

NGS sequencing technologies



Swiss Institute of
Bioinformatics

Sylvain Pradervand
March 21, 2015

Course outline

1. History of sequencing technologies
2. Illumina sequencing
3. PacBio sequencing
 - Technology description
 - Data characteristics and quality assessment
 - PacBio applications
4. Oxford Nanopore sequencing
 - Technology description
 - Data from MinION access program

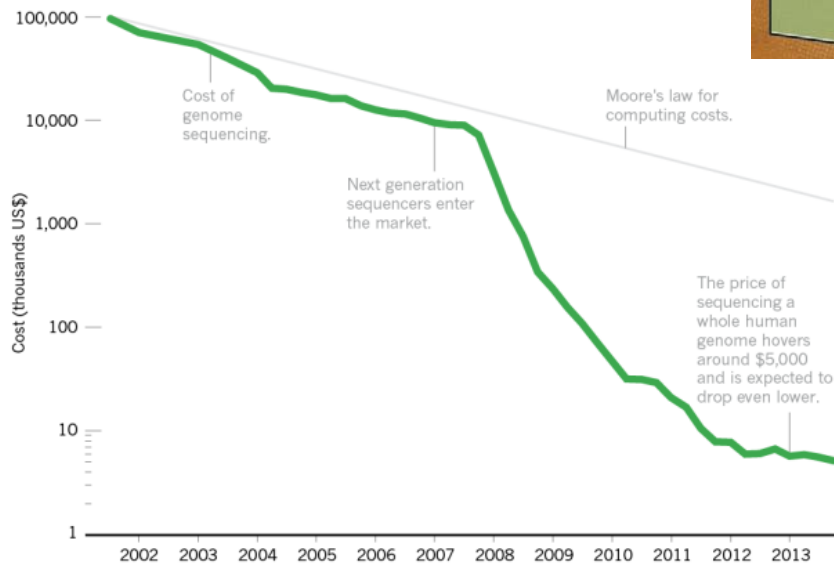


Swiss Institute of
Bioinformatics

The \$1,000 genome

Falling fast

In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore's law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.



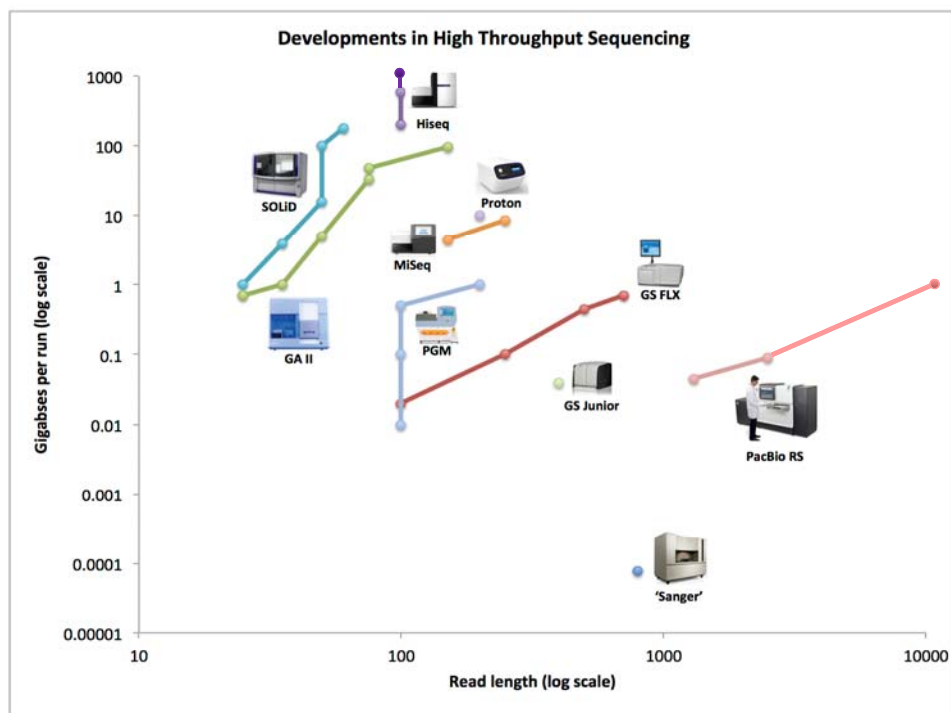
© 2009 SIB

Hayden, Nature 507, 294–295 (2014) doi:10.1038/507294a



Swiss Institute of
Bioinformatics

Sequencing technologies



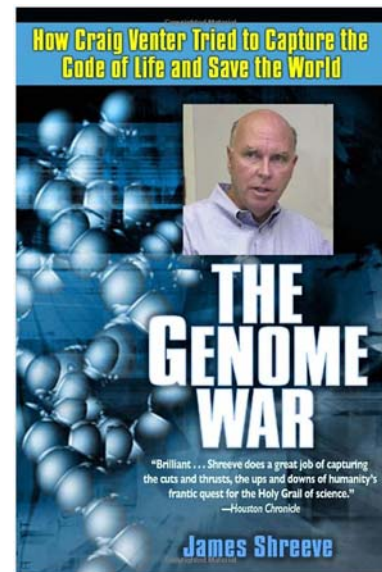
© 2009 SIB



Swiss Institute of
Bioinformatics

Sequencing landmarks (20th century)

- 1977
 - Maxam-Gilbert, Sanger sequencing
- 1986
 - ABI 370: Slab gel sequencer. ~5'000 bases/day (Hunkapiller and Hood)
- 1995
 - ABI 377: up to 96 lanes and 19'000 bases/day
- 1998
 - ABI 3700: 96 capillary sequencer. Over 400'000 bases/day

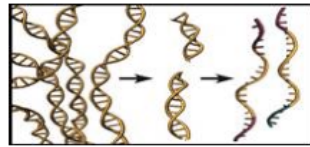


Current sequencing rate: > 60 Gbases /day (= 150'000 x 1998 rate)

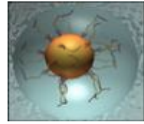
Sequencing landmarks (21th century)

- 2005
 - 454 Life Sciences, Pyrosequencing, 1M reads, 250bp, O/N
- 2006
 - Solexa 1G, 33M reads, 30bp, 3 days
- 2010
 - Ion Torrent, 1M, 100bp, 2 hours
- 2010
 - HiSeq, 600M reads, 100bp, 4 days
- 2011
 - PacBio, 25'000 reads, 2000bp, 90 min
- 2014
 - Oxford Nanopore MinION, 20'000 reads, 5000bp, 6 hours

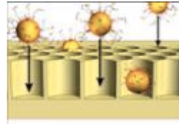
The 454 Life Sciences* technology



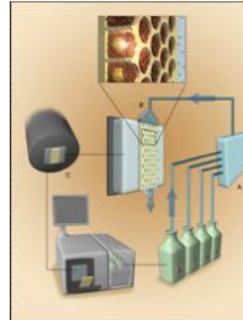
1) Prepare Adapter Ligated ssDNA Library



2) Clonal Amplification on 28 µm beads



3) Load beads and enzymes in PicoTiterPlate™



4) Perform Sequencing by synthesis on the 454 Instrument

- sequential flows of dA, dC, dG and dT
- base incorporation(s) produce light
- light intensity from each bead detected by CCD camera

454 sequencers will be phased-out mid 2016

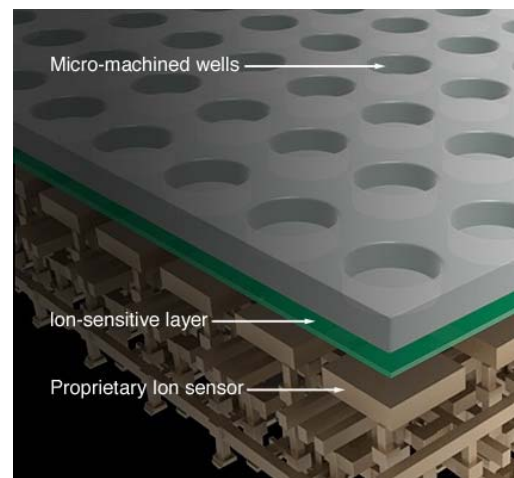
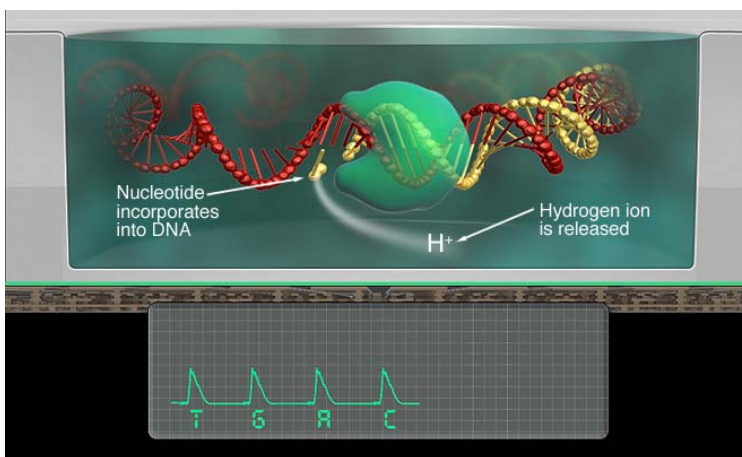
© 2009 SIB

**Company purchased by Roche in March 2007*



Swiss Institute of
Bioinformatics

Ion Torrent* and benchtop sequencers



© 2009 SIB

**Company purchased by Life Technologies in August 2010*






Swiss Institute of
Bioinformatics

Ion Torrent sequencers

Ion PGM



Chip	Expected sequencing run time			Expected output*		
	35-base reads	200-base reads	400-base reads	35-base reads	200-base reads	400-base reads
 Ion 314 [™] Chip	0.5 hr	2.3 hr	3.7 hr	3 Mb	20 Mb	40 Mb
 Ion 316 [™] Chip	0.7 hr	3.0 hr	4.9 hr	30 Mb	200 Mb	400 Mb
 Ion 318 [™] Chip	0.9 hr	4.4 hr	7.3 hr	300 Mb	500 Mb–1 Gb	1–2 Gb

1 million wells

6 million wells

11 million wells

*Expected output with >99% aligned/measured accuracy. Output is dependent on read length and application.

Ion Proton



Ion Proton [™] System performance specifications with Ion PI [™] Chip	
Throughput	Up to 10 Gb (Note: Ion PI [™] Chip* will be available about six months after the Ion PI [™] Chip. Ion PI [™] Chip* will enable sample-to-variant analysis of a human genome in a single day, at up to 20x coverage.)
Read length	Up to 200-base fragment reads
Number of reads passing filter	60–80 million reads passing filter
Sequencing run time	2–4 hours

PI: 165 million wells, PII: 660 million wells

© 2009 SIB



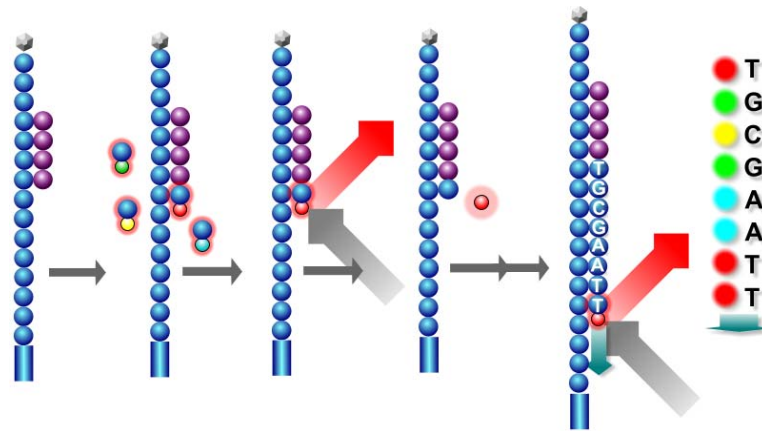
Course outline

1. History of sequencing technologies
2. Illumina sequencing
3. PacBio sequencing
 - Technology description
 - Data characteristics and quality assessment
 - PacBio applications
4. Oxford Nanopore sequencing
 - Technology description
 - Data from MinION access program

© 2009 SIB



The Solexa* technology



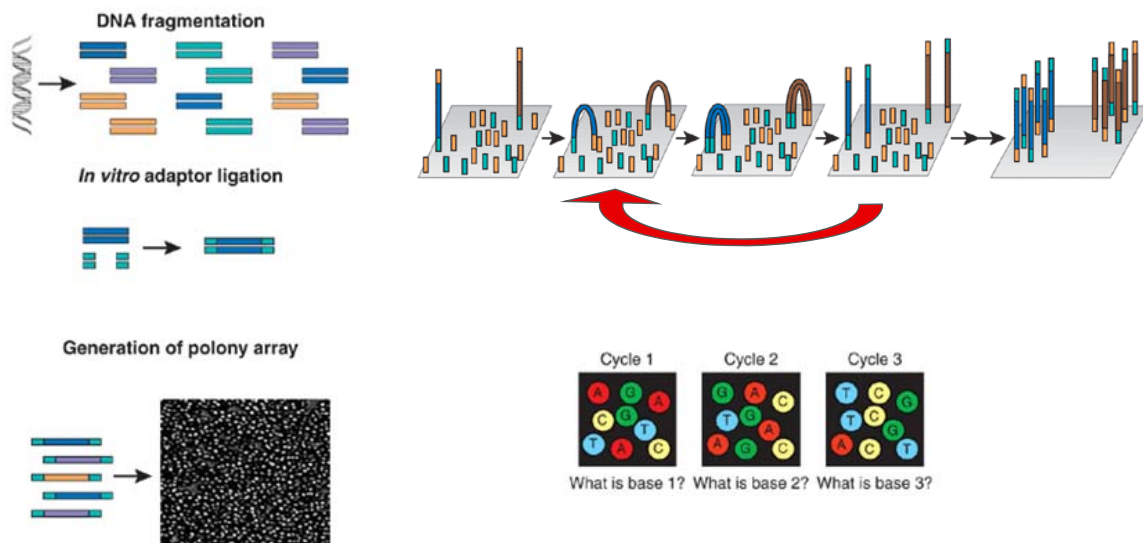
© 2009 SIB

*Company purchased by Illumina in November 2006



Swiss Institute of
Bioinformatics

Illumina/Solexa technology advantage: solid-support cluster amplification



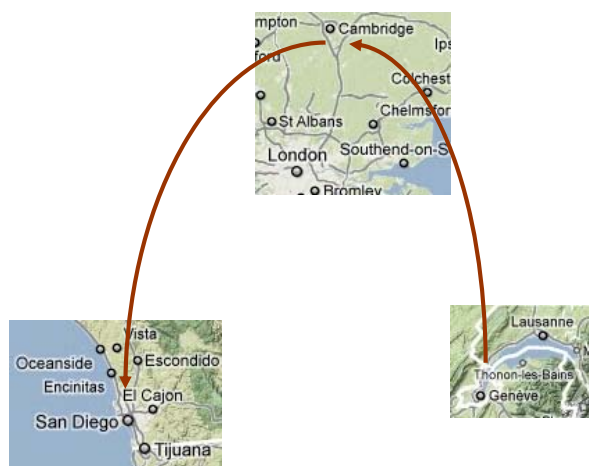
From Jay Shendure & Hanlee Ji
Nature Biotechnology 26 (2008)

© 2009 SIB



Swiss Institute of
Bioinformatics

- Cluster amplification developed by L.Farinelli et al. at Glaxo in Geneva !
- Technology developed by Manteia (spin-off of Serono)
- Manteia technology was sold to Solexa Ltd in 2003



USPTO PATENT FULL-TEXT AND IMAGE DATABASE

[Home](#)
[Quick](#)
[Advanced](#)
[Pat Num](#)
[Help](#)
[Bottom](#)
[View Cart](#)
[Add to Cart](#)
[Images](#)

(1 of 1)

United States Patent

7,985,565

Mayer, et al.

July 26, 2011

Method of nucleic acid amplification

Abstract

A nucleic acid molecule can be annealed to an appropriate immobilized primer. The primer can then be extended and the molecule and the primer can be separated from one another. The extended primer can then be annealed to another immobilized primer and the other primer can be extended. Both extended primers can then be separated from one another and can be used to provide further extended primers. The process can be repeated to provide amplified, immobilized nucleic acid molecules. These can be used for many different purposes, including sequencing, screening, diagnosis, in situ nucleic acid synthesis, monitoring gene expression, nucleic acid fingerprinting, etc.

Inventors: Mayer; Pascal (Geneva, CH), Farinelli; Laurent (Vevey, CH), Kawashima; Eric H. (Geneva, CH)

Assignee: Illumina, Inc. (San Diego, CA)

Appl. No.: 10/449,010

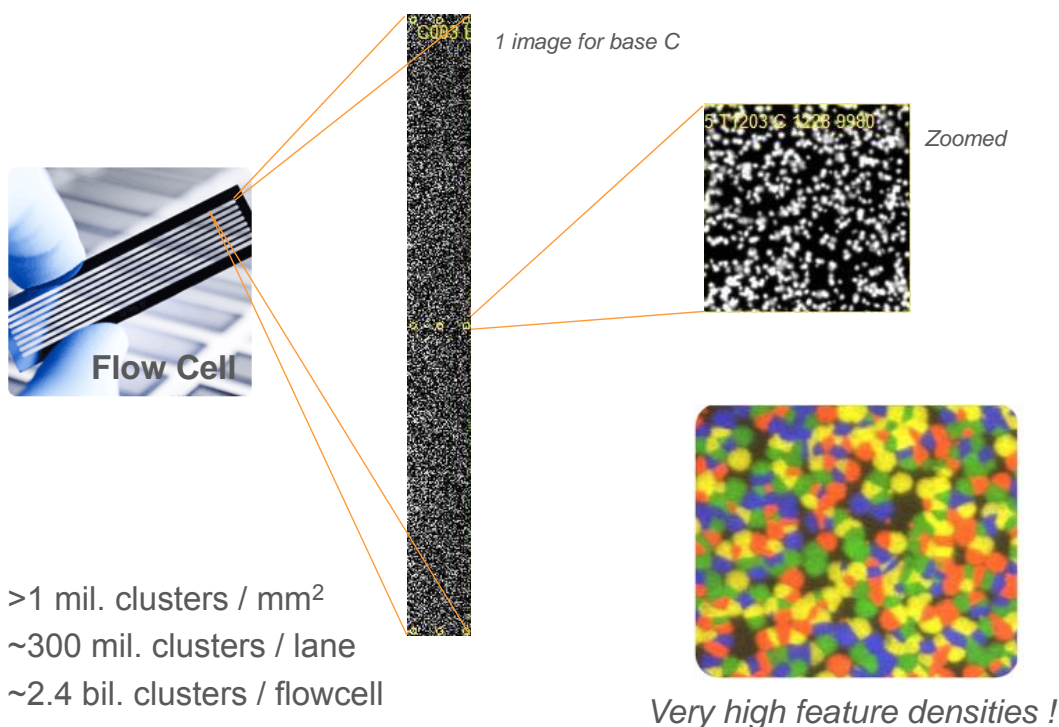
Filed: June 2, 2003

© 2009 SIB

SIB

Swiss Institute of
Bioinformatics

The real massively parallel sequencing

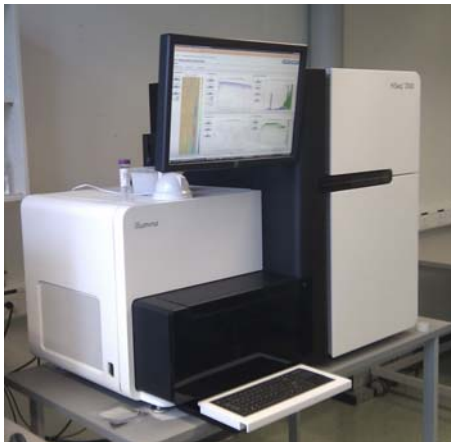


© 2009 SIB

SIB



Swiss Institute of
Bioinformatics

1 HiSeq run 2x125 bp = 6.5x the data published for the Human genome



- Human genome published in 2001:
65 fold mean coverage = $65 \times 3.4 \text{ Gb}$
= 221 Gb
- 1 HiSeq run (2 flow cells) in 2015:
1000 Gb in 6 days

2015 Illumina benchtop sequencers

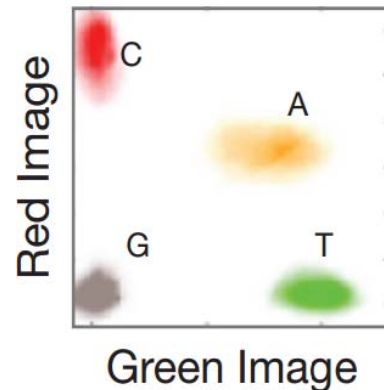
Key Methods	Small genome, amplicon, and targeted gene panel sequencing.	Everyday genome, exome, transcriptome sequencing, and more.	
	 MiSeq	 NextSeq 500	
Run Mode	N/A	Mid-Output	High-Output
Flow Cells per Run	1	1	1
Output Range	0.3-15 Gb	20-39 Gb	30-120 Gb
Run Time	5-55 hours	15-26 hours	12-30 hours
Reads per Flow Cell†	25 million‡	130 million	400 million
Maximum Read Length	2 x 300 bp	2 x 150 bp	2 x 150 bp

NextSeq: 2 Channels imaging

- HiSeq, MiSeq use four-channels sequence-by-synthesis (SBS) with a unique fluorescent dye for each of the 4 bases. Four images are necessary at each cycle.
- NextSeq use two-channels SBS with only two dyes used and 2 images necessary at each cycle.



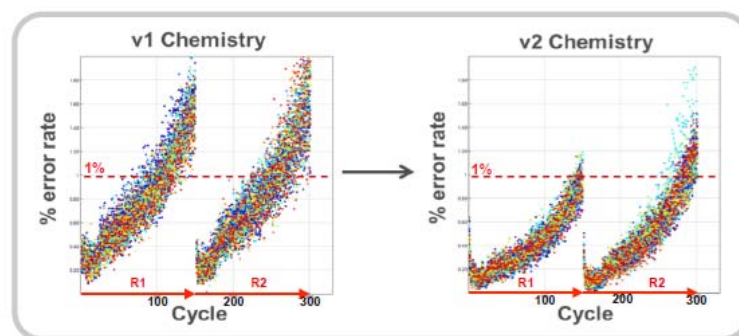
- Different basecalling
- System needs to be properly benchmarked



© 2009 SIB



NextSeq 500 chemistry








	v1 Chemistry	v2 Chemistry	HiSeq v4 Chemistry
% >Q30	81.6	87.9	% >Q30 = 90-95%
Error rate (%), PhiX	0.70	0.37	Error rate=0.28
SNV sensitivity	95.8%	95.8%	
SNV precision	99.8%	99.8%	

NHL-16 prepared with TruSeq Nano 550 bp insert for WGS; also works on 350 bp inserts

- Contact an Illumina representative for access to these data sets
- Above analysis used Illumina Sequence Analysis Viewer and BWA Whole Genome Sequencing BaseSpace App 1.0
- Sensitivity/Precision metrics calculated relative to NIST Genome In a Bottle reference samples

2015 Illumina HiSeq sequencers

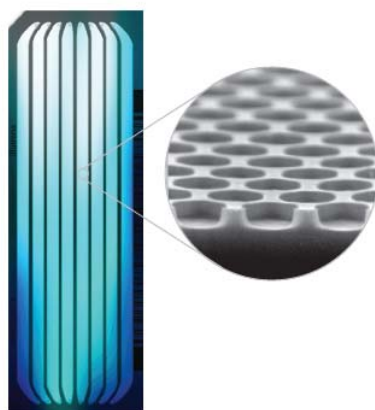
Key Methods	Production-scale genome, exome, transcriptome sequencing, and more.				Population- and production-scale human whole-genome sequencing.	
	 HiSeq 2500	 HiSeq 3000	 HiSeq 4000		 HiSeq X Five*	 HiSeq X Ten*
Run Mode	Rapid Run	High-Output	N/A	N/A	N/A	N/A
Flow Cells per Run	1 or 2	1 or 2	1	1 or 2	1 or 2	1 or 2
Output Range	10-300 Gb	50-1000 Gb	125-750 Gb	125-1500 Gb	900-1800 Gb	900-1800 Gb
Run Time	7-60 hours	<1-6 days	<1-3.5 days	<1-3.5 days	<3 days	<3 days
Reads per Flow Cell†	300 million	2 billion	2.5 billion	2.5 billion	3 billion	3 billion
Maximum Read Length	2 x 250 bp	2 x 125 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp	2 x 150 bp

© 2009 SIB

<http://www.illumina.com/systems/sequencing.html>



Patterned Flow Cell



- Equipped HiSeq 3000, 4000, X Five, X Ten
- Nanowells at fixed locations
- Even cluster spacing, uniform feature size
 - Extremely high densities
- Exclusion amplification clustering
 - Only one single DNA template per well

Figure 2: Advanced Patterned Flow Cell Design Enables Ultra-High Throughput. Patterned flow cells contain billions of nanowells at fixed locations, providing even cluster spacing and uniform feature size to deliver extremely high cluster density.

© 2009 SIB



Course outline

1. History of sequencing technologies
2. Illumina sequencing
3. **PacBio sequencing**
 - Technology description
 - Data characteristics and quality assessment
 - PacBio applications
4. Oxford Nanopore sequencing
 - Technology description
 - Data from MinION access program

What is PacBio Sequencing?

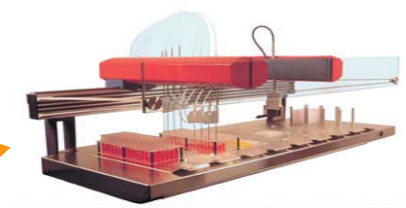
- **Single molecule sequencing**
 - no local amplification of template DNA required
 - sequences reach >35kb maximum
- **Real time monitoring** of the sequencing reaction
 - incorporation of fluorochrome labelled nucleotides by the polymerase is recorded on a movie over 240 min



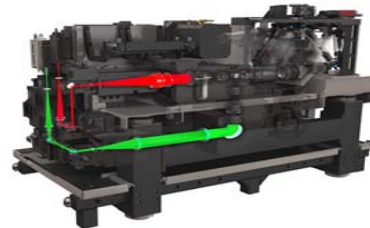
PacBio RS II sequencer



Blade server
(Automated primary
analysis pipeline)



Liquid handler



Laser bench

© 2009 SIB



Swiss Institute of
Bioinformatics

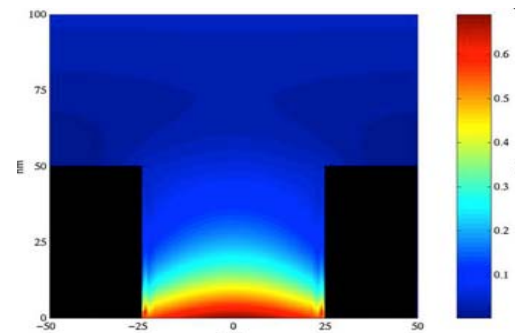
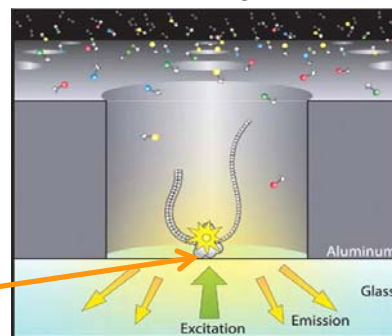
Single Molecule Real-Time



The sequencing
unit **SMRT Cell**

A single DNA
polymerase at
bottom

150'000
Zero Mode Waveguides



window to observe
DNA sequencing in real-time



incorporation rate 3nt/sec (Illumina : 1nt/h)

© 2009 SIB



Swiss Institute of
Bioinformatics

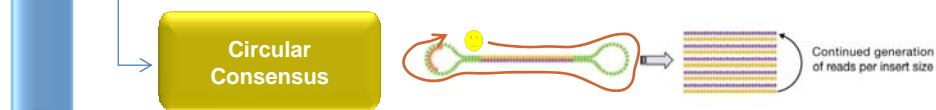
PacBio reads

CLR – Continuous long read



- Large insert sizes (2kb-35kb)
- Generates **one pass** on each molecule sequenced

CCS – Circular consensus sequence

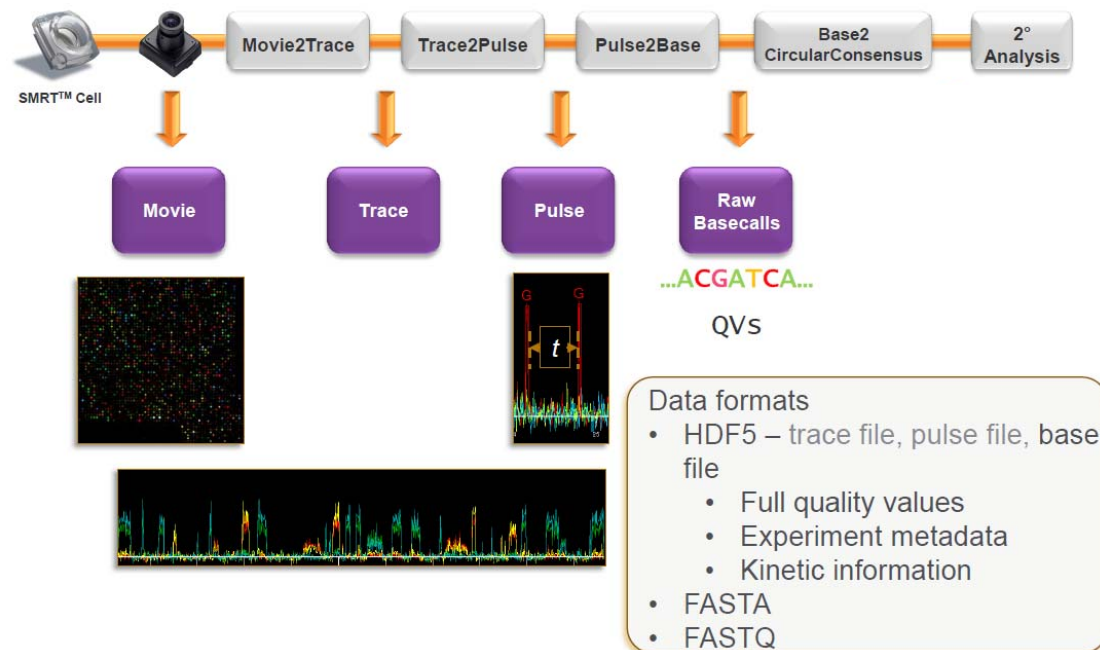


- Small insert sizes < 1kb
- Generates **multiple passes** on each molecule sequenced
- Error corrected consensus sequence

Course outline

1. History of sequencing technologies
2. Illumina sequencing
3. **PacBio sequencing**
 - Technology description
 - **Data characteristics and quality assessment**
 - PacBio applications
4. Oxford Nanopore sequencing
 - Technology description
 - Data from MinION access program

Primary data processing



© 2009 SIB

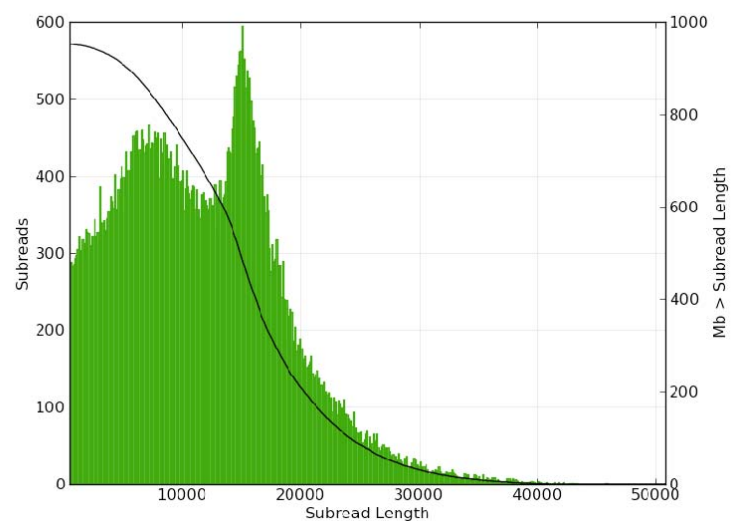


Swiss Institute of Bioinformatics

Sequencing data throughput

SMRT Cells: 1 Movies: 1

Job Metric	Value
Polished Contigs	1
Adapter Dimers (0-10bp)	0.0%
Short Inserts (11-100bp)	0.0%
Number of Bases	958,423,236
Number of Reads	59,834
N50 Read Length	21,712
Mean Read Length	16,018
Mean Read Score	0.85
Mapped Reads	57,458
Mapped Read Length of Insert	10,740
Average Reference Length	3,652,649
Average Reference Bases Called	100.0%
Average Reference Consensus Concordance	100.0%
Average Reference Coverage	230.17



Mean mapped subread length: 11,152 bases

© 2009 SIB



Swiss Institute of Bioinformatics

PacBio Sequencing Accuracy

- PacBio single-pass sequence reads in SMRT Sequencing are error-prone, with a median error of ~11%, predominantly deletions or insertions.
- However, PacBio sequencing achieves highly accurate sequencing results, exceeding 99.999% (Q50) accuracy, regardless of the DNA's sequence context or GC content. This is possible because:
 1. Consensus accuracy
 2. Sequence context bias
 3. Mapability of sequence reads



Swiss Institute of
Bioinformatics

Consensus accuracy

- i. Generate sequence read:

◀ CGAATTCCTTAACGTCTGAGACACGACATCGACCTCTGCACCGGACTCGTCGGCGTTCTTTGGCAATCGGGATCAGCTTCGGGAGATGCGCGCAGCTTGGGATGGATAGCGGACCAATGC

- ii. Map to reference:

<GGAATTCCTAAAGTCTGAGACACGACAGCGAAGCTCTGCGGAGCTGCTGGCGGTCTTTGGCAATCGGGATCTCAGCTTGCGGAGATGTCGGGCGCACTTGCGGAGATATAGCGAGCAATGCG
 <GGAATTCCTAAAGTCTGAGACACGACAGCGAAGCTCTGCGGAGCTGCTGGCGGTCTTTGGCAATCGGGATCTCAGCTTGCGGAGATGTCGGGCGCACTTGCGGAGATATAGCGAGCAATGCG

- iii. Generate consensus (10x coverage):

Figure 1: Schematic representation of the reference match, heterozygous SNP, and homozygous SNP. The figure shows three DNA sequences with corresponding labels below. The first sequence is labeled 'Reference match' and shows a green box around a 'G' in the first position. The second sequence is labeled 'Heterozygous SNP' and shows a red box around a 'G' in the first position and a 'C' in the second position. The third sequence is labeled 'Homozygous SNP' and shows a red box around a 'C' in the first position. The sequences are: 1. GGAATCTTAAAGTCTGTTGAGACGACAGCGACGACCTCTGAGGTTGCTGCTTCGCGCTTTTGGACAATCGGGATTCAGACTTGGGATATGCGGCGGCGAGCGCTTGGGATGATAGTAGGCGGCAATGCG. 2. GGAATCTTAAAGTCTGTTGAGACGACAGCGACCTCTGAGGTTGCTGCTTCGCGCTTTTGGACAATCGGGATTCAGACTTGGGATATGCGGCGGCGAGCGCTTGGGATGATAGTAGGCGGCAATGCG. 3. GGAATCTTAAAGTCTGTTGAGACGACAGCGACCTCTGAGGTTGCTGCTTCGCGCTTTTGGACAATCGGGATTCAGACTTGGGATATGCGGCGGCGAGCGCTTGGGATGATAGTAGGCGGCAATGCG.

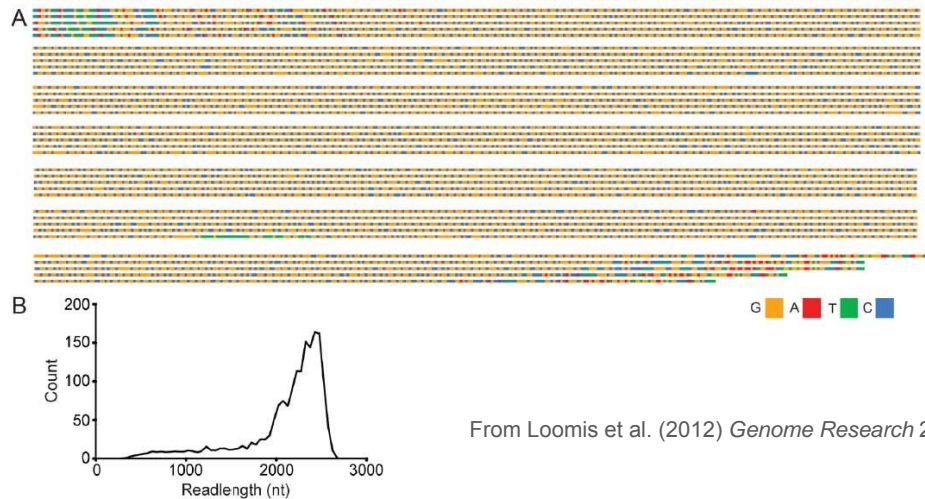
- **Systematic** error in a sequencing method will affect whether the consensus sequence can be determined correctly
- In SMRT Sequencing, errors are distributed **randomly**, which means that they wash out very rapidly upon building consensus



Swiss Institute of
Bioinformatics

Sequence context bias

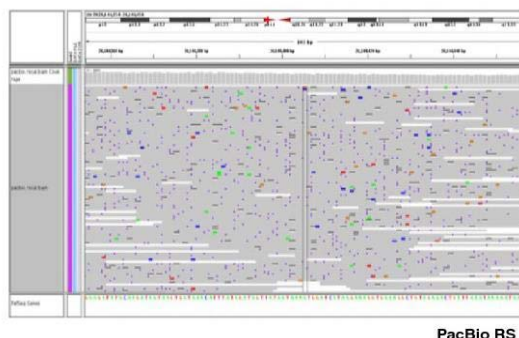
- Many sequencing systems have difficulties sequencing through extremely AT-rich or GC-rich regions, highly repetitive sequences, or long homonucleotide stretches.
- SMRT Sequencing does not exhibit such sequence context bias and performs very uniformly, even through regions previously considered difficult to sequence



Mappability of sequence reads



SYSTEMATIC ERROR



RANDOM ERROR

Same region on both dataset

Carneiro et al. (2012) *BMC Genomics* 13: 375-383

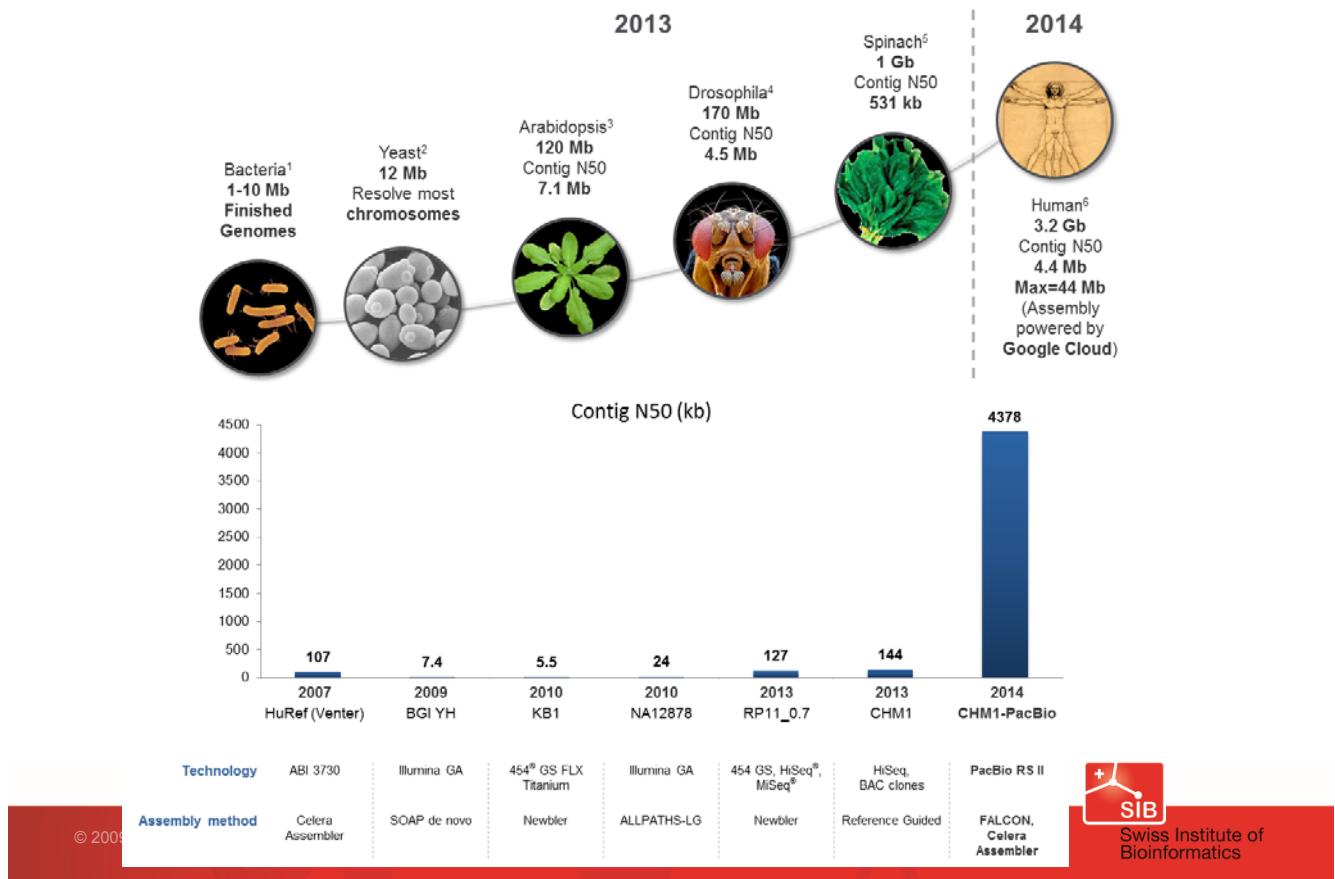
Course outline

1. History of sequencing technologies
2. Illumina sequencing
3. **PacBio sequencing**
 - Technology description
 - Data characteristics and quality assessment
 - **PacBio applications**
4. Oxford Nanopore sequencing
 - Technology description
 - Data from MinION access program

PacBio Sequencing Applications

- *de novo* genome sequencing
- Genome contigs scaffolding and gap closing
- Resequencing (amplicons)
- Epigenetic base modifications and methylation
- Transcriptome isoforms characterization (Iso-seq)

De novo genome assembly with PacBio



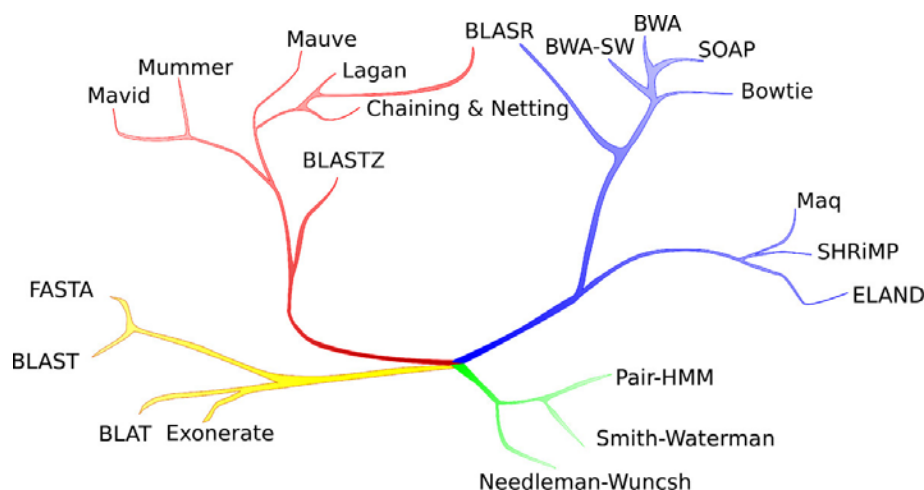
PacBio for human WGS

- Short-read sequencing struggles with many regions in the genome: repetitive, high GC, highly homopolymeric
- With PacBio, major histocompatibility complex region was entirely contained in one 9Mb contig
- Asian-specific reference genome (Asian Genome Project):
 - *de novo* sequencing using PacBio in combination with BAC clones worked best
- PacBio able to detect recombinations and break-point with *de novo* sequencing of breast cancer cell line
- J. Korlach: “by the end of the year, with improvements to read length and throughput, the cost will drop to \$10,000 to generate a reference medical grade *de novo* genome”

Amplicons sequencing with PacBio

- Pros
 - No systematic errors: with sufficient coverage less false positive
 - Get even coverage: no GC bias, all regions sequenced equivalently
 - Very long read: get structural variants and haplotypes
 - Orthogonal method to Illumina sequencing
- Cons
 - Need high sample quality (purity AND integrity)
 - High error rate with low coverage
 - Lower throughput than other Illumina orthogonal methods (e.g. Ion Torrent)
 - Bioinformatics tools not as mature as for short reads
 - Problem of false positive indels

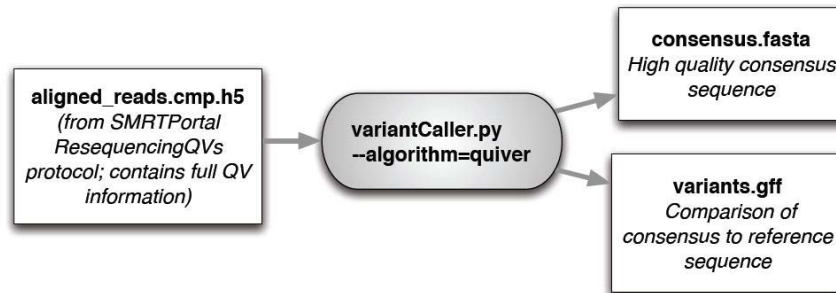
BLASR: Basic Local Alignment with Successive Refinement



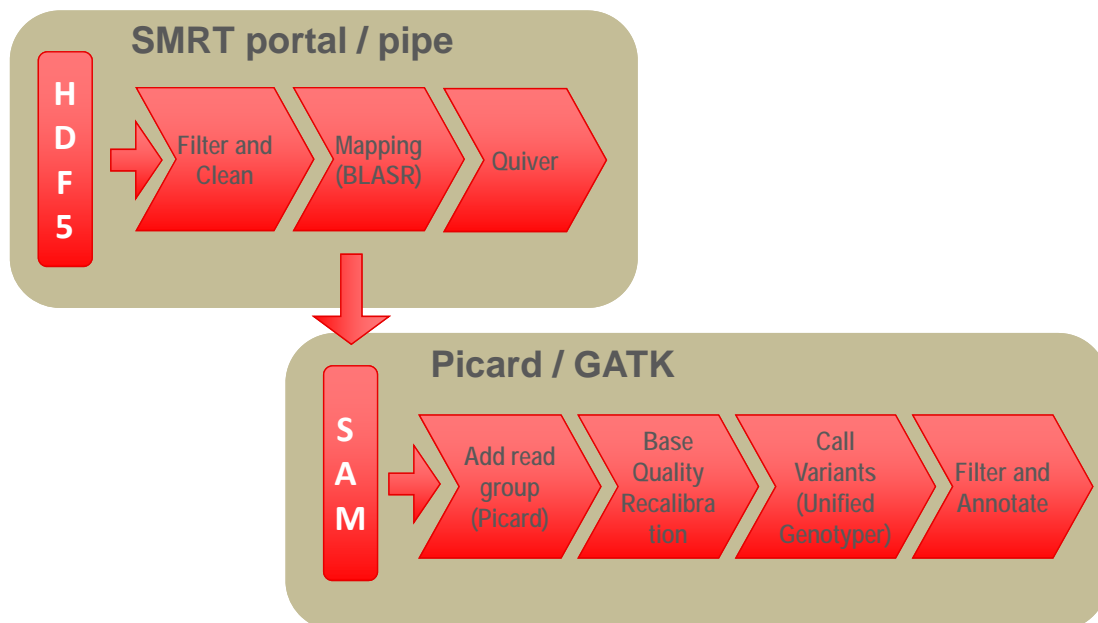
- Optimized for high indels rate
- Combines data structures from short read alignment with optimization methods from whole genome alignment

Quiver

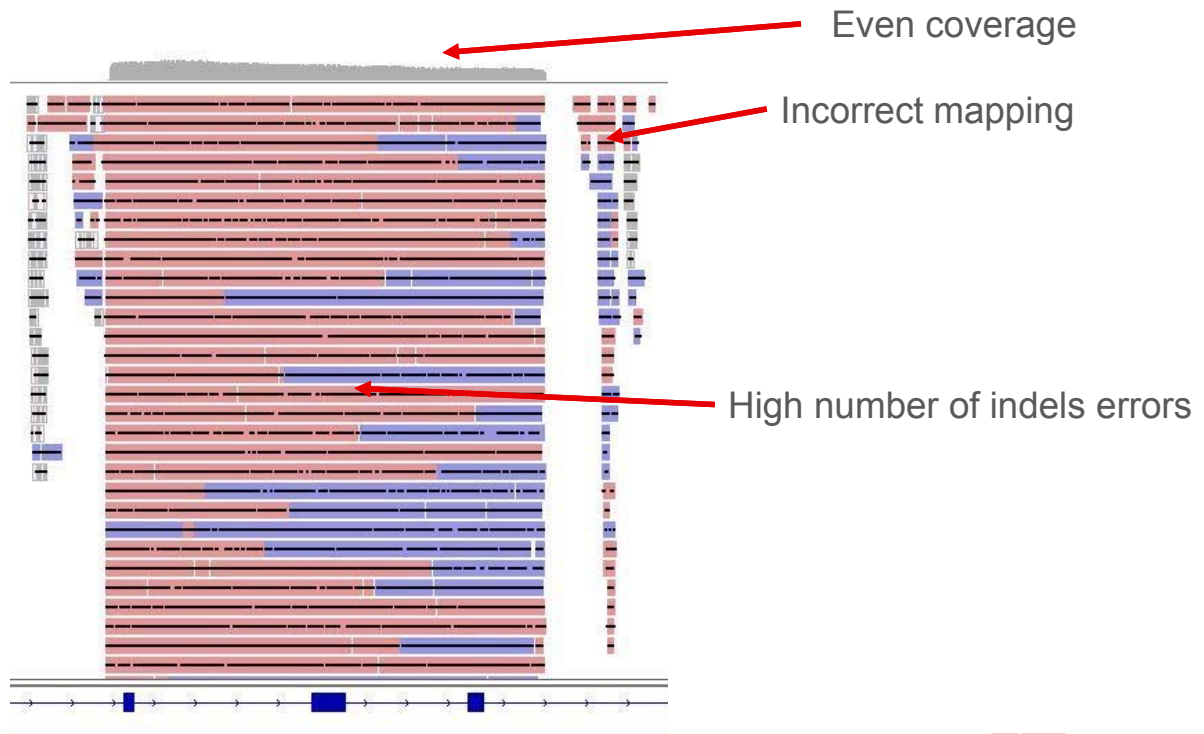
- Multiple-read consensus calling algorithm for PacBio reads
- Takes multiple reads of a given DNA template, outputs best guess of template's identity
- Hidden Markov model to model sequencing errors
- Quiver's consensus calls are completely independent of the reference—only use the reads (*Variant* calls still require reference for comparison.)



PacBio variants calling workflows



PacBio reads mapping



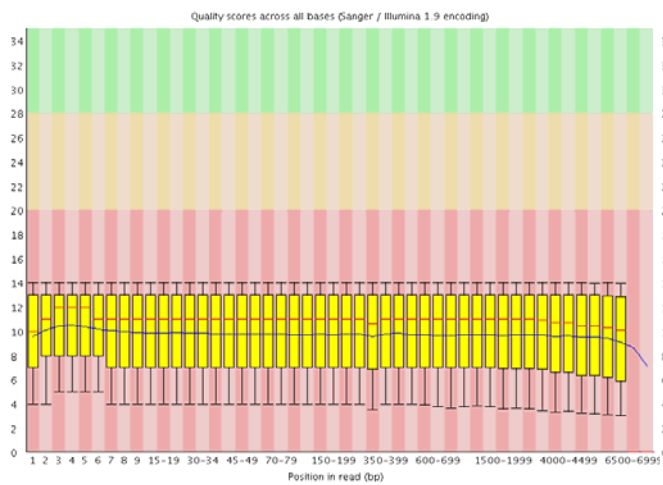
© 2009 SIB



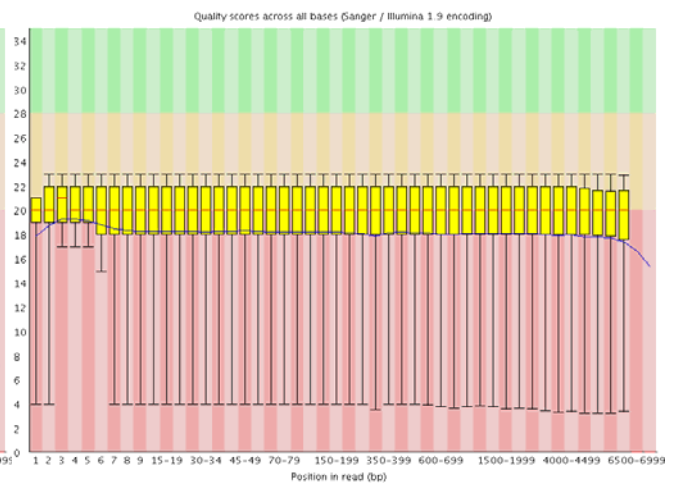
Swiss Institute of
Bioinformatics

Quality score re-calibration

Before recalibration



After recalibration



© 2009 SIB



Swiss Institute of
Bioinformatics

Example of an amplicons sequencing project

- 22 amplicons, total length:87 kb, total exonic length (17 kb)
- GATK VariantFiltration using recommended parameters
- SNVs:
 - 6 SNVs found in coding regions, all true positives
 - One variant spiked at a frequency of 11.5% among the variants identified
- Indels:
 - 17 indels found in coding regions, all likely false positives

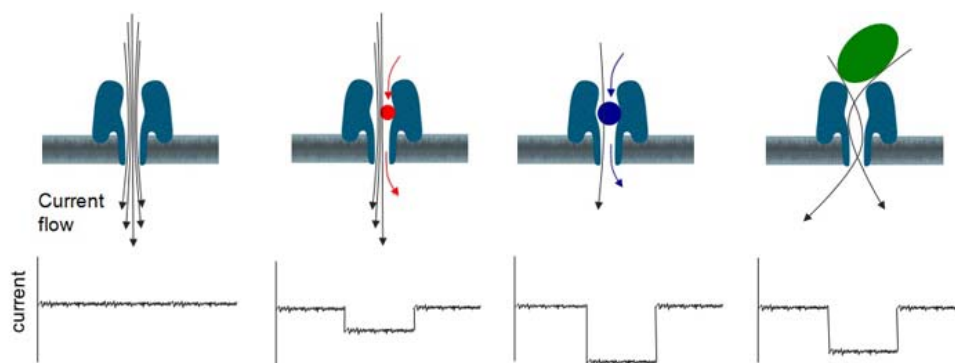
➔ Indels calling with PacBio data needs specific tuning



Course outline

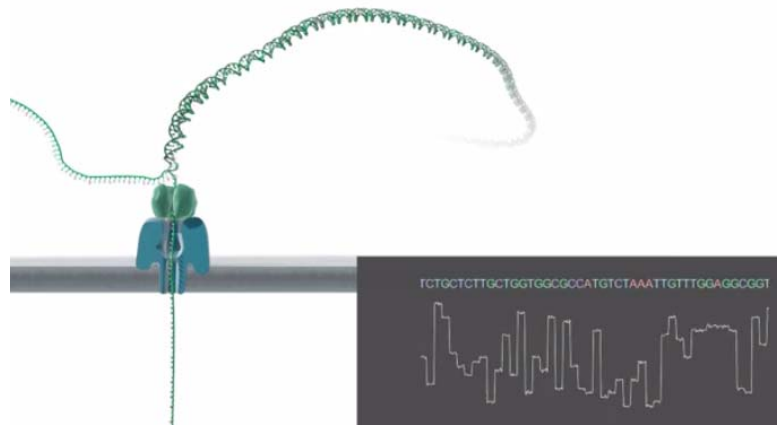
1. History of sequencing technologies
2. Illumina sequencing
3. PacBio sequencing
 - Technology description
 - Data characteristics and quality assessment
 - PacBio applications
4. Oxford Nanopore sequencing
 - Technology description
 - Data from MinION access program

Oxford Nanopore: nanopore sensing



- Nano-scale hole set in an electrically resistant membrane
- Ionic current is passed through the nanopore by setting a voltage across the membrane
- A molecule that passes through the pore or near its aperture creates a characteristic disruption in current

Oxford Nanopore: DNA sequencing



- Intact DNA polymers are sequenced in real time as the DNA passes through the pore
- Challenges
 - Many DNA bases occupy the pore at any time. Need to identify the sequence of individual bases within this strand
 - Need to controlled translocation of the strand through the nanopore

Oxford Nanopore systems



The GridION system

- operates with a single-use cartridge
- multiple nodes can be aggregated together into co-operating units



The MinION system

- portable, disposable device plugged directly into a laptop computer



The PromethION system

- benchtop, small number of samples on a very large number of nanopores or multiple samples in parallel
- modular flow cells number, no need to occupy the full capacity of the instrument

MinION reads

- Template
 - From the 1st of 2 strands presented to the pore
 - Slowed down by a proprietary processive motor enzyme which is ligated to the leader adapter
- Complement
 - Present if a hairpin has been successfully ligated
 - Slowed down by a second enzyme termed the HP motor
- 2D
 - **Normal 2D** with fewer events in the complement than in the template strand
 - **Full 2D** with more or equal complement events than template events (**Highest quality**)

First Data presented at AGBT meeting on February 14, 2014

Bacteria	Genome Size	Obtained coverage	Mean read length
E. coli	4.7 Mb	6-fold	5.4 kb
Scardovia	1.6 Mb	13-fold	4.9 kb

- Sequenced speed of 25 bases per second
- Measuring signal of 6 bases
- Accuracy:
 - 84% of reads ≥ 5 kb had at least one perfect 50-mer
 - 100% of reads ≥ 5 kb had at least one perfect 25-mer
- Systematic errors?
 - Systematic deletions
 - 'Error-blocks' regions with error prone bases

Course outline

1. History of sequencing technologies
2. Illumina sequencing
3. PacBio sequencing
 - Technology description
 - Data characteristics and quality assessment
 - PacBio applications
4. Oxford Nanopore sequencing
 - Technology description
 - Data from MinION access program

MinION access program (MAP)

Your journey through MAP

MAP Phase	What Happens	
Invitation issued	<ul style="list-style-type: none">• Review 'your guide to MAP'• Review Terms and Conditions• Review Laptop/PC, equipment and consumables requirements	<ul style="list-style-type: none">• If you wish to continue, log in to MAP website to provide deposit and initiate your first cycle of MAP
Registration & Deposit	<ul style="list-style-type: none">• Agree Terms and Conditions• Pay deposit and delivery charges• Provide biohazard information for returns	<ul style="list-style-type: none">• Receive shipping schedule
MAP cycle: Configuration phase	<ul style="list-style-type: none">• Receive Configuration Pack (MinION device, test flow cell, software)• Install and check MinION and software (requires completion to continue in MAP)	<ul style="list-style-type: none">• Await delivery of experimental pack
Burn-in experiments	<ul style="list-style-type: none">• Receive experimental pack (two flow cells, sufficient sequencing kits)• Conduct Burn-in experiments• Using existing flow cells, conduct experiments using own samples• Receive additional pack of two flow cells	<ul style="list-style-type: none">• Acknowledge with Oxford Nanopore that you are happy with results of Burn-in, including the results using your own samples
Additional experiments with own samples	<ul style="list-style-type: none">• Start to return flow cells	<ul style="list-style-type: none">• Continue using NanoporeOnline.com
End MAP cycle	<ul style="list-style-type: none">• Return MAP cycle package to Oxford Nanopore	<ul style="list-style-type: none">• Decision on future involvement

MAP one year after

Publications from MinION Access Programme Participants

[A complete bacterial genome assembled de novo using only nanopore sequencing data](#), Nicholas J. Loman, Joshua Quick, Jared T. Simpson, *bioRxiv*, doi: 10.1101/015552 (2015)

[Improved data analysis for the MinION nanopore sequencer](#), Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, Mark Akeson, *Nature Methods*, doi:10.1038/nmeth.3290 (2015)

[Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes](#), Ron Ammar, Tara A. Paton, Dax Torti, Adam Shlien, Gary D. Bader, *F1000Research*, doi: 10.12688/f1000research.6037.1 (2015)

[Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome](#), Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael Schatz, W Richard McCombie, *bioRxiv*, doi: <http://dx.doi.org/10.1101/013490> (2015)

[Nanopore Sequencing: From Imagination to Reality](#), Hagan Bayley, *Clinical Chemistry*, doi: 10.1373/clinchem.2014.223016 (2014)

[MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island](#), Philip M Ashton, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain & Justin O'Grady, *Nature Biotechnology*, doi:10.1038/nbt.3103 (2014)

[A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer](#), Joshua Quick, Aaron R Quinlan & Nicholas J Loman, *GigaScience*, doi:10.1186/2047-217X-3-22 (2014)

[poRe: an R package for the visualization and analysis of nanopore sequencing data](#), Mick Watson, Marian Thomson, Judith Risse, Richard Talbot, Javier Santoyo-Lopez, Karim Gharbi & Mark Blaxter, *Bioinformatics*, doi: 10.1093/bioinformatics/btu590 (2014)

[Poretools: a toolkit for analyzing nanopore sequence data](#), Nicholas J. Loman & Aaron R. Quinlan, *Bioinformatics*, doi: 10.1093/bioinformatics/btu555 (2014)

© 2009 SIB

<https://nanoporetech.com/technology/publications>



Swiss Institute of
Bioinformatics

MOLECULAR ECOLOGY RESOURCES

Molecular Ecology Resources (2014) 14, 1097–1102

doi: 10.1111/1755-0998.12324

OPINION

A first look at the Oxford Nanopore MinION sequencer

ALEXANDER S. MIKHEYEV and MANDY M. Y. TIN

Ecology and Evolution Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa 904-0495, Japan

Samples: Lambda phage genome (48 kb)
Protobothrops flavoviridis (snake) venom transcriptome

Runs: One flowcell for Lambda phage with chemistry R6. 36h
One flowcell for snake cDNA with chemistry R6. 24h

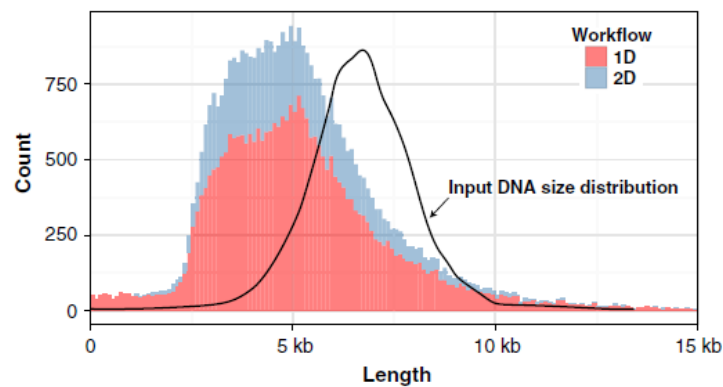
© 2009 SIB



Swiss Institute of
Bioinformatics

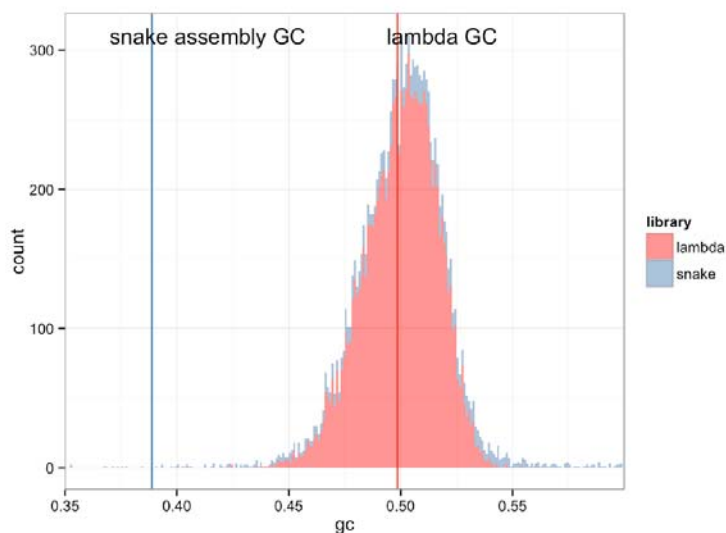
Lambda phage data

	Workflow	
	1D	2D
Reads	29 458	11 094
Total sequenced bases	155 370 698	55 854 289
Reads mapped by BLASTN	7997 (27%)	2746 (25%)
Reads mapped by BLASR	3472 (12%)	909 (8%)



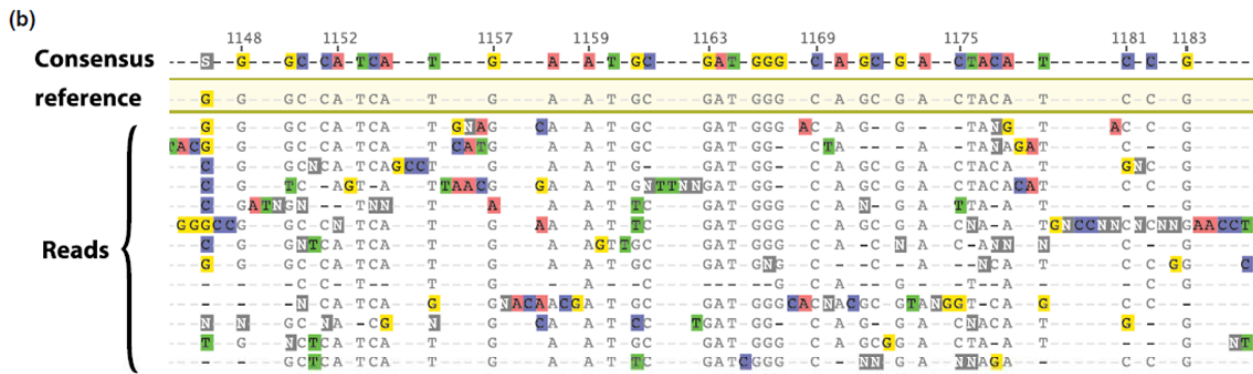
Snake cDNA data

Nb. Reads: 1429 1D reads (corresponding to 1Mb), 16 2D reads.
Overall alignment rate: 1.4%



Base caller model may be over-trained on lambda genomic sequence.

Lambda data alignment



- Insertions/Deletions introduce random spurious data.
- They were able to call the consensus sequence with 16x coverage data.
- BUT: Because of its extraordinarily high error rates, the current iteration of Oxford Nanopore technology is close to useless for genotyping applications.

DATA NOTE

Open Access

A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer

E. coli K-12

Joshua Quick^{1,2}, Aaron R Quinlan³ and Nicholas J Loman^{1*}

Table 1 Yields for each nanopore run in reads

	Run	Filename	Files	Template	Complement	All 2D	Full 2D
1	R7	Ecoli_NONI.tgz	47195	43656	23338	20087	1598
2	R7.3	Ecoli_R73.tgz	42316	39819	18889	11823	9563

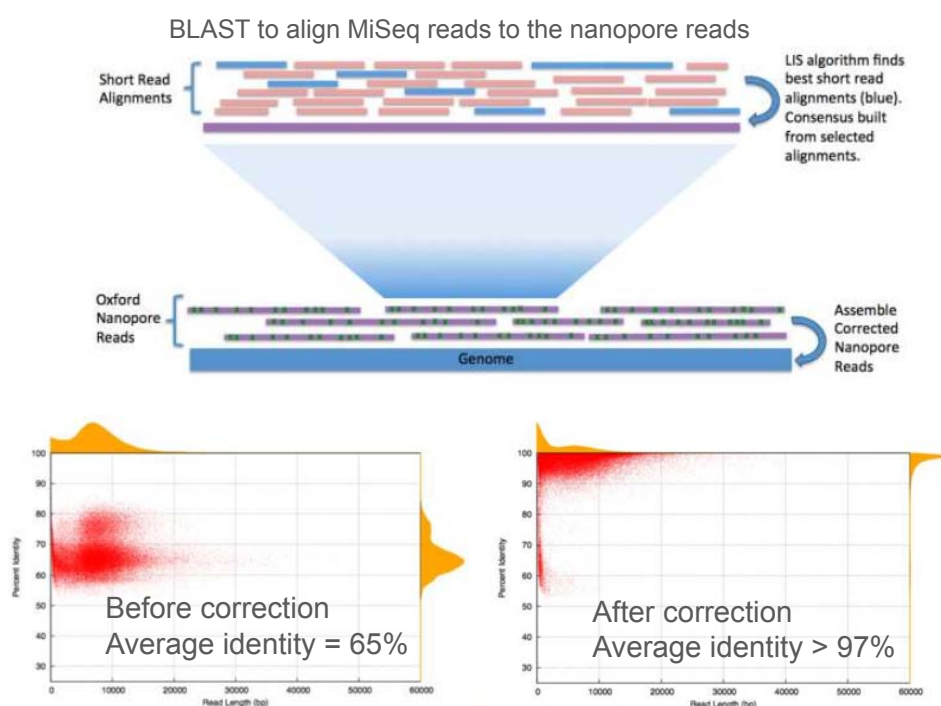
Read length: 5,458 bp (mean 2D)

- MinION is able to sequence entire bacterial genomes in a single run
- Subsequent analysis, not in the paper, showed that MinION data decreased the number of Illumina-only contigs from 96 to 4
- 6 misassemblies: transposon repeat units different between the *E. coli* batches
- N. Loman: "Importantly, the data are quite usable"

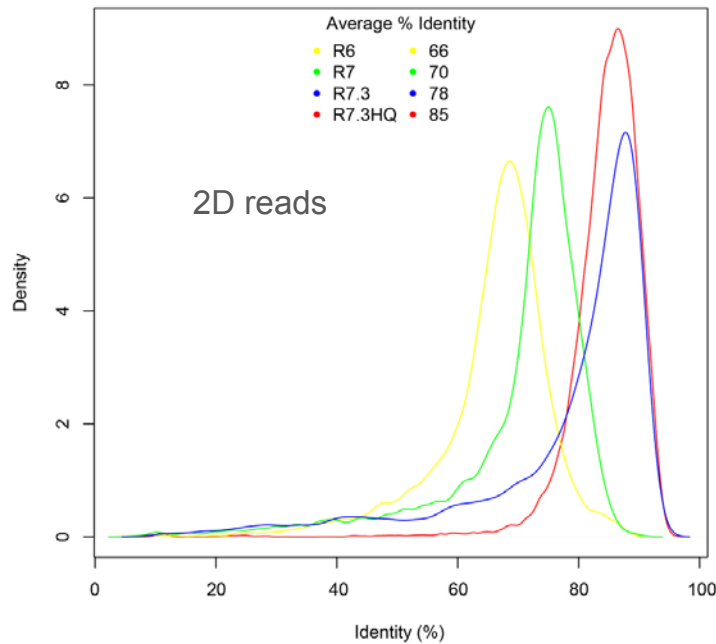
Tools for nanopore data

- Poretools: Data extraction and QC (Python) *Bioinformatics*, doi: 10.1093/bioinformatics/btu555 (2014)
- poRe : Data extraction and QC (R) *Bioinformatics*, doi: 10.1093/bioinformatics/btu590 (2014)
- BLAST, LAST (last.cbrc.jp), BWA-MEM ont2d: Aligners
- nanocorrect (github.com/jts/nanocorrect): Reads error correction pipeline *bioRxiv*, doi: 10.1101/015552 (2015)
- Nanocorr: Reads error correction using Illumina reads *bioRxiv*, doi: <http://dx.doi.org/10.1101/013490> (2015)
- Celera Assembler *Science* 287, 2196–2204 (2000)
- marginAlign: mapping and variants calling *Nature Methods*, doi:10.1038/nmeth.3290 (2015)

Nanocorr to error correct the reads for de novo genome assembly



MinION chemiseries improvement



Identity: proportion of bases in a read that align to a matching base in a reference sequence

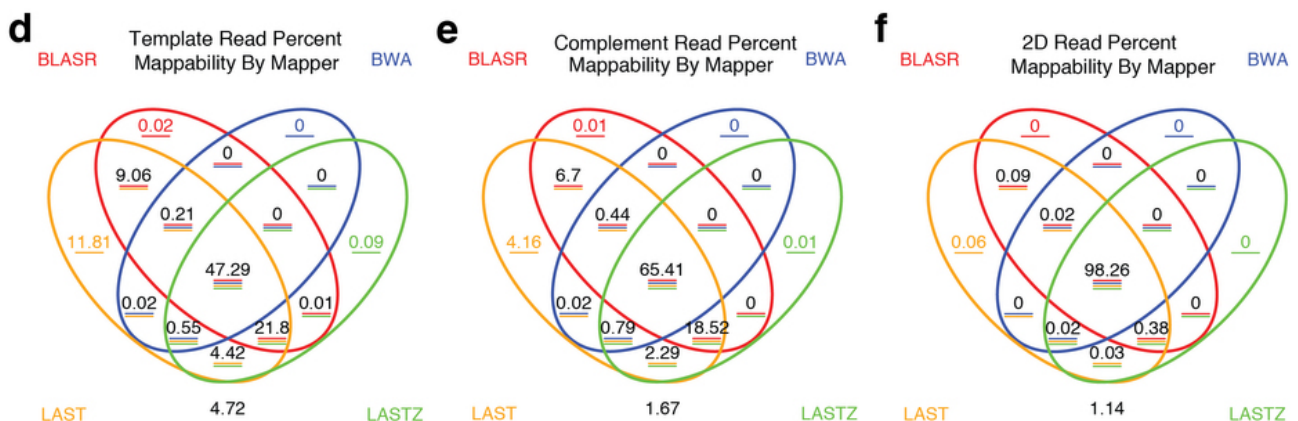
© 2009 SIB

Jain, et al., *Nature Methods*, doi:10.1038/nmeth.3290 (2015)



Swiss Institute of
Bioinformatics

Comparison of alignment programs



Proportion of reads that can be aligned to either the M13 or the phage λ DNA control using the tuned parameters for each mapper

© 2009 SIB

Jain, et al., *Nature Methods*, doi:10.1038/nmeth.3290 (2015)



Swiss Institute of
Bioinformatics

D. Sanglard's lab application for MinION early access program

- Genome sequencing of two strains of *Candida Glabrata* from one patient:
 - One strain sensitive to the drug
 - One strain resistant to the drug and more virulent
- Project performed by Luis Vale Silva



MinION runs

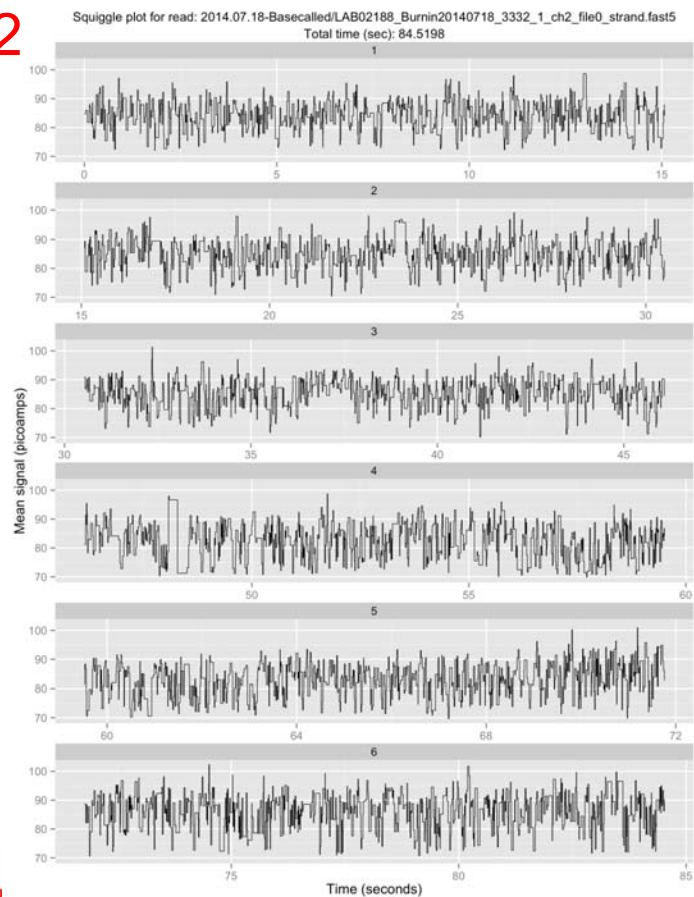
Date	MinION	Flow cell	Pore chemistry	Sample	Run time	Channels w/ reads	Basecalled fast5 files	Reads	Base pairs
24.06	MN02574	#1 MN-20-46467	R6	lambda	6h	52	14 (error w/ base calling!)	8	8'992
18.07	MN02603	#2 MN-20-46630	R6	lambda	6h	408	20'817	18'726	72'178'267
29.07	MN02603	#3 MN-20-46636	R6	lambda (same lib.; 11 days 4°C)	6h	138	2'593	2'588	12'652'844
30.07	MN02574	#4 MN-20-46617	R6	lambda (same lib.; 12 days 4°C)	6h	60	581	573	2'368'110
20.08	MN02603	#5 MN-20-68057	R7	DSY562	32h	234	3'222	2'398	9'017'131
03.09	MN02603	#6 MN-20-68111	R7	DSY562	4h30	59	134	86	270'286
03.09	MN02603	#7 MN-20-68183	R7	DSY562	48h	214	1'026	884	3'085'370
30.09	MN02574	#8 MN-20-68030	R7	DSY562	32h	184	755	868	2'889'720

Example of flowcell #2

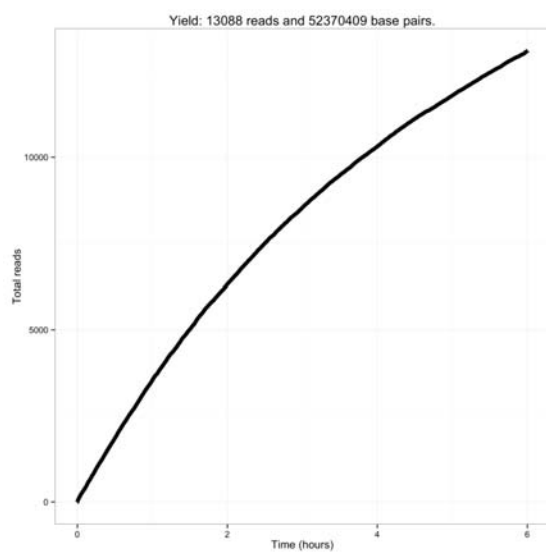
- Total reads: 18'726
- Total base pairs: 72'178'267
- Mean: 3'854.44
- Median: 3'776
- Min: 5
- Max: 84'419

Lambda phage genome size = 48 kb

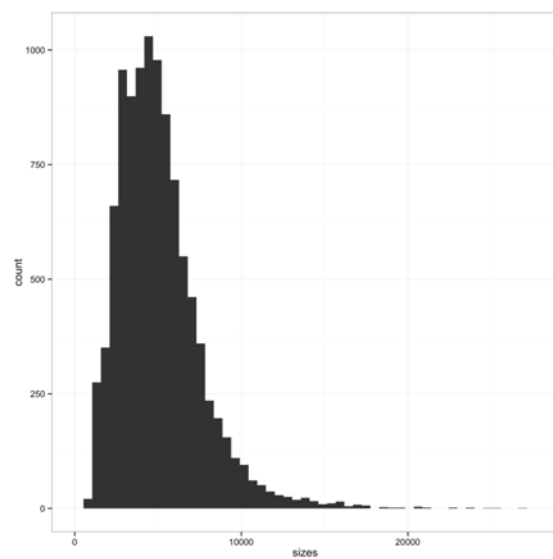
read squiggle plot



Example of flowcell#2



Yield plot



Read length histogram

Data processing

- Per-read fast5 files (hdf5 format) generated during run
- Fast5 files uploaded for cloud-based base-calling
- Read extraction into FASTA format using *Poretools* (python scripts) or R package *poRe*
- Reference-based alignment with *LAST* that can find weak similarities, with many mismatches and gaps
- Get Consensus sequence with *SAMtools*

LAST alignment

2D reads

All reads

Summary					Summary				
Globals					Globals				
Reference size	48,502				Reference size	48,502			
Number of reads	1,943				Number of reads	13,279			
Mapped reads	1,943 / 100%				Mapped reads	13,279 / 100%			
Unmapped reads	0 / 0%				Unmapped reads	0 / 0%			
Paired reads	0 / 0%				Paired reads	0 / 0%			
Read min/max/mean length	58 / 14,328 / 2,989.21				Read min/max/mean length	58 / 31,714 / 3,307.47			
Clipped reads	1,943 / 100%				Clipped reads	13,261 / 99.86%			
Duplication rate	5.87%				Duplication rate	18.42%			
ACGT Content					ACGT Content				
Number/percentage of A's	1,300,386 / 26.03%				Number/percentage of A's	10,664,100 / 25.58%			
Number/percentage of C's	1,172,054 / 23.46%				Number/percentage of C's	10,024,565 / 24.04%			
Number/percentage of T's	1,248,793 / 25%				Number/percentage of T's	10,278,461 / 24.65%			
Number/percentage of G's	1,274,497 / 25.51%				Number/percentage of G's	10,729,605 / 25.73%			
Number/percentage of N's	0 / 0%				Number/percentage of N's	0 / 0%			
GC Percentage	48.97%				GC Percentage	49.77%			
Coverage					Coverage				
Mean	124.32				Mean	1,041.5			
Standard Deviation					Standard Deviation				
Mapping Quality					Mapping Quality				
Mean Mapping Quality	210.89				Mean Mapping Quality	210.43			
Mismatches and indels					Mismatches and indels				
General error rate	34.03%				General error rate	37.86%			
Insertions	217,282				Insertions	1,165,017			
Deletions	341,530				Deletions	4,596,579			
Homopolymer indels	23.5%				Homopolymer indels	26.04%			
Chromosome stats					Chromosome stats				
Name	Length	Mapped bases	Mean coverage	Standard deviation	Name	Length	Mapped bases	Mean coverage	Standard deviation
Complete_genome_Lambda_phase:_ONT_provided_burn-in_sequence	48502	6029931	124.32		Complete_genome_Lambda_phase:_ONT_provided_burn-in_sequence	48502	50515027	1,041.5	

Oxford Nanopore Conclusions

- Multiple updates over the course of the MAP: 3 different flow cells versions and a number of sample preparation protocols
- Variability in performance between individual flow cells was considerable. Too many flow cells of bad quality.
- Data quality is poor, need 1000x coverage to get the lambda phage correct (with one ambiguity)
- Very long reads not always useful because they often do not align well: quality of the reads or inadequate alignment method?
- Systematic errors:
 - pentamers with low GC-content underrepresented
 - substitutions errors not uniform
 - predominance of single base indels
- Huge capacity for parallelization: PromethION with estimated throughput of 300 to 400 gigabases per day!

Thank You