Biochem 3BP3

Advances in DNA Sequencing

Week of Nov 22, 2021

Sequencing Trade-Offs

- Two sequencing paradigms are currently prevalent
    - Sanger DNA sequencing (low volume) 500-1200 bp via sequencing by synthesis and fragment migration technologies (gel or capillary)
    - Illumina DNA sequencing (high volume) 250 bp via massively parallel sequencing by synthesis and optical technologies (i.e. labeled bases)
- Choice of technology involves trade-offs:
    - Length of sequencing reads
    - Ability to generate mate-pair reads
    - Volume of sequencing reads
    - Speed of DNA sequencing
    - Cost of DNA sequencing
    - Quality of sequencing reads

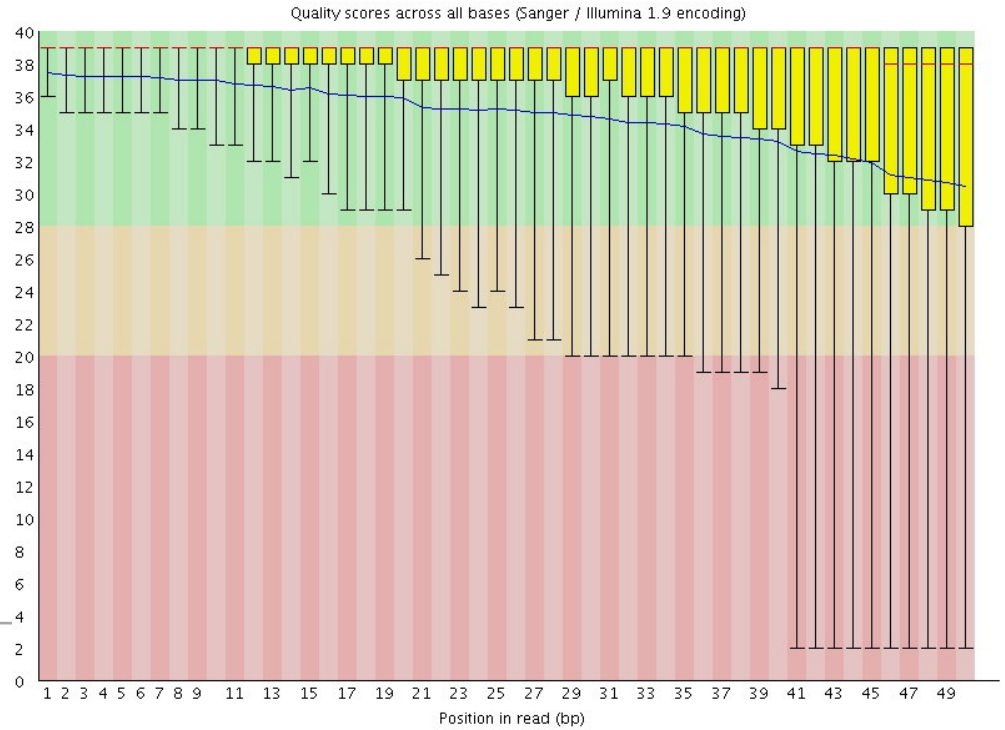
Sequencing Trade-Offs

- PHRED quality is a very important concept
- Length of read and quality of read are inter-dependent in both Sanger and Illumina sequencing approaches – can this relationship be decoupled?

Phred qualities

Quality value	Chance it is wrong	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

- $Q = -10 \log_{10} P \quad <=> \quad P = 10^{-Q/10}$
  - Q = Phred quality score
  - P = probability of base call being incorrect

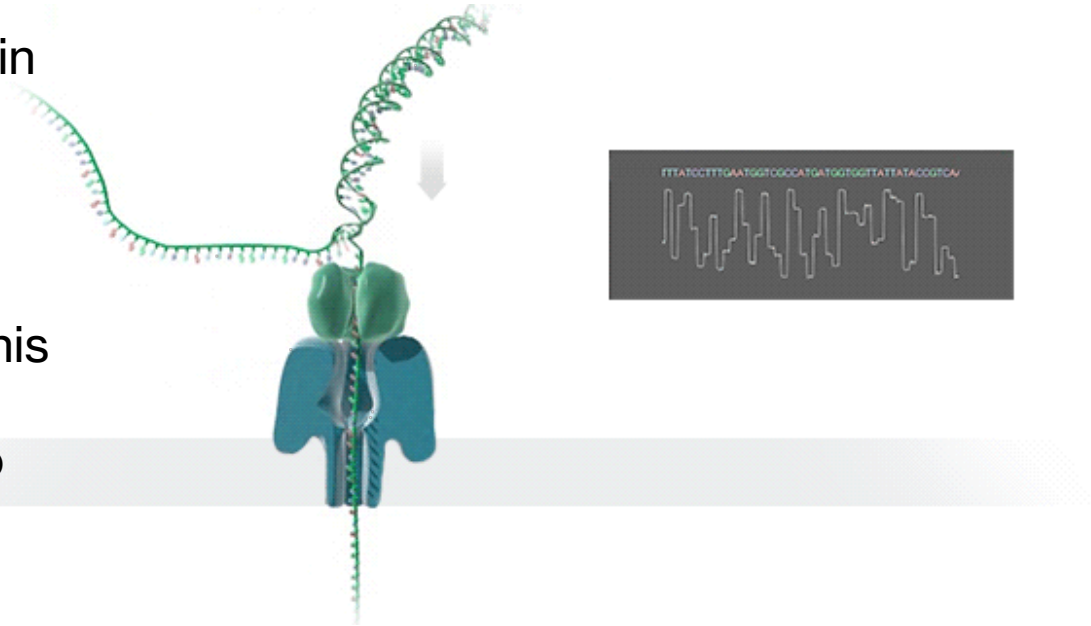


Sequencing Trade-Offs

- Illumina is the dominant technology currently
  - Short reads, but good quality
  - Massively parallel and thus high volume
  - Mate paired
  - Confidence obtained by high coverage
- But Sanger sequencing still considered the gold standard for verification of results, e.g. confirmation of SNPs discovered by exome sequencing
- Illumina short read sequencing is not the optimal data for every question
  - It is perfect for 16S microbiome profiling
  - It makes genome or transcript assembly very difficult and very noisy
- A $1000 human genome is within reach
- Exome sequencing of clinical samples is already affordable
- But fast, real-time clinical sequencing is not yet possible – speed and infrastructure remain issues

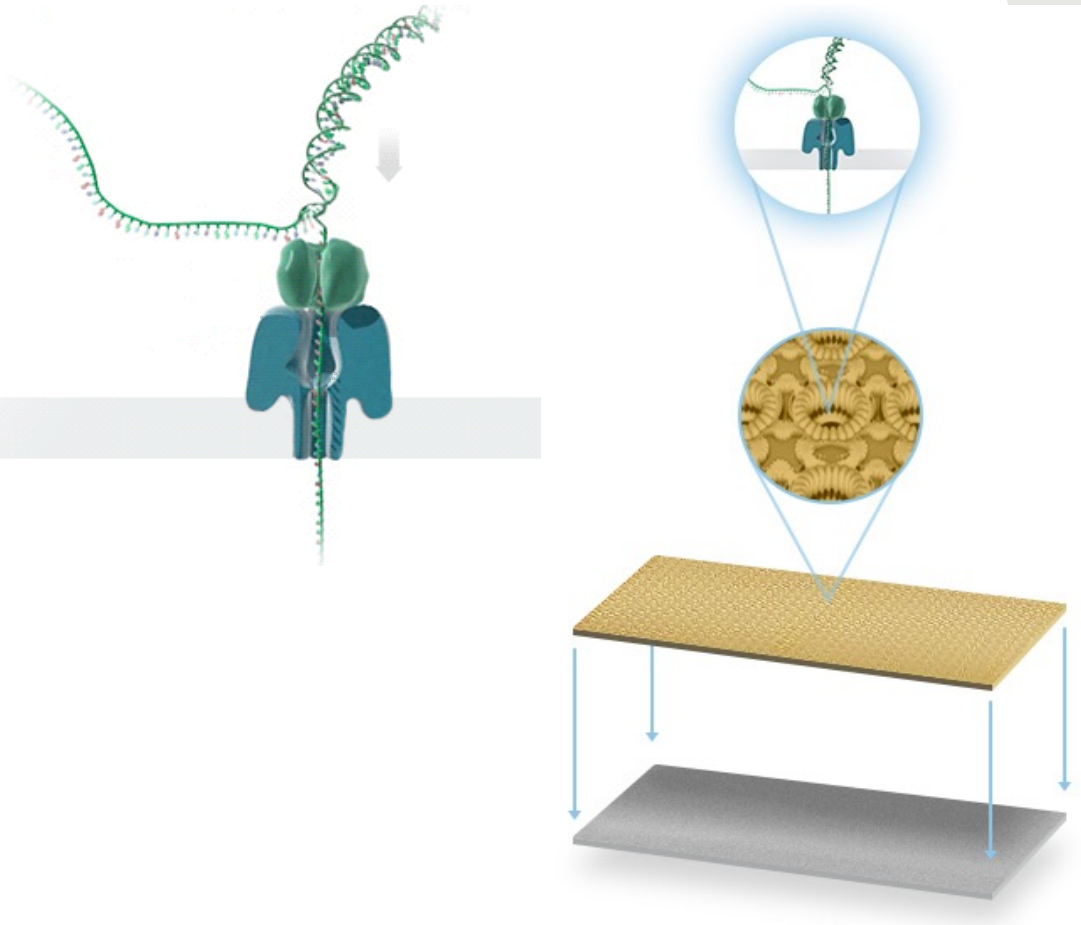
What's Next – Nanopore MinION

- A protein nanopore is set in an electrically resistant polymer membrane
- An ionic current is passed through the nanopore by setting a voltage across this membrane
- Change in current used to identify the molecule passing through



What's Next – Nanopore MinION

- Changes in current can identify bases but also modified bases
- Nanopores arrayed on a single chip using microscaffolds and microelectrodes for each nanopore
- Each nanopore sequences a single DNA molecule from start to finish – no fixed read length! (longest reported is 200 kb)



What's Next – Nanopore MinION

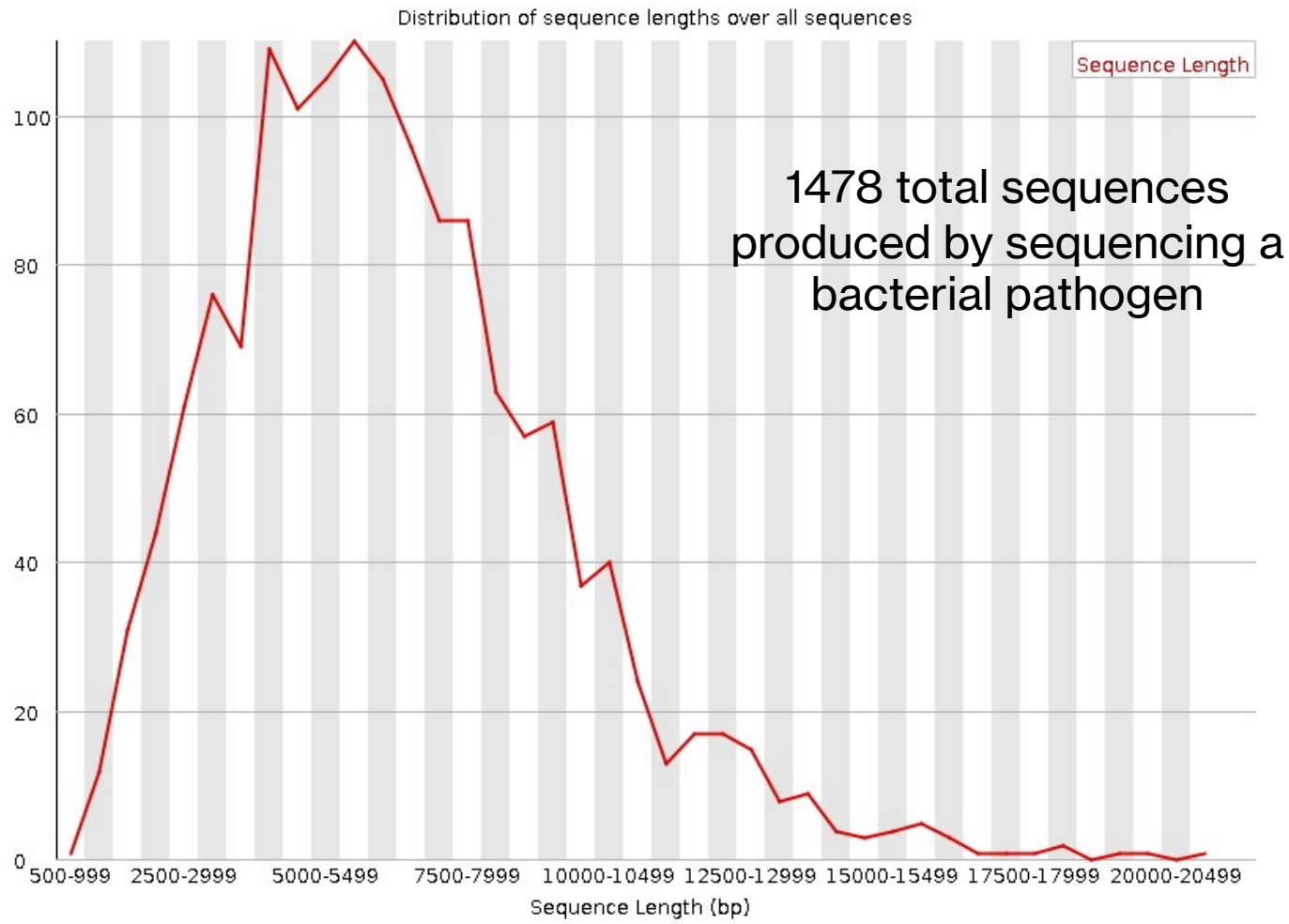
- Flowcell and overall device very small and portable
- Easy to use – no PCR step involved, submit simple DNA extracts

- DNA strand moves rapidly at the rate of 1-5 μs per base through the nanopore
- Processing of fast signal not yet mature – higher DNA sequencing error rate than other technologies - per base accuracy of the MinION has been reported as 65 - 80%

- Long MinION reads combined with higher quality Illumina read coverage a boon for genome assembly

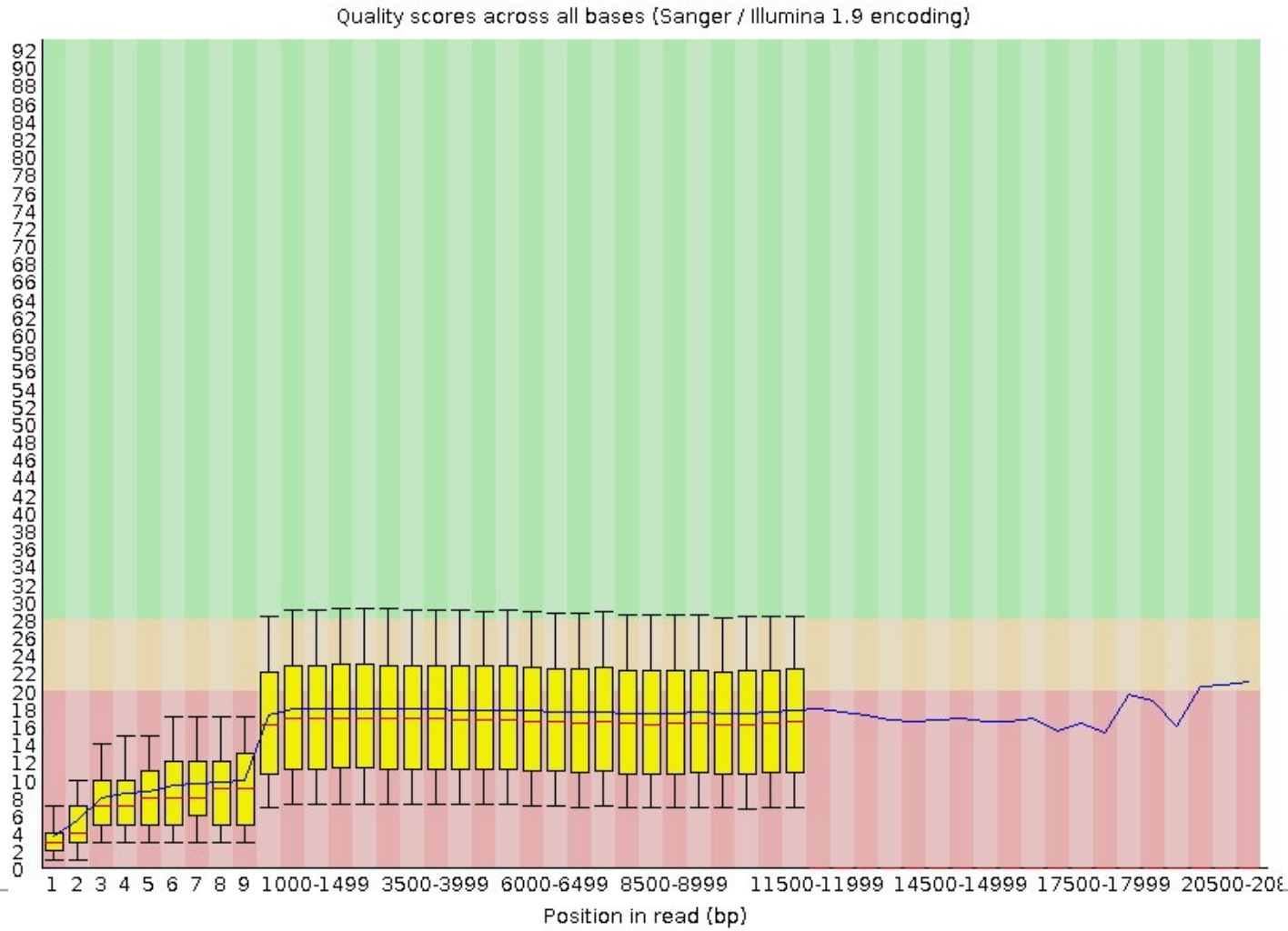
- Is the standard technology for SARS-CoV-2 sequencing and outbreak surveillance

Nanopore Sequence Lengths



1478 total sequences produced by sequencing a bacterial pathogen

Nanopore Sequence Quality



Nanopore Sequencing - Implications

- minION produces 25-42 Gb in 48 hours; proposed PromethION will produce 7-12 Tb in 48 hours (can you hear Kryder weeping?)

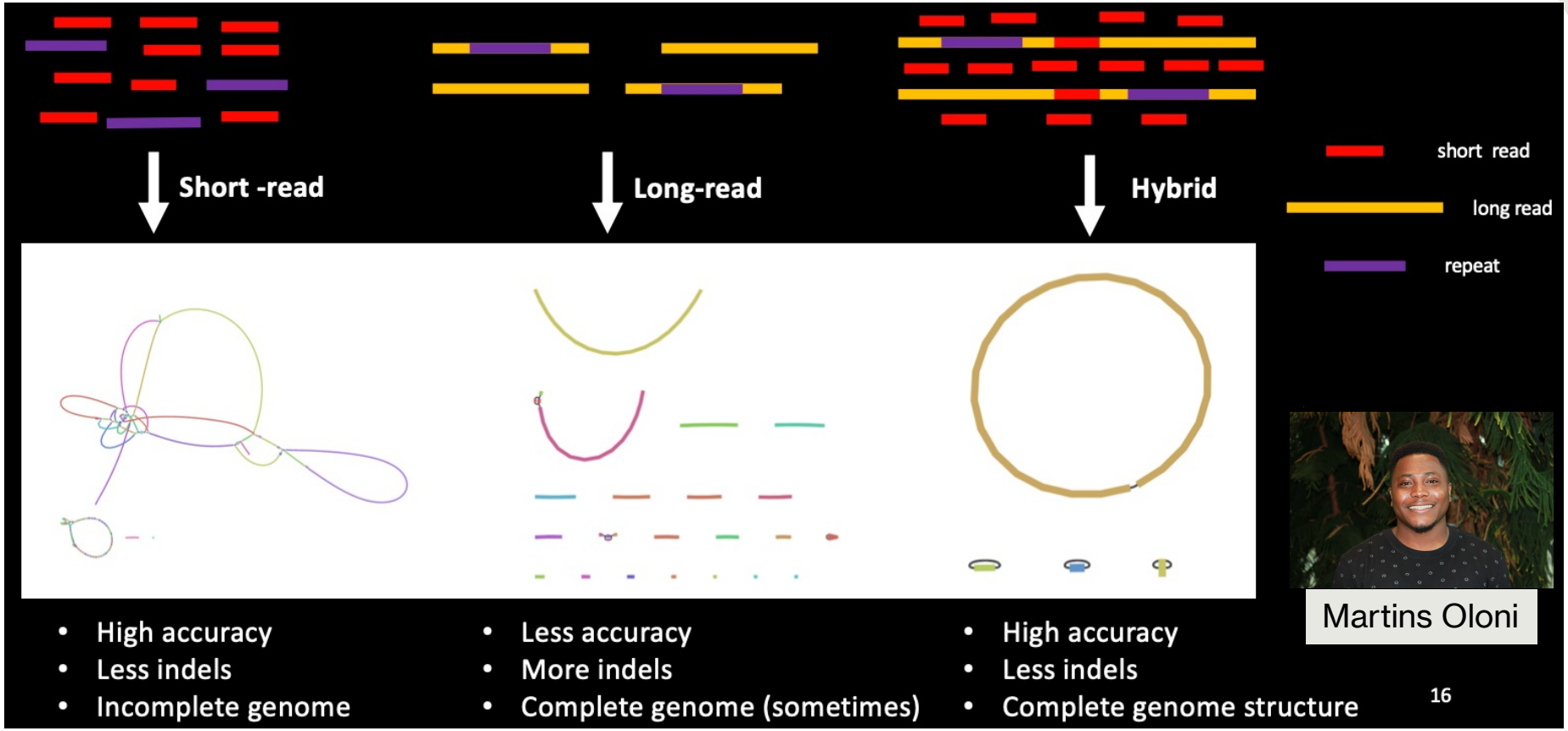
- Read length and PHRED scores decoupled. Can the PHRED performance be improved?

- Complete single pass genome or chromosome sequencing may be possible
  - Assembly algorithms may not be needed
  - No PCR or sequencing by synthesis – few consumables – cheaper!
  - Fast technology – clinical application

- Density of nanopores may allow sequencing of complete microbiomes
  - Bacteria & Viruses
  - Both pathogen identification and gene content

- PromethION sequencing of an entire transcriptome?

Nanopore Sequencing - Implications



Martins Oloni

Nanopore Sequencing – Miniaturization!

SmidgION

Increasing Use of GPU Technologies

CPU– Central Processing Unit

- “Swiss Army Knife” of computing - can run a variety of programs quickly
- Ubiquitous – your computer, your fridge, your car, your phone, etc.
- Computes in series, i.e. one calculation after another

GPU – Graphic Processing Unit

- “Video card” for display; driven by the gaming industry
- Run one kind of computation simultaneously, very quickly
- Can be re-purposed for bioinformatics

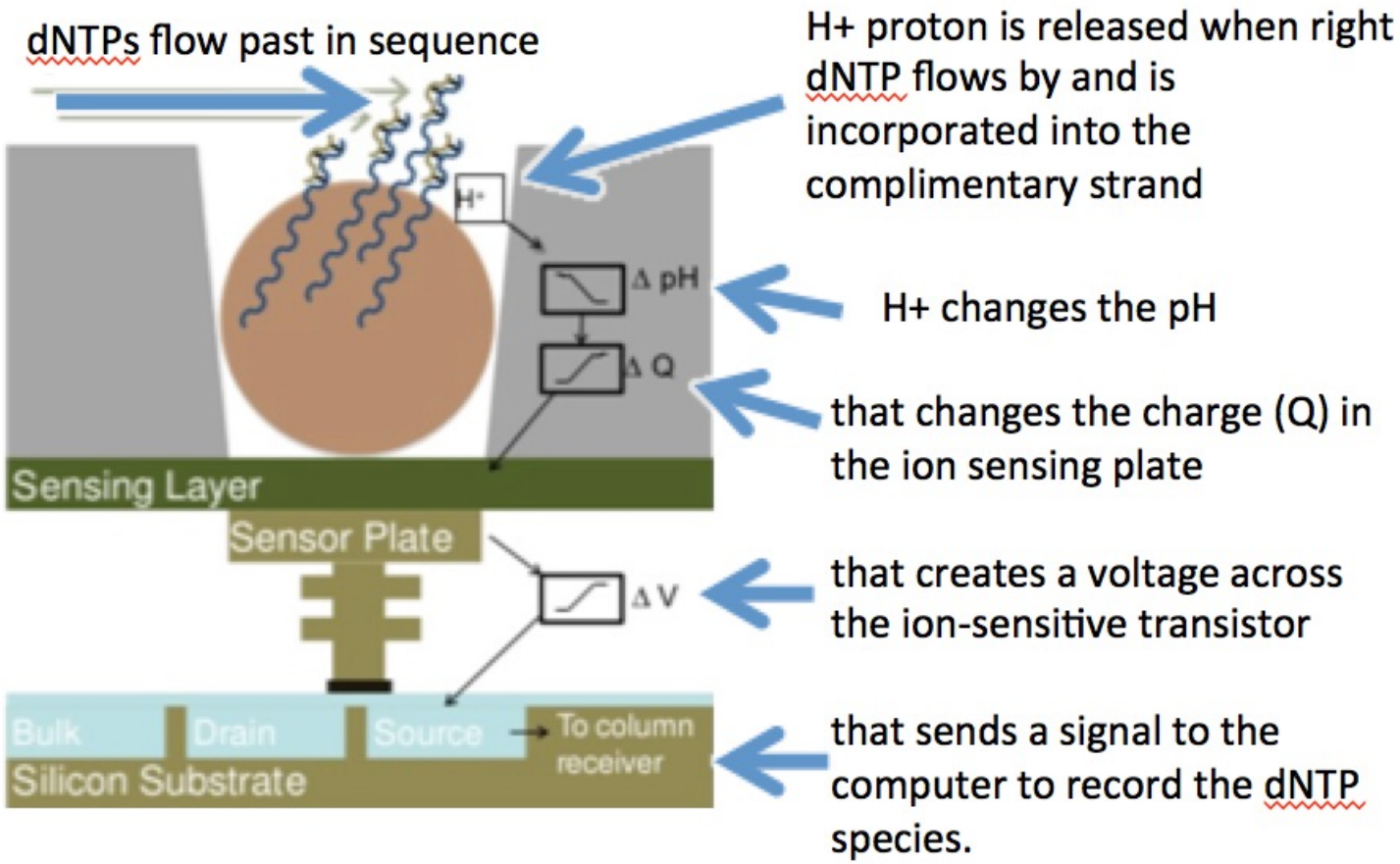
CPU versus GPU

- GPU does not do well for sequence similarity (i.e. BLAST) or read mapping
- GPU does very well for signal processing – Nanopore reads!
- GPU be cost effective for large projects, but custom software or re-programming required

A silly demonstration: https://www.youtube.com/watch?v=-P28LKWTzrI

What's Next – Ion Torrent

- Semi-conductor DNA sequencing instead of Illumina optical detection
- Still involves sequencing via synthesis so unlikely to be as affordable as nanopore technologies



What's Next – Ion Torrent

- Volume and data are similar to Illumina but…
  - Mate pairs not possible
  - PHRED scores not as high
- Lower cost per base and faster run time but technology is analogous to Illumina – massively parallel short reads

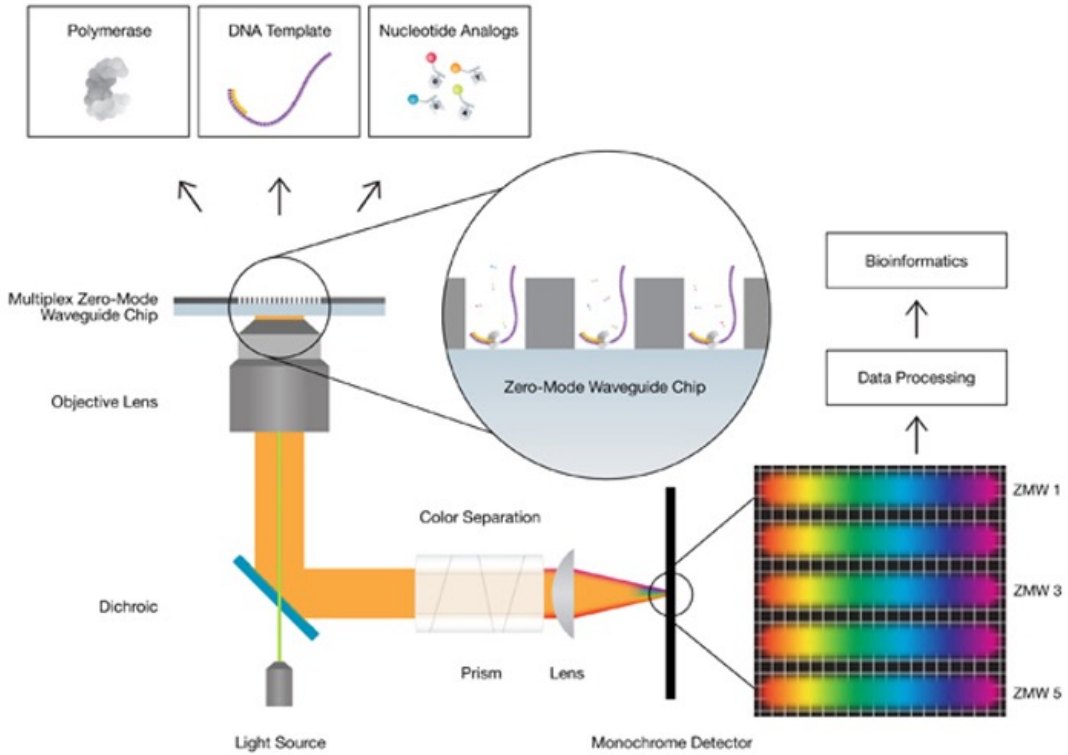
- The need for massively parallel short read technologies is not yet going away, particularly for large complicated RNA/DNA samples
- Which technology will persist – optical detection or semi-conductor sequencing?

Ion Torrent vs. Illumina Sequencing

	Ion Torrent Proton Sequencer	Illumina Hiseq 1500
Cost / Bp	MEDIUM	HIGH
Base Quality	~25	~35
Read Length	250	125
Paired End Sequencing?	NO	YES
Run Time	12 hours	40 hours

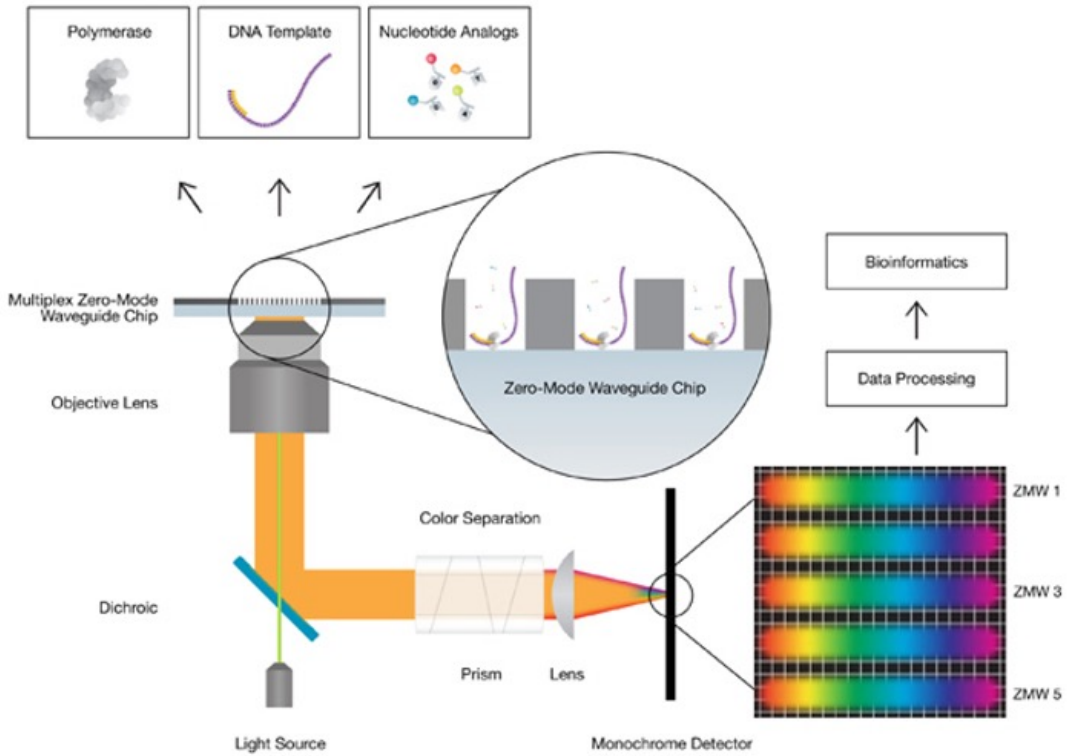
What's Next – PacBio

- Single Molecule, Real-Time (SMRT) technology - claimed long reads, uniform coverage, and high consensus accuracy
- Also a sequence by synthesis method with phospholinked nucleotides but has a novel sensor: zero-mode waveguides (ZMWs)
- A single DNA polymerase enzyme is affixed at the bottom of a ZMW with a single molecule of DNA as a template



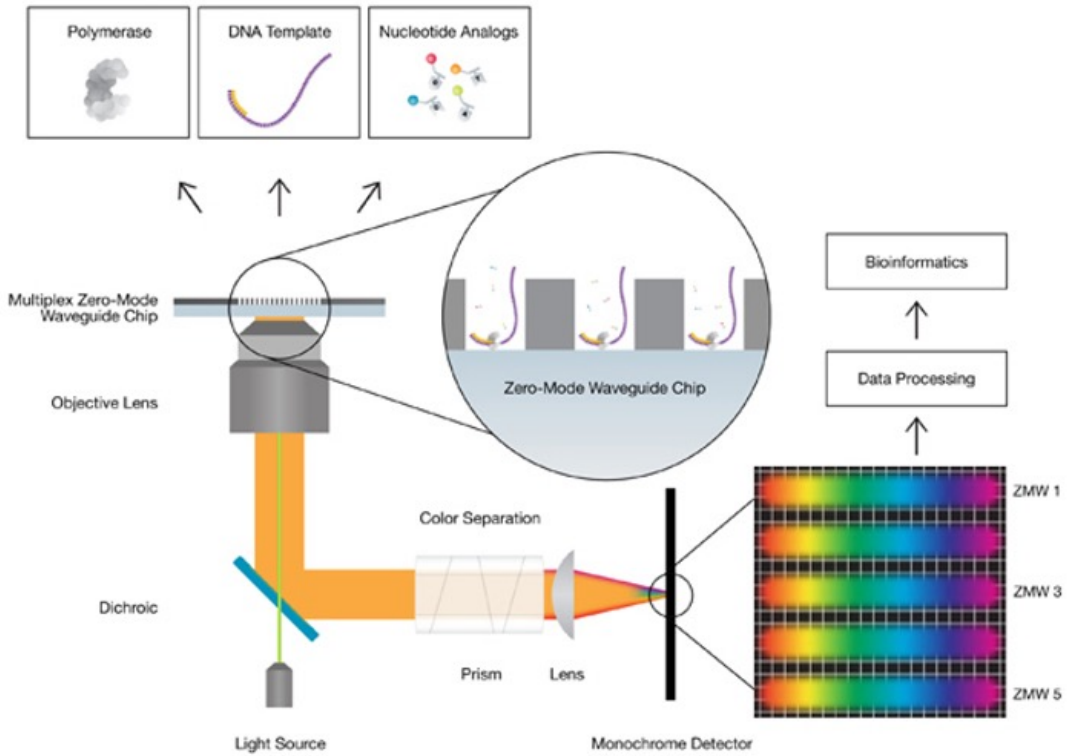
What's Next – PacBio

- As nulceotides are incorporated, the ZMW detects the release of the flourescent tag – sequencing by synthesis
- Immobilized DNA polymerase is stable enough for long reads: average > 10,000 bp, some reads > 60,000 bp
- Single polymerase – single template sequencing has orders of magnitude less consumable needs than Illumina sequencing by synthesis



What's Next – PacBio

- Cost is still high due to technology costs, but that is expected to drop
- PHRED error rates remain higher but each SMRT cell generates ~55,000 reads so error is offset by coverage (except for bias!)
- Greatly simplifies the genome sequencing problem, particularly for bacteria
- Combined with low volume Illumina sequencing, bacterial genomes can be much more easily resolved

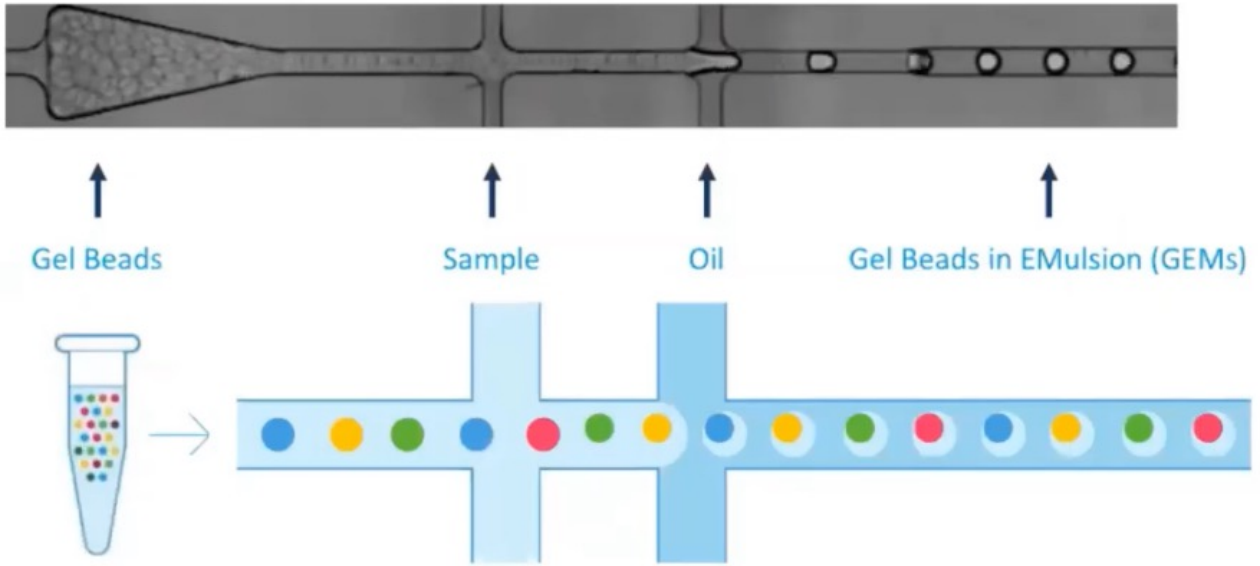


10x Genomics single-cell sequencing platform

- Combine cell sorting with DNA sequencing
- Small footprint – fit it in the BSL3 lab
- Single cell transcriptomics!
- Big implications for cancer research (i.e. stem cells) and infectious disease (i.e. virus impact by cell type)

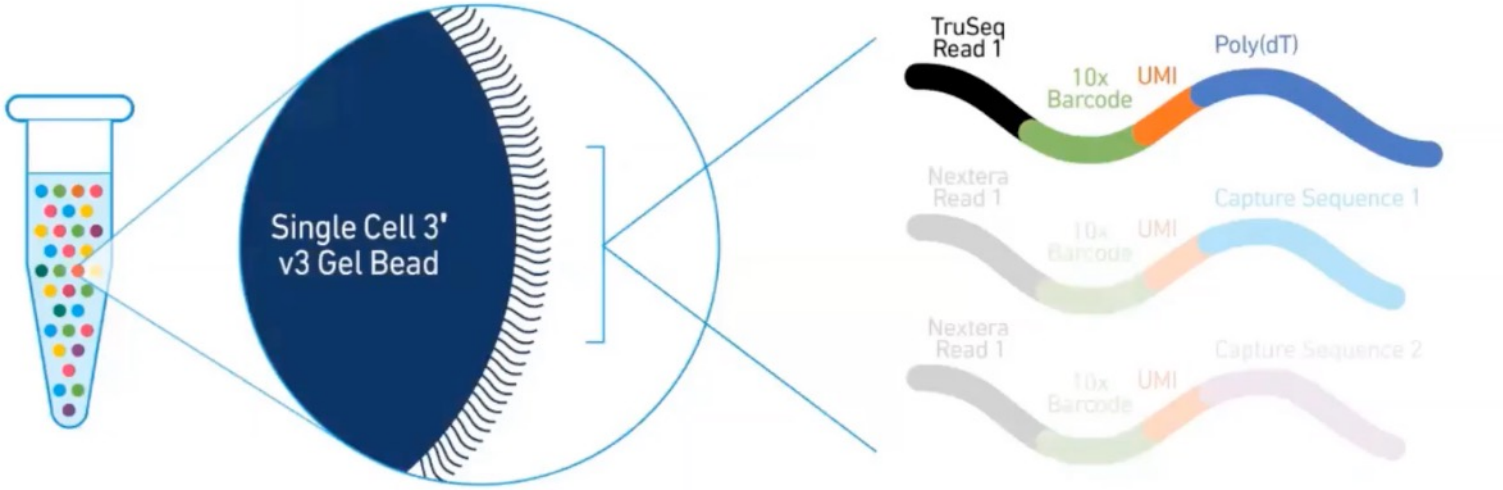
10x Genomics single-cell sequencing platform

- Combine cell sorting with DNA sequencing
- Small footprint – fit it in the BSL3 lab
- Single cell transcriptomics!
- Big implications for cancer research (i.e. stem cells) and infectious disease (i.e. virus impact by cell type)



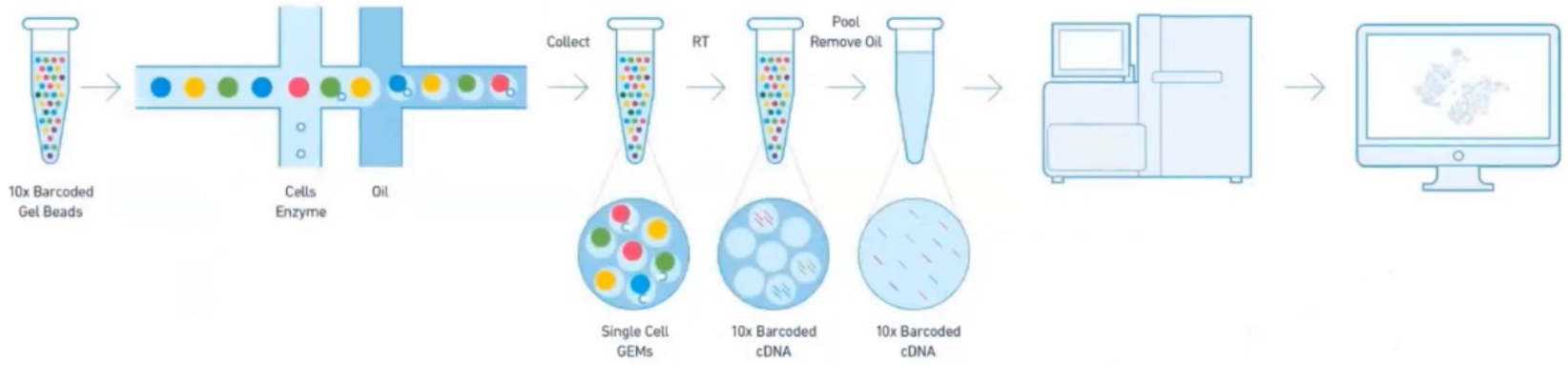
10x Genomics single-cell sequencing platform

- Combine cell sorting with DNA sequencing
- Small footprint – fit it in the BSL3 lab
- Single cell transcriptomics!
- Big implications for cancer research (i.e. stem cells) and infectious disease (i.e. virus impact by cell type)



10x Genomics single-cell sequencing platform

- Combine cell sorting with DNA sequencing
- Small footprint – fit it in the BSL3 lab
- Single cell transcriptomics!
- Big implications for cancer research (i.e. stem cells) and infectious disease (i.e. virus impact by cell type)



10x Genomics single-cell sequencing platform

- Combine cell sorting with DNA sequencing
- Small footprint – fit it in the BSL3 lab
- Single cell transcriptomics!
- Big implications for cancer research (i.e. stem cells) and infectious disease (i.e. virus impact by cell type)

Ion Gene Studio S5 Prime & Ion Chef

- Ion Torrent Single Molecule, Real-Time (SMRT) technology with maximum throughput and rapid turnaround time
- 2–130 million reads per run
- Low sample input
- Less than 24 hour turnaround time
- Low error rate (99.5% accuracy)

- Pandemics - real time outbreak response; 2-50 Gb of sequencing per day; 1500+ SARS-CoV-2 genomes per day!

- Ion Chef full automates library construction, template preparation, and chip loading in less than 15 minutes!

Illumina NovaSeq 6000 versus NextSeq 1000 & 2000

NextSeq 1000 & 2000

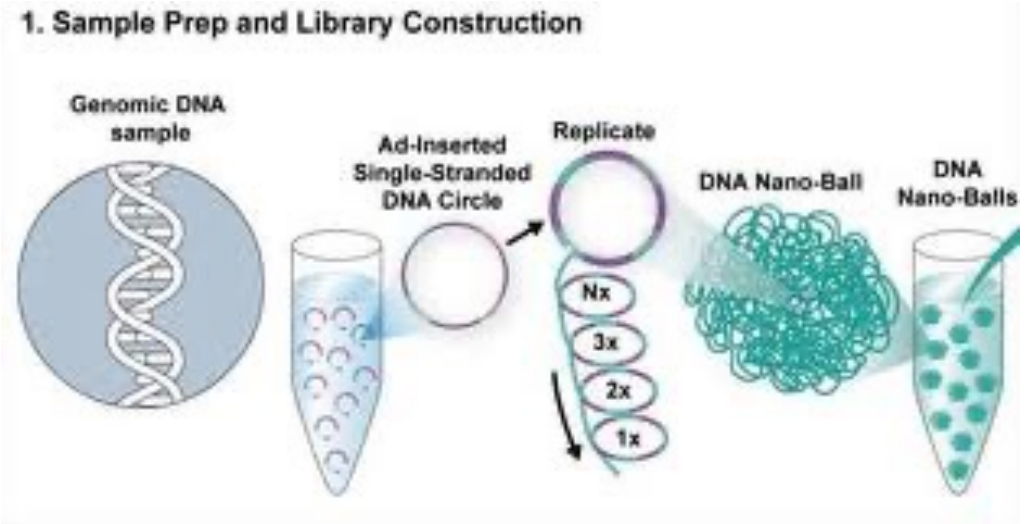
330 Gb output
11-48 hours
1.1 billion reads per run
2 x 150 bp

NovaSeq 6000

6000 Gb output
13-44 hours
20 billion reads per run
2 x 250 bp

MGI DNA Sequencers

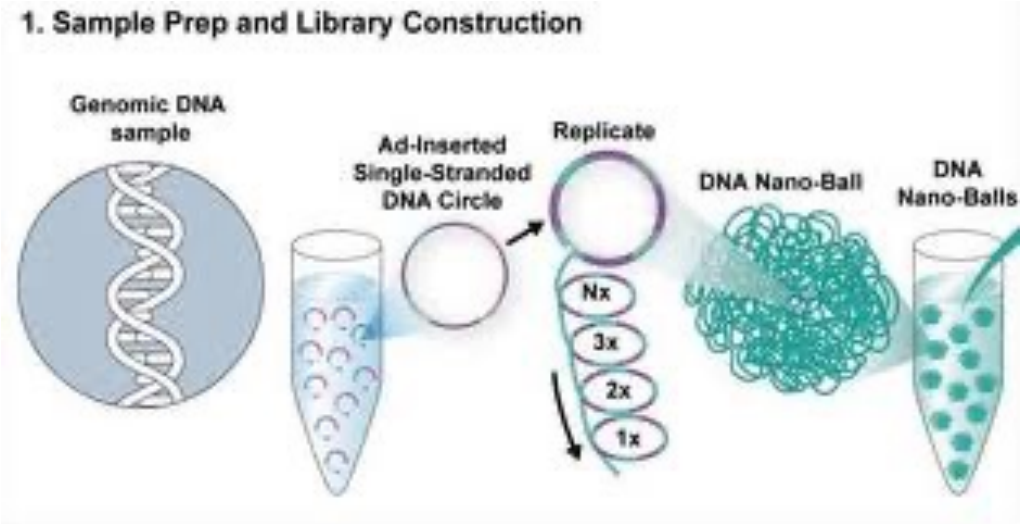
- DNA nanoball and combinatorial probe anchor synthesis



Rolling circle amplification leads to creation of DNA nano-balls (DNBs)

MGI DNA Sequencers

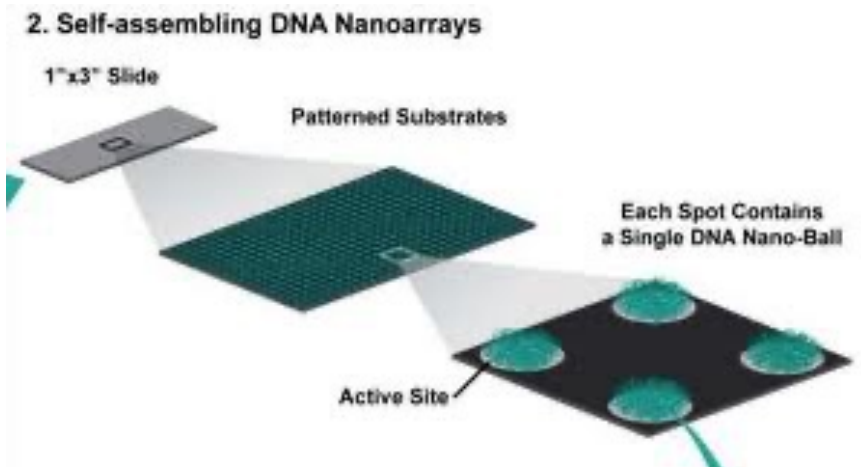
- DNA nanoball and combinatorial probe anchor synthesis
- No amplification bias



Rolling circle amplification leads to creation of DNA nano-balls (DNBs)

MGI DNA Sequencers

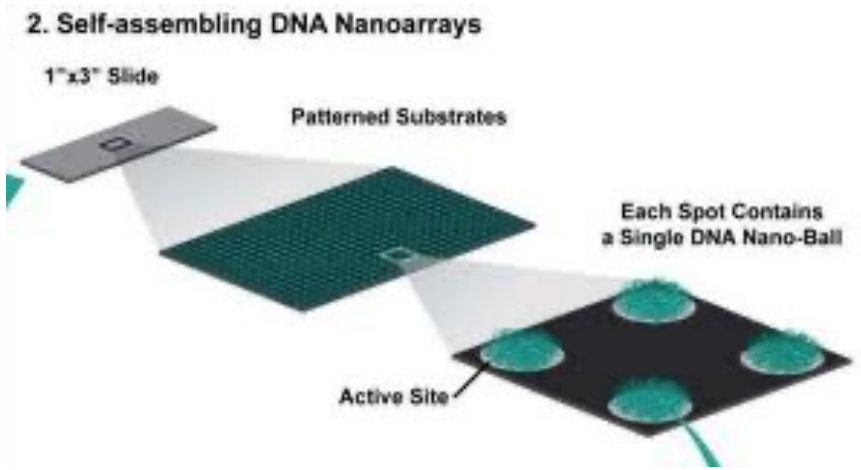
- DNA nanoball and combinatorial probe anchor synthesis
- No amplification bias



Array has 2+ billion spots – one DNB per spot via self-assembly

MGI DNA Sequencers

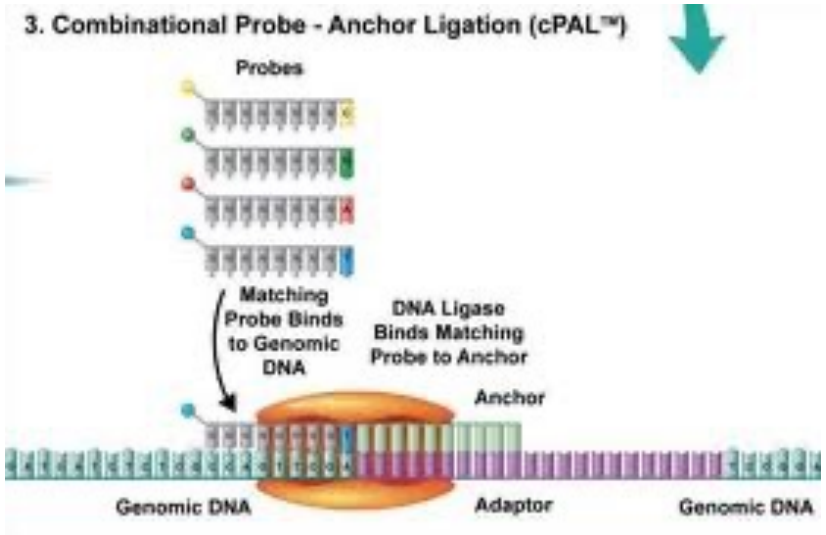
- DNA nanoball and combinatorial probe anchor synthesis
- No amplification bias; no ligation bias



Array has 2+ billion spots – one DNB per spot via self-assembly

MGI DNA Sequencers

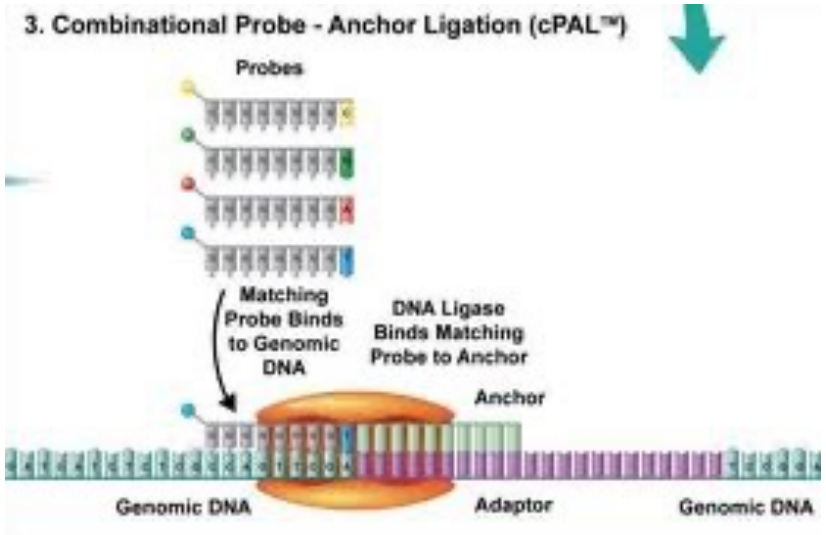
- DNA nanoball and combinatorial probe anchor synthesis
- No amplification bias; no ligation bias



Combinatorial Probe-Anchor Ligation (cPAL) sequencing, i.e. hybridization of dye-labelled probes

MGI DNA Sequencers

- DNA nanoball and combinatorial probe anchor synthesis
- No amplification bias; no ligation bias
- Novel base & read calling bioinformatics – PHRED scores?



Combinatorial Probe-Anchor Ligation (cPAL) sequencing, i.e. hybridization of dye-labelled probes

MGI DNA Sequencers

- DNA nanoball and combinatorial probe anchor synthesis
- No amplification bias; no ligation bias
- Novel base & read calling bioinformatics – PHRED scores?

DNBSEQ-T7
- 6 Tb of data
- 5 billion reads
- 24 hours
- 2 x 150 bp

DNBSEQ-G400
- 1440 Gb of data
- 1.8 billion reads
- 30-100 hours
- 2 x 200 bp or 1 x 400 bp

Conclusions

- Throughput is going to continue to increase - Kryder's law is going to be an increasing problem
- Long read technologies *should* lead to easier bioinformatics as they overcome many of the problems of genome or transcript assembly (particularly repeats)
- The PHRED paradigm will continue to be important but we will need new PHRED prediction algorithms for each new technology
- Are we prepared for bioinformatics where we can completely sequence a human transcriptome or thousands of bacterial genomes in a day? Every undergraduate thesis could generate 'Big Data'