Long read sequencing

The good, the bad, and the really cool.

A/Prof Torsten Seemann

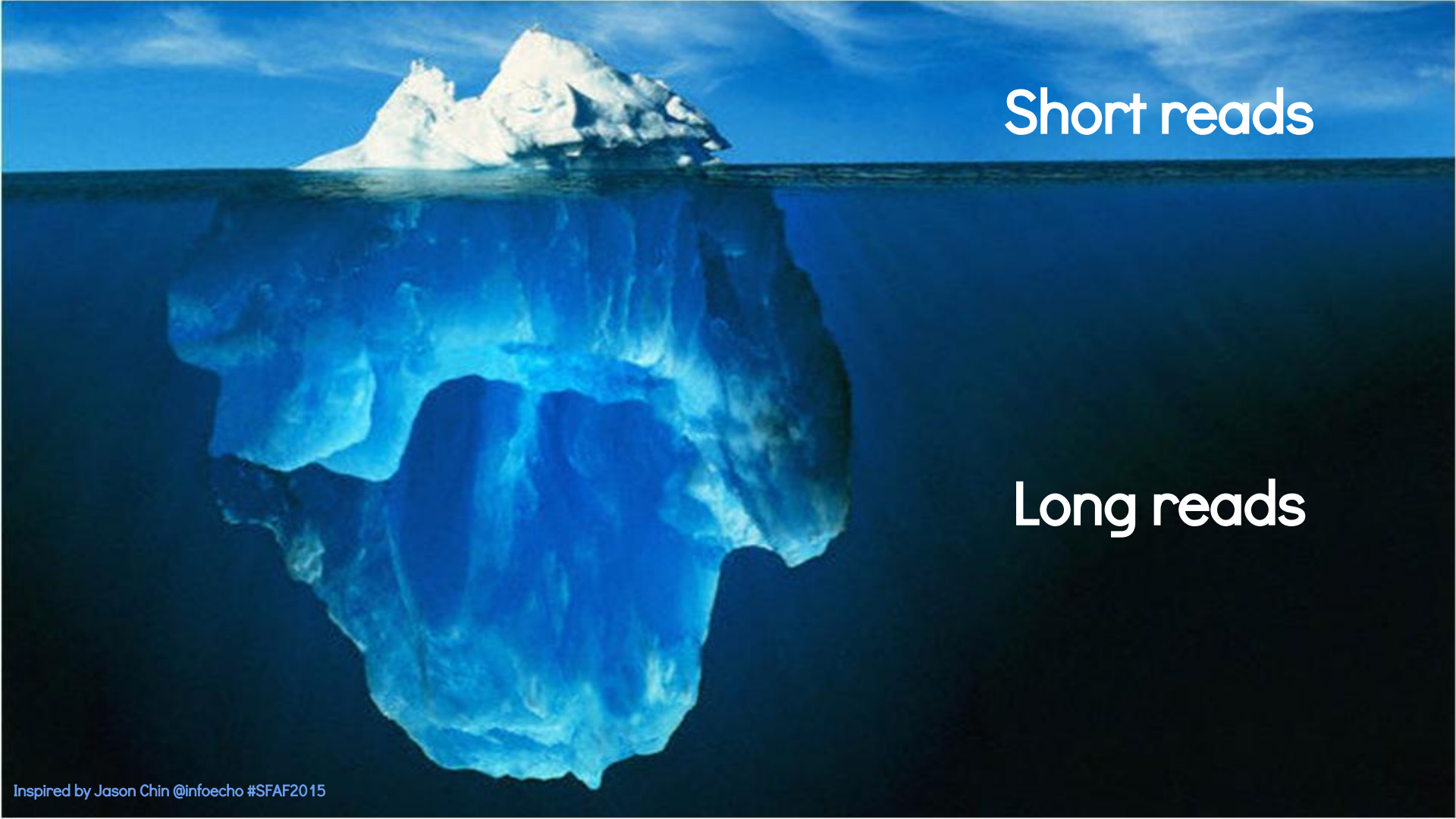
ANGUS 2015 Workshop - KBS, Michigan, USA - Sat 15 Aug 2015

Melbourne, Australia

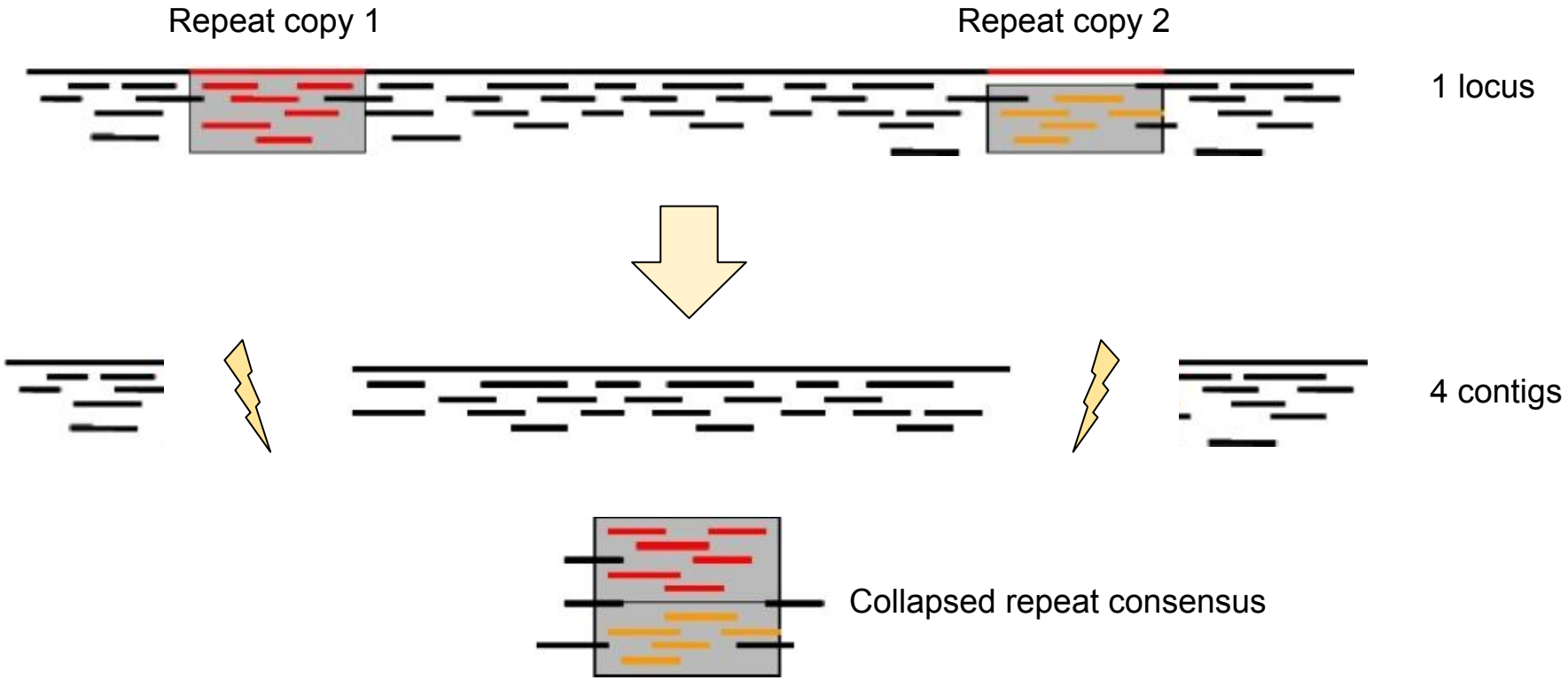
Microbial genomics

Doherty Centre for
Applied Microbial Genomics

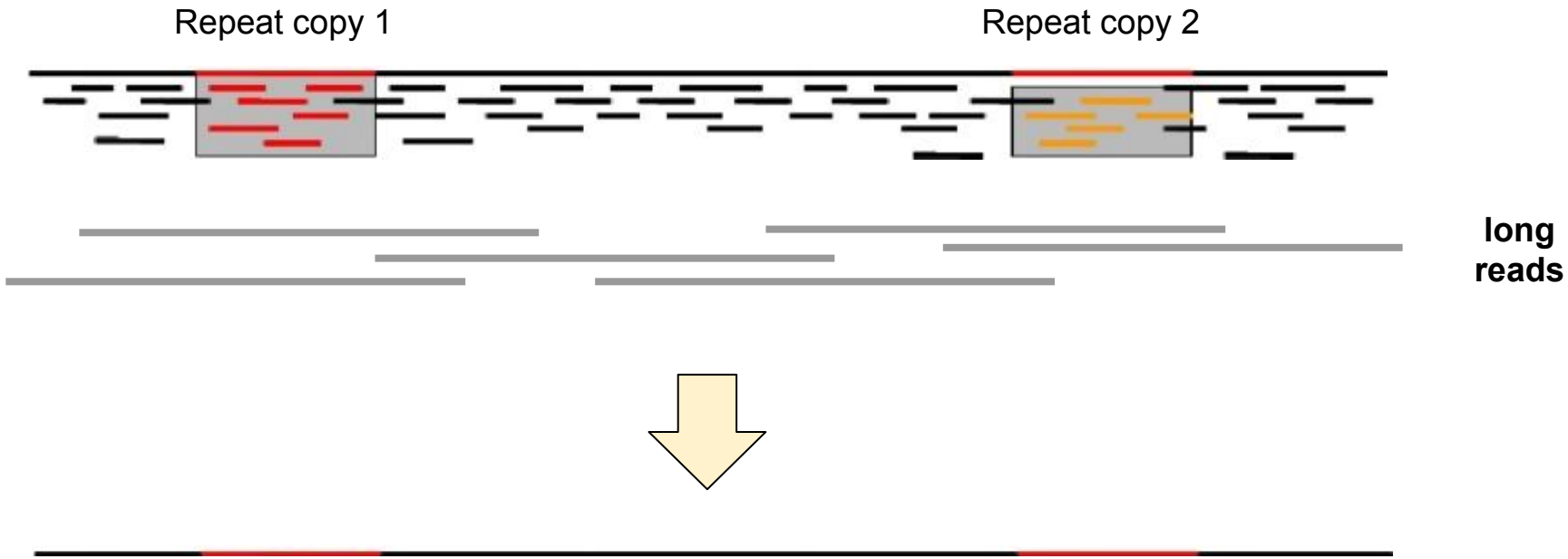
Why do we need long reads?


Short reads
Long reads
Inspired by Jason Chin @infoecho #SFAF2015

Repeats

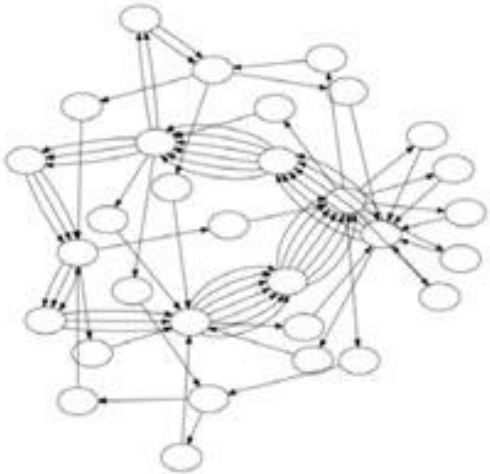

Repeat copy 1
Repeat copy 2
1 locus
4 contigs
Collapsed repeat consensus

Long reads can span repeats

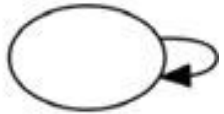


Long reads untangle graphs

100 bp

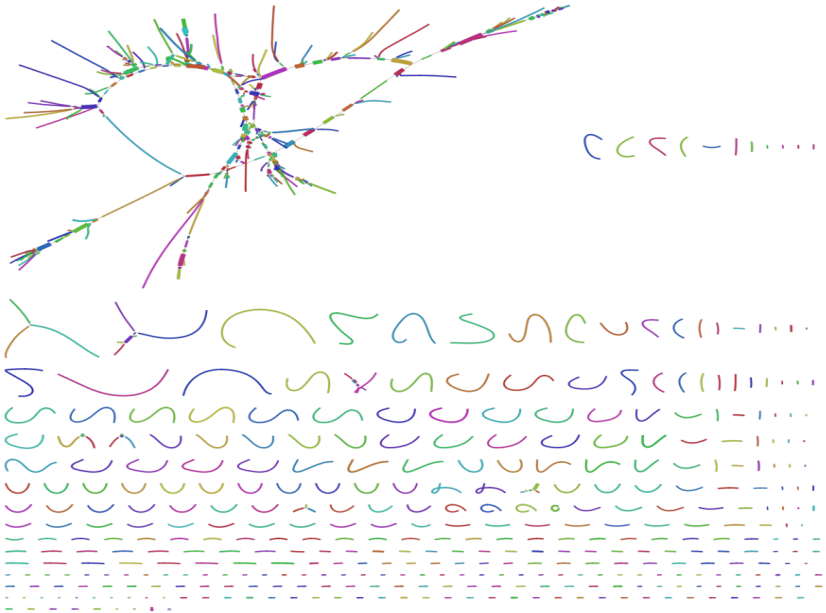


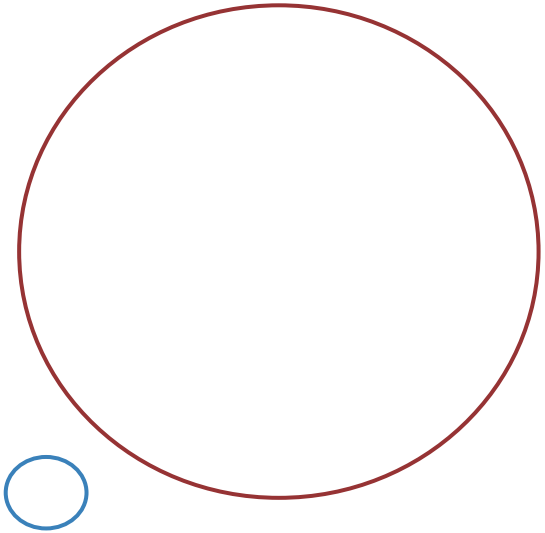
1000 bp



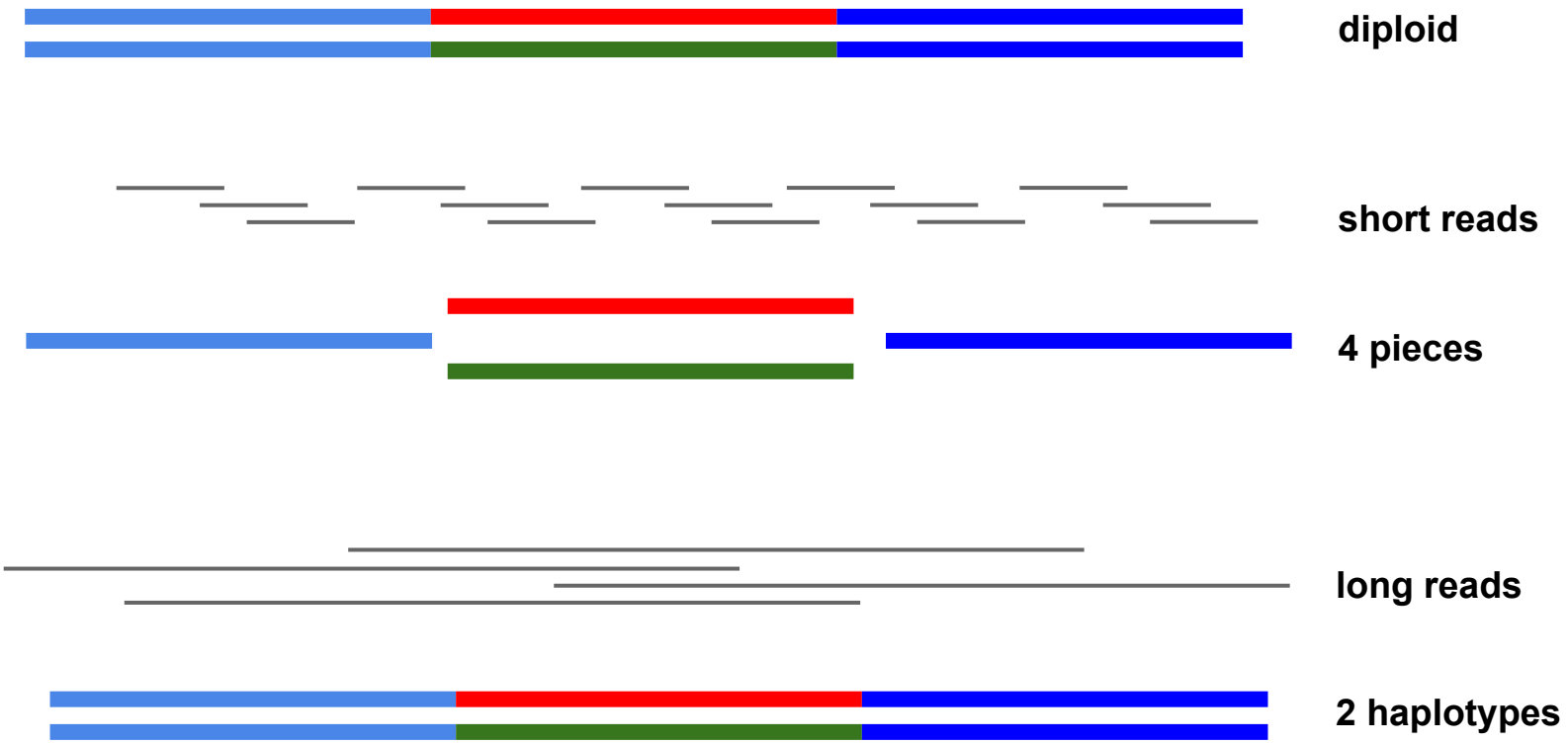
5000 bp

Completed genomes





Heterozygosity



Phased haplotypes

Maternal ATGCTACGATCGCTCG

Paternal ATGGTACGATCGATCG

Unphased: ATG(C/G)TACGATCG(C/A)TCG

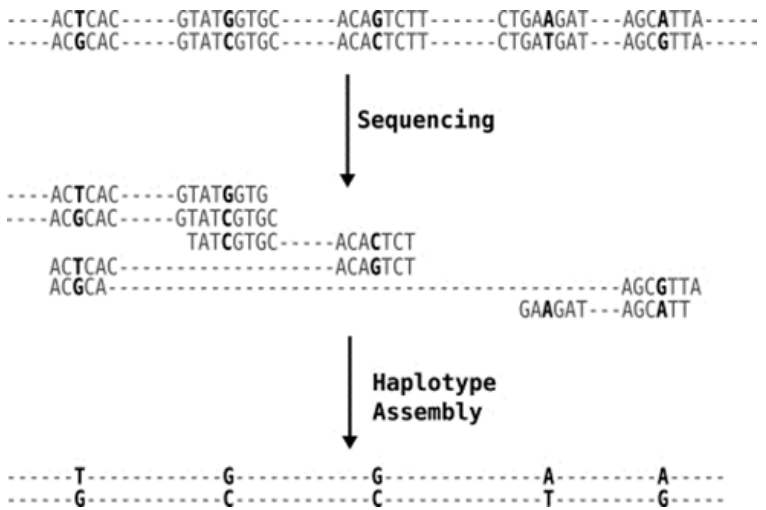
A few phasing possibilities

Maternal ATGCTACGATCGCTCG
Paternal ATGGTACGATCGATCG

Maternal ATGGTACGATCGATCG
Paternal ATGCTACGATCGCTCG

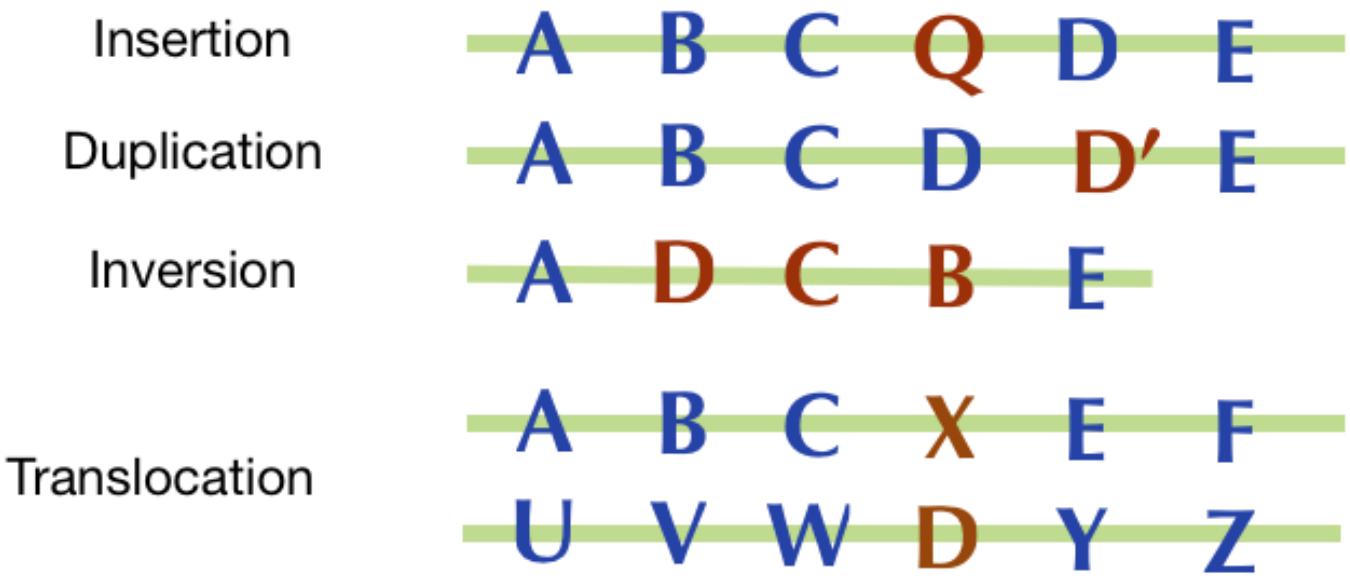
Maternal ATGCTACGATCGATCG
Paternal ATGGTACGATCGCTCG

Maternal ATGGTACGATCGCTCG
Paternal ATGCTACGATCGATCG



Structural variation

The missing heritability - not just SNPs



Long read technologies

Two flavours

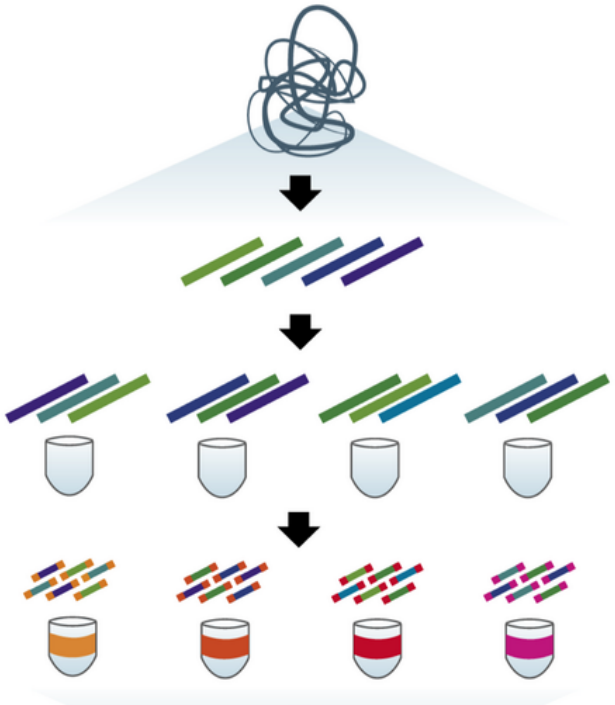
:: Synthetic long reads
   : Still needs an Illumina short-read sequencer
   : Molecular biology tricks + local assembly / constraints
   : *Illumina SLR, 10X Genomics, Dovetail*

:: Genuine long reads
   : The real deal - no tricks!
   : *Pacific Biosciences, Oxford Nanopore*

Illumina SLR

"Moleculo" synthetic long reads

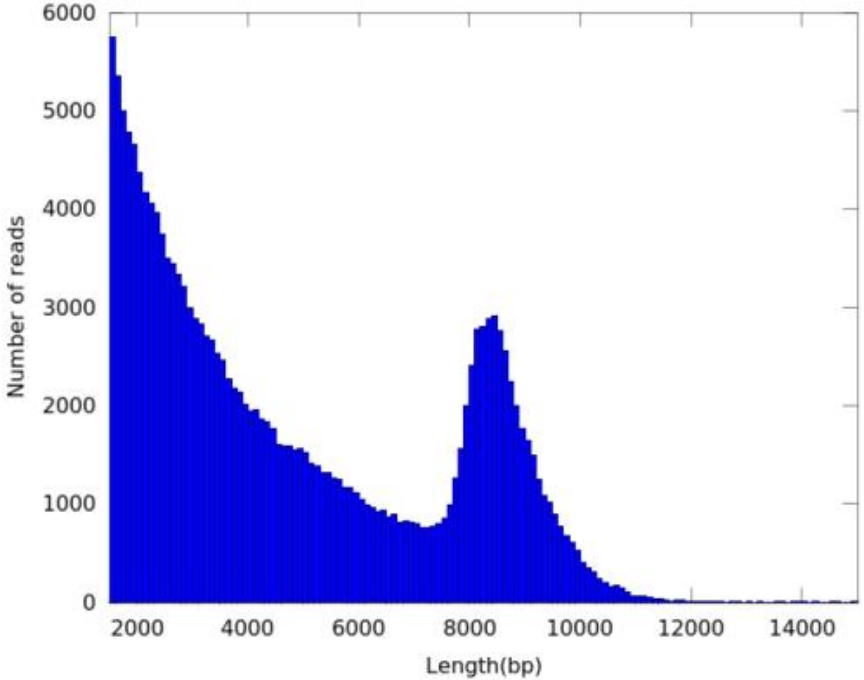
Synthetic long reads

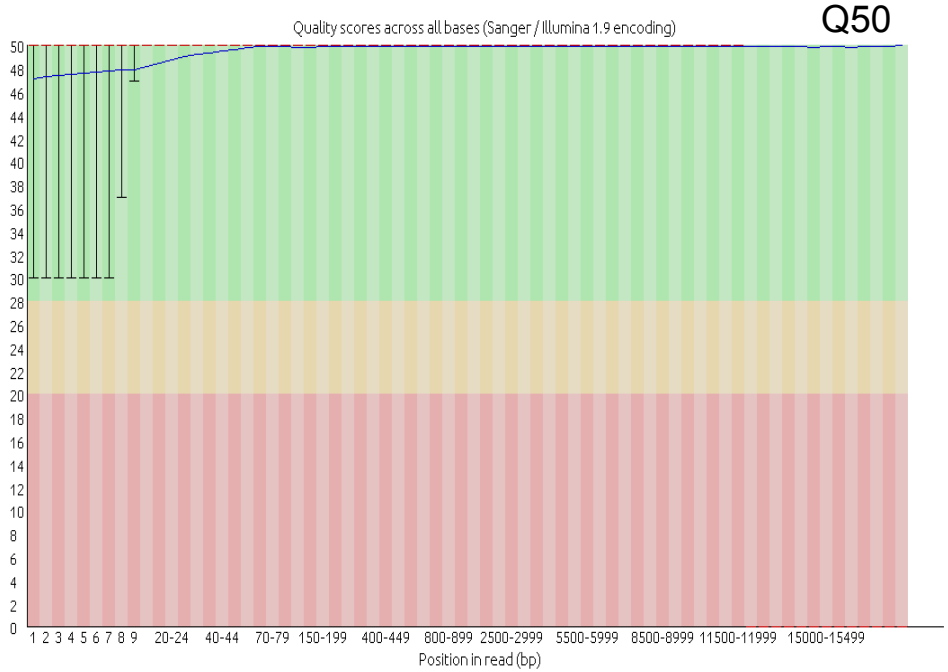


1. Genomic DNA
2. Shear ~10 kbp fragments
3. Dilute ~500 fragments per well
4. Amplify
5. Shear to ~ 500 bp fragments
6. Barcode (384)
7. Short read sequencing
8. De-barcode into pools
9. *De novo* assemble each pool
10. Get ~500 x 10 kbp "long reads"

Adapted from http://www.nature.com/nbt/journal/v32/n3/abs/nbt.2833.html

Illumina SLR - “reads”





Pacific Biosciences

It's already here and it works.

Pacific Biosciences RSII

Operator*

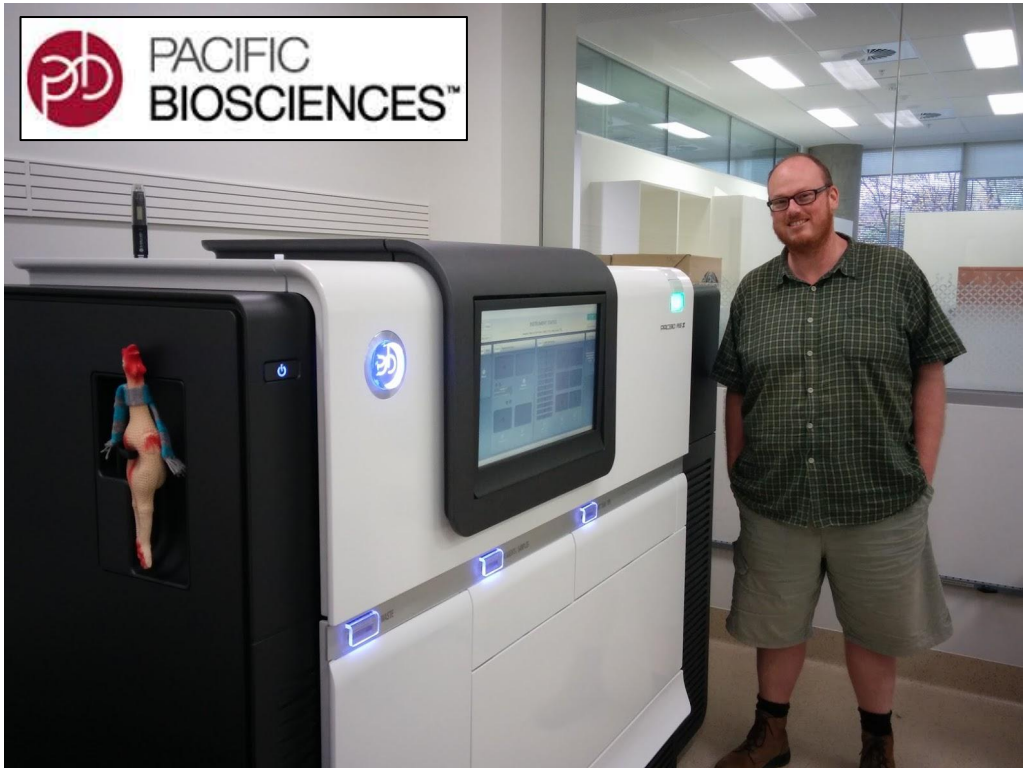
Robotics

Compute

Sequencing

** not included*

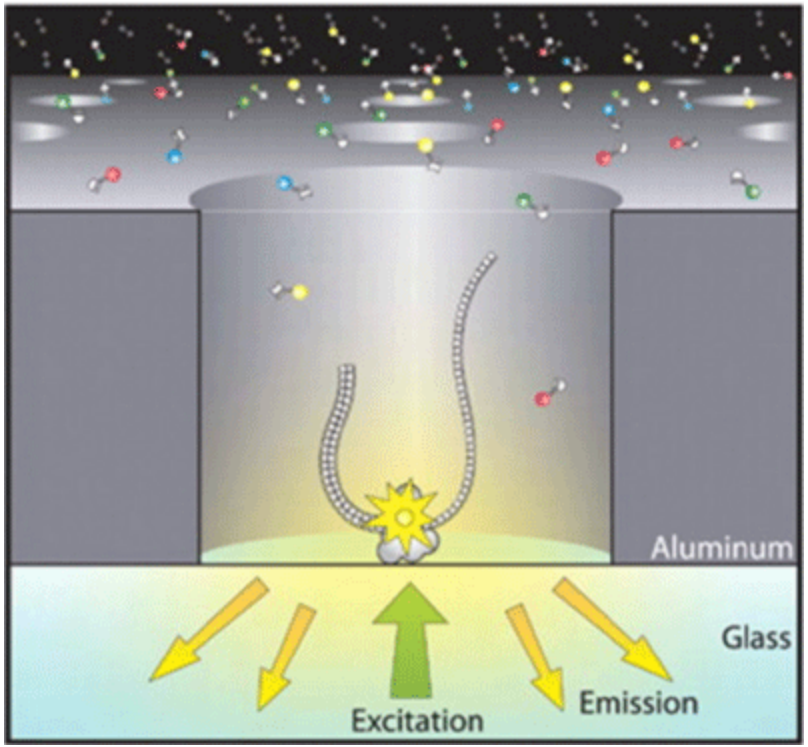
Pacific Biosciences RSII



Installed June 2015
at Doherty Institute

Just started running!

PacBio: technology

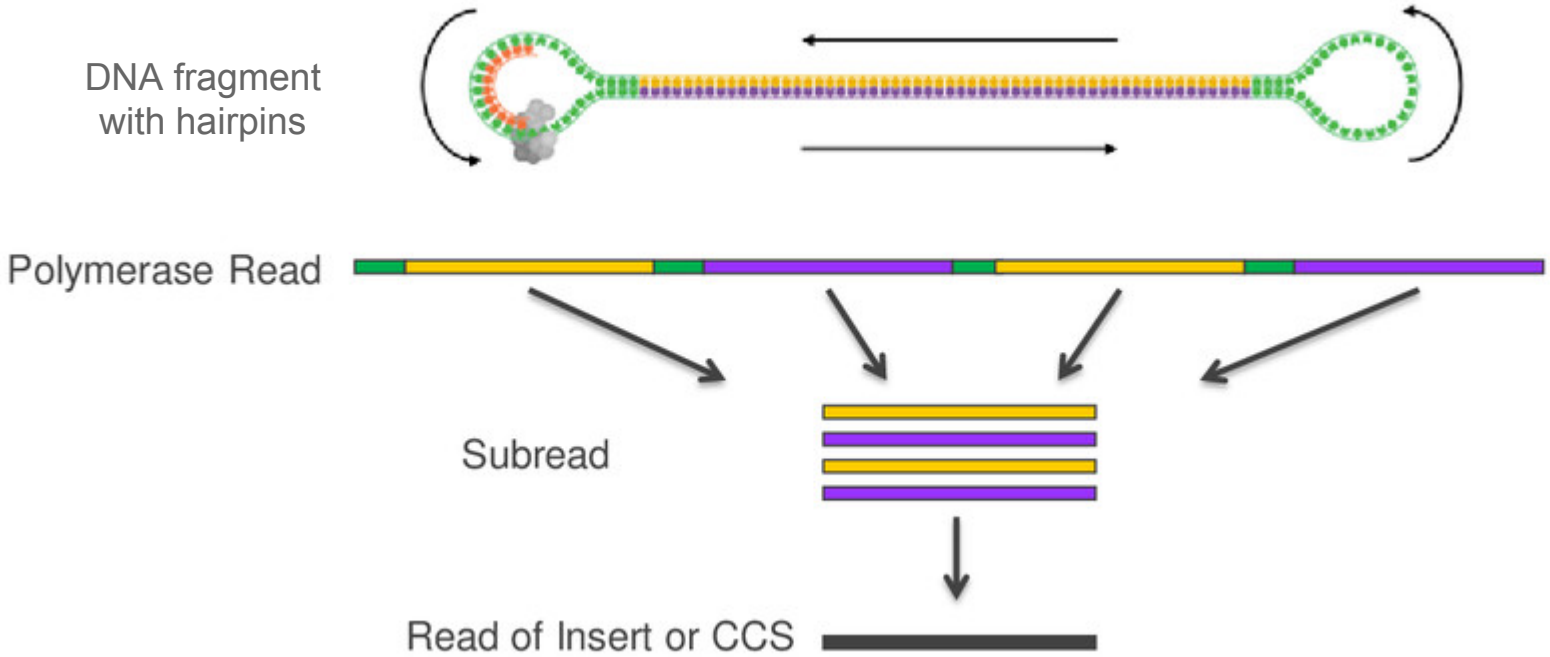


:: Polymerase bound to bottom of ZMW μ-well

:: Incorporation of fluorescent nucleotides measured in real time

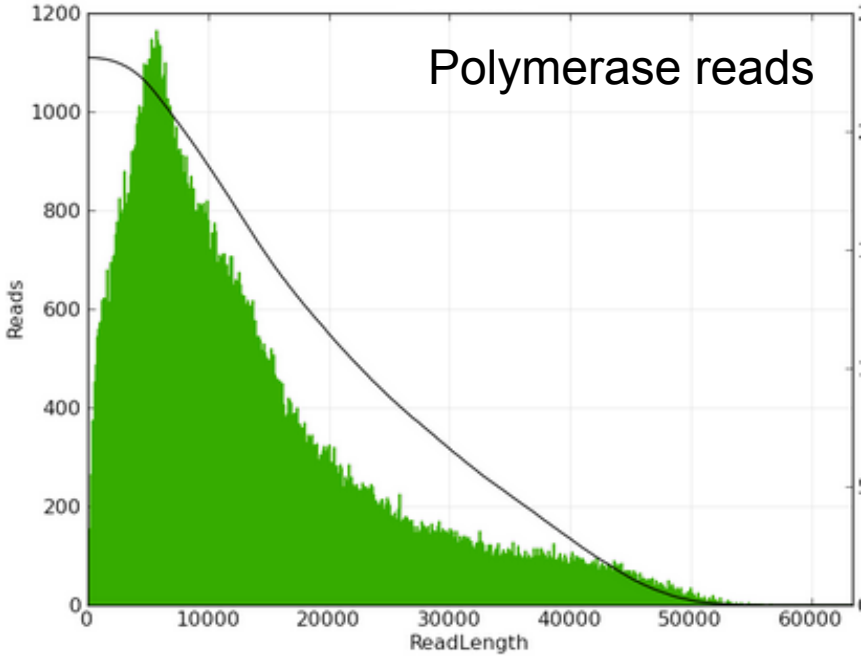
:: 3 hour “movies”

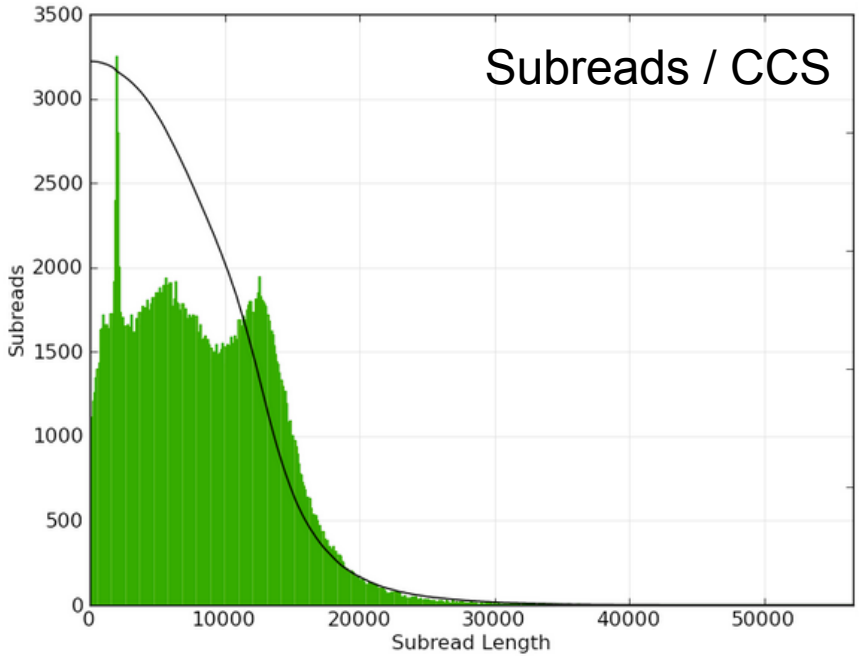
Pacbio: reads



Adapted from https://speakerdeck.com/pacbio/track-1-de-novo-assembly

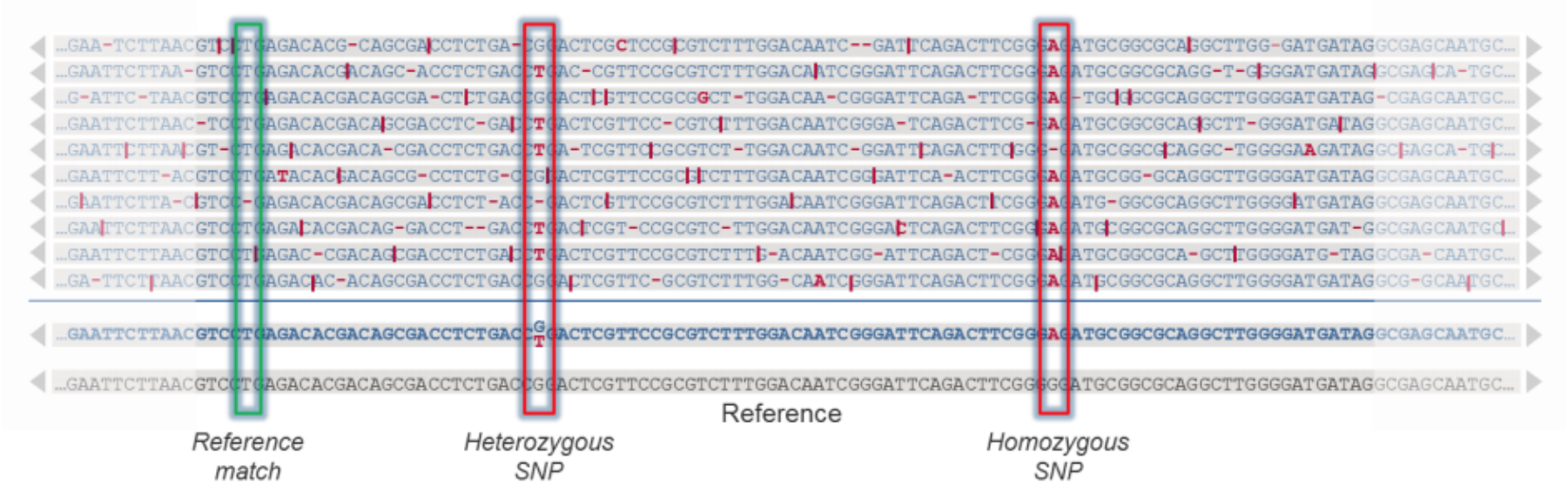
PacBio: our first two SMRT cells





Yield: 2.3 Gbp **No. reads:** 275,906 **Mean length:** 8387 bp **N50 length:** 11782 bp

PacBio: error rate

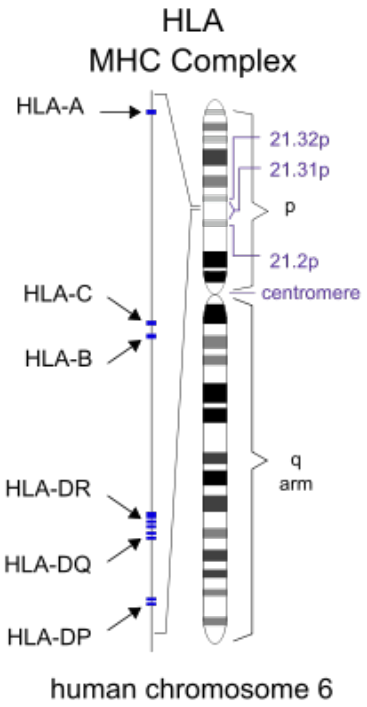


Single read: 86%

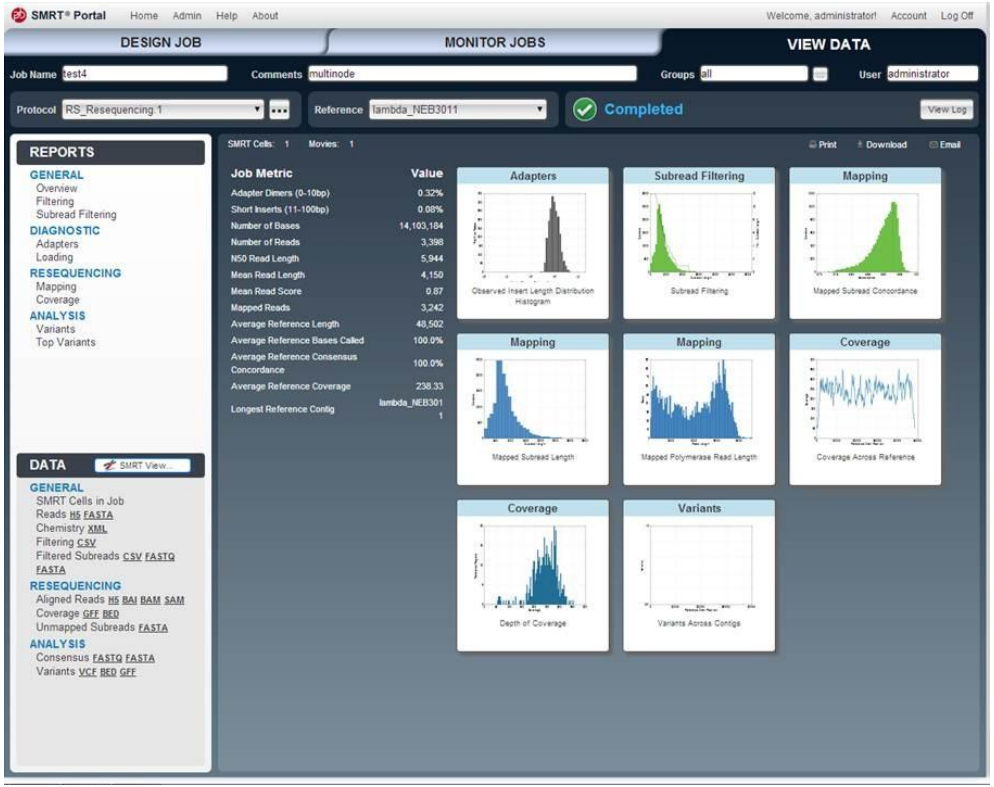
30x Consensus: 99.999%

PacBio: main applications

- Finished genomes
- Full length cDNA (mRNA isoforms)
- Extreme GC sequence
- HLA / MHC / KIR haplotyping
- Base modifications (methylation)



PacBio: bioinformatics

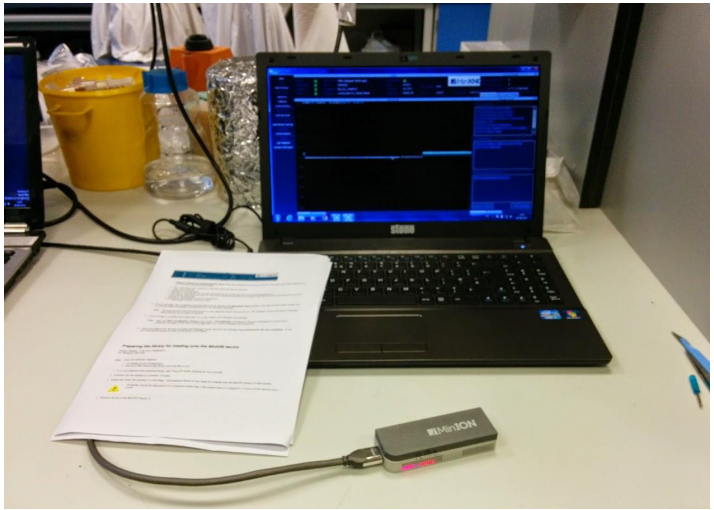


:: All in GitHub
:: SMRT Portal
  : Nice GUI
  : Cloud ready
  : Linux backend
  : Cluster ready
:: Cmdline tools
:: Good docs

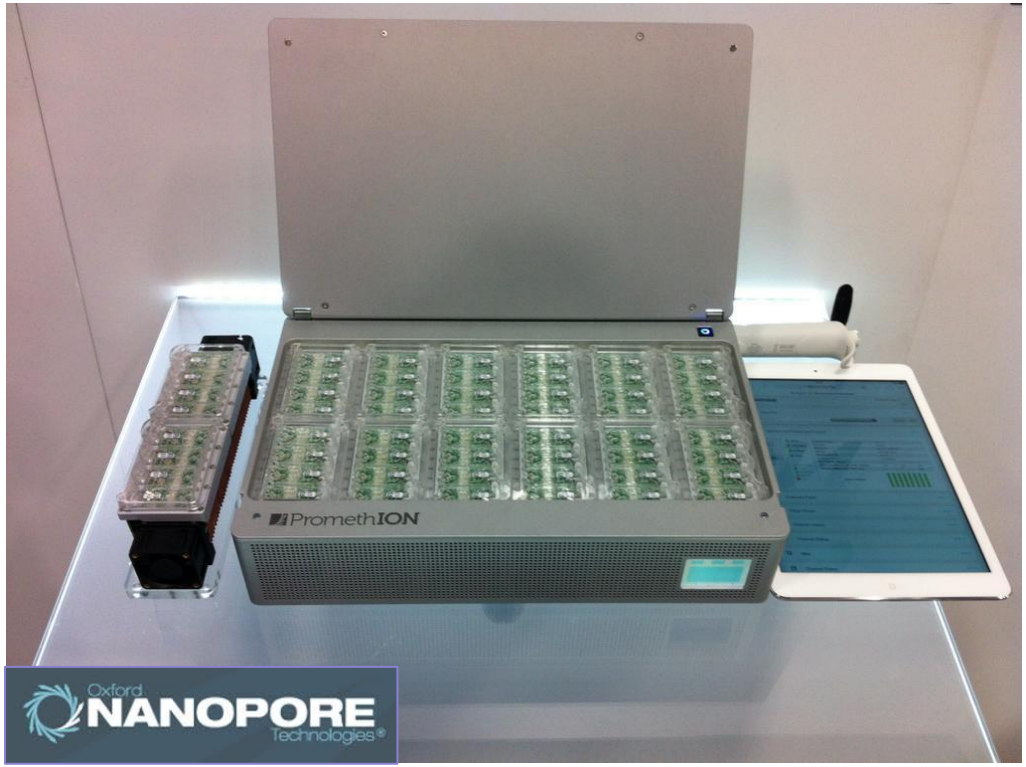
Oxford Nanopore

The new kid on the block.

MinION - the device

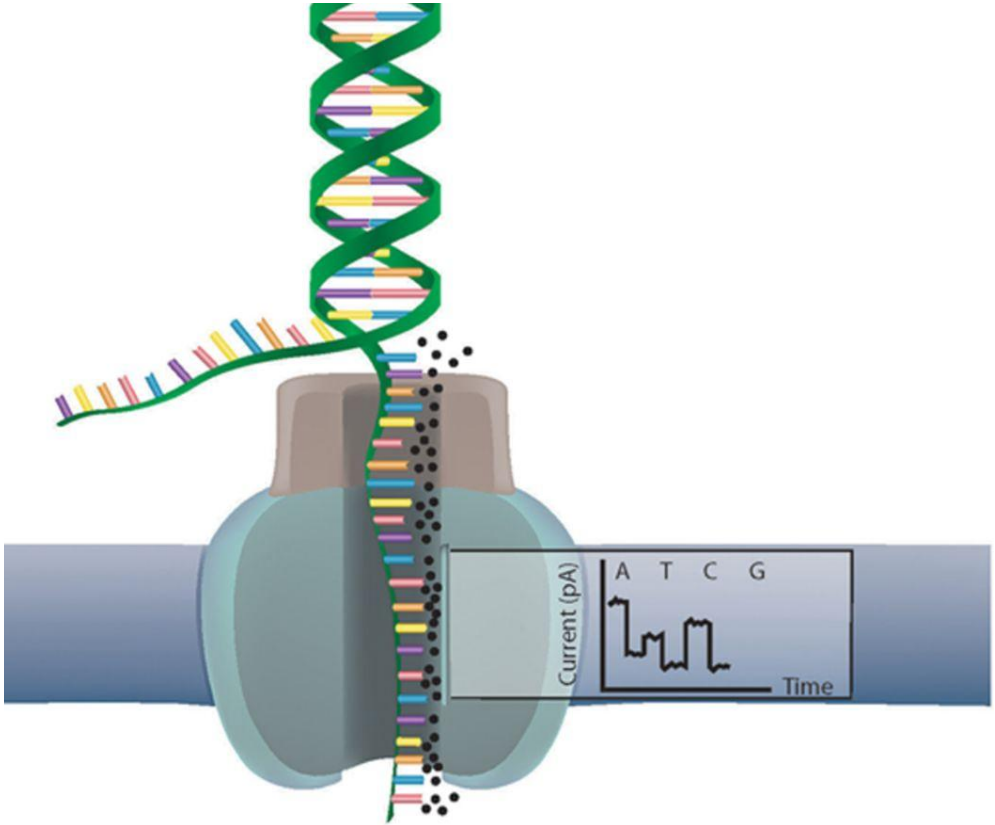


PromethION - large scale device



- 48 independent flow cells
- On board ASIC
- Runs Python
- Optional compute

Nanopore - technology

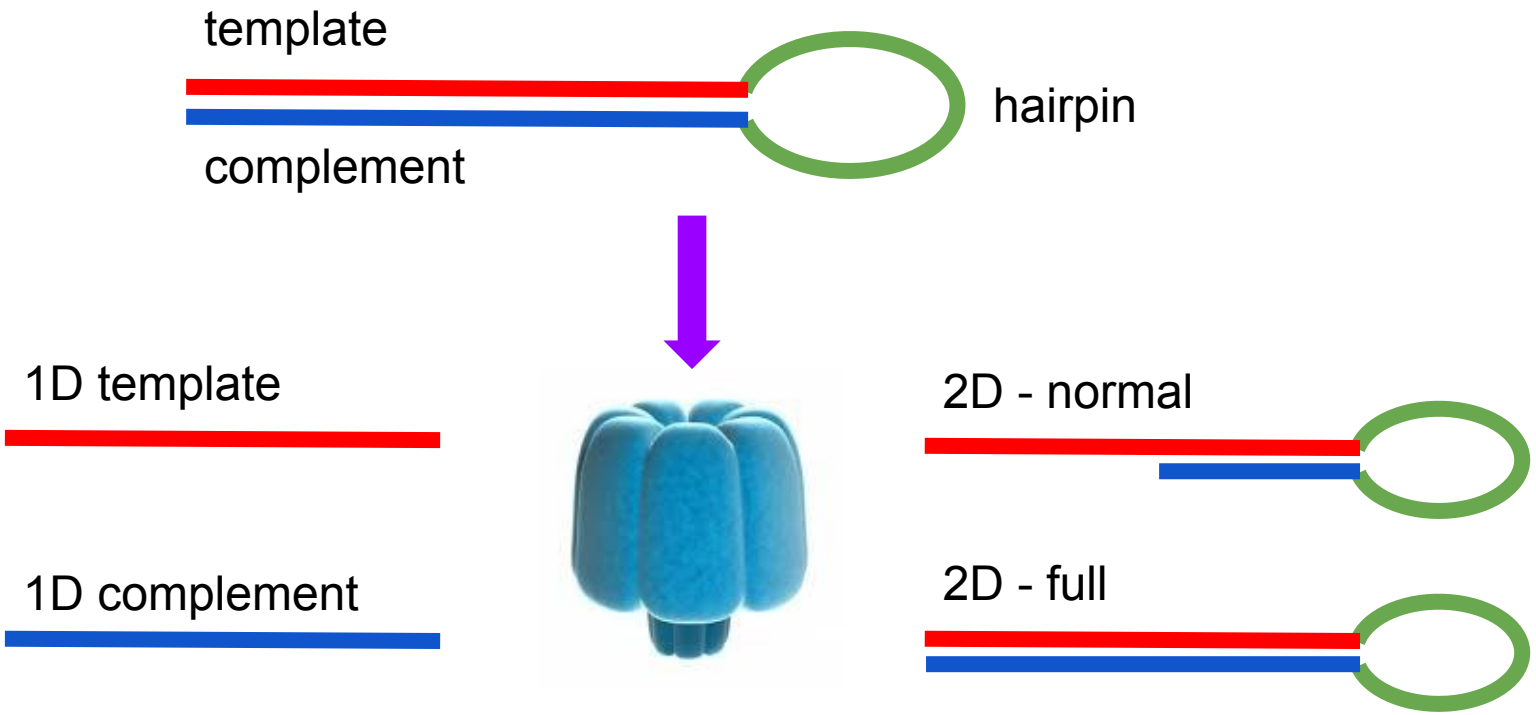


Signal is measured from 5 bases

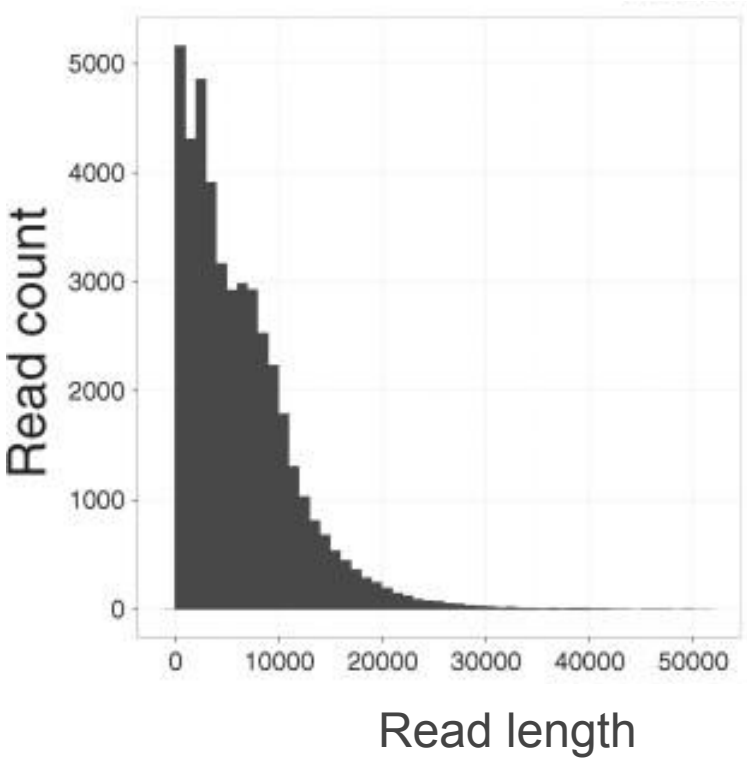
Timing is irregular

Base modifications do alter the signal

Nanopore - reads



Nanopore - read lengths

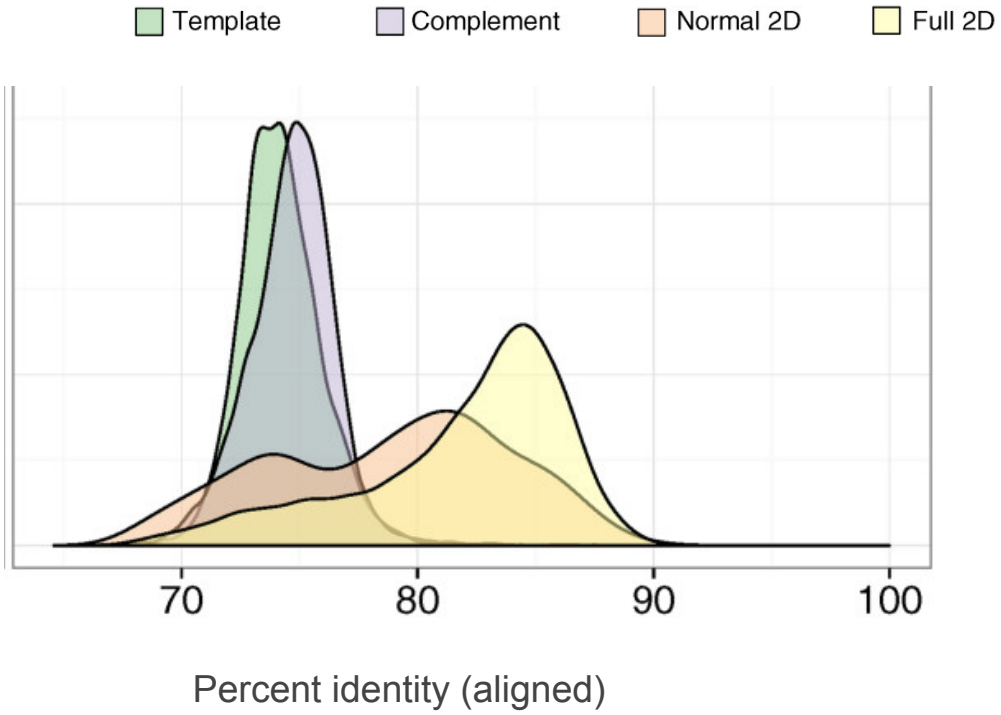


Read length is not limited by technology but by library preparation.

Can get >100kbp reads.

But not trivial to do so!

Nanopore - error rate



- 5-mer errors
- Homopolymer issues
- Not modelling base mods yet
- Changes with pore & motor enzyme

MinION - applications

:: Same as PacBio plus....

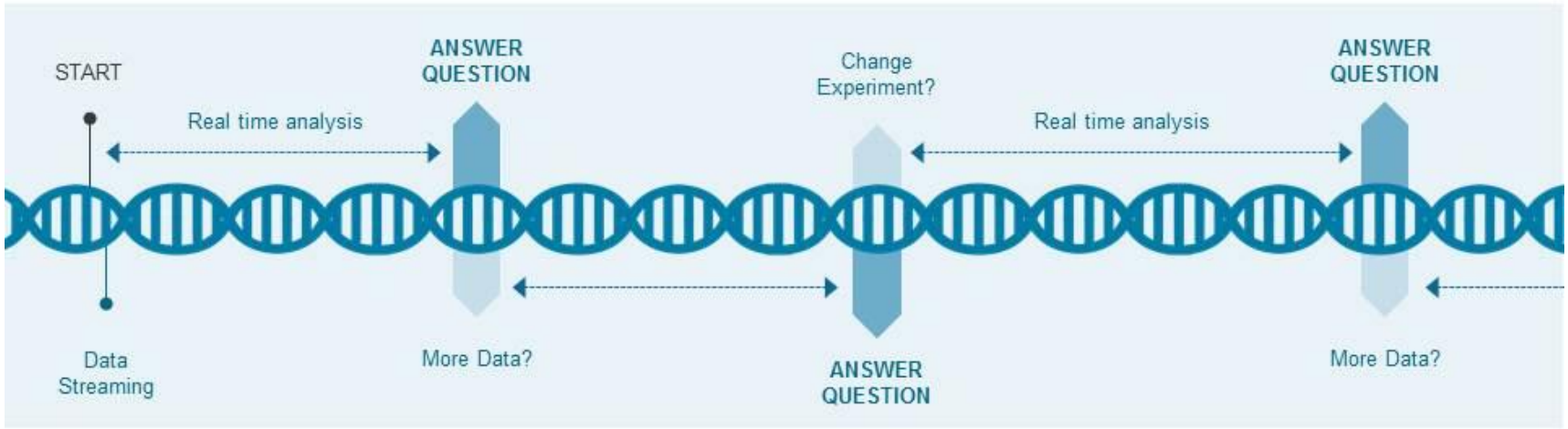
:: Portable sequencing
  : in the field eg. Josh Quick in Guinea for Ebola
  : in hospitals - infection control
  : monitoring - water/food supply, production facilities
  : at the GP - pathogen test in 10 min from blood prick?
  : spit in a home device every morning?

Disruptive technology

Or just another sequencer?

“Run until”

Dynamically adjust sequencing yield

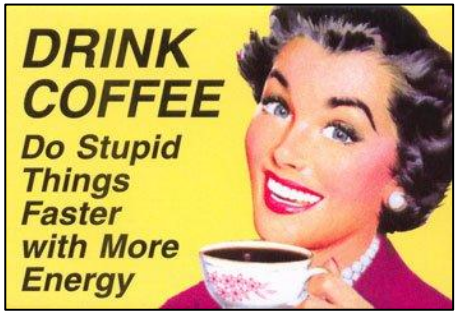


“Read until”

:: Can access events/bases during reading
  : remember reads are long 40 kbp
  : examine first 100 bp say (40 bp/sec currently)
  : can decide to stop reading and eject molecule!

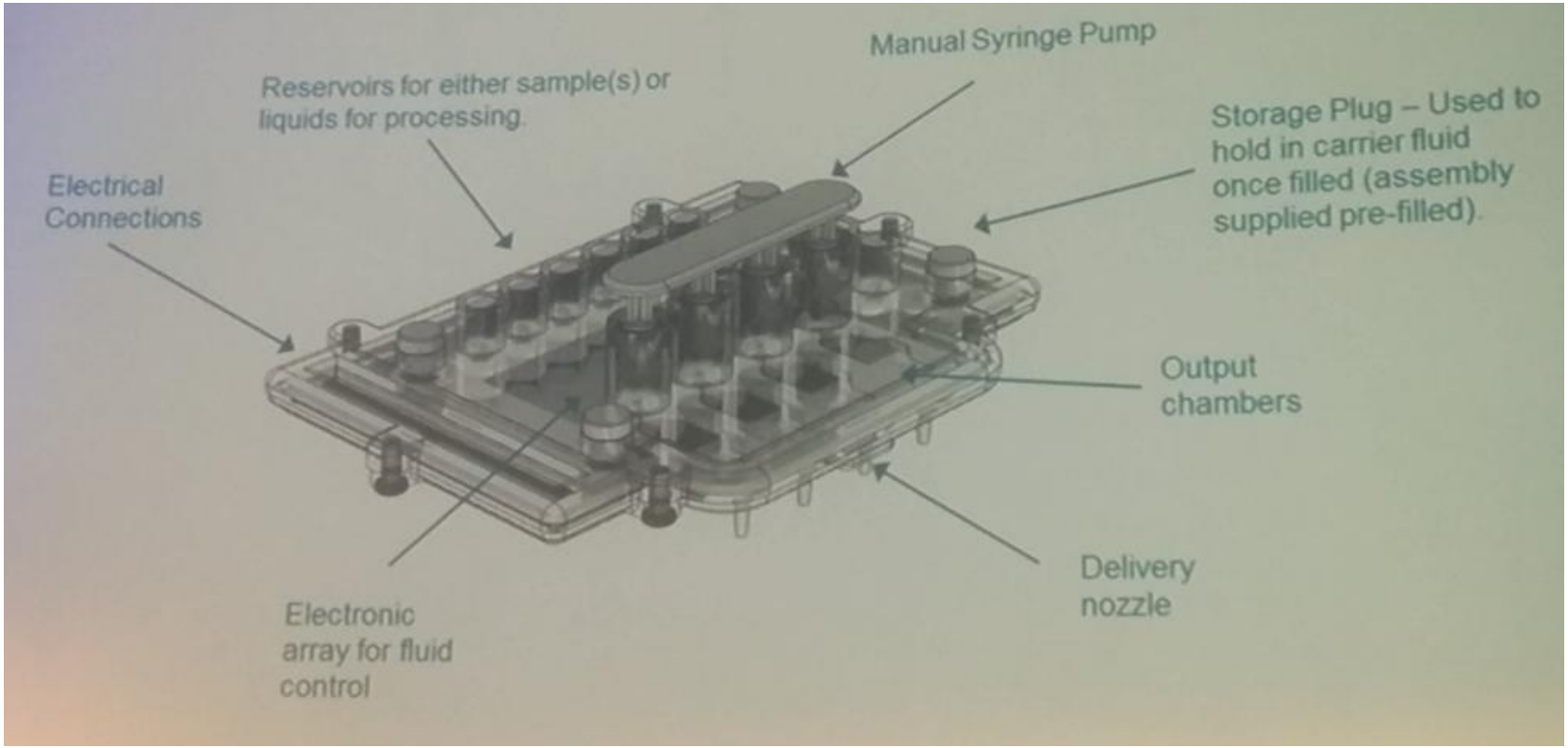
:: This is a killer app!
  : only want pathogens? eject if human DNA
  : only want exome? eject if not exonic looking
  : controlled with Python code

“Fast mode”



:: 2015 / MkI
  : enzyme deliberately slowed down - ASIC can’t keep up!
  : ~40 bases / sec / channel
  : 40 bp x 24 h x 512 channels ~ 2 Gbp
:: 2016 / MkII
  : new ASIC with ~3000 channels x 500 bp / sec
  : 500 bp x 1 week x 3000 channels ~ 900 Gbp
:: PromethION
  : has 48 flow cells . . . ~ 400 Tbp / week !?!

VolTRAX - library prep



A new business model

- No capital or reagent costs
  - Instrument will be free
  - Flow cells will be free
  - Only pay for what you want to sequence
  - Min. $20 and ~$1000 for a 100x human genome

- But I'll scam the system!
  - Flowcell stats sent back to base
  - Won't send you new flow cells if they look unused

How will bioinformatics change?

Some things never change

:: Don't worry!

: 50% of our job will always be converting file formats ☺

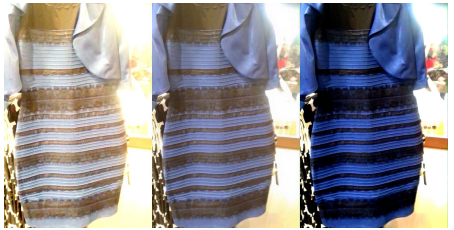
:: HDF5 - Hierarchial Data Format

: groups, multi-dimensional, indexed, random access

: Pacbio and Nanopore produce .h5 files

:: Can extract FASTQ from HDF5 easily

New work patterns



- Less aligning to reference, more *de novo*
- Learning to work with haplotypes
- More graph-based methods
- Complex structural variant information: `VCF4.2`
- Smaller data files ?

Read alignment

- Reads are 15% error, mainly indels
- PacBio
  - BLASR, Daligner, MHAP
  - BWA MEM: `bwa mem -x pacbio`
- Nanopore
  - BWA MEM: `bwa mem -x ont`
  - MarginAlign - sum over possible alignments

De novo assembly

- Pacbio
  - Very modular system: Overlap, Layout, Consensus
  - Polylploid aware assembly possible (Falcon)
- MinION
  - Higher error rate, but rapid community development
  - NanoCorrect + Celera Assembler + NanoPolish
- For existing assemblies
  - Gap-filling, scaffolding/breaking, hybrid assembly

Streaming analysis

- We are not going to keep all this data
  - Need to think *streaming analyses*
- Extract info we need and discard
  - Cheaper to resequence?
- Lots of new applications
  - Much scope for method development
  - Even more scope for biological discovery

Conclusion

Exciting times!

:: Genomics is changing all the time

: but science will press on

:: Pipelines are often short lived

: except maybe clinical / accredited ones

:: Bioinformaticians need to be able to adapt

: focus on key skills not specific apps

Acknowledgments

Doherty Institute

:: Tim Stinear
:: Ben Howden

VLSCI

:: Andrew Lonie
:: Helen Gardiner
:: Dieter Bulach
:: Simon Gladman

Twitter/Blogs

:: Nick Loman
:: Lex Nederbragt
:: Keith Robison
:: C. Titus Brown

Oxford Nanopore

:: Clive Brown
:: Gordon Sanghera

Millennium Science

:: Paul Lacaze
:: Matthew Frazer
:: Rubber Chicken

Pacific Biosciences

:: Siddarth Singh
:: Jason Chin
:: Stephen Turner

Australian Government
National Health and Medical Research Council
N H M R C

Australian Government
Australian Research Council

All the other CIs on the LIEF grant

THE UNIVERSITY OF MELBOURNE

Contact

http://tseemann.github.io

tseemann@unimelb.edu.au

@torstenseemann

The End

Thank you for listening.