Older and edited lecture slides are available for lectures 1-12

https://courses.grainger.illinois.edu/bioe310/sp2024/Lectures_1_to_12.pdf

Genome Assembly

Whole-genome "shotgun" sequencing starts by copying and fragmenting the DNA

("Shotgun" refers to the random fragmentation of the whole genome; like it was fired from a shotgun)

Input: GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
35bp

Copy by PCR:
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Fragment:
GGCGTCTA TATCTCGG CTCTAGGCCCTC ATTTTTT
GGC GTCTATAT CTCGGCTCTAGGCCCTCA TTTTTT
GGCGTC TATATCT CGGCTCTAGGCCCT CATTTTTT
GGCGTCTAT ATCTCGGCTCTAG GCCCTCA TTTTTT

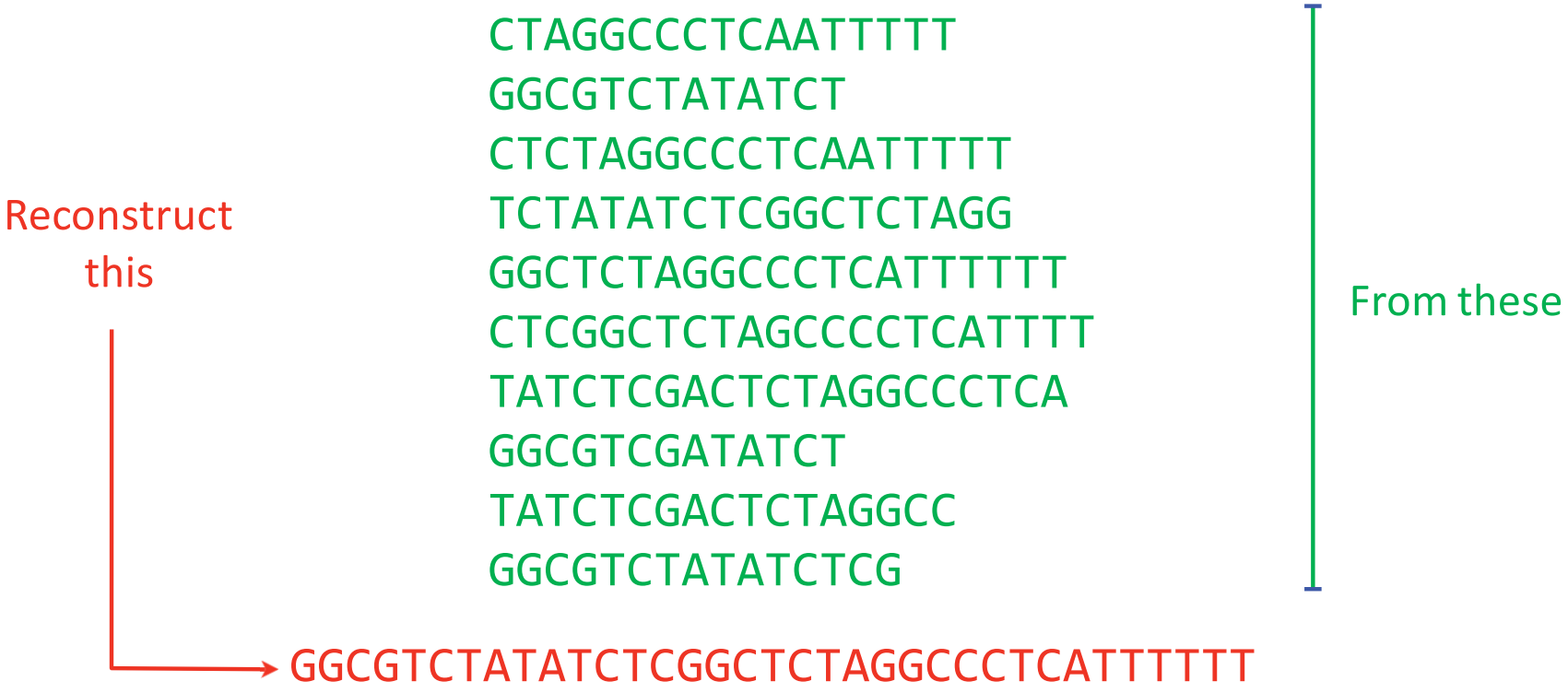
Courtesy of Ben Langmead. Used with permission.

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

Assembly

Assume sequencing produces such a large # fragments that almost all genome positions are *covered* by many fragments...

...but we don't know what came from where

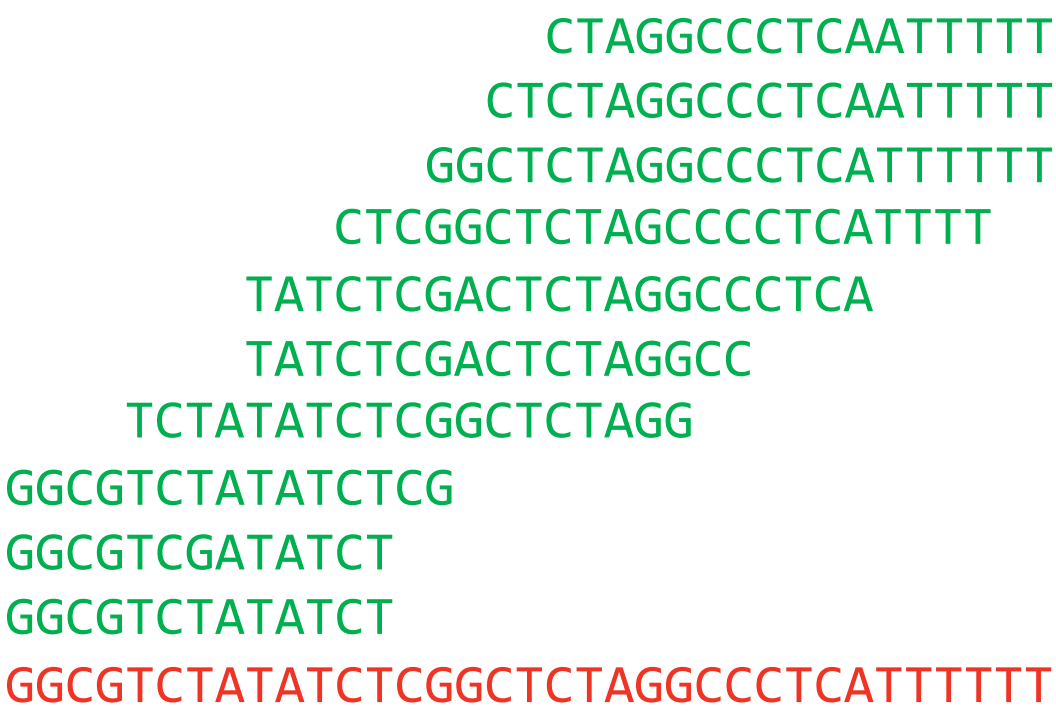


Courtesy of Ben Langmead. Used with permission.

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

Assembly

Overlaps between short reads help to put them together



Courtesy of Ben Langmead. Used with permission.

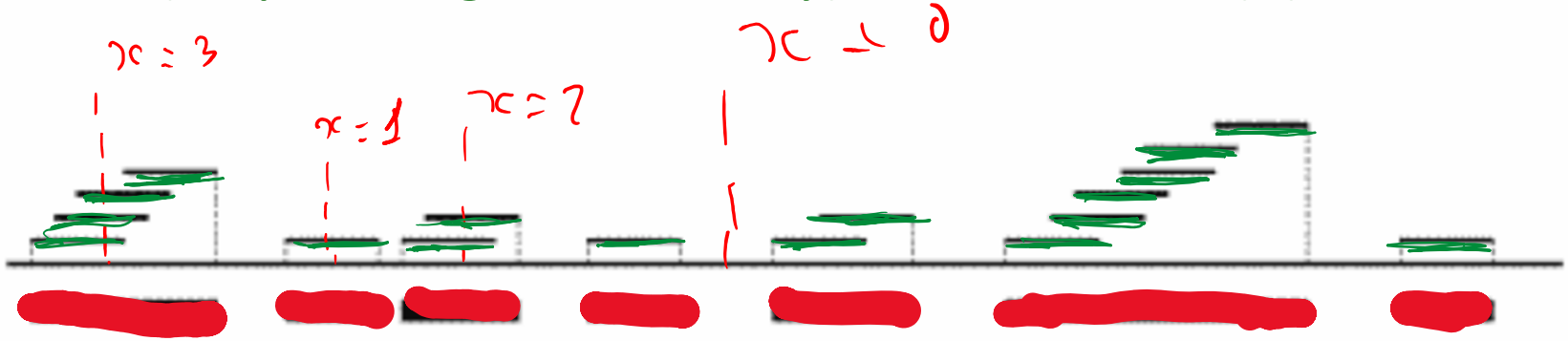
http://www.langmead-lab.org/teaching-materials/

Where is the Poisson?

- *G - genome length (in bp)*
- *L - short read average length*
- *N – number of short read sequenced*
- *λ – sequencing coverage redundancy = LN/G*
- *x- number of short reads covering a given site on the genome*

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Poisson as a limit of Binomial: For a given site on the genome for each short read Prob(site covered): p=L/G is very small. Number of attempts (short reads): N is very large. Their product (sequencing redundancy): λ = NL/G is O(1).



Ewens, Grant, Chapter 5.1

Estimate

- Human genome is $3x10^9$ bp long
- Chromosome 1 is about $G=0.25x10^9$ bp
- Illumina generates short reads L=100 bp long
- What number of reads *N* are needed to completely assemble the 1st chromosome*?*
- The formula to use is: $1=N_{contigs}=Ne^{-\lambda}=Ne^{-NL/G}$
- Answer: $N=4.4x10^7$ short (100bp) reads
  Test: 4.4e7*exp(-4.4e7*100/0.25e9)=0.9997
- What coverage redundancy λ will it be?
  Answer: $\lambda = NL/G = 17.6$ coverage redundancy

How much would it cost to assemble human genome now?

- Human Genome Project: $2.7 billion in 1991 dollars.
- Now a de novo full assembly of the whole human genome would now cost $3 \times 10^9 \times 17.6 / 10^6 \times 0.1\$/MB = \$ 5300$
- 2nd genome (and after) would be even cheaper as we would already have a reference genome to which we can map short reads. (Puzzle: picture on the box)
- But, this is a naïve estimate. In reality there are complications. See next slides:

What spoils these estimates?

>gi|224514922|ref|NT_024477.14| Homo sapiens chromosome 12 genomic contig, GRCh37.p13 Primary Assembly (displaying 3′ end)

CGGGAAATCAAAAGCCCCTCTGAATCCTGCGCACCGAGATTCTCCCCAGCCAAGGTGAGGCGGCAGCAGT
GGGAGATCCACACCGTAGCATTGGAACACAAATGCAGCATTACAAATGCAGACATGACACCGAAAATATA
ACACACCCCATTGCTCATGTAACAAGCACCTGTAATGCTAATGCACTGCCTCAAAACAAAATATTAATAT
AAGATCGGCAATCCGCACACTGCCGTGCAGTGCTAAGACAGCAATGAAAATAGTCAACATAATAACCCTA
ATAGTGTTAGGGTTAGGGTCAGGGTCCCGGTCCGGGTCGGGGTCCGGGTCCGGGGTCCGGGTCAGGGTGA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAG

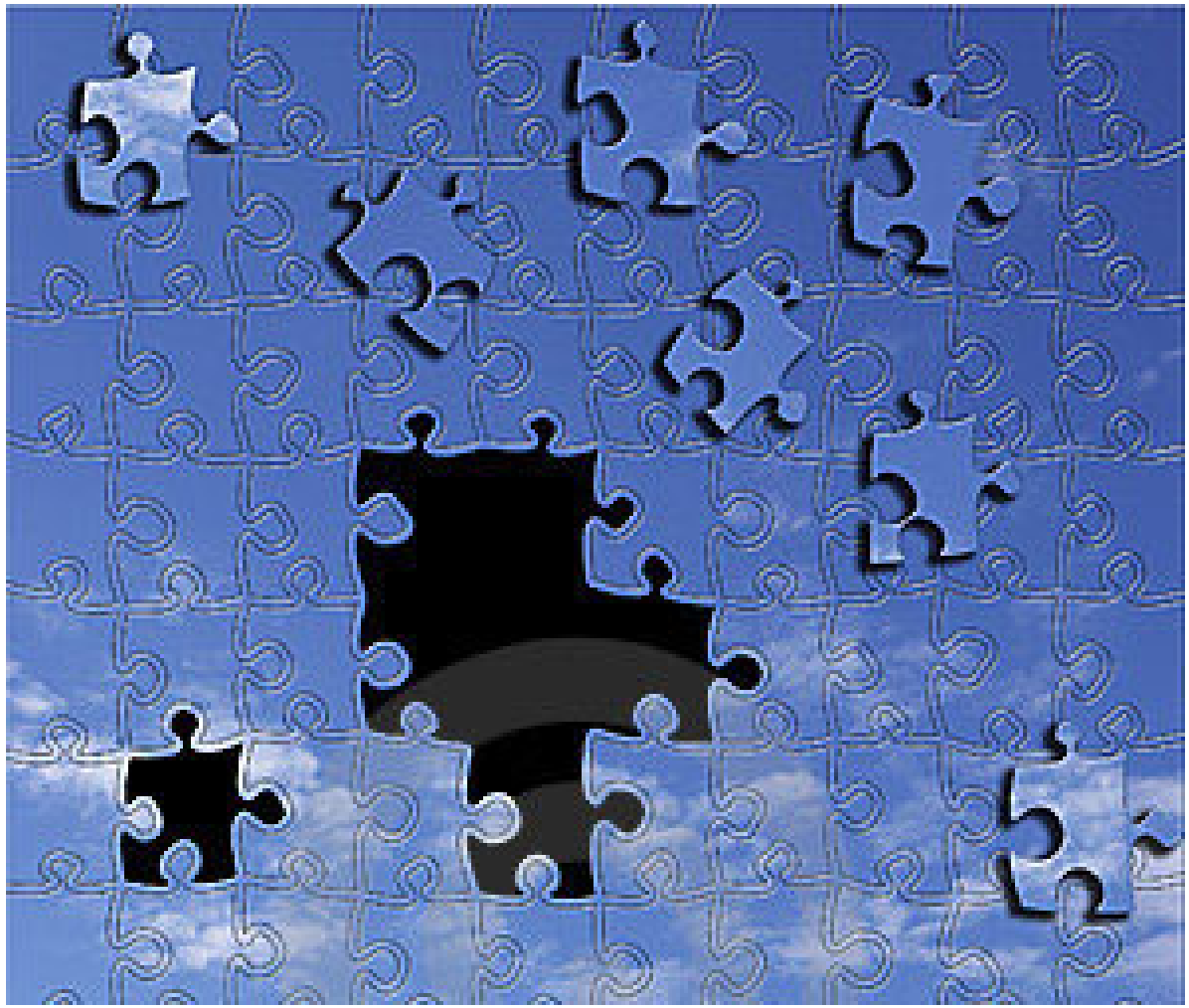
FIGURE 8.11 A BLASTN search of the human genome (all assemblies) database was performed at the NCBI website using TTAGGGTTAGGGTTAGGG as query (i.e., three TTAGGG repeats). There were matches to hundreds of genomic scaffolds. This figure shows an example (NT_024477.14) assigned to the telomere of chromosome 12q having many dozens of TTAGGG repeats. These occurred at the 3′ end of the genomic contig sequence.

There were 100s of matches while one expects << 1 match:

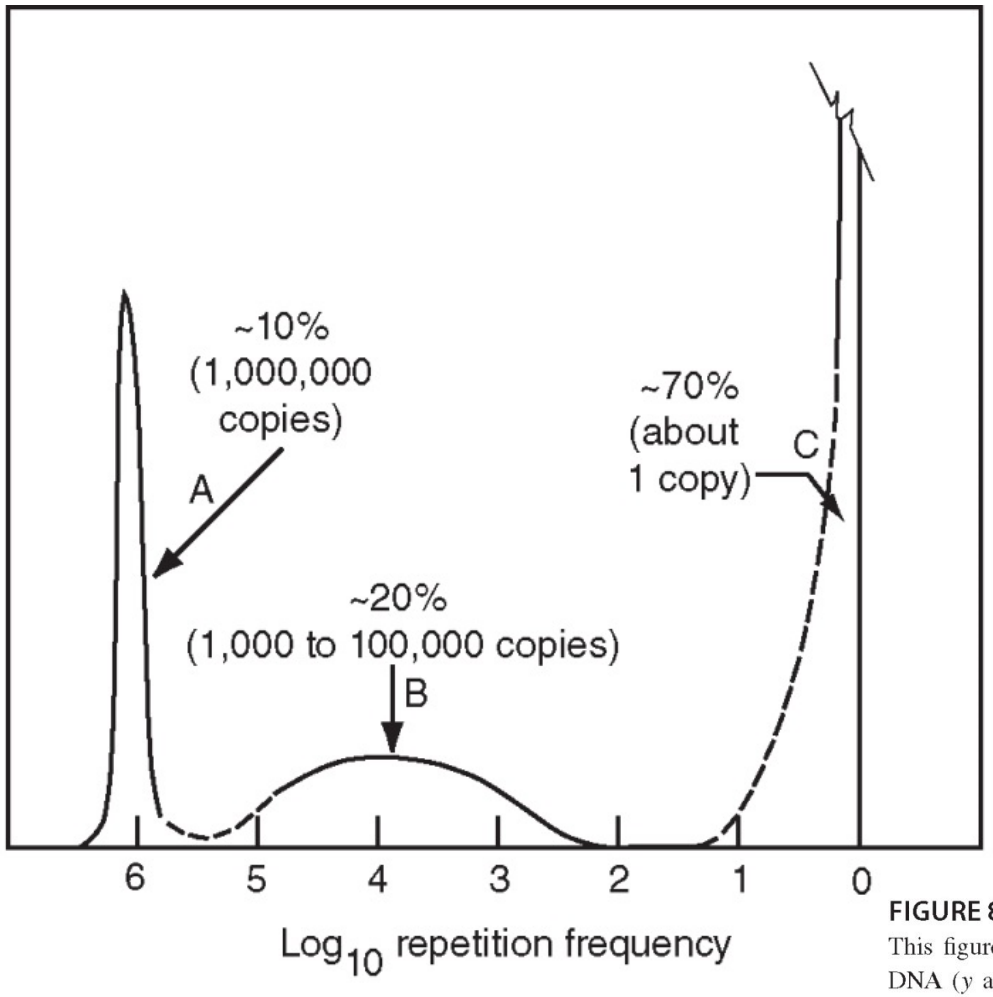
$$2 \cdot 3x10^{9} \cdot 4^{-18}=0.08<<1$$

DNA repeats make assembly difficult

Repeats are like sky puzzle pieces

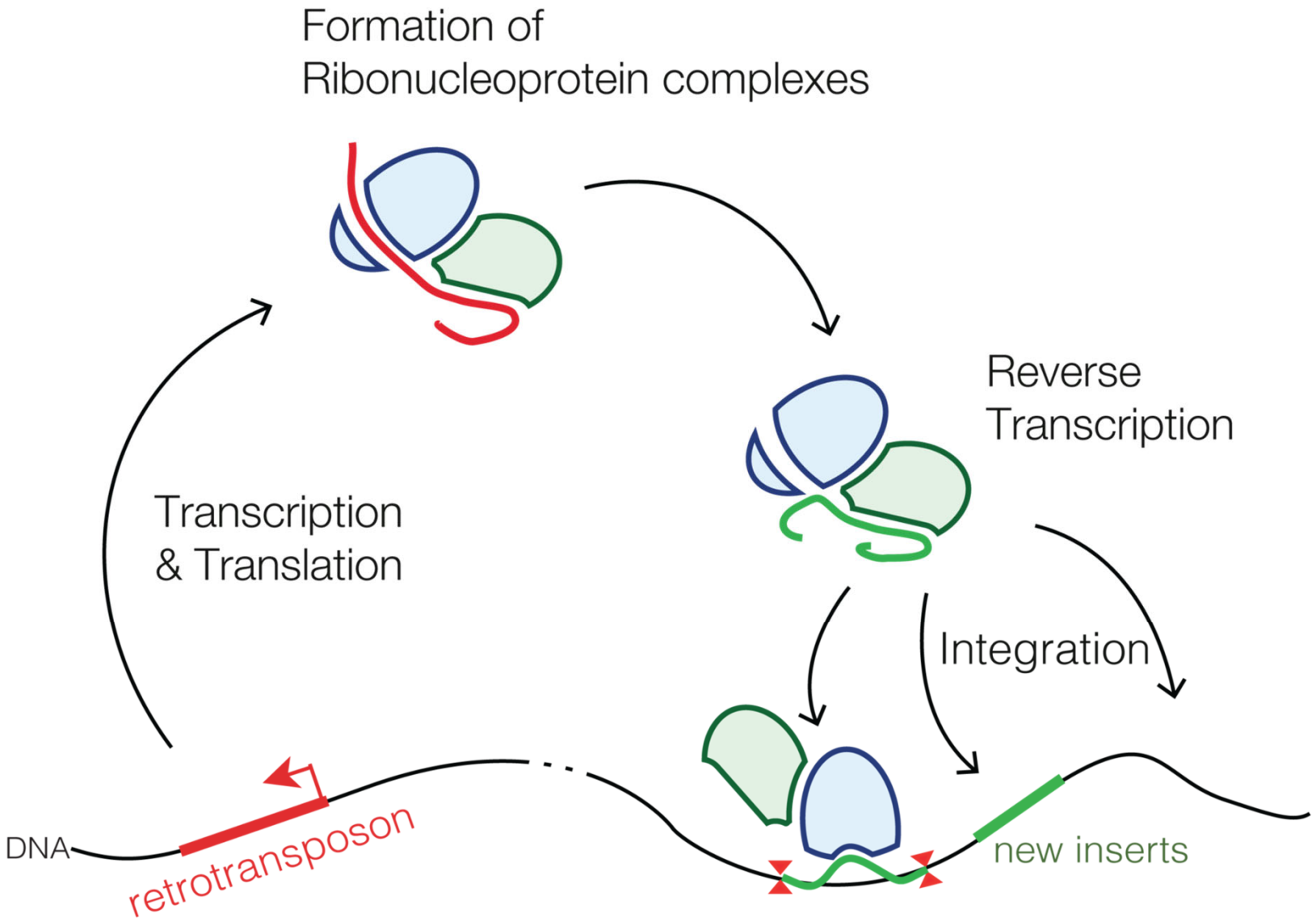


How many repeats are in eukaryotic genomes?

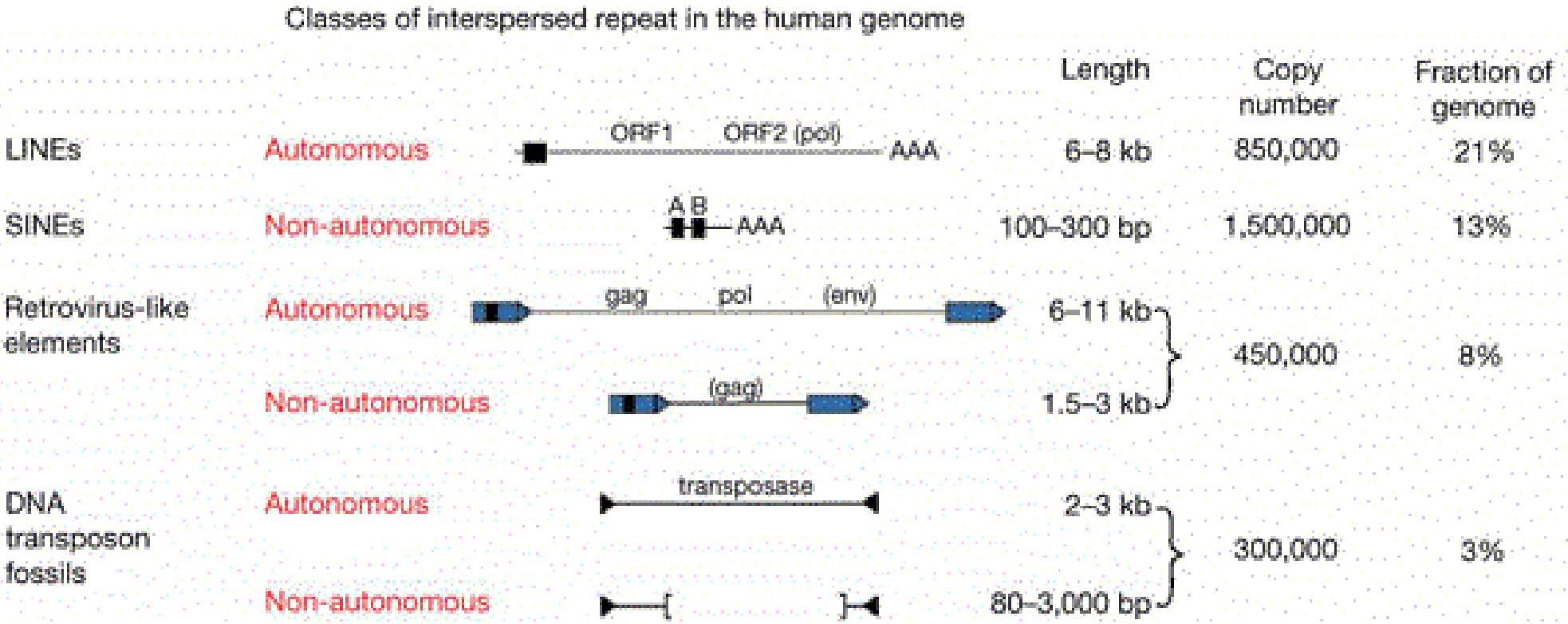


Data for mouse genome obtained in 1961 (sic!) using DNA denaturation and renaturation curves

FIGURE 8.6 The complexity of genomic DNA can be estimated by denaturing then renaturing DNA. This figure (redrawn from Britten and Kohne, 1968) depicts the relative quantity of mouse genomic DNA (y axis) versus the logarithm of the frequency with which the DNA is repeated. The data are derived from a $C_0\,t_{1/2}$ curve, which describes the percent of genomic DNA that reassociates at particular times and DNA concentrations. A large $C_0\,t_{1/2}$ value implies a slower reassociation reaction. Three classes are apparent. The fast component accounts for 10% of mouse genomic DNA (arrow **A**), and represents highly repetitive satellite DNA. An intermediate component accounts for about 20% of mouse genomic DNA and contains repeats having from 1000 to 100,000 copies. The slowly reassociating component, comprising 70% of the mouse genome, corresponds to unique, single-copy DNA. Britten and Kohne (1968) obtained similar profiles from other eukaryotes, although distinct differences were evident between species. Used with permission.


Formation of
Ribonucleoprotein complexes
Reverse
Transcription
Integration
new inserts
Transcription
& Translation
DNA
retrotransposon

Almost all transposable elements in mammals fall into one of four classes



Slide by Ross Hardison, Penn State U.

How to assemble a real genome with repeats?

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

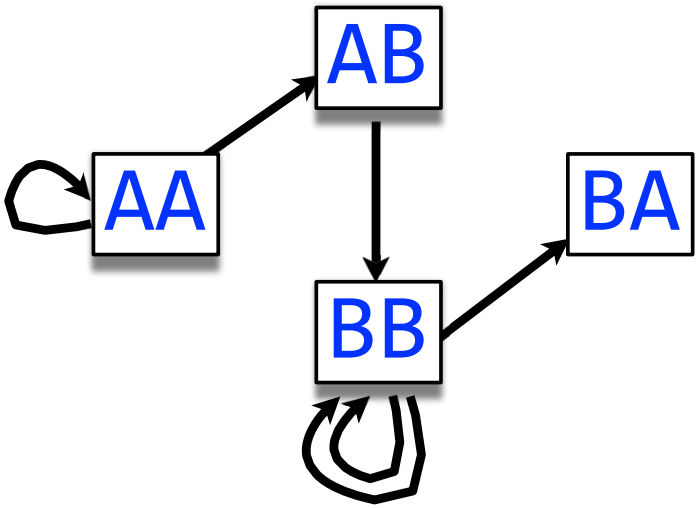
Nicolaas Govert de Bruijn (1918 – 2012) was a Dutch mathematician, noted for his many contributions in the fields of graph theory, analysis, number theory, combinatorics and logic

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



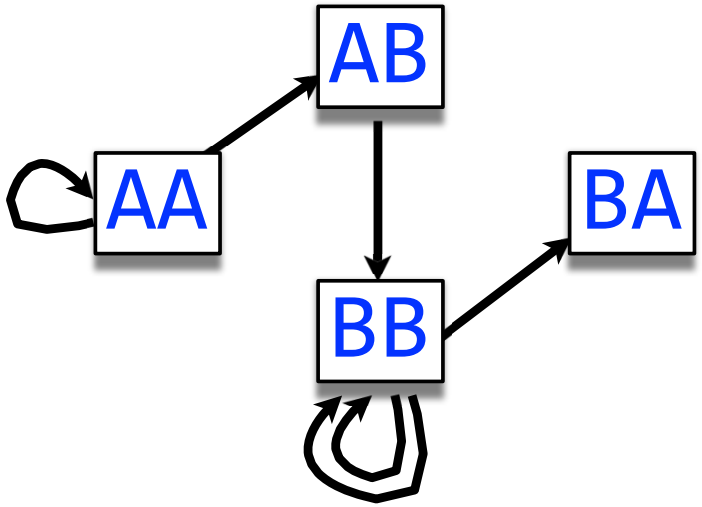
One edge per every *k*-mer

One node per distinct *k*-1-mer

Courtesy of Ben Langmead. Used with permission.

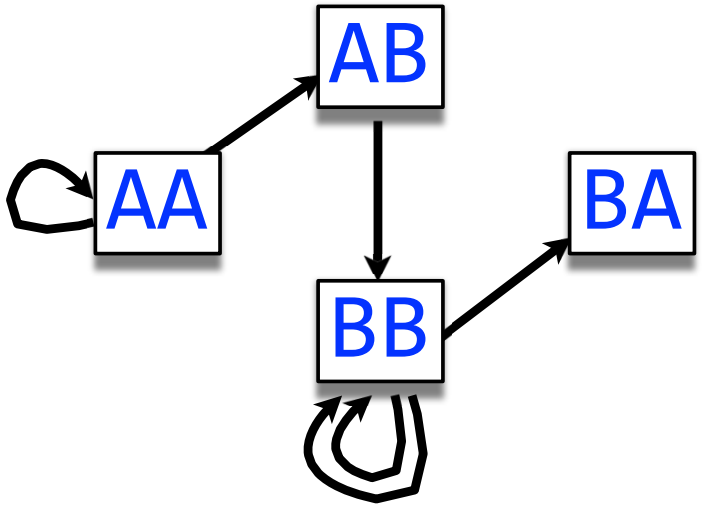
http://www.langmead-lab.org/teaching-materials/

De Bruijn graph



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

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*.

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

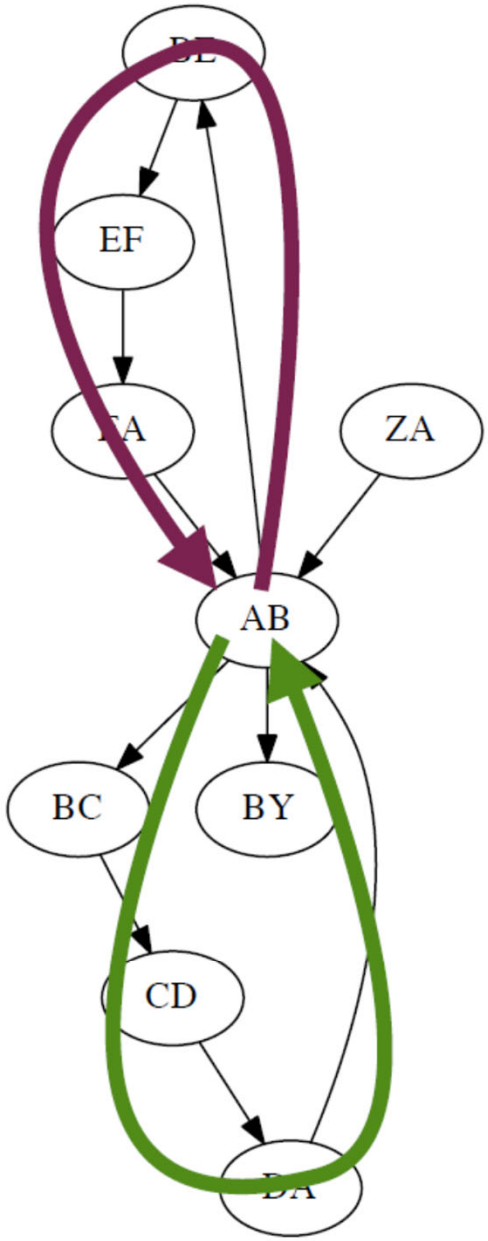
Alternative Eulerian walks:

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

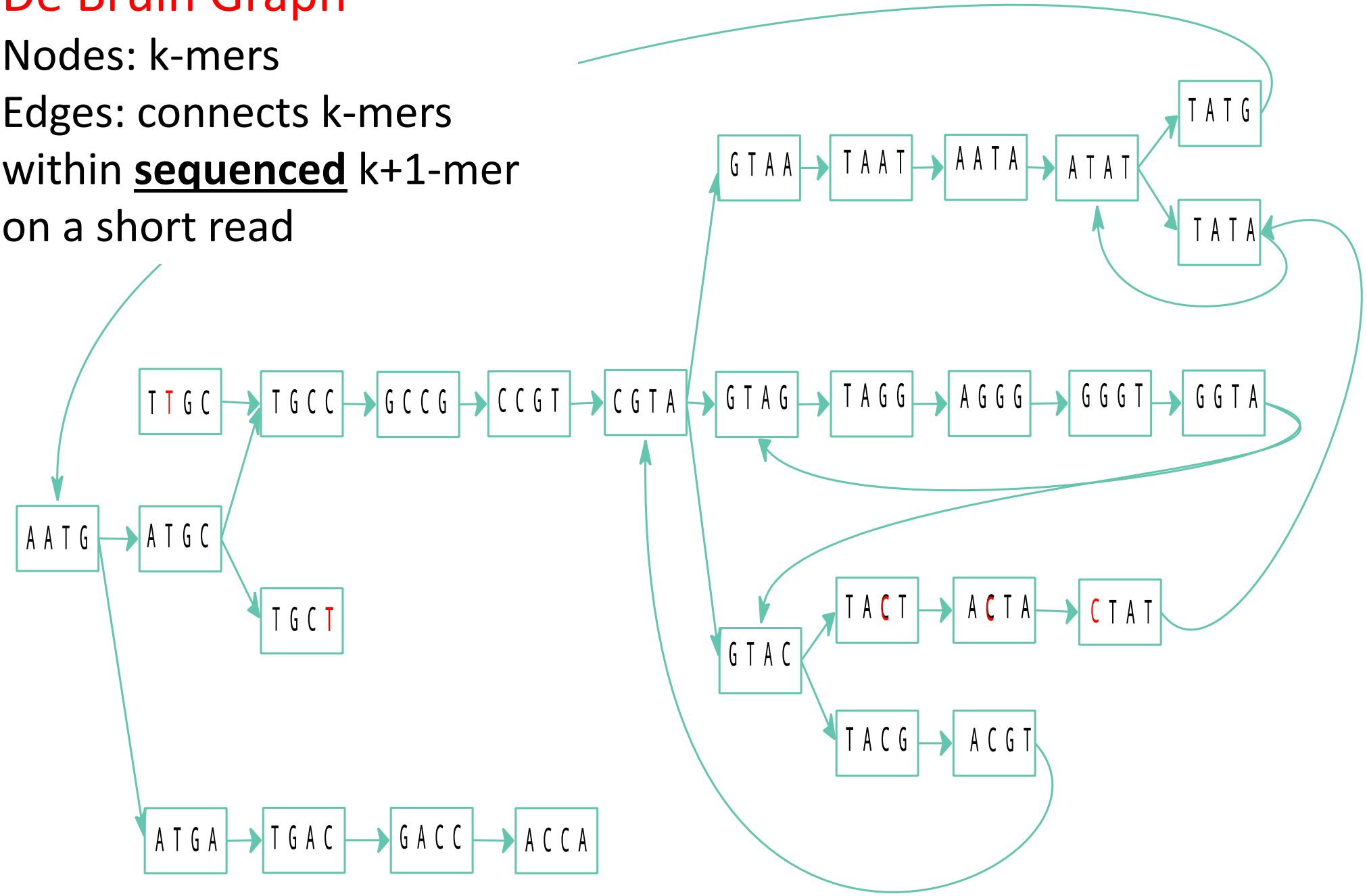


Adapted from a slide by Ben Langmead, Johns Hopkins U.

De Bruin Graph

Nodes: k-mers

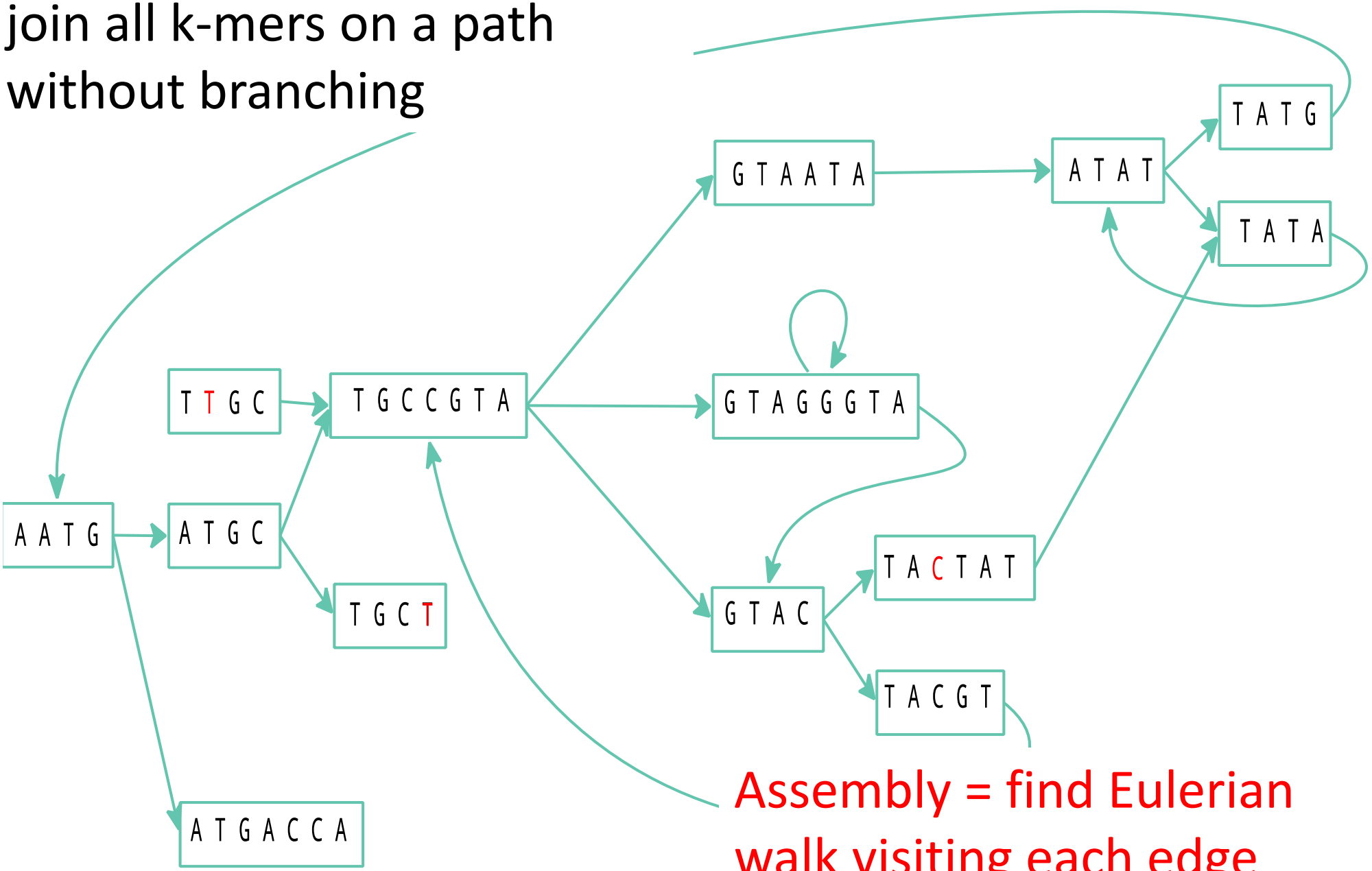
Edges: connects k-mers within **sequenced** k+1-mer on a short read



Slide by Sorin Istrail, Brown U.

Simplified De Bruin Graph

join all k-mers on a path without branching



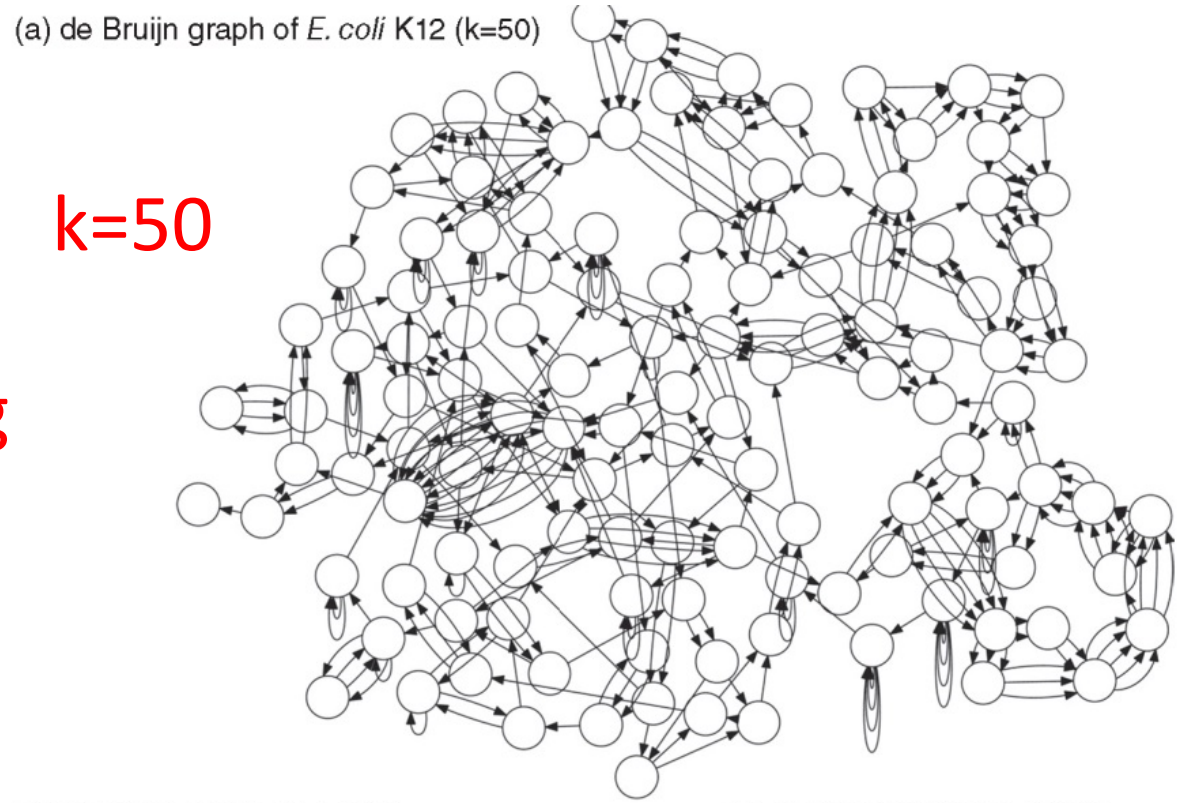
Assembly = find Eulerian walk visiting each edge once

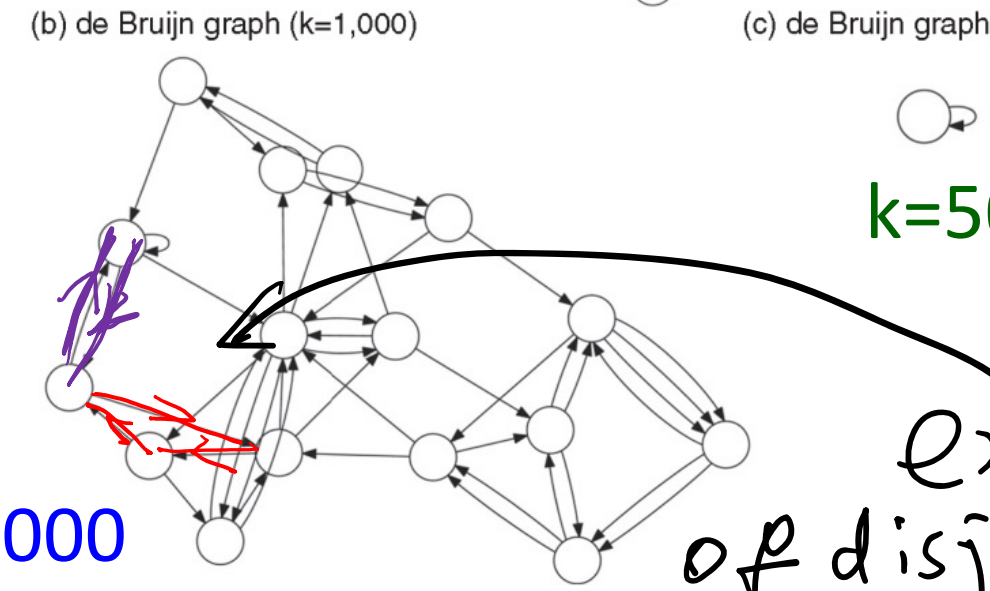
Slide by Sorin Istrail, Brown U.

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	200
Pacific Biosciences	2900
Sanger	400–900



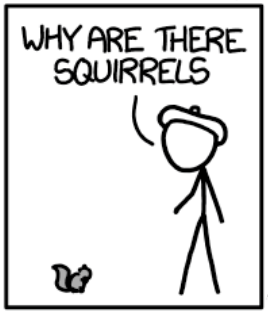


Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY ARE THERE SQUIRRELS



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE

WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD

WHY IS SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT

WHY AREN'T MY ARMS GROWING

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA

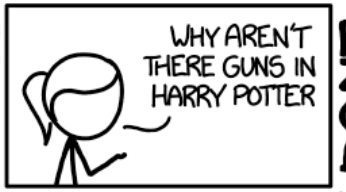
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

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, … , 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*,

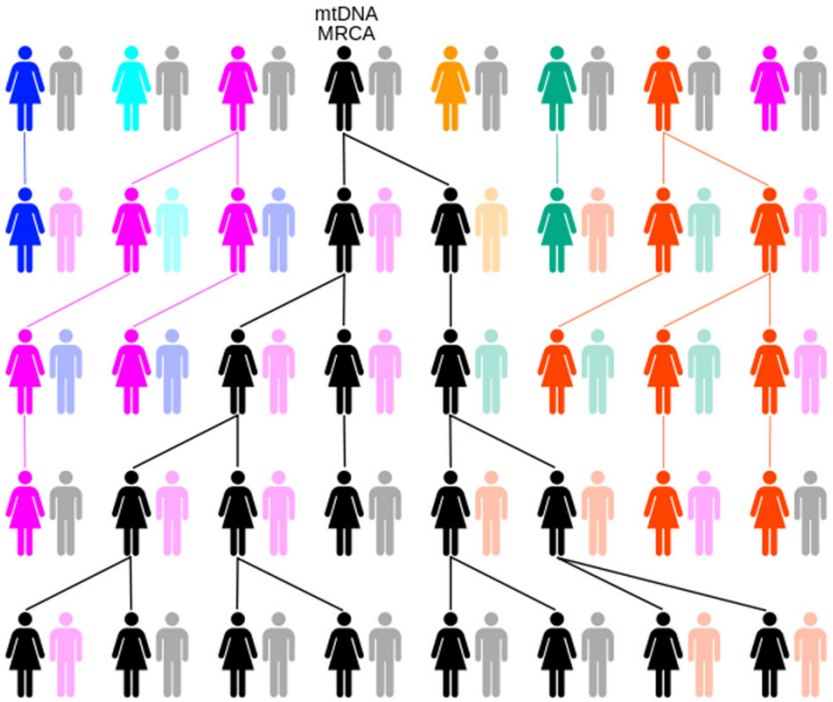
$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \qquad (3\text{-}10)$$

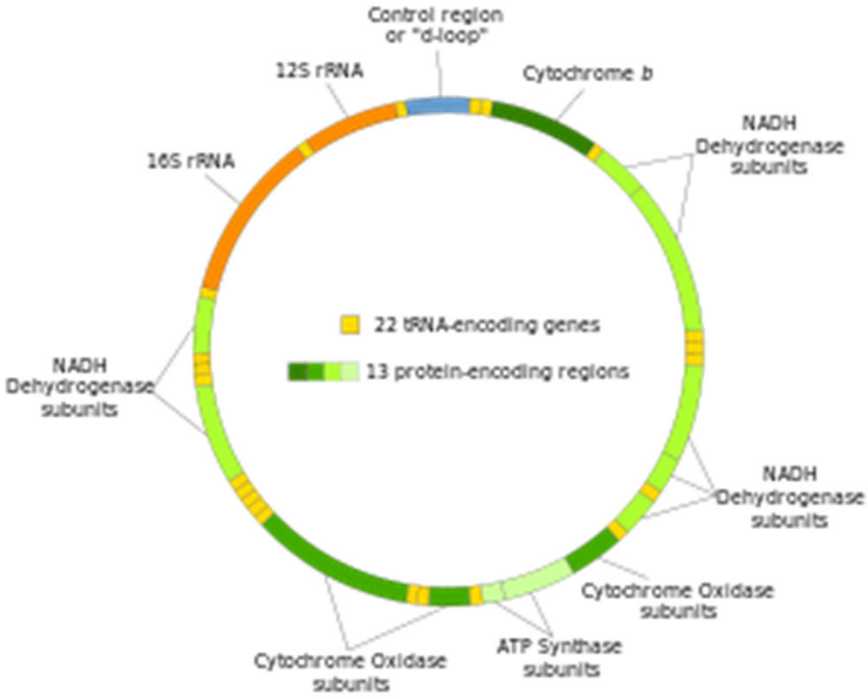
- For small *p* the standard deviation $=(1\text{-}p)^{0.5}/p$ ~= mean=1/p

- Very different from Binomial and Poisson, where

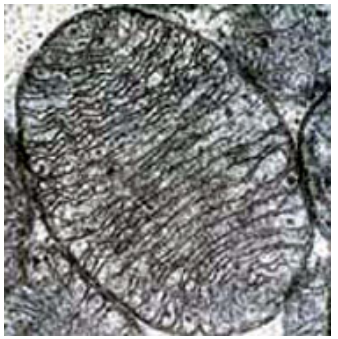
variance = mean and standard deviation = $\text{mean}^{1/2}$

Geometric distribution in biology



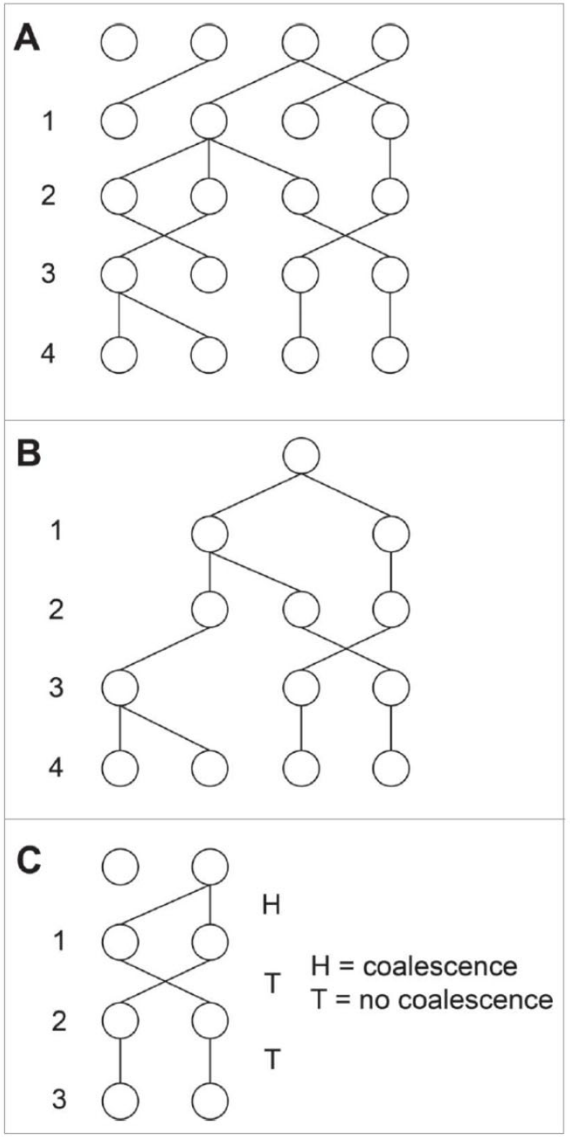


- Each of our cells has mitochondria with 16.5kb of mtDNA **inherited only from our mother**
- Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA
- Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeaon (of UIUC's Carl 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

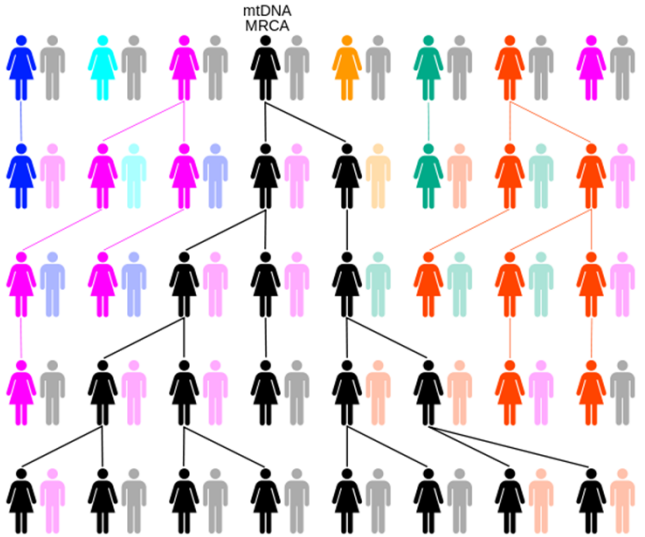
- Constant population of N women
- Random number of (female) offsprings. Average is 1 (but can be 0 or 2)
- Randomly pick two women.
  Question: how many generations T since their last maternal ancestor?
- T is a random variable What is its PMF: P(T=t)?
  Answer: P(T=t) follows a geometric distribution
- 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)

- 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 $N \to N-1$,
- so time until the next
- merger is $\frac{2}{(N-1)(N-2)}$



Most Recent Common Ancestor (MRCA)

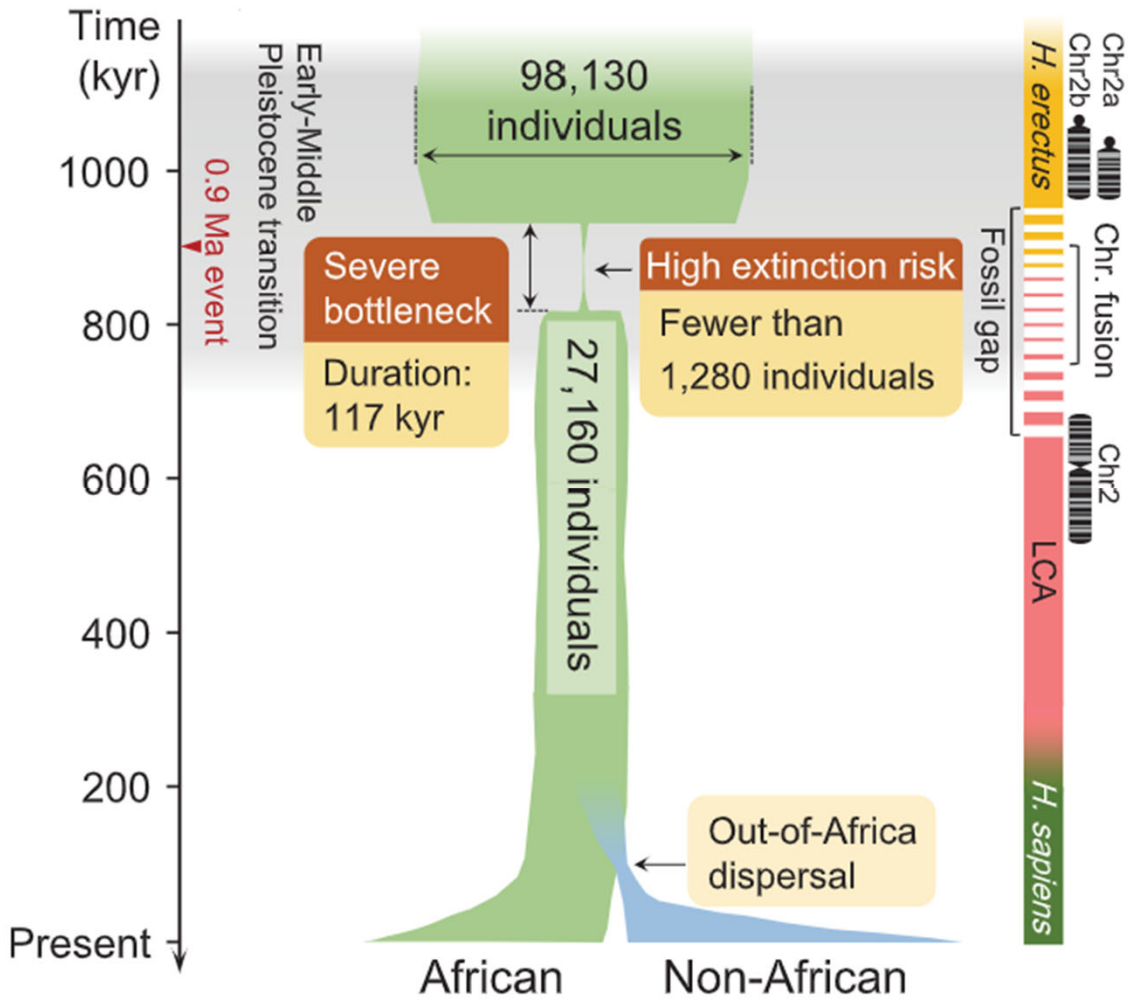
Total time until the MRCA

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

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

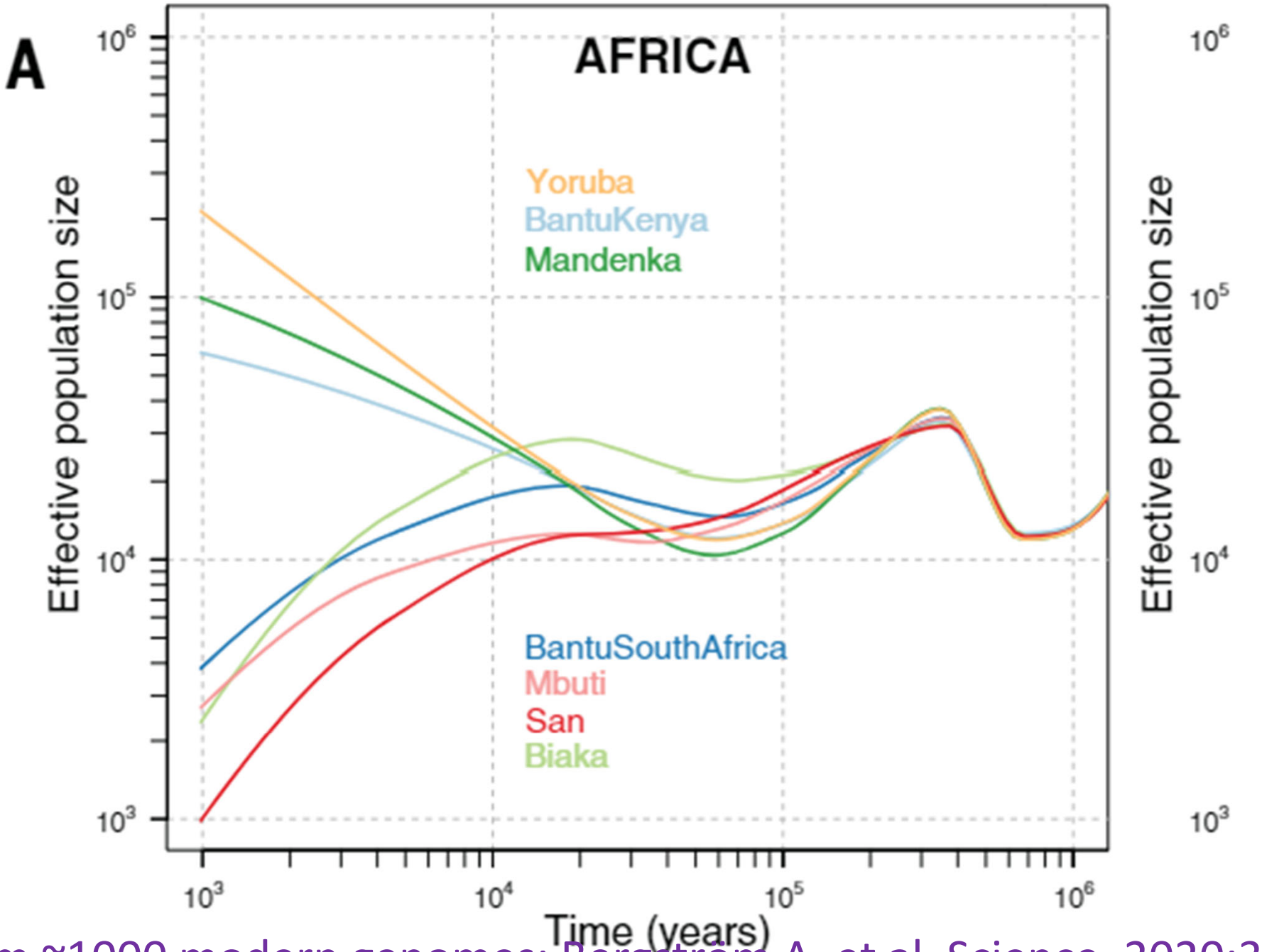
- There are about $N=3.5x10^{9}$ women living today
- Most Recent maternal Common Ancestor (MRCA)
  of all people living today lived $T_{MRCA}=2N$ generations ago
- $T_{MRCA} = 2 \cdot 3.5x10^{9}$ generations
- If the generation time 20 years it is 140 billion years > 10 times the time since the Big Bang.
- Something is wrong here!

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



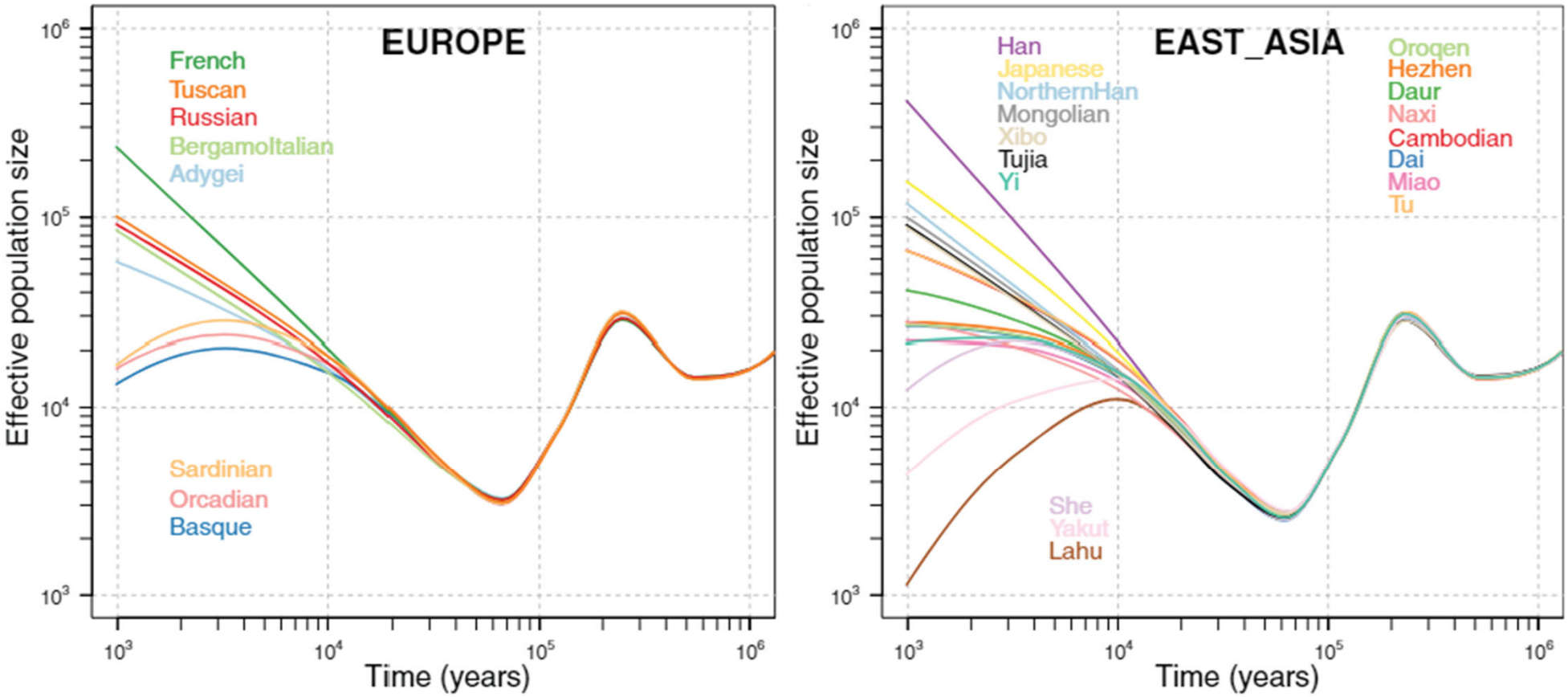
Hu W, et al. Science. 2023;381: 979–984

Effective human population size ~10,000



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

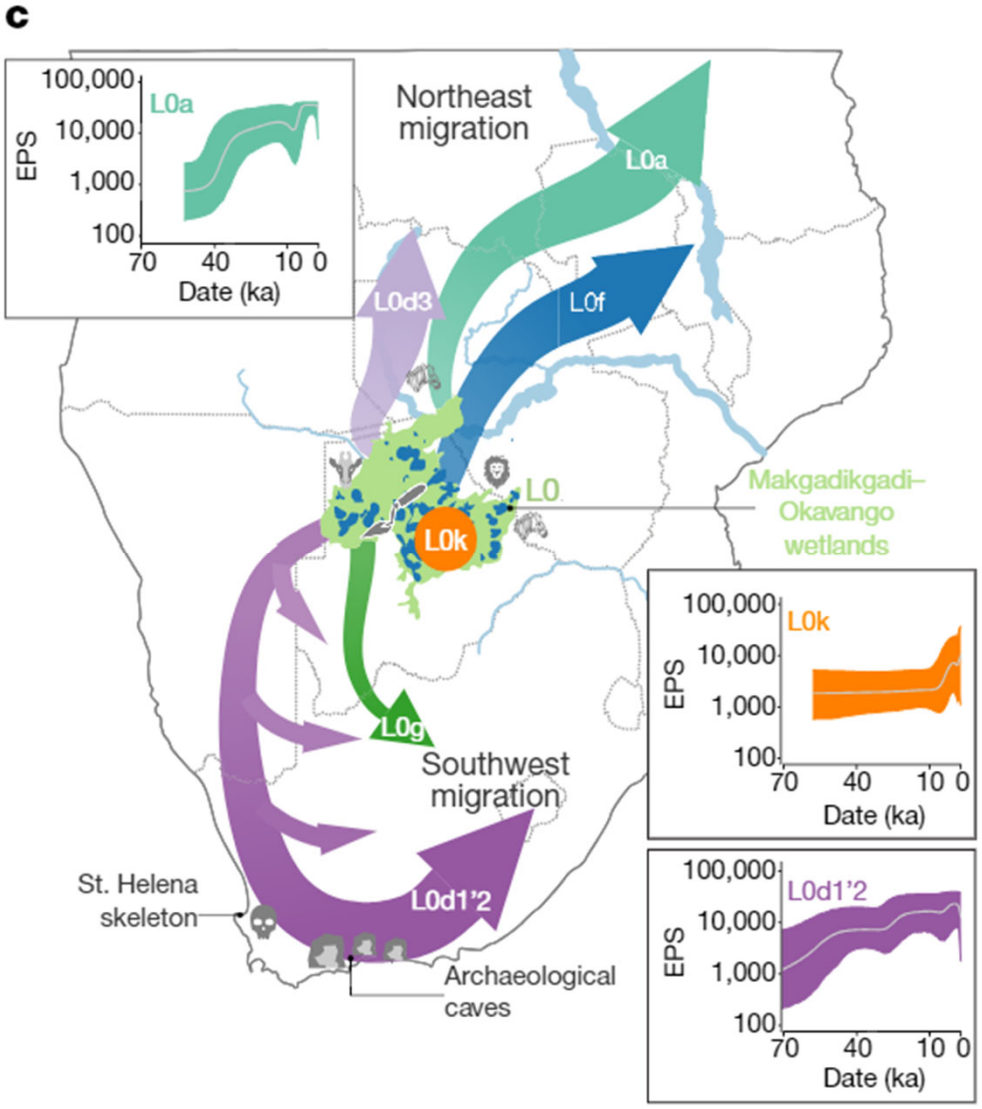
Effective human population size in Europe and Asia ~3000 people ~60,000 years ago



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
- Current thinking is that for all of us including people of African ancestry N_e~10,000 people
- For humans of European + Asian ancestry N_e~ 3000 people
- Mito Eve lived in Africa ~2*(Ne/2)*20 years=10,000*20 years= 200,000 years ago

“Mitochondrial Eve” lived in Africa

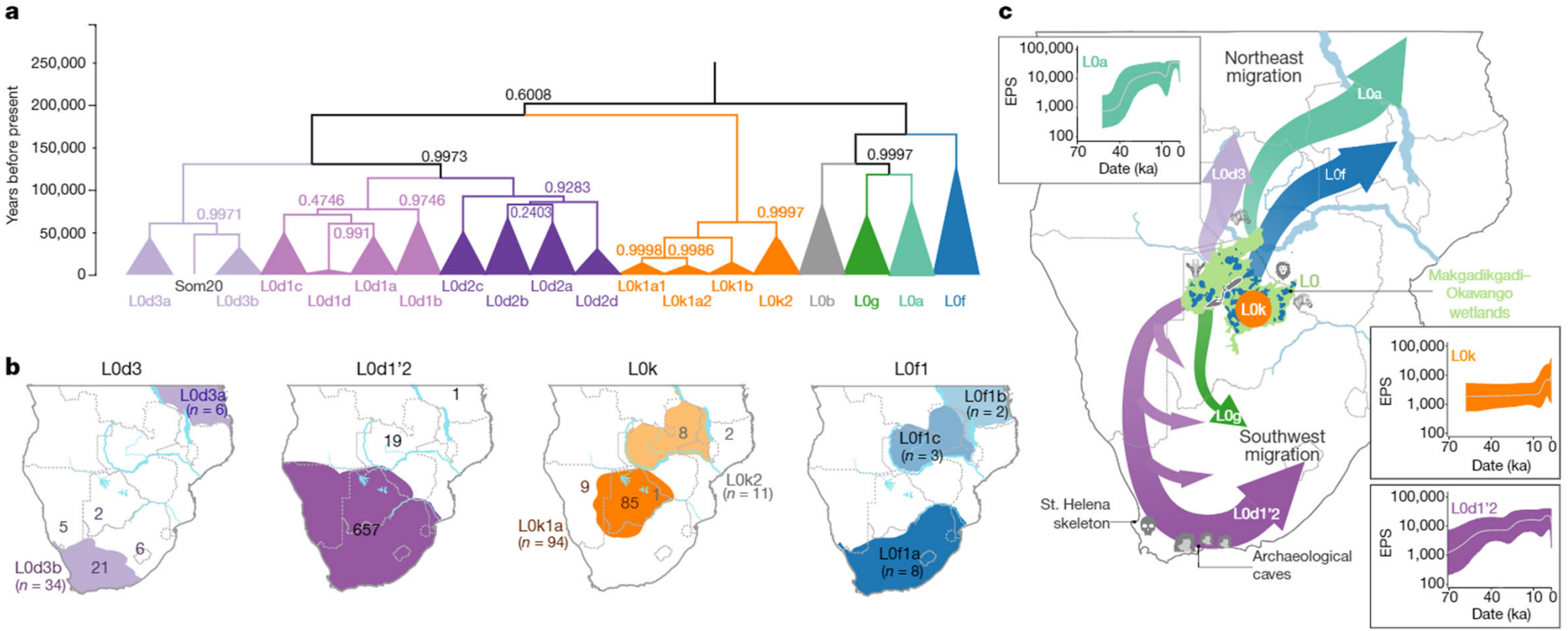


“Mitochondrial Eve” lived in Makgadikgadi–Okavango paleo-wetland of southern Africa ~200,000 years ago (between 165,000 and 240,000 years ago)

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

Okavango Delta now

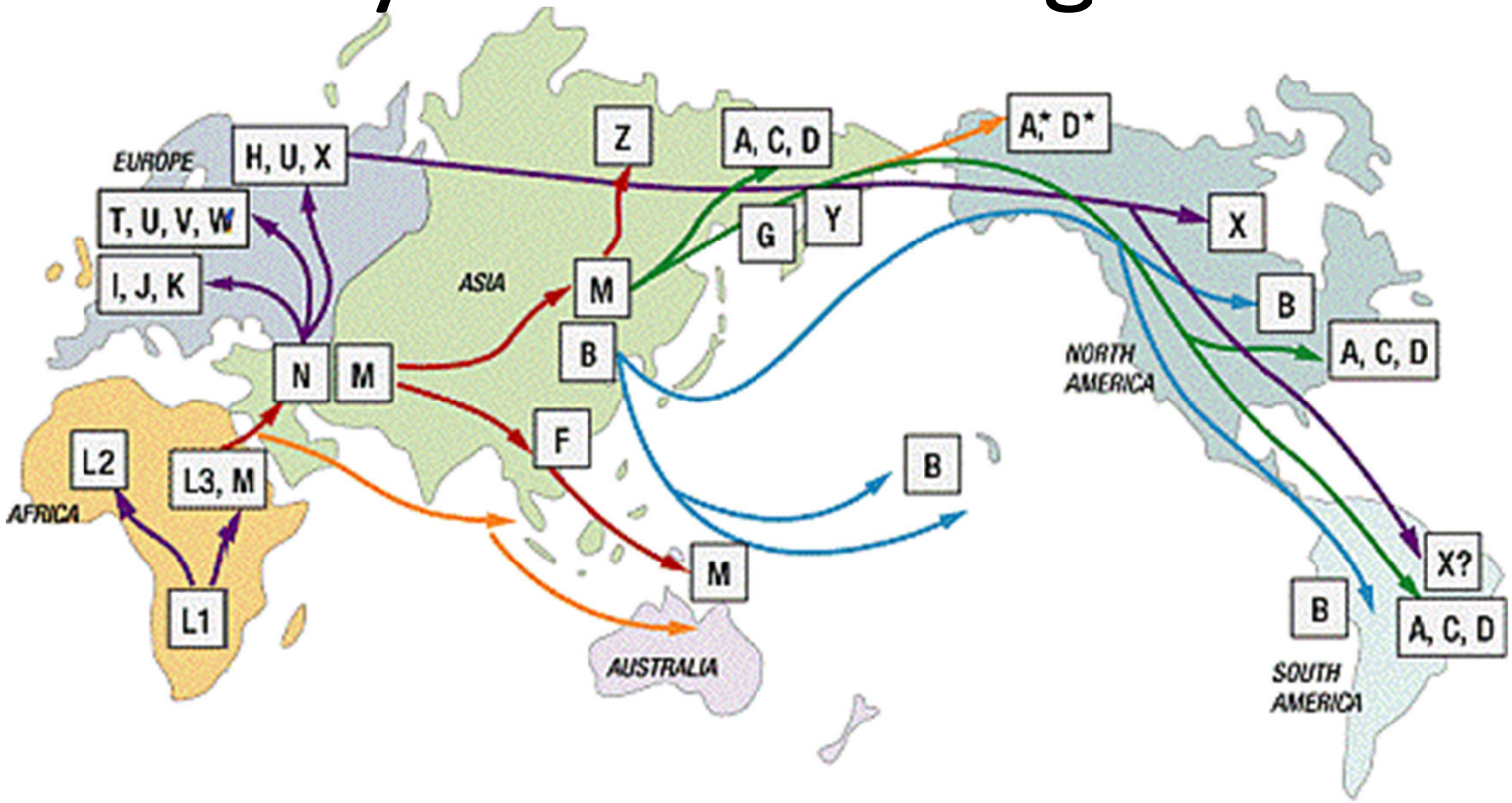
“Mitochondrial Eve” lived in Africa



“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.

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).*

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

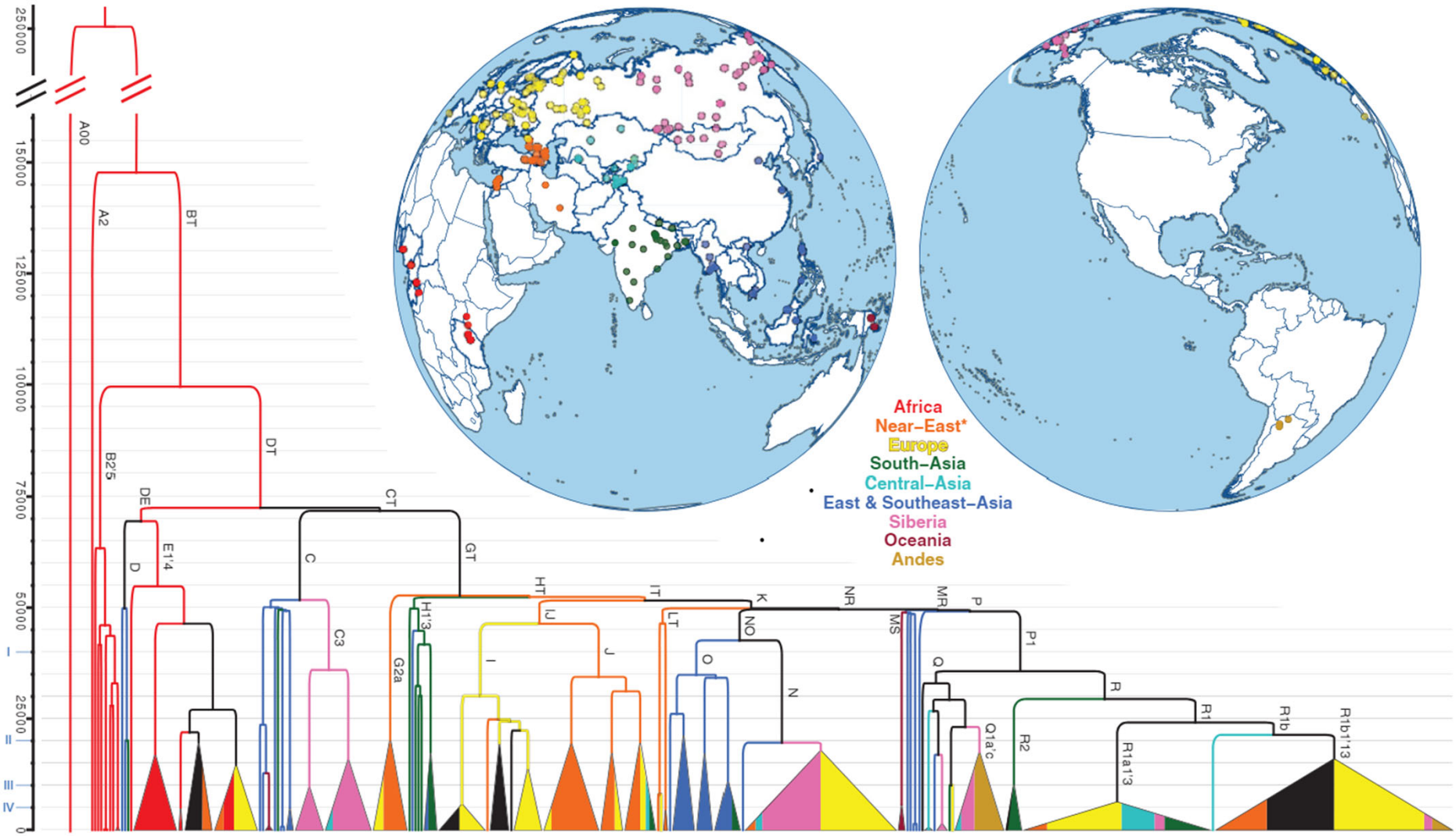
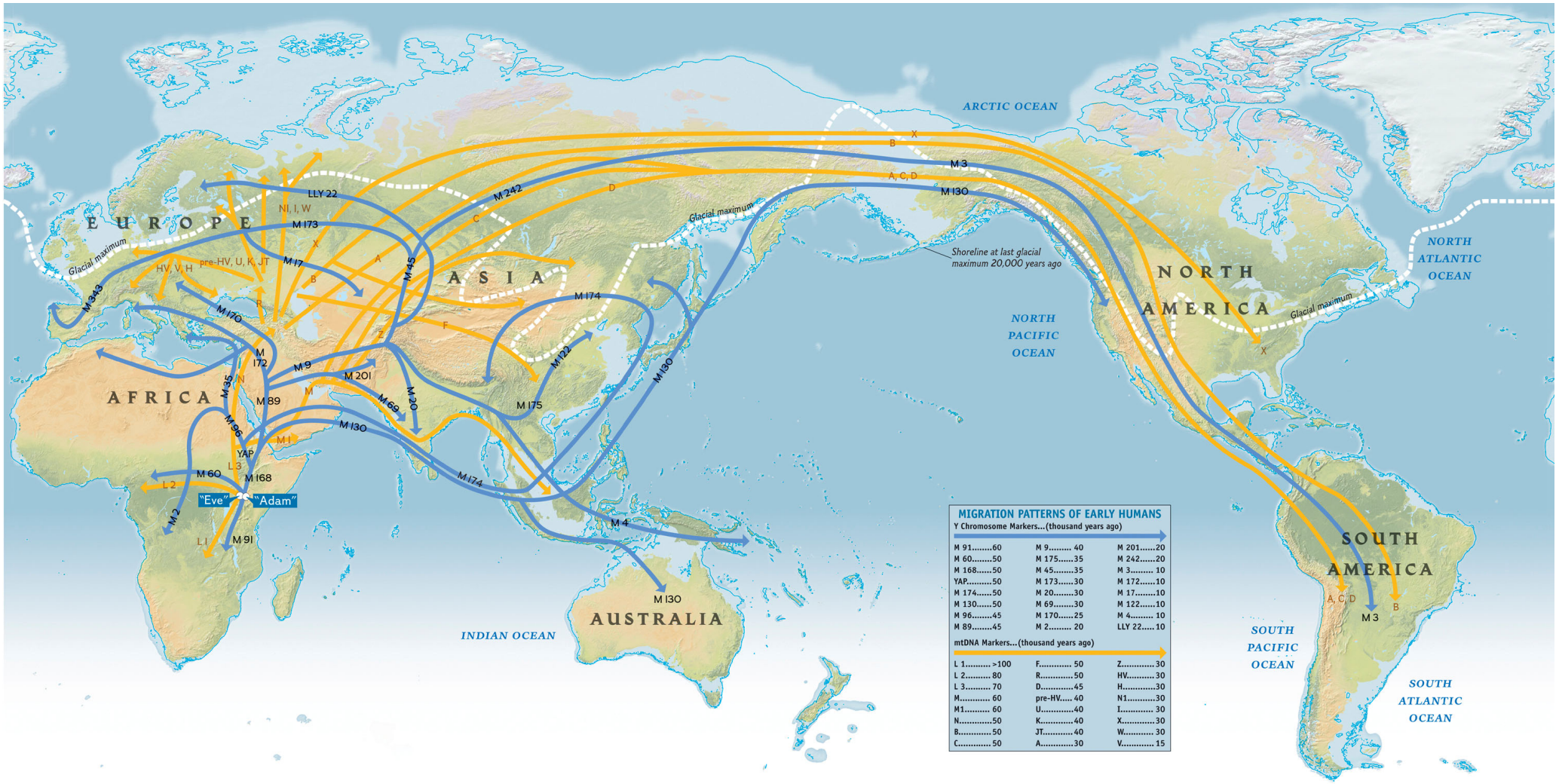


Figure 1. The phylogenetic tree of 456 whole Y chromosome sequences and a map of sampling locations. The phylogenetic tree is reconstructed using BEAST. Clades coalescing within 10% of the overall depth of the tree have been collapsed. Only main haplogroup labels are shown (details are provided in Supplemental 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).*