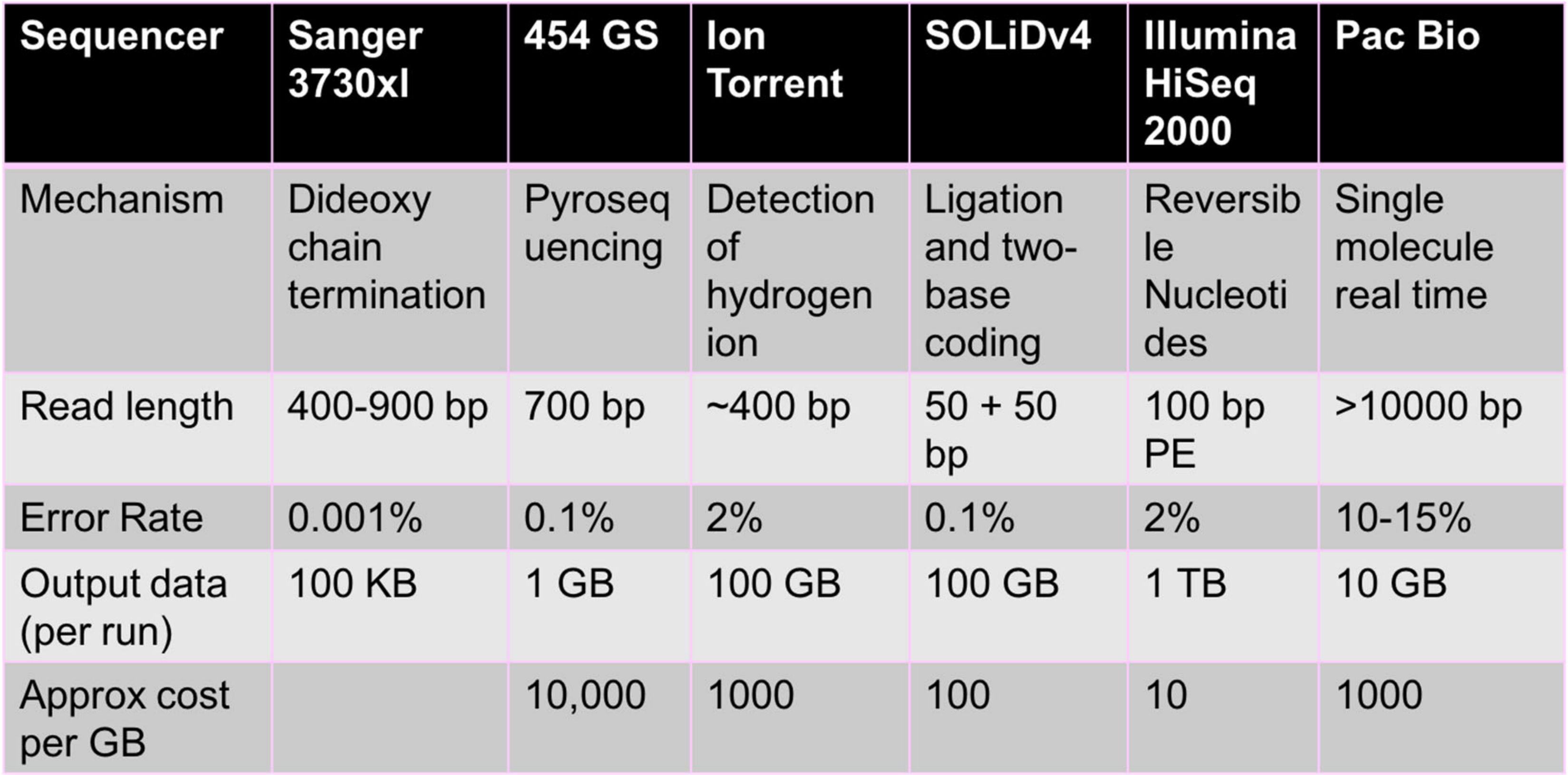
Poisson Example: Genome Assembly

- Goal: DNA sequence (ACTG) of the entire genome
- Problem: Sequencers generate random short reads



Sequencer	Sanger 3730xl	454 GS	Ion Torrent	SOLiDv4	Illumina HiSeq 2000	Pac Bio
Mechanism	Dideoxy chain termination	Pyrosequencing	Detection of hydrogen ion	Ligation and two-base coding	Reversible Nucleotides	Single molecule real time
Read length	400-900 bp	700 bp	~400 bp	50 + 50 bp	100 bp PE	>10000 bp
Error Rate	0.001%	0.1%	2%	0.1%	2%	10-15%
Output data (per run)	100 KB	1 GB	100 GB	100 GB	1 TB	10 GB
Approx cost per GB		10,000	1000	100	10	1000

- Solution: assemble genome from short reads using computers. Whole Genome Shotgun Assembly.

Table from the course EE 372 taught by David Tse at Stanford

Current sequencing technologies

	Second gen. (Illumina)	Oxford Nanopore (MinIon)	PacBio
read length (bases)	100-500	10K-100K	10K-20K
error rates	< 1%	10-15%	10-15%
speed (time/base)	6 mins/base/strand	250 bases/s	3 bases/s
# of reads in parallel	10^9	2000	150K
throughput (total # of bases/s)	3M	500K	450K

Table from the course EE 372: Data Science for High-Throughput Sequencing.
taught by David Tse at Stanford

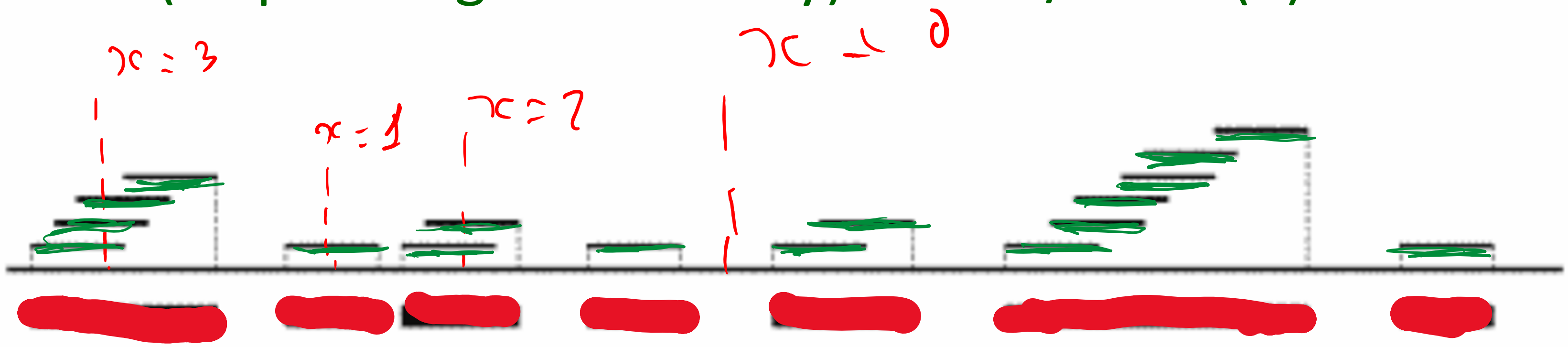
MinION, a palm-sized gene sequencer made by
UK-based Oxford Nanopore Technologies

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).

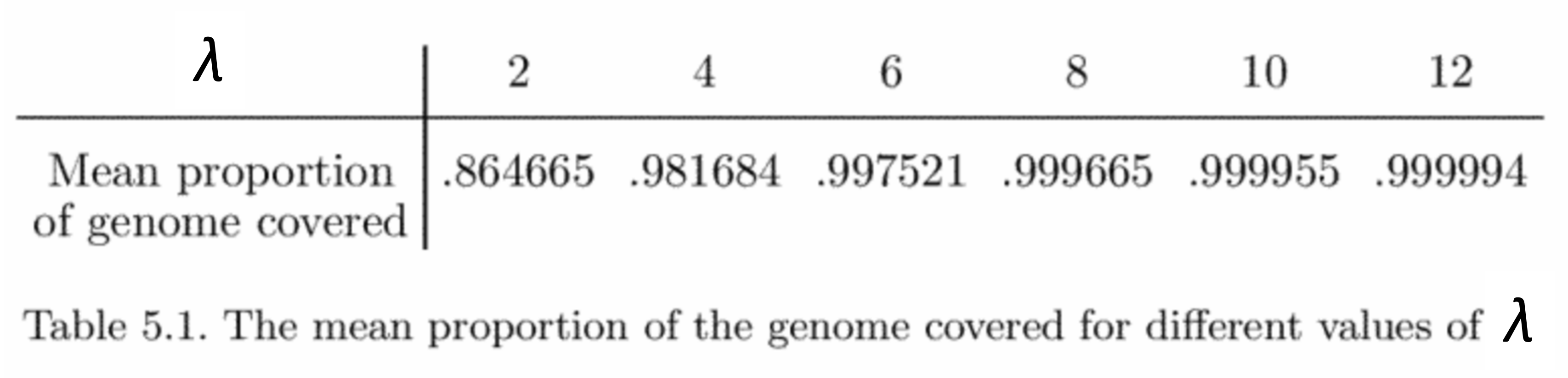


Ewens, Grant, Chapter 5.1

What fraction of genome is covered?

What fraction of genome is covered?

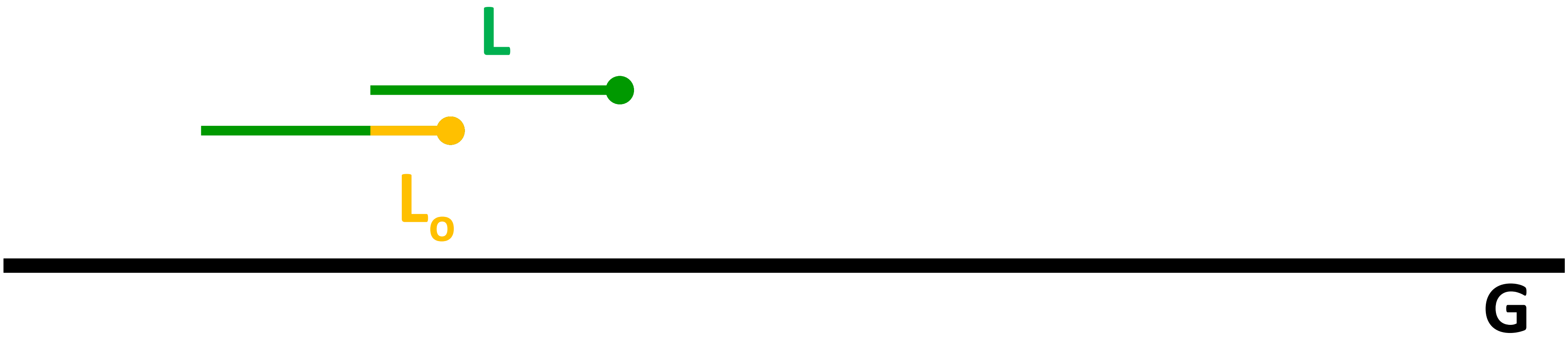
- Coverage: $\lambda=NL/G$,
  X – random variable equal to the number of times a given site is covered by short reads.
  Poisson: $P(X=x)=\lambda^{x}\exp(-\lambda)/x!$
  $P(X=0)=\exp(-\lambda)$, $P(X>0)=1-\exp(-\lambda)$
- *Total length covered:* $G*[1-\exp(-\lambda)]$



λ	2	4	6	8	10	12
Mean proportion of genome covered	.864665	.981684	.997521	.999665	.999955	.999994

Table 5.1. The mean proportion of the genome covered for different values of λ

How long should the overlap be to connect two short reads?



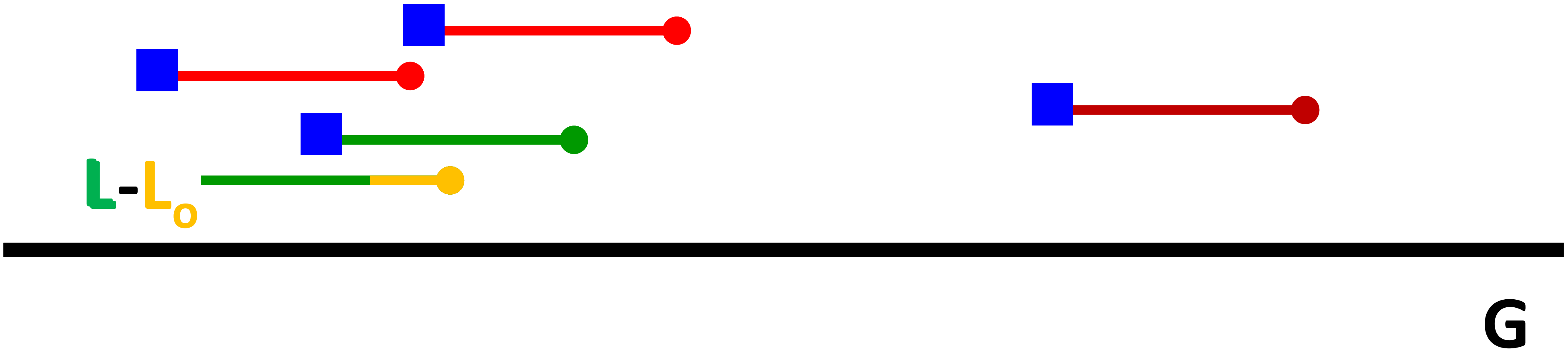
If DNA was a random chain with $p_A = p_C = p_G = p_T = 1/4$

L_o~16-20 would be enough

$$2 \cdot G \cdot 4^{-Lo} = 2 \cdot 3x10^9 \cdot 4^{-16} = 1.4$$

$$2 \cdot 3x10^9 \cdot 4^{-20} = 0.0055 << 1$$

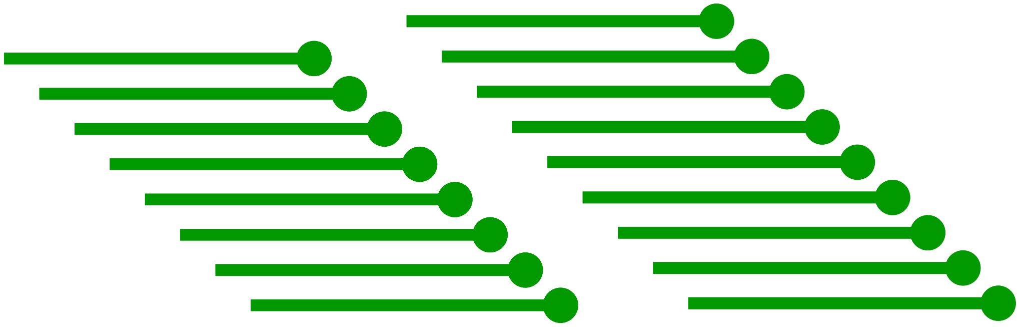
How many contigs?



$$\text{P(short read can be extended by another short read)}=\frac{L - L_o}{G}=\text{p}$$

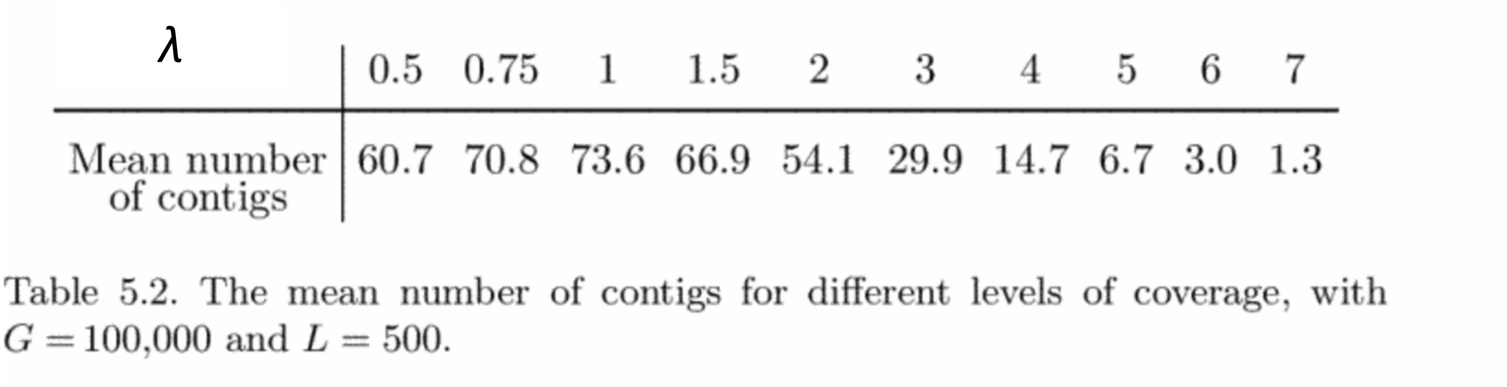
$$\text{P(short read cannot be extended by any short reads)}=e^{-pN} \approx Ne^{-\lambda}$$

$$\text{number of contigs}=Ne^{-pN} \approx Ne^{-\lambda}$$



How many contigs?

- A given short read is the right end of a contig if and only if no left ends of other short reads fall within first $L\text{-}L_{overlap}$ base pairs
- The left end of another short read has the probability $p=(L\text{-}L_{overlap})/G$ to fall within a given read. There are $N\text{-}1$ other reads.
- The expected number of left ends inside a given short read is $p\cdot (N\text{-}1)=(N\text{-}1)\cdot(L\text{-}L_{overlap})/G \approx \lambda$ (if $L>>L_{overlap}$)
- Probability that no left ends fall inside a given short read is $exp(- \lambda)$. Thus, the Number of contigs is $N_{contigs}=Ne^{-\lambda}$:



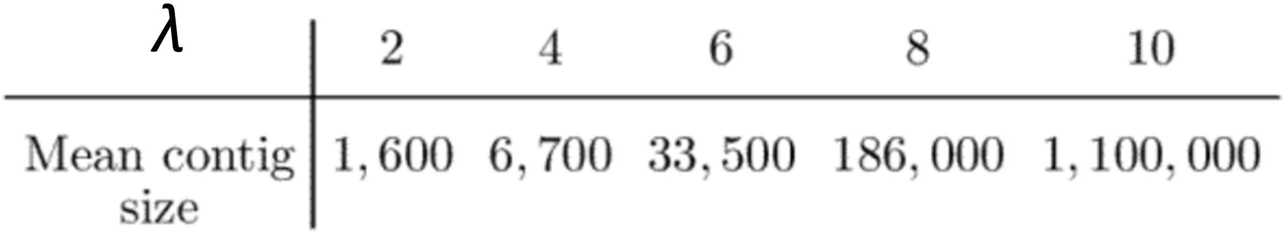
λ	0.5	0.75	1	1.5	2	3	4	5	6	7
Mean number of contigs	60.7	70.8	73.6	66.9	54.1	29.9	14.7	6.7	3.0	1.3

Table 5.2. The mean number of contigs for different levels of coverage, with $G = 100{,}000$ and $L = 500$.

Average length of a contig?

- Length of a genome covered: $G_{covered} = G \cdot P(X>0) = G \cdot (1 - \exp(-\lambda))$
- *Number of contigs* $N_{contigs} = N \cdot e^{-\lambda}$
- Average length of a contig = $\langle L \rangle = \sum_i L_i / N_{contigs} = G_{covered} / N_{contigs} =$

$$G \cdot (1 - \exp(-\lambda)) / N \cdot e^{-\lambda} = L \cdot (1 - \exp(-\lambda)) / \lambda \cdot e^{-\lambda}$$



λ	2	4	6	8	10
Mean contig size	1,600	6,700	33,500	186,000	1,100,000

Table 5.3. The mean contig size for different values of a for the case $L = 500$.

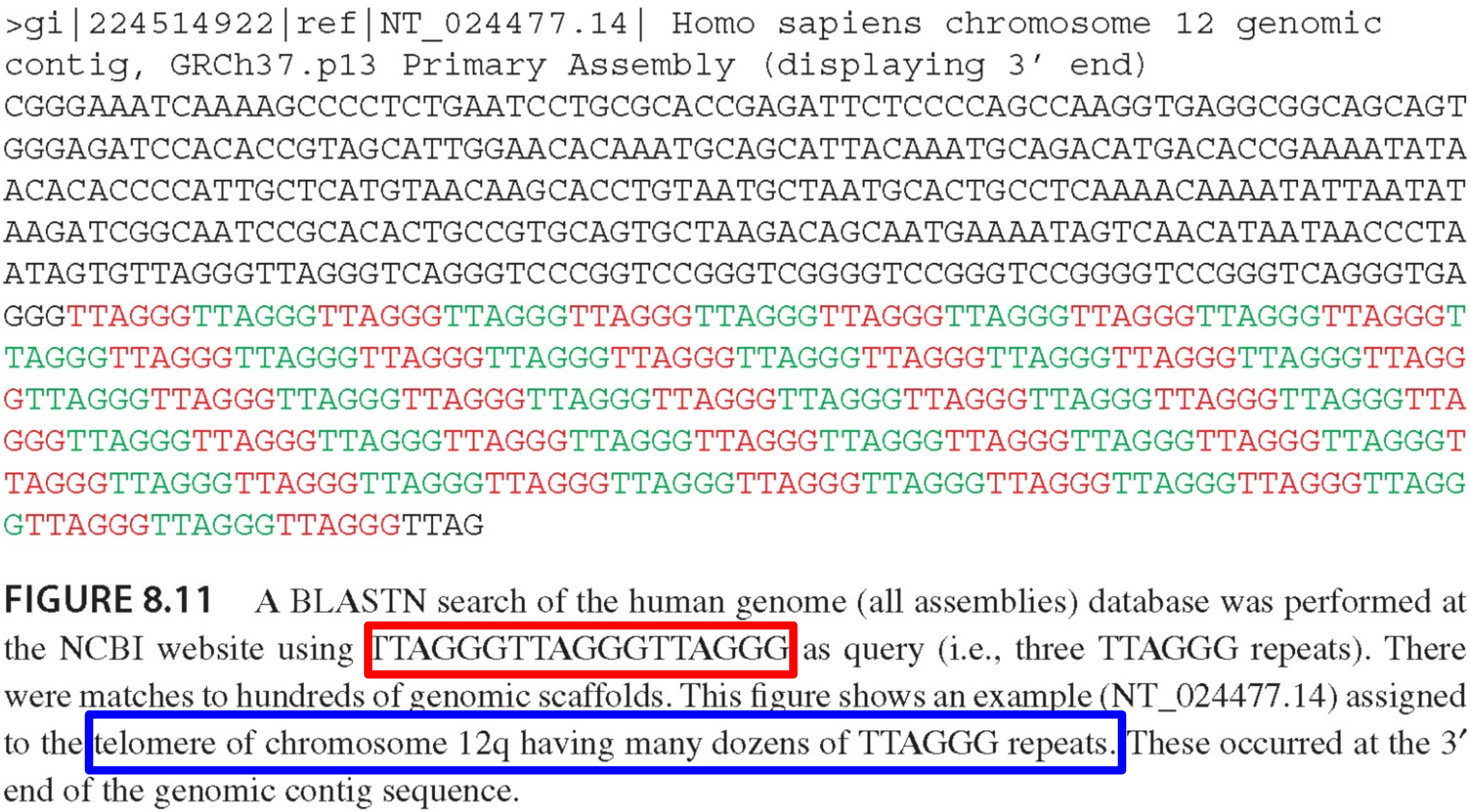
Estimate

- Human genome is 3×10^9 bp long
- Chromosome 1 is about $G=0.25 \times 10^9$ 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: $N=4.4 \times 10^7$ 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 3×10^9 x 17.6 $/10^6$ x 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
```

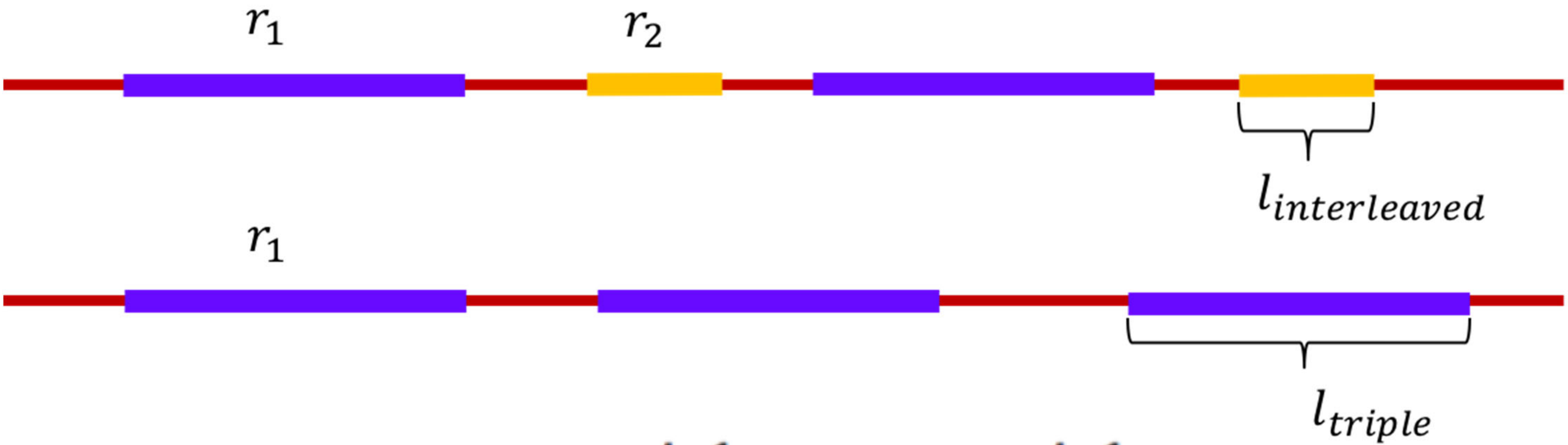
FIGURE 8.11 A BLASTN search of the human genome (all assemblies) database was performed at the NCBI website using TTAGGGTTAGGGTTAGGG as query (i.e., three TTAGGG repeats). There were matches to hundreds of genomic scaffolds. This figure shows an example (NT_024477.14) assigned to the telomere of chromosome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

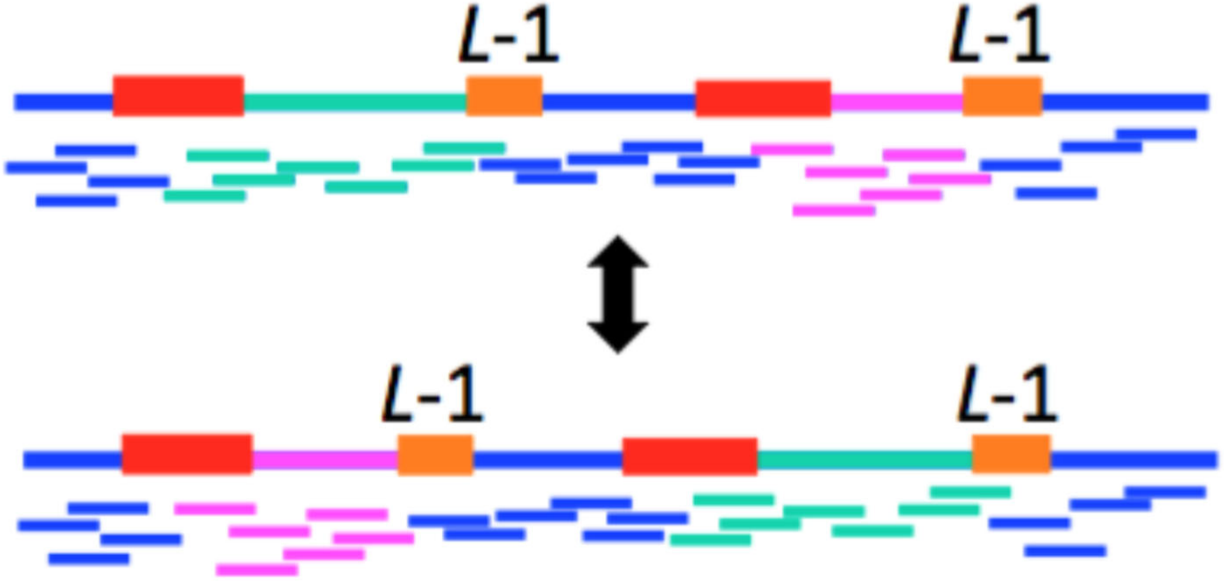
There were 100s of matches while one expects << 1 match:

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

DNA repeats make assembly difficult

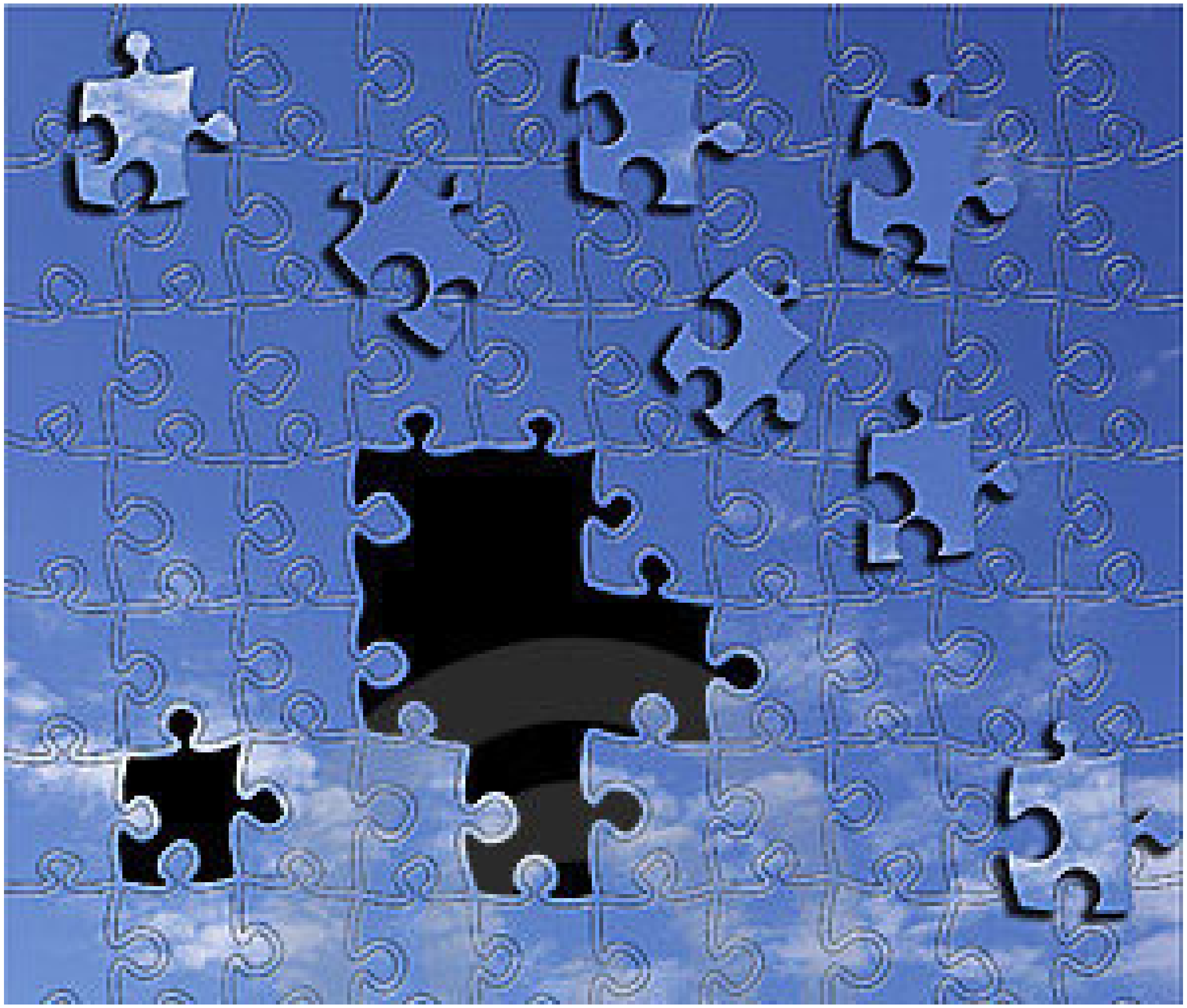
Why repeats make assembly difficult?



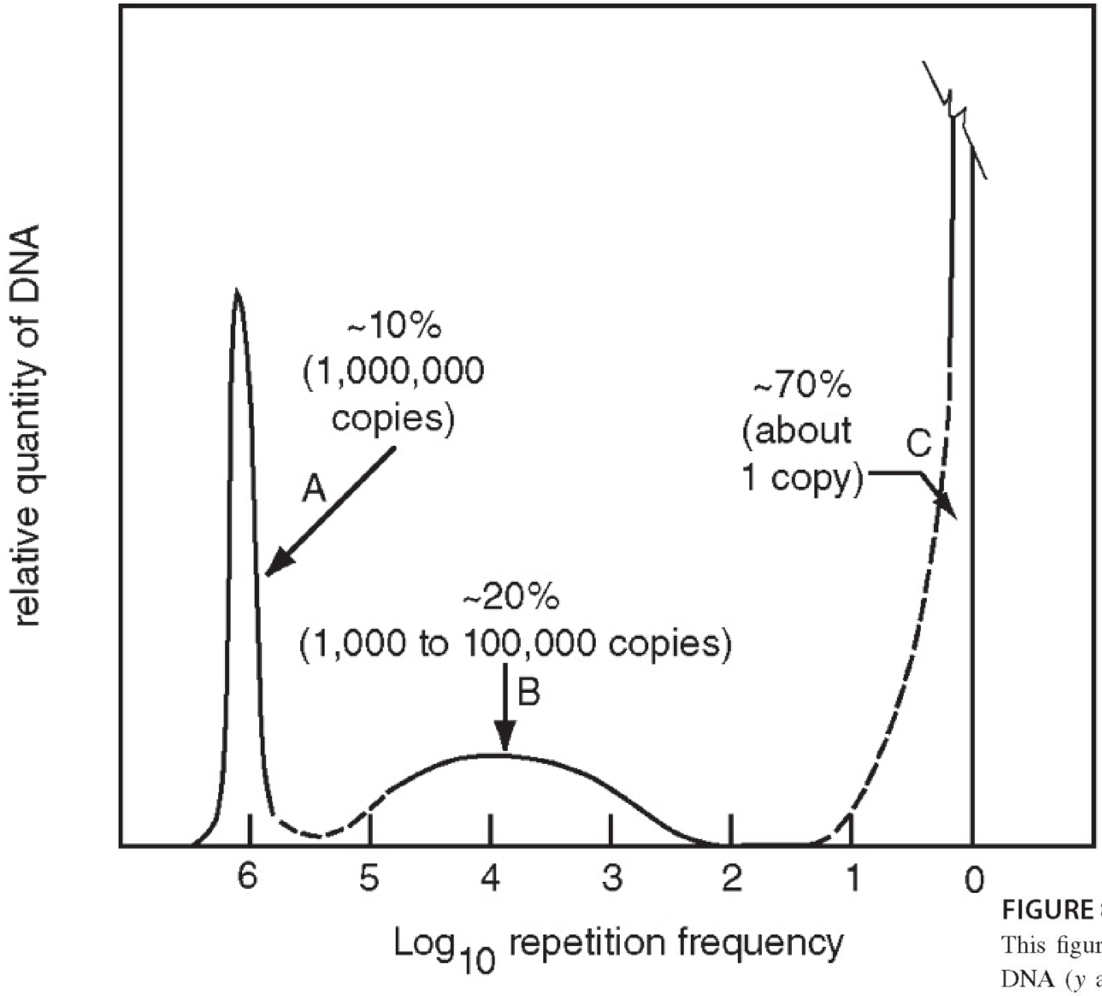


Images from the course EE 372: Data Science for High-Throughput Sequencing. taught by David Tse at Stanford

Repeats are like sky puzzle pieces



How many repeats are in eukaryotic genomes?

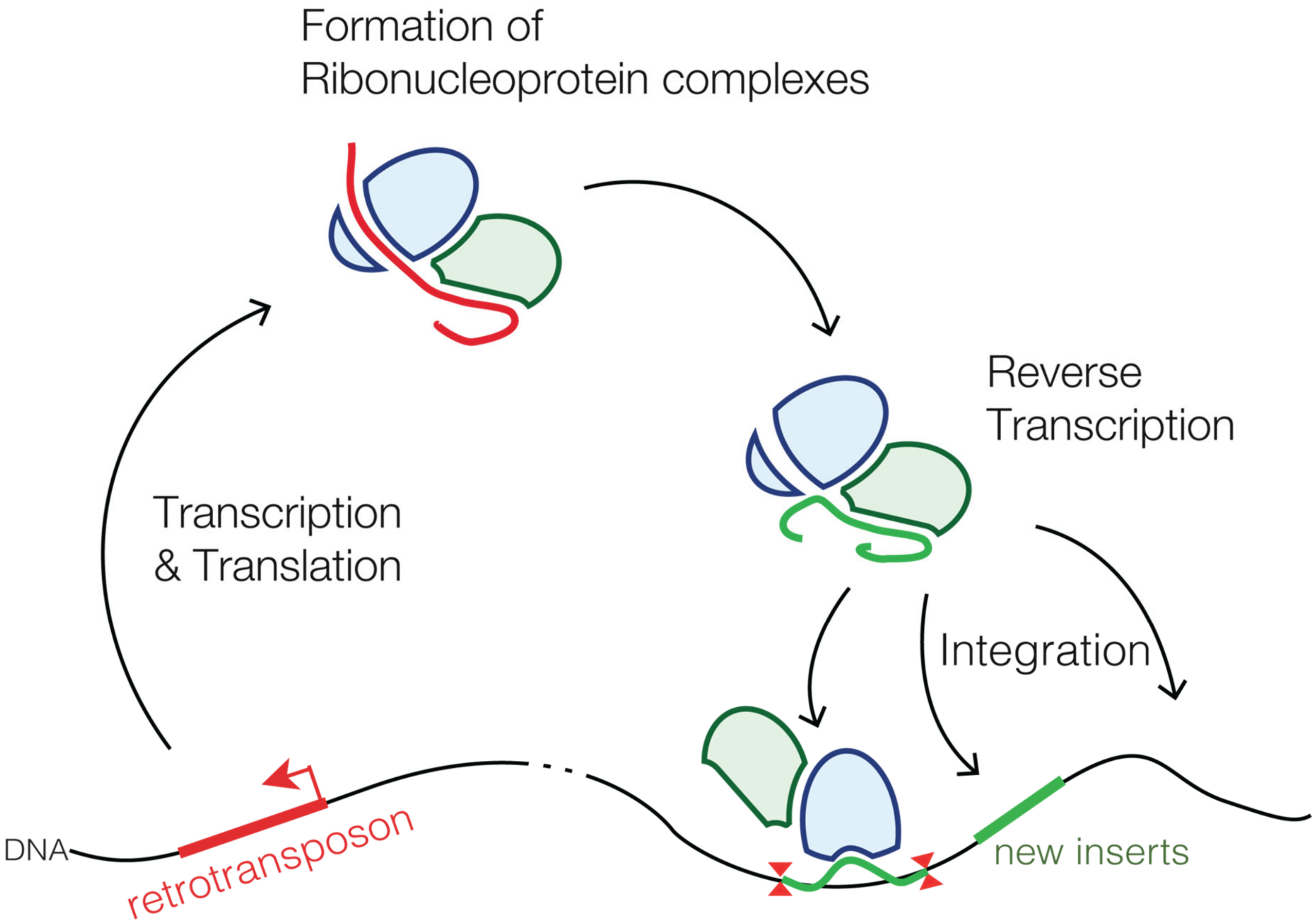


Data for mouse genome obtained in 1961 (sic!) using DNA denaturation and renaturation curves

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.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.

Companion Website: www.wiley.com/go/pevsnerbioinformatics



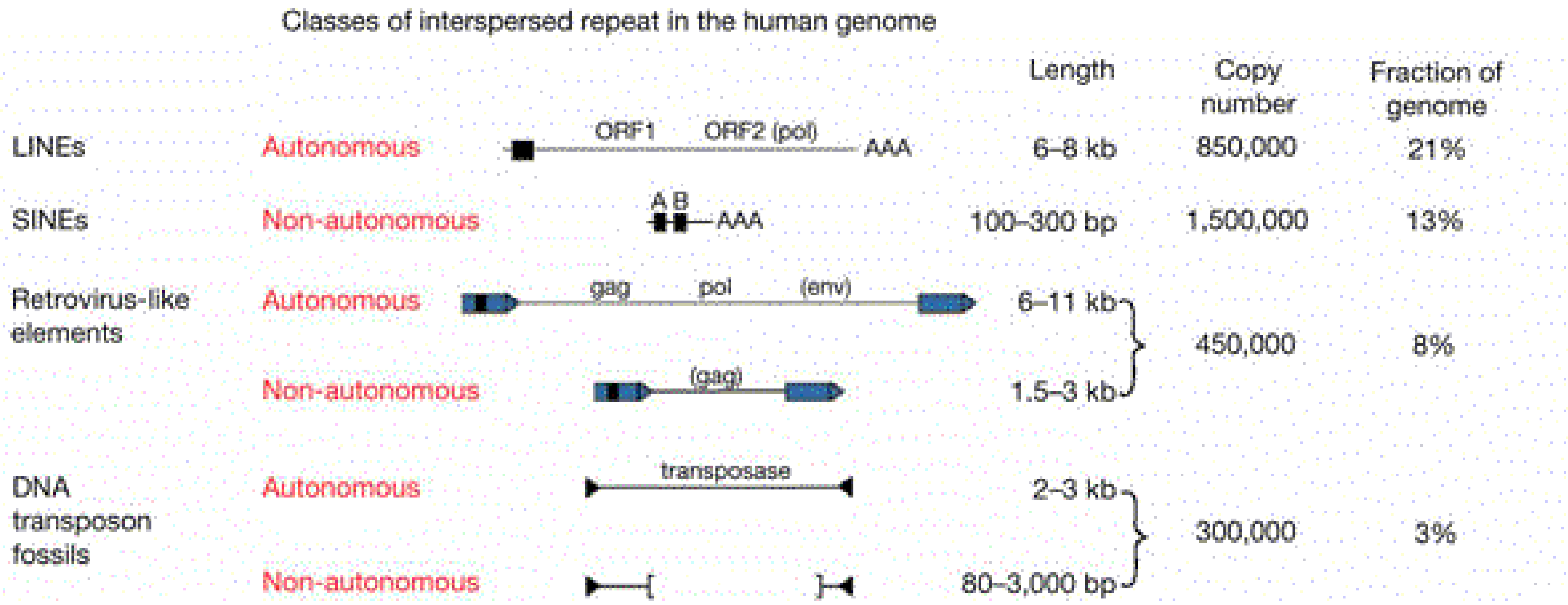
Formation of Ribonucleoprotein complexes

Reverse Transcription

Integration

Transcription & Translation

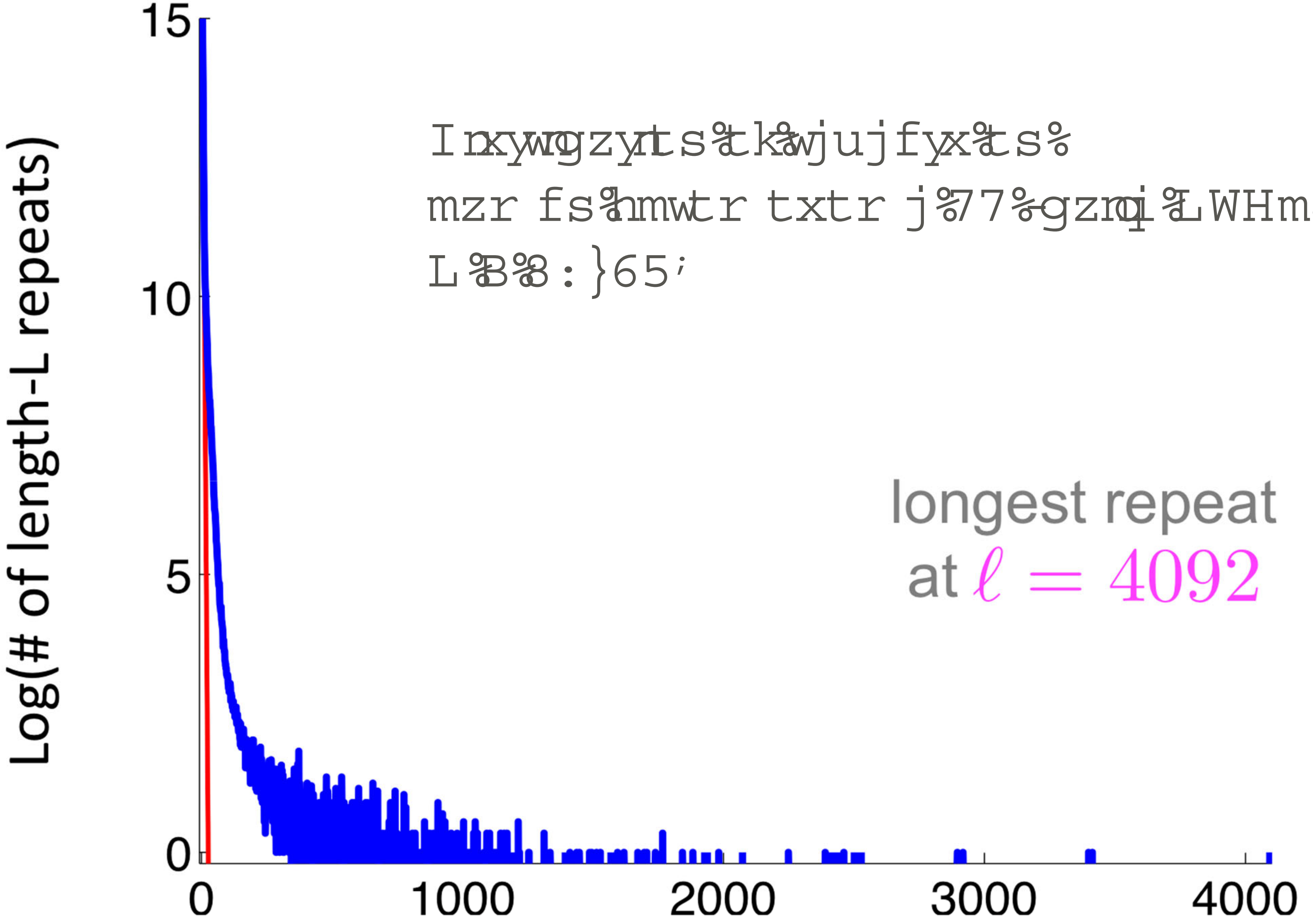
Almost all transposable elements in mammals fall into one of four classes



Classes of interspersed repeat in the human genome

			Length	Copy number	Fraction of genome
LINEs	Autonomous	ORF1 ORF2 (pol) AAA	6–8 kb	850,000	21%
SINEs	Non-autonomous	A B AAA	100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	gag pol (env)	6–11 kb	450,000	8%
	Non-autonomous	(gag)	1.5–3 kb		
DNA transposon fossils	Autonomous	transposase	2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

Slide by Ross Hardison, Penn State U.



Inxywngzynts%tk%wjujfyx%ns%
mzrfs%hmwtrtxtrj%77%gznqi%LWHm8<.%
L%B%8:}65;

longest repeat at $\ell = 4092$

Images from the course EE 372: Data Science for High-Throughput Sequencing.
taught by David Tse at Stanford

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.

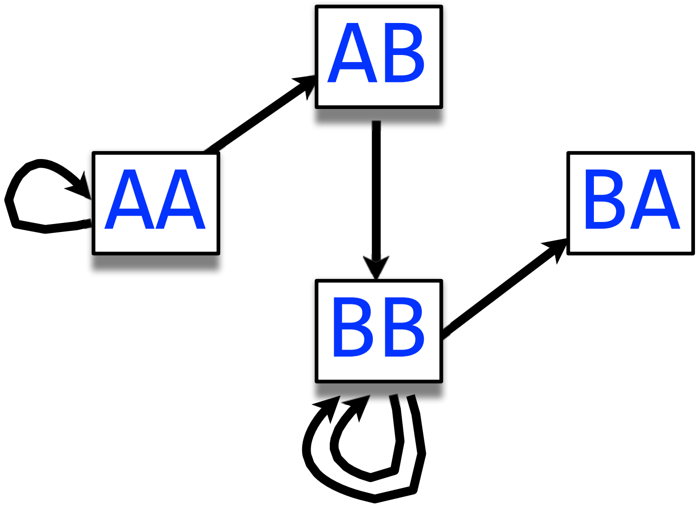
http://www.langmead-lab.org/teaching-materials/

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

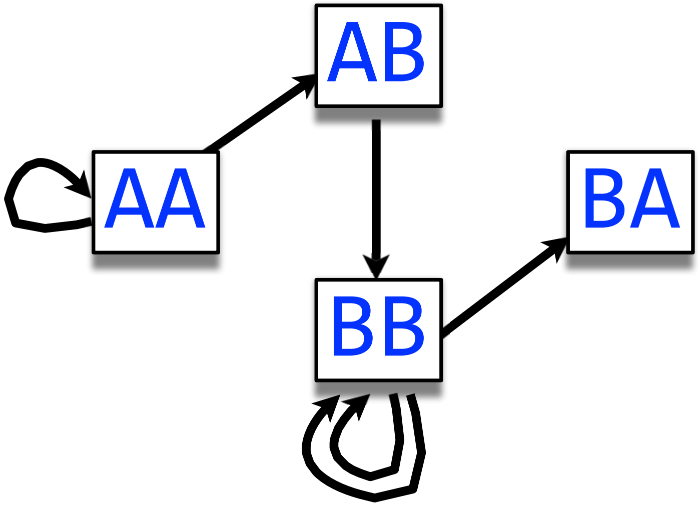


One edge per every *k*-mer

One node per distinct *k*-1-mer

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

De Bruijn graph

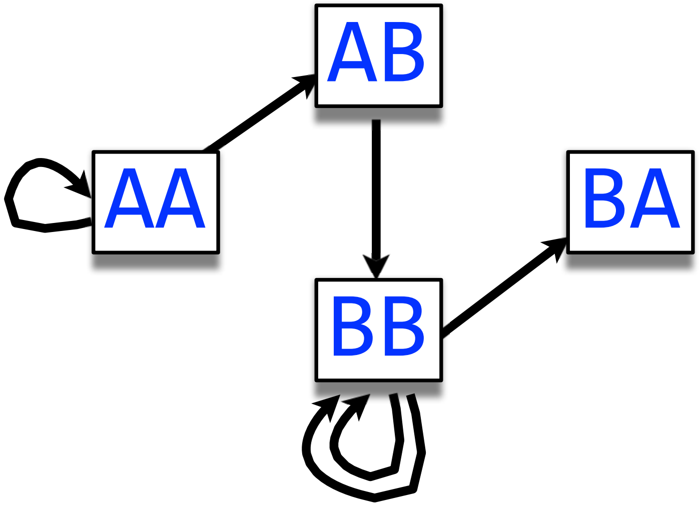


Walk crossing each edge exactly once gives a reconstruction of the genome

Courtesy of Ben Langmead. Used with permission.

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

Assembly = Eulerian walk on De Bruijn graph

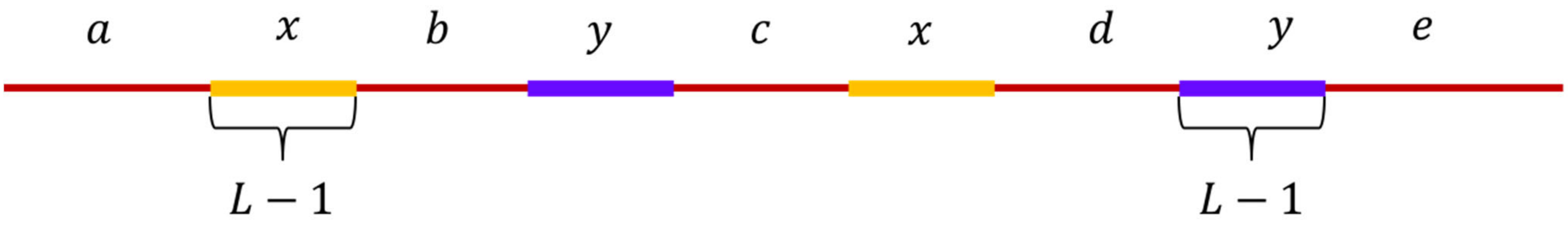


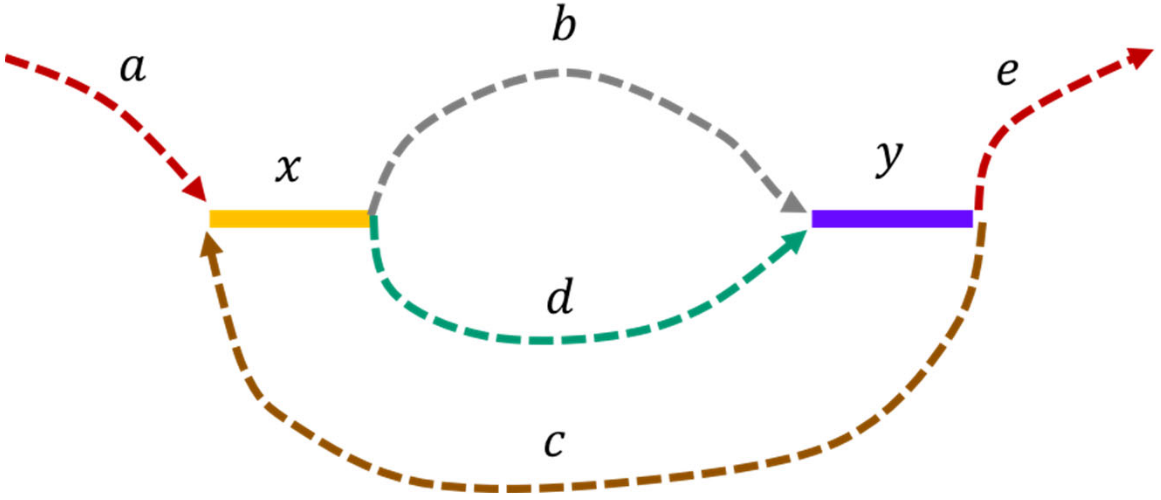
AAABBBBA

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/

Why interleaved repeats are dangerous?

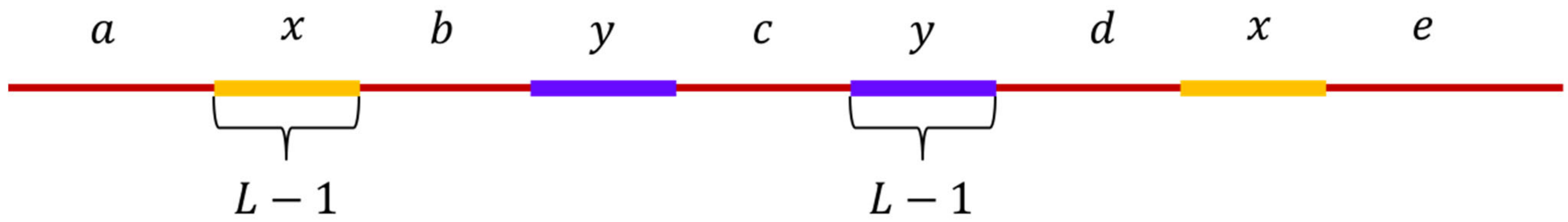


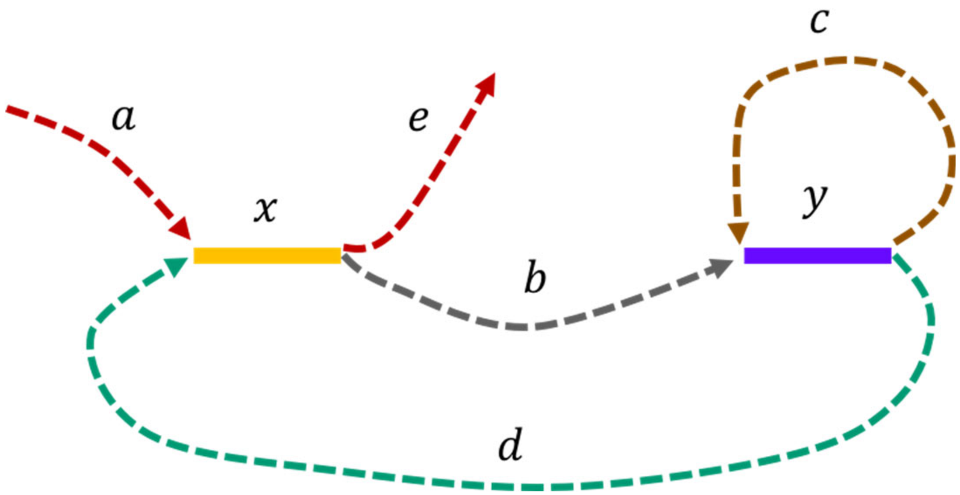


The two Eulerian paths that are on the graph:
a–x–b–y–c–x–d–y–e and a–x–d–y–c–x–b–y–e

Images from the course EE 372: Data Science for High-Throughput Sequencing.
taught by David Tse at Stanford

Why non-interleaved repeats are safe?

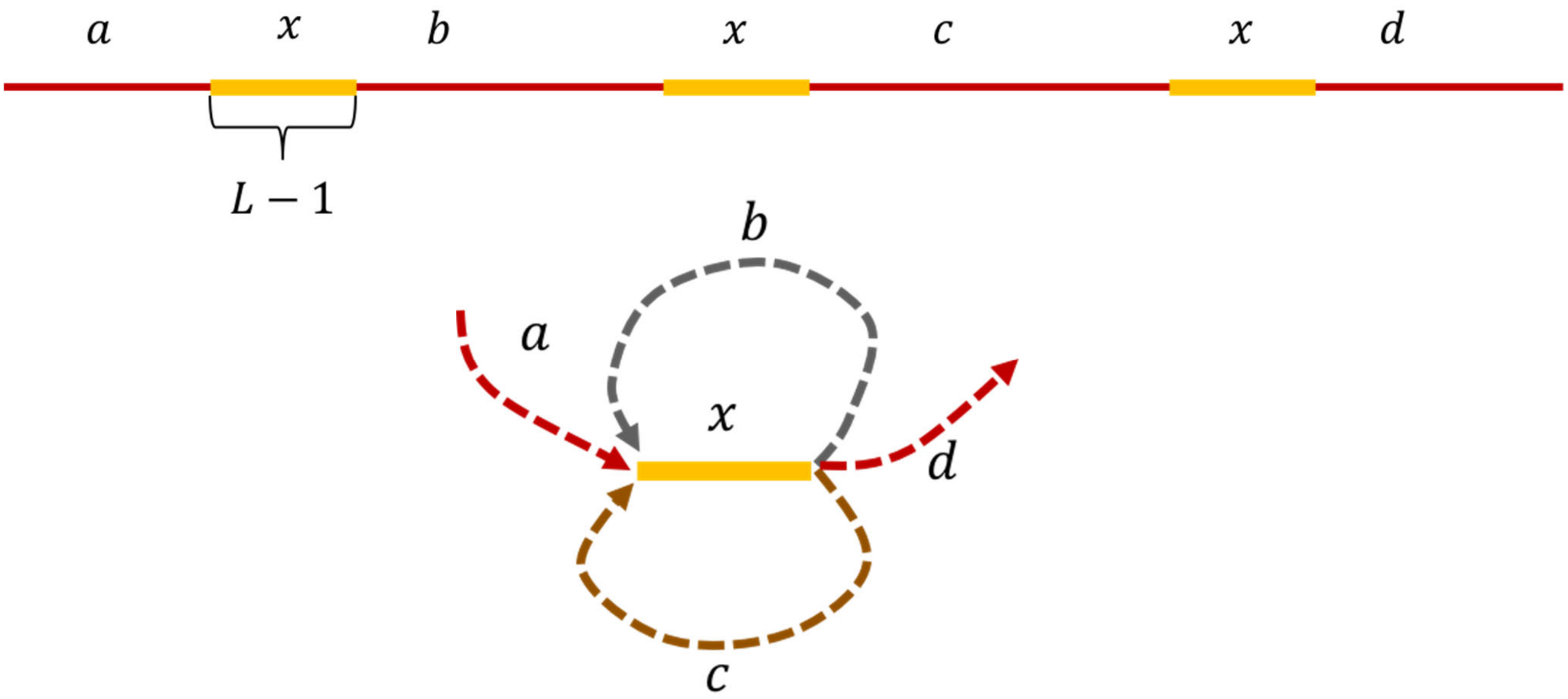




The only Eulerian path is: a–x–b–y–c–y–d–x–e

Images from the course EE 372: Data Science for High-Throughput Sequencing. taught by David Tse at Stanford

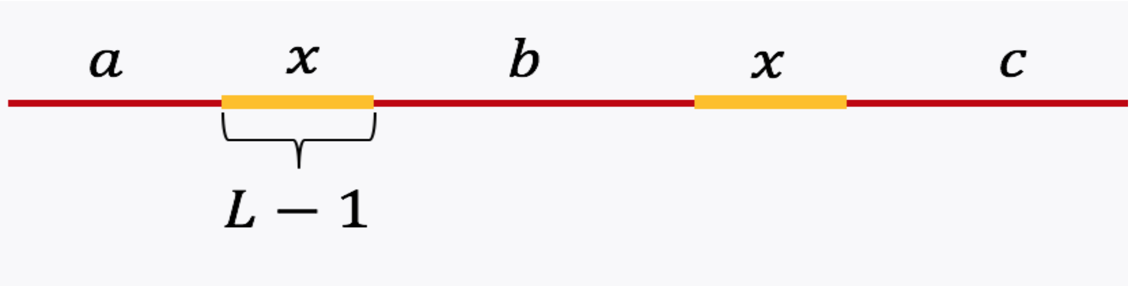
Why triple repeats are dangerous?

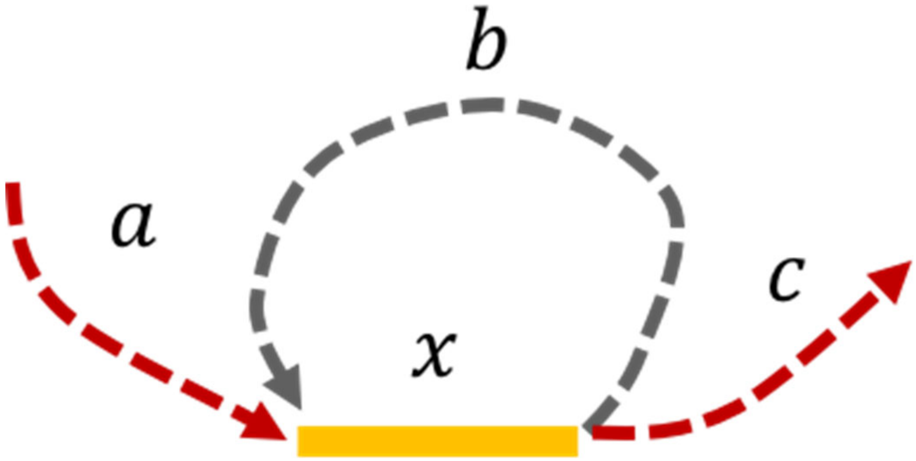


The two Eulerian paths that are on the graph:
a−x−b−x−c−x−d   and a−x−c−y−b−x−d

Images from the course EE 372: Data Science for High-Throughput Sequencing. taught by David Tse at Stanford

Why double repeats are safe?





The only Eulerian path is: a–x–b–x–c

Images from the course EE 372: Data Science for High-Throughput Sequencing. taught by David Tse at Stanford

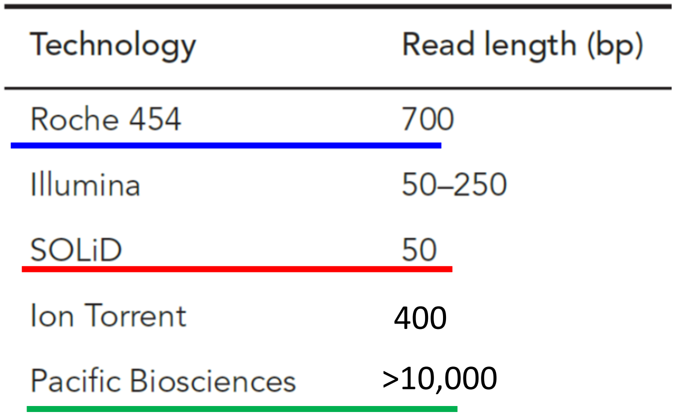
Pavel Pevzner's theorem

- **Theorem [Pevzner 1995]**: If L, the read length, is strictly greater than max($\ell_{\text{interleaved}}$,$\ell_{\text{triple}}$), then the de Bruijn graph has a unique Eulerian path corresponding to the original genome.

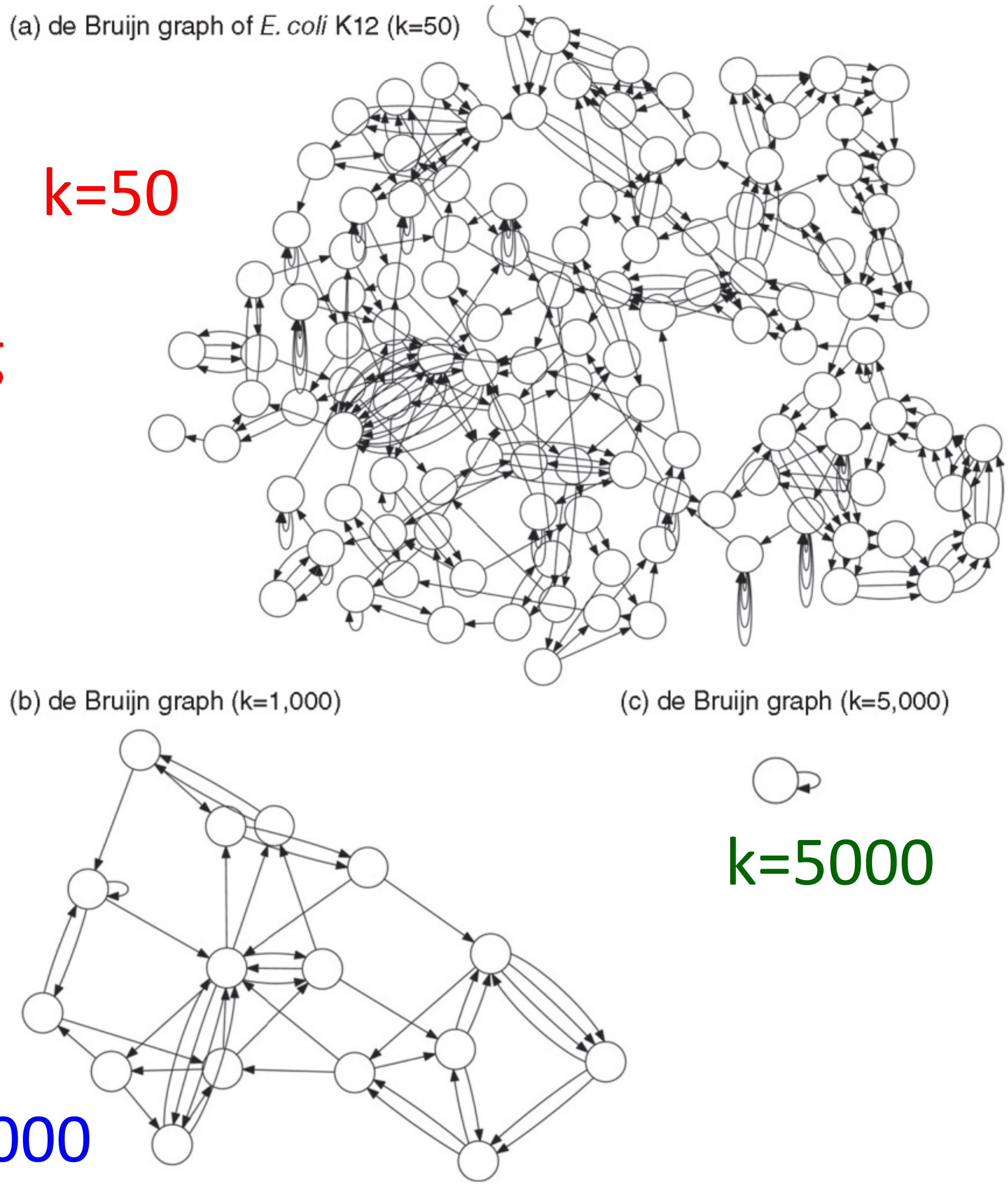
Pavel Pevzner
is the Ronald R. Taylor Chair and
Distinguished Professor of
Computer Science and Engineering
at University of California, San Diego.
His Alma Mater is
Moscow Institute of
Physics and Technology
in Russia.

How to assemble a genome with repeats?

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



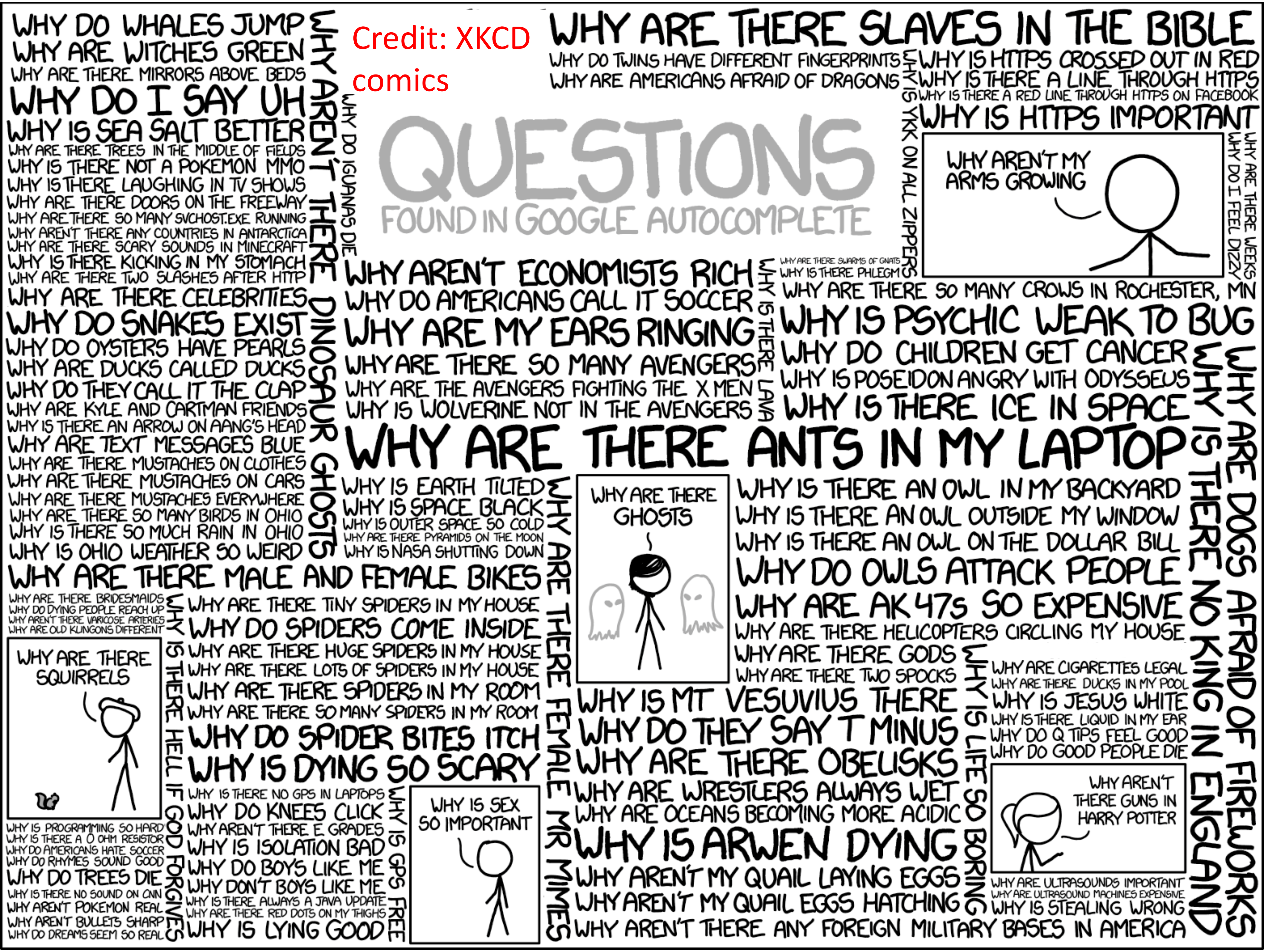
Technology	Read length (bp)
Roche 454	700
Illumina	50–250
SOLiD	50
Ion Torrent	400
Pacific Biosciences	>10,000



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

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

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


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 AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
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 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 SWARMS OF GNATS
WHY IS THERE PHLEGM
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 MALE AND FEMALE BIKES
WHY ARE THERE GHOSTS
WHY ARE THERE FEMALE MR MIMES
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 THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
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 IS THERE HELL IF GOD FORGIVES
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 GPS FREE
WHY IS SEX SO IMPORTANT
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

Credit: XKCD comics

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

Geometric Mean & Variance

$$P(X=x) = p(1-p)^{x-1} = p \cdot q^{x-1}$$

$$S(p,q) = \sum_{x=1}^{\infty} P(X=x) = \frac{p}{1-q} = \frac{p}{p} = 1$$

$$q\frac{\partial S}{\partial q} = \sum (x-1) P(X=x) = \frac{pq}{(1-q)^2} = \frac{q}{p}$$

$$\langle x \rangle = \sum (x-1) P(X=x) + 1 = \frac{1-p}{p} + 1 = \frac{1}{p}$$

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 ~= mean

- Very different from Poisson, where it is

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

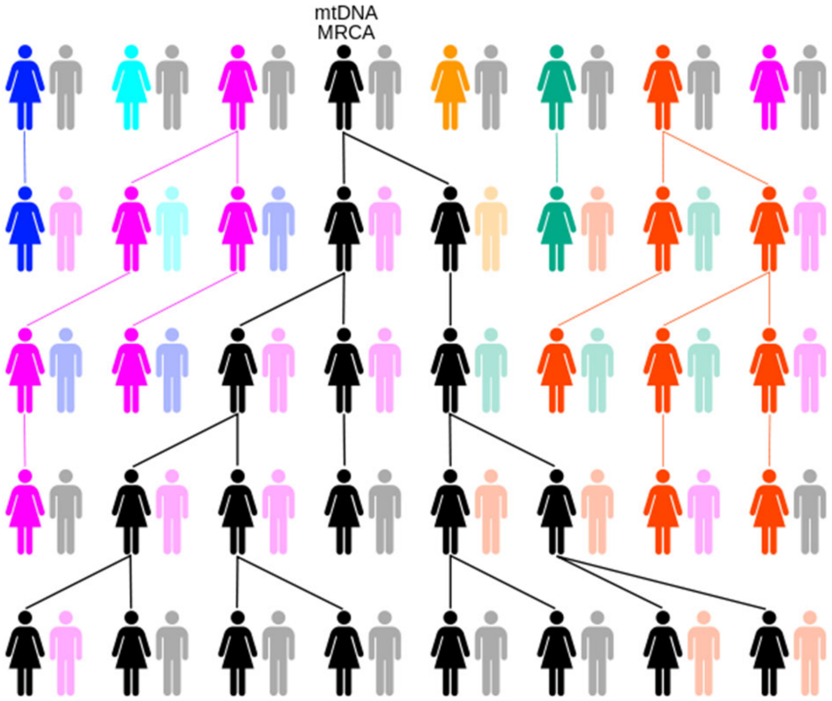
Matlab exercise

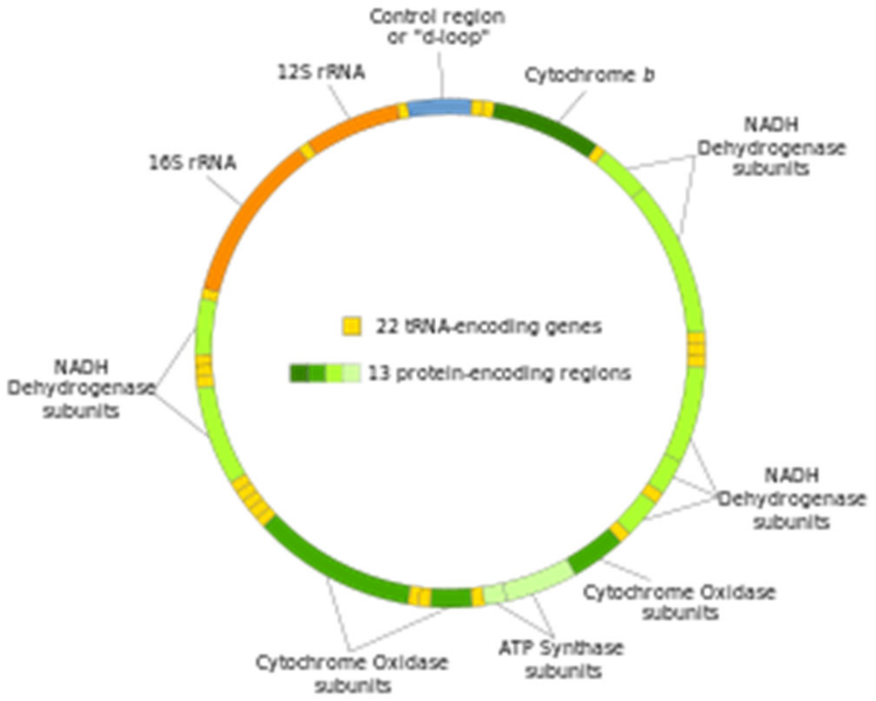
- Find mean, variance, and histogram of 100,000 geometrically-distributed numbers with p=0.1
- Hint: Use help page for **random** command on how to generate geometrically-distributed random numbers

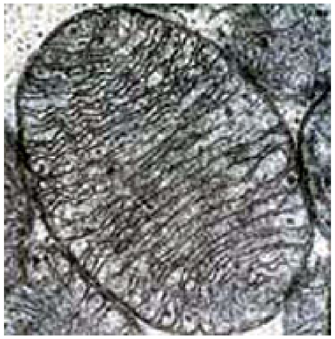
Matlab: Geometric distributions

- **Stats=100000;**
- **p=0.1;**
- **r2=random('Geometric',p,Stats,1);**
- **r2=r2+1;**
- **disp(mean(r2));**
- **disp(var(r2));**
- **disp(std(r2));**
- **[a,b]=hist(r2, 1:max(r2));**
- **p_g=a./sum(a);**
- **figure; semilogy(b,p_g,'ko-');**

Geometric distribution in biology



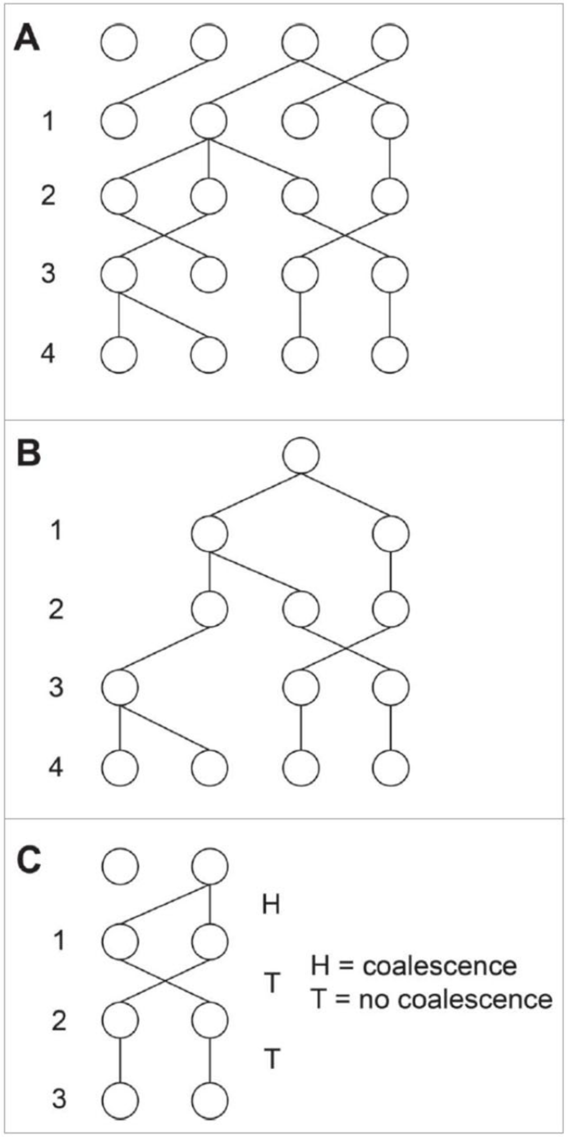




- Each of our cells has mitochondria with 16.5kb of mtDNA **inherited only from our mother**
- Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA
- Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeaon (of UIUC’s Carl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