Introduction to Sequencing

BCB 511: Applied Bioinformatics

Matt Settles

University of Idaho

January 22, 2015

Outline

History

Roche 454 Pyrosequencing

Illumina/Solexa

Life Sciences Ion Torrent

Pacific Biosystems

Oxford Nanopore

History

It would take a few more decades after the discovery of the double helix in 1953 before we could readily analyze fragment of DNA. RNA sequencing actually preceded DNA sequencing when Walter Friers from the University of Ghent published the first complete gene and genome of Bacteriophage MS2 in 1972 and 1976 respectively.

Location specific primer extension: Raw Wu (1970), using DNA polymerase catalysis and specific nucleotide labeling.

chain-terminating inhibitors: Frederick Sanger (1977), aided in speeding up the process

Automation

1. Leroy E. Hood's laboratory at the California Institute of Technology announced the first semi-automated DNA sequencing machine in 1986.
2. Applied Biosystems' produced the first fully automated sequencing machine, the ABI 370, in 1987, followed by the ABI Prism 373, (1990), ABI Prism 377 (1995), ABI Prism 310 (also 1995) represented the first capillary sequencer, ABI Prism 3700 (1999, the workhorse of the human genome project), ABI 3730xl DNA analyzer (2002) @ 2M bases per day.

primer walking to *de novo* sequencing

Primer Walking

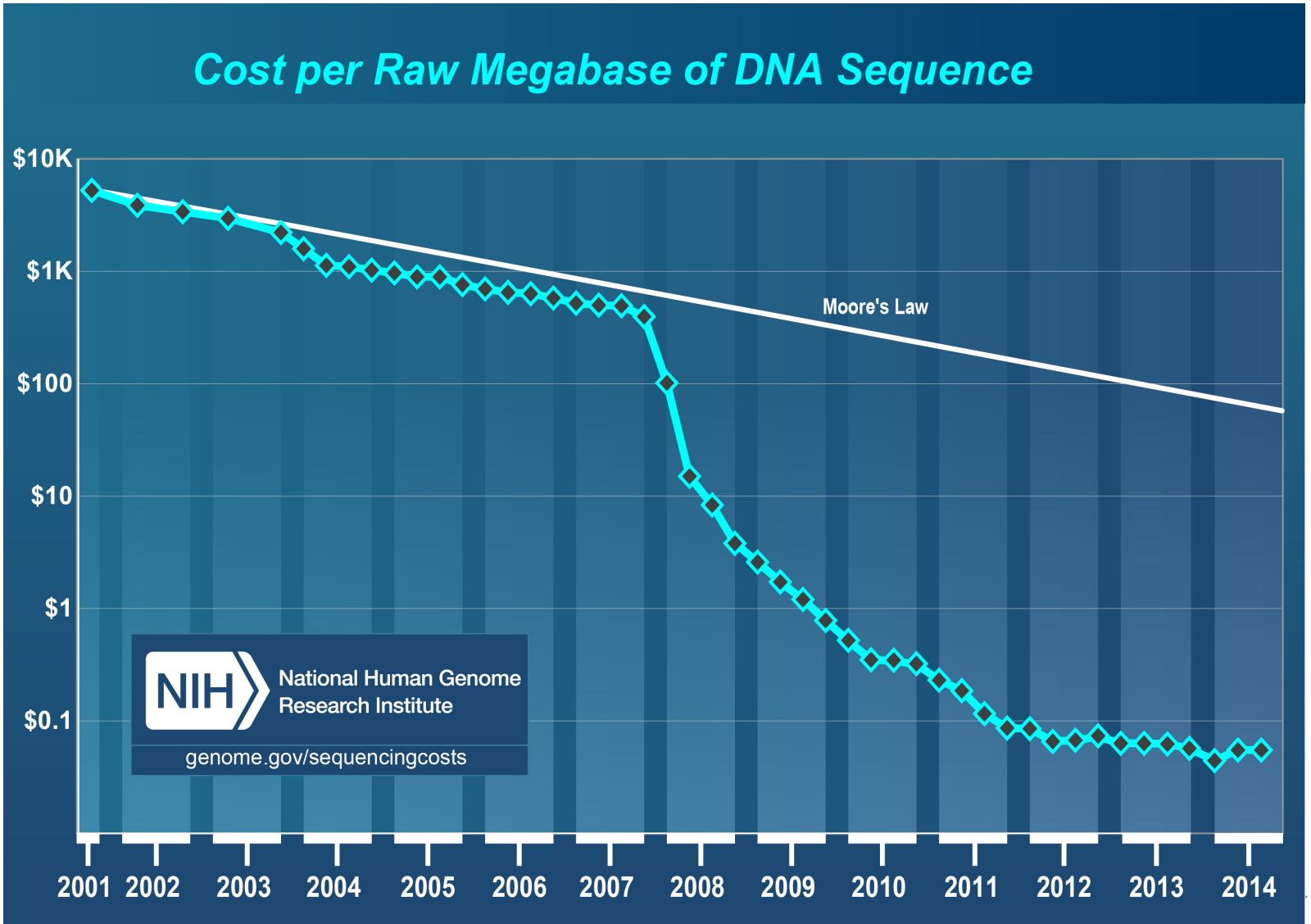
1. Design a primer that matches the sequence neighboring the unknown sequence
2. Sequence the short DNA strand using the Sanger method
3. The new sequenced portion is used to design a new primer and repeated

de novo sequencing or Shotgun sequencing

1. High-molecular weight DNA is sheared into random fragments
2. Shorter fragments are cloned into a vectors
3. clones are sequenced from both ends, creating two "reads"
4. original sequence is reconstitutes by "assembling" the reads

Evolution of DNA Sequencing

July - 2014: $0.05 per Megabase, $4,905 per Human Sized Genome (30x coverage)



Roche 454

The first massively parallel method to become commercially available was developed by 454 Life Sciences in 2005 (acquired by Roche in 2007) and is based on the pyrosequencing technique. Similar to the Sanger method, sequencing is carried out using primed synthesis by DNA polymerase. However in the 454 pyrosequencing method, the DNA fragments are presented with each of the four dNTPs sequencially and without a dye-terminator, as is done with Sanger sequencing, allowing for multiple incorportation in the same flow. The amount of the incorporation is monitored by luminometric detection of the pyrophosphate released (hence the name "pyrosequencing").

Roche 454 platforms

Roche 454 has 2 platforms the GS Junior System (a "benchtop" system) and the GS FLX+ System (what we have on campus).

GS FLX+ System		
Sequencing Kit	**New! GS FLX Titanium XL+**	**GS FLX Titanium XLR70**
Read Length	Up to 1,000 bp	Up to 600 bp
Mode Read Length	700 bp	450 bp
Throughput Profile	- 85% of total bases from reads >500 bp - 45% of total bases from reads >700 bp	- 85% of total bases from reads > 300 bp - 20% of total bases from reads > 500 bp
Typical Throughput	700 Mb	450 Mb
Reads per Run	~1,000,000 shotgun	~1,000,000 shotgun, ~700,000 amplicon
Consensus Accuracy*	99.997%	99.995%
Run Time	23 hours	10 hours
Sample Input	gDNA or cDNA	gDNA, cDNA, or amplicons (PCR products)
Multiplexing	Multiplex Identifiers (MIDs): 132 Gaskets: 2, 4, 8, 16 regions	

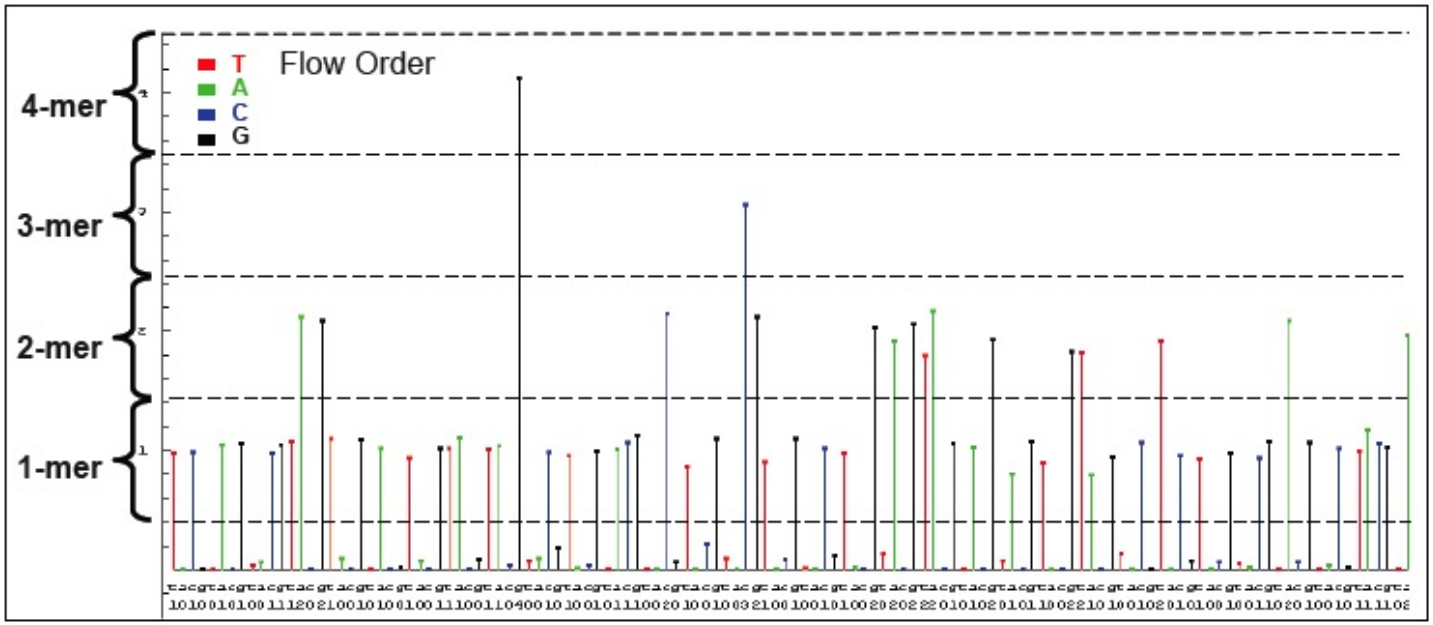
Roche 454 Workflow

- Library Construction
- QA - Library Quantification (Titration)
- emulsion PCR (emPCR)
- Picotiter Plate Loading
- Sequencing
- Image extraction
- Flowgram extraction

Roche 454 Workflow Video

454 Video

Roche 454 Flowgrams



454 Read Naming Convections

Roche 454 raw data are stored in SFF files (standard flowgram format), but fasta and qual (or fastq) files can be extracted from them

>EBO6PME01EGNVK

Timestamp EB06PM

Randomized E

Plate Region 01

X,Y coord EGNVK

The timestamp, hash character and X,Y location use a base-36 encoding (where values 0-25 are the letters 'A'-'Z' and the values 26-35 are the digits '0'-'9'). An accession thus consists only of letters and digits, and is case-insensitive.

Illumina Solexa

The second next-generation sequencing technology to be released (in 2006) was Illumina Solexa sequencing. A key difference between Roche 454 and Illumina sequencing was the use of chain-terminating nucleotides. The fluorescent label on the terminating base can be removed to leave an unblocked 3' terminus, mating the chain termination a reversible process. The method thus sequences one base at a time, rather than 0 or more bases as does Roche 454.

Illumina Platforms

Illumina has multiple systems, in three classes
MiSeq, NextSeq and HiSeq

Illumina Systems

Illumina Workflow

- Library Construction
- Cluster Generation
- Sequencing
- image extraction

Illumina Workflow Video

Illumina Video

Illumina Read Naming Conventions

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Ion Torrent Workflow Video

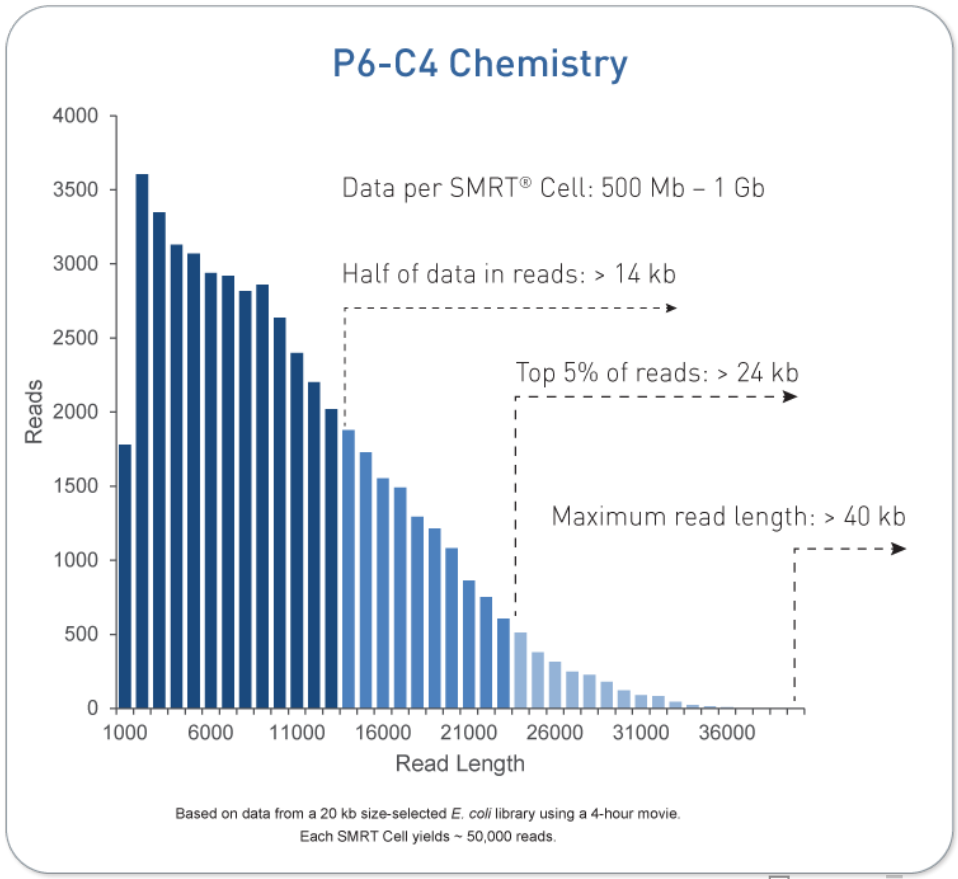
Ion Torrent PGM, first available in 2011, generates up to 400bp reads (reported) and up to 2Gb (5.5m reads) per run. Cheap fast runs. Ion Proton system can generate up to 10Gb per run. Generates flowgrams and SFF files similar to Roche 454 data as well as the standard fastq files.

Ion Torrent Video

Pacific Biosciences Workflow Video

Pacific Biosystems is so far the most successful third generation DNA sequencing system. Key differences are that its a single molecule, real time (SMRT) technology and capable of producing sequences of multi-kilobases.

Pacific Biosciences Video



Announced in 2012, Oxford Nanopore sent a ripple through the sequencing community but has yet to live up to expectations. Promises "tens of kbs".

Minion Video