# CS 68: BIOINFORMATICS

Prof. Sara Mathieson

Swarthmore College

Spring 2018

# Outline: Mar 19

- Population Genetics
- Genetic drift
- Introduction to the Coalescent

Notes:
- Office hours TODAY 3-5pm
- Lab 5 due Wednesday

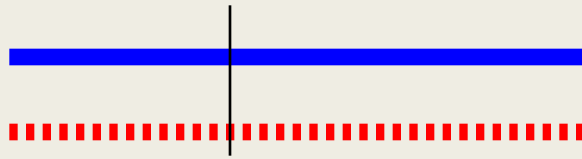# What is population genetics?

# Moving on from phylogenetics...

- **UPGMA algorithm** for constructing ultrametric phylogenetic trees

- **Neighbor Joining algorithm** for constructing phylogenetic trees

- **Fitch's algorithm** for reconstructing ancestral states (unweighted parsimony)

- **Sankoff's algorithm** for reconstructing ancestral states (weighted parsimony)

- **Gusfield's algorithm** for determining if there is a perfect phylogeny

Note: runtime for perfect phylogeny will be part of Lab 6

Although we can apply these algorithms to individuals from the same population, in practice we usually apply them to individuals from distantly related species.

# Recombination: one reason we can't just use phylogenetic models for the same species

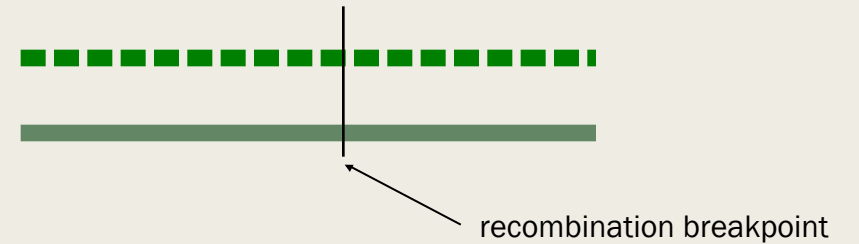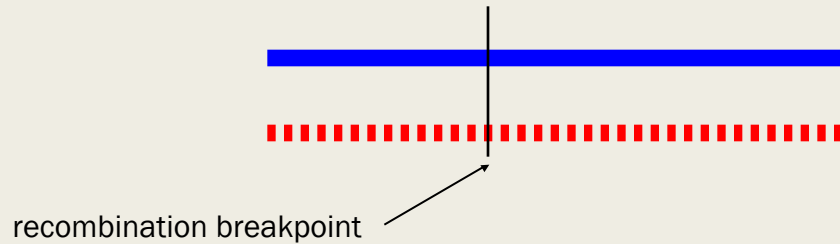# Recombination: one reason we can't just use phylogenetic models for the same species
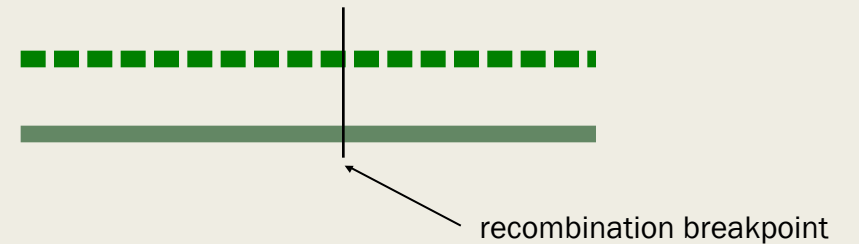


recombination breakpoint

recombination breakpoint

# Recombination: one reason we can't just use phylogenetic models for the same species



recombination breakpoint

recombination breakpoint

# Recombination: one reason we can't just use phylogenetic models for the same species
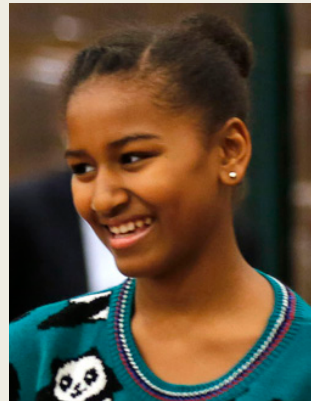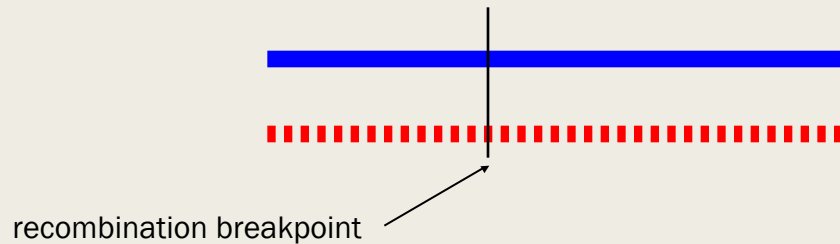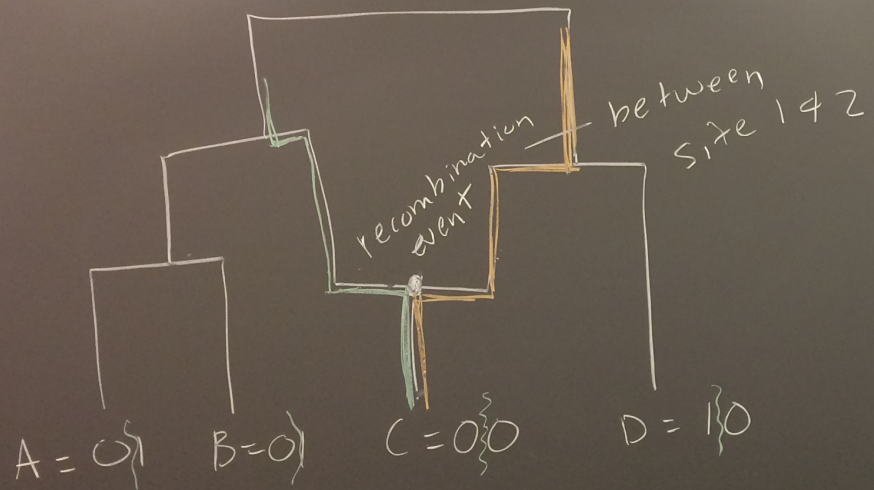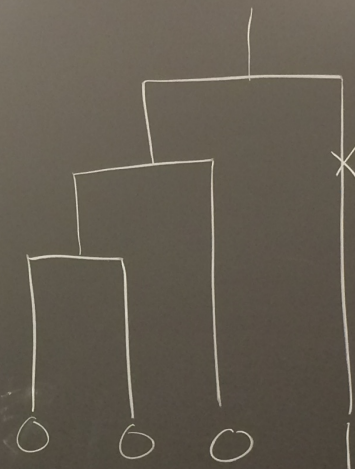
recombination breakpoint

recombination breakpoint

Claymation recombination video
https://www.youtube.com/watch?v=3qgBKrAZCLg

A = 0,1   B = 0,1   C = 0,0   D = 1,0

**ARG**: ancestral recombination graph

recombination event — between site 1 & 2

site 1

site 2

not one single tree!

# Population Genetics

- For individuals from the same population, we need to model evolution differently

- We will "zoom in" on one population and make the assumption of random mating within the population

- What are our goals in population genetics?

# Goals of population genetics

- Understand the evolutionary history of a population or related populations

# Goals of population genetics

- Understand the evolutionary history of a population or related populations

- What population size changes have occurred in the past? Bottlenecks? Exponential growth?

# Goals of population genetics

- Understand the evolutionary history of a population or related populations

- What population size changes have occurred in the past? Bottlenecks? Exponential growth?

- What regions of the genome are under natural selection? How strong is the selection? When did it start/end?

# Goals of population genetics

- Understand the evolutionary history of a population or related populations

- What population size changes have occurred in the past? Bottlenecks? Exponential growth?

- What regions of the genome are under natural selection? How strong is the selection? When did it start/end?

- Have populations split due to migration? When did the split happen? Has there been admixture since the split?
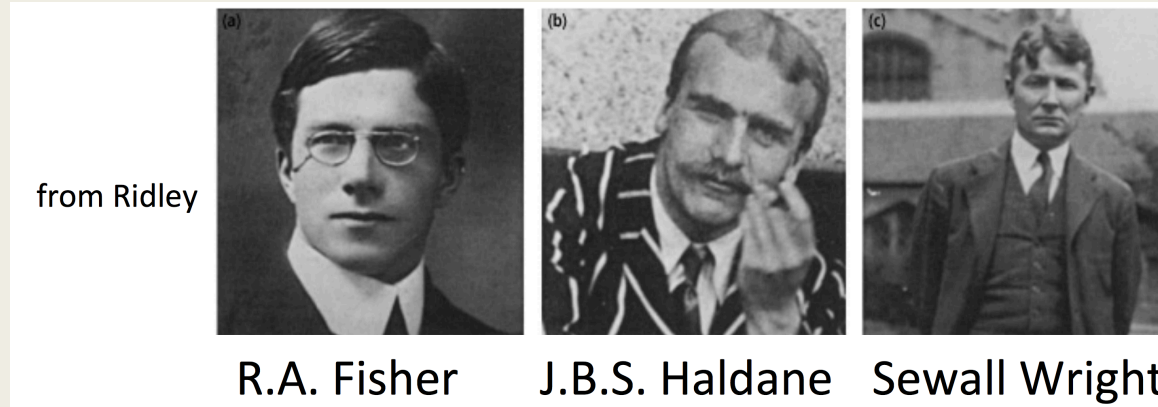
# Goals of population genetics

- Understand the evolutionary history of a population or related populations

- What population size changes have occurred in the past? Bottlenecks? Exponential growth?

- What regions of the genome are under natural selection? How strong is the selection? When did it start/end?

- Have populations split due to migration? When did the split happen? Has there been admixture since the split?

- What is the mutation rate? What is the recombination rate? Have these changed over time? Do they differ in different parts of the genome?

# Goals of population genetics

- Understand the evolutionary history of a population or related populations

- What population size changes have occurred in the past? Bottlenecks? Exponential growth?

- What regions of the genome are under natural selection? How strong is the selection? When did it start/end?

- Have populations split due to migration? When did the split happen? Has there been admixture since the split?

- What is the mutation rate?  What is the recombination rate? Have these changed over time? Do they differ in different parts of the genome?

- Which regions of the genome (e.g. genes) are associated with disease? Do these differ across populations?

# History of Population genetics

- 1930's: "modern synthesis", development of population genetics theory



from Ridley

R.A. Fisher          J.B.S. Haldane          Sewall Wright

From: Graham Coop, https://gcbias.org/2016/09/21/population-genetics-undergrad-class/

- 1950's: structure of DNA discovered, many theoretical models validated

- 2000's: deluge of data, we can finally apply theory

- Recently: too much data, theory is too slow, COMPROMISE

*"Nothing in biology makes sense except in light of evolution."*
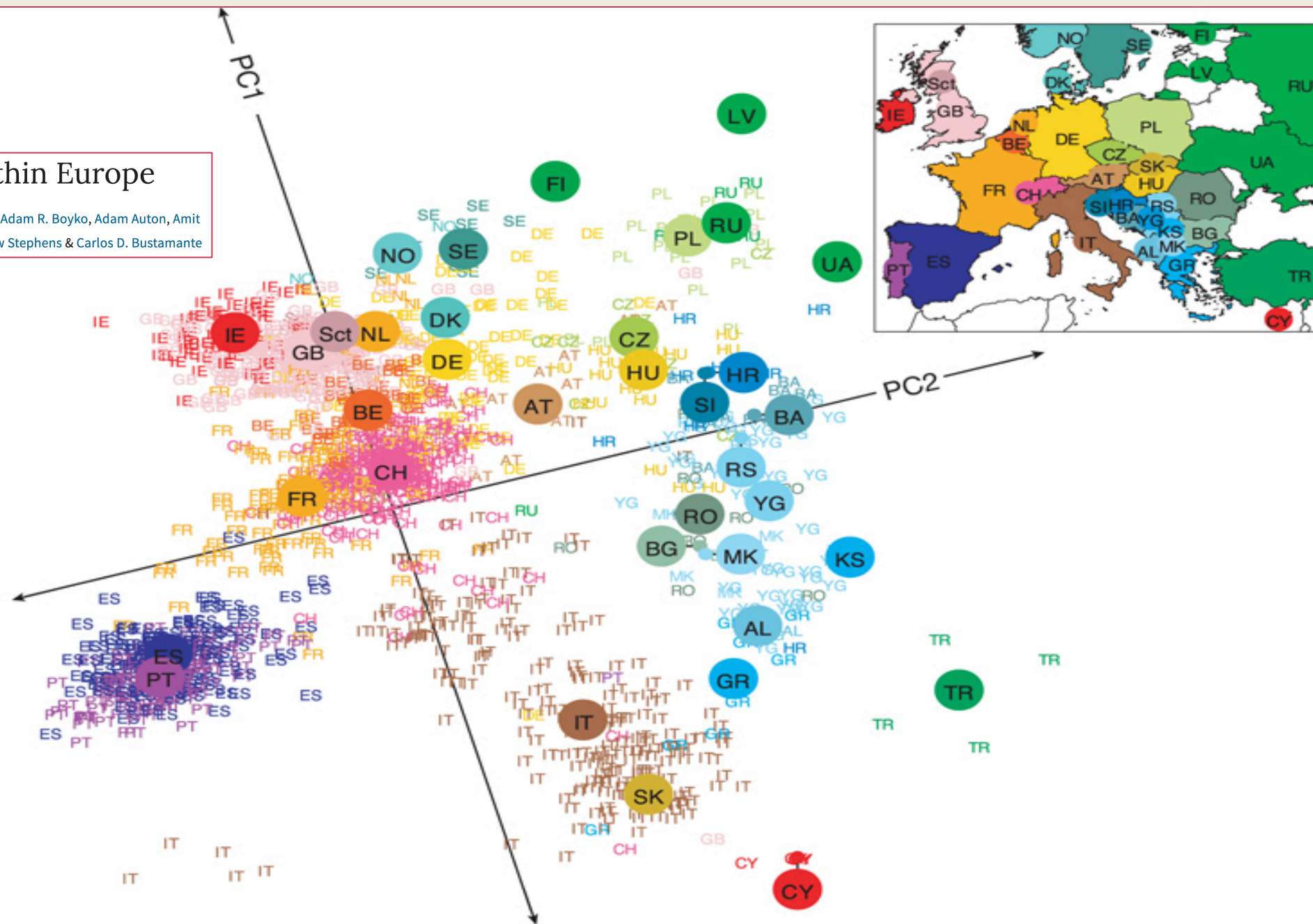T. Dobzhansky (1973)

*"Nothing in evolution makes sense except in light of population genetics."*
M. Lynch (2005)

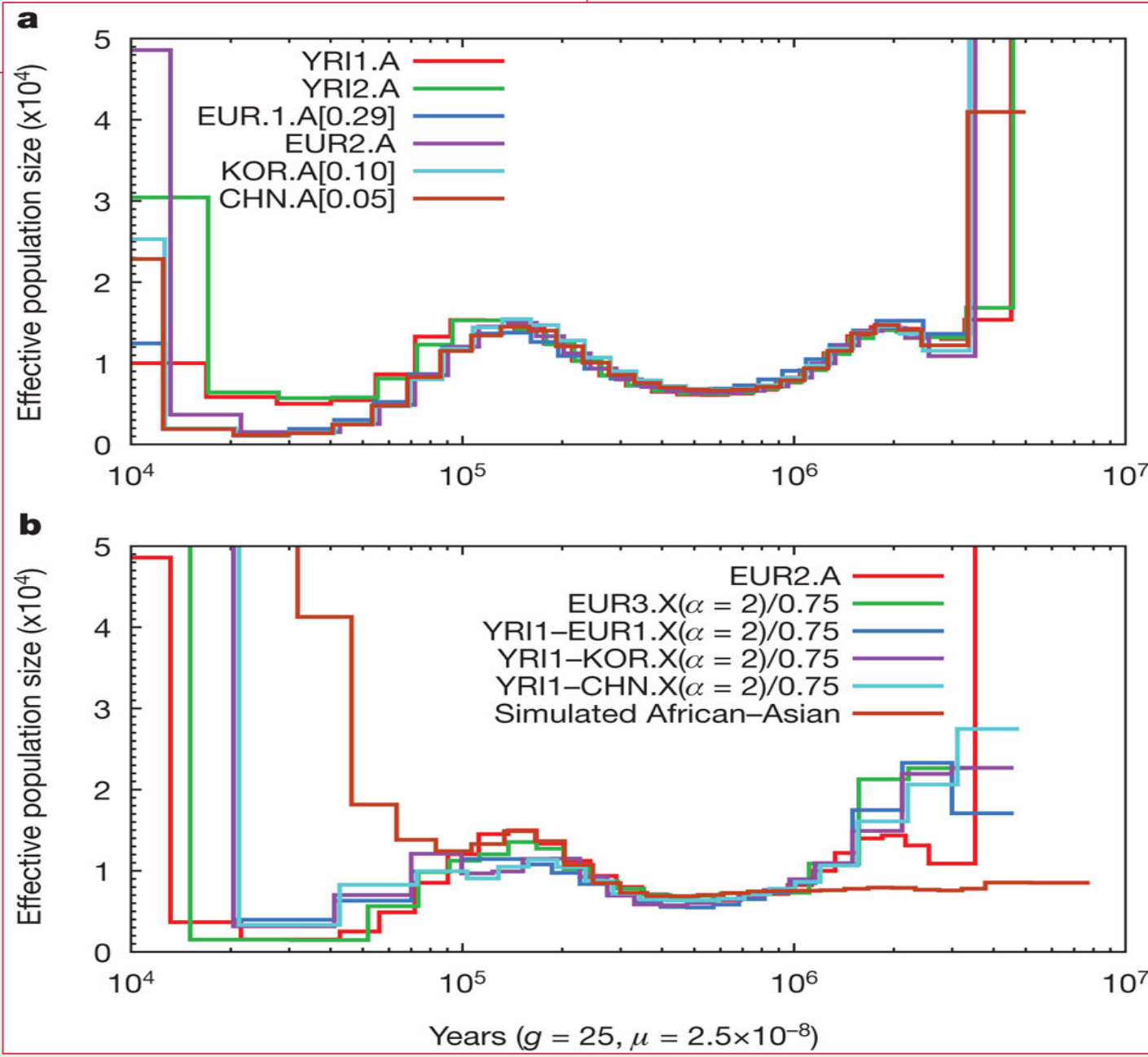# Examples of modern population genetics

# Genes mirror geography within Europe

John Novembre ✉, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens & Carlos D. Bustamante
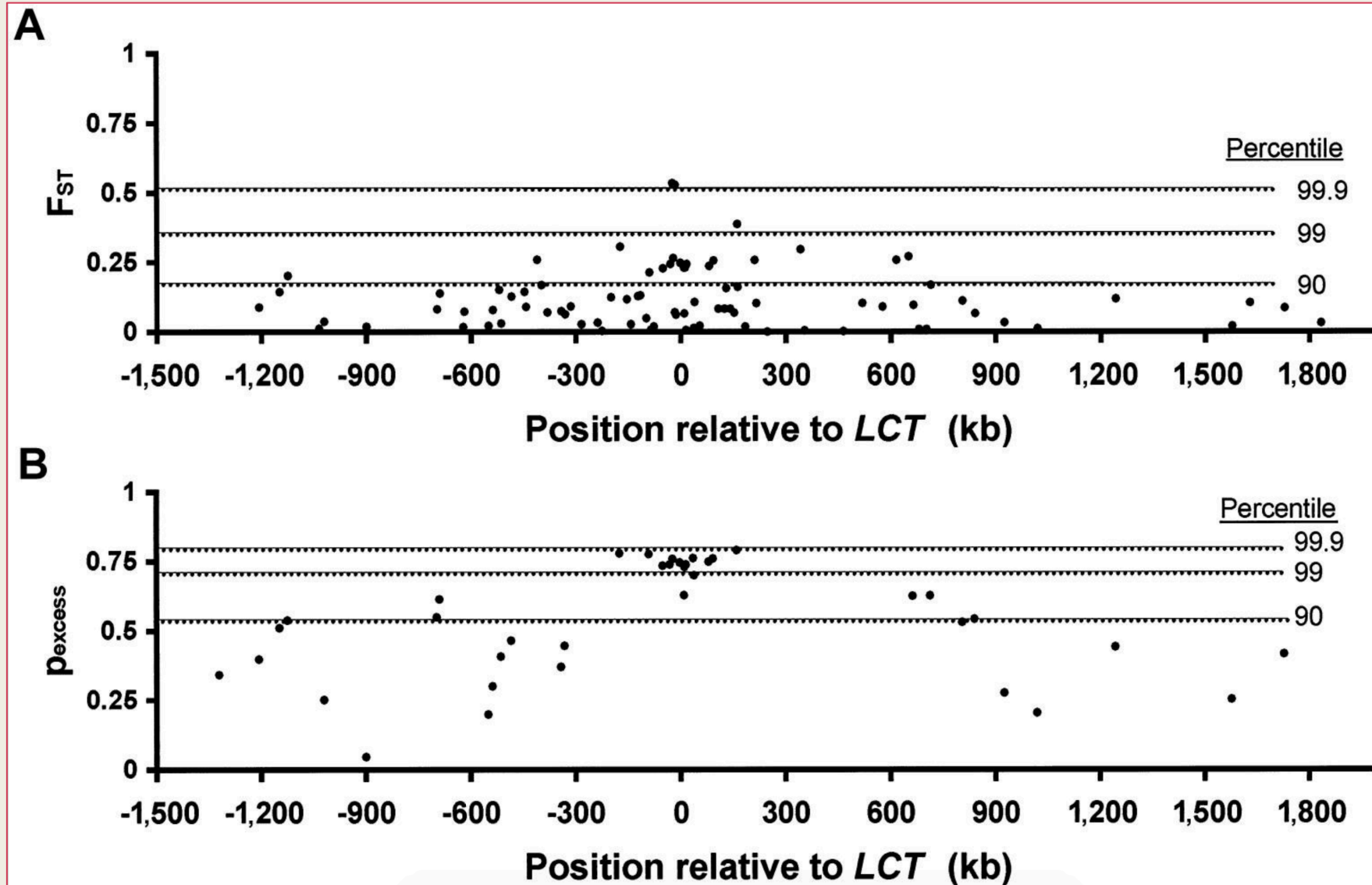


2008

# Inference of Human Population History From Whole Genome Sequence of A Single Individual
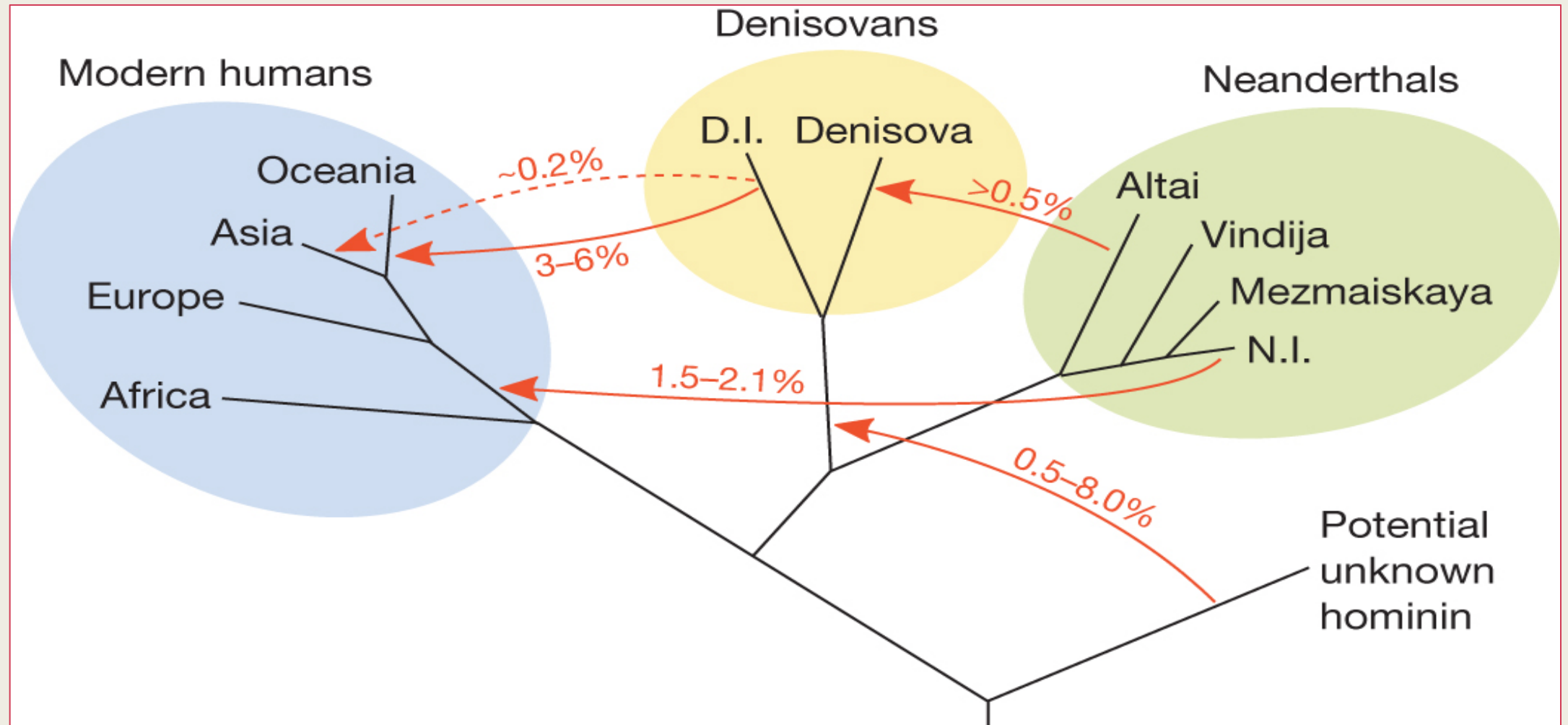
Heng Li[1,2] and Richard Durbin[1]



2009

# Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene

Todd Bersaglieri,[1] Pardis C. Sabeti,[3] Nick Patterson,[3] Trisha Vanderploeg,[1] Steve F. Schaffner,[3] Jared A. Drake,[1] Matthew Rhodes,[1,*] David E. Reich,[2,3] and Joel N. Hirschhorn[1,2,3]

2004

The complete genome sequence of a
Neanderthal from the Altai Mountains
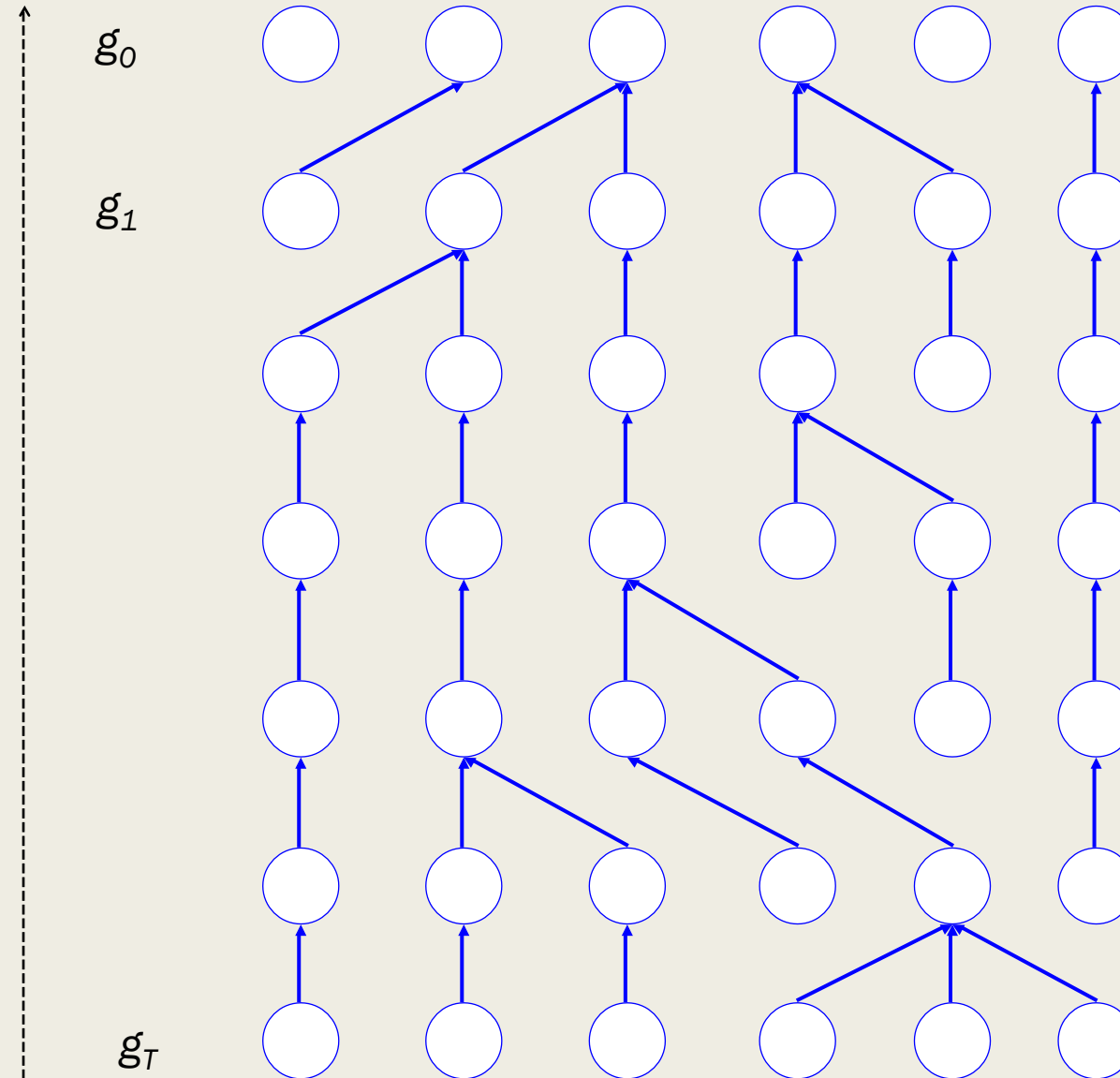
Kay Prüfer, Fernando Racimo […] Svante Pääbo ✉



2013

# Wright-Fisher Model

# Wright-Fisher Model



Generations back in time

$g_0$

$g_1$

$g_T$

- Imagine each child choosing their parent at random

- When two descendants choose the same parent, they "coalesce"

- From then on, they have the same ancestry and follow the same lineage

Constant population size: 2N

# Wright-Fisher Model

Generations back in time

$g_0$

$g_1$

$g_T$

- First: model variation but not any new mutations
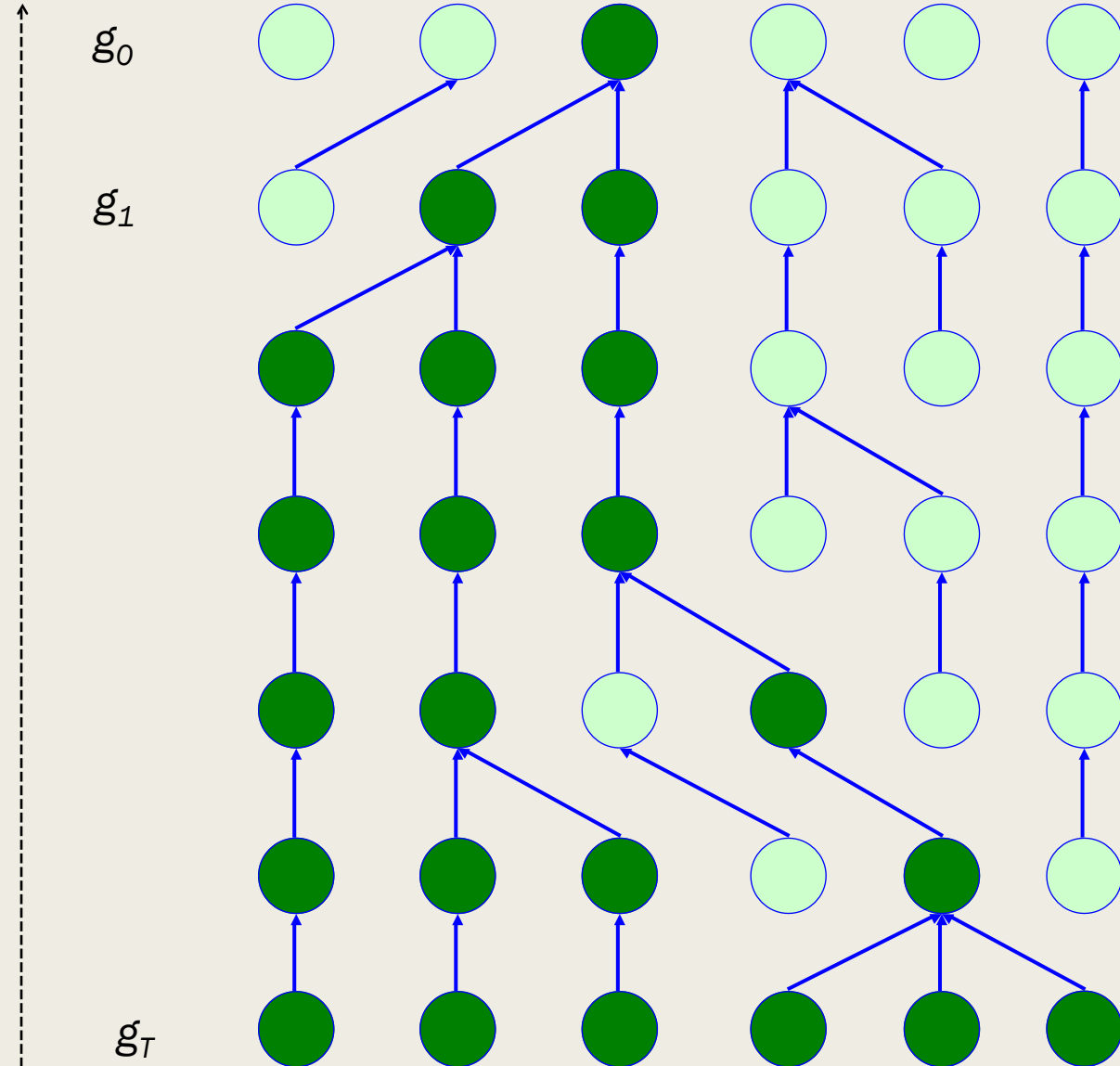
- Blue is the "A" allele

- White is the "a" allele

Constant population size: 2N

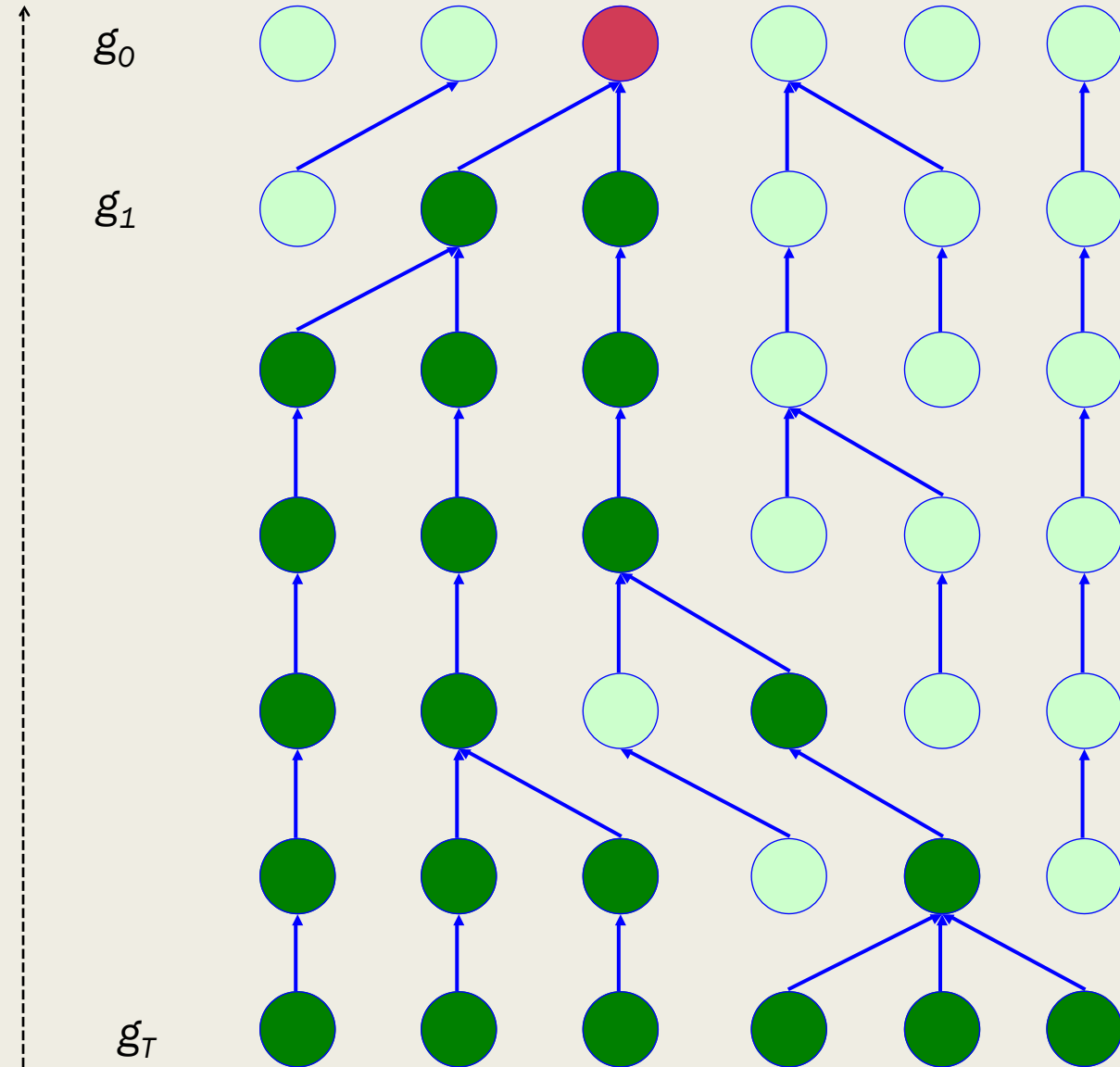# Wright-Fisher Model

Generations
back in time

- Viewed another way, track which individuals pass on genetic material that is observable at the present

- Dark green: contributes to genetic material at present

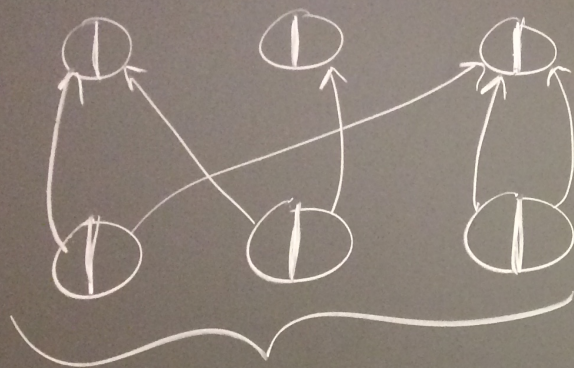- Light green: does not contribute genetic material to the present

$g_0$

$g_1$

$g_T$

Constant population size: 2N

# Wright-Fisher Model
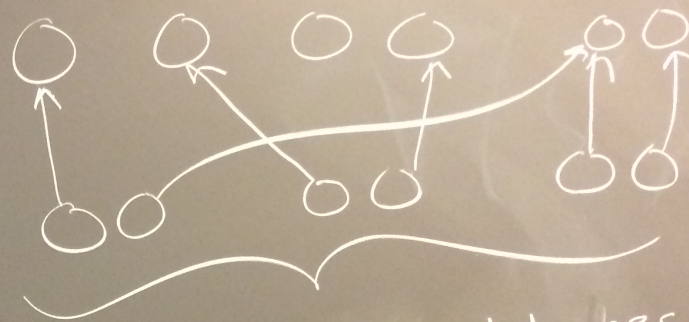


Generations back in time

$g_0$

$g_1$

$g_T$

- Eventually, all the present-day individuals will "coalesce" and share one common ancestor
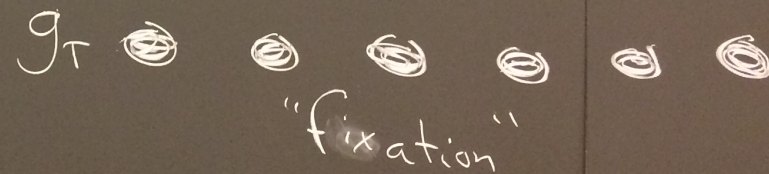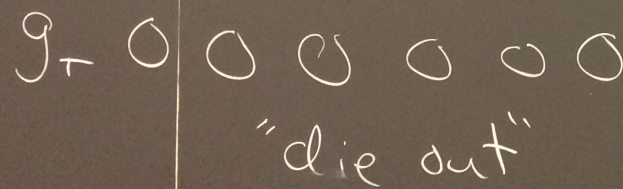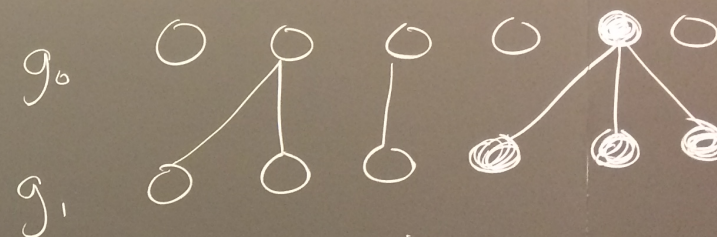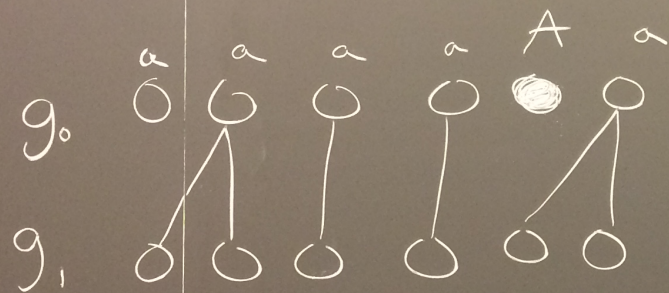
- Common ancestor in red

Constant population size: 2N

N = # of individuals

2N haplotypes
(samples)
(chromosomes)

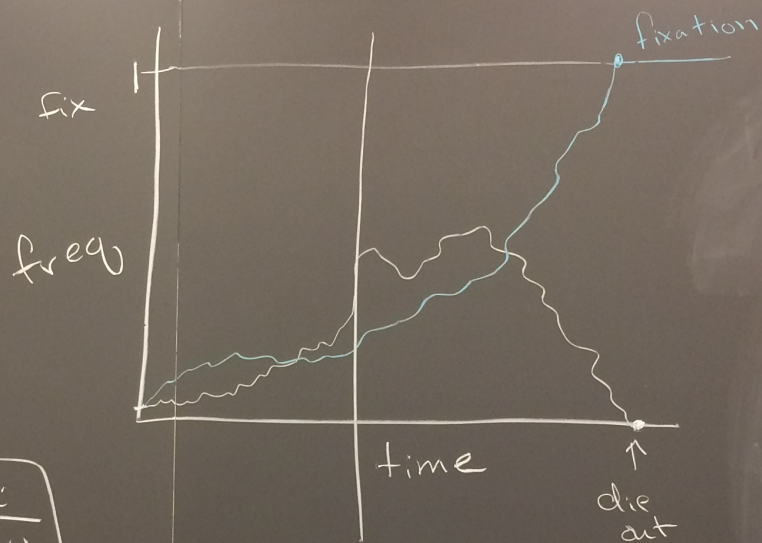# Question: given a new mutation, what is the probability it will "fix" in the population?

$g_0$ a a a a A a

$g_1$

$g_T$   "die out"

$g_0$

$g_1$

$g_T$   "fixation"

Observations

① mutation will either fix
   or die out

② eventually all
   lineages will coalesce

$f_i =$ freq of mutation $= \boxed{\dfrac{x_i}{2N}}$
   at generation $i$

$x_i =$ count of mutation
   at generation $i$

freq

fixation

time

← die
   out

$f_0 = \dfrac{1}{2}$

$P(\text{fixation}) = f_0$

AA
Aa
aA
aa

$= f_0$

$P(\text{fixation})$

$= \dfrac{1}{2N}$

new mutation

$1 - \dfrac{1}{2N}$