CS4330: Combinatorial Methods in Bioinformatics

Primer on genome sequencing technologies

Wong Limsoon
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].
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=HMyCqWhwB8E

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/WneZp3fSJlk

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=CGWZvHli3i0
**GC-bias in sequencing data**

Relative coverage = \[
\frac{\text{coverage of a given reference base in a genome}}{\text{mean coverage of all reference bases}}.
\]

---

**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.
Cost per human genome

Moore’s Law

NIH National Human Genome Research Institute

genome.gov/sequencingcosts
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

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

*Log of prob of incorrect base call, $P$*

Base quality score, $Q$

*Confidence of a nucleotide base call, assigned by sequencing machine. Typically, $Q = 10^{-Q_{\text{Phred}}/10}$*

$Q_{\text{Phred}} = 20 \implies 1\%$ chance of incorrect base call

$Q_{\text{Phred}} = 30 \implies 0.1\%$ chance of incorrect base call

$Q_{\text{Phred}} = 40 \implies 0.01\%$ chance of incorrect base call
Quality control in sequencing data

Base quality score (Phred score)

Read length distribution

GC content

Adapter contamination

Error rate

- 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.

- 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.

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

- 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
- Coverage uniformity
- Mapping quality
- Ref genome consistency

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

* 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.

* 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.

* 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

Consider the reads shown in the box. 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?
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.
uencing.pdf

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