Introduction to Sequencing

BCB 504: Applied Bioinformatics

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

February 29, 2012

Outline

History

Roche 454 Pyrosequencing

Illumina Solexa

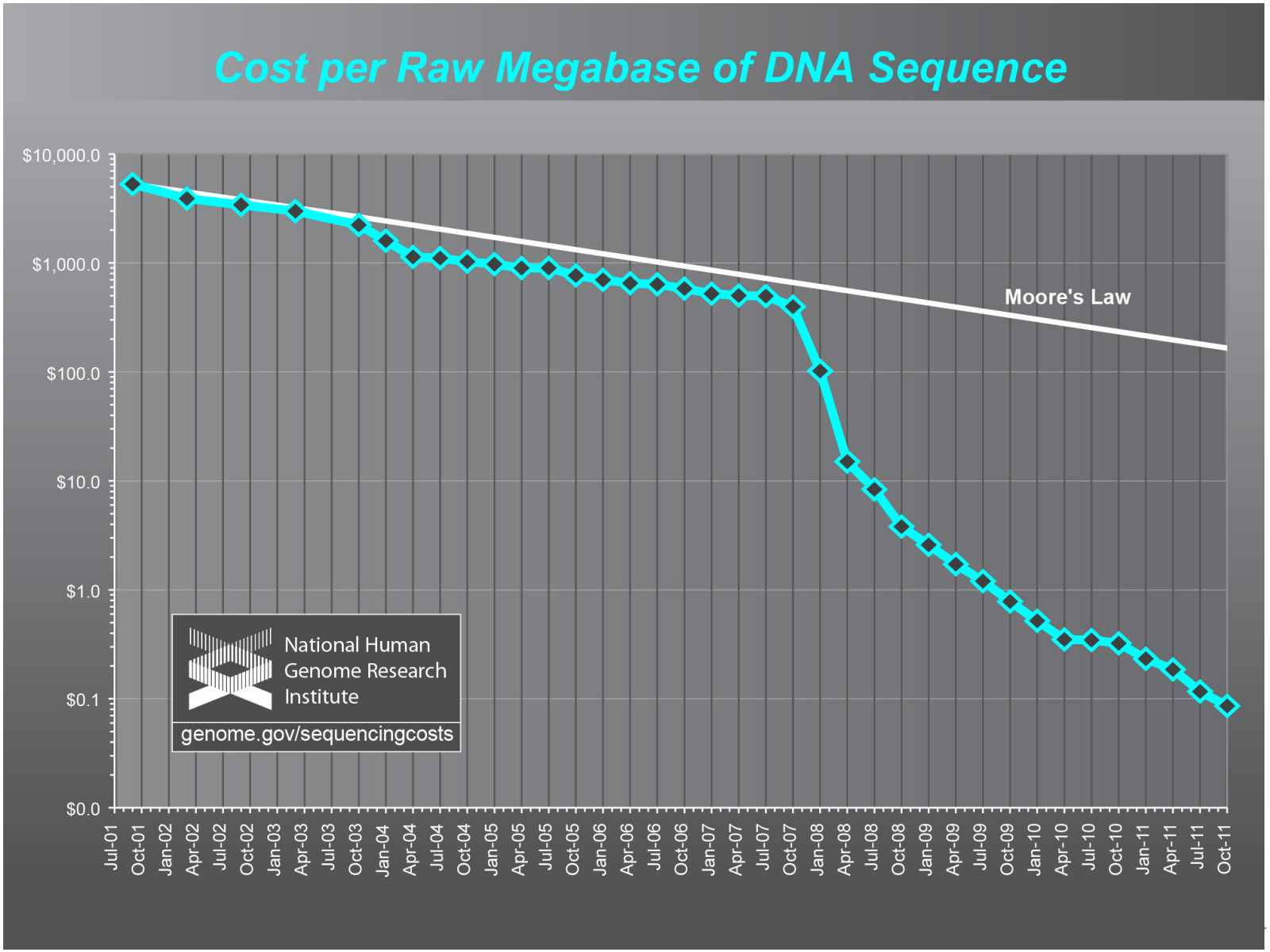
Sequence Data

Library Preparation

Other sequencing technologies

Evolution of DNA Sequencing

Jan - 2012: $0.09 per Megabase, $7,950 per Genome (30x coverage)



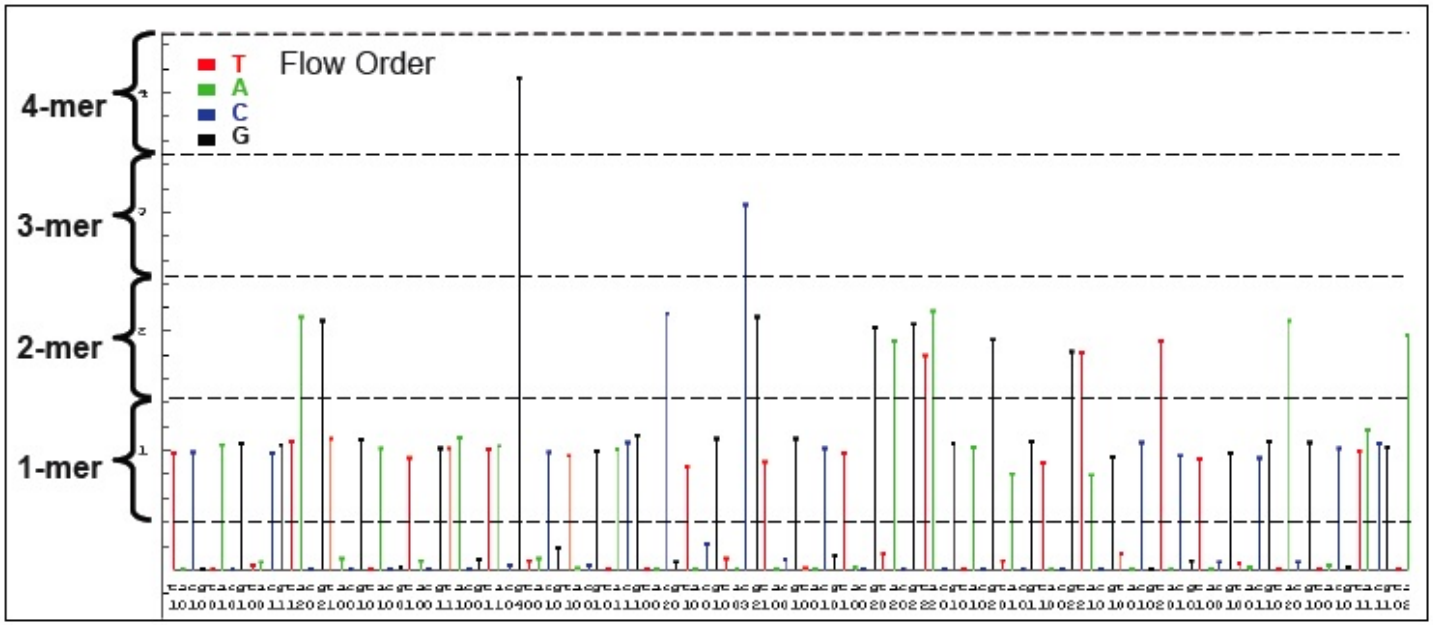
Roche 454 Workflow

- 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



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

454 Read Naming Convections

>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.

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
- Amplicons
- Paired end / Mate pair

Ion Torrent Workflow Video

Ion Torrent Video

Pacific Biosciences Workflow Video

Pacific Biosciences Video