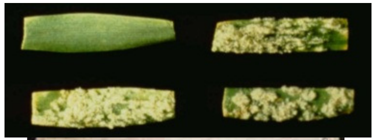
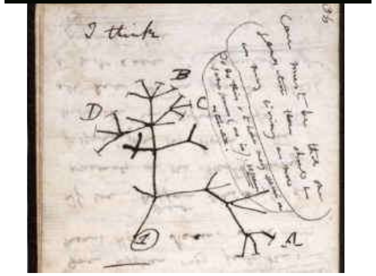
Wright-Fisher model and neutral theory

S. chilense





$$f_i = \frac{\theta}{i}\binom{n-1}{i}^{-1}\sum_{k=2}^{n-i+1}\binom{k}{2}\binom{n-k}{i-1}E(T_k), \quad 1 \le i \le n-1.$$

S. peruvianum

$$E(T_k) = \sum_{j=k}^{n}(-1)^{j+k}\frac{(2j-1)n!(n-1)!(j+k-2)!}{k!(k-1)!(n-j)!(n+j-1)!(j-k)!}\int_0^\infty \exp\left(-\beta_1^2\binom{j}{2}\int_0^t 1/\rho(s)ds\right)dt.$$

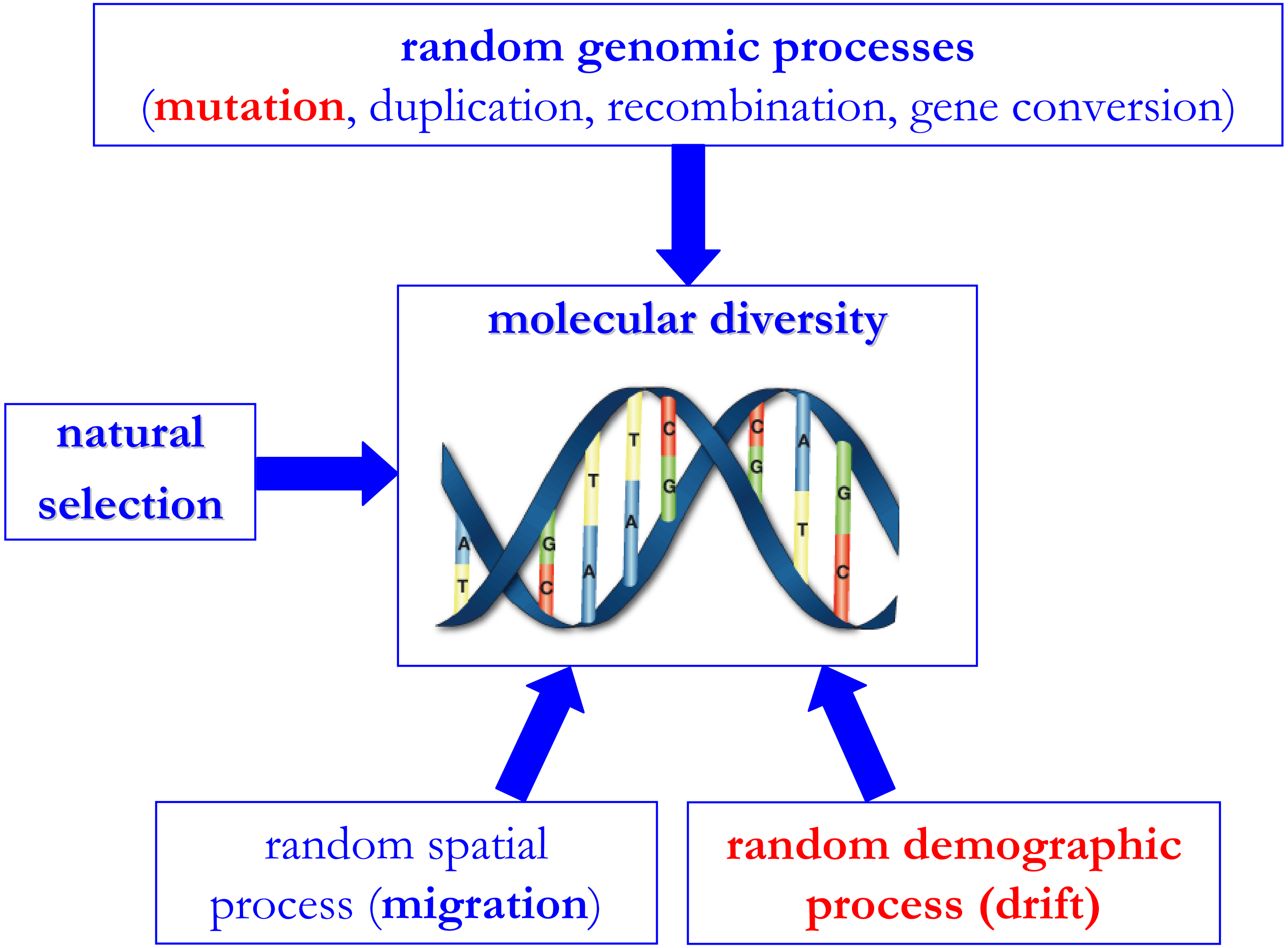
Synbreed Winter School 2012

Prof Aurélien Tellier

FG Populationsgenetik

Technische Universität München

Population genetics: 4 evolutionary forces



random genomic processes
(**mutation**, duplication, recombination, gene conversion)

molecular diversity

natural selection

random spatial process (**migration**)

random demographic process (drift)

Wright-Fisher model, genetic drift and neutral theory

Neutral theory

- Introduced by Motoo Kimura in 1960s, big controversy at the time.
- *Can we explain all polymorphism data without the action of natural selection?*
- Kimura: Most polymorphisms that occur do not influence the fitness of an individual,
- thus these polymorphisms are not subjected to selection
- **these mutations would evolve neutrally**

- **mutations at silent or degenerate sites do not change the Amino Acid BUT may or may not evolve neutrally**
- **some non-synonymous mutations do not affect fitness (change in Amino Acid does not affect the fitness) BUT may or may not evolve neutrally**

Neutral theory

- Neutral theory = most changes in allele frequencies in a population can be attributed to **genetic drift**
- Why?
- When a mutation arise in a gamete of an individual, many things can happen:
  - the carrier of the gamete must survive to reach the reproductive age,
  - that gamete must be fertilized and develop an embryo,
  - the embryo has to be viable to be at the next generation.

- **Genetic drift means that mutation creates new alleles which by chance**
  - **can rise in frequency and spread in a population,**
  - **or they can get lost.**

Neutral theory

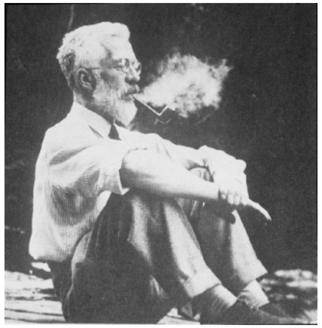
- We need a model to explain how genetic drift occurs
- and then use it to derive expectations on what polymorphism we should observe in DNA sequences

- This model is based on how a population of individuals reproduce over time

- **Important: to demonstrate that a given trait or polymorphism pattern is due to selection, you MUST disprove alternative neutral explanations!**

The Wright-Fisher model

The Wright – Fisher model



- Fundamental model in population genetics
- ***Assumptions (check list):***
  - **Constant population size**
  - **Discrete and non-overlapping generations**
  - **Random mating (= panmixia)**
  - **(Equal sex-ratio)**
  - **2N haploid individuals = N diploid individuals (with two allele each)**
  - **One locus**
  - **No intra-locus recombination**

The Wright – Fisher model

- How does it work?
- Let us assume 10 haploid individuals at generation *t*
- The offspring generation is obtained from the parents as follows:
  - constant population => 10 individuals at generation *t+1*
  - each individual from the offspring picks a parent at random from generation *t*
  - connect parent and child by a line
  - each offspring inherits the genetic information of the parents

The Wright – Fisher model

- all individuals have the same fitness (= expected number of offsprings)
- this means that there is an equal probability for all individuals to be picked as a parent

- ecah individual has **2N** chances to become ancestor to generation *t+1*
- and the probability to be picked at each draw is **1/2N**

- **you are the offspring choose one parent at random from previous generation (2N=6) => throw dice**
- **you are the parents, for each offspring you have 1/2N chance to have it as yours => throw dice**

The Wright – Fisher model

present

past

Figure 2.1: The 0th generation in a Wright-Fisher Model.

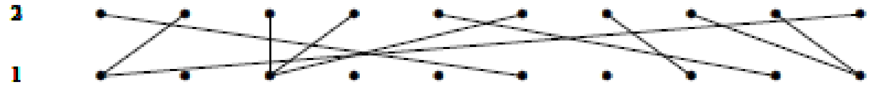


Figure 2.2: The first generation in a Wright-Fisher Model.

The Wright – Fisher model

present

past

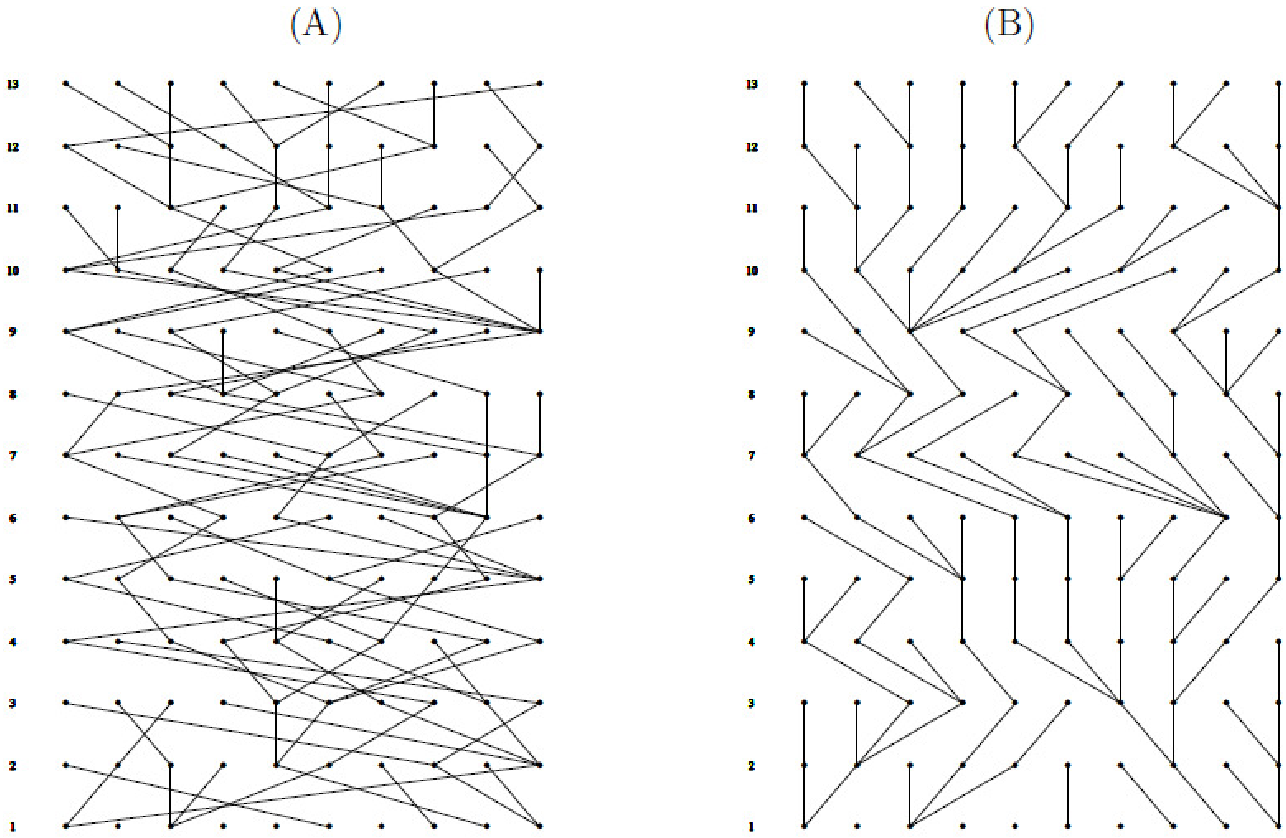


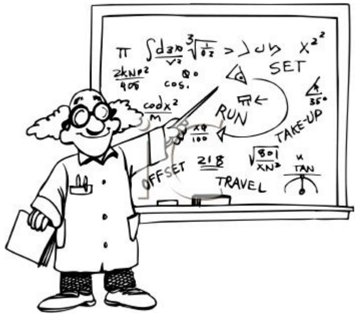
Figure 2.3: The tangled and untangled version of the Wright-Fisher Model after some generations.

The Wright – Fisher model

- all individuals have the same fitness (= expected number of offsprings)
- this means that there is an equal probability for all individuals to be picked as a parent

- each individual has **2N** chances to become ancestor to generation *t+1*
- and the probability to be picked at each draw is **1/2N**

The Wright – Fisher model



Maths 1: Binomial distribution

If a random variable X is binomial distributed with parameters ***p*** and ***n***

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$E(X) = np$$

$$Var(X) = np(1-p)$$

Where ***n***=2*N* and ***p***=1/2*N*

Maths 2: Poisson distribution

If a random variable X is binomial distributed with parameters ***p*** and ***n*** and **λ=*np*** has a reasonable size

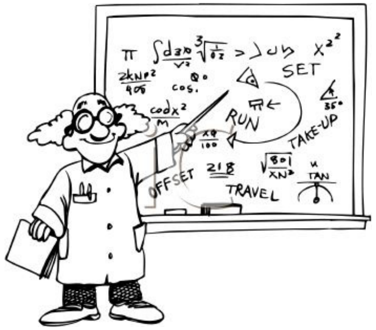
$$P(X = k) \approx e^{-\lambda} \frac{\lambda^k}{k!}$$

$$E(X) = np$$

$$Var(X) = np$$

If 2*N* is large, the distribution of number of offsprings is almost Poisson distributed with **λ** =***np*** = ***1***

The Wright – Fisher model



Maths 1: Binomial distribution

If a random variable X is binomial distributed with parameters ***p*** and ***n***

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

$$E(X)=np$$

$$Var(X)=np(1-p)$$

Where ***n***=2*N* and ***p***=1/2*N*

The variance of offspring number per individual = 1 – 1/2*N*

Maths 2: Poisson distribution

If a random variable X is binomial distributed with parameters ***p*** and ***n*** and ***λ=np*** has a reasonable size

$$P(X=k)\approx e^{-\lambda}\frac{\lambda^k}{k!}$$

$$E(X)=np$$

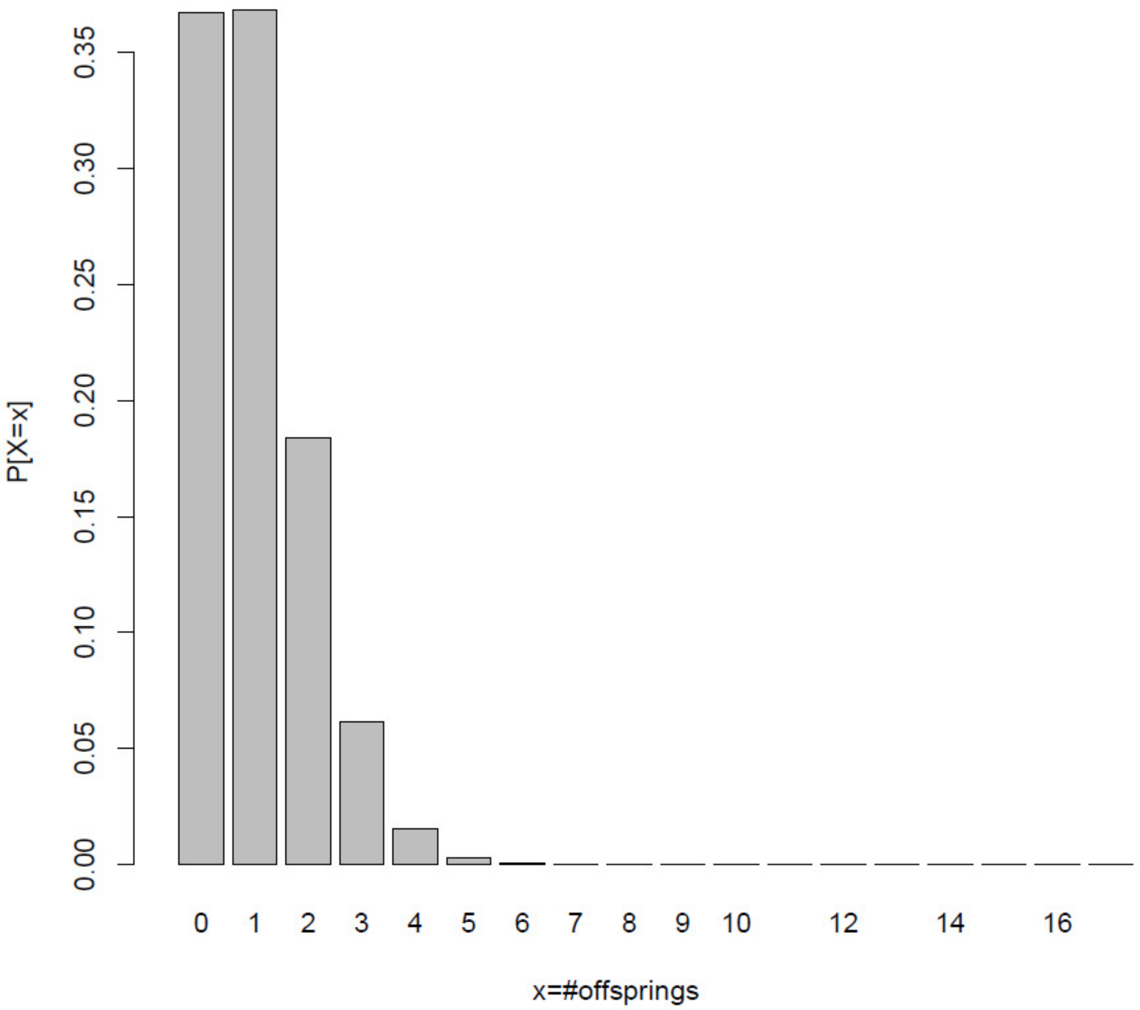
$$Var(X)=np$$

If 2*N* is large, the distribution of number of offsprings is almost Poisson distributed with ***λ*** =***np*** = ***1***

Offsprings distribution in WF

Binomial distribution: dbinom in R (here with $2N$=500)

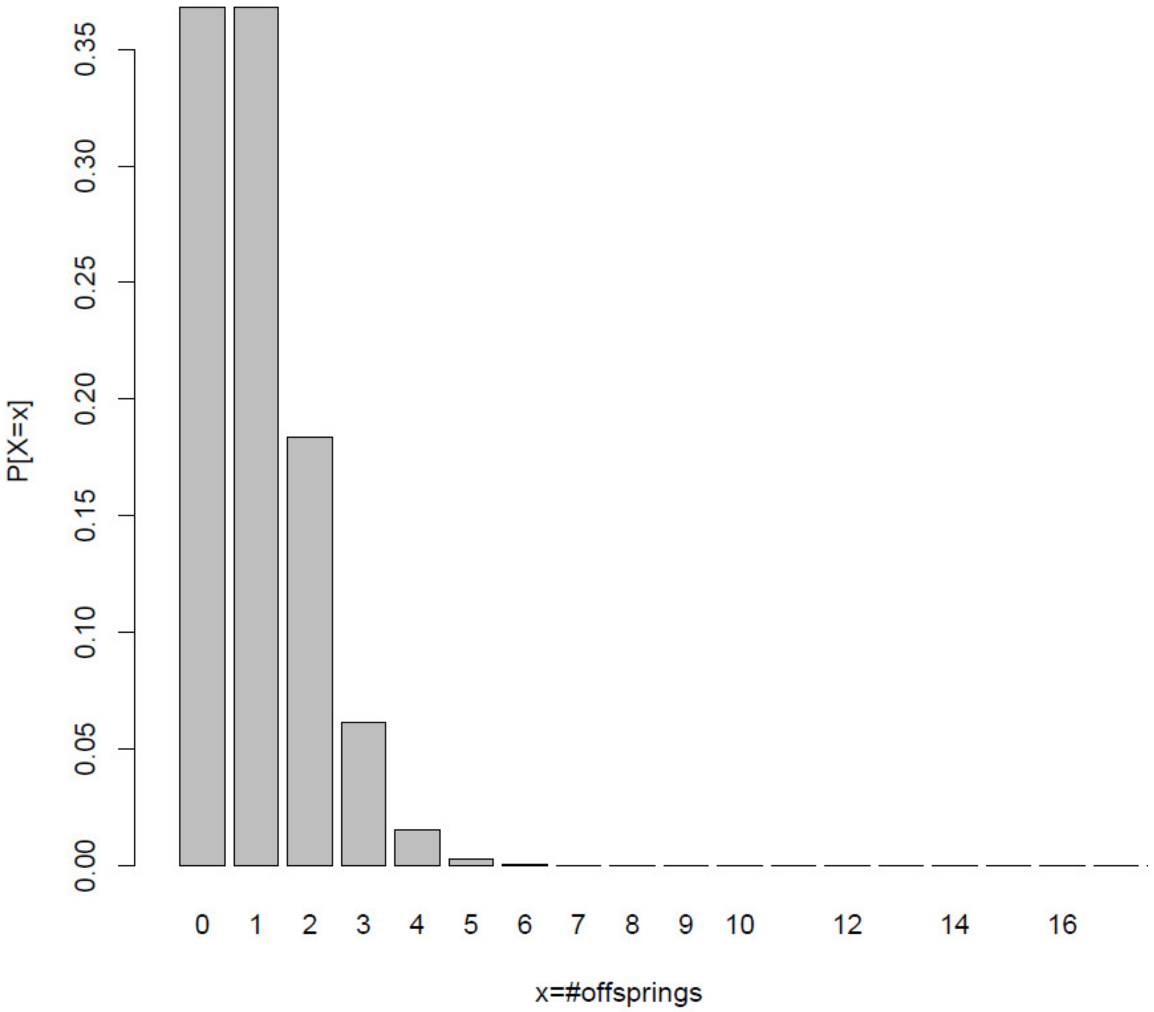
The variance of offspring distribution is finite and small compared to 2N



Offsprings distribution in WF

Poisson distribution: dpois in R (here with $2N$=500, λ=1)

The variance of offspring distribution is finite and small compared to 2N



Probability in the Wright – Fisher model

- From the Poisson distribution, the probability of an individual not to leave descendants is:
- **$P[X=0] = e^{-1} \approx 0.37$**

- A fraction $1 - 0.37 = 63\%$ of all individuals have descendants at each generation
- in a randomly mating population, the present day population descent from a small fraction of individuals a few generations ago
- this fraction is approximately $\approx 0.63^t$

- **For a population fo size $2N = 10{,}000$**
- **The population comes from: $10{,}000 \times 0.63^{15} \approx 10$ individuals from 15 generations ago**
- **The 9,990 other individuals did not leave descendants today**

Genetic drift

Genetic drift

- It is a random stochastic process!!!!
- **What it means for evolutionary biology: even if we know everything about a population and ist biology, we cannot predict the state of the population in the future**

- In fact we do not need to pick a random parent for each individual one by one:
- we can pick up this number from the **binomial distribution**, and use this as the frequency at next generation

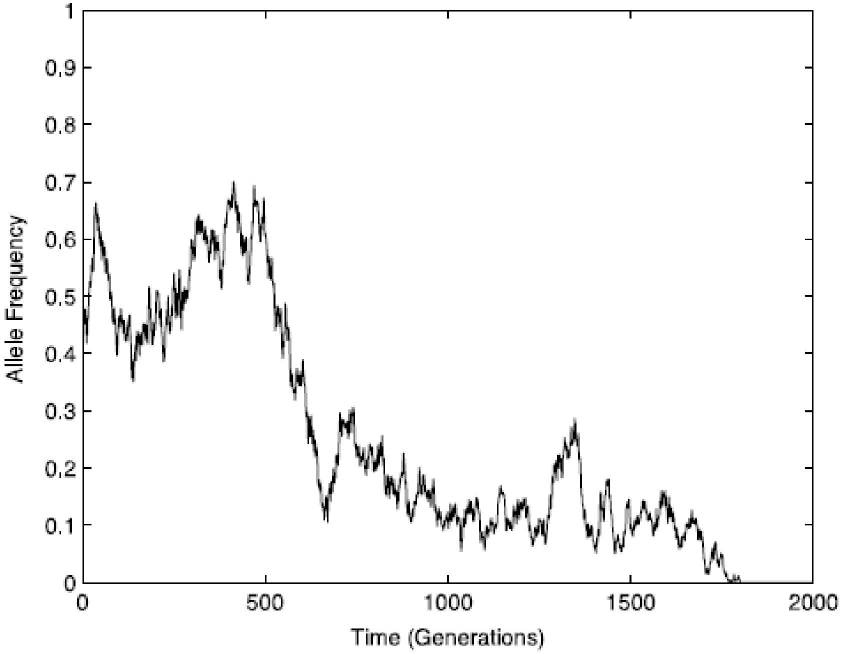
p = frequency of allele A $\quad\quad$ $1-p$ = frequency of allele a

- frequencies at last generation in population with $2N$ individuals

- if more than two alleles we use the multinomial distribution

Genetic drift

- Genetic drift = random change of allele frequency between generations
- The probability of loosing/fixing alleles is HIGHER for small *N* (population size)
- If loci are independent in the genome (physically not linked), each locus has independent changes in allele frequency under genetic drift!!!!



(a) Wright–Fisher.

Der et al. 2011 TPB

The coalescent

Back to the Wright – Fisher model

- Until now we have predicted the state of population at *t+1* based on time *t*

  This is the process forward in time

- Useful because it is logical and intuitive

- **However, we can also follow the genealogy backward in time from present to past**
- Why????
- because most data we collect come from present day populations that we can sample

- **The question becomes: what are the forces that have shaped the observed patterns of diversity in our data (SNPs,…)??????????**
- **These forces have acted in the history of the population, so we look at the geneaolgy**

Back to the Wright – Fisher model

present

past

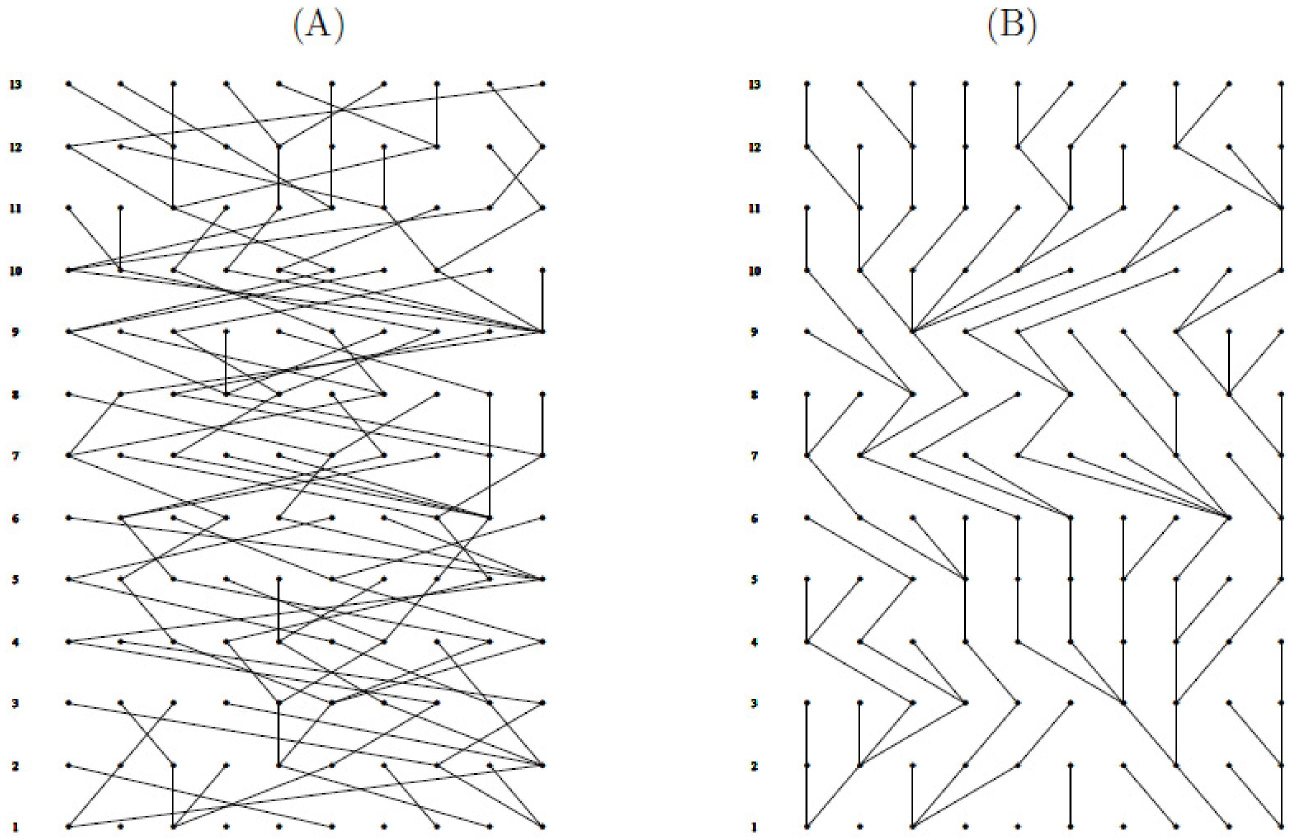


Figure 2.3: The tangled and untangled version of the Wright-Fisher Model after some generations.

Back to the Wright – Fisher model

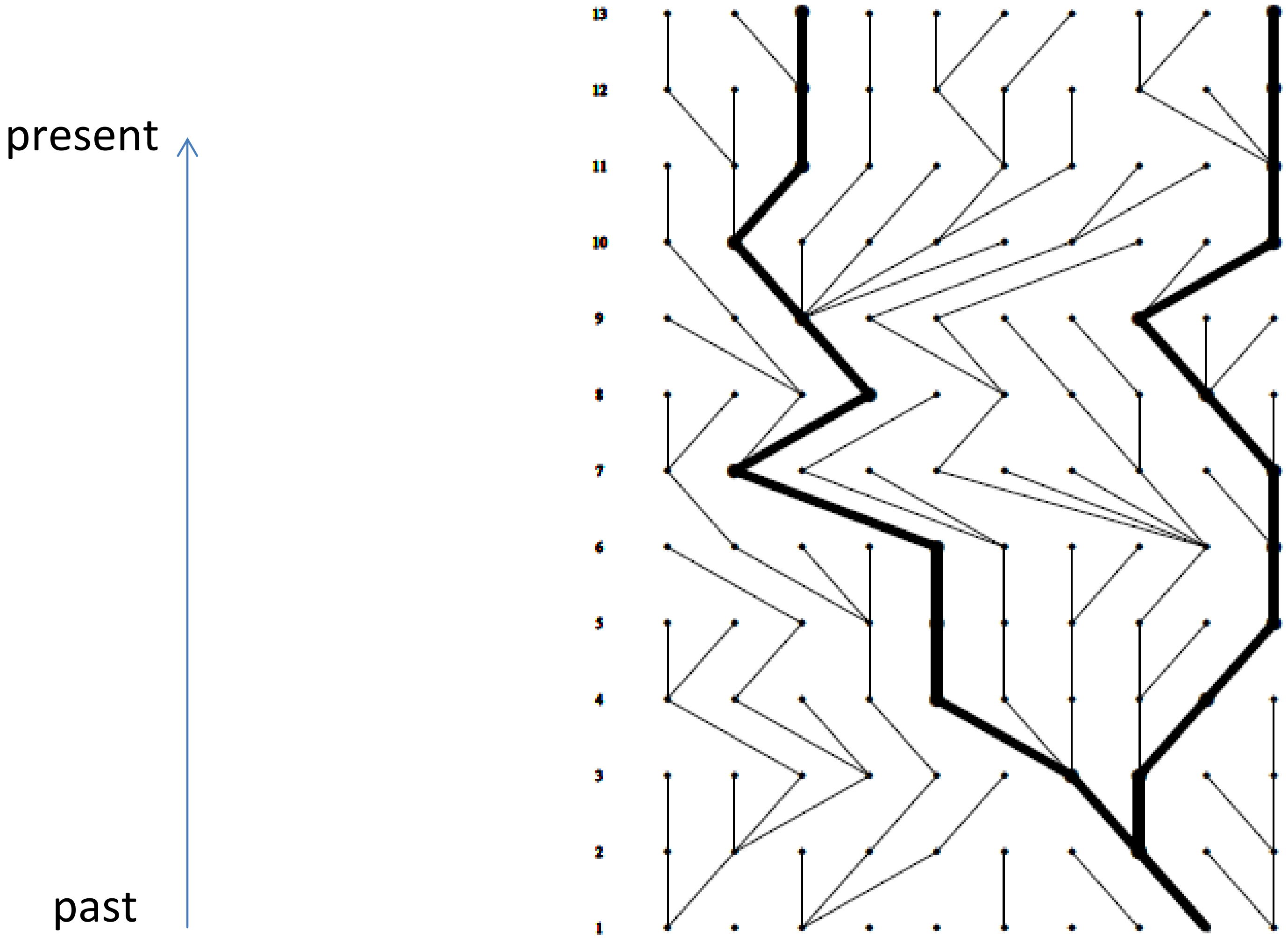


Figure 2.6: The coalescent of two lines in the Wright-Fisher Model

Back to the Wright – Fisher model

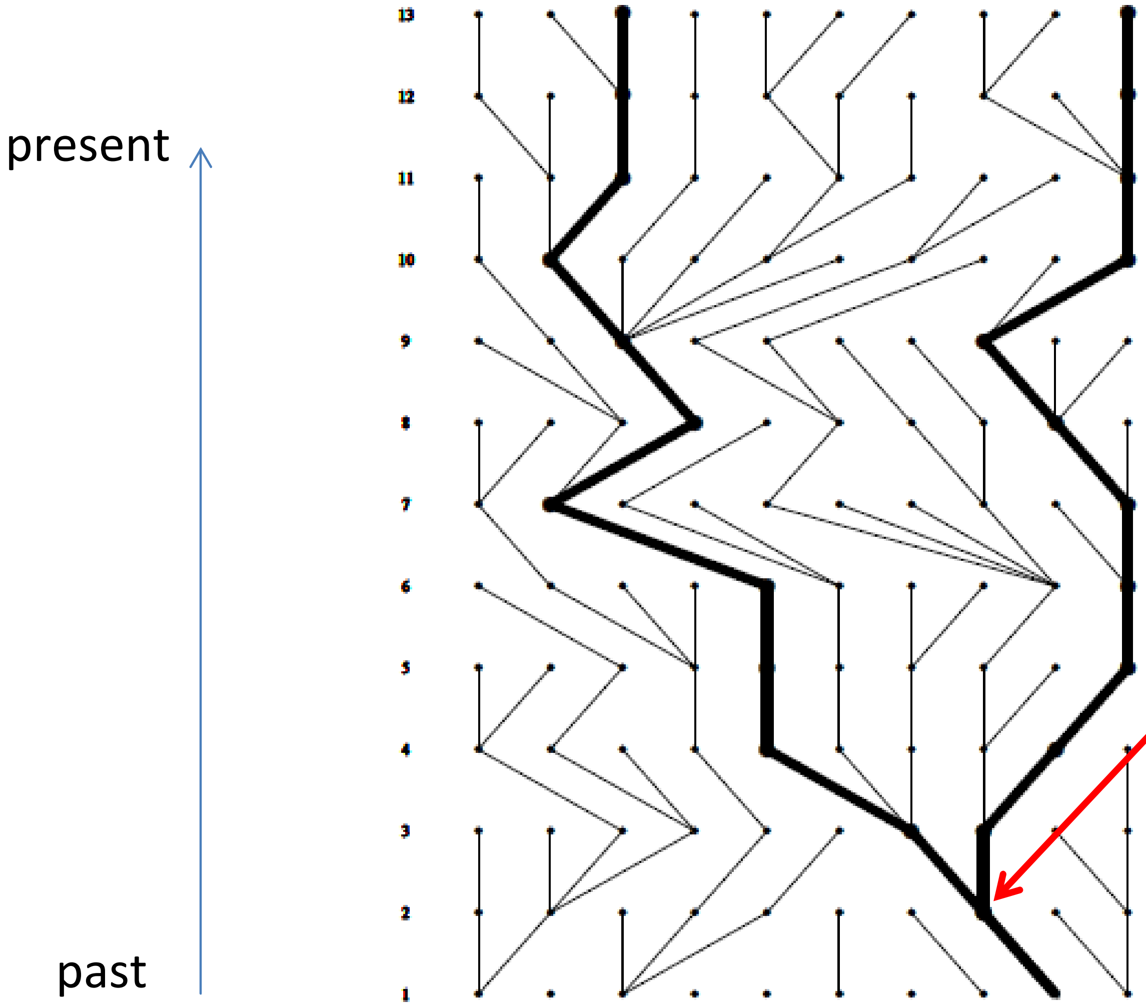


Figure 2.6: The coalescent of two lines in the Wright-Fisher Model

The genealogy of a sample of the population
Back to the so-called:

MRCA =
Most Recent Common Ancestor

of the sample

Because only a limited proportion of individuals leave descendants

The coalescent

- **Important assumption of the coalescent model:**

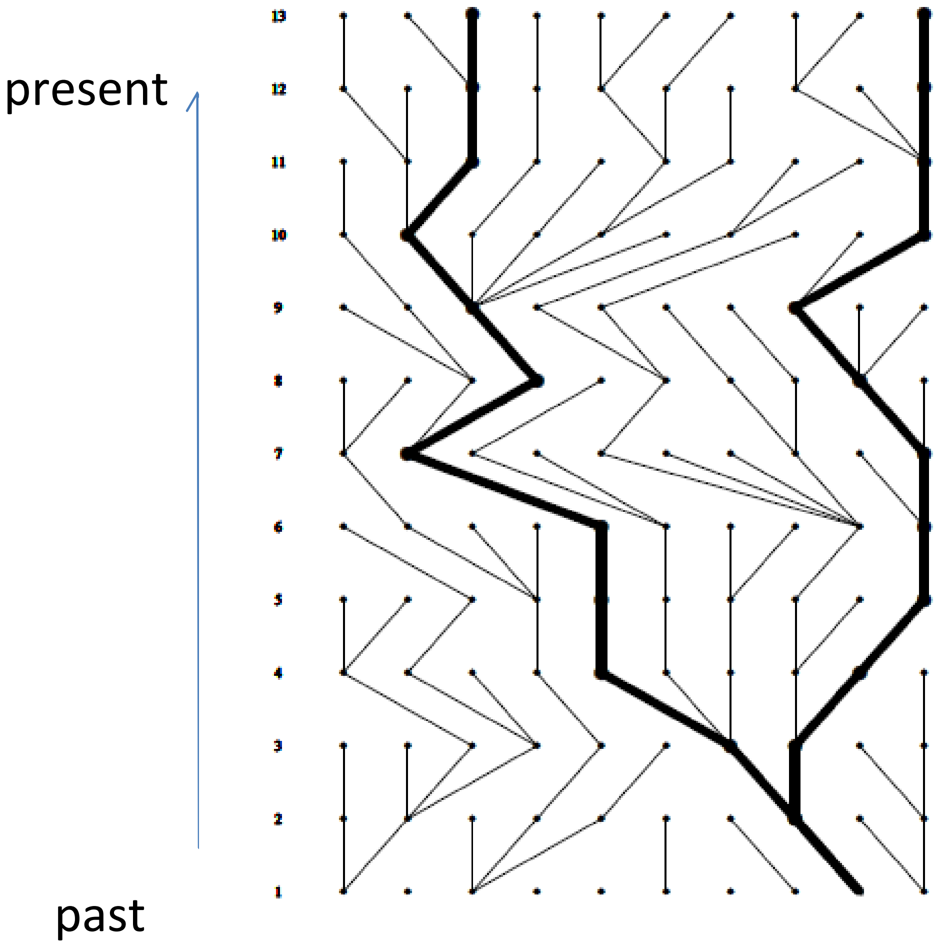


Figure 2.6: The coalescent of two lines in the Wright-Fisher Model

We assume that $n << 2N$

So that the probability of more than 2 lineages to coalesce is very small

P[3 lineages to coalesce] = $(1/2N)^2$

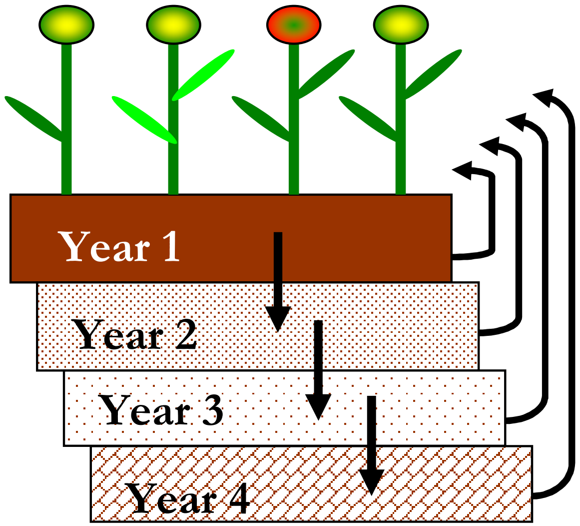
This is small enough to be be neglected if N is big, but not always true

More to come tomorrow on the coalescent….

Alternative models to Wright - Fisher

WF model with seed banks

- A model where some seeds (or eggs or spores) stay hidden for several generations



Germination
rate=b

$$N_e = N_{cs} \times (1/b)^2$$

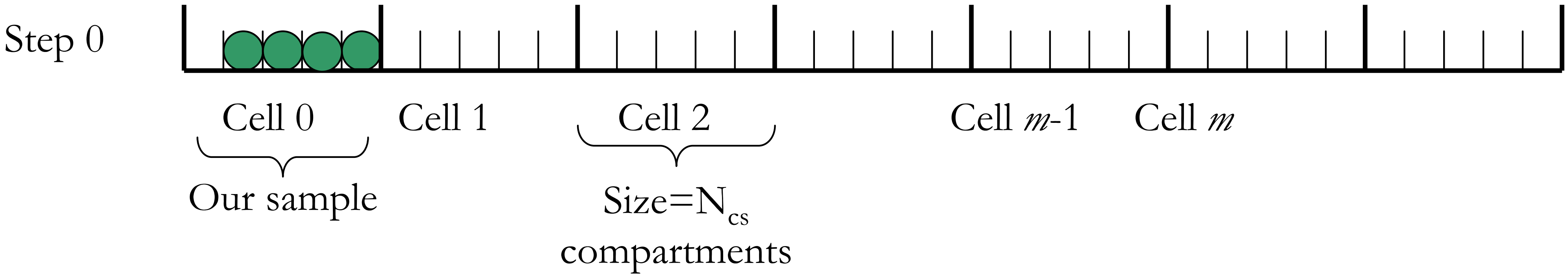
Number of individuals
per population

Germination
rate

- There is storage of diversity in the soil for several years *(Kaj et al. 2001, Vitalis et al. 2004)*
- This amounts to rescale the coalescent tree as follows:

Model for seed banks (1)

If we have no information about what lies in the seed bank, here **m = 5**



Step 0

Cell 0

Our sample

Cell 1

Cell 2

Size=N_{cs} compartments

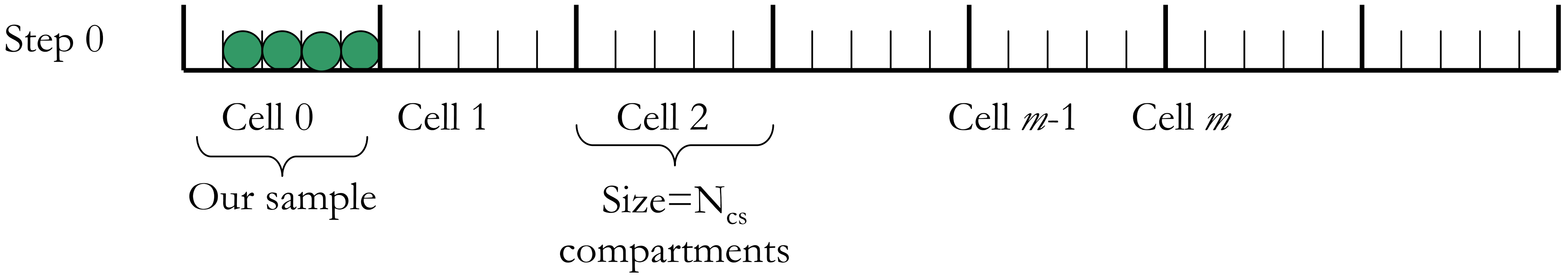
Cell m-1

Cell m

Kaj, Krone, Lascoux, J. Appl. Proba. 2001

Model for seed banks (2)

If we have no information about what lies in the seed bank, here ***m* = 5**

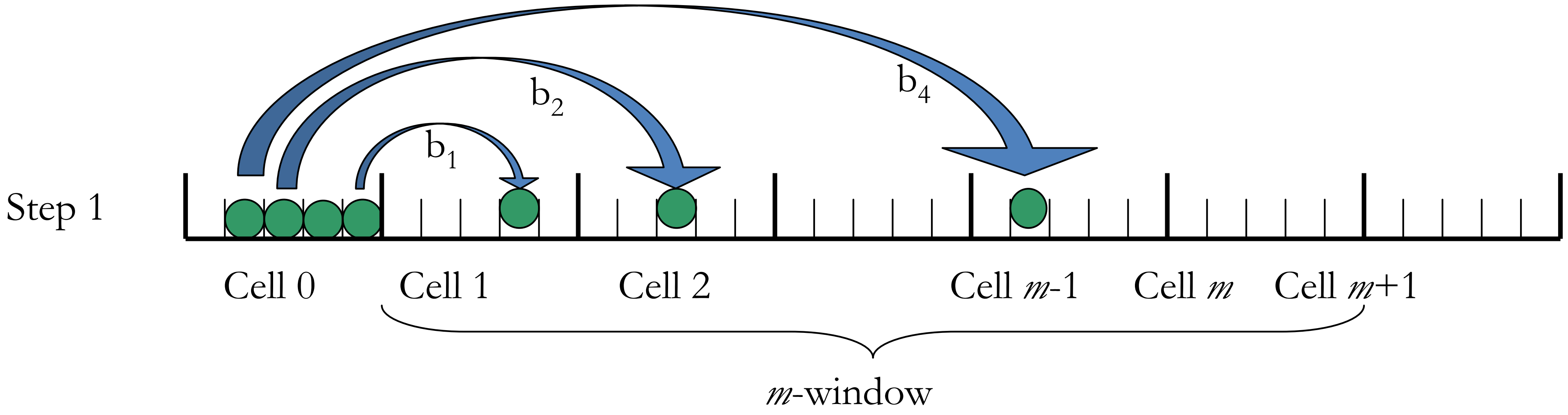


Step 0

Cell 0 Cell 1 Cell 2 Cell m-1 Cell m

Our sample

Size=N_{cs} compartments



Step 1

b_1 b_2 b_4

Cell 0 Cell 1 Cell 2 Cell m-1 Cell m Cell m+1

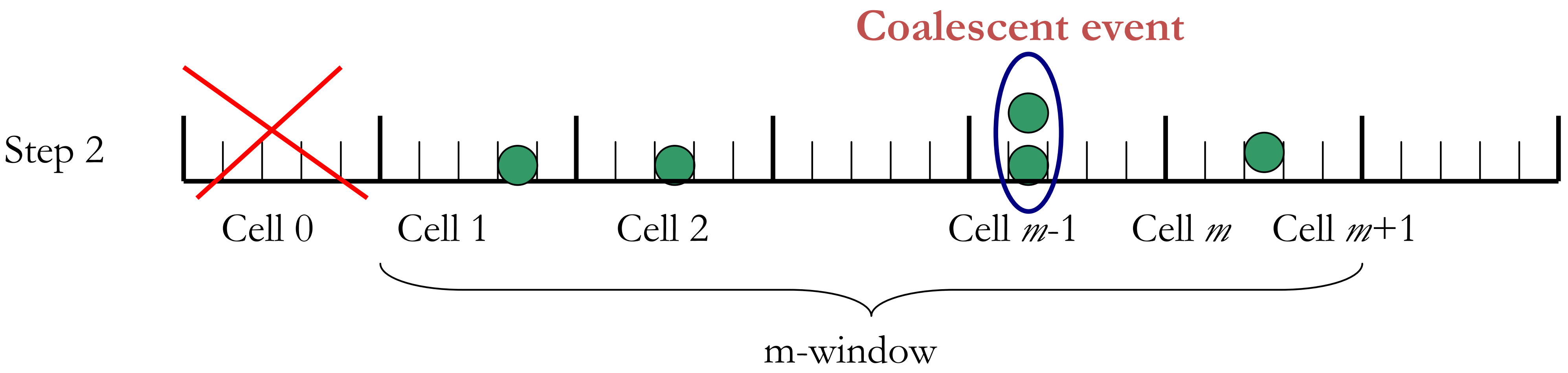
m-window

For all plants in cell 0

Kaj, Krone, Lascoux, J. Appl. Proba. 2001

Model for seed banks (3)

$m = 5$, if two seeds fall on the same compartment within a cell, there can be a coalescent event



Coalescent event

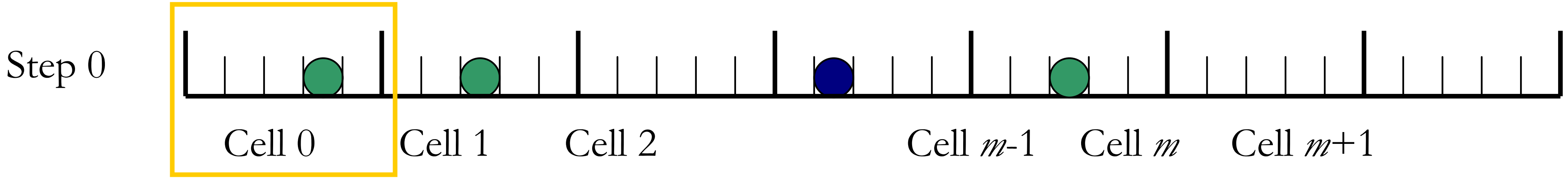
Step 2

Cell 0 Cell 1 Cell 2 Cell m-1 Cell m Cell m+1

m-window

At any moment, there are $\boldsymbol{r}$ lineages in the m-window

Then move every cell to the left, and start again step 0



Step 0

Cell 0 Cell 1 Cell 2 Cell m-1 Cell m Cell m+1

Kaj, Krone, Lascoux, J. Appl. Proba. 2001

Our model of coalescence

- The germination process is memoryless => the germination rate decrease geometrically with age of seeds
  - b = probability for a seed to germinate after one generation
  - $b(1-b)$ = probability for a seed to germinate after two generations
  - $b(1-b)^2$ = probability for a seed to germinate after three generations

- Mutation does not increase with age of seeds

- The rate of coalescence is rescaled by β^2 (the size of the genealogy is affected)

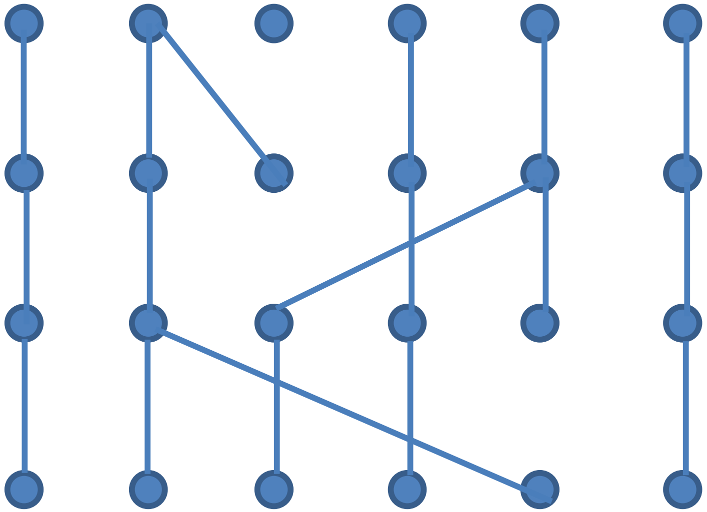
- The scaled mutation rate is scaled by β along a given ancestral line

- The recombination rate and migration rate between demes are also scaled by β

- β depends on b and m $$\beta = \frac{b(1-(1-b)^{m+1})}{1-(1+bm)(1-b)^m}$$

Moran model

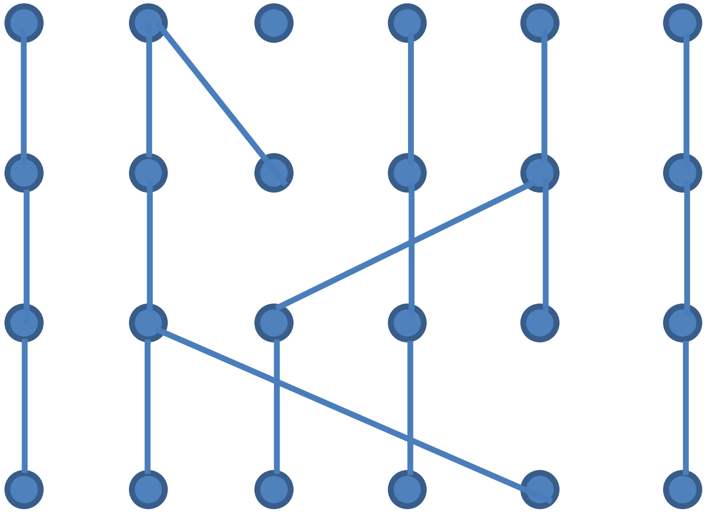
- In each generation, an individual is chosen to give birth to a new one, and one individual is chosen to die.



- As only one gene reproduces at each generation, the time scale is different from the Wright-Fisher model

Moran model

- In each generation, an individual is chosen to give birth to a new one, and one individual is chosen to die.



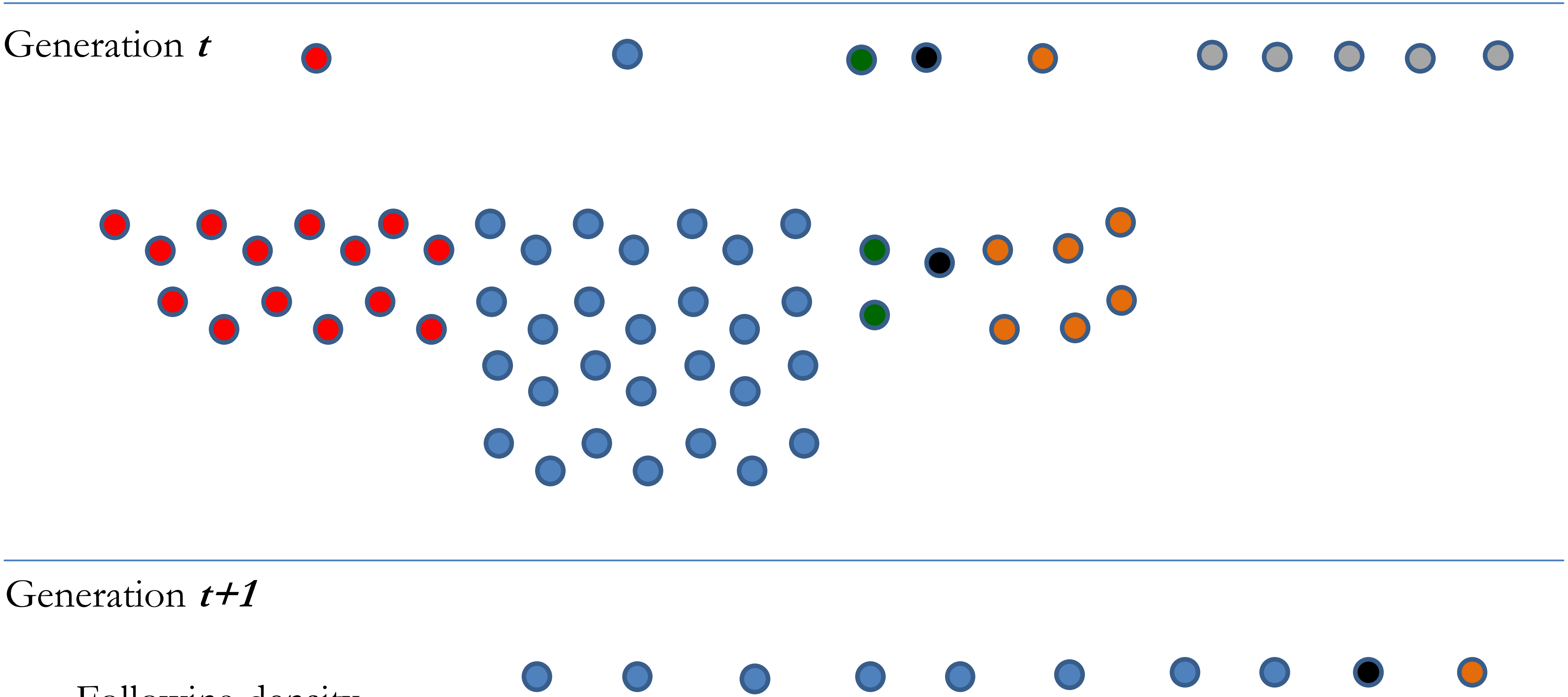
- Probability of two individuals to share a common ancestor at the previous generation = $1 / (N(2N-1))$
- The natural time scale is in units of $N(2N-1)$
- This is sometimes more convenient mathematically
- This model rules out multiple coalescent events

Cannings model - Schweinsberg

- ***Question:*** **is it realistic to assume that the variance in offspring number per individual has to be 1 – 1/2N (in the Wright – Fisher model)??**
- The Cannings model (is a general set-up from which can be derived the Moran and Wright-Fisher models)
- We look here at the model by Schweinsberg *(2003, Stoch Proc App)*

Cannings model - Schweinsberg

2N = 10



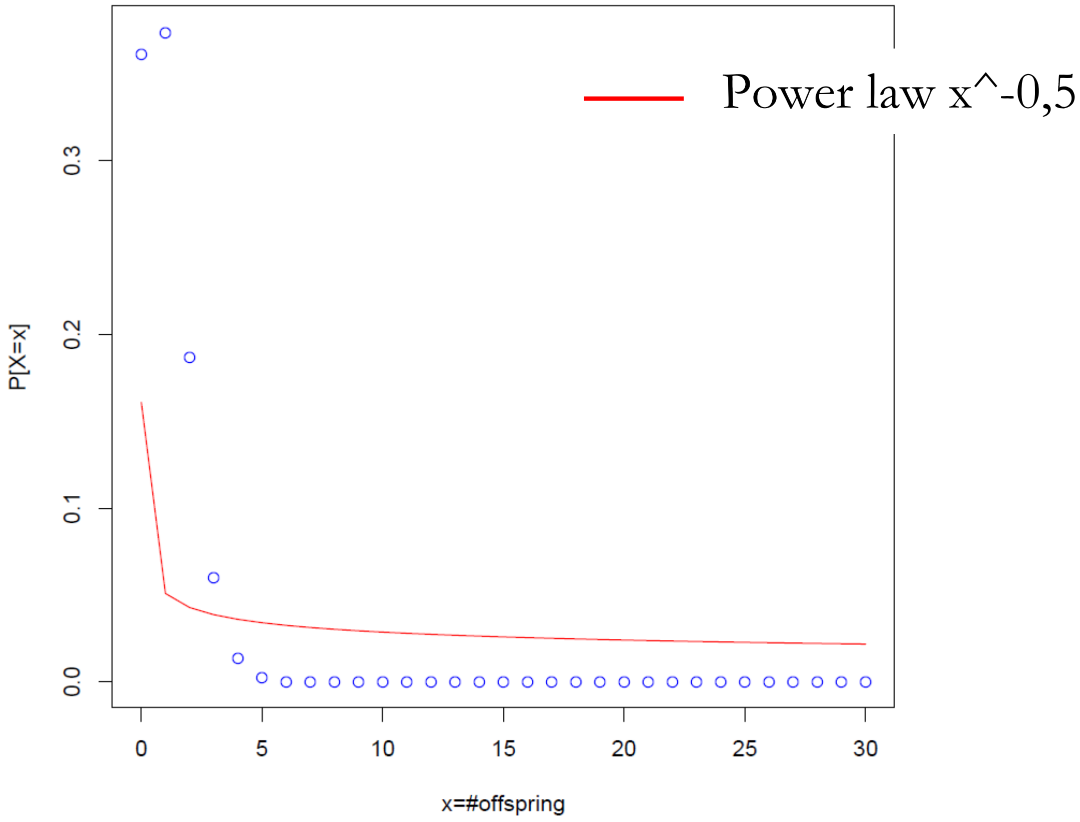
Generation *t*

Generation *t+1*

Following density-dependence in the population
Population of constant size

Cannings model - Schweinsberg

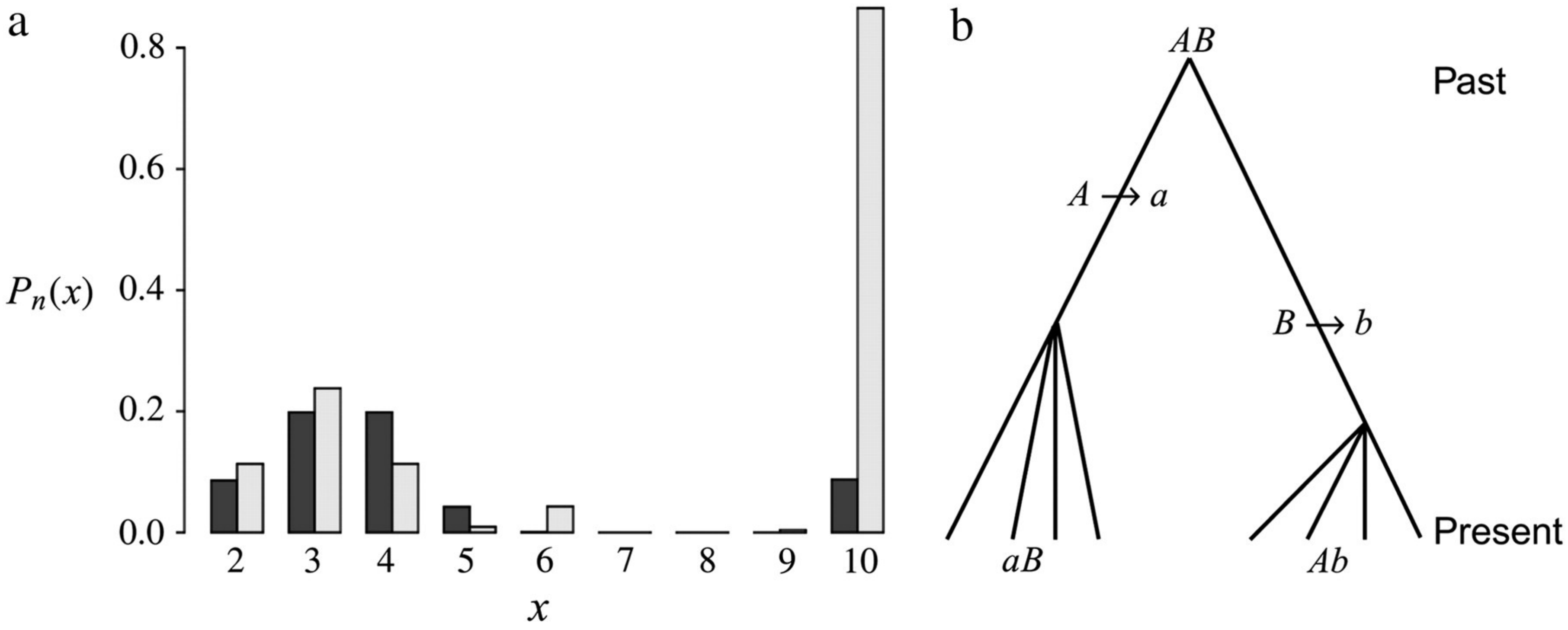
- We look here at the model by Schweinsberg (2003)
- Potentially the whole population can be made up of the offspring of one individual
- This behaviour depends on the distribution of offsprings per individual
- *Blue = binomial distribution as in Wright-Fisher*



Cannings model - Schweinsberg

- ➢ The variance in offspring production can be big
- ➢ This generates potentially multiple merger coalescent = multiple merging of lineages at one generation

Rates of coalescence and gene genealogy under a multiple-mergers coalescent.

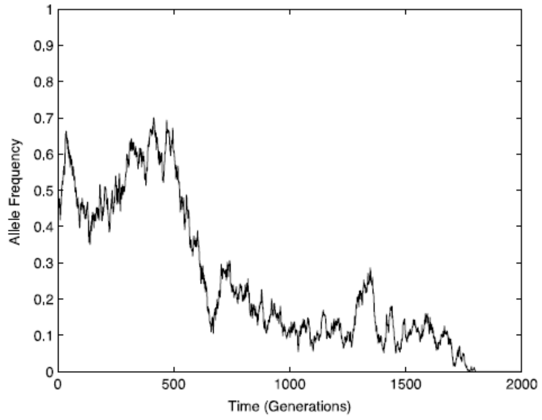


Eldon B , Wakeley J Genetics 2008;178:1517-1532

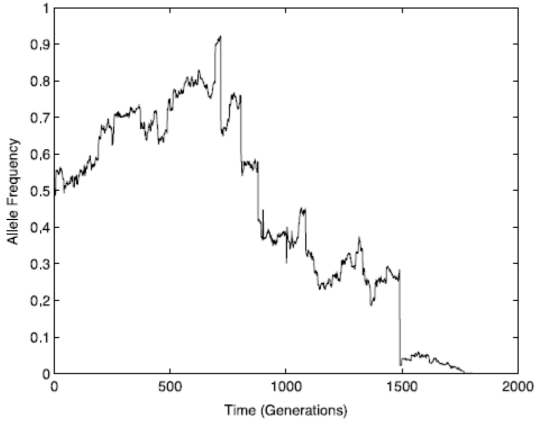
GENETICS

Cannings model - Schweinsberg

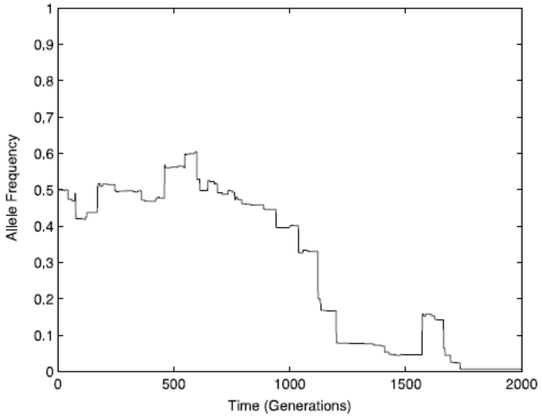
- Different types of multiple merger coalescent:
    - Λ-coalescent and Beta-coalescent, Bolthausen –Snitzman coalescent
    - Ξ-coalescent
- These models may be useful when hypotheses from the Wright-Fisher model are violated
- The rate of drift is also affected



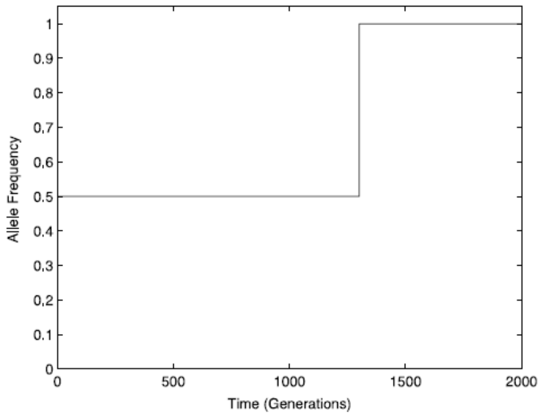
(a) Wright–Fisher.



(b) Power-law process ($\alpha = 2.5$).



(c) Power-law process ($\alpha = 1.5$).



(d) Λ_1 process.

Der et al. 2011 TPB

More to come about the effect of population structure, inbreeding, demography and selection on genealogies ….