How to assemble a real genome with repeats?

Here we assume a "de novo" assembly without help from the previously assembled genomes

Nicolaas Govert de Bruijn (1918 – 2012) was a Dutch mathematician, noted for his many contributions in the fields of graph theory, analysis, number theory, combinatorics and logic

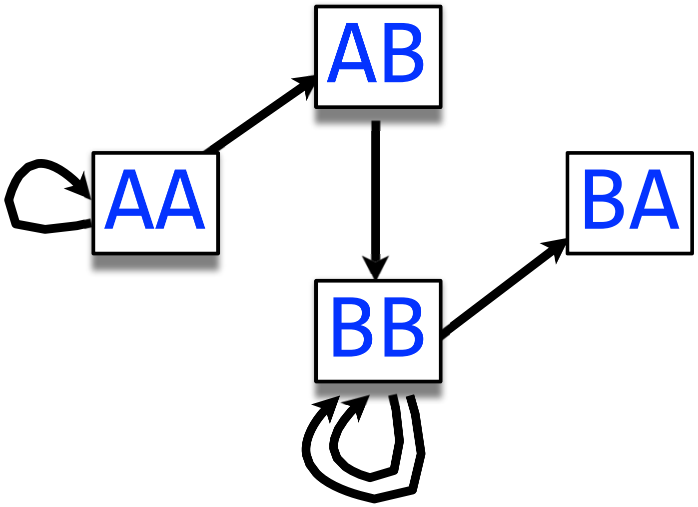
Courtesy of Ben Langmead. Used with permission.
http://www.langmead-lab.org/teaching-materials/

De Bruijn graph

genome: AAABBBBA

3-mers: AAA, AAB, ABB, BBB, BBB, BBA

L/R 2-mers: AA, AA AA, AB AB, BB BB, BB BB, BB BB, BA

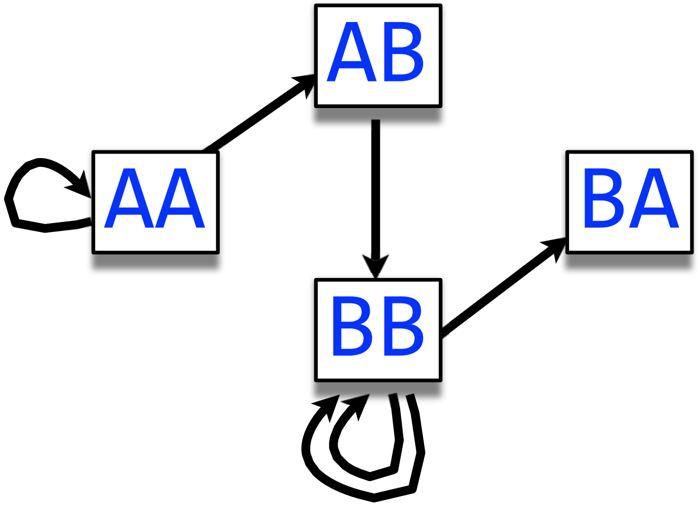


One edge per every *k*-mer

One node per distinct *k*-1-mer

Courtesy of Ben Langmead. Used with permission.
http://www.langmead-lab.org/teaching-materials/

De Bruijn graph

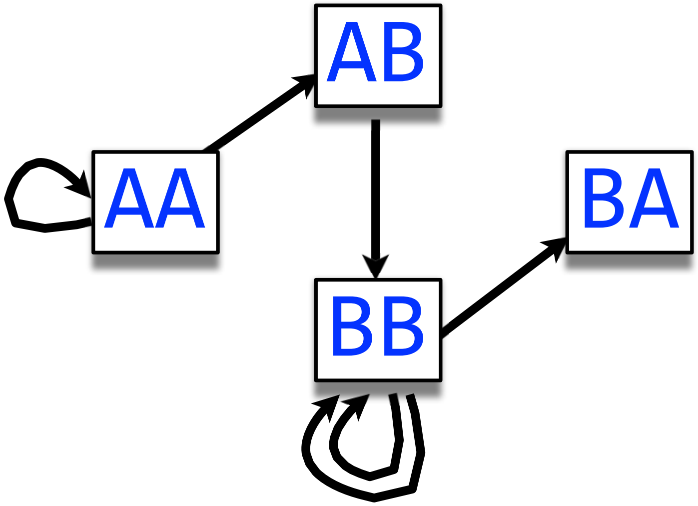


Walk crossing each edge exactly once gives a reconstruction of the genome

Courtesy of Ben Langmead. Used with permission.

http://www.langmead-lab.org/teaching-materials/

Assembly = Eulerian walk on De Bruijn graph



AAABBBBA

Walk crossing each edge exactly once gives a reconstruction of the genome. This is an *Eulerian walk*.

Courtesy of Ben Langmead. Used with permission. http://www.langmead-lab.org/teaching-materials/

Edge-disjoint loops are a problem: multiple solutions

graph can have multiple Eulerian walks, only one of which corresponds to original superstring

Right: graph for ZABCDABEFABY, k=2

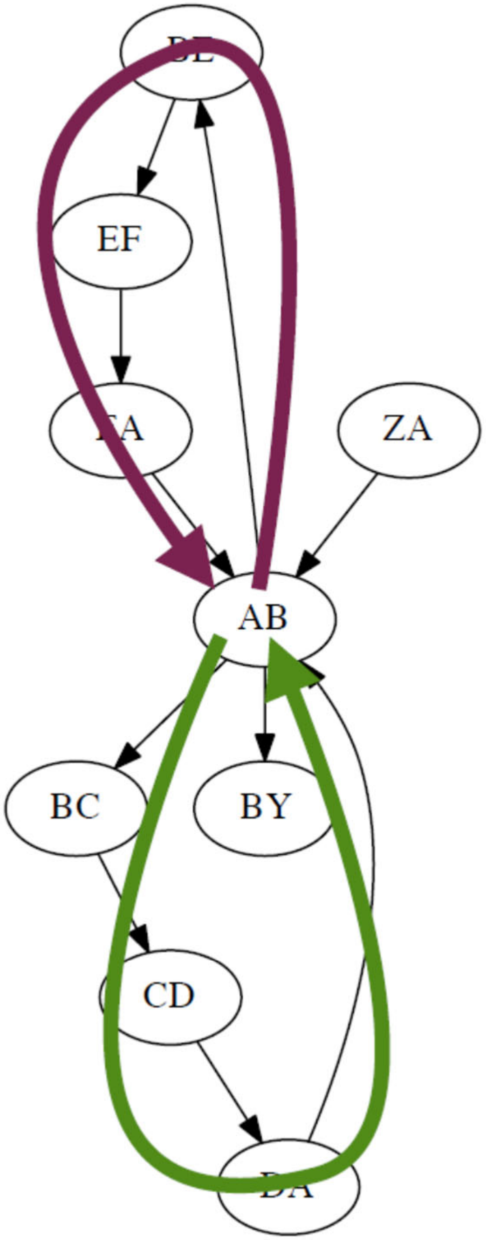
Alternative Eulerian walks:

ZA → AB → BE → EF → FA → AB → BC → CD → DA → AB → BY

ZA → AB → BC → CD → DA → AB → BE → EF → FA → AB → BY

These correspond to two edge-disjoint directed cycles joined by node AB

AB is a repeat: ZABCDABEFABY

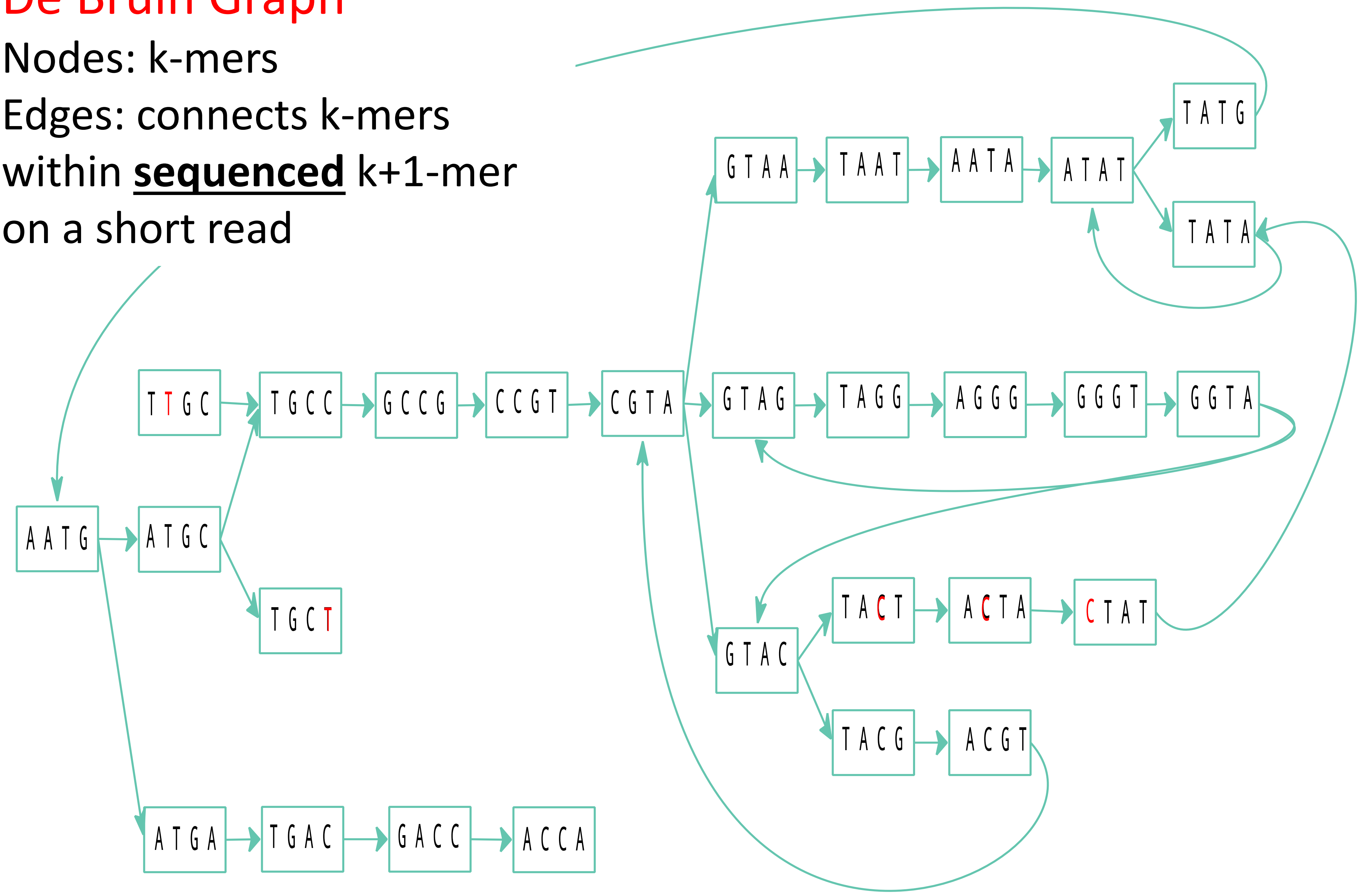


Adapted from a slide by Ben Langmead, Johns Hopkins U.

De Bruin Graph

Nodes: k-mers

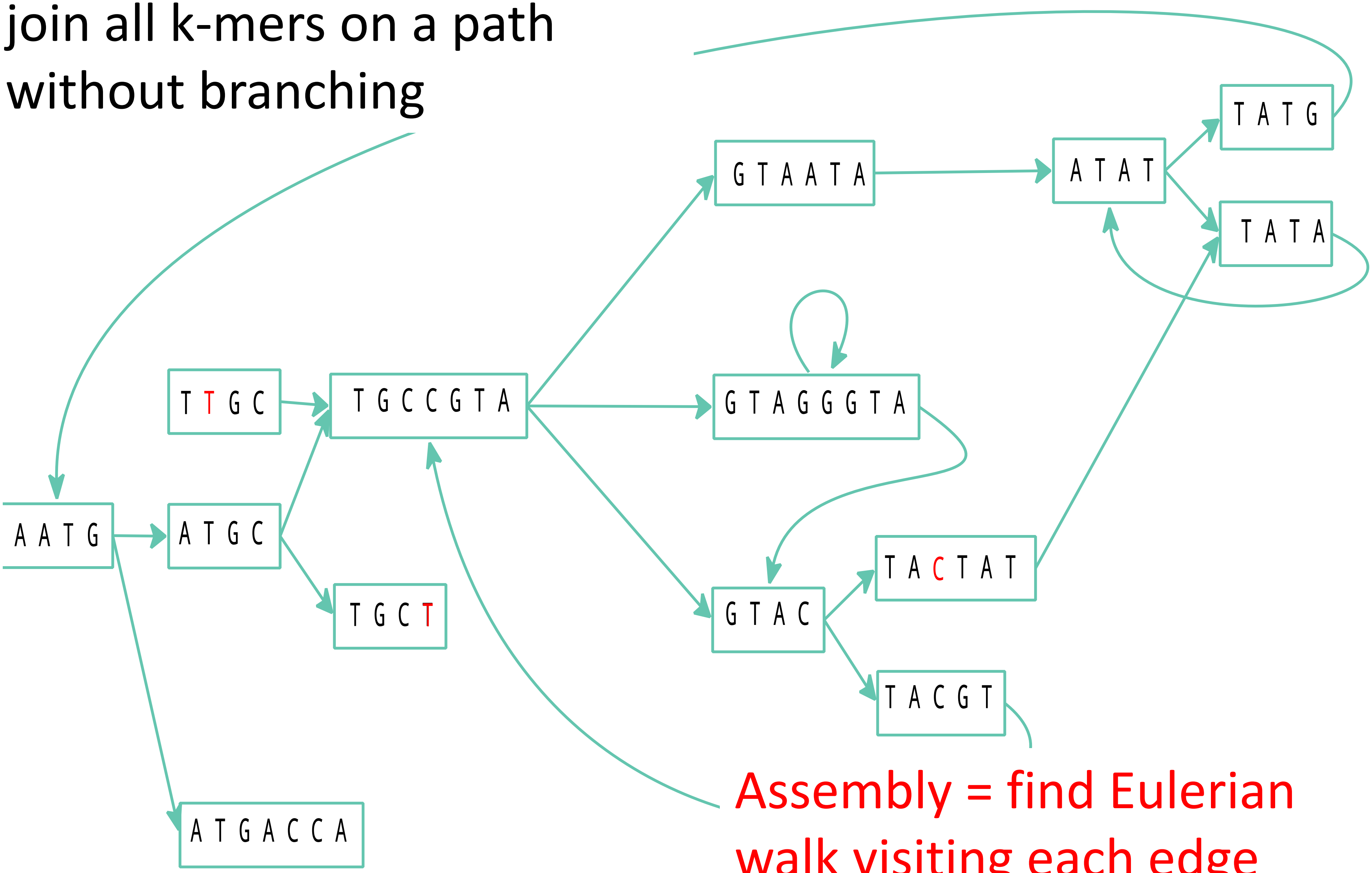
Edges: connects k-mers within **sequenced** k+1-mer on a short read



Slide by Sorin Istrail, Brown U.

Simplified De Bruin Graph

join all k-mers on a path without branching



Assembly = find Eulerian walk visiting each edge once

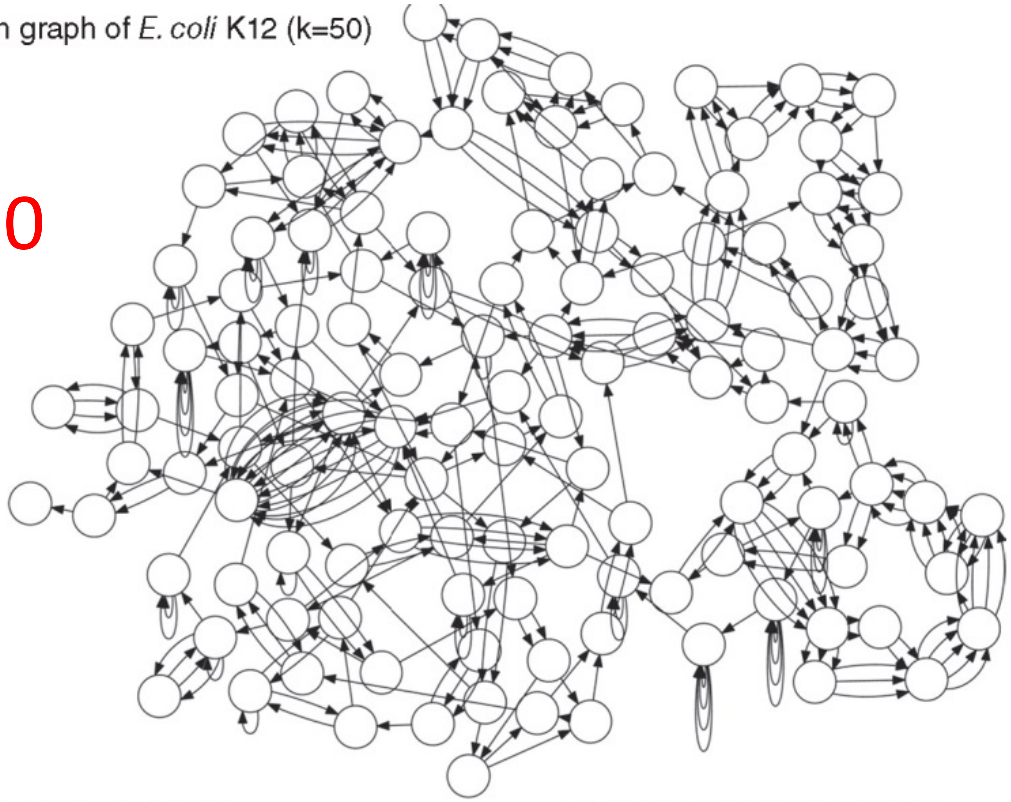
Slide by Sorin Istrail, Brown U.

How to assemble a genome with repeats?

- Answer:
  longer reads
- But:
  cheap sequencing
  =
  short reads

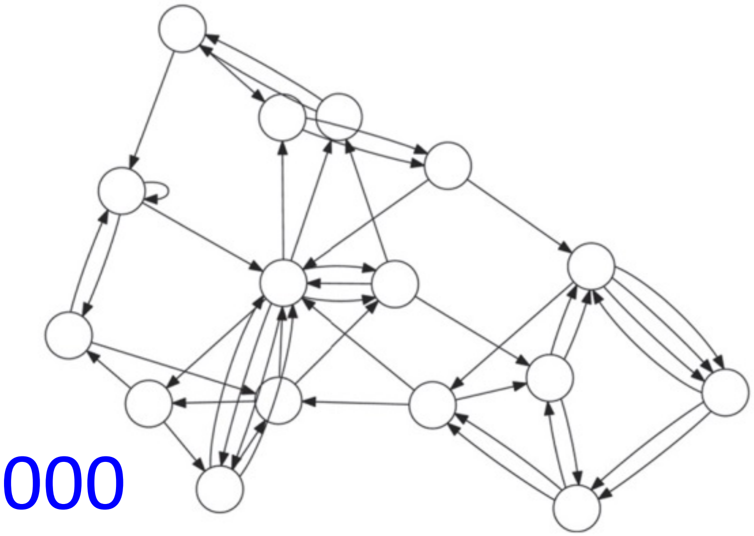
Technology	Read length (bp)
Roche 454	700
Illumina	50–250
SOLiD	50
Ion Torrent	400
Pacific Biosciences	>10,000

(a) de Bruijn graph of *E. coli* K12 (k=50)



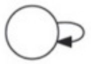
k=50

(b) de Bruijn graph (k=1,000)



k=1000

(c) de Bruijn graph (k=5,000)



k=5000

A gallery of useful
discrete probability distributions

Geometric Distribution

- A series of Bernoulli trials with probability of success =*p*. continued **until the first success**. X is the number of trials.
- Compare to: Binomial distribution has:
  - Fixed number of trials =n.
  - Random number of successes = x.

$$P(X=x)=C_x^n p^x (1-p)^{n-x}$$

- Geometric distribution has reversed roles:
  - Random number of trials, *x*
  - Fixed number of successes, in this case 1.
  - Success always comes in the end: so no combinatorial factor C_x^n
  - $P(X=x) = p(1-p)^{x-1}$ where:

    $x-1 = 0, 1, 2, \ldots$, the number of failures until the 1st success.
- NOTE OF CAUTION: Matlab, Mathematica, and many other sources use x to denote the number of failures until the first success. We stick with Montgomery-Runger notation

Geometric Mean & Variance

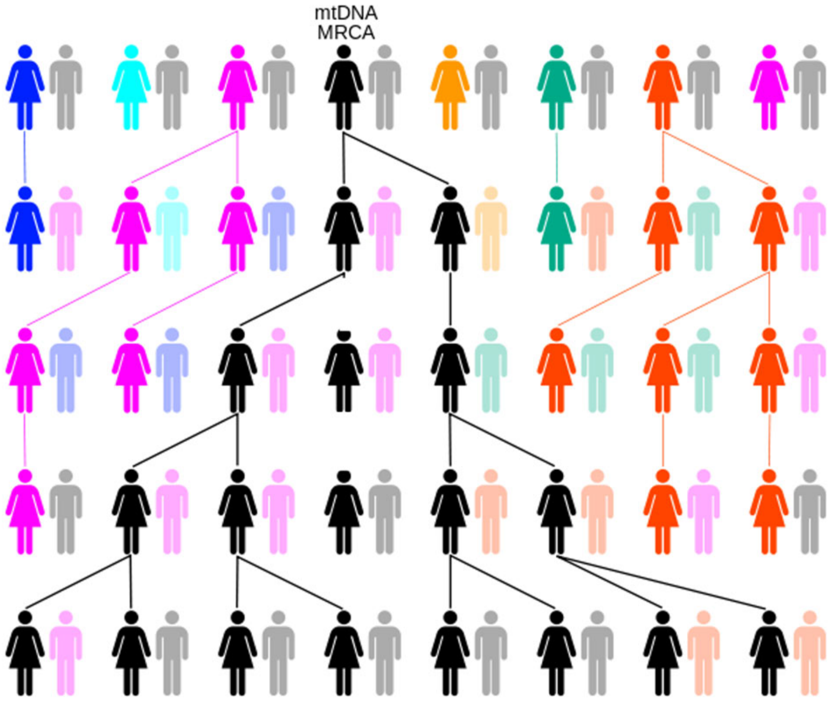
- If *X* is a geometric random variable (according to Montgomery-Bulmer) with parameter *p*,

$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \qquad \text{(3-10)}$$

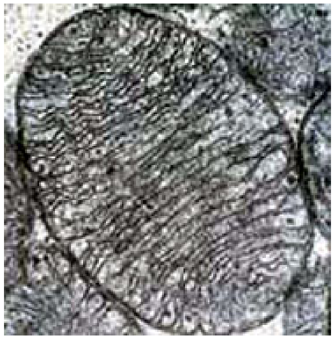
- For small *p* the standard deviation =$(1-p)^{0.5}/p$ ~= mean=1/p

- Very different from Binomial and Poisson, where variance = mean and standard deviation = $\text{mean}^{1/2}$

Geometric distribution in biology

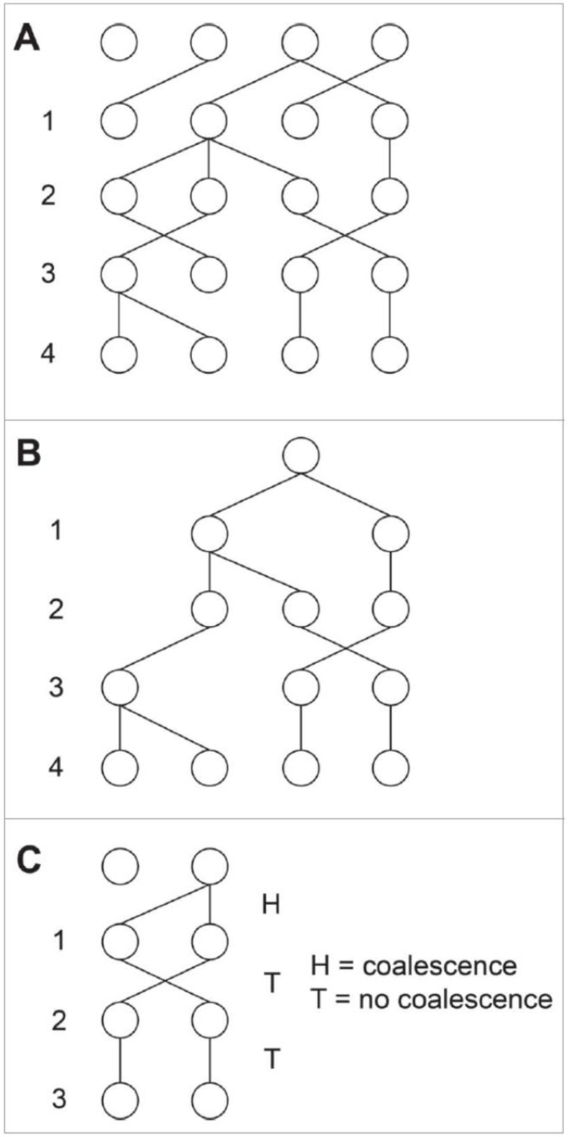


- Each of our cells has mitochondria with 16.5kb of mtDNA **inherited only from our mother**
- Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA
- Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeaon (of UIUC's Carl R. Woese fame)
- Since that time most mitochondrial genes were transferred into the nucleus
- Plants also have plastids with genomes related to cyanobacteria



Time to the last common (maternal) ancestor follows geometric distribution

- Constant population of N women
- Random number of (female) offsprings. Average is 1 (but can be 0 or 2)
- Randomly pick two women.
  Question: how many generations T since their last maternal ancestor?
- T is a random variable What is its PMF: $P(T=t)$?
  Answer: $P(T=t)$ follows a geometric distribution
- Do these two women have the same mother? Yes: "success" in finding their last common ancestor ($p=1/N$). $P(T=1)=1/N$.
- No? "failure" ($1-p=1-1/N$). Go to their mothers and repeat the same question.
- $P(T=t)=(1-1/N)^{t-1}(1/N) \approx (1/N)\exp(-(t-1)/N)$
- t can be inferred from the density of differences on mtDNA $=2\mu t$

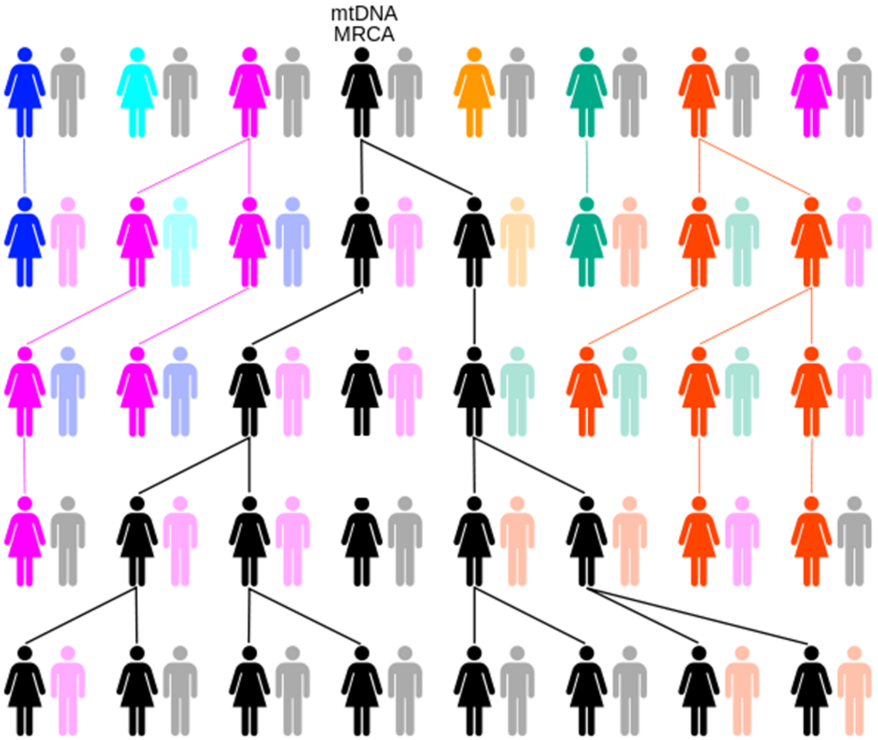


Maddamsetti R, MOBILE GENETIC ELEMENTS, 6, e1137380(2016)

Most Recent Common Ancestor (MRCA)

- Consider N women living today. Let $U(t)$ be the number of maternal ancestors t generations before present connected to all modern women
- At $t=0$, $U(0) = N$. Any of $\frac{U(0)(U(0)-1)}{2}$ pairs of woman can coalesce with probability $\frac{1}{N}$
- We expect $\frac{1}{N}\frac{U(0)(U(0)-1)}{2}$ pairs to coalesce

  $U(1) = U(0) - \frac{1}{N}\frac{U(0)(U(0)-1)}{2}$
- Now the expected number of coalescing pairs $\frac{1}{N}\frac{U(1)(U(1)-1)}{2}$



Most Recent Common Ancestor (MRCA)

$$\frac{dU(t)}{dt} = -\frac{\overbrace{1}^{\text{probability of coalescence}}}{N} \times \frac{\overbrace{U(t)(U(t)-1)}^{\text{number of pairs}}}{2} \approx \frac{U(t)^2}{2N}$$

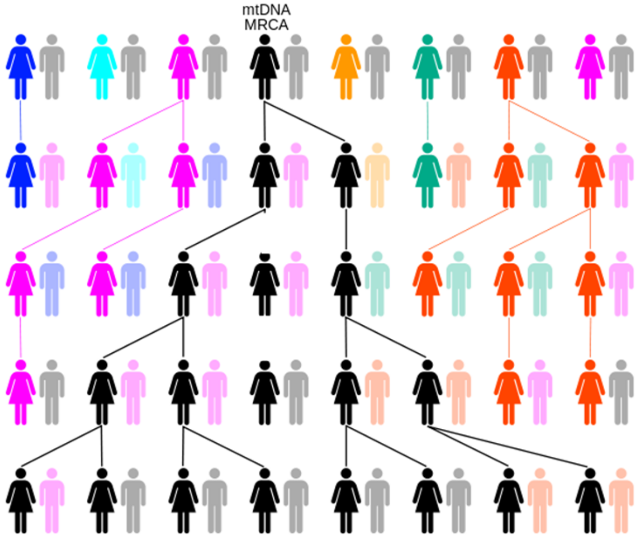
Solution: $U(t) = \frac{C}{t+2}$

Let's check: $\frac{dU}{dt}$=- $\frac{C}{(t+2)^2} = \frac{U^2}{C} \rightarrow C = 2N$

$$T_{MRCA} = 2(N-1) \approx 2N$$

Most Recent Common Ancestor (MRCA)

- Start with N individuals. Unit of time is N generations (time for one pair to merge) since $E(T)=\sum_{t=1}^{\infty} t \cdot (1/N)\, exp(-t/N)=N$
- Any of $\frac{N(N-1)}{2}$ pairs can merge first. The average time for the first pair to merge is $\frac{2}{N(N-1)}$
- After merger $N \rightarrow N-1$,
- So, the time until the next merger is longer $\frac{2}{(N-1)(N-2)}$



Most Recent Common Ancestor (MRCA)

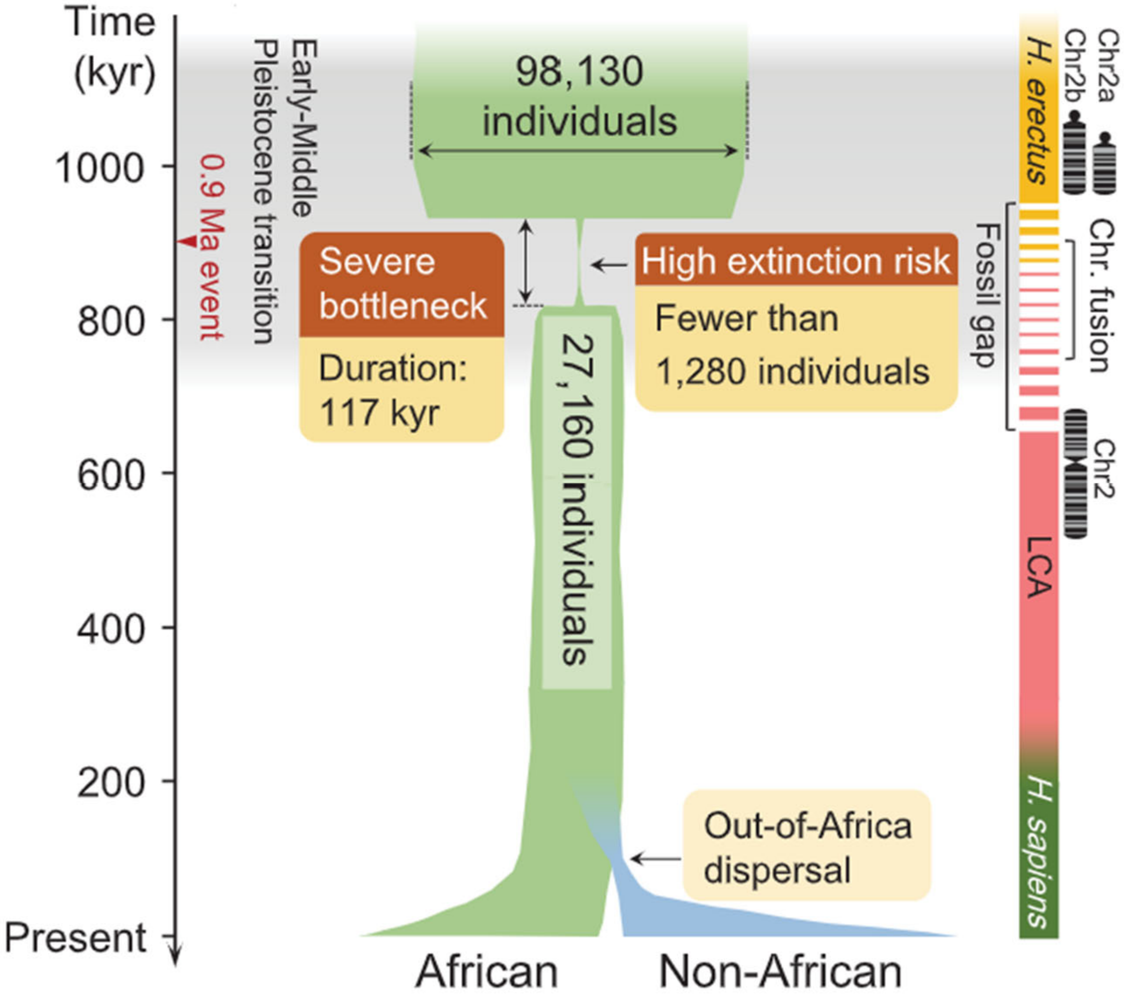
Total time until the MRCA

$$T_{MRCA} = N \cdot \sum_{k=2}^{N} \frac{2}{k(k-1)}$$

$$= 2N \sum_{k=2}^{N} \left(\frac{1}{k-1} - \frac{1}{k}\right) = 2N\left(1 - \frac{1}{N}\right) \approx 2N$$

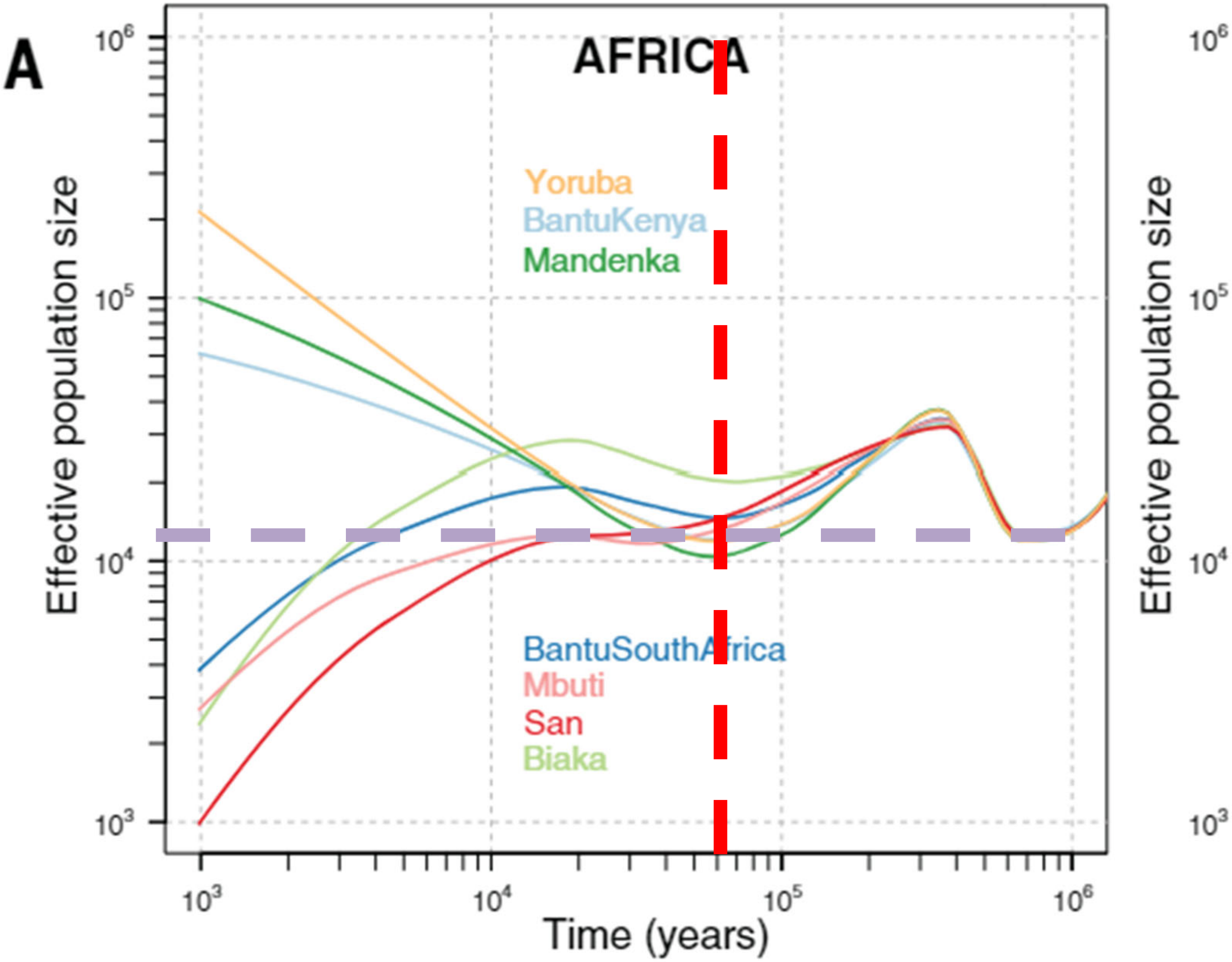
- There are about $N = 8x10^{9} /2 = 4x10^{9}$ women living today
- Most Recent maternal Common Ancestor (MRCA)
  of all people living today lived $T_{MRCA} = 2N$ generations ago
- $T_{MRCA} = 2 \cdot 4\,x10^{9}$ generations
- If the generation time 20 years it is 160 billion years > 10 times the time since the Big Bang.
- Something is wrong here!

Hot off the press: human ancestors almost got extinct about 1M years ago



Hu W, et al. Science. 2023;381: 979–984

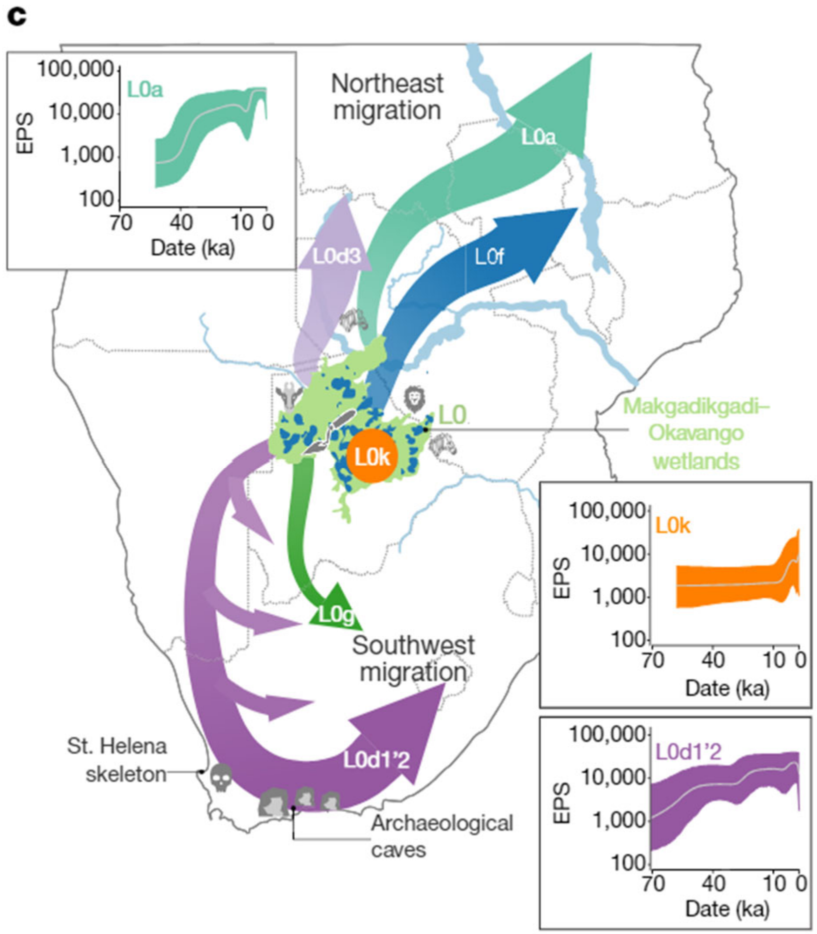
Bottleneck for human population in Africa around 10,000 individuals reached around 60,000 years ago



From ~1000 modern genomes: Bergström A, et al. Science. 2020;367

- Population is not constant and for a long time was very low
- Change N to the “effective” size N_e reached during the bottleneck
- Current thinking is that for all of humankind N_e~10,000 people
- Mito Eve lived in Africa ~2*(Ne/2)*20 years=10,000*20 years= 200,000 years ago

“Mitochondrial Eve” lived in Africa

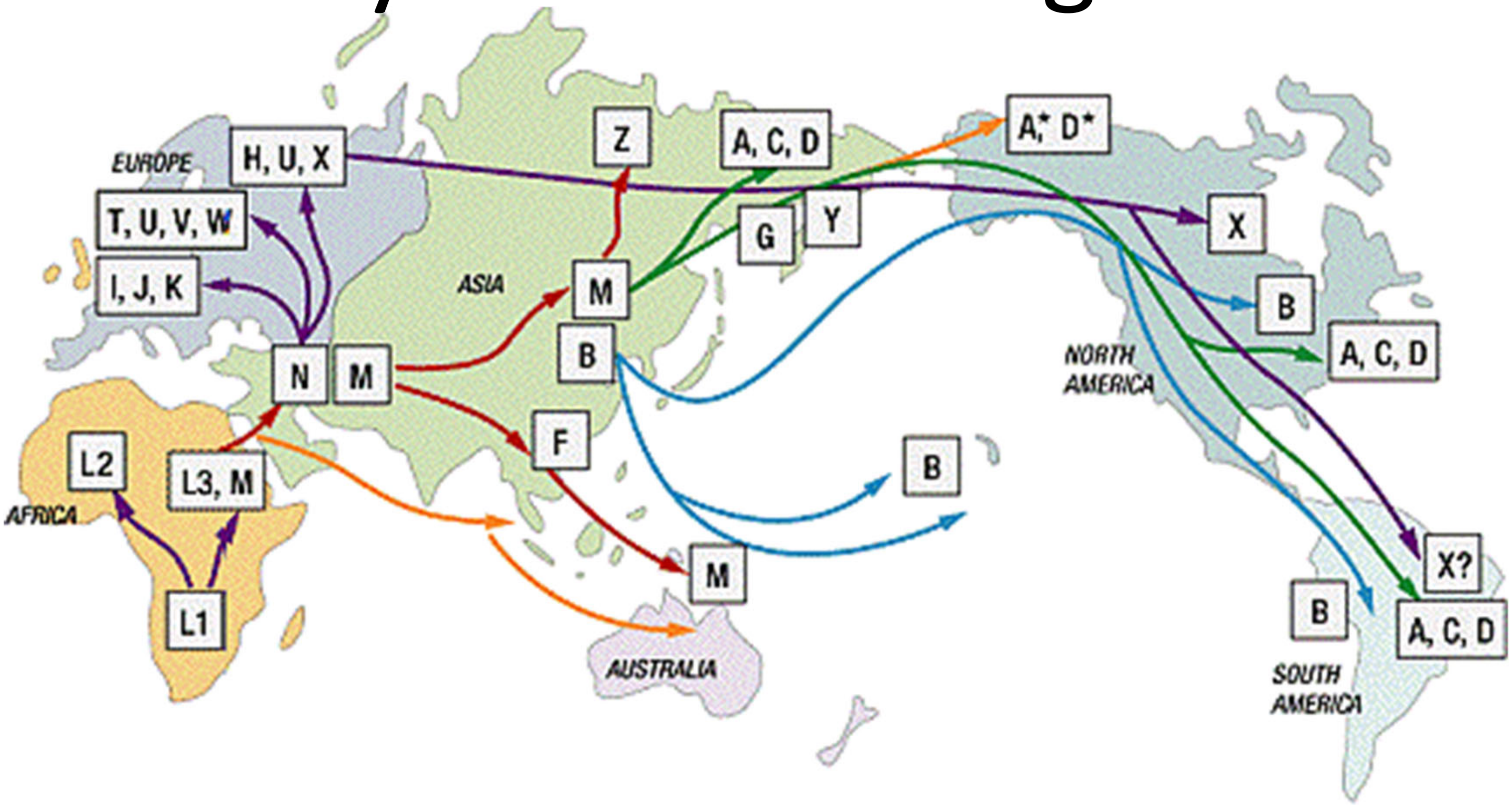


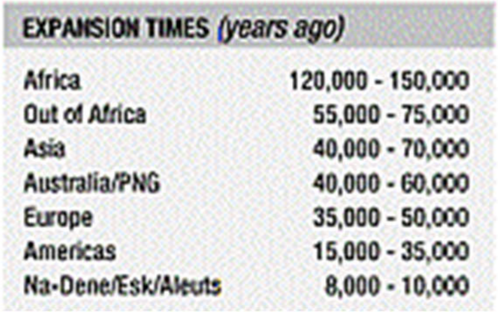
“Mitochondrial Eve” lived in Makgadikgadi–Okavango paleo-wetland of southern Africa ~200,000 years ago (between 165,000 and 240,000 years ago)

Chan EKF, et al. Nature. 2019; 575: 185–189.

Okavango Delta now

Modern mitochondrial DNA contains history of human migrations





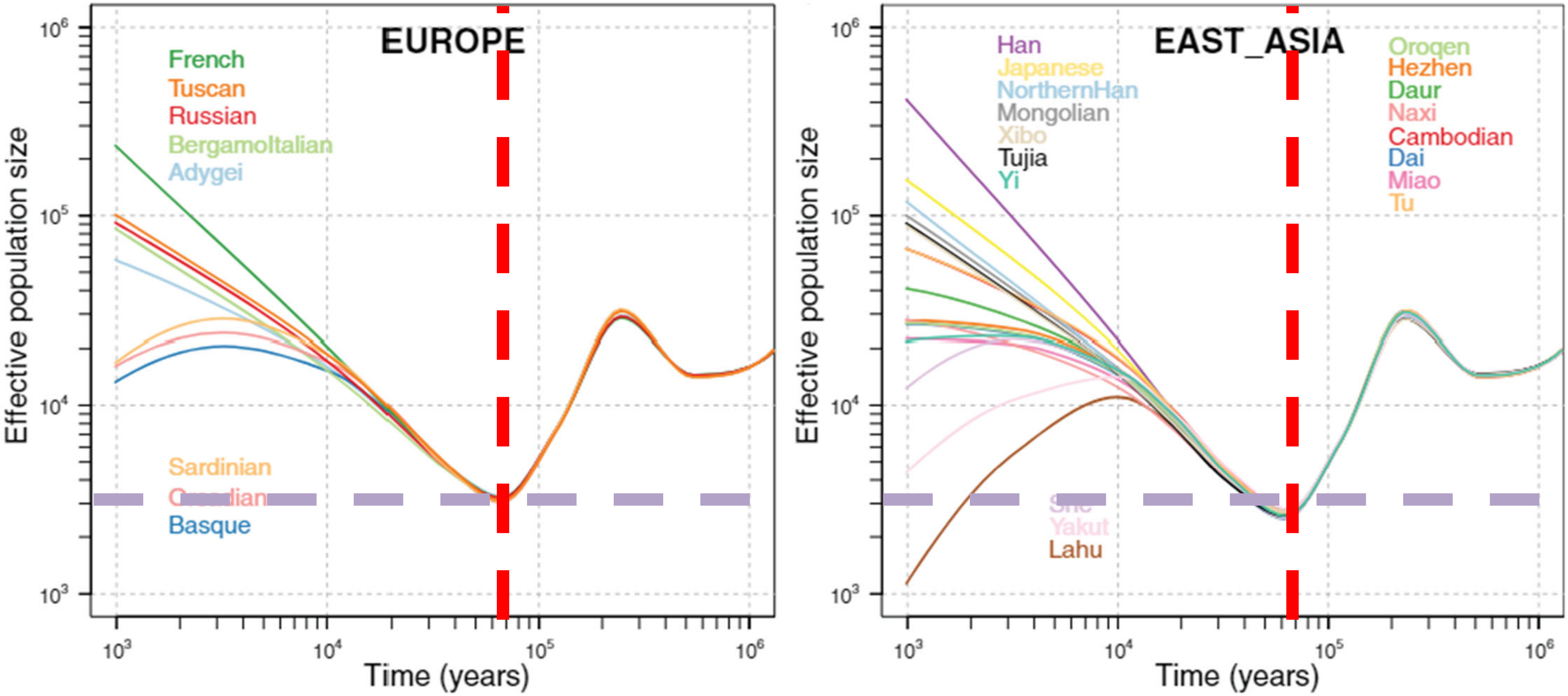
EXPANSION TIMES *(years ago)*	
Africa	120,000 - 150,000
Out of Africa	55,000 - 75,000
Asia	40,000 - 70,000
Australia/PNG	40,000 - 60,000
Europe	35,000 - 50,000
Americas	15,000 - 35,000
Na-Dene/Esk/Aleuts	8,000 - 10,000

FamilyTreeDNA
mtDNA Migrations Map

Poznik GD, et al (Carlos Bustamante lab in Stanford), Science ***341****: 562 (August 2013).*

- Bottleneck for human population in Europe and Asia was around 3,000 individuals reached around 70,000 years ago

- Non-African Eve lived ~60,000 years ago



From ~1000 modern genomes: Bergström A, et al. Science. 2020;367

What about men?

- Y-chromosome is transferred from father to son
- Like mitochondria it can be used to trace ancestry of all men to the “Y-chromosome Adam”
- Where did “Adam’” live? Did he meet the “mitochondrial Eve”?

Y-chromosomal Adam also lived in Africa

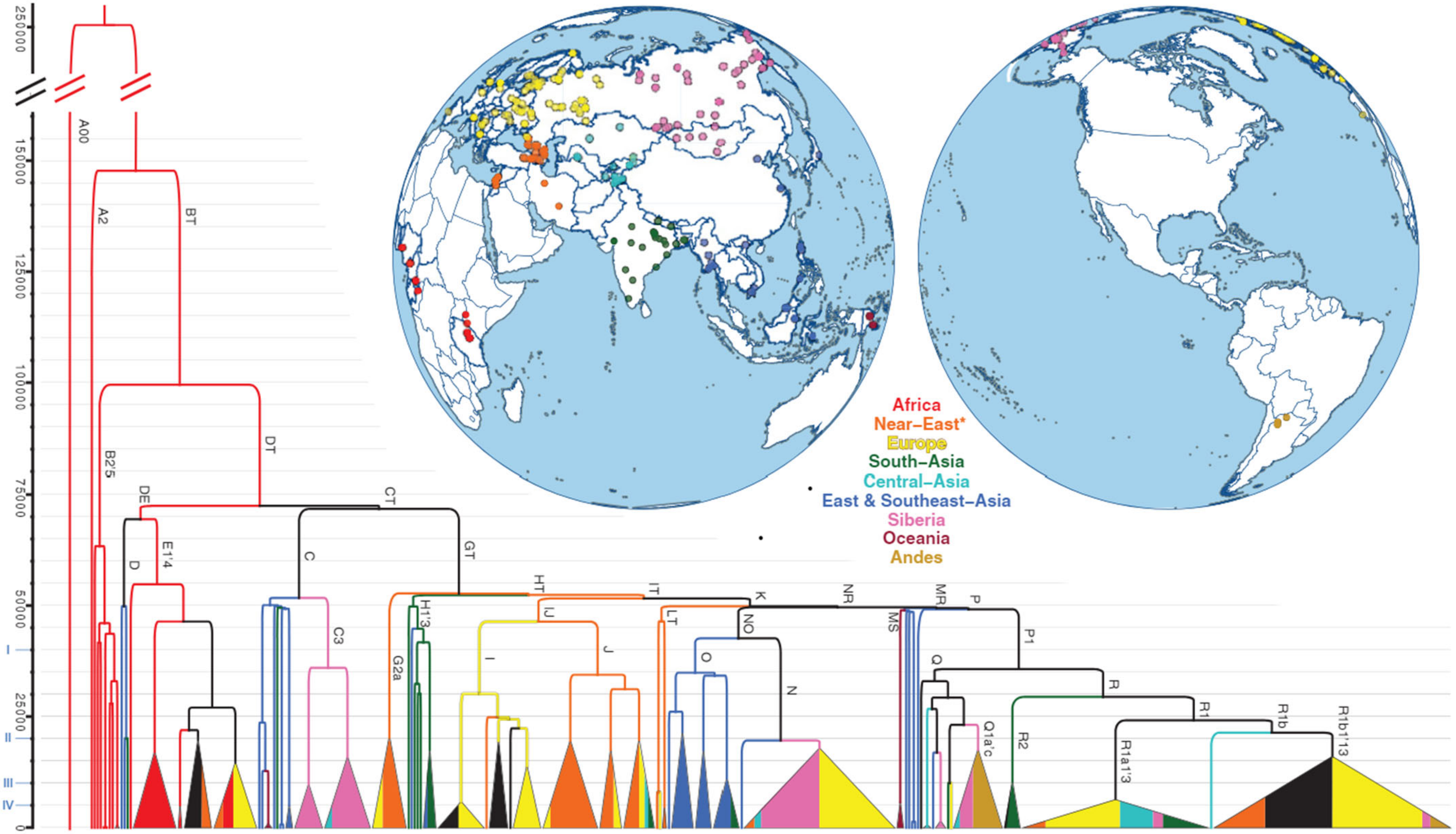
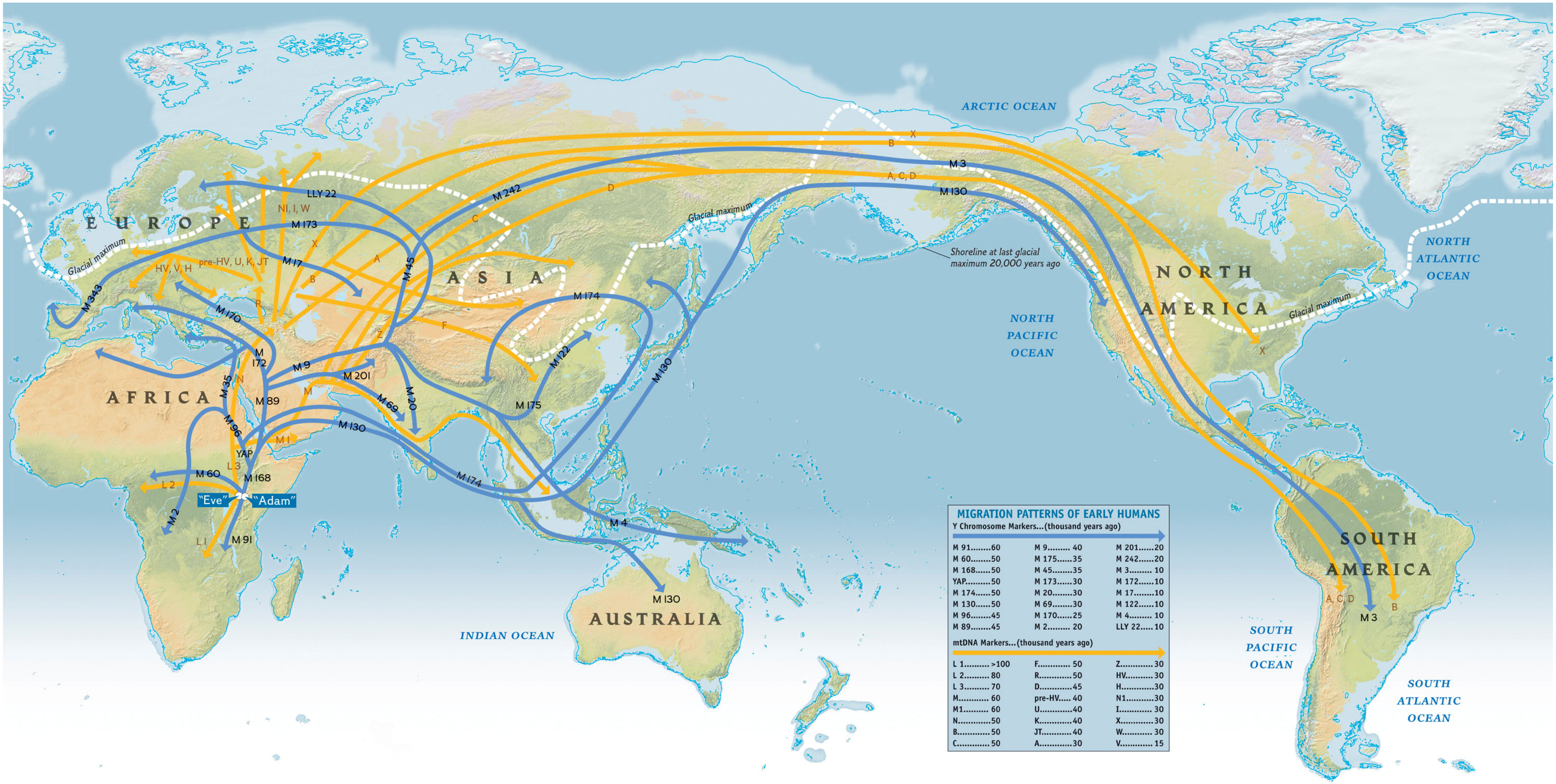


Figure 1. The phylogenetic tree of 456 whole Y chromosome sequences and a map of sampling locations. The phylogenetic tree is reconstructed using BEAST. Clades coalescing within 10% of the overall depth of the tree have been collapsed. Only main haplogroup labels are shown (details are provided in Supplemental Information 6). Colors indicate geographic origin of samples (Supplemental Table S1), and fill proportions of the collapsed clades represent the proportion of samples from a given region. Asterisk (*) marks the inclusion of samples from Caucasus area. Personal Genomes Project (http://www.personalgenomes.org) samples of unknown and mixed geographic/ethnic origin are shown in black. The proposed structure of Y chromosome haplogroup naming (Supplemental Table S5) is given in Roman numbers on the *y*-axis.

Karmin M, Saag L, Vicente M, Sayres MAW, Järve M, Talas UG, et al. Genome Res. 2015;25: 459–466.

“Adam” and “Eve” both lived in Africa



- “Mitochondrial Eve” lived in Africa between 100,000 and 240,000 years ago
- “Y-chromosome Adam” also lived in Africa between 120,000 and 160,000 years ago
- *Poznik GD, et al (Carlos Bustamante lab in Stanford), Science **341**: 562 (August 2013).*

Mitochondrial Eve (maternally transmitted ancestry)
Y-chromosome Adam (paternally transmitted ancestry)
lived ~200,000 years ago.

When lived the latest common ancestor shared by all of us based on nuclear DNA?

A. 1 million years ago
B. 200,000 years ago
C. 3400 years ago
D. 1320 years ago
E. Yesterday, I really have no clue

Get your i-clickers

Mitochondrial Eve (maternally transmitted ancestry)
Y-chromosome Adam (paternally transmitted ancestry)
lived ~200,000 years ago.

When lived the latest common ancestor shared by all of us based on nuclear DNA?

A. 1 million years ago
B. 200,000 years ago
[✓ C. 3400 years ago]
D. 1320 years ago
E. Yesterday, I really have no clue

Get your i-clickers

Last common ancestor in nuclear (non Y-chr) DNA is another matter

- Unlike Mito or Y-chromosome, **nuclear DNA gets mixed with every generation**
  - Each of us gets 1/2 of nuclear DNA from the father
    and 1/2 from the mother
  - Each of us has 2 parents, 4 grandparents, 8 great-grand parents …
- If one assumes:
  - Well-mixed marriages (not true: mostly local marriages)
  - Constant size population (not true: much smaller in the past)
  - In 33 generations the number of ancestors:
    2^{33} =8 billion = 8 billion people living today
- Every pair of us living today should have at least one shared ancestor who lived
  - 33 generations * 20 years/generation=660 years ago ~1360 AD
- **Assuming $T_{MRCA} = 2\ T_{\text{average pairwise ancestor}}$ = 2· 660 years ago =700 AD**

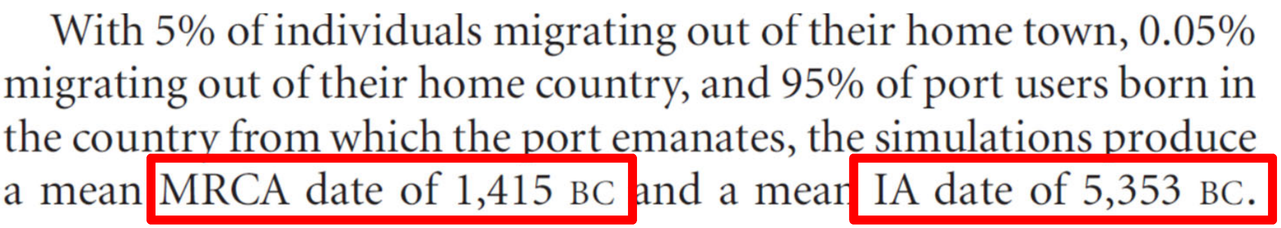
Corrected for (mostly) local marriages and rare migrations

562 NATURE | VOL 431 | 30 SEPTEMBER 2004 |

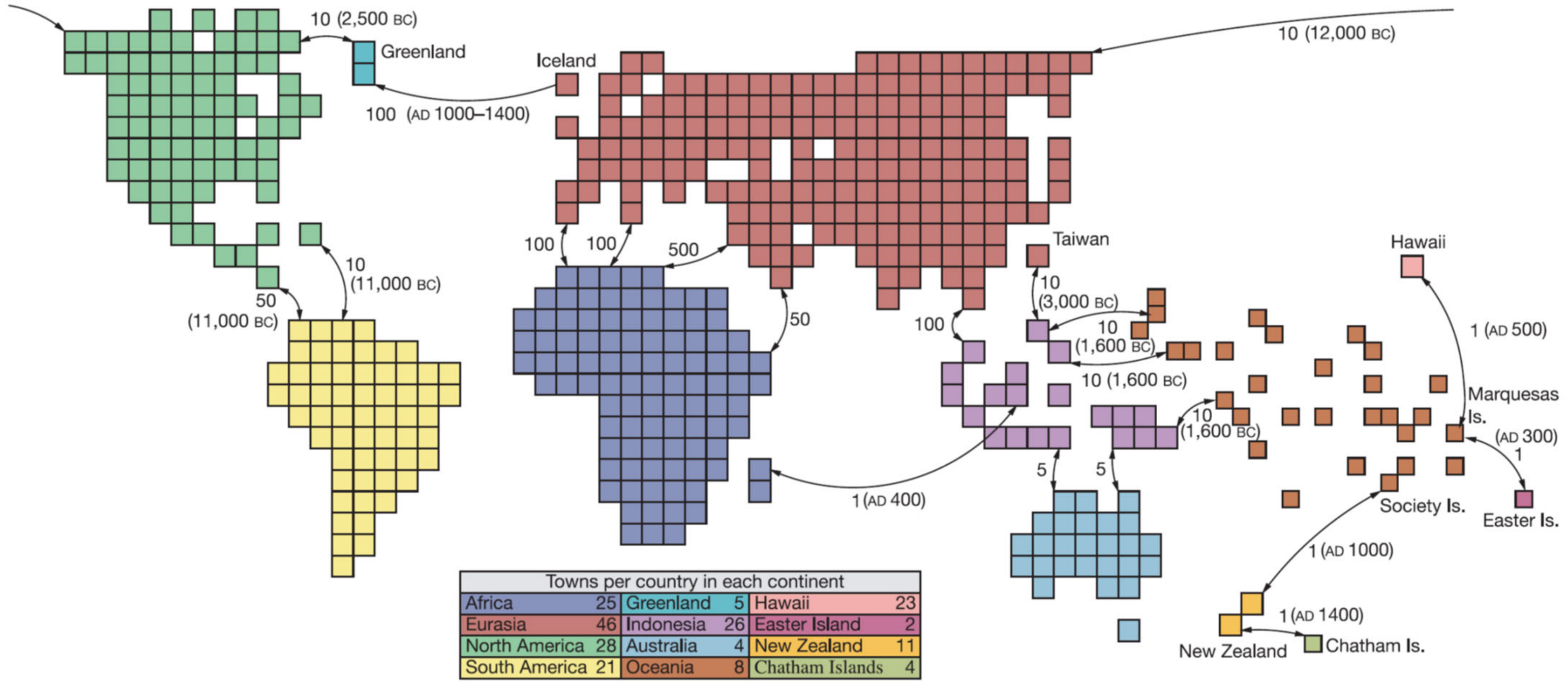
Modelling the recent common ancestry of all living humans

Douglas L. T. Rohde[1], **Steve Olson**[2] & **Joseph T. Chang**[3]

[1] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[2] *7609 Sebago Road, Bethesda, Maryland 20817, USA*
[3] *Department of Statistics, Yale University, New Haven, Connecticut 06520, USA*



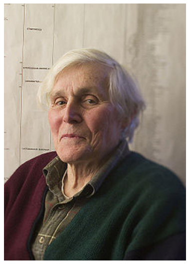
With 5% of individuals migrating out of their home town, 0.05% migrating out of their home country, and 95% of port users born in the country from which the port emanates, the simulations produce a mean MRCA date of 1,415 BC and a mean IA date of 5,353 BC.



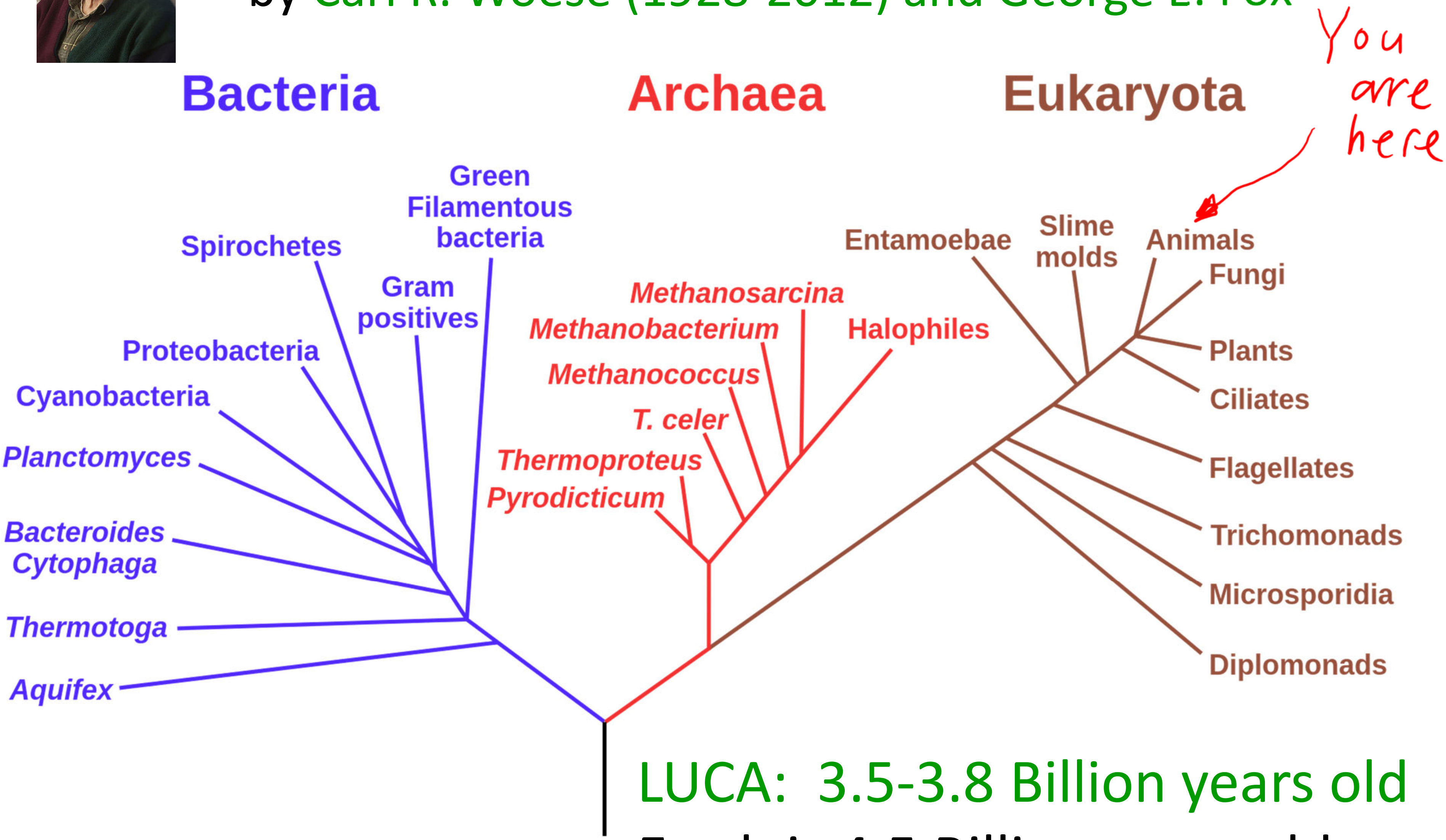
Towns per country in each continent					
Africa	25	Greenland	5	Hawaii	23
Eurasia	46	Indonesia	26	Easter Island	2
North America	28	Australia	4	New Zealand	11
South America	21	Oceania	8	Chatham Islands	4

Figure 2 Geography and migration routes of the simulated model. Arrows denote ports and the adjacent numbers are their steady migration rates, in individuals per generation. If given, the date in parentheses indicates when the port opens. Upon opening, there is usually a first-wave migration burst at a higher rate, lasting one generation.

Last Universal Common Ancestor (LUCA)

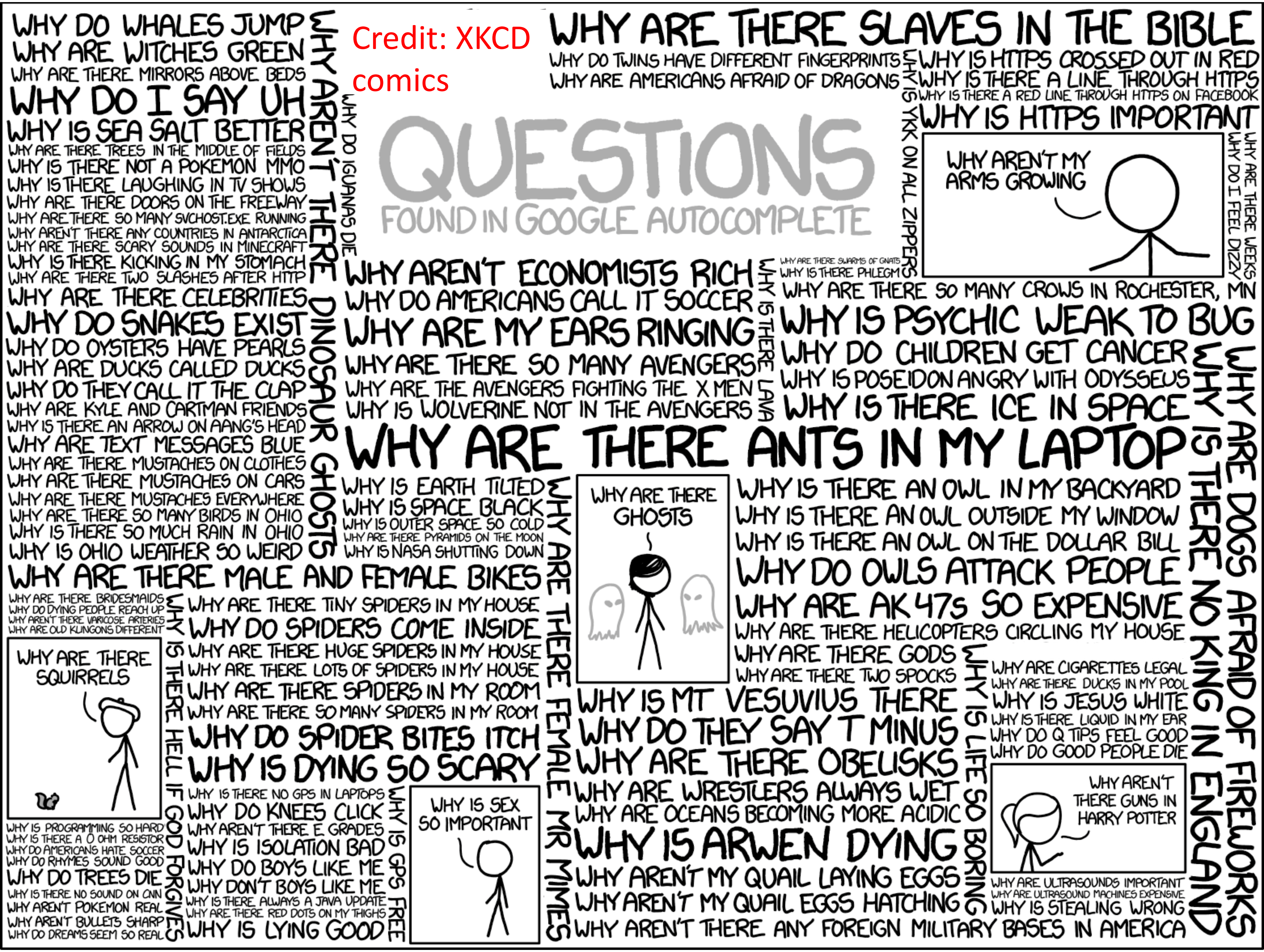


Archaea were discovered here at UIUC in 1977
by Carl R. Woese (1928-2012) and George E. Fox



LUCA: 3.5-3.8 Billion years old

Earth is 4.5 Billion years old



Credit: XKCD comics