FACOLTÀ DI
SCIENZE MATEMATICHE, FISICHE E NATURALI
Università degli Studi di Verona

NGS Technologies for Genomics and Transcriptomics

Massimo Delledonne

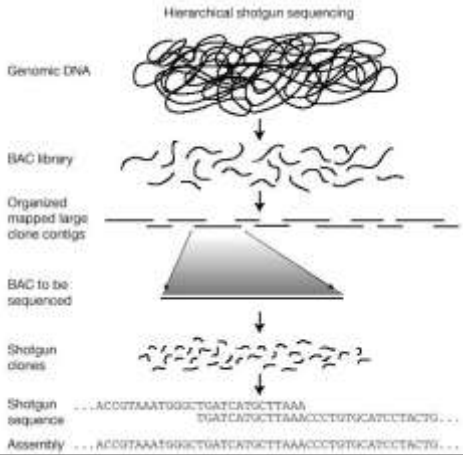
Department of Biotechnologies - University of Verona

http://profs.sci.univr.it/delledonne

articles

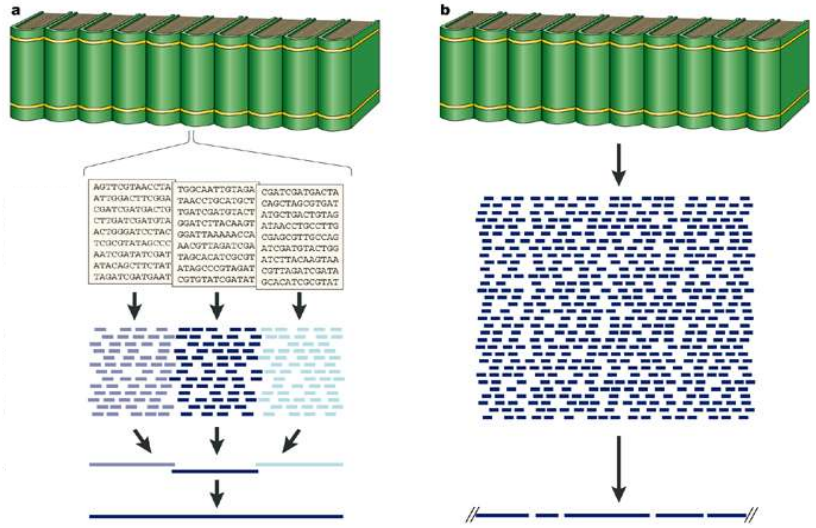
Initial sequencing and analysis of the human genome

13 years and **$3 billion** required for the Human Genome Project's reference genome



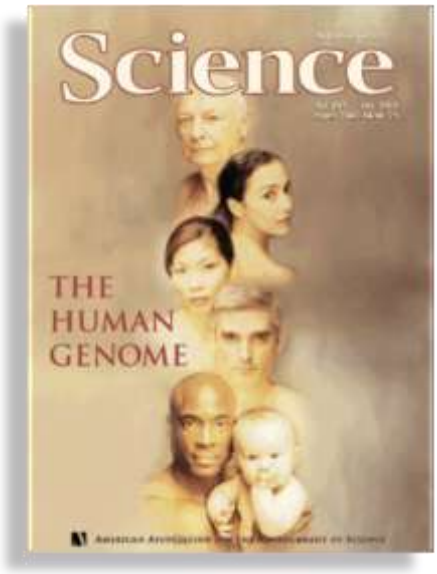
Clone-by-clone shotgun sequencing

Whole-genome shotgun sequencing

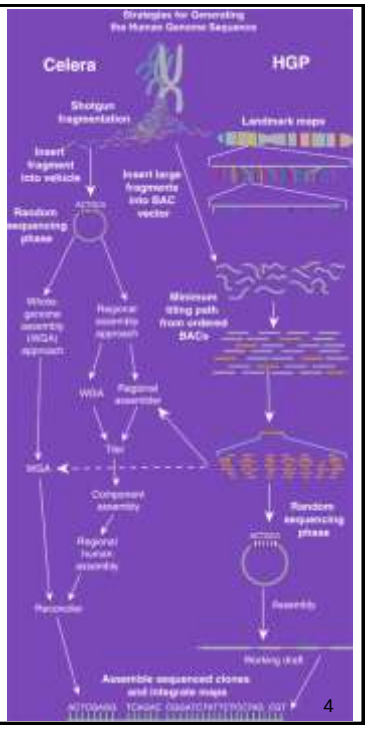


Nature Reviews | Genetics

3



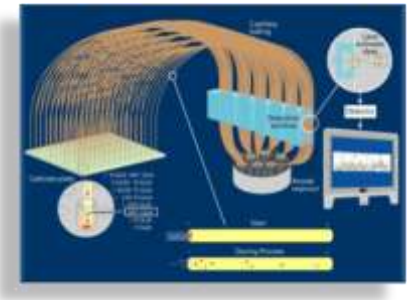
The $300,000,000 Celera effort was intended to proceed at a faster pace and at a fraction of the cost..



4

Sanger sequencing

(invented in the early 1970s)

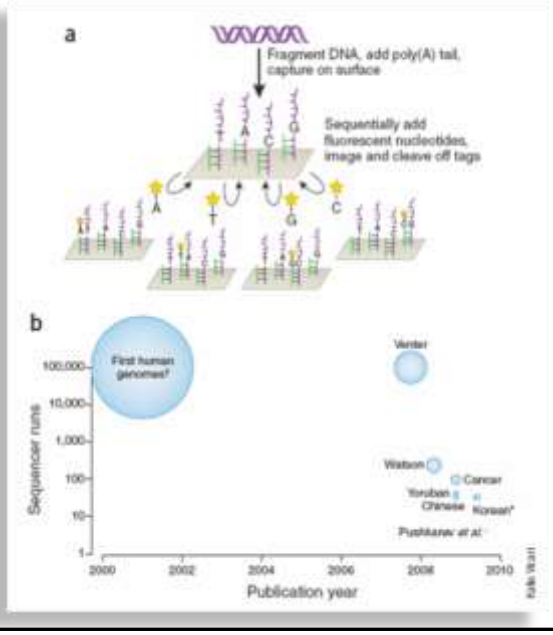


- 96 samples
- 800 nt / sample
- 1.600.000 nt / day

- Human genome: 3.000.000.000 bp
- Minimum coverage for an accurate analysis: 8X = 24.000.0000.000 nt

$$\frac{24.000.000.000}{1.600.000} = \textbf{15.000 days!}$$

"No Generation" Sequencing



New Generation Sequencing!

GS20 (2005)

100 bp x 200.000 Reads = 20 Mbp

x5

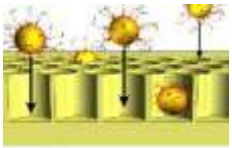
GS FLX (2007)

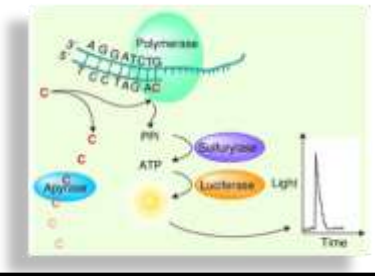
200 bp x 400.000 Reads = 100 Mbp

x5

GS FLX Titanium (2008)

400 bp x 1 M. Reads = 500 Mbp





454 "Revolution"

GS20 (2005)

100 bp x 200.000 Reads = 20 Mbp

x5

GS FLX (2007)

200 bp x 400.000 Reads = 100 Mbp

x5

GS FLX Titanium (2008)

400 bp x 1 M. Reads = 500 Mbp

```
>00405_2045_2005
CAGTCTCGTCGTCGTACGATCGTACGTAGCTTCACTTACGTACGCGGCG
GGGCGCATCTGCGCGCGCGATTATATCATCATCATACTCAGCATCGTCC
ATC

>00343_3489_2007
CAGTCTCGTCGTCGTACGATCGTACGTAGCTTCACTTACGTACGCGGCG
GGGCGCATCTGCGCGCGCGATTATATCATCATCATACTCAGCATCGTCC
ATCGATCGATCGATGATCGACGATCGTAGTCTACGTAGTACGTAGCTAG
CTTCGATCGATCGTACGTACGTCGTACGTAGTCAGACGTCAGCTACAGT
CATCTACGTAGCTCTACGTCGTGCATGCTAGCTATCGATCACGACTTAT
GCATC

>01384_3992_2008
CAGTCTCGTCGTCGTACGATCGTACGTAGCTTCACTTACGTACGCGGCG
GGGCGCATCTGCGCGCGCGATTATATCATCATCATACTCAGCATCGTCC
ATCGATCGATCGATGATCGACGATCGTAGTCTACGTAGTACGTAGCTAG
CTTCGATCGATCGTACGTACGTCGTACGTAGTCAGACGTCAGCTACAGT
CATCTACGTAGCTCTACGTCGTGCATGCTAGCTATCGATCACGACTTAT
GCATCCTCGGTATTCGGCGTACGATCGCTGACTGCTCGATTTTCGATCG
TACGTACCGTCAGCTAGCTAAAAAAAAAAGCTGCGCGGTCAGCATCGTTG
CATGCGCTACTGCTAGTACGTAGTAGTGCTACGTTTTATATCGTCGCCA
AATCTCTTCGGCTTTCG
```

nature

LETTERS

April 2008

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[1], Wen He[1], Yi-Ju Chen[1], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[1]†, Faheem Niazi[1], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[1], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†

2 months and 2.000.000 USD with 454 Life Sciences

Table 3 | SNPs matching HGMD mutations causing disease or other phenotypes

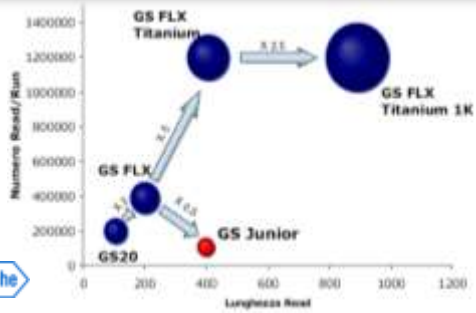
HGMD accession	Chromosome	Coordinate	HUGO symbol	Gene name	Cytogenetic	Phenotype	Zygosity
CM003589	1	97937679	DPYD	Dihydropyrimidine dehydrogenase	1p22	Dihydropyrimidine dehydrogenase deficiency	Heterozygous
CM950484	1	157443978	FY	Duffy blood-group antigen	1q	Duffy blood group antigen, absence	Homozygous*
CM942034	4	619702	PDE6B	Phosphodiesterase 6B, cGMP-specific, rod, beta	4p16.3	Retinitis pigmentosa 40	Heterozygous
CM021718	9	36208221	GNE	UDP-N-acetylglucosamine 2-epimerase	9p	Myopathy, distal, with rimmed vacuoles	Heterozygous
CM980633	10	50348375	ERCC6	Excision repair cross-complementing rodent repair deficiency, complementation group 6 protein (CSB)	10q	Cockayne syndrome	Homozygous†
CM050716	11	76533431	MYO7A	Myosin VIIA	11q13.5	Usher syndrome 1b	Homozygous†
CM950928	12	46812979	PFKM	Phosphofructokinase, muscle	12q13.3	Glycogen storage disease 7	Homozygous*
CM032029	14	20859880	RPGRIP1	Retinitis pigmentosa GTPase regulator interacting protein 1	14q11	Cone-rod dystrophy	Heterozygous
CM984025	19	18047618	IL12RB1	Interleukin-12 receptor, beta 1	19p13.1	Mycobacterial infection	Heterozygous
CM034138	19	41014441	NPHS1	Nephrosis-1, congenital, Finnish type	19q	Congenital nephrotic syndrome, Finnish type	Heterozygous
CM910052	22	49410955	ARSA	Arylsulphatase A	22q	Metachromatic leukodystrophy	Heterozygous

*Coverage at these SNP positions is less than 5. However, both produce benign phenotypes.
†Coverage at these SNP positions is greater than 5. Both would produce severe phenotypes if they were truly homozygous.

apoE unknown

Now Generation Sequencing

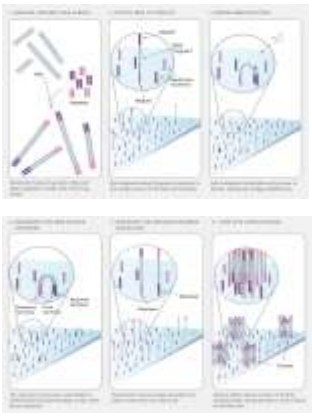
GS FLX+ System Performance	
Read Length	Up to 1,000 bp
Mode Read Length	700 bp
Throughput Profile	85% of total bases from reads >500 bp 45% of total bases from reads >700 bp
Typical Throughput	700 Mb
Reads per Run	1,000,000
Consensus Accuracy*	99.997%
Run Time	23 hours



GSFLX
1.100.000 reads (-800 nt)

GSJunior
100.000 reads (400-500 nt)

Genome Analyzer II (x)



End 2007: 30 million reads, 33 nucleotides each (1 Gbase)

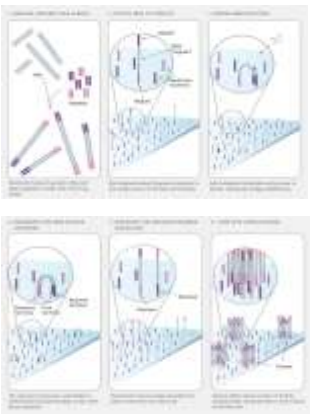
End 2008: 50 million reads, 50 nucleotides each (2.5 Gbases)

2010: 300 million single reads, 150 nucleotides each
600 million paired reads, 150 nucleotides each (90 Gbases)

illumina

11

HiSeq 2000



HIGHEST OUTPUT
600 Gb per run

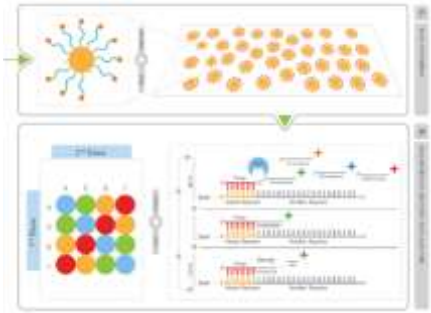
HIGHEST NUMBER OF READS
Three billion (100 or 2 x 100) reads

illumina

12

5500xl SOLiD™ System

Throughput/Run	Microbeads: Up to 180 Gb, or more than 2.9 B reads (paired-end or mate-paired runs) Nanobeads: Up to 300 Gb, or more than 4.8 B reads (paired-end or mate-paired runs)	
Samples/Run[2]	Microbeads: • 2 genomes • 24 exomes • 12 transcriptomes	Nanobeads: • 3 genomes • 40 exomes • 20 transcriptomes
Read Length	• 75 bp (fragment) • 75 bp x 35 bp (paired-end) • Up to 60 bp x 60 bp (mate-paired)	
Run Time**	• 1 day for 35 bp, 1 lane • 7 days for 75 bp x 35 bp or 60 bp x 60 bp, 12 lanes	



Applied Biosystems

applied biosystems

Complete genomics

Figure 7. Complete Genomics Sequencing Instrument.

Complete genomics

Mar 8, 2012

Complete Genomics Reports Fourth Quarter and Fiscal Year 2011 Results

MOUNTAIN VIEW, Calif. – Mar. 8, 2012 – Complete Genomics, Inc. (Nasdaq: GNOM), the whole human genome sequencing company, today announced financial results for the fourth quarter and fiscal year ended December 31, 2011.

Fiscal Year 2011 Results

- We recognized revenue with respect to approximately 2,000 genomes, compared to approximately 650 genomes in 2010.
- Revenue was $15.3 million, compared to $9.4 million in 2010.
- Cash and operating expenses were $69.6 million, compared to $57.0 million in 2010.
- Net loss was $72.3 million, compared to $57.7 million in 2010. The net loss for 2010 had included a $7.2 million non-cash expense resulting from an accounting adjustment to the market value of certain convertible equity securities.

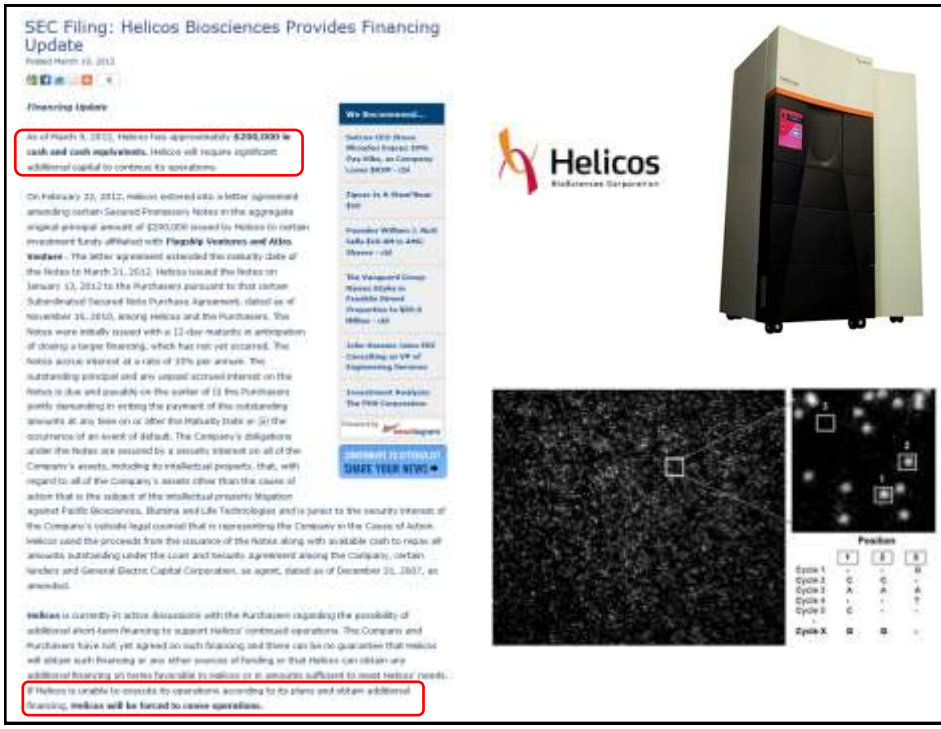
"In 2011, we clearly demonstrated the value of our service, signing orders for approximately 6,000 genomes representing aggregate revenue potential of $30 million, and delivering data to approximately 65 customers," said Dr. Clifford Reid, chairman, president and CEO of Complete Genomics. "We entered 2012 with a strong order book of about 5,000 genomes, representing aggregate revenue potential of approximately $25 million."

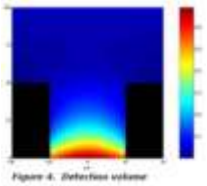
More Press Releases

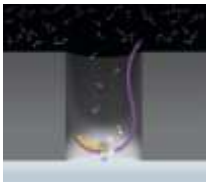
See additional press releases

About BGI

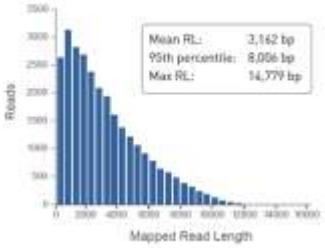
BGI was founded in 1999 as the Beijing Genomics Institute. It now has several branches and subsidiaries including: BGI-Shenzhen, a nonprofit research institute; BGI-Hong Kong, a private institute that manages international collaborations and transfers profits to BGI; and BGI-Americas, located in Boston, which just celebrated its one-year anniversary and announced new joint projects with the Broad Institute and the United Kingdom. BGI has about 4,000 employees and the capacity to sequence the equivalent of 1,600 complete human genomes each day.







Read Length Distribution

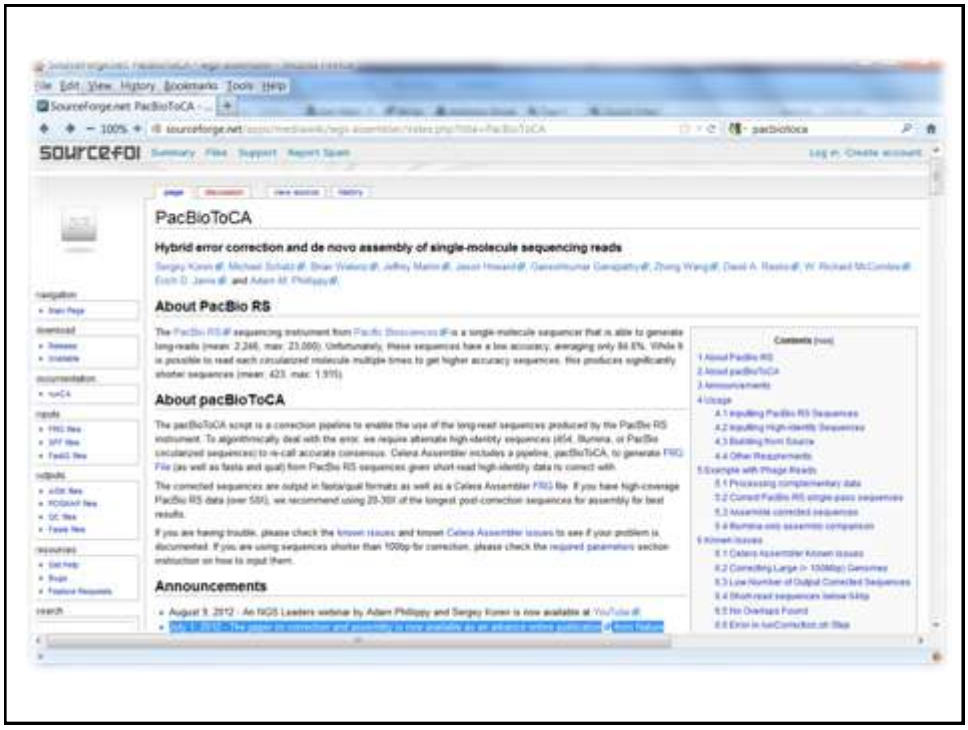


The commercial PacBio RS instrument generates **average read lengths of 1,000 bases, with a small fraction of reads longer than 10 kilobases, and have a typical run time of 30 minutes**. The commercial instrument runs chips with **75,000 wells**, or zero-mode waveguides. The commercial raw read accuracy is between 85 and 90 percent.

September 27, 2011

With the C2 chemistry, the Expression Analysis team achieved average read lengths of 2,715 base pairs, with a mean maximum read length of 13,091 base pairs. "Our personal best is just over 14,000 bases," Hurban said.

Despite the potential of PacBio's platform to produce very long reads, which would enable better assemblies, adoption of the instrument since its launch in the second quarter of this year has been slower than expected. As a result, **PacBio last week announced that it would have to lay off around 130 employees**


PacBioToCA
Hybrid error correction and de novo assembly of single-molecule sequencing reads
About PacBio RS
About pacBioToCA
Announcements


December, 2010
ion torrent
by life technologies
90-minute run times

MiSeq

January, 2011

Approximate Run Duration and Output

Cluster Generation and Sequencing

Read Length	Total Time for Amplification and Sequencing*	Output**
1 x 36 bp	~4 hours	175-245 Mb
2 x 25 bp	~5 hours	250-350 Mb
2 x 100 bp	~19 hours	1.0-1.4 Gb
2 x 150 bp	~27 hours	1.5-2.0 Gb

*Includes paired-end read, if applicable.
** Install specifications for MiSeq with an Illumina PhiX library and cluster densities between between 725–889 k/mm2 that pass filtering. Performance may vary based on sample quality, cluster density, and other experimental factors.

23

illumina

Next Generation Sequencing

Ion Proton™ Sequencer

The Benchtop Genome Center

January 10th, 2012

- Supports Ion Proton™ I and Proton™ II chips
- €209,800 system list price (sequencer and server)
  - $164,900 for Ion PGM™ owners
- State-of-the-art electronics to support highest throughput

Proton I

Proton II

January 10th, 2012

MiSeq System Performance Parameters

Read Length	Total Time from Prepped Library through Sequencing*	Output
1 × 35 bp	2.2–3.5 hours	440–660 Mb
2 × 25 bp	4.8–6.0 hours	640–800 Mb
2 × 100 bp	14.0–16.0 hours	2.5–3.1 Gb
2 × 150 bp	20.7–24.0 hours	3.7–4.6 Gb
2 × 250 bp**	>35 hours	5.0–7.0 Gb
Reads	12.0–15.0 million clusters passing filter, and 24.0–30.0 million paired-end reads	
Performance†	>90% bases higher than Q30 at 1 × 35 bp >90% bases higher than Q30 at 2 × 25 bp >85% bases higher than Q30 at 2 × 100 bp >75% bases higher than Q30 at 2 × 150 bp	

* Includes paired-end read, if applicable.

**Performance, output, simplification, and sequencing time for 2 × 250 bp read length depends on instrument upgrade, commercially available in the third quarter of 2012. Customers will be notified of upgrade availability. Upgrade dates are subject to change.

†The percentage of bases >Q30 is averaged across the entire run, not on a per-read or per-cycle basis.

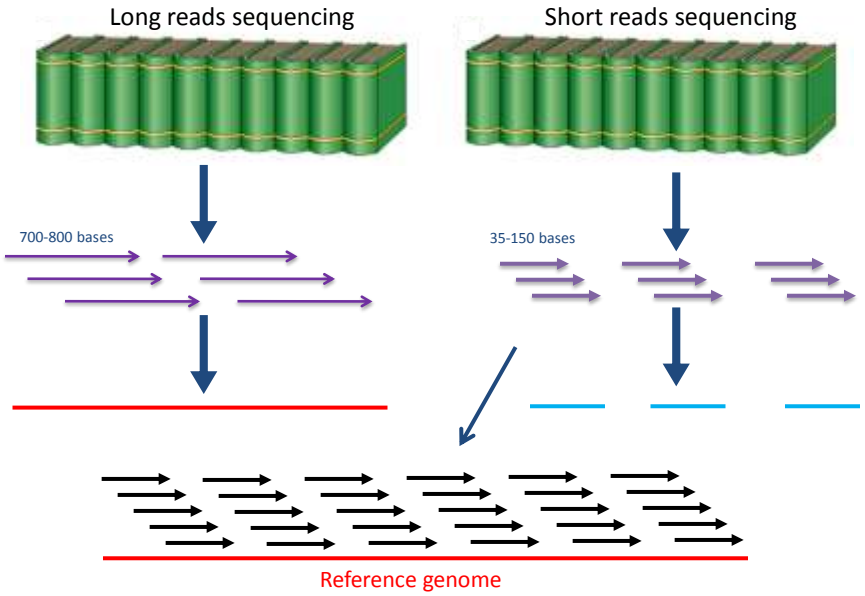
HiSeq 2500/1500

Two run modes. High output or rapid run.

Flexibility to batch process multiple samples with high output in a single run, or get rapid results with fewer samples for time-critical studies.

	HiSeq 2500		HiSeq 1500	
Run Mode	High Output	Rapid Run*	High Output	Rapid Run*
Output (2 × 100 bp)	600 Gb	120 Gb	300 Gb	60 Gb
Run Time (2 × 100 bp)	~11 days	~27 hours	~6.5 days	~27 hours
Cluster Generation	cBot	On board	cBot	On board
Paired-end Reads	4 billion	1.2 billion	2 billion	600 Million
Single Reads	3 billion	800 Million	1.5 billion	300 Million
Maximum Read Length**	2 × 140 bp	2 × 150 bp	2 × 100 bp	2 × 150 bp
Quality Scores***	> 80% (2 × 50 bp) > 80% (2 × 100 bp)			

Resequencing vs. de-novo assembling



Resequencing is good only if the reference genome is good

Next Next (Next) Generation Sequencing Technologies for haplotyping and de-novo assembling

Biological nanopores

genia

Electronic BIOSCIENCES

Oxford NANOPORE Technologies®

http://www.nanoporetech.com/sequences

Oxford Nanopore megaton announcement: "Why do you need a machine?" – exclusive interview for this blog!

By Nick Loman on February 17, 2012

Sometimes this genome blogging lark really pays off. Yest… …eak preview of the big announcement at AGBT, and 20 minutes to speak with Ox… …f Applications), Clive Brown (Chief Technical Officer) and Zoe McDougall (Director of … …nt I couldn't tell anyone what they said until the embargo was lifted at 5pm today!

I asked as many questions as I could without knowing the contents of the AGBT talk. I probably should have asked a bunch more. I do remember saying "wow" quite a lot.

First, go and read the press release!

…hod, made poss…

…es in contact with the pore are detected through electrochem…ory

…olecules (lambda phage) – no theoretical read length limit

…mple preparation

- Two products announced:
  - MinIon – USB disposable sequencer for ~ $900 has 512 nanopores – target 150mb/hour
  - MinIon can run at 120-1000 bases/minute per pore for up to 6 hours
  - GridIon – two versions of rack-mountable sequencer with 2000 nanopores (2nd half 2012), 8000 nanopores (2013)
  - GridIons can be racked in parallel, 20 could do a whole human genome in 15 minutes
  - Each GridIon can do "tens of gigabases" over 24 hours
- Both machines commercially available 2nd half 2012
- Sequencing can be paused, sample recovered, replaced, started again
- Accuracy is 96%, errors are deletions, error profile will improve through software

Current Status: alpha testing (beta testing to start by end of 2012)
Projected commercial availability: 2013

Sequencing the Human Genome

Human Genome Project	James Watson	Personal genome
		
1988 – 2001 13 years 3.000.000.000 dollars	2007 2 months 2.000.000 dollars	2010 7 days 10.000 dollars

PacBio Developing Chip Technology for Highly Multiplexed 'V2' Sequencers; Launch Planned for 2014

By Julia Karow

As it prepares to roll out its first single-molecule real-time sequencing instrument to early-access customers this summer, Pacific Biosciences is already working on a new, highly multiplexed version of its technology that integrates sequencing reactions, optical detection, and signal processing into a microchip, In Sequence has learned.

PacBio believes that its "version 2" or V2 technology "is going to completely replace any second-gen [sequencing] device," according to CEO Hugh Martin, whereas the current version 1 will likely be used in conjunction with high-throughput short-read instruments, for example in genome assemblies.

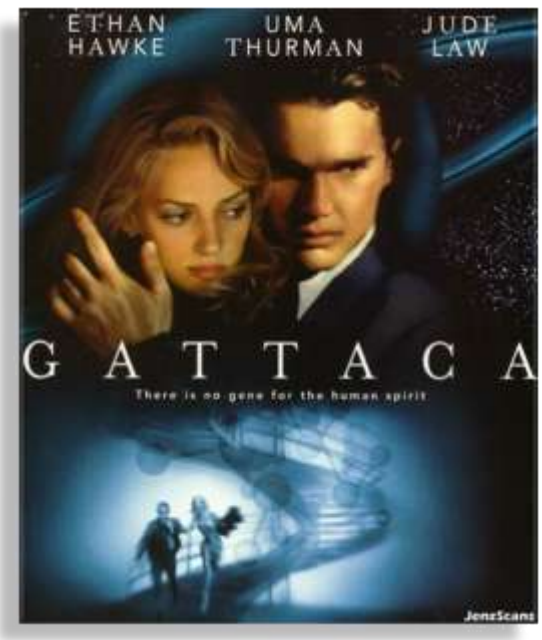
V2 will enable the company to multiplex several million zero-mode waveguide reaction wells and to use polymerase enzymes that synthesize DNA at a speed of up to 50 bases per second, resulting in a throughput of more than 100 megabases per second.

Starting in 2014, PacBio plans to launch two RS II-workflow systems that are based on the new technology: a high-throughput instrument that will be able to sequence a human genome with several fold coverage in 15 minutes or less, and a cost- and size-optimized box that will be "the size of a copier and have a lower throughput."

While the high-throughput system will be geared at core labs and genome centers, the newer throughput instrument — which will possibly be portable — will be aimed at small clinical labs "or even doctors' offices" and is designed to be "a delivery vehicle for some of the applications developed on V1 around diagnostics," Martin said.

Prior to that, in 2011 and 2012, the company plans to improve the performance of its "version 1" system by increasing the read length, polymerase speed, and number of productive ZMWs, lowering the cost per base, upgrading software, and adding new applications such as methylation sequencing, direct RNA sequencing, and protein-translation analysis. None of these improvements will require hardware changes, according to Martin.

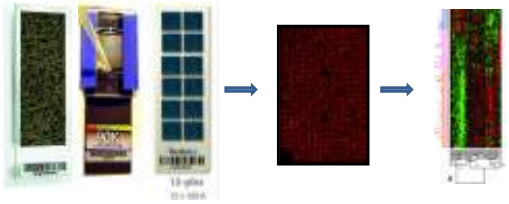
The first commercial V1 systems, scheduled for rollout later this year (see related story, this issue), will likely be used mostly in research. While "you are not going to see that in a doctor's office," Martin said, "you will see large pharma companies or diagnostic companies developing applications that are going to be deployed.



FACOLTÀ DI
SCIENZE MATEMATICHE, FISICHE E NATURALI
Università degli Studi di Verona

FONDAZIONE Cariverona

Center for Functional Genomics



Grid-ION-Proton ?

Illumina HiSeq 1000

PowerEdge™ R900
- 4 Intel Xeon® 7450 6-core
- 128 Gb RAM

Cluster HP DL380 G7
- 6 Xeon® 5645 6-core
- 144 Gb RAM

FACOLTA DI
SCIENZE MATEMATICHE, FISICHE E NATURALI
Università degli Studi di Verona

FONDAZIONE Cariverona

Alberto Ferrarini

Luca Venturini

Luciano Xumerle

Andrea Minio

Giovanni Malerba

Paola Tononi

Gianpiero Zamperin

Genny Buson

Elisa Zago

Anne-Marie Digby

Plant Biology
Mario Pezzotti (UniVR)
Grant Cramer (USA)

Medicine
Giovanni Martinelli (UniBO)
Paolo Garagnani (UniBO)
Marina Noris (Mario Negri)
Alberto Zamò (UniVR)
Elisabetta Trabetti (UniVR)
Domenico Girelli (UniVR)

EMBO