STAT588/BIOL588: Genomic Data Science
Lecture16: Next Sequencing Data Anlaysis
(Chapter 5 in Gondro's book)

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

Next Generation Sequencing (NGS) Data Analysis: An Overview

- Introduction
- NGS and Microarray
- Study Design
- Explore and Download Data from SRA
- Quality Assessment

Introduction

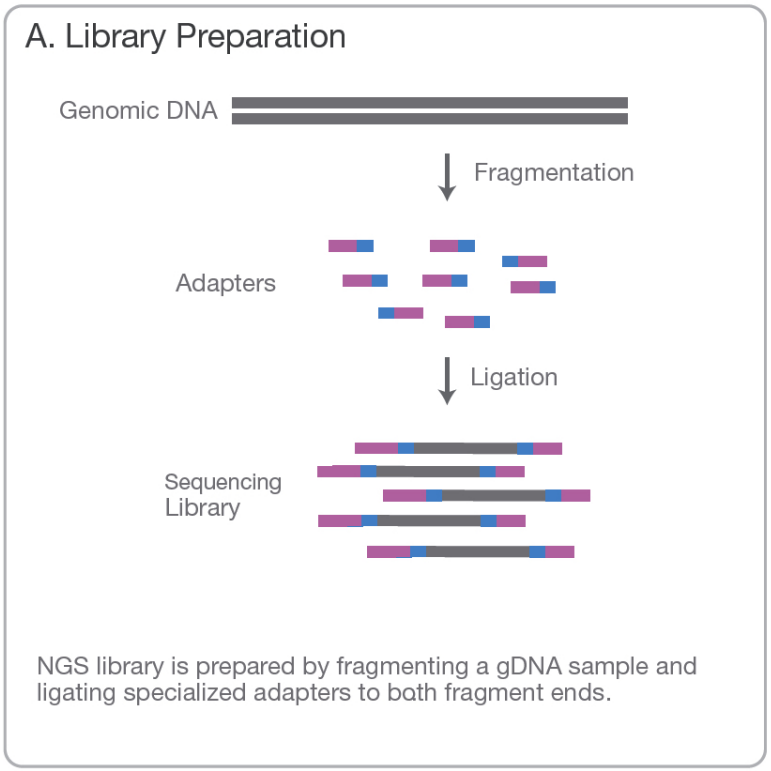
Over the past 10 years or so there has been rapid development of methods for next generation sequencing. There are 3 major producers of platforms for next generation sequencing:

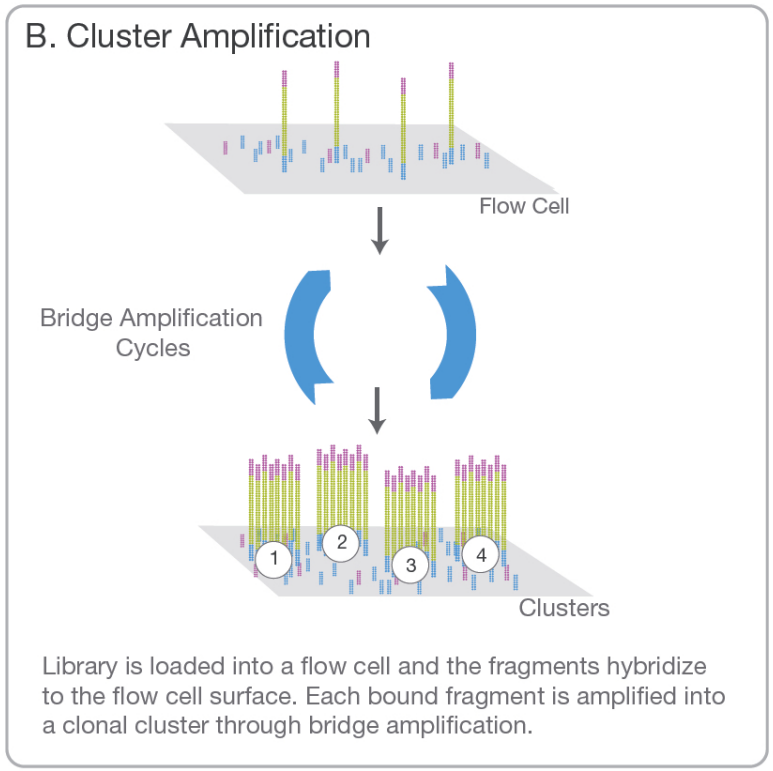
- Illumina: HiSeq, MiSeq
- LifeTech: SoLiD, Ion Torrent
- Roche 454
- Others: Pacific Bioscience, Complemte Genomics, Helicos, OxfordNanopores.

We will use Illumina technology's experimental process as an example.

Experimental Procedure

Sample prep $\rightarrow$ cluster amplification$\rightarrow$ sequencing $\rightarrow$ data analysis





An Illumina flow cell has 8 lanes and is covered with oligonucleotides that bind to the adaptors. Then bridge-amplifications creates a collection of hundreds of millions of sequence read clusters.

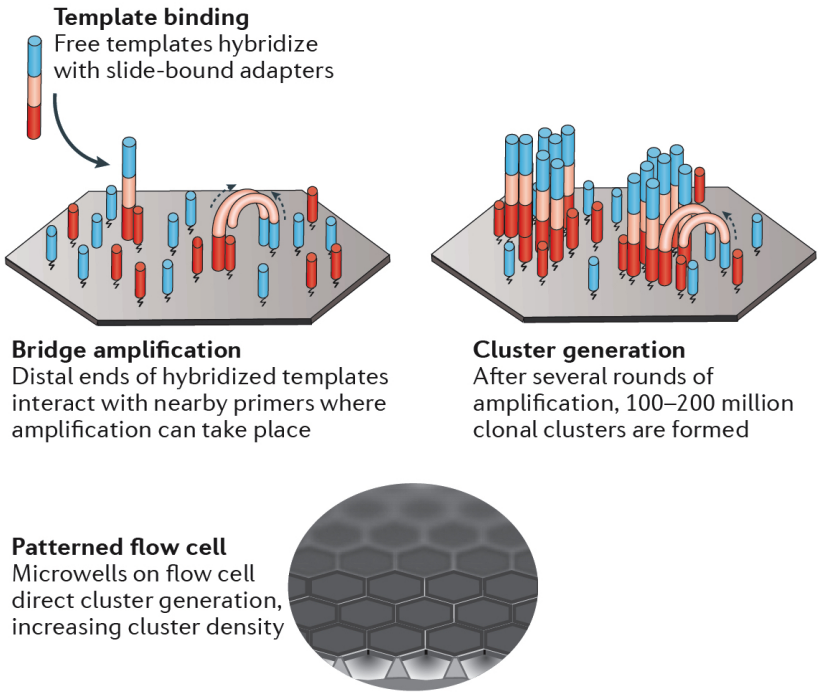
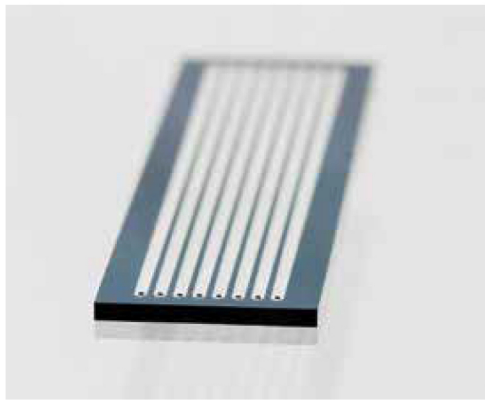


Figure 1: Illumina Flow Cell

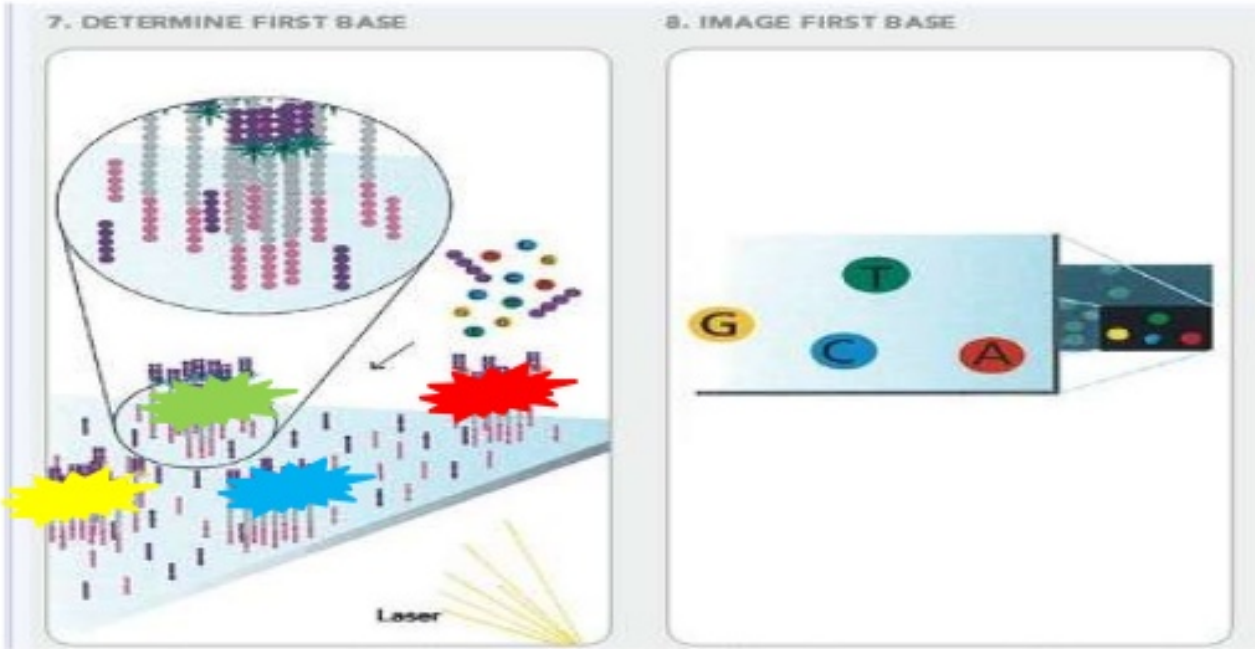


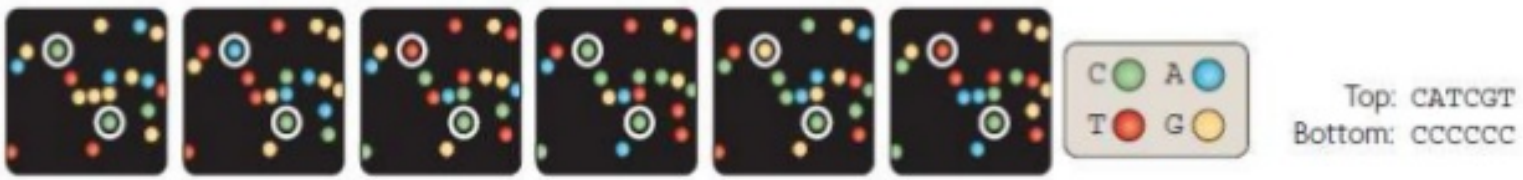
Several samples can be loaded onto the eight-lane flow cell for simultaneous analysis on an Illumina Sequencing System.

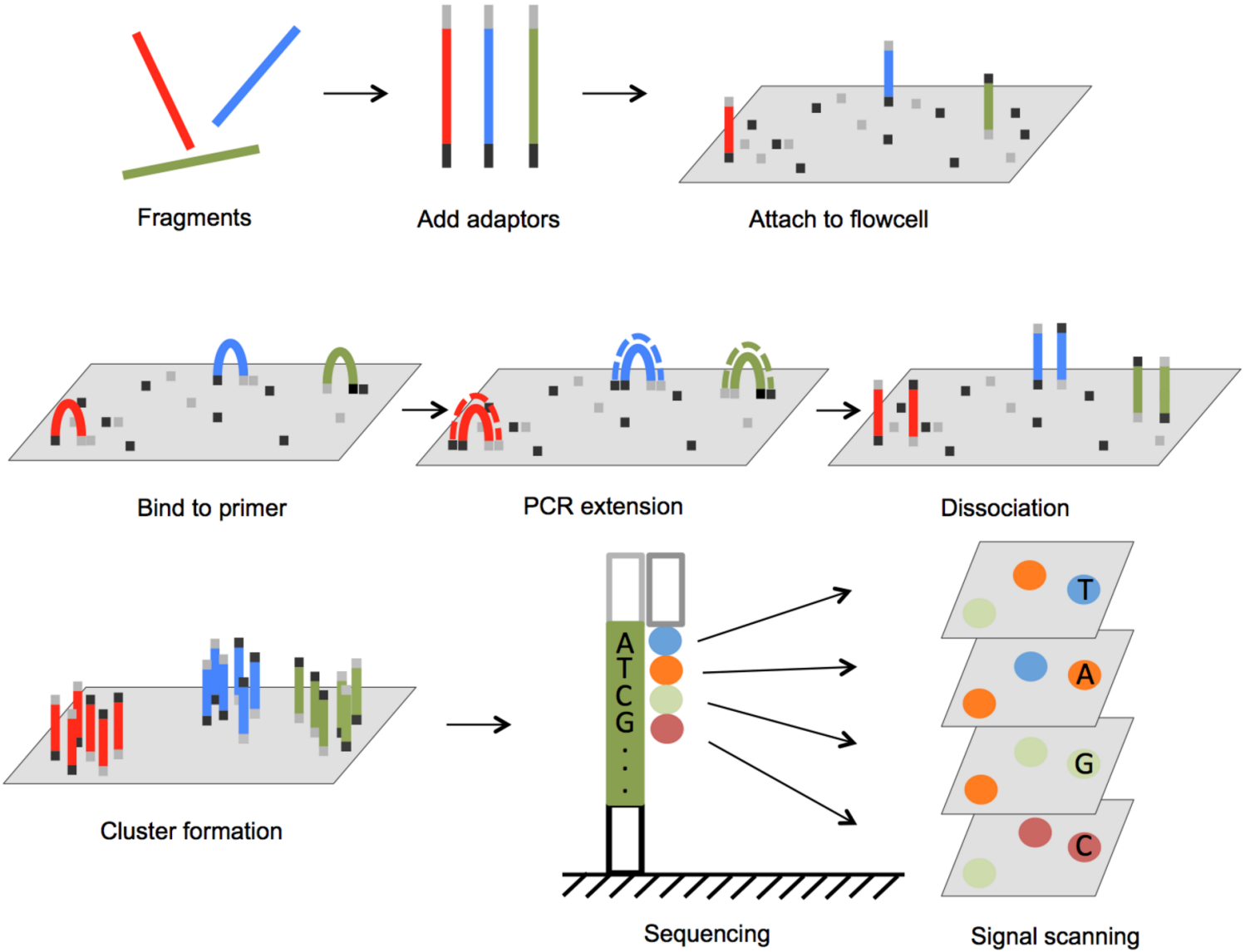
video at `https://www.youtube.com/watch?v=pfZp5Vgsbw0`

`https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=27s`

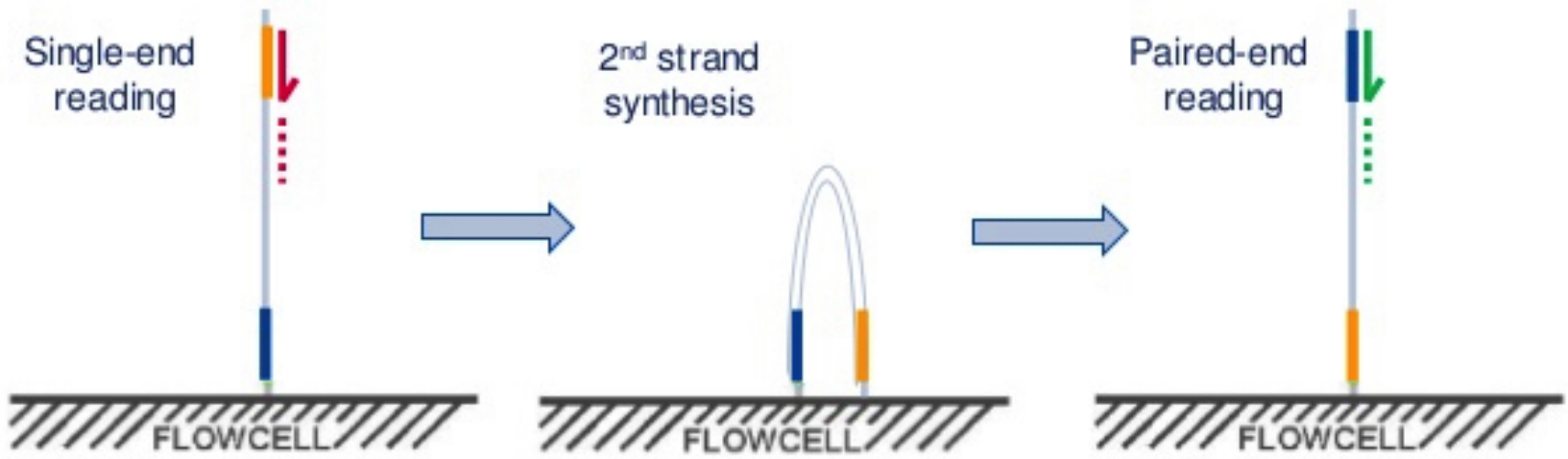
SEQUENCING BY SYNTHESIS






Fragments
Add adaptors
Attach to flowcell
Bind to primer
PCR extension
Dissociation
Cluster formation
ATCG...
Sequencing
T
A
G
C
Signal scanning

Illumina single-end vs. paired-end

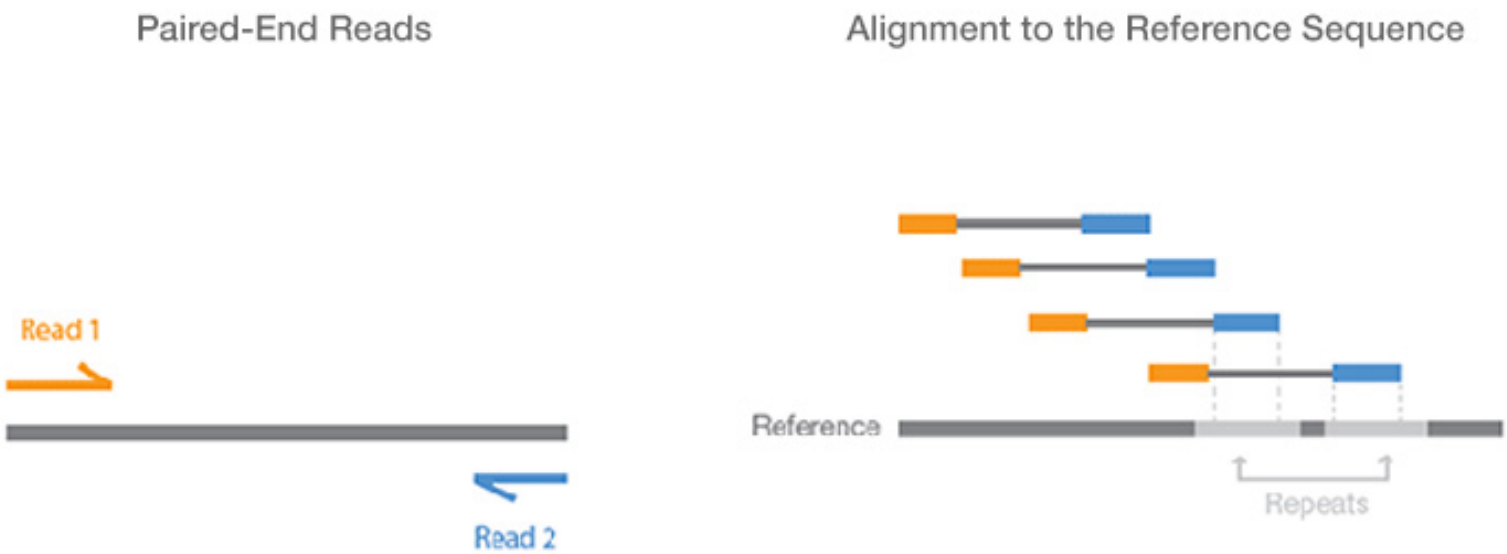


Single-end reading (SE):

- Sequencer reads a fragment from only one primer binding site

Paired-end reading (PE):

- Sequencer reads both ends **of the same fragment**
- More sequencing information, reads can be more accurately placed ("mapped")
- May not be required for all experiments, more expensive and time-consuming
- Required for high-order multiplexing of samples (indexes on both sides)



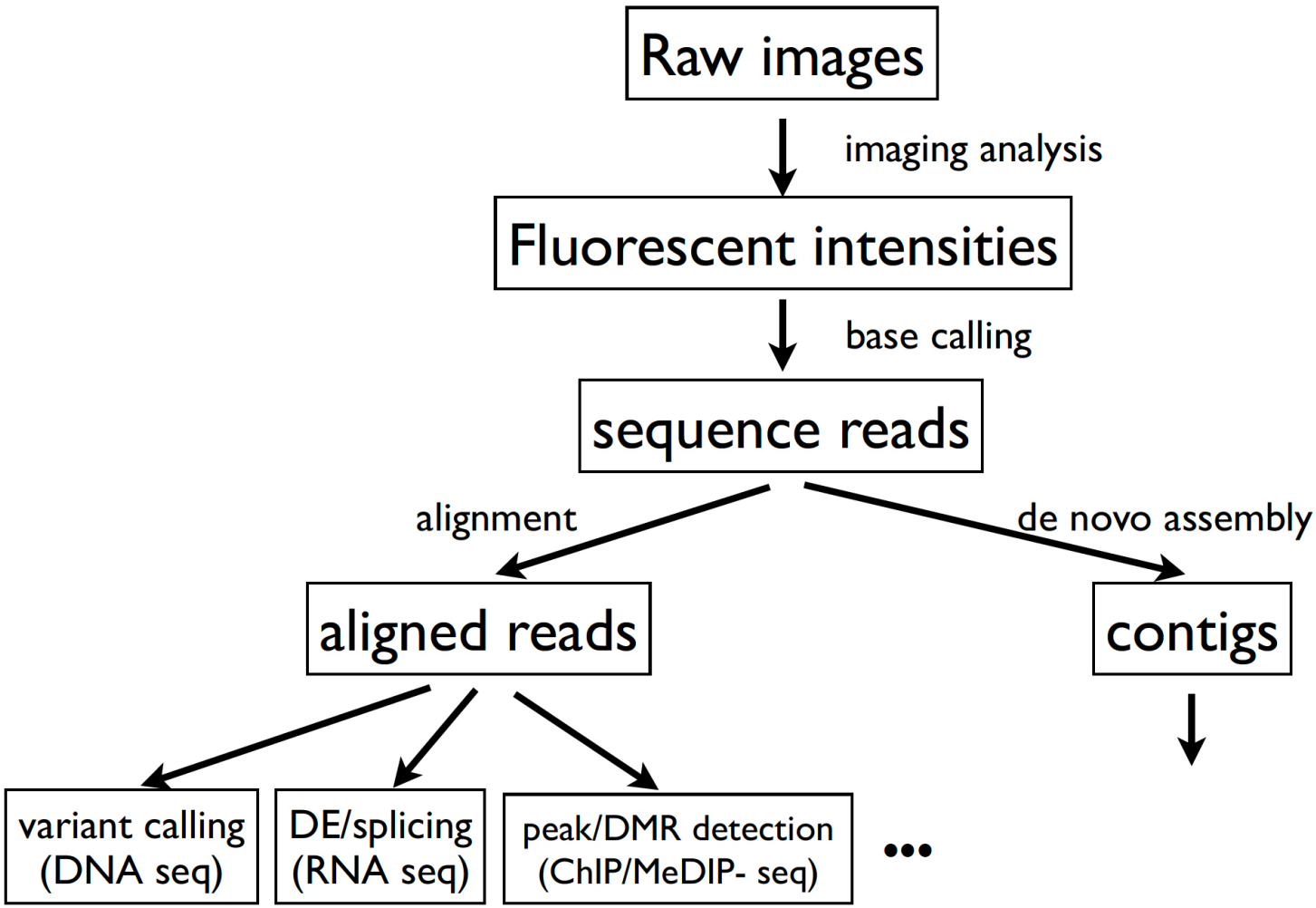
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Example: Illumina HiSeq 2000/2500 Output

Read Length	Run Time	Reads Passing Filter	Filtered Output
Single-end 50bp	$\sim$ 2 day	$\sim$ 250 M/lane	$\sim 12.5 GB/lane$
Paired-end 2$\times$ 50 bp	$\sim$ 2.5 day	$\sim$ 500 M/lane	$\sim 25 GB/lane$
Single-end 100bp	$\sim$ 3.5 day	$\sim$ 250 M/lane	$\sim 25 GB/lane$
Paired-end 2$\times$ 100 bp	$\sim$ 5 day	$\sim$ 500 M/lane	$\sim 50 GB/lane$

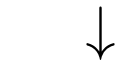
If our goal is to characterize copy number variations then getting paired-end sequences is crucial.

NGS Data Analysis Work Flow



Base Calling

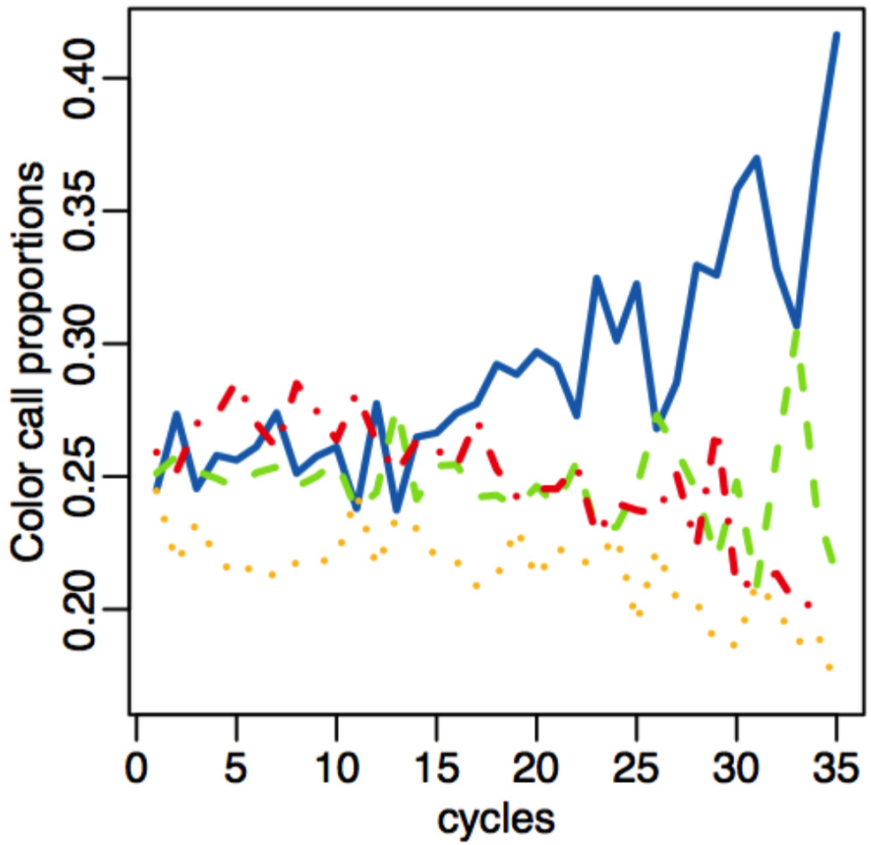
	Cycle1	Cycle2	Cycle3	Cycle4	Cycle5
A	0.05	0.08	0.31	0.41	0.14
C	0.47	0.12	0.28	0.40	0.30
G	0.05	0.43	0.17	0.01	0.39
T	0.42	0.37	0.24	0.18	0.16



CGAAG

Base Calling

Bases called are unbalanced toward the end of the reads.

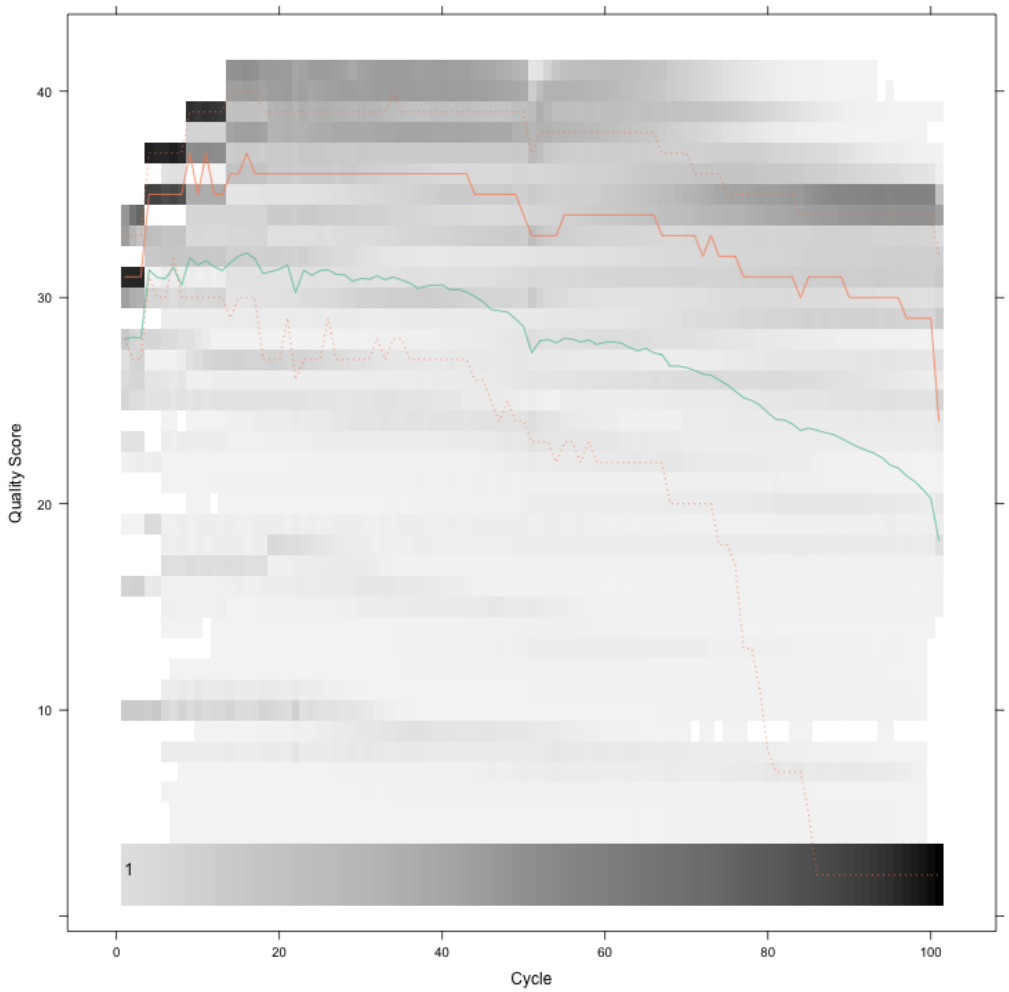


PHRED score

The quality of base called is measured on the PHRED scale, so if there is an estimated probability of an error of p, the PHRED based score is $10 \times -\log_{10} p$.

Phred quality score	Probability that the base is called wrong	Accuracy of the base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

ShortRead: QCreport



Read quality starts to drop too low (score 30) after 80bp.

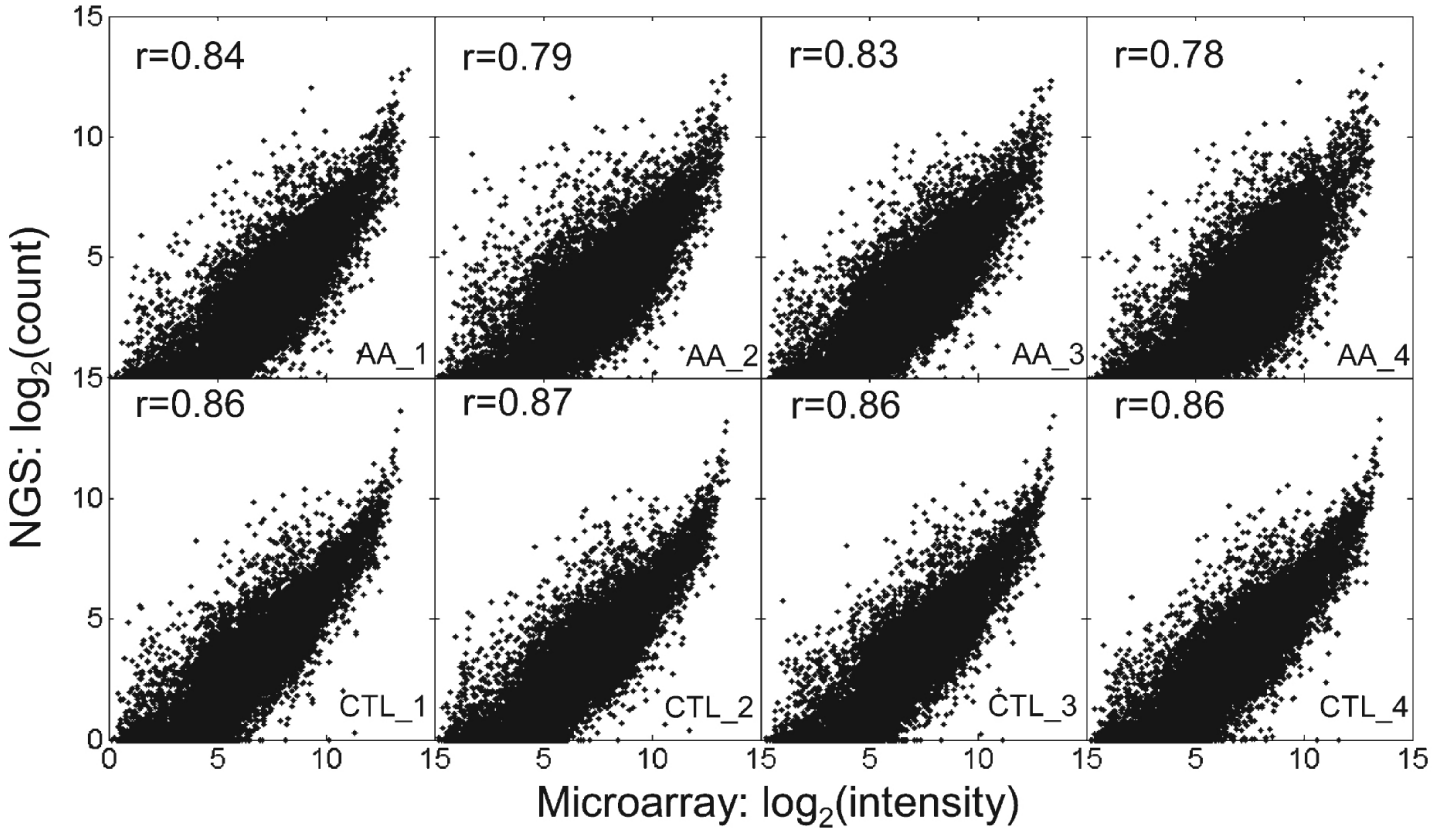
NGS Applications

- resequencing (DNA-seq)
- gene expression (RNA-seq)
- miRNR discovery and quantification
- DNA methylation
- Transcription factor binding site: Chip-seq
- metagenomics, microbiome
- ultra-deeop sequencing of viral genomes

We will focus on RNA-seq in this course

NGS vs Microarray

Correlation between NGS and microarray



Study Design

One key consideration in sequencing study is sequencing depth and coverage. The depth of coverage is how many overlapping fragments will cover a random location.

coverage = (length of reads $\times$ number of reads)/ (length of genome)

By deep sequencing, we mean coverage of 30X to 40X, whereas coverage of 4X or so would be considered low coverage.

Study Design

For a given sequencing budget, there is a tradeoff between the depth of coverage and the number of subjects on can sequence.

If the goal is to accurately sequence a tumor cell, for example, then deep coverage is appropriate.

However if the goal is to identify rare variants, then lower coverage is acceptable. Larger sample size will increase the chances to detect rare variants.

FASTA format

The raw sequence reads generated by the sequencer are stored in FASTA or FASTQ format. Format varies a little between platforms.

- first line starts with @ and sequence information
- second line are the raw sequence letters
- third line s usually just a plus symbol (+)
- forth (last) line contains the quality scores (PHRED or other scoring scheme) for each nucleotide in the second line

RNAseq for 1 sample usually requires around 2Gb storage space.

FASTA format

RNAseq.fastq

```
@DFM3XBQ1:197:D0F53ACXX:2:2305:15538:37703 2:N:0:TGTCAC
CCGCGAGCTACAGGCCCAGCTTCAGGAACAGCAGGTCCAGGTGGAGATGGACATGTCGAAACCAGACCTCACAGCTGCCCTCAGGGACATCCGTGCTCAGT
+
CCCFFDFFHHHHHHIGIIIIJJIJIJJGIIGJJJHIBFHGEI8BHEHHIJEGHIIJCEHHHFFFDDEDDDDDDDDDDDCCBBCDDDCBBBDCA?><<@?C:?4
@DFM3XBQ1:197:D0F53ACXX:2:2305:15528:37720 2:N:0:TGTCAC
TTCAAGTTCTGACCCACTTCAAGGTTGCATCTCCAAGGCAGGGTCTTTGATTGGGCTGCCATCAATAAACTGCAGCCATGAGTTCAGACCAGGAAATGGCT
+
@?@DDDBDDFFH>@FFHBDEGAHGEEDH>ABDFAE@@CFHEHJ0?DGIJB?FH@GEEEDFHIHIFIIGHEHACHCDFFBCAC>ACDCDDDDDDDBCDCCDA
@DFM3XBQ1:197:D0F53ACXX:2:2305:15822:37578 2:Y:0:TGTCAC
GTCACATTCGAGTGGCGATACGGGTGCATGACATTCCAGGTCACAATGCGGGCAGTTTTGTCTATGTCCATACGGGGACAAGGAAGTCCTAGACGATAACC
+
##########################################################################################################
@DFM3XBQ1:197:D0F53ACXX:2:2305:15961:37617 2:N:0:TGTCAC
AAAACATGAATCTTAAAAAAAACGAAAAACTGGCTTTCAGACTTTAAAATAAGCCTCCTCGTCCTTCACAGCTATCCTTCAAATATTTTAAGCAGAAAATT
+
CCCFFFFFHGGHHJJJJJJJJJJJJJIJJJJIIJJGIIJGGCEGHIGIJJGGEEHHGHHFFBBCDDDDDDDABCCDDDDCCDDDDDDECDDDDCDCCDDDD@
@DFM3XBQ1:197:D0F53ACXX:2:2305:16032:37567 2:N:0:TGTCAC
GTTTGACAAAGGCTTTTGCCGGGCCATGATCGAGGAAATTGGCCCAGAACTTGAAGGGCAAATCCCAGTGTCTAGAGGGTTACCCATCGTTAAGGTGTTCA
+
;B?:=ABDD:<A+<3C@GCCF3BFAFHIGC4:BBAGIGEHDB3=;FCAF3=C@ECEEF(5??3(6>6;@;;35(5;@2/,(,5@>53988?(<@3(:>@A#
@DFM3XBQ1:197:D0F53ACXX:2:2305:16110:37713 2:N:0:TGTCAC
CAAAAATCCTTGATGACATCTTTGCCTCTTTGGTACAAAATAAGACCTGCTCTGATTTATCAAAGAGGCCCCAAGGACTGACCATCAAGCCCAGCATTCTG
+
BCCFFFFFHGHHHJJIJJJJIJJJJJJJJJJJJJJFHHIIGIJJJJJIJJIIGIIJHIIJGJIIIIJJJJJJJJJGHHHFFFFEEEEEC@CABDDDB?CDDDD
@DFM3XBQ1:197:D0F53ACXX:2:2305:16370:37546 2:Y:0:TGTCAC
CGCACCCATTTCATCACTGCCTCAAATACTGCTTCTTCCTTTTCTACATTTAGGCTGTCATTGGAAATGACTGAGATCAGCTCCTGCGGAGACAGCTGCAT
+
?=1D:)ADFDHFCCC>+A3:CC3AEHFCC>+*?1:?9:<:**:*9?4**:*9BD@3?BB<DGCB<<F@@F78@@77C;DA######################
@DFM3XBQ1:197:D0F53ACXX:2:2305:16438:37550 2:Y:0:TGTCAC
CGAGAAGGTGCTGGCTGCTGTCTACAAGGCTCTGAGTGACCACCACATCTACCTGGAAGGCACCTTGCTGAAGCCCAATATGGTAACCCCAGGACACGCCT
+
?8?DDDFA+AFF<<C?BEHDF<H>DEA?;FB?CC>1C19BBBD@=88*0?/BFFB>@DE@@A;@7.7?;;;)7.;=AD<CC####################
```

Read name
Read sequence
Separator
Quality scores

Read fasta format file into R: ShortRead

```
>library(ShortRead)
>seq2<-readFastq("RNAseq.fastq")
>summary(seq2)
    Length      Class       Mode
    153557 ShortReadQ         S4
>slotNames(seq2)
[1] "quality" "sread"   "id"
>head(sread(seq2))
  A DNAStringSet instance of length 6
    width seq
[1]   101 CCGCGAGCTACAGGCCCAGCTT...CTCAGGGACATCCGTGCTCAGT
[2]   101 TTCAAGTTCTGACCCACTTCAA...GAGTTCAGACCAGGAAATGGCT
[3]   101 GTCACATTCGAGTGGCGATACG...AAGGAAGTCCTAGACGATAACC
[4]   101 AAAACATGAATCTTAAAAAAAA...CAAATATTTTAAGCAGAAAATT
[5]   101 GTTTGACAAAGGCTTTTGCCGG...TTACCCATCGTTAAGGTGTTCA
[6]   101 CAAAAATCCTTGATGACATCTT...GACCATCAAGCCCAGCATTCTG
```

ShortRead: QC Report

```
> head(quality(seq2),3)
class: FastqQuality
quality:
  A BStringSet instance of length 4
    width seq
[1]   101 CCCFFDFFHHHHHIGIIIIJJIJI...BBCDDDCBBBDCA?><<@?C:?4
[2]   101 @?@DDDBDDFFH>@FFHBDEGAHG...BCAC>ACDCDDDDDDDBCDCCDA
[3]   101 ##########################...########################
> head(id(seq2))
  A BStringSet instance of length 6
    width seq
[1]    54 DFM3XBQ1:197:D0F53ACXX:2...5538:37703 2:N:0:TGTCAC
[2]    54 DFM3XBQ1:197:D0F53ACXX:2...5528:37720 2:N:0:TGTCAC
[3]    54 DFM3XBQ1:197:D0F53ACXX:2...5822:37578 2:Y:0:TGTCAC
[4]    54 DFM3XBQ1:197:D0F53ACXX:2...5961:37617 2:N:0:TGTCAC
[5]    54 DFM3XBQ1:197:D0F53ACXX:2...6032:37567 2:N:0:TGTCAC
[6]    54 DFM3XBQ1:197:D0F53ACXX:2...6110:37713 2:N:0:TGTCAC
> encoding(quality(seq2))
 !  "  #  $  %  &  '  (  )  *  +  ,  -  .  /  0  1  2  3  4
 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
 5  6  7  8  9  :  ;  <  =  >  ?  @  A  B  C  D  E  F  G  H
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
 I  J
40 41
```

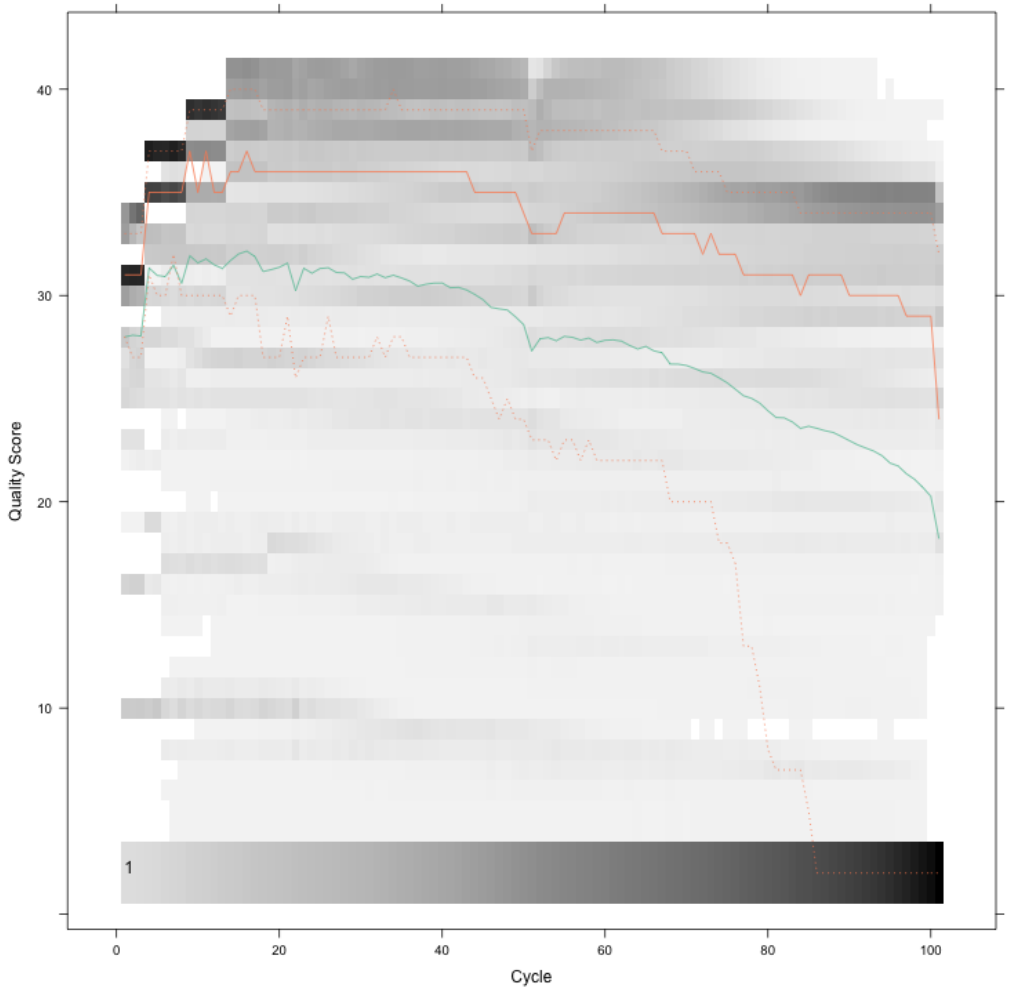
ShortRead: QC Report

```
> tbls<-tables(seq2)
> names(tbls)
[1] "top"          "distribution"
> tbls$top[1:5]
> tbls$distribution[1:3,]
  nOccurrences nReads
1            1 128065
2            2   4256
3            3   1244
> seq2
class: ShortReadQ
length: 153557 reads; width: 101 cycles
> sum(tbls$distribution[,1]*tbls$distribution[,2])
[1] 153557
```

ShortRead: QC Report

```
>seqQC<-qa("RNAseq.fastq")
> report(seqQC, dest="/Users/yen-yiho/Desktop/BIOL599/
Notes/LectureRNAseq1/index.html")
[1] "/Users/yen-yiho/Desktop/BIOL599/Notes/LectureRNAseq1/index.html/index.html
```

ShortRead: QCreport



Read quality starts to drop too low (score 30) after 80bp.

QuasR

The function **preprocessRead** in R package *QuasR* can be used to prepare the input sequences before alignment to the reference genome.

- **Truncate reads**: remove nucleotides from the start and/or end of each read.
- **Trim adapters**: remove nucleotide at the beginning and/or end of each read that match to a defined (adapter) sequence.
- **Filter out low quality reads**: filter out reads that contain more than *nBases* N bases, shorter than *minLength* or low complexity sequence.

Another popular command line tools is **Trimmomatic**.