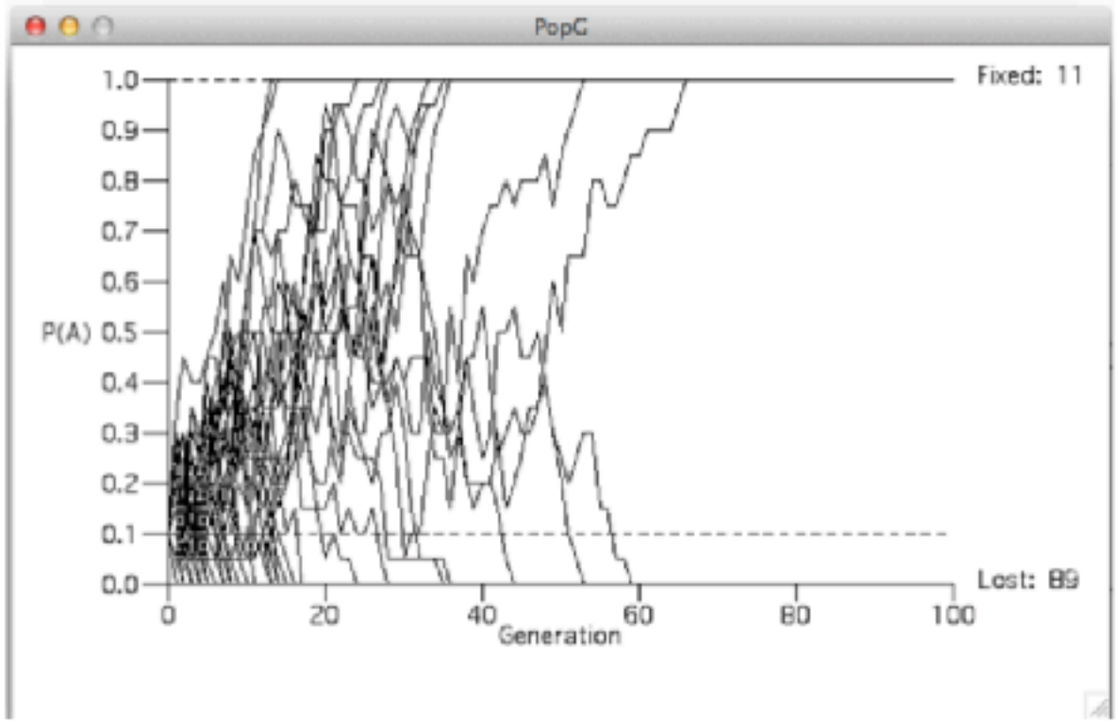


FIGURE 3.4 Random genetic drift in 107 actual populations of *Drosophila melanogaster*. Each of the initial 107 populations consisted of 16 bw^{75}/bw heterozygotes ($N = 16$; bw = brown eyes). From among the progeny in each generation, eight males and eight females were chosen at random to be the parents of the next generation. The horizontal axis of each curve gives the number of bw^{75} alleles in the population, and the vertical axis gives the corresponding number of populations. (Data from Buri 1956.)

Wright-Fisher model

Simulating the trajectories of neutral mutations



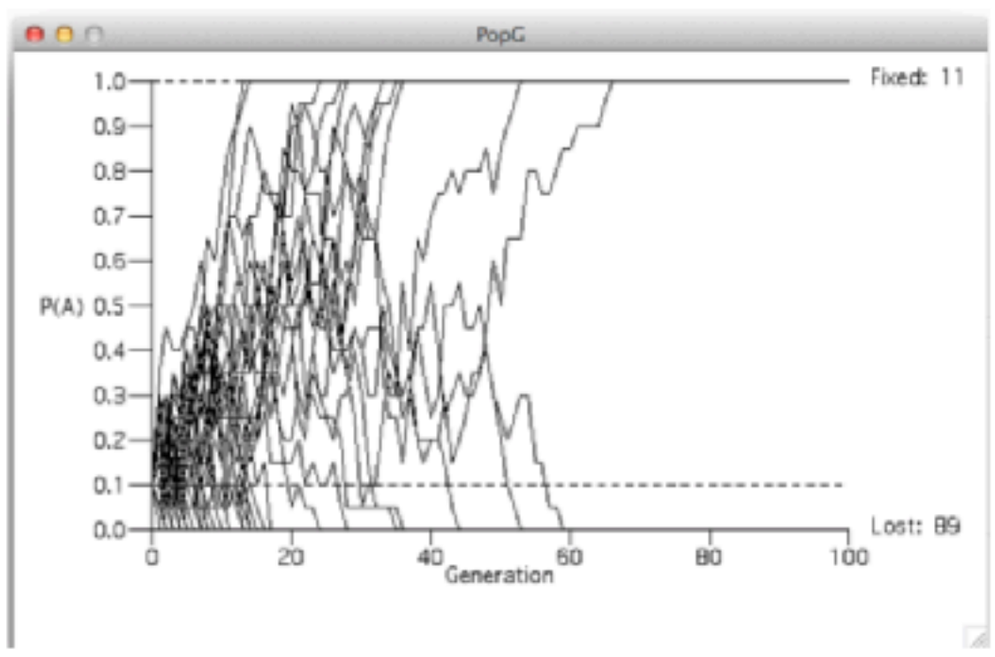
- 100 identical populations, 10 individuals each
- New mutant allele *A* starts with frequency 1/10
- 90% of time, random walk ends at frequency 0. 10% of time, it ends at frequency 1, potentially creating a *fixed difference* between populations

Wright-Fisher model

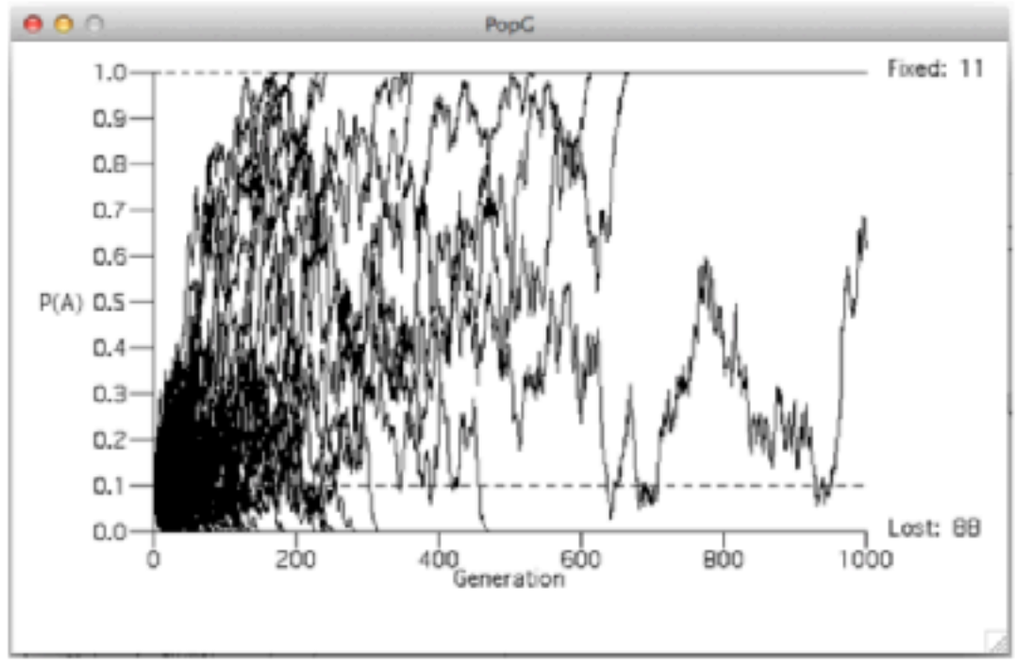
Genetic drift creates fixed differences very slowly

Using the theory of random walks, Kimura calculated that a neutral allele takes $2N$ generations to fix on average

(intuition: allele frequency changes by $1/N$ each generation)

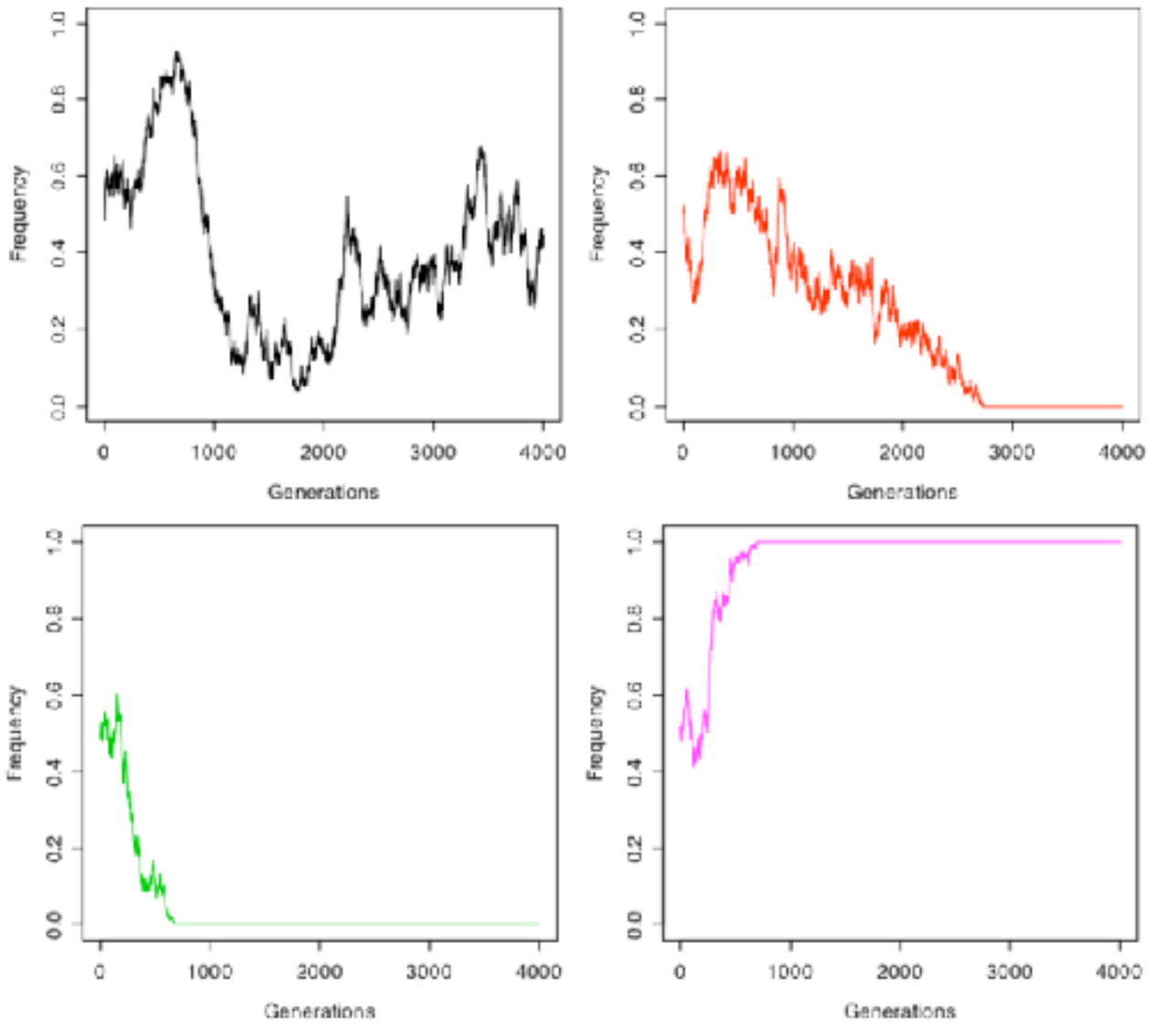


$N = 10$



$N = 100$

times to fixation or loss



Time to fixation and time to loss are random variables.

variation is eventually lost

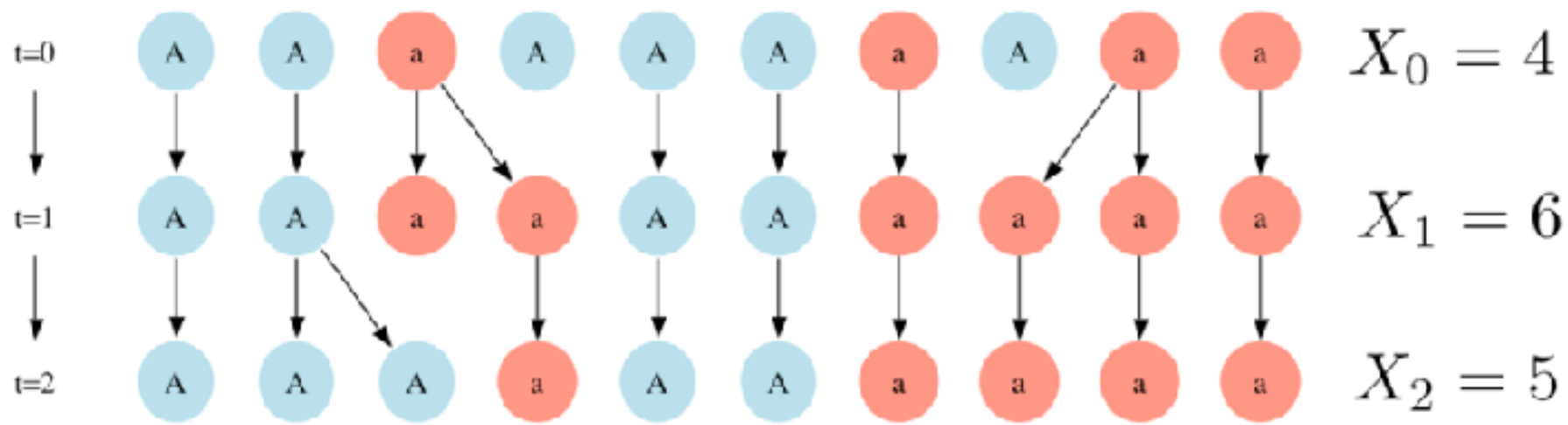
- In a model of genetic drift alone, alleles are either fixed or lost as time goes on.
- The random drunkard will eventually fall into a ditch...
- With genetic drift, genetic variation is lost through time.

identity by state

Definition

Let G_t be the probability of identity-by-state for two randomly sampled alleles from generation t.

- If two alleles are descended from the same gene copy in the previous generation, they must be *identical-by-state*. (Note: we are still ignoring mutation).
- If two alleles are not descended from the same gene copy in the previous generation, they will be *identical-by-state* with probability G_{t-1}



- Recall, G_t = probabilty of identity-by-state of two randomly sampled alleles
- Interestingly then:
  - G_t is equivalent to the expected proportion of homozygotes in a population, *the homozygosity*.
  - $1 - G_t$ is equivalent to the expected proportion of heterozygotes in a population, *the heterozygosity, or* H_t.

By considering two alleles sampled at random, we can derive:

$$G_t = \frac{1}{2N} + (1 - \frac{1}{2N})G_{t-1}$$

and we can write this as:

$$(1 - G_t) = (1 - G_{t-1})(1 - \frac{1}{2N})$$

$$H_t = H_{t-1}(1 - \frac{1}{2N})$$

$$H_t = H_0(1 - \frac{1}{2N})^t$$

$$H_t \approx H_0 e^{-\frac{t}{2N}}$$

Pr(2 children having same parent)

Pr(2 children not having same parent)

Pr(IBS if 2 children not have same parents)

By considering two alleles sampled at random, we can derive:

$$G_t = \frac{1}{2N} + (1 - \frac{1}{2N})G_{t-1}$$

and we can write this as:

heterozygosity

$$(1 - G_t) = (1 - G_{t-1})(1 - \frac{1}{2N})$$

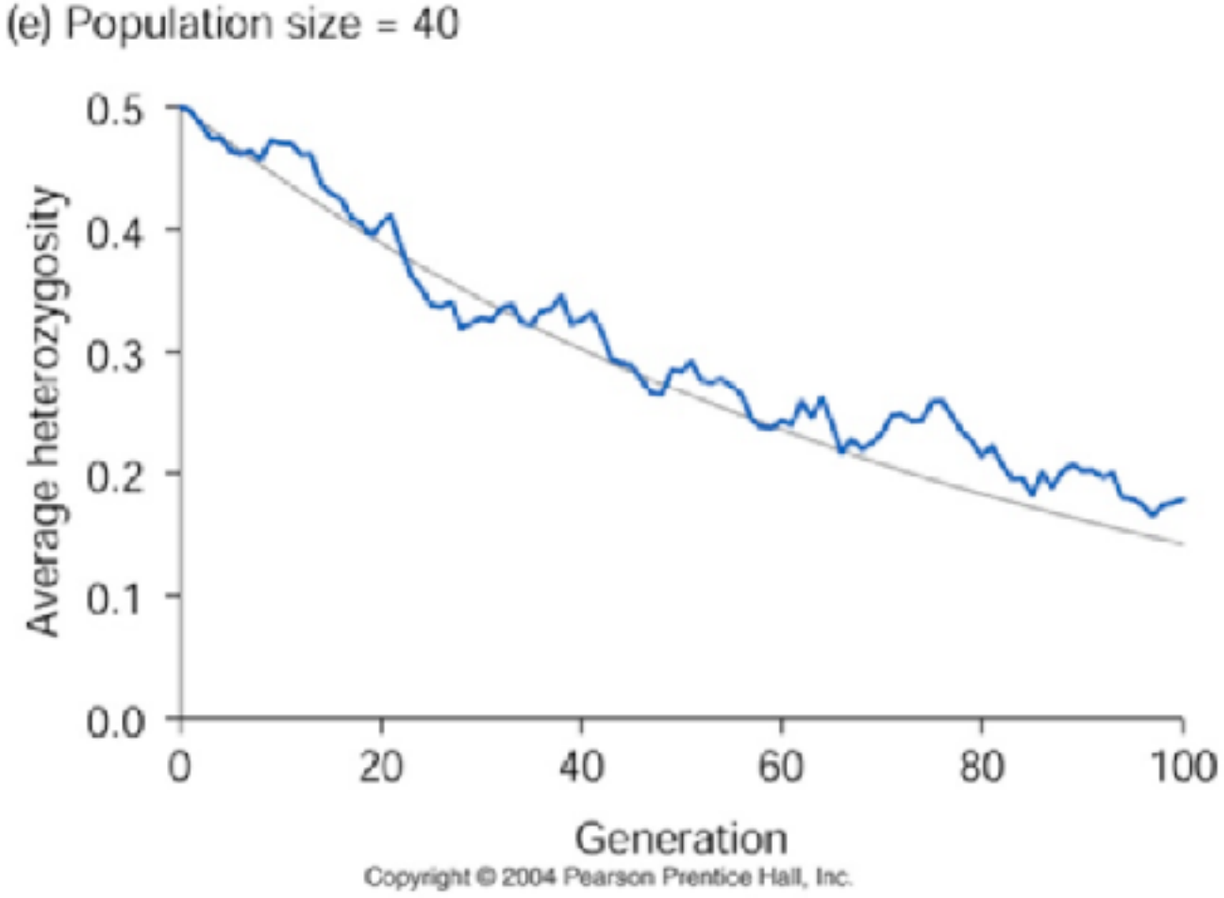
$$H_t = H_{t-1}(1 - \frac{1}{2N})$$

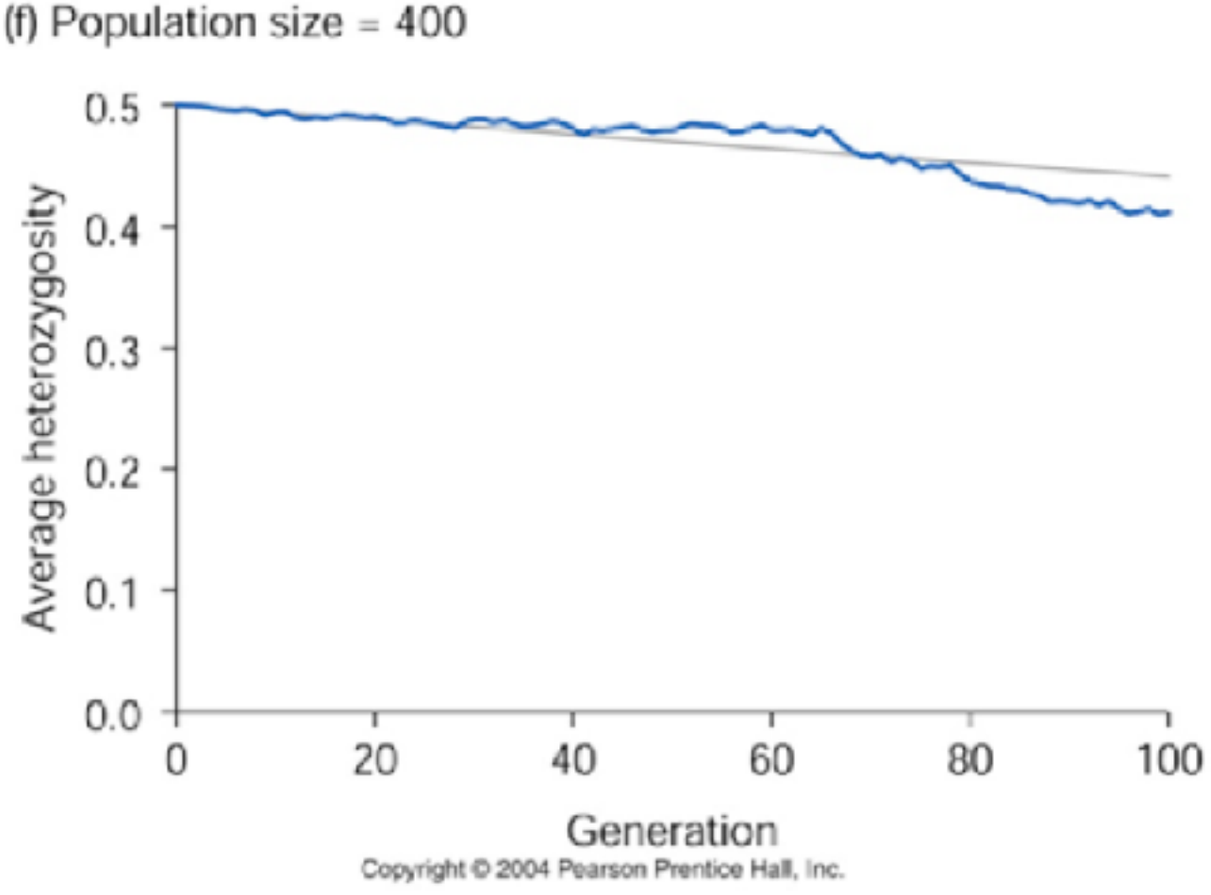
$$H_t = H_0(1 - \frac{1}{2N})^t$$

$$H_t \approx H_0 e^{-\frac{t}{2N}}$$

heterozygosity decreases on a time-scale proportional to *2N* generations

$$H_t \approx H_0 e^{-\frac{t}{2N}}$$



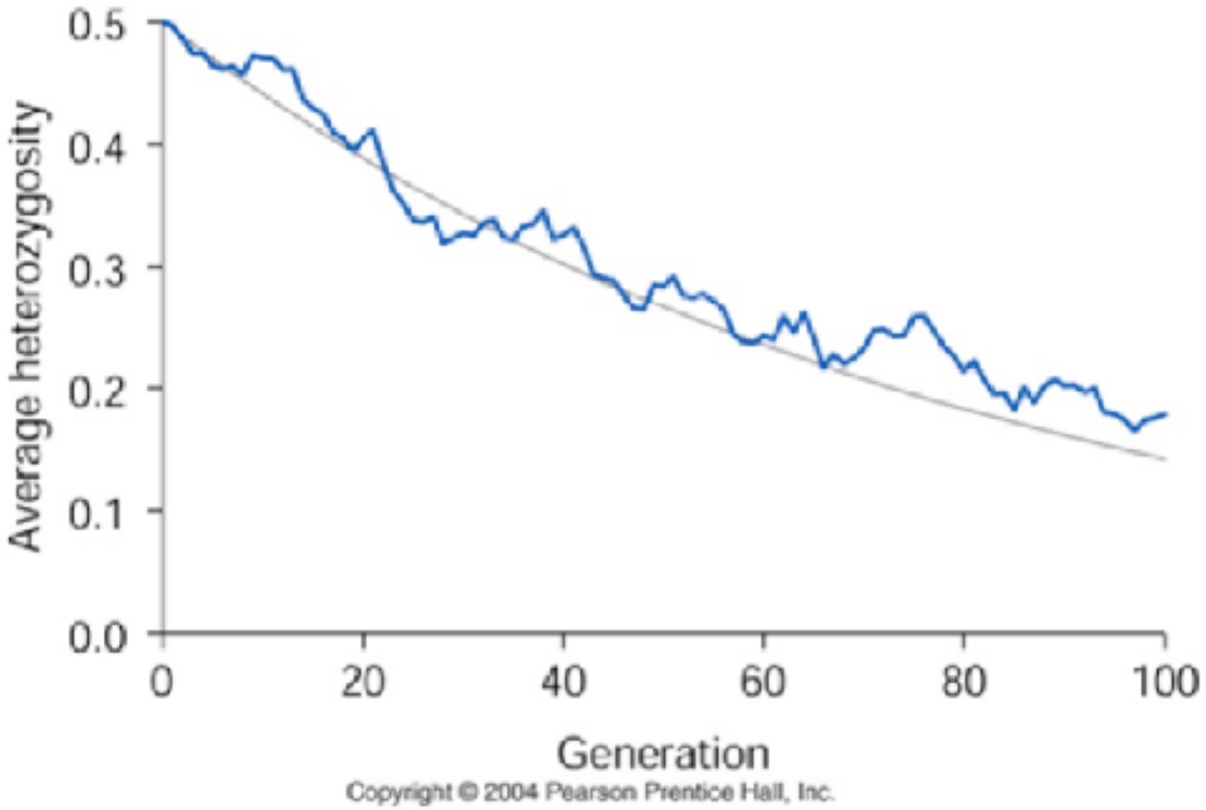


heterozygosity decreases on a time-scale proportional to *2N* generations

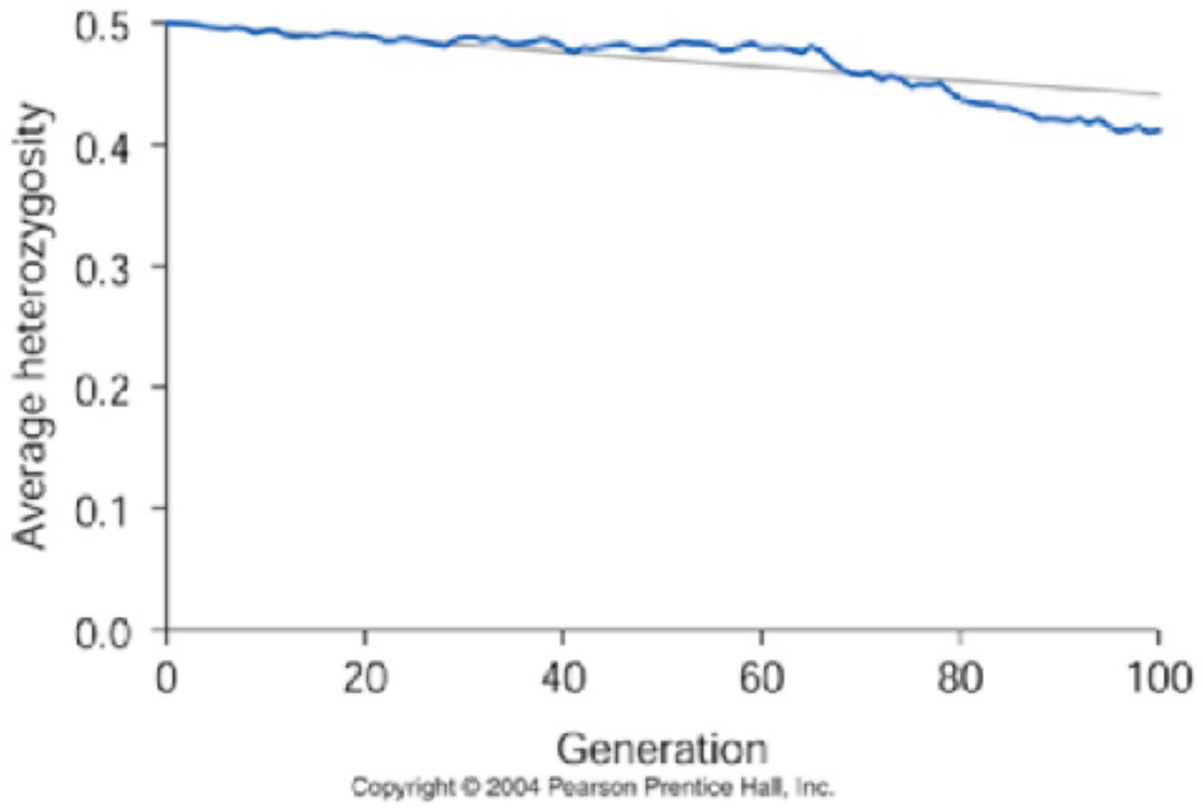
plug in H_0 and *t* in the future and *N* to predict H_t at time *t*

$$H_t \approx H_0 e^{-\frac{t}{2N}}$$

(e) Population size = 40



(f) Population size = 400



effective population sizes

- All these theoretical results are for an idealized, simple model.
- Yet, we can force it to work for many non-standard models.

Example

If population sizes vary through time as: $N_0, N_1, N_2, \ldots, N_t$, then if I plug in the harmonic mean, i.e.

$$N_e = 1/\left[\frac{1}{t}\left(\frac{1}{N_0} + \frac{1}{N_1} + \ldots + \frac{1}{N_{t-1}}\right)\right]$$

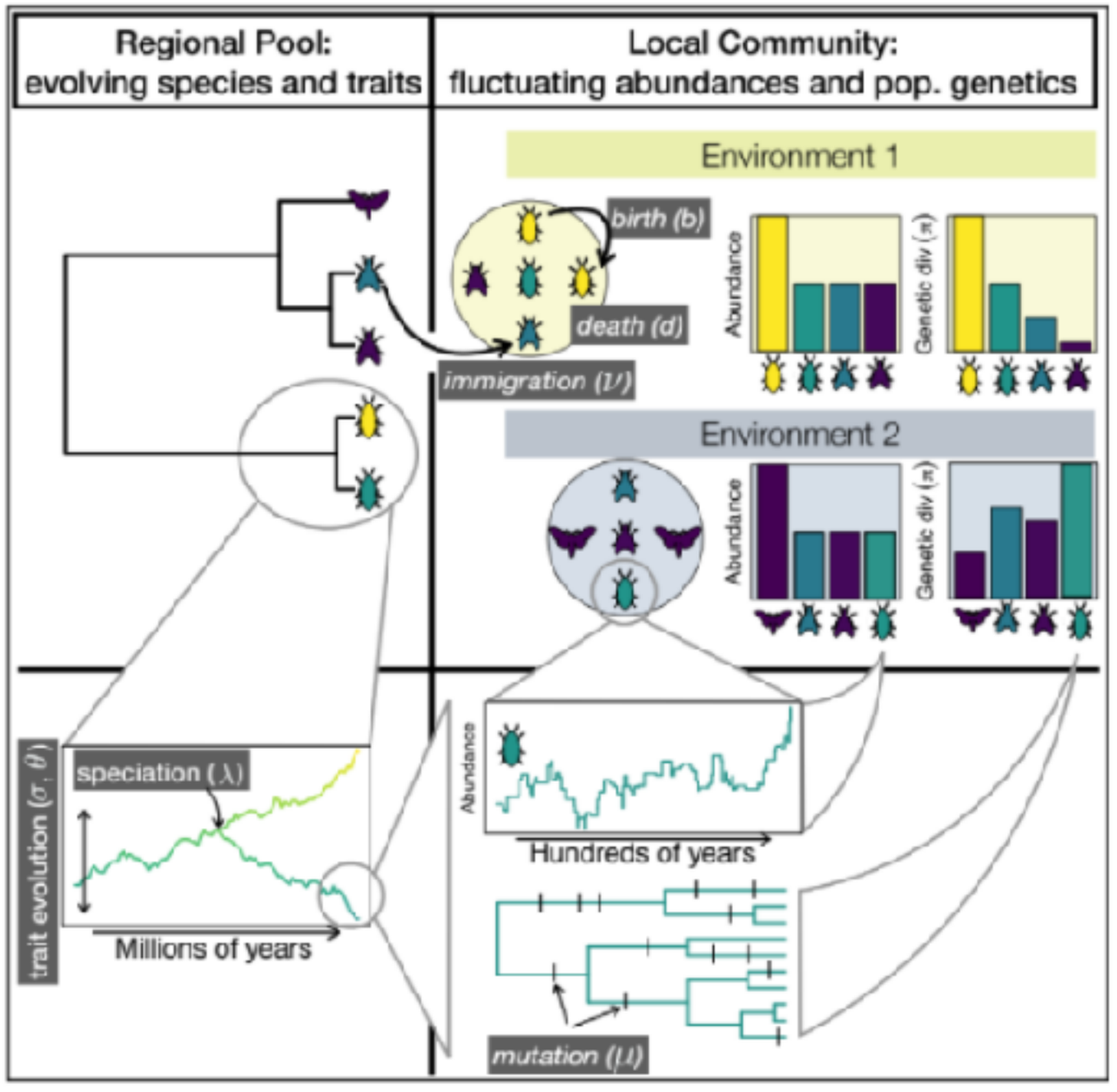
to the equations I can still predict, for example, the decrease in heterozygosity through time perfectly.

sEcoEvo Working Group

iDiv
German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig

Isaac Overcast
CUNY

Megan Ruffley
U of Idaho



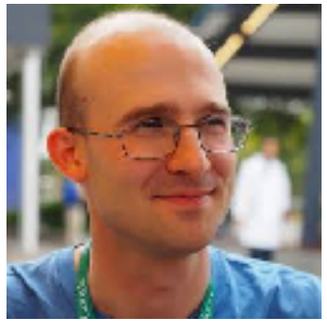
Massive
Eco-
Evolutionary
Synthesis
Simulations

Andy Rominger
(SFI)

Rosemary Gillespie

Luke Harmon

James Rosindell



Rampal Etienne

effective population sizes

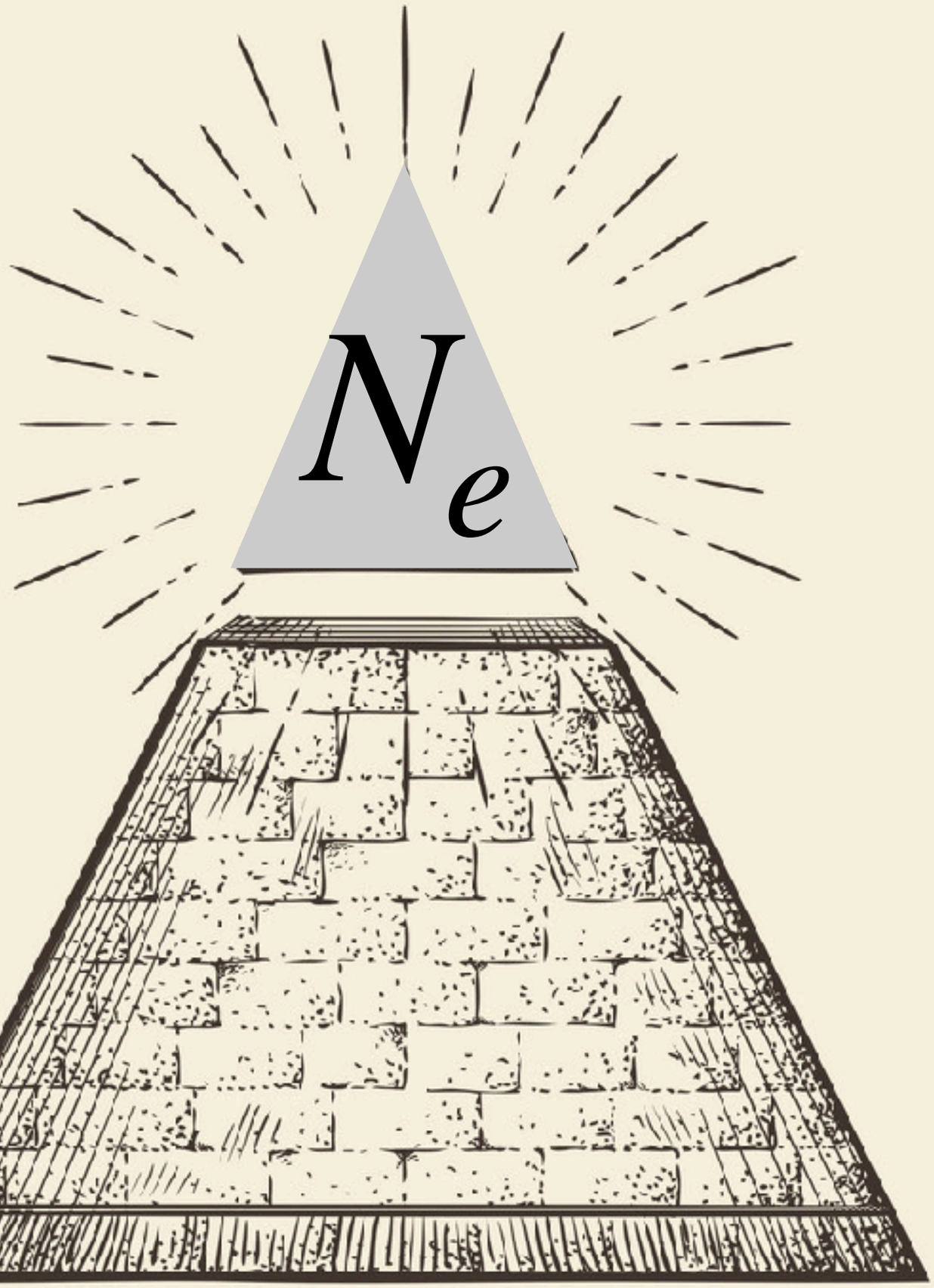
- All these theoretical results are for an idealized, simple model.
- Yet, we can force it to work for many non-standard models.

Example

If the number of males (N_m) and females (N_f) is not equal in a population, I can calculate and plug in :

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

and still do perfectly well.


N_e

low genetic variation - Conservation strategies

Low genetic diversity and resulting poor sperm quality has made breeding and survivorship difficult for cheetahs.

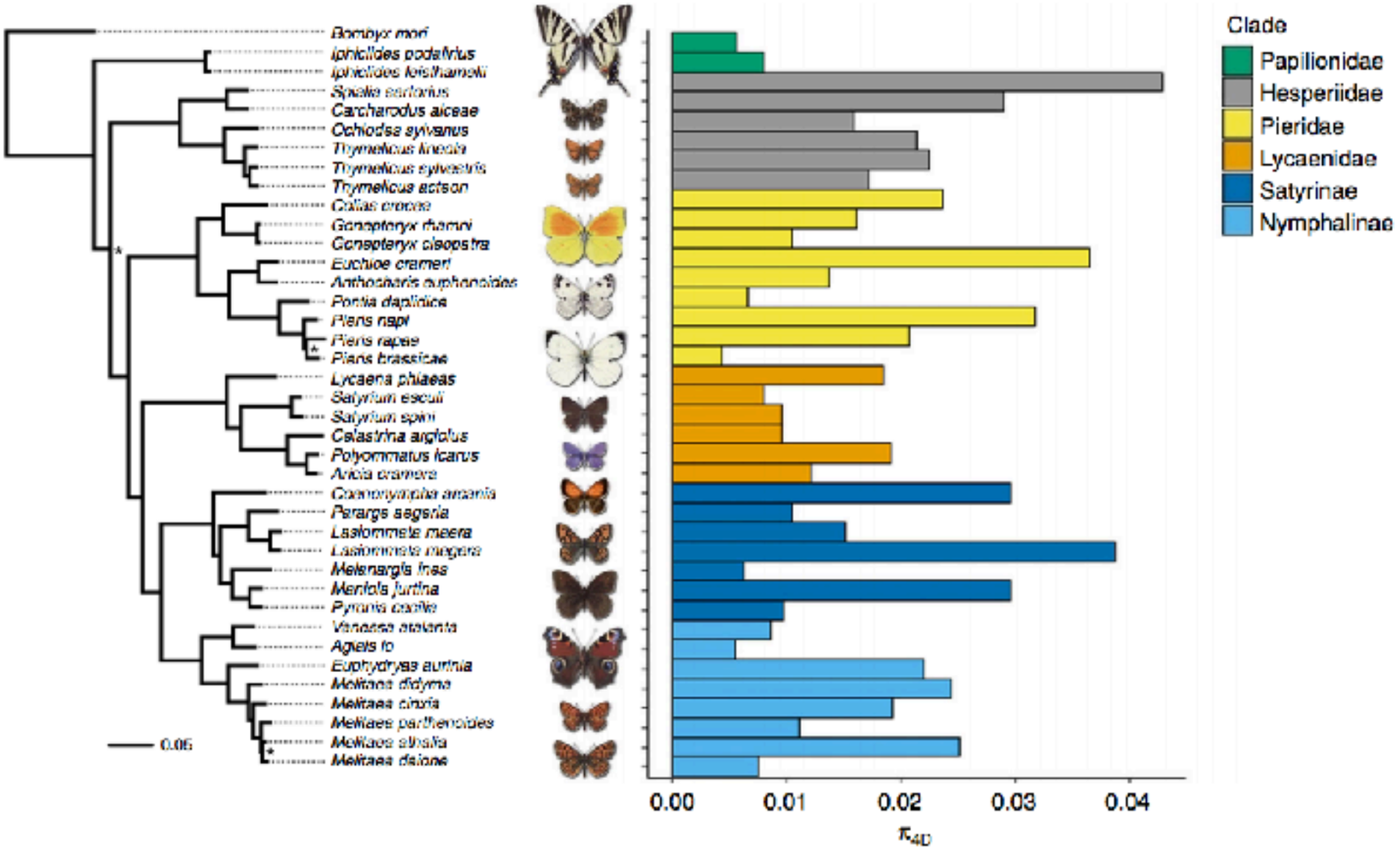
female cheetahs can mate with more than one male per litter of cubs. They undergo induced ovulation, which means that a new egg is produced every time a female mates. By mating with multiple males, the mother increases the genetic diversity within a single litter of cubs

https://doi.org/10.1038/s41467-019-11308-4 OPEN

The determinants of genetic diversity in butterflies

Alexander Mackintosh[1], Dominik R. Laetsch[1], Alexander Hayward[2], Brian Charlesworth[1], Martin Waterfall[1], Roger Vila[3] & Konrad Lohse[1]

Instead, neutral genetic diversity is negatively correlated with body size and positively with the length of the genetic map. This suggests that genetic diversity is determined both by differences in long-term population size and the effect of selection on linked sites.



. 1 Neutral genetic diversity (π_{4D}) across European butterfly species. The phylogeny is based on 218 single-copy orthologues and rooted with the moth *Bombyx mori* as an outgroup. All nodes have 100% bootstrap support unless marked with an asterisk (70–99%). The barplot on the right shows ıome-wide estimates of π_{4D} for 38 focal species sampled from the six major groups of Papilionoidea present in Europe. The phylogeny explains very little :he variation in π_{4D} in butterflies. Source data are provided as a Source Data file

Is genomic diversity a useful proxy for census population size? Evidence from a species-rich community of desert lizards

Maggie R. Grundler ✉, Sonal Singhal, Mark A. Cowan, Daniel L. Rabosky

First published: 09 February 2019 | https://doi.org/10.1111/mec.15042

summary

Our models of genetic drift provides us:

- Indication of the importance of population size, and more generally demography, on genetic variation
- We now understand a major force that *destroys* genetic variation. Prompts us to ask: *why is there so much variation then?*
- A more realistic background expectation for how allele frequencies change through time

back to the Moran Model, …

Trajectory of allele frequencies starts a $1/N$ and ends when it reaches 0 or 1

Given a neutral allele whose frequency is k/N, next frequency can be either $(k-1)/N$ or $(k+1)/N$ with equal probability

Probability of fixation (ultimate frequency = 1 instead of 0) is $1/N$

back to the Moran Model, but ….

Trajectory of allele frequencies starts a $2/N$ and ends when it reaches 0 or 1

Given a neutral allele whose frequency is $k/2N$, next frequency can be either $(k-1)/2N$ or $(k+1)/2N$ with equal probability

Probability of fixation (ultimate frequency = 1 instead of 0) is $2/N$

I'm so confused…is it 1/*N* or 1/2*N* ?

Thomas Mailund @ThomasMailund · Jul 25, 2013

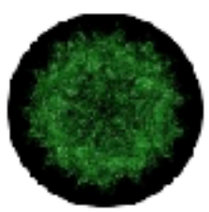
Why am I always off by a factor of two on all coalescence parameters? just my brain that can't deal with coalescence theory?

2 1

Graham Coop @Graham_Coop · Jul 25, 2013

@ThomasMailund popgen factors of two are the bane of my life. Wondered about setting up a "popgen don't care..." meme specifically f this.

6 3



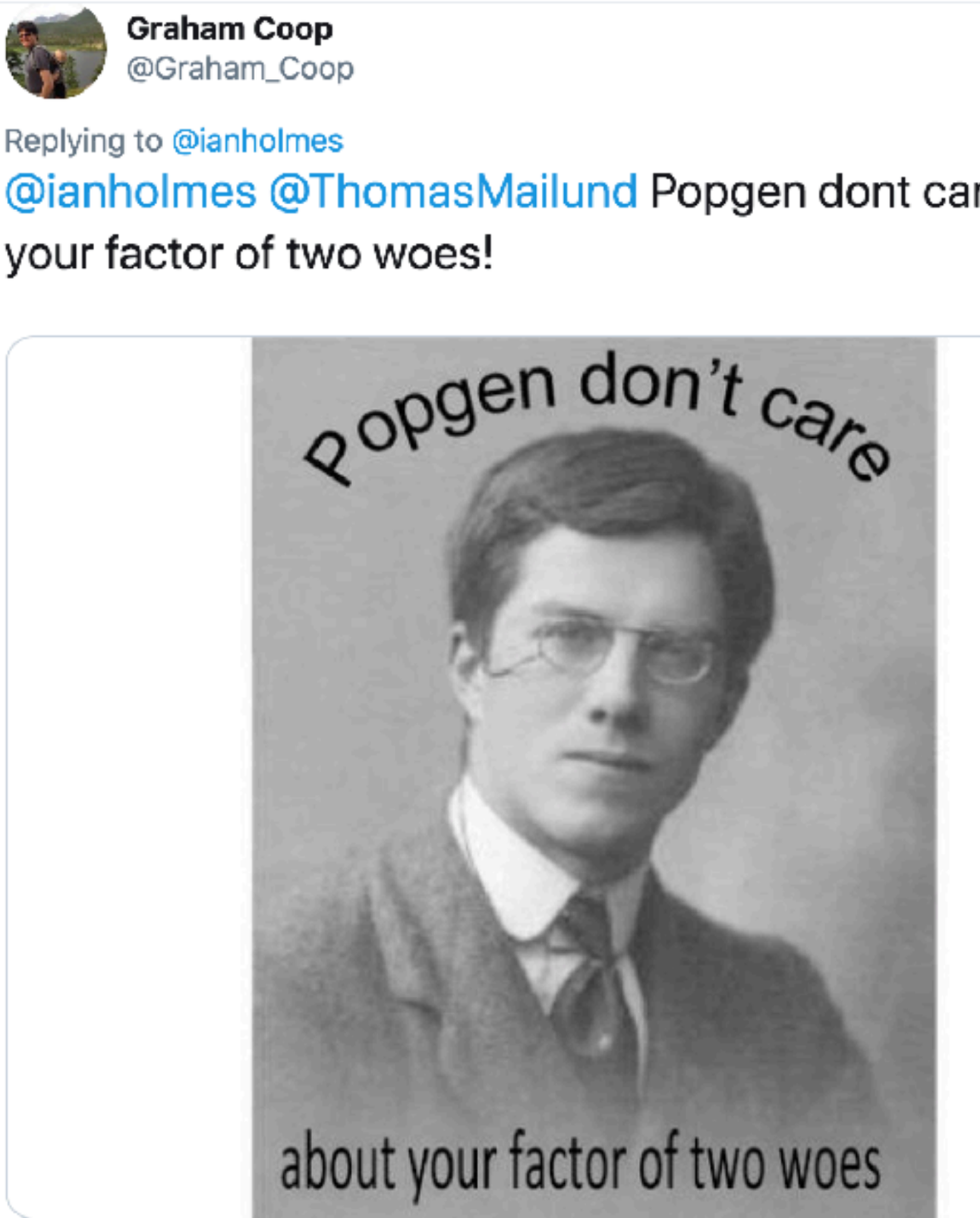
Ian Holmes @ianholmes · Jan 9, 2016

@Graham_Coop @ThomasMailund you made this meme at one point didn't you? I am working through your notes & feel like I need this. For morale

4 1

I'm so confused…is it 1/*N* or 1/2*N* ?

depends whether you're talking about diploids or not



Moran Model

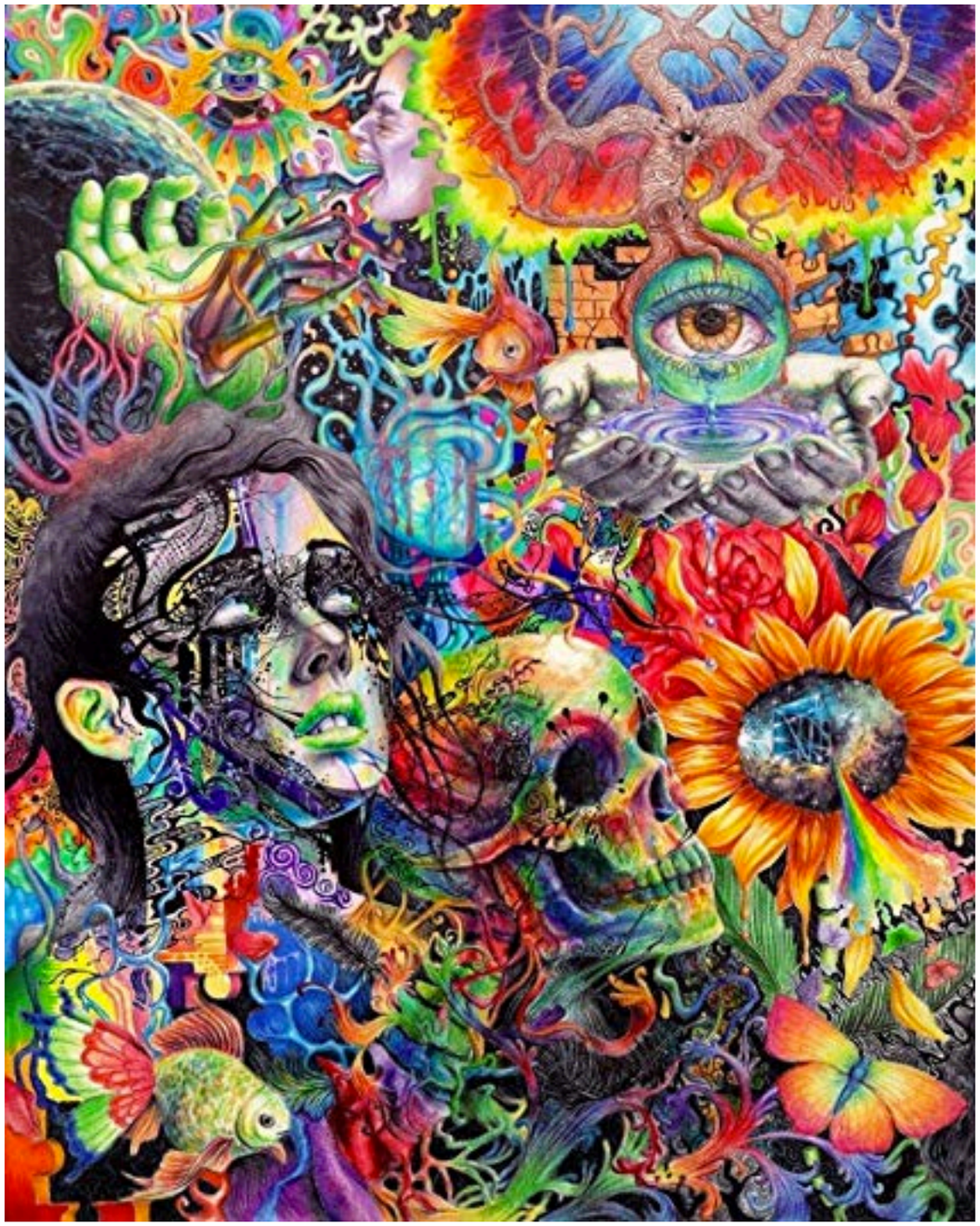
repeat birth/death process for 2N time points and you have
1 generation in the Wright-Fisher Model

and …

Moran Model

repeat birth/death process for 2N time points and you have 1 generation in the Wright-Fisher Model

and …



Moran model of genetic drift to loss/ fixation

In a population of N individuals, a new mutation starts with frequency $1/N$

Each generation, one individual is chosen to reproduce and one is chosen to die

Moran Model

repeat birth/death process for 2N time points and you have 1 generation in the Wright-Fisher Model

and …

$$E(p_{t+1}) = p_t$$

$$Var(p_{t+1}) = 2p_tq_t/2N$$

variance is 2x as large in Moran model b/c individuals can reproduce at multiple time points due to the 2N (and drift is 2x as fast).

An estimator of effective population size

- Heterozygosity θ = empirical probability that two human genomes differ at a random site (~1/1000)

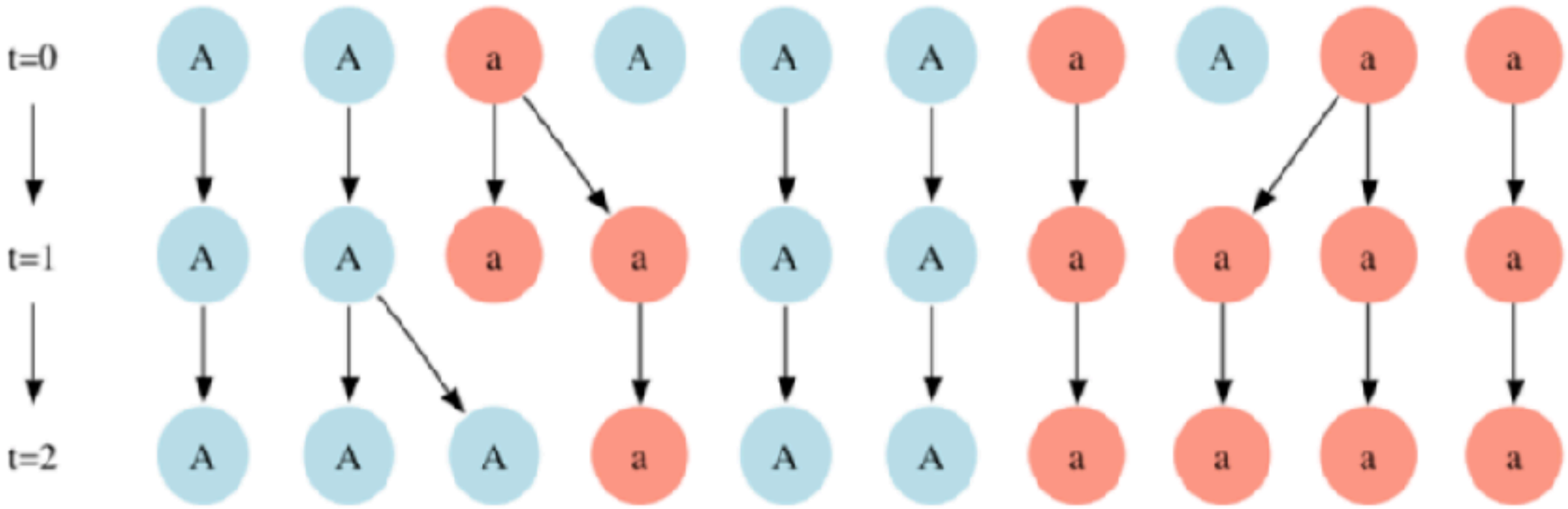
- Depends on rate of germ line mutations and length of time each mutation stays variable before fixing or dying out

- Humans accumulate about 1 germ line mutation per 10^8 base pairs per generation (mutation rate $\mu=10^{-8}$)

- We can use the Moran model to estimate the effective population size from heterozygosity measured in present-day humans

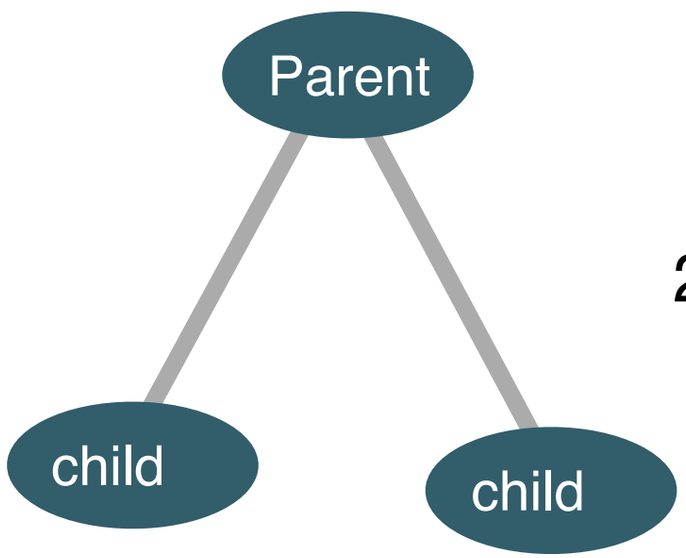
Deriving the relationship between heterozygosity and the effective population size

- What is the probability that two alleles share a common ancestor 1 generation ago? 1/2N



Deriving the relationship between heterozygosity and the effective population size

- What is the probability that two alleles share a common ancestor 1 generation ago? 1/2N
- What is the probability that one of these alleles originated as a new mutation last generation? 2μ



2 * per generation rate

Deriving the relationship between heterozygosity and the effective population size

- What is the probability that two alleles share a common ancestor 1 generation ago? 1/2N

- What is the probability that one of these alleles originated as a new mutation last generation? 2μ

- What is the probability that these two alleles are identical and also share a common ancestor exactly *t* generations ago? $(1-2\mu)^{t-1}*(1-1/2N)^{t-1}*1/2N$

Deriving the relationship between heterozygosity and the effective population size

- What is the probability that two alleles share a common ancestor 1 generation ago? 1/2N

- What is the probability that one of these alleles originated as a new mutation last generation? 2μ

- What is the probability that these two alleles are identical and also share a common ancestor exactly *t* generations ago? $(1-2\mu)^{t-1}*(1-1/2N)^{t-1} *1/2N \sim 1/2N *\exp(-t (2\mu+1/2N))$

Deriving the relationship between heterozygosity and the effective population size

---

The probability that our two alleles are identical is:

$$\frac{1}{2N}\int_0^\infty e^{-t(2\mu+1/(2N))}dt = \frac{1/(2N)}{1/(2N)+2\mu} = \frac{1}{1+4N\mu}$$

Deriving the relationship between heterozygosity and the effective population size

---

The probability that our two alleles are identical is:

$$\frac{1}{2N}\int_0^\infty e^{-t(2\mu+1/(2N))}dt = \frac{1/(2N)}{1/(2N)+2\mu} = \frac{1}{1+4N\mu}$$

The probability that the individual is heterozygous, i.e. the two alleles are not identical, is

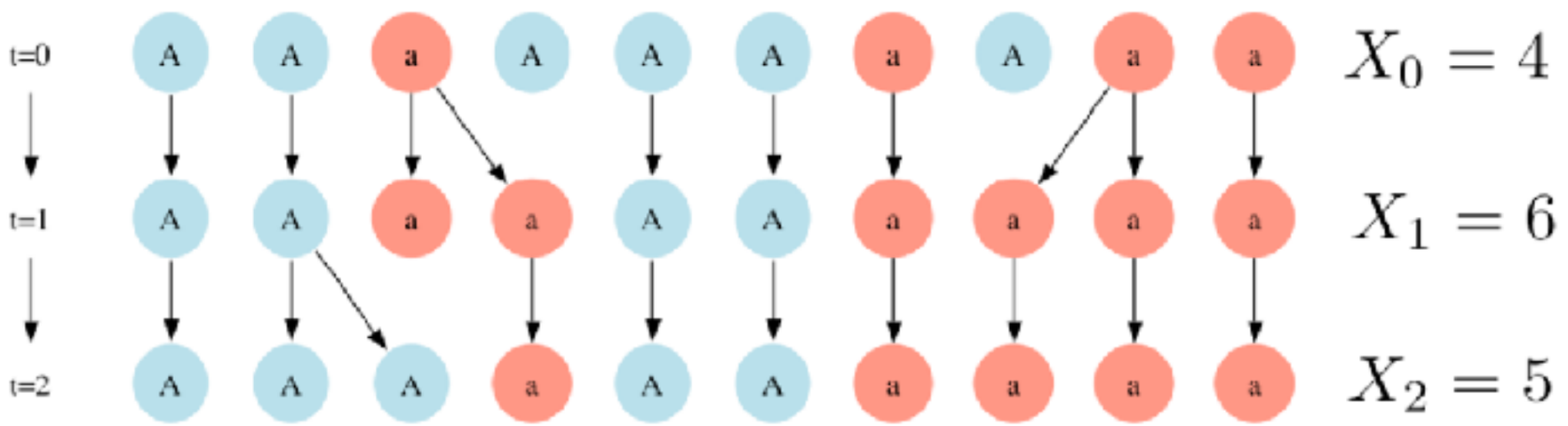
$$\frac{4N\mu}{1+4N\mu}$$

identity by descent

Let's consider a population, at some fixed time-point (t=0), which has $2N$ allele copies in a population. Call this the *reference population*.

Definition

In subsequent generations, we will ask whether two random alleles are descendant from the same reference allele in the reference population. If they are, we call them *identical-by-descent* (IBD).



the probability of identity by descent through time

- Let F_t equal the probability two randomly sampled alleles are IBD in the previous generation.
- If in generation t, we sample one random allele from a population, the probability the second allele is identical by descent is:
  - With probability $\frac{1}{2N}$, they are descendant from the same allele in the previous generation. They are IBD w/ prob 1.
  - WIth probability $1 - \frac{1}{2N}$, they are descendant from different alleles in the previous generation. They are IBD w/ prob F_{t-1}

Thus:

$$F_t = \frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}$$

Which, because $F_0 = 0$ here, this can be solved as:

$$F_t = 1 - (1 - \frac{1}{2N})^t$$

the probability of identity by descent through time

Let's look carefully:

$$F_t = 1 - (1 - \frac{1}{2N})^t$$

As t gets large, $(1 - \frac{1}{2N})^t$ goes to zero, and so F_t goes to 1.

NOTE

If the probability of identity by descent for two random alleles goes to 1, this implies all alleles must be descendant from the same reference allele.

Two important implications:

1. All alleles trace back to a common ancestor allele (i.e. prelude to *coalescent theory*).
2. With no mutation we eventually expect all alleles to be identical!