Bernoulli distribution

The simplest non-uniform distribution

p – probability of success (1)

1-p – probability of failure (0)

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

Jacob Bernoulli
(1654-1705)
Swiss mathematician (Basel)

- Law of large numbers
- Mathematical constant e=2.718…

Bernoulli distribution

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = 0(1 - p) + 1(p) = p$$

$$\text{Var}(X) = E(X^2) - (EX)^2 = [0^2(1 - p) + 1^2(p)] - p^2 = p - p^2 = p(1 - p)$$

Bernoulli distribution

The simplest non-uniform distribution

p – probability of success (1)

1-p – probability of failure (0)

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

Jacob Bernoulli
(1654-1705)
Swiss mathematician (Basel)

- Law of large numbers
- Mathematical constant e=2.718…

Bernoulli distribution

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = 0(1 - p) + 1(p) = p$$

$$\text{Var}(X) = E(X^2) - (EX)^2 = [0^2(1 - p) + 1^2(p)] - p^2 = p - p^2 = p(1 - p)$$

Refresher: Binomial Coefficients

$$\binom{n}{k} = C_k^n = \frac{n!}{k!(n-k)!}\ ,\ \text{called } n \text{ choose } k$$

$$\binom{10}{3} = C_3^{10} = \frac{10!}{3!7!} = \frac{10\cdot 9\cdot 8\cdot 7!}{3\cdot 2\cdot 1\cdot 7!} = 120$$

Number of ways to choose k objects out of n
without replacement and where the order does not matter.
Called binomial coefficients because of the binomial formula

$$(p+q)^n = (p+q)\times(p+q)\ldots\times(p+q) = \sum_{x=0}^{n} C_x^n p^x q^{n-x}$$

Binomial Distribution

- Binomially-distributed random variable *X* equals sum (number of successes) of n independent Bernoulli trials
- The probability mass function is:

$q = 1 - p$

$$f(x) = C_x^n p^x (1-p)^{n-x} \quad \text{for } x = 0,1,\ldots n \qquad (3\text{-}7)$$

- Based on the binomial expansion:

$$1 = (p+q)^n = \sum_{x=0}^{n} C_x^n p^x q^{n-x}$$

Binomial variance and standard deviation

Let X be a binomial random variable with parameters p and n

Variance:

$\sigma^2 = V(X) = np(1-p)$

Standard deviation:

$\sigma = \sqrt{np(1-p)}$

Poisson Distribution

- Limit of the binomial distribution when
  - n , the number of attempts, is very large
  - p , *the probability of success* is very small
  - $E(X)=np=\lambda$ is O(1)

The annual numbers of deaths from horse kicks in 14 Prussian army corps between 1875 and 1894

Number of deaths	Observed frequency	Expected frequency
0	144	139
1	91	97
2	32	34
3	11	8
4	2	1
5 and over	0	0
Total	280	280

From von Bortkiewicz 1898

Siméon Denis Poisson (1781–1840) French mathematician and physicist

Let $\lambda = np = E(x)$, so $p = \dfrac{\lambda}{n}$

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

$$= \frac{n(n-1)\ldots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \sim \frac{n^x}{x!}\left(\frac{\lambda}{n}\right)^x = \frac{\lambda^x}{x!};$$

$$\sum_x \frac{\lambda^x}{x!} = e^{\lambda}.$$

Normalization requires $\sum_x P(X = x) = 1$.

Thus $P(X = x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$

Poisson Mean & Variance

If X is a Poisson random variable, then:

- Mean: $\mu = E(X) = \lambda$ $\simeq n \cdot p$
- Variance: $\sigma^2 = V(X) = \lambda$ $\simeq n \cdot p \cdot (1-p) \approx n \cdot p$
- Standard deviation: $\sigma = \lambda^{1/2}$

Note: Variance = Mean

Note: Standard deviation/Mean = $\lambda^{-1/2}$
decreases with λ

Matlab exercise: Poisson distribution

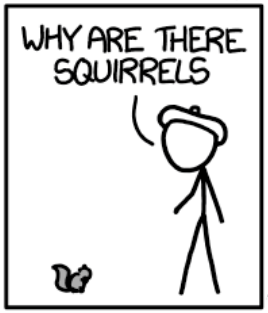
- Generate a sample of size 100,000 for Poisson-distributed random variable X with $\lambda = 2$
- Plot the approximation to the Probability Mass Function based on this sample
- Calculate the mean and variance of this sample and compare it to theoretical calculations:
  $E[X] = \lambda$ and $V[X] = \lambda$

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 IS LYING GOOD

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 SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
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 IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

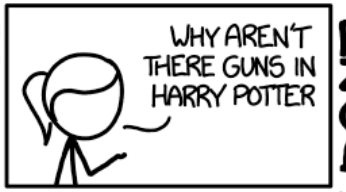
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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

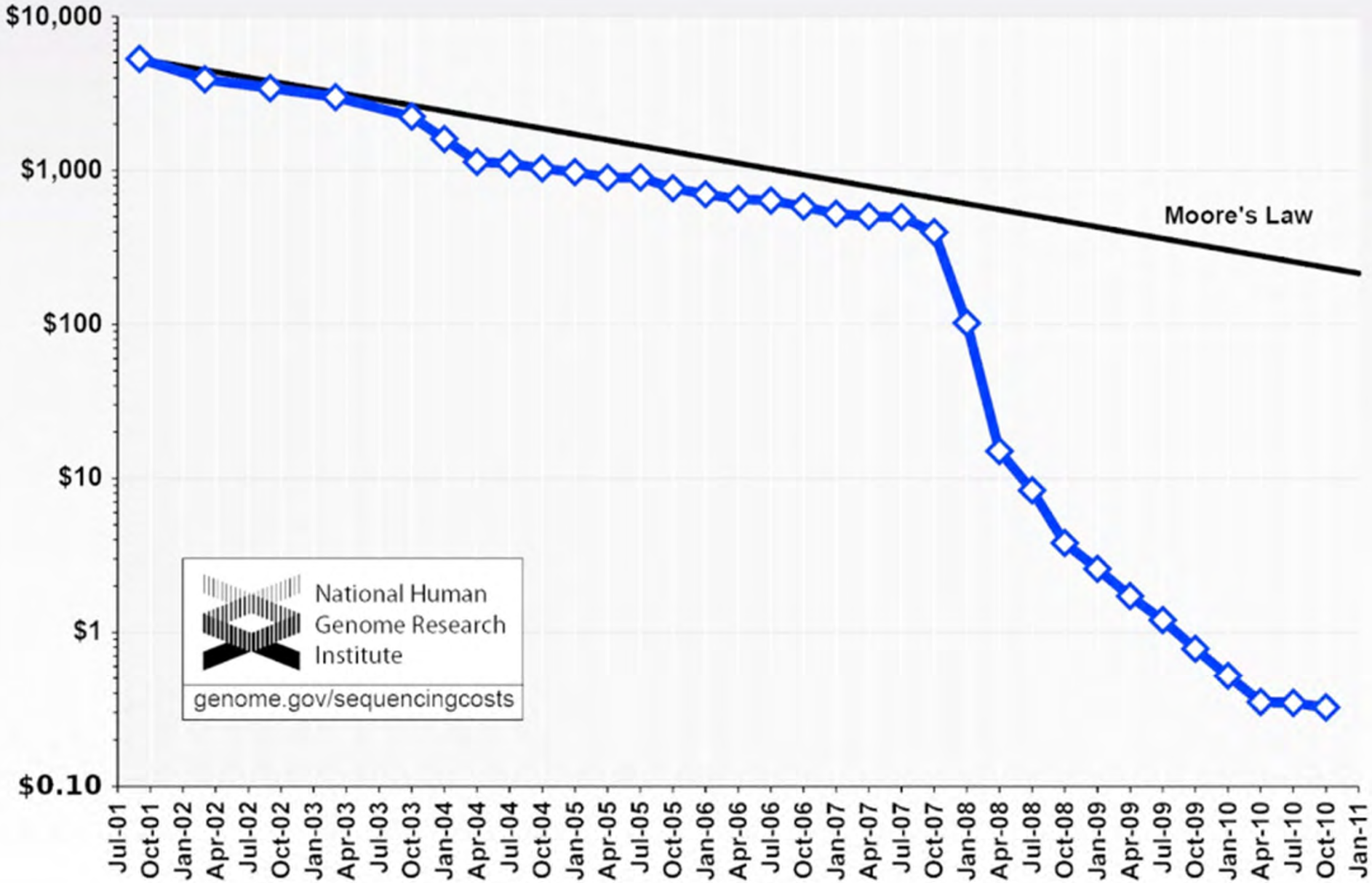
WHY AREN'T THERE GUNS IN HARRY POTTER



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

Poisson Distribution in Genome Assembly

Cost per Megabase of DNA Sequence



Poisson Example: Genome Assembly

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

TABLE 9.1 Next-generation sequencing technologies compared to Sanger sequencing. Adapted from the companies' websites, http://en.wikipedia.org/wiki/DNA_sequencer, and literature cited for each technology.

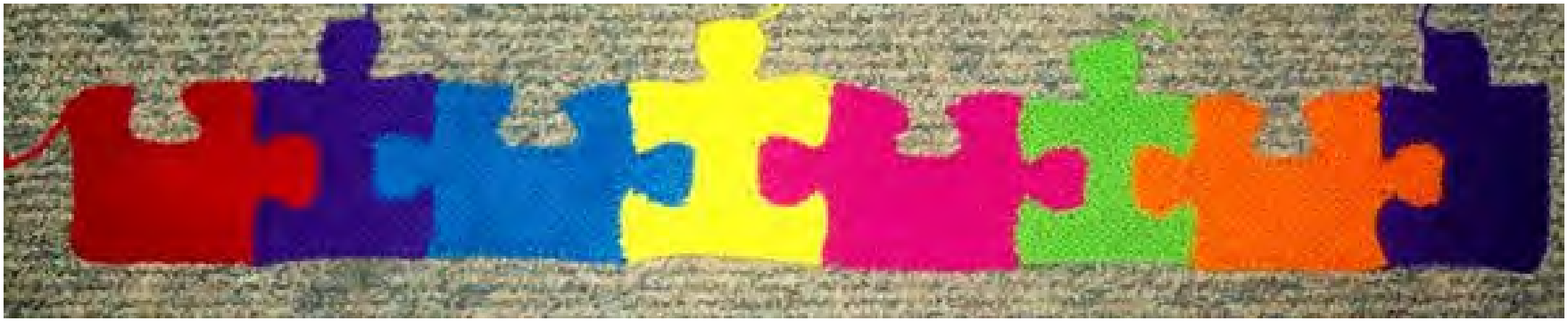
Technology	Read length (bp)	Reads per run	Time per run	Cost per megabase (US$)	Accuracy (%)
Roche 454	700	1 million	1 day	10	99.90
Illumina	50–250	<3 billion	1–10 days	~0.10	98
SOLiD	50	~1.4 billion	7–14 days	0.13	99.90
Ion Torrent	200	<5 million	2 hours	1	98
Pacific Biosciences	2900	<75,000	<2 hours	2	99
Sanger	400–900	N/A	<3 hours	2400	99.90

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

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

Short Reads assemble into Contigs

Figure 5.1.

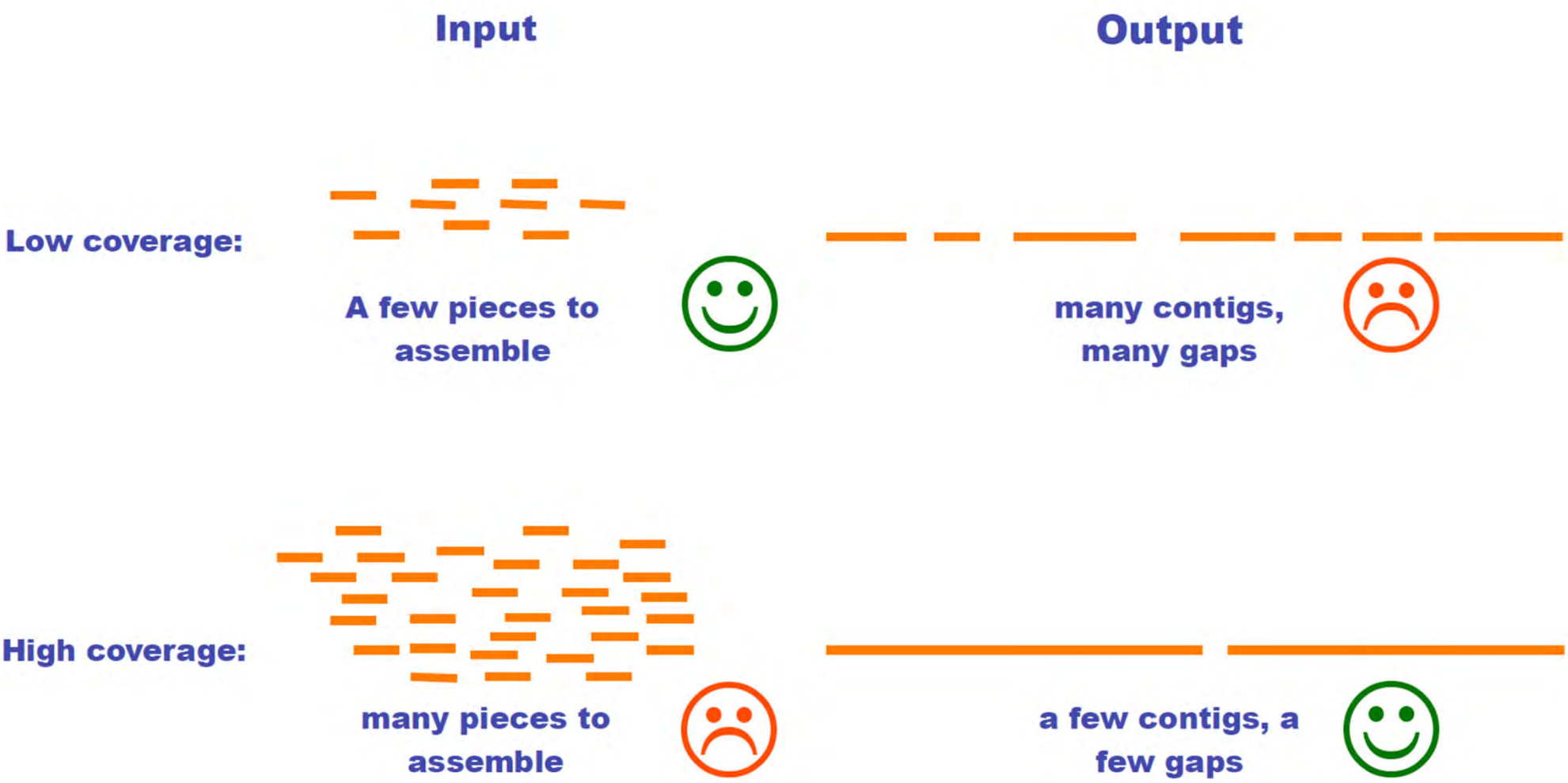


Promise of Genomics

Drew Sheneman, New Jersey -- The Newark Star Ledger, E-mail Drew.

I think I found the corner piece!

How many short reads do we need?



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

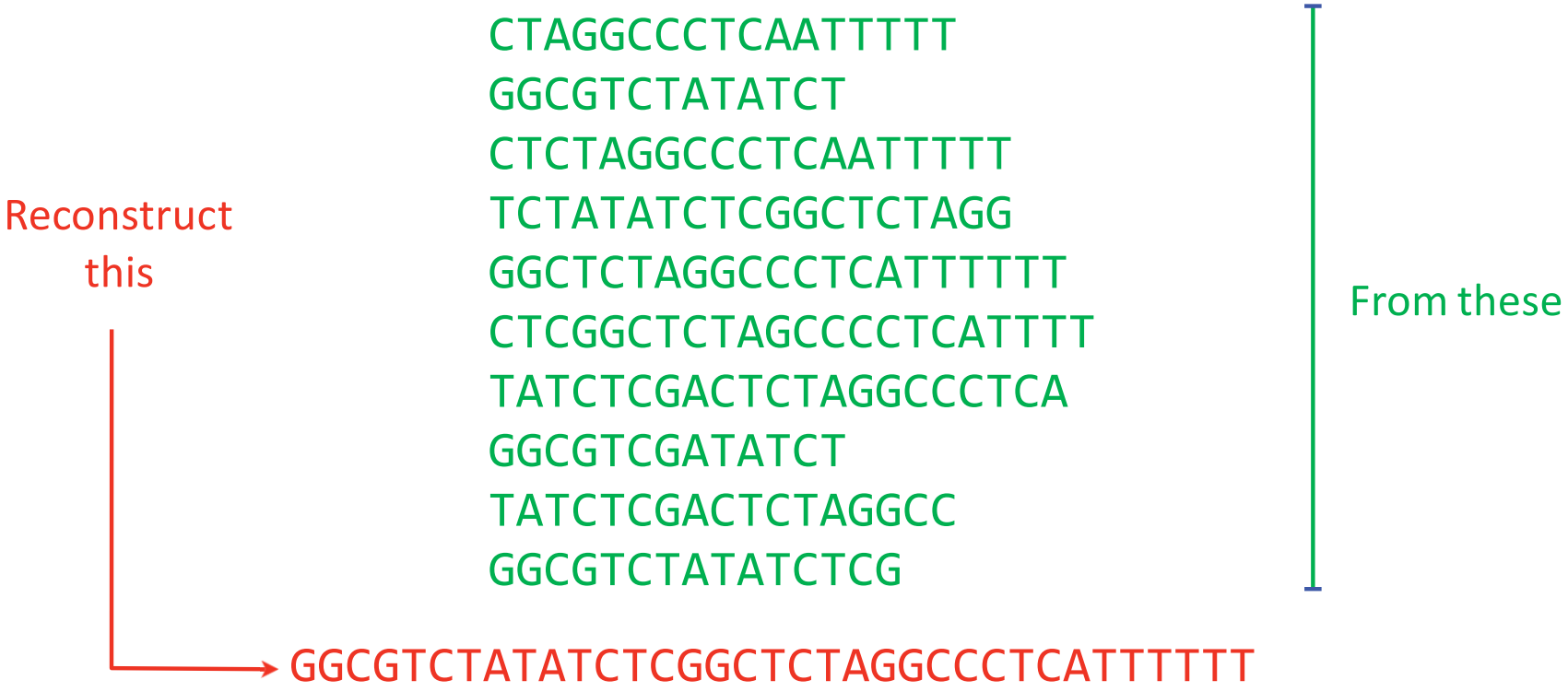
Copy by PCR:
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

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

Assembly

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

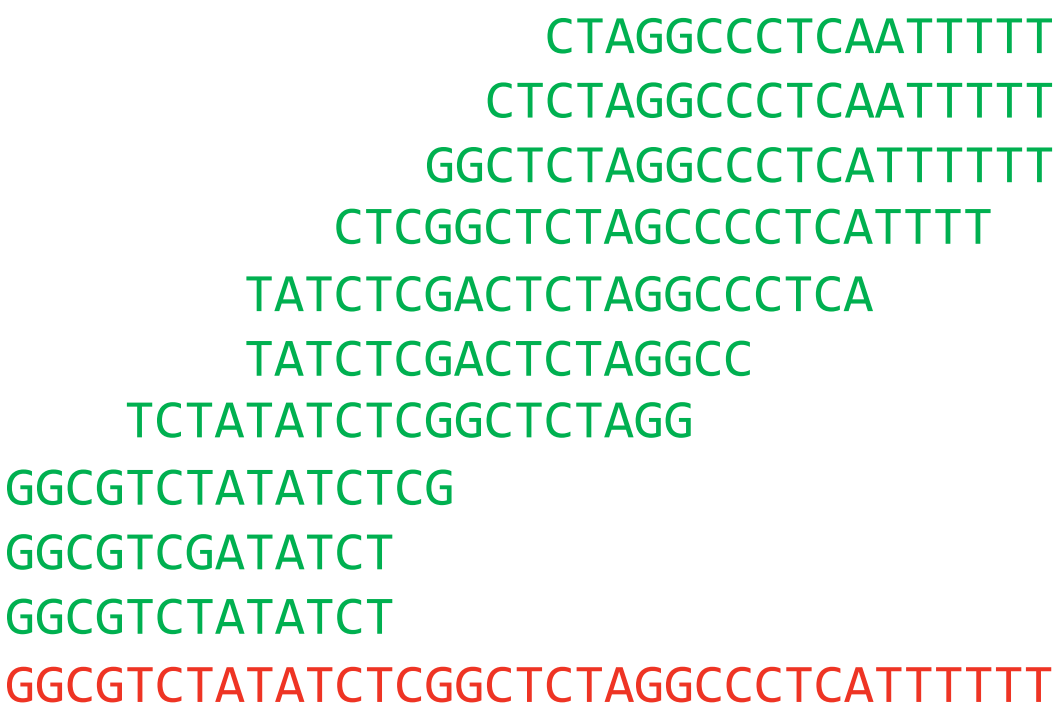
...but we don't know what came from where



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

Assembly

Overlaps between short reads help to put them together

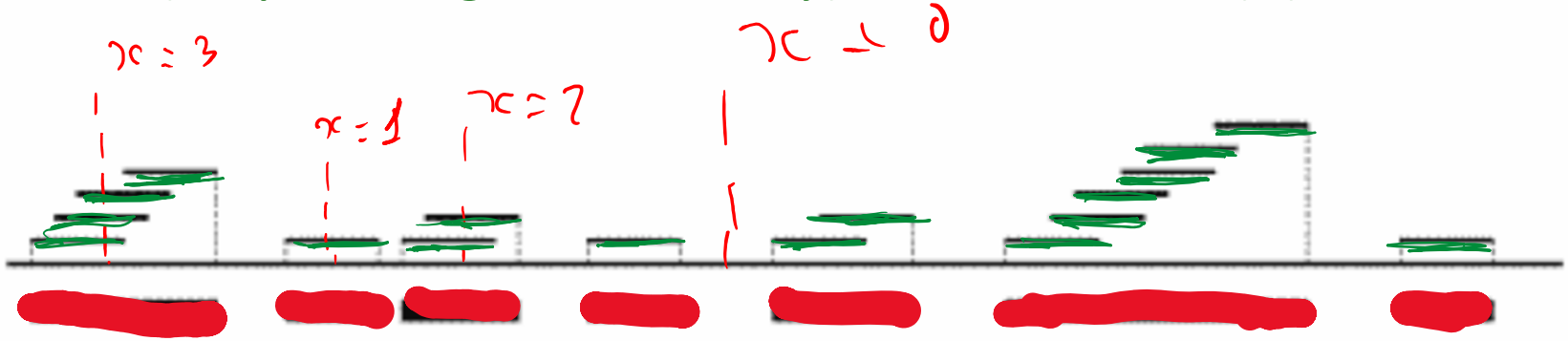


Where is the Poisson?

- *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?

- Coverage: *λ=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(- λ)]*

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 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 it.
- The left end of another short read has the probability $p=(L-1)/G$ to fall within a given read. There are $N-1$ other reads. Hence the expected number of left ends inside a given shot read is $p\cdot (N-1)=(N-1)\cdot(L-1)/G \approx \lambda$
- If significant overlap required to merge two short reads is L_{ov}, modified λ is given by $(N-1)\cdot(L-L_{ov})/G$
- Probability that no left ends fall inside a 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 $L = 500$.

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

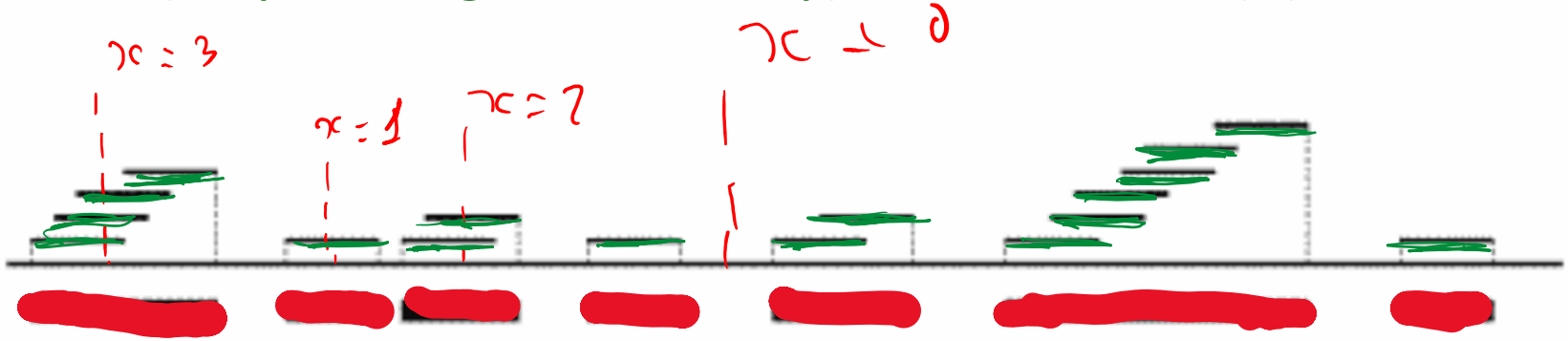
MinION, a palm-sized gene sequencer made by UK-based Oxford Nanopore Technologies

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

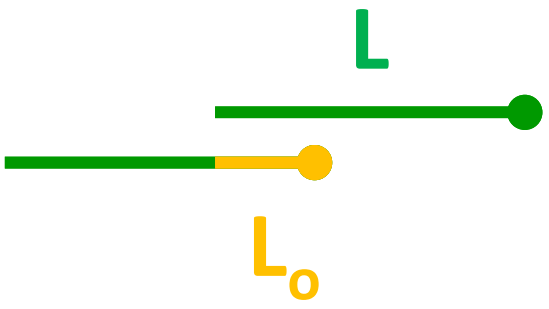
What fraction of genome is covered?

- Coverage: *λ=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(- λ)]*

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?



G

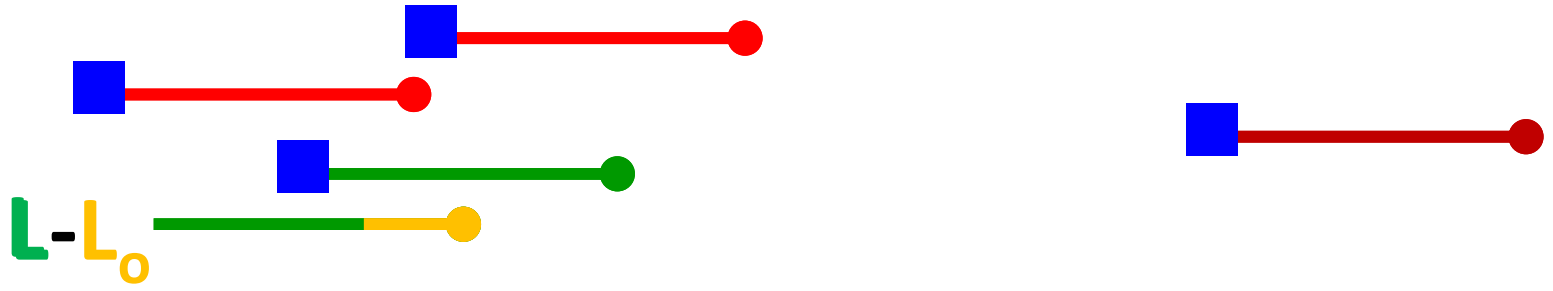
If DNA was a random chain with $p_A = p_C = p_G = p_T = 1/4$

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?

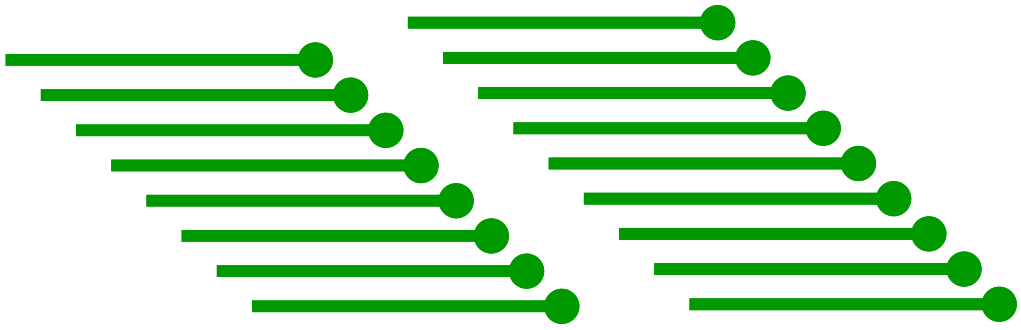


G

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

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

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



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 $L-L_{overlap}$ base pairs
- The left end of another short read has the probability $p=(L-L_{overlap})/G$ to fall within a given read. There are $N-1$ other reads.
- The expected number of left ends inside a given short read is $p\cdot (N-1)=(N-1)\cdot(L-L_{overlap})/G \approx \lambda$ (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 $3x10^9$ 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

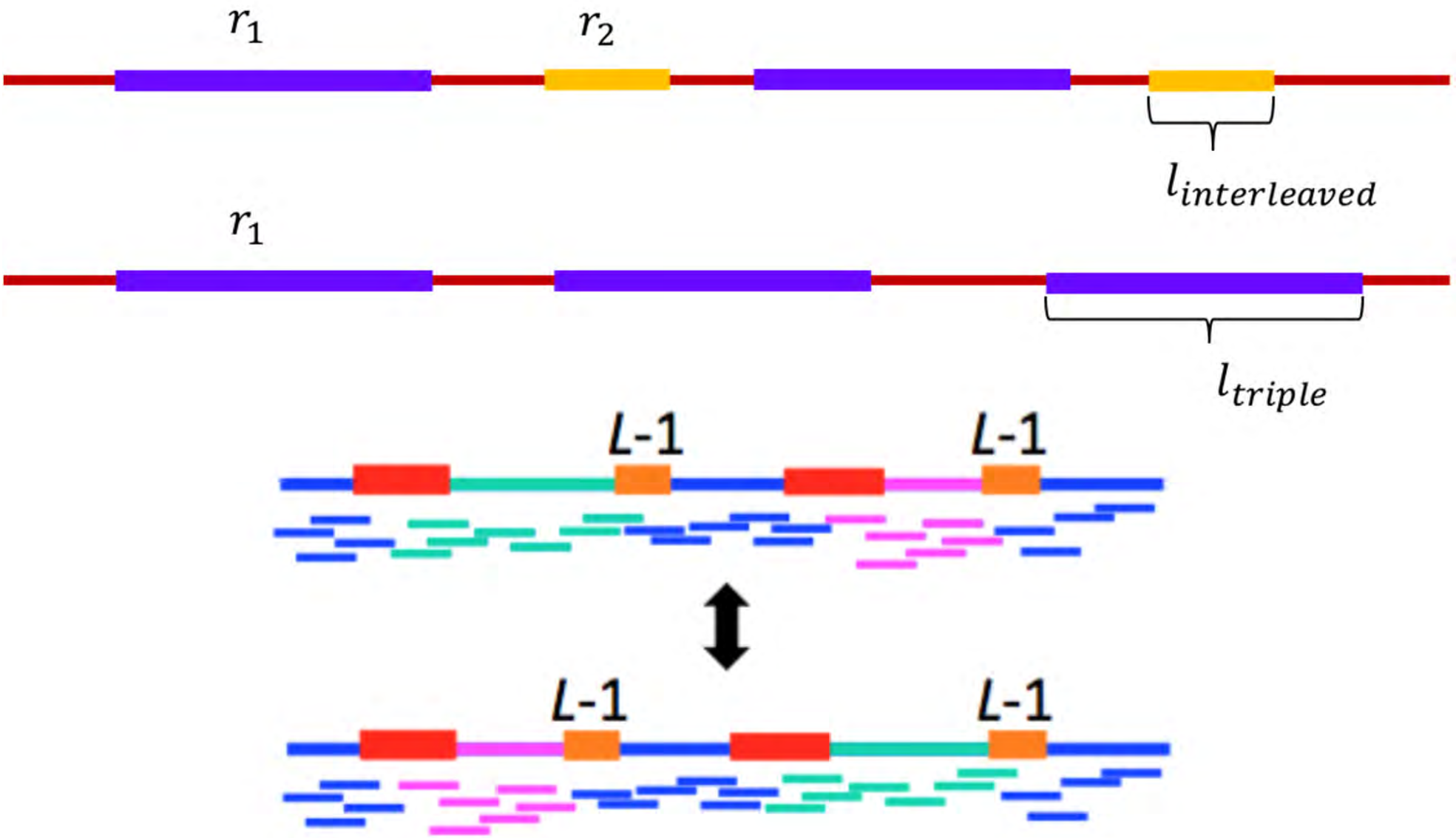
FIGURE 8.11 A BLASTN search of the human genome (all assemblies) database was performed at the NCBI website using TTAGGGTTAGGGTTAGGG as query (i.e., three TTAGGG repeats). There were matches to hundreds of genomic scaffolds. This figure shows an example (NT_024477.14) assigned to the telomere of chromosome 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:

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

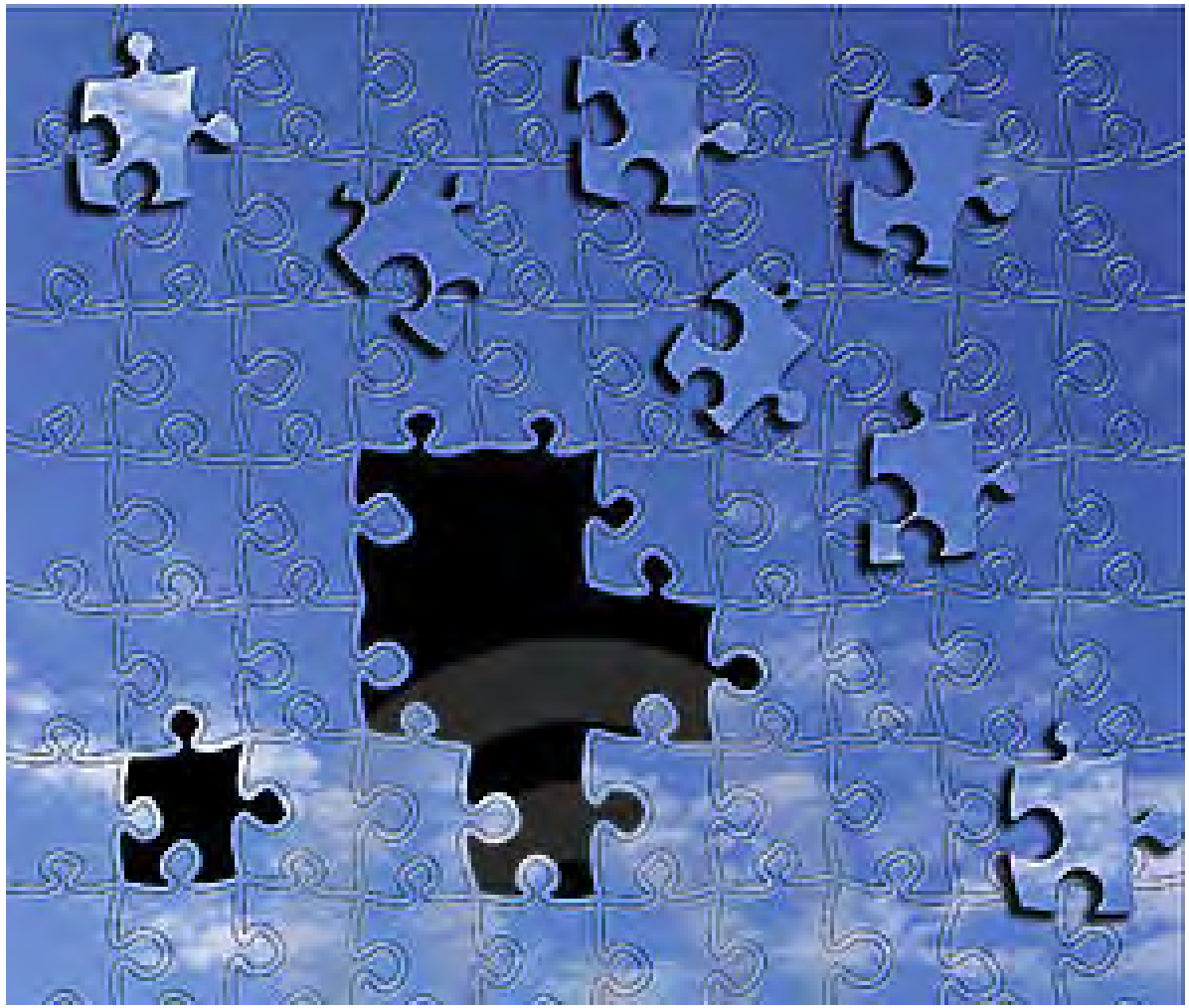
DNA repeats make assembly difficult

Why repeats make assembly difficult?

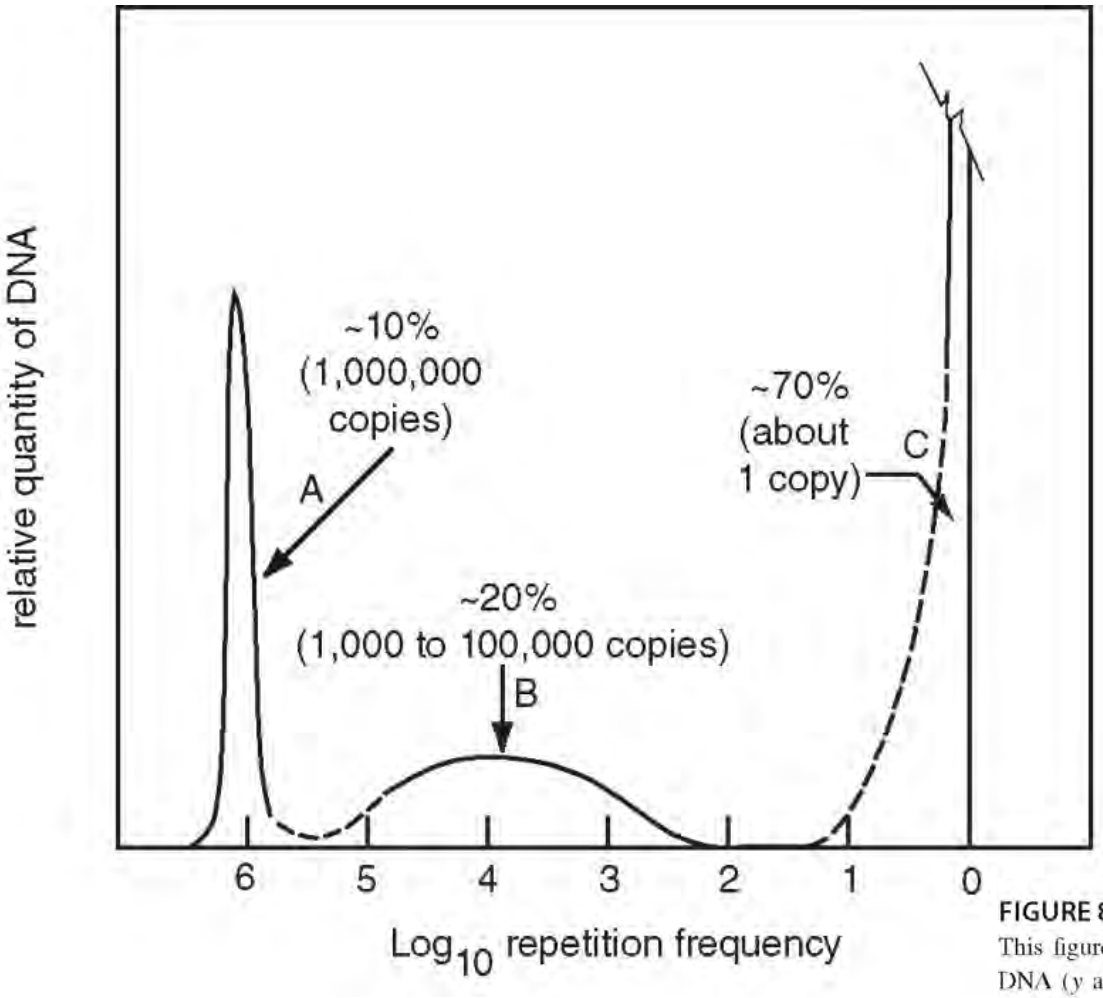


Images from the course EE 372: Data Science for High-Throughput Sequencing. taught by David Tse at Stanford

Repeats are like sky puzzle pieces

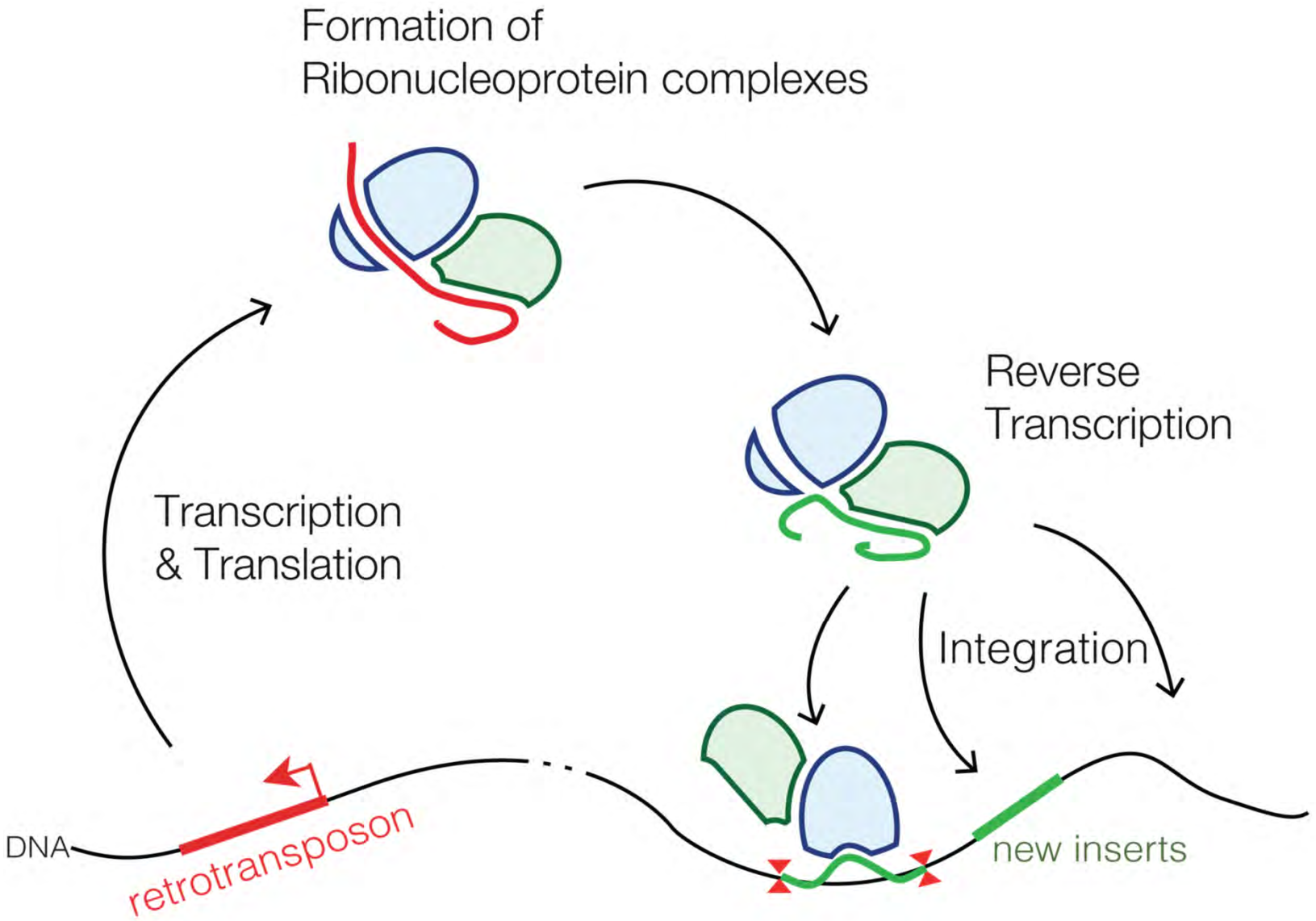


How many repeats are in eukaryotic genomes?

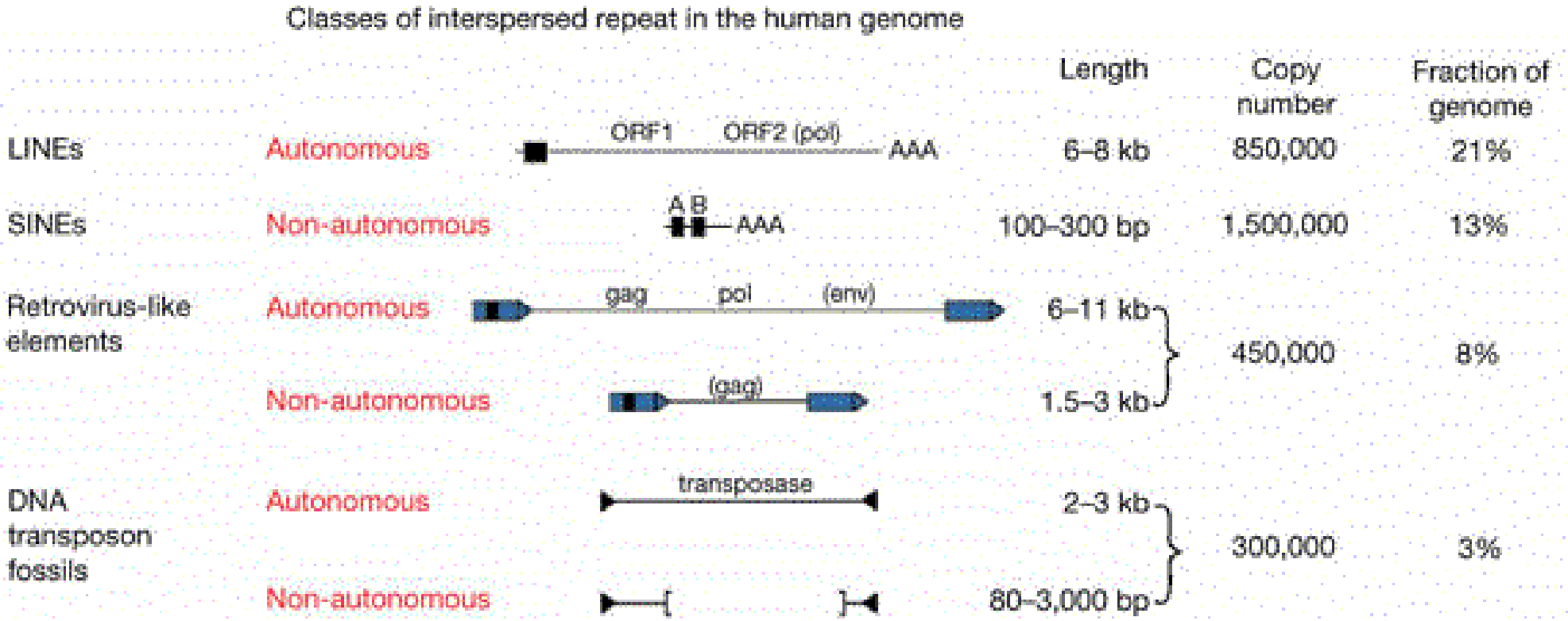


Data for mouse genome obtained in 1961 (sic!) using DNA denaturation and renaturation curves

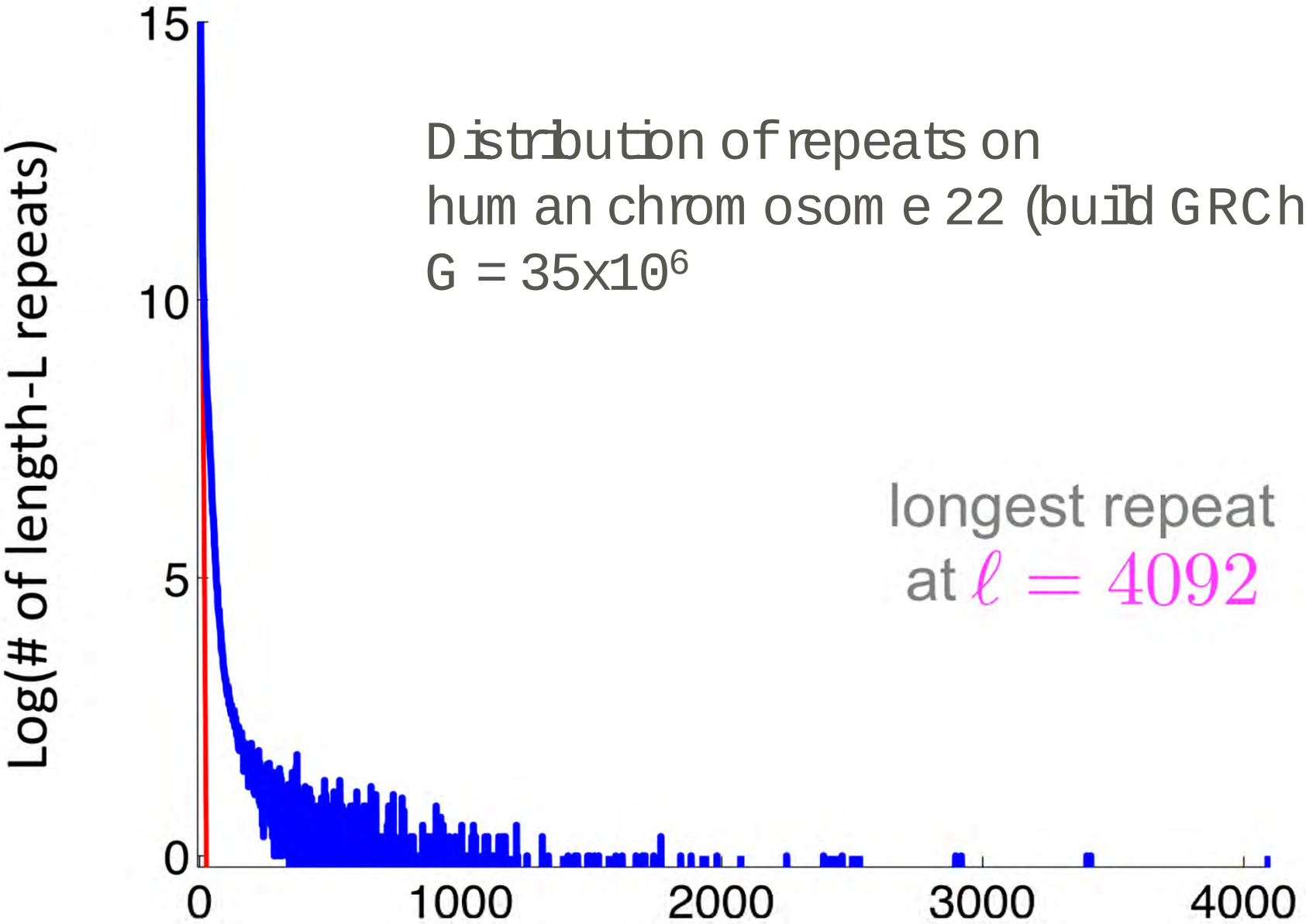
FIGURE 8.6 The complexity of genomic DNA can be estimated by denaturing then renaturing DNA. This figure (redrawn from Britten and Kohne, 1968) depicts the relative quantity of mouse genomic DNA (y axis) versus the logarithm of the frequency with which the DNA is repeated. The data are derived from a $C_0 t_{1/2}$ 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.


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.



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

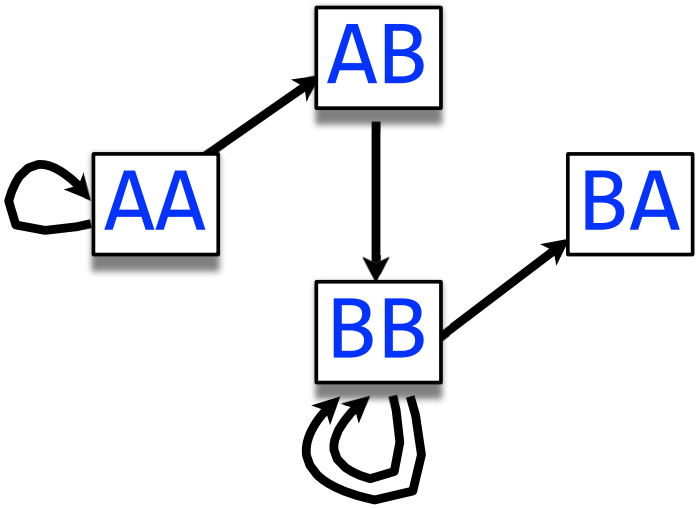
Nicolaas Govert de Bruijn (1918 – 2012) was a Dutch mathematician, noted for his many contributions in the fields of graph theory, analysis, number theory, combinatorics and logic

De Bruijn graph

genome: AAABBBBA

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

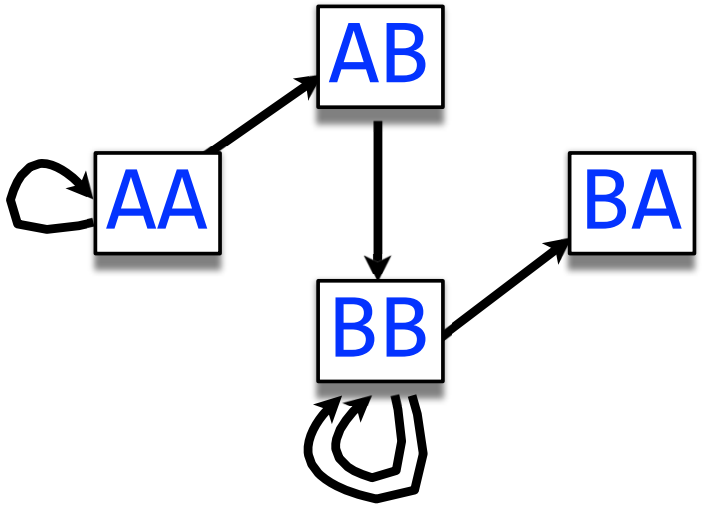
L/R 2-mers: AA, AA AA, AB AB, BB BB, BB BB, BB BB, BA



One edge per every *k*-mer

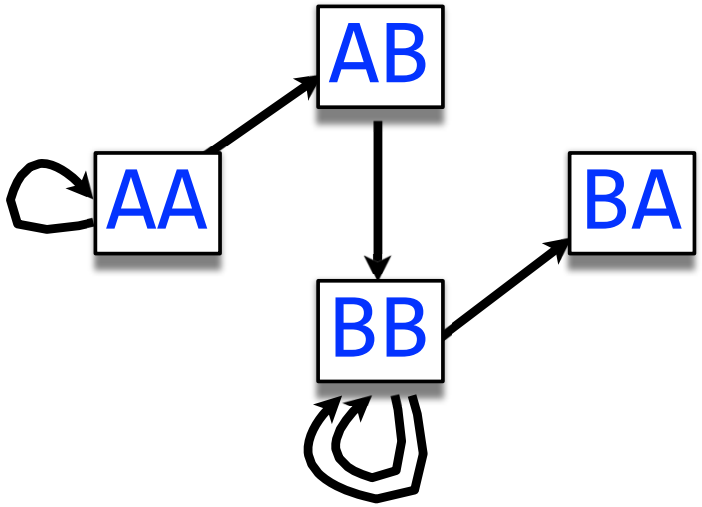
One node per distinct *k*-1-mer

De Bruijn graph



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

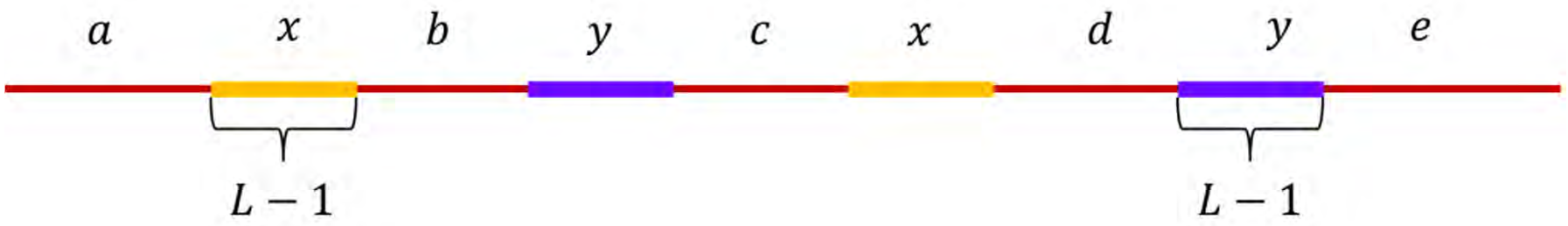
Assembly = Eulerian walk on De Bruijn graph

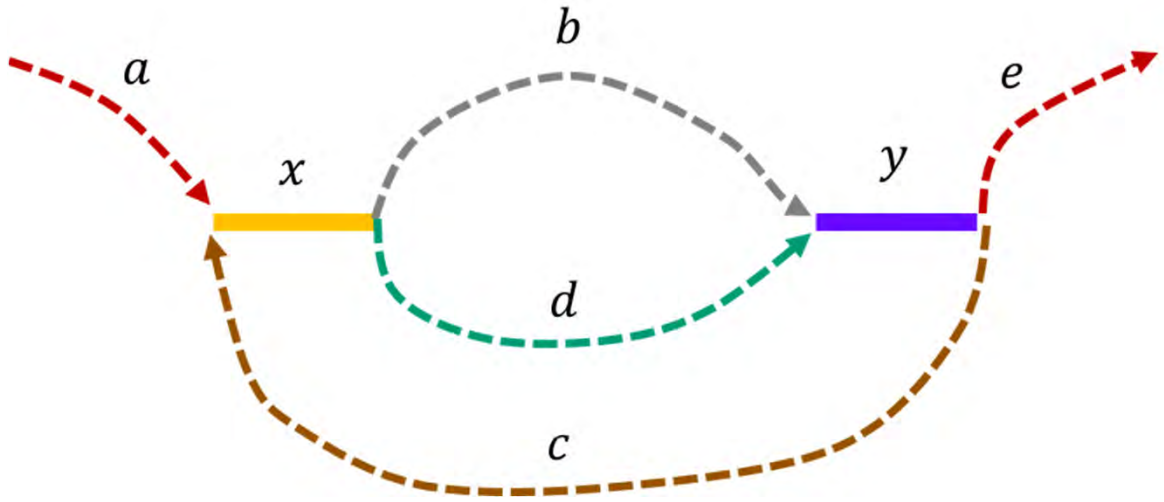


AAABBBBA

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

Why interleaved repeats are dangerous?

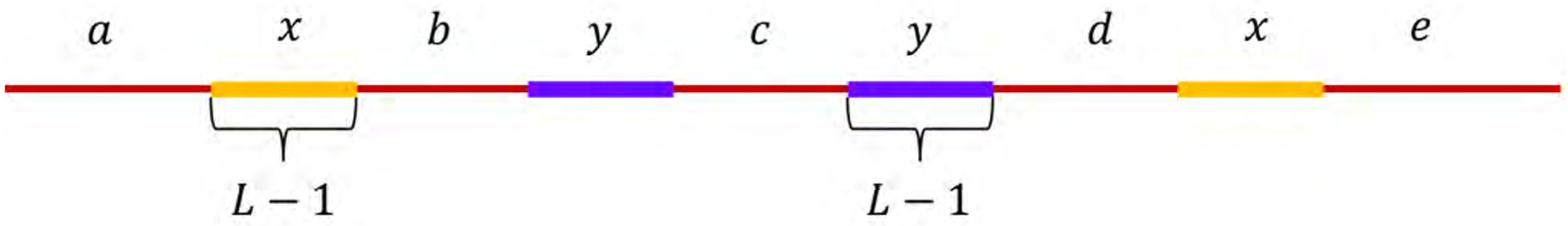


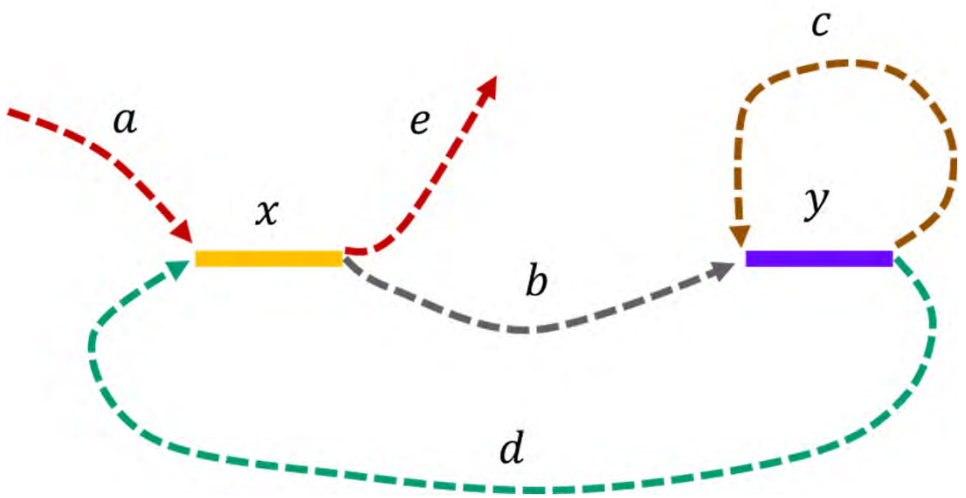


The two Eulerian paths that are on the graph:
a–x–b–y–c–x–d–y–e and a–x–d–y–c–x–b–y–e

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

Why non-interleaved repeats are safe?

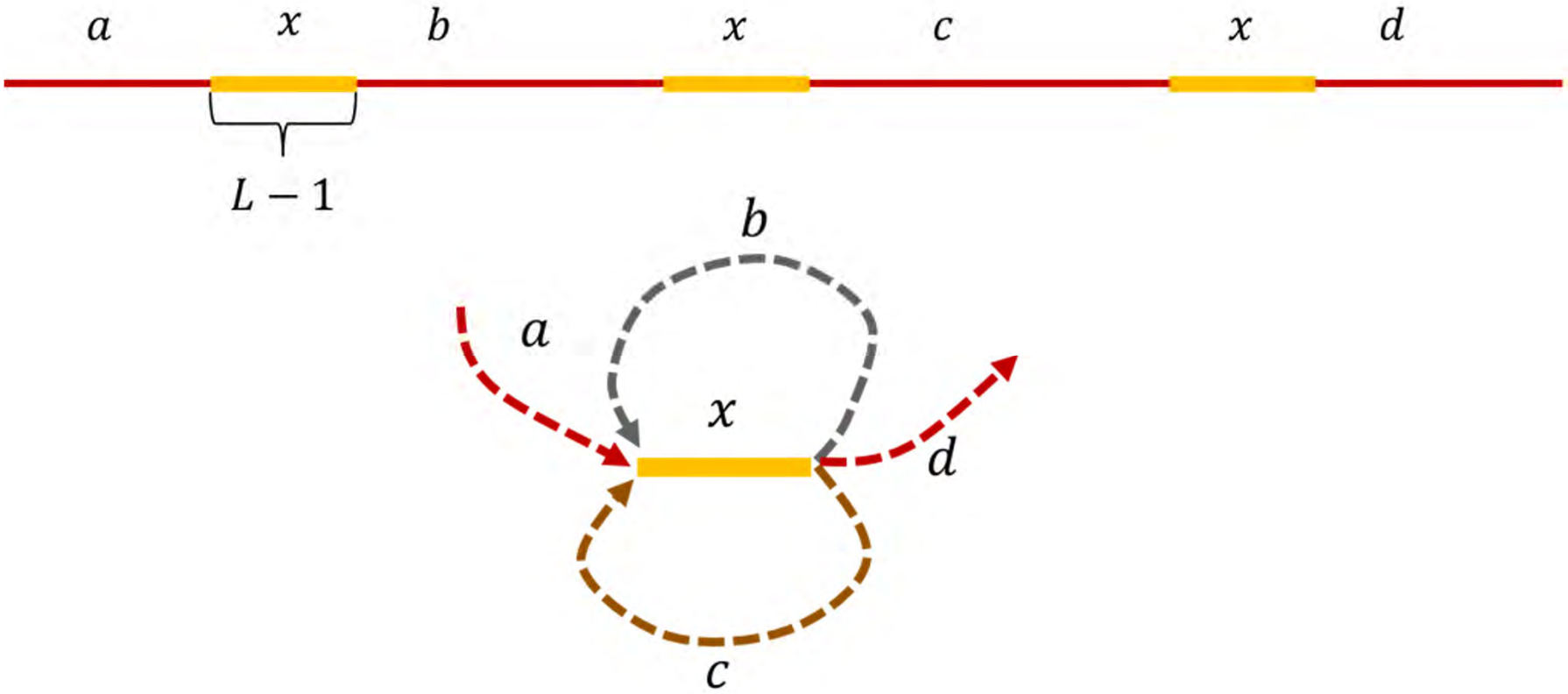




The only Eulerian path is: a–x–b–y–c–y–d–x–e

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

Why triple repeats are dangerous?

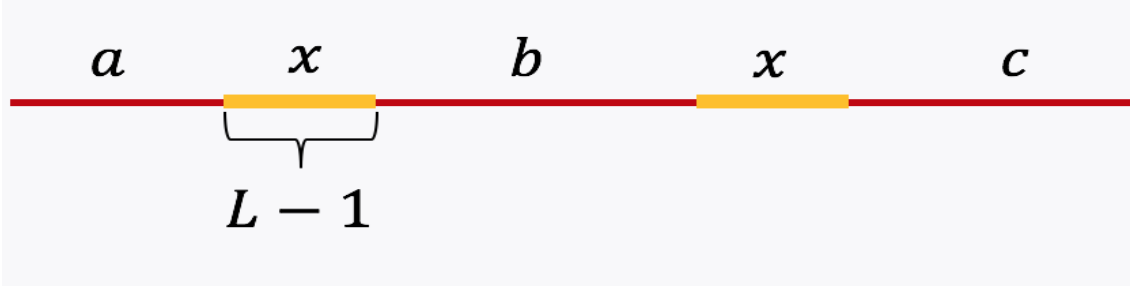


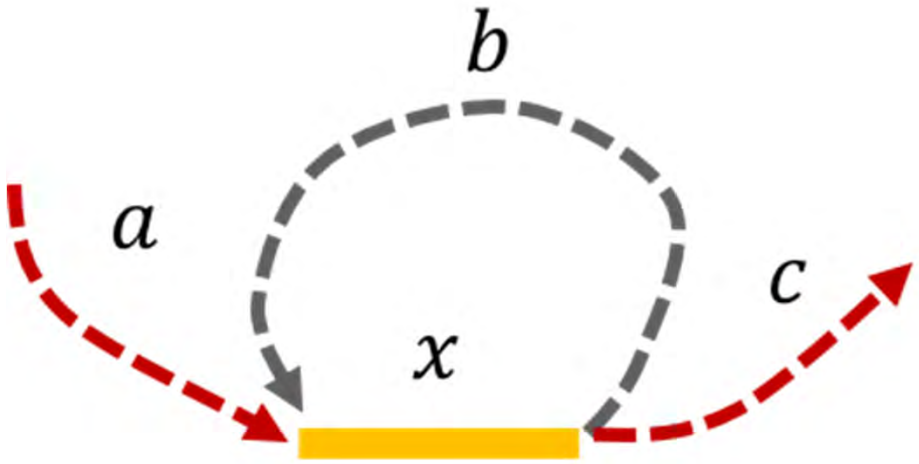
The two Eulerian paths that are on the graph:

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

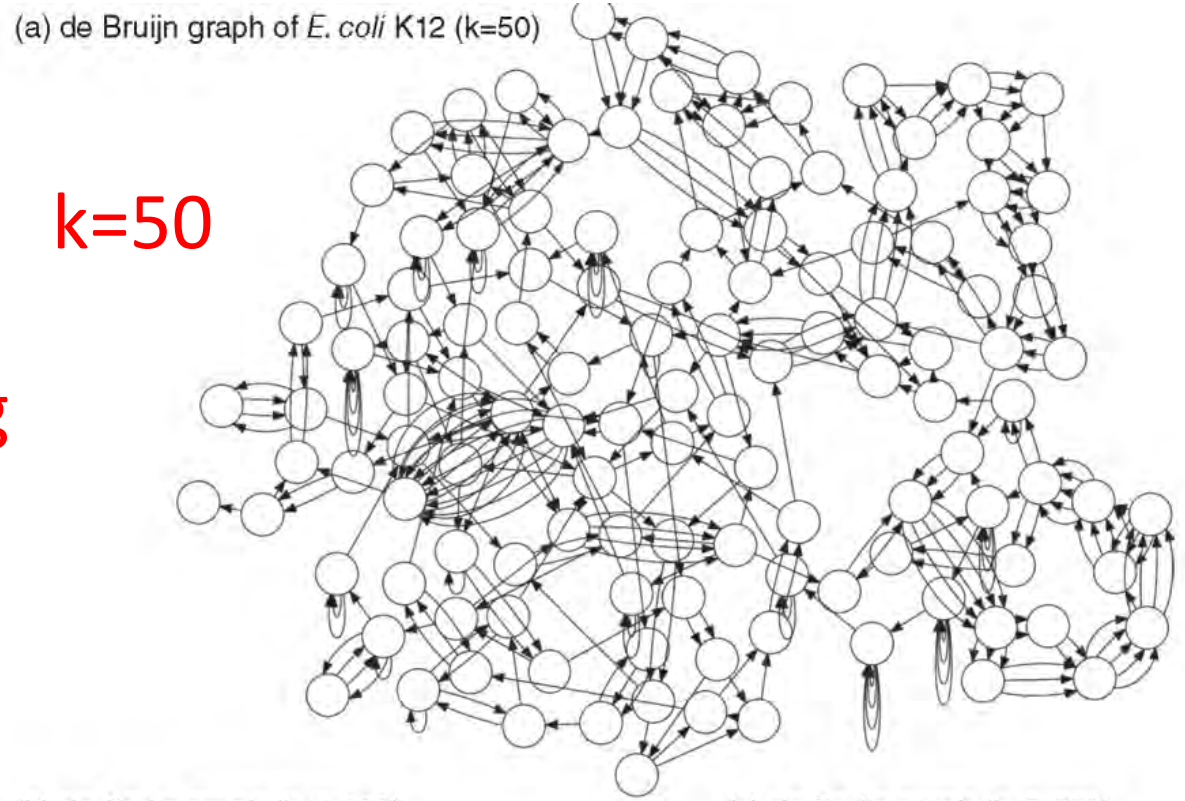
- **Theorem [Pevzner 1995]**:
  If L, the read length, is strictly greater than $\max(\ell_{interleaved}, \ell_{triple})$, then the de Bruijn graph has a unique Eulerian path corresponding to the original genome.

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.

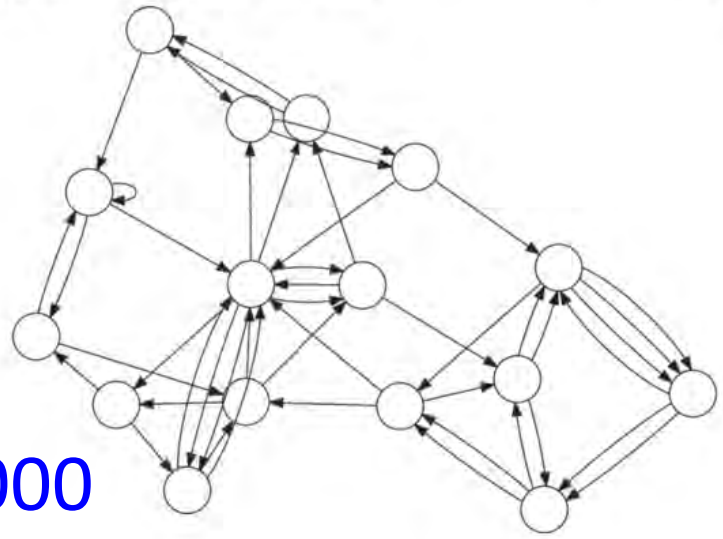
How to assemble a genome with repeats?

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

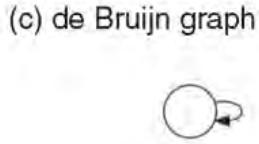
Technology	Read length (bp)
Roche 454	700
Illumina	50–250
SOLiD	50
Ion Torrent	400
Pacific Biosciences	>10,000



k=50



k=1000



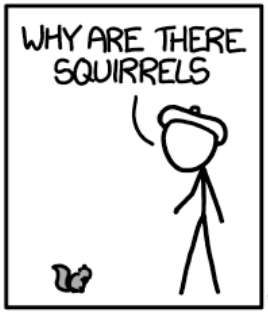
k=5000

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 IS LYING GOOD

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 SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
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 IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

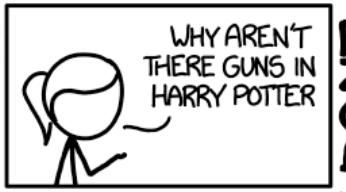
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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

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

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

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

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

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

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

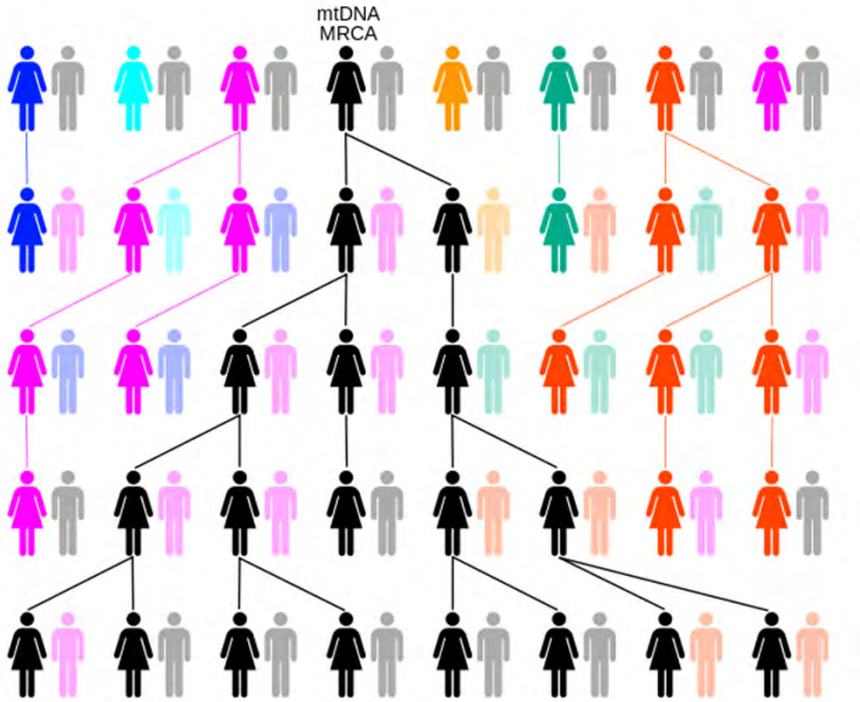
- Very different from Poisson, where it is

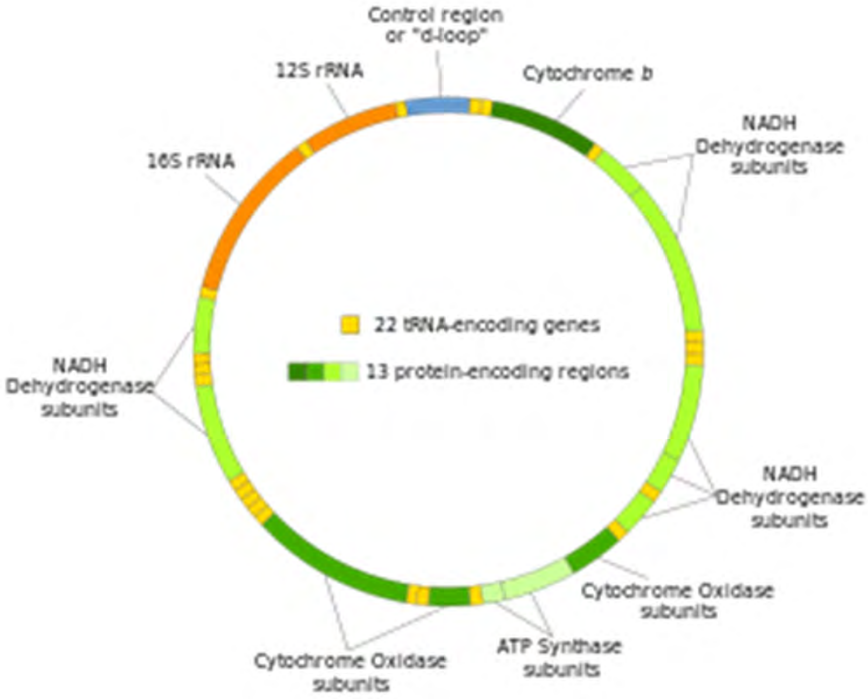
variance = mean and standard deviation = mean$^{1/2}$

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

Geometric distribution in biology

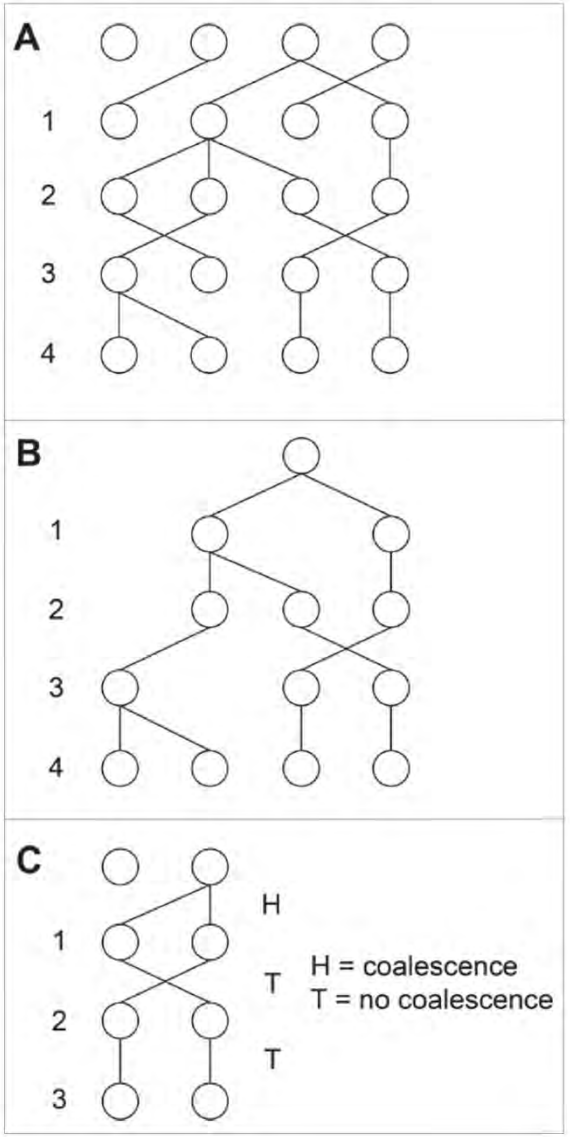




- Each of our cells has mitochondria with 16.5kb of mtDNA **inherited only from our mother**
- Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA
- Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeaon (of UIUC's Carl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

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.

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

- 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*,

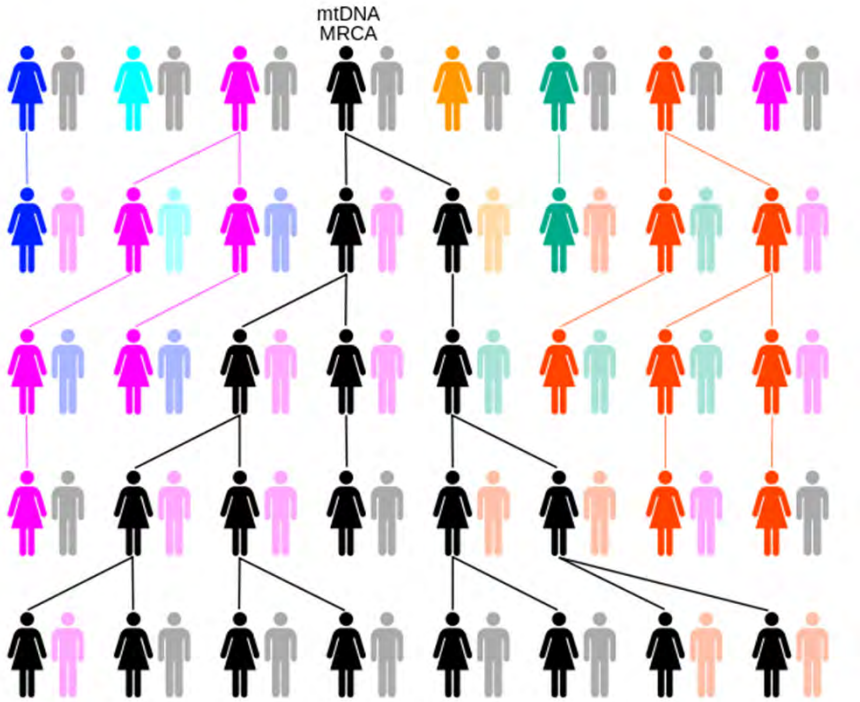
$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \qquad (3\text{-}10)$$

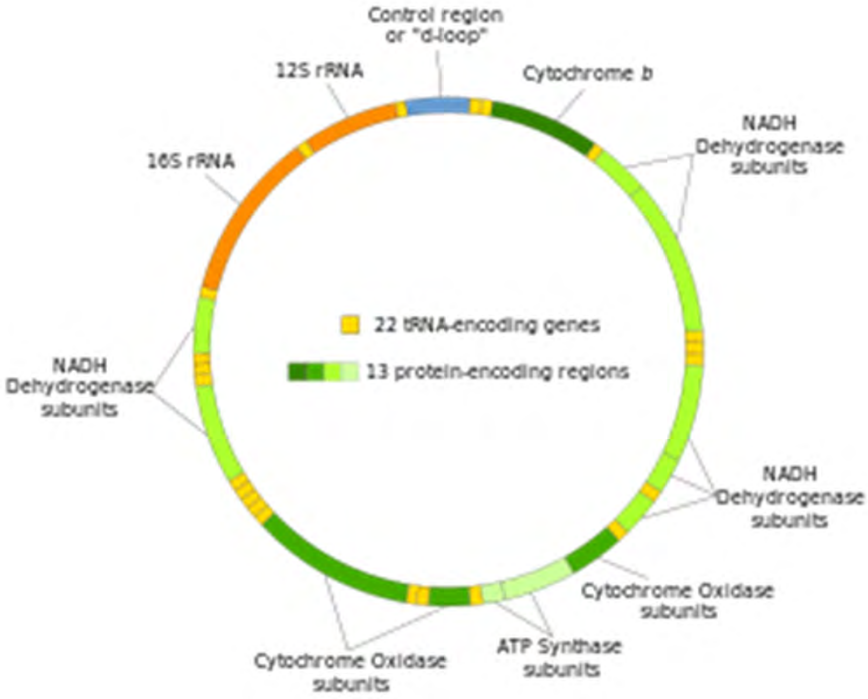
- For small *p* the standard deviation $=(1\text{-}p)^{0.5}/p$ ~= mean=1/p

- Very different from Binomial and Poisson, where

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

Geometric distribution in biology

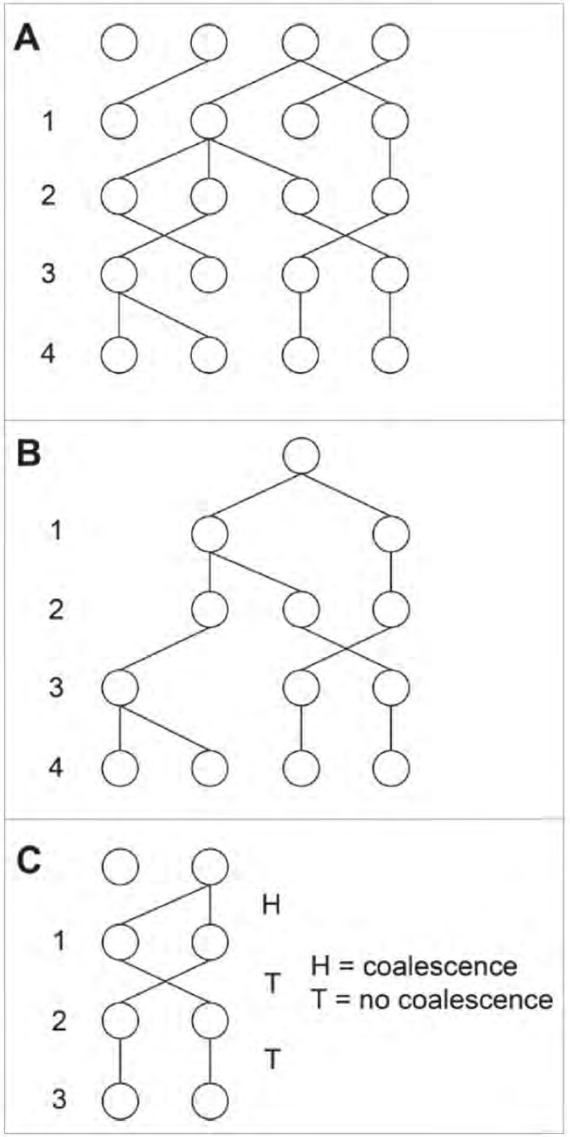




- Each of our cells has mitochondria with 16.5kb of mtDNA **inherited only from our mother**
- Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA
- Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeaon (of UIUC's Carl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

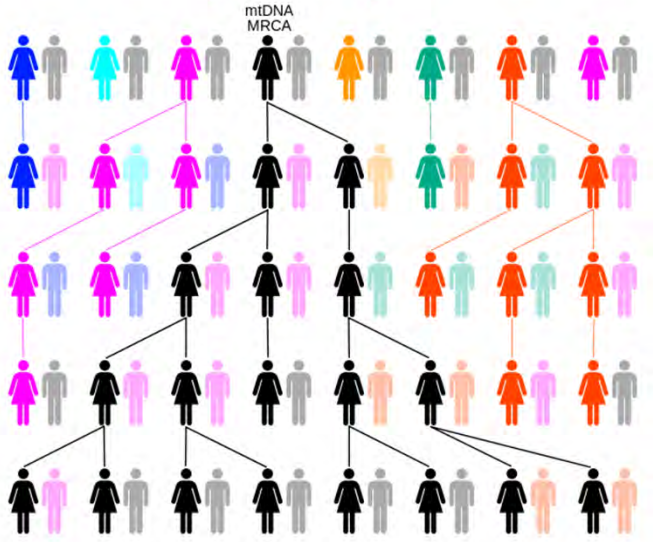
- Constant population of N women
- Random number of (female) offsprings. Average is 1 (but can be 0 or 2)
- Randomly pick two women.
  Question: how many generations T since their last maternal ancestor?
- T is a random variable What is its PMF: P(T=t)?
  Answer: P(T=t) follows a geometric distribution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

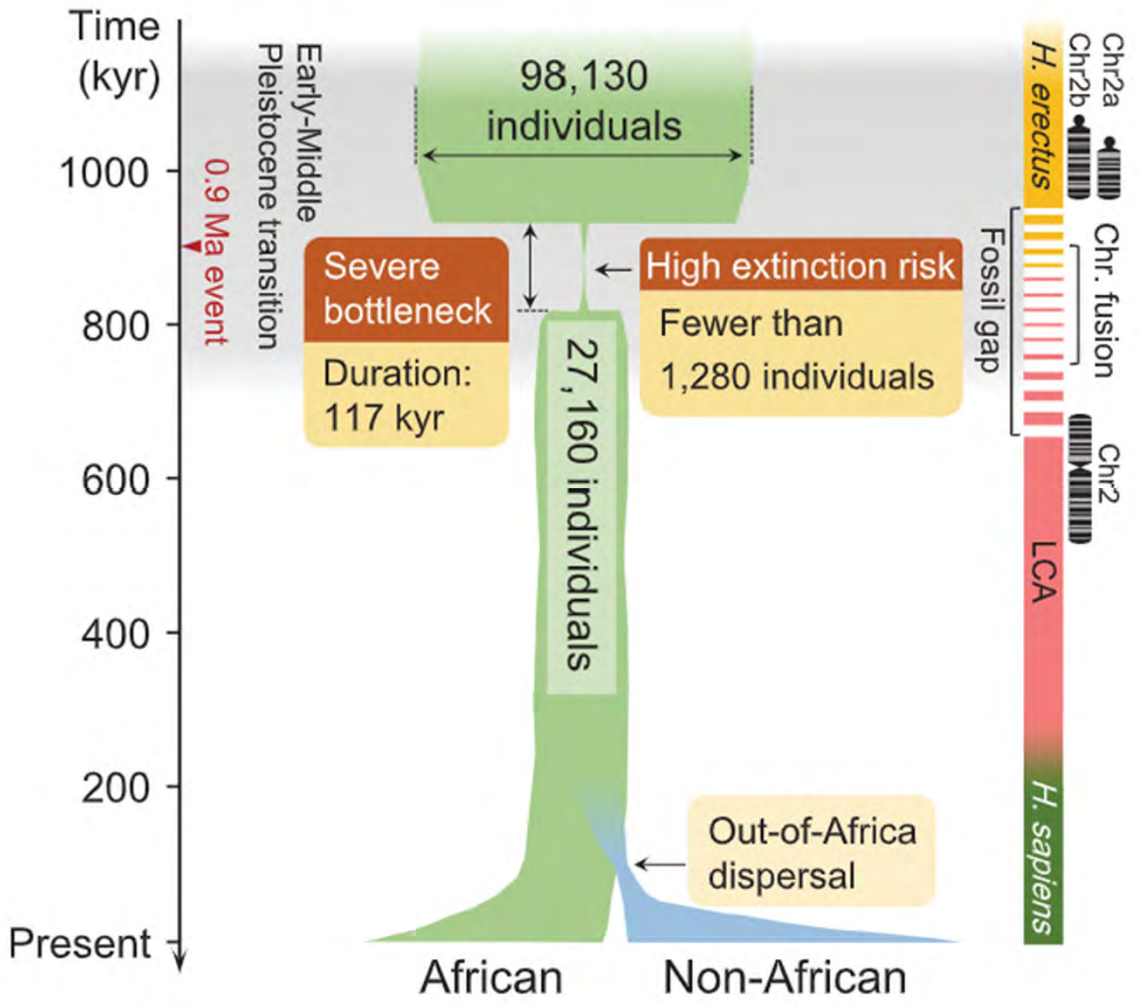
Total time until the MRCA

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

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

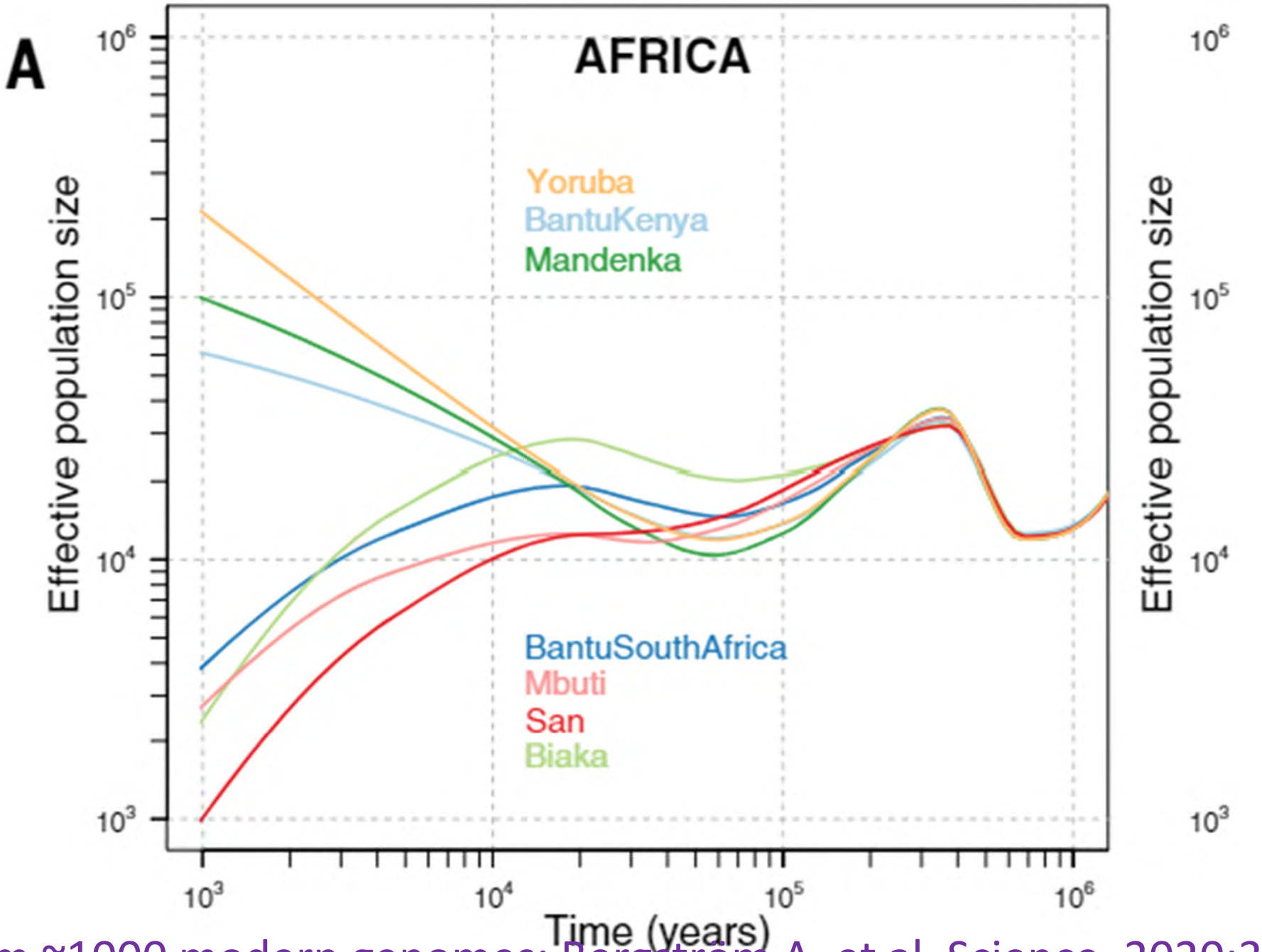
- There are about $N=3.5x10^{9}$ women living today
- Most Recent maternal Common Ancestor (MRCA)
  of all people living today lived $T_{MRCA}=2N$ generations ago
- $T_{MRCA} = 2 \cdot 3.5x10^{9}$ generations
- If the generation time 20 years it is 140 billion years > 10 times the time since the Big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

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



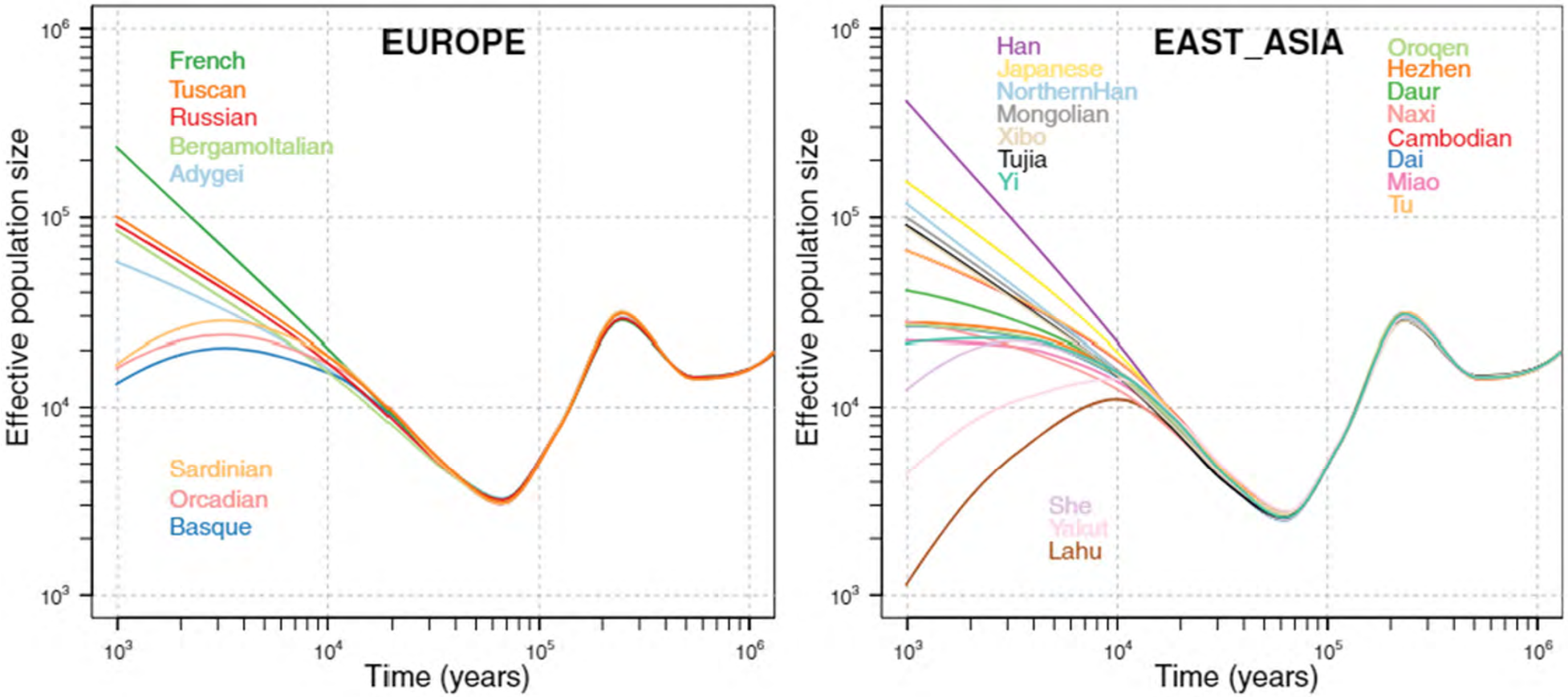
Hu W, et al. Science. 2023;381: 979–984

Effective human population size ~10,000



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

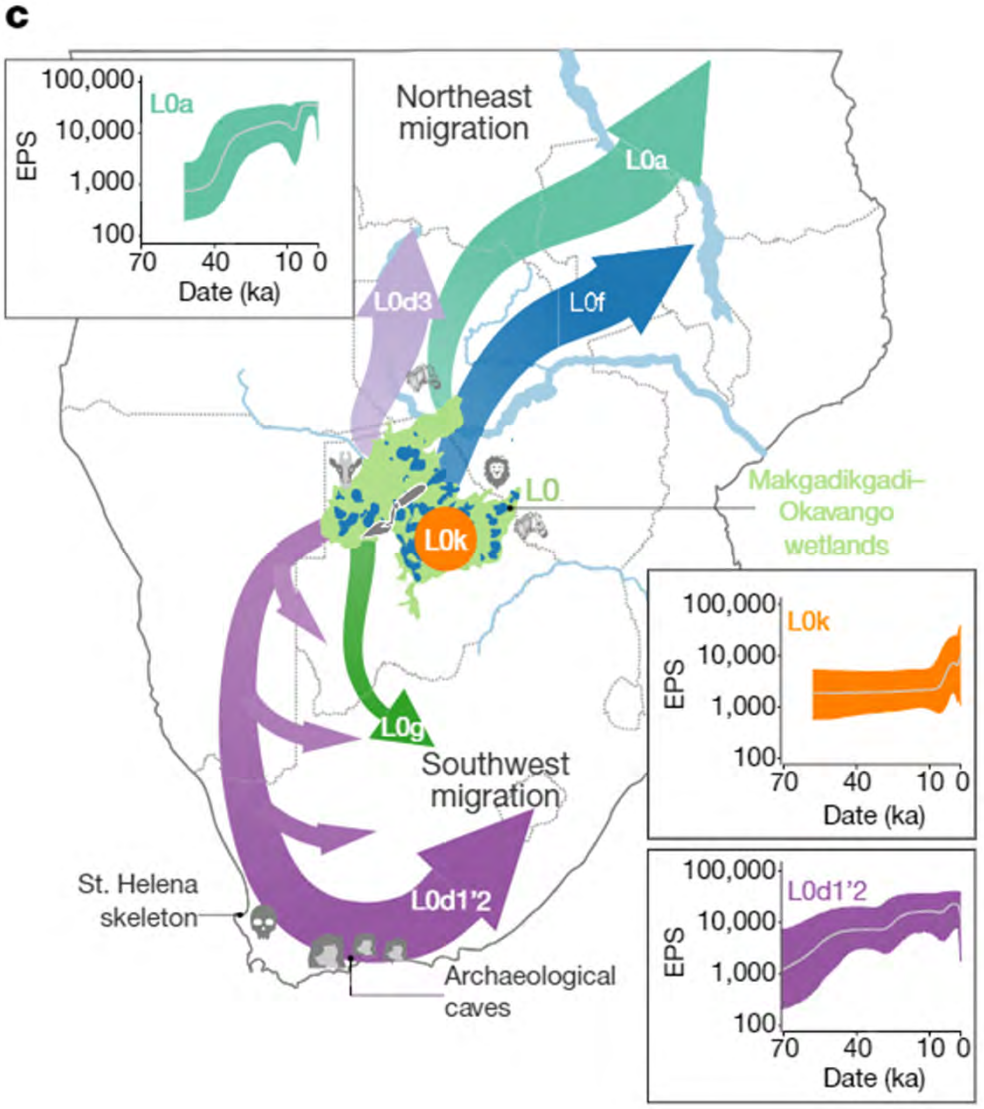
Effective human population size in Europe and Asia ~3000 people ~60,000 years ago



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

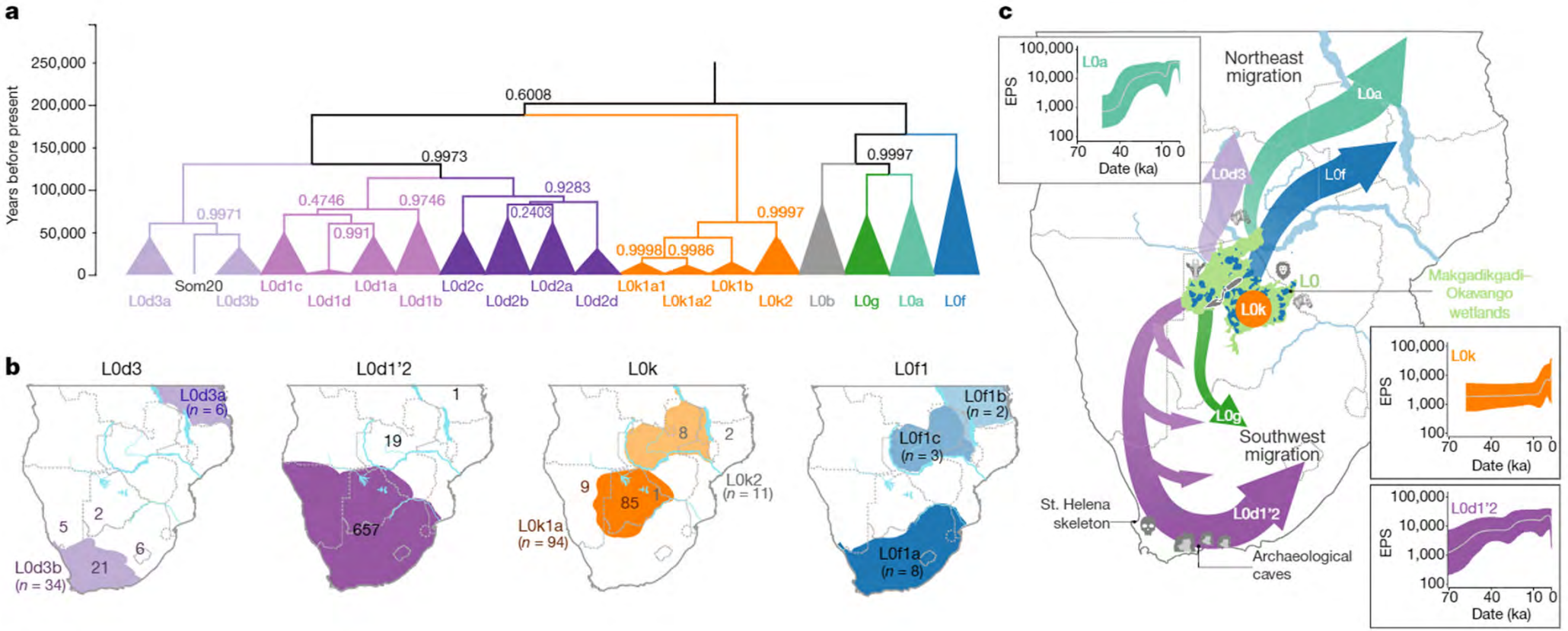


“Mitochondrial Eve” lived in Makgadikgadi–Okavango paleo-wetland of southern Africa ~200,000 years ago (between 165,000 and 240,000 years ago)

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

Okavango Delta now

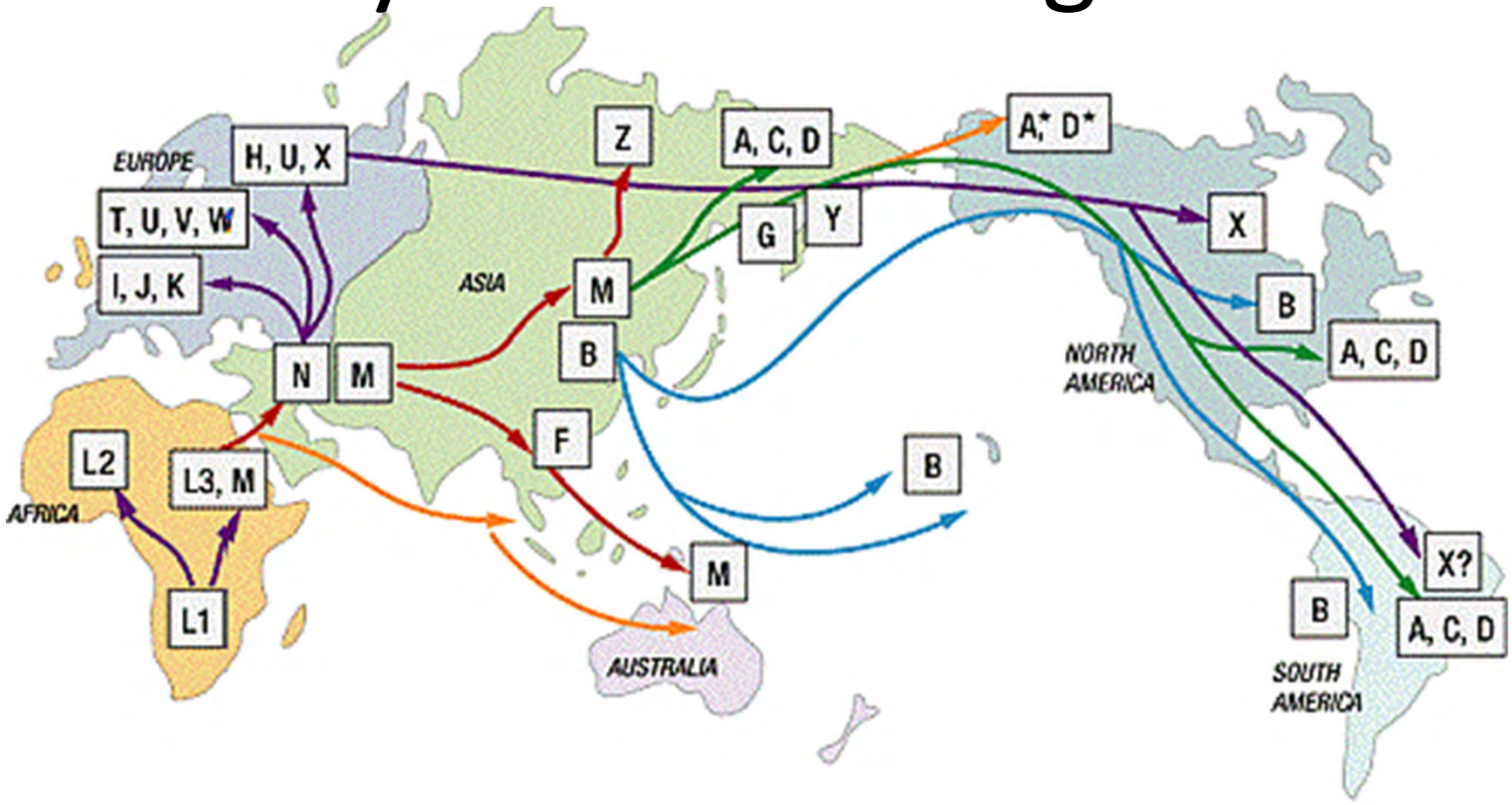
“Mitochondrial Eve” lived in Africa



“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

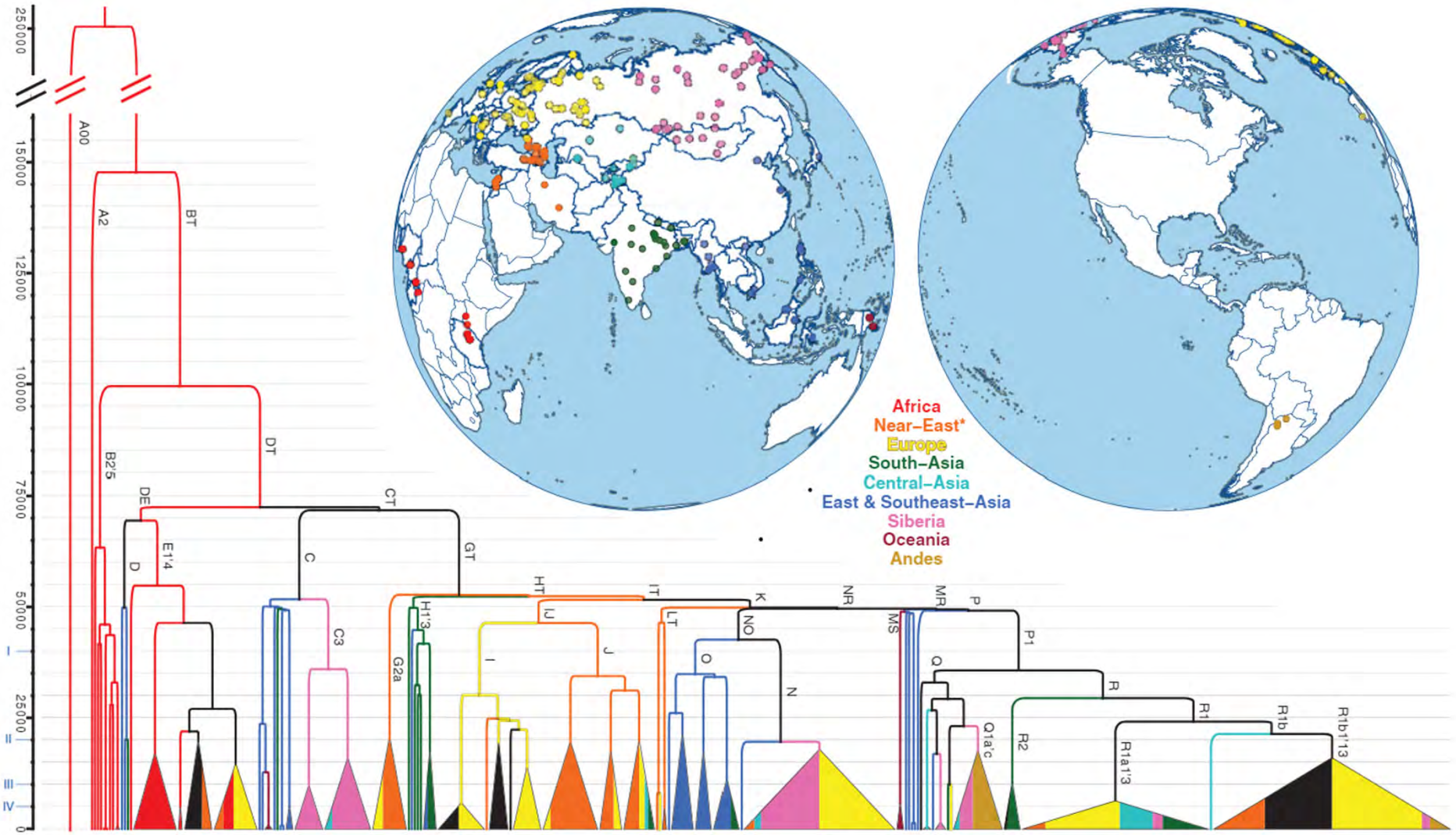
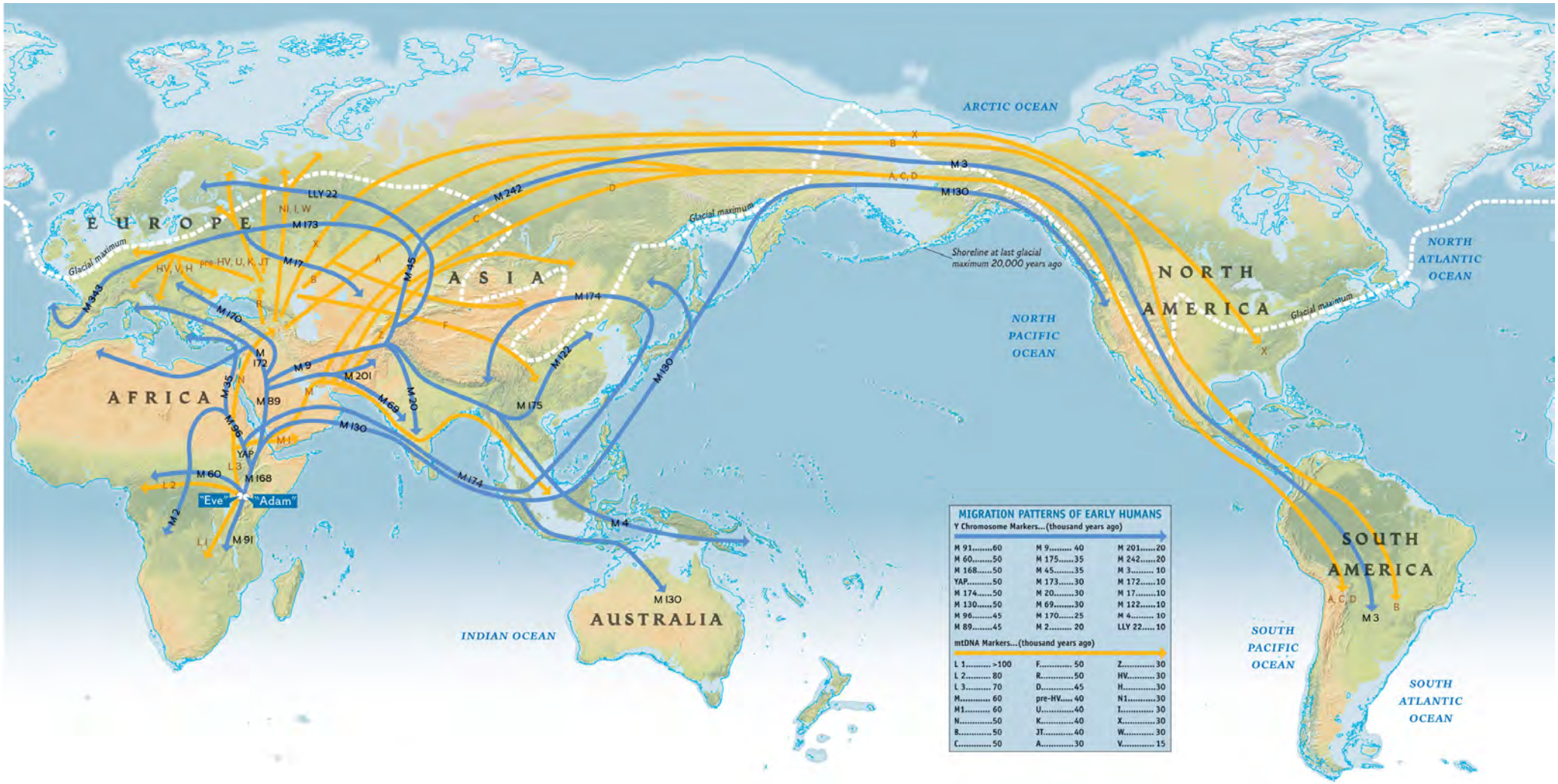


Figure 1. The phylogenetic tree of 456 whole Y chromosome sequences and a map of sampling locations. The phylogenetic tree is reconstructed using BEAST. Clades coalescing within 10% of the overall depth of the tree have been collapsed. Only main haplogroup labels are shown (details are provided in Supplemental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

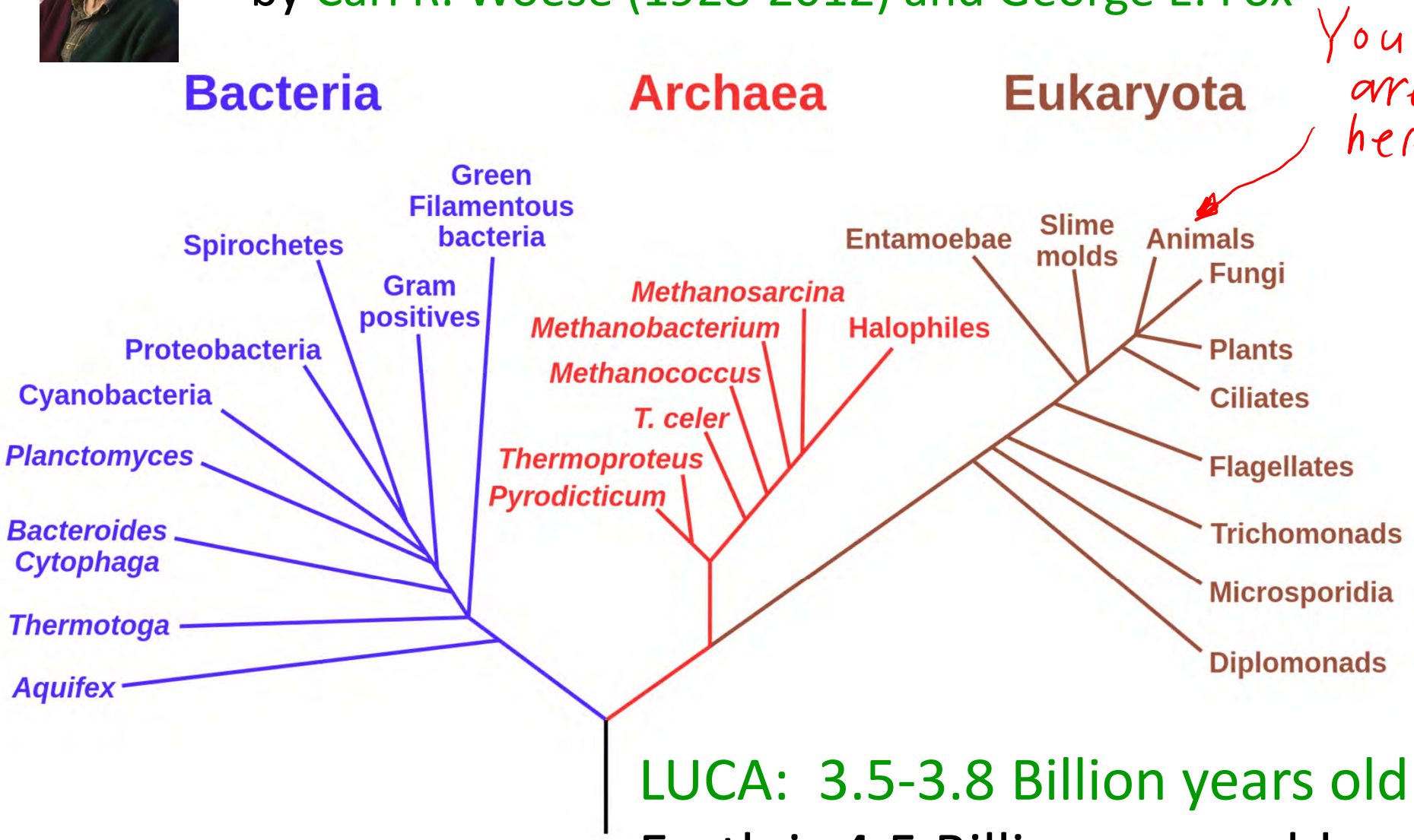
“Adam” and “Eve” both lived in Africa



- “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).*

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

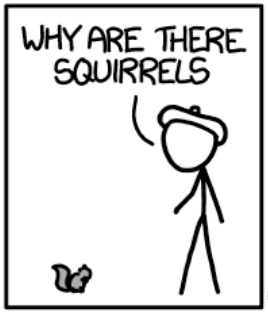
Earth is 4.5 Billion years old

Credit: XKCD comics

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 IS LYING GOOD

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 SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
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 IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

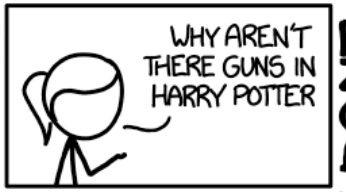
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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

Negative Binomial Definition

- In a series of independent trials with constant probability of success, p, let the random variable X denote the number of trials until r successes occur. Then X is a negative binomial random variable with parameters:
  $0 < p < 1$ and r = 1, 2, 3, ...
- The probability mass function is:

$$f(x) = C_{r-1}^{x-1} p^r (1-p)^{x-r} \quad \text{for } x = r, r+1, r+2... \qquad (3\text{-}11)$$

- Compare it to binomial

$$f(x) = C_x^n p^x (1-p)^{n-x} \quad \text{for } x = 1, 2, ... \text{ n}$$

NOTE OF CAUTION: Matlab, Mathematica, and many other sources use x to denote the number of failures until one gets r successes.
We stick with Montgomery-Runger.

Negative Binomial Mean & Variance

- If *X* is a negative binomial random variable with parameters *p* and *r*,

$$\mu = E(X) = \frac{r}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{r(1-p)}{p^2} \qquad (3\text{-}12)$$

- Compare to geometric distribution:

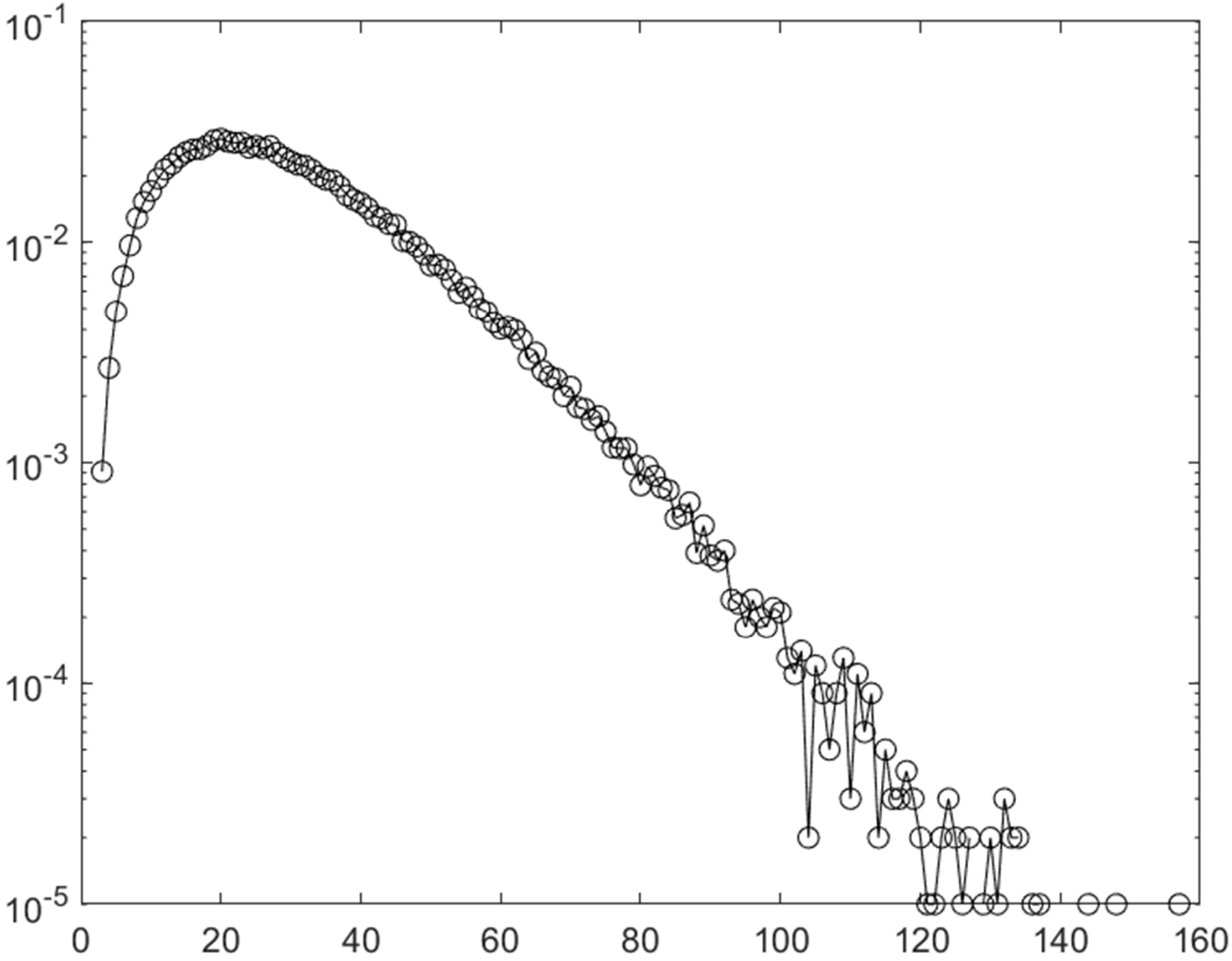
$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \qquad (3\text{-}10)$$

Matlab exercise

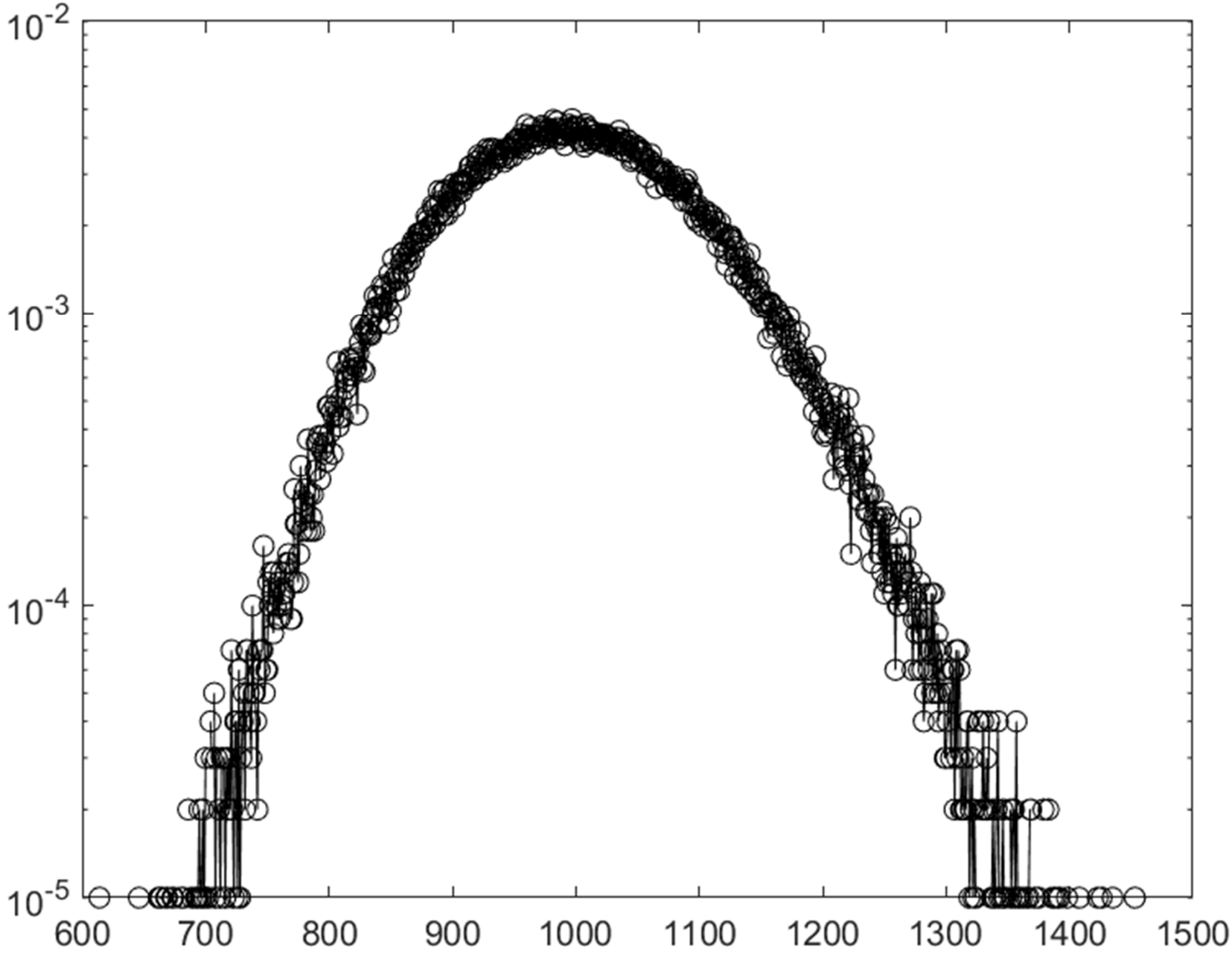
- Estimate mean, variance, and PMF based on 100,000 random variables drawn from a negative binomial distribution with p=0.1, r=3

- Repeat with negative binomial distribution with p=0.1, r=100

Negative binomial PMF, p=0,1 r=3



Negative binomial PMF, p=0,1 r=100



Cancer is scary!

- Approximately 40% of men and women will be diagnosed with cancer at some point during their lifetimes (source: NCI website)

TABLE 21.2 Leading causes of death in United States in 2010. Cause of death is based on the International Classification of Diseases, Tenth Revision, 1992.

Rank	Cause of death	Number	Percent of all deaths
–	All causes	2,468,435	100.0
1	Diseases of heart	597,689	24.2
2	Malignant neoplasms	574,743	23.3
3	Chronic lower respiratory diseases	138,080	5.6
4	Cerebrovascular diseases	129,476	5.2
5	Accidents (unintentional injuries)	120,859	4.9
6	Alzheimer's disease	83,494	3.4
7	Diabetes mellitus	69,071	2.8
8	Nephritis, nephrotic syndrome, and nephrosis	50,476	2.0
9	Influenza and pneumonia	50,097	2.0
10	Intentional self-harm (suicide)	38,364	1.6

Source: National Vital Statistics Reports, 62(6) (http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_06.pdf)

Table from
J. Pevsner
3rd edition

- "War on Cancer" – president Nixon 1971.
  "Moonshot to Cure Cancer" – vice-president Joe Biden 2016

“War on Cancer” progress report

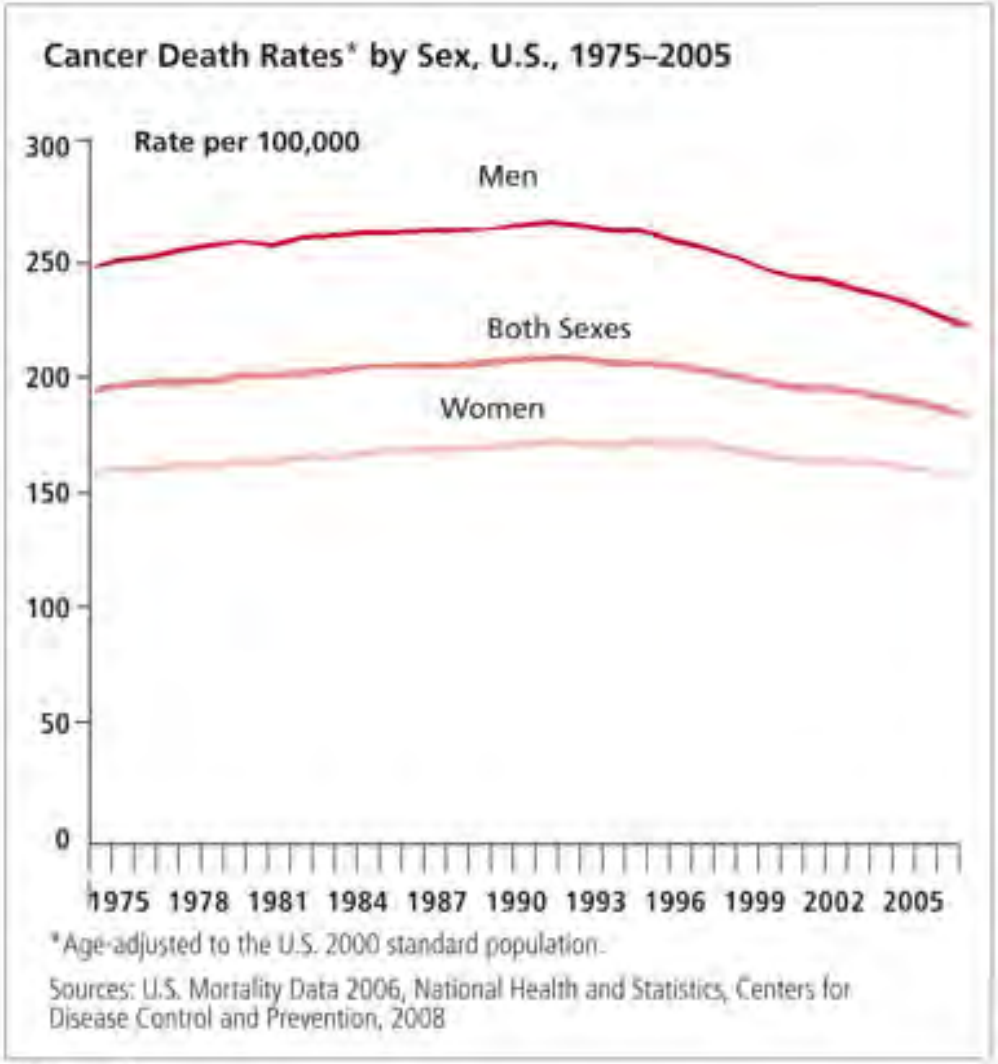


Figure 2

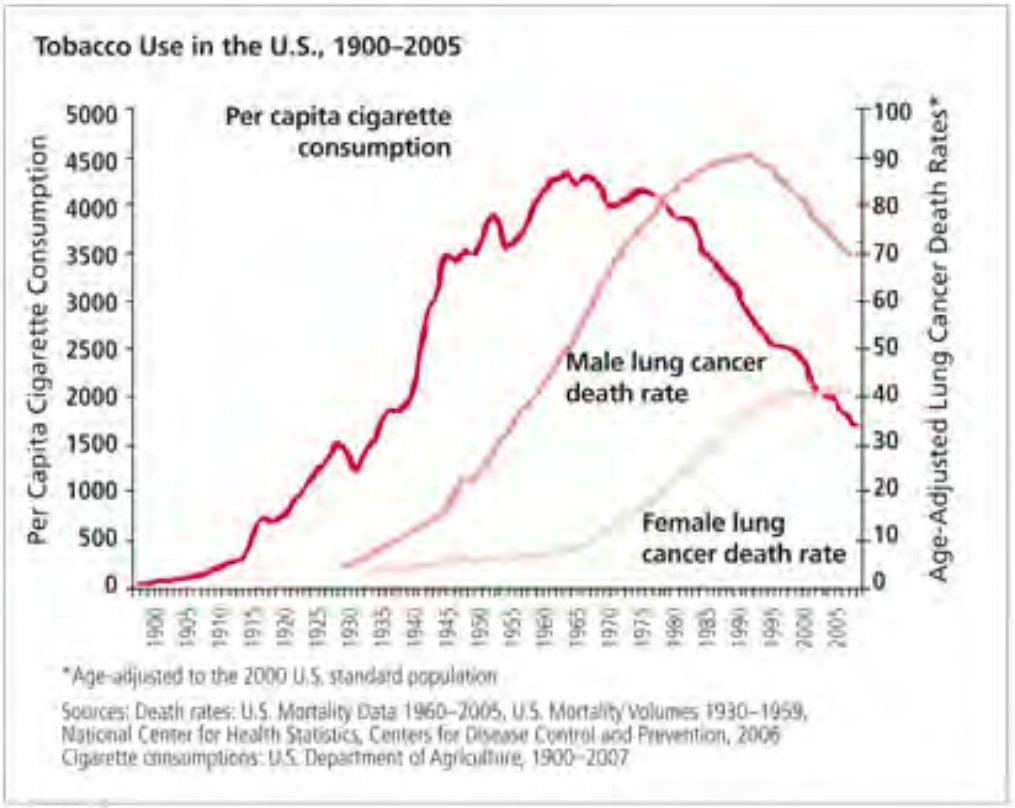


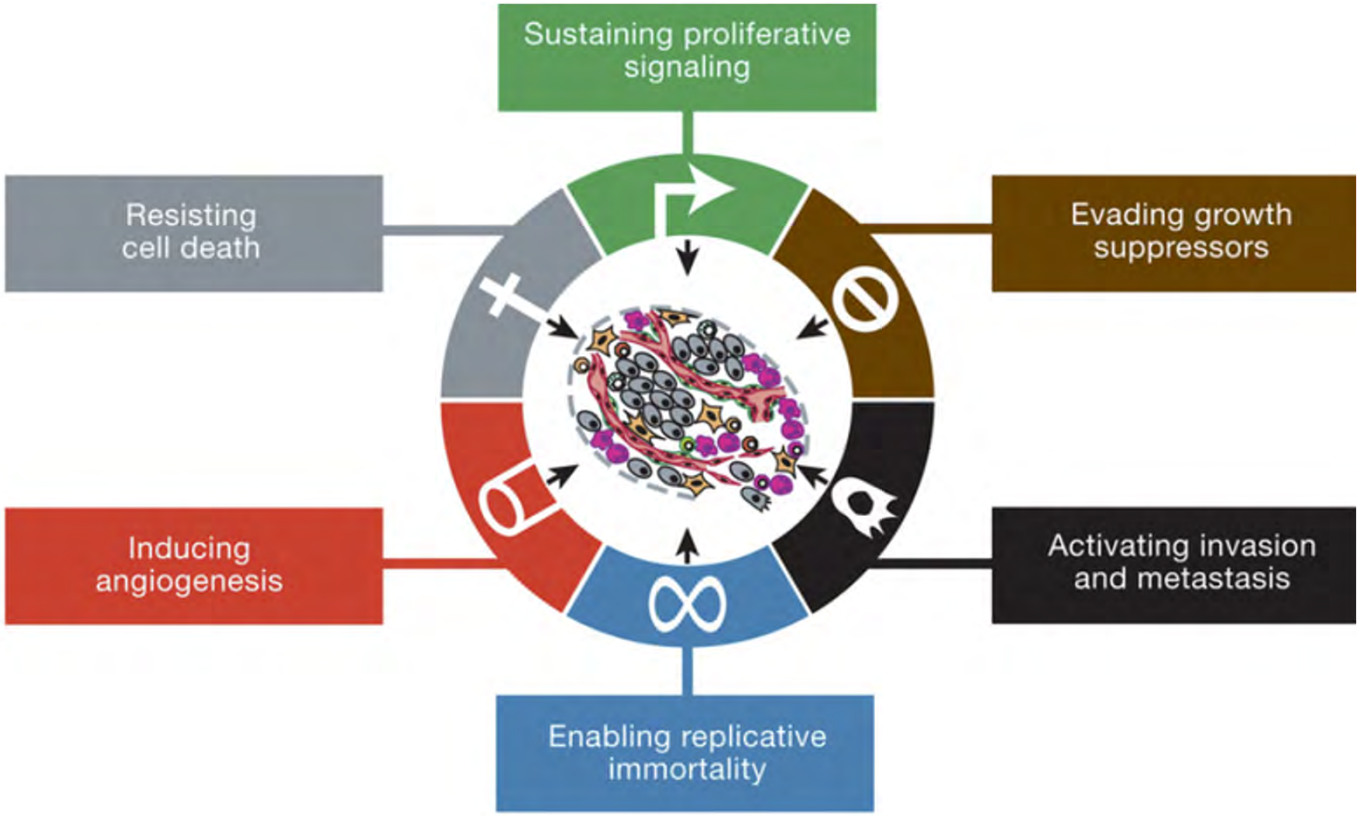
Figure 3

Probability theory and statistics
is a powerful tool to
learn new cancer biology

“Driver genes” theory

- Progression of cancer is caused by accumulation of mutations in a handful of “driver” genes
- Mutations in driver genes boost the growth of a tumor
- Oncogenes: expression needs to be elevated for cancer
- Tumor suppressors (e.g. p53) need to be turned off in cancer

Douglas Hanahan and Robert A. Weinberg **Hallmarks of Cancer**: The Next Generation Cell 144, 2011



Statistics of cancer incidence vs age

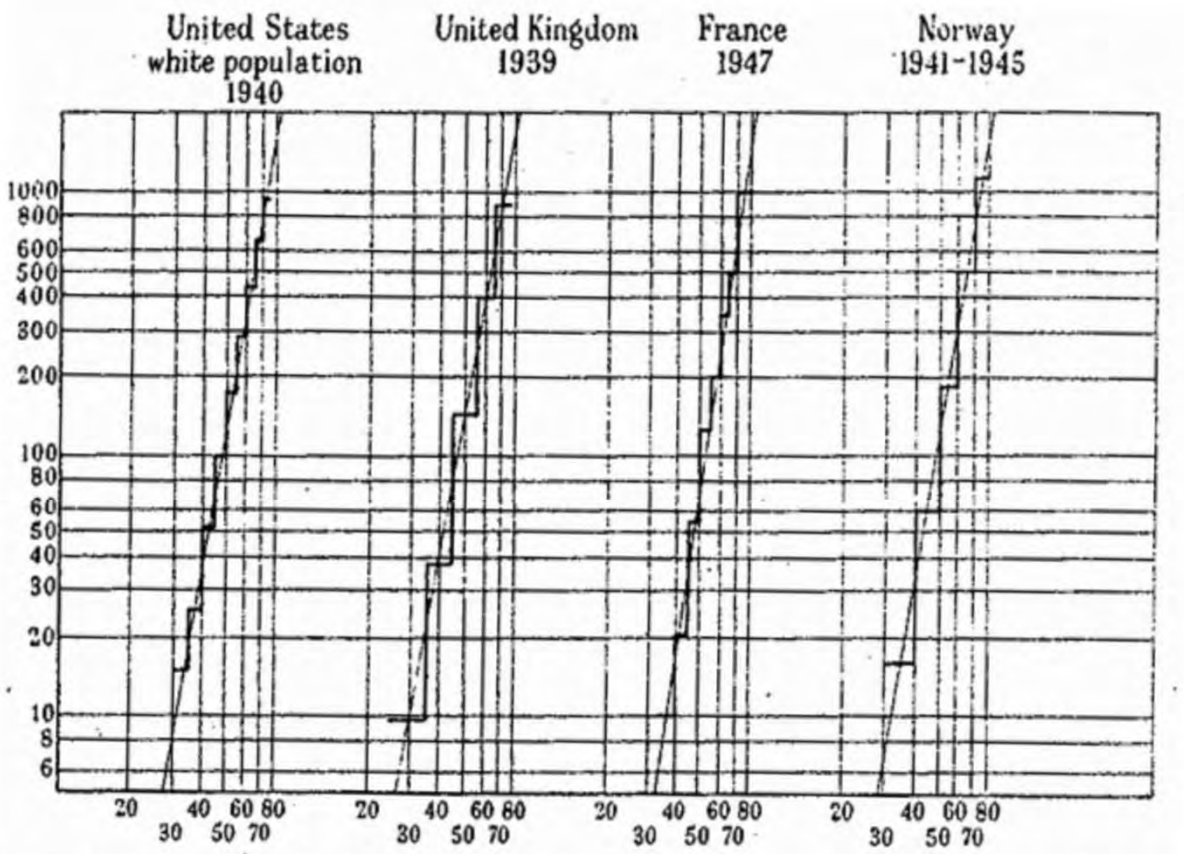


FIG. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.

Multi-mutation theory of cancer:
Carl O. Nordling (British J. of Cancer, March 1953):

Cancer death rate
~ (patient age)^6

It suggests the existence of
k=7 driver genes

$$P(T_{cancer} \le t) \sim (u_1 t)(u_2 t)..(u_k t) \sim u_1 u_2..u_k\, t^k$$

$$P(T_{cancer} = t) \sim \frac{d}{dt}(u_1 t)(u_2 t)..(u_k t) \sim k\, u_1 u_2..u_k\, t^{k-1}$$

How many driver gene mutations for different types of cancer?

PNAS

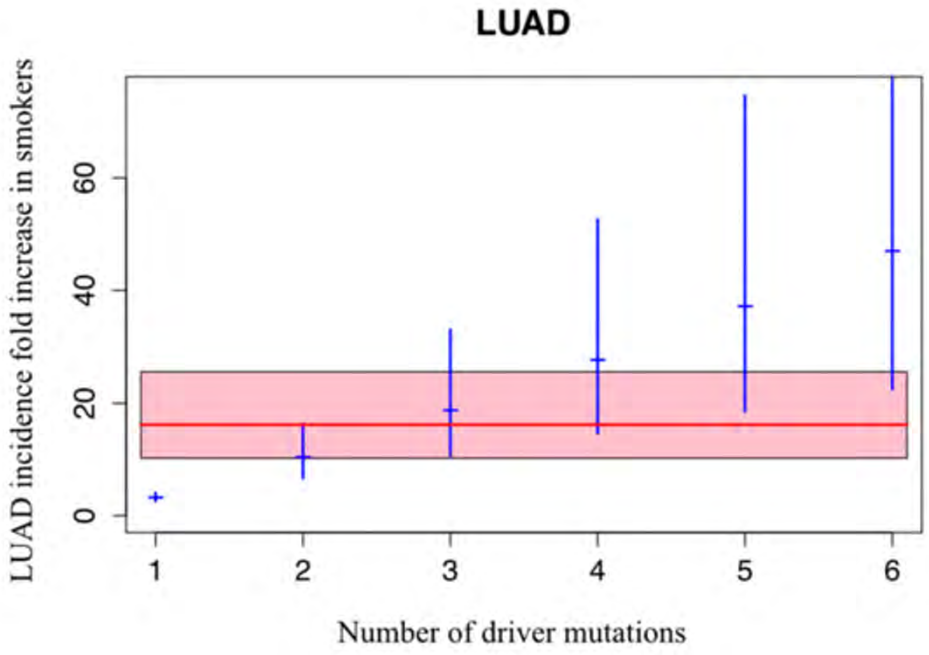
Only three driver gene mutations are required for the development of lung and colorectal cancers

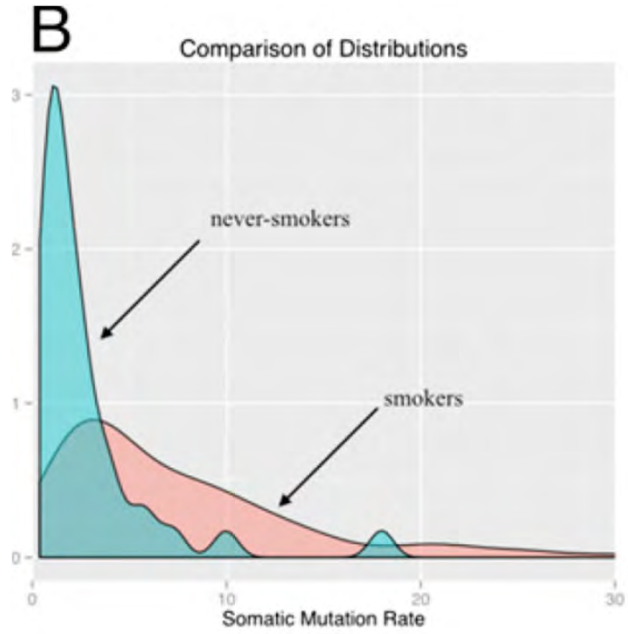
Cristian Tomasetti[a,b,1], Luigi Marchionni[c], Martin A. Nowak[d], Giovanni Parmigiani[e], and Bert Vogelstein[f,g,1]

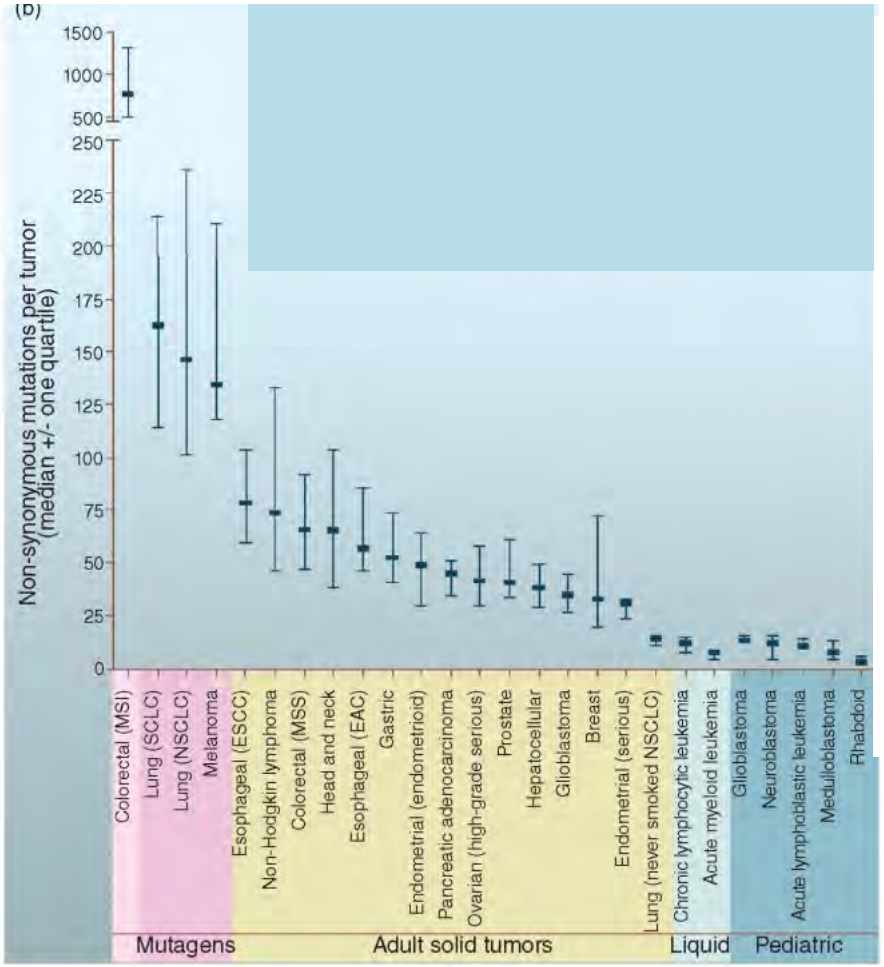
[a]Division of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, and [b]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; [c]Cancer Biology Program, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [d]Program for Evolutionary Dynamics, Department of Mathematics, Harvard University, Cambridge, MA 02138; [e]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02215; and [f]Ludwig Center for Cancer Genetics and Therapeutics and [g]Howard Hughes Medical Institute, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205

Contributed by Bert Vogelstein, November 21, 2014 (sent for review July 31, 2014; reviewed by Zvia Agur)

Smokers have 3.23 times more mutations in lungs







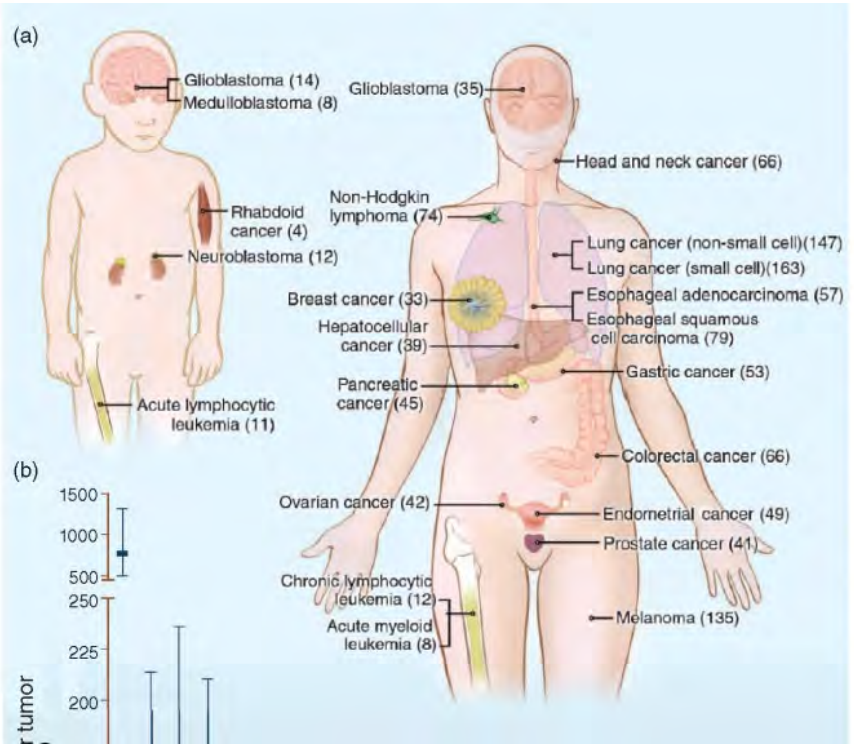


FIGURE 21.10 Somatic mutations in representative human cancers, based on genome-wide sequencing studies. (a) The genomes of adult (right) and pediatric (left) cancers are represented. Numbers in parentheses are the median number of nonsynonymous mutations per tumor. Redrawn from Vogelstein *et al.* (2013). Reproduced with permission from AAAS. (b) Median number of nonsynonymous substitutions per tumor. Horizonal bars indicate the 25% and 75% quartiles. MSI: microsatellite instability; SCLC: small cell lung cancers; NSCLC: non-small cell lung cancers; ESCC: esophageal squamous cell carcinomas; MSS: microsatellite stable; EAC: esophageal adenocarcinomas.

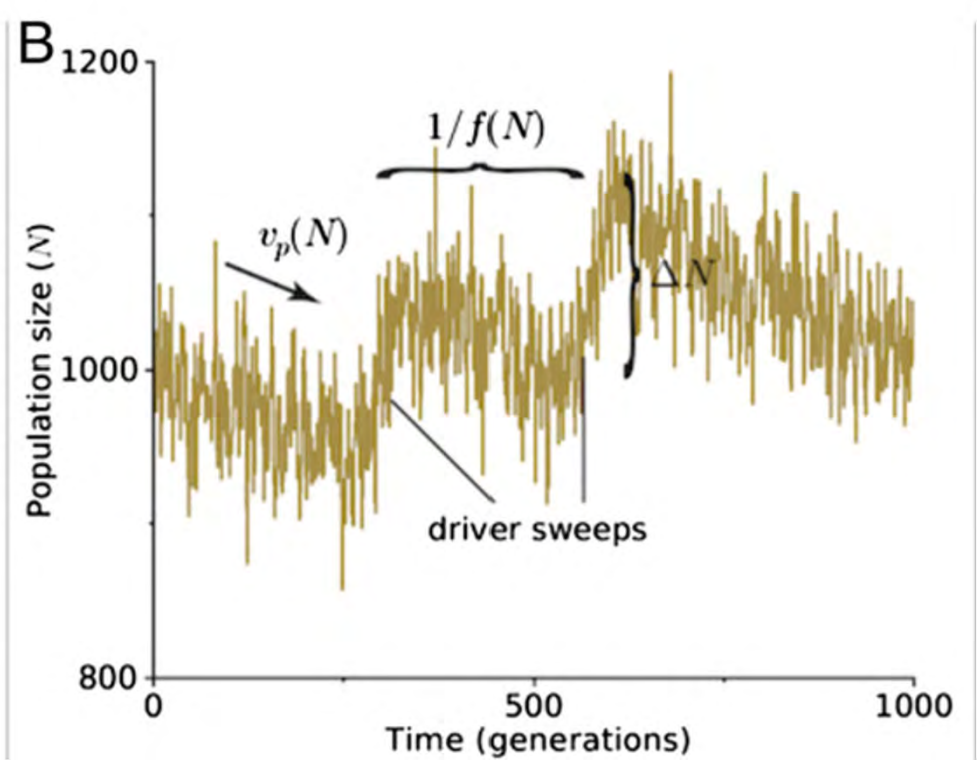
- Cancer cells carry both "Driver" and "Passengers" mutations
- Passenger mutations cause little to no harm (see later for how even little harm matters)
- Both are common as cancers elevate mutation rate

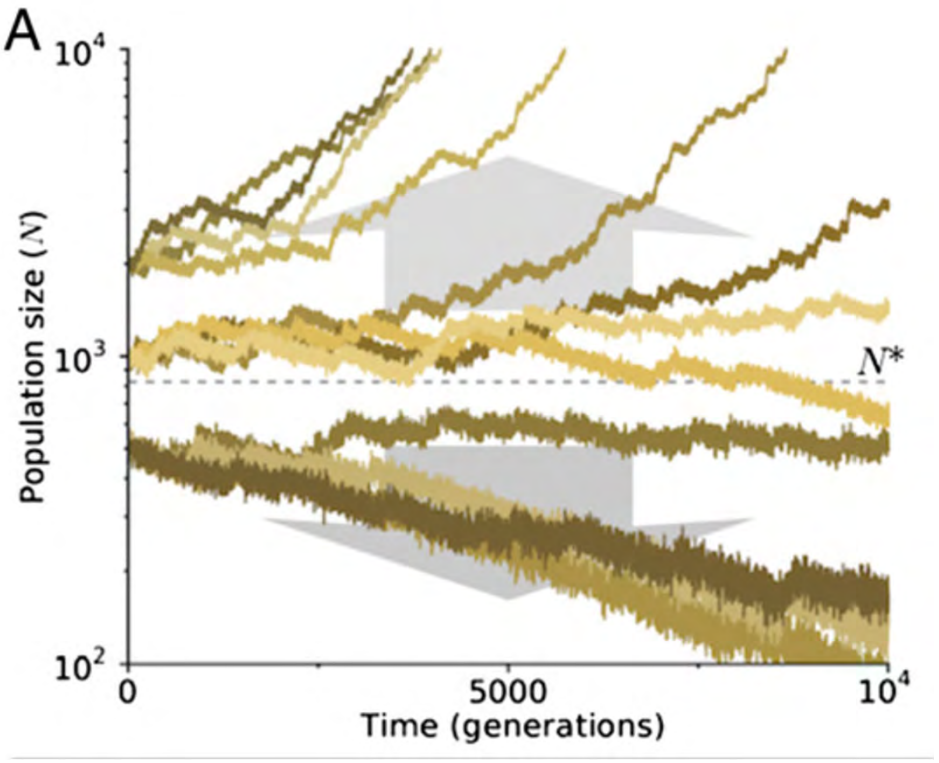
Number of passenger+driver mutations follows negative binomial distribution

- What is the probability to have n_p passenger mutations or (n_p+k) total mutations by the time you are diagnosed with cancer requiring k driver mutations?
- Let p is the probability that a mutation is a driver (p= Genome_target_of_driv/ (Genome_target_of_driv+Genome_target_of_pass)
  (1-p) – it is a passenger mutation

$$P(n_p + k \mid p, k) = \binom{n_p + k - 1}{n_p} (1-p)^{n_p} \, p^k$$

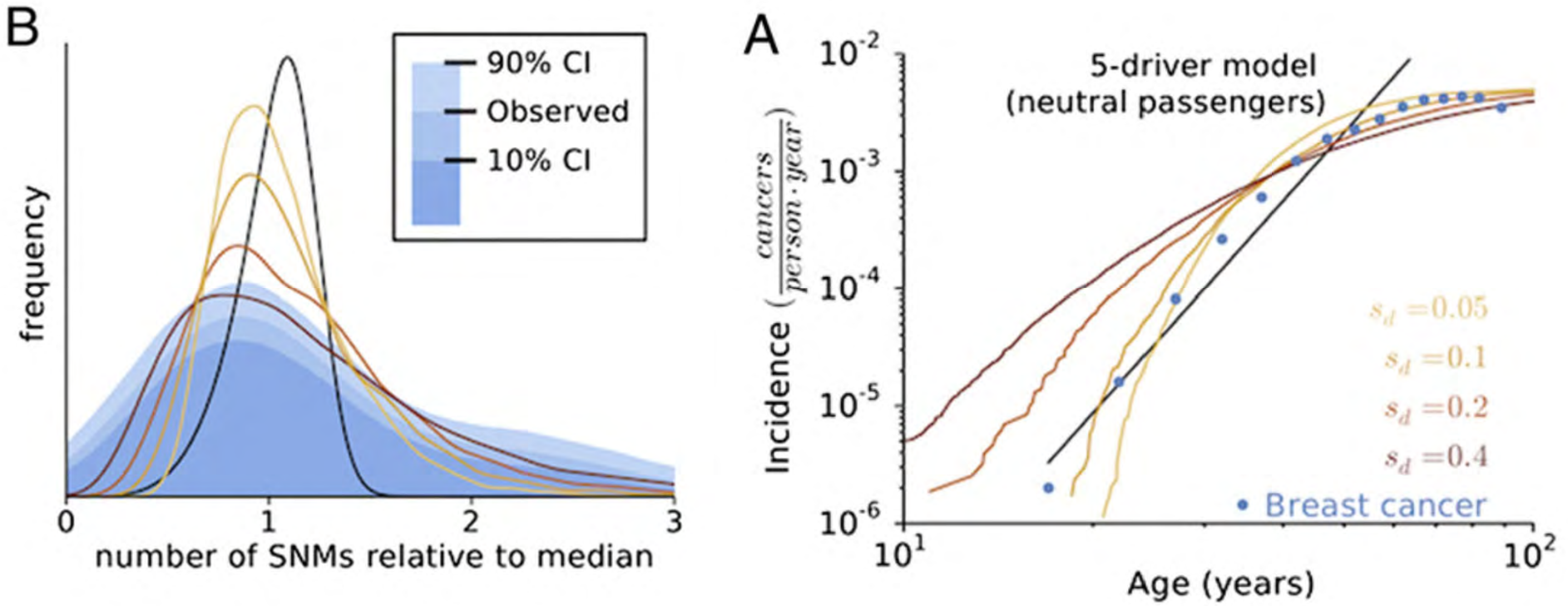
What if passenger mutations slow down the growth of cancer tumors?





McFarland CD, Mirny L, Korolev KS, PNAS 2014

Can we prove/quantify it using statistics?



Assume: growth rate of cancer=$(1+s_d)^{Nd}/(1+s_p)^{Np}$

$\mu=10^{-8}$, $Target_d=1{,}400$, $Target_p=10^7$, $s_d=0.05$ to 0.4, $s_p=0.001$

s_p/s_d for breast: 0.0060±0.0010;

melanoma: 0.016±0.003; lung: 0.0094±0.0093;

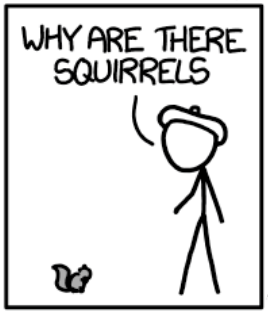
Blue - data on breast cancer: incidence; non-synonymous mutations

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 IS LYING GOOD

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 SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
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 IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

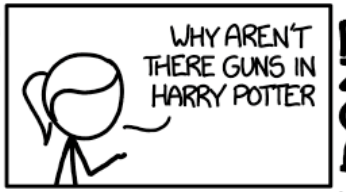
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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

Important terms & concepts for **discrete random variables**

- Probability Mass Function (PMF)
- ***Cumulative Distribution Function (CDF)***
- ***Complementary Cumulative Distribution Function (CCDF)***
- ***Expected value***
- ***Mean***
- ***Variance***
- ***Standard deviation***

Boldface and underlined are the same for continuous distributions

Name	Probability Distribution	Mean	Variance
Discrete			
Uniform	$\frac{1}{n}, a \le b$	$\frac{(b+a)}{2}$	$\frac{(b-a+1)^2-1}{12}$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n, 0 \le p \le 1$	np	$np(1-p)$
Geometric	$(1-p)^{x-1}p$, $x = 1, 2, \ldots, 0 \le p \le 1$	$1/p$	$(1-p)/p^2$
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r}p^r$ $x = r, r+1, r+2, \ldots, 0 \le p \le 1$	r/p	$r(1-p)/p^2$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$	λ	λ

What distributions we learn

- Uniform distribution
- Bernoulli distribution/trial
- Binomial distribution
- Poisson distribution
- Geometric distribution
- Negative binomial distribution

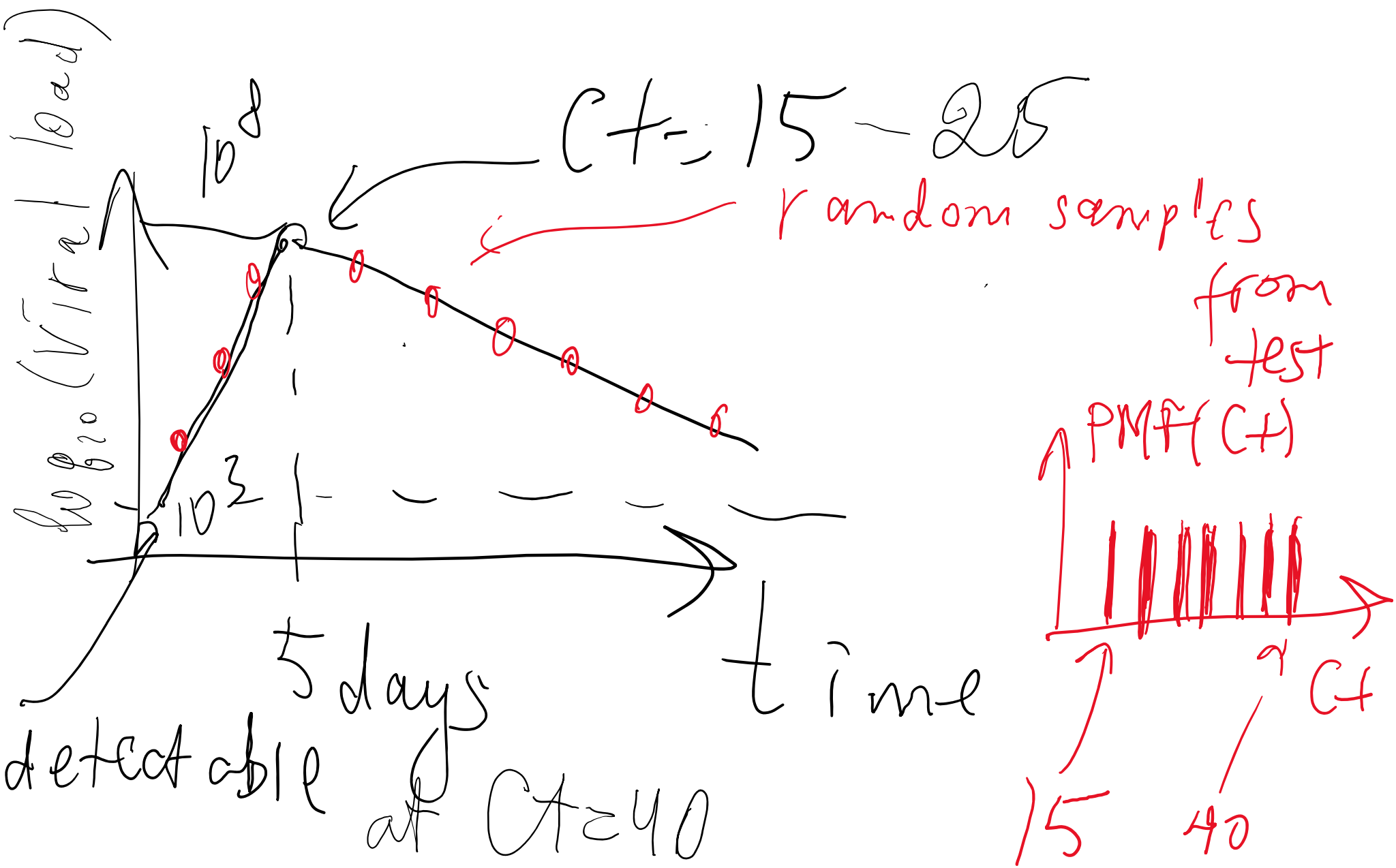
Why do we need to know these simple distributions?

Ways to use statistics

- To process your experimental data
  - What do you need? Mean, Variance, Standard deviation. No need to know any textbook distributions
- To plan experiments
  - Need to know distributions, e.g., Poisson to plan how much redundancy to use for genome assembly
- To learn biological processes behind your data
  - Need to know distributions to compare empirical distributions in your data to what you expect based on a simple hypothesis

Uniform distribution

Why Ct distribution should it be uniform?



Examples of uniform distribution: Ct value of PCR test of a virus

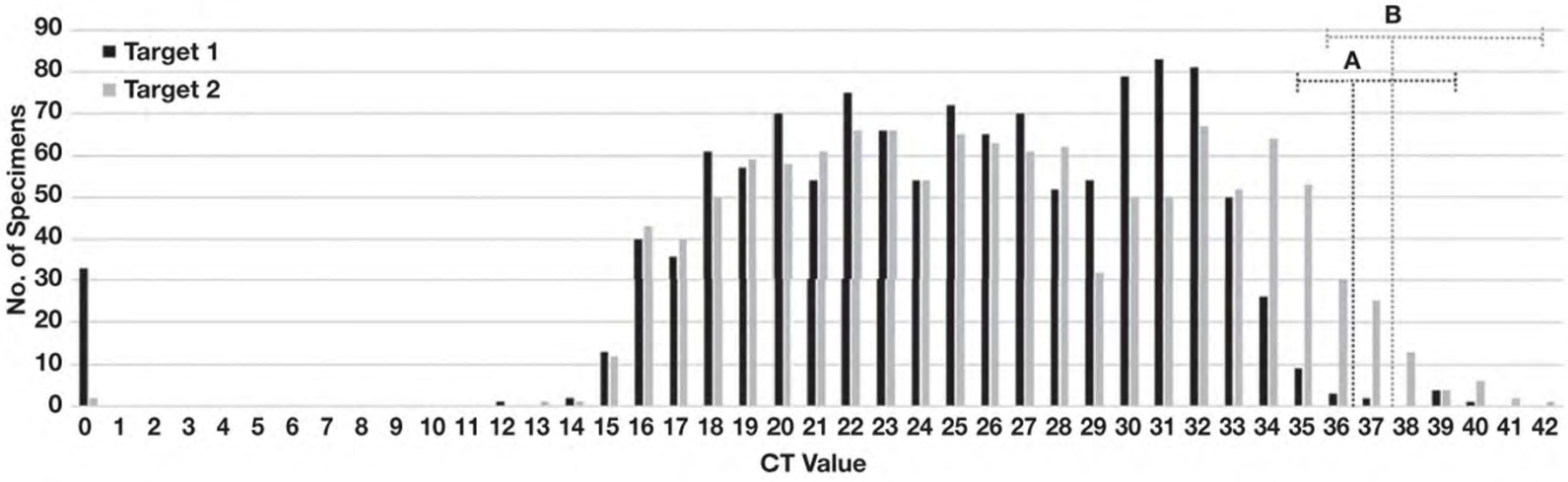


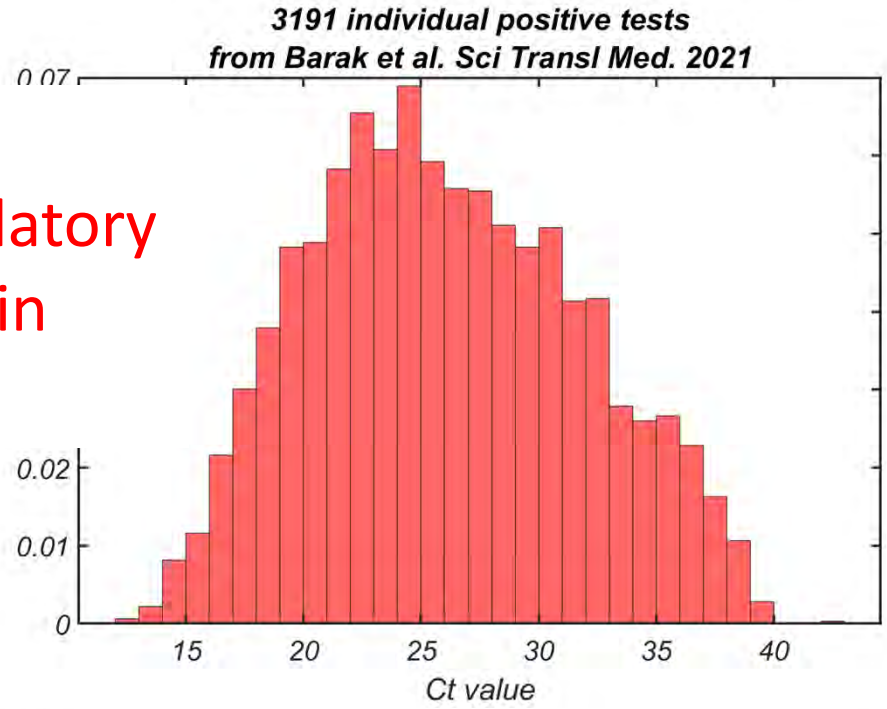
Figure 3 Distribution of cycle threshold (CT) values. The total number of specimens with indicated CT values for Target 1 and 2 are plotted. The estimated limit of detection for (A) Target 1 and (B) Target 2 are indicated by vertical dotted lines. Horizontal dotted lines encompass specimens with CT values less than 3× the LoD for which sensitivity of detection may be less than 100%. This included 19/1,180 (1.6%) reported CT values for Target 1 and 81/1,211 (6.7%) reported CT values for Target 2. Specimens with Target 1 or 2 reported as "not detected" are denoted as a CT value of "0."

Distribution of SARS-CoV-2 PCR Cycle Threshold Values Provide Practical Insight Into Overall and Target-Specific Sensitivity Among Symptomatic Patients

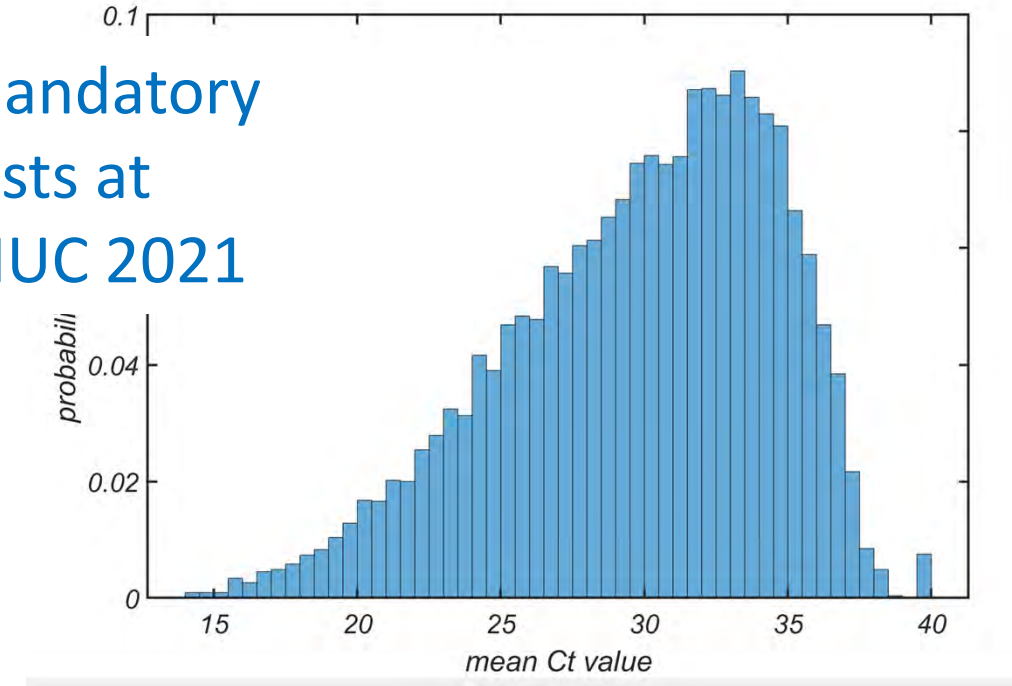
Blake W Buchan, PhD, Jessica S Hoff, PhD, Cameron G Gmehlin, Adriana Perez, Matthew L Faron, PhD, L Silvia Munoz-Price, MD, PhD, Nathan A Ledeboer, PhD *American Journal of Clinical Pathology*, Volume 154, Issue 4, 1 October 2020, https://academic.oup.com/ajcp/article/154/4/479/5873820

Why should we care?

Non-mandatory tests in Israel



Mandatory tests at UIUC 2021



- High Ct value means we identified the infected individual early, hopefully before transmission to others
- When testing is mandatory, and people are tested frequently – Ct value is skewed towards high values

Negative binomial distribution

Statistics of cancer incidence vs age

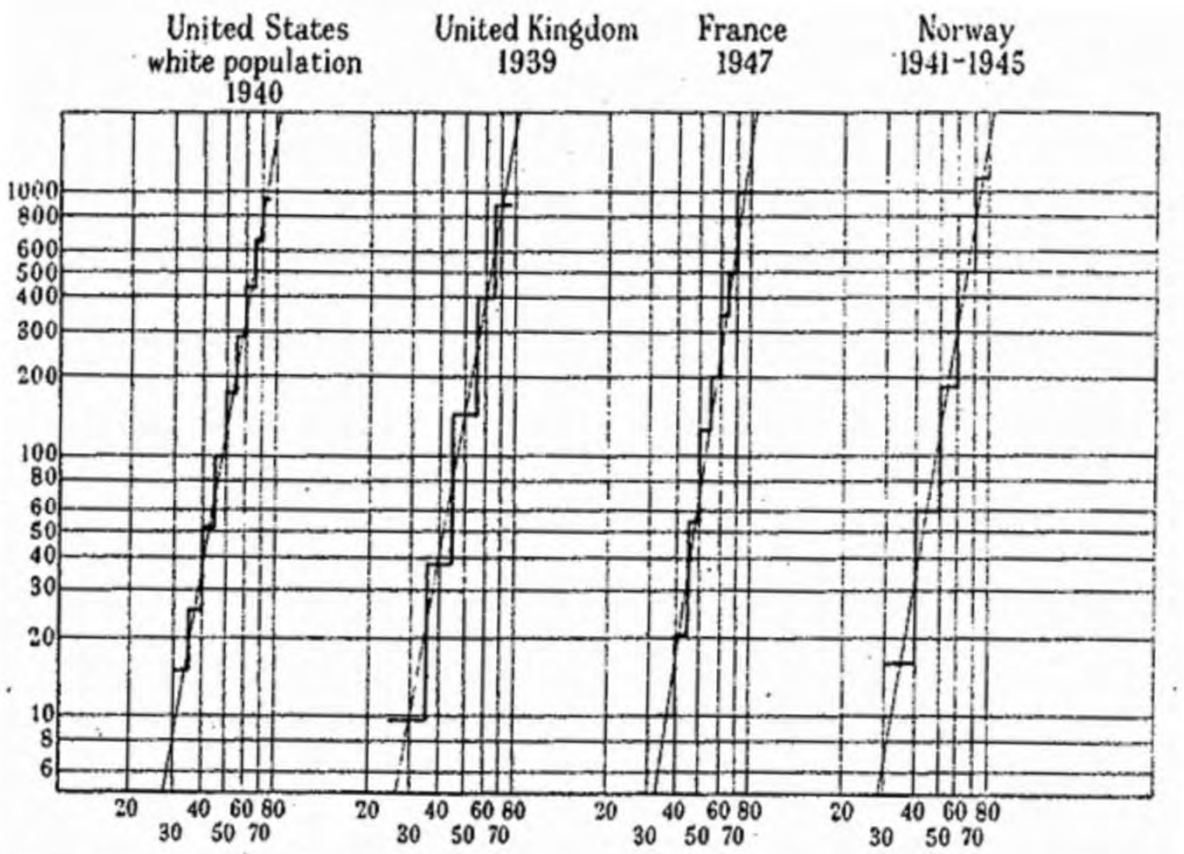


FIG. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.

Multi-mutation theory of cancer:
Carl O. Nordling (British J. of Cancer, March 1953):

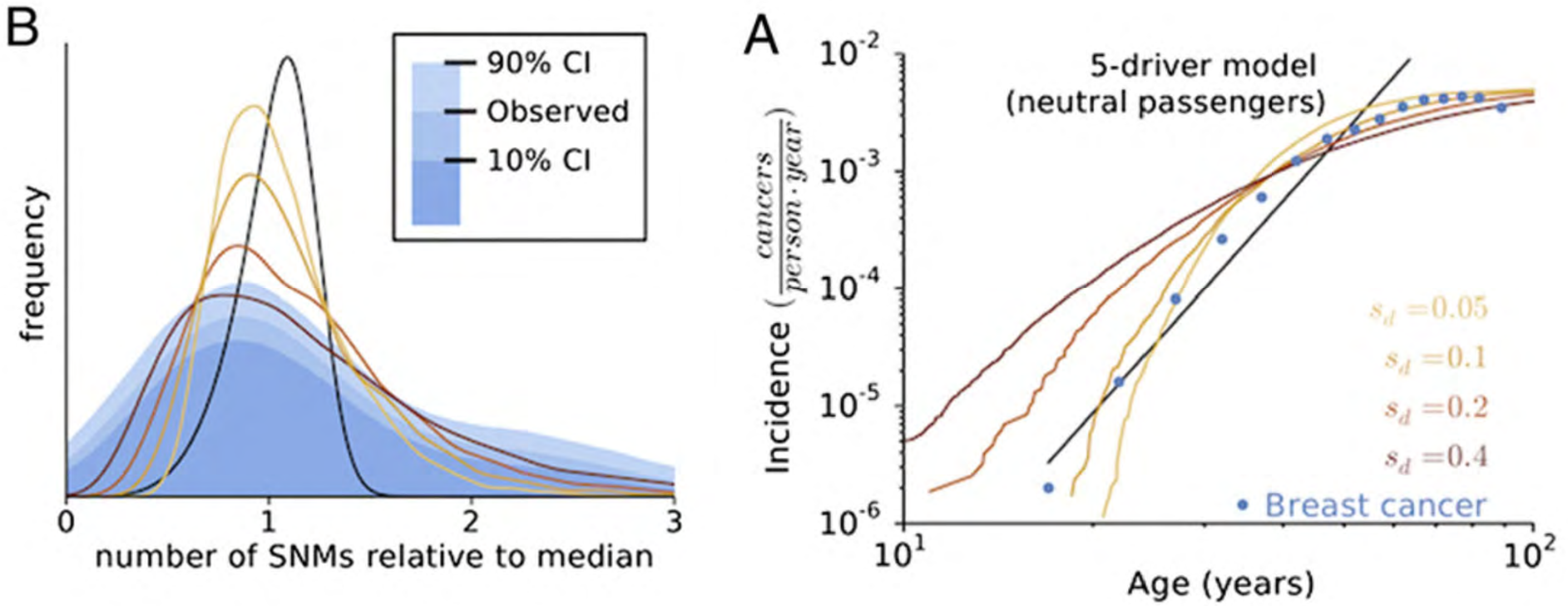
Cancer death rate
~ (patient age)^6

It suggests the existence of
k=7 driver genes

$$P(T_{cancer} \le t) \sim (u_1 t)(u_2 t)..(u_k t) \sim u_1 u_2..u_k\, t^k$$

$$P(T_{cancer} = t) \sim \frac{d}{dt}(u_1 t)(u_2 t)..(u_k t) \sim k\, u_1 u_2..u_k\, t^{k-1}$$

Can we prove/quantify it using statistics?



Assume: growth rate of cancer=$(1+s_d)^{Nd}/(1+s_p)^{Np}$

$\mu=10^{-8}$, $Target_d=1{,}400$, $Target_p=10^7$, $s_d=0.05$ to 0.4, $s_p=0.001$

s_p/s_d for breast: 0.0060±0.0010;

melanoma: 0.016±0.003; lung: 0.0094±0.0093;

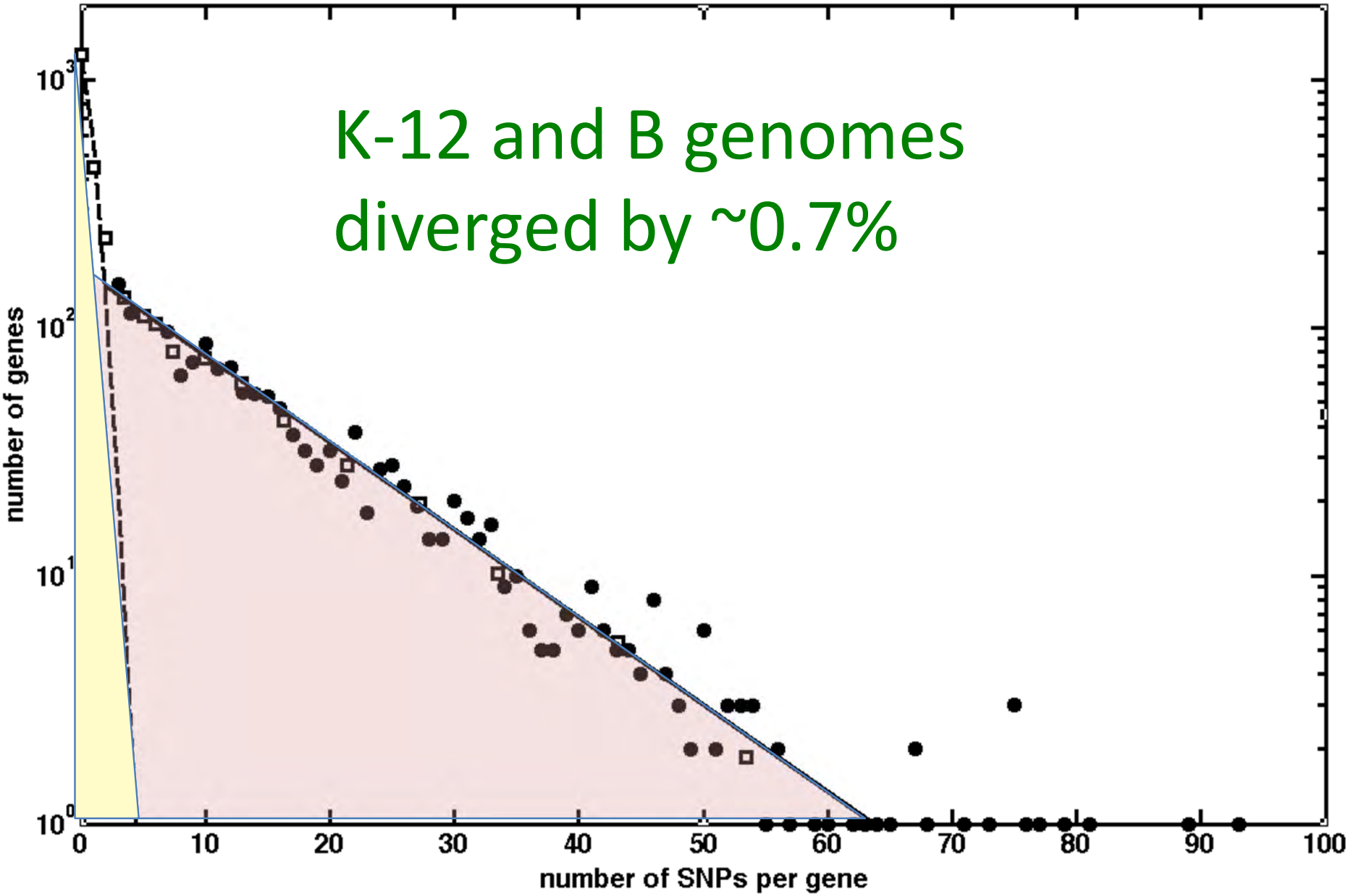
Blue - data on breast cancer: incidence; non-synonymous mutations

Poisson and Exponential Distributions

F. William Studier

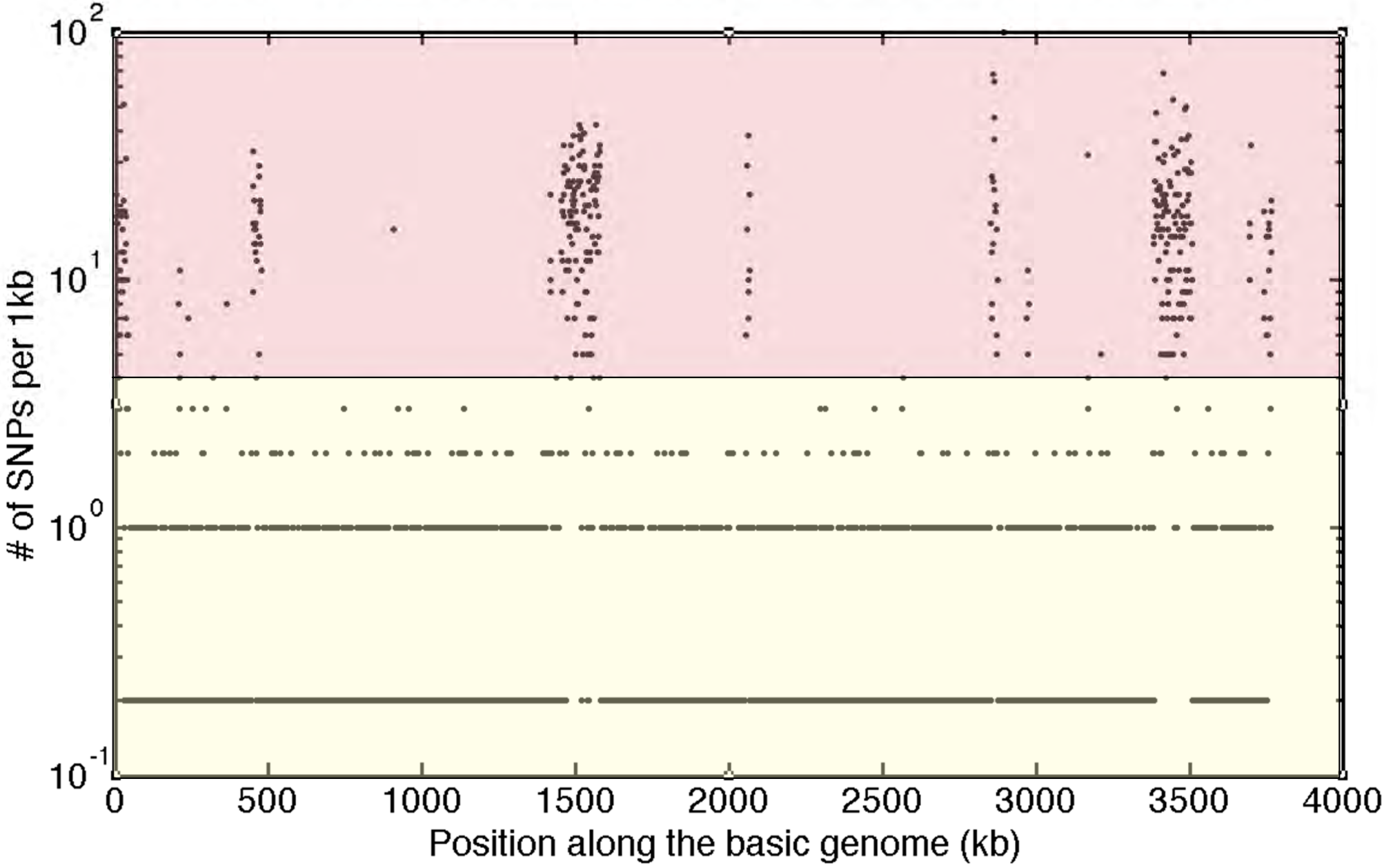
- Worked at Brookhaven National Laboratory, Long Island, NY since 1964
- Inventor of **slab gel electrophoresis in 1970** (not patented- back then no incentive to patent work if you are supported by the US givernment)
- Inventor of T7 phage expression system for **fast production of proteins**. Licensed by over 900 companies, generated over $55 million for the lab https://en.wikipedia.org/wiki/T7_expression_system

K-12 vs BL21(DE3) strains of E. coli



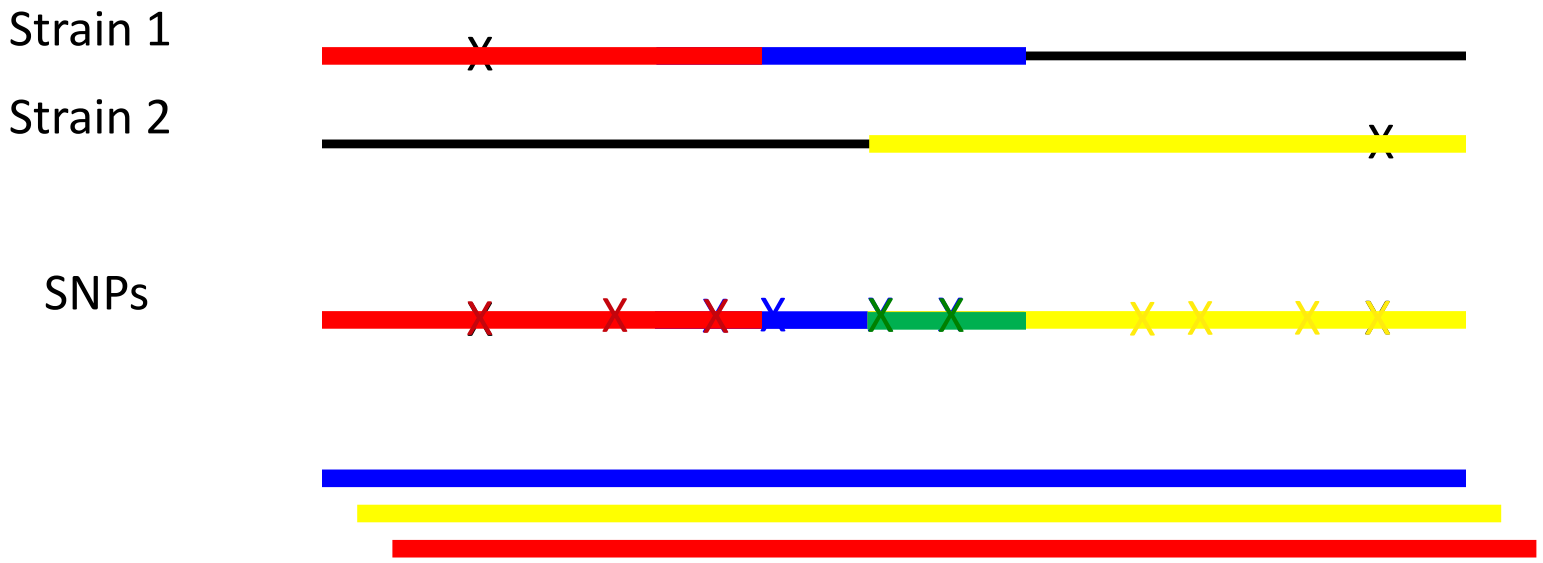
Studier FW, Daegelen P, Lenski RE, **Maslov S**, Kim JF, J. Mol Biol. (2009)

Highly variable segments are clustered



K-12 vs UMNF18 diverged by ~0.18%

Model of bacterial evolution by mutations and homologous recombination



- Mutation rate μ (bp/generation)
- Recombination rate ρ (bp/generation)
- l_R- average length of recombined segments
- $\theta=2\mu N_e$ depending on N_e – (effective) population size
- δ_{TE} transfer efficiency: Prob(successful transfer + recombination): ~ $\exp(-\delta/\delta_{TE})$

Dixit P, Pang TY, Studier FW, **Maslov S,** PNAS (2015); arXiv:1405.2548

Why exponential tail?

- Empirical data for E. coli: Prob(δ)=exp(-δ/0.01)
  Similar slopes in other species as distant as B. subtilis
- Theory 1: PopGen 101 coalescence time distribution:
  - Prob(T) ~ exp(-T/N_e) →
    Prob(δ) ~ exp(- δ/ 2μN_e) = exp(- δ/θ)
    θ =2μN_e~0.01, μ~10^{-10} → N_e~10^8
- Theory 2: biophysics of homologous recombination:
  - Requires perfect matches of L=30bp on each side →
    Prob(δ)=$(1- \delta)^{2L}$=exp(-60•δ)=exp(-δ/0.016)=exp (-δ/δ_{TE})
- Both mechanisms likely to work together:
  biophysics of recombination affects the effective population size

Continuous Probability Distributions

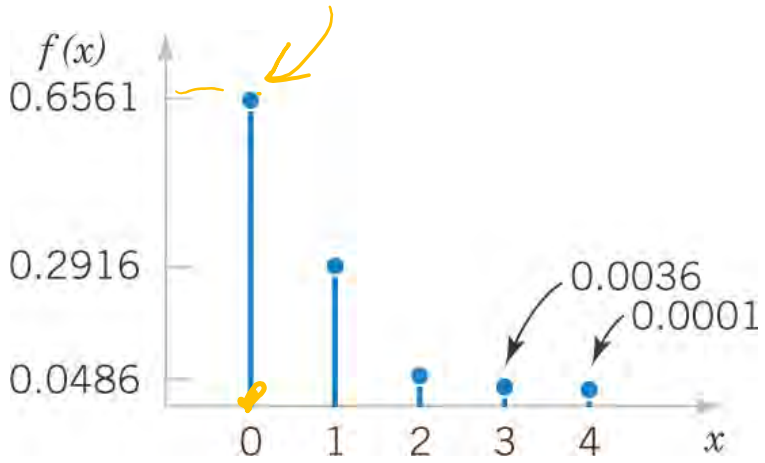
Uniform Distribution

Continuous & Discrete Random Variables

- A discrete random variable is usually integer number
  - N – the number of proteins in a cell
  - D- number of nucleotides different between two sequences
- A continuous random variable is a real number
  - C=N/V – the concentration of proteins in a cell of volume V
  - Percentage D/L*100% of different nucleotides in protein sequences of different lengths L (depending on set of L's may be discrete but dense)

Probability Mass Function (PMF)

- X – discrete random variable



Probability Mass Function for the # of mismatches in 4-mers

- Probability Mass Function: f(x)=P(X=x) – the probability that X is exactly equal to x

P(X =0) =	0.6561
P(X =1) =	0.2916
P(X =2) =	0.0486
P(X =3) =	0.0036
P(X =4) =	0.0001
$\sum_x$ P(X=x)=	1.0000

Probability Density Function (PDF)

Density functions, in contrast to mass functions, distribute probability continuously along an interval

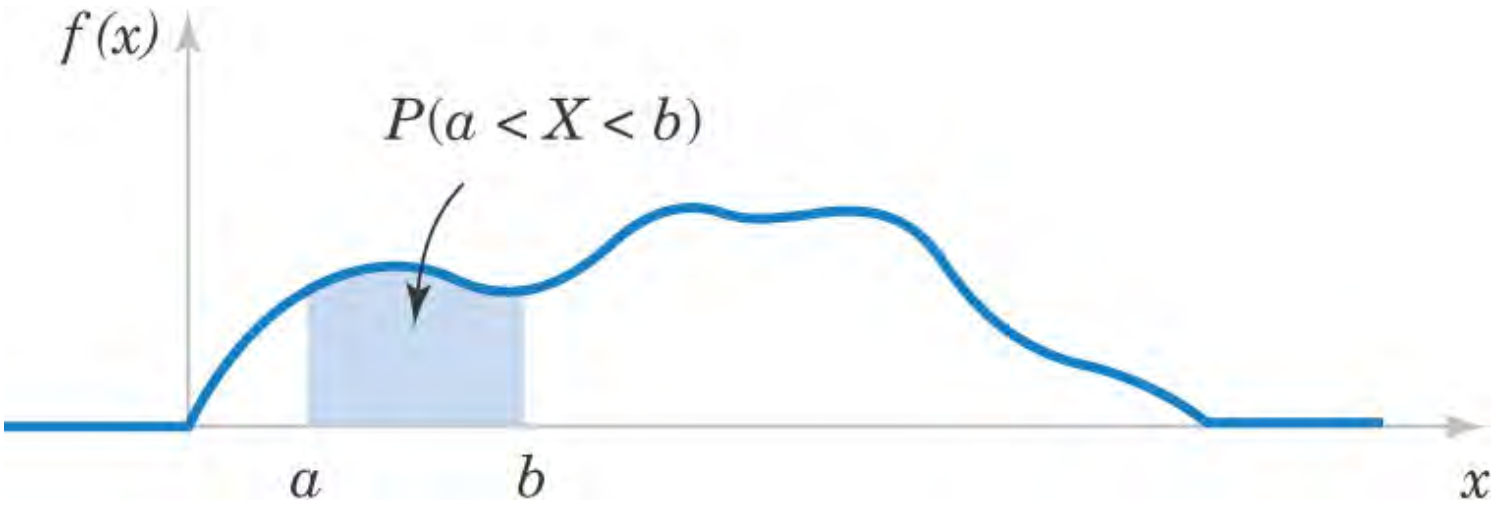


Figure 4-2 Probability is determined from the area under $f(x)$ from a to b.

Probability Density Function

For a continuous random variable X,
a probability density function is a function such that

(1) $f(x) \geq 0$ means that the function is always non−negative.

(2) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

(3) $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx =$ area under $f(x)dx$ from a to b

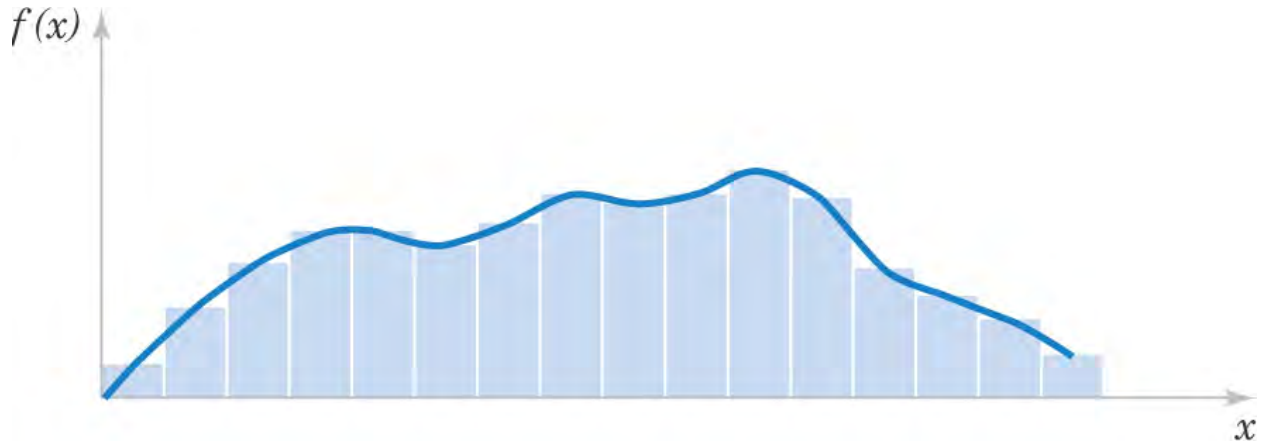
Normalized histogram approximates PDF

A histogram is graphical display of data showing a series of adjacent rectangles. Each rectangle has a base which represents an interval of data values. The height of the rectangle is a number of events in the sample within the base.

When base length is narrow, the histogram could be normalized to approximate PDF (f(x)):

height of each rectangle =

=(# of events within base)/(total # of events)/width of its base.



Normalized histogram approximates a probability density function.

Cumulative Distribution Functions (CDF & CCDF)

The cumulative distribution function (CDF)
of a continuous random variable X is,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du \text{ for } -\infty < x < \infty \quad (4-3)$$

One can also use the inverse cumulative distribution function
or complementary cumulative distribution function (CCDF)

$$F_{>}(x) = P(X > x) = \int_{x}^{\infty} f(u)du \text{ for } -\infty < x < \infty$$

Definition of CDF for a continous variable is the same
as for a discrete variable

Density vs. Cumulative Functions

- The probability density function (PDF) is the derivative of the cumulative distribution function (CDF).

$$f(x) = \frac{dF(x)}{dx} = -\frac{dF_{>}(x)}{dx}$$

as long as the derivative exists.

Mean & Variance

Suppose X is a continuous random variable with probability density function $f(x)$. The mean or expected value of X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx \qquad (4-4)$$

The variance of X, denoted as $V(X)$ or σ^2, is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

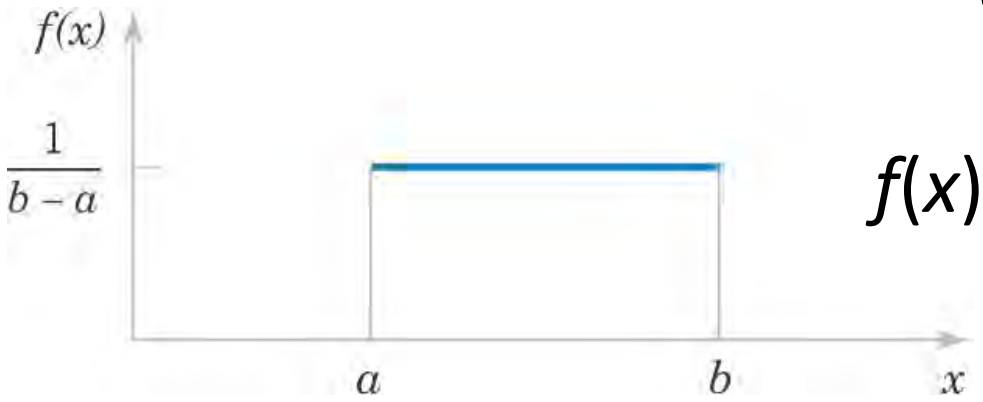
The standard deviation of X is $\sigma = \sqrt{\sigma^2}$.

Gallery of Useful Continuous Probability Distributions

Continuous Uniform Distribution

- This is the simplest continuous distribution and analogous to its discrete counterpart.
- A continuous random variable *X* with probability density function

$$f(x) = 1 / (b-a) \text{ for } a \le x \le b \qquad (4\text{-}6)$$



Compare to discrete

$f(x) = 1/(b-a+1)$

Figure 4-8 Continuous uniform PDF

Comparison between Discrete & Continuous Uniform Distributions

Discrete:

- PMF: $f(x) = 1/(b-a+1)$
- Mean and Variance:

  $\mu = E(x) = (b+a)/2$

  $\sigma^2 = V(x) = [(b-a+1)^2-1]/12$

Continuous:

- PMF: $f(x) = 1/(b-a)$
- Mean and Variance:

  $\mu = E(x) = (b+a)/2$

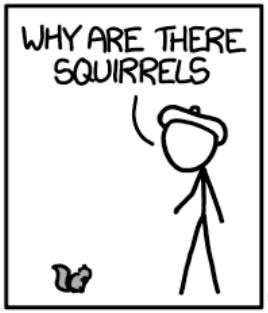
  $\sigma^2 = V(x) = (b-a)^2/12$

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 IS LYING GOOD

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 SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
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 IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

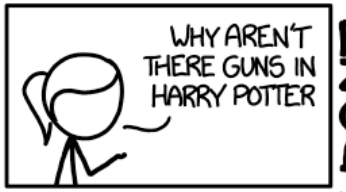
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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

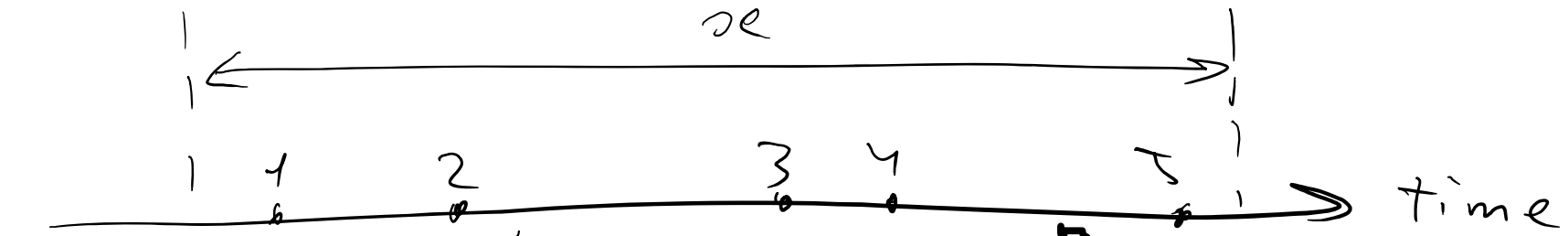
WHY AREN'T THERE GUNS IN HARRY POTTER



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

Constant rate (Poisson) process

Constant rate (Poisson) process



discrete events happen at rate Γ

Expected number of events in time x is Γx

The actual number of events N_x is a Poisson distributed discrete random variable

$$P(N_x = n) = \frac{(\Gamma x)^n}{n!} e^{-\Gamma x}$$

Why Poisson? Divide x into many tiny intervals of length Δx

$p = \Gamma \Delta x$

$L = x/\Delta x$

$E(N_x) = pL = \Gamma x$

$$\text{Prob}(N = n) = \binom{L}{n} p^n (1-p)^{L-n}$$

$\downarrow$ $p \sim \Delta x \to 0$, $L \sim 1/\Delta x \to \infty$

Poisson

Constant rate (AKA Poisson) processes

- Let's assume that proteins are produced by ribosomes in the cell at a rate r per second.
- The expected number of proteins produced in x seconds is r·x.
- The actual number of proteins N_x is a discrete random variable following a Poisson distribution with mean r·x:

$$P_N(N_x=n)=\exp(-r\cdot x)(r\cdot x)^n/n! \qquad E(N_x)= rx$$

- Why Discrete Poisson Distribution?
  - Divide time into many tiny intervals of length $\Delta x \ll 1/r$
  - The probability of success (protein production) per internal is small: $p\_success=r\Delta x \ll 1$,
  - The number of intervals is large: $n= x/\Delta x \gg 1$
  - Mean is constant: $r=E(N_x)=p\_success \cdot n= r\Delta x \cdot x/\Delta x = r\cdot x$
  - In the limit $\Delta x \ll x$, p_success is small and n is large, thus Binomial distribution → Poisson distribution

Exponential Distribution Definition

Exponential random variable X describes interval between two successes of a constant rate (Poisson) random process with success rate r per unit interval.

The probability density function of X is:

$$f(x) = re^{-rx} \quad \text{for} \quad 0 \le x < \infty$$

Closely related to the discrete geometric distribution

$$f(x) = p(1-p)^{x-1} = p\, e^{(x-1)\ln(1-p)} \approx pe^{-px} \quad \text{for small p}$$

To summarize constant rate processes:

r – rate per unit of time length — 1D

$N(x)$ — discrete number of events in time x

Poisson: $P(N(x)=n) = \frac{(r \cdot x)^n}{n!} e^{-r \cdot x}$

Time interval X between successive events is a continuously distributed random variable

Its PDF if $f(x) = e^{-rx}$

What is the interval X between two successes of a constant rate process?

- X is a continuous random variable
- CCDF: $P_X(X>x) = P_N(N_X=0)=exp(-r \cdot x)$.
  - Remember: $P_N(N_X=n)=exp(-r \cdot x)\ (r \cdot x)^n/n!$
- PDF: $f_X(x)=-dCCDF_X(x)/dx = r \cdot exp(-r \cdot x)$
- We started with a discrete Poisson distribution where time x was a parameter
- We ended up with a continuous exponential distribution

Exponential Mean & Variance

If the random variable X has an exponential distribution with rate r,

$$\mu = E(X) = \frac{1}{r} \quad \text{and} \quad \sigma^2 = V(X) = \frac{1}{r^2} \qquad (4-15)$$

Note that, for the:

- Poisson distribution: mean= variance
- Exponential distribution: mean = standard deviation = variance$^{0.5}$

Biochemical Reaction Time

- The time x (in minutes) until all enzymes in a cell catalyze a biochemical reaction and generate a product is approximated by this CCDF:

$$F_{>}(x) = e^{-2x} \text{ for } 0 \leq x$$

Here the rate of this process is r=2 min^{-1} and 1/r=0.5 min is the average time between successive products of these enzymes

- What is the PDF?

$$f(x) = -\frac{dF_{>}(x)}{dx} = -\frac{d}{dx}e^{-2x} = 2e^{-2x} \text{ for } 0 \leq \text{x}$$

- What proportion of reactions will not generate another product within 0.5 minutes of the previous product?

$$P(X > 0.5) = F_{>}(0.5) = e^{-2*\,0.5} = 0.37$$

We observed our cell for 1 minute
and no product has been generated:
The product is "overdue"

What is the probability that
a product will not appear
during the next 0.5 minutes?

$$F_{>}(x) = e^{-2x}$$
$$F_{>}(0.5) \approx 0.37$$
$$F_{>}(1.5) \approx 0.05$$
$$F_{>}(1.0) \approx 0.13$$

A. 0.32
B. 0.37
C. 0.08
D. 0.24
E. I have no idea

Get your i-clickers

Memoryless property of the exponential distribution

$$P(X>t+s \mid X>s) = P(X>t)$$

$$P(X>t+s \mid X>s) = \frac{P(X>t+s,\ X>s)}{P(X>s)} =$$

$$= \frac{\exp(-\lambda(t+s))}{\exp(-\lambda s)} = \exp(-\lambda t) =$$

$$= P(X>t)$$

Exponential is the only memoryless distribution

Matlab exercise:

- Generate a sample of 100,000 variables from Exponential distribution with r =0.1
- Calculate mean and compare it to 1/r
- Calculate standard deviation and compare it to 1/r
- Plot semilog-y plots of PDFs **and CCDFs.**
- Hint: read the help page (better yet documentation webpage) for random('Exponential'…) one of their parameters is different than r

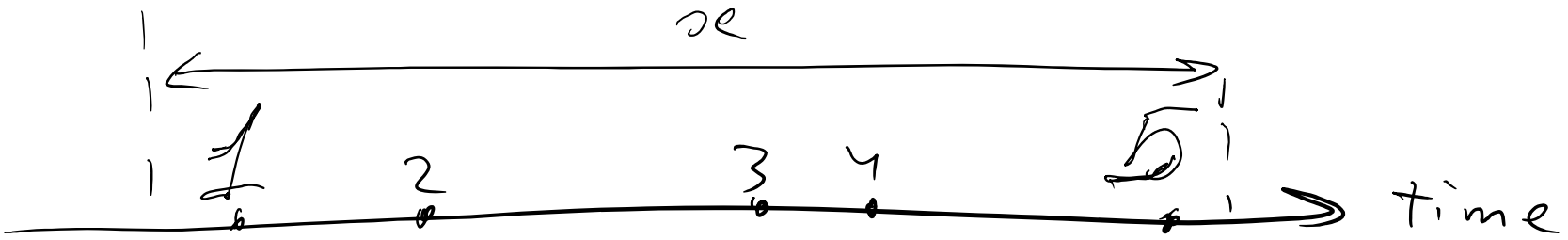
Matlab exercise: Exponential

- **Stats=100000; r=0.1;**
- **r2=random('Exponential', 1./r, Stats,1);**
- **disp([mean(r2),1./r]); disp([std(r2),1./r]);**
- **step=1; [a,b]=hist(r2,0:step:max(r2));**
- **pdf_e=a./sum(a)./step;**
- **subplot(1,2,1); semilogy(b,pdf_e,'rd-’);**
- **x=0:0.01:max(r2);**
- **for m=1:length(x);**
- **ccdf_e(m)=sum(r2>x(m))./Stats;**
- **end;**
- **subplot(1,2,2); semilogy(x,ccdf_e,'ko-');**

Erlang Distribution

- The Erlang distribution is a generalization of the exponential distribution.
- The exponential distribution models the time interval to the 1^{st} event, while the
- Erlang distribution models the time interval to the k^{th} event, i.e., a sum of k exponentially distributed variables.
- The exponential, as well as Erlang distributions, is based on the constant rate (or Poisson) process.

Constant rate (Poisson) process



Events happen <u>independently</u> from each other at <u>constant rate = r</u> $\therefore E[N_x] = rx$

- X follows Erlang distribution

$$P(X > x) = \sum_{n=0}^{r-1} P(N_x = n) =$$

$$= \sum \frac{(rx)^n}{n!} e^{-rx} \quad n = 0$$

Erlang Distribution

Generalizes the Exponential Distribution:

waiting time until k's events

(constant rate process with rate=r)

$$P(X > x) = \sum_{m=0}^{k-1} \frac{e^{-rx}(rx)^m}{m!} = 1 - F(x)$$

Differentiating $F(x)$ we find that all terms in the sum except the last one cancel each other:

$$f(x) = \frac{r^k x^{k-1} e^{-rx}}{(k-1)!} \quad \text{for } x > 0 \text{ and } k = 1, 2, 3, \ldots$$

Gamma Distribution

The random variable *X* with a probability density function:

$$f(x) = \frac{r^k x^{k-1} e^{-rx}}{\Gamma(k)}, \text{ for } x > 0 \qquad (4\text{-}18)$$

has a gamma random distribution with parameters r > 0 and *k* > 0. If *k* is a positive integer, then *X* has an Erlang distribution.

$$f(x)=\frac{r^k x^{k-1} e^{-rx}}{\Gamma(k)},\ \text{for}\ x>0$$

$$\int_0^{+\infty} f(x)\,dx = 1\,,\quad \text{Hence}$$

$$\Gamma(k)=\int_0^{+\infty} r^k x^{k-1} e^{-rx}\,dx=\int_0^{+\infty} y^{k-1} e^{-y}\,dy$$

Comparing with Erlang distribution

for integer k one gets

$$\Gamma(k)=(k-1)!$$

Gamma Function

The gamma function is the generalization of the factorial function for r > 0, not just non-negative integers.

$$\Gamma(k) = \int_0^\infty y^{k-1} e^{-y} dy, \quad \text{for } r > 0 \qquad (4\text{-}17)$$

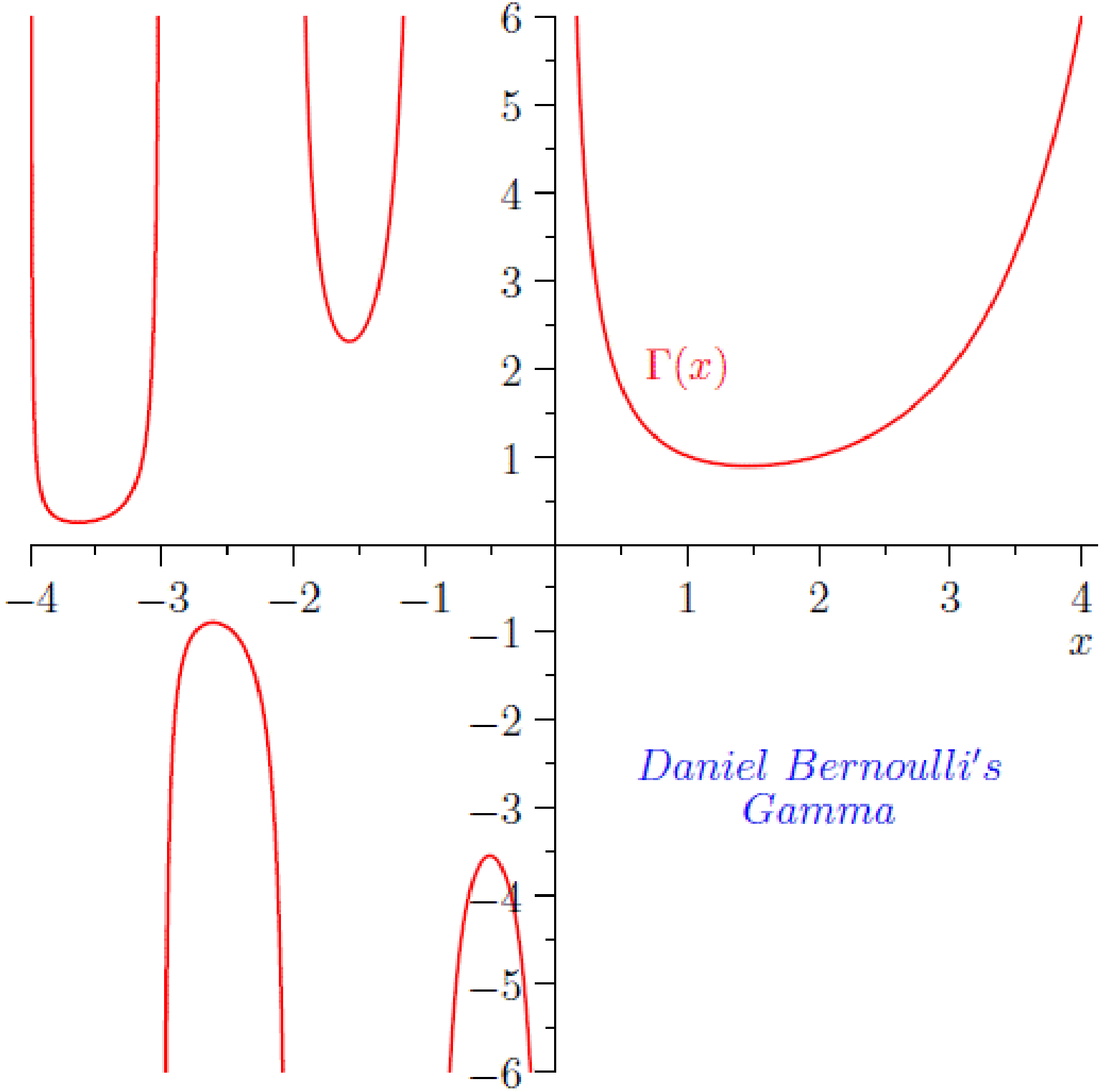
Properties of the gamma function

$$\Gamma(1) = 1$$

$$\Gamma(k) = (k-1)\Gamma(k-1) \quad \text{recursive property}$$

$$\Gamma(k) = (k-1)! \quad \text{factorial function}$$

$$\Gamma(1/2) = \pi^{1/2} = 1.77 \quad \text{interesting fact}$$


6
5
4
3
2
1
−1
−2
−3
−4
−5
−6
−4
−3
−2
−1
1
2
3
4
x
Γ(x)
Daniel Bernoulli's
Gamma

SOLO

BERNOULLI FAMILY

Bernoulli trials

SOLO HERMELIN

http://www.solohermelin.com

The Bernoulli family

Nicolaus
1623-1708

Jacob	Nicolaus	Johann
1654-1705	1662-1716	1667-1748
	Nicolaus (I) 1687-1759	

Nicolaus (II)	Daniel	Johann (II)
1695-1726	1700-1782	1710-1790

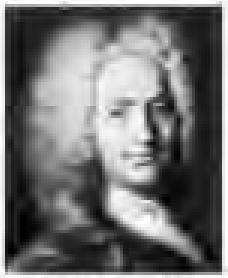
Johann (III)	Daniel (II)	Jacob (II)
1744-1807	1751-1834	1759-1789

Those shown in **bold** above are in our archive

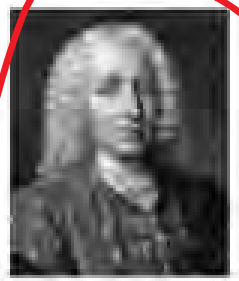
Row This

Jacob
1654-1705

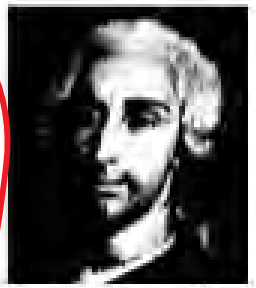
Johann
1667-1748



Nicolaus II
1695-1720



Daniel
1700-1782



Johann II
1710-1790

Johann III
1744-1807

Jacob II
1759-1789

Gamma function

Mean & Variance of the Erlang and Gamma

- If X is an Erlang (or more generally Gamma) random variable with parameters r and k,

  $\mu = E(X) = k/r$ and $\sigma^2 = V(X) = k/r^2$ (4-19)

- Generalization of exponential results:

  $\mu = E(X) = 1/r$ and $\sigma^2 = V(X) = 1/r^2$ or

  Negative binomial results:

  $\mu = E(X) = k/p$ and $\sigma^2 = V(X) = k(1-p) / p^2$

Matlab exercise:

- Generate a sample of 100,000 variables with "Harry Potter" Gamma distribution with r =0.1 and k=9 ¾ (9.75)
- Calculate mean and compare it to k/r (Gamma)
- Calculate standard deviation and compare it to sqrt(k)/r (Gamma)
- Plot semilog-y plots of PDFs **and CCDFs.**
- Hint: read the help page (better yet documentation webpage) for random('Gamma'…): one of their parameters is different than r

Matlab exercise: Gamma

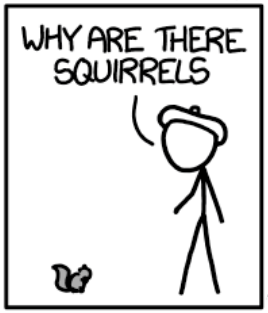
```
Stats=100000; r=0.1; k=9.75;
r2=random('Gamma', k,1./r, Stats,1);
disp([mean(r2),k./r]);
disp([std(r2),sqrt(k)./r]);
step=0.1; [a,b]=hist(r2,0:step:max(r2));
pdf_g=a./sum(a)./step;
figure;
subplot(1,2,1); semilogy(b,pdf_g,'ko-'); hold on;
x=0:0.01:max(r2); clear cdf_g;
for m=1:length(x);
    cdf_g(m)=sum(r2>x(m))./Stats;
end;
subplot(1,2,2); semilogy(x,cdf_g,'rd-');
```

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 IS LYING GOOD

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 SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
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 IS YKK ON ALL ZIPPERS

WHY AREN'T MY ARMS GROWING

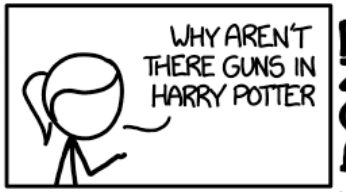
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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

Continuous Probability Distributions

Normal or Gaussian Distribution


PAY
ATTENTION

Normal or Gaussian Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$-\infty < x < \infty$

is a normal random variable

with mean μ,

and standard dewviation σ

sometimes denoted as

$N(\mu, \sigma)$

Carl Friedrich Gauss (1777 –1855)
German mathematician

Normal Distribution

- The location and spread of the normal are independently determined by mean (μ) and standard deviation (σ)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

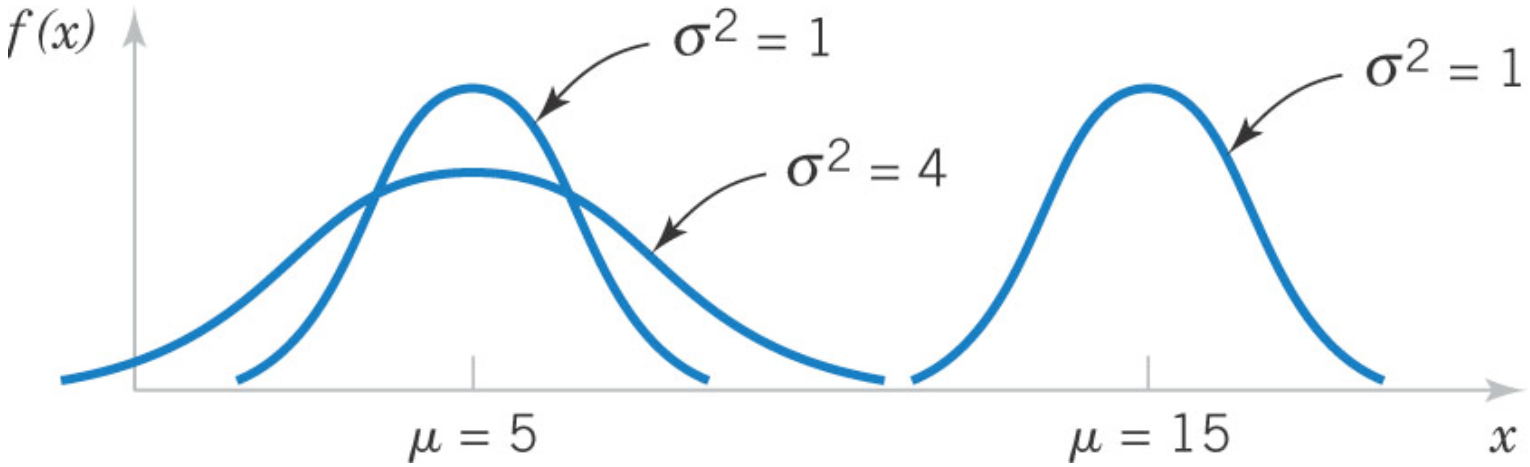


Figure 4-10 Normal probability density functions

Matlab exercise:

plot PDF of the Gaussian distribution

with **mu=3; sigma=2**

calculate mean, standard deviation and variance,

Linear-y and Semilog-y plots of PDF

Hint:

Generate Standard normal distribution using

randn(Stats,1) then

multiply and add using sigma, mu

Matlab exercise solution

- **Stats=100000;**
- **mu=3; sigma=2;**
- **r1=sigma.*randn(Stats,1)+mu;**
- **step=0.1;**
- **[a,b]=hist(r1,(mu-10.*sigma):step:(mu+10.*sigma));**
- **pdf_n=a./sum(a)./step;**
- **figure; subplot(1,2,1); plot(b,pdf_n,'ko-');**
- **subplot(1,2,2); semilogy(b,pdf_n,'ko-');**

Gaussian (Normal) distribution is very important because any sum of many independent random variables can be approximated with a Gaussian

Standard Normal Distribution

- A normal (Gaussian) random variable with

  $$\mu = 0 \text{ and } \sigma^2 = 1$$

  is called a standard normal random variable and is denoted as Z.
- Thed cumulative distribution function of a standard normal random variable is denoted as:

  $$\Phi(z) = P(Z \leq z)$$
- Values are found in Appendix A Table III to Montgomery and Runger textbook

Standardizing

If X is a normal random variable with $E(X) = \mu$ and $V(X) = \sigma^2$, the random variable

$$Z = \frac{X - \mu}{\sigma} \qquad (4\text{-}10)$$

is a normal random variable with $E(Z) = 0$ and $V(Z) = 1$. That is, Z is a standard normal random variable.

Suppose X is a normal random variable with mean μ and variance σ^2.

Then, $P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P(Z \le z)$ (4-11)

where Z is a standard normal random variable, and

$z = \frac{(x-\mu)}{\sigma}$ is the z-value obtainedby standardizing x.

The probability is obtained by using Appendix Table III

$P(X < \mu - \sigma) = P(X > \mu + \sigma) = (1-0.68)/2=0.16=16\%$

$P(X < \mu - 2\sigma) = P(X > \mu + 2\sigma) = (1- 0.95)/2=0.023=2.3\%$

$P(X < \mu - 3\sigma) = P(X > \mu + 3\sigma) = (1- 0.997)/2=0.0013=0.13\%$

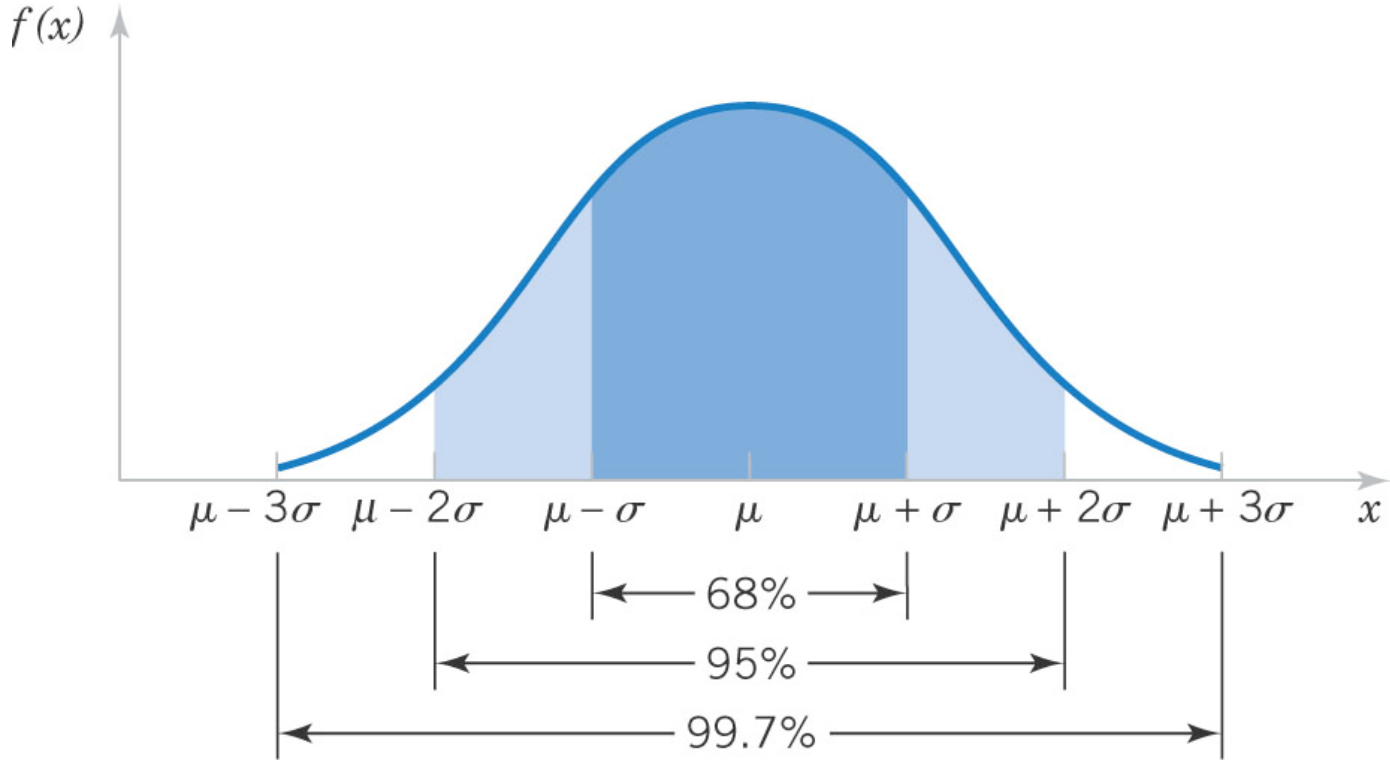


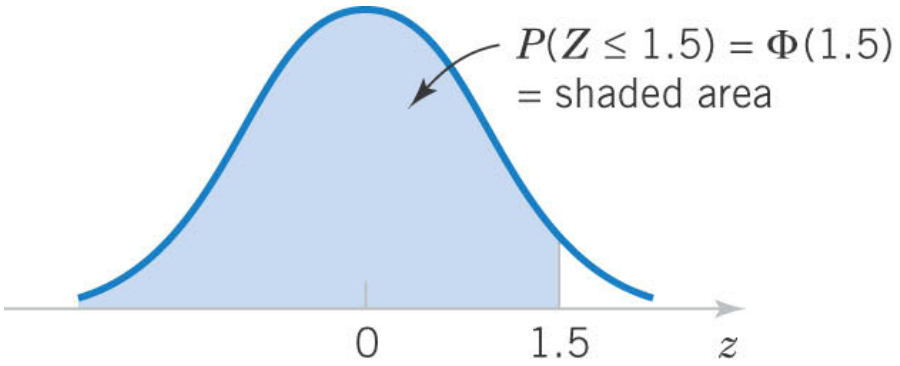
Figure 4-12 Probabilities associated with a normal distribution – well worth remembering to quickly estimate probabilities.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

Standard Normal Distribution Tables

Assume Z is a standard normal random variable.

Find $P(Z \leq 1.50)$. Answer: 0.93319



z	0.00	0.01	0.02	0.03
0	0.50000	0.50399	0.50398	0.51197
⋮		⋮		
1.5	0.93319	0.93448	0.93574	0.93699

Figure 4-13 Standard normal PDF

Table III from, Appendix A in Montgomery & Runger

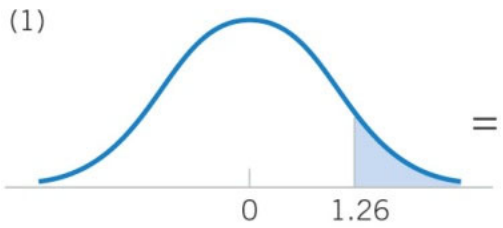
Find $P(Z \leq 1.53)$. Answer: 0.93699

Find $P(Z \leq 0.02)$. Answer: 0.50398

Standard Normal Exercises

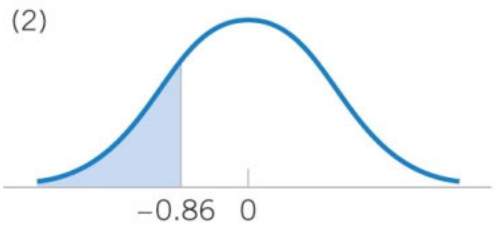
1. $P(Z > 1.26) = 1 - P(Z < 1.26) = 1 - 0.8962 =$

   $= 0.1038$

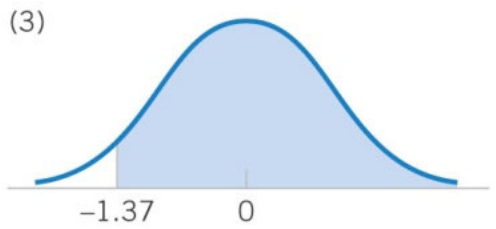


2. $P(Z < -0.86) = P(Z > 0.86) = 1 - P(Z < 0.86) =$

   $1 - 0.815 = 0.195$



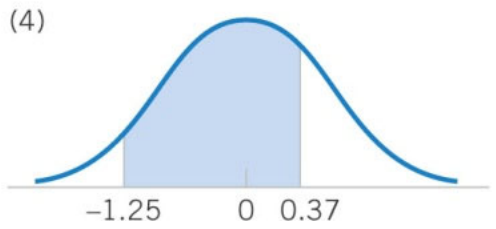
3. $P(Z > -1.37) = P(Z < 1.37) = 0.915$



4. $P(-1.25 < Z < 0.37) = P(Z < 0.37) - P(Z < -1.25)$

   $= P(Z < 0.37) - P(Z > 1.25) = P(Z < 0.37) -$

   $(1 - P(Z < 1.25)) = 0.6443 - (1 - 0.8944) = 0.5387$



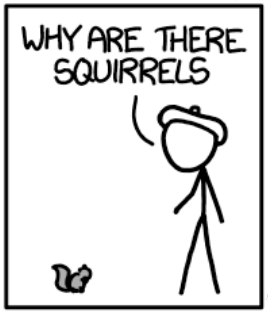
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

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 IS LYING GOOD
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE LAVA
WHY IS YKK ON ALL ZIPPERS
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY IS LIFE SO BORING

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 SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
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

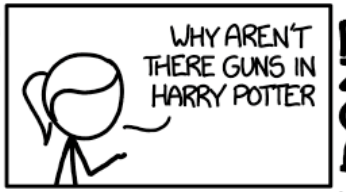
WHY AREN'T MY ARMS GROWING

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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 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 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

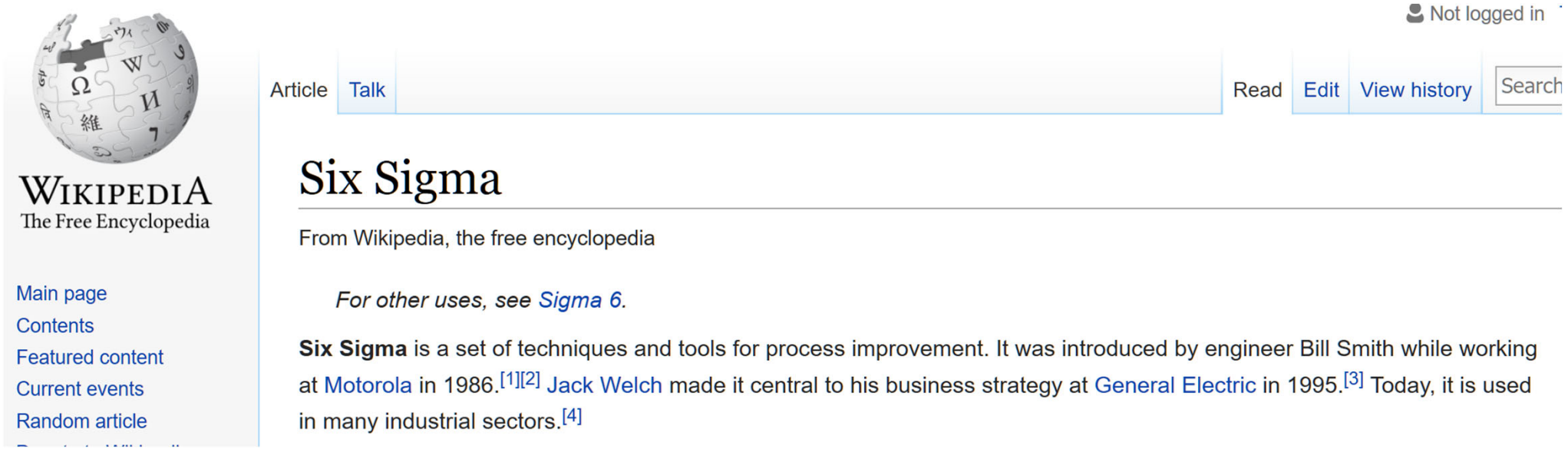


WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Range	The expected fraction of population inside the range	Approximate expected frequency outside the range	The approximate frequency for daily event
μ ± 0.5σ	0.382924922548026	2 in 3	Four or five times a week
μ ± 1σ	0.682689492137086	1 in 3	Twice a week
μ ± 1.5σ	0.866385597462284	1 in 7	Weekly
μ ± 2σ	0.954499736103642	1 in 22	Every three weeks
μ ± 2.5σ	0.987580669348448	1 in 81	Quarterly
μ ± 3σ	0.997300203936740	1 in 370	Yearly
μ ± 3.5σ	0.999534741841929	1 in 2149	Every six years
μ ± 4σ	0.999936657516334	1 in 15787	Every 43 years (twice in a lifetime)
μ ± 4.5σ	0.999993204653751	1 in 147160	Every 403 years (once in the modern era)
μ ± 5σ	0.999999426696856	1 in 1744278	Every 4776 years (once in recorded history)
μ ± 5.5σ	0.999999962020875	1 in 26330254	Every 72090 years (thrice in history of modern humankind)
μ ± 6σ	0.999999998026825	1 in 506797346	Every 1.38 million years (twice in history of humankind)
μ ± 6.5σ	0.999999999919680	1 in 12450197393	Every 34 million years (twice since the extinction of dinosaurs)
μ ± 7σ	0.999999999997440	1 in 390682215445	Every 1.07 billion years (four times in history of Earth)

Source: Wikipedia

Business buzzword: Six Sigma



Not logged in

WIKIPEDIA
The Free Encyclopedia

Article | Talk | Read | Edit | View history | Search

Main page
Contents
Featured content
Current events
Random article

Six Sigma

From Wikipedia, the free encyclopedia

For other uses, see Sigma 6.

Six Sigma is a set of techniques and tools for process improvement. It was introduced by engineer Bill Smith while working at Motorola in 1986.[1][2] Jack Welch made it central to his business strategy at General Electric in 1995.[3] Today, it is used in many industrial sectors.[4]

Business literature defined six sigma as no more than 3.4 defective products per million

Matlab group exercise 3

- $P(X-\mu>z\cdot\sigma)=P(Z>z)=(1-\text{erf}(z./\text{sqrt}(2)))/2$
- You can also use 1-normcdf(z)
- Calculate Prob(X-μ>6σ) and compare with expected 3.4 errors per million
- Find z such that Prob(X-μ>z · σ)=3.4 errors per million

What Six Sigma should be really called if $P(X-\mu>z\cdot\sigma) = 3.4e\text{-}6$

A. 6 sigma
B. 7 sigma
C. 3 sigma
D. 4.5 sigma
E. I could not figure it out

Get your i-clickers

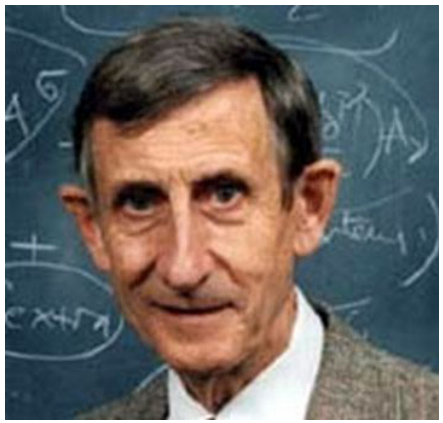
Appendix Table III is no good for 6-sigma How to calculate in Matlab?

- Matlab has a built-in function normcdf
- 1-normcdf(z) is the Prob[X-μ>z·σ]
- I expected: P(Z>6)= 3.4e-6
- Matlab says 1-normcdf(6)~ 1e-9
- Six sigma is not 6σ at all !!!
- Let's find out how many simas are in six sigma
- Matlab says: invnorm(3.4e-6)=4.5
- Six sigma should be called 4.5σ
- Does not have the same buzz

What's wrong with Six Sigma?

- Motorola has determined, through years of process and data collection, that processes vary and drift over time – what they call the Long-Term Dynamic Mean Variation. This variation typically falls between 1.4 and 1.6. They shifted their sigma down by 1.5.
- The statistician Donald J. Wheeler has dismissed the 1.5 sigma shift as "goofy" because of its arbitrary nature.
- A *Fortune* article stated that "of 58 large companies that have announced Six Sigma programs, 91 percent have trailed (performed below) the S&P 500 index since"

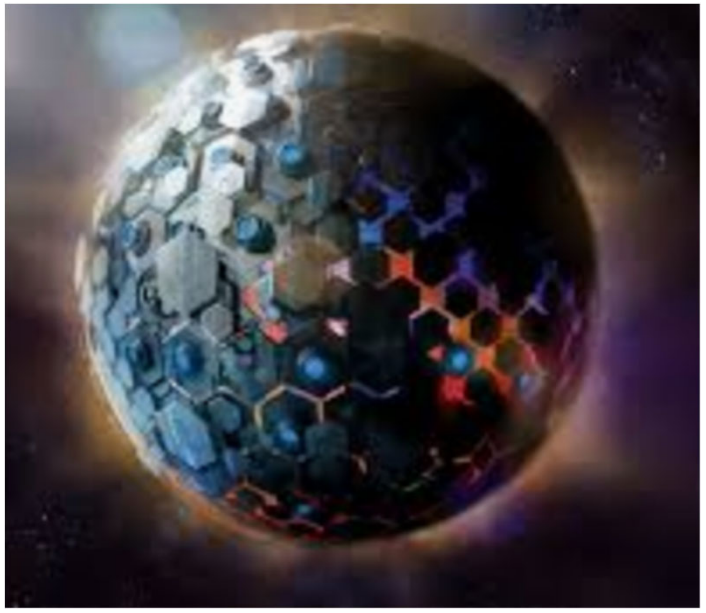
- Freeman Dyson (a famous theoretical physicist) once sat on a committee reviewing Department of Energy Joint Genomics Institute (DOE JGI)
- Motorola sent their six-sigma preacher Freeman Dyson asked him:
  - D: Can you explain me what is six–sigma?
    - P: Mumbling something about it being the gold standard of reliability
  - D: Can you at least define one-sigma?
    - P: Silence
- Six-sigma was never implemented at JGI



Born:
December 15, 1923,
Crowthorne, UK
Died:
February 28, 2020
Princeton, NJ USA

Dyson's legacy

- **Seminal contributions to quantum mechanics**
- The Origin of Life:
  Cells → Enzymes → DNA/RNA later
  First proposed by Alexander Oparin in 1922
- Dyson sphere:
  Completely
  captures light from a star
- Dyson tree:
  genetically engineered
  tree growing inside a
  comet

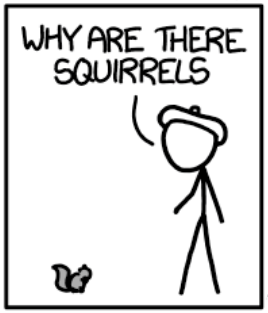


Credit: XKCD comics

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 IS LYING GOOD
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE LAVA
WHY IS YKK ON ALL ZIPPERS
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY IS LIFE SO BORING

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 SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
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

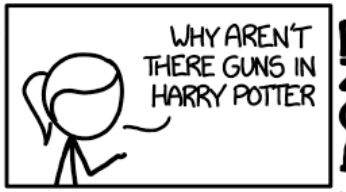
WHY AREN'T MY ARMS GROWING

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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 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 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

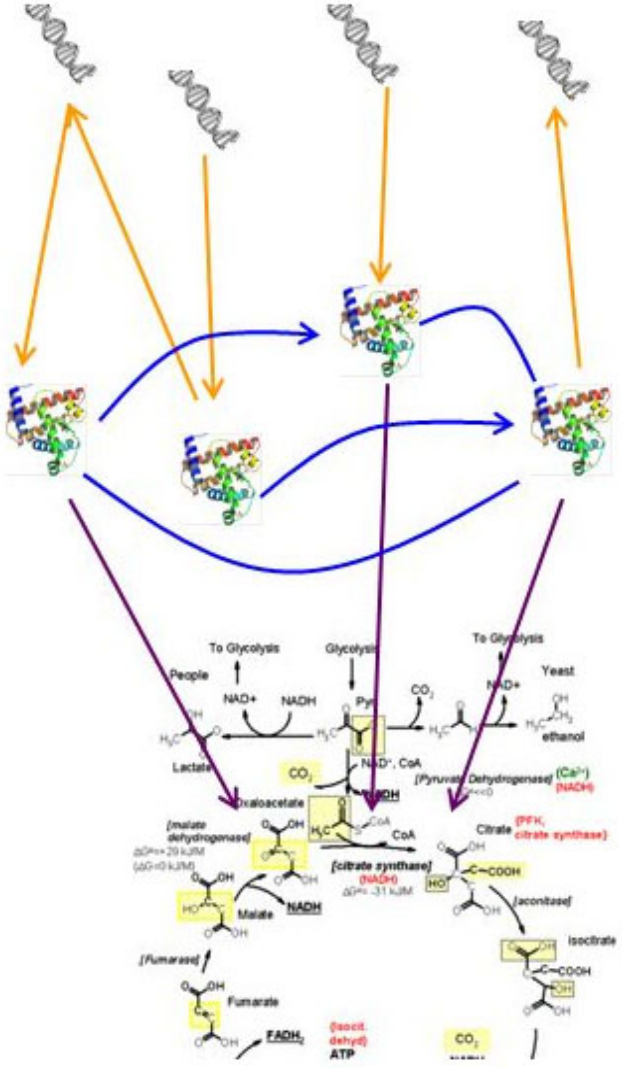


WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Fitting a Gaussian distribution: a biological example

Molecular binding is used at multiple levels

Each level has its own molecular interaction network



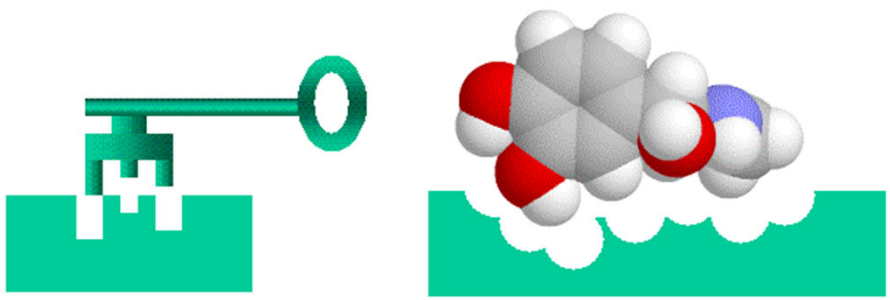
Regulatory network:
RNA-level regulation
By DNA-binding Proteins

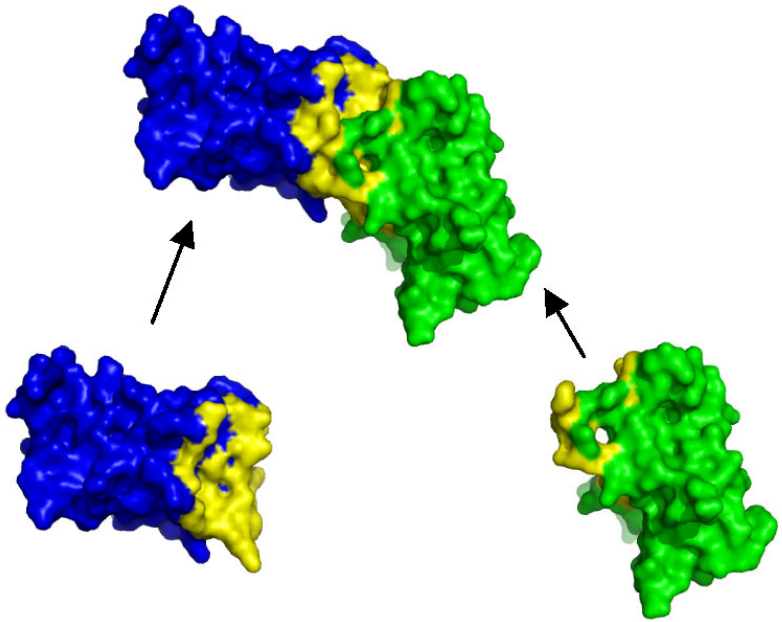
Protein-Protein (binding) Interaction Network

Protein-Metabolite Interactions: Metabolic network

Biological example of a Gaussian: Energy of Protein-Protein Binding Interactions

- Proteins and other biomolecules (metabolites, drugs, DNA) specifically (and non-specifically) bind each other
- For specific bindings: Lock-and-Key theory
- For non-specific bindings: random contacts





A simple physical model for scaling in protein–protein interaction networks

Eric J. Deeds*, Orr Ashenberg†, and Eugene I. Shakhnovich‡§

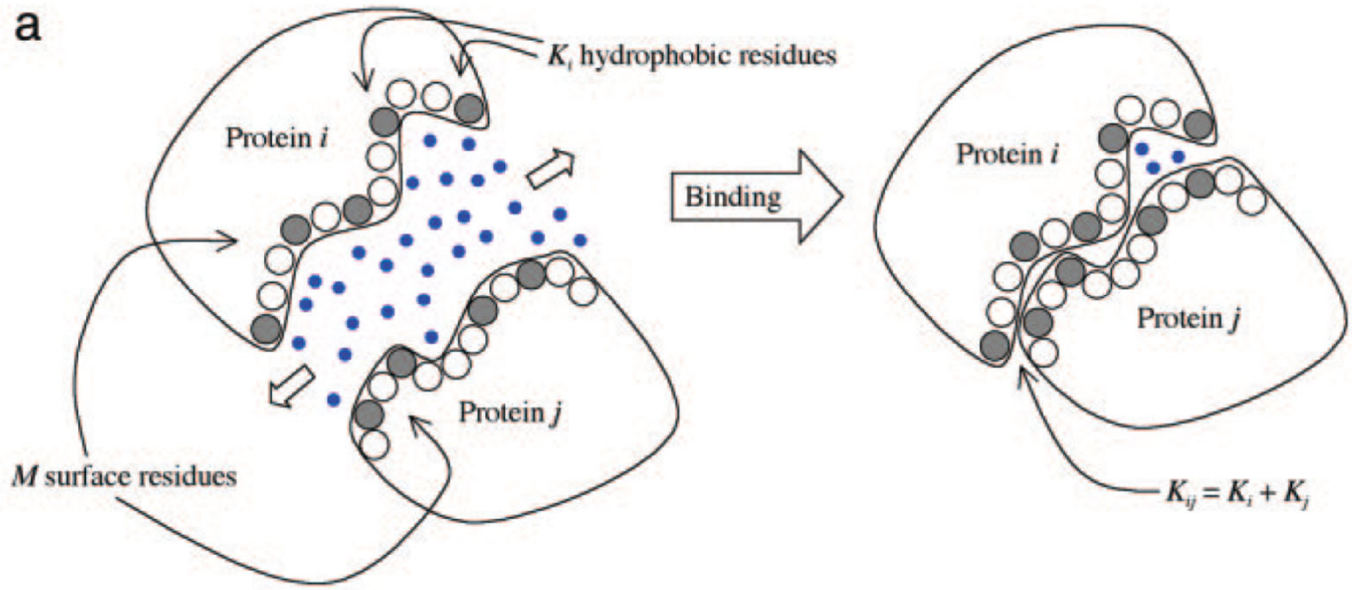
*Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138; †Harvard College, 12 Oxford Street, Cambridge, MA 02138; and ‡Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

Communicated by David Chandler, University of California, Berkeley, CA, November 10, 2005 (received for review September 23, 2005)

It has recently been demonstrated that many biological networks exhibit a "scale-free" topology, for which the probability of observing a node with a certain number of edges (k) follows a power law: i.e., $p(k) \sim k^{-\gamma}$. This observation has been reproduced by

(19–22). Indeed, when the two major *S. cerevisiae* p protein interaction (PPI) experiments are compared w another, one finds that only ≈150 of the thousands of tions identified in each experiment are recovered in th

Most Binding energy is due to hydrophobic amino-acid residues being screened from water

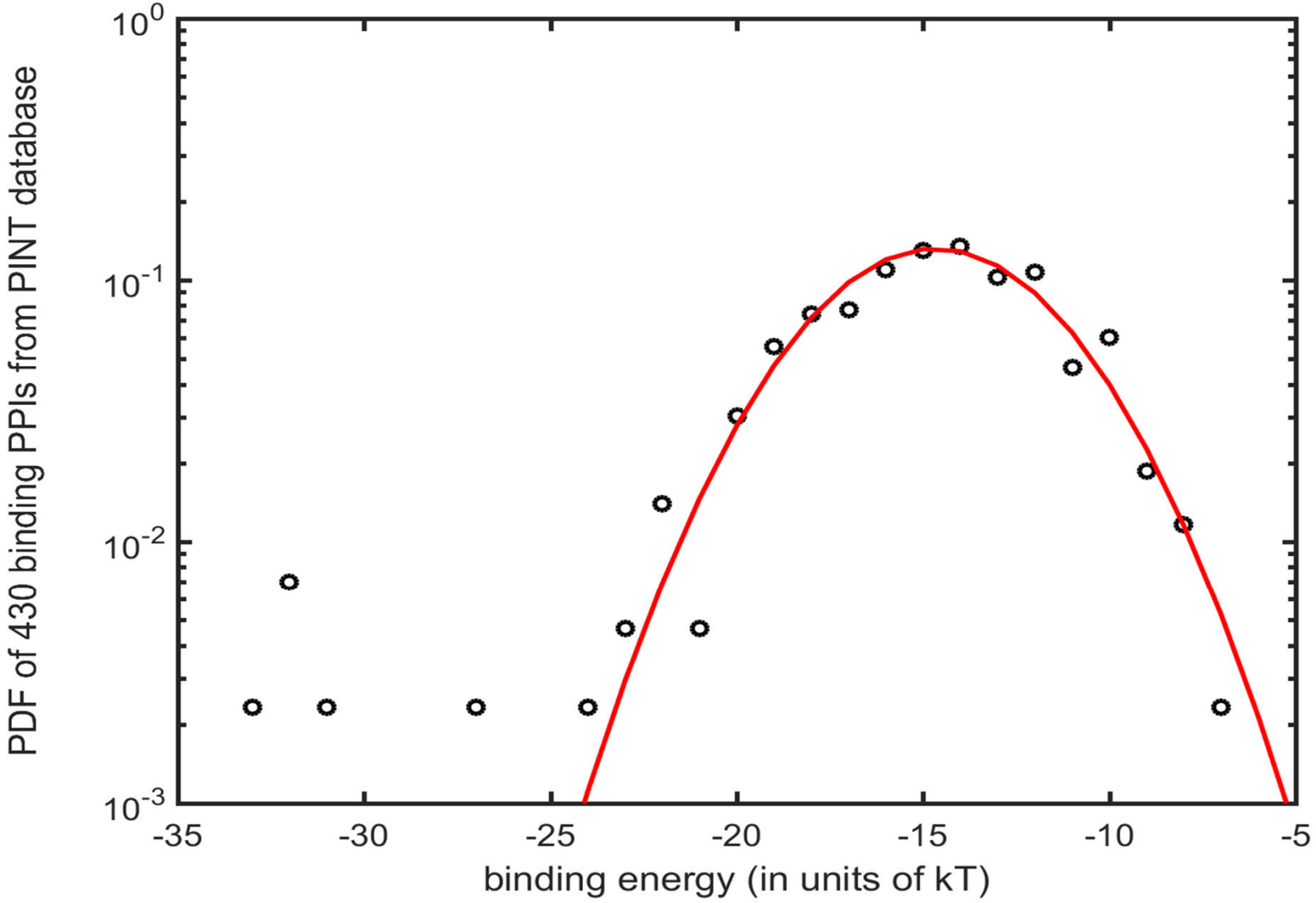


Predicted Gaussian distribution: PDF(E_{ij}=E)– because E_{ij} – sum of hydrophobicities of many independent residues

Matlab exercise

- In Matlab load PINT_binding_energy.mat with binding energy E_{ij} (in units of kT at room temperature) for 430 pairs of interacting proteins from human, yeast, etc.
- Data collected in 2007 from the PINT database http://www.bioinfodatabase.com/pint/
  and analyzed in J. Zhang, **S. Maslov**, E. Shakhnovich, Molecular Systems Biology (2008)
- Fit Gaussian to the distribution of E_{ij} using dfittool
- Use "Exclude" button to generate the new exclusion rule to drop all points with X<-23 from the fit
- Use "New Fit" button to generate the new "Normal" fit with the exclusion rule you just created
- Find mean (mu) and standard deviation (sigma)
- Select "probability plot" from "Display type" dropdown menu to evaluate the quality of the plot. Where does the probability plot deviate from a straight line?

How does it compare with the experimental data ?



J. Zhang, **S. Maslov**, E. Shakhnovich,
Nature/EMBO Molecular Systems Biology (2008)

Data on binding interactions
from PINT database

Dissociation constant

- Interaction between two molecules (say, proteins) is usually described in terms of dissociation constant
  $K_{ij}=1M \exp(-E_{ij}/kT)$
- Law of Mass Action: the concentration D_{ij} of a heterodimer formed out of two proteins with free (monomer) concentrations C_i and C_j : $D_{ij}=C_iC_j/K_{ij}$
- What is the distribution of K_{ij}?
- Answer: it is called log-normal since the logarithm of K_{ij} is the binding energy $-E_{ij}/kT$ which is normally distributed

Lognormal Distribution

- Let *W* denote a normal random variable with mean of θ and variance of ω^2, i.e., $E(W) = \theta$ and $V(W) = \omega^2$
- As a change of variable, let $X = e^W = \exp(W)$ and $W = \ln(X)$
- Now X is a lognormal random variable.

$$F(x) = P[X \le x] = P[\exp(W) \le x] = P[W \le \ln(x)]$$

$$= P\left[Z \le \frac{\ln(x) - \theta}{\omega}\right] = \Phi\left[\frac{\ln(x) - \theta}{\omega}\right] = \quad \text{for} \quad x > 0$$

$$= 0 \quad \text{for} \quad x \le 0$$

$$f(x) = \frac{dF(x)}{dx} = \frac{1}{x\omega\sqrt{2\pi}} e^{-\left[\frac{\ln(x)-\theta}{2\omega}\right]^2} \qquad \text{for } 0 < x < \infty$$

$$E(X) = e^{\theta + \omega^2/2} \qquad \text{and} \qquad V(X) = e^{2\theta+\omega^2}\left(e^{\omega^2} - 1\right) \qquad (4\text{-}22)$$

Lognormal Graphs

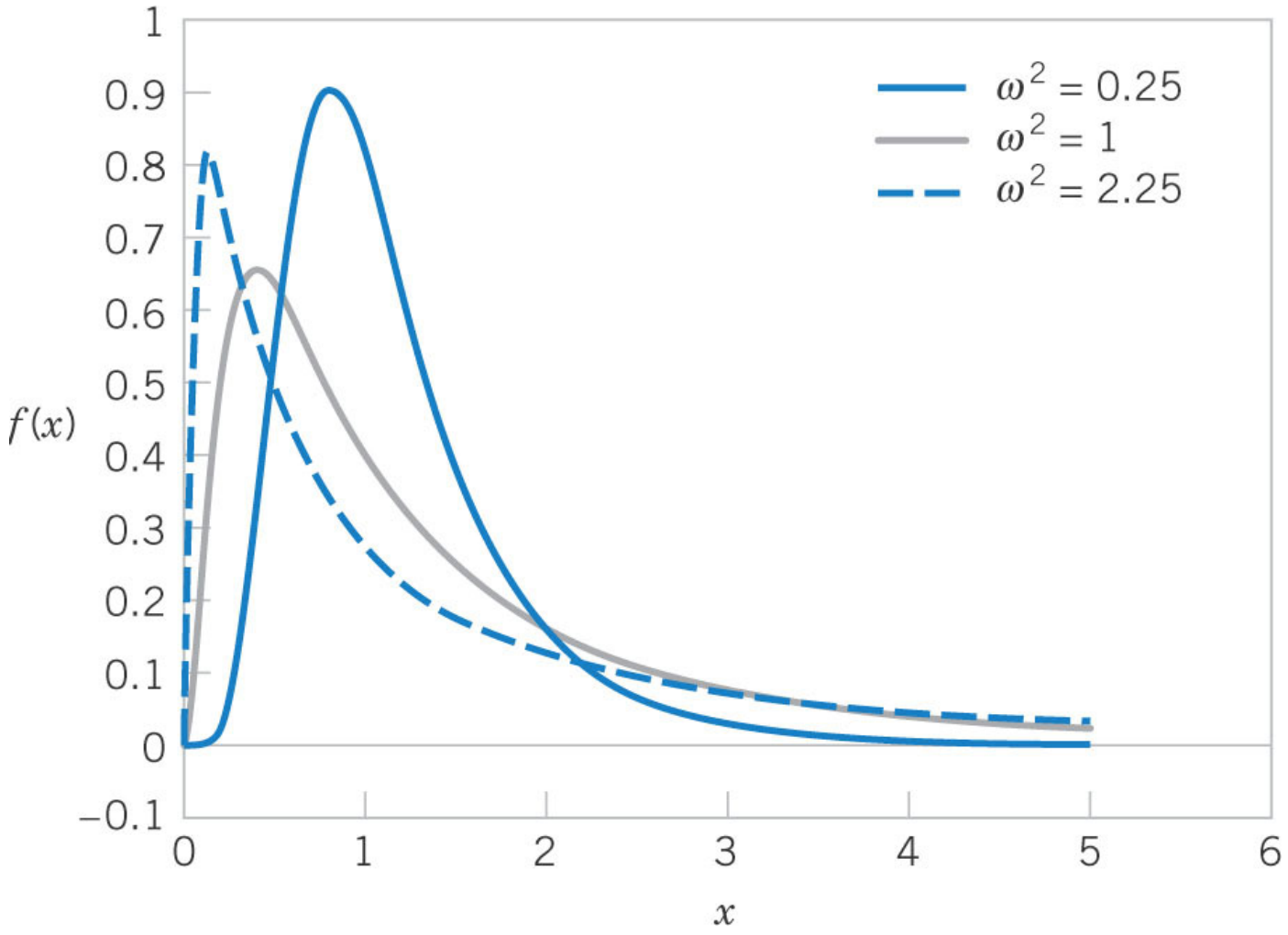


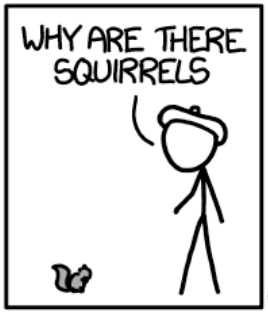
Figure 4-27 Lognormal probability density functions with θ = 0 for selected values of ω^2.

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 IS LYING GOOD
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE LAVA
WHY IS YKK ON ALL ZIPPERS
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY IS LIFE SO BORING

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 SEX SO IMPORTANT

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
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

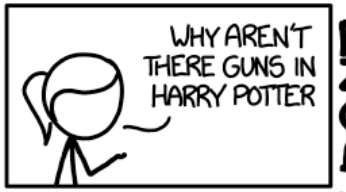
WHY AREN'T MY ARMS GROWING

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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 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 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