Introduction to Bioinformatics

Next Gen Sequencing

Stevan Radanović

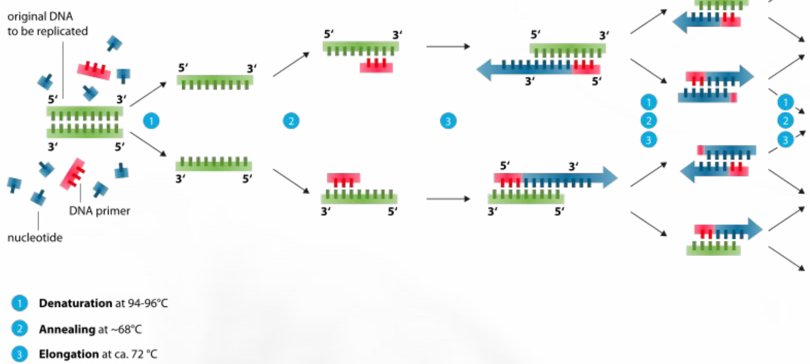
Resources

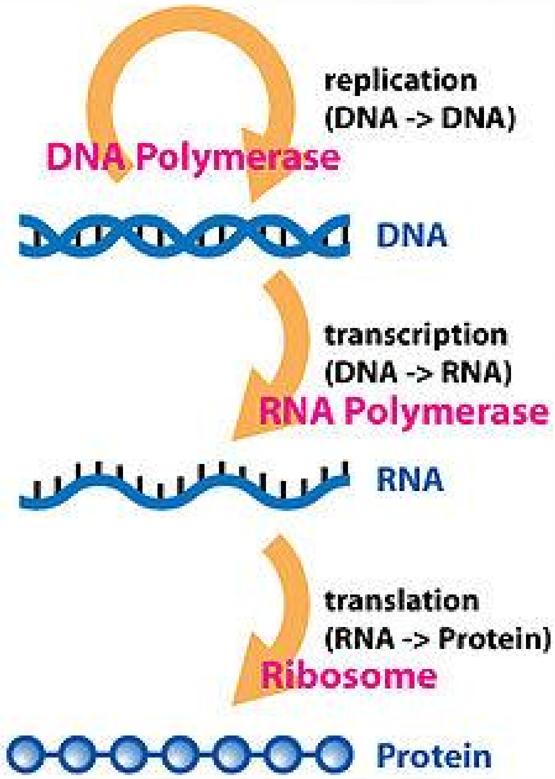
- Illumina inc., https://www.youtube.com/channel/UCxWMU29FF4kIG8YmQf6Zv0g
- AWS https://aws.amazon.com/documentation/ec2/

Refresher

- Central dogma: DNA makes RNA makes Proteins
- PCR

Polymerase chain reaction - PCR

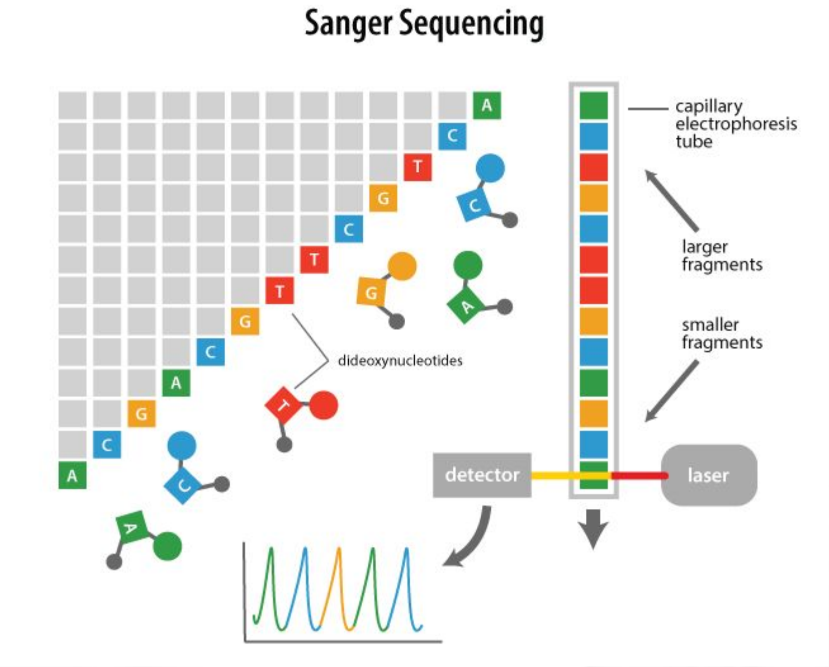




SevenBridges
genomics

Refresher

- Sanger sequencing



Sequencing

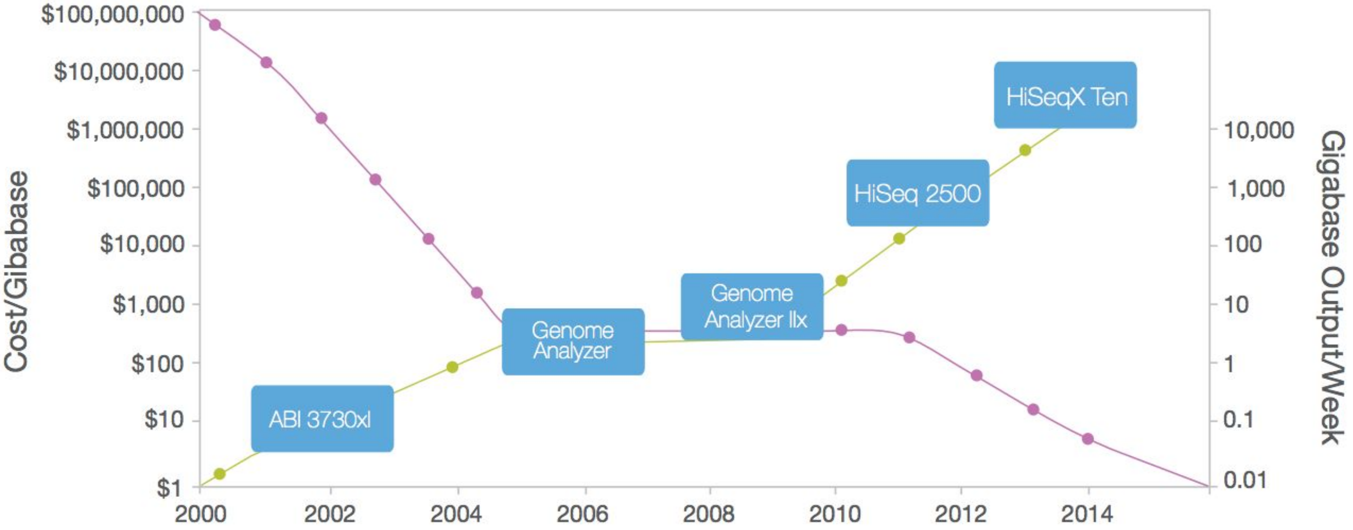
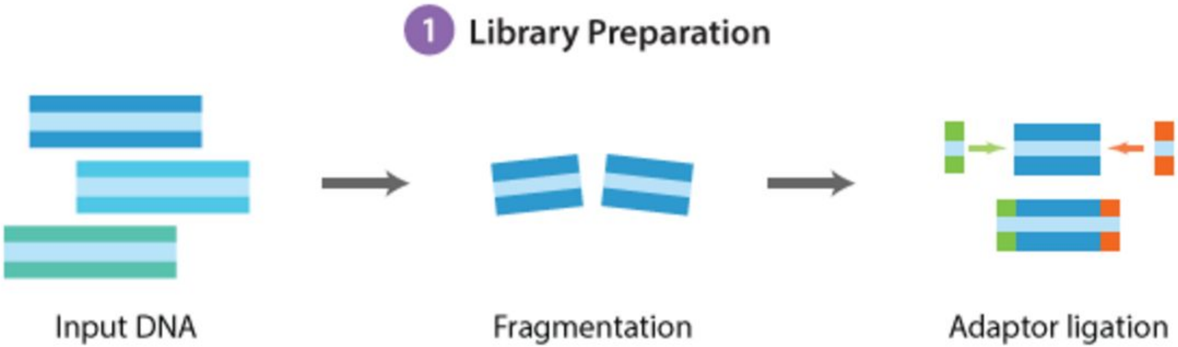
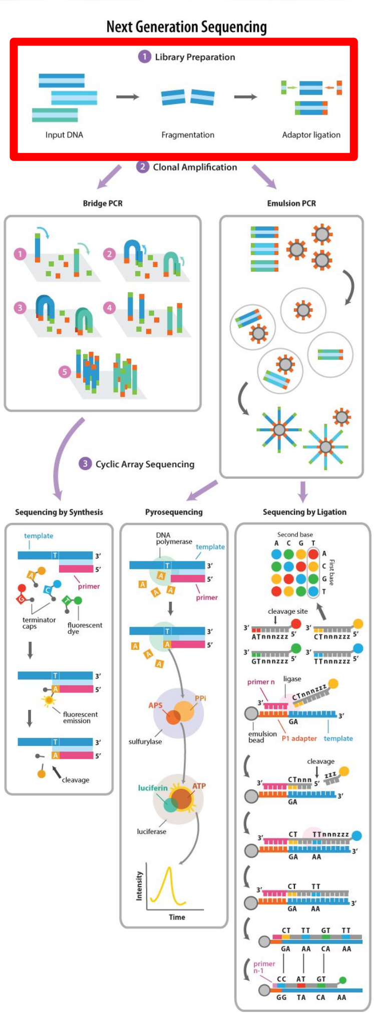


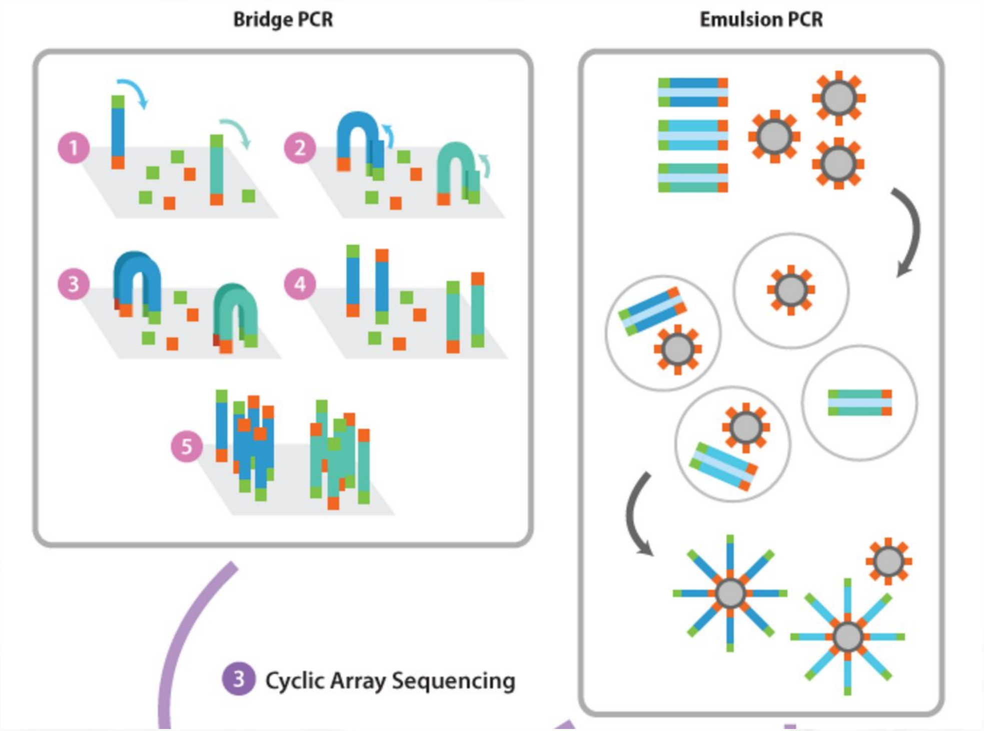
Figure 1: Sequencing Cost and Data Output Since 2000—The dramatic rise of data output and concurrent falling cost of sequencing since 2000. The Y-axes on both sides of the graph are logarithmic.

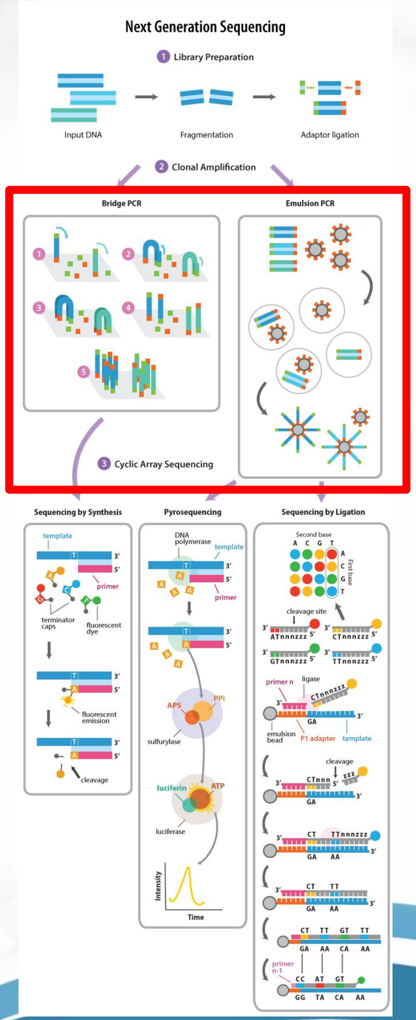
Next Gen Sequencing



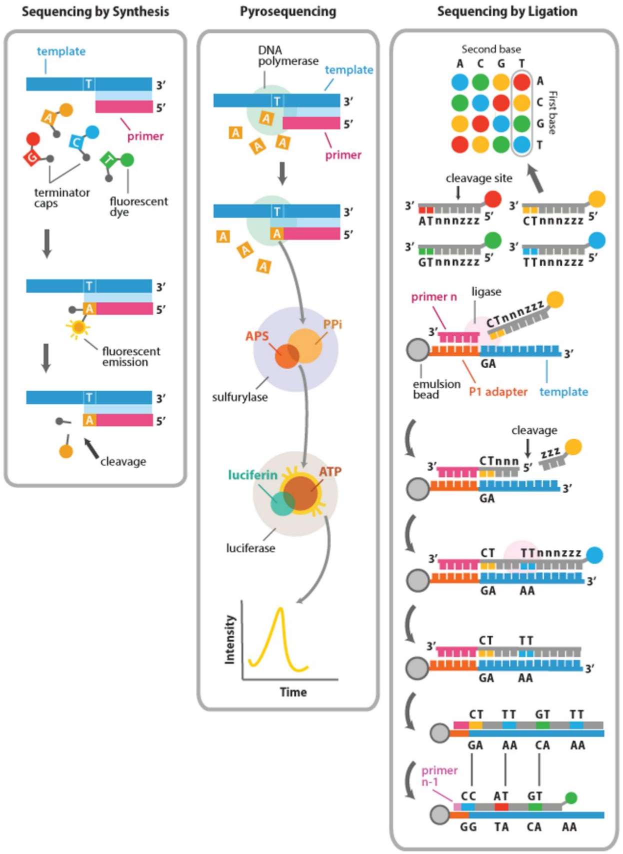


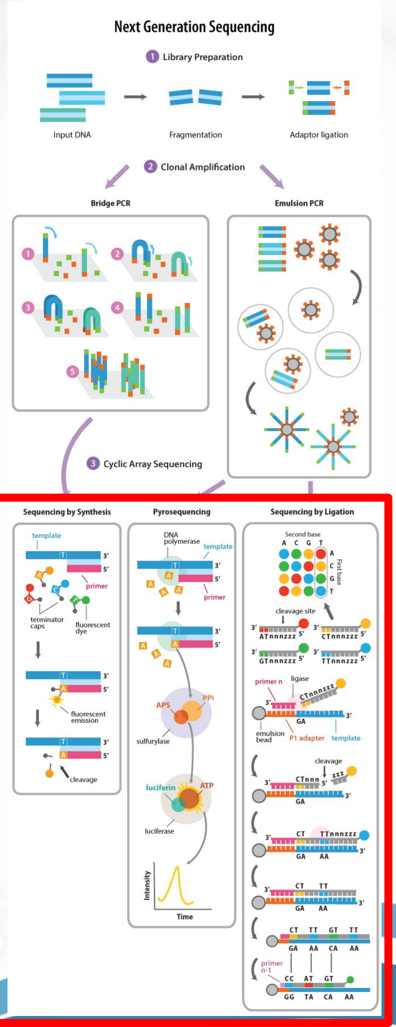
Next Gen Sequencing





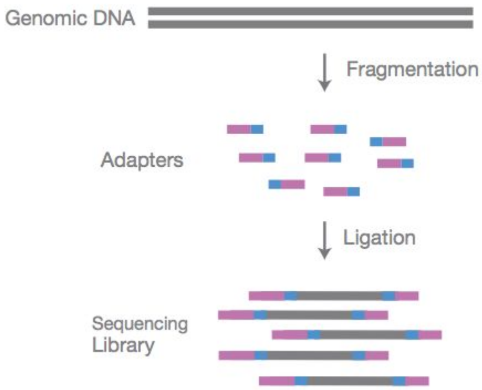
Next Gen Sequencing





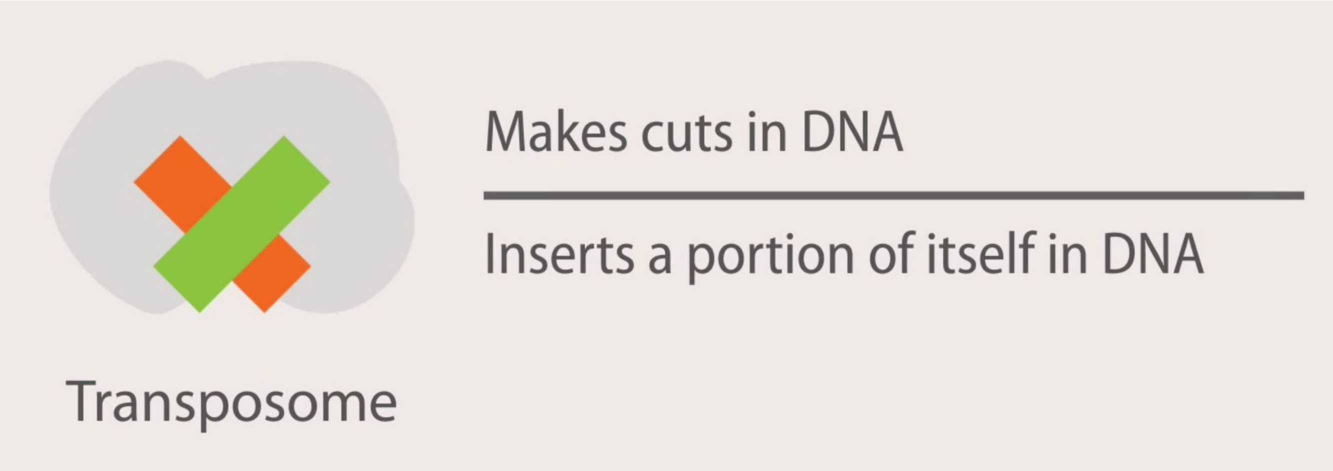
Illumina Sequencing

A. Library Preparation



NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

Illumina Sequencing

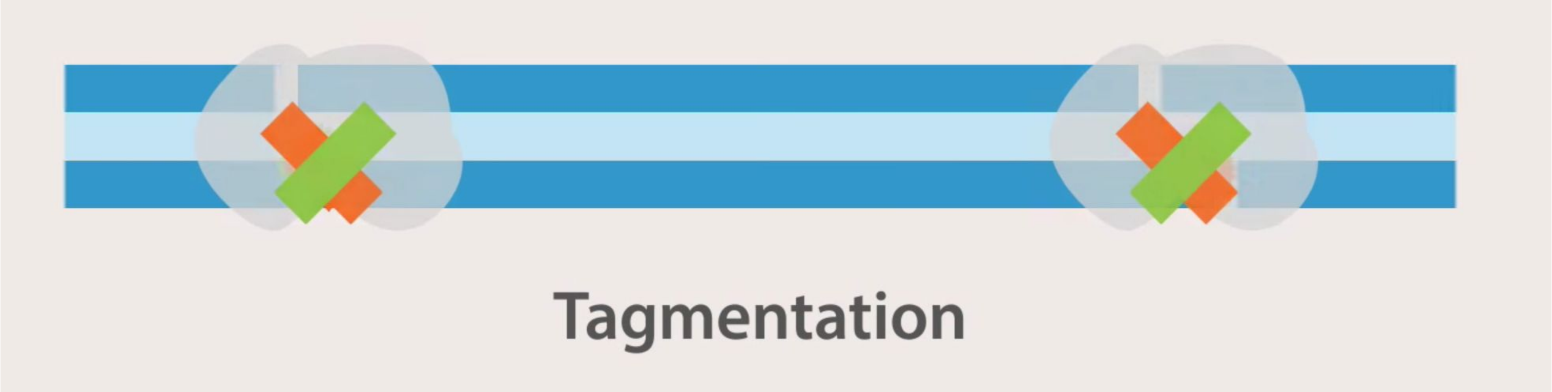


Makes cuts in DNA

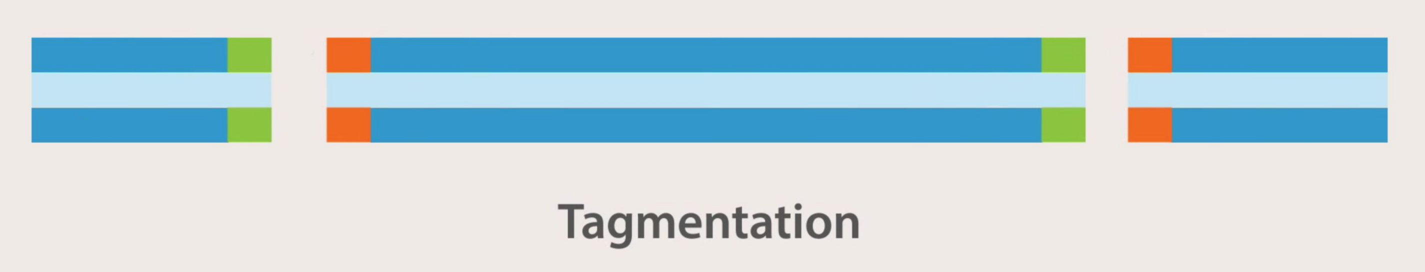
Inserts a portion of itself in DNA

Transposome

Illumina Sequencing

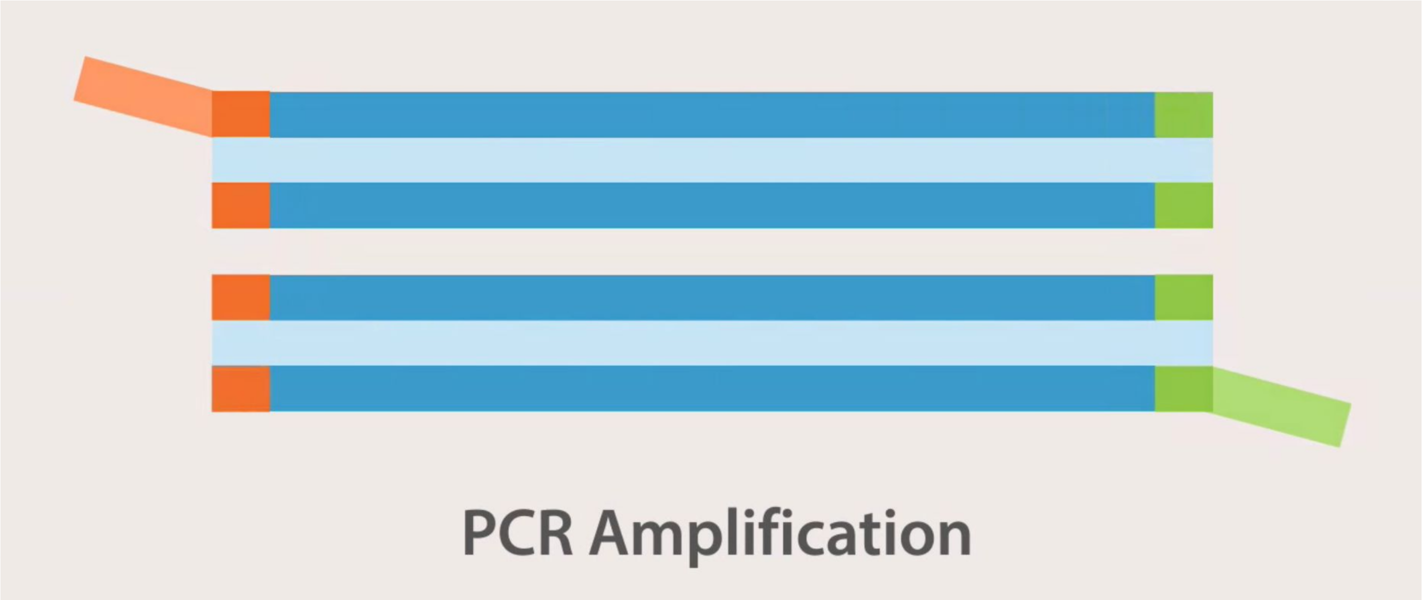


Illumina Sequencing



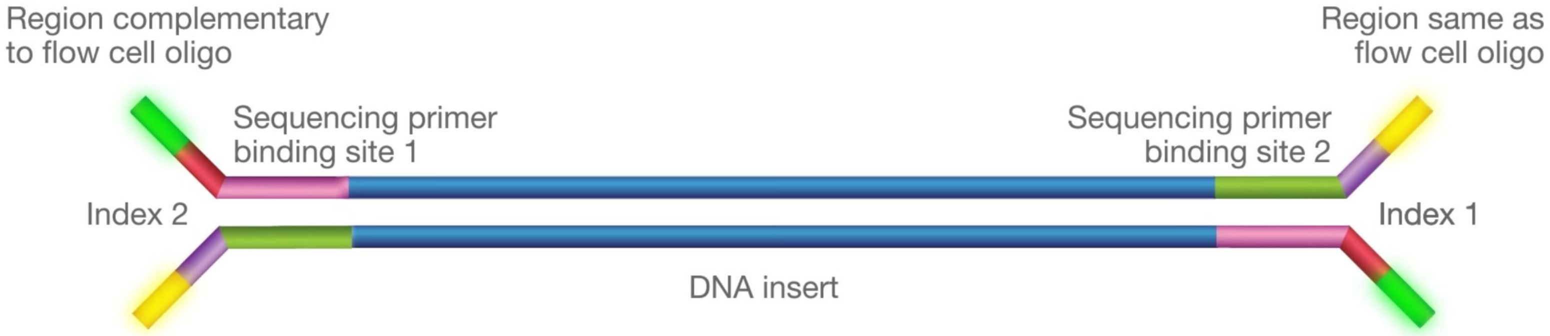
Tagmentation

Illumina Sequencing



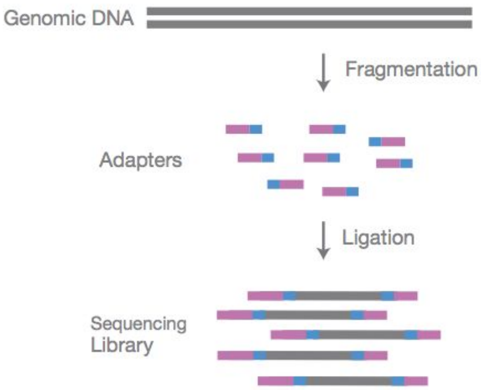
PCR Amplification

Illumina Sequencing


Region complementary to flow cell oligo
Region same as flow cell oligo
Sequencing primer binding site 1
Sequencing primer binding site 2
Index 2
Index 1
DNA insert

Illumina Sequencing

A. Library Preparation



Genomic DNA

Fragmentation

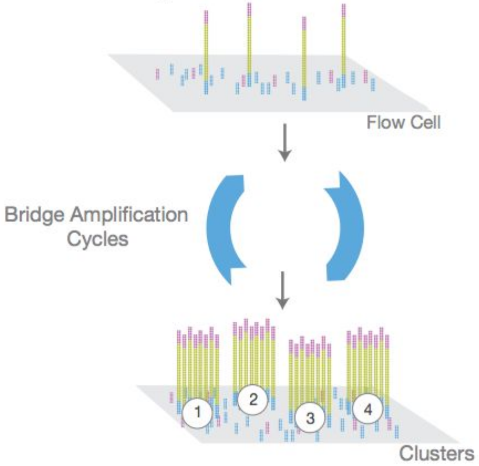
Adapters

Ligation

Sequencing Library

NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

A. Cluster Amplification



Flow Cell

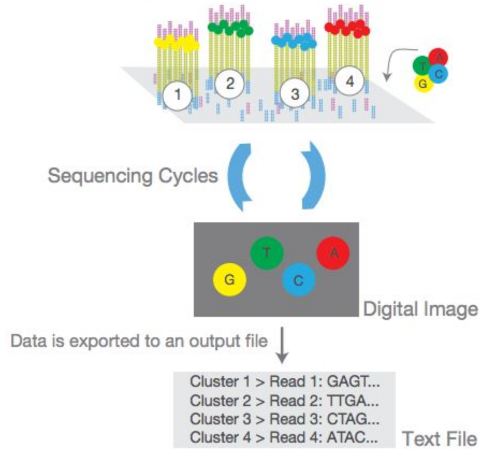
Bridge Amplification Cycles

Clusters

Library is loaded into a flow cell and the fragments hybridize to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

Illumina Sequencing

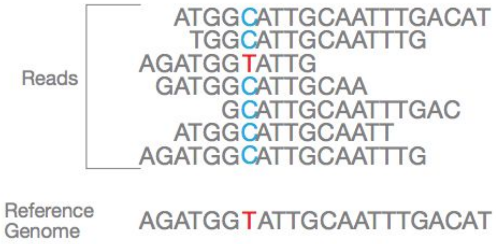
C. Sequencing



Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...

Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

D. Alignment & Data Anaylsis

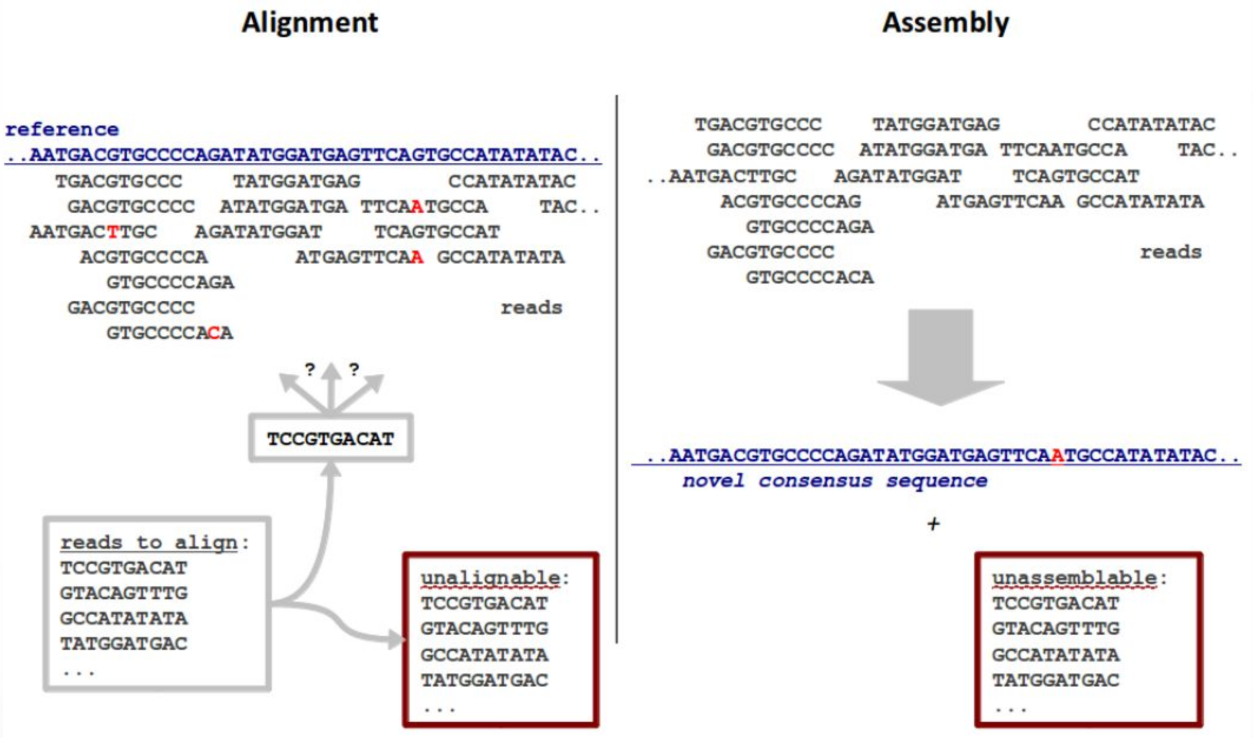


```
  ATGGCATTGCAATTTGACAT
   TGGCATTGCAATTTG
AGATGGTATTG
 GATGGCATTGCAA
     GCATTGCAATTTGAC
  ATGGCATTGCAATT
AGATGGCATTGCAATTTG

AGATGGTATTGCAATTTGACAT
```

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Data Analysis



SevenBridges
genomics

Illumina Sequencing

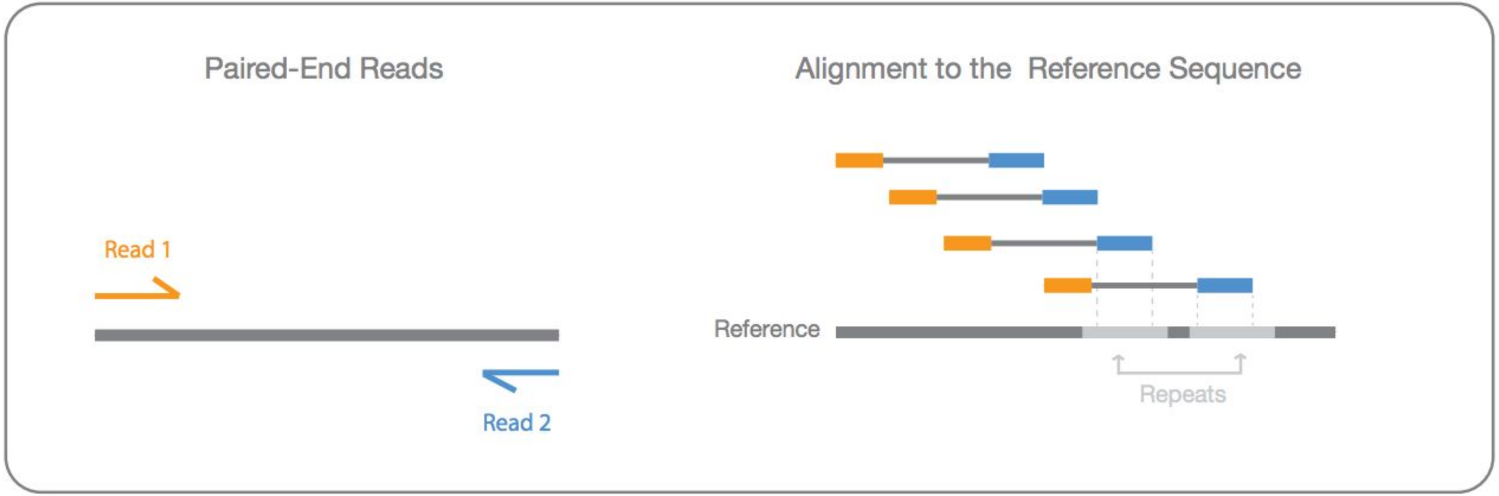


Figure 4: Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Illumina Sequencing

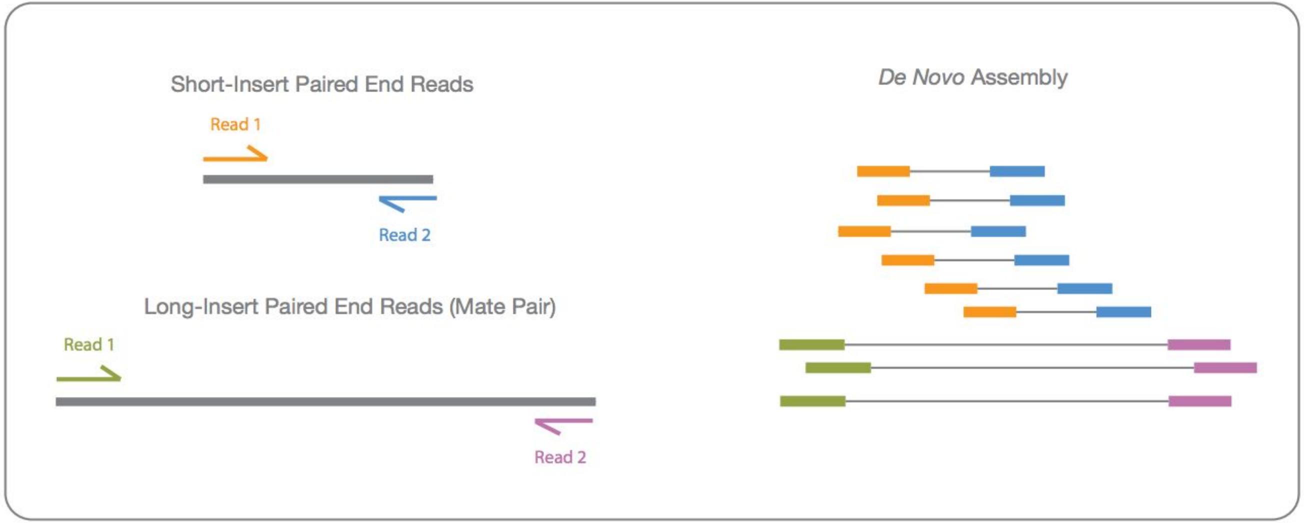

Short-Insert Paired End Reads
Read 1
Read 2
Long-Insert Paired End Reads (Mate Pair)
Read 1
Read 2
De Novo Assembly

Figure 7: *De Novo* Assembly with Mate Pairs—Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for *de novo* assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occurred. Shorter inserts sequenced at higher depths can fill in gaps missed by larger inserts sequenced at lower depths. Thus a diverse library of short and long inserts results in better *de novo* assembly, leading to fewer gaps, larger contigs, and greater accuracy of the final consensus sequence.

Illumina Sequencing

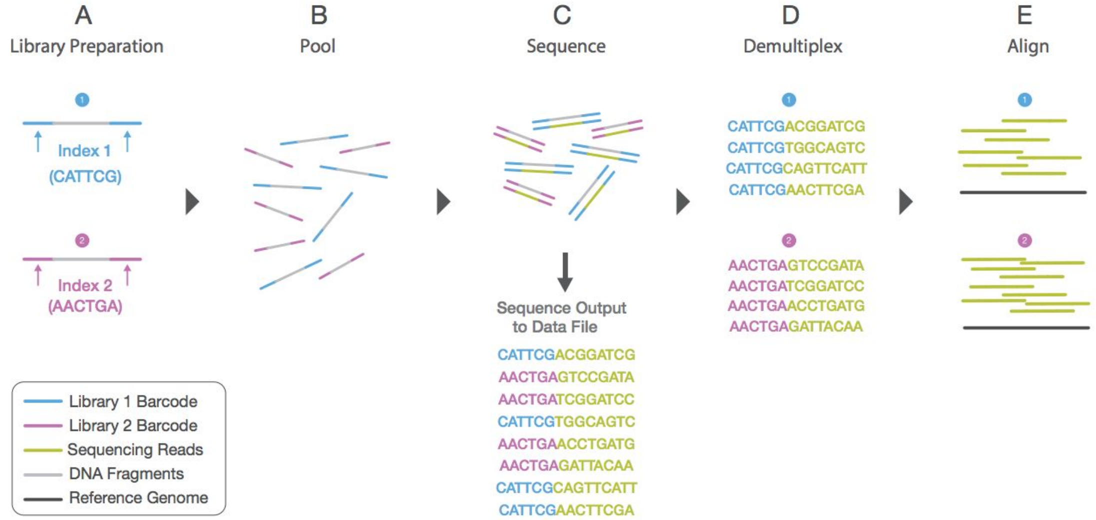


Figure 5: Library Multiplexing Overview.

a. Two distinct libraries are attached to unique index sequences. Index sequences are attached during library preparation.

b. Libraries are pooled together and loaded into the same flow cell lane.

c. Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file.

d. A demultiplexing algorithm sorts the reads into different files according to their indexes.

e. Each set of reads is aligned to the appropriate reference sequence.

Illumina Sequencing

MiSeq Series

Small genome, amplicon and targeted gene panel sequencing.

NextSeq Series

Everyday genome, exome transcriptome sequencing, and more.

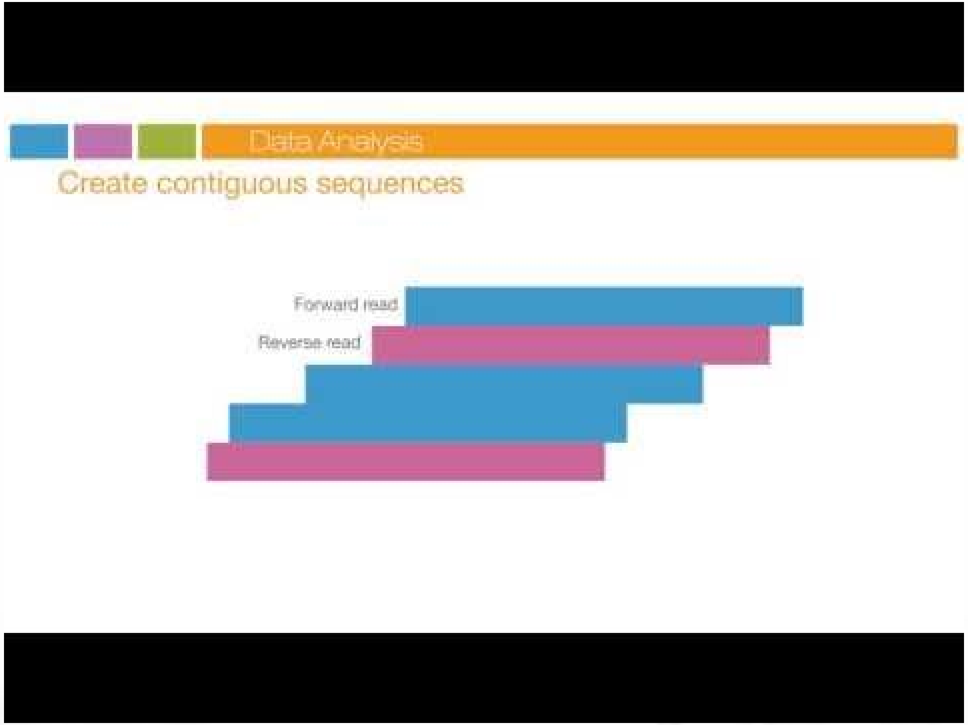
HiSeq Series

Production-scale genome, exome, transcriptome sequencing and more.

HiSeq X Series

Population- and production-scale human whole-genome sequencing.

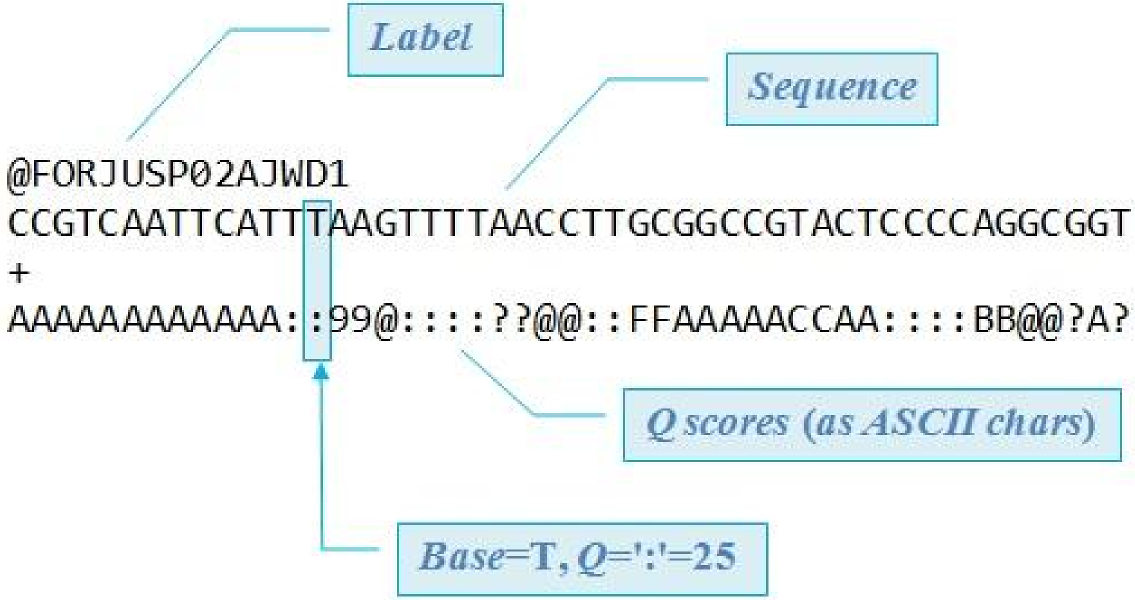
Illumina Sequencing



Data Analysis

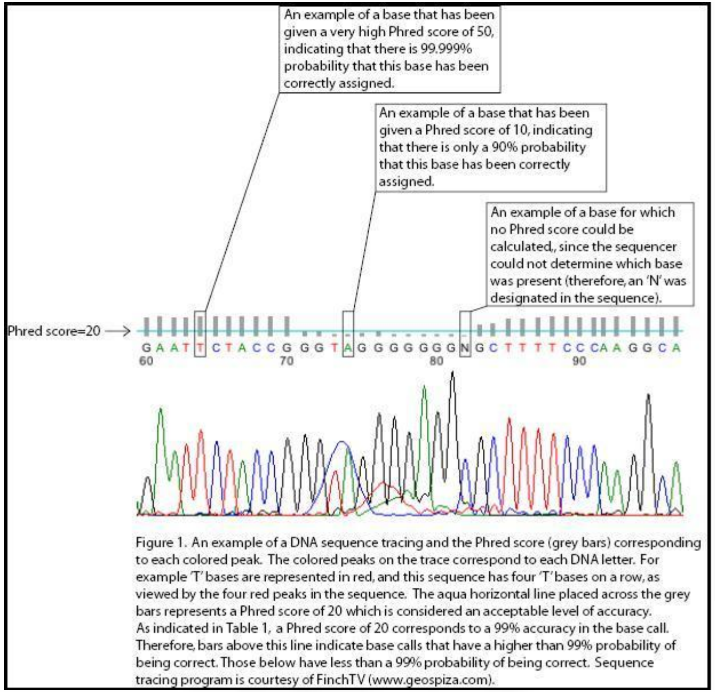
Create contiguous sequences

FASTQ File Format



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

FASTQ File Format

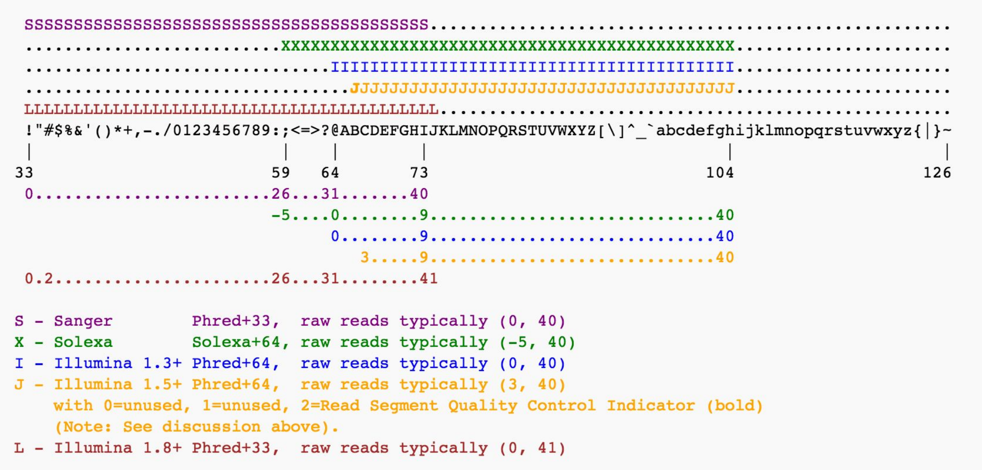


An example of a base that has been given a very high Phred score of 50, indicating that there is 99.999% probability that this base has been correctly assigned.

An example of a base that has been given a Phred score of 10, indicating that there is only a 90% probability that this base has been correctly assigned.

An example of a base for which no Phred score could be calculated,, since the sequencer could not determine which base was present (therefore, an 'N' was designated in the sequence).

Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0.........................26...31.......40
                          -5....0........9..............................40
                                0........9..............................40
                                   3.....9..............................40
 0.2.......................26...31........41

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)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```