

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 deaths	of 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 $p = \frac{\lambda}{n}$

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

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

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

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

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

Poisson Mean & Variance

If X is a Poisson random variable, then:

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

Note: Variance = Mean

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

Matlab exercise: Poisson distribution

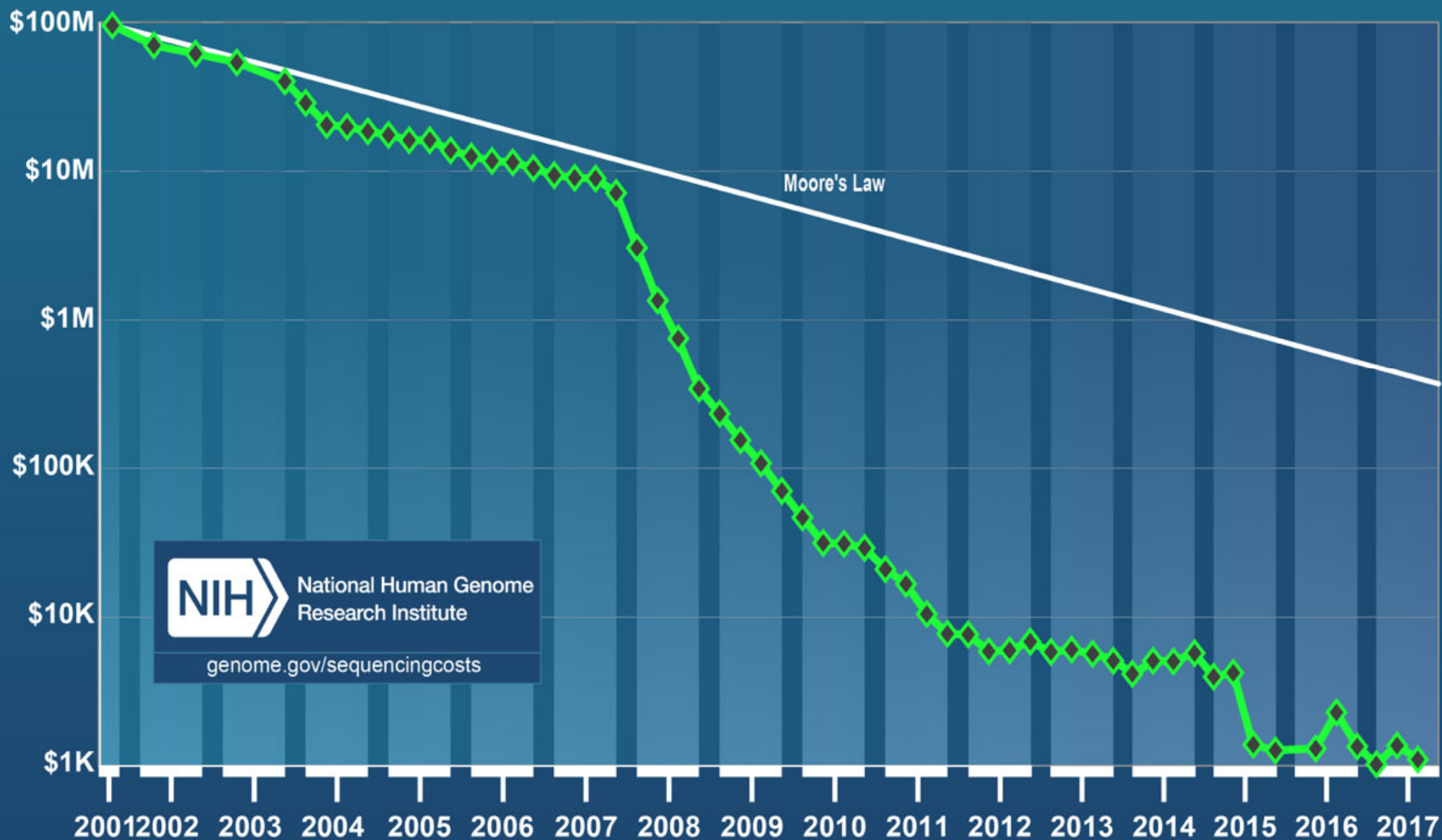
- Generate a **sample of size 100,000** for Poisson-distributed random variable X with $\lambda = 2$
- Plot the approximation to the **Probability Mass Function** based on this sample
- Calculate the mean and variance of this sample and compare it to **theoretical calculations**:
 $E[X] = \lambda$ and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

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-')**

Poisson Distribution in Genome Assembly

Cost per Genome



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 (Minlon)	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

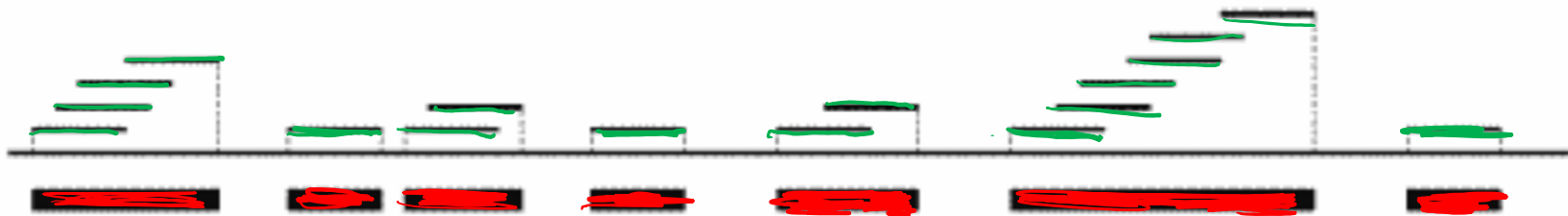
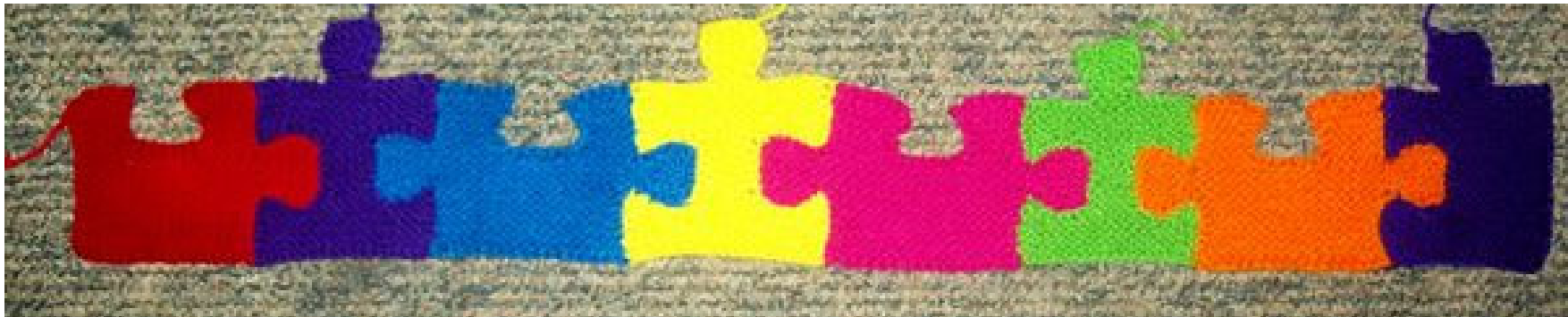


Figure 5.1.



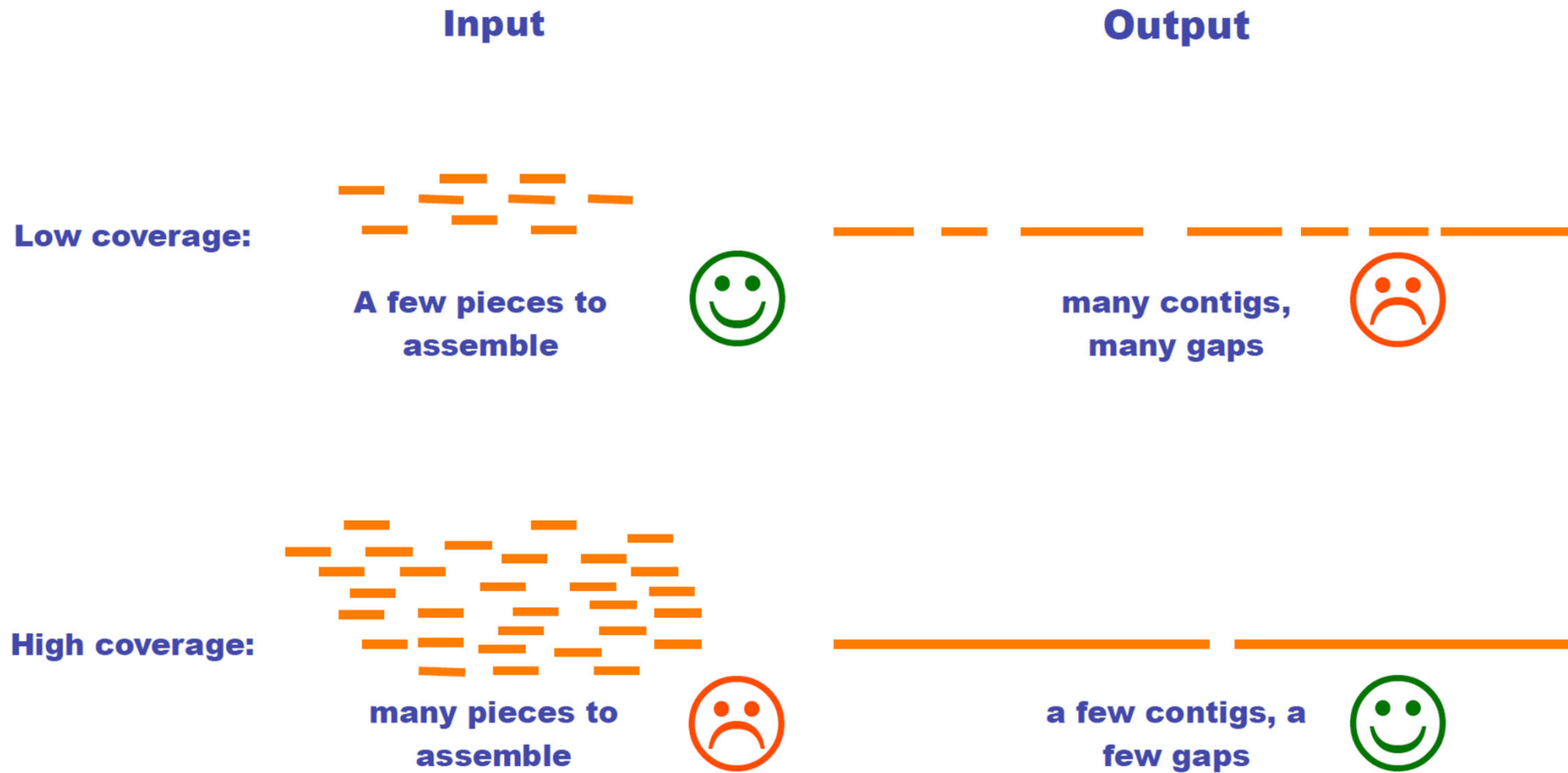
Promise of Genomics



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?



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 GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
by GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
PCR: GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

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

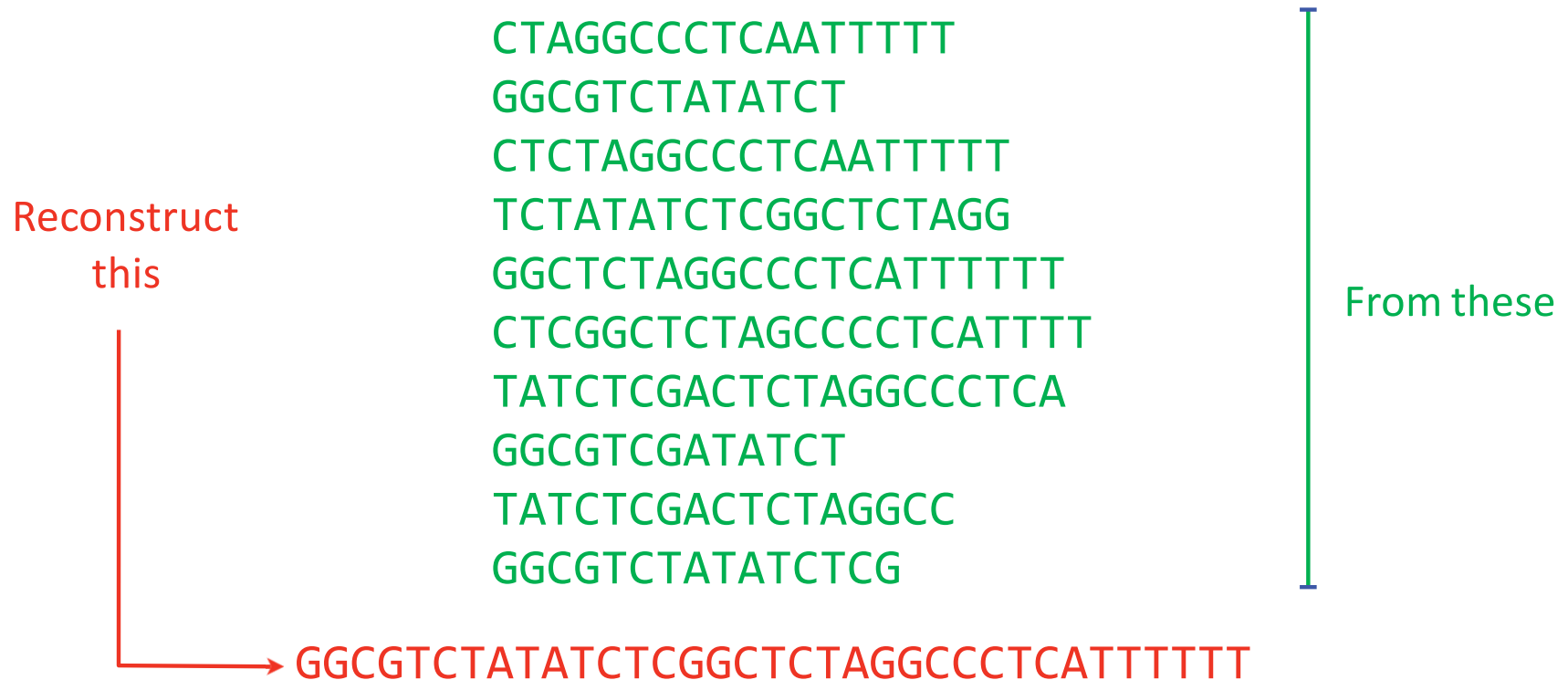
Courtesy of [Ben Langmead](http://www.langmead-lab.org/teaching-materials/). Used with permission.

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

Assembly

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

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



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

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

Assembly

Overlaps between short reads help to put them together

```
          CTAGGCCCTCAATTTTT
        CTCTAGGCCCTCAATTTTT
      GGCTCTAGGCCCTCATTTTT
    CTCGGCTCTAGCCCCTCATTTT
  TATCTCGACTCTAGGCCCTCA
TATCTCGACTCTAGGCC
TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCGATATCT
GGCGTCTATATCT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTT
```

177 nucleotides

35 nucleotides

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): $\lambda = NL/G$ is $O(1)$.



What fraction of the genome is missing?

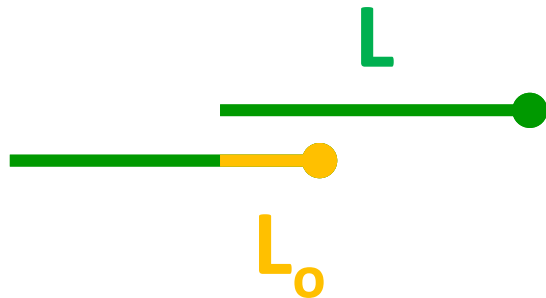
What fraction of genome is covered?

- 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?



G

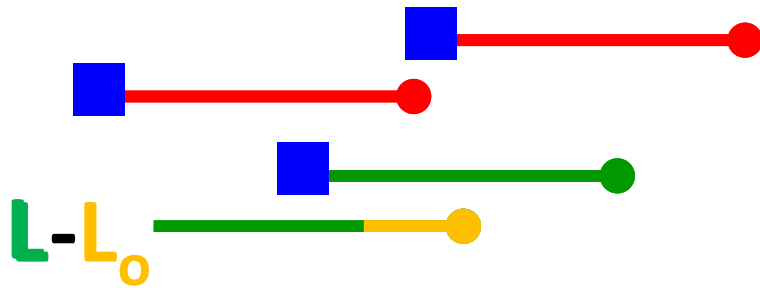
If DNA was a random chain with $p_A = p_C = p_G = p_T = 1/4$

$L_o \sim 16-20$ would be enough

$$2 \cdot G \cdot 4^{-L_o} = 2 \cdot 3 \times 10^9 \cdot 4^{-16} = 1.4$$

$$2 \cdot 3 \times 10^9 \cdot 4^{-20} = 0.0055 \ll 1$$

How many contigs?



G

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

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

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



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 =

$$\langle L \rangle = \sum_i L_i / N_{contigs} = G_{covered} / N_{contigs} =$$

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

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

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

Estimate

- Human genome is 3×10^9 bp long
- Chromosome 1 is about $G = 0.25 \times 10^9$ bp
- Illumina generates short reads $L = 100$ bp long
- What number of reads N are needed to completely assemble the 1st chromosome?
- The formula to use is: $1 = N_{\text{contigs}} = N e^{-\lambda} = N e^{-NL/G}$
- Answer: $N = 4.4 \times 10^7$ short (100bp) reads
Test: $4.4 \times 10^7 * \exp(-4.4 \times 10^7 * 100 / 0.25 \times 10^9) = 0.99997$
- What coverage redundancy λ will it be?
Answer: $\lambda = NL/G = 17.6$ coverage redundancy

How much would it cost to assemble the 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 $3 \times 10^9 \times 17.6 / 10^9 \times 10\$/\text{GBase} = \$ 530$
- **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 the 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  
ATAGTGTTAGGGTTAGGGTCAGGGTCCCGGTCCGGGTCCGGGTCCGGGTCCGGGTCCGGGTCCGGGTCA  
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT  
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG  
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA  
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT  
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG  
GTTAGGGTTAGGGTTAGGGTTAG
```

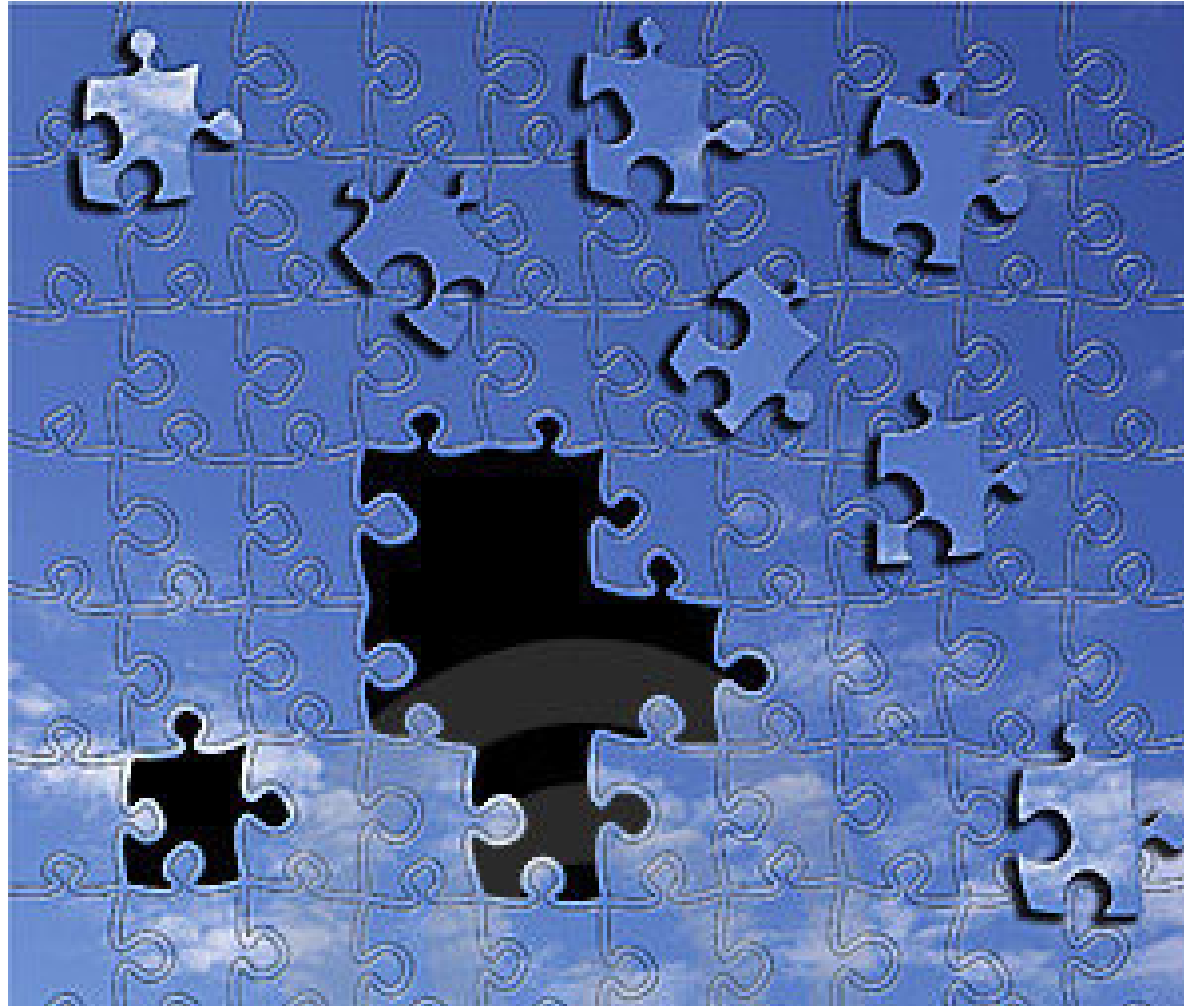
FIGURE 8.11 A BLASTN search of the human genome (all assemblies) database was performed at the NCBI website using **TTAGGGTTAGGGTTAGGG** as query (i.e., three TTAGGG repeats). There were matches to hundreds of genomic scaffolds. This figure shows an example (NT_024477.14) assigned to the **telomere of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

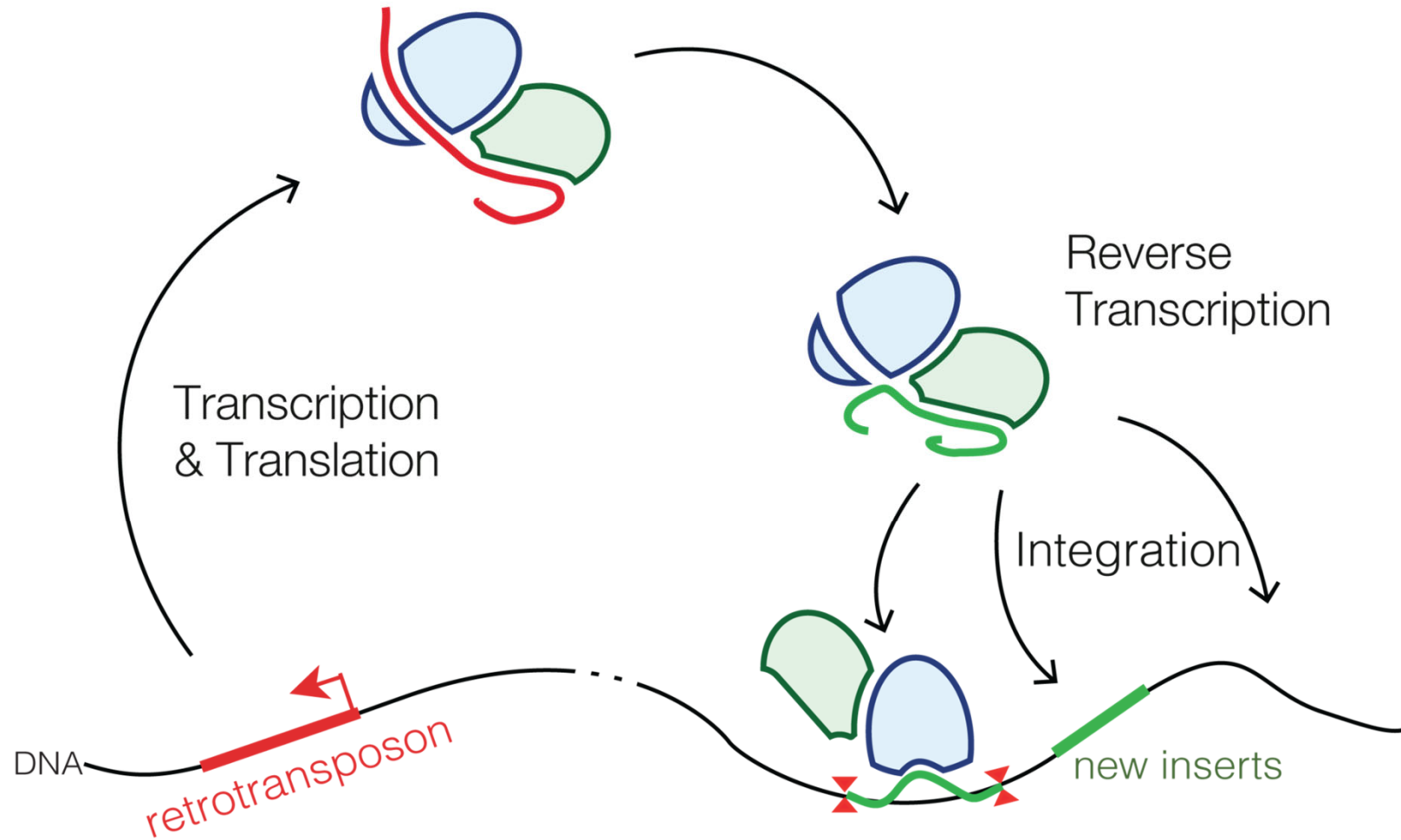
$$2 \cdot 3 \times 10^9 \cdot 4^{-18} = 0.08 \ll 1$$

DNA repeats make assembly difficult

Repeats are like sky puzzle pieces



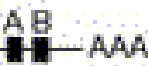
Formation of
Ribonucleoprotein complexes



For each of three classes of **parasitic (selfish) DNA** there is a **shorter version** parasitizing the **longer version**

Classes of interspersed repeat in the human genome

Long Interspersed Nuclear Elements

			Length	Copy number	Fraction of genome
LINEs	Autonomous		6–8 kb	850,000	21%
	Non-autonomous		100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Non-autonomous		1.5–3 kb		
DNA transposon fossils	Autonomous		2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

Adapted from slide by Ross Hardison, Penn State U.

How to assemble a real genome with repeats?

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



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

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

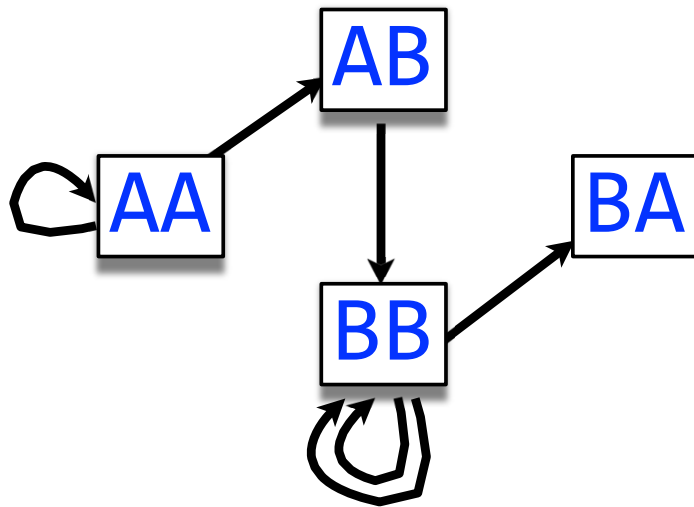
<http://www.langmead-lab.org/teaching-materials/>

De Bruijn graph

genome: **AAABBBBA**

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

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



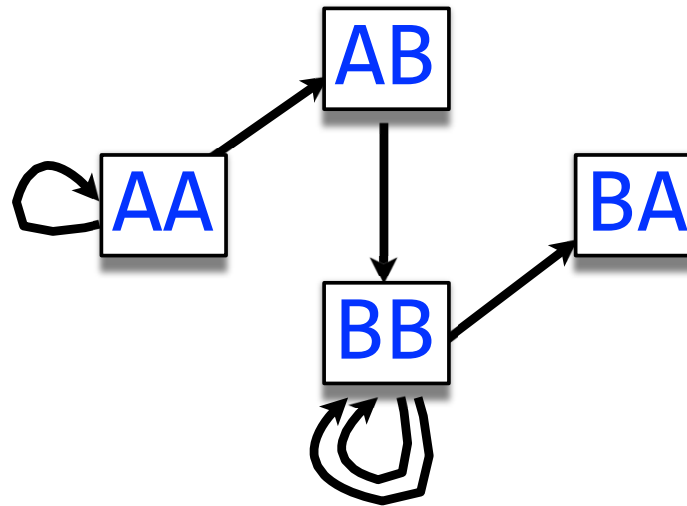
One edge per **every** k -mer

One node per **distinct** $k-1$ -mer

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

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

De Bruijn graph

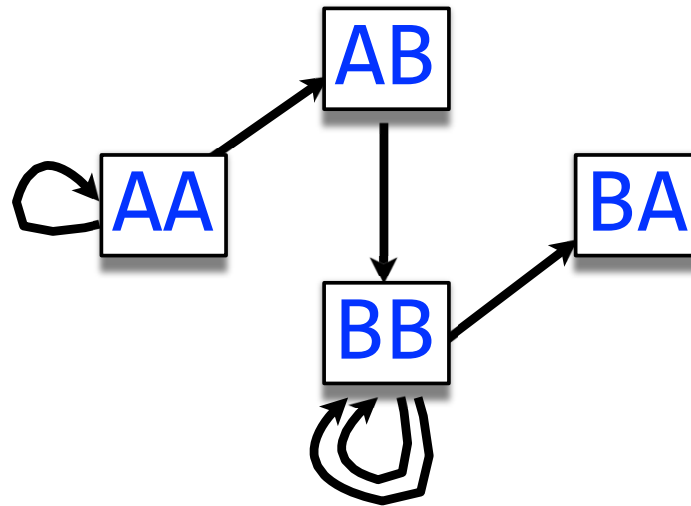


Walk crossing each edge exactly once gives a reconstruction of the genome

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

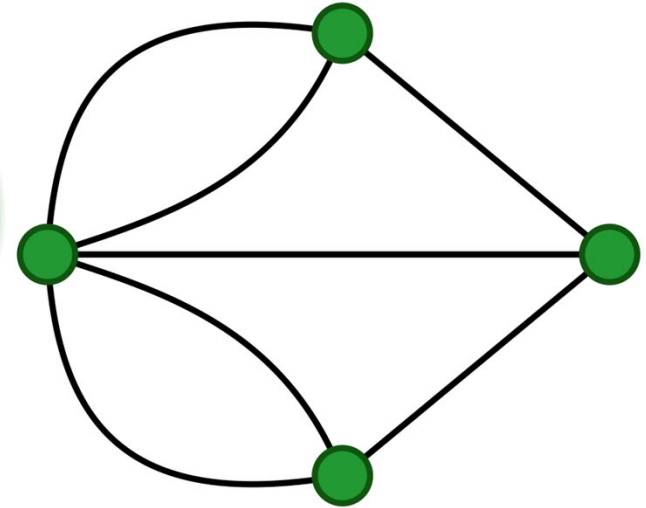
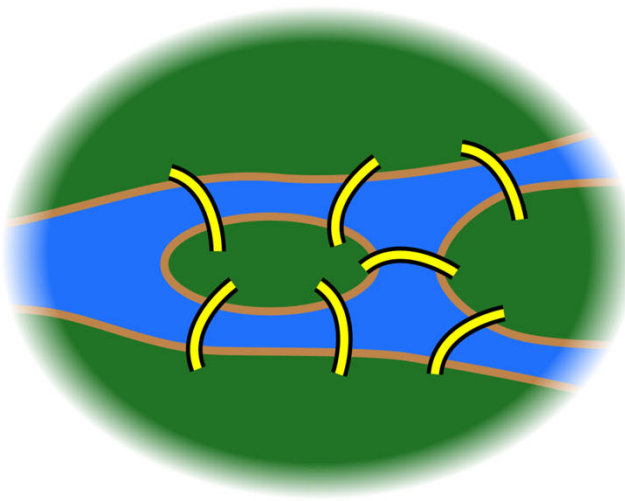
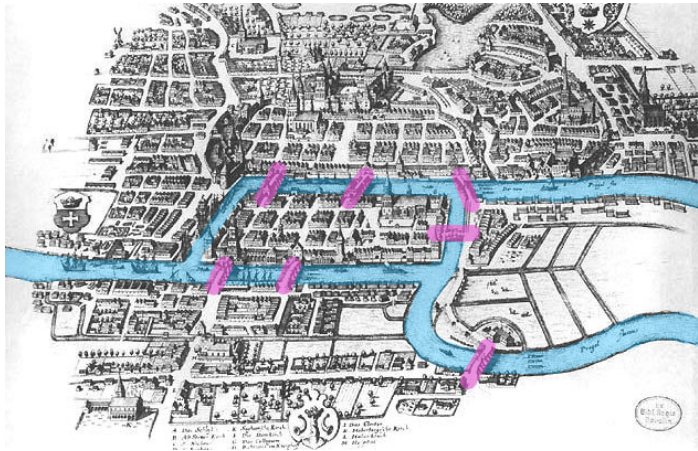
<http://www.langmead-lab.org/teaching-materials/>

Assembly = Eulerian walk on De Bruijn graph



AAABBBBA

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

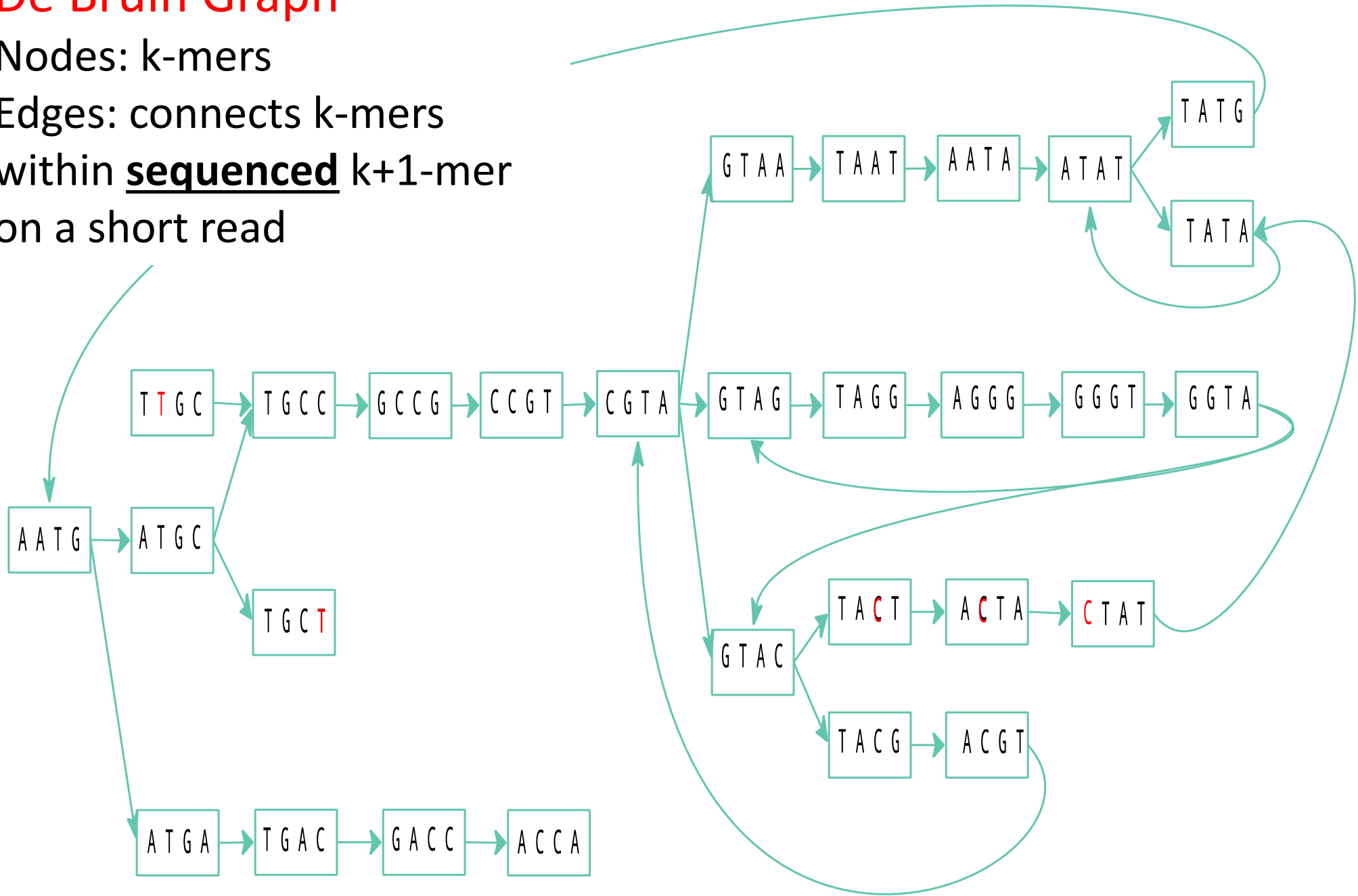


- The city of **Königsberg in Prussia** (now Kaliningrad, Russia) has
 - **two large islands**: Kneiphof and Lomse
 - **two mainland portions**: Altstadt and Vorstadt
 - Connected by **seven bridges**
- **Seven Bridges of Königsberg problem** formulated by **Leonhard Euler** in **1736**
 - Is it possible to make a **walk** (not necessarily a loop) walking **each bridge exactly once**?
 - **Euler** (why the base of natural logarithm is **e**): Basel → Berlin → St. Petersburg
- Euler **proved** that **such walk is impossible**
 - Graph has **three nodes** with **odd degrees**: 3,3,and 5
 - **Only two nodes with odd degrees are allowed**: start and finish of the walk
 - This laid the **foundations of graph theory** and prefigured the idea of **topology**

De Bruin Graph

Nodes: k-mers

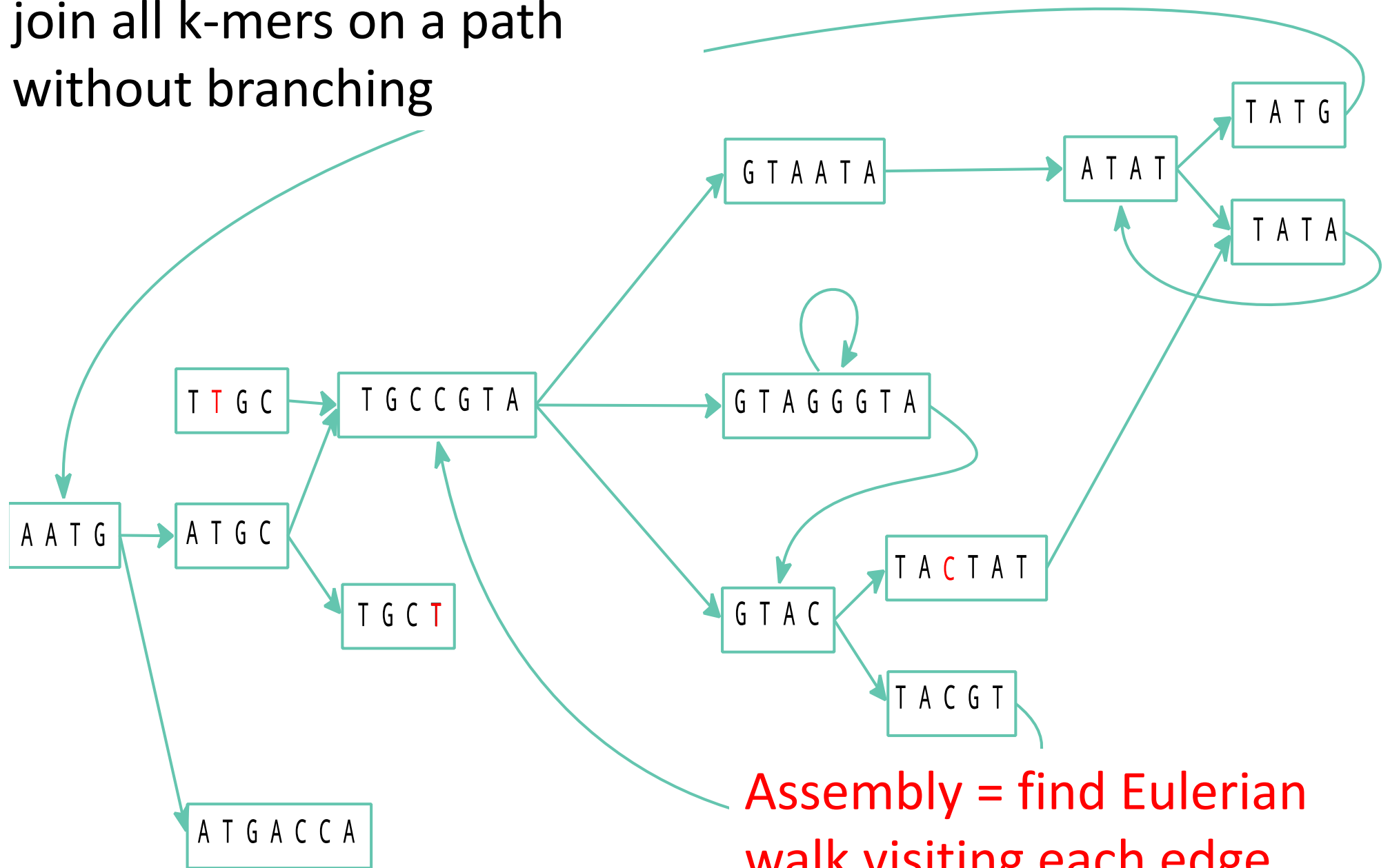
Edges: connects k-mers
within **sequenced** k+1-mer
on a short read



Slide by Sorin Istrail, Brown U.

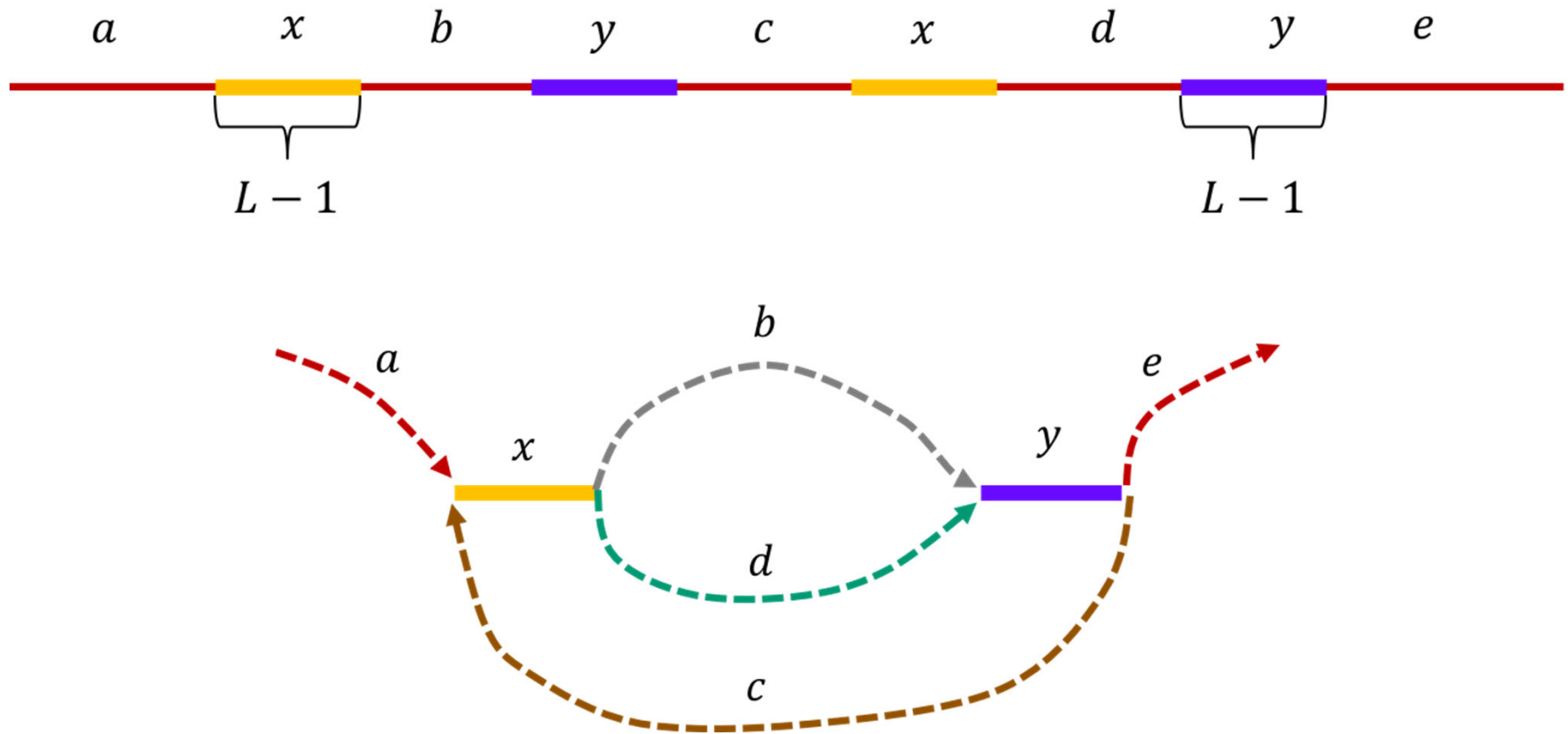
Simplified De Bruin Graph

join all k-mers on a path
without branching



Slide by Sorin Istrail, Brown U.

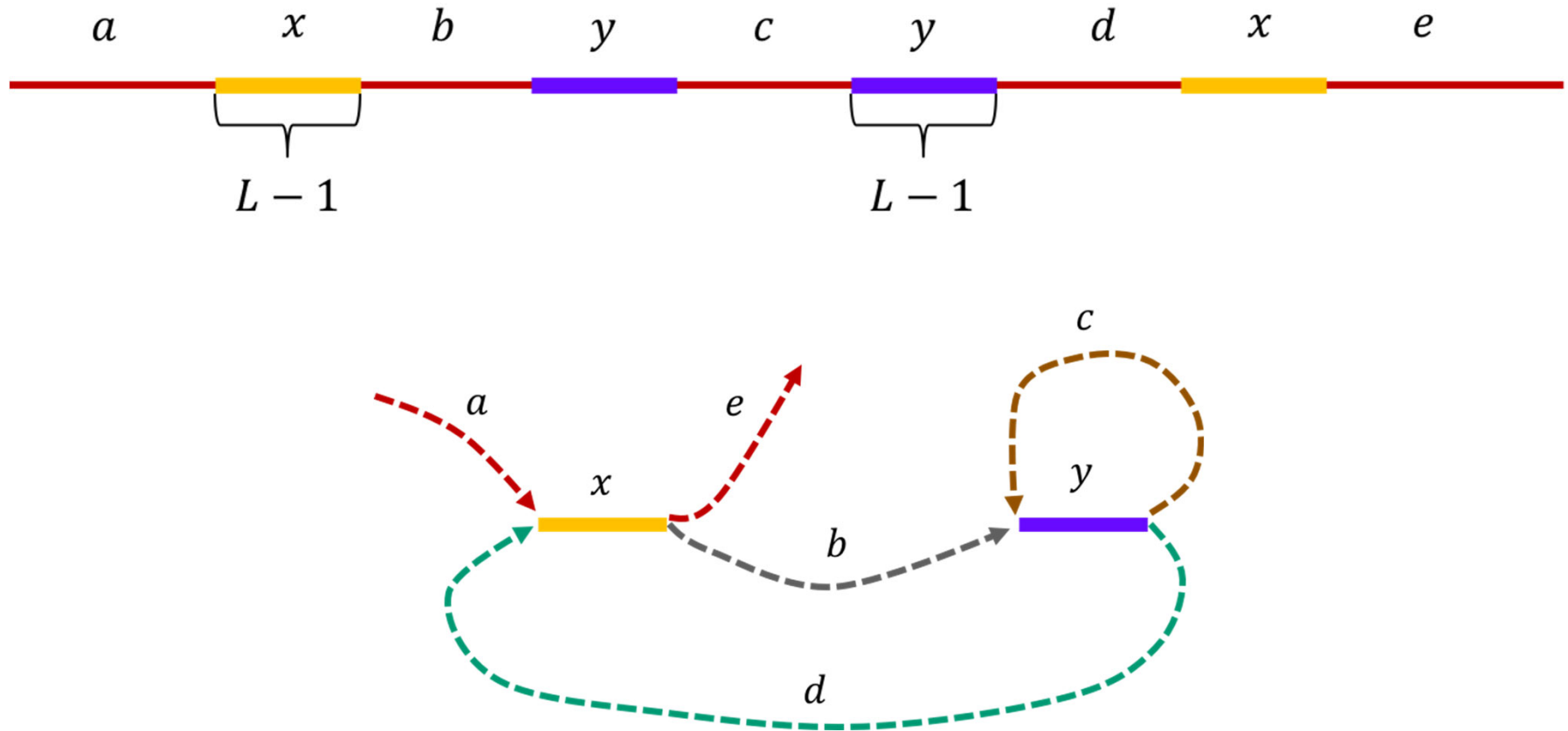
Why interleaved repeats are dangerous?



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

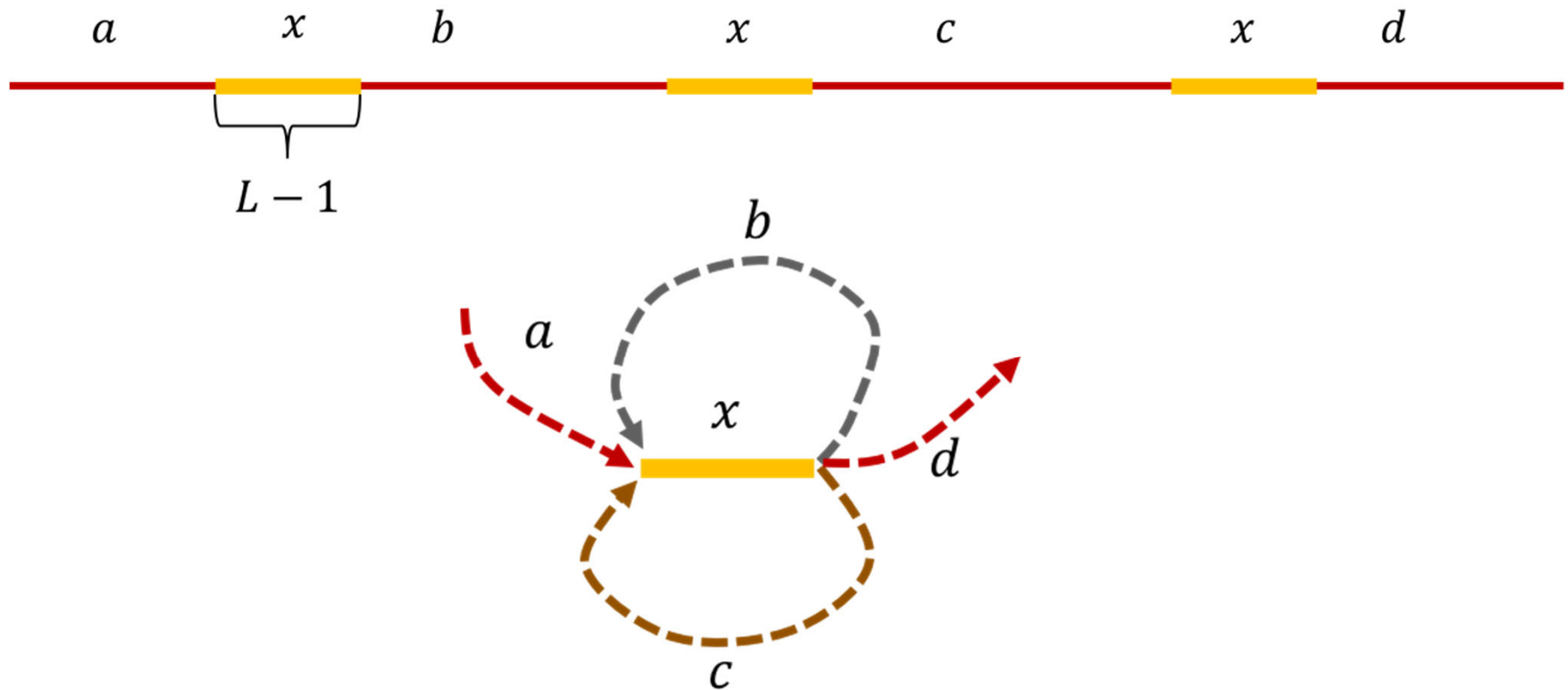
Images from the course EE 372: Data Science for High-Throughput Sequencing.
taught by David Tse at Stanford

Why non-interleaved repeats are safe?



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

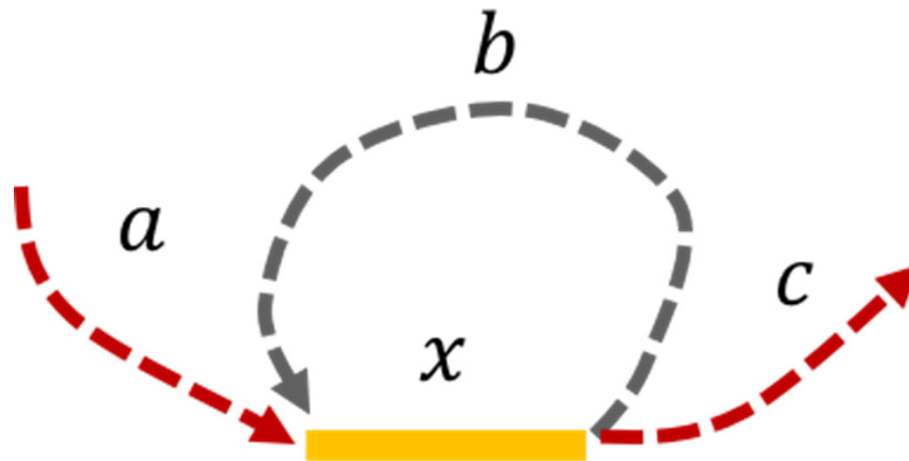
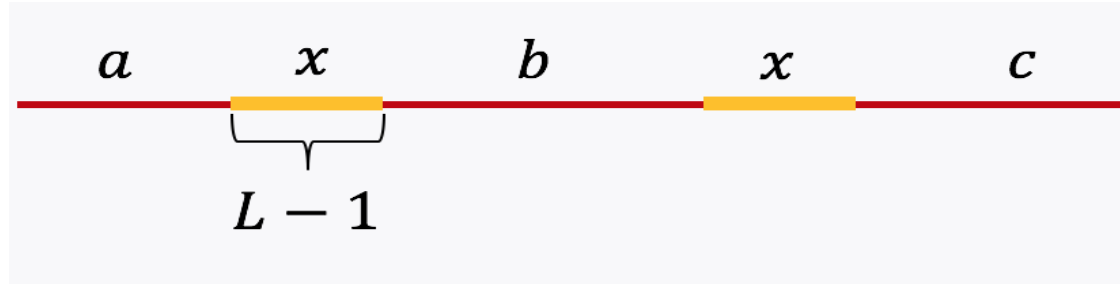
Why triple repeats are dangerous?



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

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: $a-x-b-x-c$

Pavel Pevzner's theorem

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



Pavel Pevzner
is the Ronald R. Taylor Chair and
Distinguished Professor of
Computer Science and Engineering
at University of California, San Diego

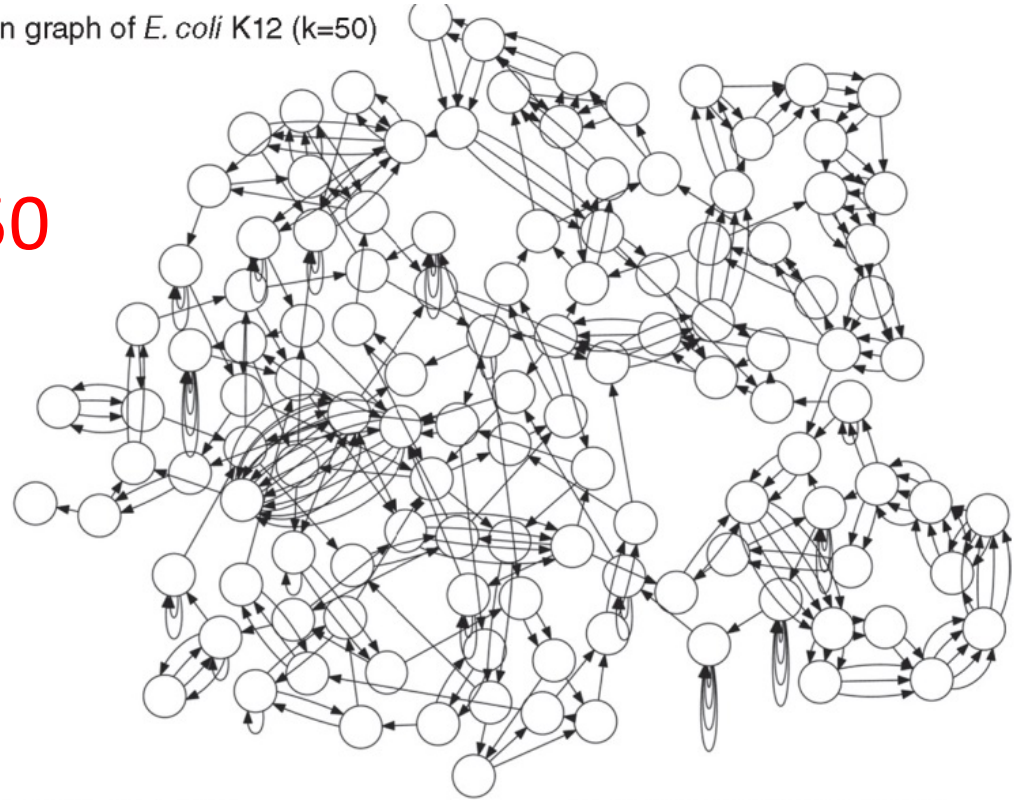
How to assemble a genome with repeats?

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

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

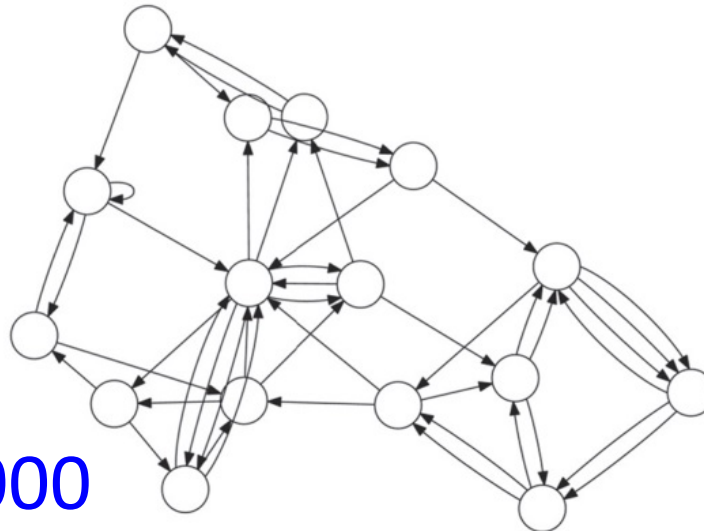
(a) de Bruijn graph of *E. coli* K12 ($k=50$)

$k=50$



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

$k=1000$



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

$k=5000$

