BIOL 502 Population Genetics Spring 2017

Week 3 Genetic Drift and Coalescence

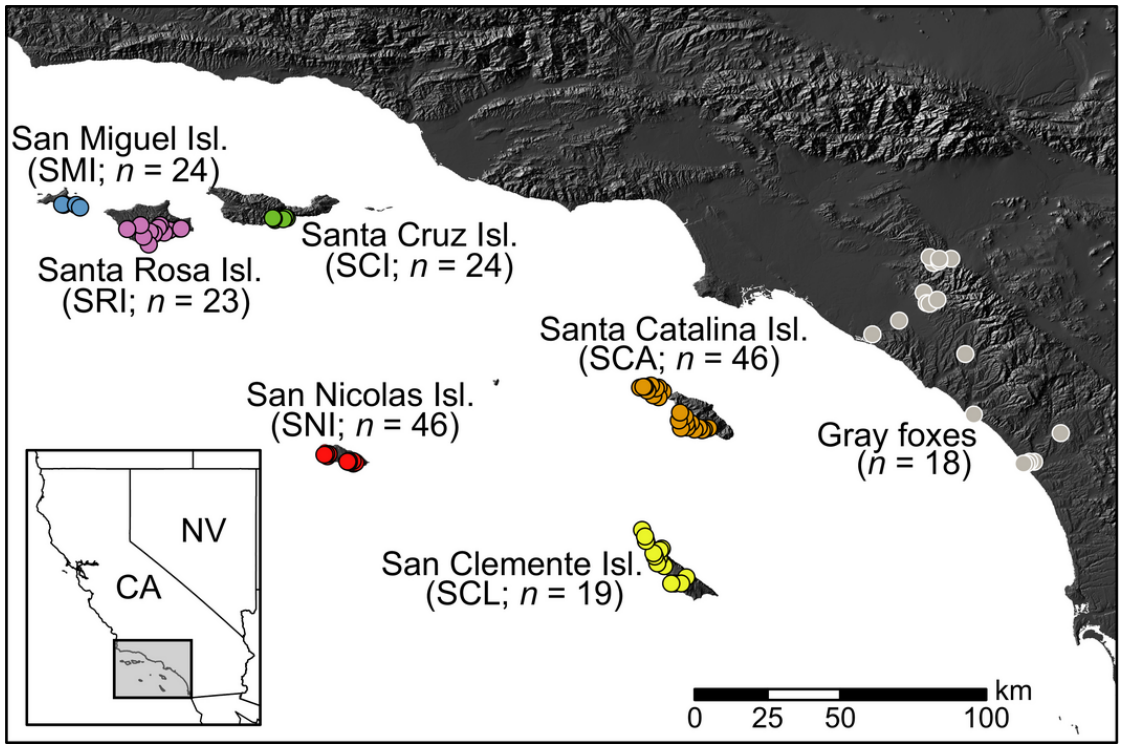
Arun Sethuraman

California State University San Marcos

Table of contents

Island Fox History - 1 - Funk et al. 2016 10.1111/mec.13605

Adaptive divergence despite strong genetic drift: genomic analysis of the evolutionary mechanisms causing genetic differentiation in the island fox (*Urocyon littoralis*)



Molecular Ecology
Volume 25, Issue 10, pages 2176-2194, 5 APR 2016 DOI: 10.1111/mec.13605
http://onlinelibrary.wiley.com/doi/10.1111/mec.13605/full#mec13605-fig-0001

Island Fox History - 2

Site	Bottleneck year	Bottleneck size	Current size in 2014	*n* (total sampled)	*n* (after filters)
Grey foxes	NA	NA	Unknown	18	16
SMI	1999–2000	15	470	24	21
SRI	1999–2000	15	826	23	23
SCI	1999–2000	50–60	2466	24	24
SCA	1999	(>90% decline)	1624	46	43
SCL	NA	NA	1230	19	17
SNI	1970s	20?	263	46	44

Island fox bottleneck year and bottleneck population size estimates from Coonan *et al*. (2010). No estimate of the population size is available for Santa Catalina Island during its 1999 bottleneck, but this population is estimated to have declined by >90%. San Nicolas island foxes may have dropped to as low as 20 individuals in the 1970s. Current adult population size estimates for 2014 from Coonan (2015).

Human Evolutionary History - Li and Durbin 2011 10.1038/nature10231

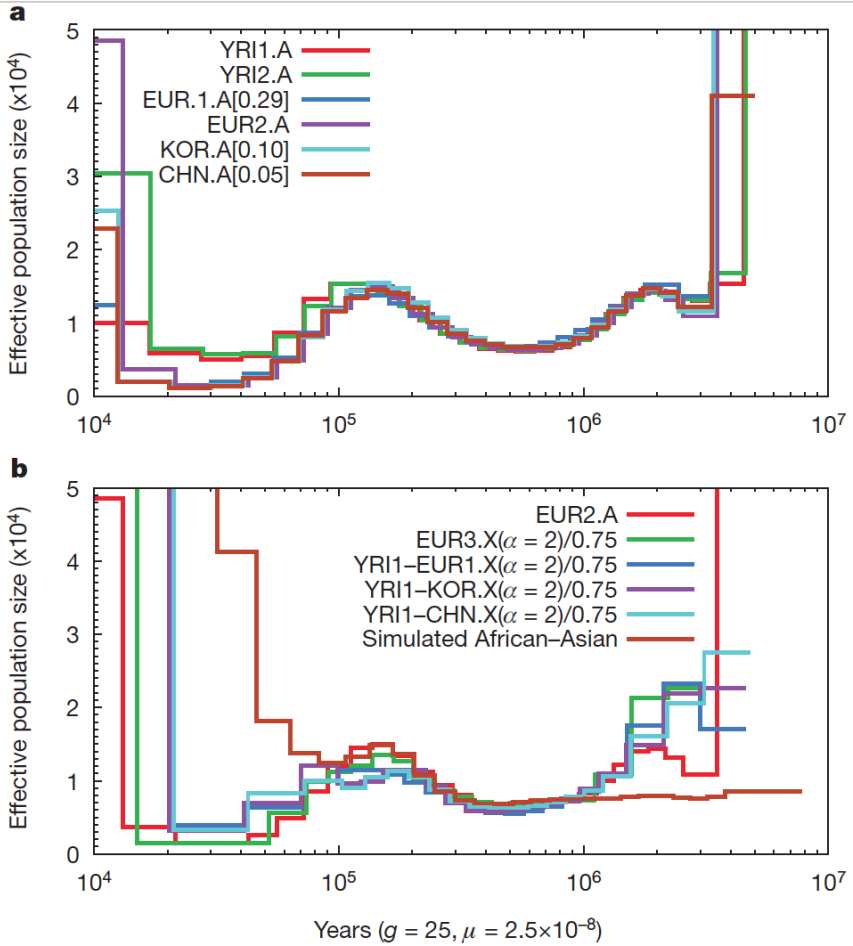


Figure 3 | PSMC estimate on real data. **a**, Population sizes inferred from autosomes of six individuals. 5%, 10% and 29% of heterozygotes are assumed to be missing in CHN.A, KOR.A and EUR1.A, respectively. **b**, Population sizes inferred from male-combined X chromosomes and the simulated African–Asian combined sequences from the best-fit model in ref. 21. Sizes inferred from X-chromosome data are scaled by 4/3. The neutral mutation rate on X, which is used in time-scaling, is estimated with the ratio of male-to-female mutation rate, α, equal to 2 (see Methods).

Genetic Drift

Drift

Genetic Drift

The chance changes in allele frequency that result from the random sampling of gametes from generation to generation in a finite population.

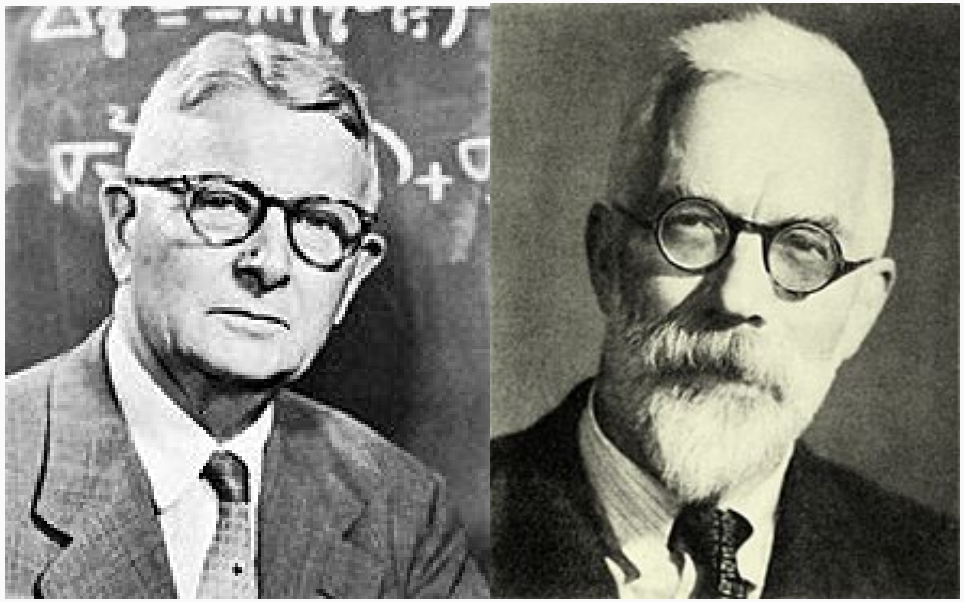
Bottleneck

Period during which only a few individuals survive to continue the existence of the population.

Founder Effect

Populations descended from a small founder group may have low genetic variation or by chance have a high or low frequency of particular alleles.

Wright-Fisher Model & Binomial Sampling



- Consider a large population at HWE, with alleles A and a at equal frequencies $p = \frac{1}{2} = q$.
- Here the genotype frequencies will be expected to be $\frac{1}{4}AA$, $\frac{1}{2}Aa$ and $\frac{1}{4}aa$.
- Let's say that this population undergoes a bottleneck, and only 4 individuals survive, randomly chosen.

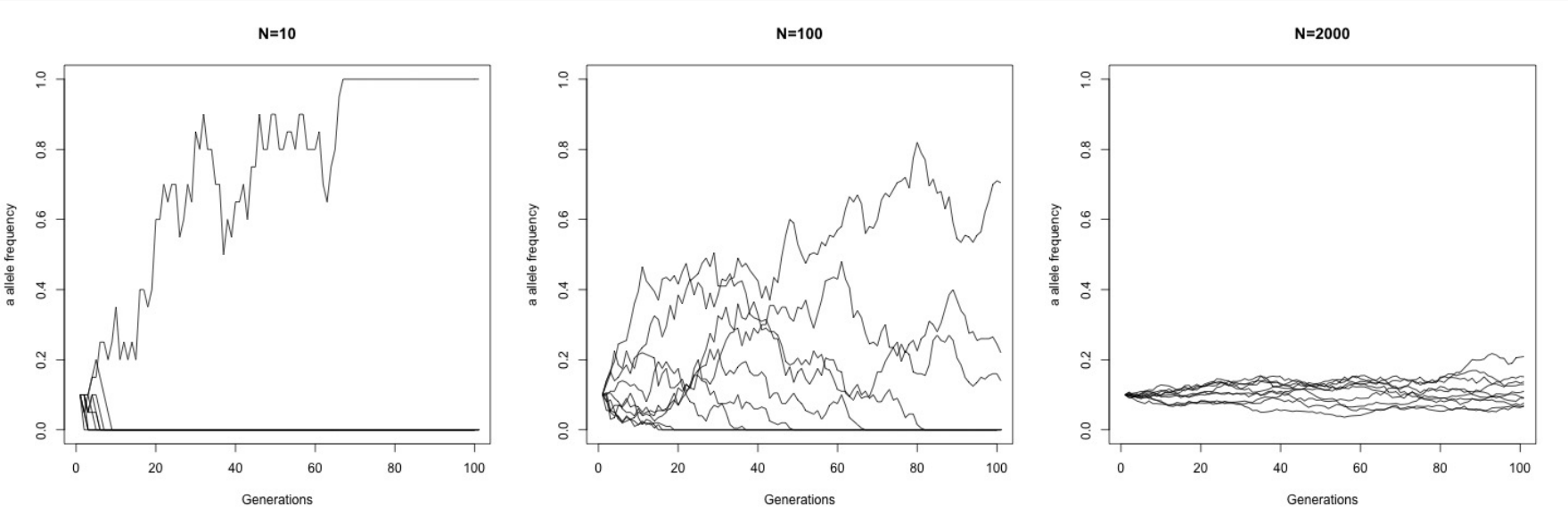
Wright-Fisher Model

- Probability that these 4 individuals will be $AA = (\frac{1}{4})^4 = \frac{1}{256}$.
- Similarly, other combinations are possible - chosen genotypes can be any combination of the 3 possible genotypes.
- This is equivalent to sampling 8 haploid gametes of type A and a (recall independent assortment).
- Ergo - binomial sampling!
- So generalizing this, in a population with N diploid individuals (i.e. $2N$ haploid gametes), probability of randomly drawing (and replacing) j gametes of type A, and $2N - j$ of type a from the parental generation to the offspring generation is:
- Pr(j alleles of type A)$= \binom{2N}{j} p^j q^{2N-j}$

Simulating genetic drift in R

```
p<-0.9 #Frequency of A allele
N<-20 #Population size
g<-100 #Number of generations
I<-10 #Number of iterations
jpeg("drift.jpg")
plot(c(1,g),c(0,1),type="n",
xlab="Generations",ylab="Allele frequency of a allele") #Empty plot
for(i in 1:I){
y1<-rep(0,round(2*N*p))#sample A alleles
y2<-rep(1,2*N-round(2*N*p))#sample a alleles
x<-c(y1,y2)#combine the two
ave<-mean(x)#allele frequency
for (j in 1:g) {
x<-sample(x,replace=T)
ave<-c(ave,mean(x))
}
points(ave,type="l")
}
dev.off()
```

Simulating drift in R



Redefining F

Drift in subpopulations

IBD v IBS

- Two alleles are Identical By Descent if they are replicas by DNA replication of a gene present in some previous generation.
- Two alleles are Identical By State if they are replicas of each other.

Allozygosity

Two alleles at a locus are allozygous if they are derived from different sources - i.e. heterozygotes are always allozygous. Homozygotes can be allozygous too, if they come from different sources (IBS).

Autozygous

Two alleles at a locus are autozygous if they are derived from the same source, i.e. IBD.

Another definition of F

F statistic

Let's redefine F as the probability that any two alleles in a diploid individual are IBD, also called the *fixation index*.

- Define at a distant point in time in the past, all alleles are "distinct" in the starting population, i.e. at this time t, $F_t = 0$.
- As times goes on, this population splits into subpopulations, where alleles drift, and become more similar to each other.
- i.e. F_t increases.

So the probability of IBD between two alleles in any randomly sampled individual can be defined as: $F_t = \frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}$

where $\frac{1}{2N}$ is the probability of drawing any two alleles, and them being from the same ancestral copy, and F_{t-1} is the probability that they were IBD in the previous generation.

Another definition of F - contd.

Multiplying both sides by -1 and adding 1 to each side leads to:

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

Rewriting this recursion:

$$1 - F_t = (1 - \frac{1}{2N})^t(1 - F_0)$$

i.e. when $F_0 = 0$, $F_t = 1 - (1 - \frac{1}{2N})^t$

Another definition of F contd.

Recall that we had previously defined $F = \frac{H_{exp} - H_{obs}}{H_{exp}}$

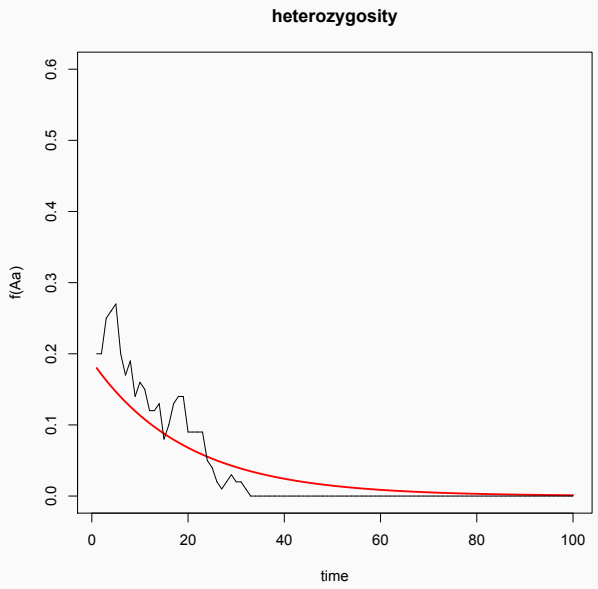
If we redefine it now, as the proportion of loss of heterozygosity from population 0 to population t due to subpopulation structure,

$$F = \frac{H_0 - H_t}{H_0}, \text{ i.e. } H_t = (1 - F_t)H_0.$$

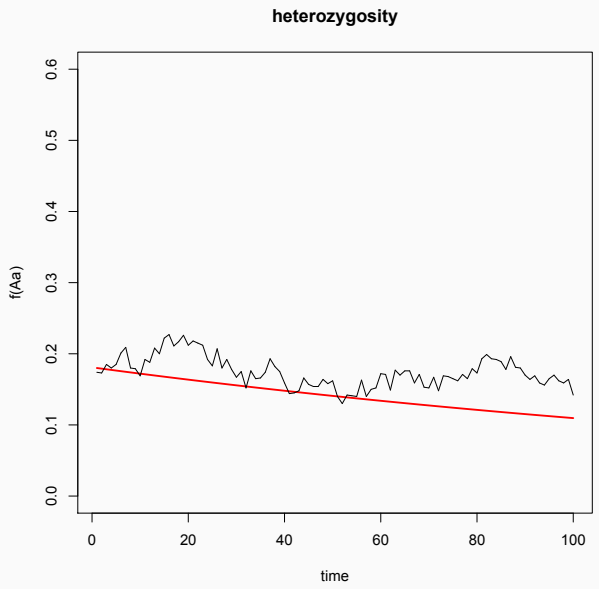
Substituting, we have

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

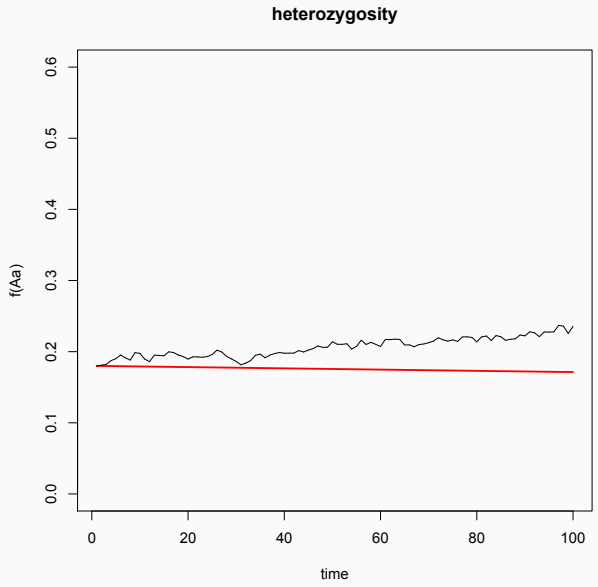
Decline in heterozygosity



N=10



N=100



N=1000

Neutral Theory

Neutrality and Drift

- The concept of drift is "coupled" with the idea of neutral evolution.
- A neutral allele is one which has no effect on fitness over other alleles at that locus.
- i.e. neutral alleles drift.
- Selected alleles can be affected by drift only if they are under weak selection (unless they are very rare).
- Recall - only about 2% of all of our genome encodes for proteins (exome).
- Changes outside exons may be entirely neutral if they don't affect any regulatory sites.
- Examples of neutral sites: synonymous mutations, non-synonymous change that replaces an amino acid with one that's functionally similar, a non-synonymous change that produces a large change in phenotype on which selection no longer acts.

Neutral Variation vs Selection

- $\approx$ 36 million substitutions have occurred between humans and chimps since they last shared a common ancestor.
- How many of these fix due to drift?
- How many due to selection?
- Why is there so much polymorphism?
- If selection and drift quickly fix alleles, why is there so much variation at all?
- Three explanations:
  - Balancing selection
  - Mutation-selection balance
  - Mutation-drift balance (Neutral theory)

Neutral Theory of Molecular Evolution


Motoo Kimura

- Most new mutations are deleterious and lost immediately.
- Most of the observed polymorphisms are neutral.
- Mutation-drift Equilibrium
- Variation lost by drift = variation introduced by mutation.

Effective Population Size

Effective Population Size N_e

N_e

Define the effective population size as the number of individuals in a theoretical population having the same amount of genetic drift, i.e the ideal population of size N_e in which all parents have an equal expectation of being parents of any progeny individual, i.e. the size of the randomly mating population.

Population size change

Recall: $1 - F_t = (1 - \frac{1}{2N})^t(1 - F_0)$ where N is the effective population size. If we break this up, $1 - F_1 = (1 - \frac{1}{2N_0})(1 - F_0)$
$1 - F_2 = (1 - \frac{1}{2N_1})(1 - F_1)$. Combining the two,
$1 - F_2 = (1 - \frac{1}{2N_1})(1 - \frac{1}{2N_0})(1 - F_0)$.

Also, using the general recursion, $1 - F_2 = (1 - \frac{1}{2N})^2(1 - F_0)$

Setting these equal to each other, $(1 - \frac{1}{2N})^2 = (1 - \frac{1}{2N_1})(1 - \frac{1}{2N_0})$

Or approximately,

$\frac{1}{N} = \frac{1}{2}(\frac{1}{N_0} + \frac{1}{N_1})$

More generally,

$\frac{1}{N_e} = \frac{1}{t}(\frac{1}{N_0} + \frac{1}{N_1} + \ldots + \frac{1}{N_{t-1}})$

Founder Effects

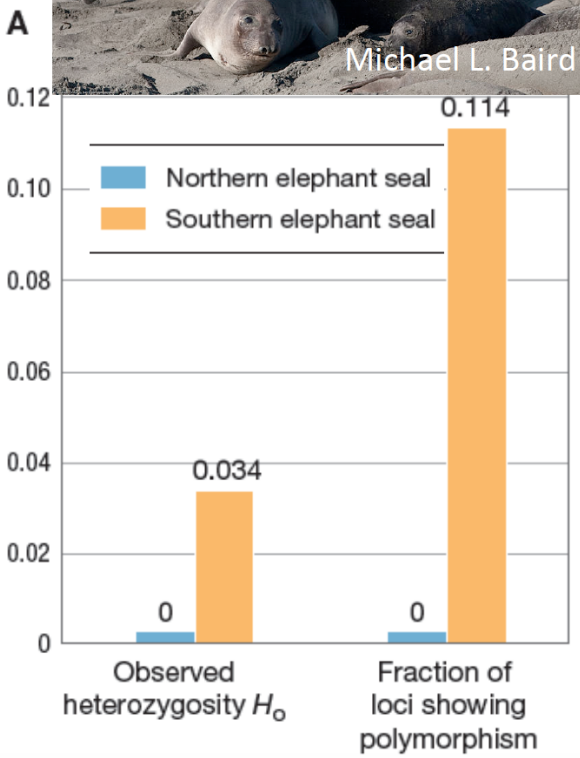
Problem 3.7

If a population went through a bottleneck such that $N_0 = 1000, N_1 = 10, N_2 = 1000$, calculate the effective size of this population across three generations.

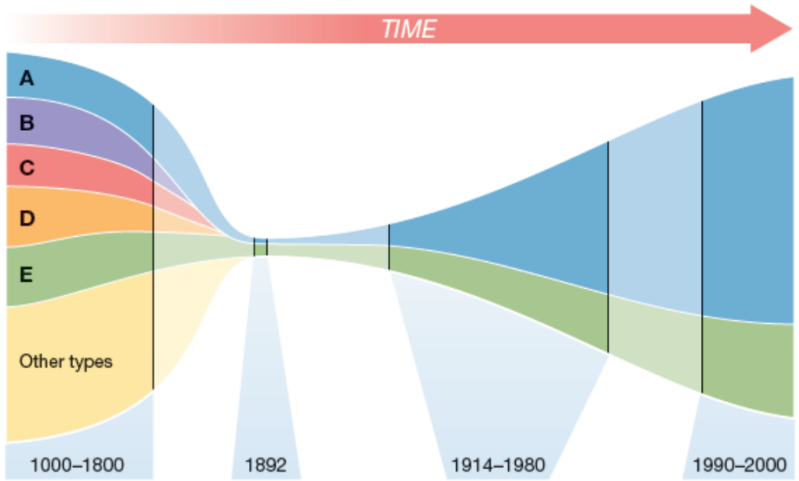
Bottleneck Effects - courtesy Graham Coop, UC Davis

Bottleneck in Northern Elephant Seals

Figures from Bergstrom and Dugatkin. Evolution. 2nd Ed



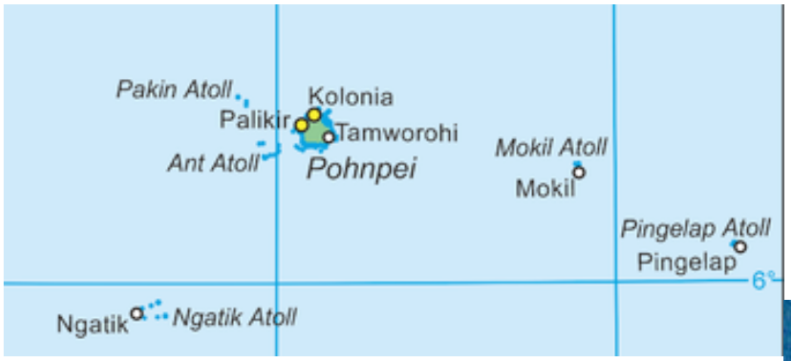
mtDNA diversity over time



	Number of Individuals			
A	1	0	4	107
B	2	0	0	0
C	1	0	0	0
D	1	0	0	0

Bottleneck Effects - courtesy Graham Coop, UC Davis

During bottlenecks rare alleles can by chance reach high frequency



5% of islanders on Pingelap have
A recessive form of achromatopsia
(total colour blindness)
The allele is at ~20% frequency

Higher incidence of polydactyl in Amish population

Using Genetic Tests, Ashkenazi Jews Vanquish a Disease

By GINA KOLATA
Published: February 18, 2003

A number of years ago, five families in Brooklyn who had had babies with a devastating disease decided to try what was then nearly unthinkable: to eliminate a terrible genetic disease from the planet.

The disease is Tay-Sachs, a progressive, relentless neurological disorder that afflicts mostly babies, leaving them mentally impaired, blind, deaf and unable to swallow. There is no treatment, and most children with the disease die by 5.

Tay-Sachs disease (a recessive disorder) is found at higher frequency in Ashkenazi Jews

Coalescence

Gene trees and Coalescence

- In an ideal world, we would be able to trace the ancestry of every allele backwards in time to the common ancestor.
- Wright-Fisher process gives us some means to do that - in an ideal neutral population.
- But what if we had a pedigree, or a genealogy for every allele?

Coalescence

Coalescence refers to the process in which, looking backward in time, the genealogies of two alleles merge at a common ancestor.

Coalescence

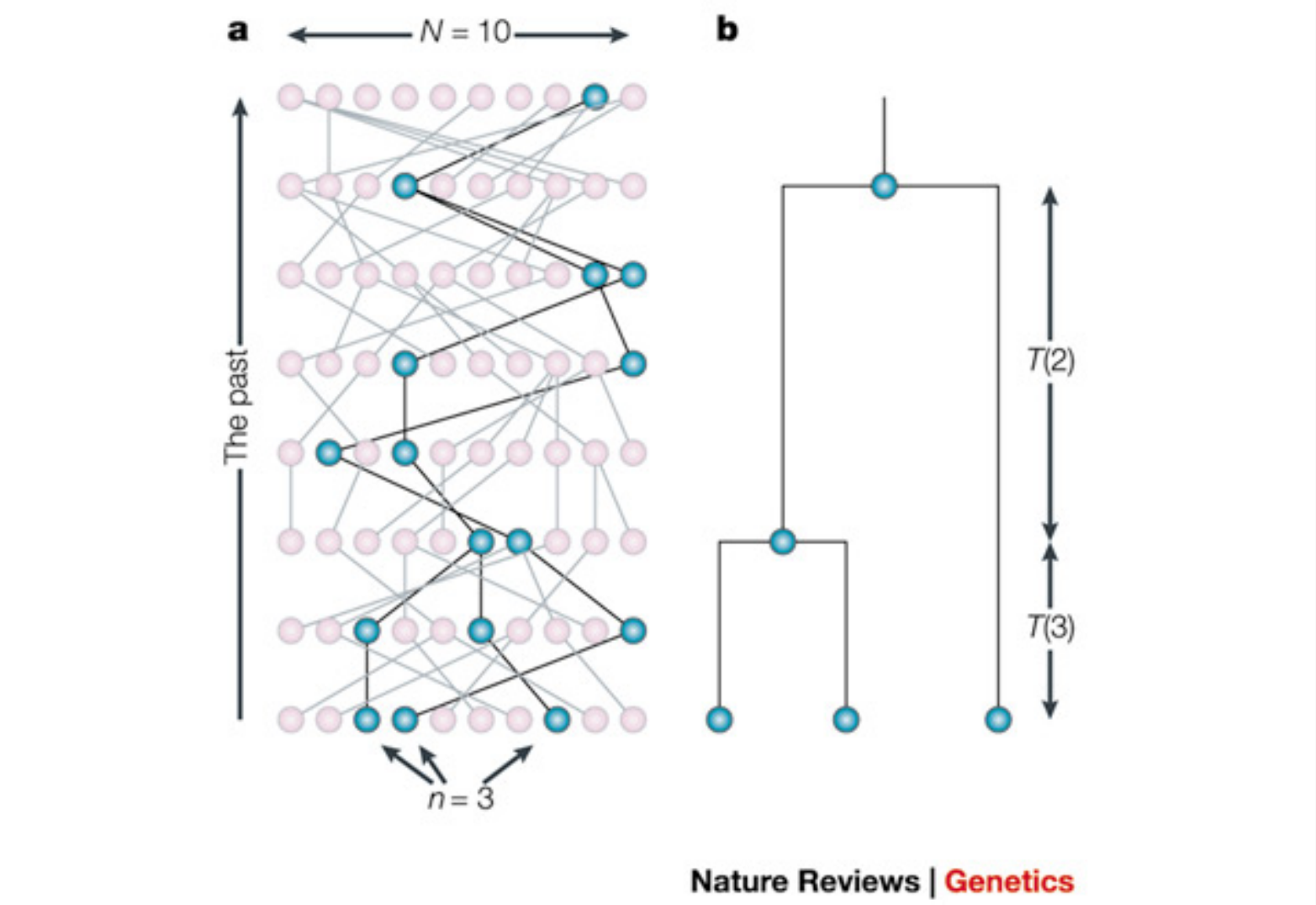


Figure from Rosenberg and Nordborg 2002 10.1038/nrg795

Coalescence Time

How long does it take for any two randomly chosen alleles in a population in the present to coalesce?

- Recall IBD probabilities - the probability that any two randomly chosen alleles are from the same ancestor in the previous generation $= \frac{1}{2N}$.
- Hence probability that they came from two distinct alleles $= 1 - \frac{1}{2N}$.
- Going backwards in time, the probability that two alleles don't coalesce for t generations, and then coalesce in generation $t+1$ can then be written as: $(1 - \frac{1}{2N})^t \frac{1}{2N} \approx \frac{1}{2N} e^{\frac{-t}{2N}}$.
- If there are k alleles in a sample (present), probability that the k alleles do not coalesce for t generations, and then one pair coalesces giving $k-1$ alleles at $t+1$ generations ago is: $Pr(k)^t(1 - Pr(k)) \approx \frac{\binom{k}{2}}{2N} \exp(-\frac{\binom{k}{2}}{2N} t)$

Coalescence Time

- Mean $= \frac{4N}{k(k-1)}$
- Variance $= \frac{16N^2}{(k(k-1))^2}$

Some key points

- Coalescent genealogies will be different within a population for each gene/set of alleles.
- Coalescence provides a means to simulate pedigrees and ancestral origins of an allele and hence a population.
- Variations of the coalescent allow simulating mutations, recombination, population structure, migration, selection.
- Provides a computational/statistical framework for inference of evolutionary history.

Conclusions

Summary

- Drift is the random sampling of alleles from one generation to the next.
- Wright-Fisher model extends binomial sampling to multiple generations.
- If only drift is acting in a population, the probability that an allele will drift to fixation is the initial frequency of the allele in the population.
- Heterozygosity decreases at an average rate of $\frac{1}{2N}$ in each generation due to drift.
- Real populations are not perfect Wright-Fisher populations - hence we use a theoretical effective population size N_e that is the size of a random mating population that drifts.
- Coalescence describes the Wright-Fisher population history of each allele backwards in time.

Questions?