COMP90016 – Computational Genomics

Sequencing DNA (and other things)

Department of Computing and Information Systems

The University of Melbourne

Administration

- Student representative.
- Volunteers?

Overview

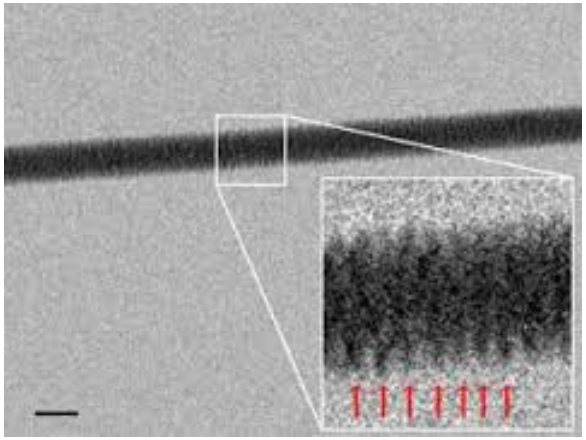
- What is sequencing for?
- How does sequencing work?
- What are the limitations of sequencing?
- How can these limitations be mitigated?

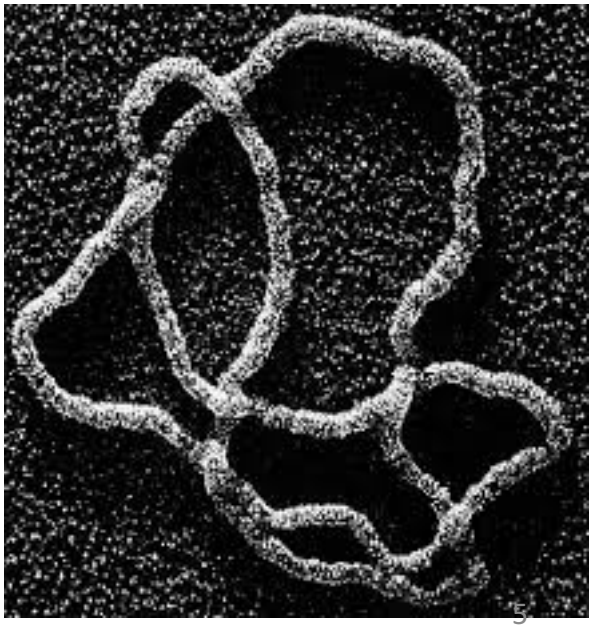
The Human Genome Project

- "The Human Genome Project was a 13-year-long, publicly funded project initiated in 1990 with the objective of determining the DNA sequence of the entire euchromatic human genome within 15 years."
- It cost roughly 3 billion dollars in public funding.
- "To map the very stuff of life; to look into the genetic mirror and watch a million generations march past. That, friends, is both our curse and our proudest achievement. For it is in reaching to our beginnings that we begin to learn who we truly are." (From the Civilization video game series)
- https://en.wikipedia.org/wiki/Human_Genome_Project

What is Sequencing For?

- DNA (and RNA, or Proteins for that matter) is too small to be broken down into its individual nucleotides (or amino acids) visually (i.e. microscopy).
- Sequencing is a class of technology that enables us to *read* the DNA of an organism.
- Reading the DNA allows us to determine the genome of an organism.
- Sequencing == (DNA->{ACGT}*)





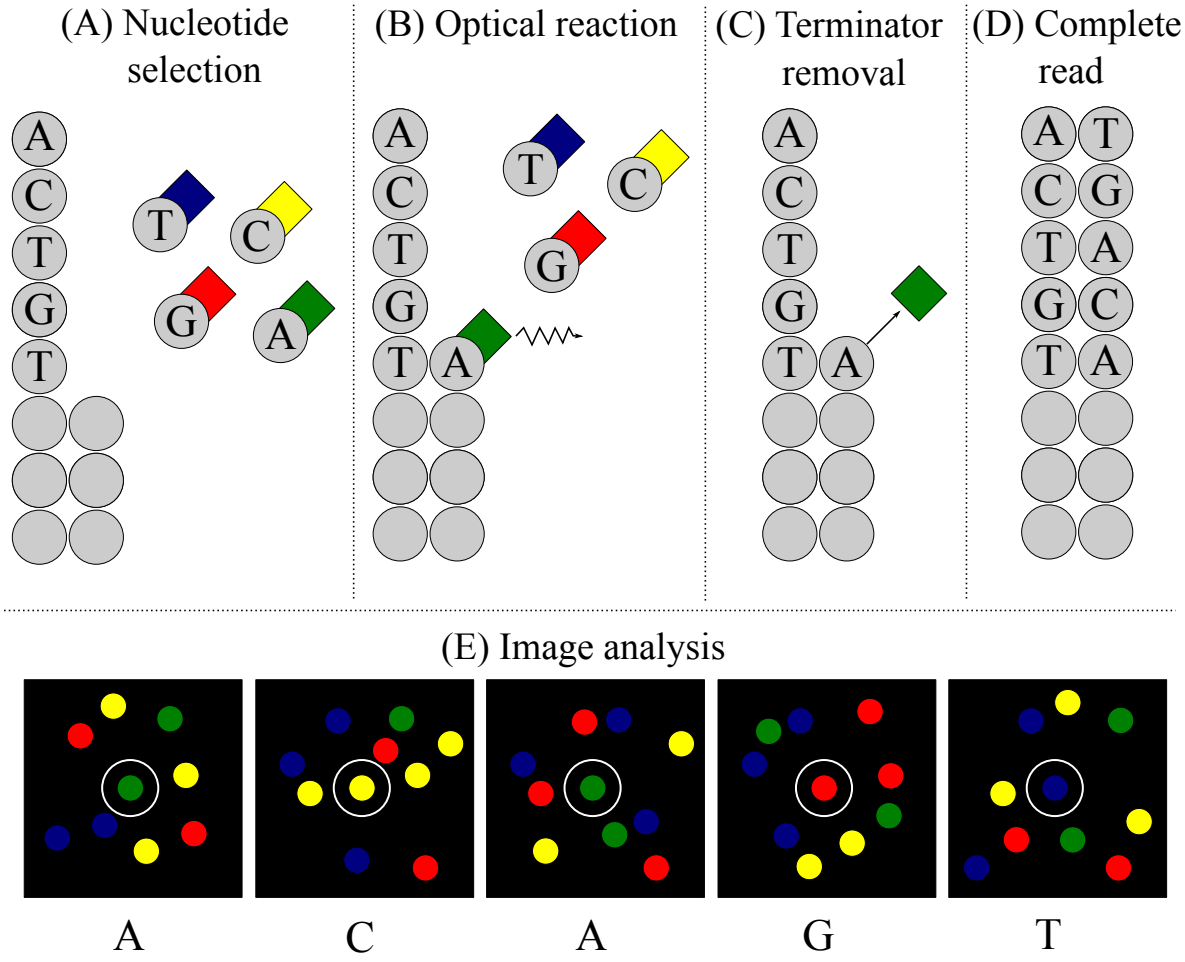
How Does Sequencing Work?

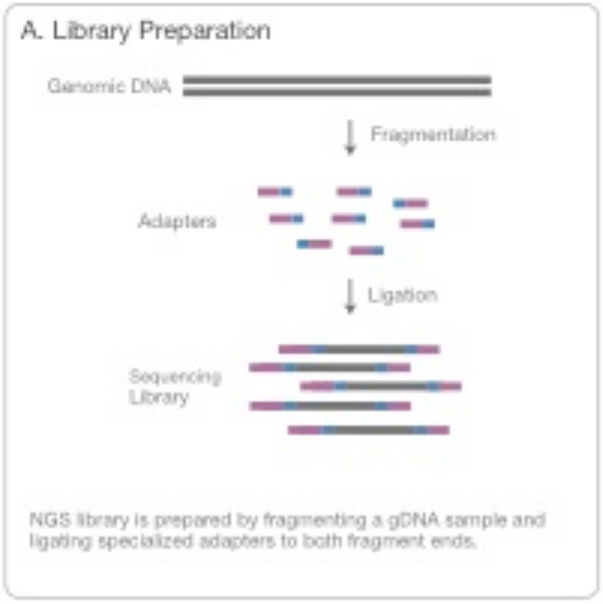
- There are numerous technologies to sequence DNA.
- The process of sequencing RNA is largely identical to that of DNA.
- Sequencing proteins is very different; we are not discussing it in this lecture.
- The following slide explains the Illumina sequencing technology (the current market leader).

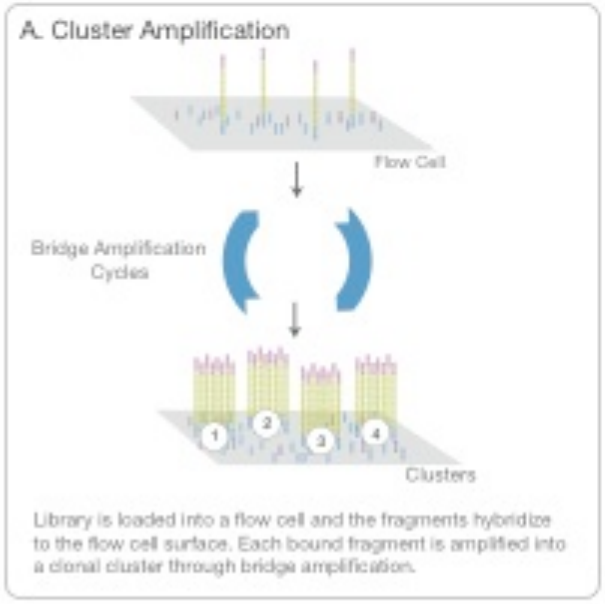
How Does Sequencing Work? Illumina Sequencing

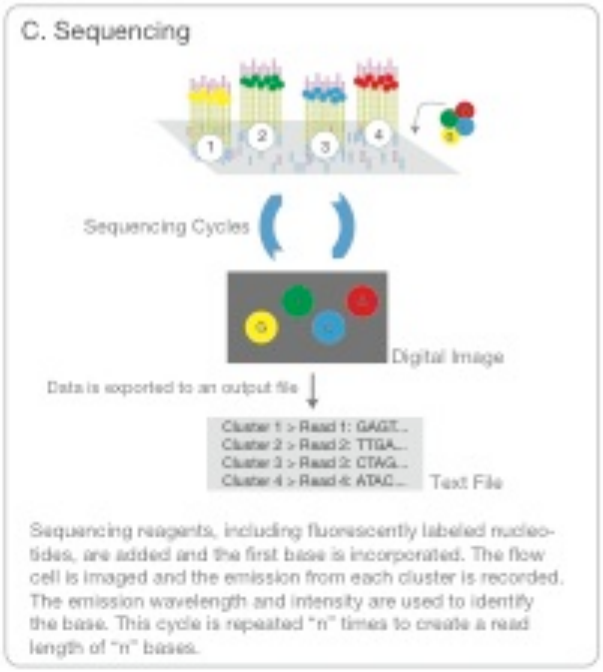
- The Illumina technology utilises nature's own polymerase to synthesize single stranded DNA into double stranded DNA in a controlled manner:
  1. Turn double-stranded DNA into single-stranded DNA
  2. Start DNA polymerase, supplying it with special nucleotides.
  3. Nucleotides have a fluorescent die attached that allow the machine to identify them.
  4. Nucleotides have a terminator attached that will stop polymerase after synthesizing a single base.

How Does Sequencing Work? Illumina Sequencing (2)









Library preparation is an art, and is the point of difference between DNA seq and RNA seq. We won't dwell on it, since it's a "wet lab" step.

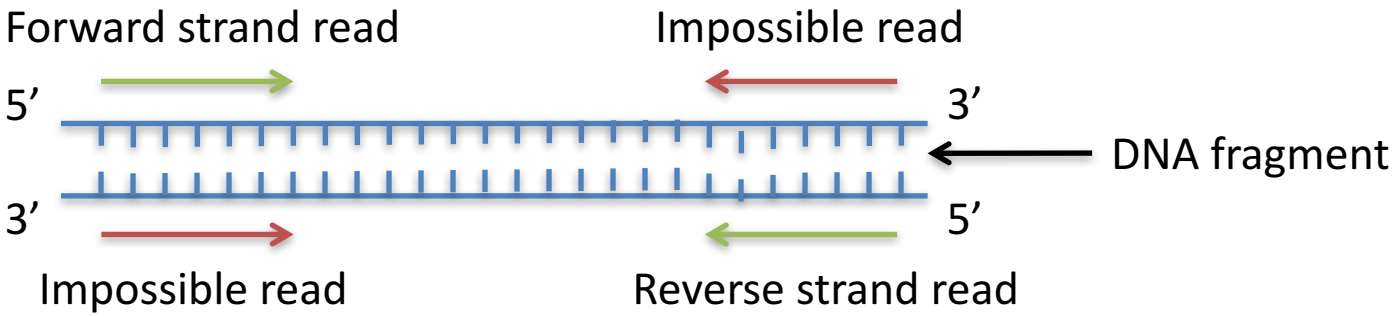
Two additional notes:

1. Reads are sequenced as clusters.
2. Reads are sequenced in parallel.

Step D is what bioinformatics is all about.

A Note About Strandedness

- Since nature's own machinery (polymerase) is used for the task of sequencing , the reading direction of DNA is preserved.
- That is, the read is sequenced in the 5' to 3' direction of its strand.
- However, which strand a read comes from cannot be controlled nor reconstructed (a priori).



The Evolution of Sequencing

- http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

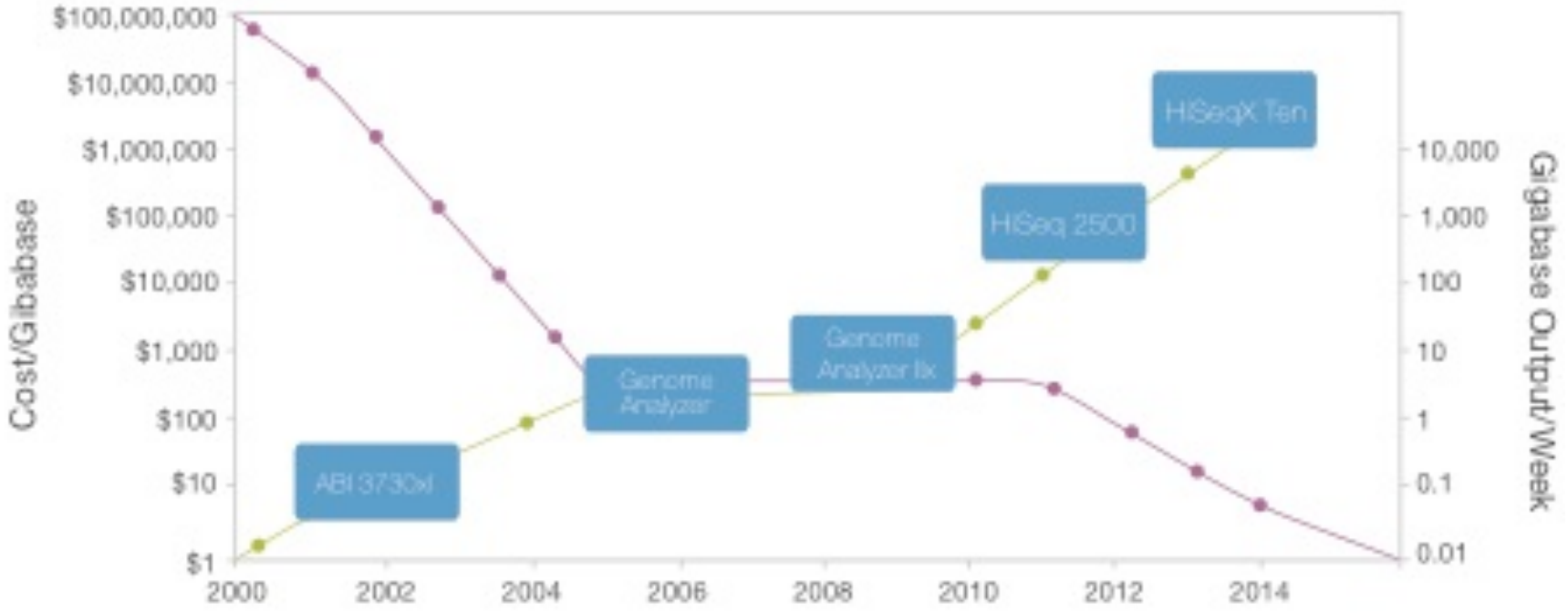
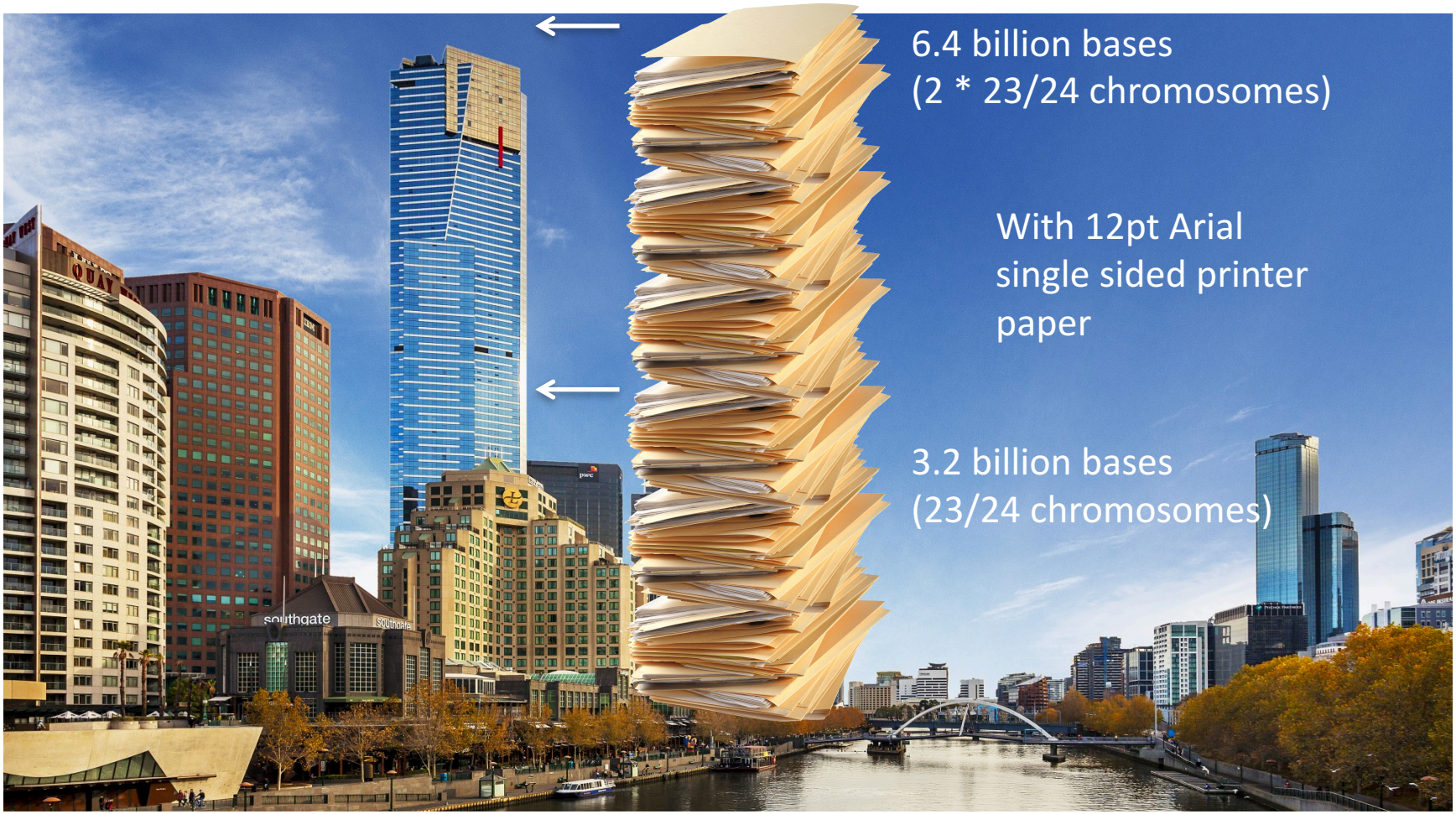


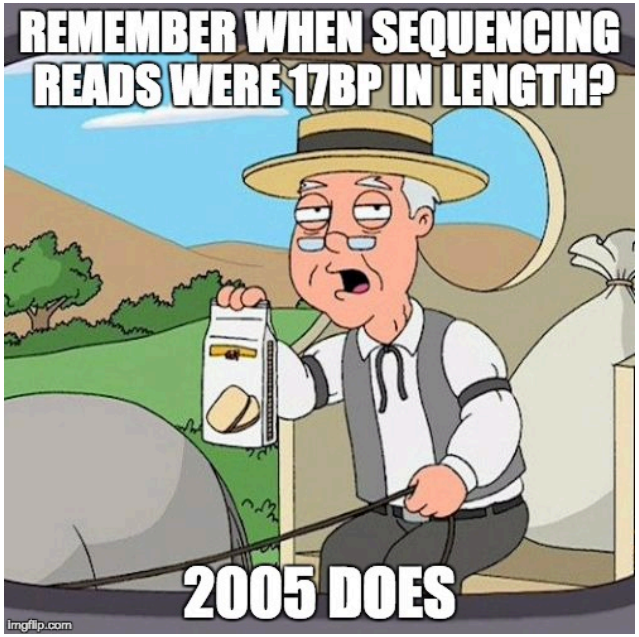
Figure 1: Sequencing Cost and Data Output Since 2000—The dramatic rise of data output and concurrent falling cost of sequencing since 2000. The Y-axes on both sides of the graph are logarithmic.

How Big is 3 x 10^9 Bases?



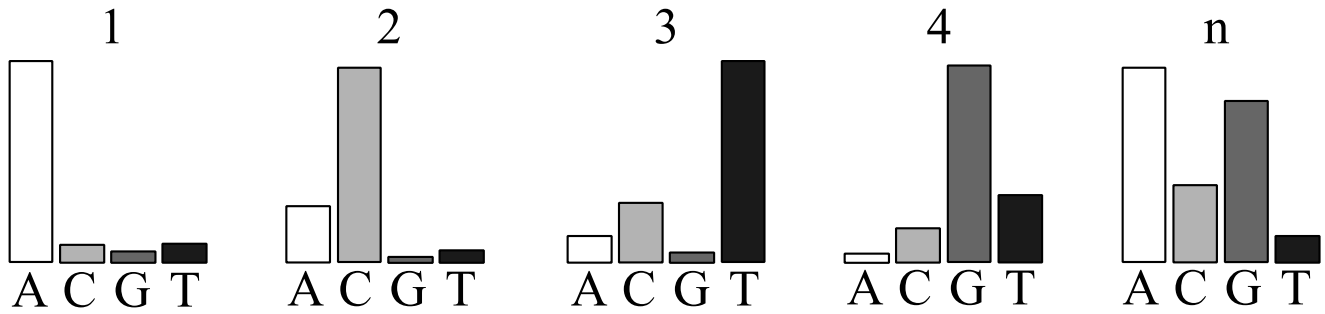
The Limitations of Sequencing (1)

- Length!
  - Reads are tiny compared to the genome sizes that we are interested in.
  - Illumina's maximum length is now 250bp.
  - Repeats in the human genome can be 1000s of bp in length.



The Limitations of Sequencing (2)

- Errors!
  - The technology is not perfect and various technical issues can change a base from one nucleotide to another, or insert unknown bases (Ns).
  - Example: Phasing



Discussion Question

- How can we tackle a large genome (such as that of humans) with tiny reads?
- How can 100bp reads determine the entire sequence of a chromosome (hundreds of millions of bases)?

Solutions to the Limitations of Sequencing Technologies

- Handling the problem of short reads:
  - Shotgun sequencing
  - Paired-end sequencing
  - Mate-pair sequencing
- Handling the problem of errors in reads:
  - Error Correction algorithms
  - Sensitive alignment algorithms

Shotgun Sequencing

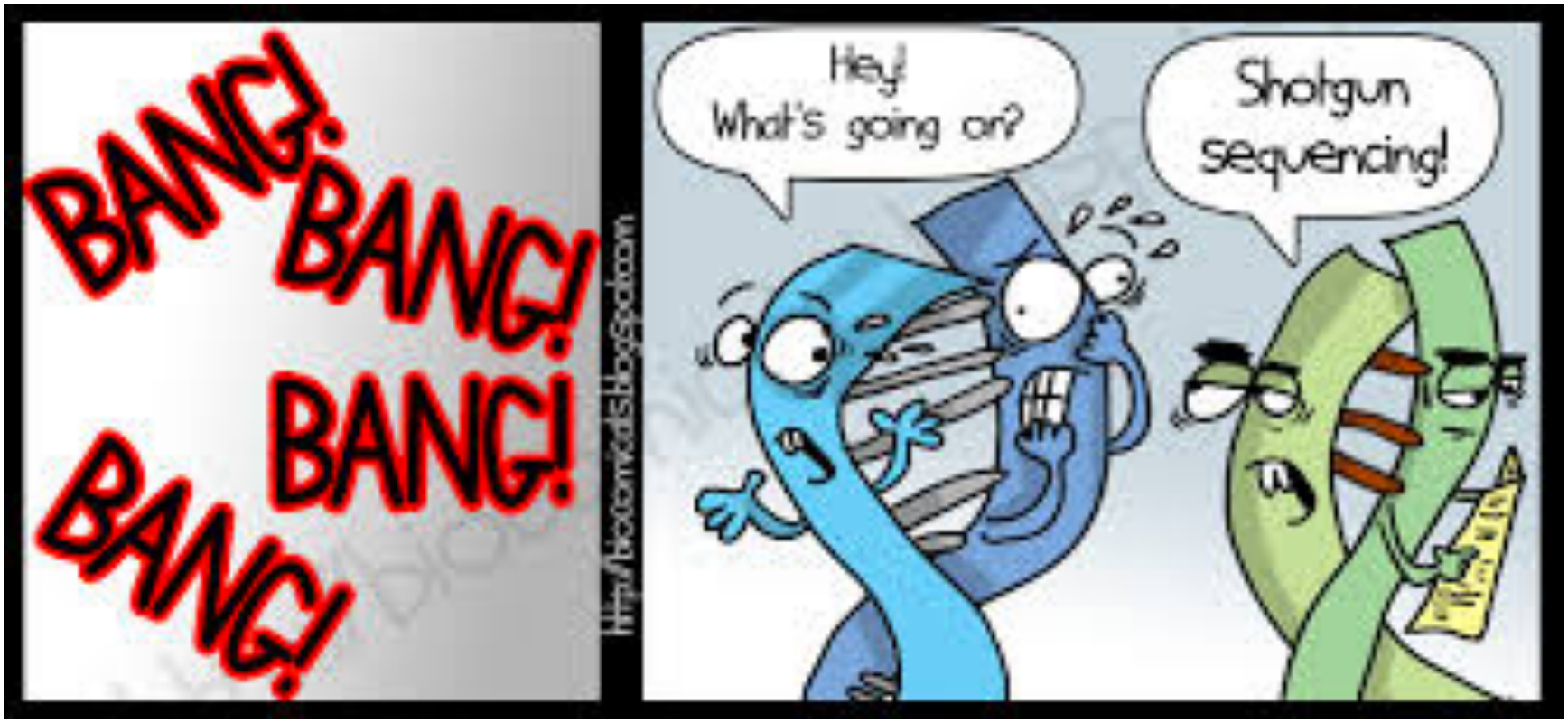
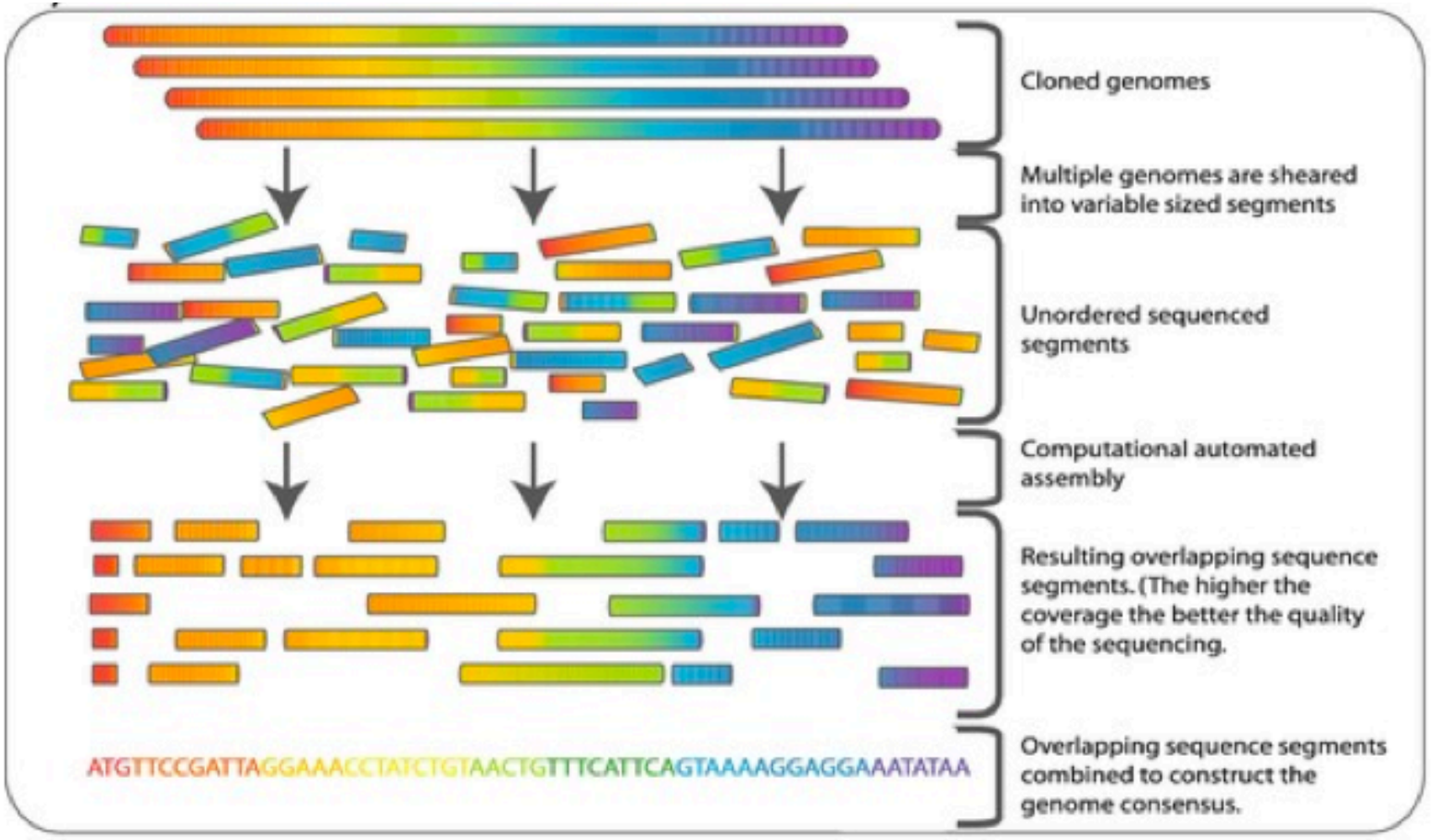


Image from biocomicals.blogspot.com

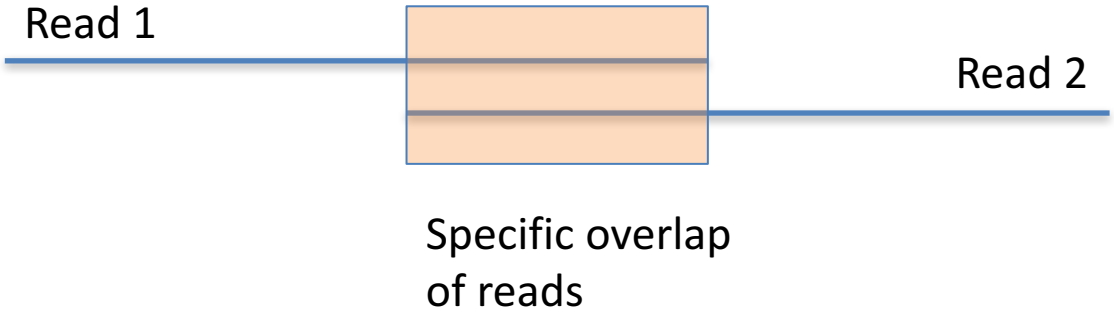
Whole Genome Shotgun Sequencing



Wikipedia

Whole Genome Shotgun Sequencing (2)

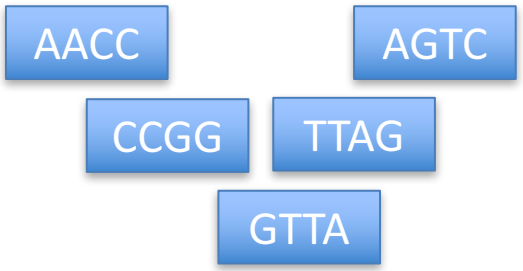
- Redundancy is the key to extending our knowledge about the genome beyond the read length:



- The principle of de novo assembly

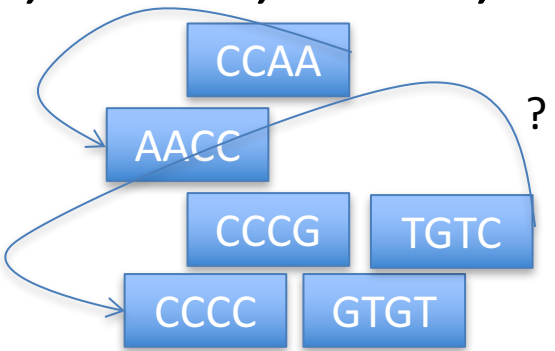
De Novo Assembly Toy Example

- Consider the “genome”:
  AACCGGTTAGTC
- Our “sequencing machine” can only read DNA of length 4.
- Consider the “read set”:
  TTAG, CCGG, GTTA, AACC, AGTC



Repeats – An Example

- Overlaps between reads are only useful as long as they are specific.
- Consider the “genome”:
  AACCCCGTGTCCCCAA
- The sequencing machine from before may produce the following reads:
  TGTC, CCCG, CCAA, GTGT, AACC, CCCC



Problems with the Shotgun Sequencing Assembly Approach

- Repetitive sequence in genomes.

- As hinted at earlier, genomes contain repetitive elements. Some more than others. With different lengths and levels of conservation.
- At the overlap of reads level, up to 10-30% of a genome cannot be discovered by the overlap method, because they are not unique.
- For further reading:
  Wentian Li and Jan Freudenberg. *Mappability and read length.* Front Genet. 2014; 5: 381.

Paired-End Sequencing to Combat Repeats

Fragments are sequenced twice:
Once from each end and strand.
This technique generates reads in a pairwise manner – with a known distance between them (depending on how well the fragment size can be controlled).
The two reads are on opposite strands of the DNA.

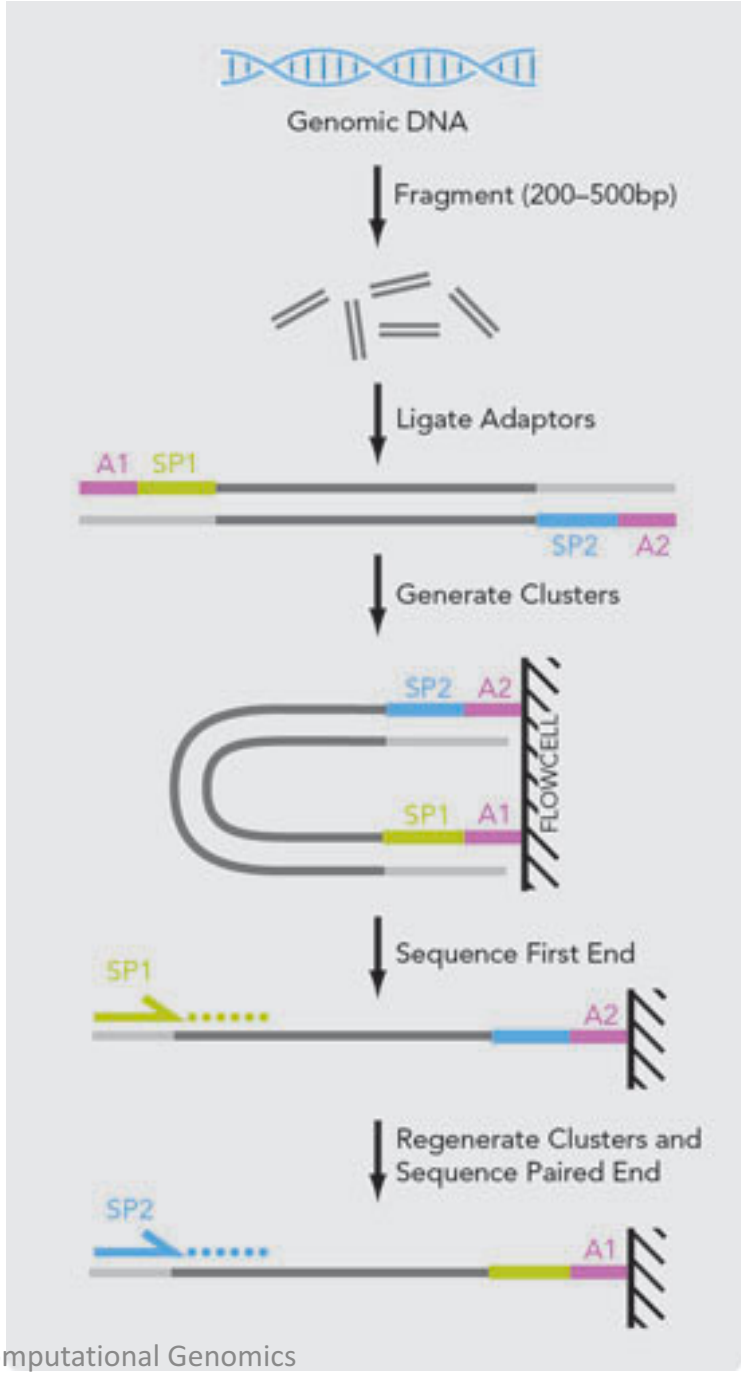
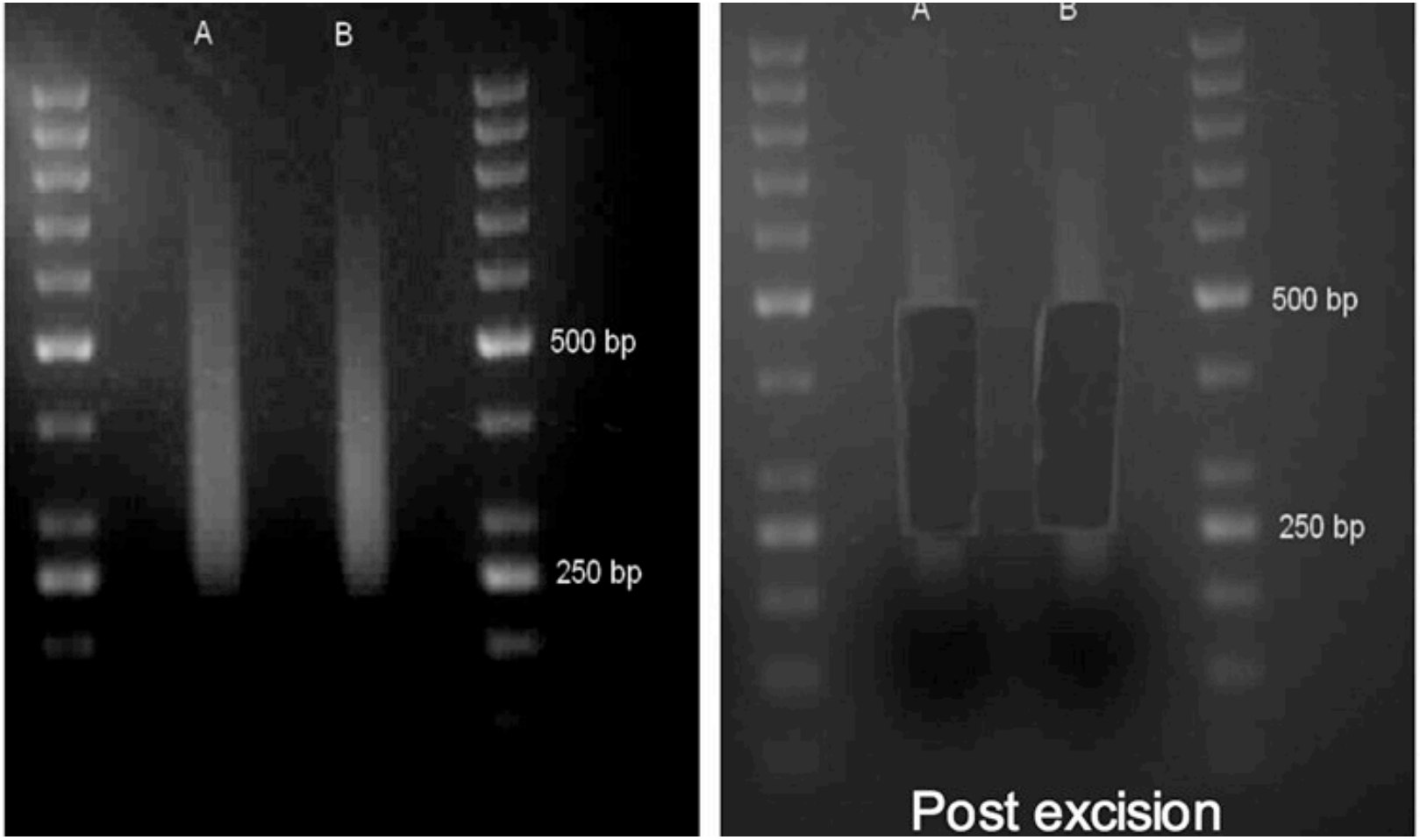


Image from Illumina.com

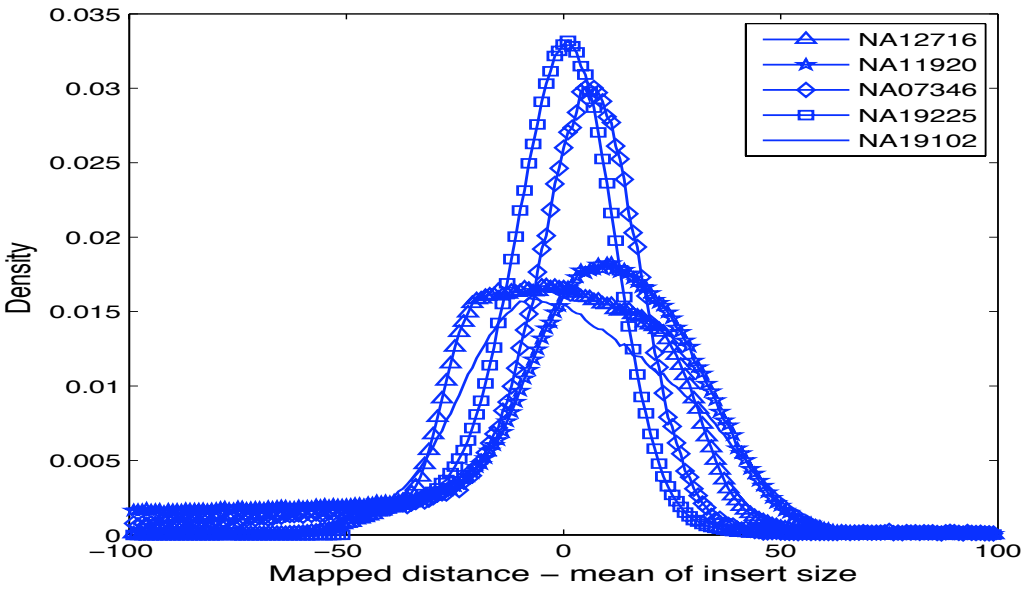
Size selection for paired-end sequencing



Size selection ensures an approximately known distance between each two reads in a read pair.

Insert sizes

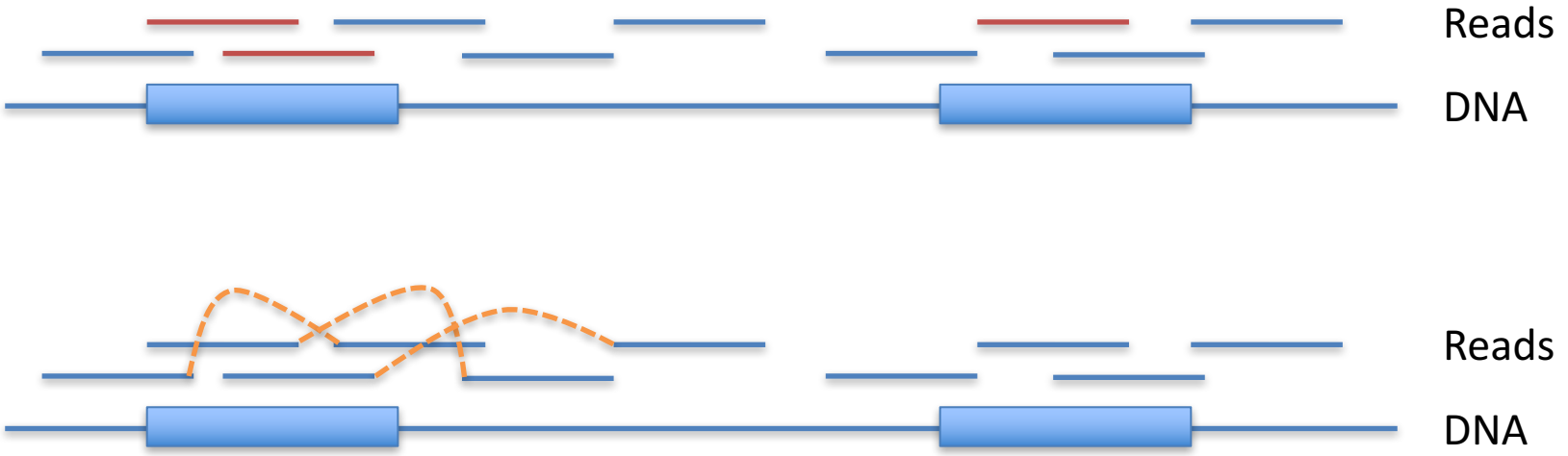
Examples of normalised insert size distributions



Any pair is expected to be within the expected distance and orientation on the same chromosome.

Discussion Question

- How does PE sequencing help in the face of repeats?



- The repeat problem is pushed out to the fragment length.

Paired-End Sequencing Summary

1. The Genome gets amplified and randomly fragmented.
2. The fragments are run on a gel (gel-electrophoreses) to separate by size.
3. A subset of DNA fragments is extracted from the gel.
4. Fragments are sequenced from both ends.

The length of a fragment is referred to as the fragment length or, curiously, as the insert size.

Resequencing

- Once a reference genome (or several) for an organism has been determined, we can infer the sequenced DNA by comparison:
  - Instead of finding overlaps of reads with each other, we identify similarity to the reference.
- This process is called *mapping* or *alignment*, and is going to be the subject of another lecture.

Resequencing Toy Example

Reference genome: AACCCCGTGTCCCCAA

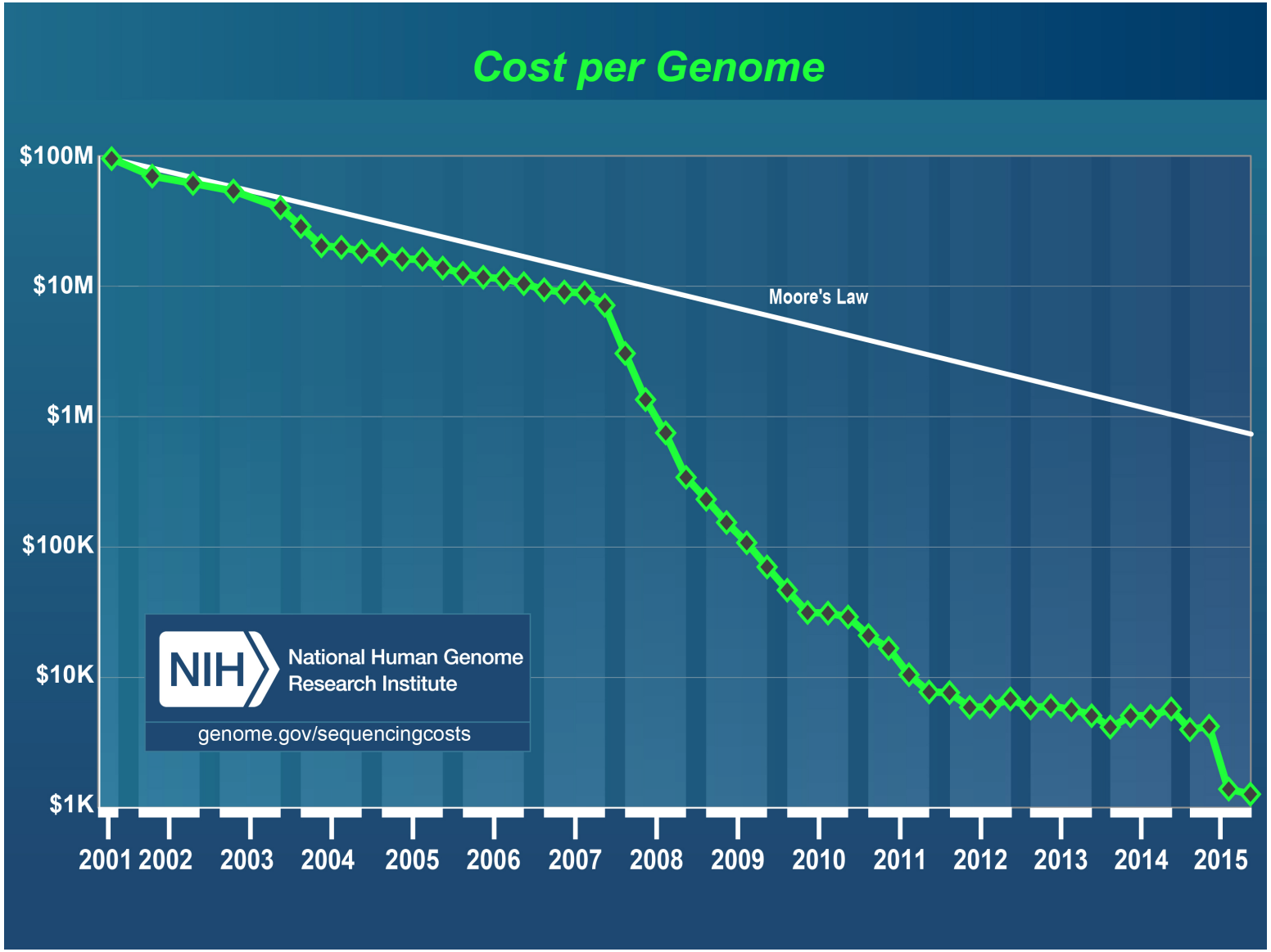
(example as before, see repeats)

Read set: TGTC, CCCG, CCAA, GTGT, AACC, CCCC

Alignment:

```
AACCCCGTGTCCCCAA
AACC
   CCCG
      GTGT                    CCCC?
       TGTC
            CCAA
```

Cost of Sequencing



Summary

- Sequencing is the operation of *reading* DNA.
- Sequencing is the most important and most ubiquitous technology for computational genomics (bioinformatics, etc.)
- Even with its limitations, it has helped to decipher the genetic codes of numerous organisms and continues to expand in application.
- There has been a staggering technology evolution for sequencing: while the first human genome cost $3bn to sequence, we can now read an individual's genome for ~$1000.

Genomics Data: FASTA files

>first line has description

sequence starts on second line and can have more lines of sequence (DNA or protein) after it

>there can be another sequence descriptor

The other sequence and as many more as you want, each preceded by an identifier and then a sequence

Genomics Data: FASTQ files (Sequencing Data)

@first line has description

Read sequence on second line

+there can be another descriptor for the quality string

The fourth line contains the sequencing quality scores (one character per read base)

For example:

- @SimSeq_1/1
- AAATGCCAAAGCTCTATTGTGAGGTCATTTGCTGTGGAGTATGTGCCGTA ATTACTGAGCCACACCCCACCCTGCCTCTCTACCATACGTTCCAAGGGGC
- +
- E6CFGDGAFGBE>FEGBGE?D>EGDBDG4DCE?EG?AFG>DE;FED5C07E C#B?FDG#AF;:?=GEF@E==5EE#1CDAAEE;DBA##?6DD#:AA##?

Accessing Sequencing Data

- Although sequencing data is simply stored in text files, it is tedious to parse the specifics of the format manually.
- There exist libraries for most contemporary programming languages to do this for us.
- Pysam is one of those libraries, which we are going to utilise throughout this course.

Pysam

- Python library that interfaces the Samtools functionality. Samtools is going to get introduces in more detail soon.
- Easy to use and integrate.
- Installed on our lab computers.
- Full API and more on :
  - http://pysam.readthedocs.org

Pysam Example

```
#!/usr/bin/env python
import pysam
f = pysam.FastxFile('reads.fastq')
for read in f:
  print "This is the first read in the file:”
  print "The read name is", read.name
  print "The quality string:", read.quality
  print "Which stands for the following Phred
values:", read.get_quality_array()
  print "The first base of the read is:",
read.sequence[0]
  break
```

Pysam Example

-bash-4.1$ python pysam_example.py
This is the first read in the file:

The read name is D81P8DQ1:112:C1BRYACXX:7:1308:19501:23517

The quality string:
CCCFFFFFGHHGHHIIJJJJJJJJJJJJJJJJJJJJJJJJIJIJJJJJJJJJJJJJJJJJJJJJJJIJIJIJJIHHH
HHFFFFDDDDDDBBBDDDDBDD

Which stands for the following Phred values: array('B', [34, 34, 34, 37, 37, 37, 37, 37, 38, 39, 39, 38, 39, 39, 40, 40, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 40, 41, 40, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 41, 40, 41, 40, 41, 40, 41, 41, 40, 39, 39, 39, 39, 39, 37, 37, 37, 37, 35, 35, 35, 35, 35, 35, 33, 33, 33, 35, 35, 35, 35, 33, 35, 35])

The first base of the read is: T

Further Reading: Long Read Sequencing

- PacBio sequencing: https://www.pacb.com/smrt-science/smrt-sequencing/
- Nanopore sequencing: https://nanoporetech.com/applications/dna-nanopore-sequencing
- These technologies are relatively expensive, but are being developed rapidly. They will become more relevant in the future.

Further Reading

- “Human Y chromosomes have more in common with gorillas than chimpanzees, researchers find”
- Read more: http://www.dailymail.co.uk/sciencetech/article-3473780/Human-Y-chromosomes-common-gorillas-chimpanzees-researchers-find.html#ixzz42CAYSuNE