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 Mean

X is a binomial random variable
with parameters *p* and *n*

Mean:

$\mu = E(X) = np$

$$\mu = \sum x C_x^n p^x q^{n-x} = p \frac{\partial}{\partial p} \sum C_x^n p^x q^{n-x} =$$

$$= p \frac{\partial}{\partial p} (p+q)^n = np$$

Binomial mean, variance and standard deviation

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

- Mean:

$\mu = np$

- Variance:

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

- Standard deviation:

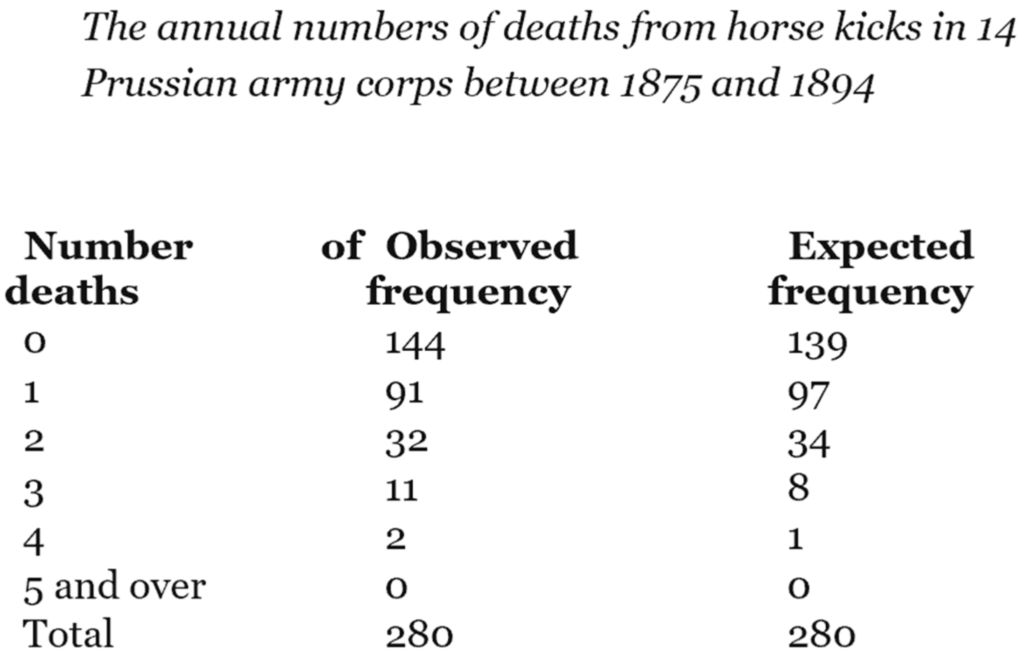
$\sigma = \sqrt{np(1-p)}$

- Standard deviation to mean ratio

$$\sigma/\mu = \sqrt{np(1-p)}/np = \frac{\sqrt{(1-\mathrm{p})/\mathrm{p}}}{\sqrt{n}}$$

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

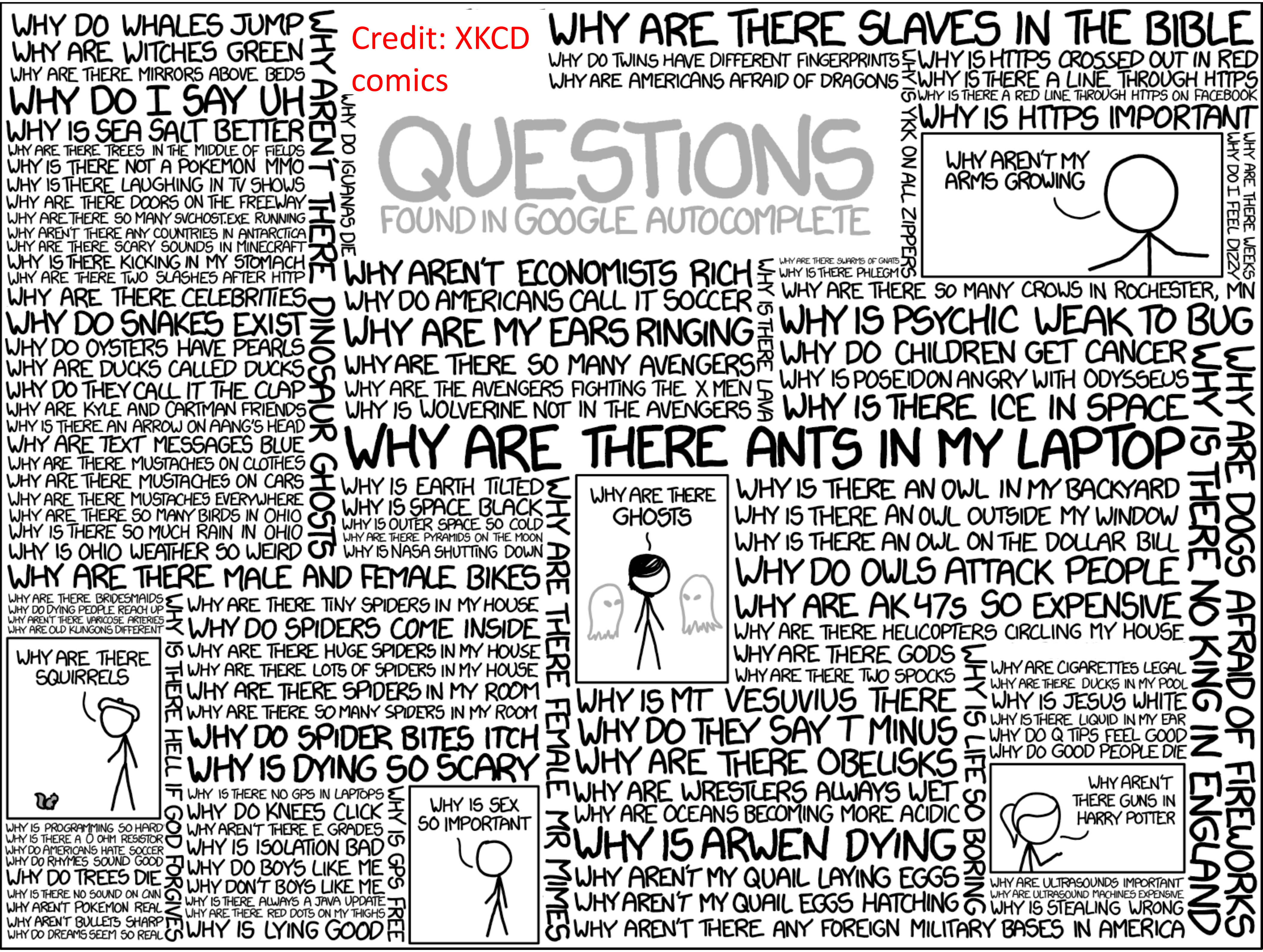
Matlab exercise: Poisson distribution

- Generate a sample of size 100,000 for Poisson-distributed random variable X with λ =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]= λ and V[X]=λ

Matlab exercise: Poisson distribution

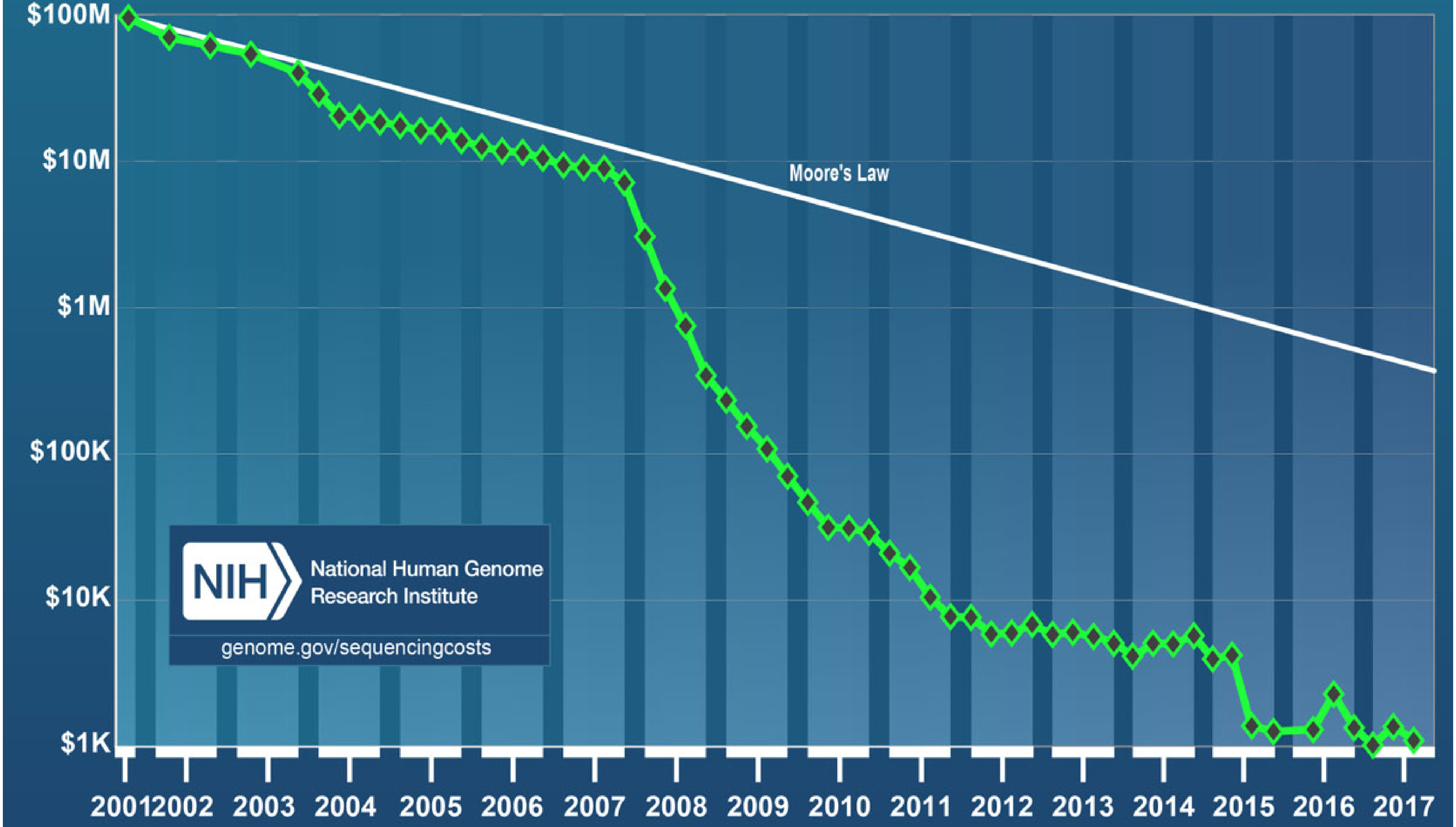
- **Stats=100000; lambda=2;**
- **r2=random('Poisson',lambda,Stats,1);**
- **mean(r2)**
- **var(r2)**
- **[a,b]=hist(r2, 0:max(r2));**
- **p_p=a./sum(a);**
- **figure; stem(b,p_p);**
- **figure; semilogy(b,p_p,'ko-')**

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 AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
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 IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
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 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 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 AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Poisson Distribution in Genome Assembly

Cost per Genome

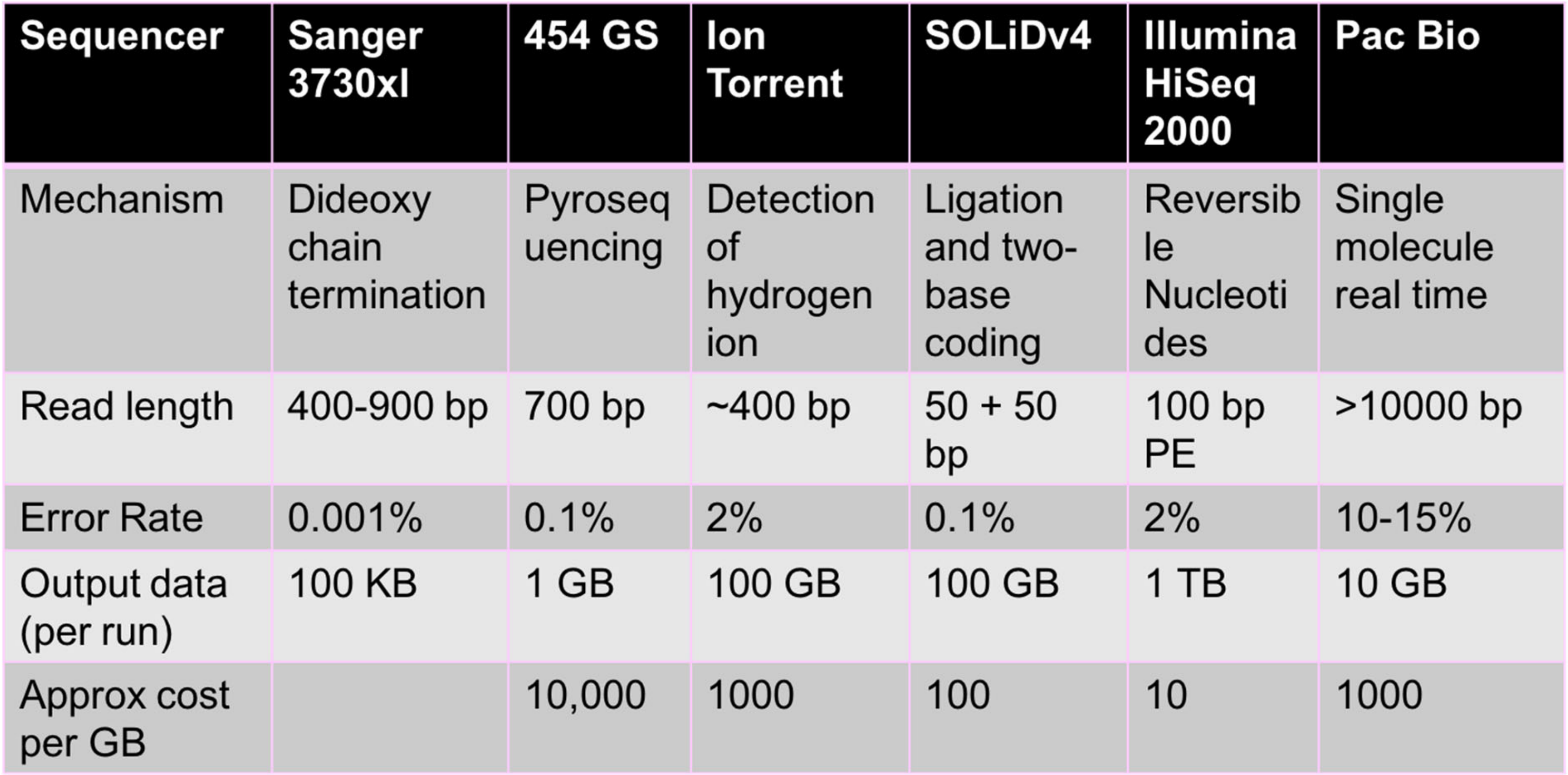


NIH National Human Genome Research Institute

genome.gov/sequencingcosts

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

Short Reads assemble into Contigs

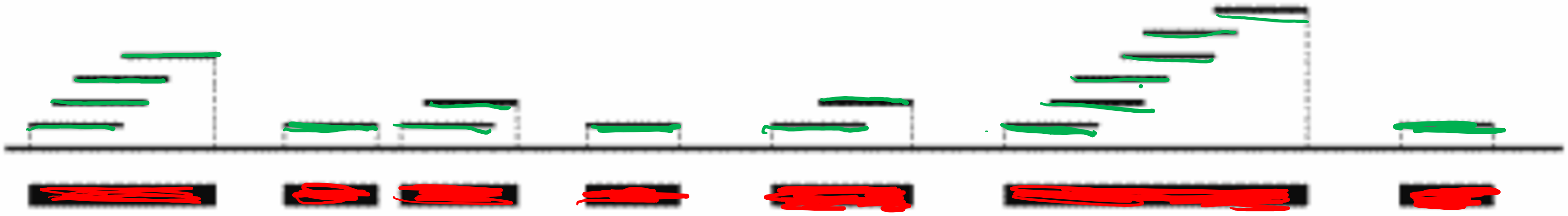
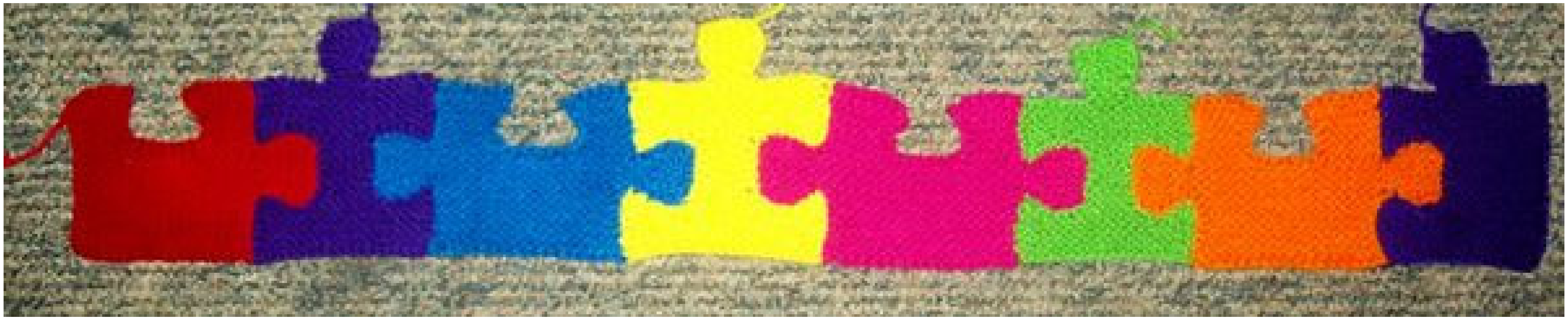
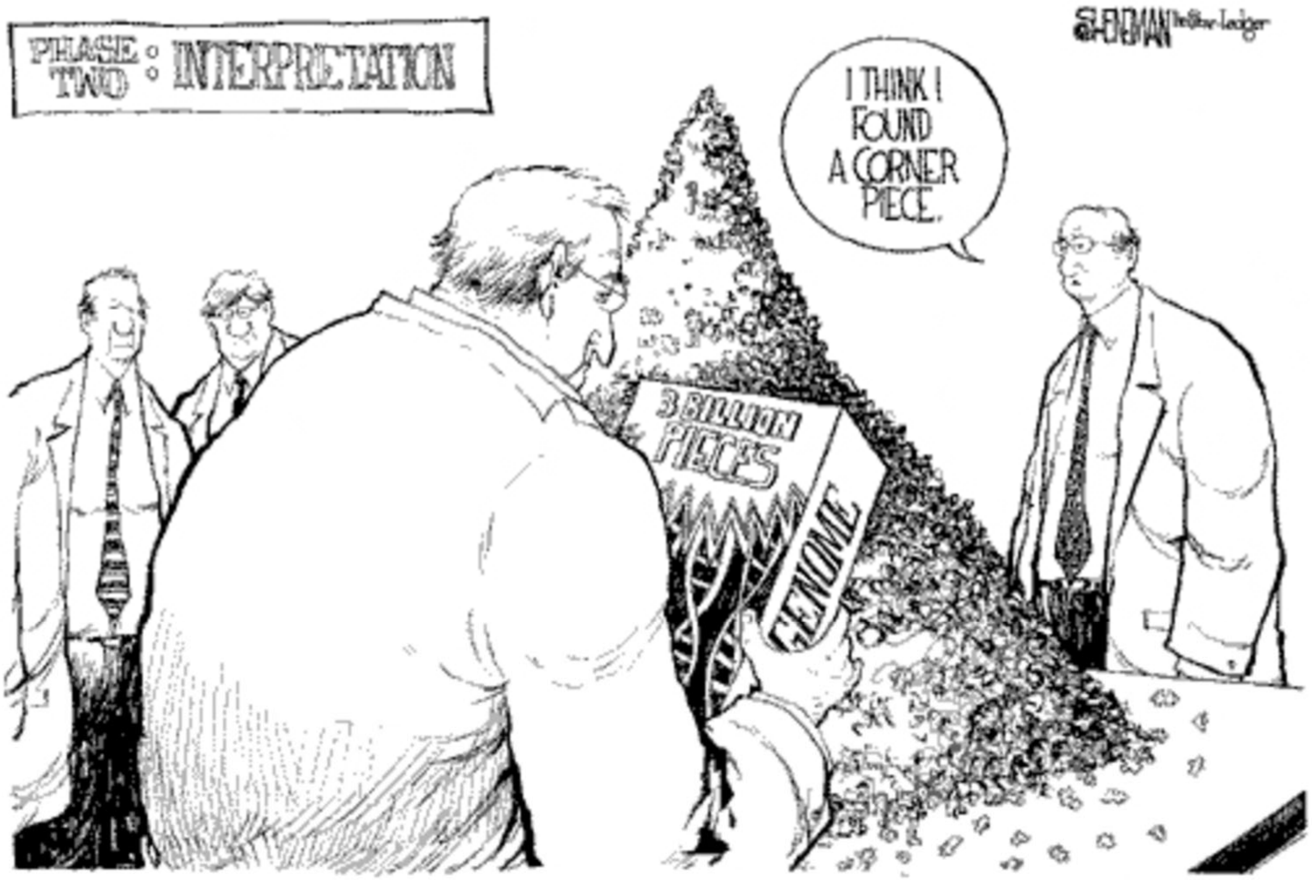


Figure 5.1.



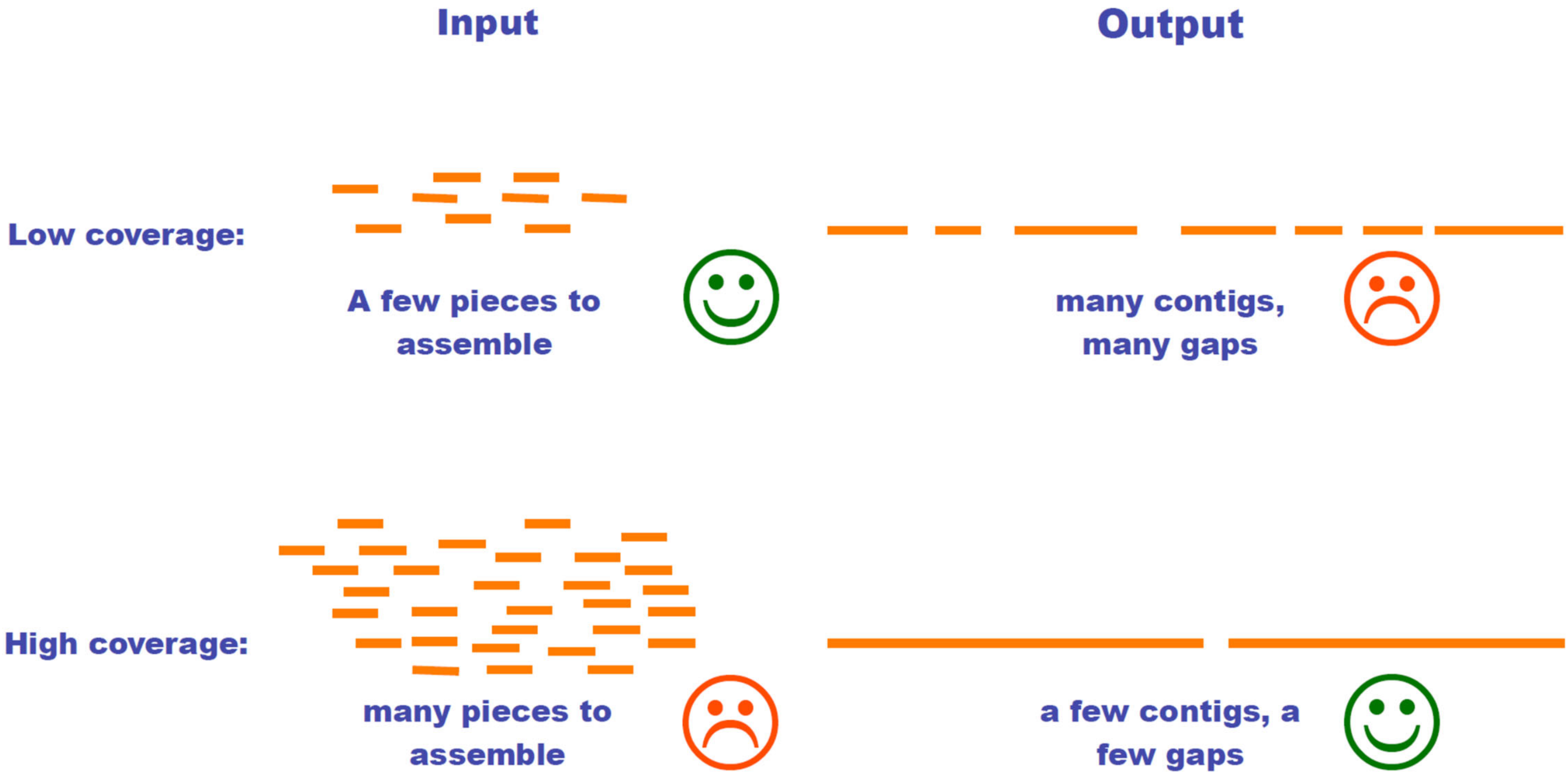
Promise of Genomics



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?



Input

Output

Low coverage:

A few pieces to assemble

many contigs, many gaps

High coverage:

many pieces to assemble

a few contigs, a few gaps

Genome Assembly

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

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

Input: GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
35bp

Copy by PCR:
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

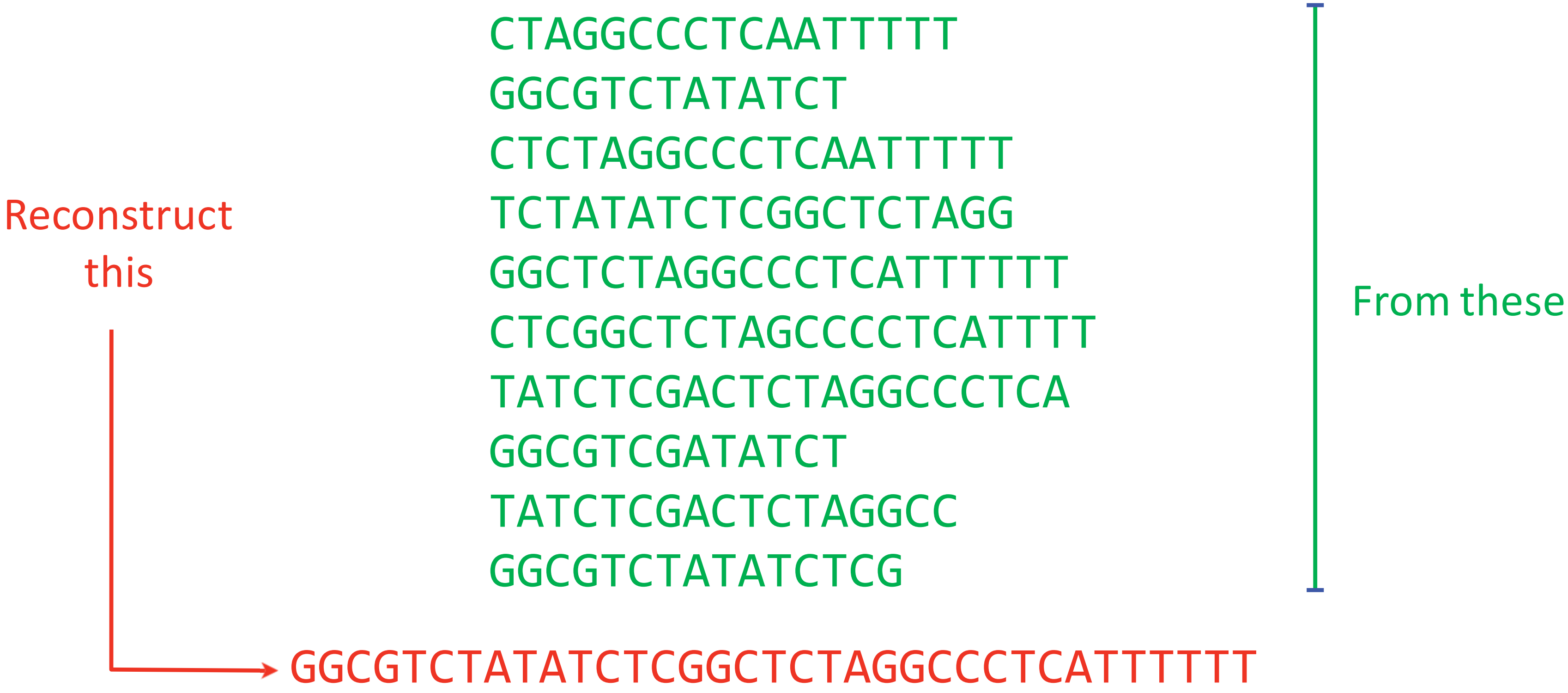
Fragment:
GGCGTCTA TATCTCGG CTCTAGGCCCTC ATTTTTT
GGC GTCTATAT CTCGGCTCTAGGCCCTCA TTTTTT
GGCGTC TATATCT CGGCTCTAGGCCCT CATTTTTT
GGCGTCTAT ATCTCGGCTCTAG GCCCTCA TTTTTT

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

Assembly

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

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



CTAGGCCCTCAATTTTT
GGCGTCTATATCT
CTCTAGGCCCTCAATTTTT
TCTATATCTCGGCTCTAGG
GGCTCTAGGCCCTCATTTTTT
CTCGGCTCTAGCCCCTCATTTT
TATCTCGACTCTAGGCCCTCA
GGCGTCGATATCT
TATCTCGACTCTAGGCC
GGCGTCTATATCTCG

From these

Reconstruct this

GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Courtesy of Ben Langmead. Used with permission.

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

Assembly

Overlaps between short reads help to put them together

```
                  CTAGGCCCTCAATTTTT
                CTCTAGGCCCTCAATTTTT
              GGCTCTAGGCCCTCATTTTTT
           CTCGGCTCTAGCCCTCATTTT
        TATCTCGACTCTAGGCCCTCA              177 nucleotides
        TATCTCGACTCTAGGCC
    TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCGATATCT
GGCGTCTATATCT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT     35 nucleotides
```

Courtesy of Ben Langmead. Used with permission.

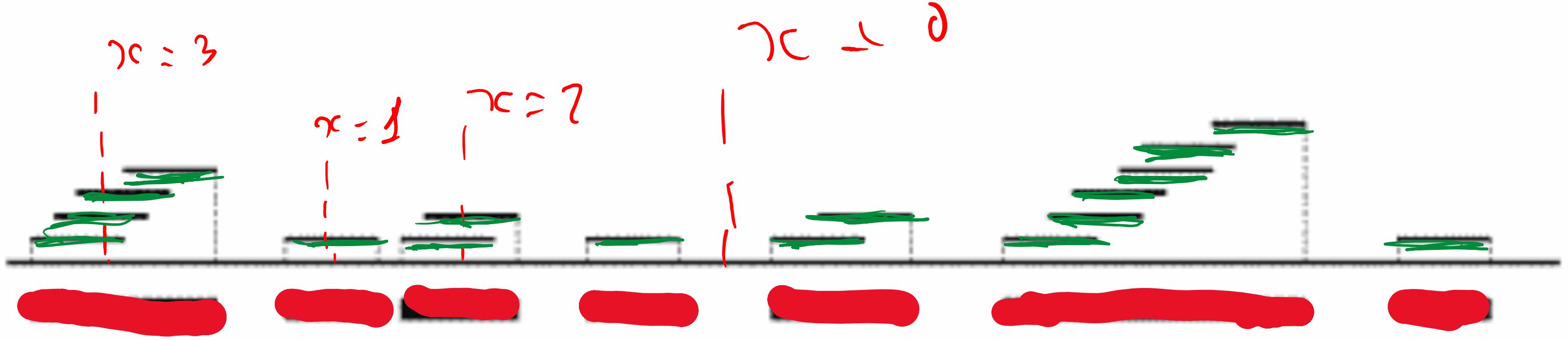
http://www.langmead-lab.org/teaching-materials/

Where is the Poisson?

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

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

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

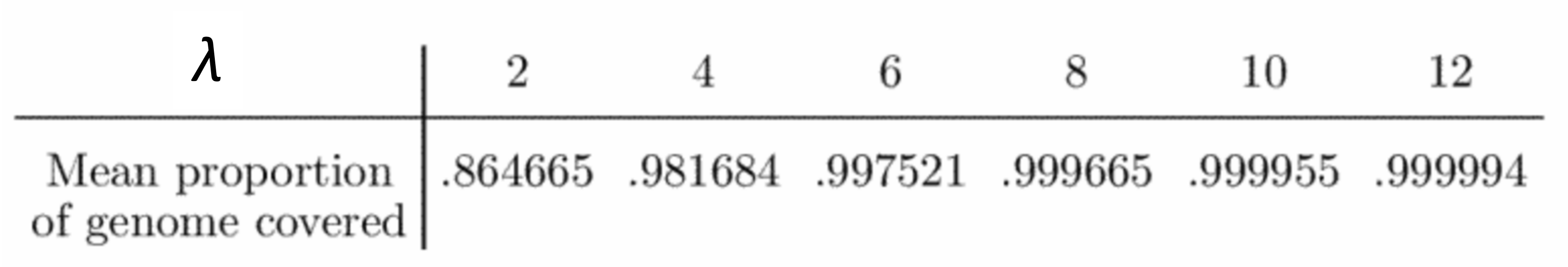


Ewens, Grant, Chapter 5.1

What fraction of the genome is missing?

What fraction of genome is covered?

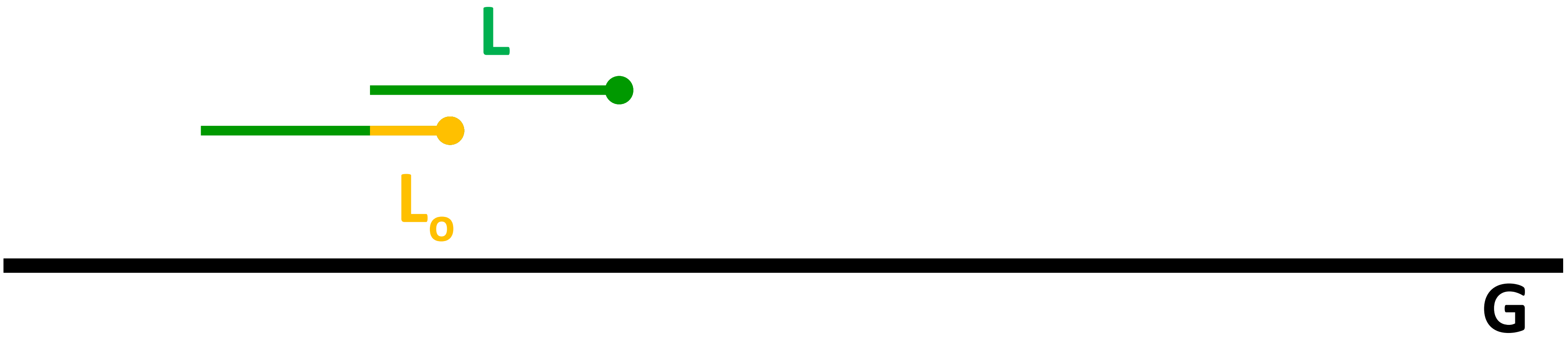
- Coverage: $\lambda=NL/G$,
  X – random variable equal to the number of times a given site is covered by short reads.
  Poisson: $P(X=x)=\lambda^{x}\exp(-\lambda)/x!$
  $P(X=0)=\exp(-\lambda)$, $P(X>0)=1-\exp(-\lambda)$
- *Total length covered:* $G*[1-\exp(-\lambda)]$



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

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

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



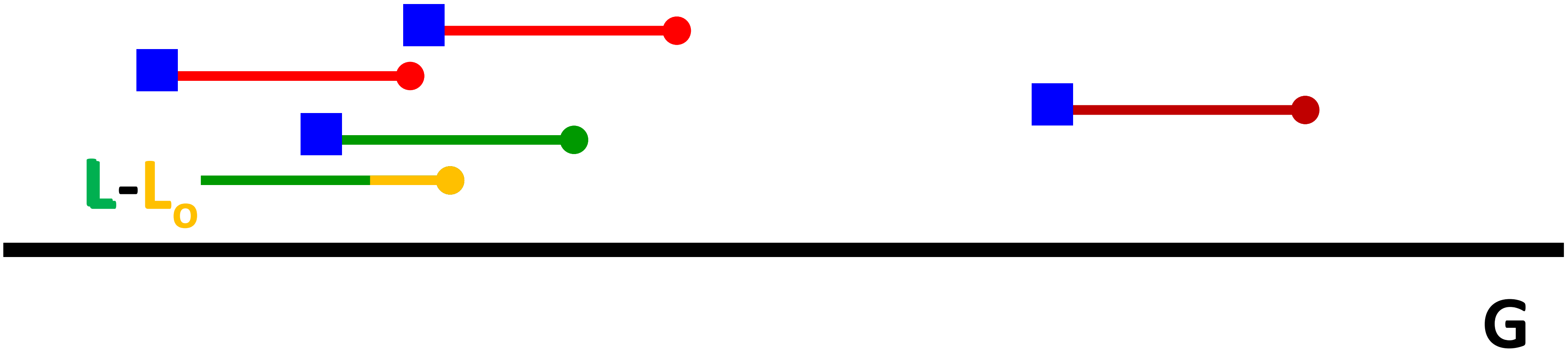
If DNA was a random chain with $p_A = p_C = p_G = p_T = 1/4$

L_o~16-20 would be enough

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

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

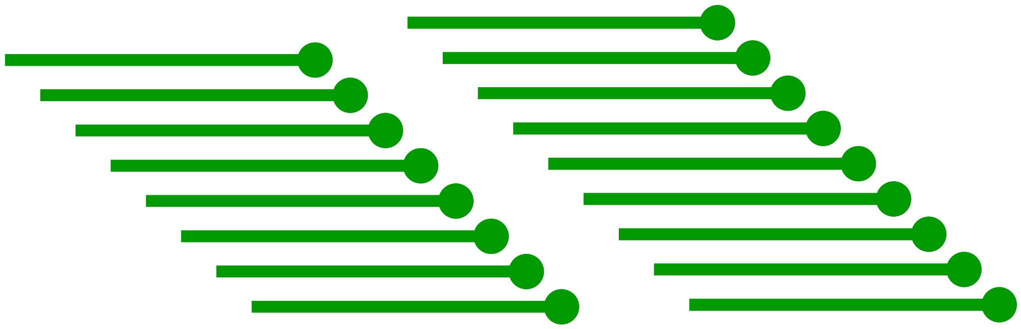
How many contigs?



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

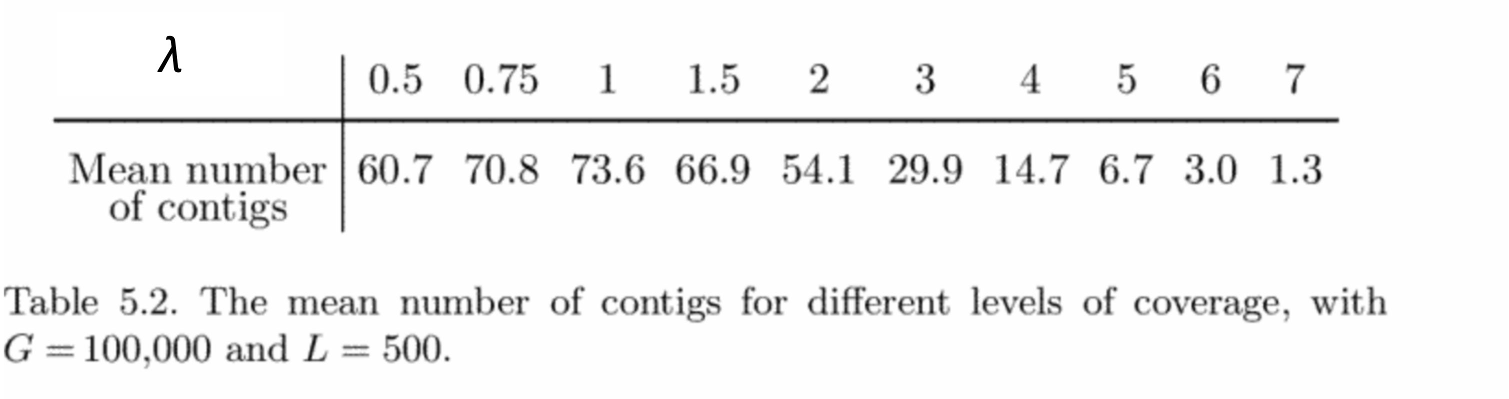
$$\text{P(short read cannot be extended by any short reads)}=e^{-pN}\approx Ne^{-\lambda}$$

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



How many contigs?

- A given short read is the right end of a contig if and only if no left ends of other short reads fall within 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 $N_{contigs}=Ne^{-\lambda}$:



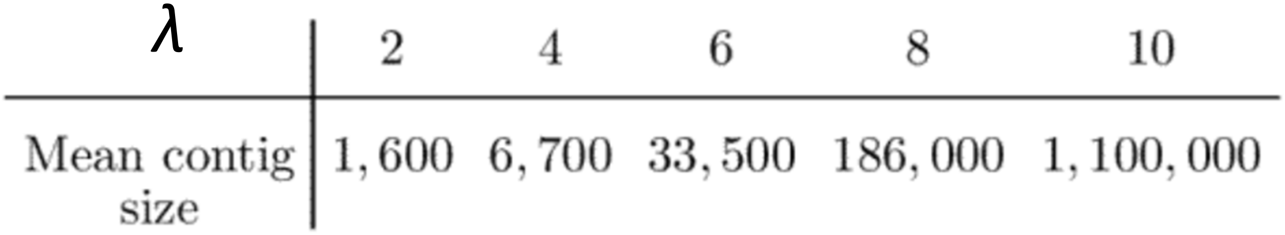
λ	0.5	0.75	1	1.5	2	3	4	5	6	7
Mean number of contigs	60.7	70.8	73.6	66.9	54.1	29.9	14.7	6.7	3.0	1.3

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

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 = $<L>=\sum_i L_i/N_{contigs}=G_{covered}/N_{contigs}=$

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



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: