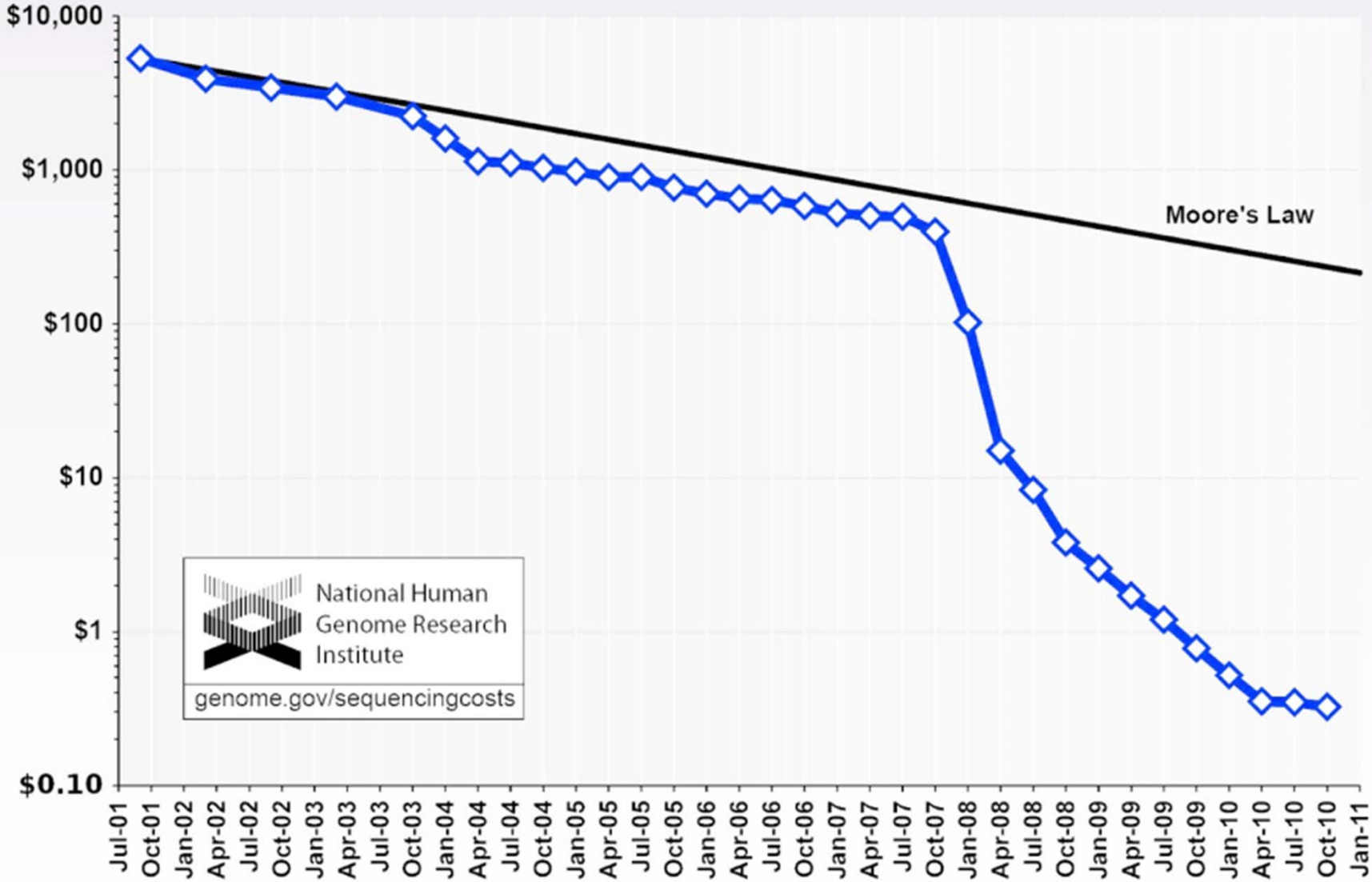
Poisson Distribution in Genome Assembly

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.

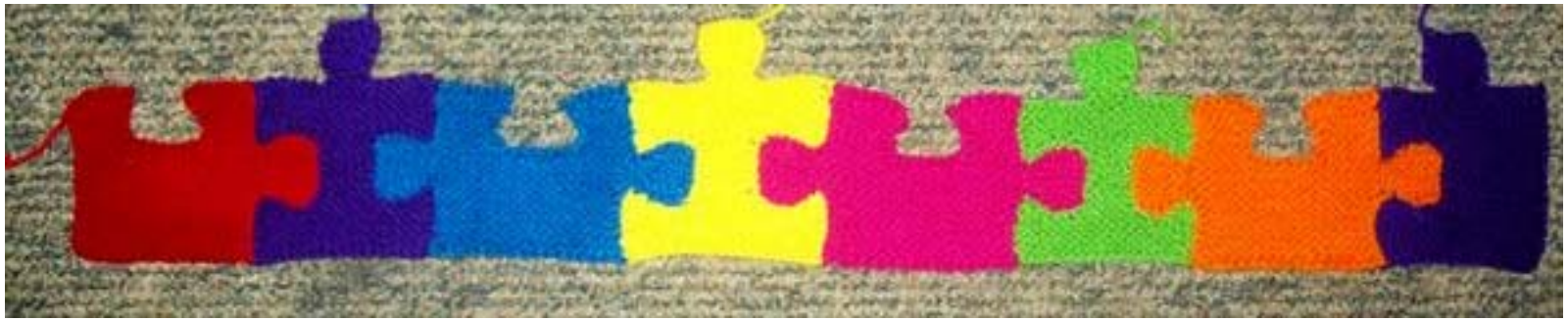
Technology	Read length (bp)	Reads per run	Time per run	Cost per megabase (US$)	Error (%)	Accuracy (%)
Roche 454	700	1 million	1 day	10	0.1	99.90
Illumina	50–250	<3 billion	1–10 days	~0.10	2	98
SOLiD	50	~1.4 billion	7–14 days	0.13	0.1	99.90
Ion Torrent	200	<5 million	2 hours	1	2	98
Pacific Biosciences	2900	<75,000	<2 hours	2	1	99
Sanger	400–900	N/A	<3 hours	2400	0.1	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.

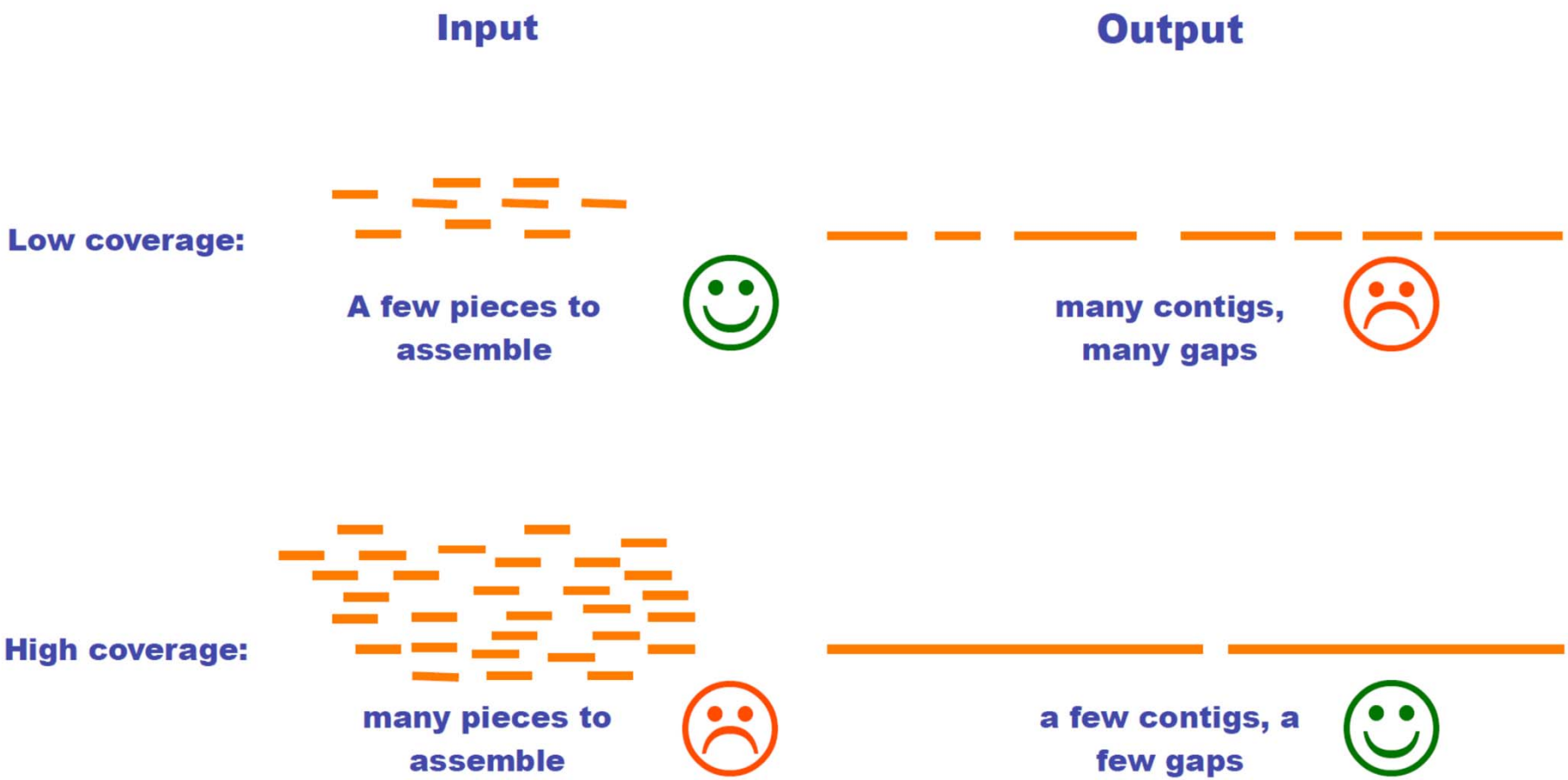


Promise of Genomics

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?



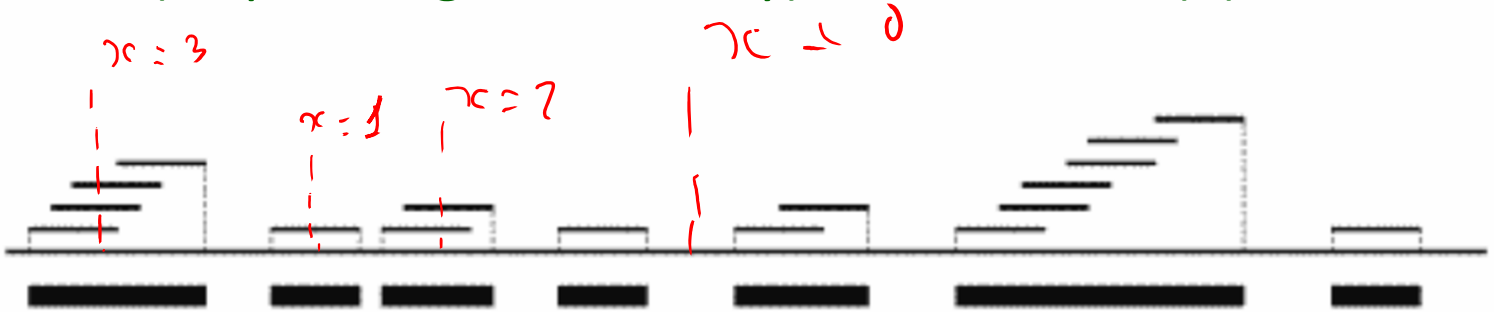
Where is the Poisson?

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

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

Ewens, Grant, Chapter 5.1

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).



What fraction of genome is covered?

- Coverage: *λ=NL/G,*
  X – r.v. equal to the number of times a given site is covered
  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 many contigs?

- Probability that a given short read is the right end of a contig = no left ends of other reads fall within it.
- Left ends of each of $N-1 \approx N$ other reads has Prob: $p=(L-1)/G \approx L/G$ to fall within given read. Expected number of short reads extending a given one is $N*p=N*L/G= \lambda$ Probability that none will extend it = $exp(-\lambda)$:
- *Number of contigs:* $N_{contigs}=Ne^{-\lambda}$:

a	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 spans about $250x10^6$ bp
- Illumina generates short reads 100 bp long
- How many short reads are needed to completely assemble the 1^{st} chromosome?

The answer is

$N=44x10^6$ short (100bp) reads

or $\lambda = 17.6$ fold redundant coverage.

At 0.1$/Mb that means that the reads for de novo full assembly of human genome would cost

$(3 x10^9 \times 17.6 / 10^6) \times 0.1\$$

$=\$5300$ /genome

In reality is cheaper as we don't need de novo assembly

What spoils these estimates?

- Try searching human genome for TTAGGGTTAGGGTTAGGG (18 bases) and you will find 100s of matches
- How many one expects by pure chance?
  $2 \cdot 3x10^{9} \cdot 4^{-18}=0.08$ << 1 match
- Genome has lots of repeats. DNA repeats make assembly difficult
- Side remark. If it was not for repeats 17 bases would be enough to specify unique position in the human genome.
  $\log(2 \cdot 3x10^{9})/\log(4)=16.2412$
- That is why microRNAs recognizing ~22 bases don't have a lot of off-target cleavage

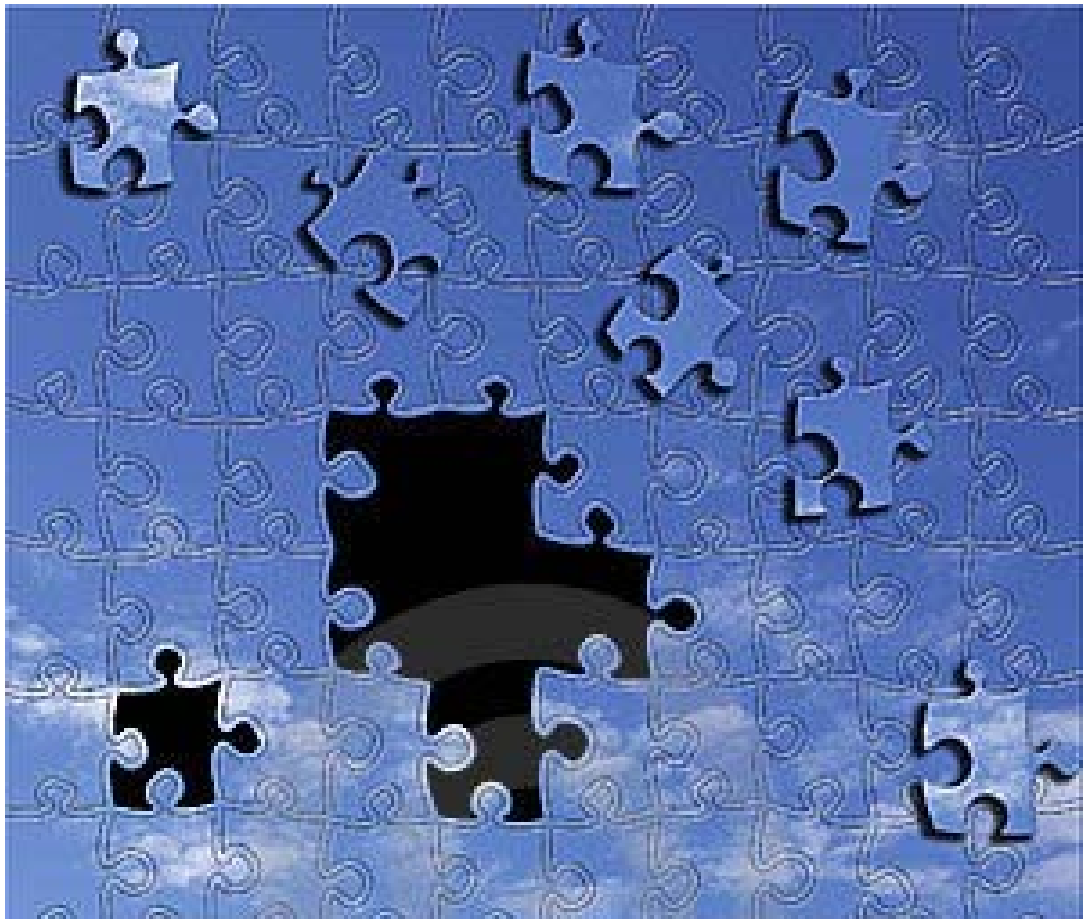
Stuttering human genome

>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.

Repeats are like sky puzzle pieces



How many repeats are in eukaryotic genomes?

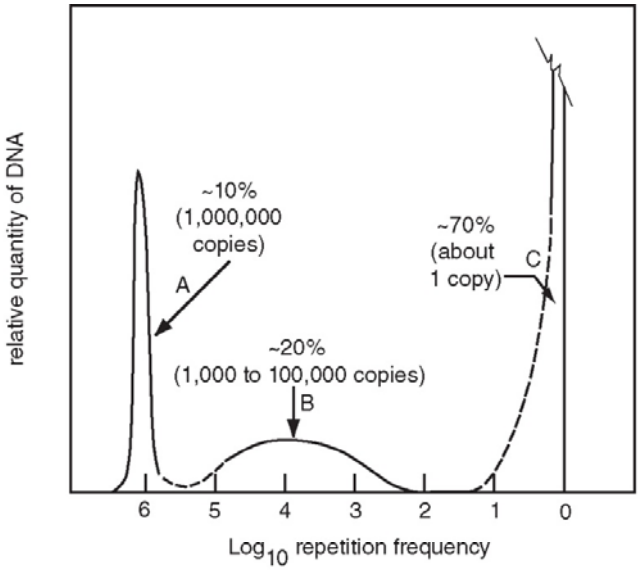


FIGURE 8.6 The complexity of genomic DNA can be estimated by denaturing then renaturing DNA. This figure (redrawn from Britten and Kohne, 1968) depicts the relative quantity of mouse genomic DNA (y axis) versus the logarithm of the frequency with which the DNA is repeated. The data are derived from a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

I showed my masterpiece to the grown-ups and asked them if my drawing frightened them.

The Little Prince, Antoine de Saint-Exupéry, 1943

Repetitive DNA and next-generation sequencing: computational challenges and solutions
Todd J. Treangen & Steven L. Salzberg
Nature Reviews Genetics **13**, 36-46 (January 2012)
doi:10.1038/nrg3117

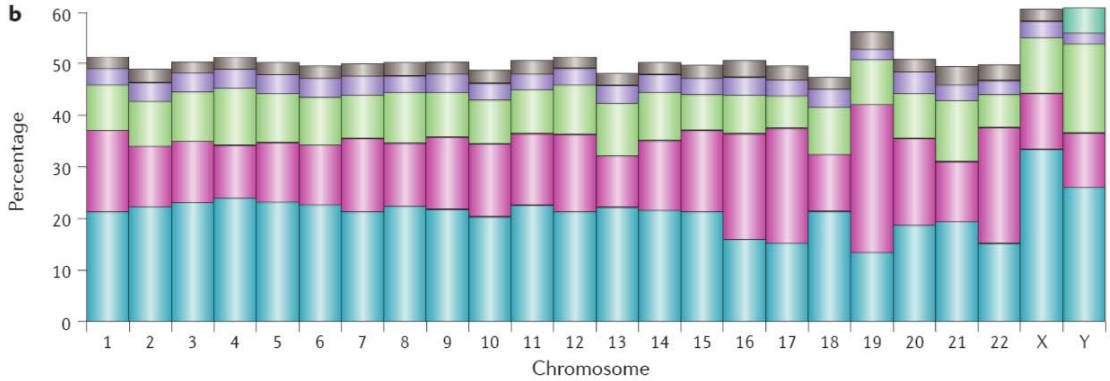
Box 1 | **Repetitive DNA in the human genome**

Approximately 50% of the human genome is comprised of repeats. The table in panel **a** shows various named classes of repeat in the human genome, along with their pattern of occurrence (shown as 'repeat type' in the table; this is taken from the RepeatMasker annotation). The number of repeats for each class found in the human genome, along with the percentage of the genome that is covered by the repeat class (Cvg) and the approximate upper and lower bounds on the repeat length (bp). The graph in panel **b** shows the percentage of each chromosome, based on release hg19 of the genome, covered by repetitive DNA as reported by RepeatMasker. The colours of the graph in panel **b** correspond to the colours of the repeat class in the table in panel **a**. Microsatellites constitute a class of repetitive DNA comprising tandem repeats that are 2–10 bp in length, whereas minisatellites are 10–60 bp in length, and satellites are up to 100 bp in length and are often associated with centromeric or pericentromeric regions of the genome. DNA transposons are full-length autonomous elements that encode a protein, transposase, by which an element can be removed from one position and inserted at another. Transposons typically have short inverted repeats at each end. Long terminal repeat (LTR) elements (which are often referred to as retrovirus-like elements) are characterized by the LTRs (200–5000 bp) that are harboured at each end of the retrotransposon. LINE, long interspersed nuclear element; rDNA, ribosomal DNA; SINE, short interspersed nuclear element.

a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000

b



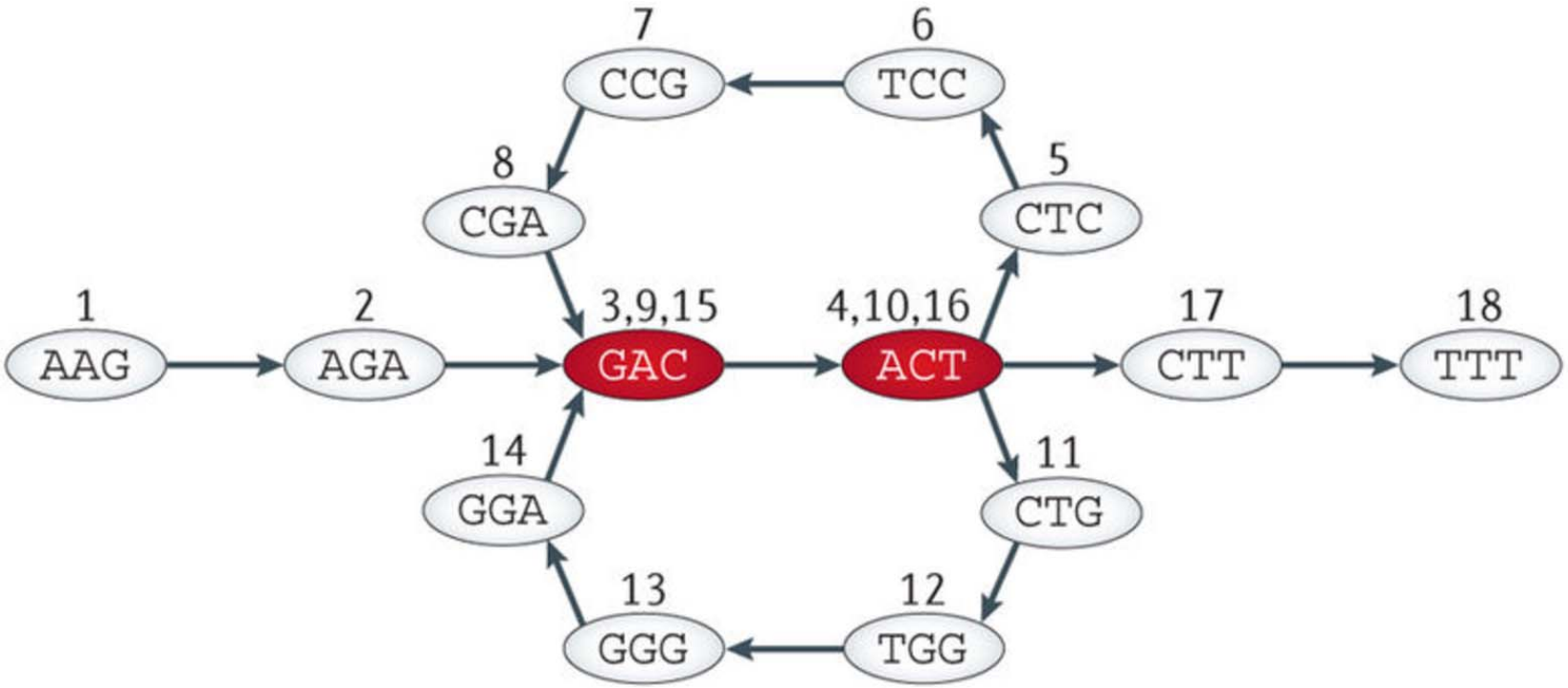
Interspersed repeats
Identical or nearly identical DNA sequences that are separated by hundreds, thousands or even millions of nucleotides in the source genome. Repeats can be spread out through the genome by mechanisms such as transposition.

Tandem repeats
DNA repeats (≥2bp in length) that are adjacent to each other and can involve as few as two copies or many thousands of copies. Centromeres and telomeres are largely comprised of tandem repeats.

Short interspersed nuclear elements
(SINEs). Repetitive DNA elements that are typically 100–300 bp in length and spread throughout the genome (such as *Alu* repeats).

Long interspersed nuclear elements
(LINEs). Repetitive DNA elements that are typically >300 bp in length and spread throughout the genome (such as L1 repeats).

De Bruijn graph



de Bruijn graph

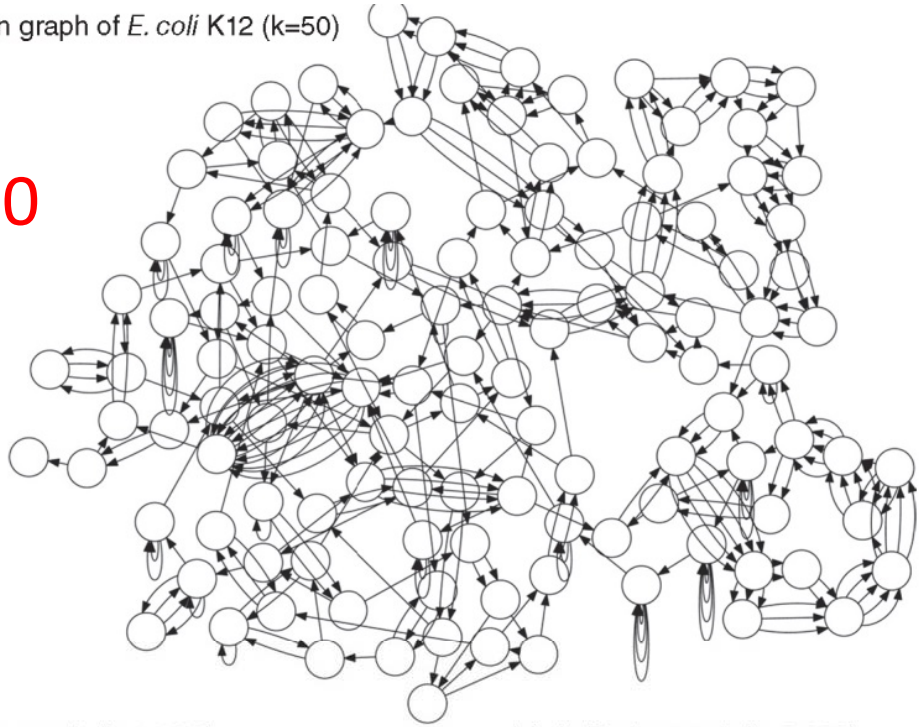
AAGACTCCGACTGGGACTTT

How to assemble genome with repeats?

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

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

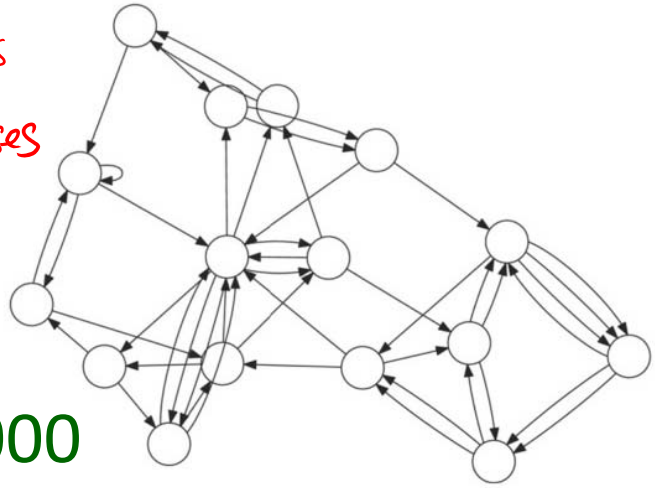
L=50



Technology	Read length (bp)
Roche 454	700
Illumina	50–250
SOLiD	50
Ion Torrent	200
Pacific Biosciences	2900
Sanger	400–900

~10 cents per 10^6 bases

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



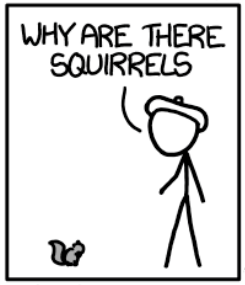
L=1000

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

L=5000

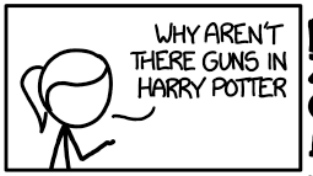
Credit: XKCD comics

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

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS 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 IS YKK ON ALL ZIPPERS
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 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
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY IS LIFE SO BORING
WHY ARE THERE FEMALE MR MIMES
WHY IS GPS FREE
WHY IS THERE HELL IF GOD FORGIVES
WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS