

RNA-Seq, SAM/BAM

BINF Programming I

Outline

- RNA-Seq and Illumina sequencing
- Using IGB

Illumina sequencing

- Applications
 - genome sequencing, transcriptome sequencing, assessing gene expression



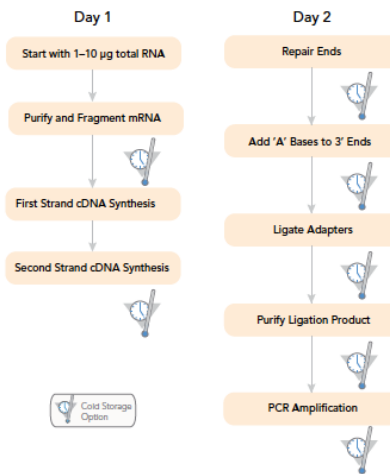
Flow Cell has 8 channels

8 "lanes"

Illumina Genome Analyzer II

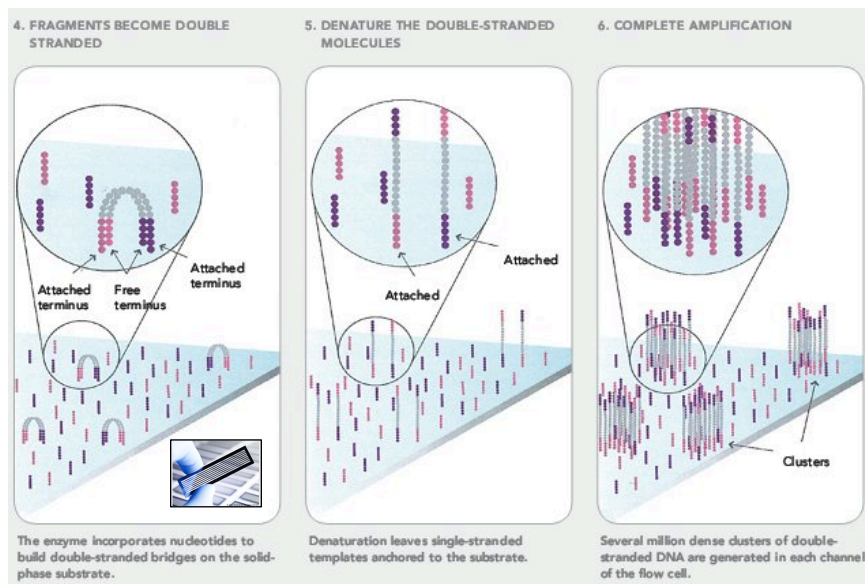


RNA Seq sample prep

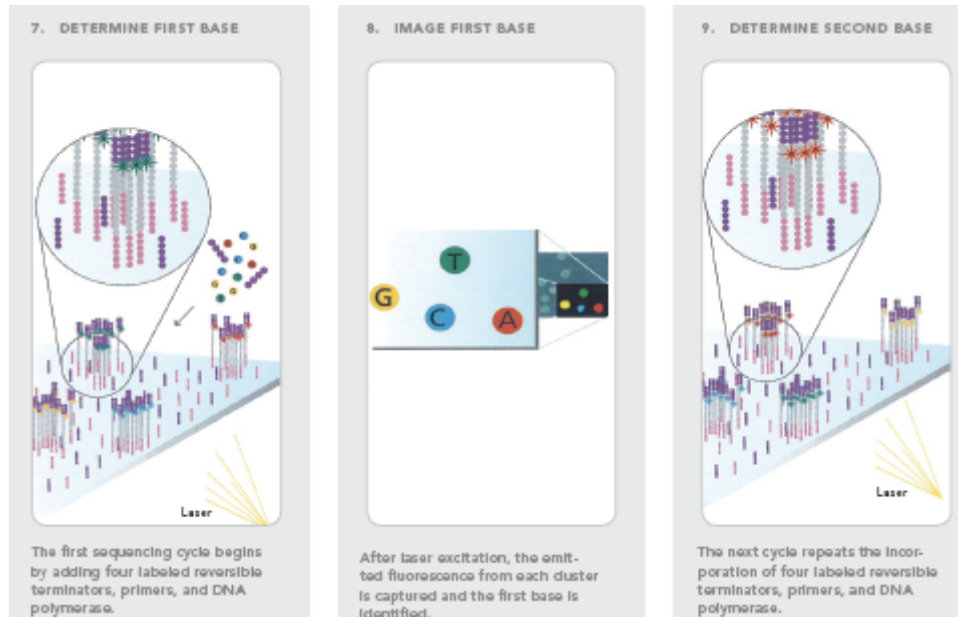


<http://www.illumina.com/applications.ilmn>

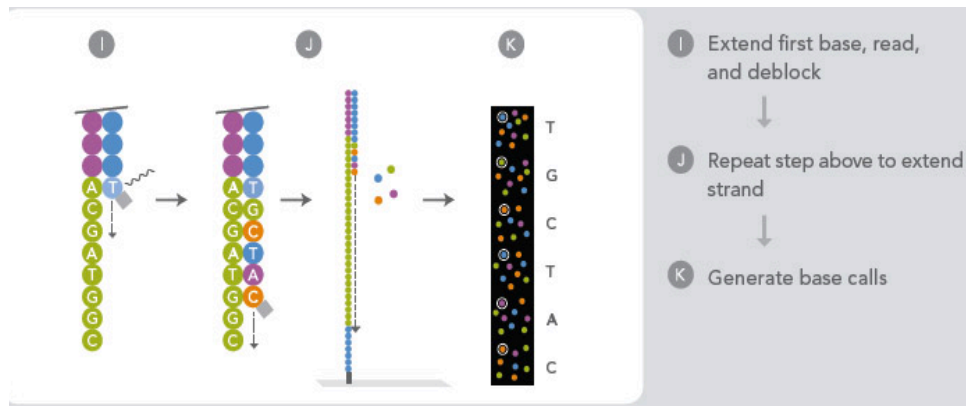
Clustering



Sequencing by synthesis



Sequencing by synthesis



FASTQ format – four lines per sequence

- Line 1 begins with "@"
– contains sequence identifier
- Line 2 is sequence letters
- Line 3 begins with '+' and contains optional description
- Line 4 contains quality values for sequence in Line 2

FASTQ – 4 lines per record

```
@61CFUAAXX:4:1:1205:16349#0/1
CTGGTGGCTGTGAAGACGAAGNAACTTAAGCANTGGNNNCAAACCTCCTTGTC
+61CFUAAXX:4:1:1205:16349#0/1
bbbbbbbbbbbbbbbbba^^B^^^^^^^^^^B\OBBBBNOONbbbbbbbbbb
```

FASTQ files are huge! contain millions of reads

Illumina sequence identifiers

[\[edit\]](#)

Sequences from the [Illumina](#) software use a systematic identifier:

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, <i>I1</i> or <i>I2</i> (<i>paired-end</i> or <i>mate-pair</i> reads only)

Versions of the Illumina pipeline since 1.4 appear to use **#NNNNNN** instead of **#0** for the multiplex ID, where **NNNNNN** is the sequence of the multiplex tag.

For RNA-Seq, align sequences onto reference genome

- See:
 - <https://wiki.transvar.org/confluence/x/woUHAQ>
- Two tools commonly used
 - bowtie
 - finds alignments for reads that don't cross intron-intron boundaries
 - tophat – a spliced alignment tool

Use samtools to sort, index alignments

- More in the next lecture
- Demo – using IGB to visualize short read alignments
- launch IGB
 - www.bioviz.org/igb