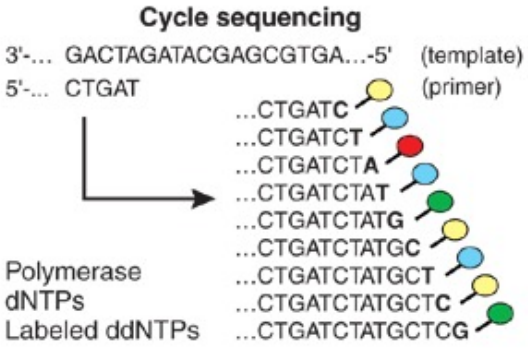
High-Throughput Sequencing

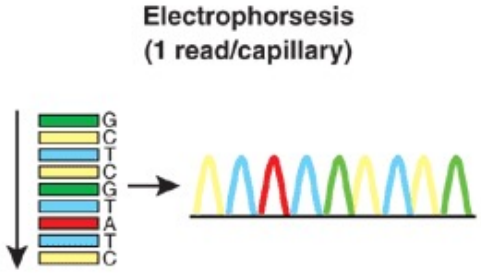
https://dbsloan.github.io/TS2022/

NextGen Sequencing

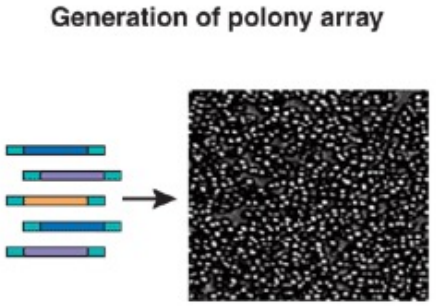
(aka NGS, aka high-throughput sequencing, aka HTS)

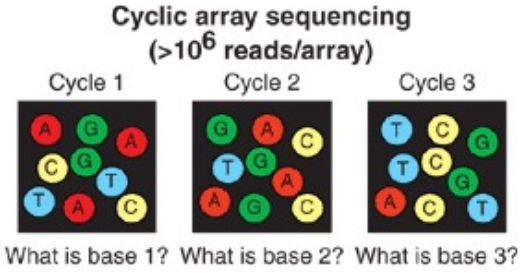
Conventional





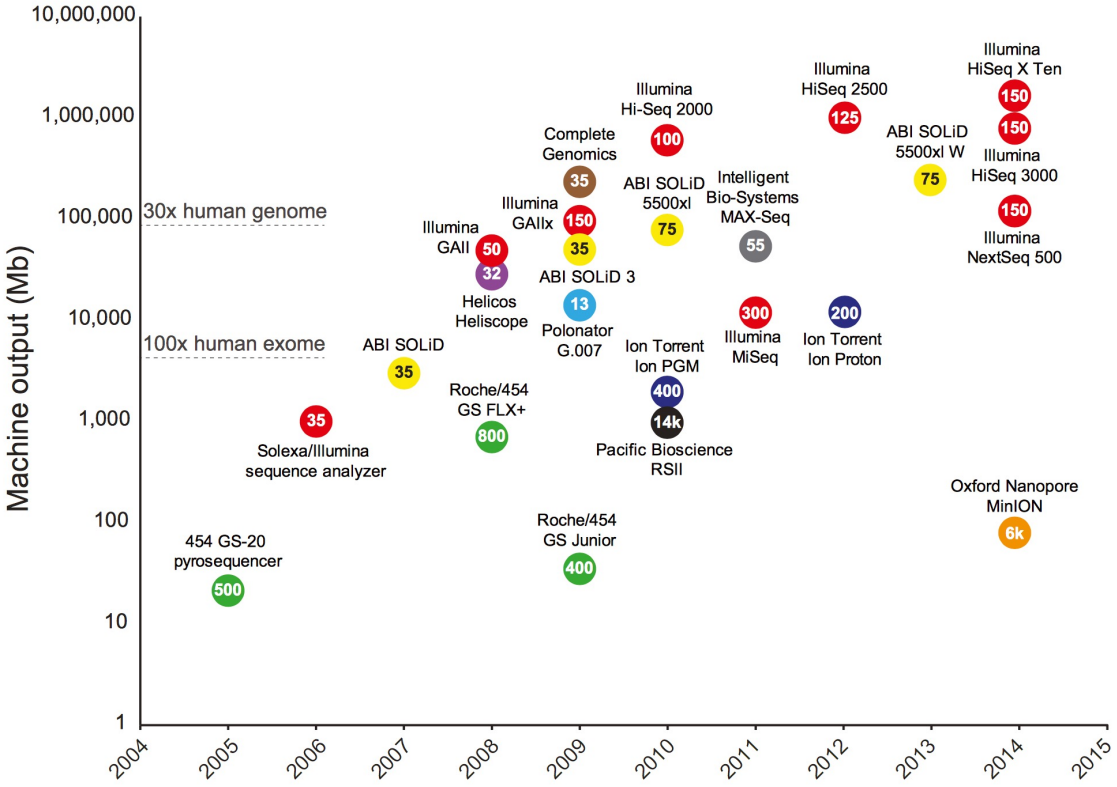
Next-Gen





Shendure and Ji (2008)

Timeline of NextGen Sequencers



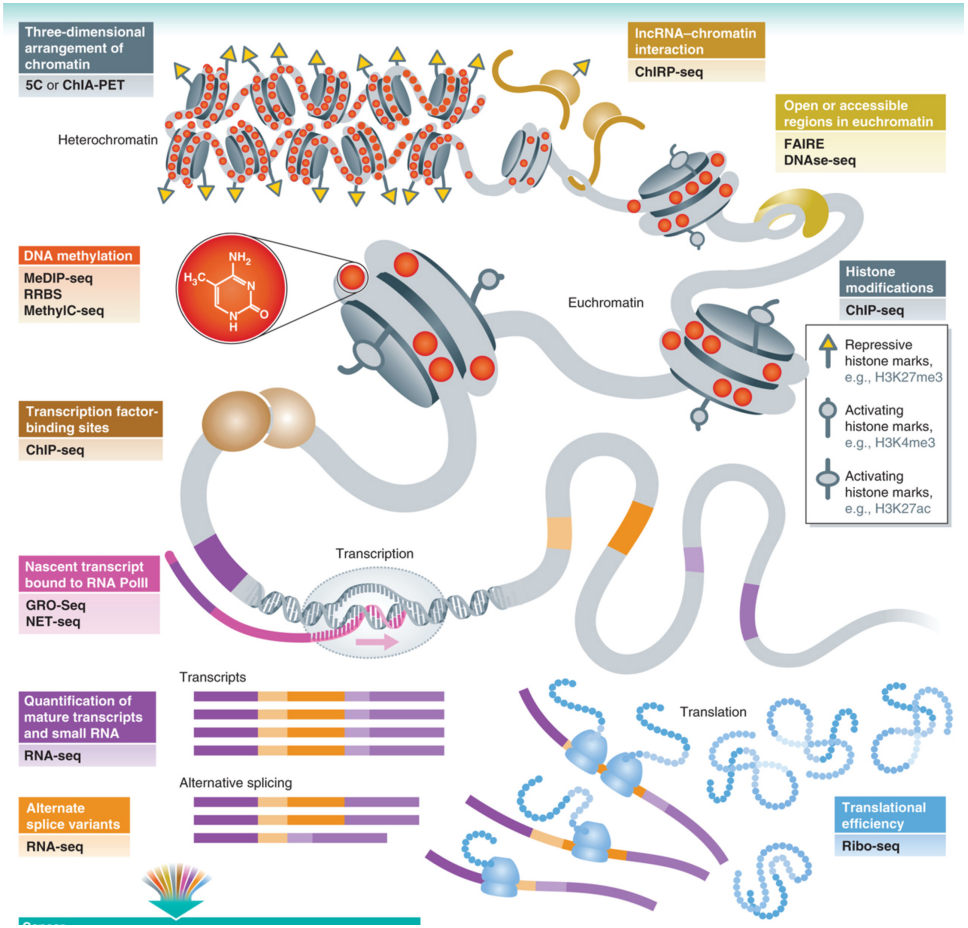
Reuter et al 2015

Illumina Platforms

Read length: 2x150
Reads: 25 million

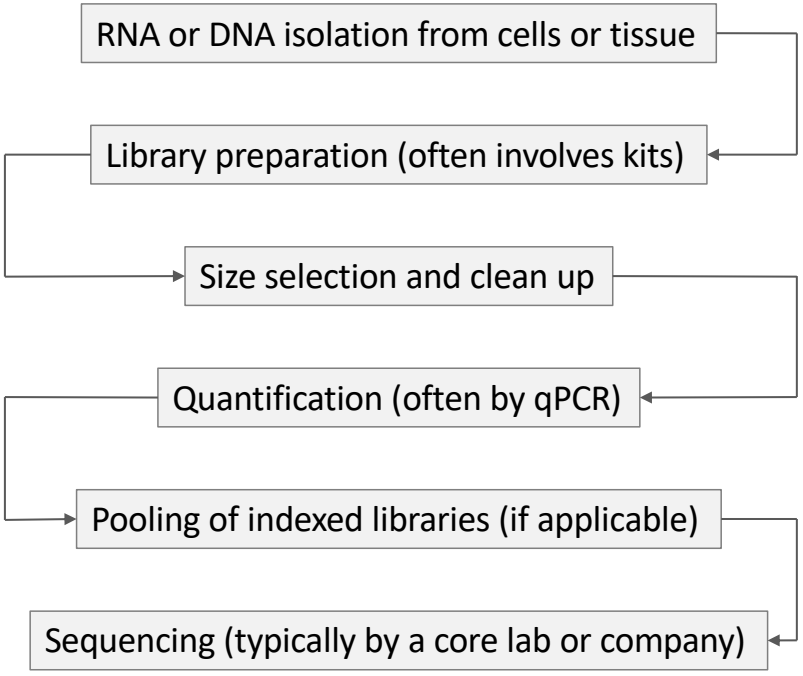
Read length: 2x150
Reads: 5 billion

NextGen Sequencing Applications



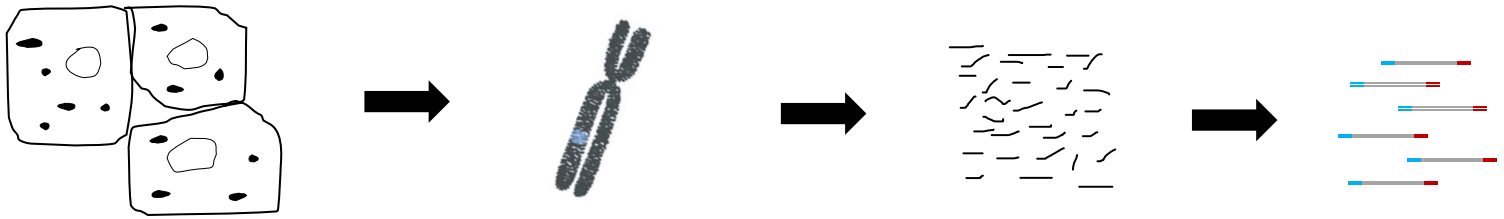
Soon et al (2013)

Sequencing Workflow



NextGen Sequencing Library Prep

The goal is to capture your molecules of interest, such as DNA, mRNA, sRNA, from your cells of interest, and prepare them for sequencing.



Collect cells

Isolate DNA or RNA

Fragment DNA or RNA

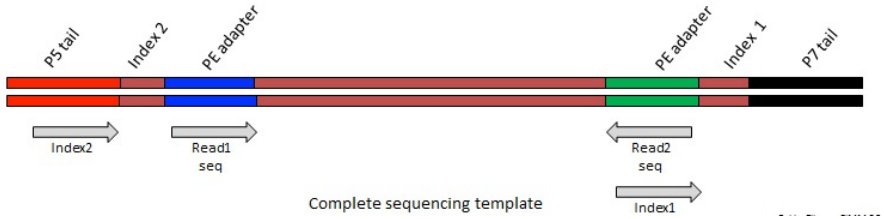
Add adapters to DNA and do PCR

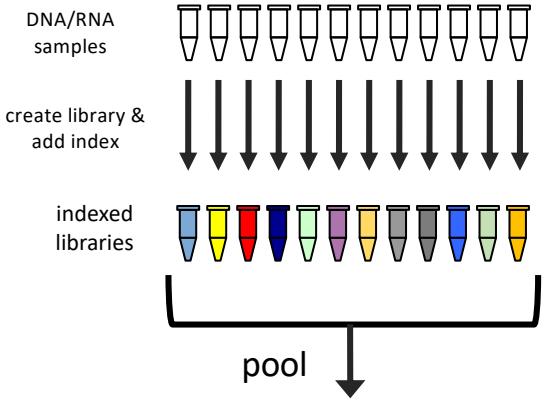
NextGen Sequencing Libraries

Libraries are collections of DNA fragments that contain additional DNA sequences called adapters on both ends that facilitate library preparation and sequencing.

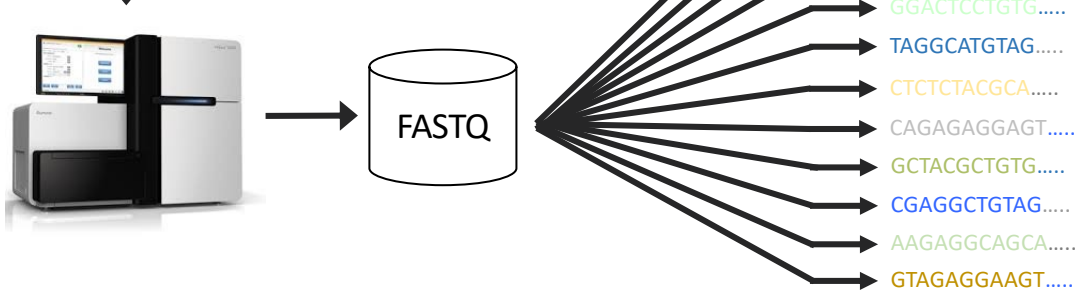
The DNA fragments themselves for Illumina sequencing are usually 100-500 nt long and can derive from DNA or RNA that has been converted to DNA.

Indexing
pooling samples



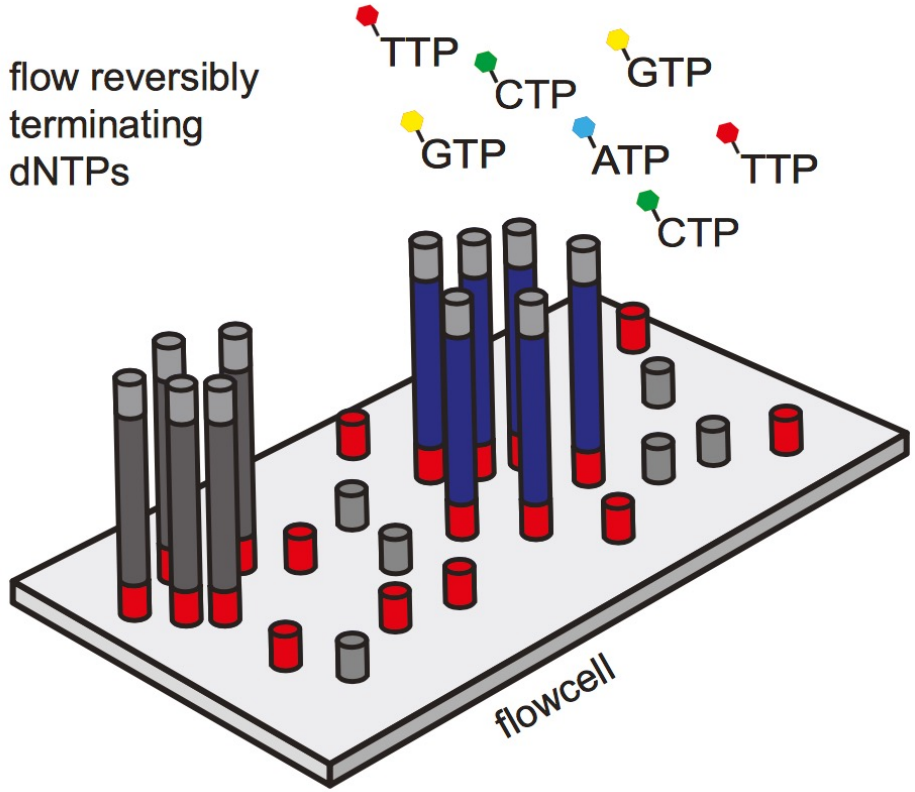


- unique sequence "barcode"
- increase throughput
- maximize capacity



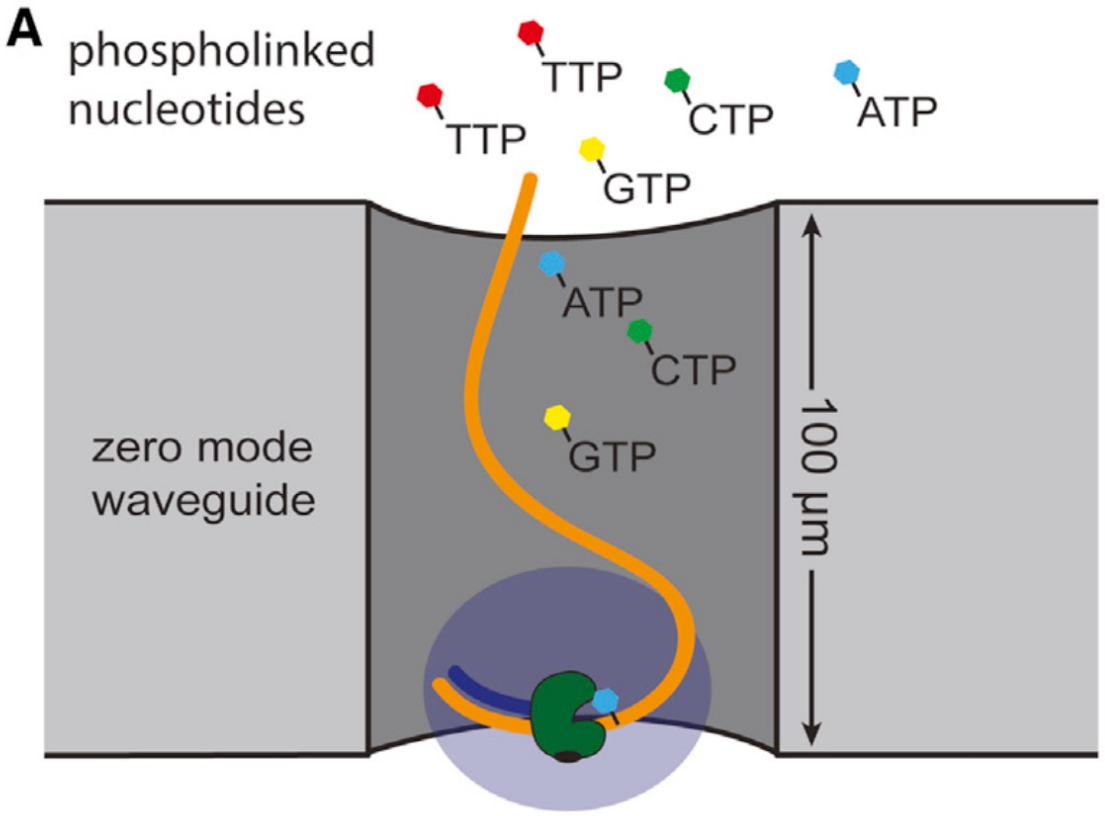
Illumina Sequencing

Short reads, low error rate, very high-throughput



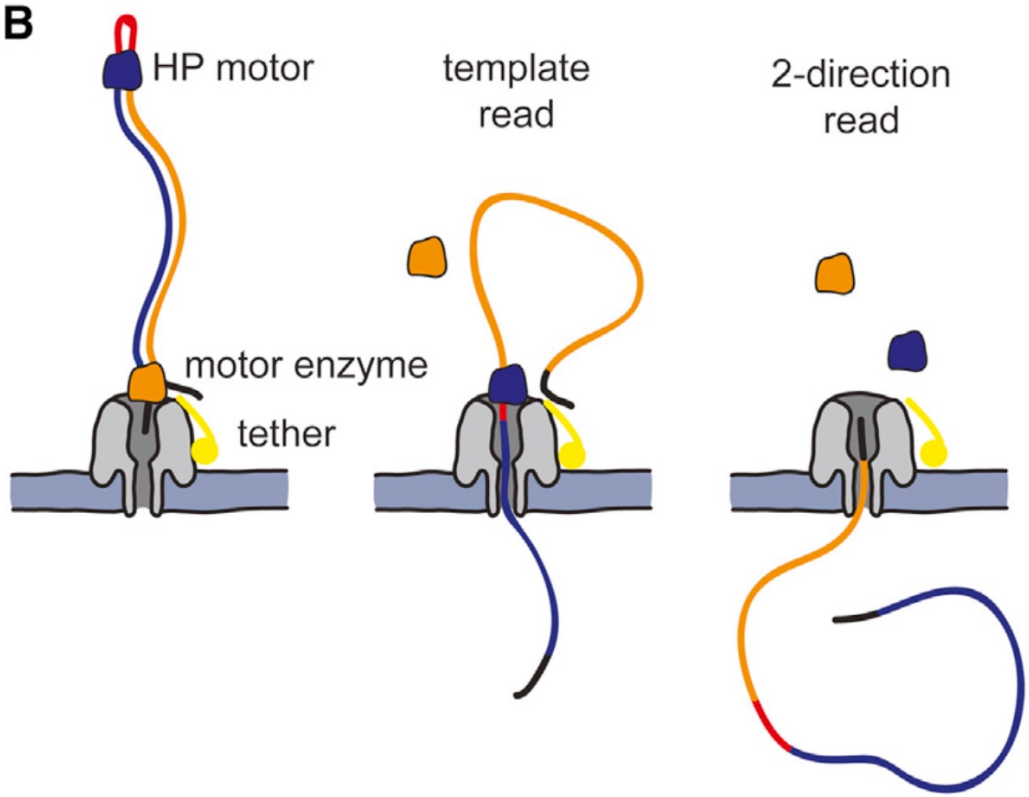
Pacific Biosciences Smart Sequencing

Trade off between read length and error rate, mid-throughput



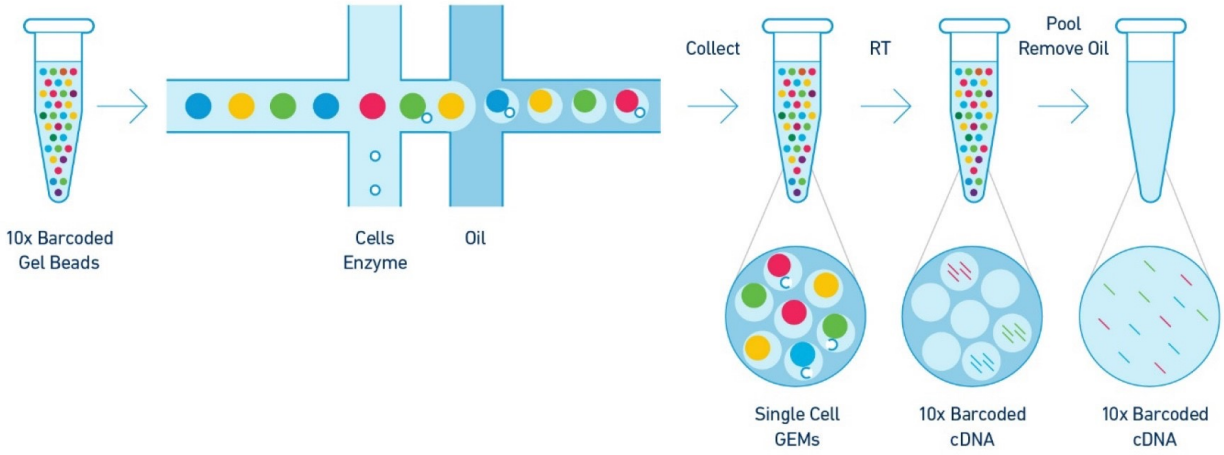
Oxford Nanopore Sequencing

Long reads, high error rate, mid-throughput, can directly sequence RNA.



Single-Cell Sequencing

Sequencing RNA from individual cells using 10xGenomics



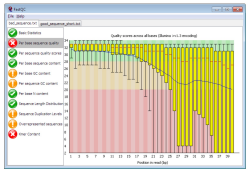
10x GENOMICS®

High-throughput Sequencing Data Analysis

EXAMPLES

mothur

QIIME



Download **fastq** files from server

data.fastq

```
@D64TDFP1:248:C50DMACXX:5:1101:1241:2095 1:N:0:ATCACG
CACCGCCCGTCGCTATCCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCA
+
CCCFFFFFHHHHHJIJGHJJJJIJJJJJGGGFFFFBABDHHHFHFF#@DD>
@D64TDFP1:248:C50DMACXX:5:1101:1371:2154 1:N:0:ATCACG
TCAATATTGCATAGGGTATCTGGAATTCTCGGGTGCCAAGGAACTCCAGT
+
CCCFFFFFHHHHHJJJJGFHIJJJJJJJJJJJJFHHIIJJHGHJFGHJJI
@D64TDFP1:248:C50DMACXX:5:1101:1461:2205 1:N:0:ATCACG
GAAAGACGTCTTCCTAGATTATGGAATTCTCGGGTGCCAAGGAACTCCAGT
+
CCCFFFFFHHHHHJJJJJJJJJJJIJJJJJJJJJHIJJJJJGIIJFGIJJJ
```

Quality assessment using **FastQC** or **fastp**

Quality filtering using (if using fastp, skip initial QA)

fastp

trimmomatic

velvet

Trinity

Combine overlapping sequences for genome/transcript assembly

Map sequences to reference genome

bowtie

bwa

star

kallisto

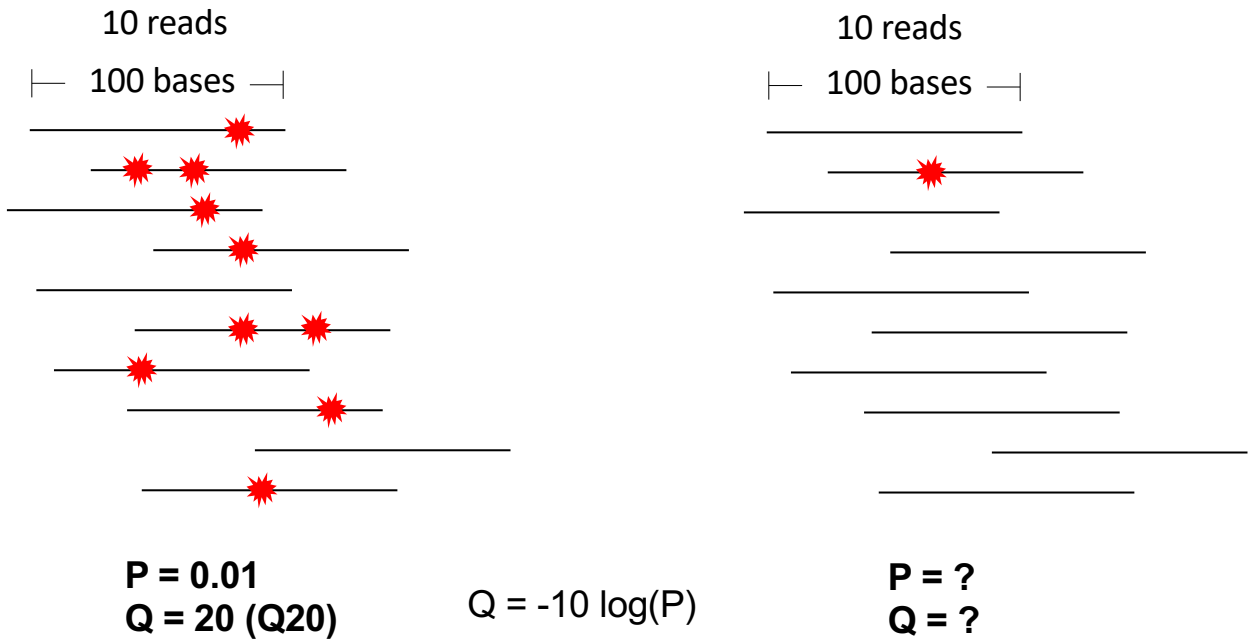
GATK

Python

R

Additional analysis and visualization

Quality Control



- Phred quality score (Q): a measure of the quality of base call during NextGen sequencing.
- P, error probability
- Q30 is a common quality threshold or quality criterion

Quality Control

fastq file snippet

Read

```
1 @D64TDFP1:248:C50DMACXX:5:1101:1241:2095 1:N:0:ATCACG
2 CACCGCCCGTCGCTATCCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCA
3 +
4 !""#FFFFHHHHHJIJGHJJJJIJJJJJGGGFFFFEABDHHHFHFF@@!!>
```

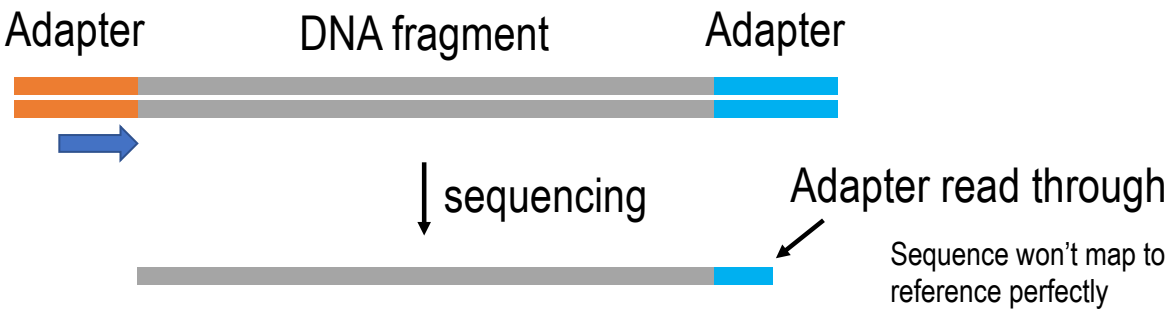
Read

```
1 @D64TDFP1:248:C50DMACXX:5:1101:1371:2154 1:N:0:ATCACG
2 TCAATATTTGCATAGGGTATCTGGAATTCTCGGGTGCCAAGGAACTCCAGT
3 +
4 CCCFFFFFHHHHHJJJJ!!!!JJJJJJJJJJJJJFHHIIJJHGHJFGHJJI
```

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Adapter Trimming

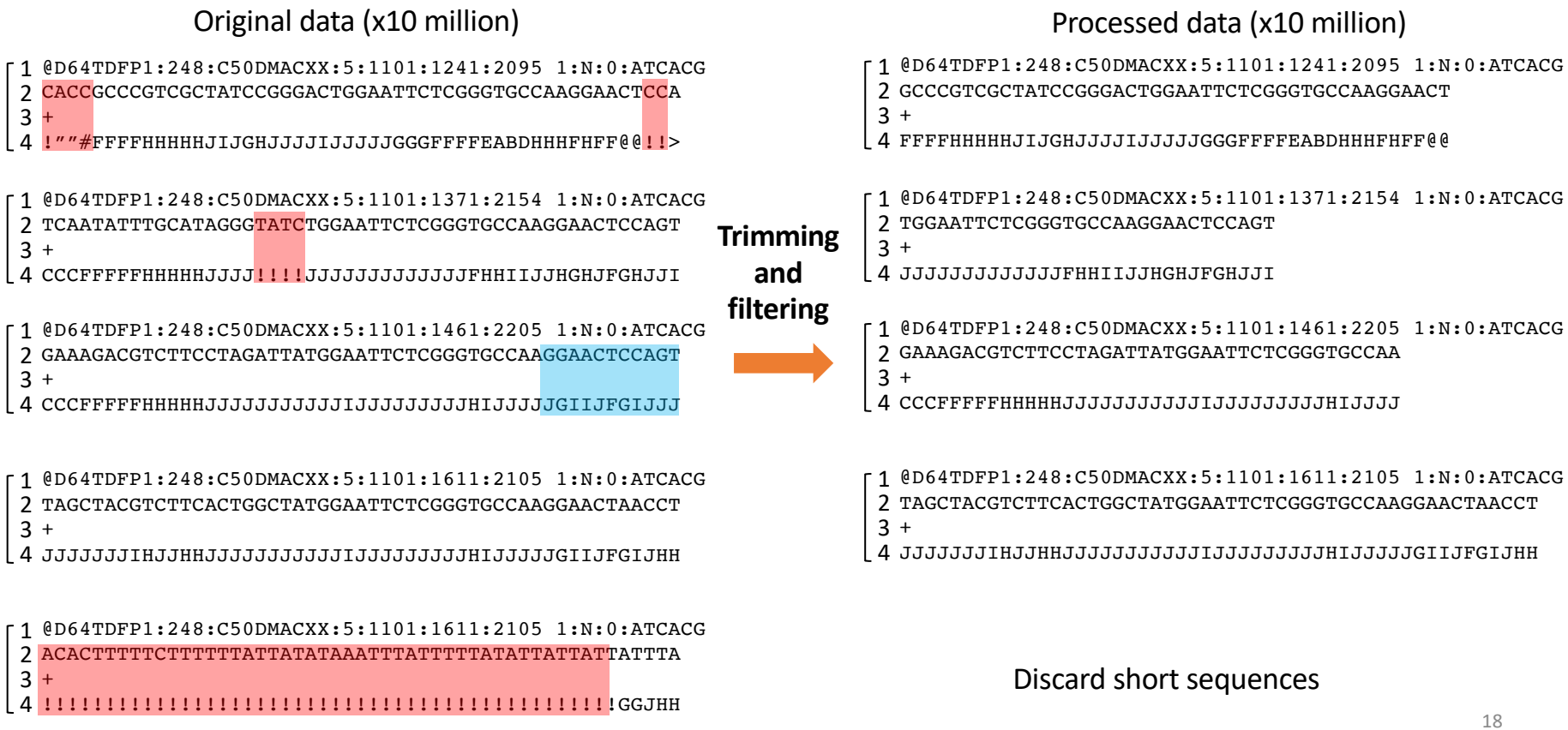


Read

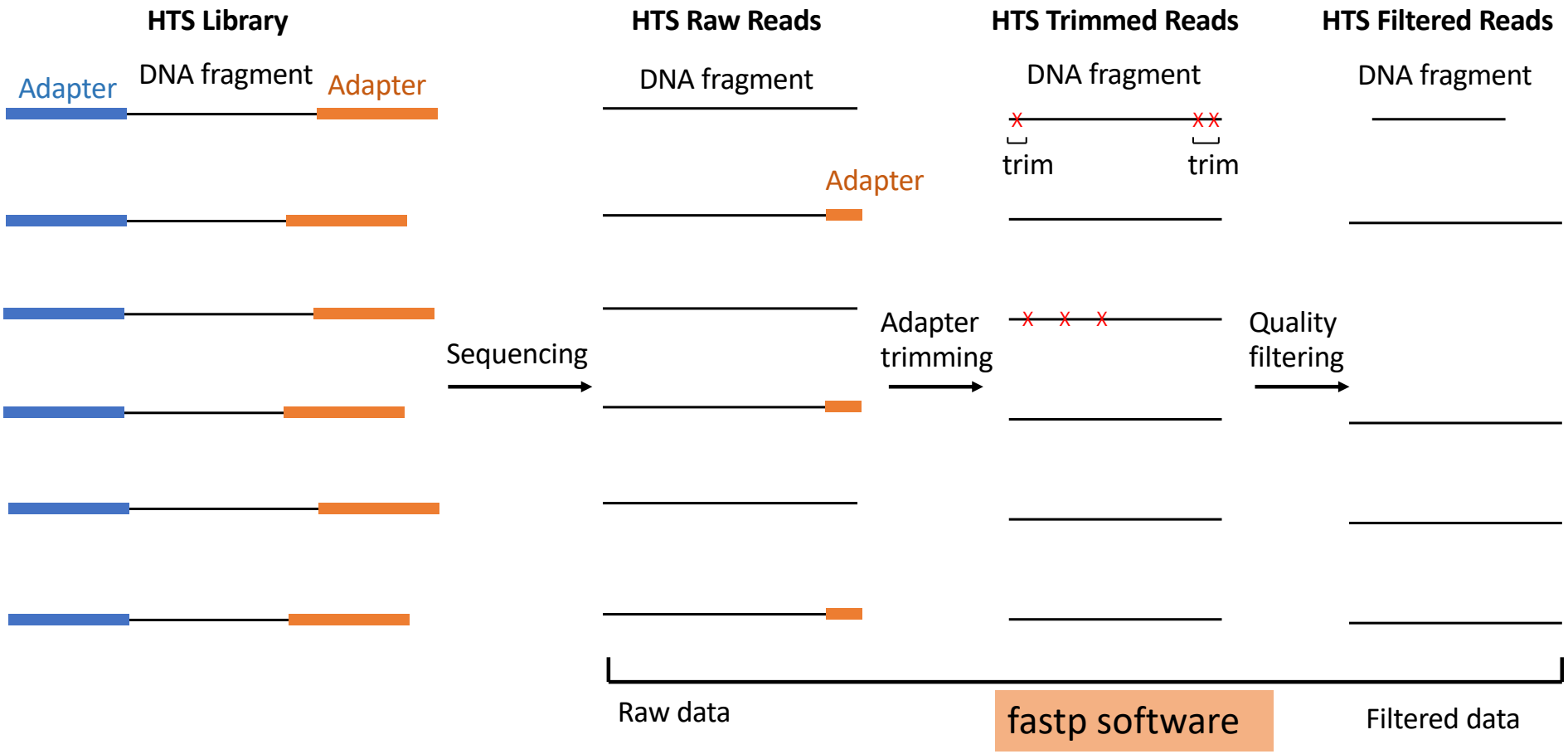
```
1 @D64TDFP1:248:C50DMACXX:5:1101:1461:2205 1:N:0:ATCACG
2 GAAAGACGTCTTCCTAGATTATGGAATTCTCGGGTGCCAAGGAACTCCAGT
3 +
4 CCCFFFFFHHHHHJJJJJJJJJJJIJJJJJJJJJJHIJJJJJGIIJFGIJJJ
```

Need to remove adapter sequence before mapping

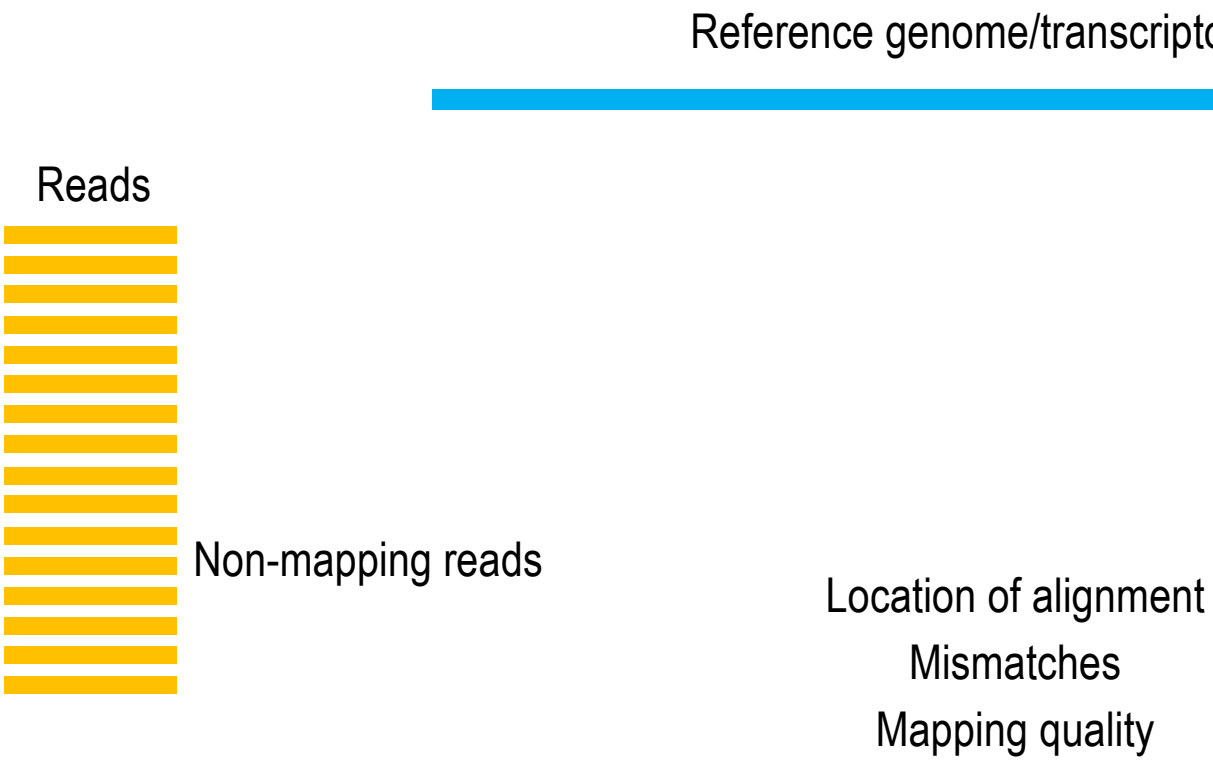
Sequence Preprocessing: trimming and filtering



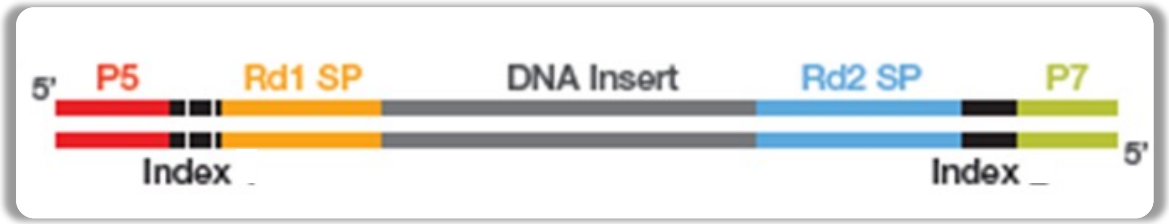
Sequence Preprocessing: trimming and filtering



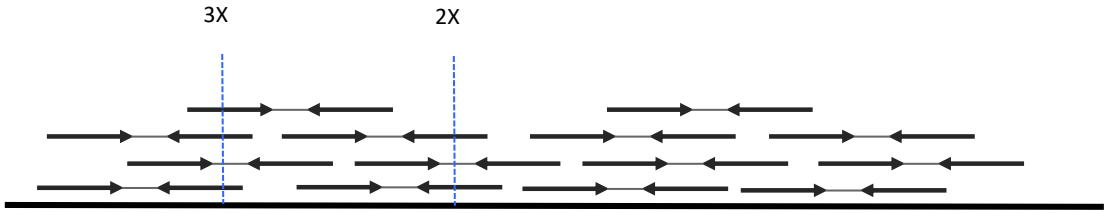
Mapping Reads is an Essential Step in Most HTS Data Analysis Workflows



Coverage



Alignment to reference ome

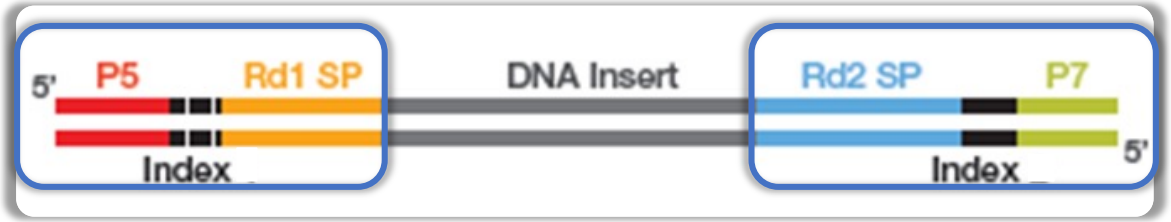


Reference sequence

Illumina Sequencing

Some slides used with
permission from Illumina

Library Prep is Critical for Successful Sequencing

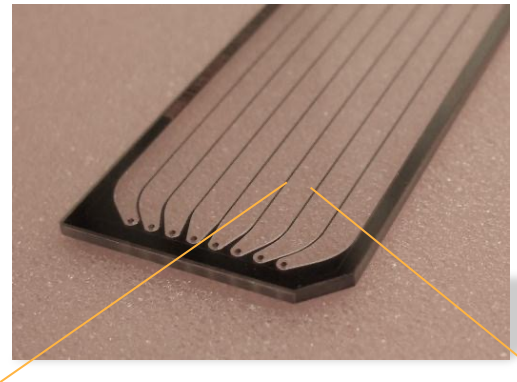


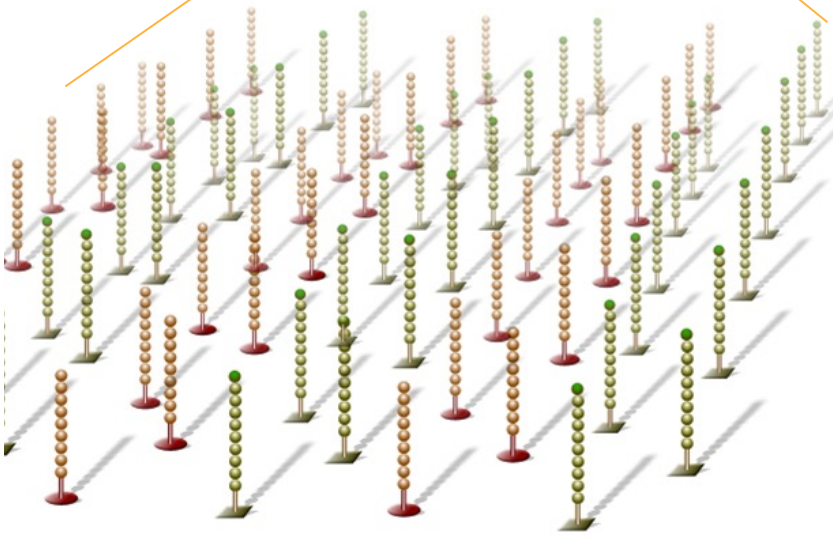
Dual Index Library shown

The aim of the sample prep step is to obtain nucleic acid fragments with adapters attached on both ends

What is a Flow Cell?

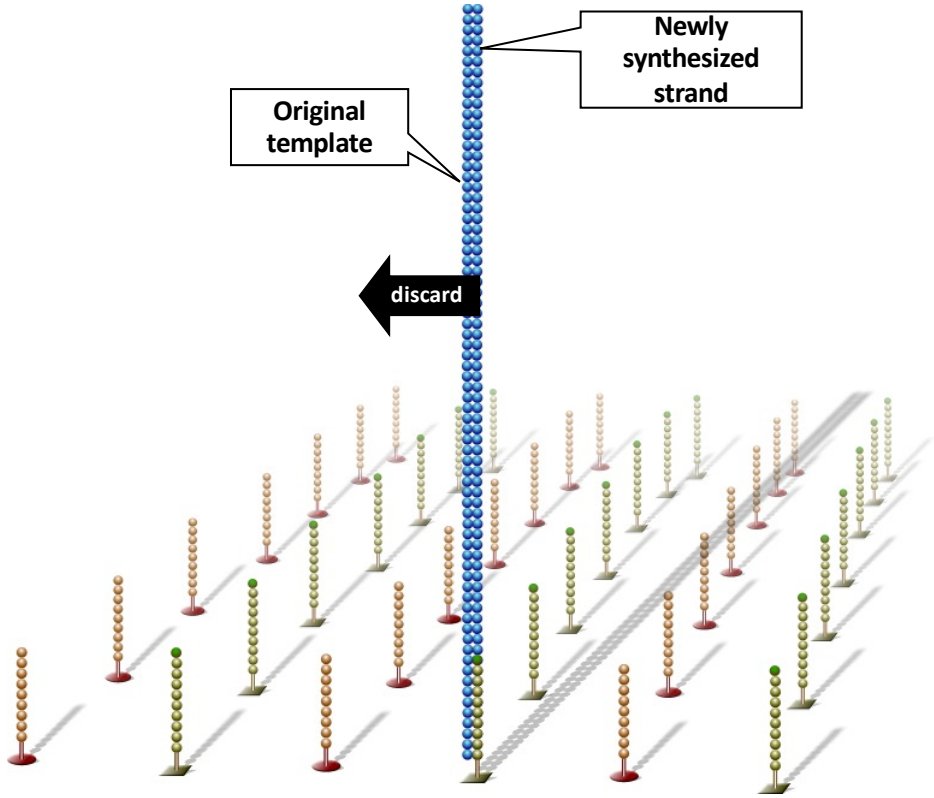
- Cluster generation occurs on a flow cell
- A flow cell is a thick glass slide with channels or lanes
- Each lane is randomly coated with a lawn of oligos that are complementary to library adapters





Denature Double-stranded DNA

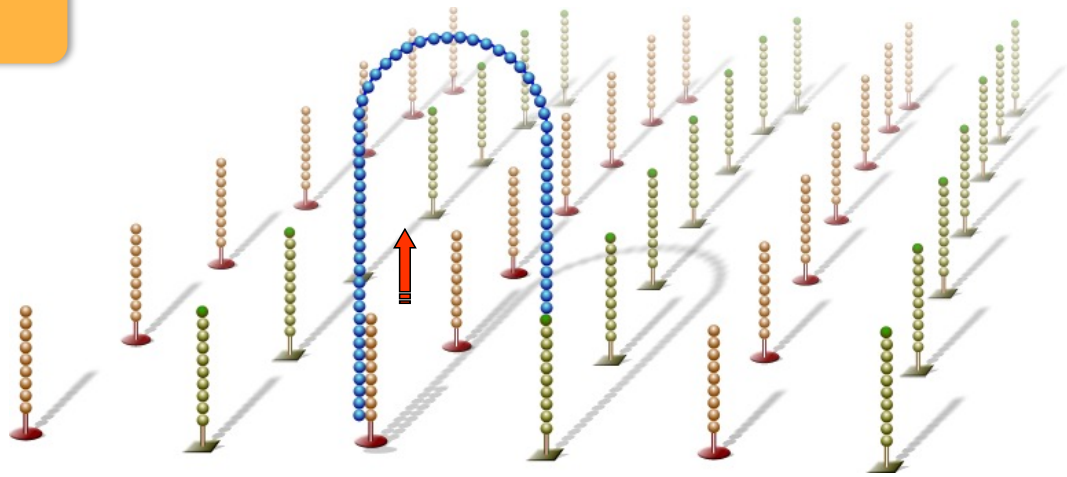
- Double-stranded molecule is denatured
- Original template washed away
- Newly synthesized strand is covalently attached to flow cell surface



Bridge Amplification

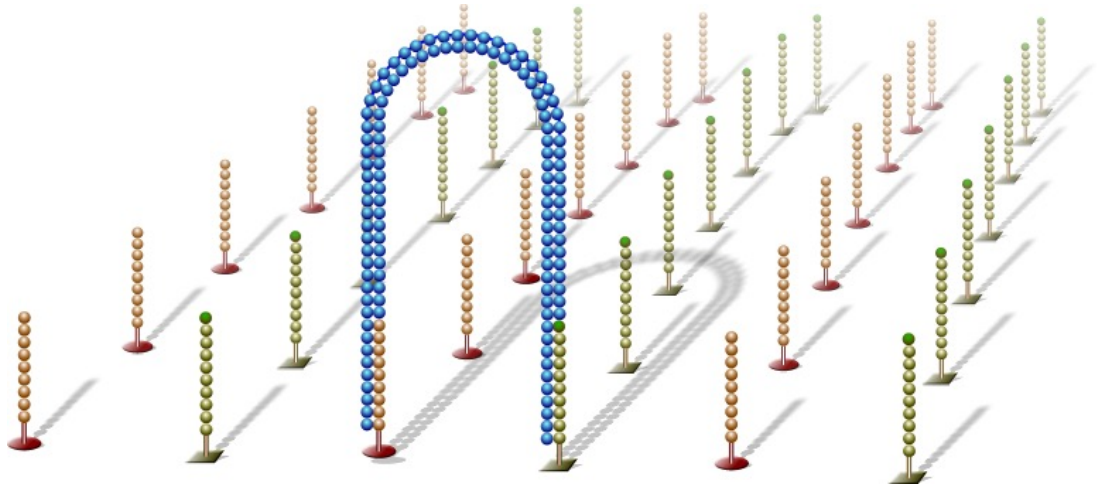
Single-stranded molecule flips over and forms a bridge by hybridizing to adjacent, complementary primer

Hybridized primer is extended by polymerases



Bridge Amplification

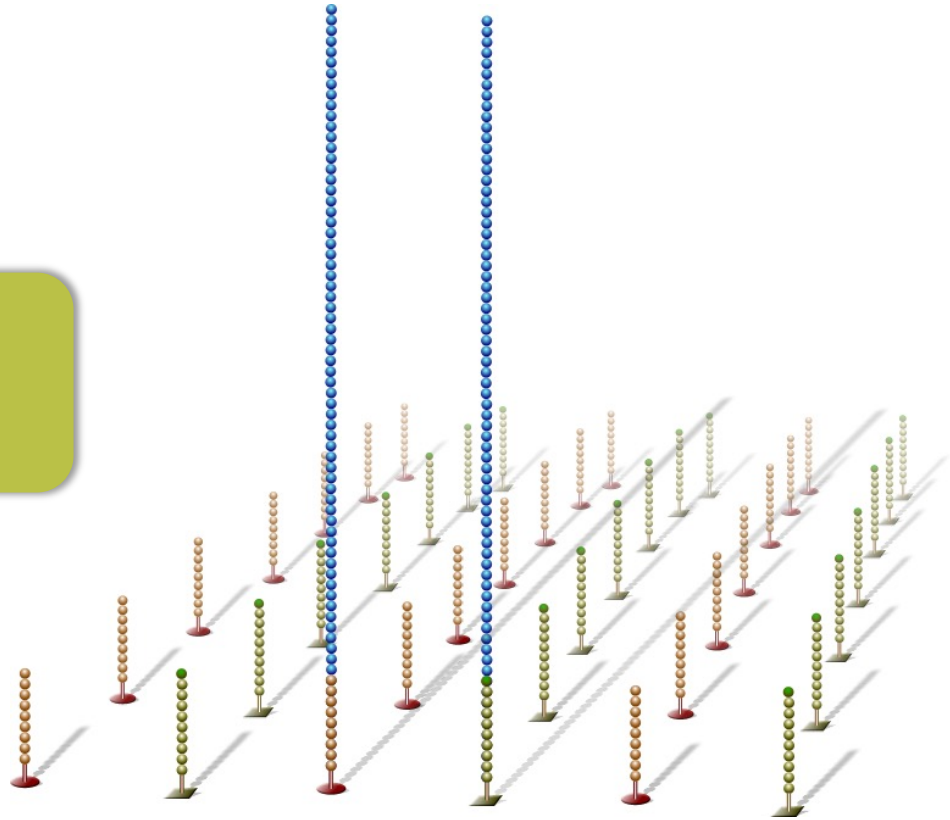
Double-stranded bridge is formed



Denature Double-stranded Bridge

Double-stranded bridge is denatured

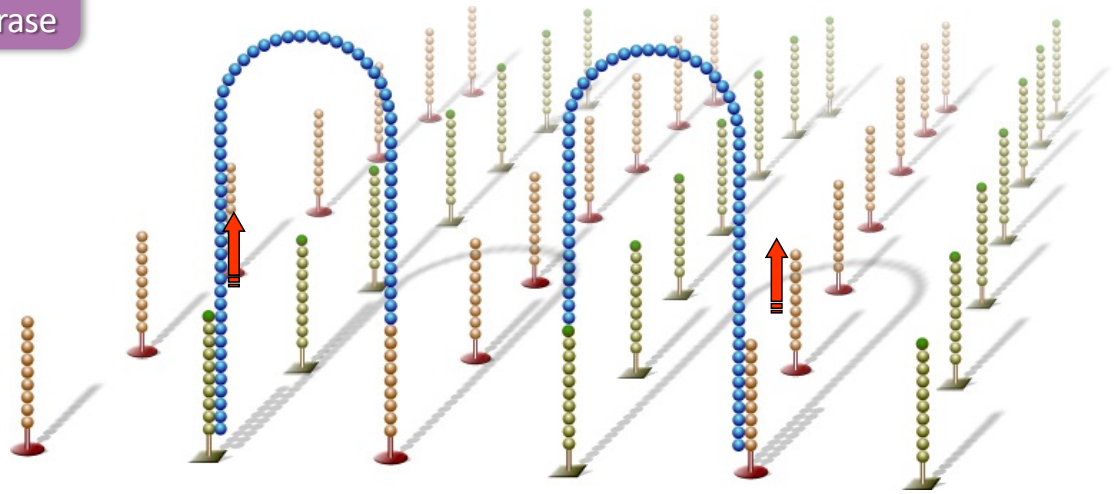
Result:
Two copies of covalently bound single-stranded templates



Bridge Amplification

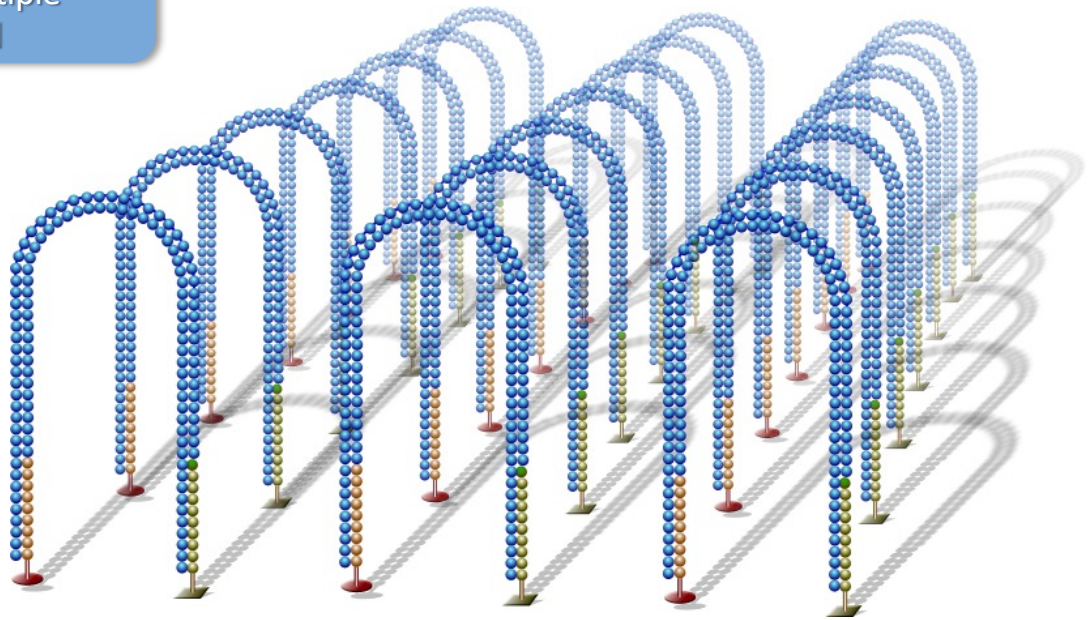
Single-stranded molecules flip over to hybridize to adjacent primers

Hybridized primer is extended by polymerase



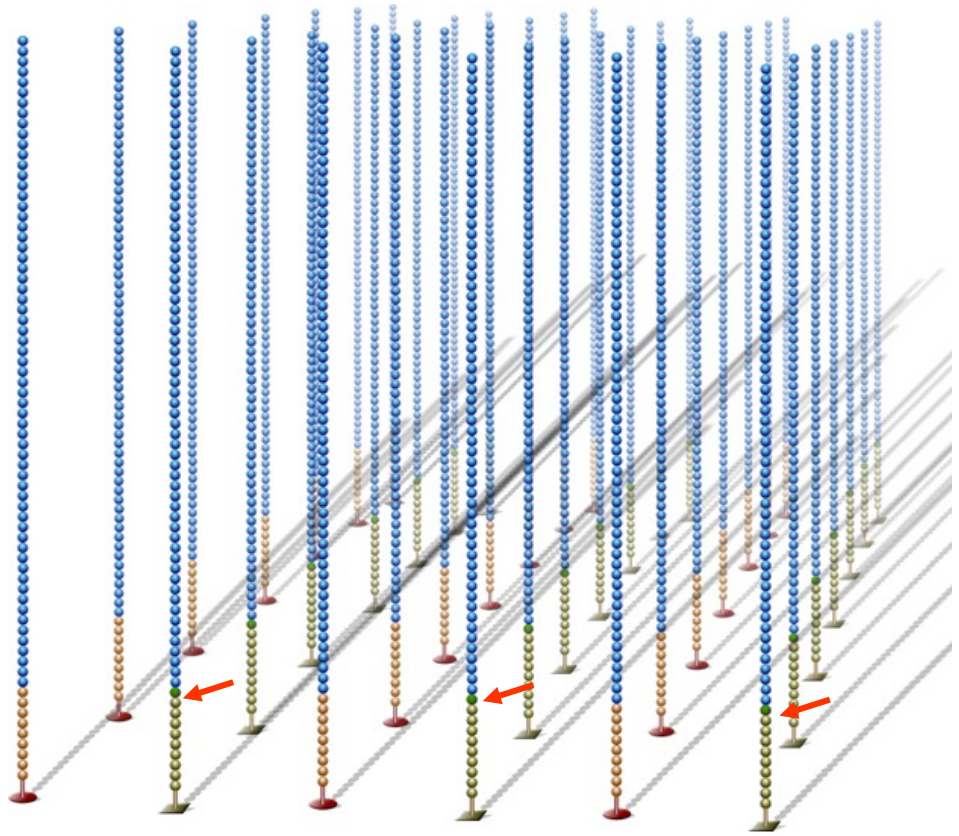
Bridge Amplification

Bridge amplification cycle repeated until multiple bridges are formed



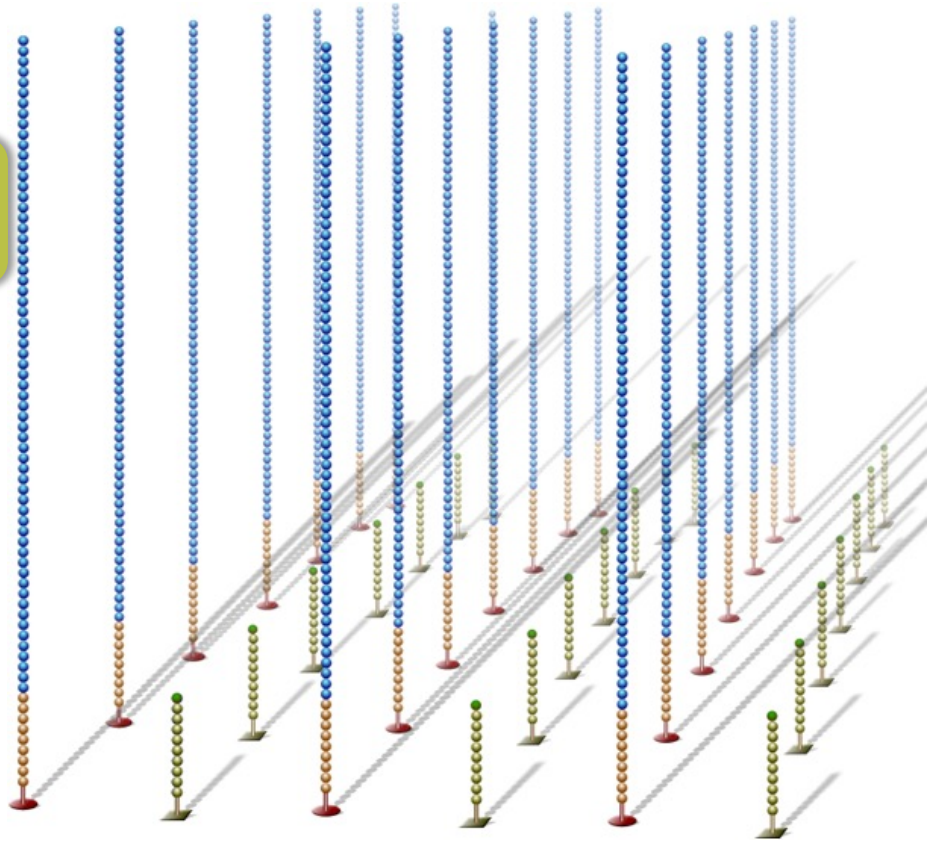
Linearization

dsDNA bridges are denatured



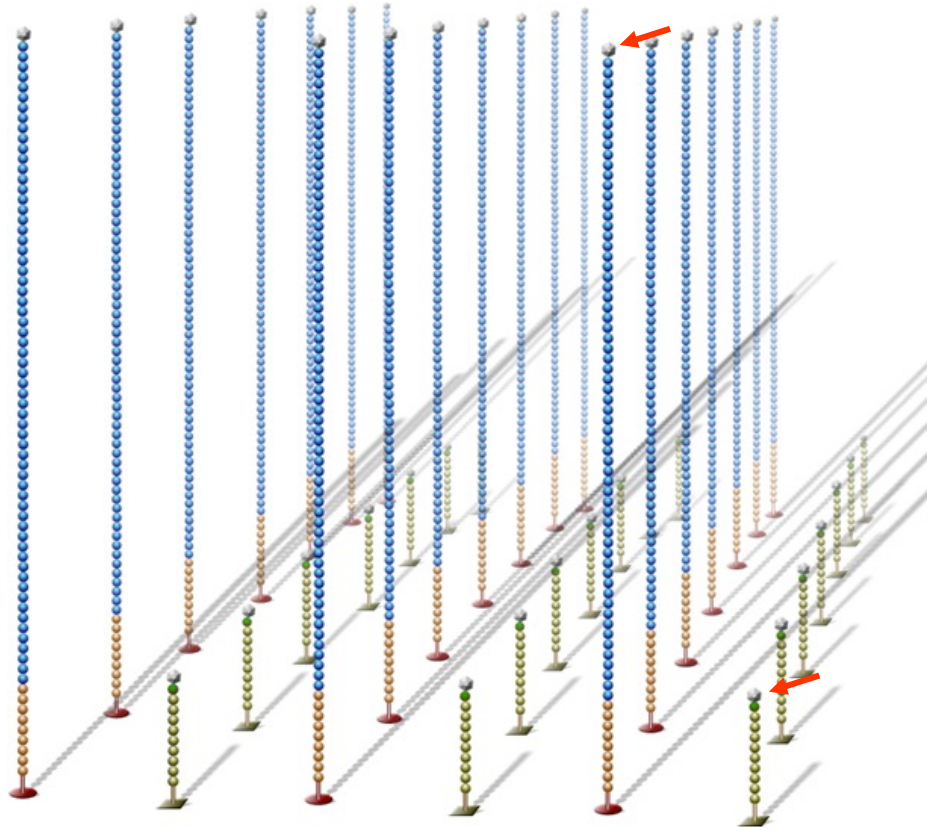
Reverse Strand Cleavage

Reverse strands cleaved and washed away, leaving a cluster with forward strands only



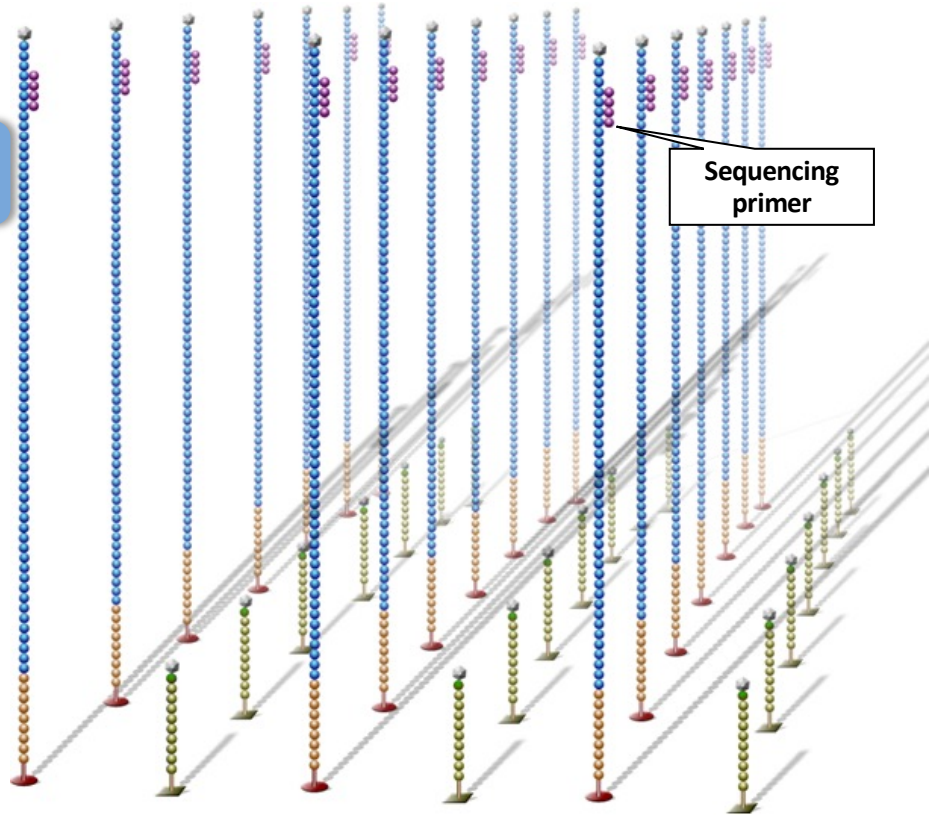
Blocking

Free 3’ ends are blocked to prevent unwanted DNA priming

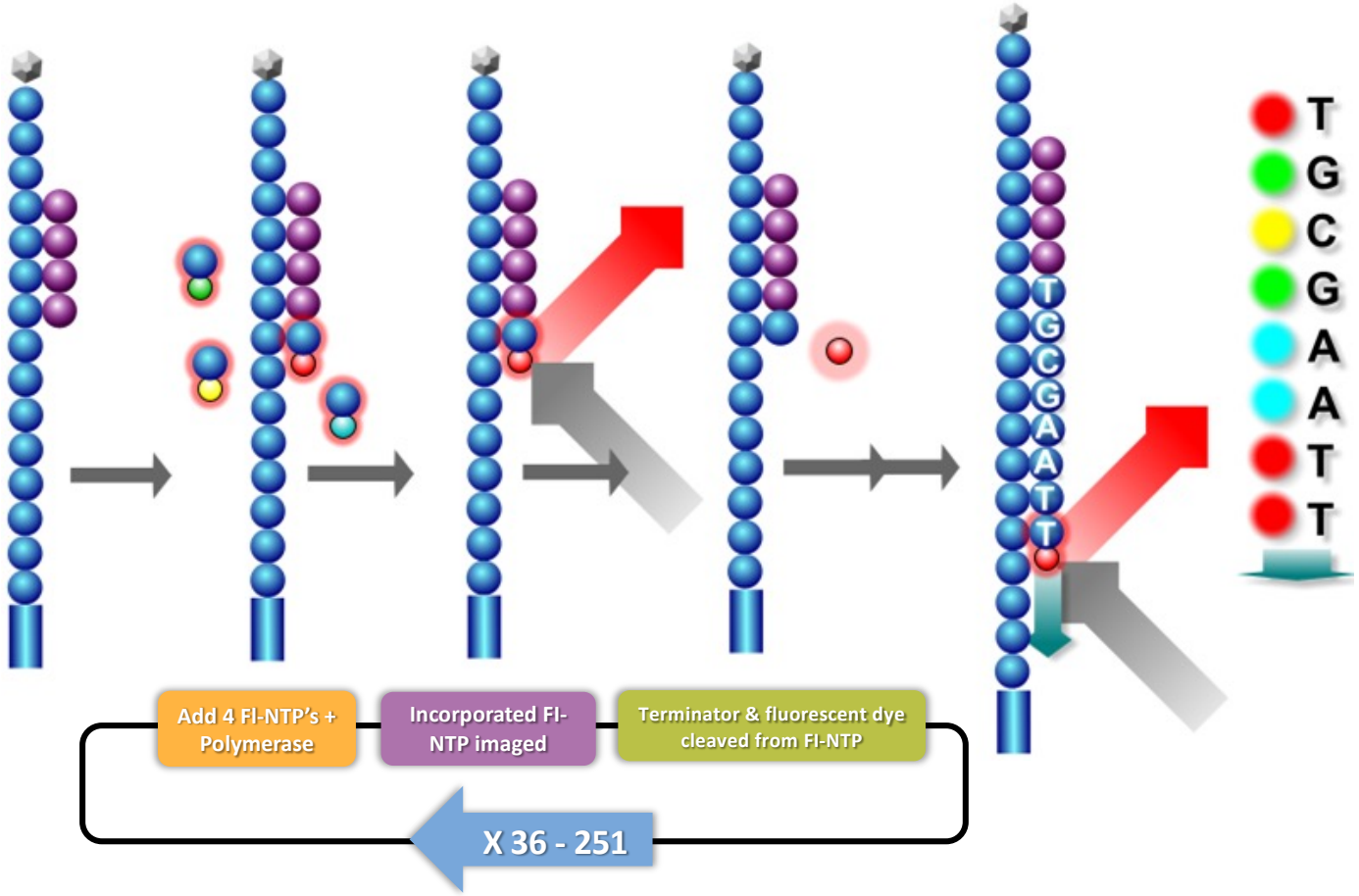


Read 1 Primer Hybridization

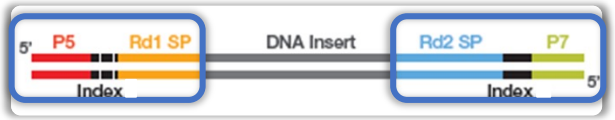
Sequencing primer is hybridized to adapter sequence



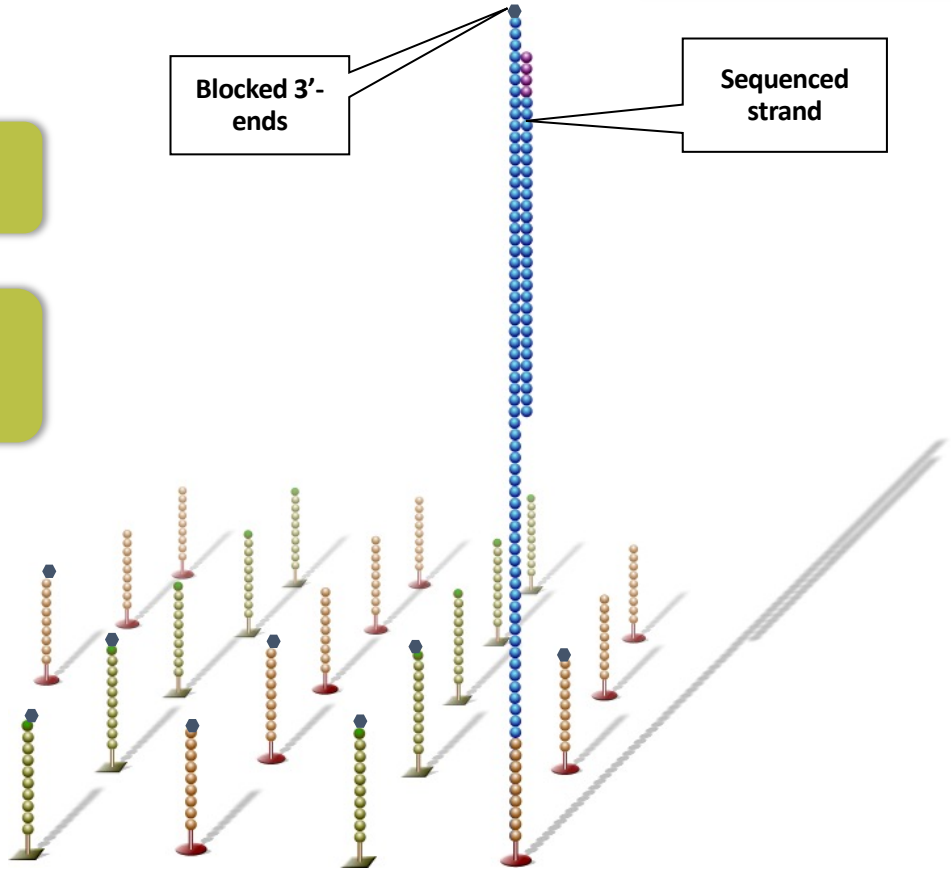
Sequencing By Synthesis



Paired End Sequencing



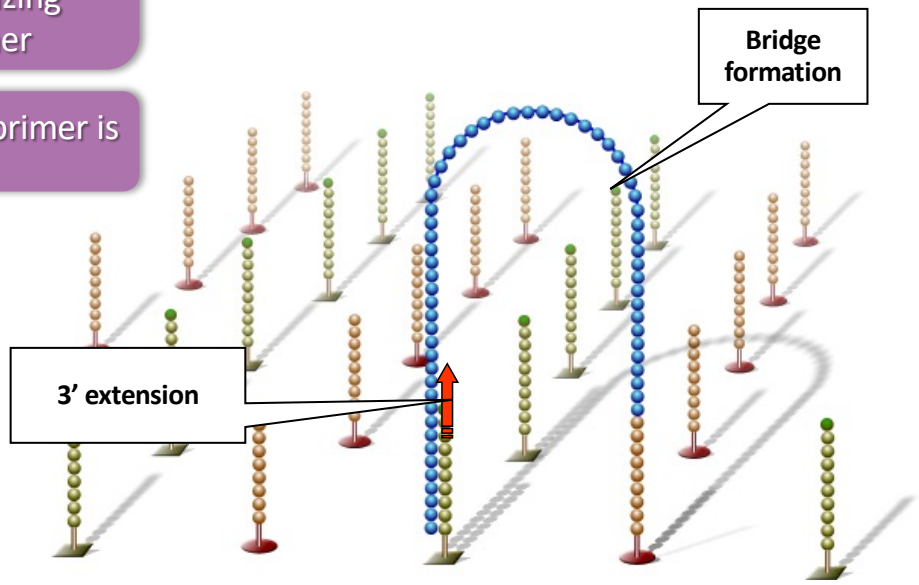
- Sequenced strand is stripped off
- 3’-ends of template strands and lawn primers are unblocked



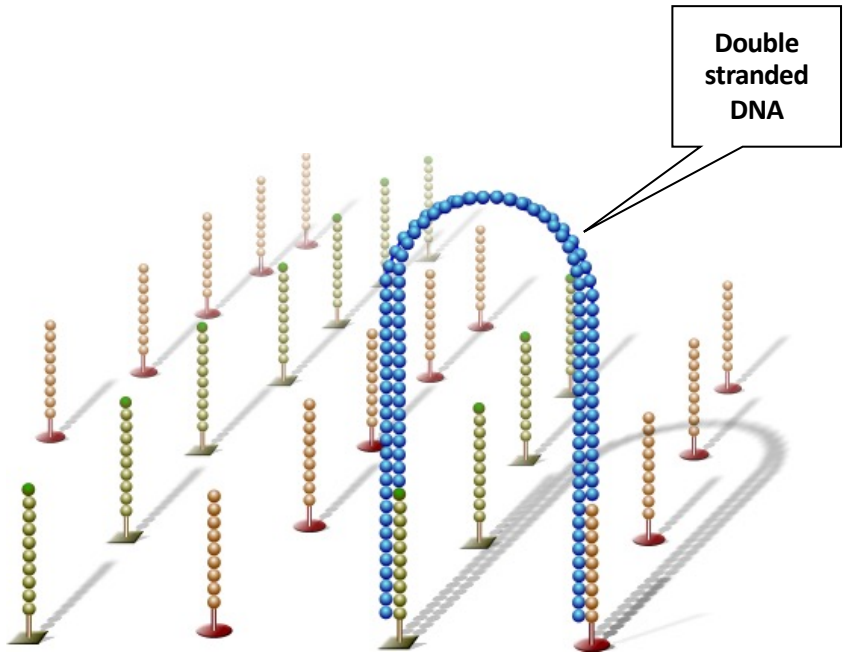
Paired End Sequencing

Single-stranded template loops over to form a bridge by hybridizing with a lawn primer

3’-ends of lawn primer is extended

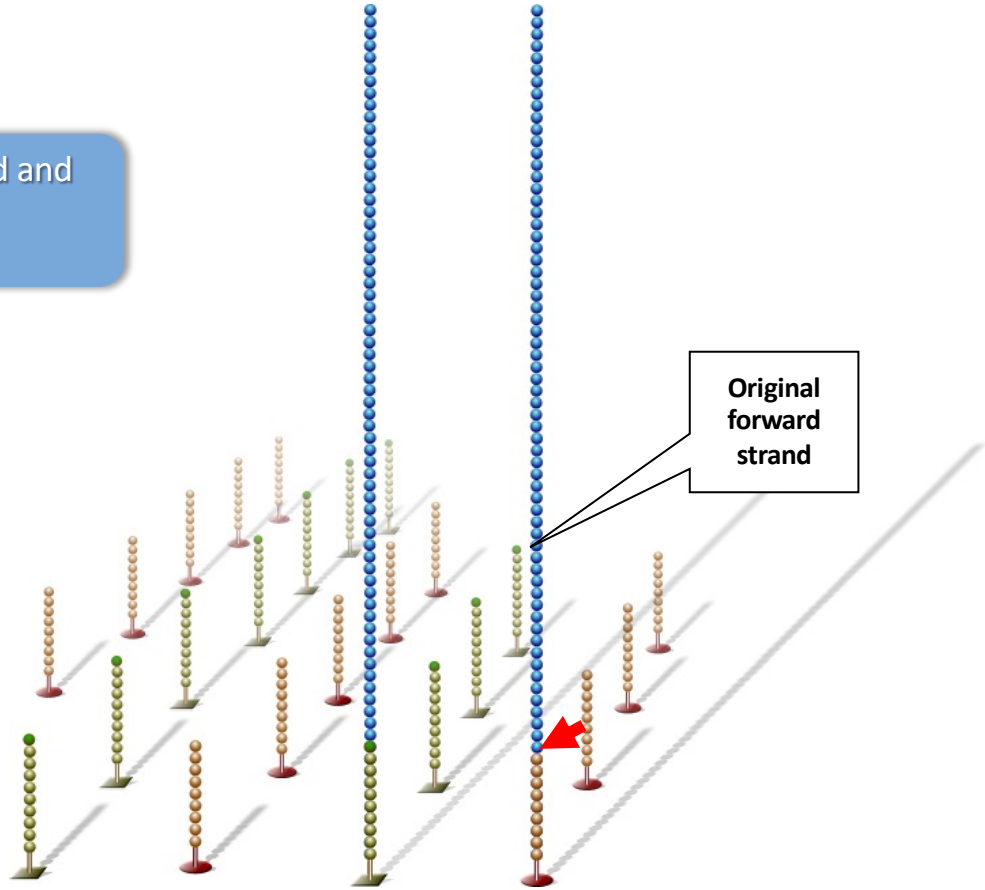


Paired End Sequencing



Paired End Sequencing

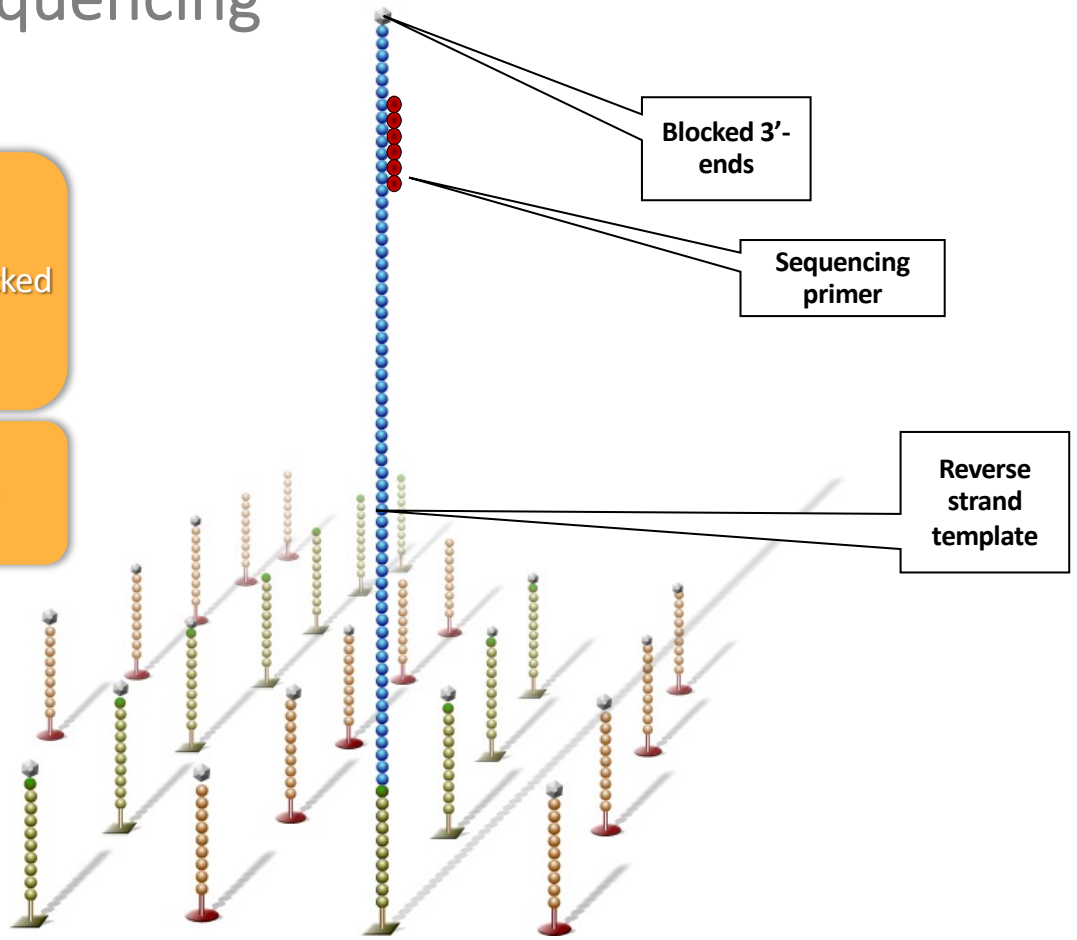
Bridges are linearized and the original forward template is cleaved



Paired End Sequencing

Free 3’ ends of the reverse template and lawn primers are blocked to prevent unwanted DNA priming

Sequencing primer is hybridized to adapter sequence



Sequencing By Synthesis 2nd Read