GENOMICS EDUCATION PARTNERSHIP

Long-Read Sequencing Technology

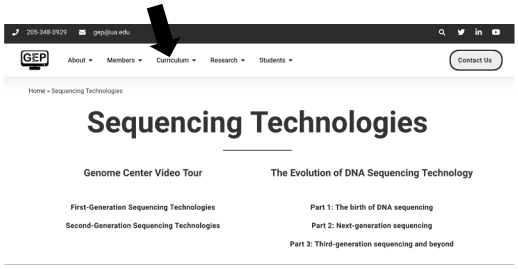
Katie M. Sandlin, M.S.

kmsandlin@ua.edu
@KatieMSandlin
linkedin.com/in/kmsandlin

June 20, 2024

1

Currently Available GEP Resources



thegep.org/sequencing/

2

Note to Instructors

- I have scripted the talk in the notes section of the PPT.
- Advanced versions of this presentation and lecture recordings were developed by Wilson Leung (Washington University in St. Louis).
  - Overview Third-generation Sequencing
    - Lecture Recording
  - Overview PacBio Iso-Seq
    - Lecture Recording

3

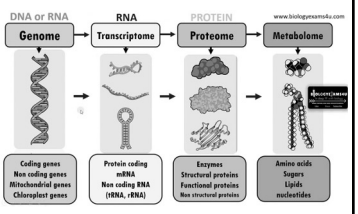
Agenda

1. Review of RNA-Seq Short-read Sequencing
2. Overview and benefits of Long-read Sequencing
3. Oxford Nanopore Sequencing
4. Pacific Biosciences Sequencing

4

RNA Sequencing (RNA-Seq)

- RNA sequencing (or RNA-seq) lets us investigate and discover the transcriptome–the total cellular content of RNAs including mRNA, rRNA and tRNA–by using next-generation sequencing to examine the quantity and sequences of RNA in a sample.
  - RNA-seq analyzes the transcriptome, indicating which of the genes encoded in our DNA are turned on or off and to what extent.



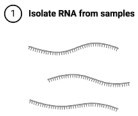
The **transcriptome** is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition.

Wang, Z., Gerstein, M. & Snyder, M. (2009).

BiologyExams4u.com

5

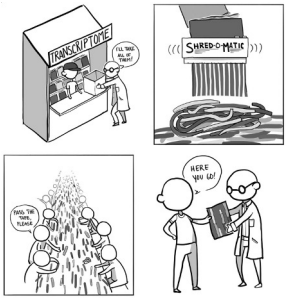
RNA-Seq: **reads are typically 30–300 bp**



Created with BioRender.com

6

Transcriptome reconstruction—akin to reassembling magazine articles after they have been through a paper shredder

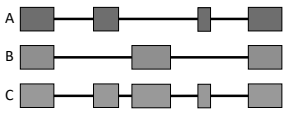


Conesa A (2013) Genomics the state of the art in RNA-seq analysis. *Nature Methods*

7

Short vs. Long Read RNA-Seq

Gene with Several Transcript Isoforms



Adapted from: Genetic Engineering & Biotechnology News

8

Agenda

1. Review of RNA-Seq Short-read Sequencing
2. Overview and benefits of Long-read Sequencing
3. Oxford Nanopore Sequencing
4. Pacific Biosciences Sequencing

9

Short vs. Long Read Sequencing

Sequencing instruments are categorized as long-read or short-read based on their underlying chemistry and the length of DNA fragments they analyze.

PacBio HiFi Sequencing - Unlock Your Next Great Discovery

10

Benefits of assembling a genome with long-read sequencing: an example using books

- A way to think about the difference between assembling a genome with short-reads versus long-read sequencing technology would be to imagine two different approaches to reconstructing a 368-page novel from randomized snippets of text.

Pacific Biosciences

11

Benefits of assembling a genome with long-read sequencing: an example using books

Those looking for serious land

the North Carolina coast

The shack sat back

hatchet and pack a buck

by a torn shoreline

Excerpts from *Where the Crawdads Sing*

12

Benefits of assembling a genome with long-read sequencing: an example using books

The marsh was guarded by a torn shoreline, labeled by early explorers as the "Graveyard of the Atlantic" because riptides, furious winds, and shallow shoals wrecked ships like paper hats along what would become the North Carolina coast. One seaman's journal read, "rang'd along the Shoar . . . but could discern no Entrance . . . A violent Storm overtook us . . . we were forced to get off to Sea, to secure Ourselves and Ship, and were driven by the Rapidity of a strong Current . . .

"The Land . . . being marshy and Swamps, we return'd towards our Ship . . . Discouragement of all such as should hereafter come into those Parts to settle."

Claiming territory hadn't changed much since the 1500s. The scattered marsh holdings weren't legally described, just staked out natural—a creek boundary here, a dead oak there—by renegades. A man doesn't set up a palmetto lean-to in a bog unless he's on the run from somebody or at the end of his own road.

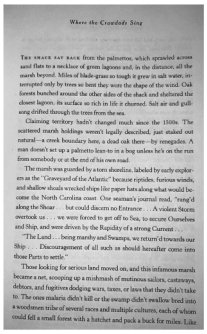
Where the Crawdads Sing

THE SHACK SAT BACK from the palmettos, which sprawled across sand flats to a necklace of green lagoons and, in the distance, all the marsh beyond. Miles of blade-grass so tough it grew in salt water, interrupted only by trees so bent they wore the shape of the wind. Oak forests bunched around the other sides of the shack and sheltered the closest lagoon, its surface so rich in life it churned. Salt air and gull-song drifted through the trees from the sea.

Those looking for serious land moved on, and this infamous marsh became a net, scooping up a mishmash of mutinous sailors, castaways, debtors, and fugitives dodging wars, taxes, or laws that they didn't take to. The ones malaria didn't kill or the swamp didn't swallow bred into a woodsmen tribe of several races and multiple cultures, each of whom could fell a small forest with a hatchet and pack a buck for miles. Like

13

Benefits of assembling a genome with long-read sequencing: an example using books


Where the Crawdads Sing

THE SHACK SAT BACK from the palmettos, which sprawled across sand flats to a necklace of green lagoons and, in the distance, all the marsh beyond. Miles of blade-grass so tough it grew in salt water, interrupted only by trees so bent they wore the shape of the wind. Oak forests bunched around the other sides of the shack and sheltered the closest lagoon, its surface so rich in life it churned. Salt air and gull-song drifted through the trees from the sea.

Claiming territory hadn't changed much since the 1500s. The scattered marsh holdings weren't legally described, just staked out natural—a creek boundary here, a dead oak there—by renegades. A man doesn't set up a palmetto lean-to in a bog unless he's on the run from somebody or at the end of his own road.

The marsh was guarded by a torn shoreline, labeled by early explorers as the "Graveyard of the Atlantic" because riptides, furious winds, and shallow shoals wrecked ships like paper hats along what would become the North Carolina coast. One seaman's journal read, "rang'd along the Shoar . . . but could discern no Entrance . . . A violent Storm overtook us . . . we were forced to get off to Sea, to secure Ourselves and Ship, and were driven by the Rapidity of a strong Current . . .

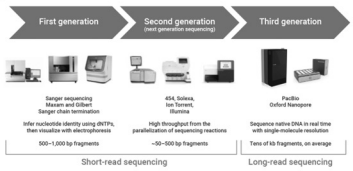
"The Land . . . being marshy and Swamps, we return'd towards our Ship . . . Discouragement of all such as should hereafter come into those Parts to settle."

Those looking for serious land moved on, and this infamous marsh became a net, scooping up a mishmash of mutinous sailors, castaways, debtors, and fugitives dodging wars, taxes, or laws that they didn't take to. The ones malaria didn't kill or the swamp didn't swallow bred into a woodsmen tribe of several races and multiple cultures, each of whom could fell a small forest with a hatchet and pack a buck for miles. Like

14

3rd Generation (Long-Read) Sequencing Emerges

- To address the limitations of next-generation sequencing, scientists began developing real-time, single-molecule DNA sequencing platforms.
- In the mid 2000s, two single-molecule sequencing technologies emerged giving rise to 3rd generation sequencing:
  - single-molecule real-time (SMRT) sequencing by Pacific Biosciences
  - nanopore sequencing by Oxford Nanopore Technologies



15

Third-generation Sequencing Technologies

- Sequencing platforms from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT)
- Key characteristics:
  - **Single-molecule**, **real-time** sequencing
    - Sequence samples directly without amplification
  - Can detect epigenetic modifications
    - DNA methylation
- Comparisons with Illumina:
  - Produce **longer reads**
  - Raw reads have **higher error rates** than Illumina

16

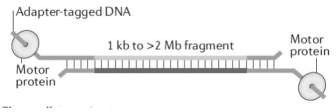
Agenda

1. Review of RNA-Seq Short-read Sequencing
2. Overview and benefits of Long-read Sequencing
3. Oxford Nanopore Sequencing
4. Pacific Biosciences Sequencing

17

ONT Sequencing – Template Topology

- Yellow – forward DNA strand
- Dark Blue – reverse DNA strand
- Light Blue – sequencing adapters

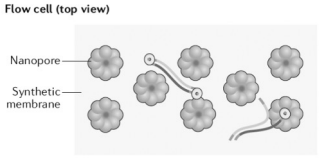


Logsdon, G.A., Vollger, M.R. & Eichler, E.E. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**, 597–614 (2020).

18

ONT Sequencing – Flow Cell

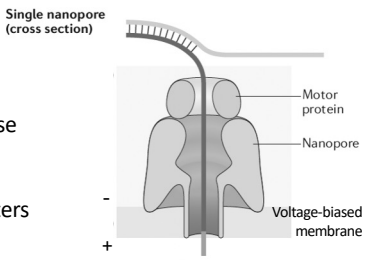
- Yellow – forward DNA strand
- Dark Blue – reverse DNA strand
- Light Blue – sequencing adapters



19

ONT Sequencing - Nanopore

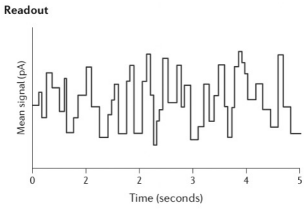
- Yellow – forward DNA strand
- Dark Blue – reverse DNA strand
- Light Blue – sequencing adapters



20

ONT Sequencing – Base Calling

- You can think of sequencing DNA or RNA like reading music — the bases being the notes, which when played in the correct order, produces a recognizable song.

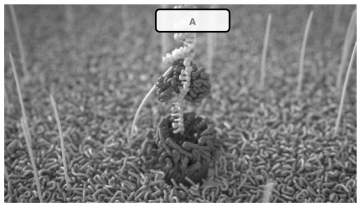


21

Check For Understanding

Raise your fingers (1-4) to share your answer.

1. Synthetic Membrane
2. Motor Protein
3. DNA or RNA
4. Nanopore

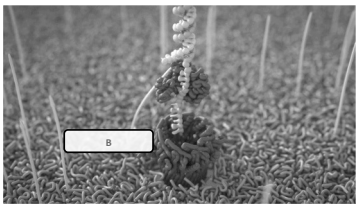


22

Check For Understanding

Raise your fingers (1-4) to share your answer.

1. Synthetic Membrane
2. Motor Protein
3. DNA or RNA
4. Nanopore

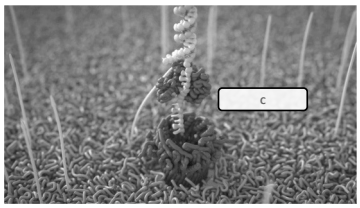


23

Check For Understanding

Raise your fingers (1-4) to share your answer.

1. Synthetic Membrane
2. Motor Protein
3. DNA or RNA
4. Nanopore

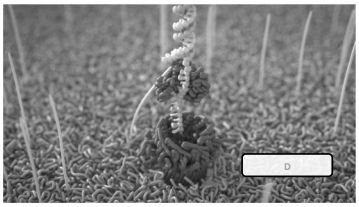


24

Check For Understanding

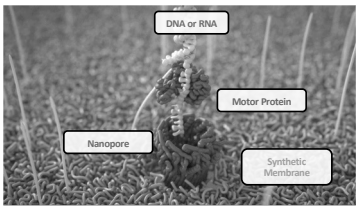
Raise your fingers (1-4) to share your answer.

1. Synthetic Membrane
2. Motor Protein
3. DNA or RNA
4. Nanopore



25

Check For Understanding

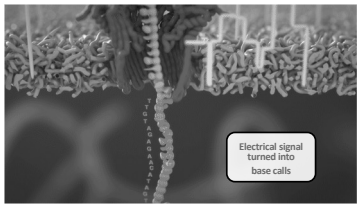


- Read length is up to 4 Mb
  - Long reads sequencing mode: 10–100kb
  - Ultra-long sequencing mode: 100–300kb

26

Review – Base Calling

- This digital method of reading DNA or RNA sequence has multiple benefits:
  - Real-time analysis
  - PCR free, no amplification bias
  - **Modified base detection**
  - Read-length agnostic
  - **Direct sequencing of DNA or RNA**



Disclaimer: The real-time base caller has lower accuracy. For many applications, it might be beneficial to do the base calling afterwards with a more powerful GPU and a more compute intensive algorithm [i.e., duplex with super accuracy (SUP)].

27

Flow Cell
Where Nanopores are Housed

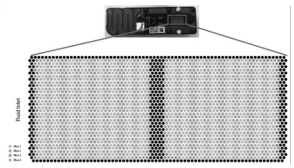
The **Flongle** Flow Cell can generate up to 2.8 Gb of data enabling direct, real-time DNA & cDNA sequencing on smaller, single-use flow cells.

The **MinION** Flow Cell can generate up to 50 Gb of data for sequencing DNA, cDNA or native RNA in real-time.

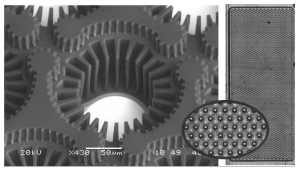
The **PromethION** Flow Cell can generate up to 290 Gb for sequencing DNA, cDNA or native RNA in real-time.

28

MinION Flow Cell



MinION Mk1B and Mk1C/GridION Flow Cell Channel Layout



SEM image of a single sensor well (left) and the hexagonal array of wells (right).

29

Nanopore Sequencing Devices

$1,999

30

ONT Sequencing Workflow

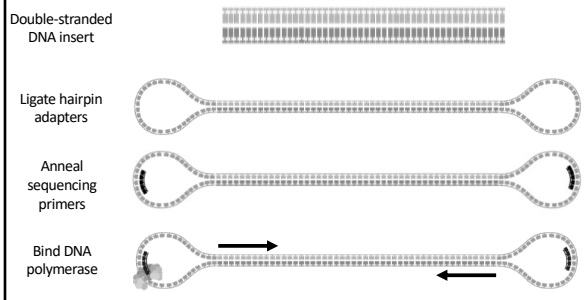
31

Agenda

1. Review of RNA-Seq Short-read Sequencing
2. Overview and benefits of Long-read Sequencing
3. Oxford Nanopore Sequencing
4. Pacific Biosciences Sequencing

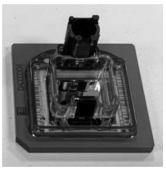
32

PacBio SMRT Sequencing – SMRTbell Template

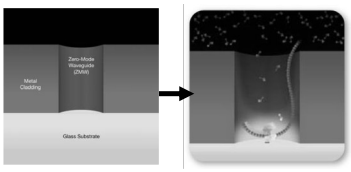


33

Key components of SMRT Sequencing



8M SMRT Cell

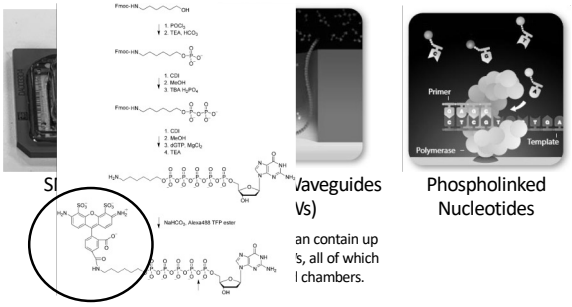


Zero-Mode Waveguides (ZMWs)

Each SMRT Cell can contain up to 8 million ZMWs, all of which are liquid-filled chambers.

34

Key components of SMRT Sequencing



Sl

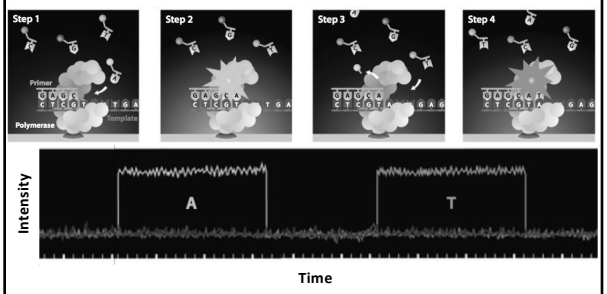
Waveguides (Vs)

an contain up
's, all of which
l chambers.

Phospholinked Nucleotides

35

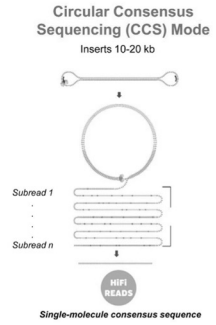
Phospholinked Nucleotide Incorporation



1 DNA molecule and 1 polymerase in each well (ZMW)
4 colors flash in real-time as polymerase acts

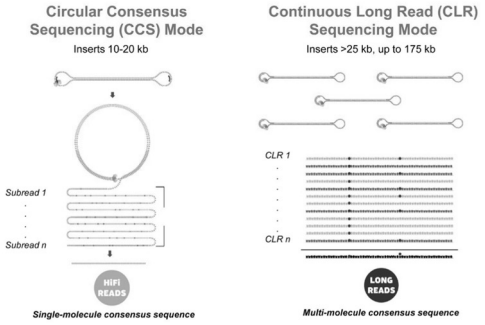
36

2 Modes of SMRT Sequencing



37

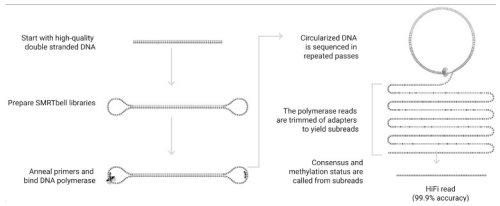
2 Modes of SMRT Sequencing



38

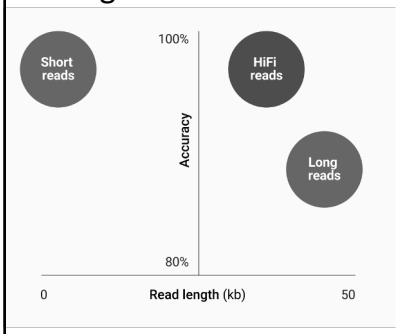
How are HiFi reads generated?

- SMRT sequencing can sequence the same DNA molecule **multiple times** generating highly accurate long reads, or **HiFi reads**.
- HiFi reads are produced using **circular consensus sequencing (CCS)** mode on PacBio long-read systems. HiFi reads provide base-level resolution with 99.9% single-molecule read accuracy.



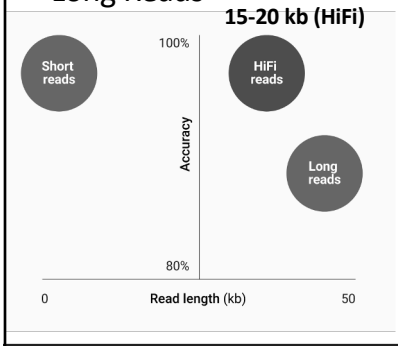
39

HiFi Reads vs. Long Reads



40

HiFi Reads vs. Long Reads



Disclaimer: PacBio HiFi reads typically have lower quality than Illumina data. They also have systematic errors (particularly with homopolymers and specific types of repetitive sequences; see Mc Cartney et al., 2022).

PacBio HiFi reads are defined as reads with an accuracy of Q20 or above (error rate of 1/100; 99% accuracy). For the Revio system, the mode for the read quality is ~Q30 (99.9% accuracy; see the Revio system brochure).

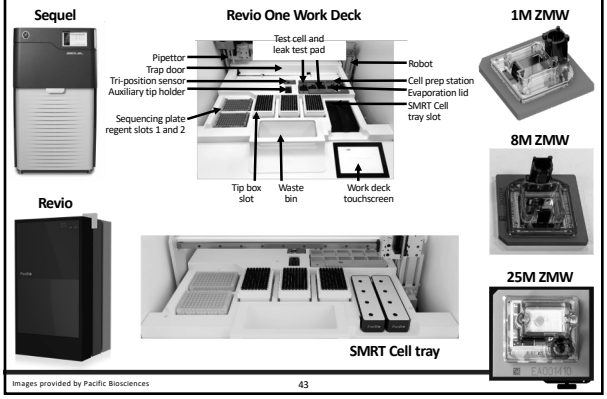
For the Illumina XLEAP-SBS chemistry, the read quality is Q40 (99.99% accuracy; see the "XLEAP-SBS chemistry enables Q40 and above data quality on NovaSeq X and NextSeq 1000/2000" blog post).

41

PacBio Sequencing – How it Works

42

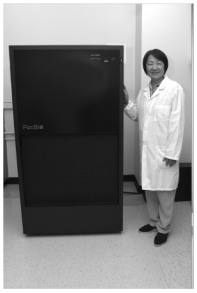
Overview of PacBio Sequencers



Images provided by Pacific Biosciences

43

PacBio Revio Sequencing System



$779,000

Weill Cornell Medicine

44

PacBio SMRT Sequencing Workflow

Although each sequencing project is unique, there are three main steps to go from DNA to discovery with SMRT sequencing.

Library prep	SMRT sequencing	Data analysis
Learn more about kits for fast and easy library preparation	Explore sequencing interactive	SMRT Link user interface and SMRT Analysis

Pacific Biosciences – Steps of SMRT Sequencing

45

Isoform Sequencing (Iso-Seq)

- Using SMRT sequencing to do RNA-Sequencing = marketing term is Iso-Seq
- Iso-Seq data is obtained via Circular Consensus Sequencing (CCS)

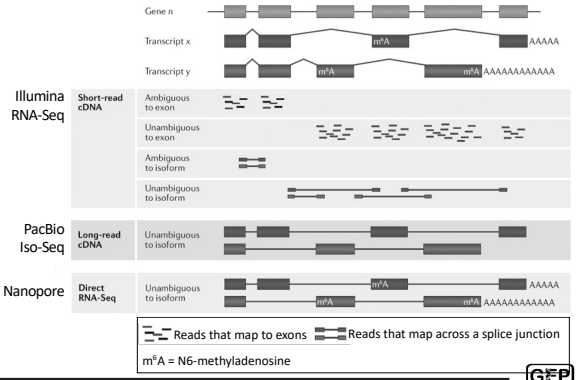
Iso-Seq

Scalable
De Novo
Isoform Discovery
from PacBio HiFi Reads

GitHub

46

Benefits of Long-read RNA-Seq



Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, *20*(11), 631-656.

Stark et al., 2019

47

Projects typically combine multiple sequencing technologies to construct genome assemblies

- The Vertebrate Genomes Project (VGP) aims to construct reference genome assemblies for ~70,000 vertebrate species.
- Sequencing technologies used by the Vertebrate Genome Lab (VGL) at the Rockefeller University:

PacBio Sequel IIe	ONT Promethion	Bionano Saphyr	Arima Hi-C
Use HiFi data to construct the initial assembly	Use ultra-long Nanopore reads to resolve highly repetitive regions	Use optical maps to determine order and orientation of the contigs (scaffolding)	Second round of scaffolding to generate chromosome-scale assemblies

- Illumina data also used to polish the consensus sequence

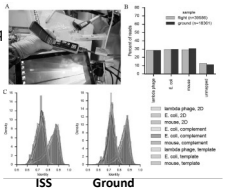
WSL Technologies

48

Applications of 3rd Generation Sequencing

- Chromosome-scale genome assemblies
- Cancer genomics, human disease research
- Population genetics (SNVs, phasing)
- Epigenomics (DNA methylation)
- Full-length RNA-Seq
  - PacBio Iso-Seq, Nanopore direct RNA-Seq
- Single-cell RNA-Seq
  - PacBio Kinnex (MAS-Seq)
- Conservation genomics
- Complete plasmid verification
  - Plasmidsaurus, Eurofins, …
- Metagenomics, microbial genomics
- … … …

Astronaut Kate Rubins on the ISS



Wallace. 2019

49

Conclusion

- Sequencing technology is the 'microscope' by which geneticists study genetic variation, and it is clear that long-read technologies have provided us with a new 'lens and objective' for understanding DNA and RNA variation, structure and organization.
- Although the two predominant long-read technologies are competitive, some of the best results have been obtained when the sequencing platforms are used to complement one another.

Logsdon, G.A., Vollger, M.R. & Eichler, E.E. Long-read human genome sequencing and its applications. *Nat Rev Genet* 21, 597–614 (2020).

50

Copy Google Doc Worksheet

	Oxford Nanopore Technologies (ONT)	Pacific Biosciences (PacBio)
Sequencing chemistry/ technology name		
Read length		
What is detected in the sequencing process?		
How is that detected?		
Enzyme used in the sequencing process		

thegep.org/lrs

Adapted from: Bowling, et al. Bringing Next-Generation Sequencing into the Classroom through a Comparison of Molecular Biology Techniques. *The American Biology Teacher*; 76 (6): 396–401.

51

Answer Key

	Oxford Nanopore Technologies (ONT)	Pacific Biosciences (PacBio)
Sequencing chemistry/ technology name	Nanopore Sequencing	Single Molecule, Real-Time (SMRT)
Read length	Up to 4 Mb; Long reads sequencing mode: 10–100kb; Ultra-long sequencing mode: 100–300kb	15-20 kb (HiFi)
What is detected in the sequencing process?	Changes in current as nucleotides pass through the pore	Fluorophore associated with the phospholinked nucleotides emit different color light
How is that detected?	Different collection of nucleotides that occupy a nanopore will cause different changes in the electric current. Machine learning algorithms (neural networks) are used to analyze the continuous changes to the electric current (squiggle) to infer the sequences that pass through the nanopore.	Each phospholinked nucleotide (A, C, G, T) is labeled with a different fluorescent dye. When a nucleotide is incorporated by the polymerase, it emits different color light that corresponds to the fluorescent dye for the nucleotide.
Enzyme used in the sequencing process	Motor protein (helicase)	DNA polymerase

Adapted from: Bowling, et al. Bringing Next-Generation Sequencing into the Classroom through a Comparison of Molecular Biology Techniques. *The American Biology Teacher*; 76 (6): 396–401.

52

Acknowledgements

- Special thanks to Wilson Leung (Washington University in St. Louis) for his suggestions and feedback.
- Thank you to Jason Williams and Anna Feitzinger at the CSHL DNA Learning Center for leading a Nanopore Sequencing Workshop.
- This presentation is based upon work supported by the National Science Foundation under Grant No. 1915544 and National Institute of General Medical Sciences of the National Institutes of Health under award number R25GM130517 to LKR.

53

Questions?

Eichler

54