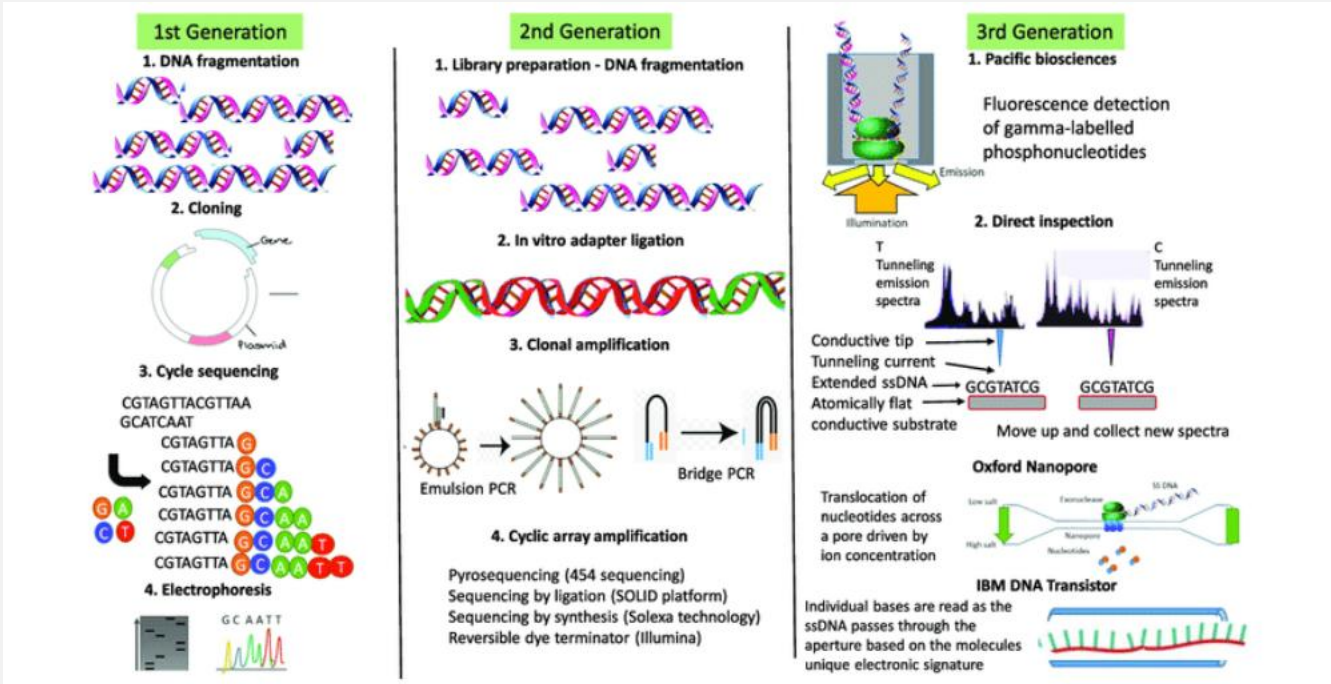
CS4330: Combinatorial Methods in Bioinformatics

Primer on genome sequencing technologies

Wong Limsoon

Genome sequencing technologies



The advancement of DNA sequencing. 1st generation sequencing or Sanger sequencing involves the fragmentation and cloning of the target DNA into plasmid vectors. The DNA is then sequenced using a cyclic chain termination method with either radio isotopically labelled or fluorescently labelled dNTPs. The 2nd generation sequencing technologies are all based on sequencing by synthesis. Two common methods used are emulsion PCR and bridge PCR. Following these methods, different platforms make use of different sequencing technologies. 3rd generation sequencing methods have been developed by many different companies and are based on different technologies. They all involve more direct examination of the target DNA [19].

Image credit:
Zodwa Dlamini

1st generation: Sequencing by cyclic chain termination

1977 – 2000s

Let's watch this video together

https://www.youtube.com/watch?v=ONGdehkB8jU

2nd or next-generation: Sequencing by synthesis

2005 – 2010s

Let's watch this video together

https://www.youtube.com/watch?v=WNM6A9h6GJI

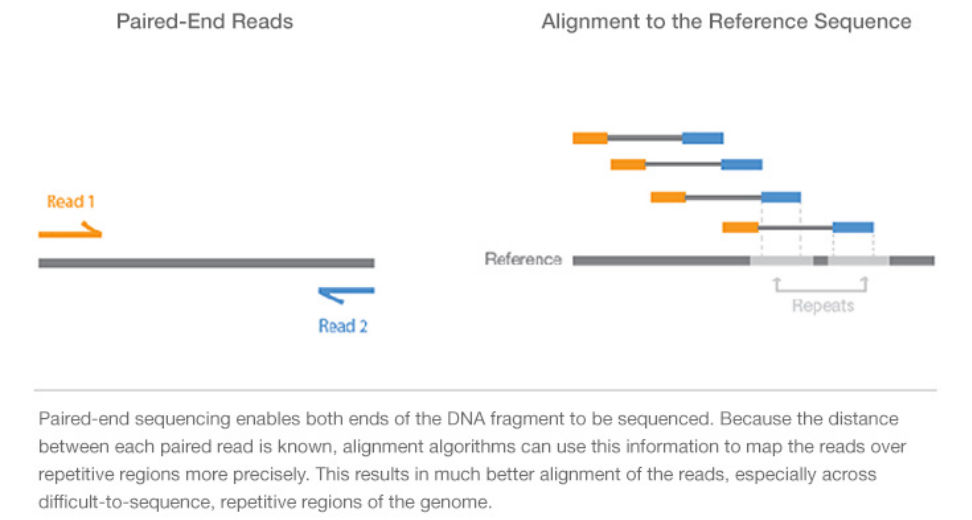
You can also watch this one on your own

https://www.youtube.com/watch?v=CZeN-IgjYCo

Important variation of 2nd-generation sequencing: Pair-end sequencing

Let's watch this video together

https://youtu.be/WneZp3fSJIk



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

3rd generation: Long-read sequencing by "direct inspection"

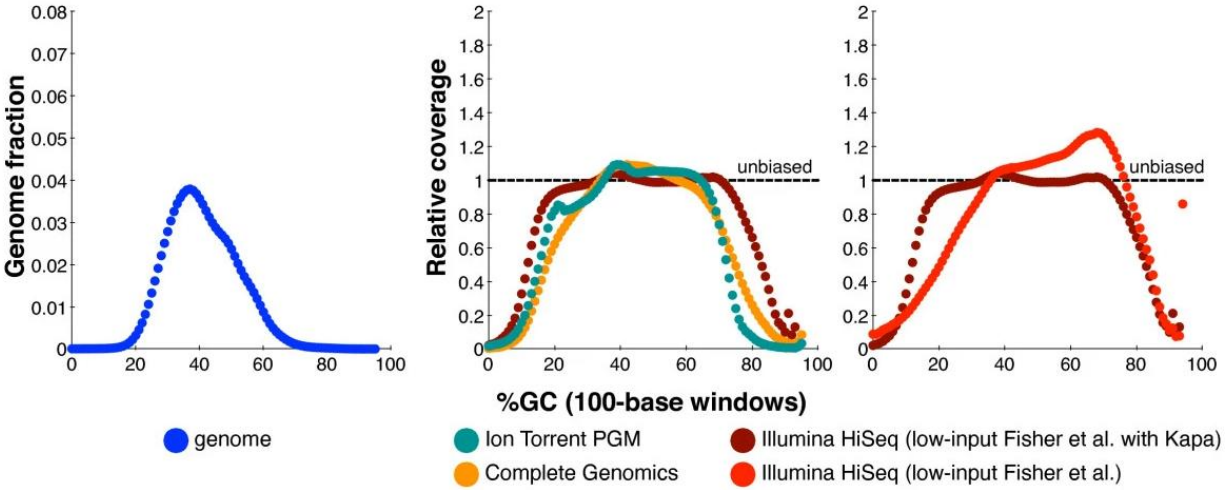
2010s and ongoing

Let's watch this video together

https://www.youtube.com/watch?v=CGWZvHIi3i0

GC-bias in sequencing data

Relative coverage = $\frac{\text{coverage of a gieven reference base in a genome}}{\text{mean coverage of all reference bases}}$.

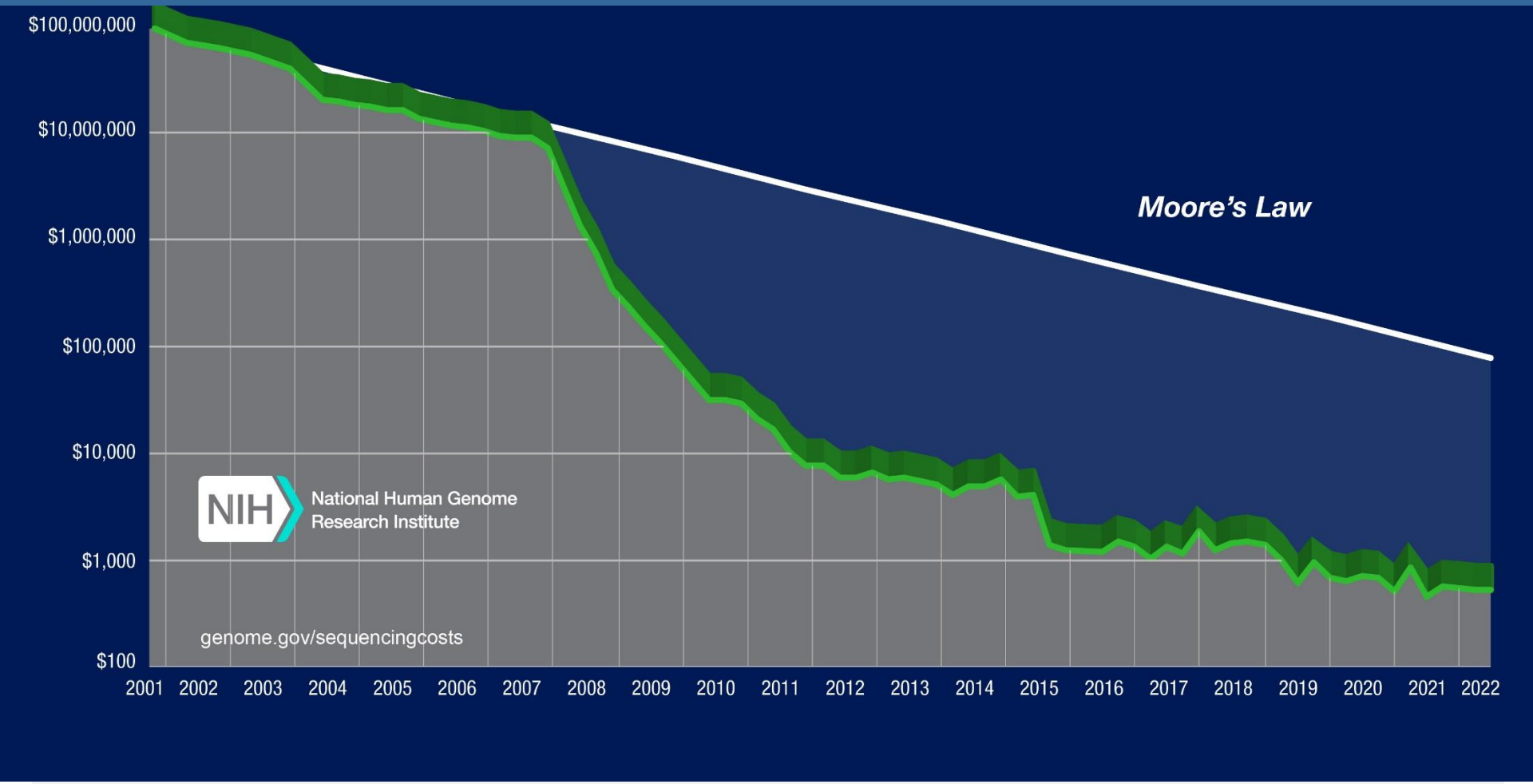


GC-bias plots for the human genome. Left: the GC composition distribution of the human genome (HG19, GRCh37). Center and right: GC-bias plots for several data sets from human NA12878. Unbiased coverage would be represented by a horizontal line at relative coverage = 1. Center: HiSeq v3 with sample-preparation reagents from Kapa Biosystems (Table 2, data set 14), Ion Torrent PGM (data set 15), and Complete Genomics data (data set 16). Right: HiSeq v3 with sample-preparation reagents from Kapa Biosystems (data set 14, as in center panel) and HiSeq v3 with the standard Fisher *et al.* [31] reagents (data set 13). Note that Illumina relative coverage exceeded the y-axis above 93% GC content. Relative coverage is only plotted for GC percentages for which there are at least 1,000 100-base windows in the genome.

Read https://doi.org/10.1146/annurev.biophys.30.1.1 to understand the chemistry likely to be underlying this phenomenon

Ross et al., Characterizing and measuring bias in sequence data, Genome Biology 14:R51, 2013

Cost per human genome



Speed per human genome

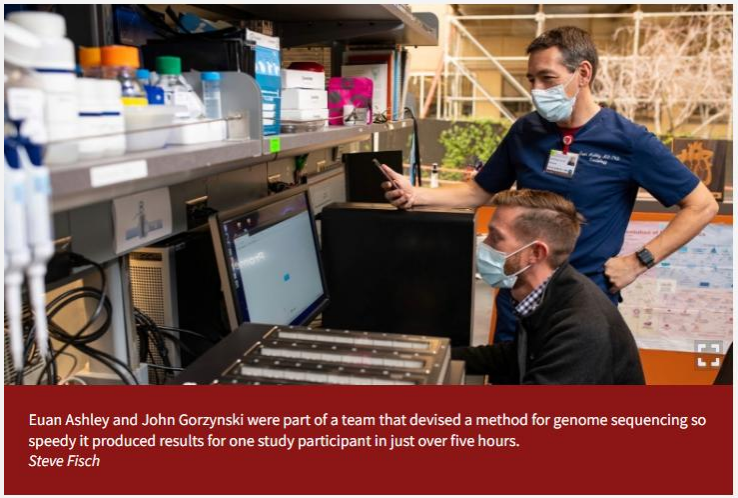
It took 13 years & $2.7 billion to sequence the 1st human genome

Now, a human genome can be sequenced in 1 day at less than $1000

Fastest DNA sequencing technique helps undiagnosed patients find answers in mere hours

A research effort led by Stanford scientists set the first Guinness World Record for the fastest DNA sequencing technique, which was used to sequence a human genome in just 5 hours and 2 minutes.

January 12, 2022 - By Hanae Armitage



Euan Ashley and John Gorzynski were part of a team that devised a method for genome sequencing so speedy it produced results for one study participant in just over five hours.
Steve Fisch

https://med.stanford.edu/news/all-news/2022/01/dna-sequencing-technique.html

Sequencing error rates & read lengths

1st-gen, e.g. Sanger
Error rate ~0.01%, read length 400 – 900 nt

2nd-gen, e.g. Illumina
Error rate ~0.1%, read length 150 – 300 nt

3rd-gen, e.g. PacBio & ONT,
Error rate ~10-15%, read length 5000 - 15000 nt

Base quality

Phred score is log of prob of incorrect base call, P

$Q_{Phred} = -10 \log_{10}(P)$

P is assigned by the sequencing machine used

Note that $P = 10^{-Q_{Phred}/10}$

Qphred	Error probability	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%

Exercise

$$\text{Phred score} = -10 \times \log_{10}(P)$$

Where P is the probability that the base call is incorrect. Typically, this probability is derived from the raw data generated during sequencing.

To compute the Phred score, follow these steps:

1. Determine the probability P that the base call is incorrect. This could be based on factors such as the intensity of the signal from the sequencing instrument, the quality of the sequencing chemistry, and other technical aspects of the sequencing process.
2. Take the negative base 10 logarithm of P.
3. Multiply the result by -10 to obtain the Phred score.

Here's an example:

If $P = 0.01$ (i.e., there's a 1% chance that the base call is incorrect),

$$\text{Phred score} = -10 \times \log_{10}(0.01) = -10 \times (-2) = 20$$

If you don't know the error rate, how would you derive or estimate the Phred score of a sequencing project?

Quality control in sequencing data

Base quality score (Phred score)

Read length distribution

- Examining the distribution of read lengths helps ensure consistency across the dataset. Deviations from the expected length may indicate issues with library preparation or sequencing.

GC content

- Analyzing the GC content distribution ensures that there are no biases that could affect downstream analyses. An uneven distribution may indicate biases in amplification during library preparation.

Adapter contamination

- Adapters are short DNA sequences used in library preparation. Detecting and removing adapter contamination is crucial to prevent artifacts and misinterpretations in downstream analyses.

Error rate

- Monitoring error rates, especially in low-complexity regions, helps identify potential sequencing or library preparation artifacts.

Additional quality control in sequencing data when there is a reference genome

Duplicate removal

- PCR amplification during library preparation can introduce duplicate reads. Identifying and removing duplicates is essential for accurate quantification and variant calling.

Coverage uniformity

- Assessing the evenness of coverage across the genome helps identify regions with low or high coverage, which can impact the reliability of variant detection and quantification.

Mapping quality

- Evaluating the mapping quality of reads to a reference genome helps ensure proper alignment. Low mapping quality may indicate issues such as contamination, misalignment, or the presence of repetitive elements.

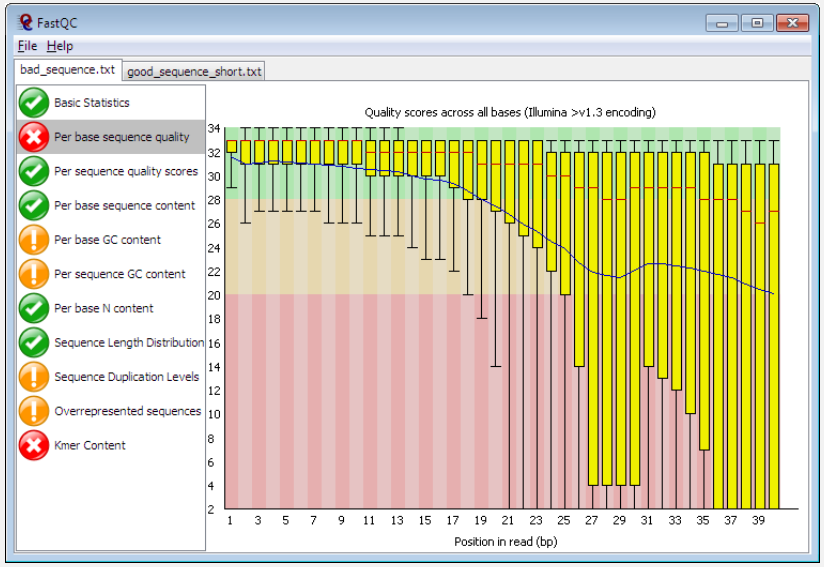
Ref genome consistency

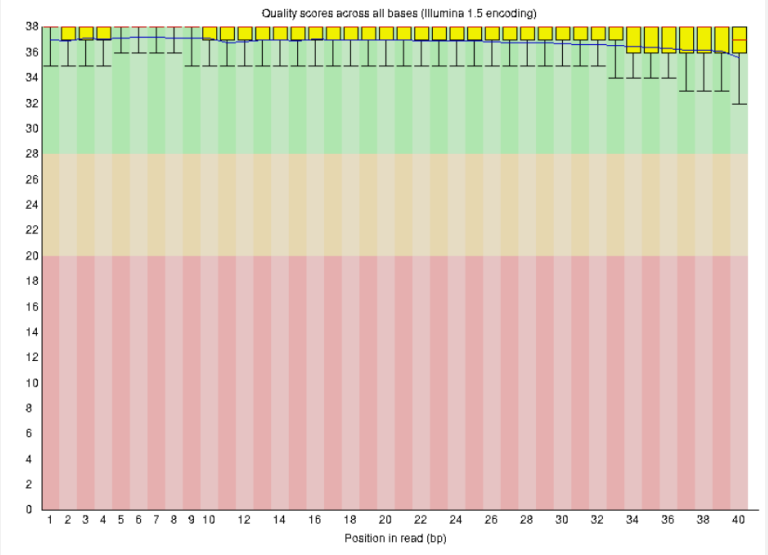
- Verifying that the sequencing data aligns well with the chosen reference genome is important to identify potential issues such as contamination or misidentification of the reference.

FastQC, a sequencing quality control tool

Get FastQC at

https://www.bioinformatics.babraham.ac.uk/projects/fastqc





Exercise

Suppose these reads are mapped to the same loci in a human genome

Is the green G likely an error?

Is the green T likely an error?

Is there likely an error in the red column?

Is the blue C likely an error?

```
ACGTAGGTACTG
ACGTACGTACTG
ACTTACGTACTG
ACGTACGTACTG
ACGTACGTACTG
ACCTACGTAGTG
ACCTACGTAGTG
ACCTACGTAGTG
ACCTACGTAGTG
ACCTACGTAGTG
ACCTACGTACTG
ACGTACGTACTG
```

Sequencing coverage

Coverage = # bases sequenced / Size of sequenced region

The value represents how many times, on average, each base in the target region has been sequenced

Good to read

Illumina sequencing technology

https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

FastQC

https://www.bioinformatics.babraham.ac.uk/projects/fastqc