Introduction to Sequencing

BCB 504: Applied Bioinformatics

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

February 11, 2013

Outline

History

Roche 454 Pyrosequencing

Illumina Solexa

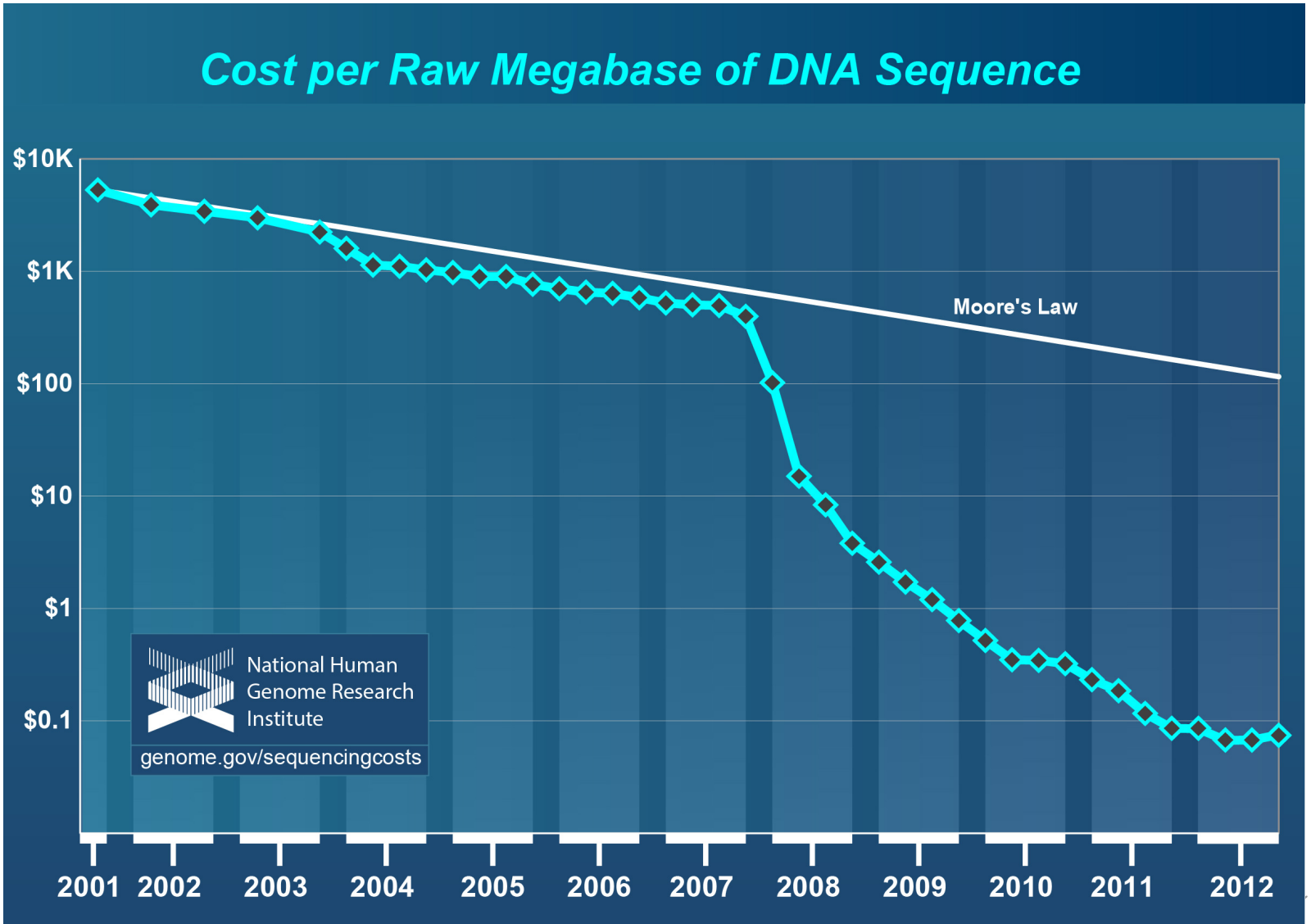
Sequence Data

Library Preparation

Other sequencing technologies

Evolution of DNA Sequencing

Oct - 2012: $0.07 per Megabase, $6,618 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	

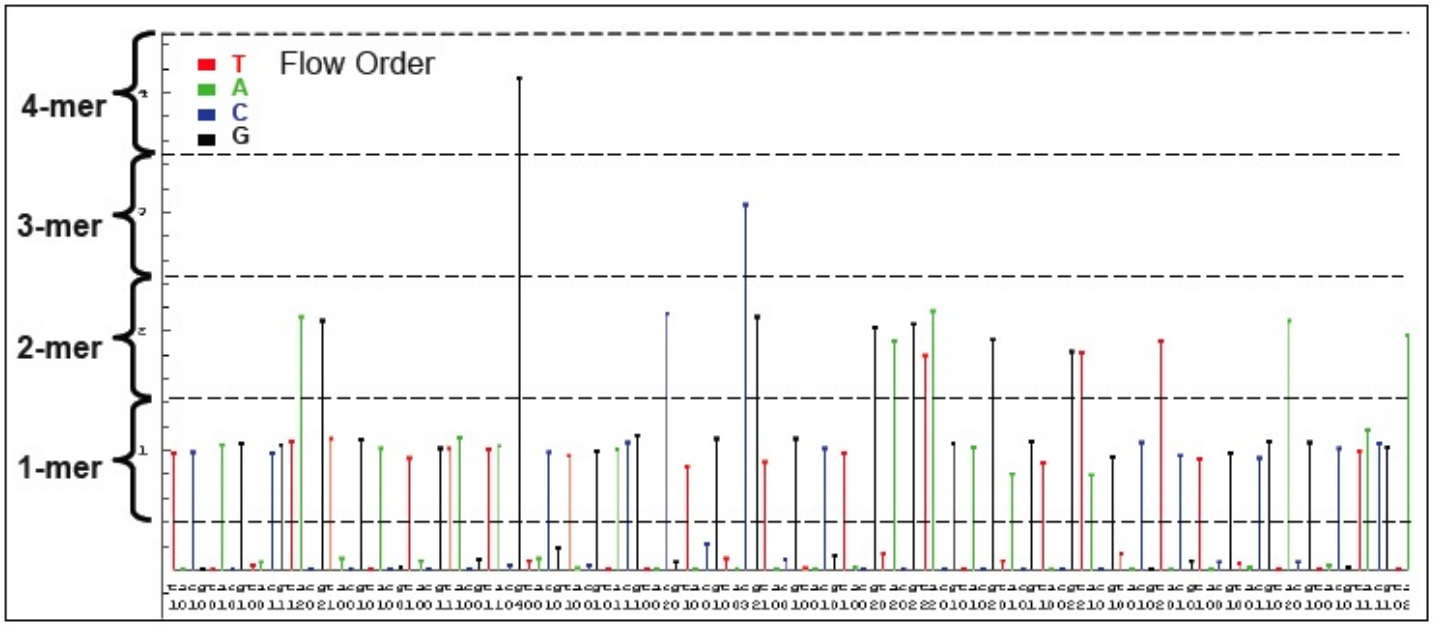
Roche 454 Workflow Video

454 Video

Roche 454 Workflow

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

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 suse 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 at a time, rather than multiple bases (in a homopolymer run) as does Roche 454.

Illumina Platforms

Illumina currently has 2 platforms, the MiSeq benchtop version (what we have on campus) and the HiSeq (with 4 variations).

	HIGH OUTPUT RUN MODE*			RAPID RUN MODE*		
Read Length	Dual Flow Cell (HiSeq 2500 only)	Single Flow Cell (HiSeq 1500 or 2500)	Dual Flow Cell Run Time	Dual Flow Cell (HiSeq 2500 only)	Single Flow Cell (HiSeq 1500 or 2500)	Dual Flow Cell Run Time
1 x 36	95-105 Gb	47-52 Gb	2 days	18-22 Gb	9-11 Gb	7 hr
2 × 50	270-300 Gb	135-150 Gb	5.5 days	50-60 Gb	25-30 Gb	16 hr
2 x 100	540-600 Gb	270-300 Gb	11 days	100-120 Gb	50-60 Gb	27 hr
2 x 150	N/A	N/A	N/A	150-180 Gb	75-90 Gb	40 hr
Reads Passing Filter	Up to 3 billion single reads or 6 billion paired-end reads	Up to 1.5 billion single reads or 3 billion paired-end reads		Up to 600 million single reads or 1.2 billion paired-end reads	Up to 300 million single reads or 600 million paired-end reads	
Quality	Greater than 85% of bases above Q30 at 2 × 50 bp Greater than 80% of bases above Q30 at 2 × 100 bp			Greater than 85% of bases above Q30 at 2 × 50 bp Greater than 80% of bases above Q30 at 2 × 100 bp Greater than 75% of bases above Q30 at 2 × 150 bp		

Illumina Workflow Video

Illumina Video

Illumina Workflow

- Library Construction
- Cluster Generation
- Sequencing
- image extraction

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

fasta,qual and fastq files

- fasta files

  >sequence1
  ACCCATGATTTGCGA
- qual files

  >sequence1
  40 40 39 39 40 39 40 40 40 40 20 20 36 39 39
- fastq files

  @sequence1
  ACCCATGATTTGCGA
  +
  IIHHIHIIII55EHH

phred scores

$$Q = -10 log_{10} P$$

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%

phred score conversion

$Q_{sanger} = -10 log_{10} P$ - based on probability (aka phred)
$Q_{solexa} = -10 log_{10} \frac{P}{1-P}$ - based on odds

S - Sanger	Phred+33,	raw reads typically (0, 40)
X - Solexa	Solexa+64,	raw reads typically (-5, 40)
I - Illumina 1.3+	Phred+64,	raw reads typically (0, 40)
J - Illumina 1.5+	Phred+64,	raw reads typically (3, 40)
L - Illumina 1.8+	Phred+33,	raw reads typically (0, 41)

Library Preparation types

- Shotgun - randomly fragmented DNA (100bp - 1kb)
- RNA - Random nanomers or 3' bias (stranded or unstranded)
- Amplicons
- Paired end / Mate pair

Ion Torrent Workflow Video

Ion Torrent, first available in 2011, generates up to 400bp reads (reported) and up to 2Gb per run. Cheap fast runs. Ion Proton system available soon(?). 200-bp fragments and up to 10Gb per run. Generates flowgrams and SFF files similar to Roche 454 data.

Ion Torrent Video

Pacific Biosciences Workflow Video

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

Pacific Biosciences Video