CSE 5370: Bioinformatics

Jacob M. Luber, Ph.D.

Lecture 5: Genome Sequencing

February 2nd, 2022

HW1 & Quiz 1

- Quiz1
  - Average 85
  - Standard Deviation 20
  - Median 83

- HW1
  - Average 91
  - Standard Deviation 23
  - Median 102.5

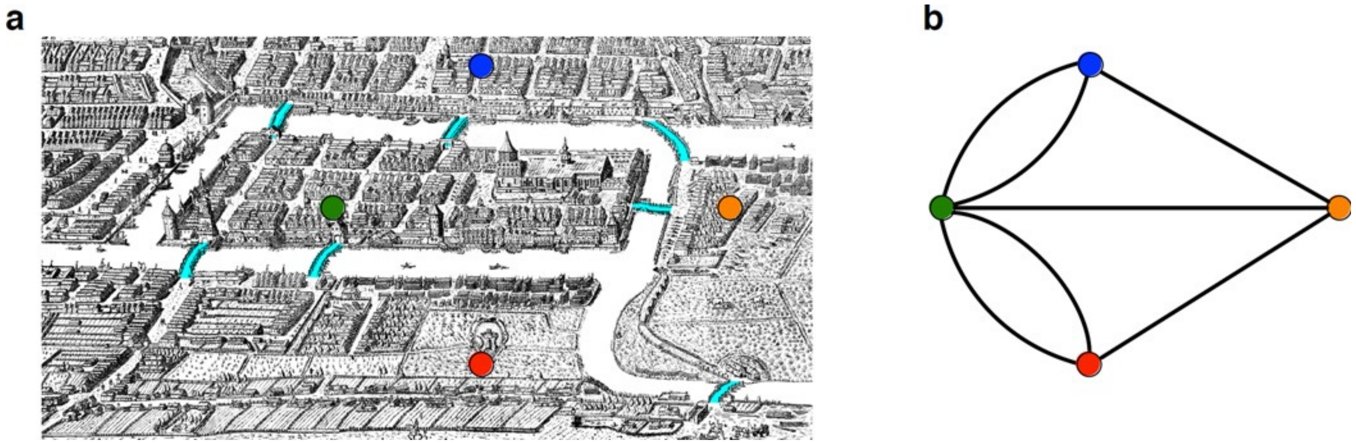
HW1 Thoughts

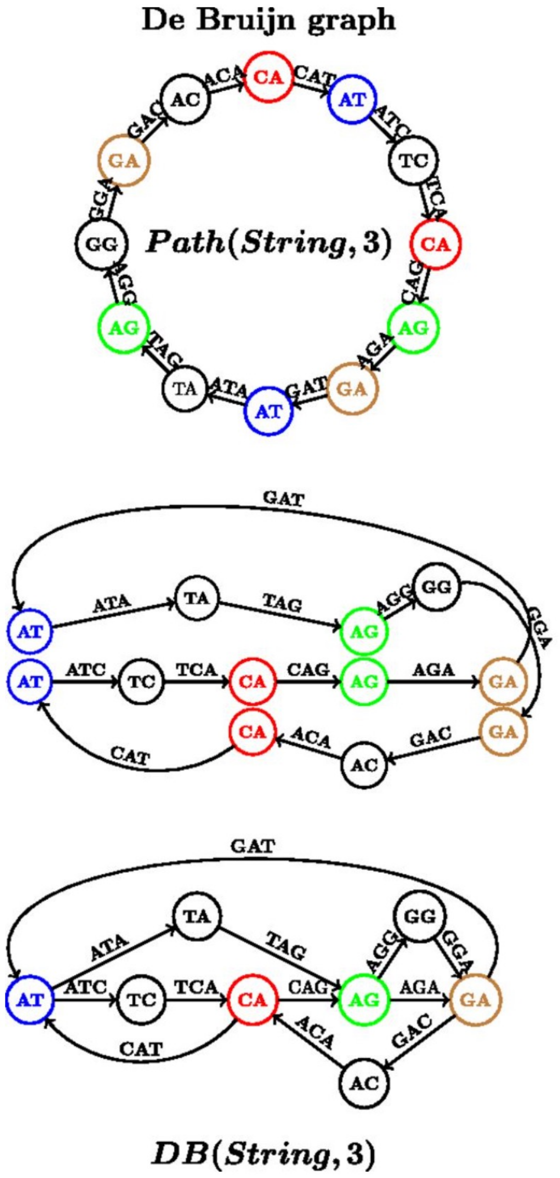
- Average was very high, but the distribution was very bi-modal
- If you received <70, please schedule an office hours appointment as coding will only get more challenging
    - Starting HW early and coming to office hours
- Assignment took an average of 6-7 hours; longer than designed but you all did well!
- HW2 is being adjusted so less time will be spent on installing packages
- Non-deterministic output of Megahit

HW2

- Will be released in next few days after adjustments are made
- First assignment that can be put in a github portfolio for job applications (all of the remaining assignments will be this way)
- We will write a simplified genome assembler:
    1. Write a brute force approach (team)
    2. Speed up brute force approach with graph algorithms (team)
    3. Compare with megahit (individual)
- Two weeks to complete

Graph Algorithms For Strings





-Bridges and Landmasses in Königsberg, Russia: Can we walk to each landmass crossing each bridge only once?

-Solved by mathematician Leonhard Euler in 1735: Eulerian Cycle

-landmass as nodes, bridges as edges

-Can be applied to genome assembly with overlaps

READS: ATCATG ATGCGC

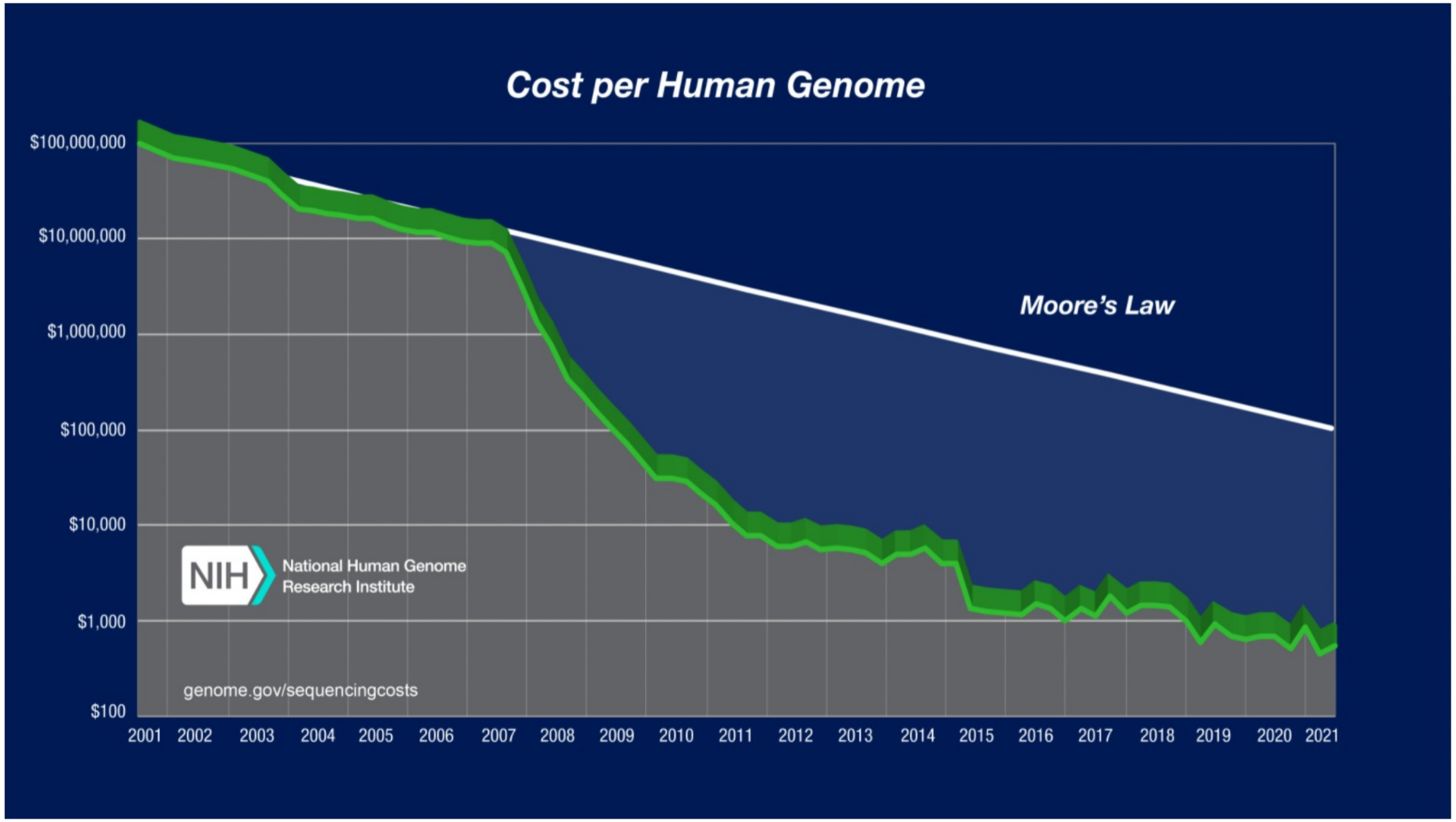
Assembly: ATCATG ATGCGC

-Genomes as *strings*

Lin et. al., Proceedings of the National Academy of Sciences Dec 2016, 113 (52) E8396-E8405; DOI: 10.1073/pnas.1604560113

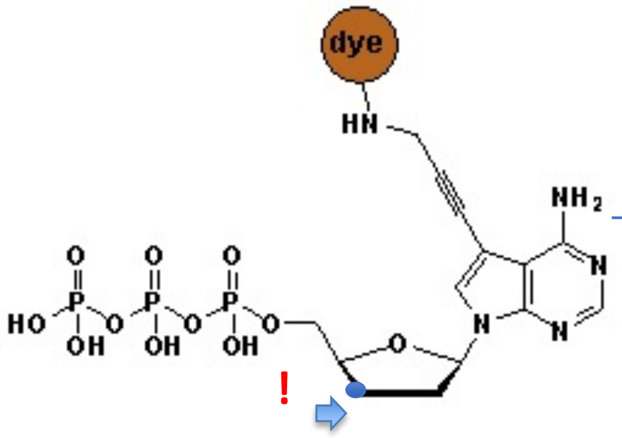
Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol*. 2011;29(11):987-991. Published 2011 Nov 8. doi:10.1038/nbt.2023

Why Study Bioinformatics?

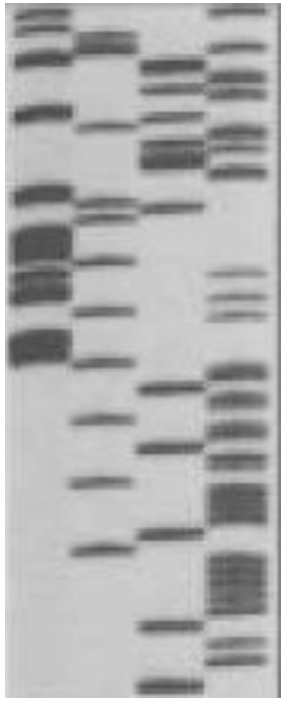


Some slides courtesy of SciLifeLab, Sweden

Sanger sequencing

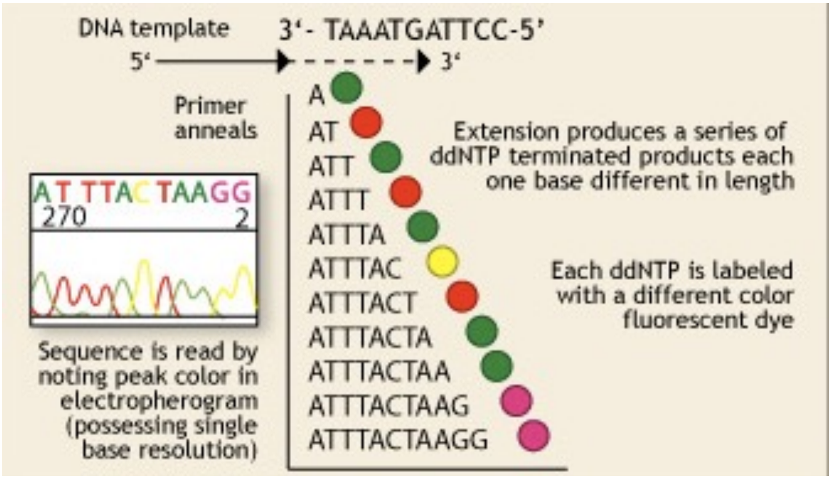


P^{32} labelled ddNTPs



Lack of OH-group at 3’ position of deoxyribose

Fluorescent dye terminators



Extension produces a series of ddNTP terminated products each one base different in length

Each ddNTP is labeled with a different color fluorescent dye

Sequence is read by noting peak color in electropherogram (possessing single base resolution)

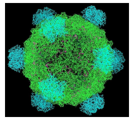
Max fragment length – 750 bp

Sequencing genomes using Sanger's method

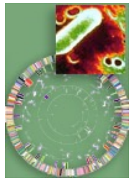
- Extract & purify genomic DNA
- Fragmentation
- Make a clone library
- Sequence clones
- Align sequencies (-> contigs -> scaffolds)
- Close the gaps

- Cost/Mb=1000 $, and it takes TIME

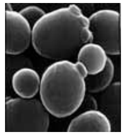
At the very beginning of genome sequencing era…



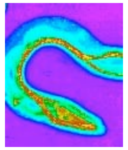
- First genome: virus ϕ X 174 - 5 368 bp (1977)



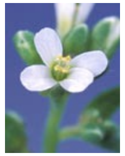
- First organism: *Haemophilus influenzae* - 1.5 Mb (1995)



- First eukaryote: *Saccharomyces cerevisiae* - 12.4 Mb (1996)



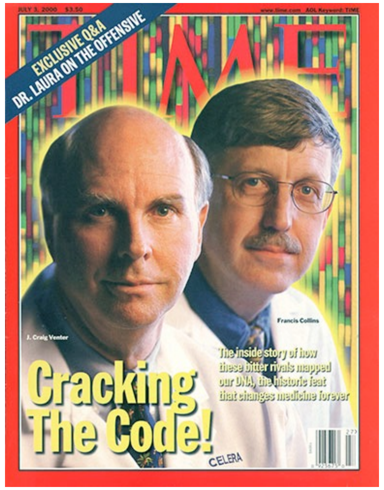
- First multicellular organism: *Cenorhabditis elegans* - 100 MB (1998-2002)

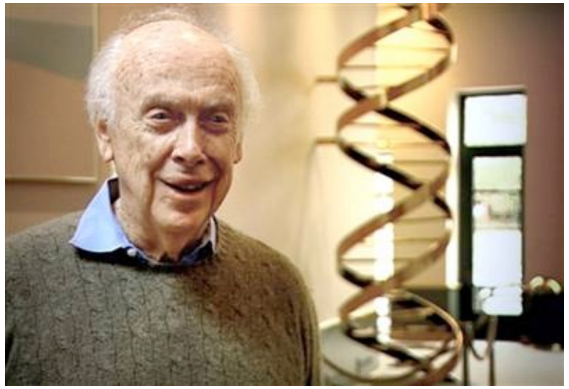


- First plant: *Arabidopsis thaliana* - 157 Mb (2000)

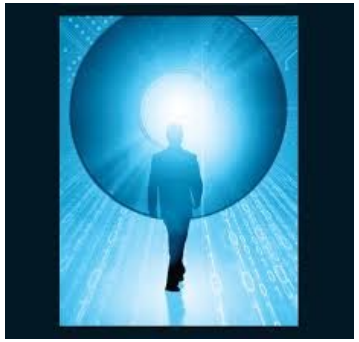
Just an interesting comparison:

- Human genome project, 2007
  - Genome of Craig Venter costs $70M
    - Sanger’s sequencing
  - Genome of James Watson costs $2M
    - 454 pyrosequencing
  - Today: 1000 $ / individual


JULY 3, 2000 $3.50
www.time.com AOL Keyword: TIME
TIME
EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
Cracking The Code!
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
CELERA



Paradigm Change



- From single genes to complete genomes
- From single transcripts to whole transcriptomes
- From single organisms to complex metagenomic pools
- From model organisms to anything

NGS technologies

Company	**Platform**	**Amplification**	**Sequencing method**
Roche	454**	emPCR	Pyrosequencing
Illumina	HiSeq MiSeq	Bridge PCR	Synthesis
LifeTech	SOLiD**	emPCR/ Wildfire	Ligation
LifeTech	Ion Torrent Ion Proton	emPCR	Synthesis (pH)
Pacific Bioscience	RSII	None	Synthesis
Complete genomics	Nanoballs	None	Ligation
Oxford Nanopore*	GridION	None	Flow

RIP technologies: Helicos, Polonator, etc.

In development: Tunneling currents, nanopores, etc.

Differences between platforms

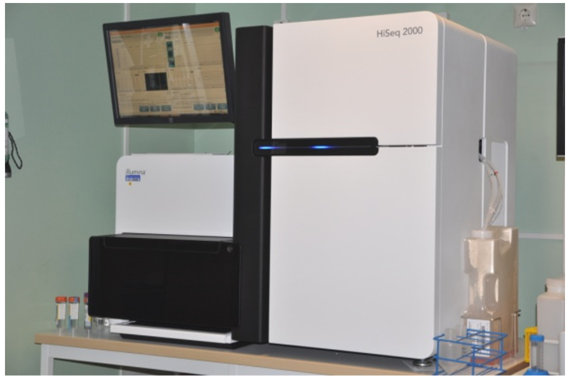
- Technology: chemistry + signal detection
- Run times vary from hours to days
- Production range from Mb to Gb
- Read length from <100 bp to > 20 Kbp
- Accuracy per base from 0.1% to 15%
- Cost per base varies

Illumina

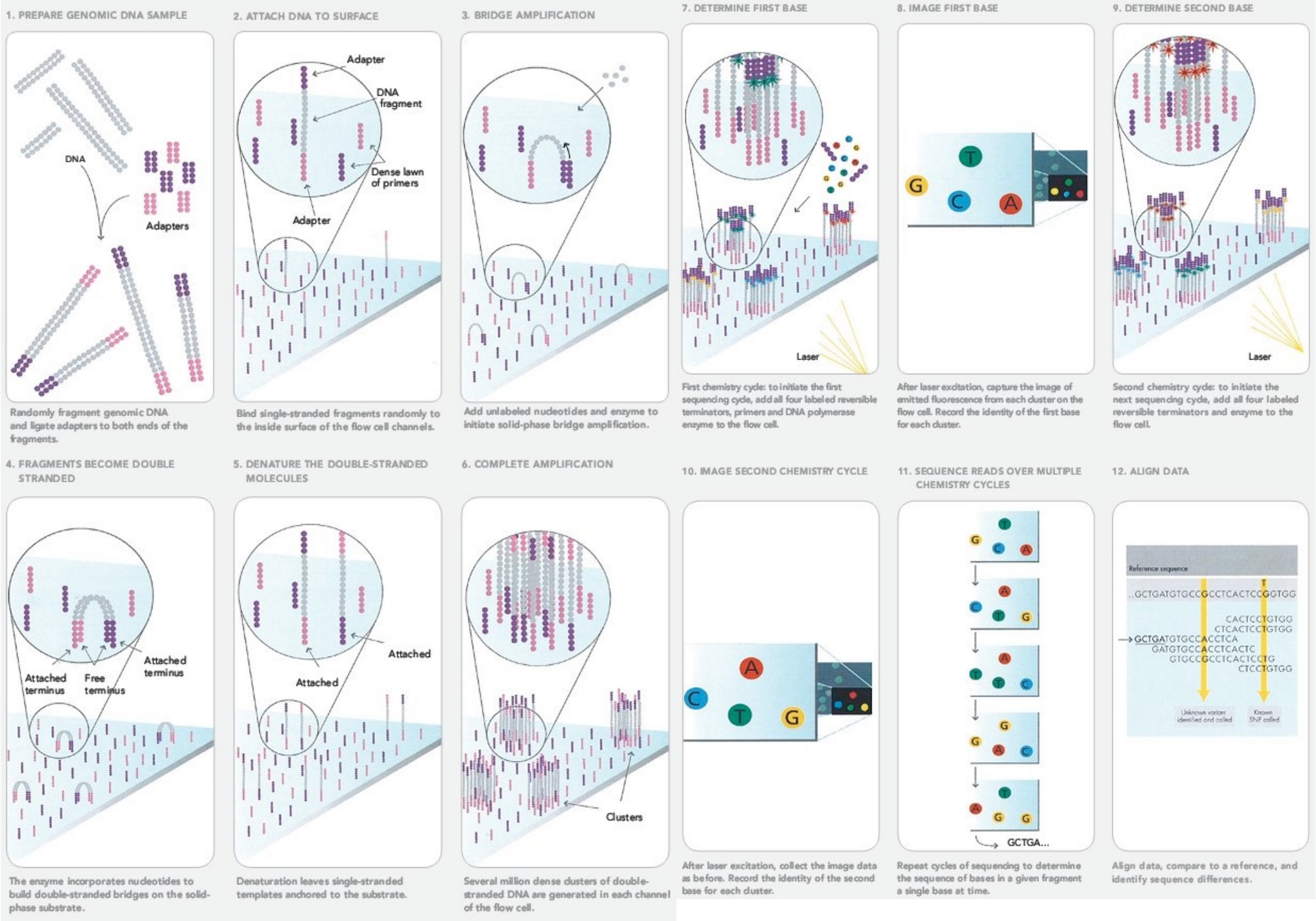
Instrument	**Yield and run time**	**Read Length**	**Error rate**	**Error type**
Upgrade HiSeq2500	120 GB in 27h or standard run	100x100	0.1%	Subst
MiSeq	540 Mb – 15 Gb (4 – 48 hours)	Upp to 350x350	0.1%	Subst

Main applications

- Whole genome, exome and targeted reseq
- Transcriptome analyses
- Methylome and ChiPSeq
- Rapid targeted resequencing (MiSeq)



Illumina



1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

7. DETERMINE FIRST BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

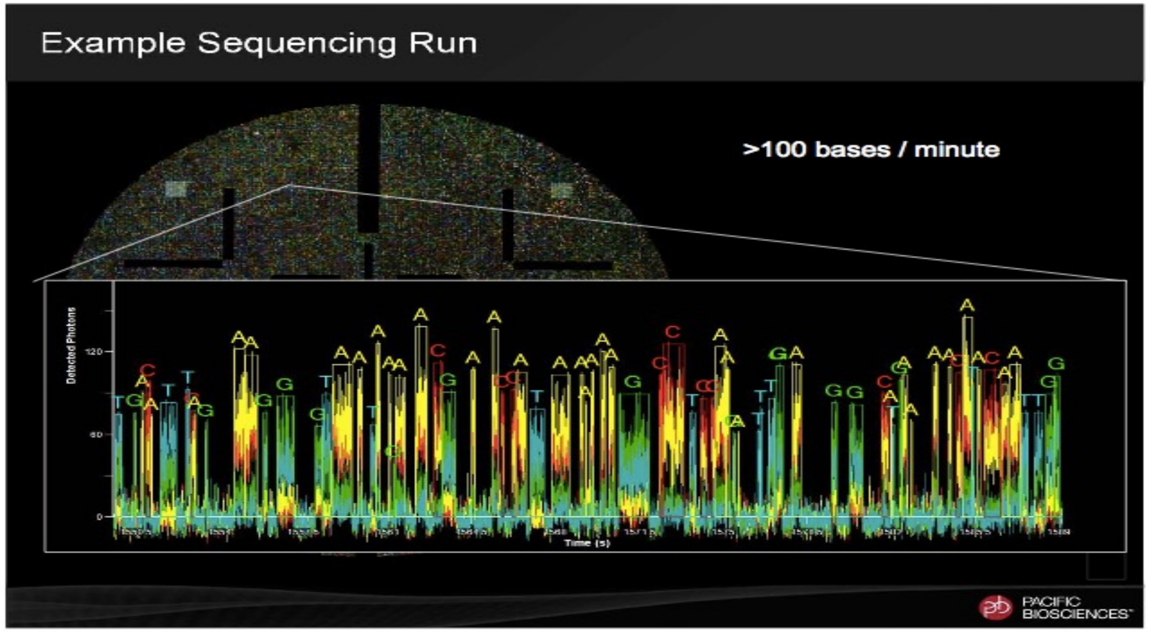
12. ALIGN DATA

Align data, compare to a reference, and identify sequence differences.

Pacific Bioscience

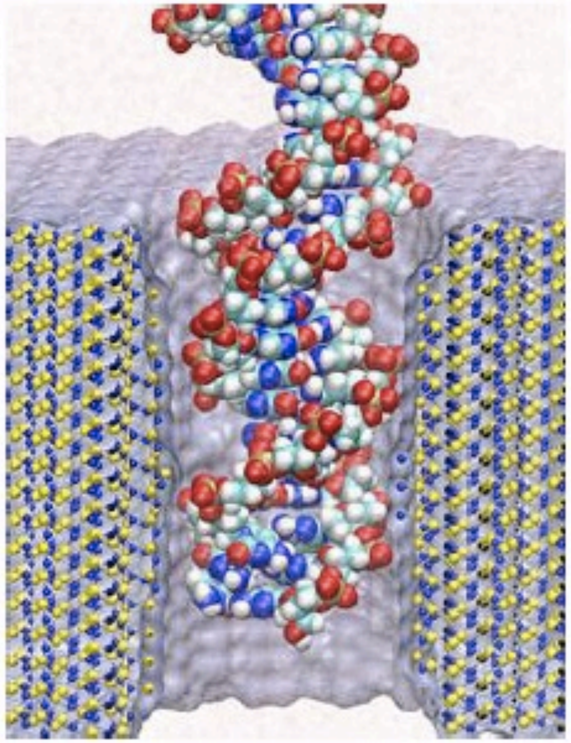
Instrument	Yield and run time	Read Length	Error rate	Error type
RS II	500 MB/180 min SMRTCell	250 bp – 20 000 bp (35 000 bp)	15% (on a single passage!)	Insertions , **random**

Single-Molecule, Real-Time DNA sequencing



Oxford Nanopore Technologies

- » Protein nanopores on silicon chip
- » DNA measured as it's pulled through
- » 125 Gb / day (8K)
- » 50–100,000 base reads
- » 4% error rate
- » As low as $10/Gb (20 x 8K)

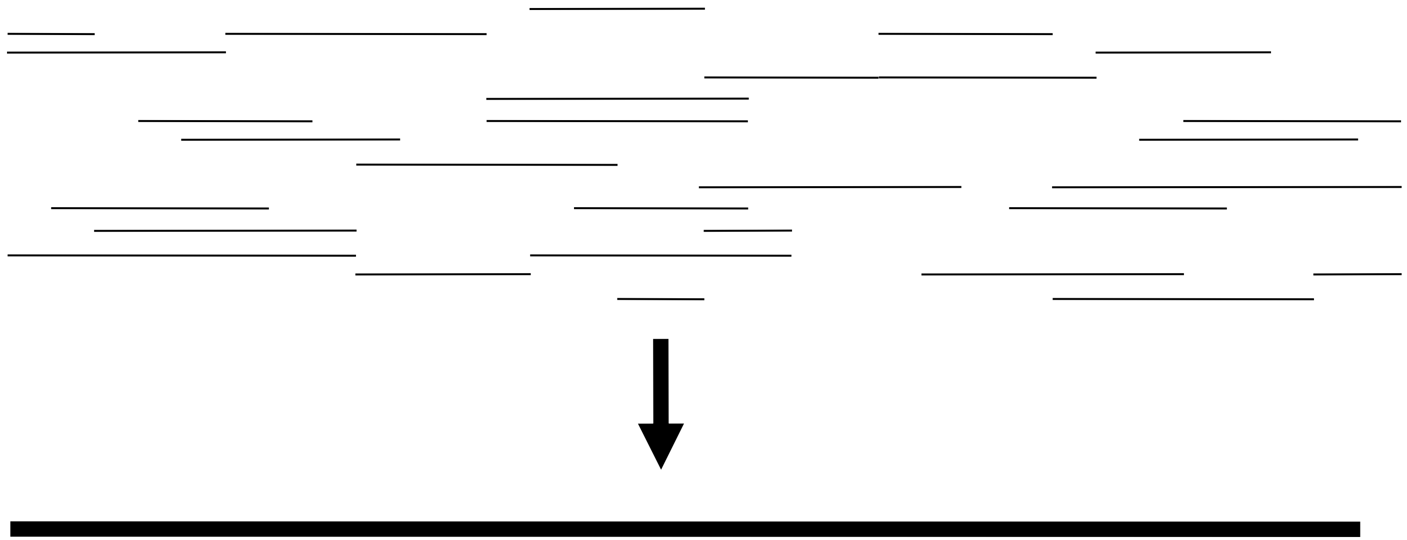


	Illumina HiSeq	Illumina MiSeq	SOLiD Wildfire	Ion Torrent	Ion Proton	PacBio
Read length	100 + 100 bp (150+150 bp)	250 + 250 bp (350+350 bp)	75 bp	200 bp 400 bp (500 bp)	150 bp 200 bp	1 – 20 Kbp
WGS:						
- human	++++		(+)		+	(+)
- small	+++	+++	(+)	++++	+++	+++++
De novo	+++	++		+++	++	+++++
RNA-seq	+++		+++		+++	+++*
miRNA	+++		+++			
ChIP	+++		++++			
Amplicon	++	+++		+++	+++	+++
Metylation	+++					++++*
Target re-seq	++	+++	(+)		+++	+++
Exome	+++		(+)		++++	(+)

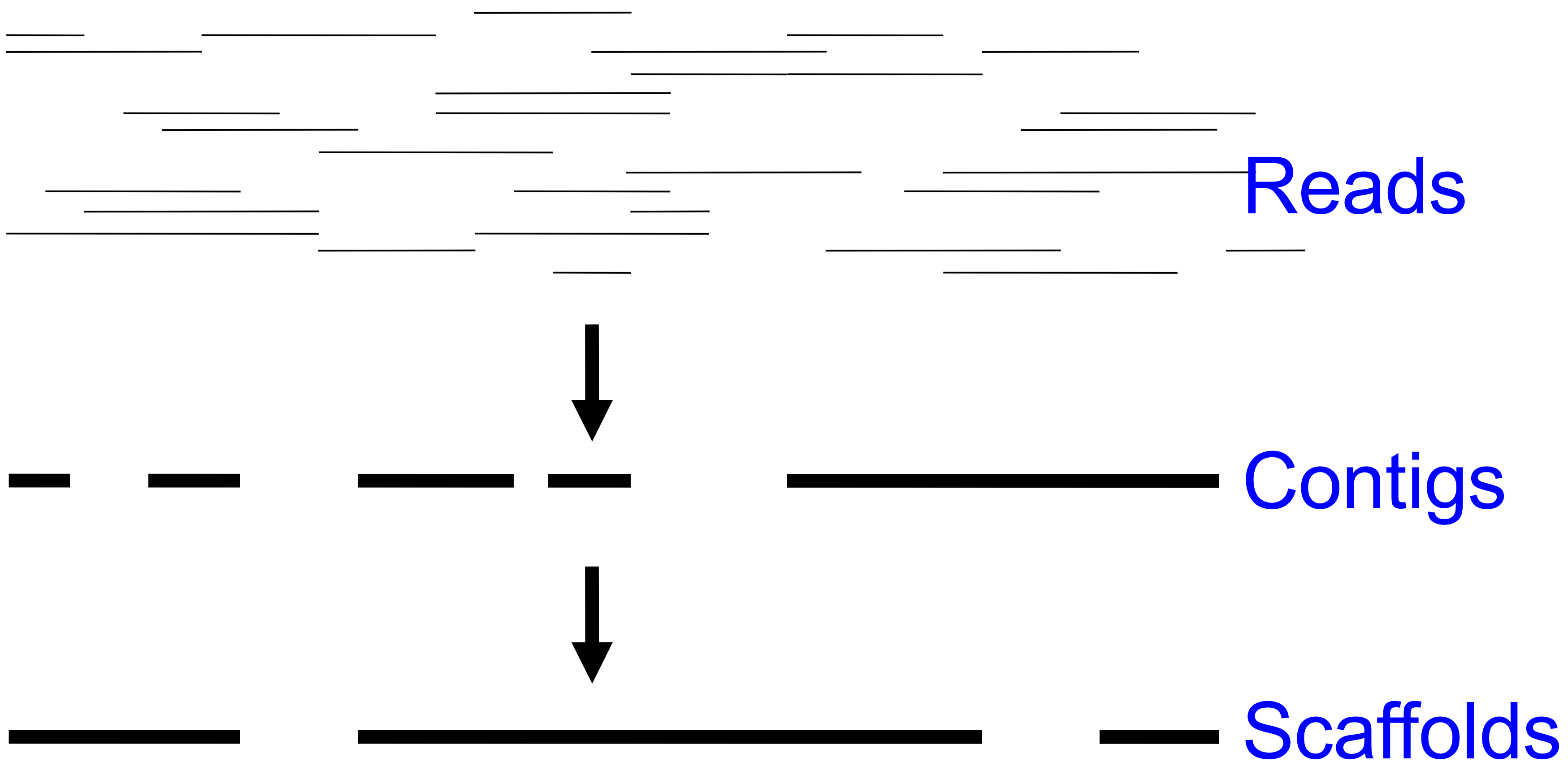
Exam Question Hint

- Given a sequencing problem, which platform would we want to use

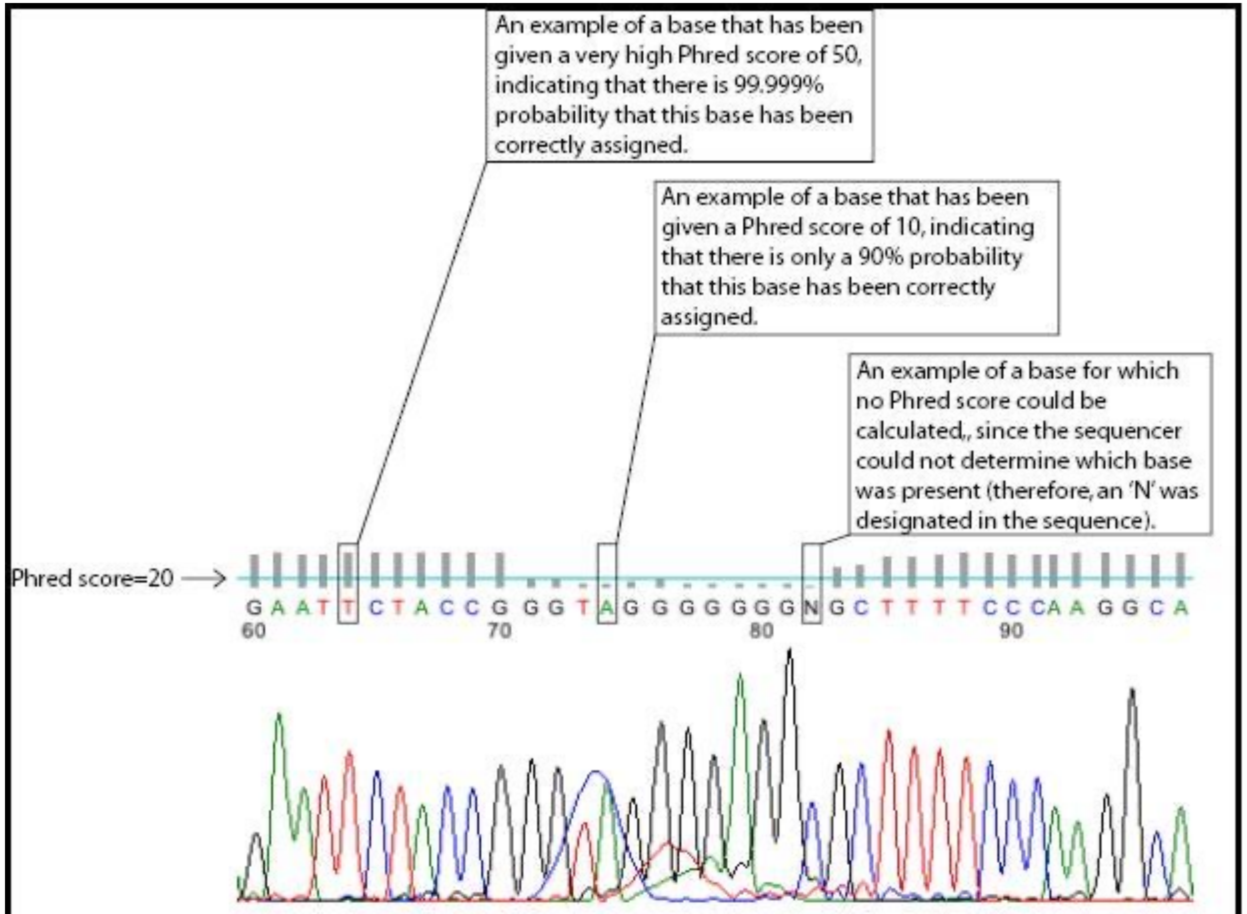
Sequence assembly



Sequence assembly



Phred



An example of a base that has been given a very high Phred score of 50, indicating that there is 99.999% probability that this base has been correctly assigned.

An example of a base that has been given a Phred score of 10, indicating that there is only a 90% probability that this base has been correctly assigned.

An example of a base for which no Phred score could be calculated,, since the sequencer could not determine which base was present (therefore, an 'N' was designated in the sequence).

Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).

Four approaches to assembly

- Naïve approach
- Greedy approach
- Overlap / Layout / Consensus
- de Bruijn Graphs

Naïve approach

- Compare every sequence to every other sequence
- Find stretches that are the same
- Need to account for phred scores – what if a base is wrong?
- How long of a sequence do you need to be unique?

Sequence composition

- 4 bases
- 4^n chance of finding a sequence if all evenly used (they are not)
- 3 bp: $4^3 = 64$
- 8 bp: $4^8 = 65,336$
- 20 bp: $4^{20} = 1,099,511,627,776$

Greedy approaches

- Start with a sequence

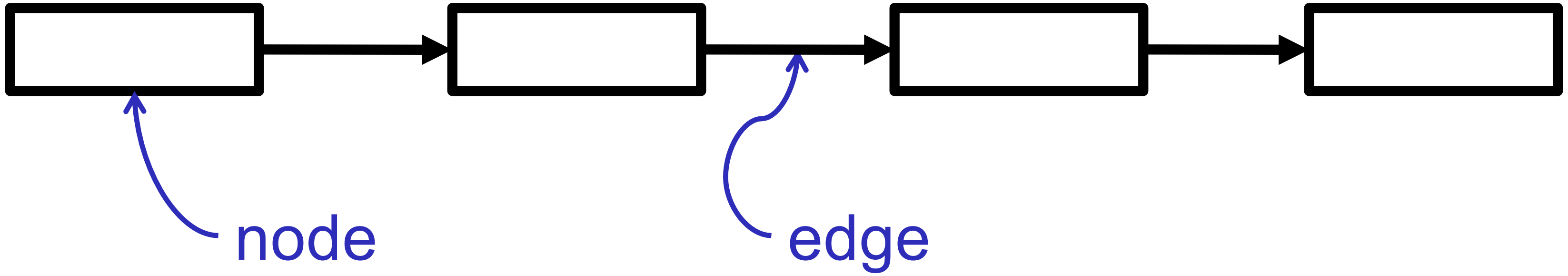
- Keep extending it while another sequence matches the end

- When can not be extended further, mark as a contig

Assembly is a “graph” problem

- Overlap/Layout/Consensus
- de Bruijn Graph
- Greedy graphs

- A graph is nodes + edges



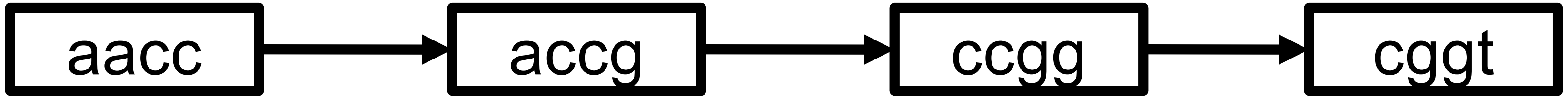
Assemble these two sequences!

```
           AACCGGT
             CCGGTTA

Consensus: AACCGGTTA
```

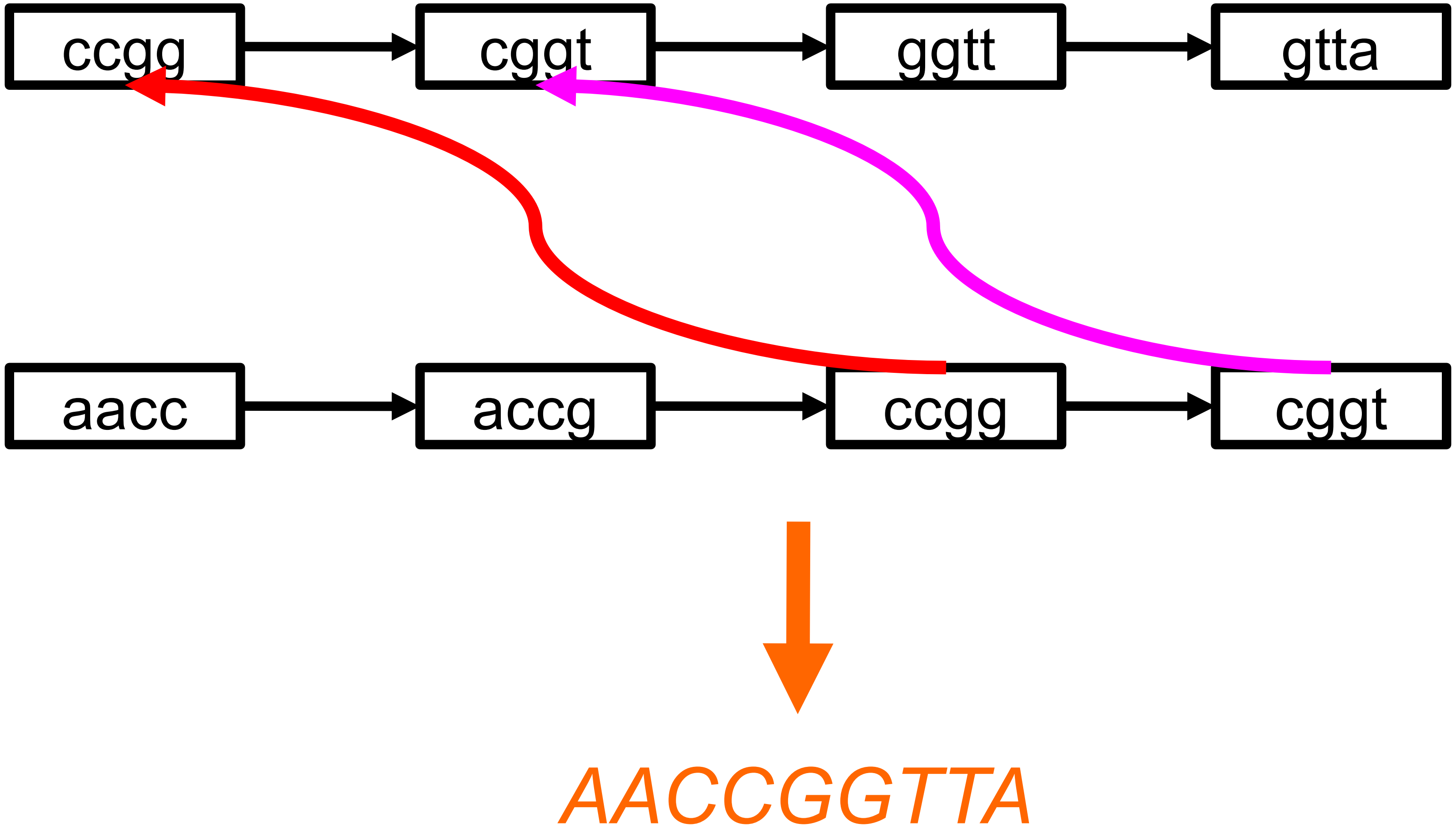
AACCGGT as graphs

Node = K-mers; edges = nodes that overlap by K-1 bases.

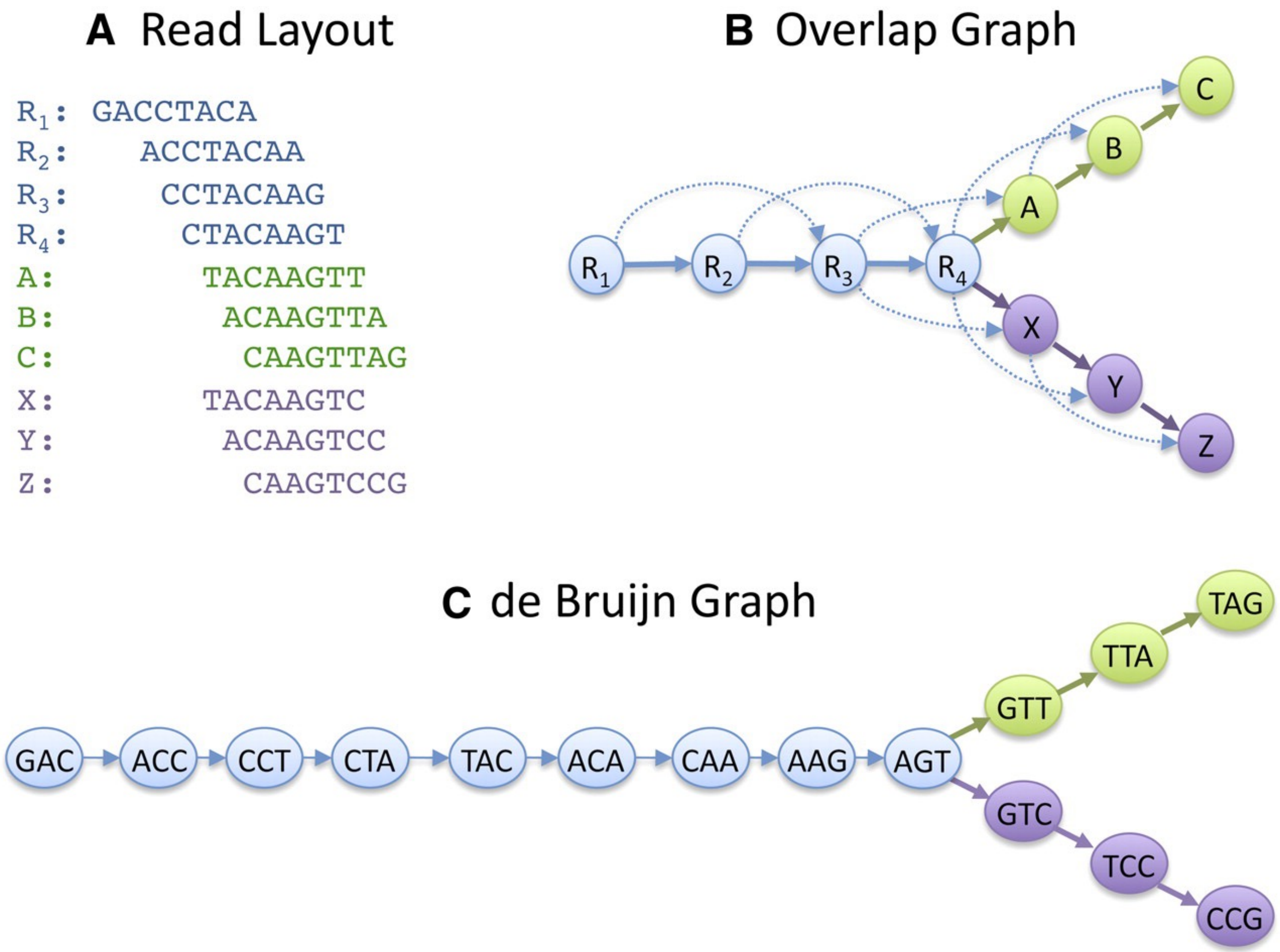


Here K = 4, but in reality K = 19 to 31

Join the two graphs



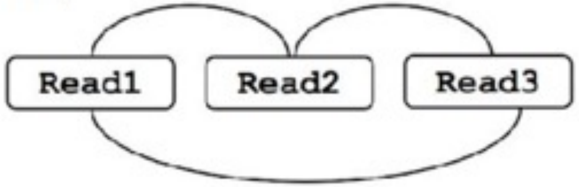
Differences between overlap graphs and de Bruijn graphs for assembly.



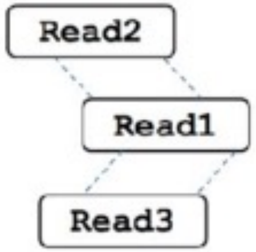
Schatz M C et al. Genome Res. 2010;20:1165-1173

(a) Overlap, Layout, Consensus assembly

(i) Find overlaps



(ii) Layout reads



(iii) Build consensus

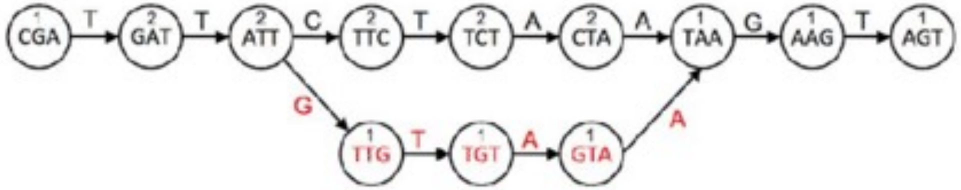
```
CGATTCTA
   TTCTAAGT
 GATTGTAA
-----------
CGATTCTAAGT
```

(b) De Bruijn graph assembly

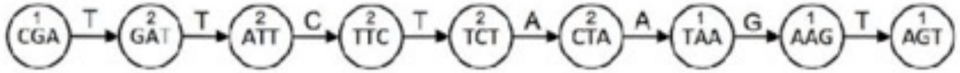
(i) Make kmers

```
Read1: TTCTAAGT   Read2: CGATTCTA   Read3: GATTGTAA
Kmers: TTC        Kmers: CGA        Kmers: GAT
        TCT               GAT               ATT
         CTA               ATT               TTG
          TAA               TTC               TGT
           AAG               TCT               GTA
            AGT               CTA               TAA
```

(ii) Build graph



(iii) Walk graph and output contigs



CGATTCTAAGT

HW2

- Will be released in next few days after adjustments are made
- First assignment that can be put in a github portfolio for job applications (all of the remaining assignments will be this way)
- We will write a simplified genome assembler:
  1. Write a brute force approach (team)
  2. Speed up brute force approach with graph algorithms (team)
  3. Compare with megahit (individual)
- Two weeks to complete

Next Class

- I will be coding, working on the document scanner on assembly problems