EEB Population Genetic Simulation and Inference

Lecture 1:

Models of Evolution, intro, etc

Professor Mike Hickerson

Department of Biology - City College

CUNY GC subprogram in EEB

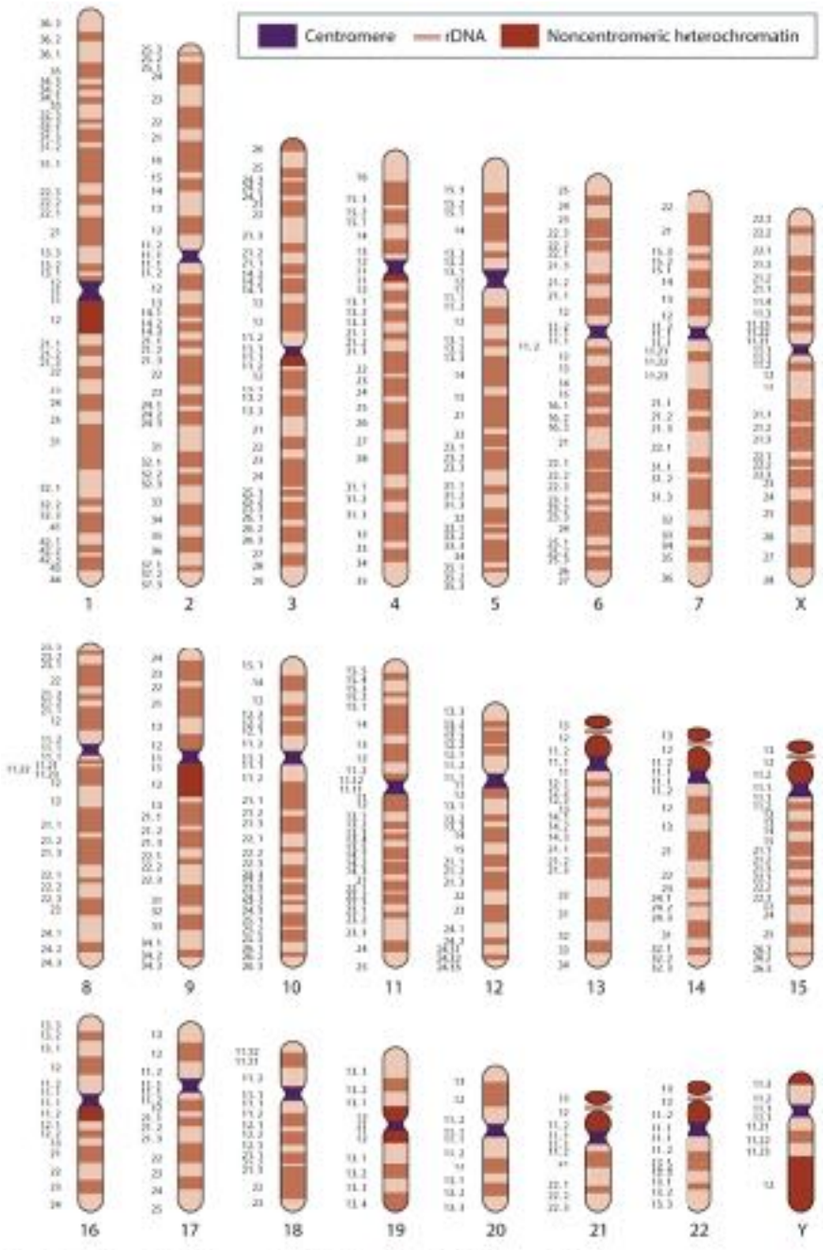


Figure 2.12 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

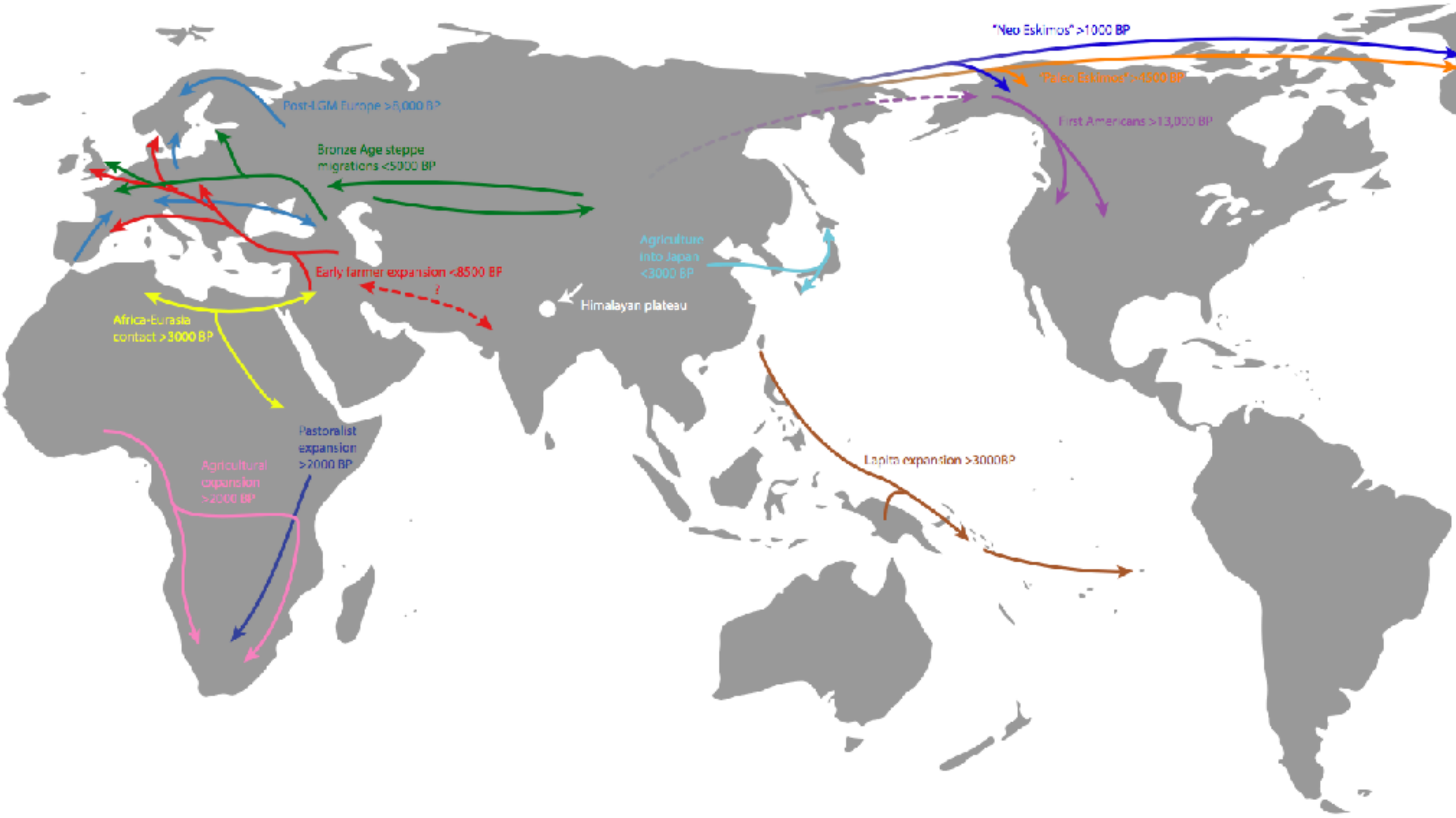
The history in your DNA

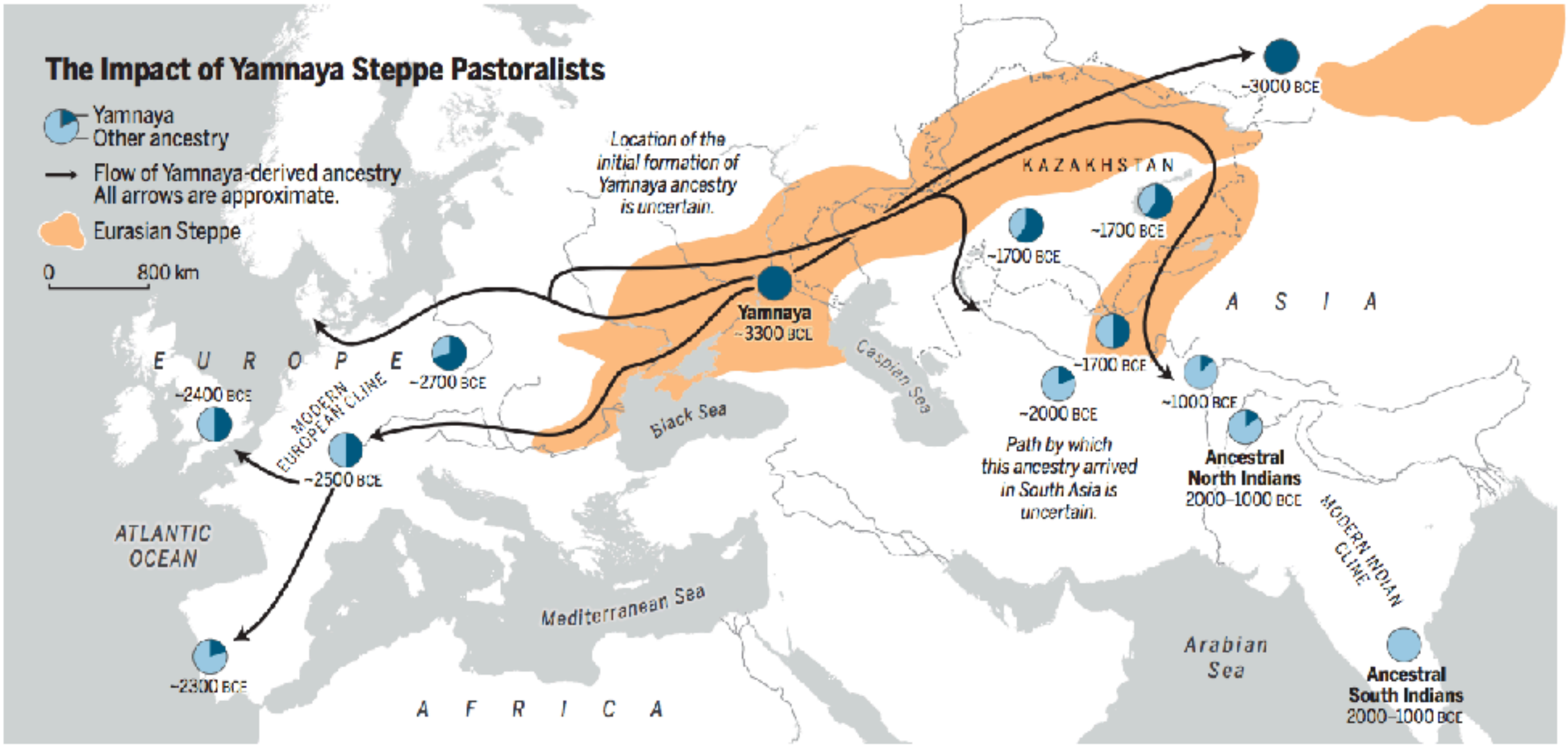
immediate family history

recent human history (hundreds to thousands of years)

ancient human pre-history (10s to 100s of thousands of years)

recent human history (hundreds to thousands of years)





The Bronze Age spread of Yamnaya Steppe pastoralist ancestry into two subcontinents—Europe and South Asia. Pie charts reflect the proportion of Yamnaya ancestry, and dates reflect the earliest available ancient DNA with Yamnaya ancestry in each region. Ancient DNA has not yet been found for the ANI and ASI, so for these the range is inferred statistically.

ancient human pre-history (10s of thousands of years)

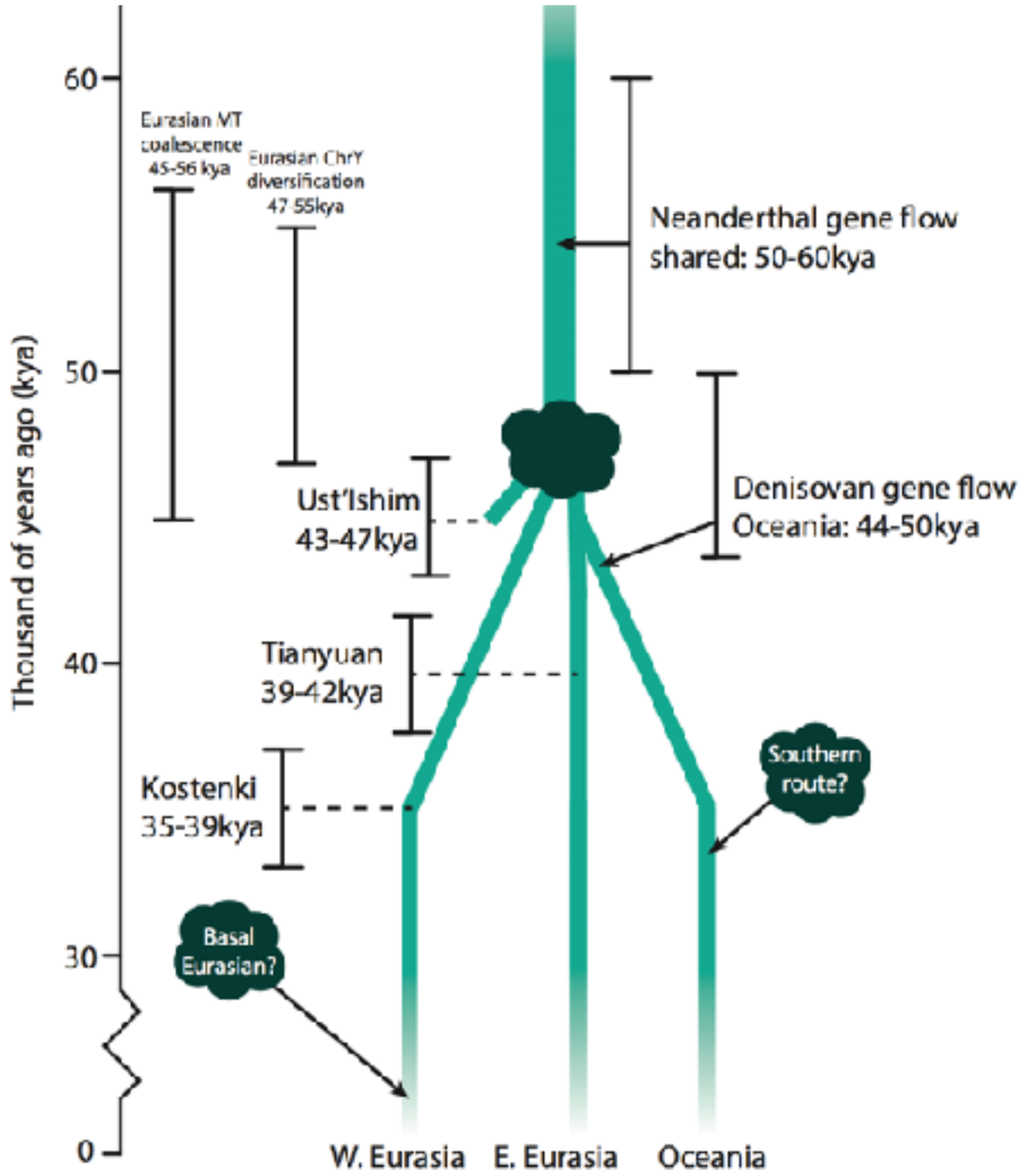
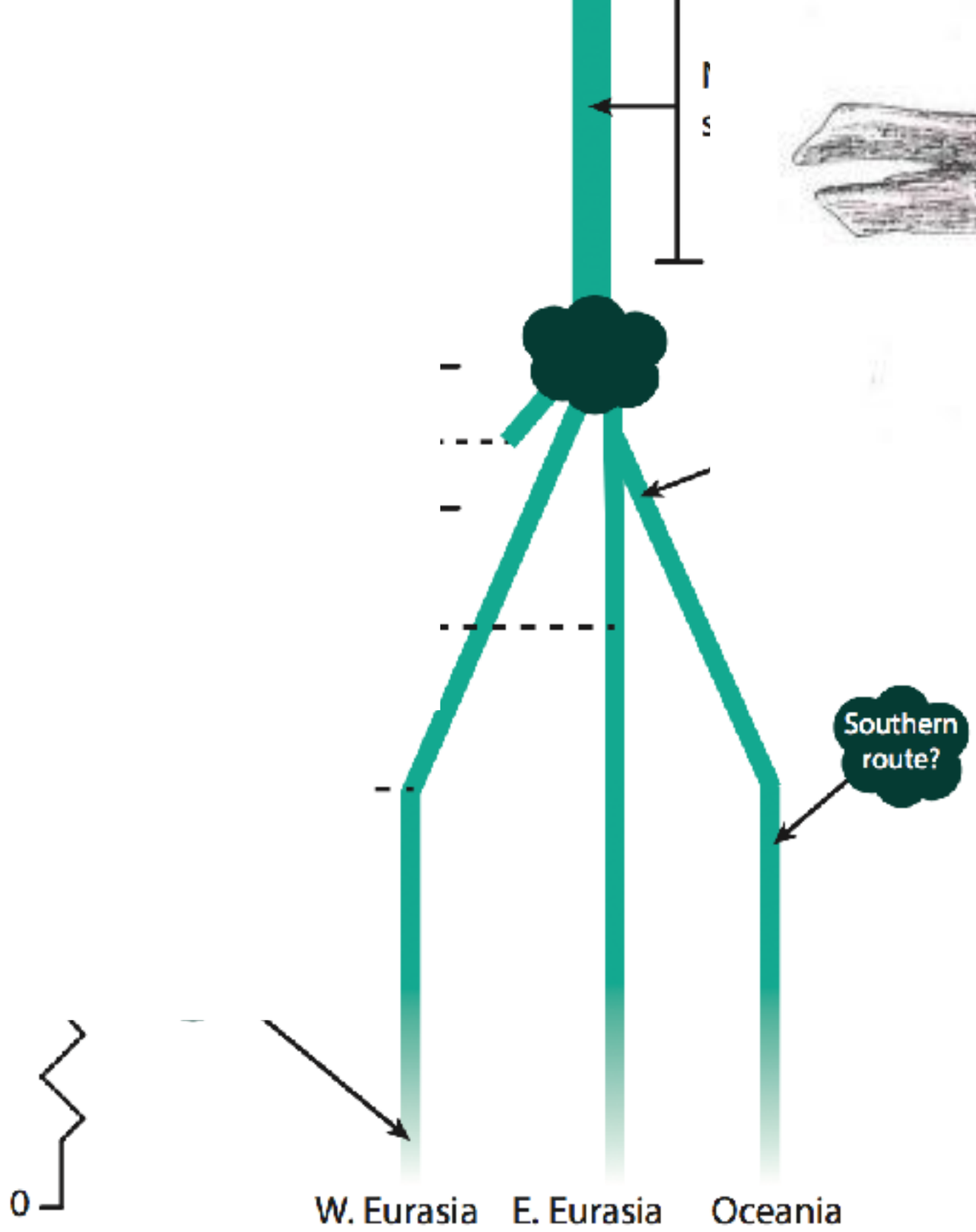
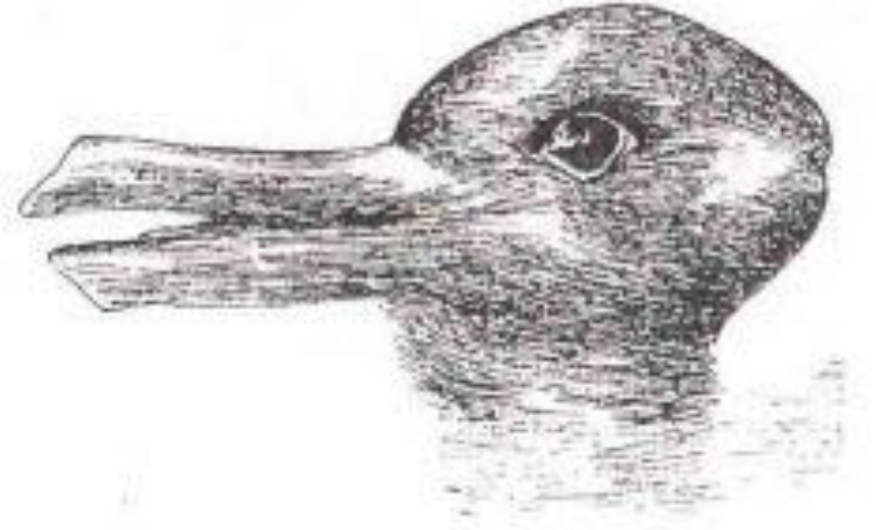


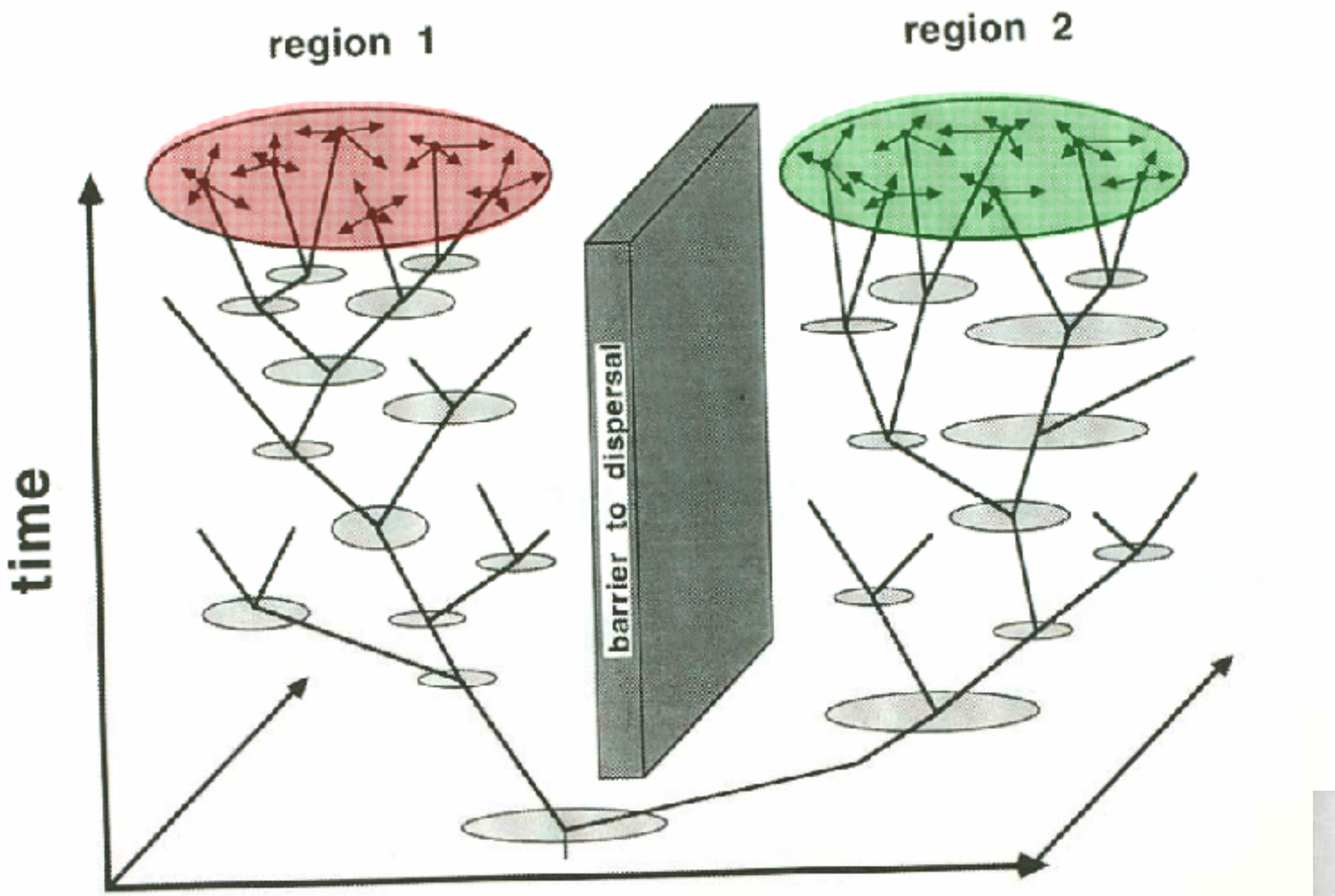
Figure 3. The origin and diversification of present-day Eurasian lineages (green). Dark green clouds represent events with uncertain dates or structures. Black bars represent time intervals for





Embrace the uncertainty!

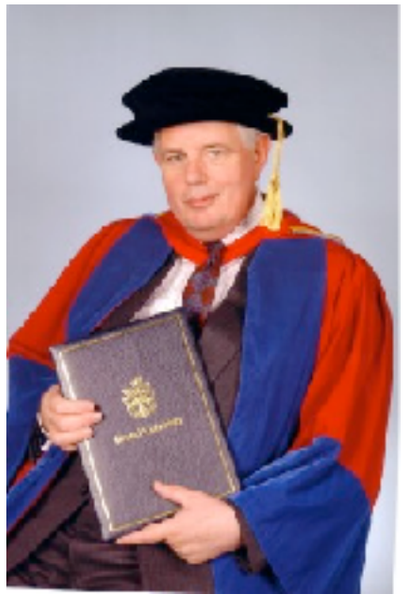
Figure 3. The origin and diversification of present-day Eurasian lineages (green). Dark green clouds represent events with uncertain dates or structures. Black bars represent time intervals for



~1987 Avise coins **“phylogeography”**

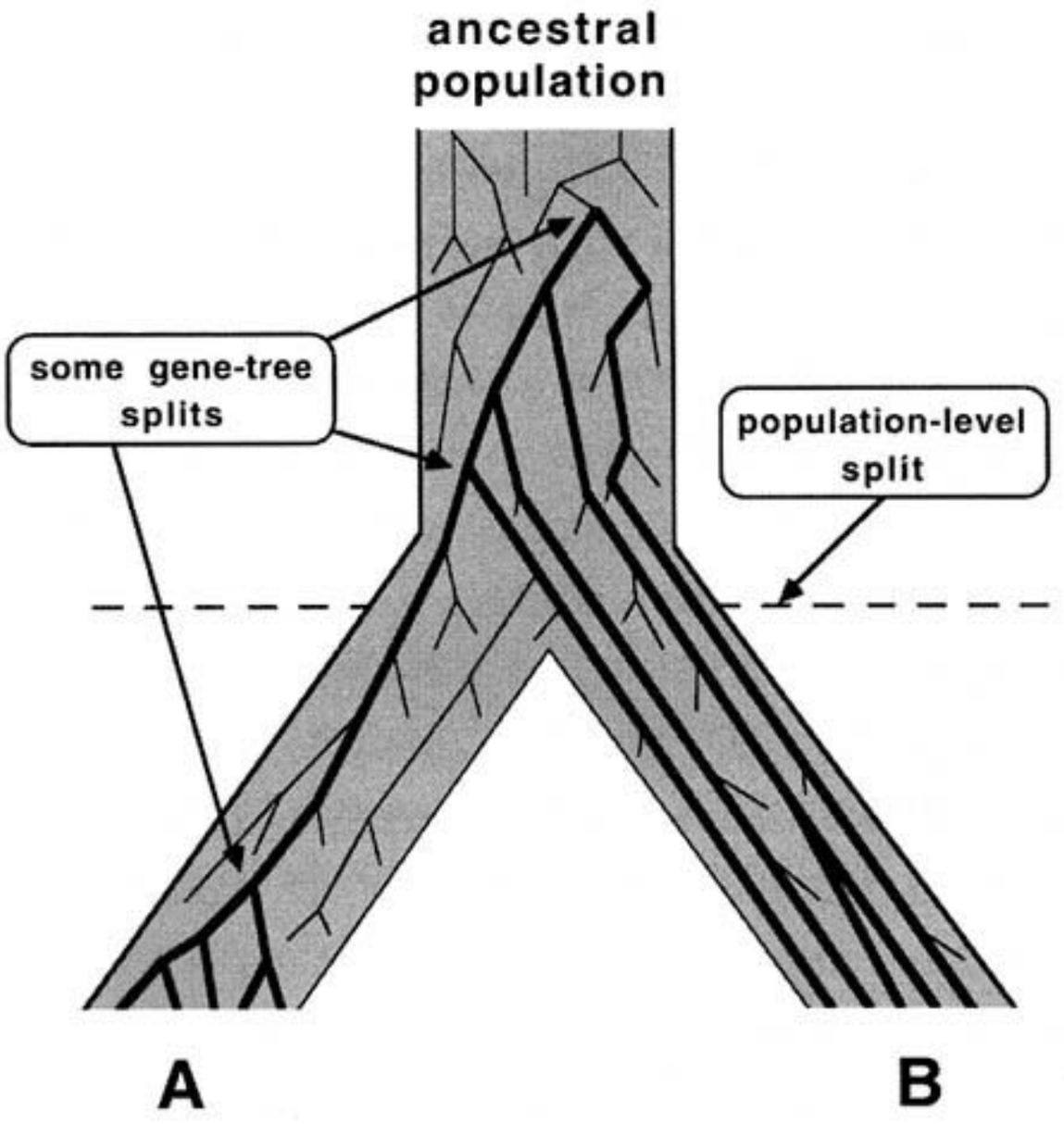
J. Avise

mtDNA Gene trees carry signature of species **demographic histories** (outgrowth of PCR revolution)

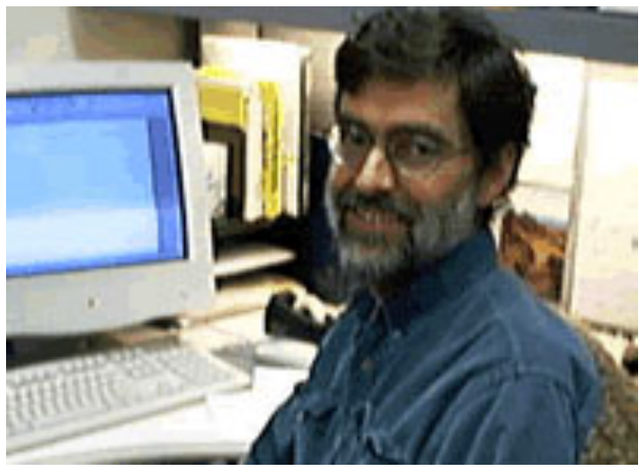


Kingman

$$P_c(t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \left(\frac{1}{2N_e}\right).$$



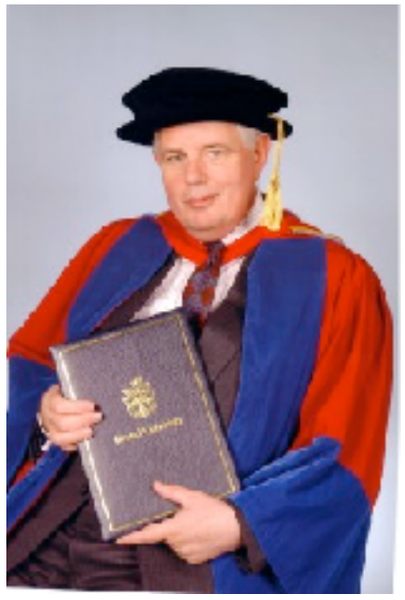
Tajima



Hudson

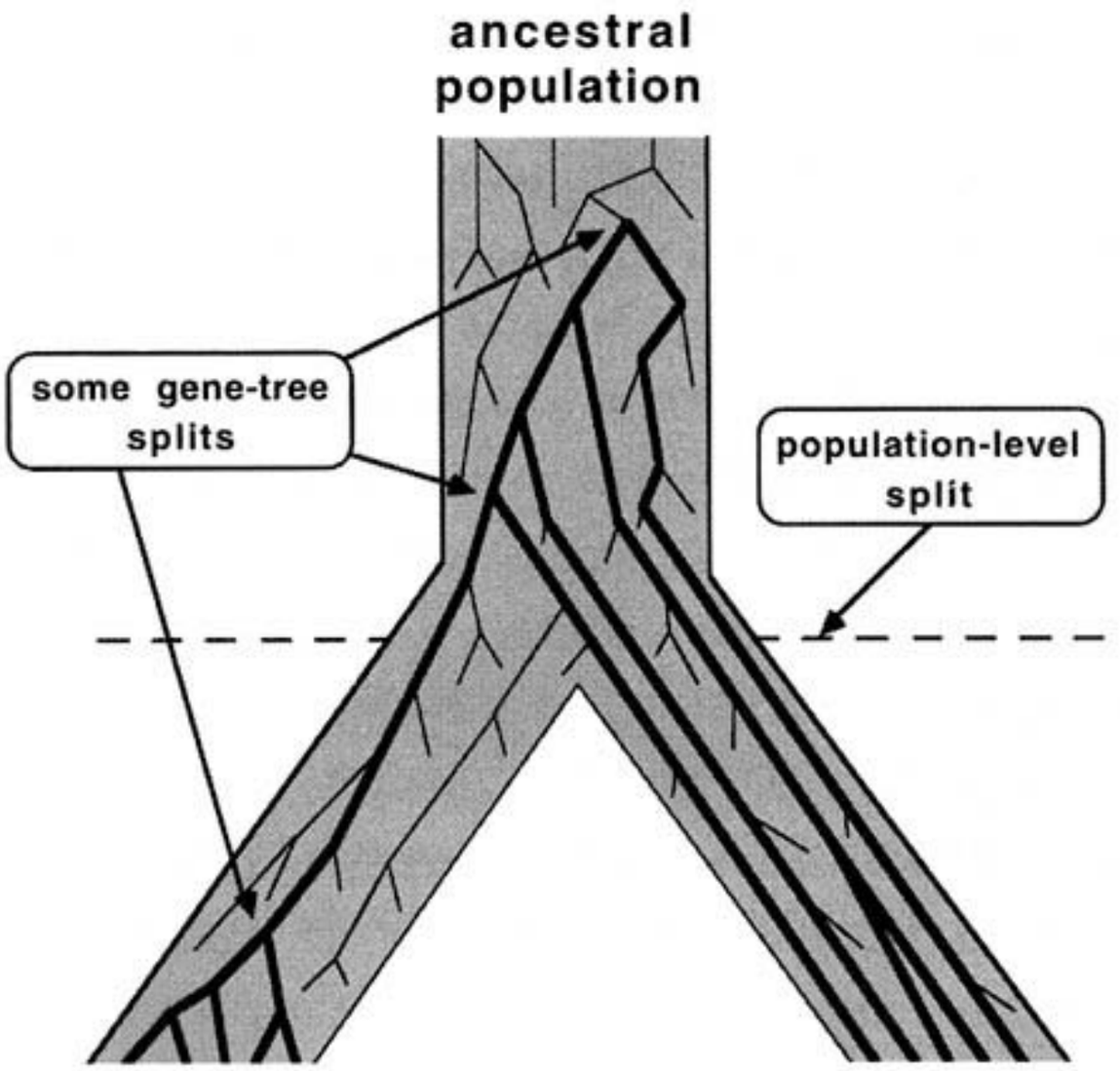
FIGURE 1.13 Fundamental distinction between a gene tree and a population tree or

Coalescent Theory (1983) slowly percolates into phylogeography

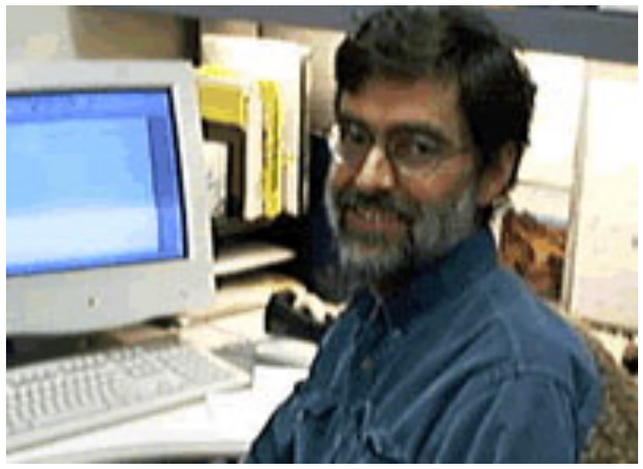


Kingman

$$P_c(t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \left(\frac{1}{2N_e}\right).$$



Tajima

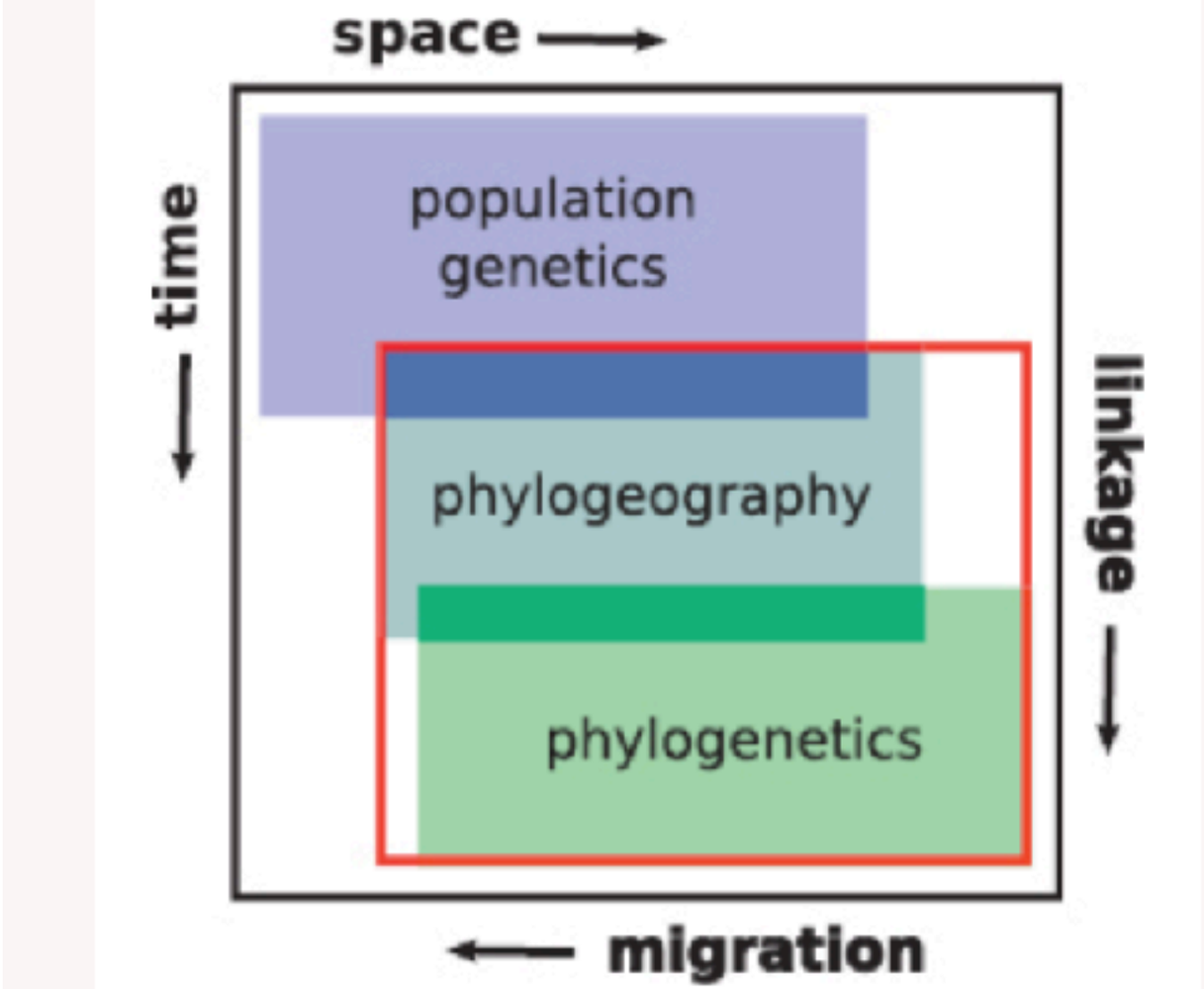


Hudson

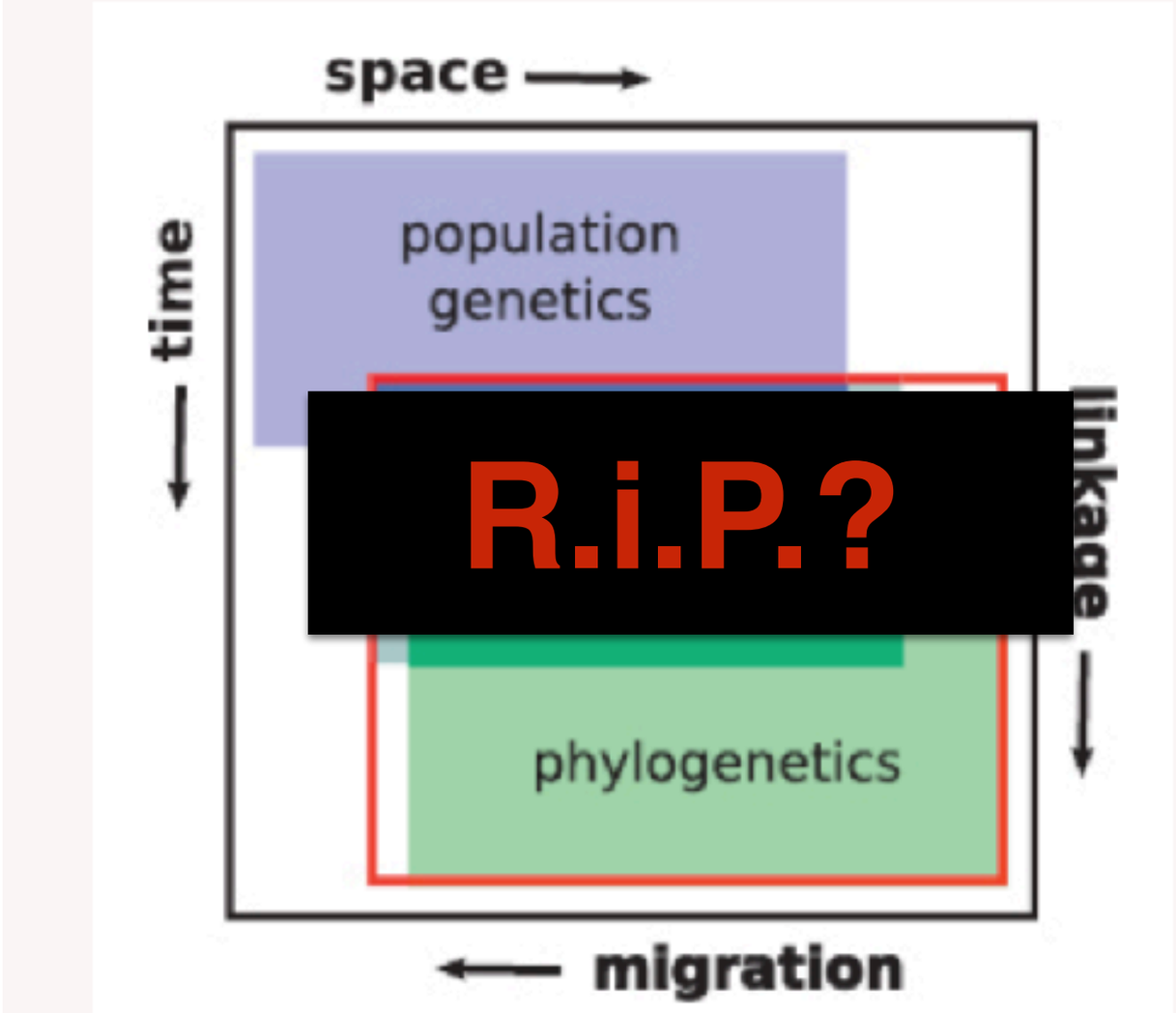
Headroom

FIGURE 1.13 Fundamental distinction between a gene tree and a population tree or

Coalescent Theory (1983) slowly percolates into phylogeography

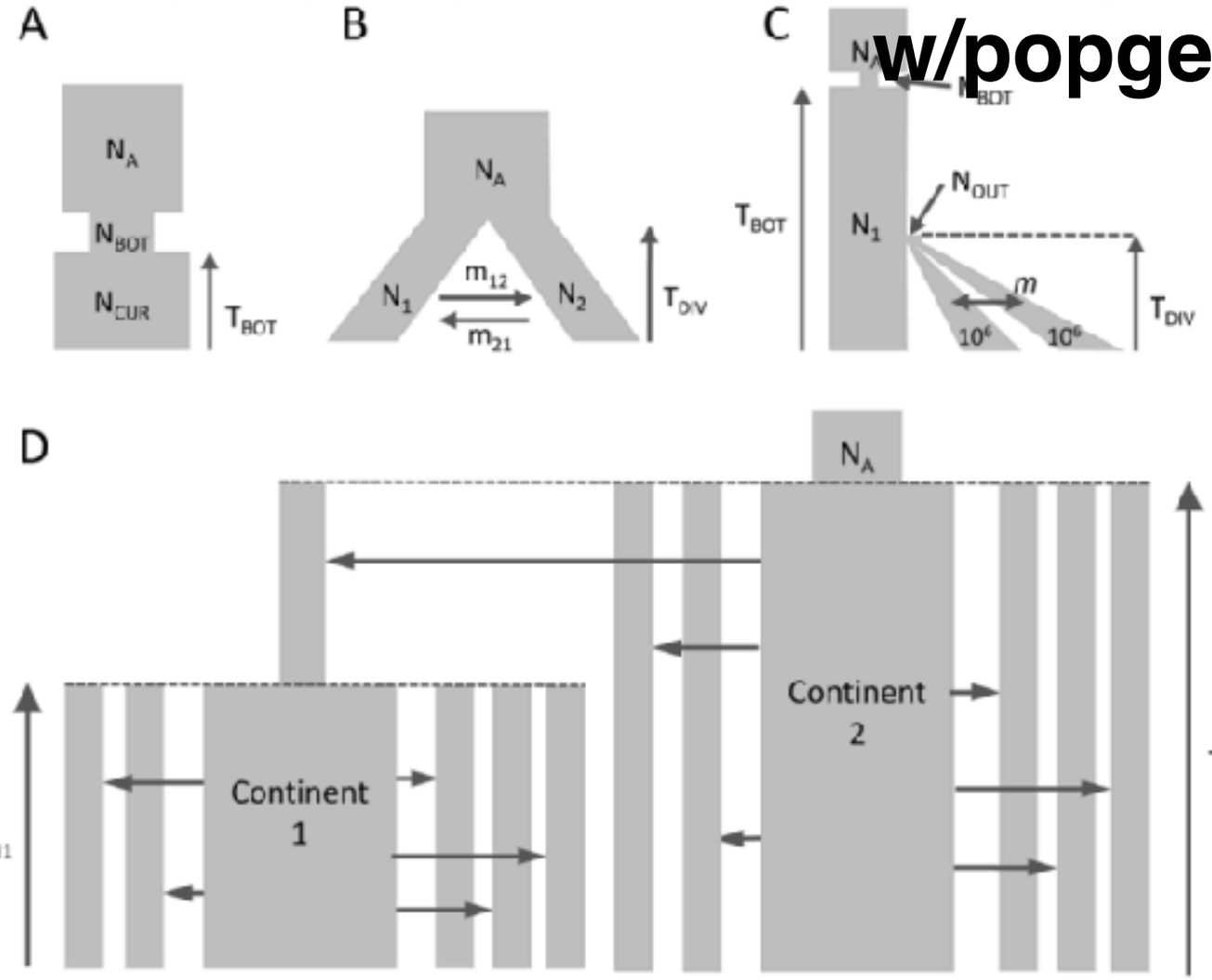


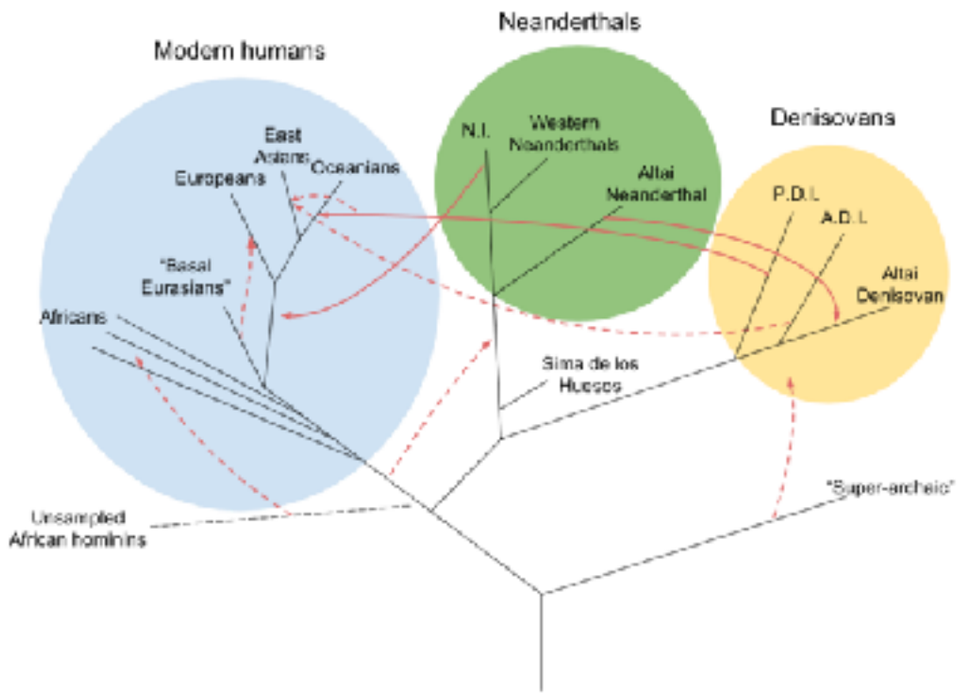
Avise et al. 1987 → *Edwards et al. 2016*



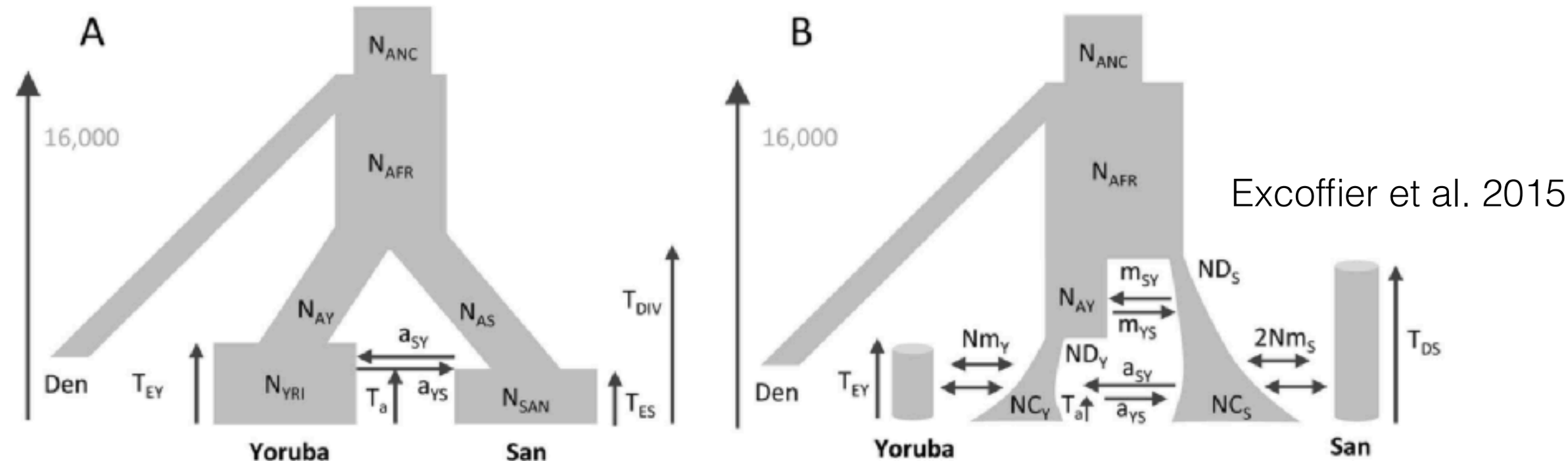
Avise et al. 1987 → *Edwards et al. 2016*

phylogeography ~= demographic historical inference w/popgen data





Dannemann & Racimo 2018



Excoffier et al. 2015

Human Genome 2001

how many genes?

how much encodes protein?

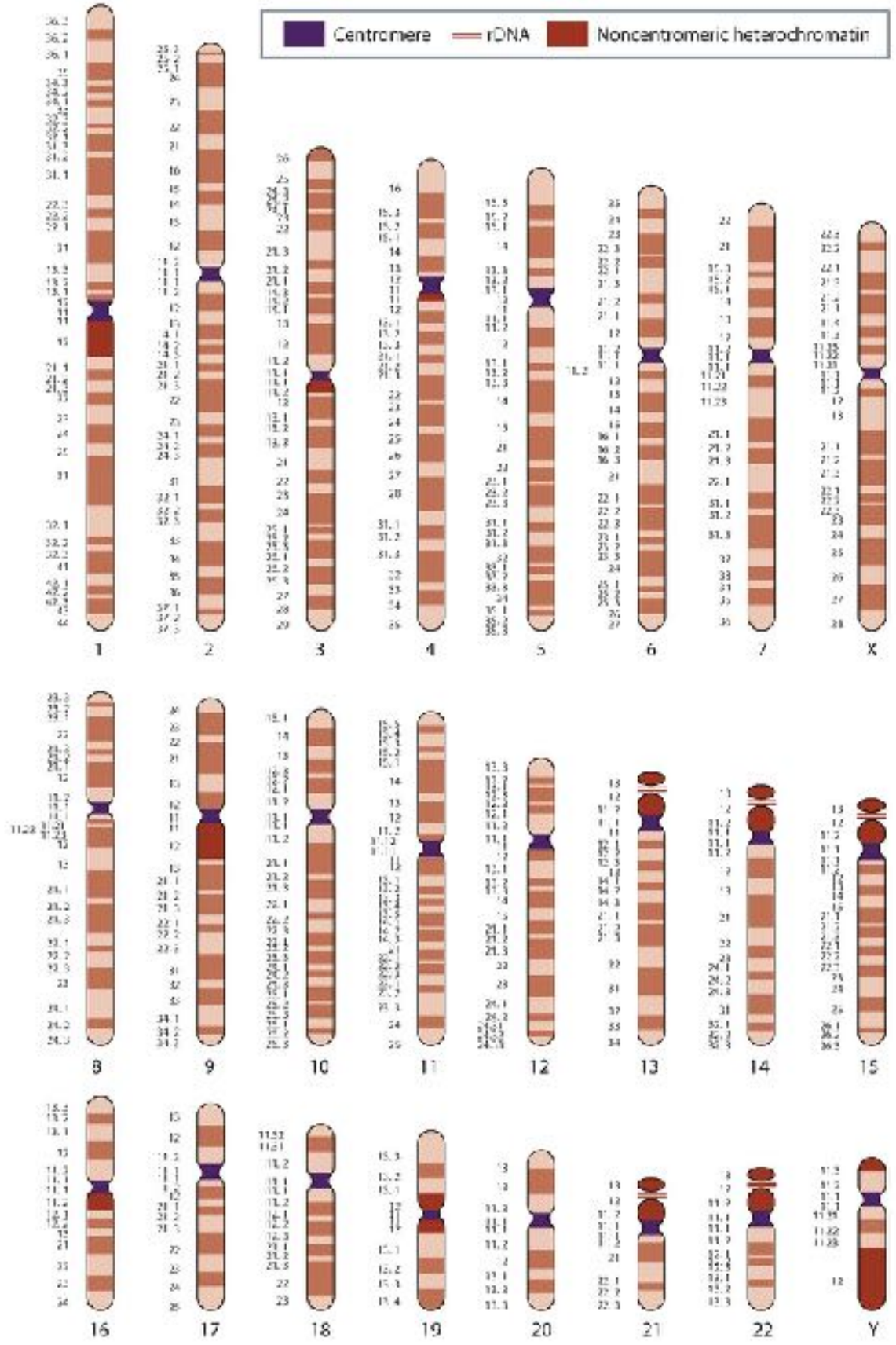


Figure 2.12 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

Human Genome 2001

how many genes?

19,000-20,000

how much encodes protein?

~1.5%

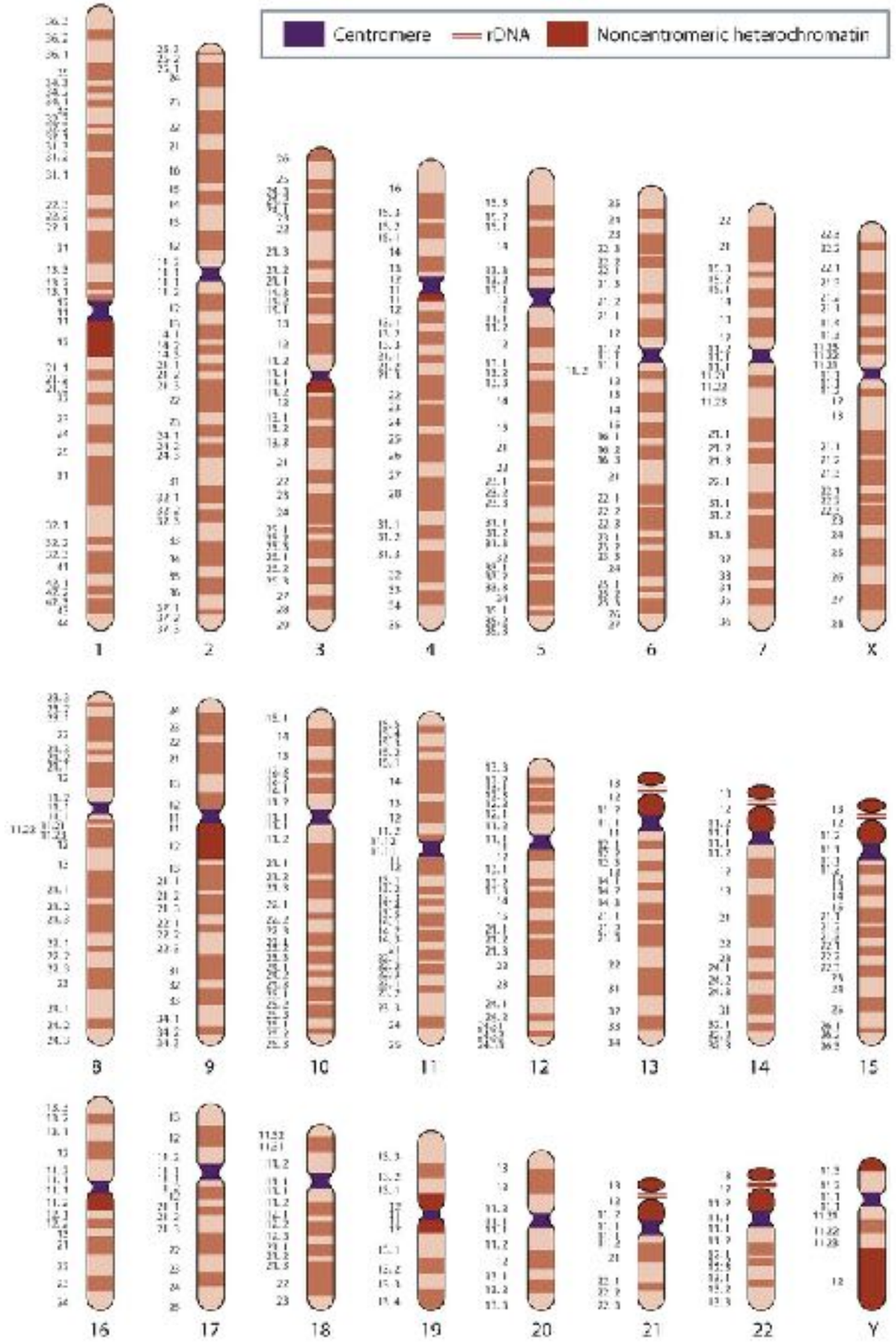


Figure 2.12 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

what about your study organism of interest?

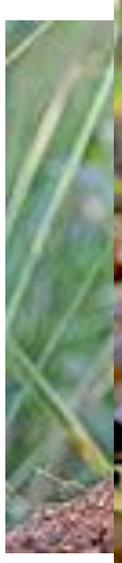
what about your study organism of interest?

what about your study organism of interest?

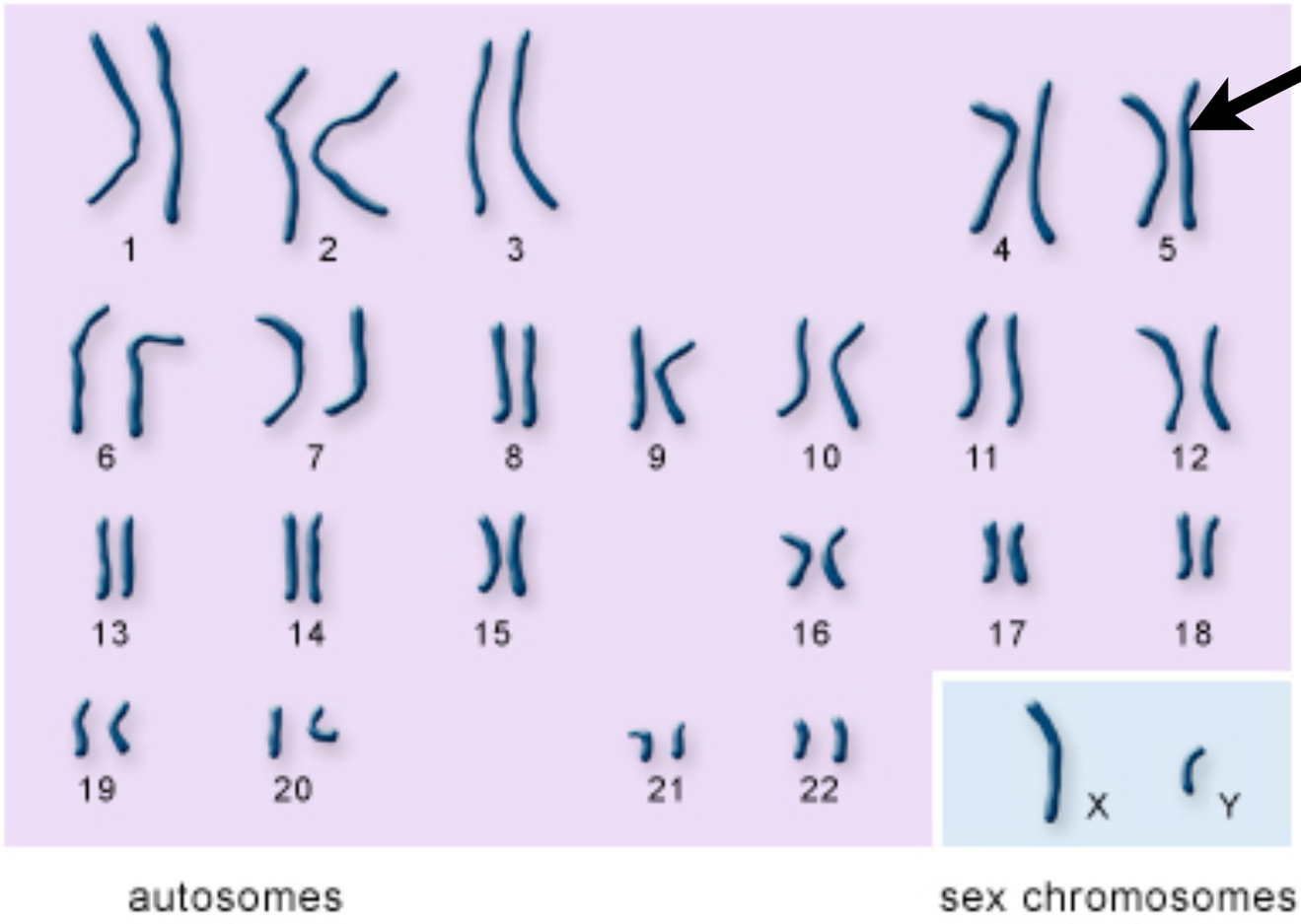
what about your study organism of interest?

what about your study organism of interest?

what about your study organism of interest?



A question

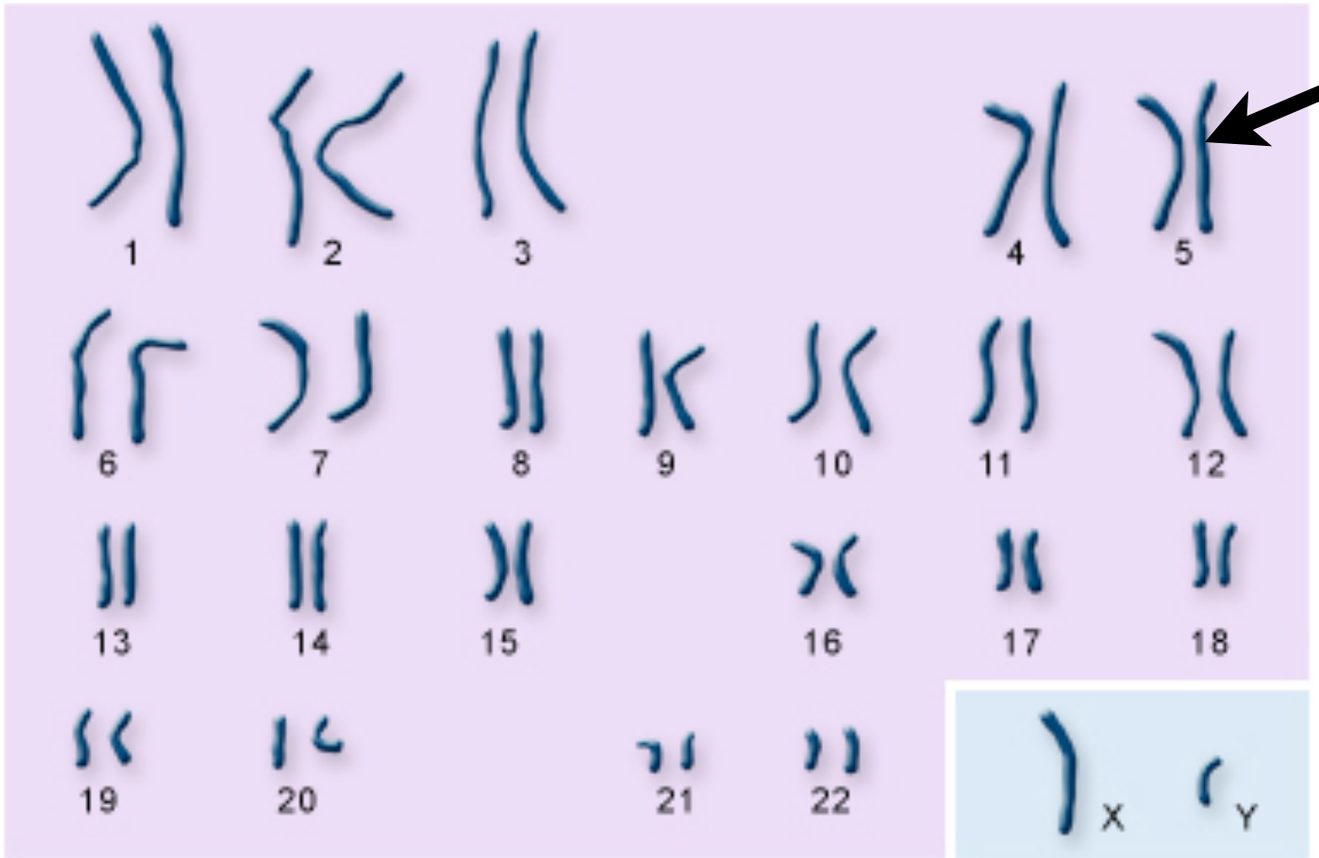


U.S. National Library of Medicine

Consider a single portion of a your genome.

At that locus, are you homozygous or heterozygous?

Intuitive answers...



U.S. National Library of Medicine

These are just some of the possible answers...

Why might you be homozygous?

1) Your mother and father both gave you identical copies they received from a very recent common ancestor (identity-by-descent - inbreeding)

2) There is only one version of the DNA that is viable (purifying selection)

Why might you be heterozygous?

1) Mutation rates at that locus are relatively high

2) Your parents are very distantly related (migration or large population size)

3) Being heterozygous for that locus is common because it is selectively favored (heterozygote advantage)

population genetics

- The study of genetic diversity
  - How much exists?
  - Why does it exist?
  - What can it tell us about the populations that carry it?
  - What can it tell us about the structure of genomes?
- A field that integrates multiple scales of biology: molecular-level processes (mutation and recombination) and population-level processes (migration, genetic drift, natural selection)

Preliminaries

- Overview of the syllabus and course structure

goto Course Website

overview

introductions

Simulators - SLIM and weeklong workshop (Nov 4-8)

Simulation assignments

Official discussions

Final Projects

Schedule

introductions

name

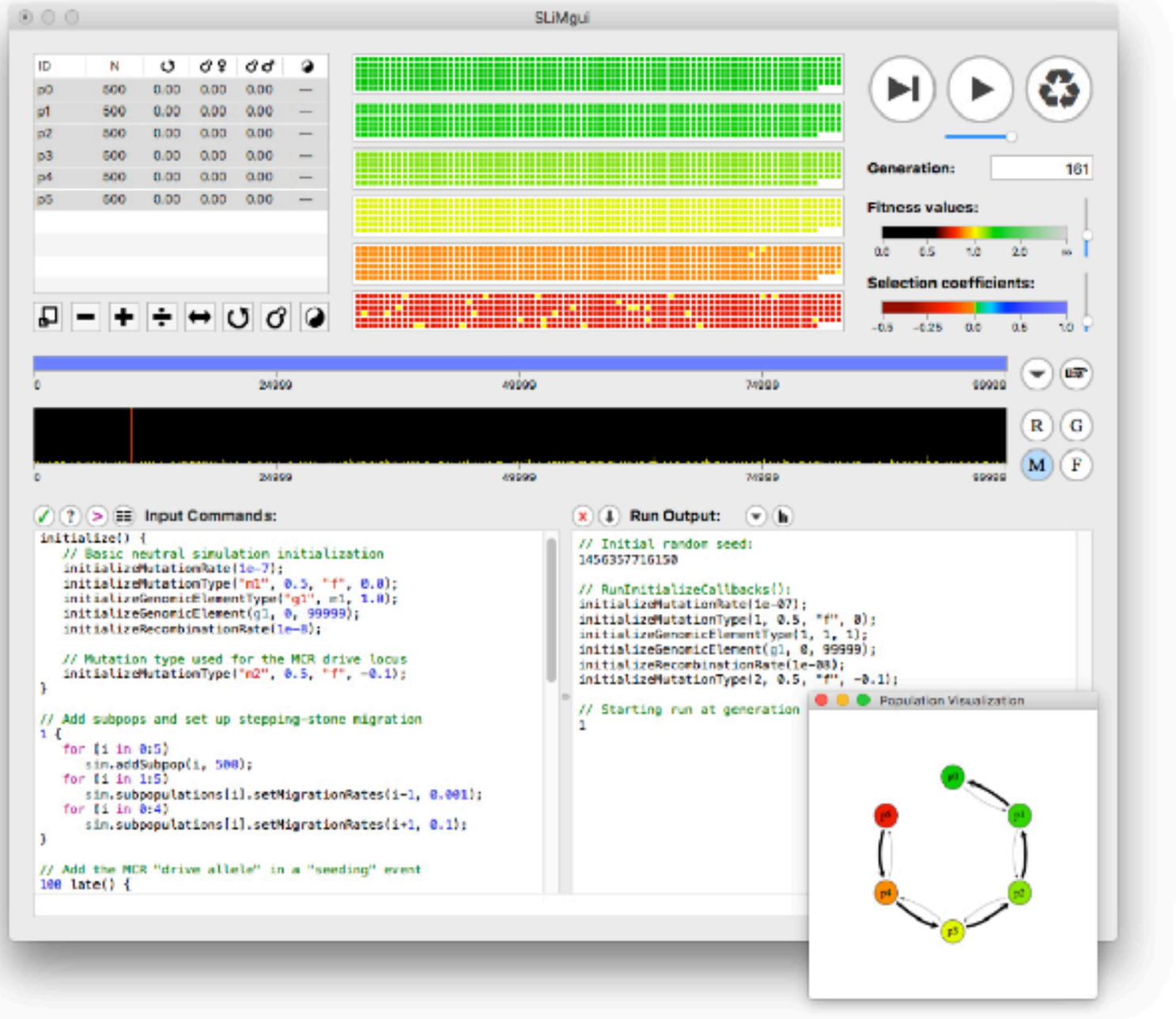
research interest

hobby / outside interest

Simulators - SLIM and weeklong workshop (Nov 4-8)

SLiMgui

With SLiMgui for Mac OS X you can visualize your simulation as it runs and examine its parameters in real-time, allowing for much easier simulation development.



But, we'll look at other ones (like Pipe-Master)
and you can use anything that is best for your final project (more on that)

Simulation assignments

simulation assignments due at beginning of class.

briefly demonstrate your simulation results at the start of class

You are free to work in groups, but each individual is responsible for designing and running simulation in addition to analyzing and/or visualizing the results.

e.g. co-demographic history

Coalescent Model

Prediction

Inference

data

DNA barcodes (ie mtDNA)

multiple taxa

"whole" genomes

Official discussions

two randomly selected people will lead the discussion on the assigned chapter of Hahn (or paper, TBA).

Not for every week (it will be announced the week before)

After you lead two discussions, you are off the hook

Final Projects

1. Simulation-based Inference with your data or available data

or

2. stand-alone simulation experiments that answer questions about observable predictions under different hypotheses/ models

Final Projects

1. Simulation-based Inference with your data or available data

or

2. stand-alone simulation experiments that answer questions about observable predictions under different hypotheses/ models

Often done with machine learning or ABC, but there are other flavors (more on this later)

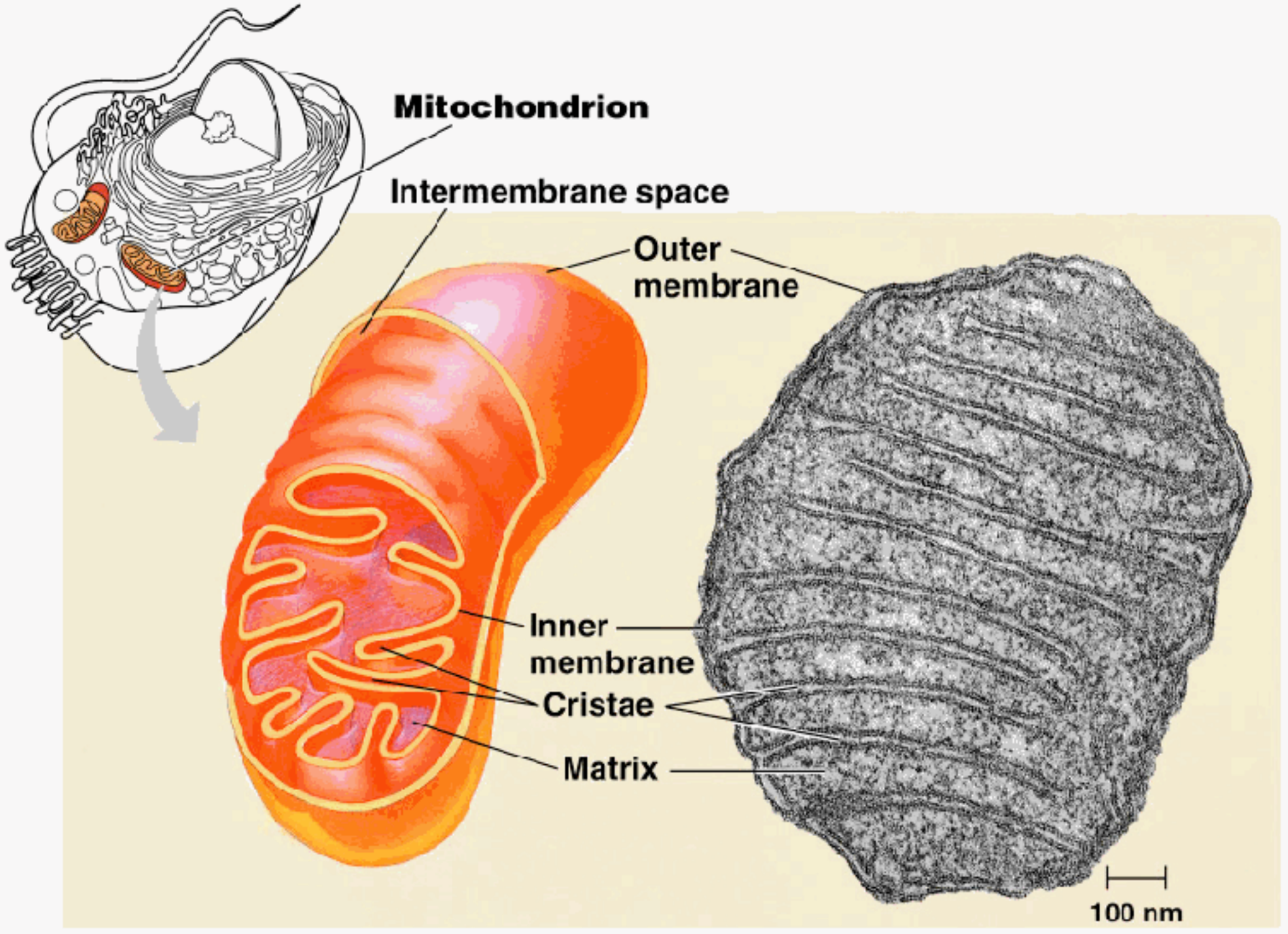
Can be modest, but the idea is to be part of your downstream publication (and part of your dissertation)

Oral presentation of final project at the end of the semester (15 min, 3 min for questions)

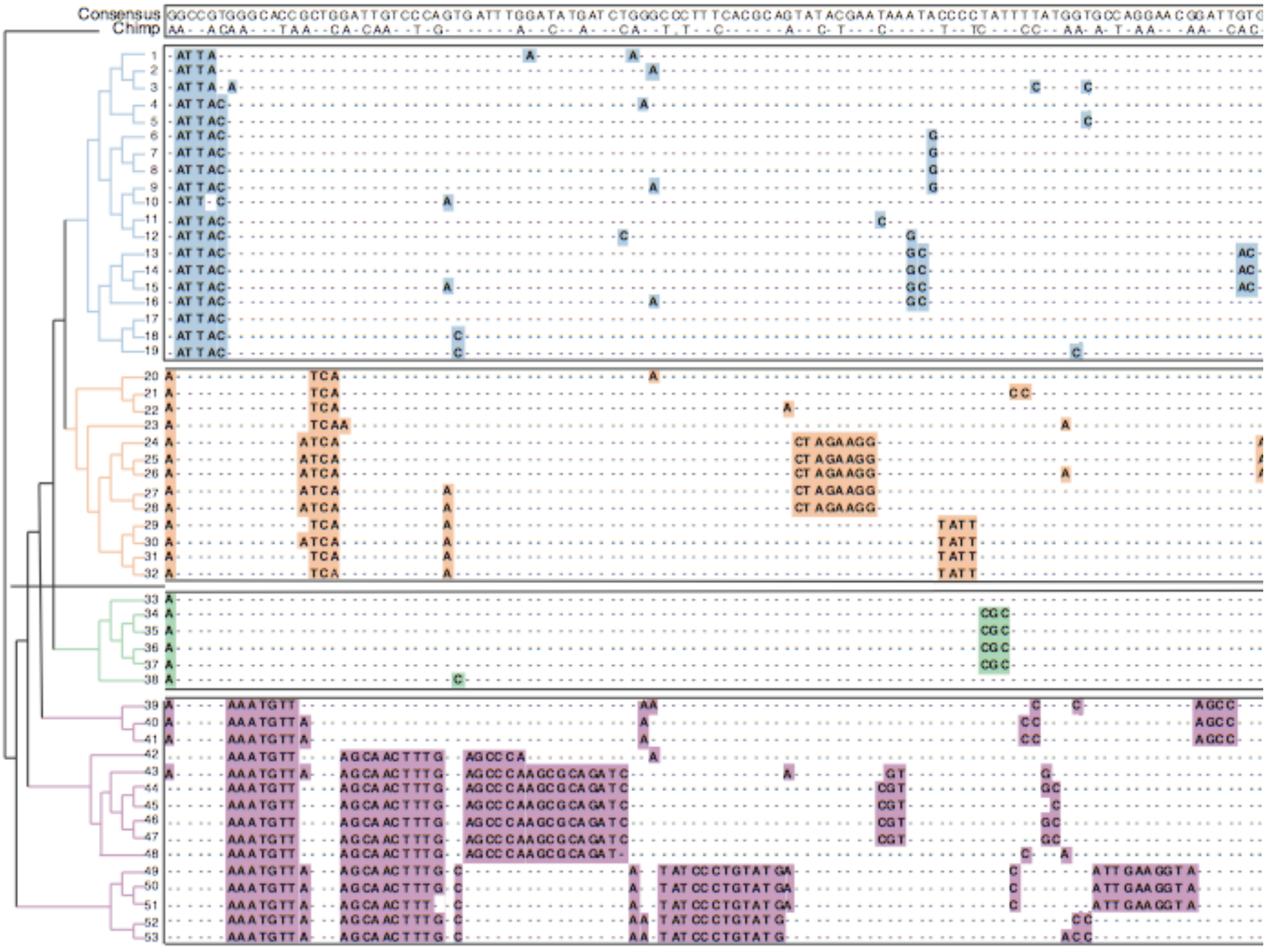
Schedule

Tentative Schedule

Date	Topic	Reading
Sept 6	Models of Evolution	Chapter 1
Sept 13°	Genetic Variation	Chapter 3
Sept 20*	The Coalescent model	Chapter 6
Sept 27*	Recombination	Chapter 4
Oct 4°	Population Structure	Chapter 5
Oct 11	MIDTERM	
Oct 18	Direct Selection	Chapter 7
Oct 25*°	Linked Selection	Chapter 8
Nov 1*°	Demographic History part 1 (guest appearence by Ariadna Morales - phrapl)	Chapt 9 (203-221)
Nov 4-8	5 day workshop by Ben Haller	SLiM
Nov 15	Demographic History part 2	Chapt 9 (222-248)
Nov 22°	Special topics	TBA
Dec 6	Project Presentations	N/A
Dec 13	Project Presentations	N/A



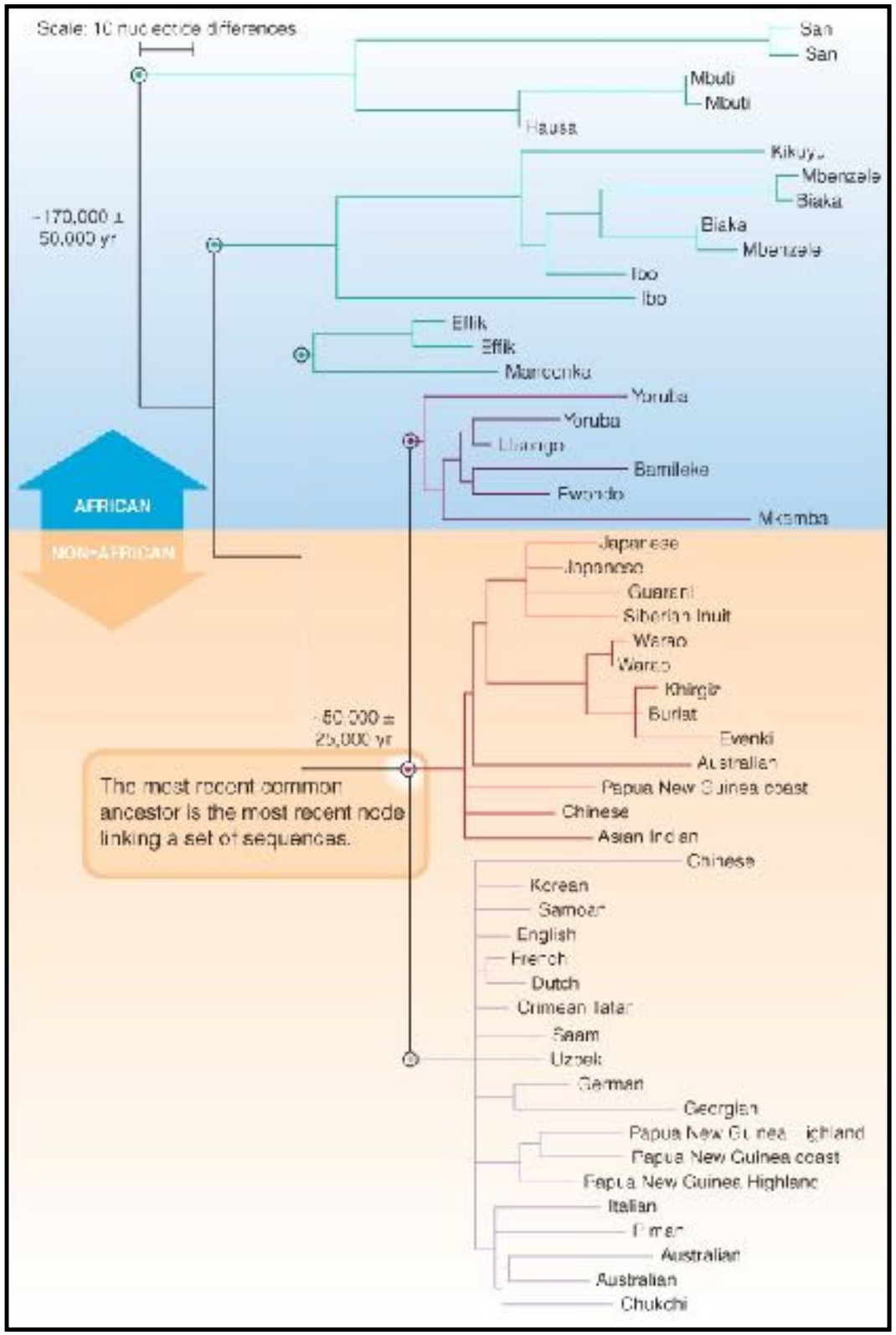
>gi|251831106|ref|NC_012920.1| Homo sapiens mitochondrion, complete genome



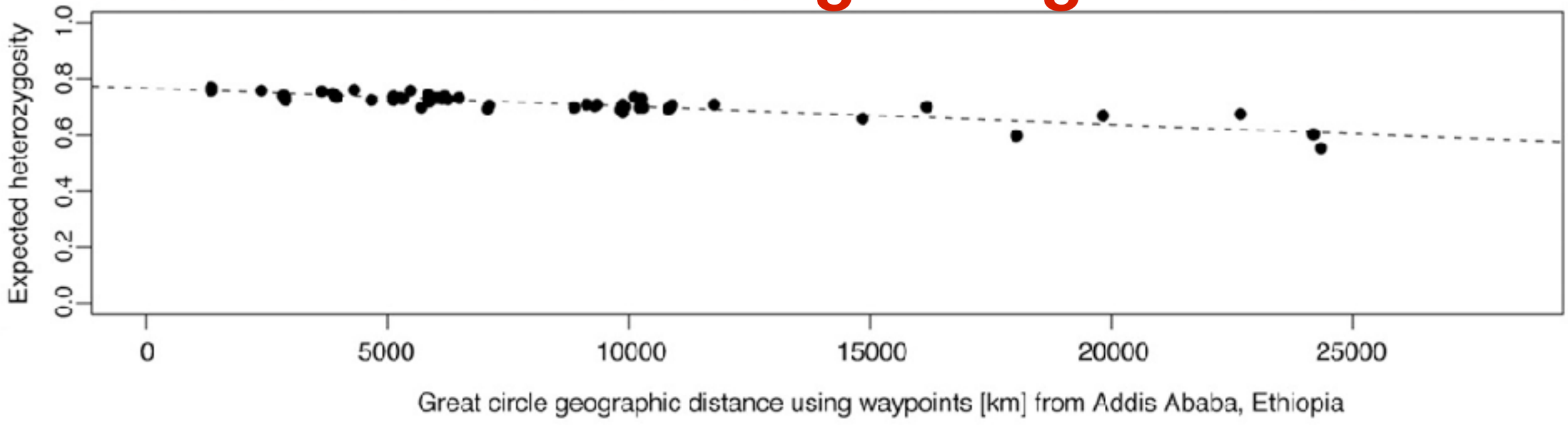
Ingman et al (2000)

All human mitochondrial DNA sequences have a common ancestor “Eve” ~170k years ago

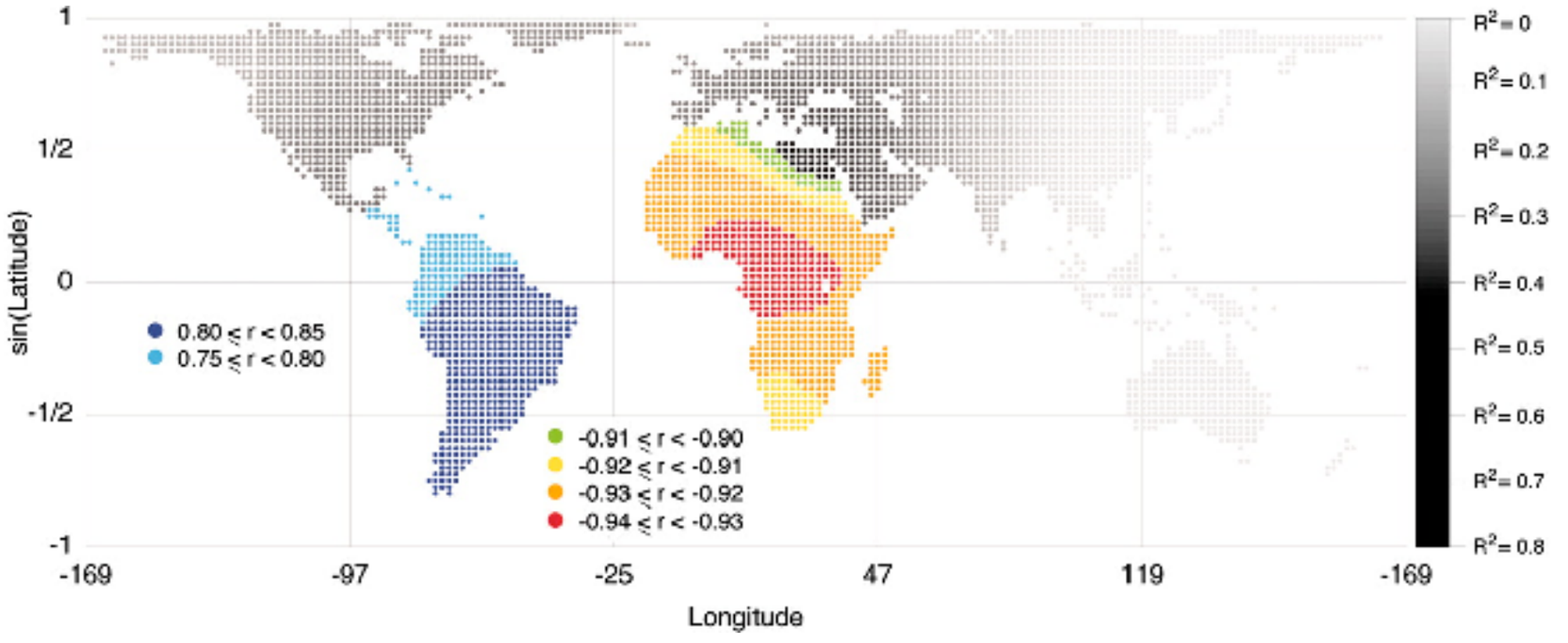
Non-African sequences have a common ancestor at ~50k years ago



Geographic patterns in genetic diversity serial founder effect originating in Africa

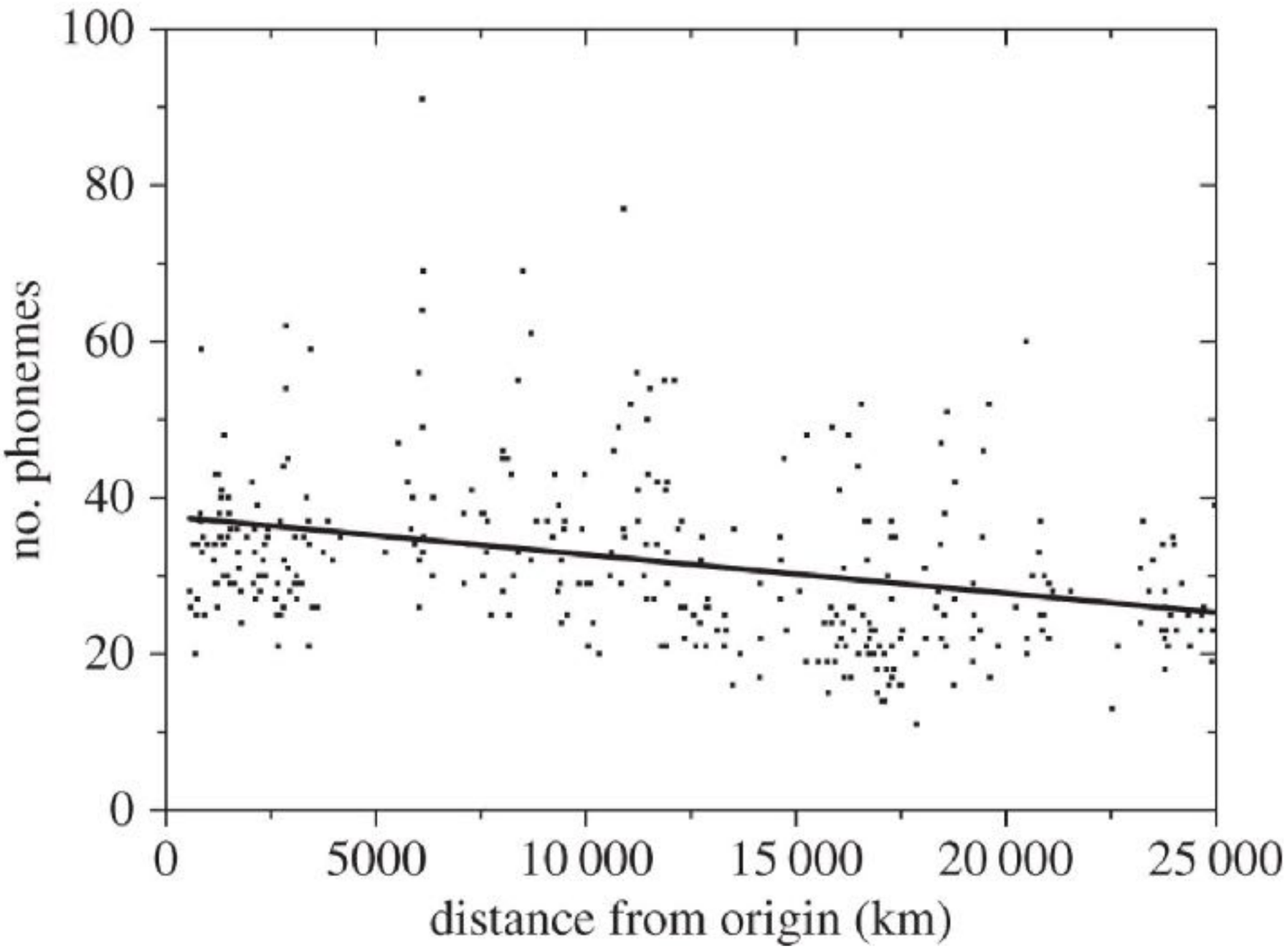


Ramachandran et al (2005)

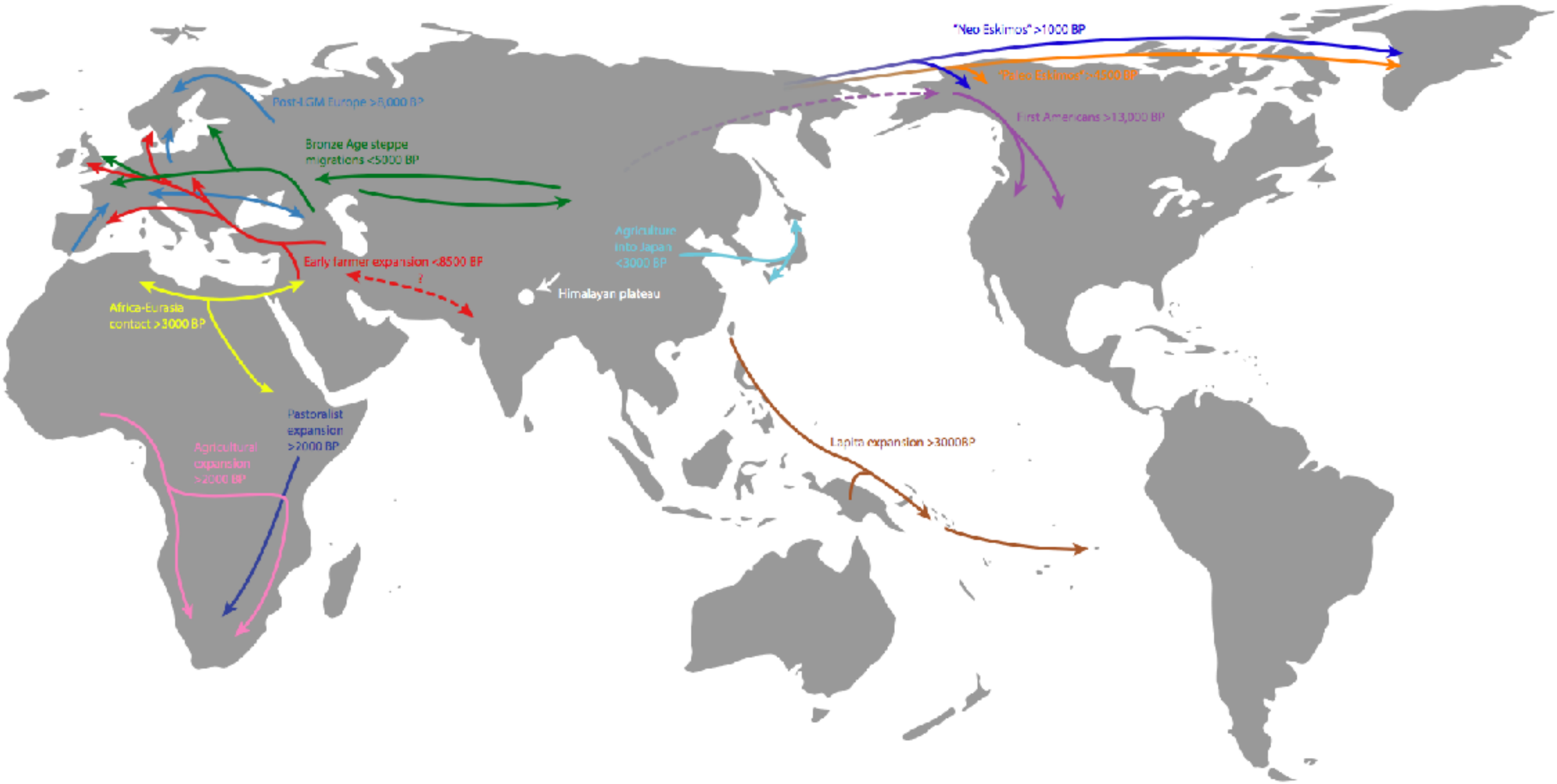


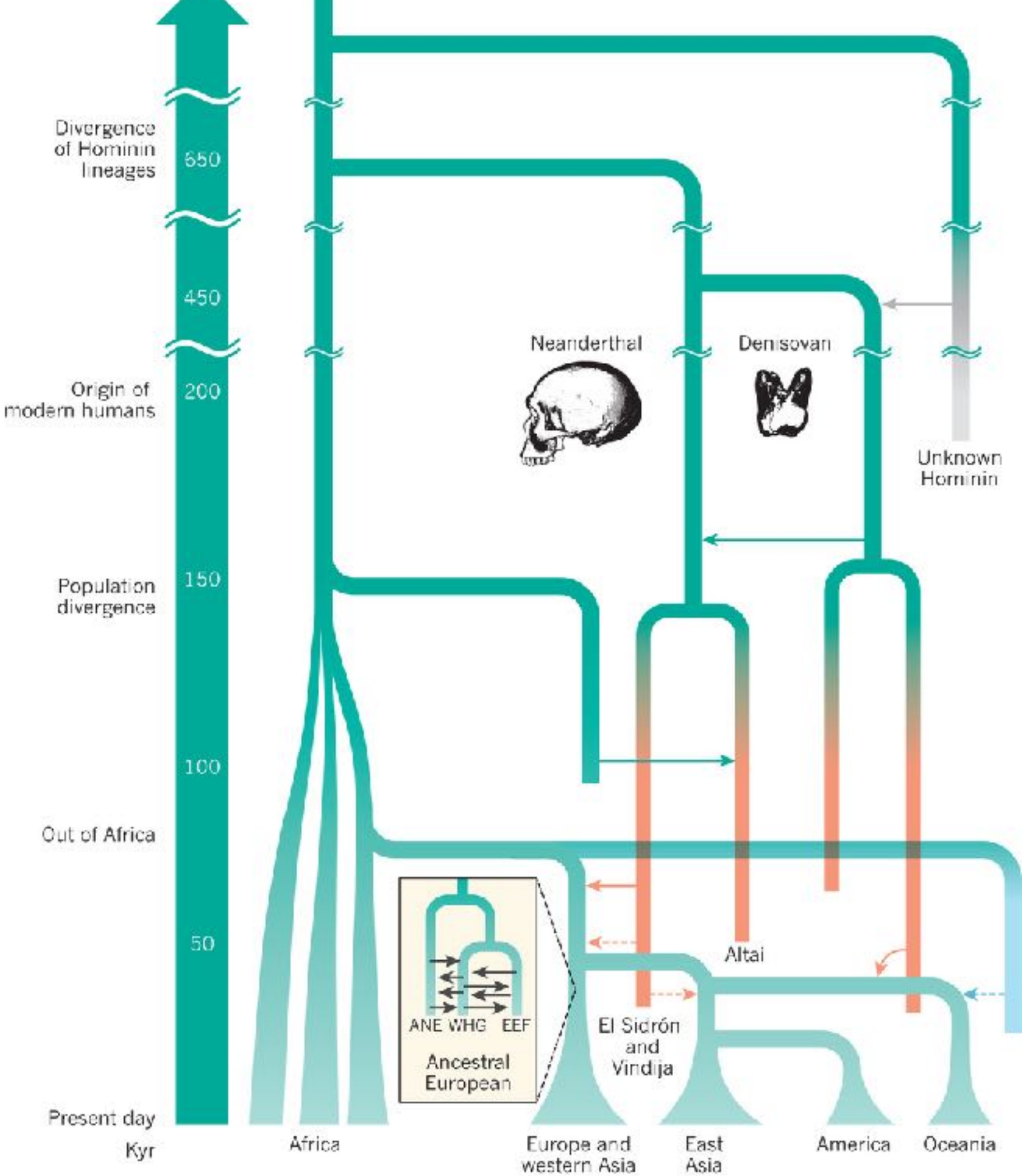
Can a linguistic serial founder effect originating in Africa explain the worldwide phonemic cline?

Fort and Pérez-Losada 2016

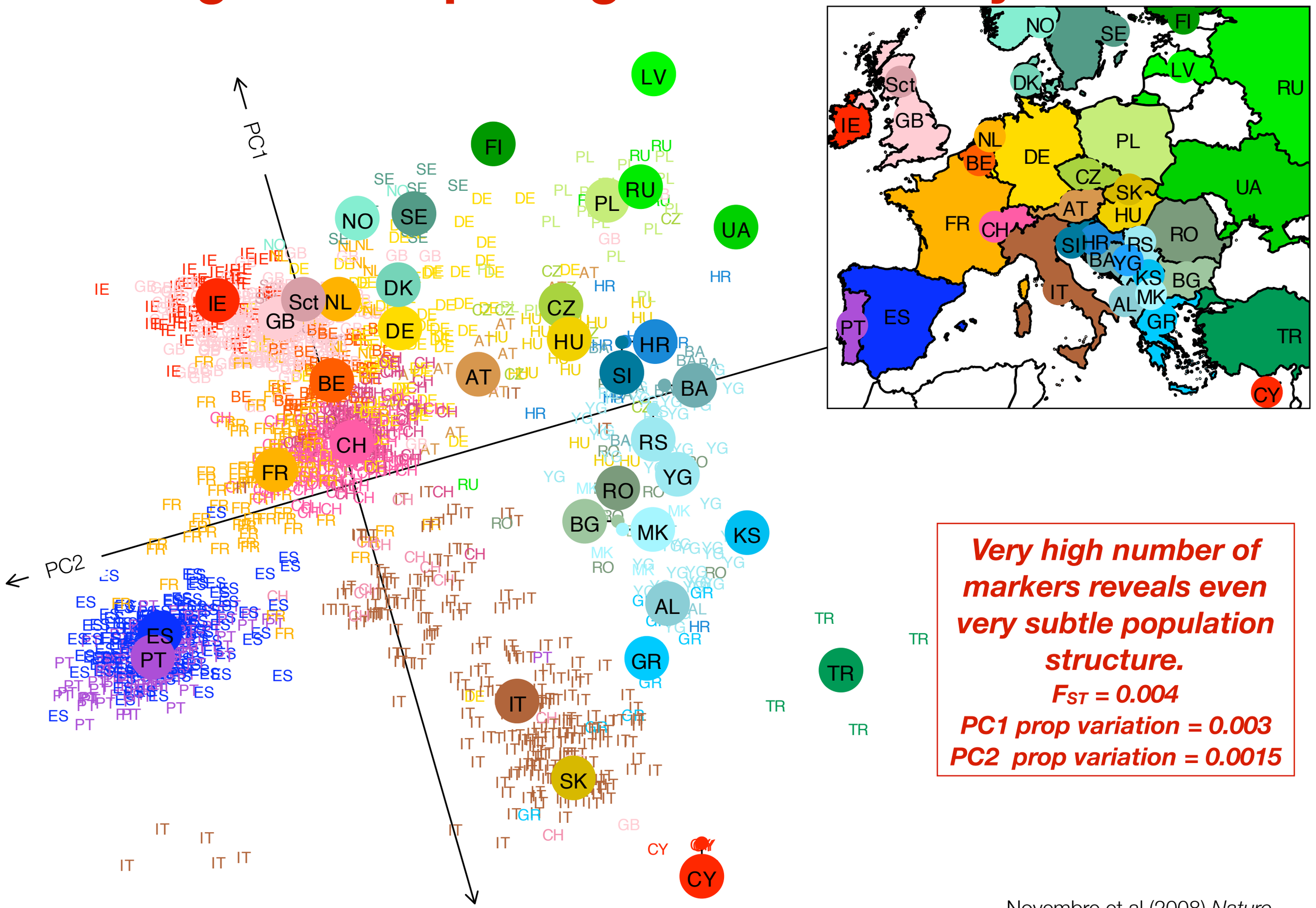


phoneme: units of sound that distinguish one word from another in a particular language.


"Neo Eskimos" >1000 BP
"Paleo Eskimos" >4500 BP
First Americans >13,000 BP
Post-LGM Europe >8,000 BP
Bronze Age steppe migrations <5000 BP
Agriculture into Japan <3000 BP
Early farmer expansion <8500 BP
Himalayan plateau
Africa-Eurasia contact >3000 BP
Pastoralist expansion >2000 BP
Agricultural expansion >2000 BP
Lapita expansion >3000BP


Divergence of Hominin lineages
650
450
Origin of modern humans
200
Population divergence
150
100
Out of Africa
50
Present day
Kyr
Neanderthal
Denisovan
Unknown Hominin
ANE WHG EEF
Ancestral European
Altai
El Sidrón and Vindija
Africa
Europe and western Asia
East Asia
America
Oceania

Zooming in: European genetic diversity



Very high number of markers reveals even very subtle population structure.

F_{ST} = 0.004

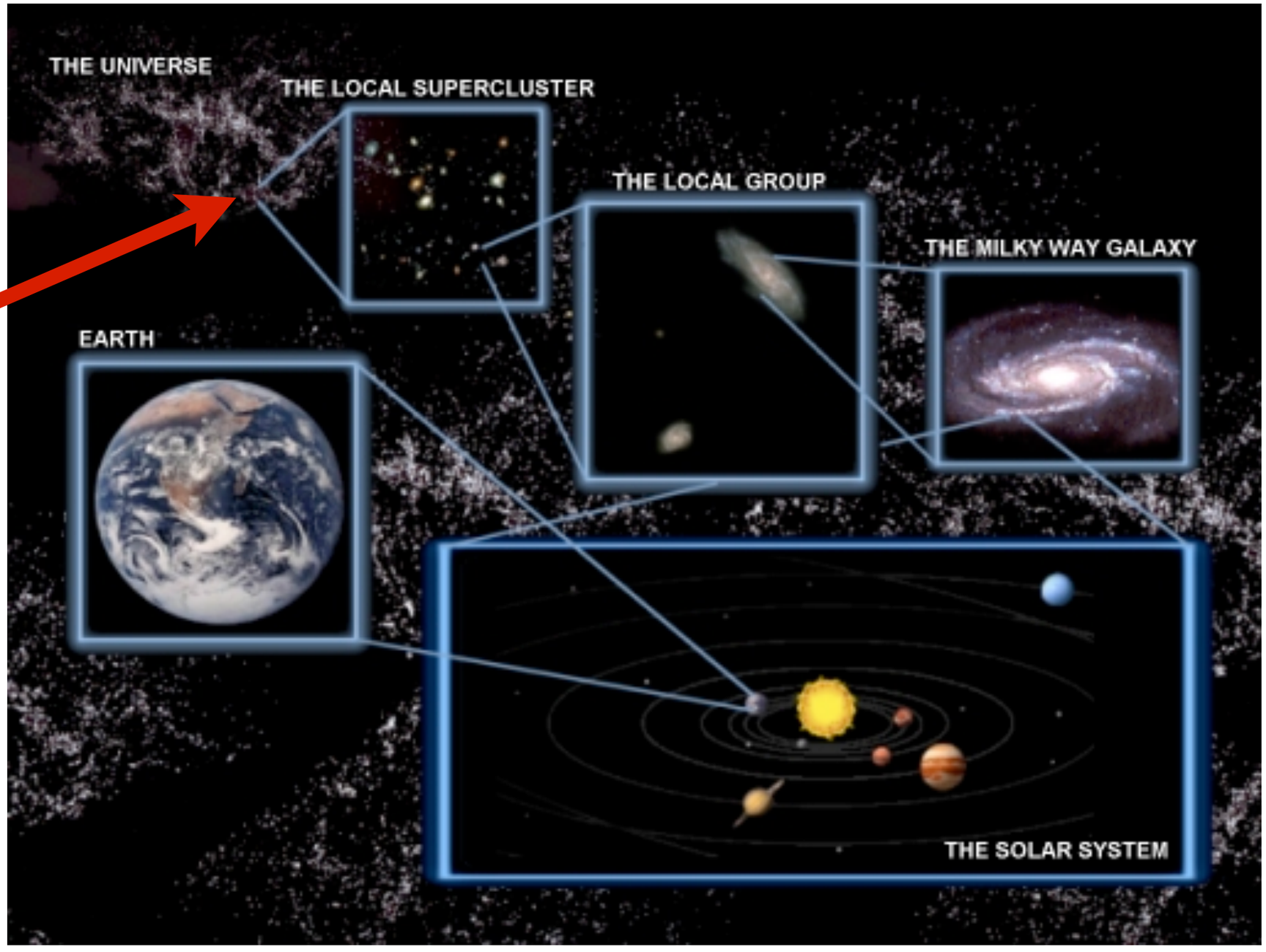
PC1 prop variation = 0.003

PC2 prop variation = 0.0015

Novembre et al (2008) *Nature*

Our cosmic address

You are here



Our evolutionary address:

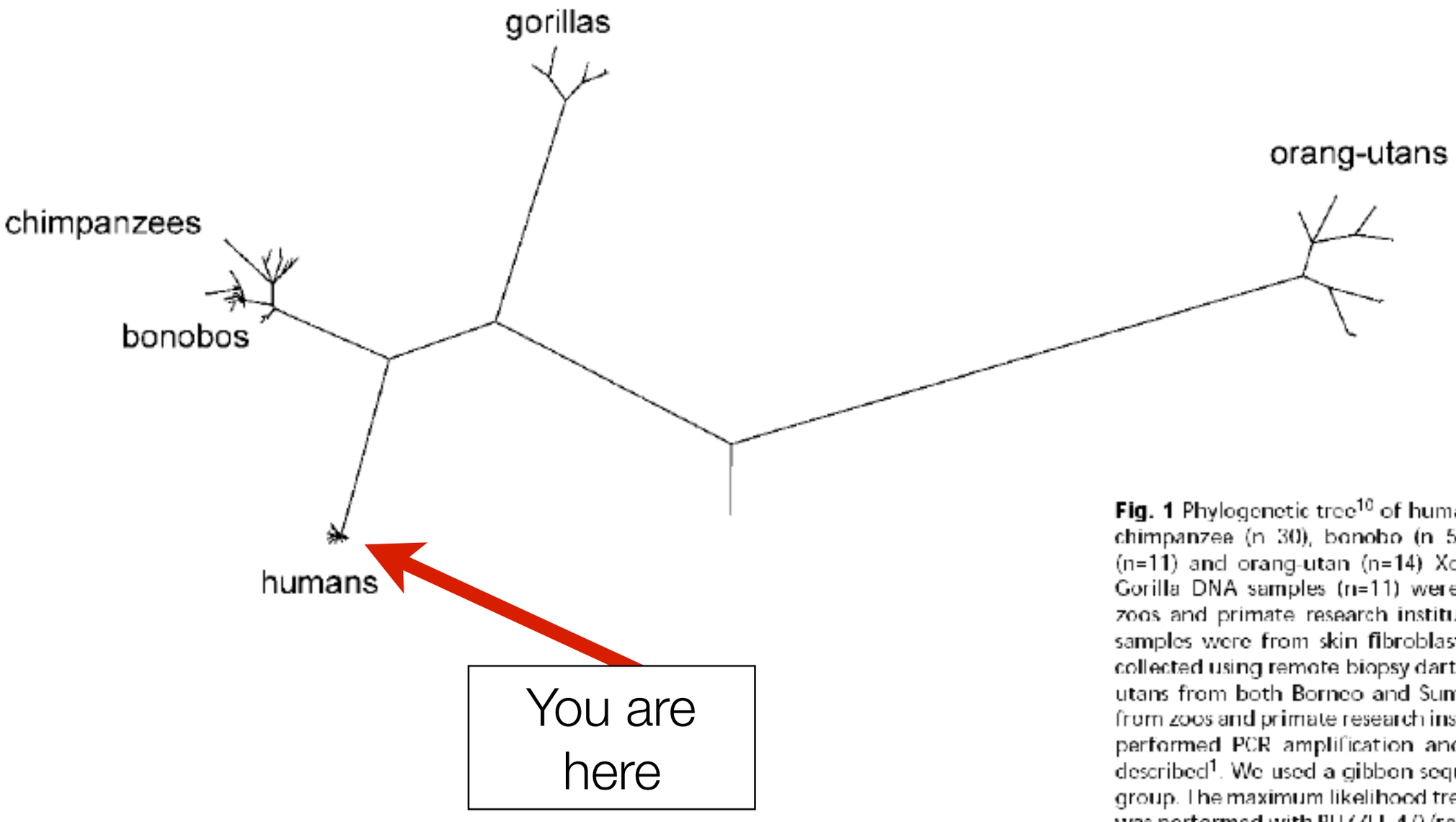
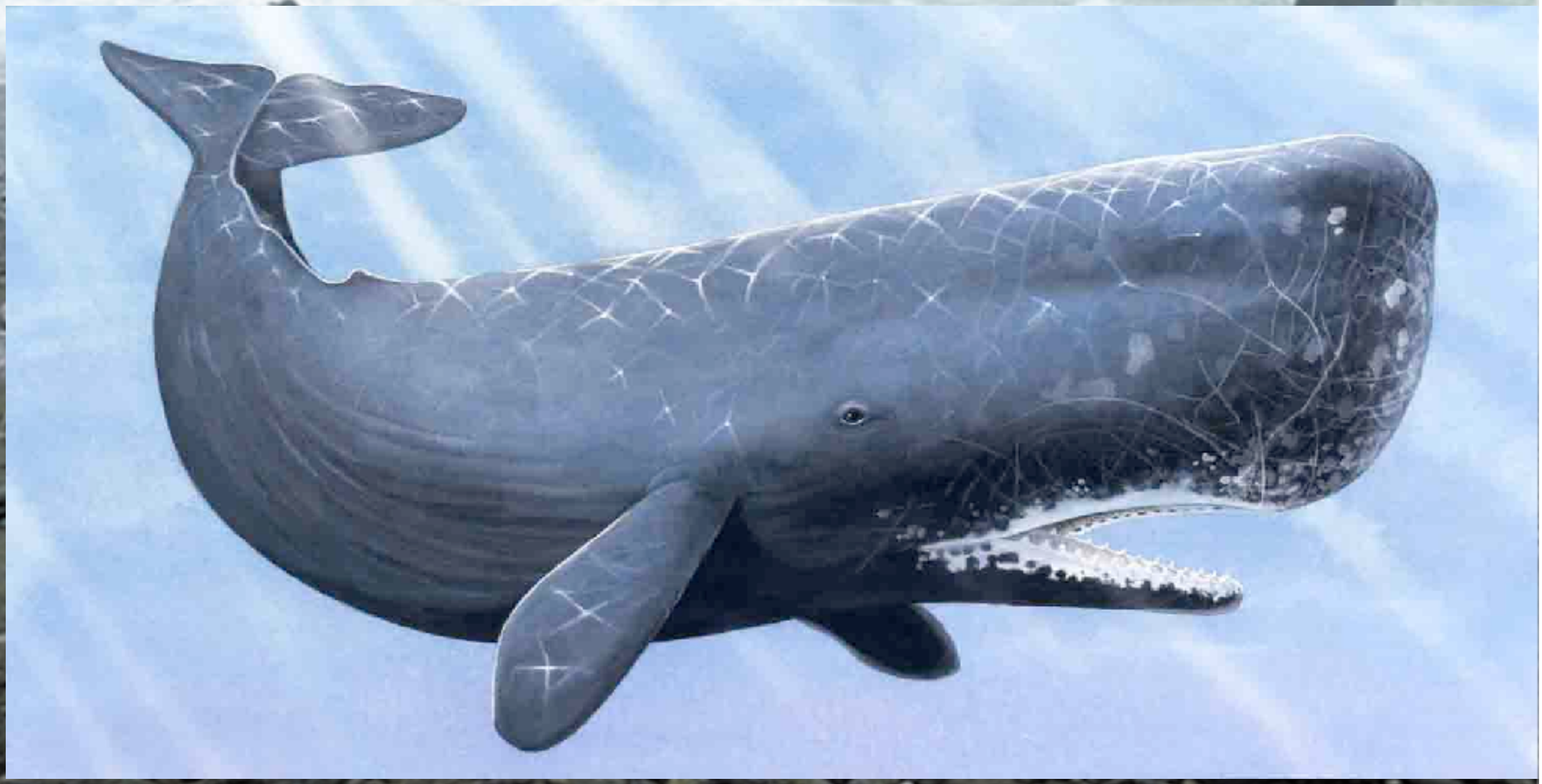


Fig. 1 Phylogenetic tree[10] of human (n=70; ref. 2), chimpanzee (n 30), bonobo (n 5; ref. 1), gorilla (n=11) and orang-utan (n=14) Xq13.3 sequences. Gorilla DNA samples (n=11) were obtained from zoos and primate research institutes. Orang-utan samples were from skin fibroblast cell lines (n=9; collected using remote biopsy darts) of wild orang-utans from both Borneo and Sumatra[4] as well as from zoos and primate research institutes (n=5). We performed PCR amplification and sequencing as described[1]. We used a gibbon sequence as an out-group. The maximum likelihood tree reconstruction was performed with PUZZLE 4.0 (ref. 10) assuming a Tamura-Nei model with γ-distributed rates.

Autosomal gene tree for Xq13.3

Kaessmann et al (2001) Nature Genetics

Identification ….. (DNA barcoding) …….
Sperm Whale (remains of)



Identification ….. (DNA barcoding) …….
Sperm Whale (remains of)

Comes in handy when tracing illegal trade in whale meat

DNA busted restaurants in LA and Seoul

Jalopy Records
315 Columbia St
Brooklyn, NY 11231

Sasquatch and Yeti Variations

Sasquatch and Yeti Variations Vol. 3

by Mike Hickerson

Sasquatch and Yeti Variations Vol. 3a1 00:00 / 22:32

Digital Album

Streaming + Download

Includes high-quality download in MP3, FLAC and more. Paying supporters also get unlimited streaming via the free Bandcamp app.

Buy Digital Album name your price

Send as Gift

1. Sasquatch and Yeti Variations Vol. 3a1
   22:32
2. Sasquatch and Yeti Variations Vol. 3a2

Sasquatch and Yeti Variations
Brooklyn, New York

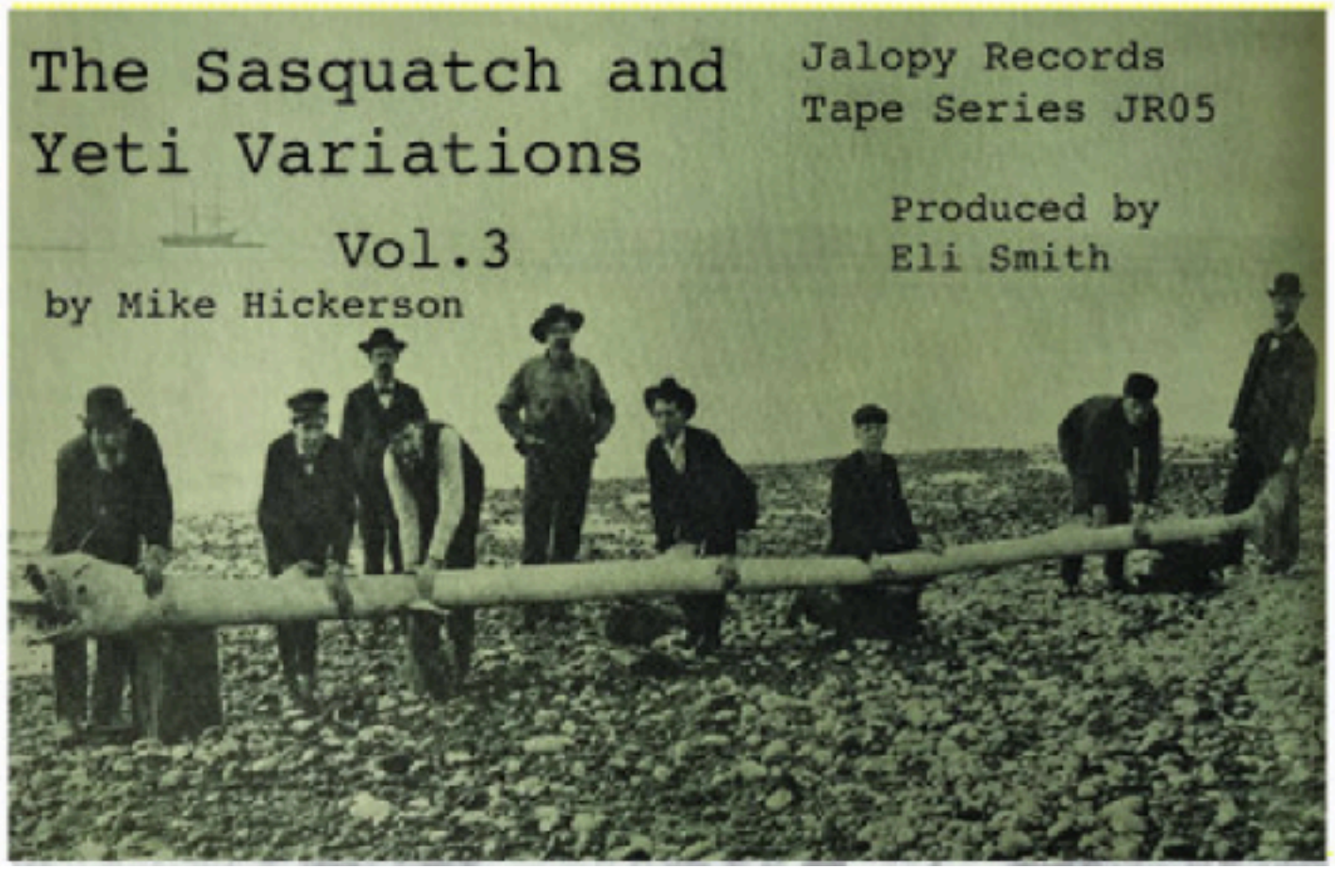
Follow

discography

Sasquatch and Yeti Variations Vol. 4
Apr 2019

Jalopy Records
315 Columbia St
Brooklyn, NY 11231

Sasquatch and Yeti Variations Vol. 3



Loch Ness DNA project may find invasive fish

7 June 2018

Share

THE LOCH NESS

THE LOCH NESS MONSTER?

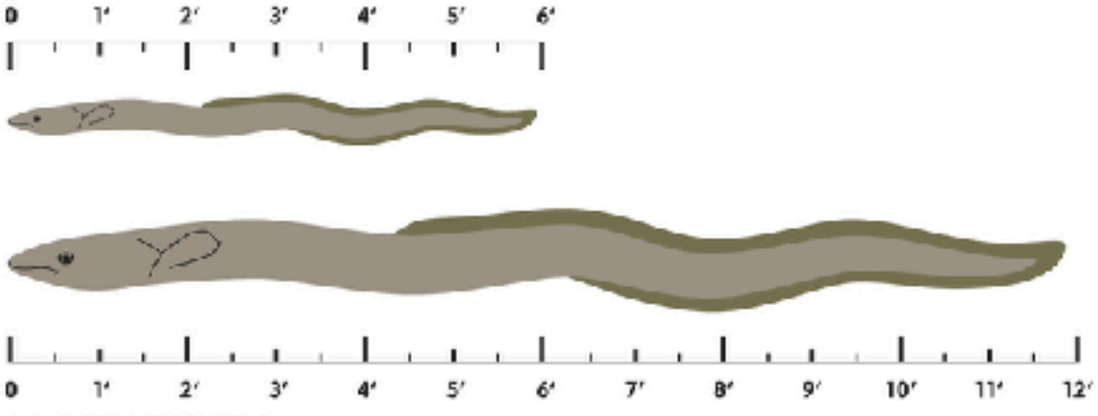
UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

Of the common theories associated with the 1,000 or so sightings of something swimming in the water at Loch Ness, the environmental DNA data obtained suggests at least one theory remains plausible.

Eels returned the largest proportion of DNA from the 250 water samples taken throughout Loch Ness.

Typically not gigantic, could an extremely large European eel be the creature people have seen moving "like a torpedo" in the water? The data obtained suggests this may be possible, although no eel of the size described in some accounts has ever been caught or found.

Infrequent visitors such as seals and possibly sturgeons may account for some sightings, but wakes, standing waves and logs are the basis of most.



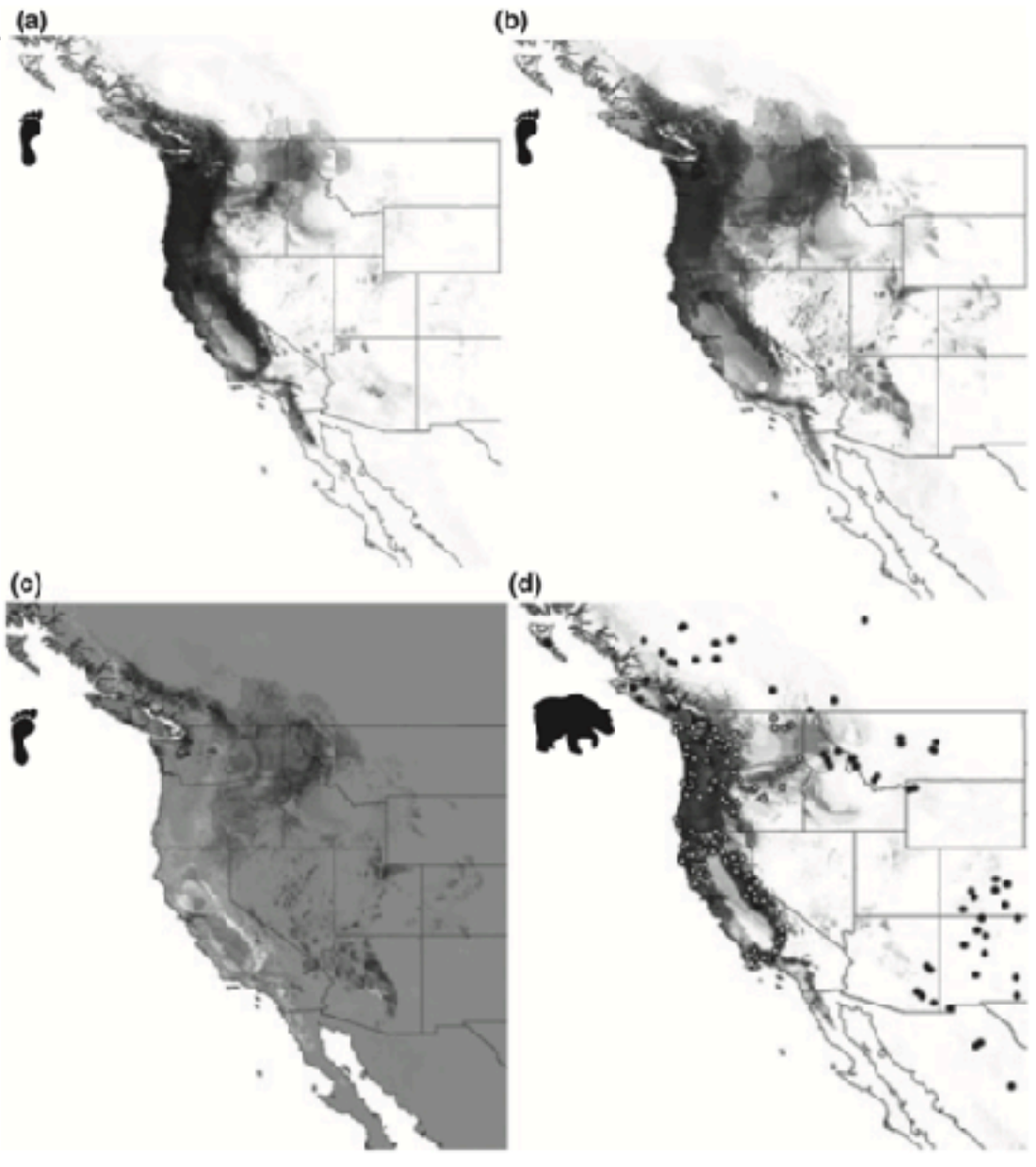
Journal of Biogeography (J. Biogeogr.) (2009) **36**, 1623–1627

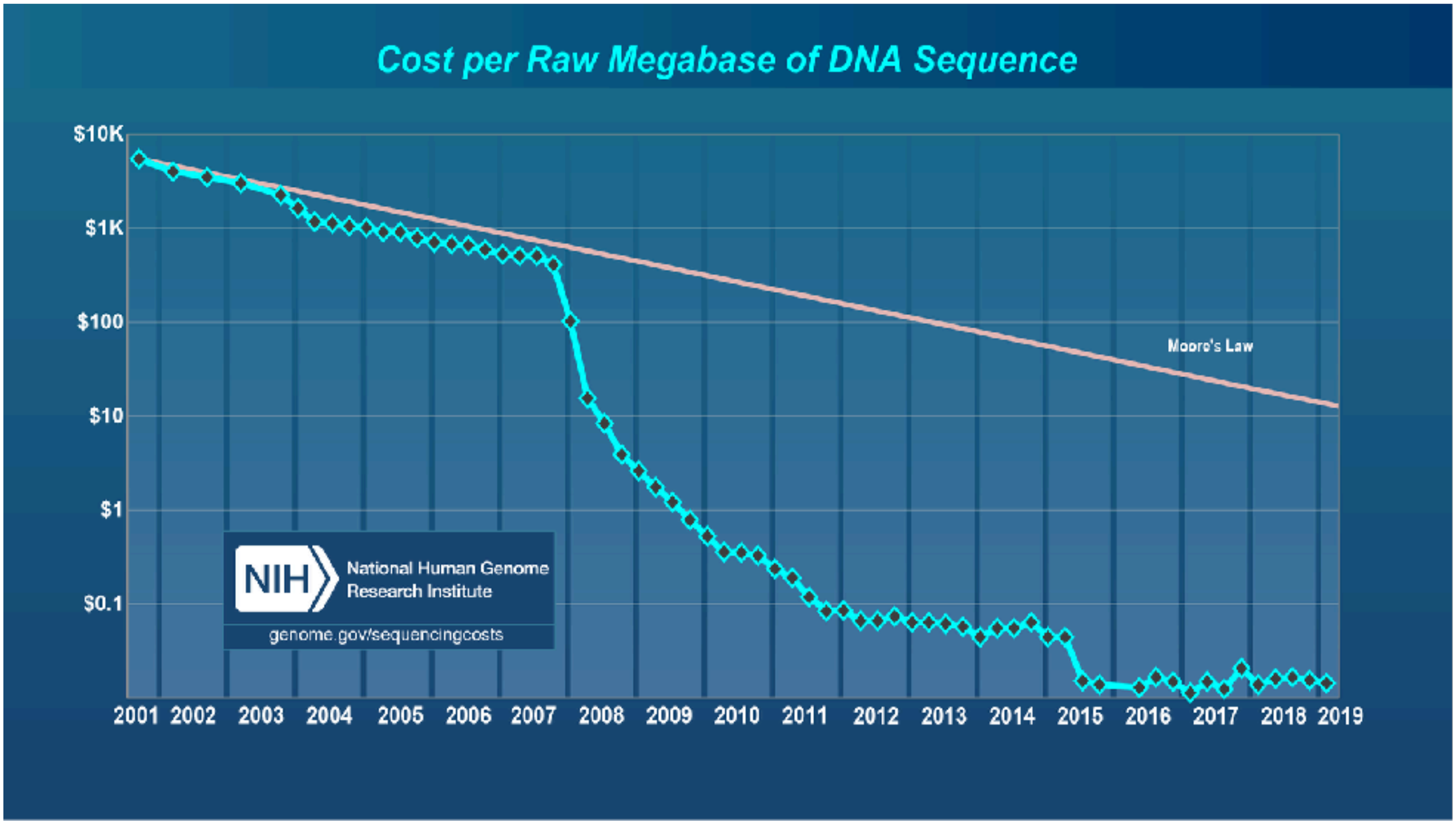
GUEST EDITORIAL

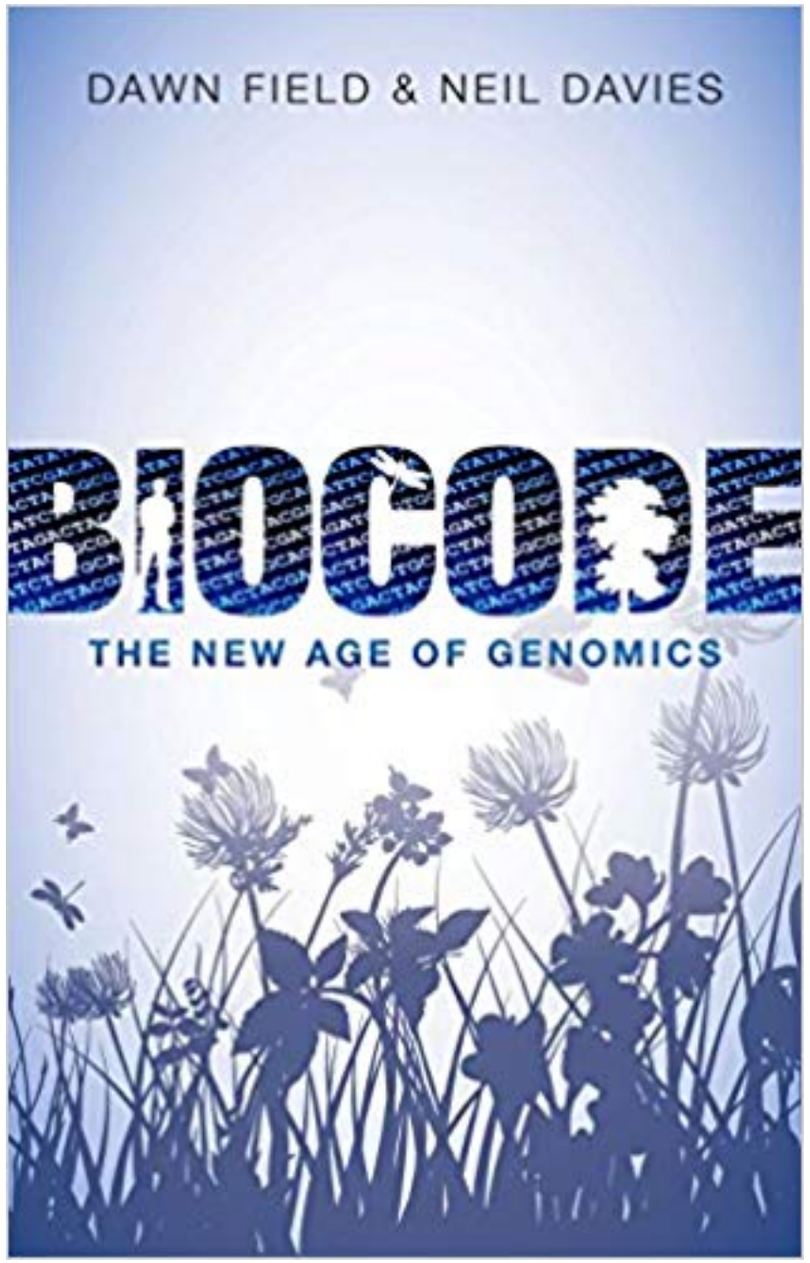
Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling

J. D. Lozier[1]*, P. Aniello[2] and M. J. Hickerson[3]

Figure 1 Map of Bigfoot encounters from Washington, Oregon and California used in the analyses. Points represent visual/audi-




Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019



A New Age of Gay Genomics Is Here. Are We Ready for the Consequences?

The scientists are presenting their findings with an abundance of caution. But that won't necessarily prevent anti-queer abuse.

By JEREMY YODER

AUG 29, 2019 • 7:21 PM

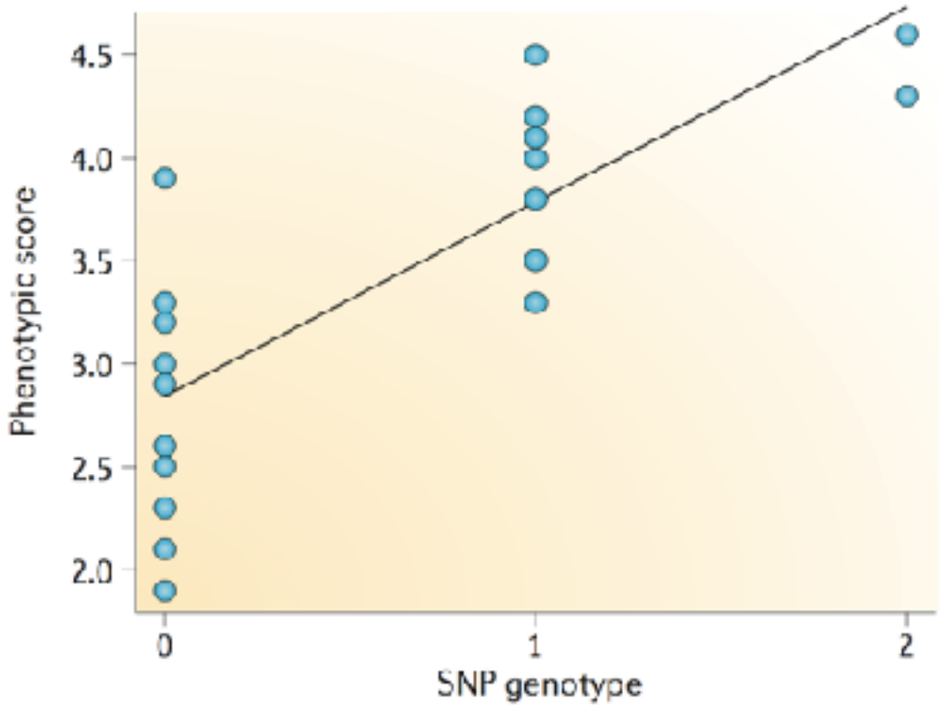
aside from reconstructing complex human history

applications: Genome-wide association mapping (medical genomics)

Example: Discrete trait

	Case	Control
AA	563	518
AG	375	403
GG	62	79

Example: Quantitative trait



- The search for genetic variation underlying a complex trait:

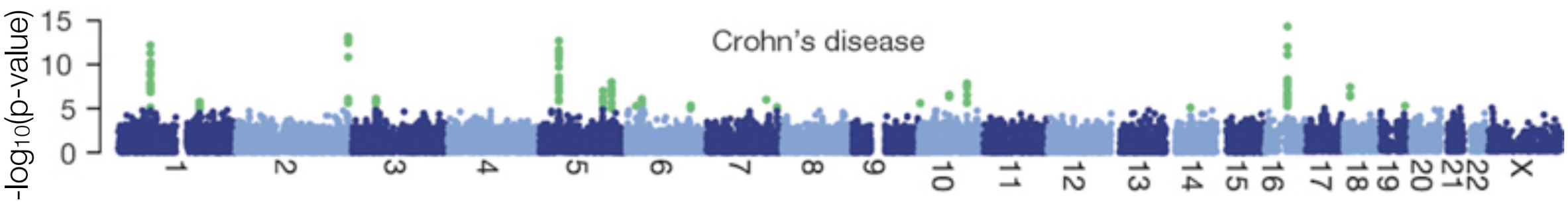
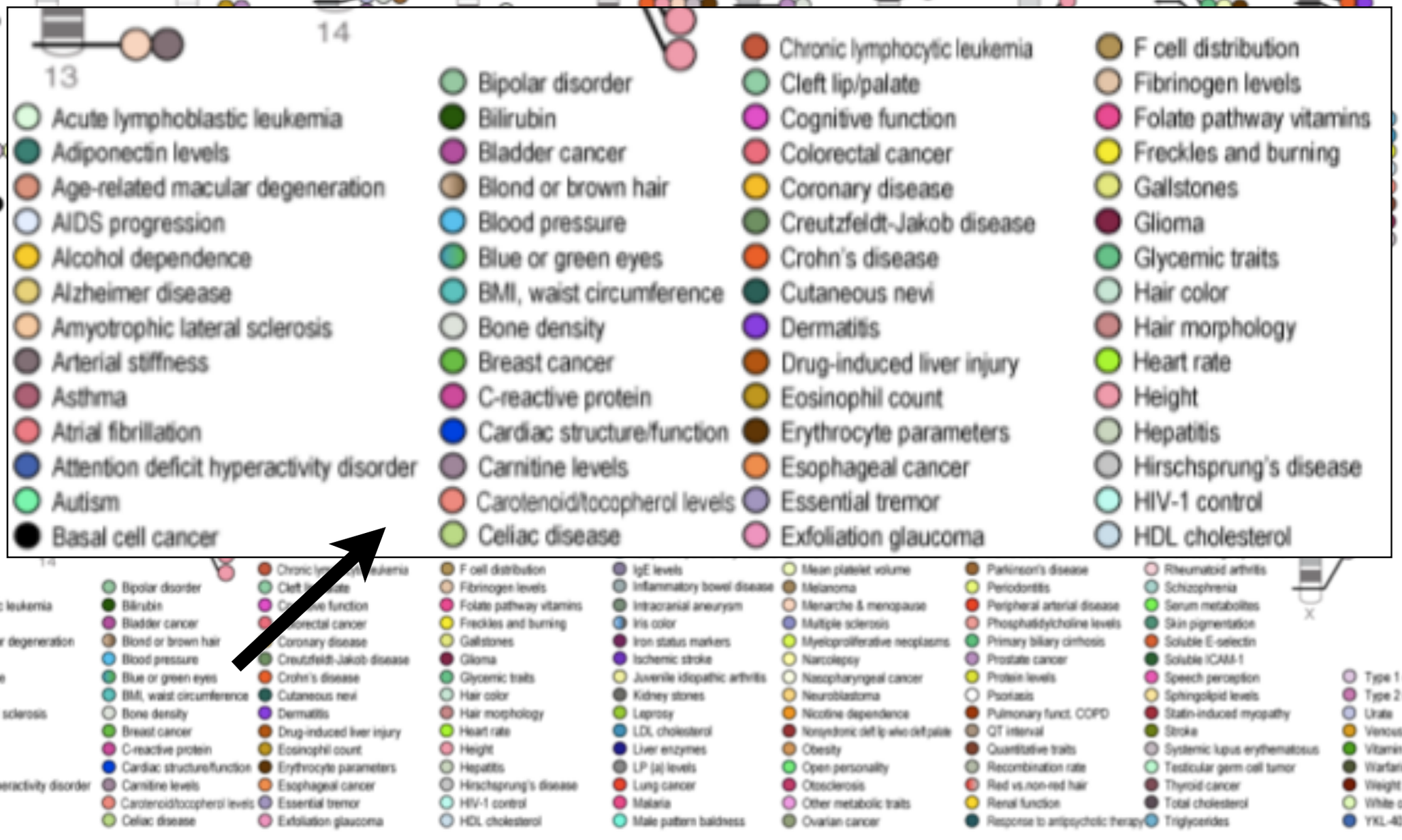


Figure from WTCCC

Published Genome-Wide Associations through 12/2009, 658 published GWA at p≤5x10^{-8}

NHGRI GWA Catalog
www.genome.gov/GWAStudies



Published Genome-Wide Associations through 12/2009, 658 published GWA at $p \leq 5 \times 10^{-8}$

NHGRI GWA Catalog

Studies

Acute lymphoblastic leukemia
Adiponectin levels
Age-related macular degenera
AIDS progression
Alcohol dependence
Alzheimer disease
Amyotrophic lateral sclerosis
Arterial stiffness
Asthma
Atrial fibrillation
Attention deficit hyperactivity disorder
Autism
Basal cell cancer
Bone density
Breast cancer
C-reactive protein
Cardiac structure/function
Carnitine levels
Carotenoid/tocopherol levels
Celiac disease
Dermatitis
Drug-induced liver injury
Eosinophil count
Erythrocyte parameters
Esophageal cancer
Essential tremor
Exfoliation glaucoma
Hair morphology
Heart rate
Height
Hepatitis
Hirschsprung's disease
HIV-1 control
HDL cholesterol
Leprosy
LDL cholesterol
Liver enzymes
LP (a) levels
Lung cancer
Malaria
Male pattern baldness
Nicotine dependence
Nonsyndromic def lp who cleft palate
Obesity
Open personality
Otosclerosis
Other metabolic traits
Ovarian cancer
Pulmonary funct. COPD
QT interval
Quantitative traits
Recombination rate
Red vs.non-red hair
Renal function
Response to antipsychotic therapy
Statin-induced myopathy
Stroke
Systemic lupus erythematosus
Testicular germ cell tumor
Thyroid cancer
Total cholesterol
Triglycerides
Urate
Venous thromboembolism
Vitamin B12 levels
Warfarin dose
Weight
White cell count
YKL-40 levels

Applications: Detecting regions of the genome that recently undergone adaptive evolution

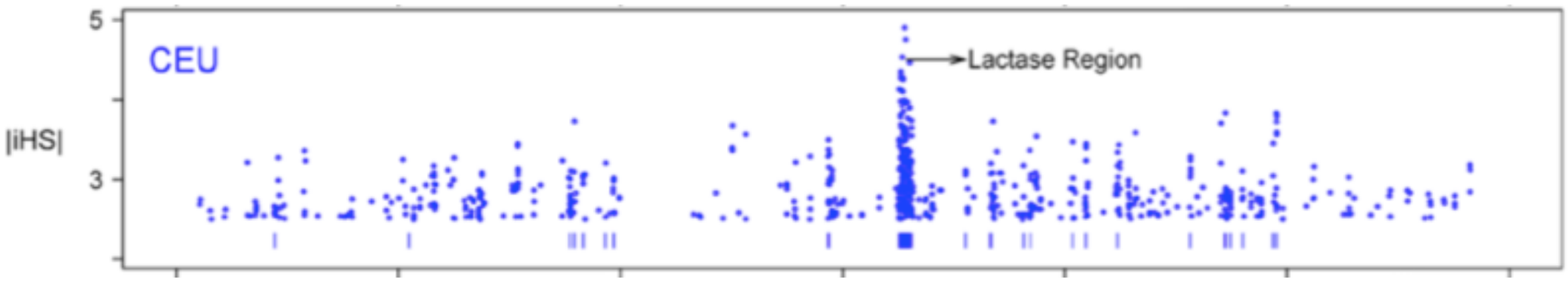


Figure 3. Plots of Chromosome 2 SNPs with Extreme iHS Values Indicate Discrete Clusters of Signals

SNPs with |iHS| >2.5 (top 1%) are plotted. The bottom plot combines signals for all three populations, plotting only SNPs with derived frequency >0.5 and iHS <−2.5. Such SNPs correspond to high-frequency-derived SNPs in the range for which our test is most powerful. The short vertical bars below each plot indicate 100-kb windows whose signals are in the top 1% of windows genome-wide.

DOI: 10.1371/journal.pbio.0040072.g003

integrated Haplotype Score (iHS): statistic sensitive to increase of linkage disequilibrium during a selective sweep

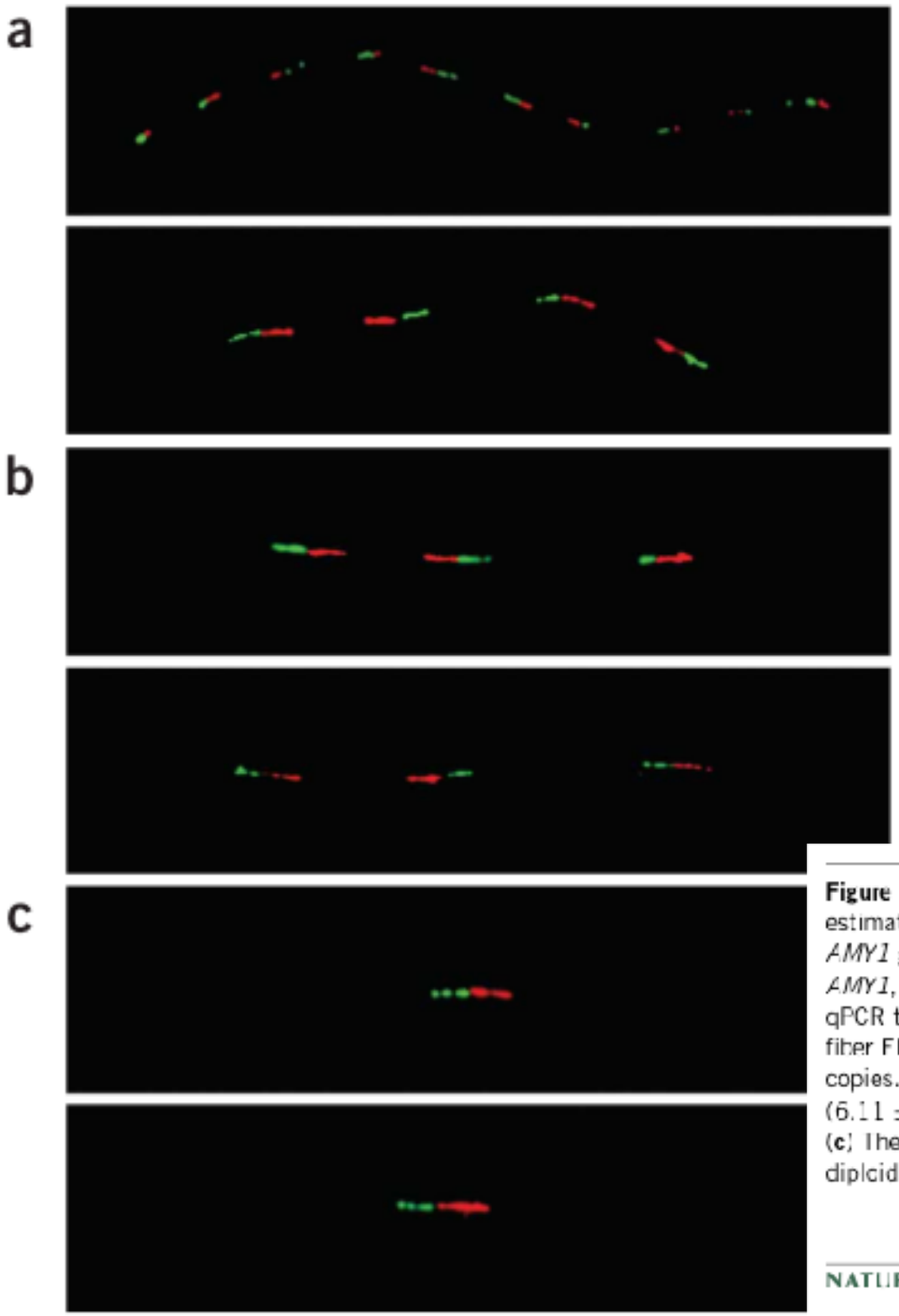
Voigt et al (2005) PLoS Biology

Salivary amylase copy number variation

Japanese individual
10 + 4 = 14 copies

Biaka individual
3 + 3 = 6 copies

1 + 1 = 2 copies
Chimpanzee



AMY1 copy number variation as seen by fluorescent in-situ hybridization (FISH)

Copy number for AMY1 is highly variable among human populations

Figure 3 High-resolution fiber FISH validation of *AMY1* copy number estimates. Red (~10 kb) and green (~8 kb) probes encompass the entire *AMY1* gene and a retrotransposon directly upstream of (and unique to) *AMY1*, respectively. (**a**) Japanese individual GM18972 was estimated by qPCR to have 14 (13.73 ± 0.93) diploid *AMY1* gene copies, consistent with fiber FISH results showing one allele with ten copies and the other with four copies. (**b**) Biaka individual GM10472 was estimated by qPCR to have six (6.11 ± 0.17) diploid *AMY1* gene copies, consistent with fiber FISH results. (**c**) The reference chimpanzee (Clint; SDC6006) was confirmed to have two diploid *AMY1* gene copies.

Applications: Personalized Genomics

23andMe

deCODEme

deCODE your ancestry

learn about your ancestral story

Ancestry is one of the most exciting areas of genetic exploration. The use of genetic data to make inferences about people's ancestors and genealogical relationships allow for some truly remarkable discoveries. Scientific research in this field has revealed that we all descend from a common ancestral group of humans that originated in Africa about 200 thousand years ago. About 70 thousand years ago, the descendants of these first humans began colonizing other parts of the world. Who were your ancestors? In what part of the human story did they play a role?

The ancestry results

included in our complete scan:

- map of kinship
- genetic atlas
- ancestral origins
- your male line y-group
- your female line mitogroup

OUR COMPLETE SCAN

Diet and Variation in salivary amylase copy number

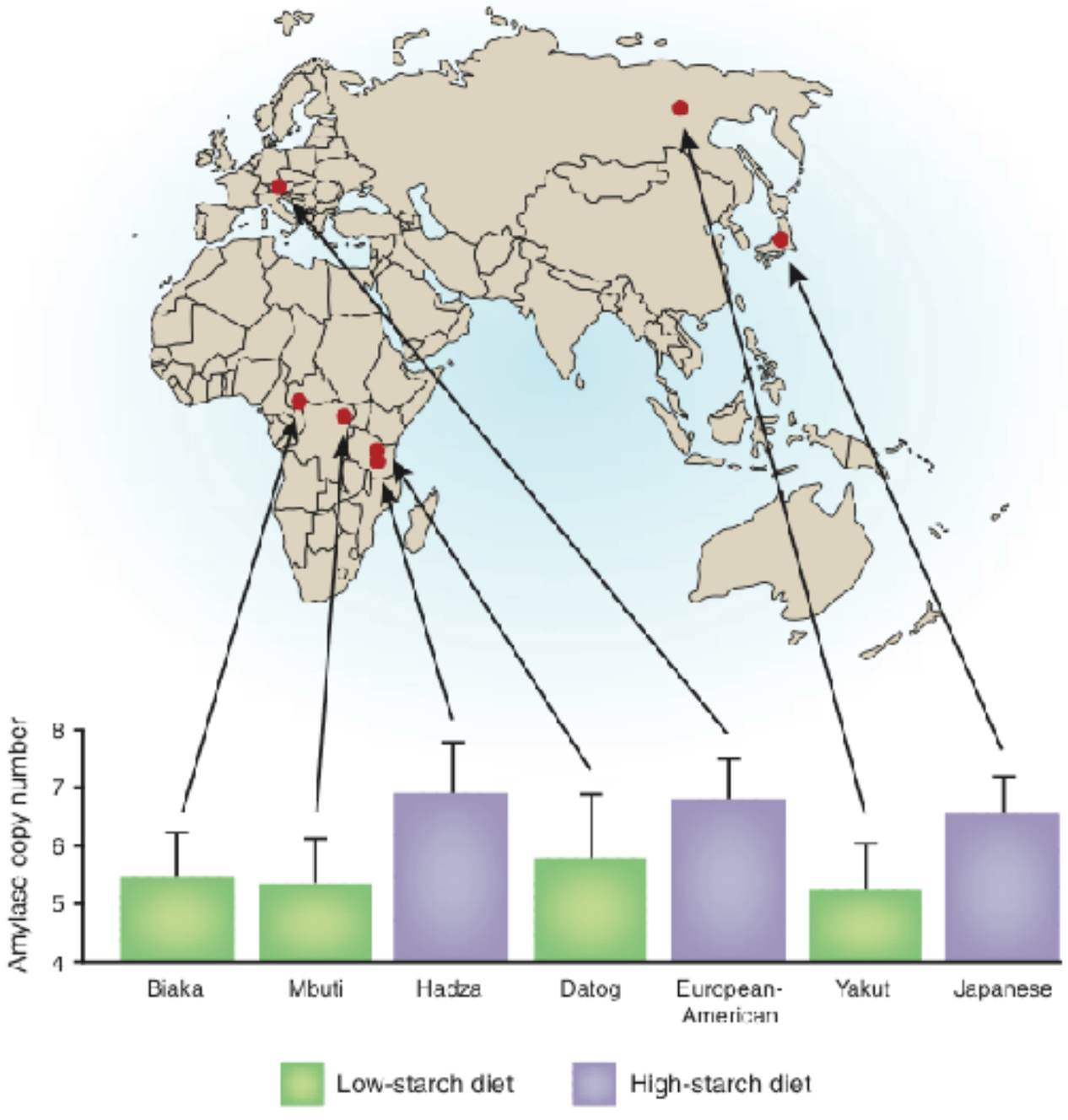
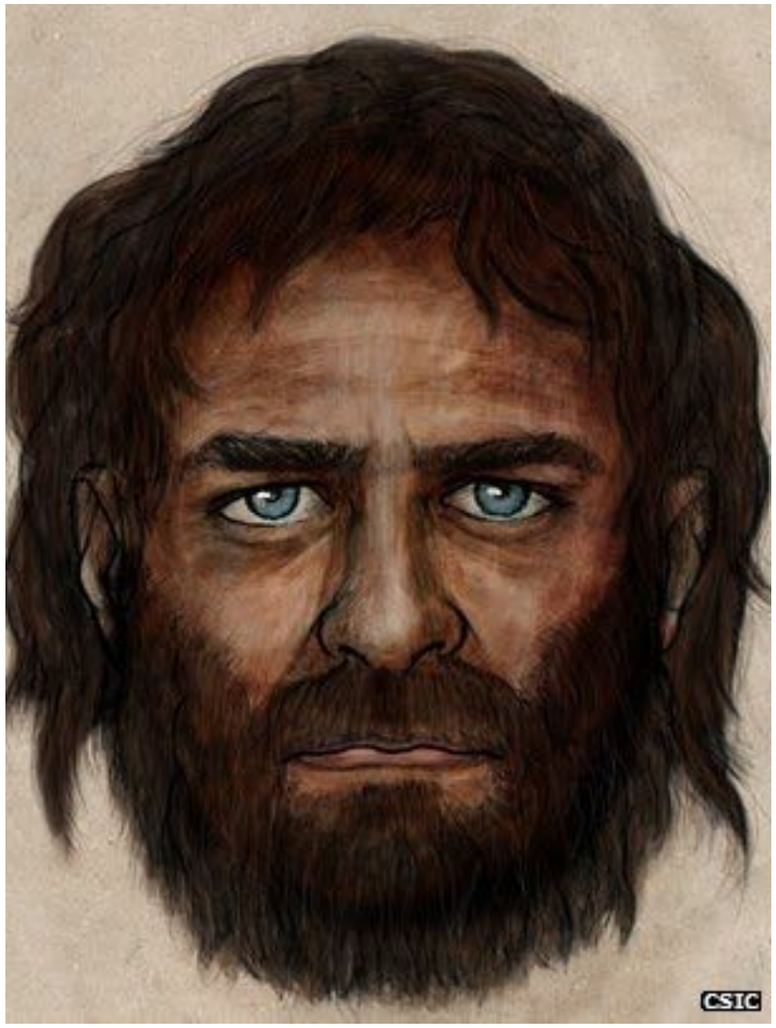


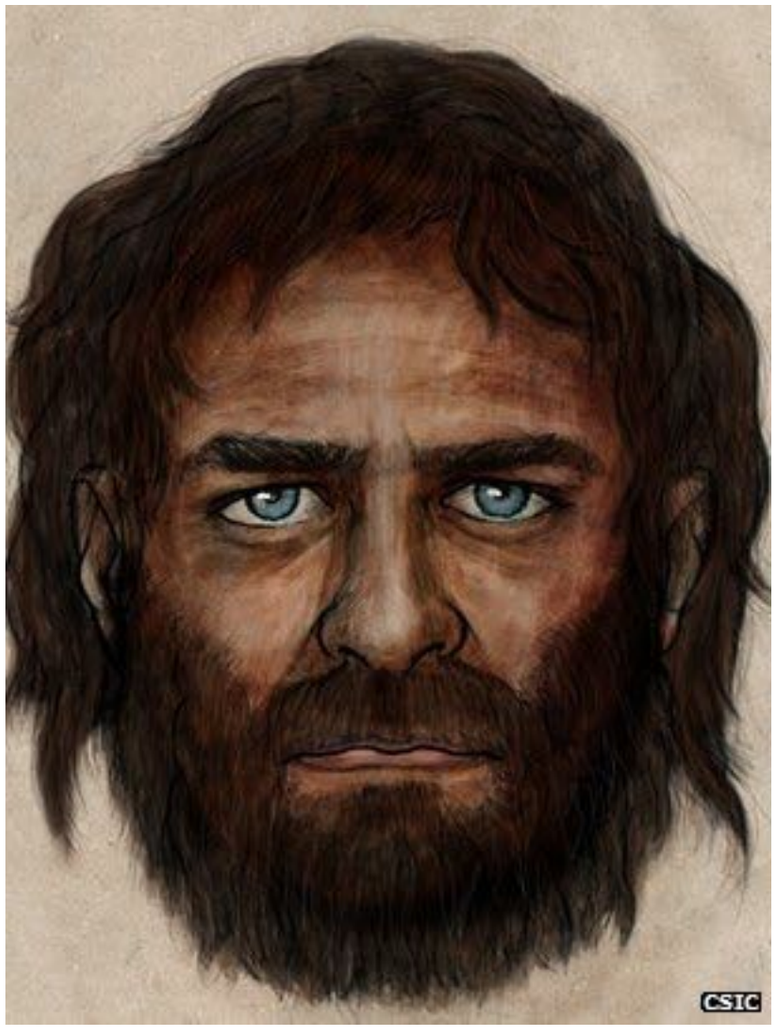
Figure 1 The distribution of salivary amylase copy number in the seven samples from Perry *et al.*[1] The bar chart depicts the mean copy number per sample, with an interval of two standard errors above the mean. Mean copy number is found to be higher in populations with high-starch diets, even when samples are relatively near one another geographically (for example, comparing Hadza and Datog or Yakut and Japanese populations).

High-starch diet populations have higher average copy number for AMY1 than low-starch diet populations



whole genome from
hunter from 8,000 years ago
(Spain)

digests milk and wheat?



whole genome from
hunter from 8,000 years ago
(Spain)

digests milk and wheat?
no, no (and has dark skin with blue eyes)

Conclusions

- Population genetics is an engaging field with numerous research challenges.
  - Exciting growth period: Increasingly large-scale data-sets are leading to an era of rapid discovery (including non-model species)
  - Detailed mathematical theory: relies heavily on probability theory and statistics
  - Relevance: Results have impacts ranging from understanding population history and trait evolution to conservation to understanding the genetic basis of major evolutionary transitions
-

Samples, populations, and species

What do we think about when we think about a population?

Samples, populations, and species

What do we think about when we think about a population?

Individuals within the population mate randomly with each other and never mate with anyone outside the population

Samples, populations, and species

What do we think about when we think about a population?

Individuals within the population mate randomly with each other and never mate with anyone outside the population

Failure of real populations to meet this ideal standard is called ***population structure***

Samples, populations, and species

What do we think about when we think about a population?

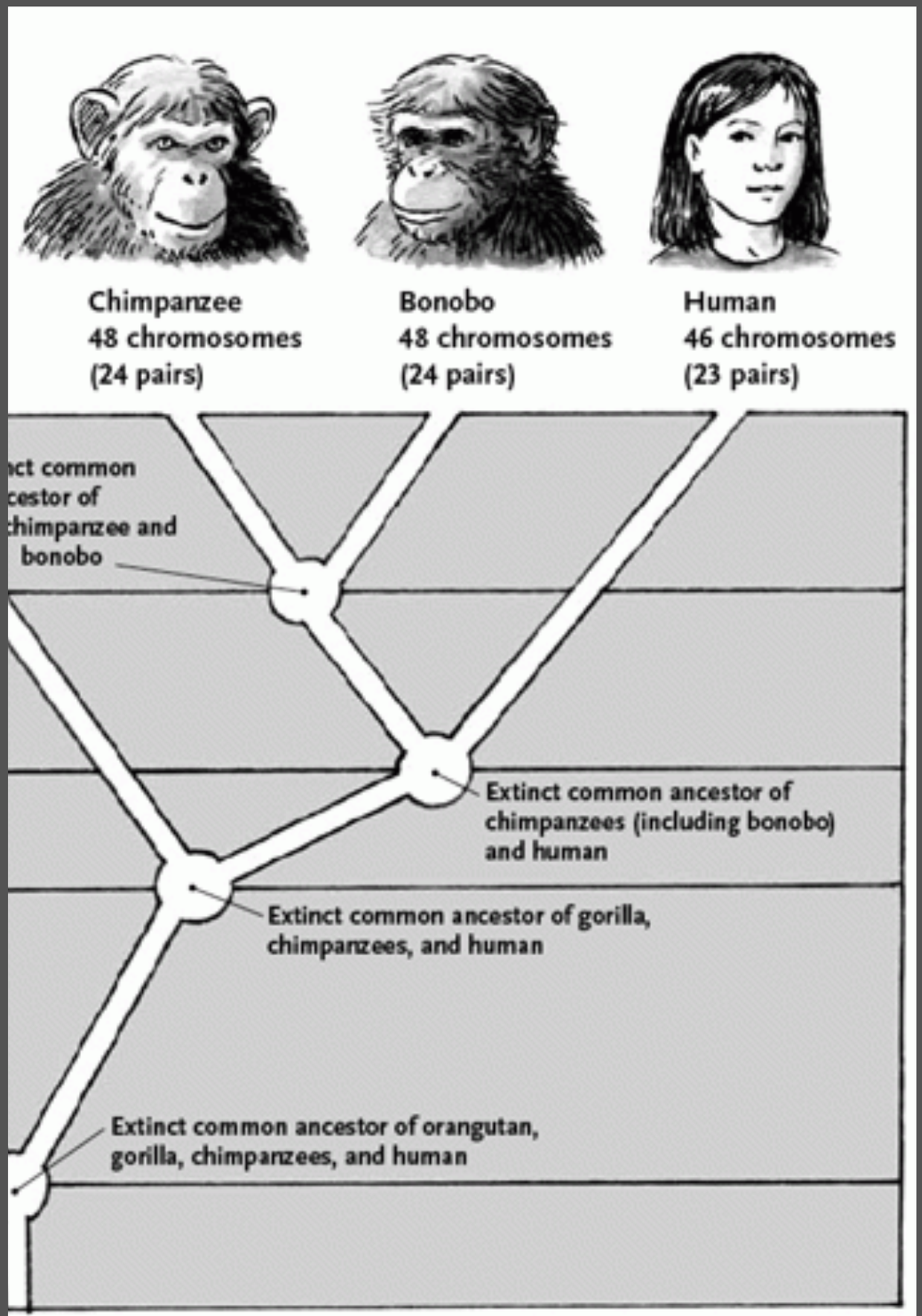
Individuals within the population mate randomly with each other and never mate with anyone outside the population

Failure of real populations to meet this ideal standard is called ***population structure***

We assume that individuals don’t mate with anyone outside their species, but we don’t assume random mating within a species

What about humans and Chimps?

Humans have one pair fewer chromosomes than other apes, with ape chromosomes 2 and 4 fused in the human genome into a large chromosome

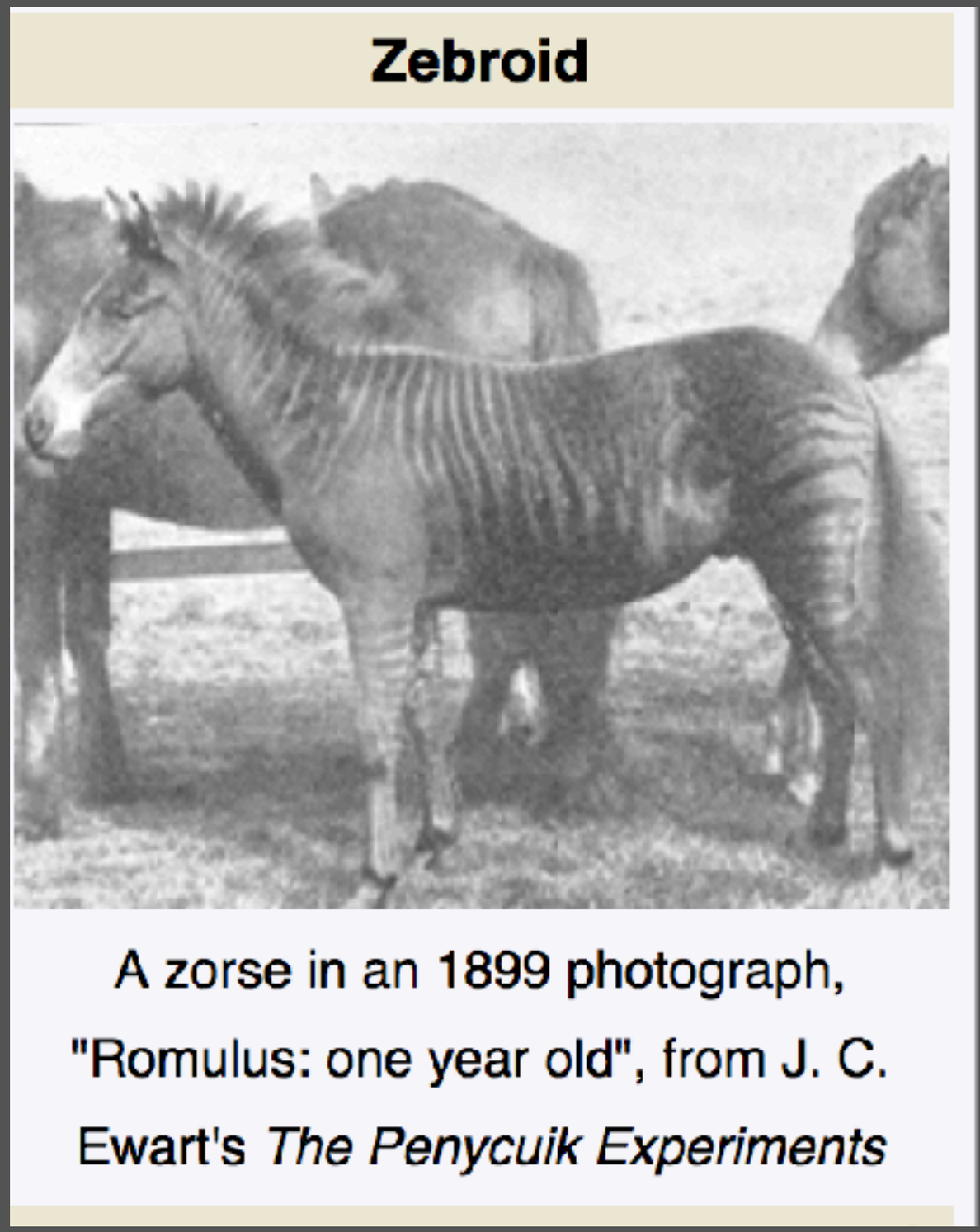


What about humans and Chimps?

similar mismatches are relatively common within existing species

Humans and chimps match at most other chromosomes (about same level as within equines that can all hybridize easily)

Similar complexities pertain to horse–zebra hybrids, whose have chromosomal disparity(horses =32 chromosome pairs & zebras = 16 and 23



A zorse in an 1899 photograph, "Romulus: one year old", from J. C. Ewart's *The Penycuik Experiments*

it has been found that human sperm binds to gorilla oocytes with almost the same ease as to human ones.

MOLECULAR REPRODUCTION AND DEVELOPMENT 31:264–267 (1992)

Hemizona Assay for Measuring Zona Binding in the Lowland Gorilla

S.E. LANZENDORF, W.J. HOLMGREN, D.E. JOHNSON, M.J. SCOBEY, AND R.S. JEYENDRAN
Section of Reproductive Endocrinology and Infertility, Department of Obstetrics and Gynecology, Northwestern University Medical School, Prentice Women's Hospital, Chicago, Illinois

This study found that gorilla sperm bound well to both gorilla and human hemizonae, with a mean of 112.5 and 81.0 tightly bound sperm, respectively. Human sperm also bound to gorilla (mean 229.5) and human (mean 236.5) hemizonae. Following incubation with intact gorilla zonae, motile human sperm were found within the perivitelline space. However, gorilla sperm were not visible within the perivitelline space of nonviable human oocytes. These

FELID HYBRIDS

This shows how many millions of years ago the species diverged from their common ancestors

DOMESTIC CAT
European wild Cat
African Wild Cat
Chinese Desert Cat
Sand Cat
Black-Footed Cat
Jungle Cat
Leyhausen's Hybrids
Chausie
Bengal
Viverral/Machbagral
Domestic x Ocelot
Bristol & Bengal Breeds
Safari
Caracat
Falkena-Rohrle's Hybrids
Savannah

Pallas Cat (Manul)
Rusty-Spotted Cat
Asian Leopard Cat
Fishing Cat
Flat-headed Cat
Viverral/Machbagral

Puma
Jaguarundi
Cheetah

Blynx
Canadian Lynx
Eurasian Lynx
Iberian Lynx
Bobcat

Ocelot
Margay
Andean Mountain Cat
Pampas Cat
Geoffroy's Cat
Kodkod
Oncilla
Ocelot
Margay
Ocelbob
Marlot
Pumapard
Domestic x Ocelot
Bristol & Bengal Breeds
Safari
Falkena-Rohrle's Hybrids
Caracat
Savannah

Caracal
African Golden Cat
Serval
Servical

Bay Cat
Asian Golden Cat
Marbled Cat

Leopard
Lion
Jaguar
Tiger
Snow Leopard
Liger, Tigon, Lepjag, Leopon etc

Clouded Leopard

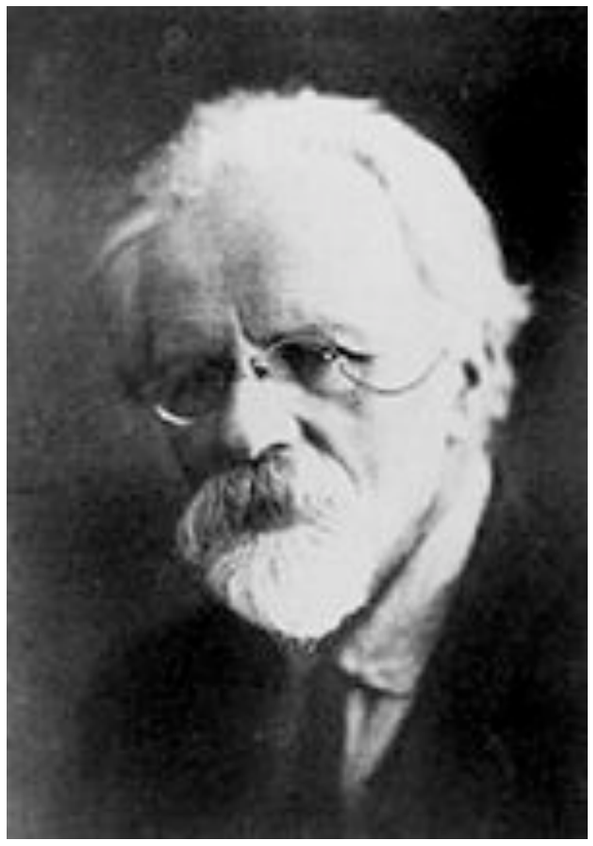
10 5 0
Millions of Years Ago (MYA) Present Day

Authenticated Hybrid
Anecdotal Only

Messybeast.com

What about humans and Chimps?

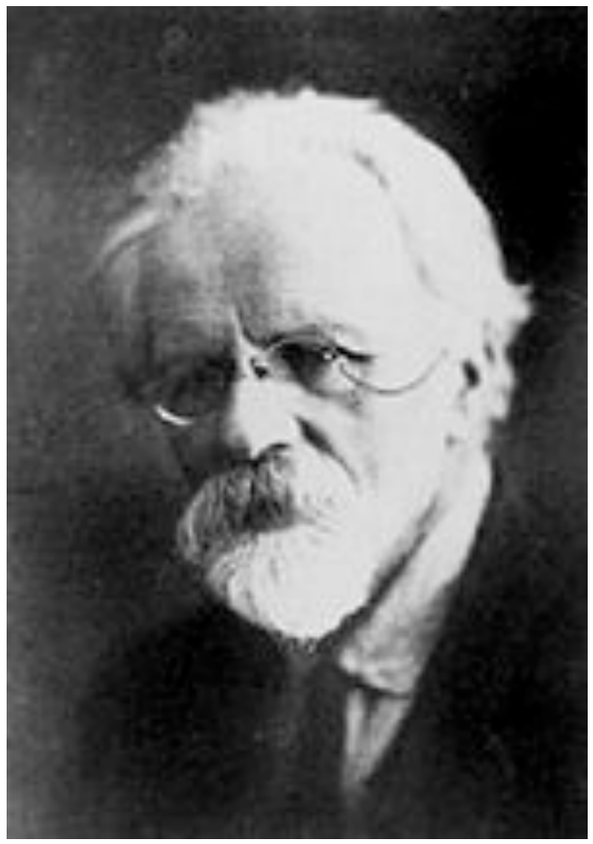
attempted to breed a "humanzee"
1927-29



Ilya Ivanov, 1927

In the 1920s, Ivanov carried out a series of experiments (in Africa), culminating in inseminating three female chimpanzees with human sperm, but he failed to achieve a pregnancy

What about humans and Chimps?



Ilya Ivanov, 1927

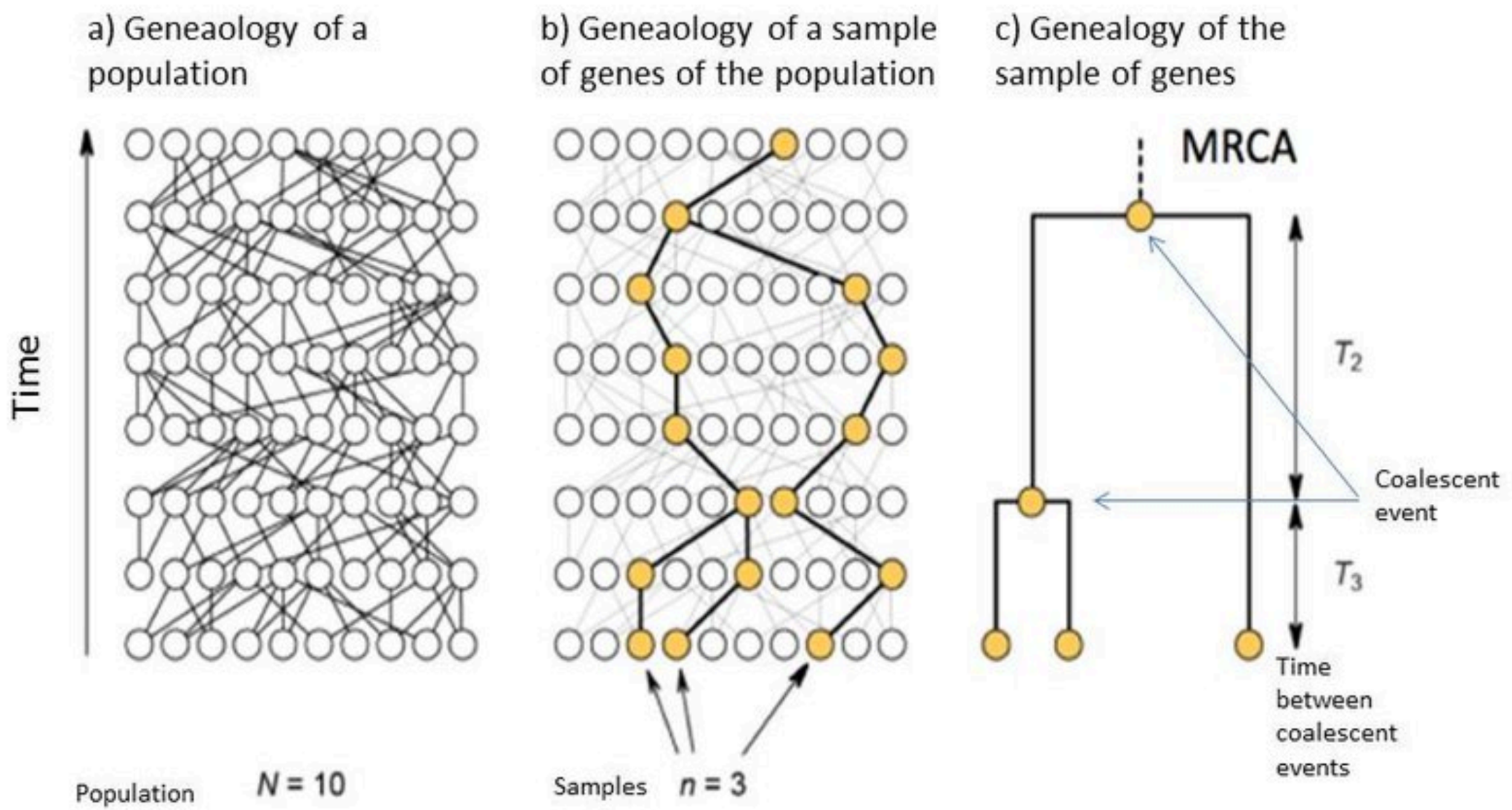
1929 - experiments involving nonhuman ape sperm and human volunteers,

delayed by the death of his last orangutan.

The next year he fell under political criticism from the Soviet government and was sentenced to exile in the Kazakh SSR;

Samples, populations, and species

It's usually impossible to collect data from an entire population, so the individuals we have sequenced/genotyped are called the sample



A statistical interlude

A frequency in a sample is just an estimate of the frequency in the population the sample came from

A statistical interlude

A frequency in a sample is just an estimate of the frequency in the population the sample came from

If you catch 5 frogs, genotype them, and calculate an allele frequency of 3/10 at a particular locus, 3/10 is just your estimate of the frog population allele frequency

A statistical interlude

A frequency in a sample is just an estimate of the frequency in the population the sample came from

If you catch 5 frogs, genotype them, and calculate an allele frequency of 3/10 at a particular locus, 3/10 is just your estimate of the frog population allele frequency

The more diverse and substructured the population is, the worse your estimate is likely to be

The Neutral Theory of Molecular Evolution

Motoo Kimura postulated in 1983 that most change at the phenotypic level is driven by selection but most change at the molecular level is caused by drift

If drift is dominant, very few features would come back if the tape were played twice

ie not deterministic

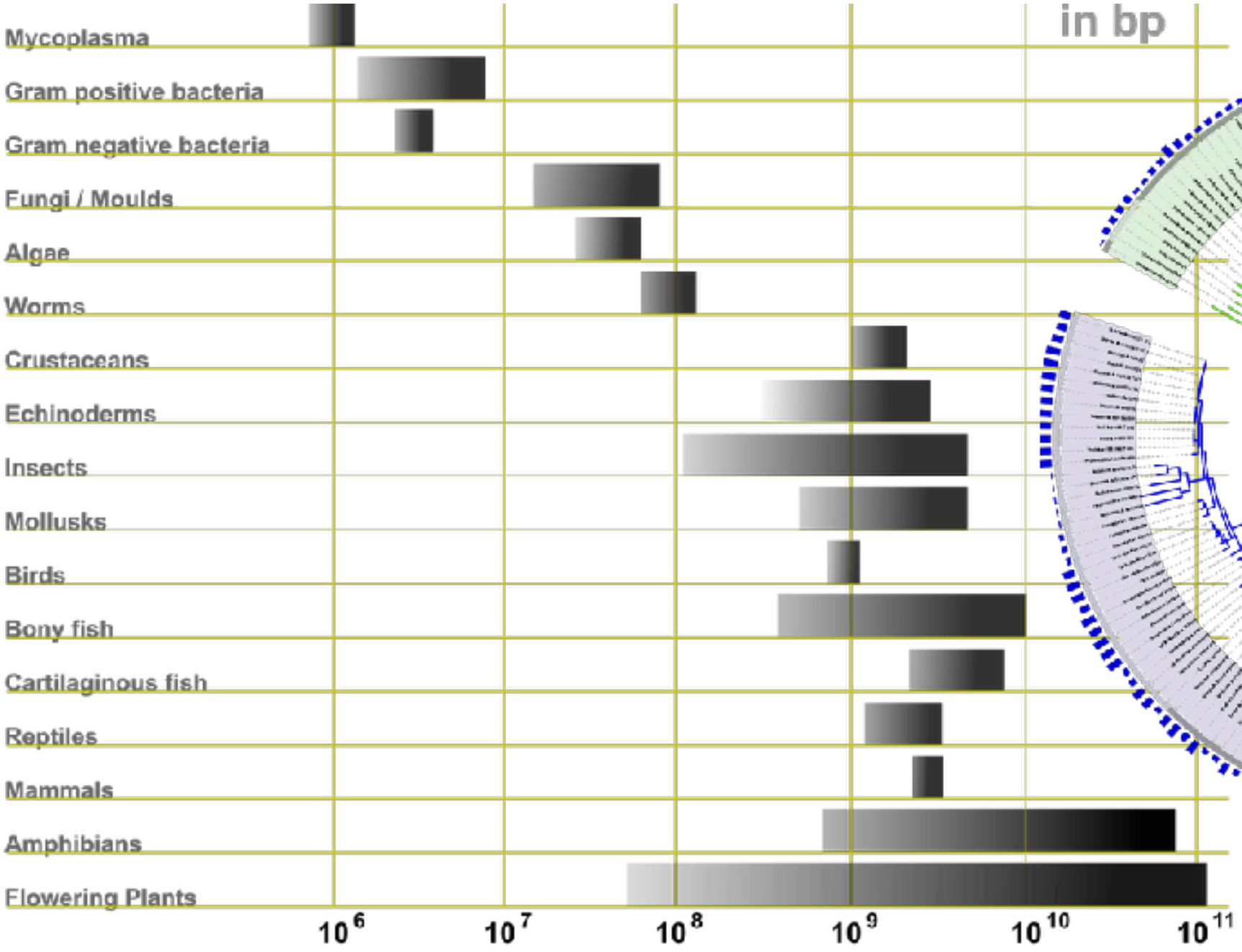
The Neutral Theory of Molecular Evolution

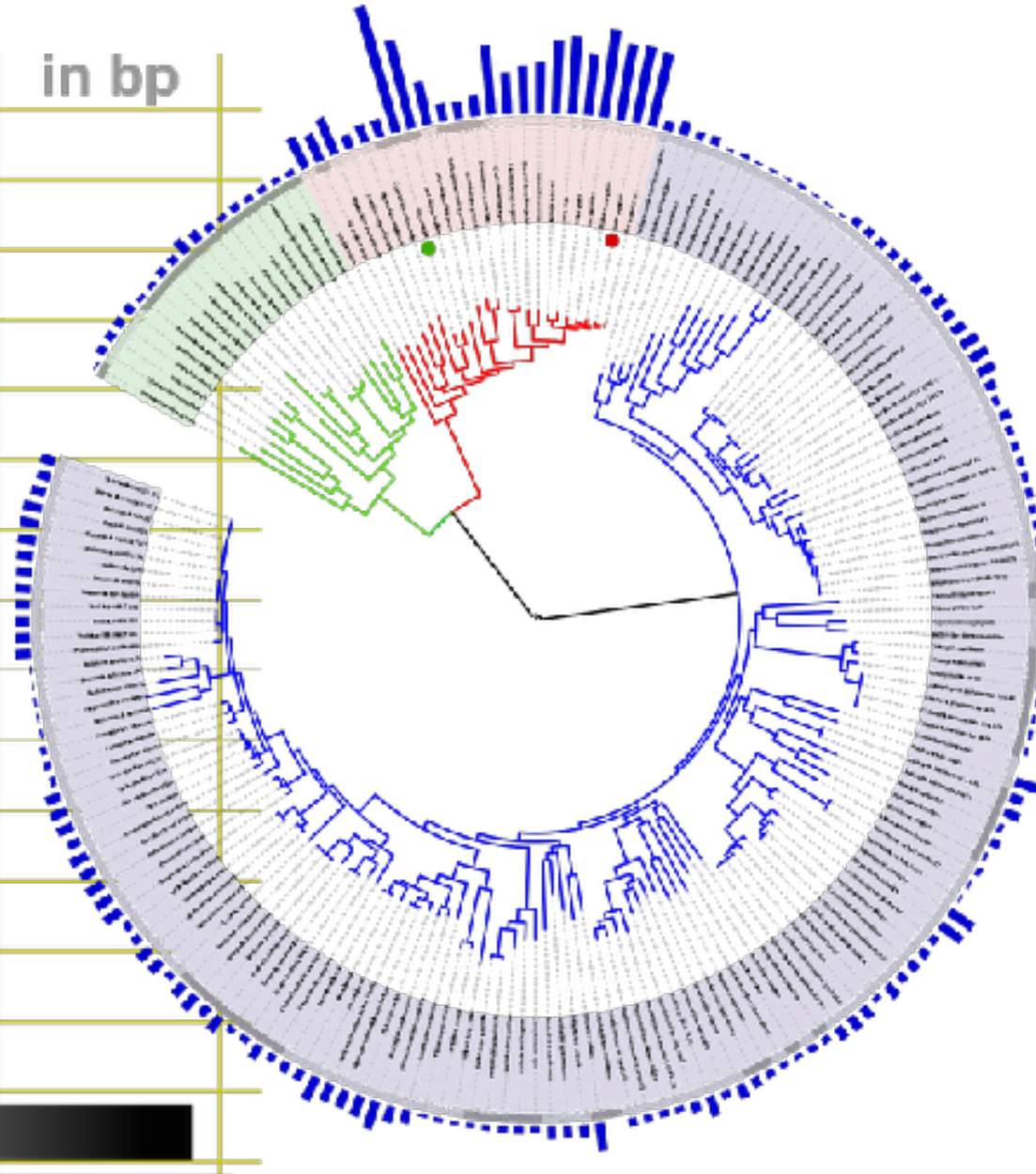
Motoo Kimura postulated in 1983 that most change at the phenotypic level is driven by selection but most change at the molecular level is caused by drift

If drift is dominant, very few features would come back if the tape were played twice

The Neutral Theory of Molecular Evolution

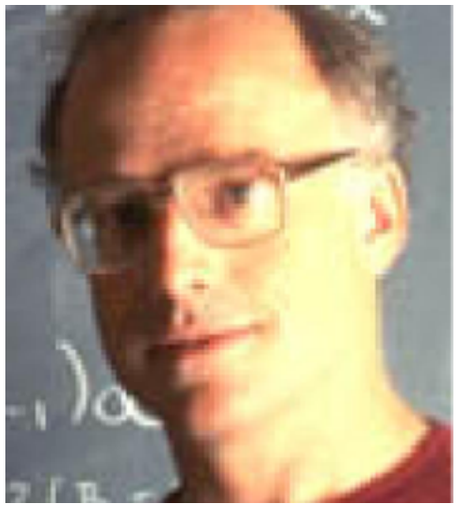
The neutral theory is clearly more applicable to large vertebrate genomes than to compact microbial genomes





The Neutral Theory of Molecular Evolution

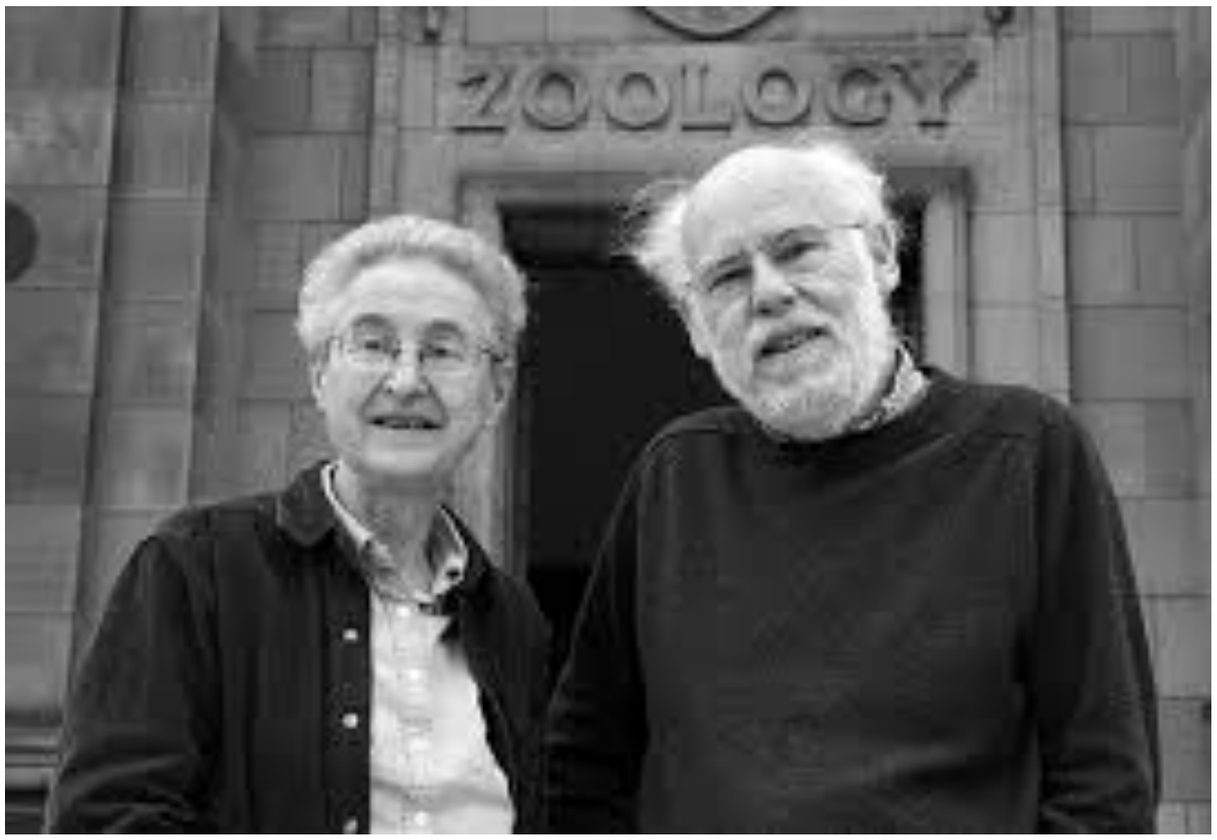
Motoo Kimura



John Gillespie

“Neutralists” and “Selectionists” still debate the exact percentage of change that is driven by selection

The Neutral Theory of Molecular Evolution



“Neutralists” and “Selectionists” still debate the exact percentage of change that is driven by selection

Formal models of neutral evolution

If we build good mathematical models of evolution under neutrality vs selection, we can predict how these types of variants should look different in a snapshot of genetic variation

Formal models of neutral evolution

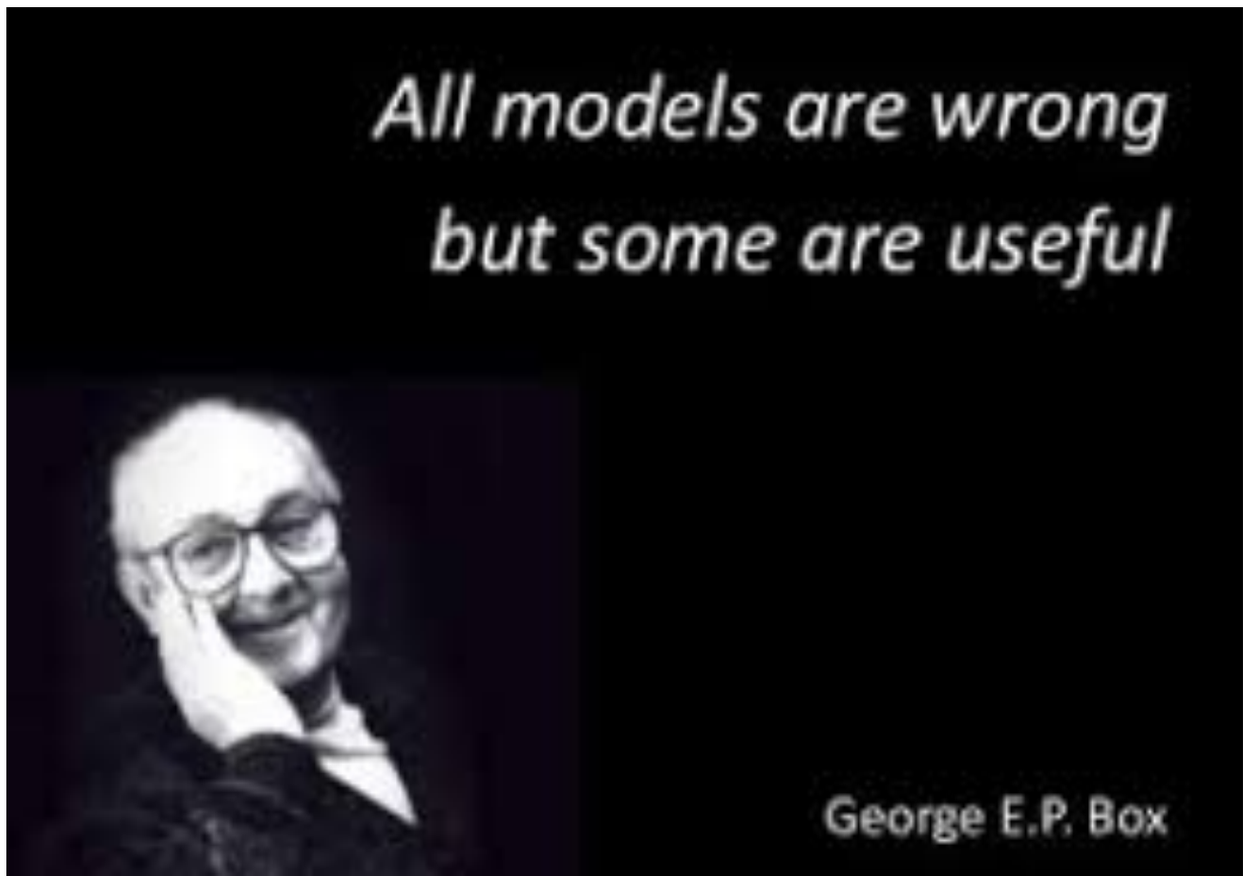
Two very similar classic models: The Wright-Fisher model and the Moran model

Models allow us to calculate how well a given hypothesis fits the data

- Nature is complex, and models are simplified versions of Nature containing essential features or variables we need to understand (useful approximations)

- Models allow prediction and/or estimation,

- Make models “as simple as possible, but no simpler” - attributed to A. Einstein

- 

Moran model of genetic drift to loss/ fixation

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

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 of genetic drift to loss/ fixation

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 $1/N$

Moran model of genetic drift to loss/ fixation

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 $1/N$

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

Wright-Fisher model

Population constant size of N diploid individuals (hermaphrodite)
= $2N$ chromosomes

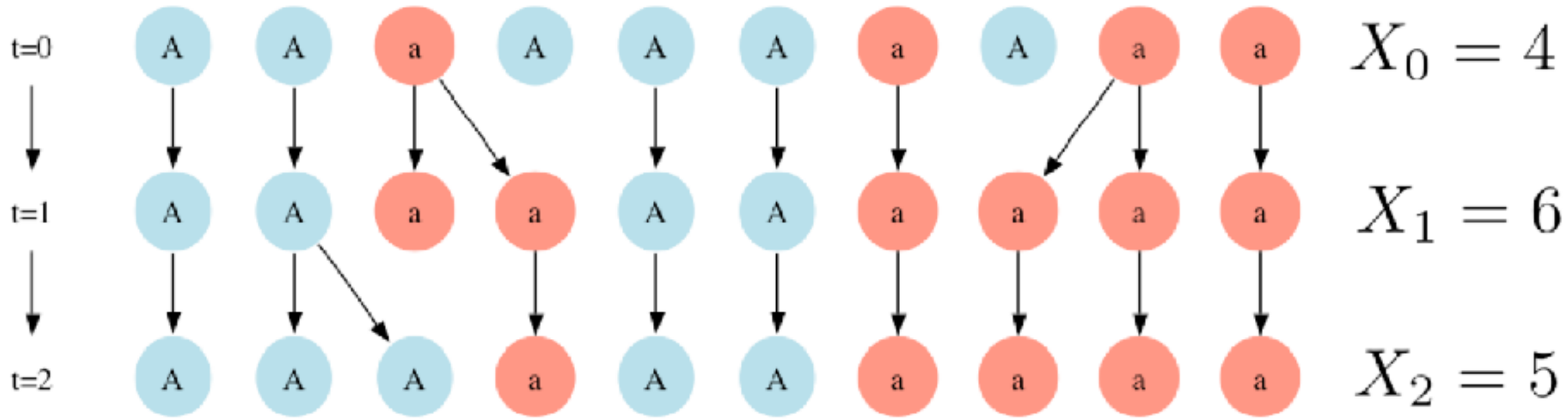
individuals mate randomly per generation

chromosomes sampled uniformly with replacement to procreate

no individuals survive per generation (ie non-overlapping generations)

Wright-Fisher model

chromosomes sampled uniformly with replacement to procreate

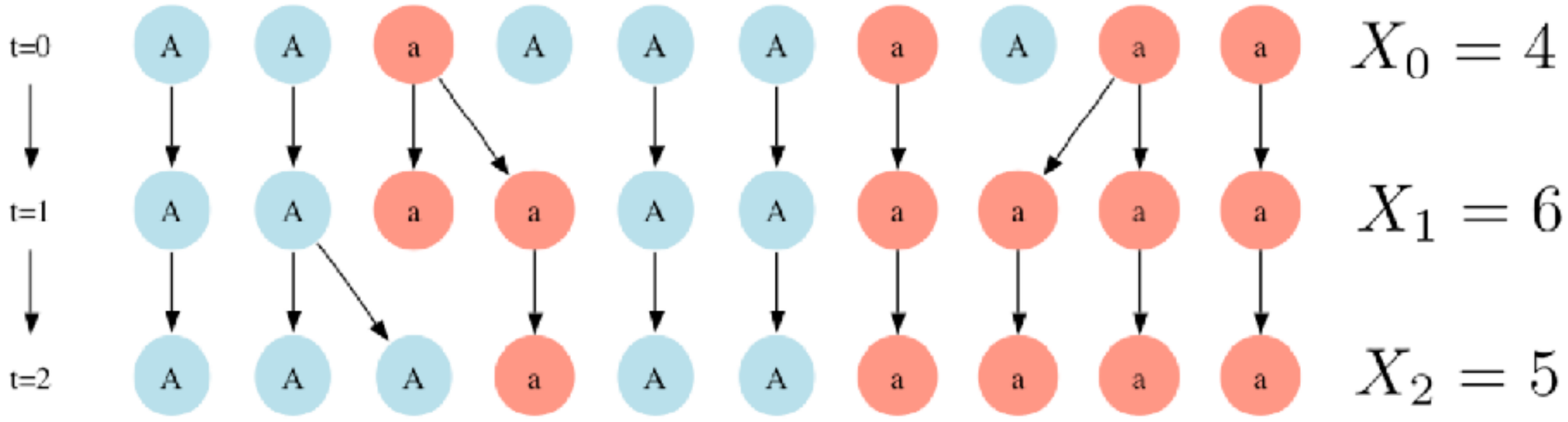


no individuals survive per generation (ie non-overlapping generations)

Wright-Fisher model

Adults produce a large pool of gametes which will be randomly joined to form offspring (i.e. random mating).

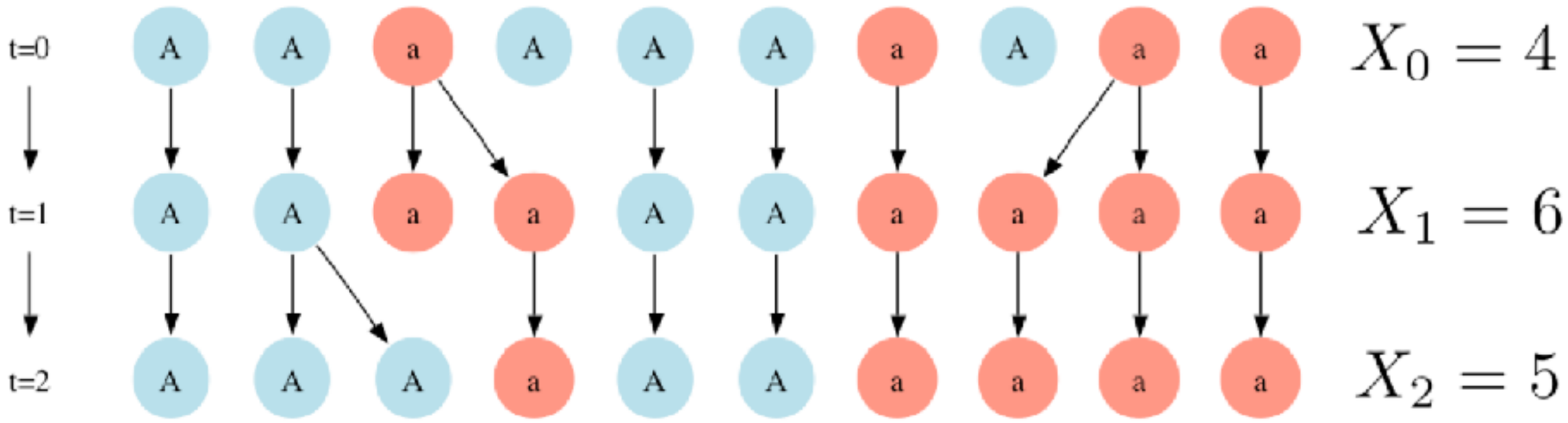
$2N$ gametes are sampled at random from the pool to make up the offspring in the next generation



Wright-Fisher model

If i out of $2N$ adult chromosomes carry allele A, then in the gamete pool, a proportion $p_t - i$ will be of type A.

$2N - i$ chromosomes carry allele a, with frequency $q_t = 1 - p_t$

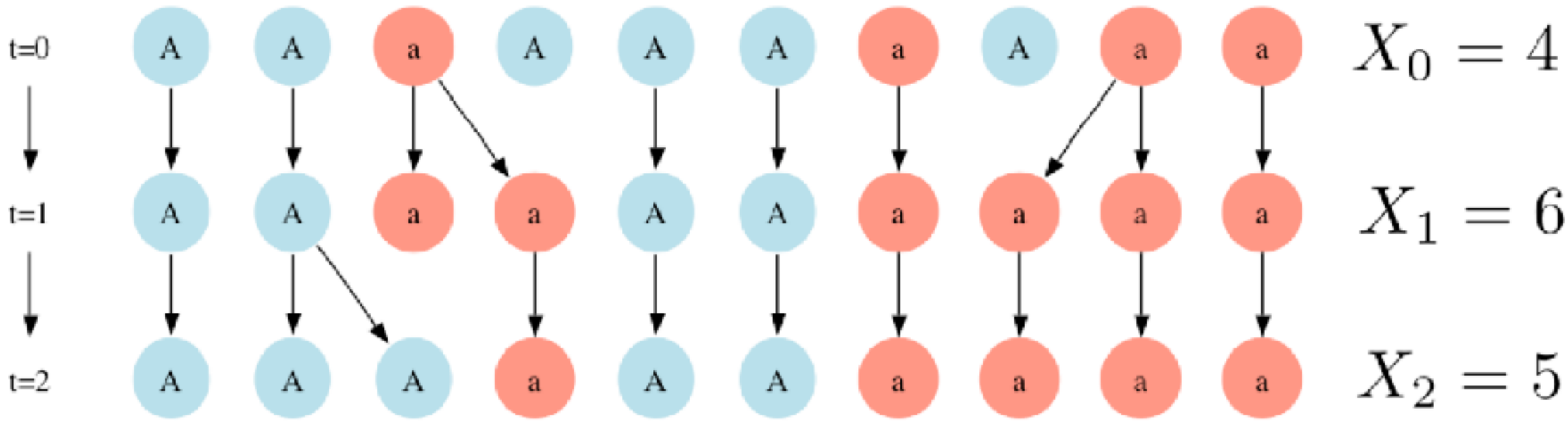


Wright-Fisher model

If i out of $2N$ adult chromosomes carry allele A, then in the gamete pool, a proportion $p_t - i$ will be of type A.

$2N - i$ chromosomes carry allele a, with frequency $q_t = 1 - p_t$

randomly sampling chromosomes each generation is like sampling from binomial distribution with parameters $2N$ and $i/2N$



Wright-Fisher model

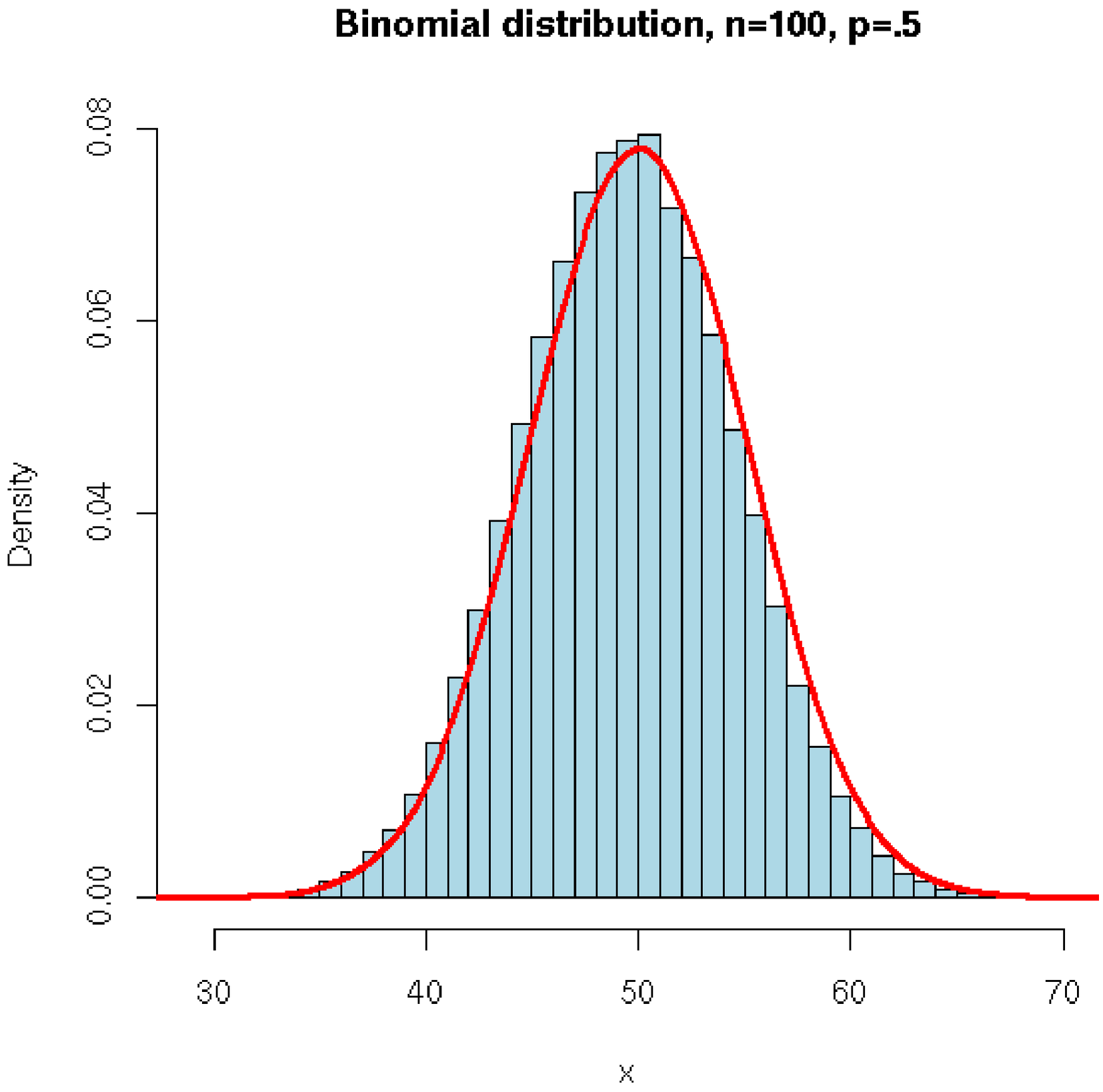
If i out of $2N$ adult chromosomes carry allele A, then in the gamete pool, a proportion $p_t - i$ will be of type A.

$2N - i$ chromosomes carry allele a, with frequency $q_t = 1 - p_t$

randomly sampling chromosomes each generation is like sampling from binomial distribution with parameters $2N$ and $i/2N$

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

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


Binomial distribution, n=100, p=.5
Density
0.00
0.02
0.04
0.06
0.08
30
40
50
60
70
x

n trials; *p* probability of success

$$\Pr[X] = \binom{n}{X} p^X (1-p)^{n-X}$$

$$\binom{n}{X} = \frac{n!}{X!(n-X)!}$$

of "successful" trials

$$P(X_{t+1} = j | X_t = i) = \binom{2N}{j} p^j (1-p)^{2N-j}$$

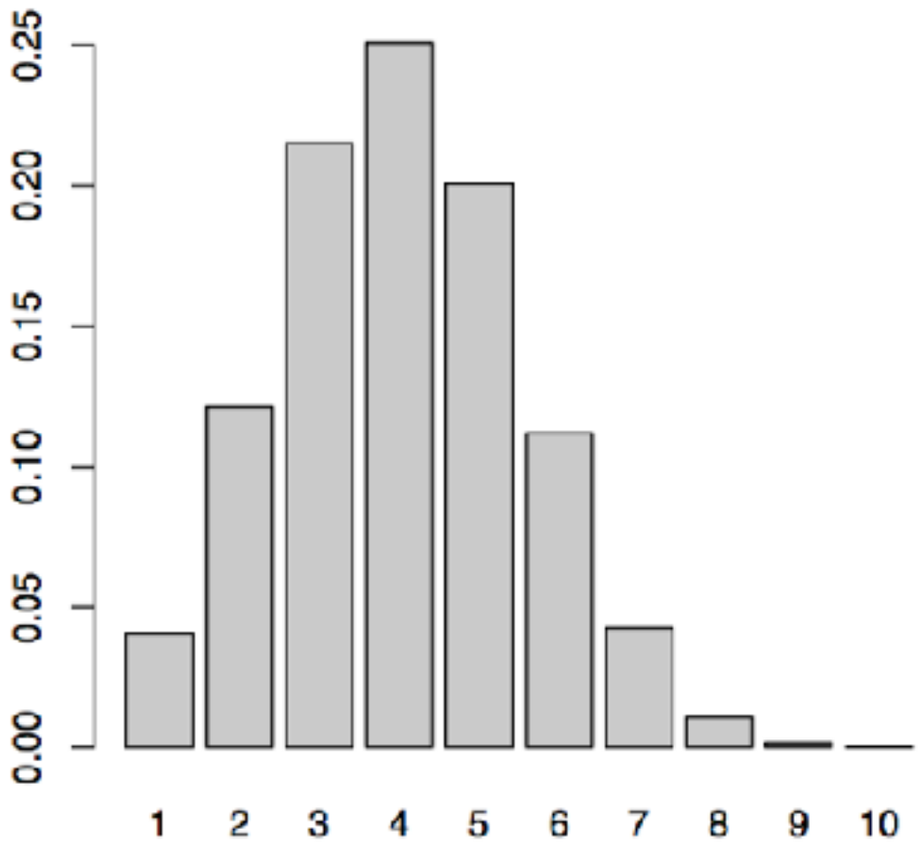
Number of ways to get j gametes of type A out of 2N gametes

Probability picking one gamete of type A

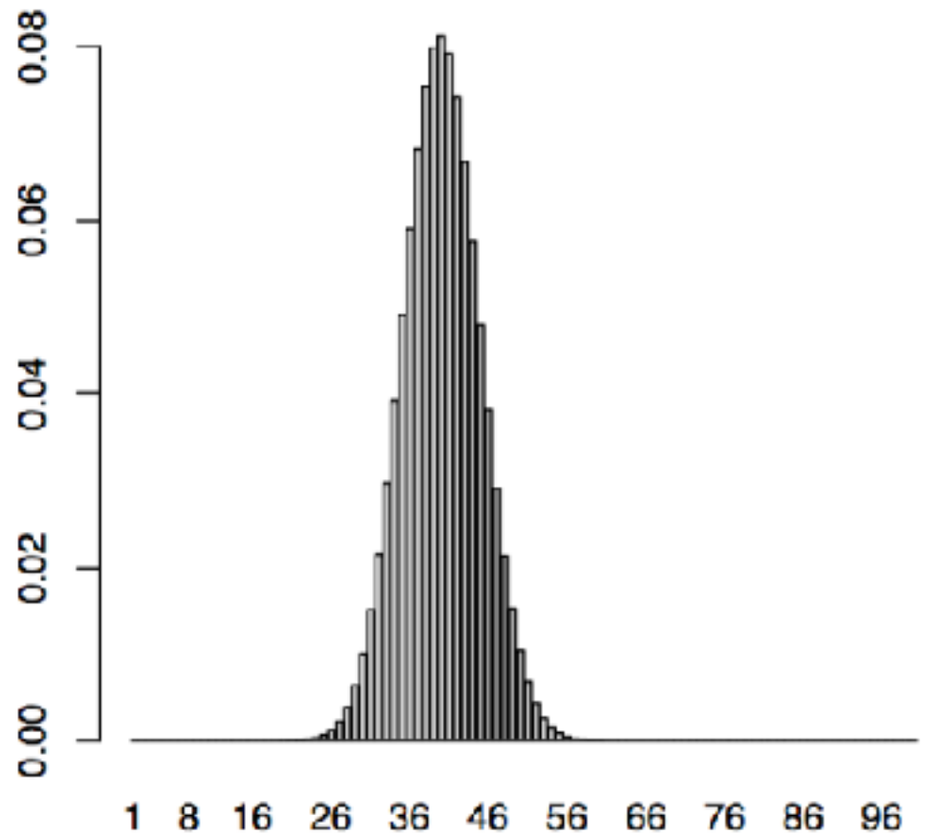
Probability not picking one gamete of type A

example: Binomial sampling probabilities

$i = 4,\ 2N = 10,\ p = 0.4$



$i = 40,\ 2N = 100,\ p = 0.4$

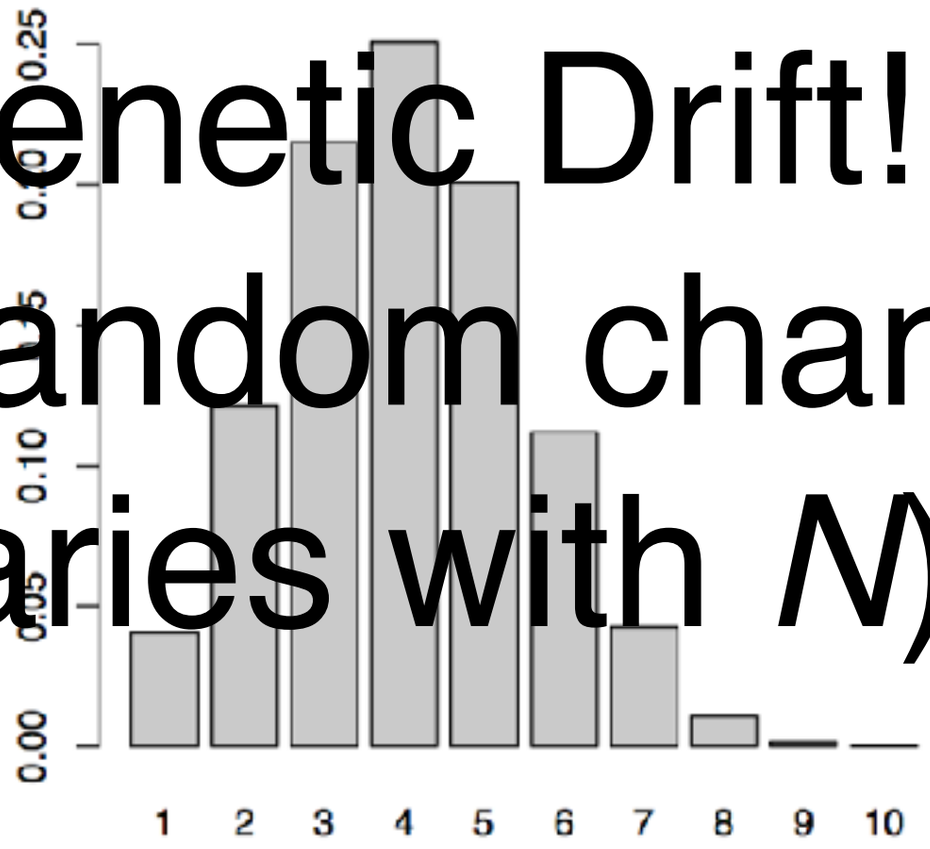


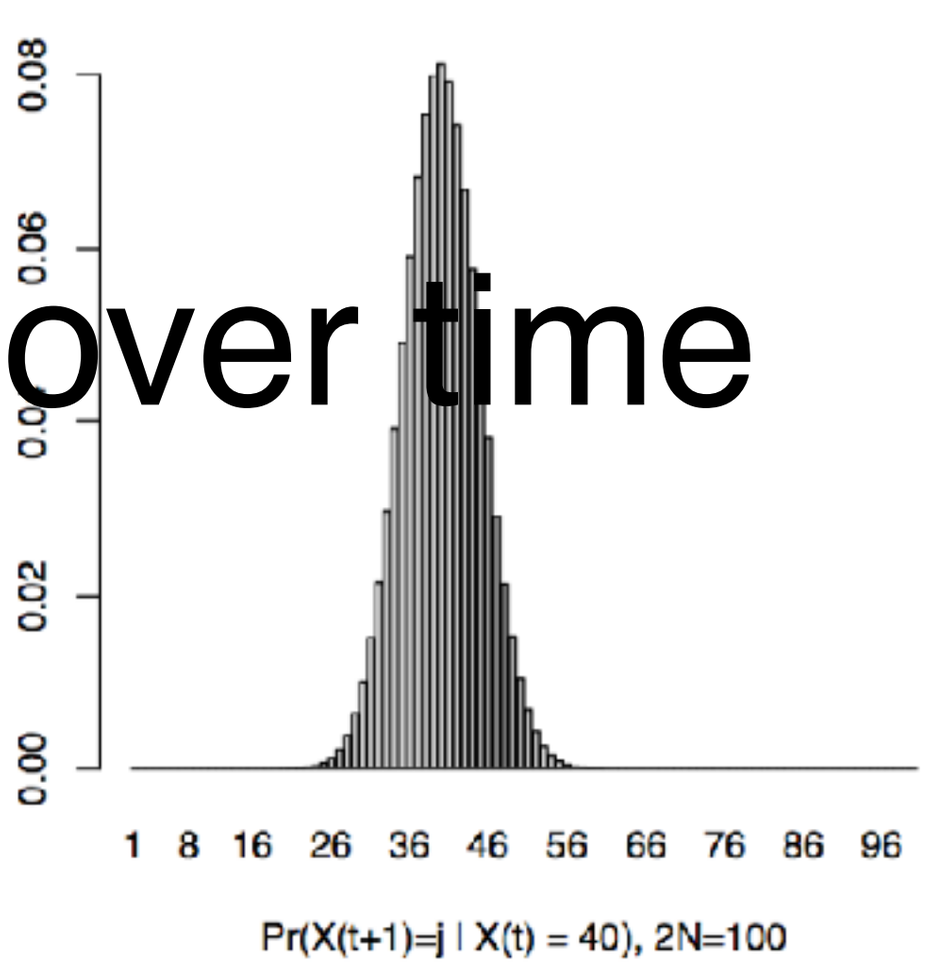
example: Binomial sampling probabilities

$i = 4,\ 2N = 10,\ p = 0.4$

$i = 40,\ 2N = 100,\ p = 0.4$

Genetic Drift!
(random change over time varies with *N*)





Wright-Fisher model

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

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

cannot predict change (E), but can predict that it ***will*** change (Var)
(and amount of change is related to *N*

Marcy K. Uyenoyama

Professor of Biology

Office:
130 Science Drive, Room 137, Duke Box 90338, Durham, NC 27708
Campus Box:
Duke Box 90338, Durham, NC 27708
Phone: (919) 660-7350
Email: marcy@duke.edu

- http://www.biology.duke.edu/marcylab/

Evolution of the sex ratio and effective number under gynodioecy and androdioecy

Marcy K. Uyenoyama [a,*], Naoki Takebayashi [b]

[a] *Department of Biology, Box 90338, Duke University, Durham, NC 27708-0338, USA*
[b] *Institute of Arctic Biology and Department of Biology and Wildlife, University of Alaska, Fairbanks, Fairbanks, AK 99775, USA*

Androdioecy: In the next generation forward in time, genotypic frequencies correspond to

$$
\begin{aligned}
z_0' &\propto \tilde{s}\tau(z_0h_0 + z_1h_1/4) + (1-\tilde{s})(z_0h_0 + z_1h_1/2)q \\
z_1' &\propto \tilde{s}\tau z_1h_1/2 + (1-\tilde{s})[(z_0h_0 + z_1h_1/2)(1-q) \\
&\quad + (z_1h_1/2 + z_2h_2)q] \\
z_2' &\propto \tilde{s}\tau(z_1h_1/4 + z_2h_2) + (1-\tilde{s})(z_1h_1/2 + z_2h_2)(1-q),
\end{aligned}
$$

for q denoting the frequency of the A allele in the pollen pool:

$$
q = \frac{h_0z_0 + h_1z_1/2 + \omega Z[(1-h_0)z_0 + (1-h_1)z_1/2]}{h_0z_0 + h_1z_1 + h_2z_2 + \omega Z[(1-h_0)z_0 + (1-h_1)z_1 + (1-h_2)z_2]}. \tag{21a}
$$

These expressions imply

$$
\begin{aligned}
Tz_0' &= s_A(z_0h_0 + z_1h_1/4) + (1-s_A)(z_0h_0 + z_1h_1/2)q \\
Tz_1' &= s_Az_1h_1/2 + (1-s_A)[(z_0h_0 + z_1h_1/2)(1-q) \\
&\quad + (z_1h_1/2 + z_2h_2)q] \\
Tz_2' &= s_A(z_1h_1/4 + z_2h_2) + (1-s_A)(z_1h_1/2 + z_2h_2)(1-q),
\end{aligned} \tag{21b}
$$

for s_A given in (9) and the normalizer by

$$
T = h_0z_0 + h_1z_1 + h_2z_2. \tag{21c}
$$

In the absence of selection on the modifier locus ($h_0 = h_1 = h_2$), recursion system (21a) indicates that allele frequency in seeds and pollen ($z_0 + z_1/2 = q$) remains at its initial value, with asymptotic convergence at rate $s_A/2$ of the frequency of heterozygotes (z_1) to

$$
2q(1-q)(1-F_{neut}),
$$

for F_{neut} the fixation index (Wright, 1933):

$$
F_{neut} = \frac{s}{2-s}, \tag{22}
$$

Gynodioecy: Genotypic frequencies in the next generation forward in time correspond to

$$
\begin{aligned}
z_0' &\propto \tilde{s}\tau(z_0h_0 + z_1h_1/4) \\
&\quad + \{(1-\tilde{s})(z_0h_0 + z_1h_1/2) + \tilde{\sigma}Z[z_0(1-h_0) \\
&\quad + z_1(1-h_1)/2]\}q \\
z_1' &\propto \tilde{s}\tau z_1h_1/2 \\
&\quad + \{(1-\tilde{s})(z_0h_0 + z_1h_1/2) + \tilde{\sigma}Z[z_0(1-h_0) \\
&\quad + z_1(1-h_1)/2]\}(1-q) \\
&\quad + \{(1-\tilde{s})(z_1h_1/2 + z_2h_2) + \tilde{\sigma}Z[z_1(1-h_1)/2 \\
&\quad + z_2(1-h_2)]\}q \\
z_2' &\propto \tilde{s}\tau(z_1h_1/4 + z_2h_2) \\
&\quad + \{(1-\tilde{s})(z_1h_1/2 + z_2h_2) + \tilde{\sigma}Z[z_1(1-h_1)/2 \\
&\quad + z_2(1-h_2)]\}(1-q),
\end{aligned}
$$

in which q represents the frequency of the A allele in the pollen pool (which derives entirely from hermaphrodites),

$$
q = \frac{h_0z_0 + h_1z_1/2}{h_0z_0 + h_1z_1 + h_2z_2}. \tag{23a}
$$

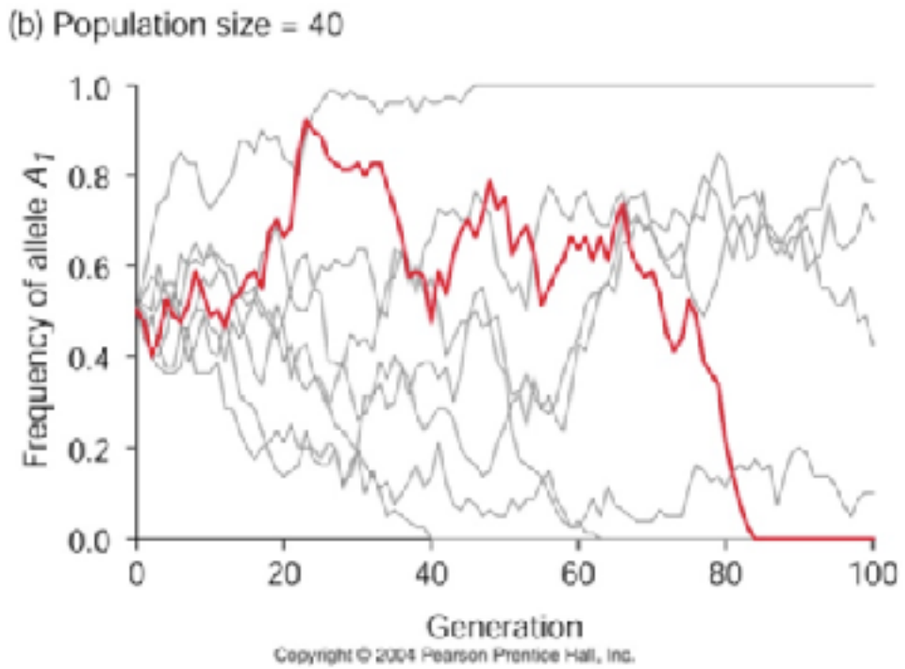
After division by $(\tilde{s}\tau + 1 - \tilde{s})$, we obtain

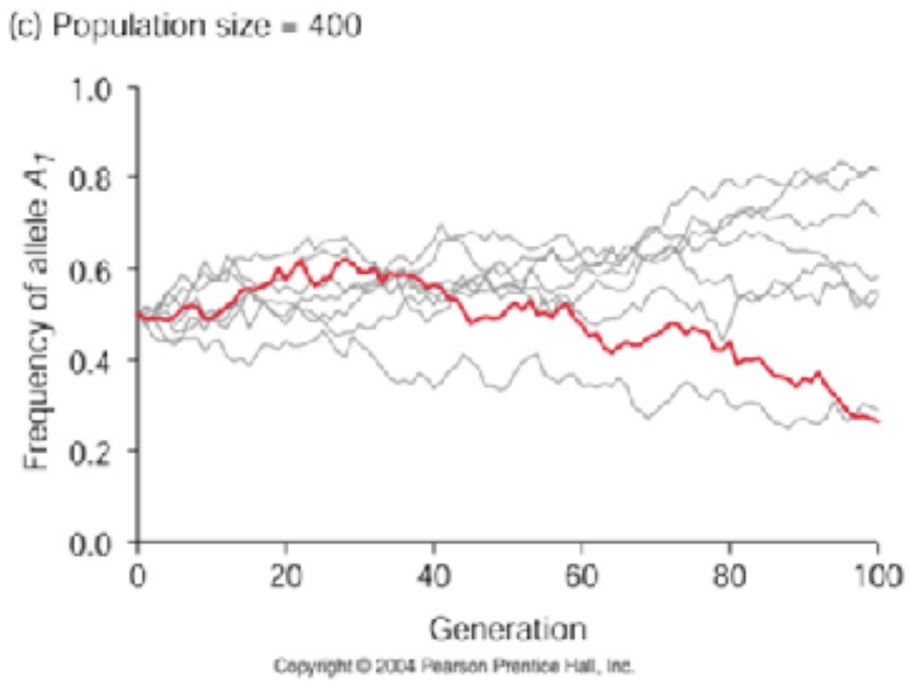
$$
\begin{aligned}
Tz_0' &= s(z_0h_0 + z_1h_1/4) \\
&\quad + \{[1-s][z_0h_0 + z_1h_1/2) + \sigma Z[z_0(1-h_0) \\
&\quad + z_1(1-h_1)/2]\}q \\
Tz_1' &= sz_1h_1/2 \\
&\quad + \{[1-s][z_0h_0 + z_1h_1/2) + \sigma Z[z_0(1-h_0) \\
&\quad + z_1(1-h_1)/2]\}(1-q) \\
&\quad + \{[1-s][z_1h_1/2 + z_2h_2) + \sigma Z[z_1(1-h_1)/2 \\
&\quad + z_2(1-h_2)]\}q \\
Tz_2' &= s(z_1h_1/4 + z_2h_2) \\
&\quad + \{[1-s][z_1h_1/2 + z_2h_2) + \sigma Z[z_1(1-h_1)/2 \\
&\quad + z_2(1-h_2)]\}(1-q),
\end{aligned} \tag{23b}
$$

for the normalizer corresponding to

$$
T = \sum_{i=0}^{2} z_i[h_i + \sigma Z(1-h_i)]. \tag{23c}
$$

Example allele frequency trajectories due to genetic drift

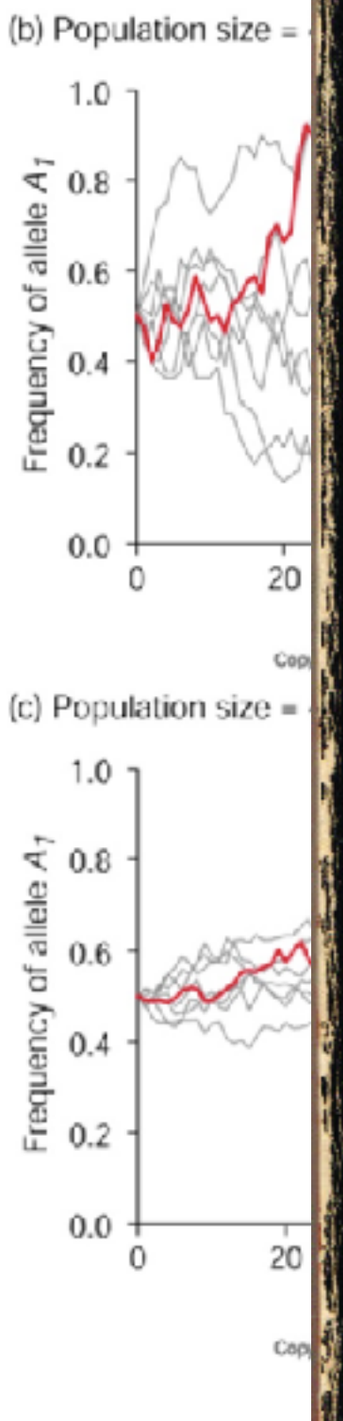


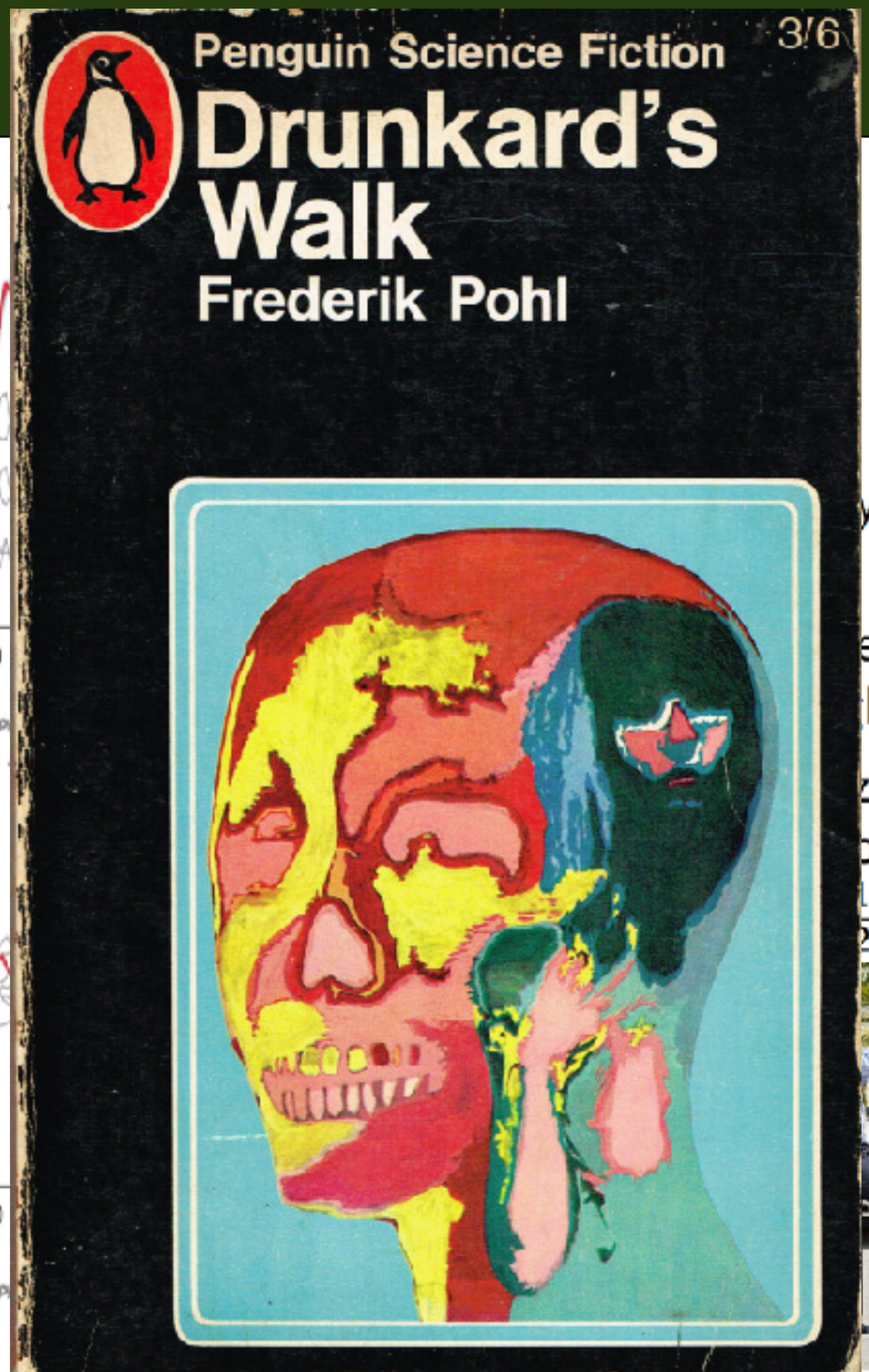


Analogy to a "drunkard's walk".

- Stepping left and right at random
- Size of the steps proportional here to: $\frac{p(1-p)}{2N}$

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*

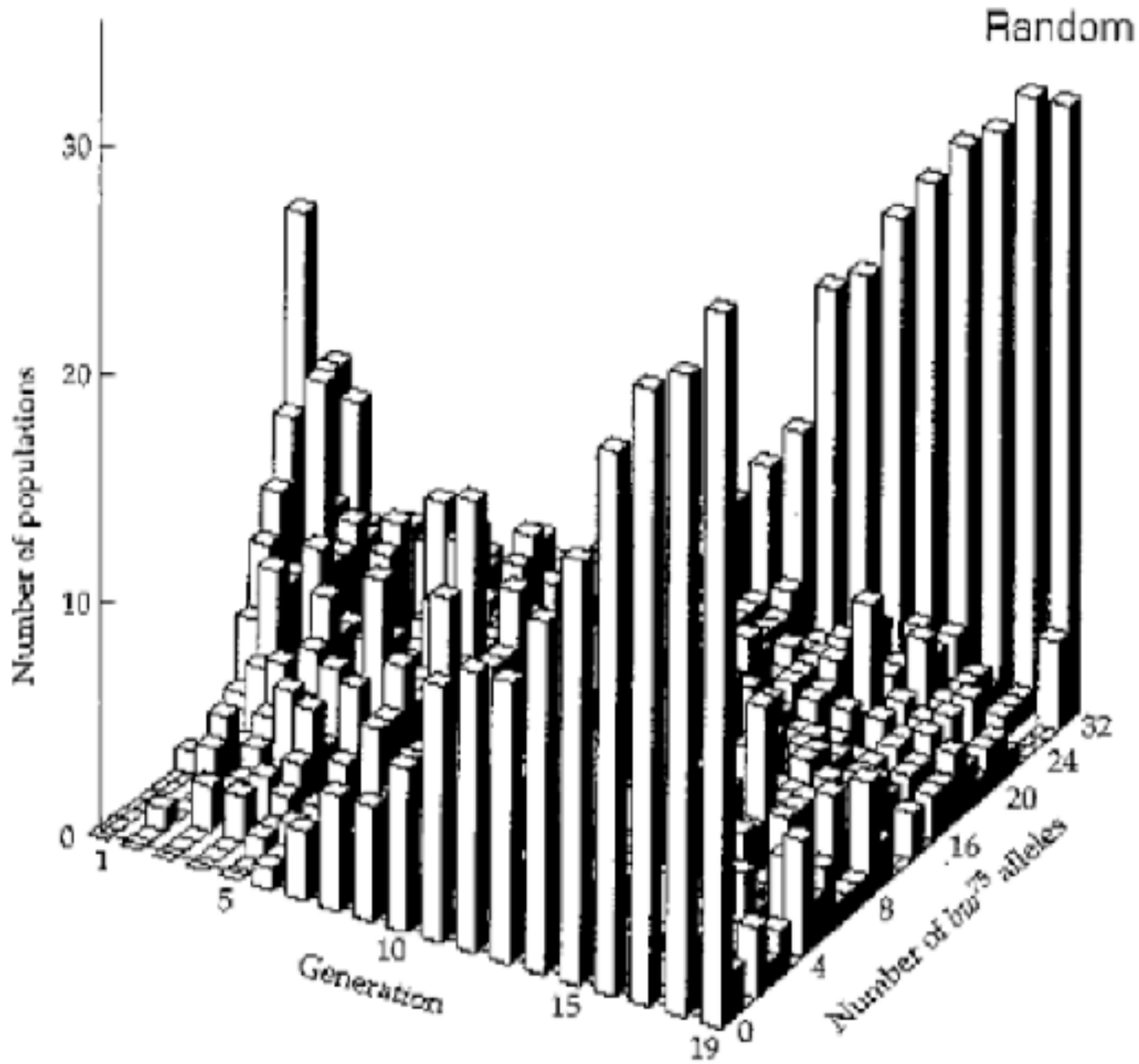


FIGURE 3.4 Random genetic drift in 107 actual populations of *Drosophila melanogaster*. Each of the initial 107 populations consisted of 16 bw^{75}/bw heterozygotes ($N = 16$; bw = brown eyes). From among the progeny in each generation, eight males and eight females were chosen at random to be the parents 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

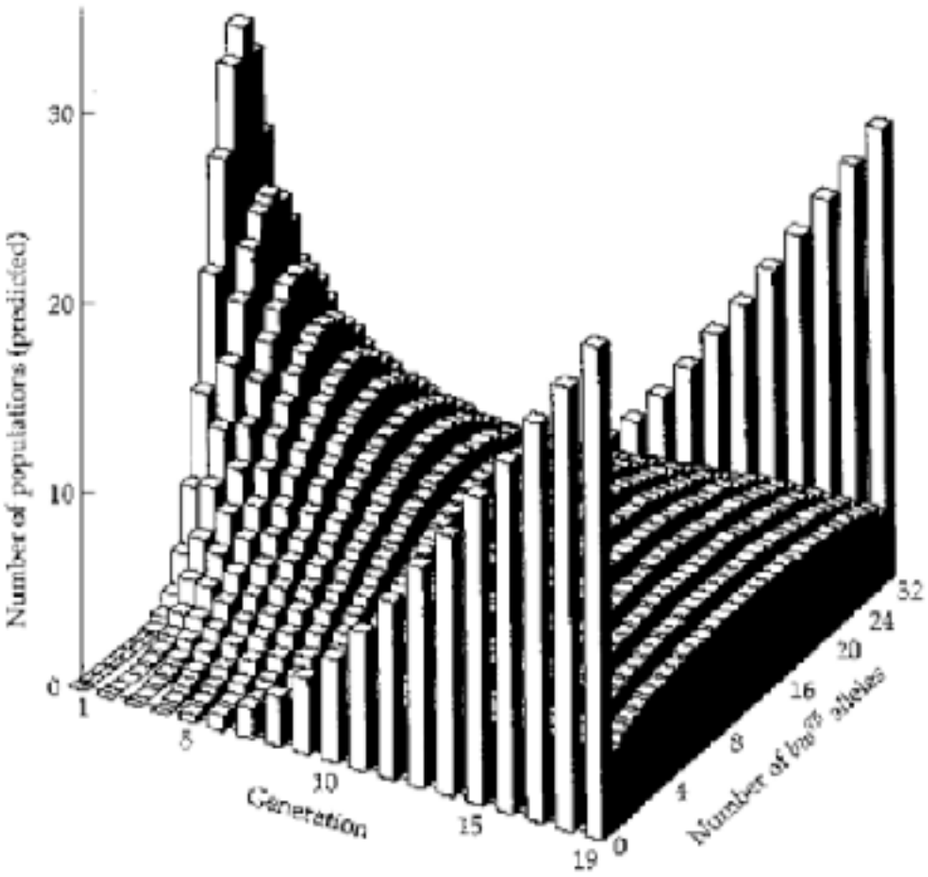


FIGURE 3.5 Prediction of the Wright-Fisher model for the distribution of allele frequencies $\phi(p, x; t)$ in subpopulations of size $N = 16$, where x represents the allele frequency in generation t. Time runs for 19 generations, and all subpopulations start with an initial allele frequency of $p = 0.5$. 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.

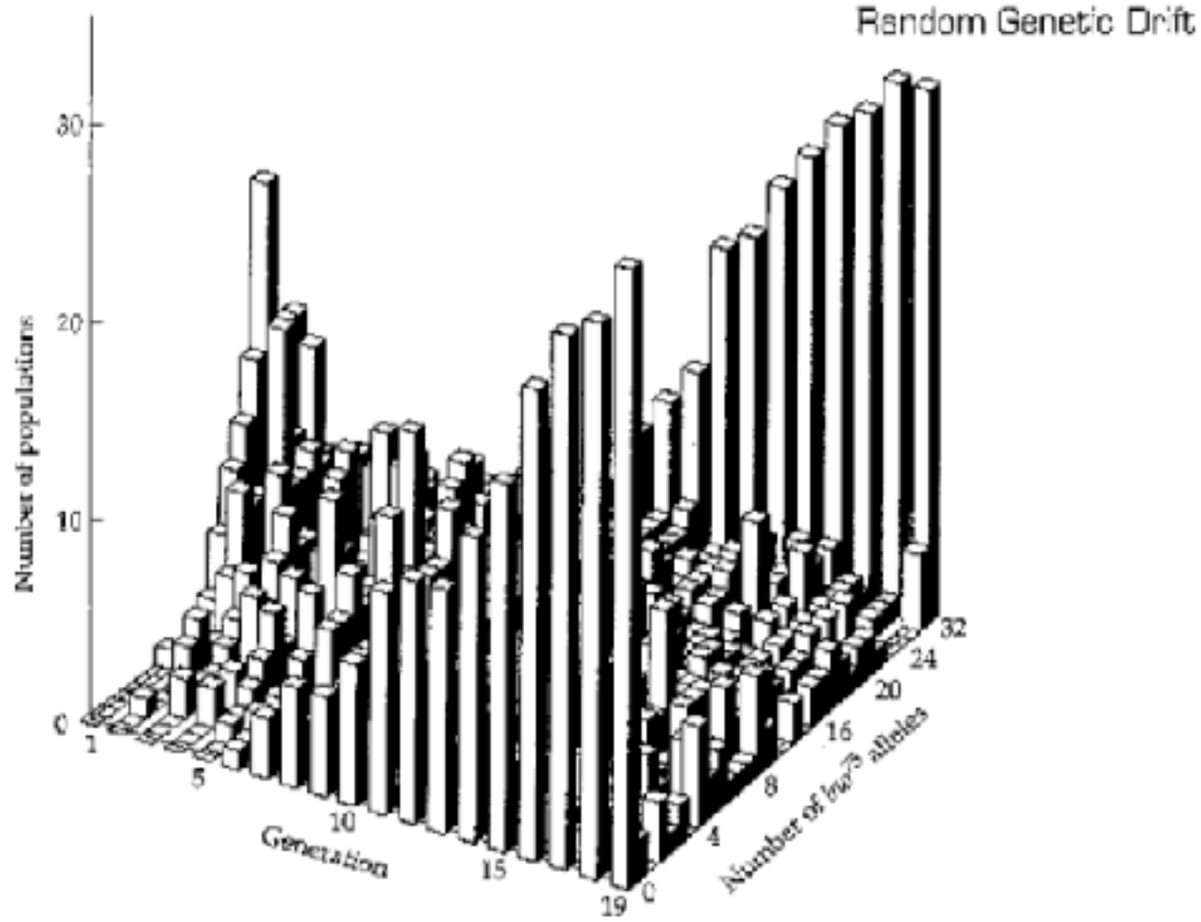
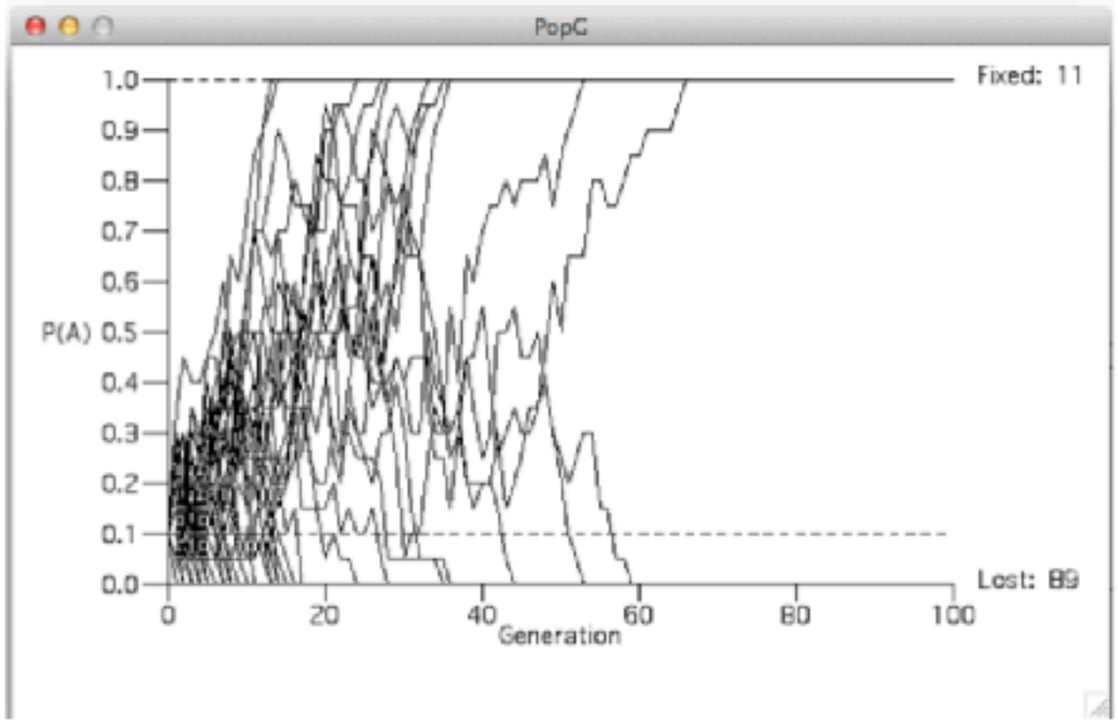


FIGURE 3.4 Random genetic drift in 107 actual populations of *Drosophila melanogaster*. Each of the initial 107 populations consisted of 16 bw^{75}/bw heterozygotes ($N = 16$; bw = brown eyes). From among the progeny in each generation, eight males and eight females were chosen at random to be the parents 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

Messer Lab — SLiM

About SLiM

SLiM is an evolutionary simulation framework that combines a powerful engine for population genetic simulations with the capability of modeling arbitrarily complex evolutionary scenarios. Simulations are configured via the integrated Eidos scripting language that allows interactive control over practically every aspect of the simulated evolutionary scenarios. The underlying individual-based simulation engine is highly optimized to enable modeling of entire chromosomes in large populations. For Mac OS X users (on OS X 10.10 or later), we also provide a graphical user interface for easy simulation set-up, interactive runtime control, and dynamical visualization of simulation output.

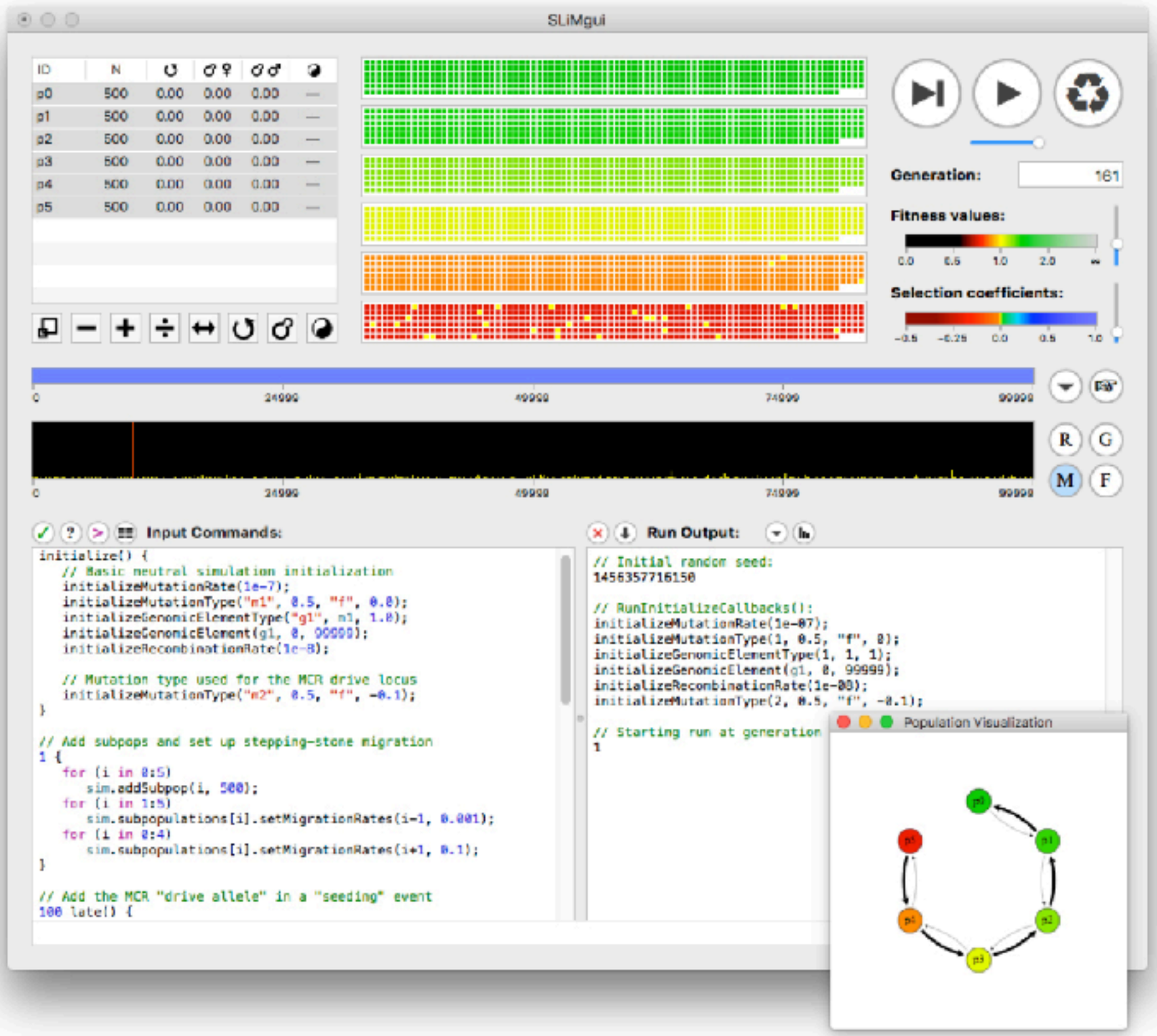
Downloads (version 3.3)

OS X Installer | Source Code | SLiM Manual | Eidos Manual | Ref Sheets

The OS X Installer will install the slim command-line tool, the SLiMgui graphical development environment, and both manuals. On other Un*x platforms, you should download the source code archive above, unzip it, and build it following the instructions in the provided

SLiMgui

With SLiMgui for Mac OS X you can visualize your simulation as it runs and examine its parameters in real-time, allowing for much easier simulation development.



Some SLiM-based simulation exercises

https://github.com/petrelharp/popbio/blob/master/notebooks/slim_intro.ipynb

SLiM manual - default demography: a randomly-mating Wright-Fisher population of fixed size.

```
basic_WF = """
// set up a simple neutral simulation
initialize()
{
    // set the overall mutation rate
    initializeMutationRate(1e-7);

    // m1 mutation type: neutral
    initializeMutationType("m1", 0.5, "f", 0.0);
    // m2 mutation type: beneficial
    initializeMutationType("m2", 0.5, "f", 0.0);

    // g1 genomic element type: uses m1 or m2 with equal prob for mutations
    initializeGenomicElementType("g1", c(m1,m2), c(1.0,1.0));

    // uniform chromosome of length 100 kb
    initializeGenomicElement(g1, 0, 99999);

    // uniform recombination along the chromosome
    initializeRecombinationRate(1e-8);
}

// create a population of 500 individuals
1 {
    sim.addSubpop("p1", 500);
}

// run to generation 10000
10000 {
    sim.simulationFinished();
}
9999 late() {
    p1.outputSample(10);
}
"""

out, logfile = slim_script(basic_WF, "basic_WF")
```

Exercise:

1- Put this in the SLiM GUI (if you have a Mac).

2 - Run it.

3 - Click on the "help" button and browse the functions.

4 - Option-click on a function to pop up the help for it.

Change the mutation type to be beneficial, like so:

initializeMutationType("m1", 0.5, "f", 0.0);

to

initializeMutationType("m1", 0.5, "f", 0.05);

and watch what happens.

if no mac (we can run same stuff from command line using jupyter notebook with conda and python)

Homework:

a. Read Chapter 1(if you haven't already)
b. Read chapter 2 and be ready for discussion next week (**2 randomly selected individuals will lead discussion**)

c. Read sections 1-4 of the SLiM Manual.

Exercise (at home?):

a. Add a beneficial mutation type to SLiM recipe 4 "Neutral simulations in a panmictic population", sampling 10 genomes from the population at the end
b. How do the two mutation types behave as a function of population size or mutation rate?

Next week: Lecture Topic: Genetic variation (Chapter 3)

Be looking into popgen simulation software (mac people -> guiSLiM). PipeMaster, msPrime etc