Early experiences with the Ion Torrent

(How fickle is the Ion Trickle?)

Dr Torsten Seemann
Victorian Bioinformatics Consortium, Monash University

Dr Tim Stinear
Dept. Microbiology & Immunology, Melbourne University

VLSCI Capacity Building Meeting - Fri 29 July 2011

MONASH University

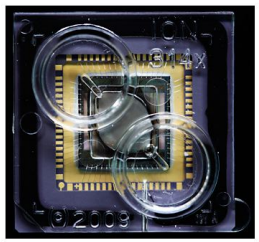
Victorian Bioinformatics Consortium

THE UNIVERSITY OF MELBOURNE

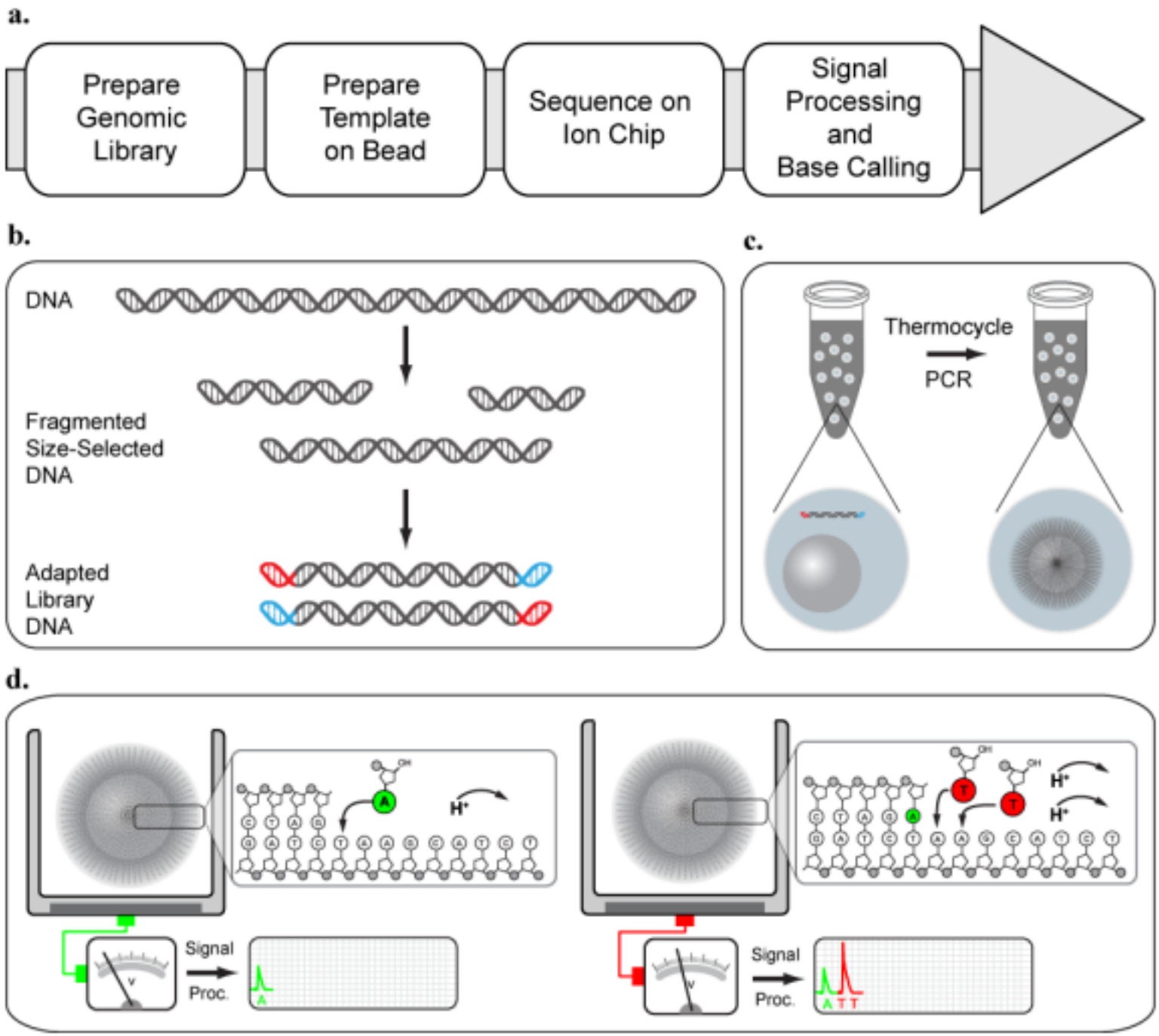
What is "Ion Torrent" ?

ion torrent
by life technologies

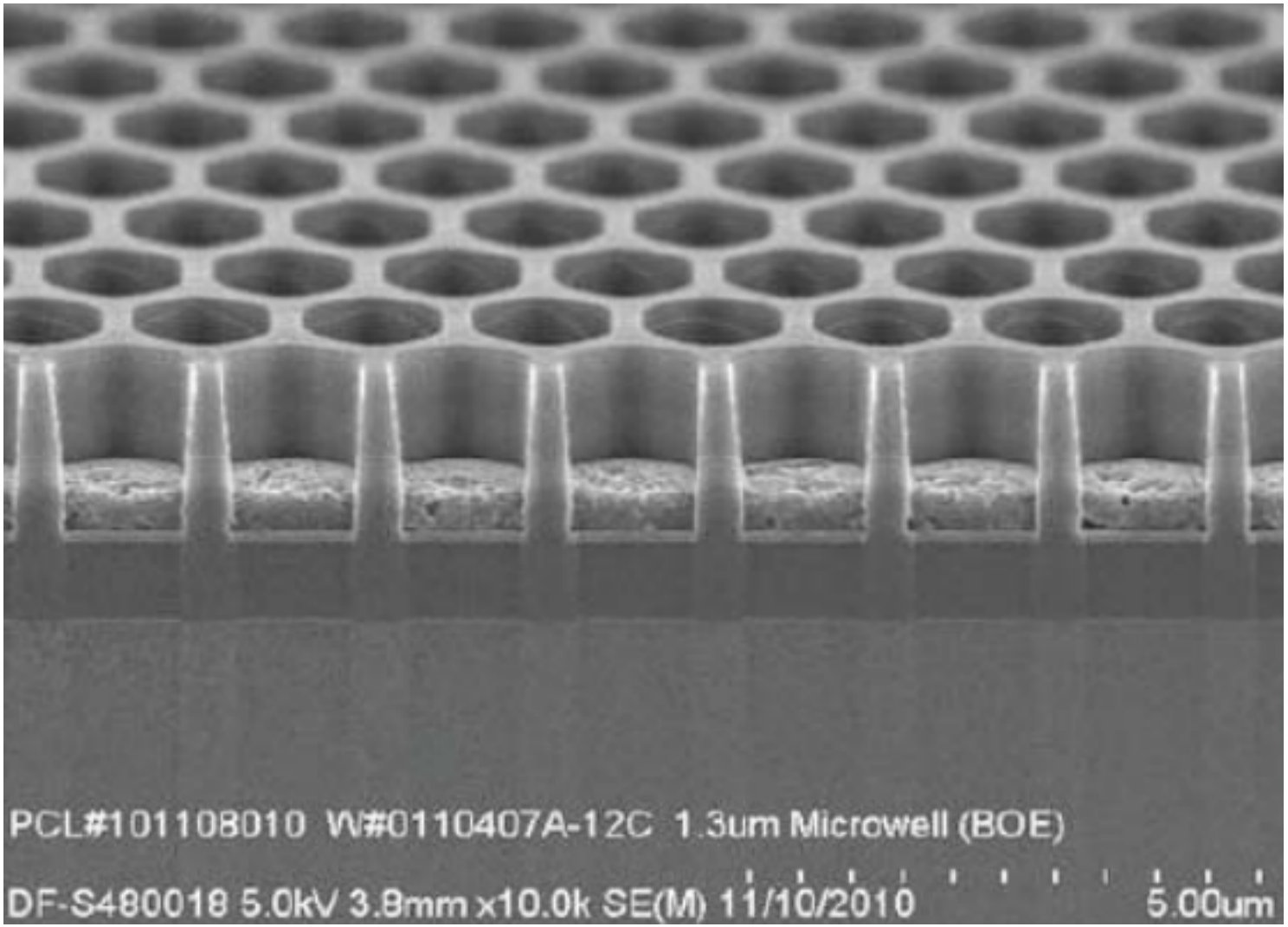
- Another next-generation DNA sequencing instrument
  - PGM - "Personal Genome Machine"




```
@SRR006511.105 8_1_663_27 length=36
ATAGCGGCACTGTTGGTTCGCTTGTTCCTTTGAGTC
+
IIII7II-9/0;+8I<03.*%-,&"+'($,#"''&"
@SRR006511.112 8_1_829_108 length=36
AGAATTTTATGTATCTGGATGCAATAAAAAATGATG
+
II6IIIIIIIIIIIIIII1DII>O<I7>069;64(*%
@SRR006511.490 8_1_351_672 length=36
AGCACCCGCCGTGTGTCCCCCATGCTCCACACCTCT
+
IO>OA,I2H):$)6)#4$.)>'.&.$)"E7*1I)&&
@SRR006511.632 8_1_79_187 length=36
ATGCCGAAAGGTATCGGTAAACCGTTGAAATTCTTC
+
IIIIII<I;II57G;II.IO++32.--)$32++9),
@SRR006511.726 8_1_300_437 length=36
ACCACGTGGACTTCCAGGACCATGAGGCCAAATTGG
+
I1B>:IIII)3,I&0-,:$(%&%1$+1"&(8%"&#"
```


a.
Prepare Genomic Library
Prepare Template on Bead
Sequence on Ion Chip
Signal Processing and Base Calling
b.
DNA
Fragmented Size-Selected DNA
Adapted Library DNA
c.
Thermocycle
PCR
d.
Signal Proc.
Signal Proc.

The chip up close



A scanning electron micrograph of a large array of 1.3 μm wells

Chip specifications

Chip	Nominal Yield	Typical Yield (on good day)	Read length	Current cost
312	1 Mbp	3 Mbp	~100	n/a
314	10 Mbp	25 Mbp	~100	$100
316	100 Mbp	220 Mbp	>100	$600
318	1 Gbp	?	~200 ?	?
320	?	?	?	?

Assessing output sequence

- Yield
  - The total number bp of sequence
- Count
  - How many reads there were
- Length
  - The distribution of read lengths: mode, mean, ...
- Quality
  - The distribution of quality scores across the read

- And then the same *after* quality filtering and trimming
  - the actual usable/useful/trustworthy sequence!

Phred qualities

- Base quality is reported as:
  - a "Q" value (also known as a "phred" quality)
  - encoded in FASTQ files using a single ASCII character

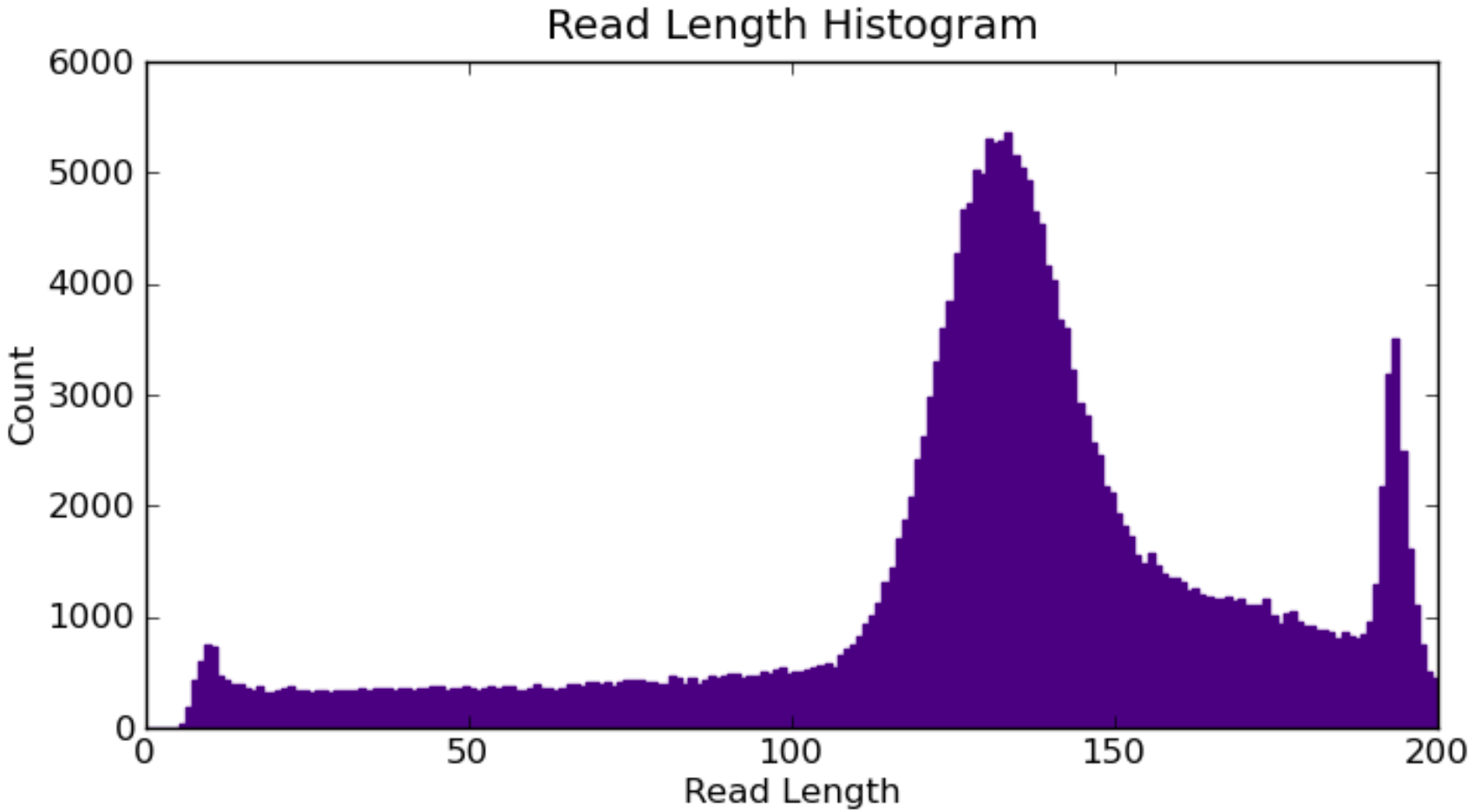
- Represents the estimated probability of error:
  - $Q = -10 \log_{10} P$
  - $P = 10 \wedge (-Q/10)$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

The 10Mb chip - Yield

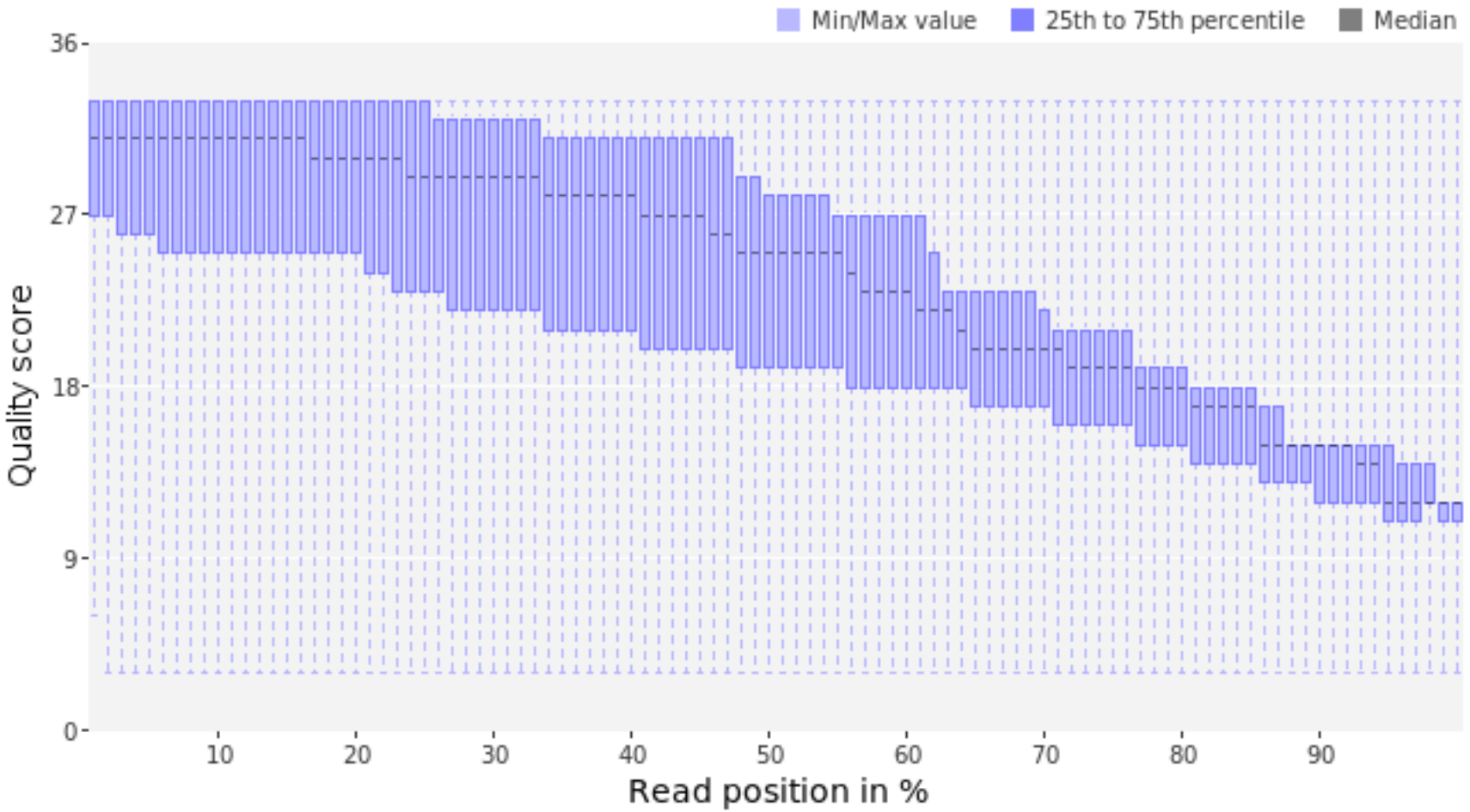
- Ion Torrent data release
  - 49 Mb for control E.coli K-12 DH10B (0.5M reads)
- IMB (Brisbane, AU)
  - typically got 20-30 Mbp (human PCR + bacteria)
- Uni. Birmingham (UK)
  - typically got ~25 Mbp too (bacteria)
- Melbourne Uni (AU)
  - Bad days: 2, 4, 3, 1, 3, 4 Mb **:-(**
  - Good days: 31, 26, 31, 60, 30, 31 Mbp **:-)**

The 10Mb chip - Length



- This is before filtering!
  - the mode of 125bp will end up about 100bp post-filtering
  - the 190bp peak is hompolymer gumpf

The 10Mb chip - Quality



- Quality profile is similar to 454 and Illumina
  - poor at 3′ end of the read
  - 5' is pretty good, but this is the *E.coli* post-filtered
  - Q10 means 1 in 10 chance of error!

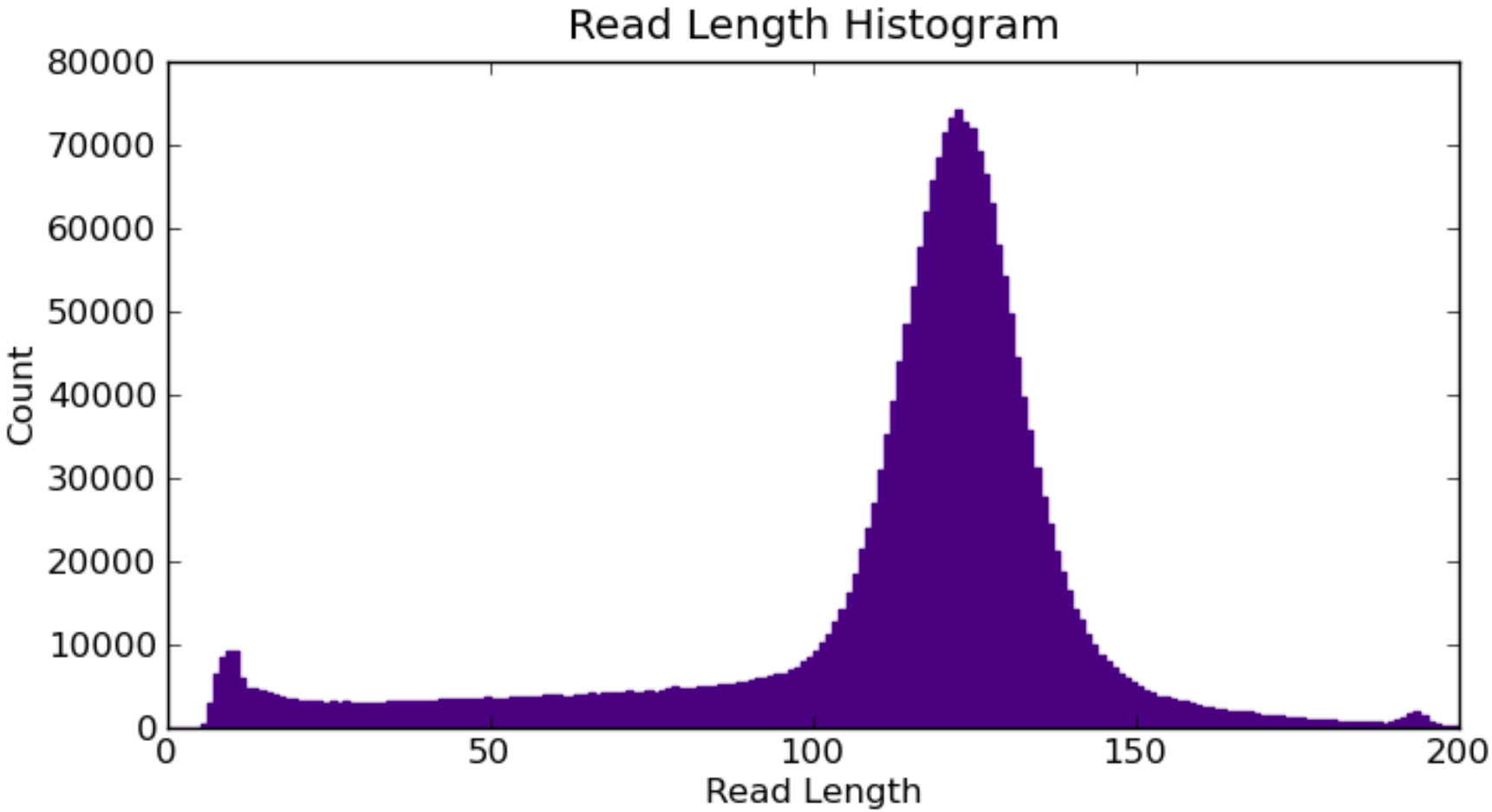
The 100Mp chip - Yield

- IMB Brisbane (AU)
  - Beta customer
  - Getting about ~220 Mb per run (cancer + bacteria)

- Uni. Birmingham (UK)
  - Run 1 - 251 Mb
  - Run 2 - 209 Mb

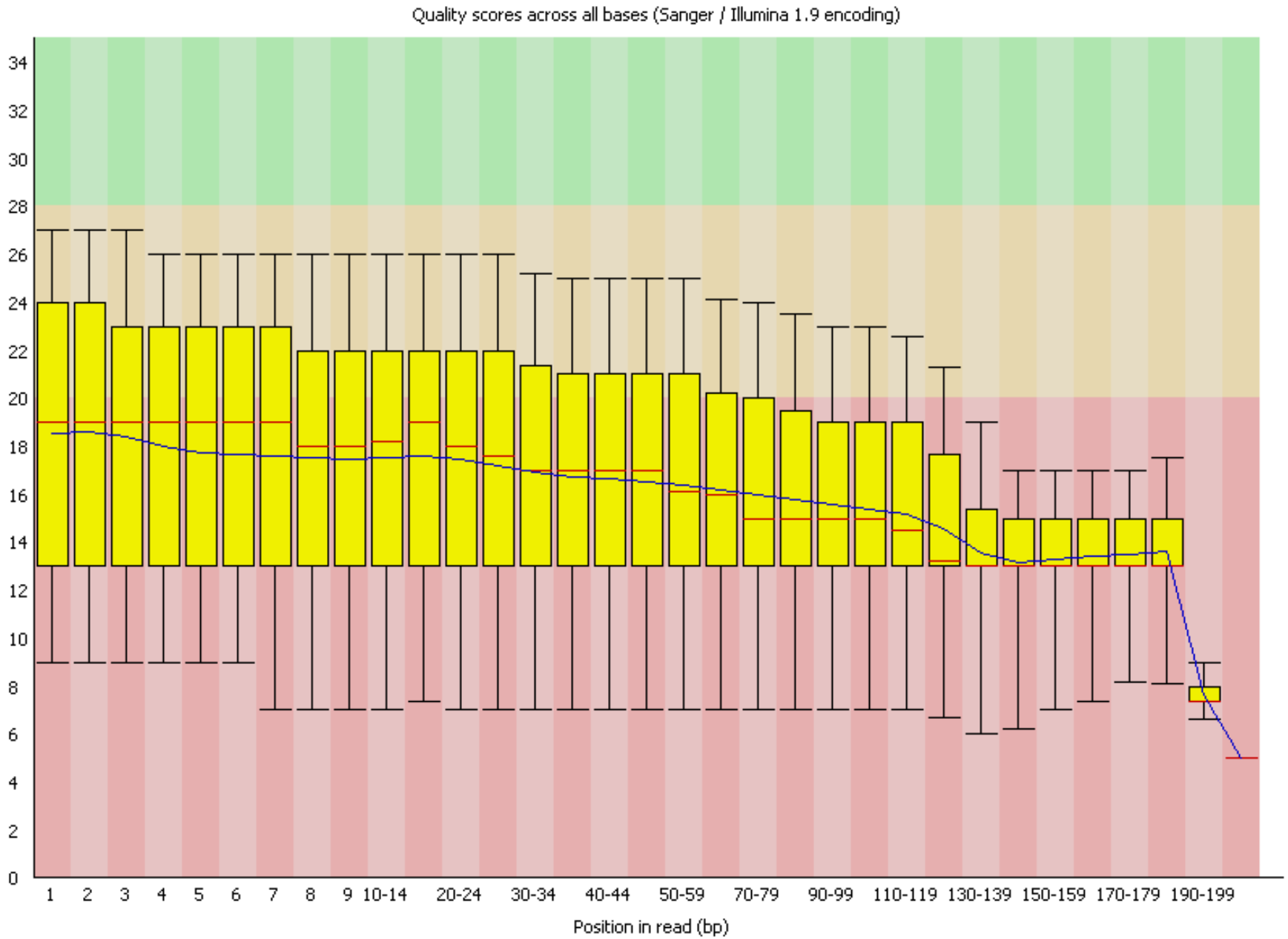
Melbourne Uni
  - Not available yet

The 100Mb chip - Length



- Mode ~ 125 bp but will be ~100bp after trimming

The 100Mb chip - Quality



Analysis - read mapping

- Need to use an aligner that:
  - handles multiple indels properly
  - works with un-paired reads
- Recommendations:
  - Work: TMAP, SHRiMP, BFAST, CLC
  - Might work: BWA, NovoAlign
  - Probably won't work: MAQ
- Nesoni
  - Our bacterial swiss-army knife tool inc. SNP calling
  - Seems to work fine with Ion data
    - uses SHRiMP v2 under the hood for alignment

Analysis - *De novo* assembly

- Very similar to original Roche GS20 data
  - ~100bp length single end reads
  - homopolymer errors
  - quality issues at read ends

- Software that "works"
  - Newbler (as Ion uses .SFF flowgrams too)
  - CLC Genomics Workbench (V4, maybe not V3)
  - Mira (author has tweaked it to handle PGM data)

- Results
  - not very good, need longer reads + paired-end protocol

Applications - pooled PCR products

- Verification of SNPs called from high(er) throughput data
- Previously, *per SNP*:
  - design oligoes around site
  - generate a PCR product
  - capillary sequence (Sanger) the PCR
- Now, *pooled*:
  - design oligoes
  - generate all PCR products and pool
  - Ion Torrent together
  - *de novo* assemble the result
  - get a contig for each PCR product!

The costs

- Hardware
  - $100k for PGM, Dell server, iPod Touch
  - $30k for oneTouch etc. (simpler lab prep)
  - $100-$800 per chip (excludes labour)
  - $5k+ for another workstation to do analysis on
- Software
  - $0 - comes with some basic mapping tools
  - $0 - open source Unix tools available
  - $5k - CLC Genomics Workbench
- Wages
  - $??k pa - cover time of your lab research assistant
  - $100k pa - bioinformatician in your lab

Conclusions

- Will democratize sequencing
- Well suited to microbial labs (due to lower yield)
- Applications in pathology

- Needs a mate-pair protocol
- Will need a multiplexing protocol

- Will be challenged by Illumina's MiSeq (Oct 2011?)
- Will be challenged by PacBio and others

References

- Nick Loman's blog
  - *http://pathogenomics.bham.ac.uk/blog/*

- Offical paper
  - Jonathan M. Rothberg et al. An integrated semiconductor device enabling non-optical genome sequencing. **Nature** 475, 348–352 (21 July 2011)

- Ion Torrent website:
  - http://www.iontorrent.com/