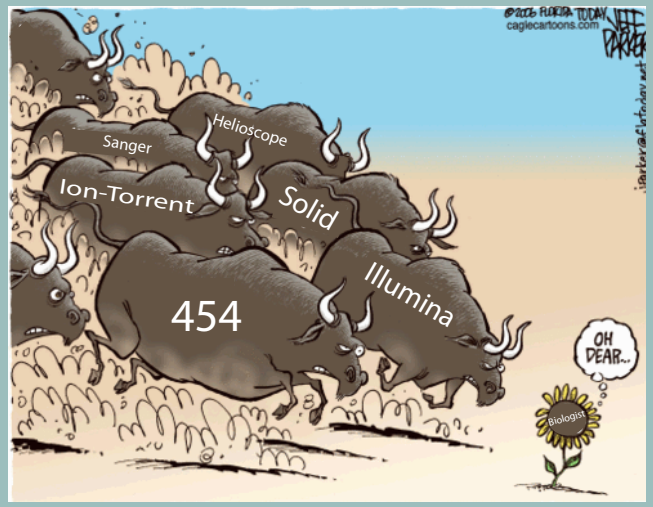
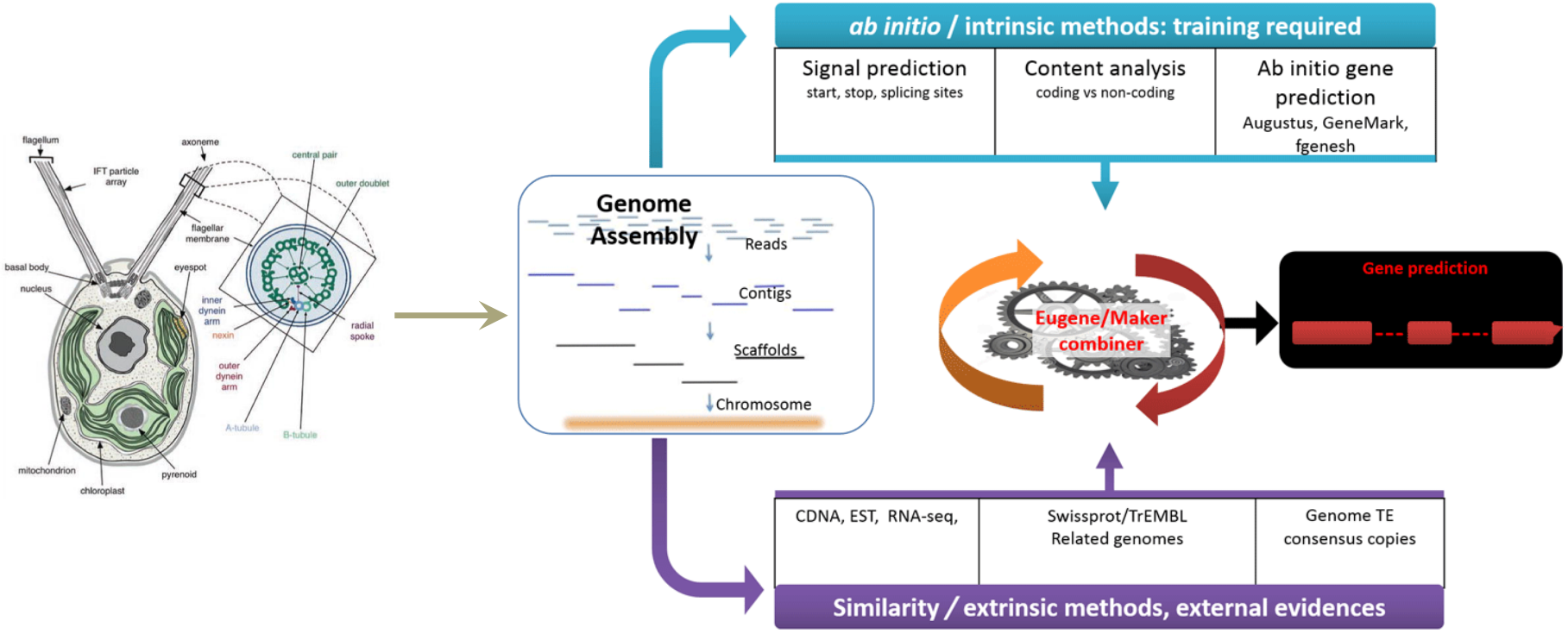


ALGORITHMS IN SEQUENCE ANALYSIS

High Throughput DNA Sequencing

GENOME SEQUENCING - THE FUNDAMENT OF GENOMICS



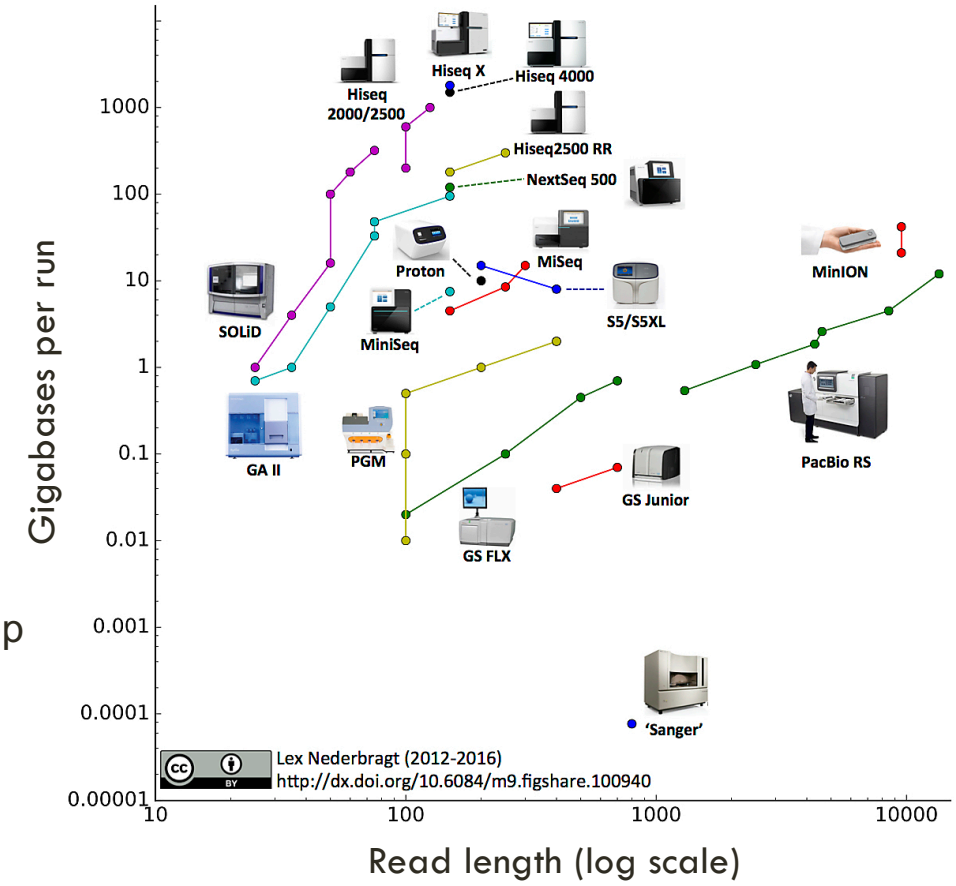
Del Angel et al. (2018) F1000 Research 7(ELIXIR):148

HOW DO WE SEQUENCE DNA?

1st generation (1977)
- **Sanger** method: Sequencing by synthesis
- **Maxam-Gilbert** method: chemical sequencing

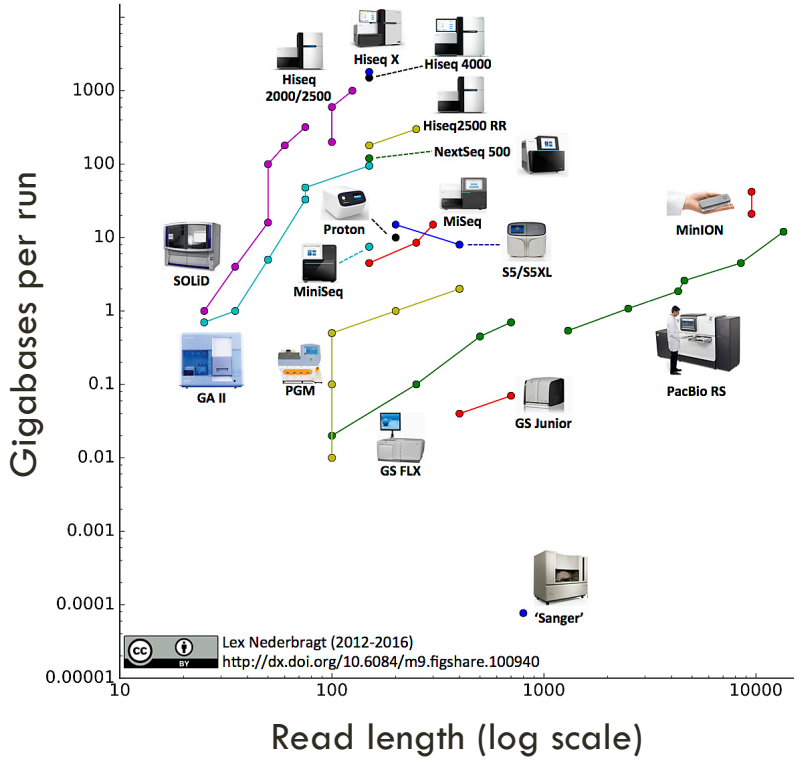
2nd generation ("next generation"; 2005)
- **454** - pyrosequencing
- **SOLiD** – sequencing by ligation
- **Illumina** – sequencing by synthesis
- **Ion Torrent** – ion semiconductor
- **Pac Bio** – Single Molecule Real-Time sequencing, 1000 bp

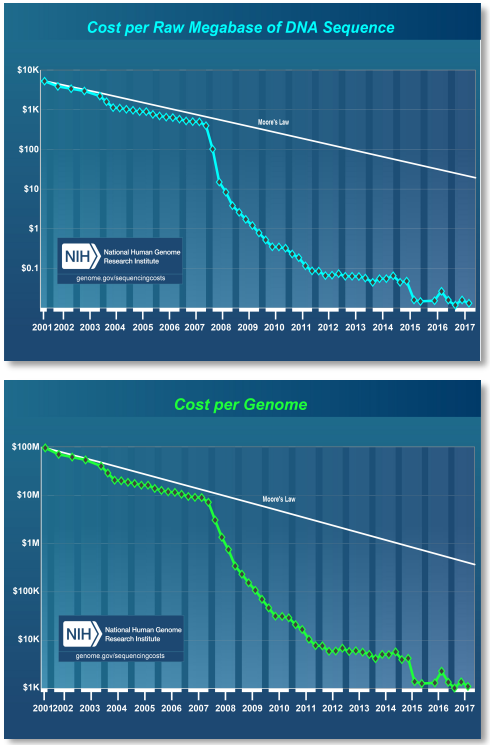
3rd generation (2015)
- **Pac Bio** – SMRT, Sequel system, 20,000 bp
- **Nanopore** – ion current detection
- **10X Genomics** – novel library prep for Illumina



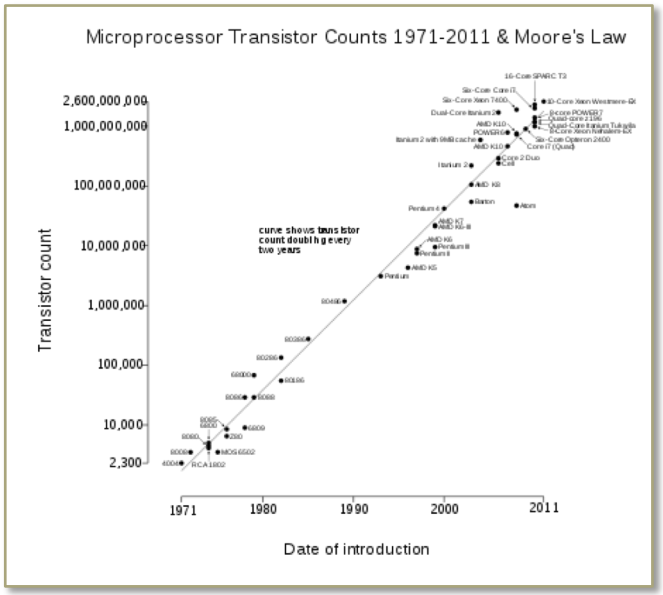
Del Angel et al. (2018) F1000 Research 7(ELIXIR):148

SEQUENCE DATA GROWS FASTER THAN COMPUTER POWER



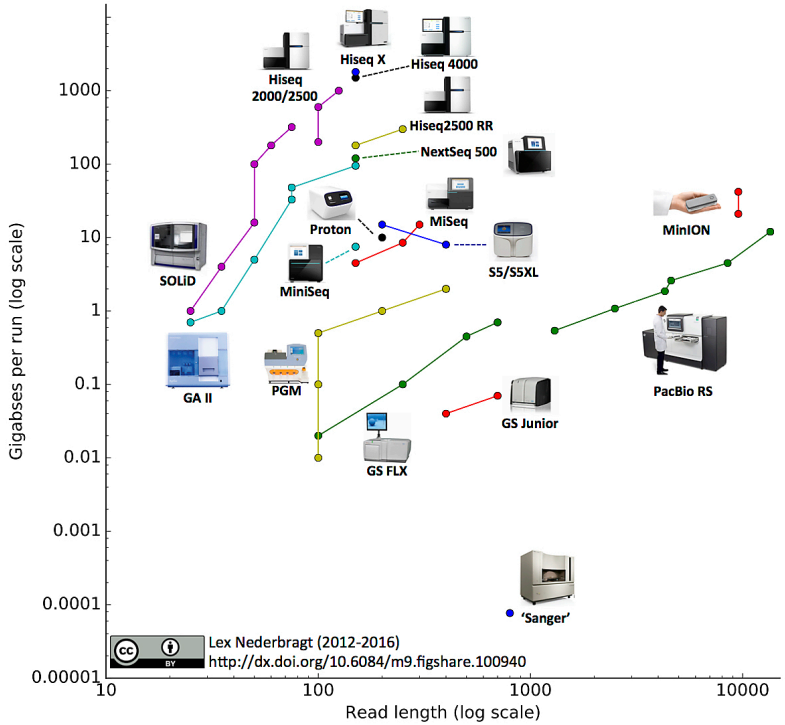


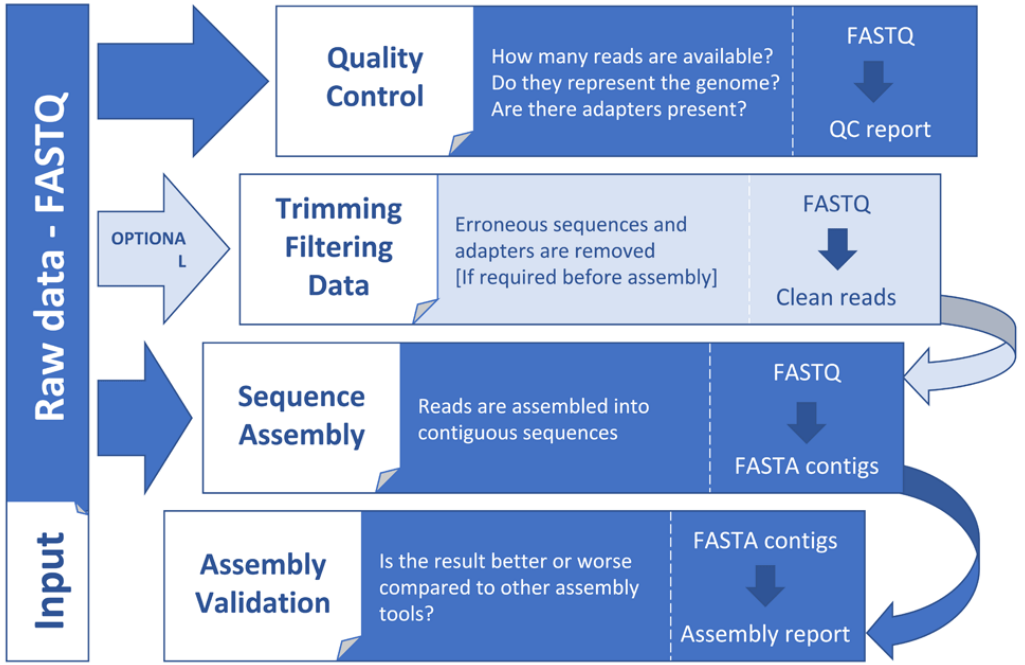
Source: https://www.genome.gov/sequencingcostsdata/



Source: wikipedia

DNA SEQUENCING TECHNOLOGIES





Del Angel et al. (2018) F1000 Research 7(ELIXIR):148

LIBRARY PREPARATION

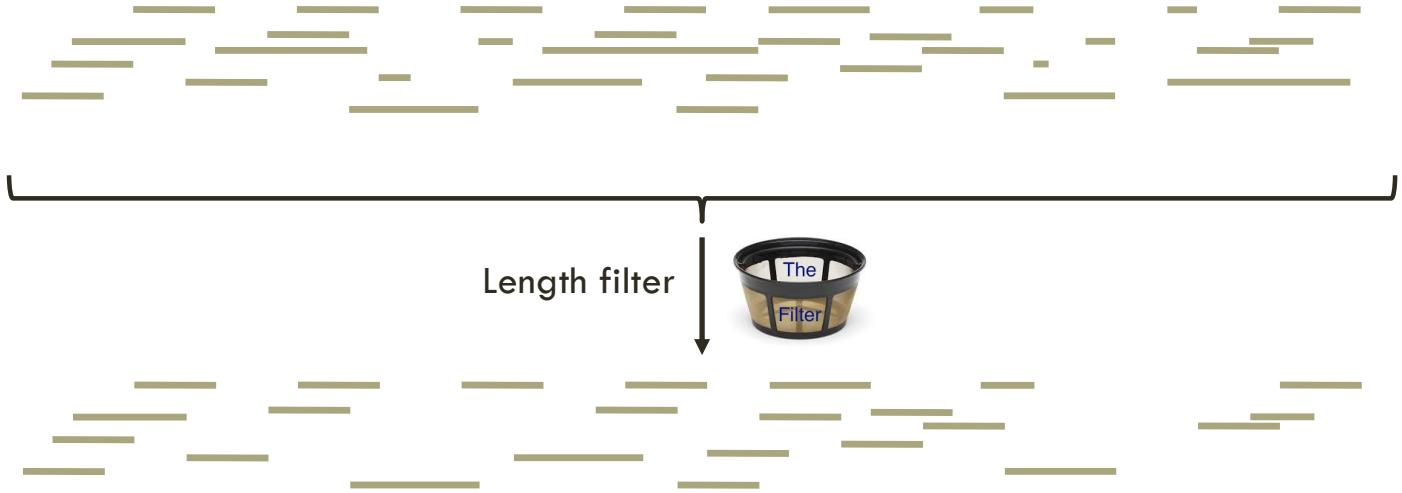
STRATEGIES TO SEQUENCE LONG DNA MOLECULES: SHOTGUN SEQUENCING

1. Randomly break template DNA into pieces

STRATEGIES TO SEQUENCE LONG DNA MOLECULES: SHOTGUN SEQUENCING

1. Randomly break template DNA into pieces

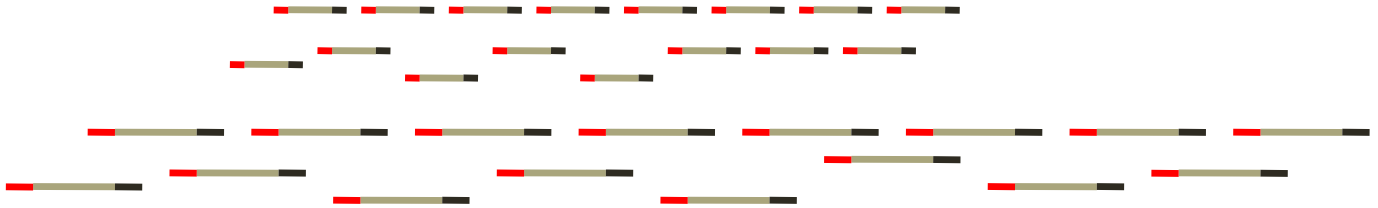
STRATEGIES TO SEQUENCE LONG DNA MOLECULES: SHOTGUN SEQUENCING



1. Processing of the template DNA
    1. Random fragmentation
    2. Size selection (-> Insert-size[1])

1 typically several 100 Bp for ‚short-read-Technologies (e.g.. Illumina)

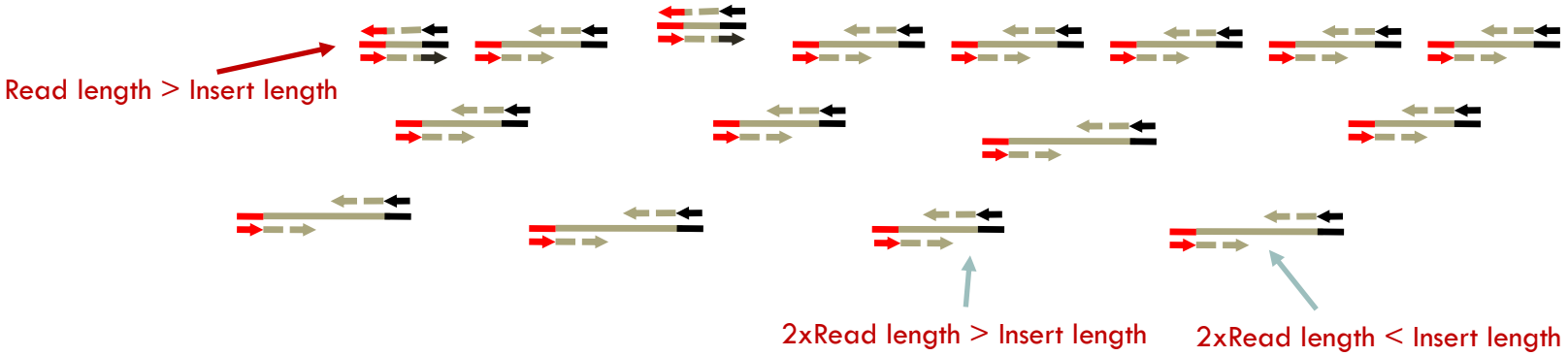
STRATEGIES TO SEQUENCE LONG DNA MOLECULES: SHOTGUN SEQUENCING



1. Processing of the template DNA
    1. Random fragmentation
    2. Size selection (-> Insert-size)
2. Append adapters[1] (DNA fragements with known sequence) that provide the necessary binding sites for downstream wet lab experiments (amplification, sequencing), as well as index sequences

1 each fragment gets the same set of adaptors

SHOTGUN SEQUENZIERUNG
EIN ANSATZ ZUR SEQUENZIERUNG LANGER DNA MOLEKÜLE

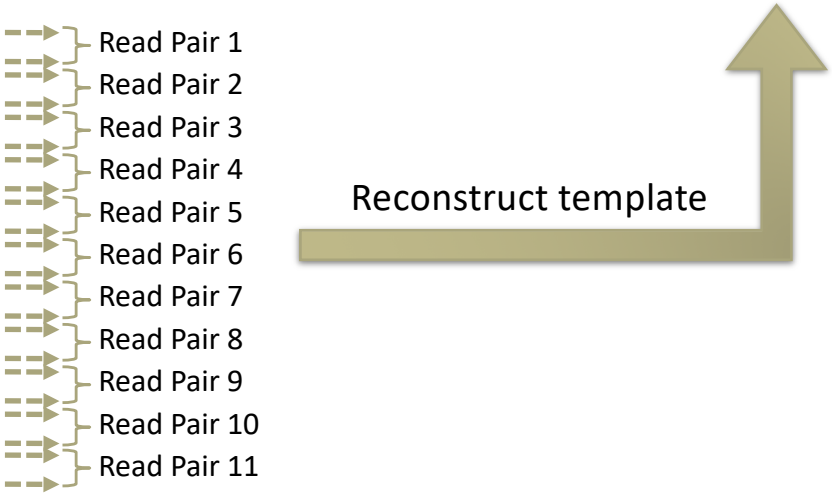


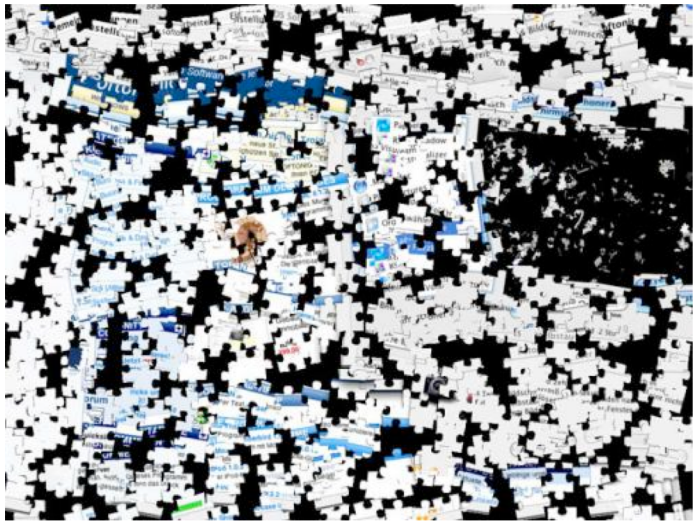
1. Processing of the template DNA
    1. Random fragmentation
    2. Size selection (-> Insert-size)
2. Append adapters[1] (DNA fragements with known sequence) that provide the necessary binding sites for downstream wet lab experiments (amplification, sequencing), as well as index sequences
3. Sequence the insert ends

1 typically, we sequence both ends of the insert -> Paired-End Reads

2 if read length > Insert length, you will seqeunce into the adapter

STRATEGIES TO SEQUENCE LONG DNA MOLECULES: SHOTGUN SEQUENCING





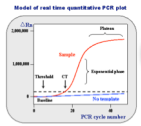
1. Randomly break template DNA into pieces
2. Add adapters of known sequence to the fragment ends
3. Sequence (typically) the ends of the fragments
4. Identify and remove adapter part from the determined sequences
5. Reconstruct template sequence from the sequence reads

Sanger Sequencing in a Nutshell (Sequencing by synthesis)

The Template:
5′–…CTGATCTATGCTCGCACT…–3′
3′–…GACTAGATACGAGCGTGA…–5′

Step1: Template amplification

single template molecule



Polymerase
Chain
Reaction
~35 cycles

Millions of identical template molecules

Step2: Cycle sequencing

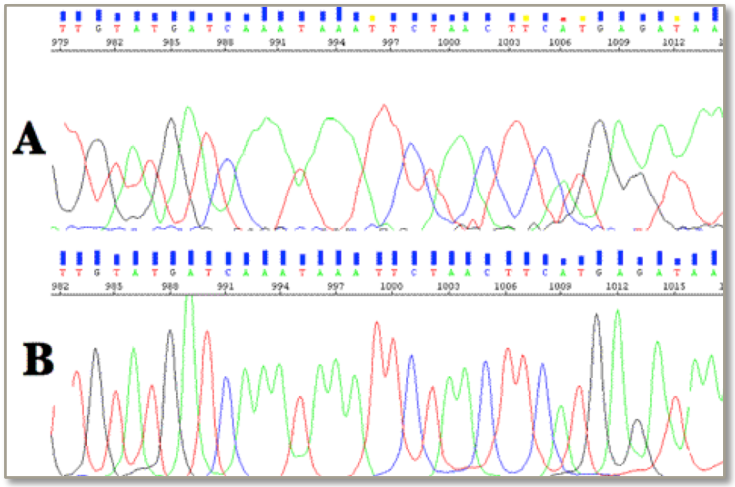
DNA-Polymerase
Primer for starting the synthesis
Desoxinucleotides:
dATP, dCTP, dTTP, dGTP
Di-Desoxinucleotides (Dye-Terminators)
ddATP, ddCTP
ddTTP, ddGTP

3′–…GACTAGATACGAGCGTGA…–5′ (template)
5′–…CTGAT→→→ (primer)

Repeat cycle of primer annealing, polymerization and strand separation *n* times

…CTGATC
…CTGATCT
…CTGATCTA
…CTGATCTAT
…CTGATCTATG
…CTGATCTATGC
…CTGATCTATGCT
…CTGATCTATGCTC
…CTGATCTATGCTCG

Step 3: Size separation via electrophoresis and **detection** of fluorescence markers



Example for a chromatogram

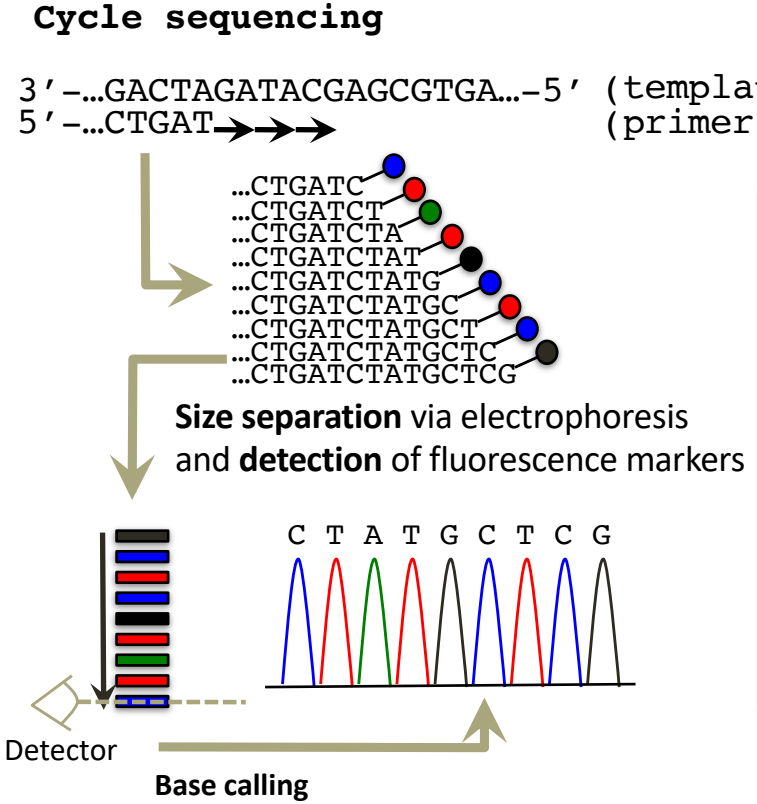
C T A T G C T C G

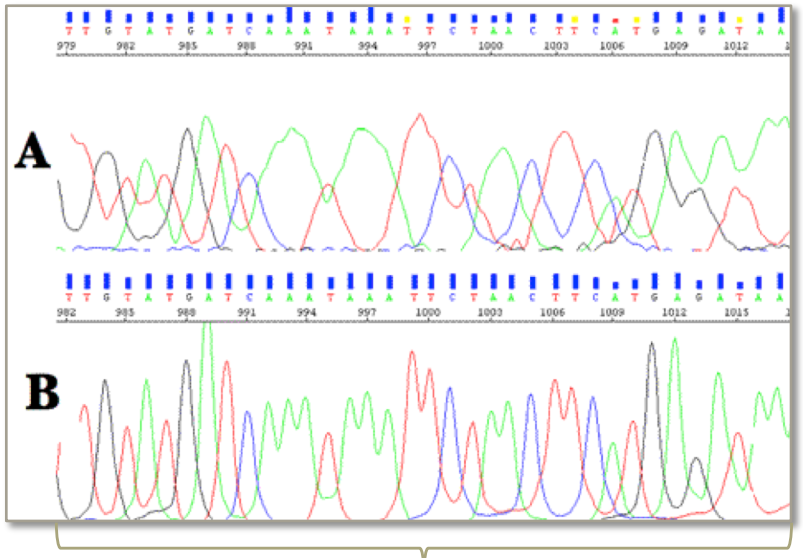
Electric field

Base calling

Detector

Sanger Sequencing in a Nutshell: Base quality values ***Q***





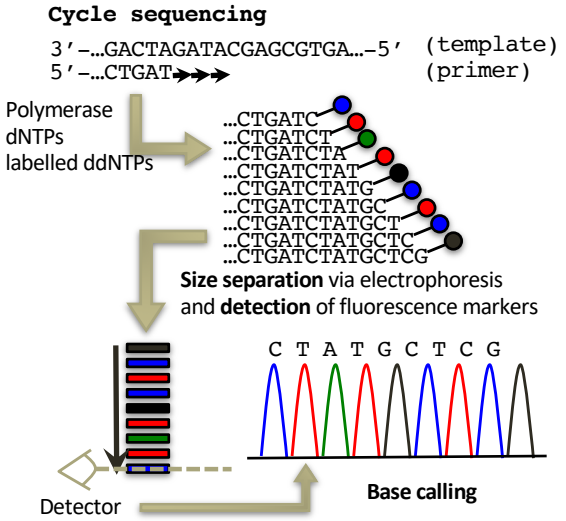
Quality Parameters
- Peak Spacing (7)
- Uncalled/Called Ratio (7)
- Uncalled/Called Ratio (3)
- Peak Resolution

Phred Base Quality

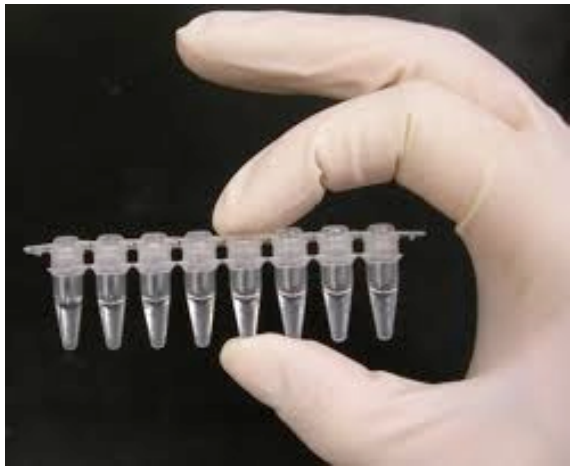
$Q = -10 \log_{10}(P_e)$*

*P_e: empirical error probability

Ewing B, Green P: Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186-194 (1998).

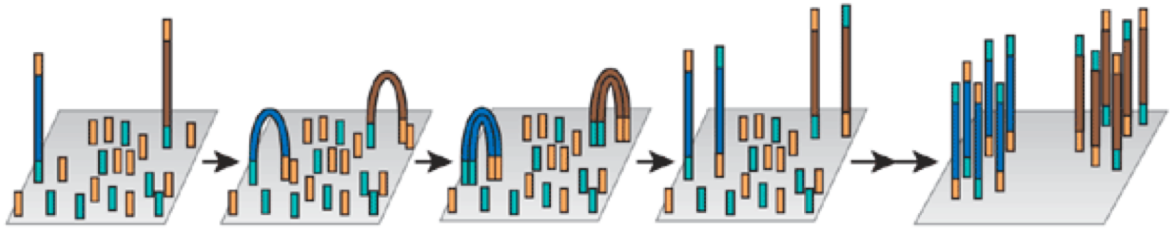


low degree of parallelization= low throughput



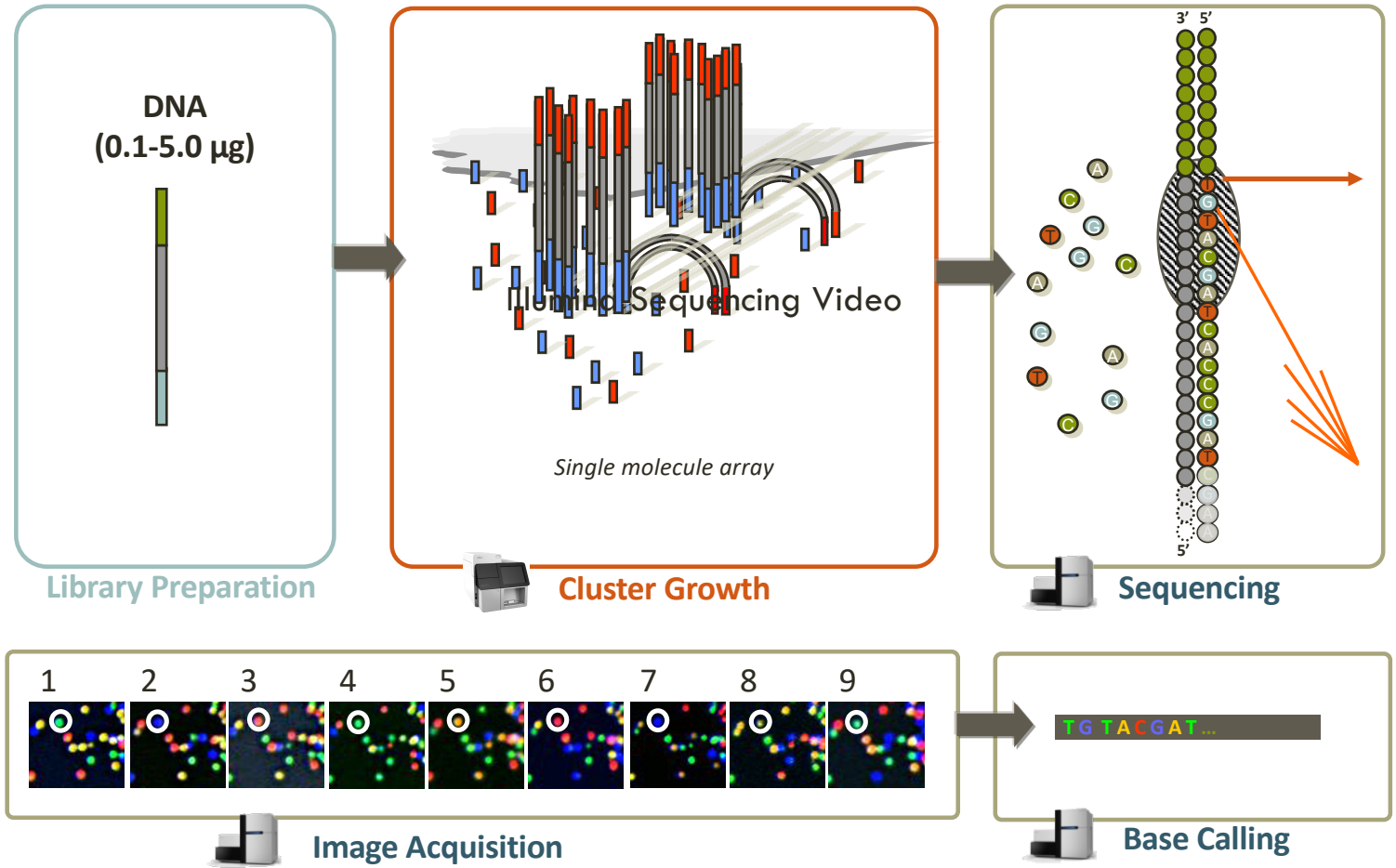
Main advantage of Next Generation Sequencing technologies: Parallelization

Bridge Amplification (Illumina)



Analyzing Millions of sequences at the same time

ILLUMINA SEQUENCING TECHNOLOGY WORKFLOW



As a first step, create the sequencing library

(There are many different kinds of libraries*)

- Single read libraries:
  - Unidirectional Sequencing
  - Single Read Flowcells ONLY
  - Counting applications: ChIP or low coverage resequencing projects

- Paired end libraries:
  - Uni- OR Bi-directional (paired reads)
  - Paired End Flowcells; Single: Unidirectional only
  - Most applications, #1 whole genome shotgun assembly
  - Tailor insert size and distribution per project:
    - Tight size distribution – Assembly, structural rearrangement detection
    - Wide distribution libraries - Resequencing, high coverage

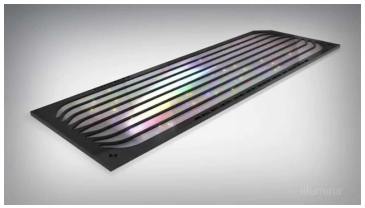
- Multiplex Paired End (aka Indexing or Barcoding)
  - Uni- OR Bi-directional
  - Allows multiple libraries per lane
  - 12 Index tags available x 8 lanes = 96 libraries per flowcell

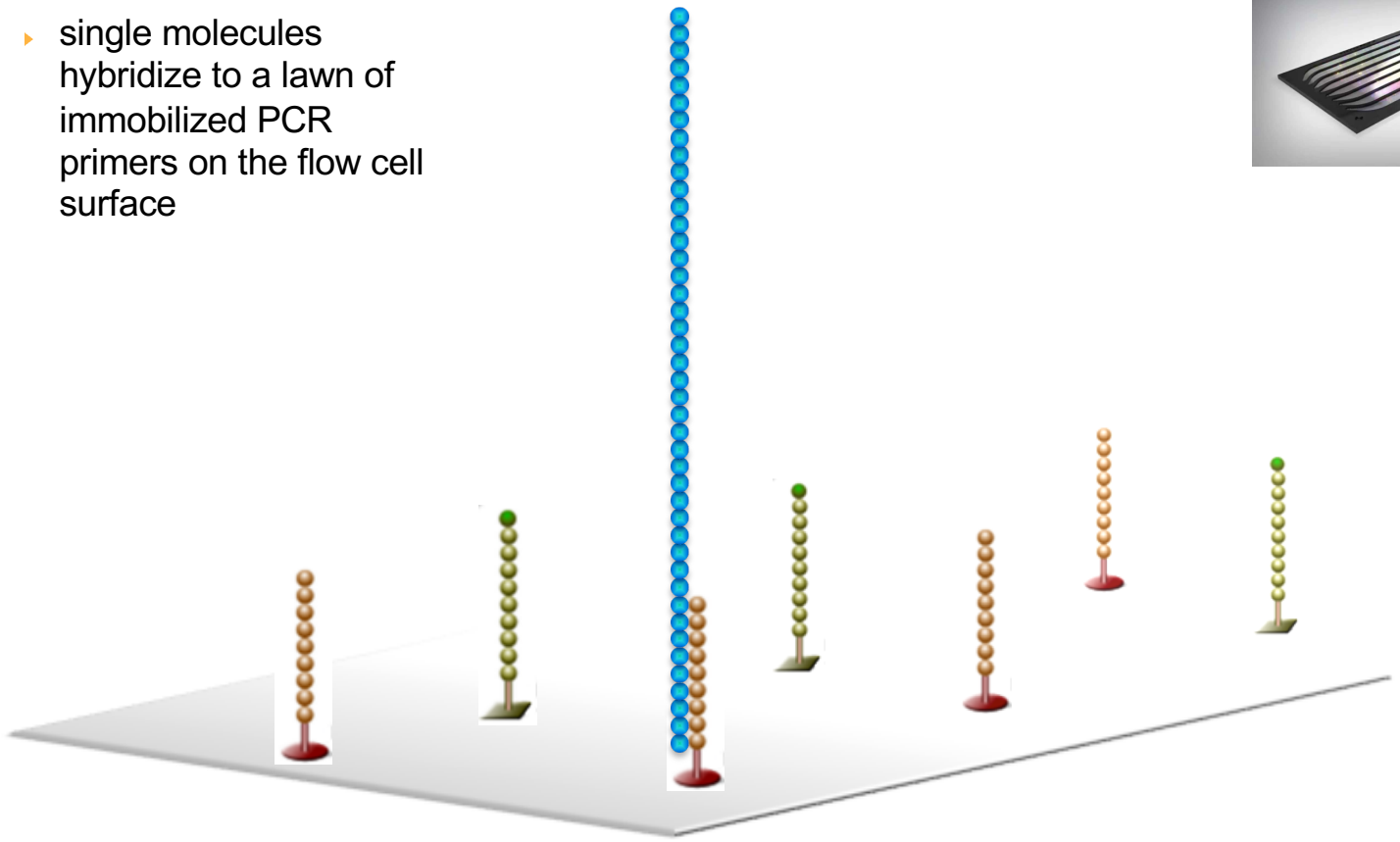
BROAD INSTITUTE illumina

*Make sure you know what kind of library you are dealing with!

TEMPLATE HYBRIDIZATION AND EXTENSION

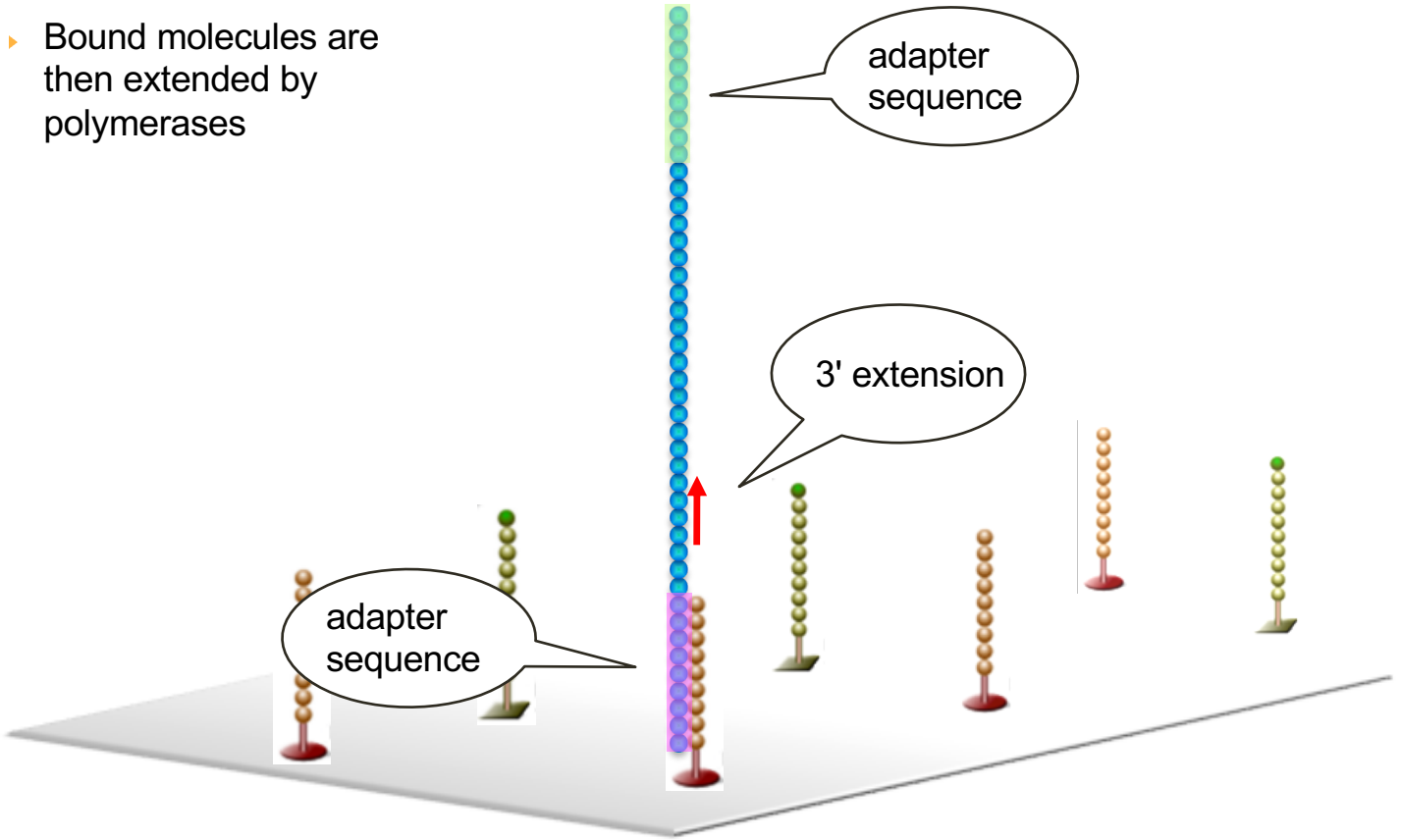
- single molecules hybridize to a lawn of immobilized PCR primers on the flow cell surface





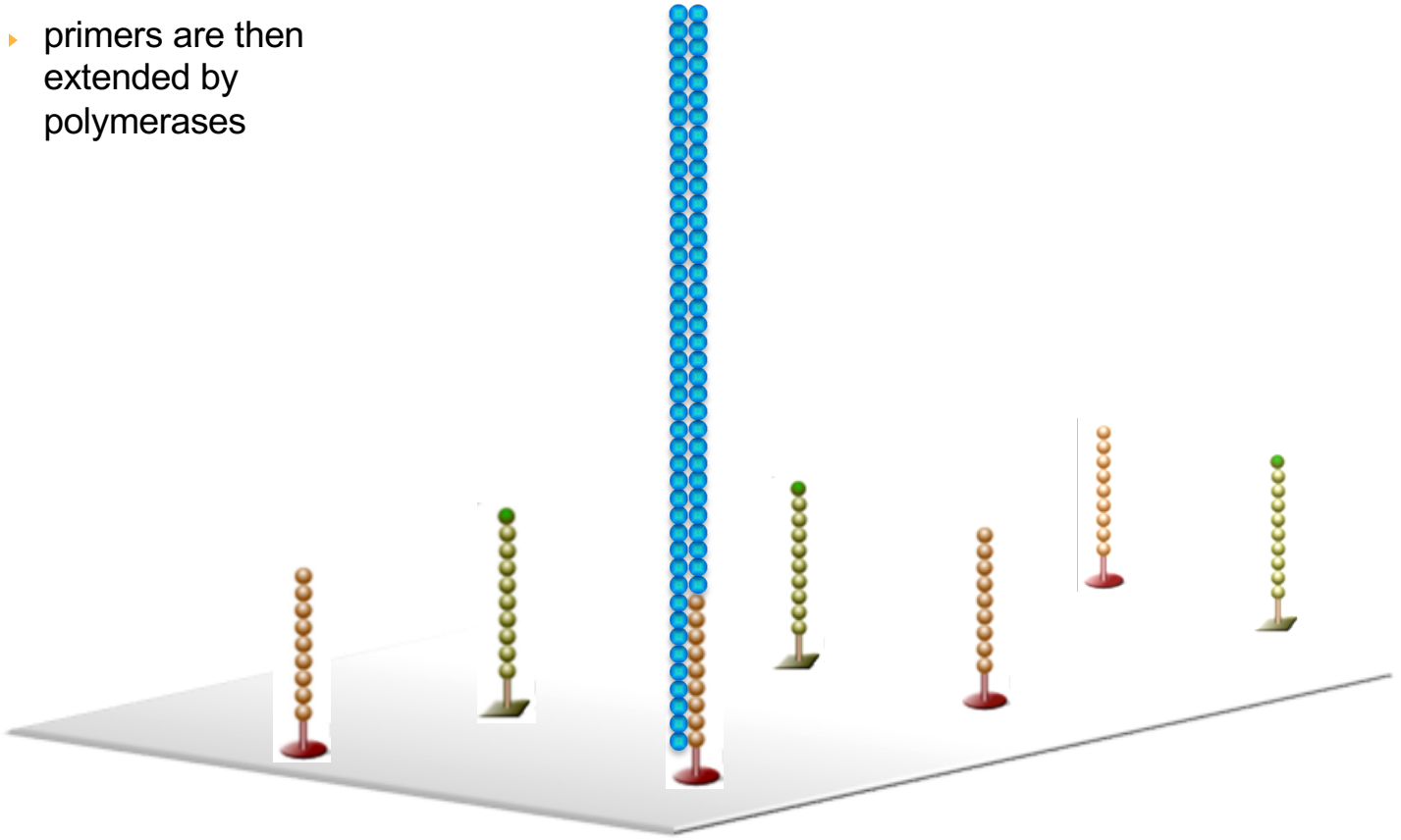
TEMPLATE HYBRIDIZATION AND EXTENSION

- Bound molecules are then extended by polymerases



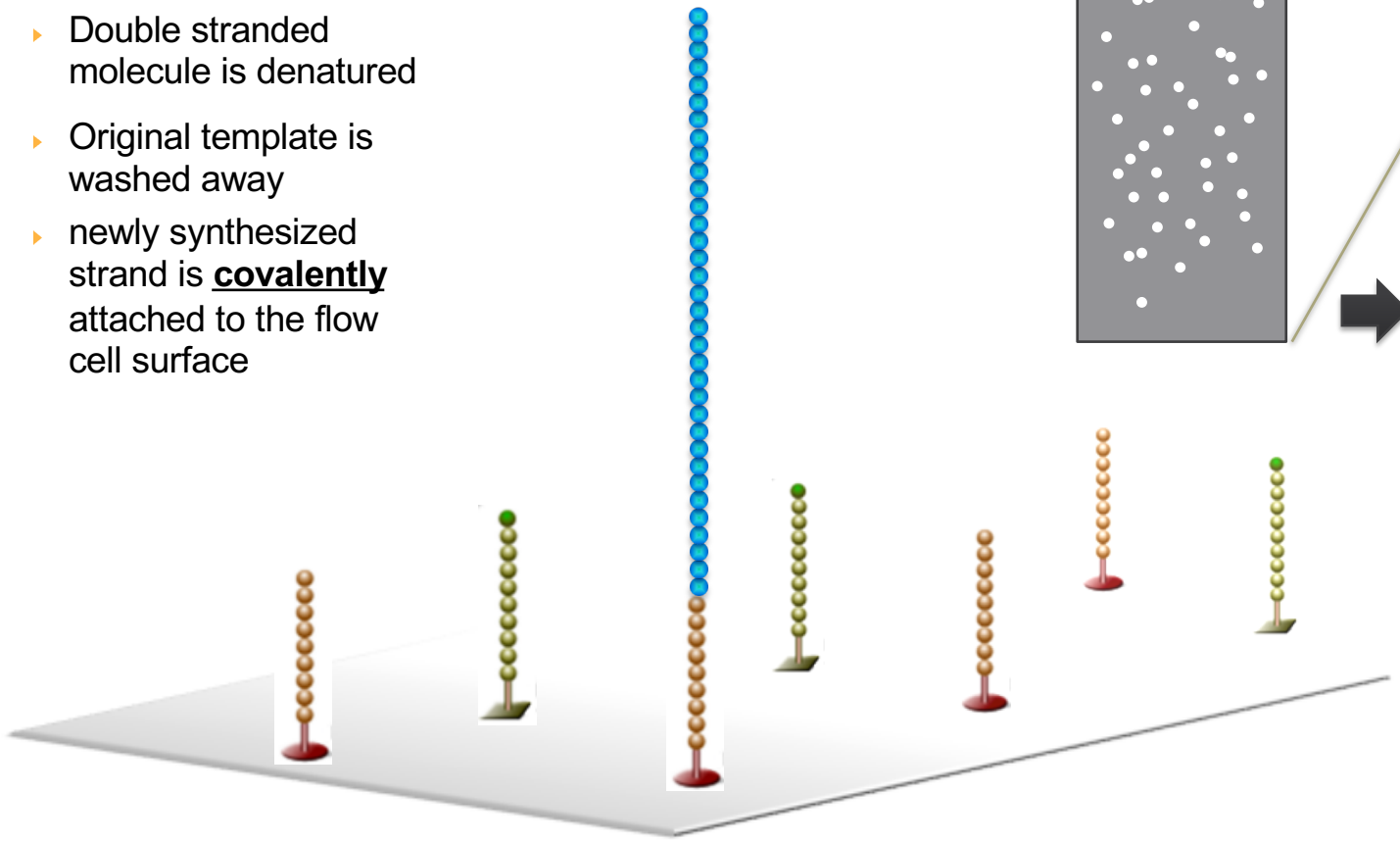
TEMPLATE HYBRIDIZATION AND EXTENSION

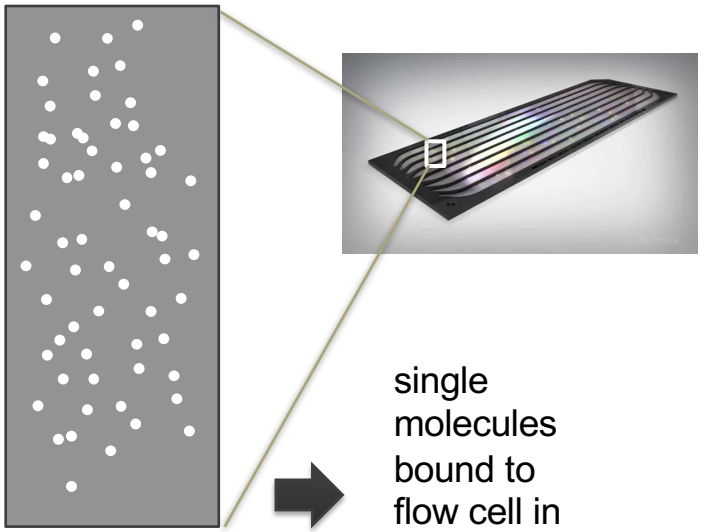
- primers are then extended by polymerases



REMOVAL OF ORIGINAL STRAND

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

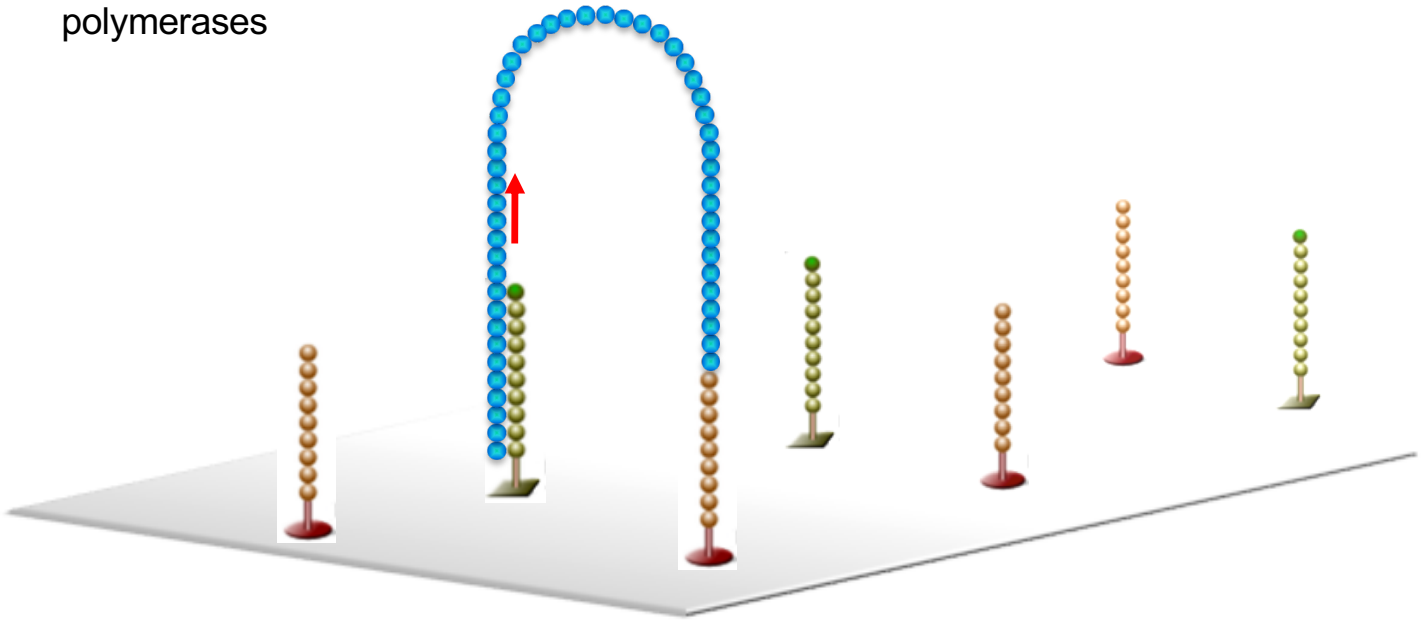




single molecules bound to flow cell in a random pattern

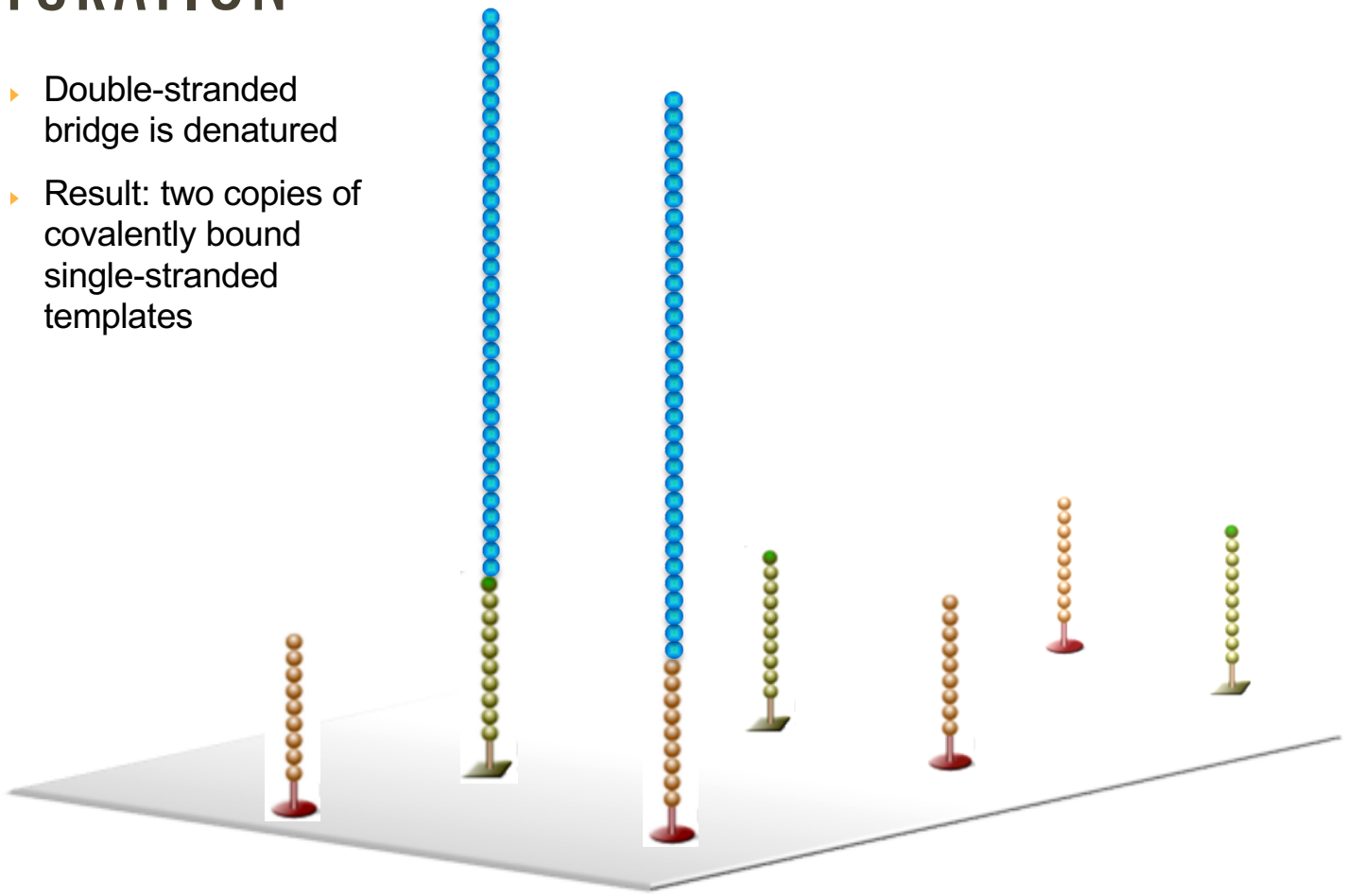
BRIDGING OVER

- Single-strand flips over to hybridize to adjacent oligos to form a bridge
- Hybridized primer is extended by polymerases



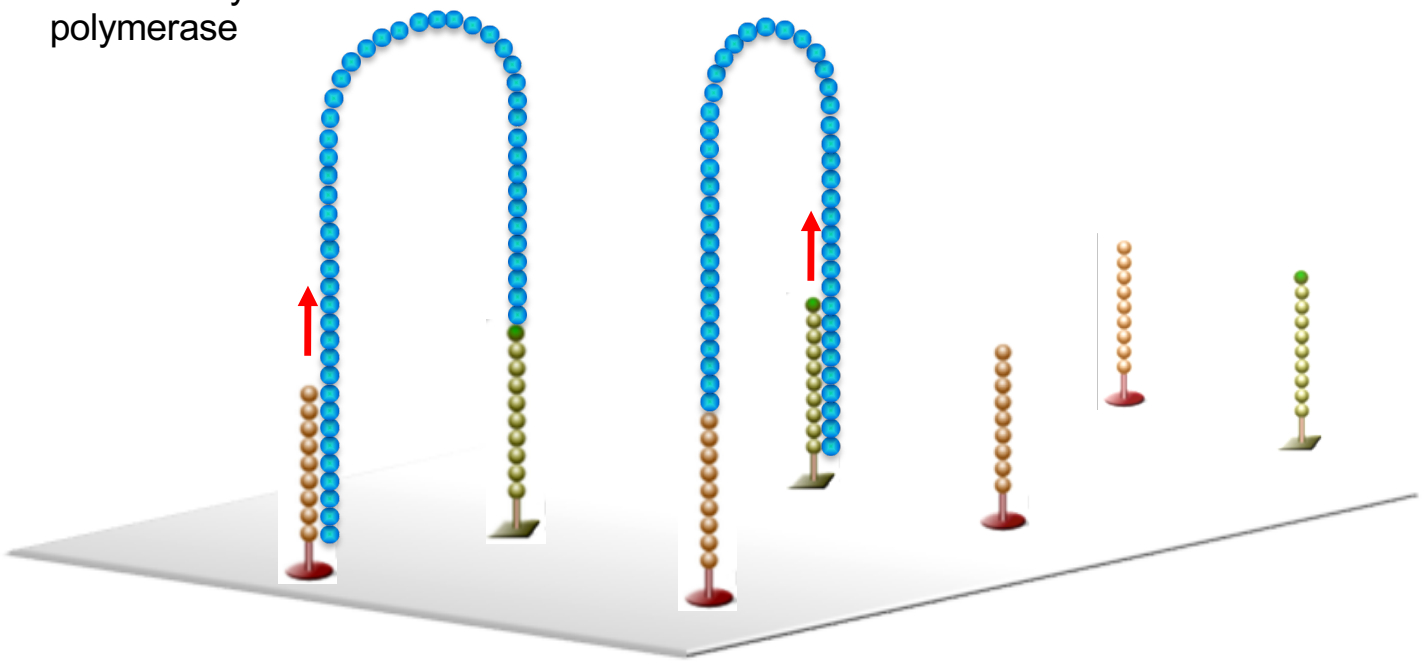
DENATURATION

- Double-stranded bridge is denatured
- Result: two copies of covalently bound single-stranded templates



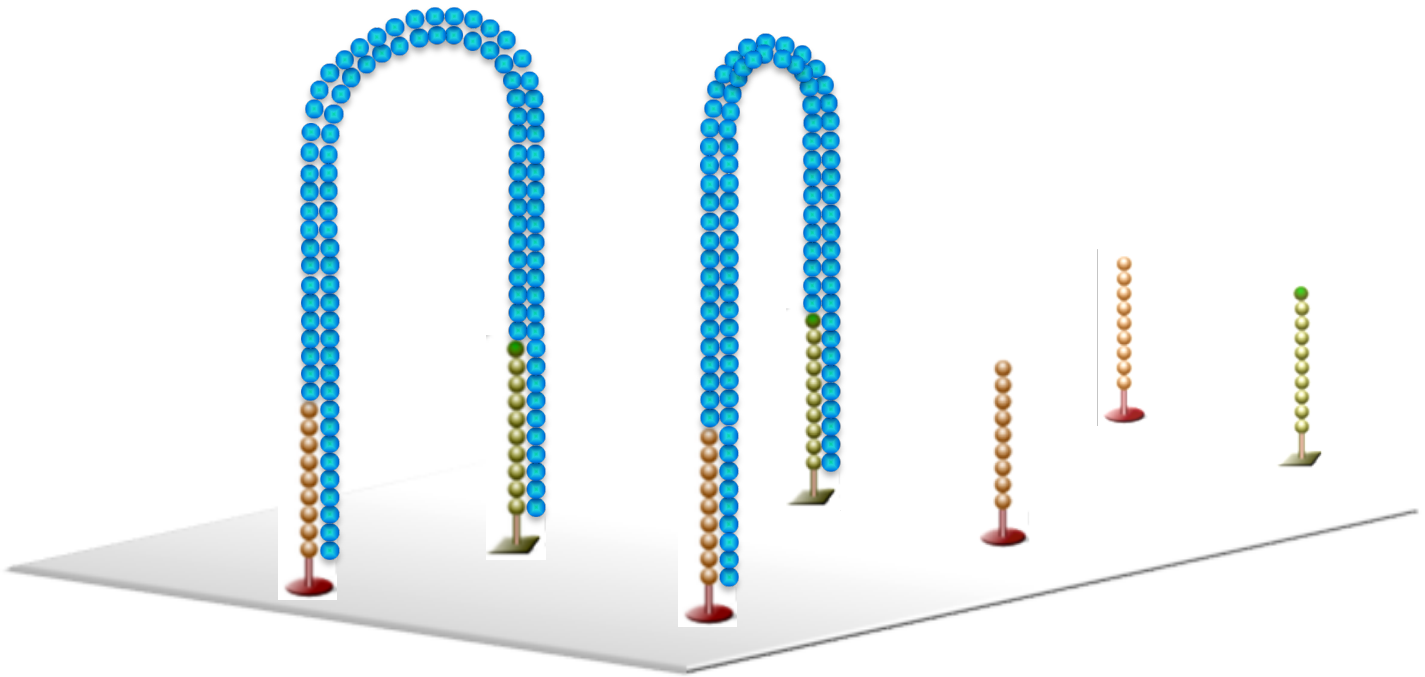
BRIDGING OVER OF TEMPLATES

- Single-strands flip over to hybridize to adjacent oligos to form bridges
- Hybridized primer is extended by polymerase



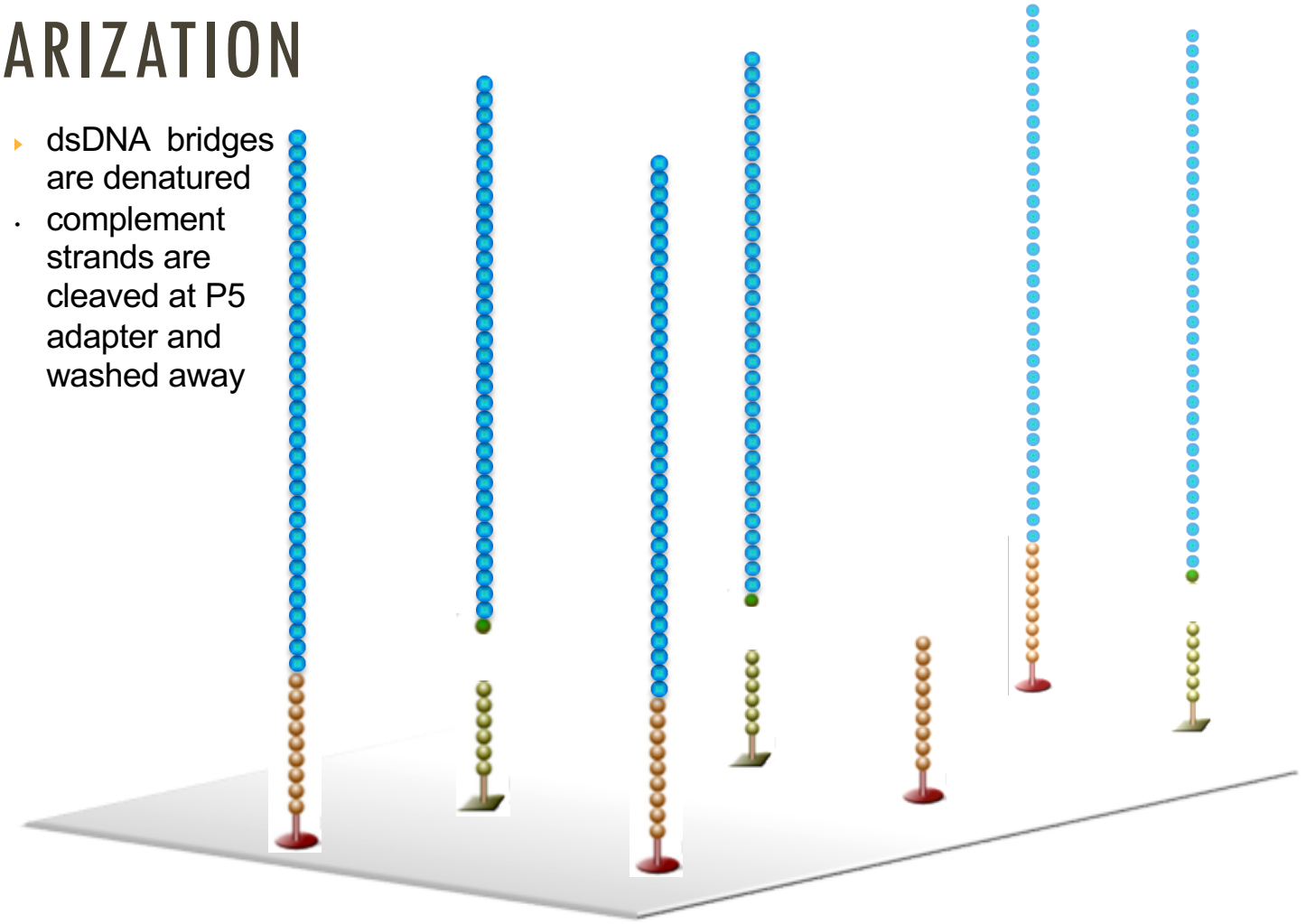
AMPLIFICATION

- Bridge amplification cycle repeated until multiple bridges are formed across the entire flow cell



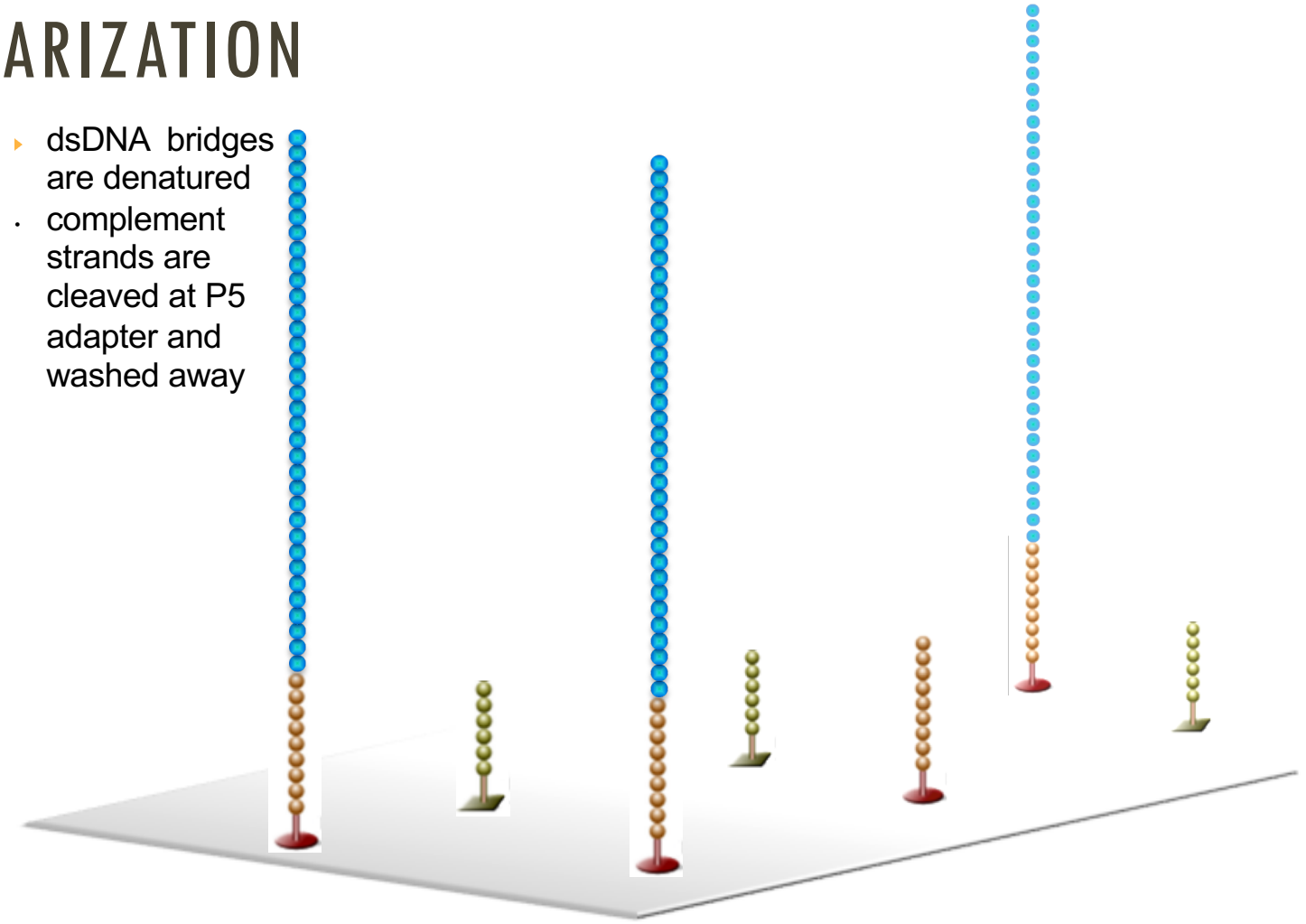
LINEARIZATION

- dsDNA bridges are denatured
- complement strands are cleaved at P5 adapter and washed away



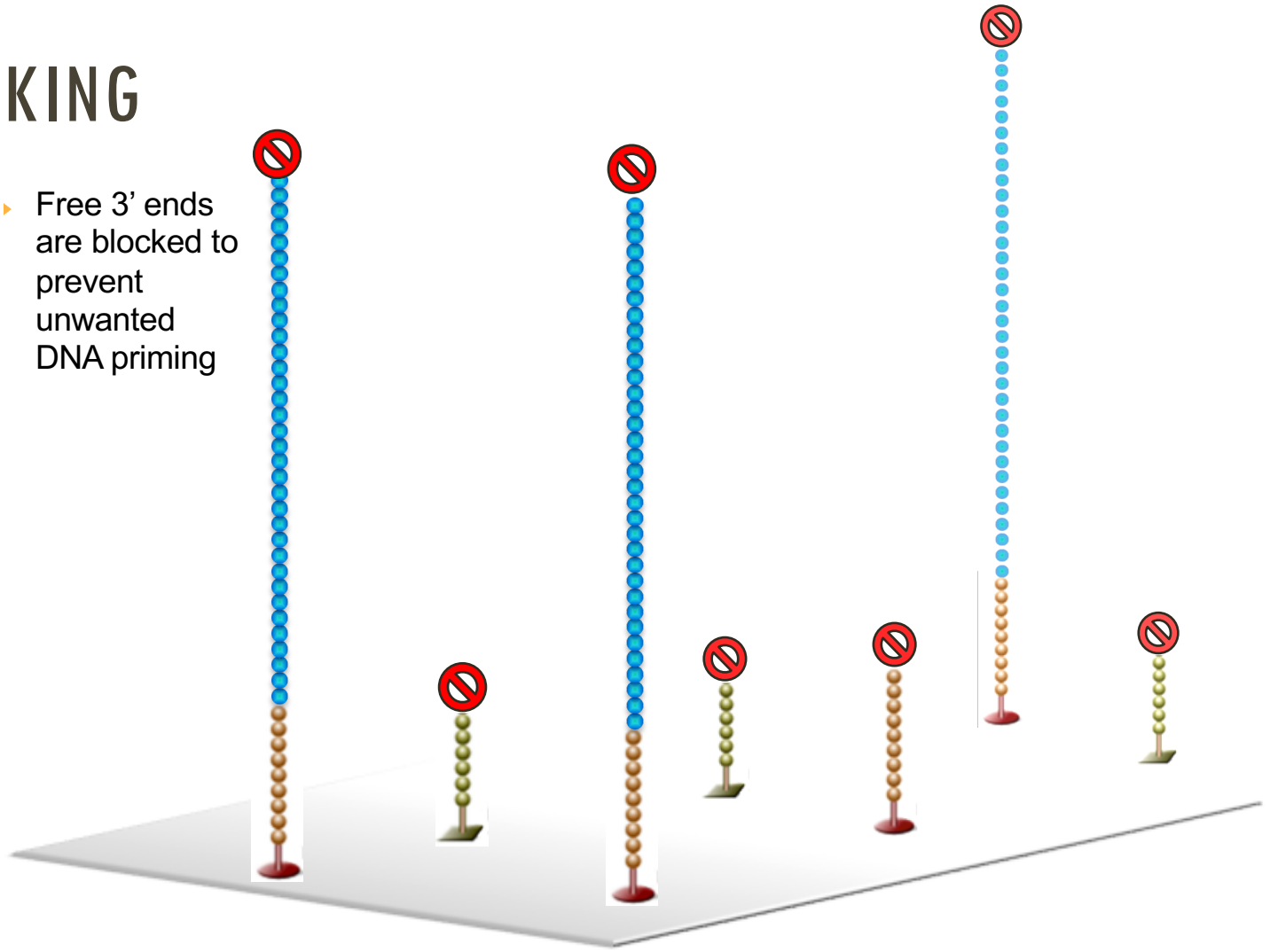
LINEARIZATION

- dsDNA bridges are denatured
- complement strands are cleaved at P5 adapter and washed away



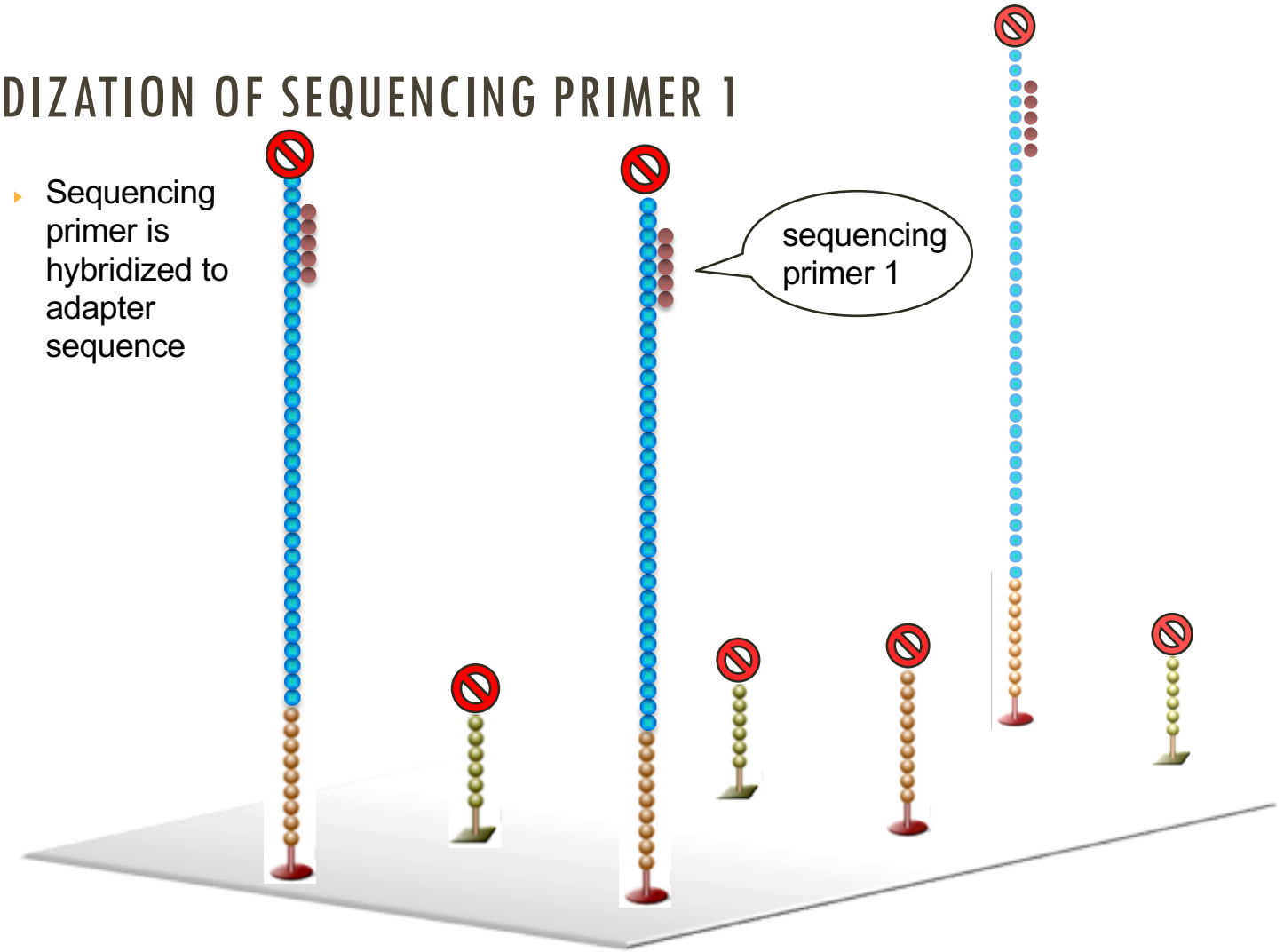
BLOCKING

- Free 3’ ends are blocked to prevent unwanted DNA priming

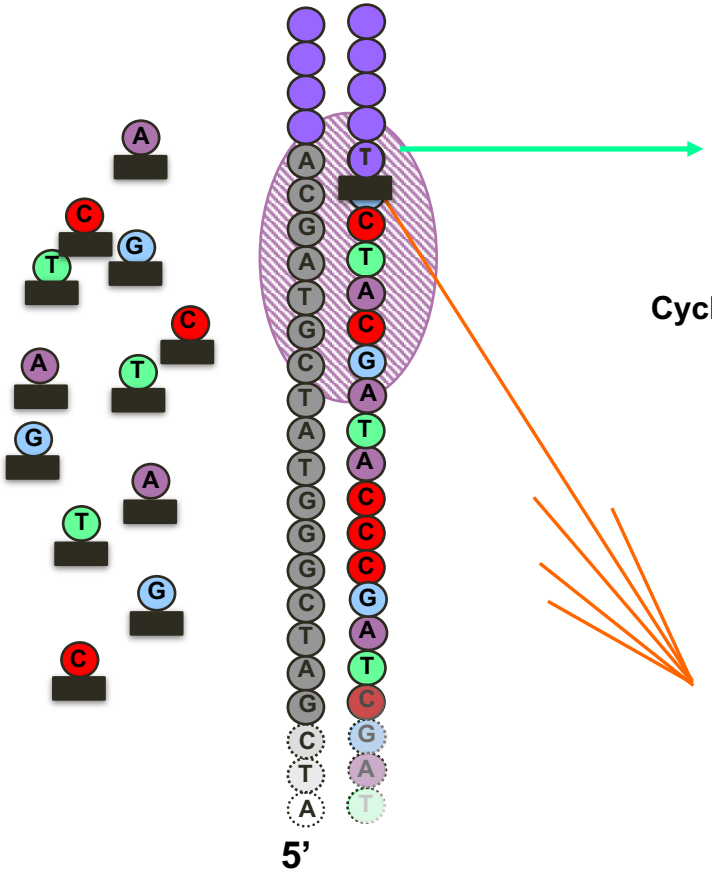


HYBRIDIZATION OF SEQUENCING PRIMER 1

- Sequencing primer is hybridized to adapter sequence



SEQUENCING BY SYNTHESIS (SBS)



Cycle 1: Add sequencing reagents (All 4 labeled nucleotides in 1 reaction)

First base incorporated (reversible dye terminator)

Remove unincorporated bases

Detect signal

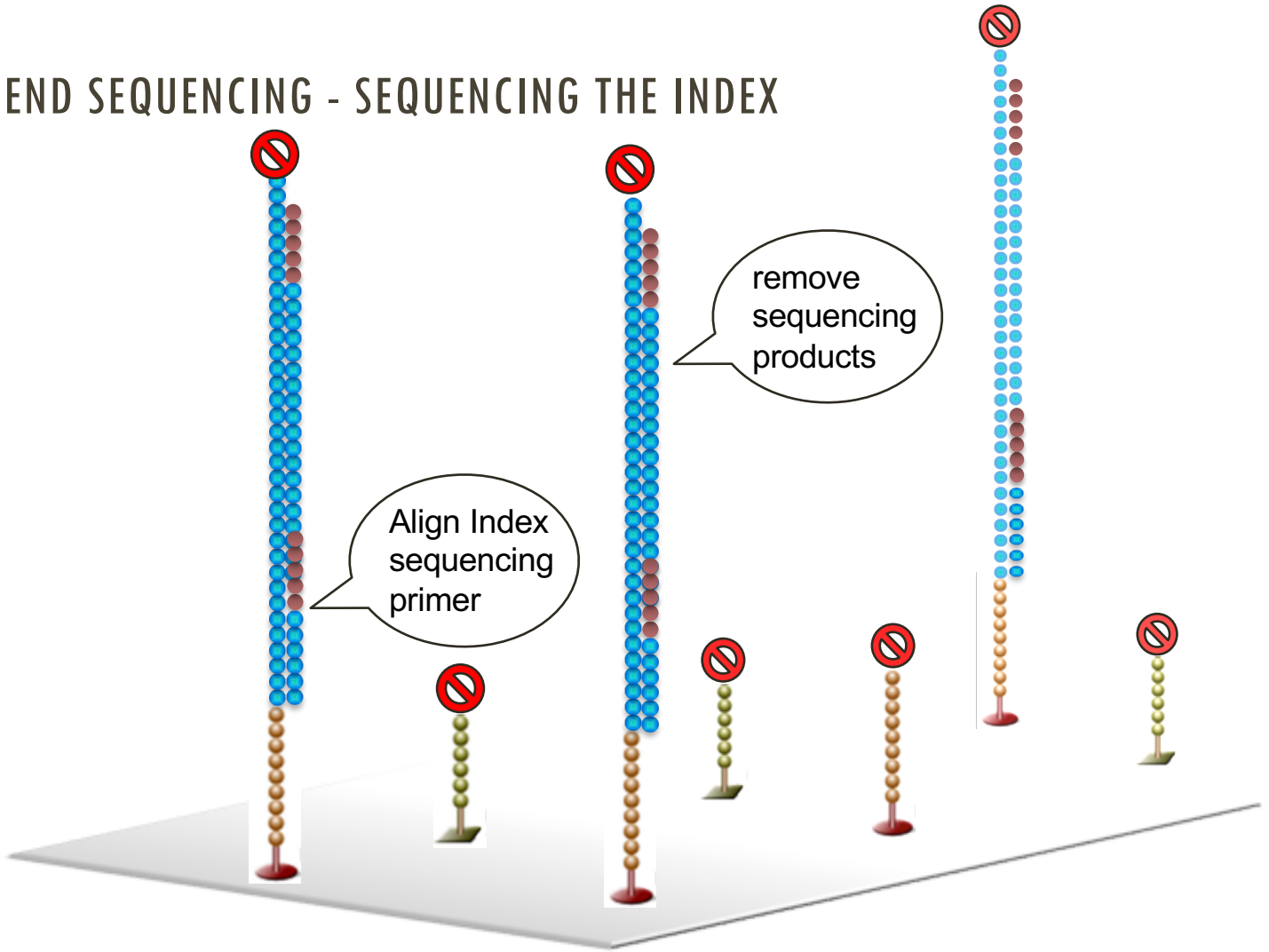
Unprotect/remove dye

Cycle 2-n: Add sequencing reagents and repeat

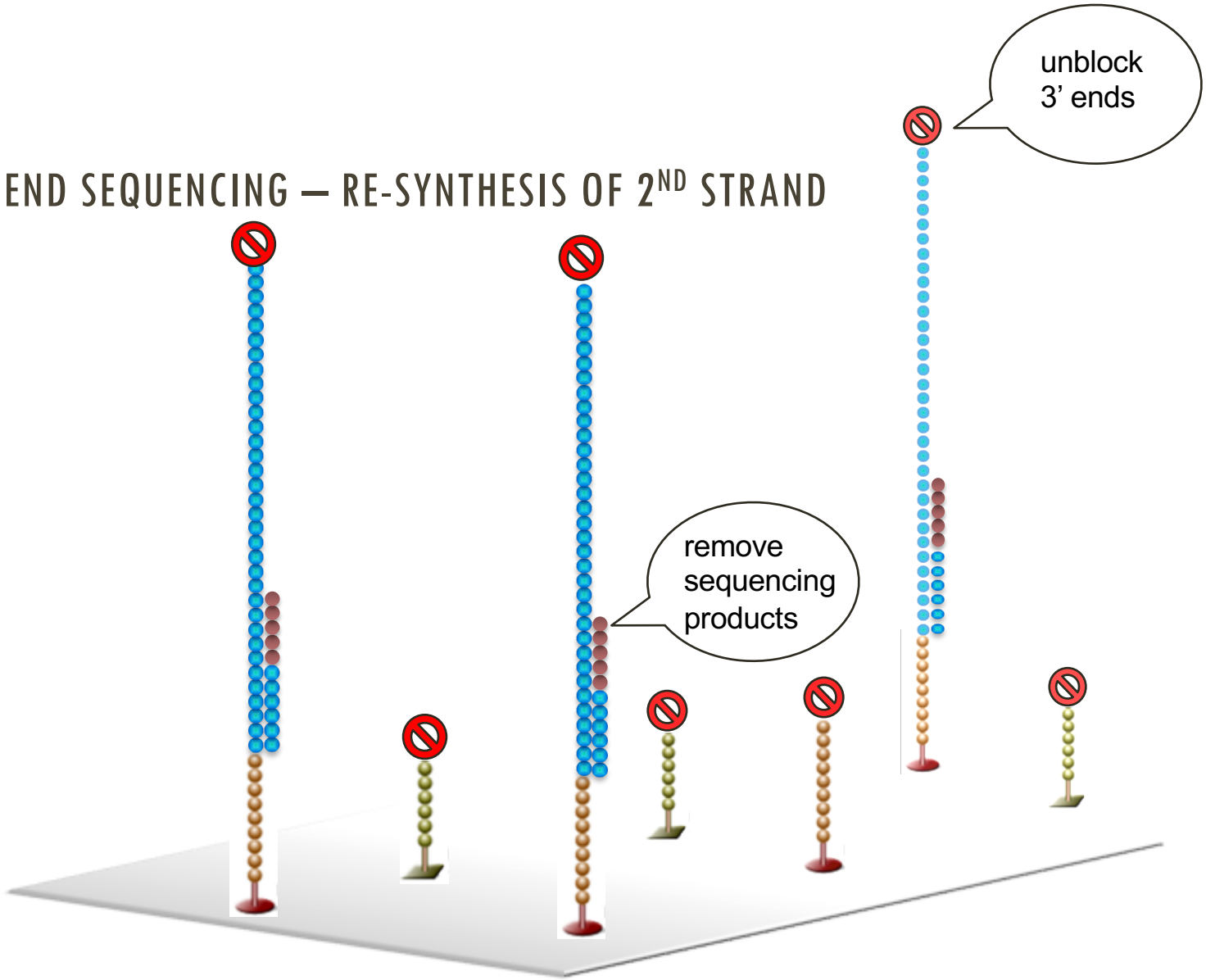
Key points

- **All four labelled nucleotides in one reaction**
- **Reversible dye terminator**
- **Base-by-base sequencing**
- **Real-time sequencing**
- **Read length is determined by the number of cycles!**

PAIRED-END SEQUENCING - SEQUENCING THE INDEX

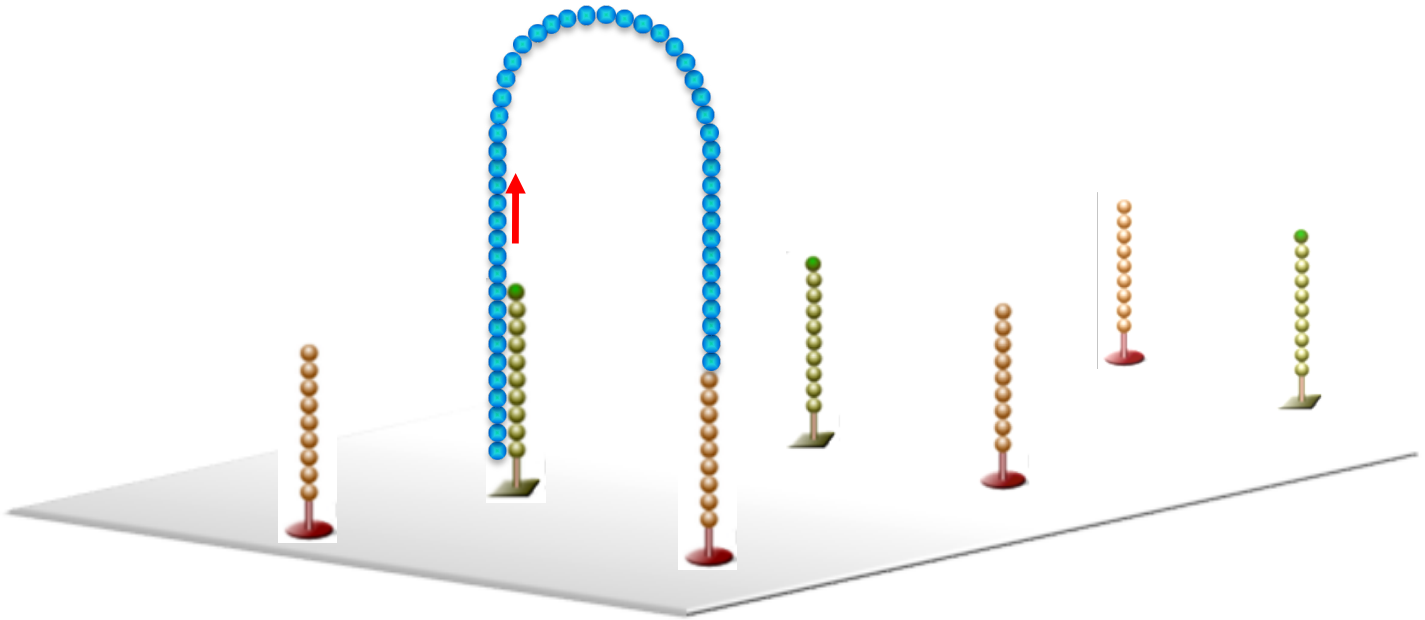


PAIRED-END SEQUENCING – RE-SYNTHESIS OF 2ND STRAND



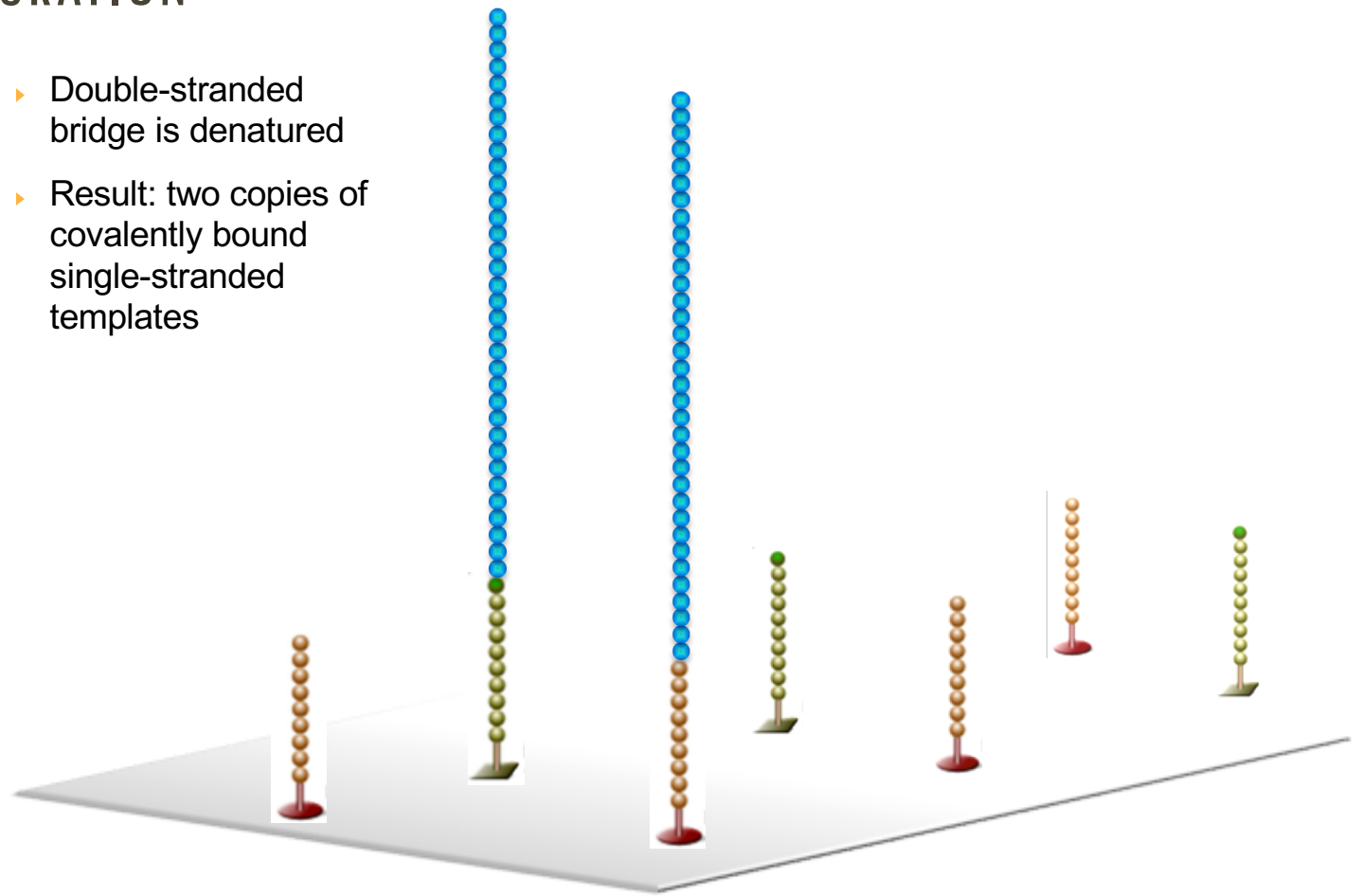
PAIRED-END SEQUENCING – RE-SYNTHESIS OF 2ND STRAND

- Bridge formation and 3' extension



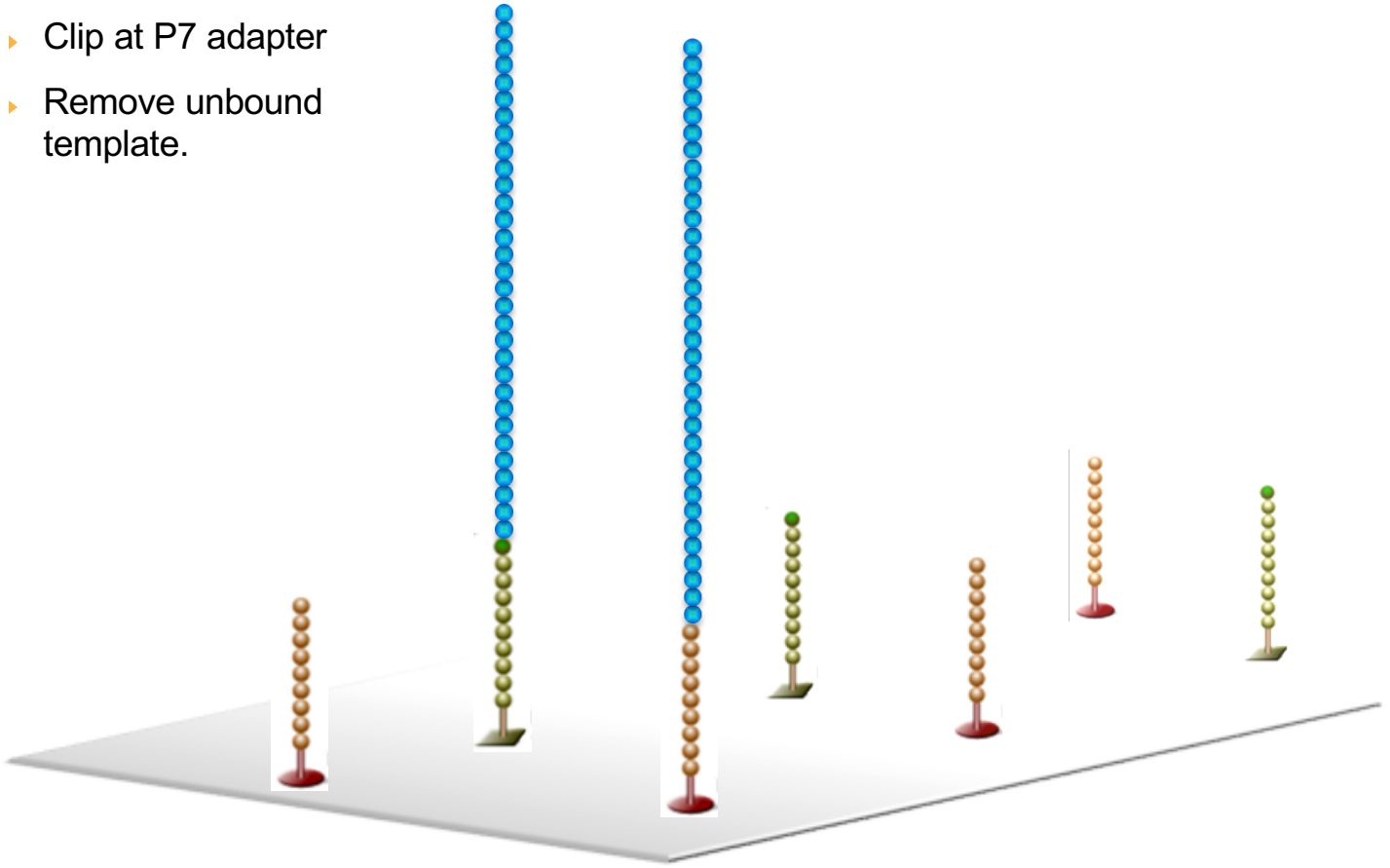
DENATURATION

- Double-stranded bridge is denatured
- Result: two copies of covalently bound single-stranded templates



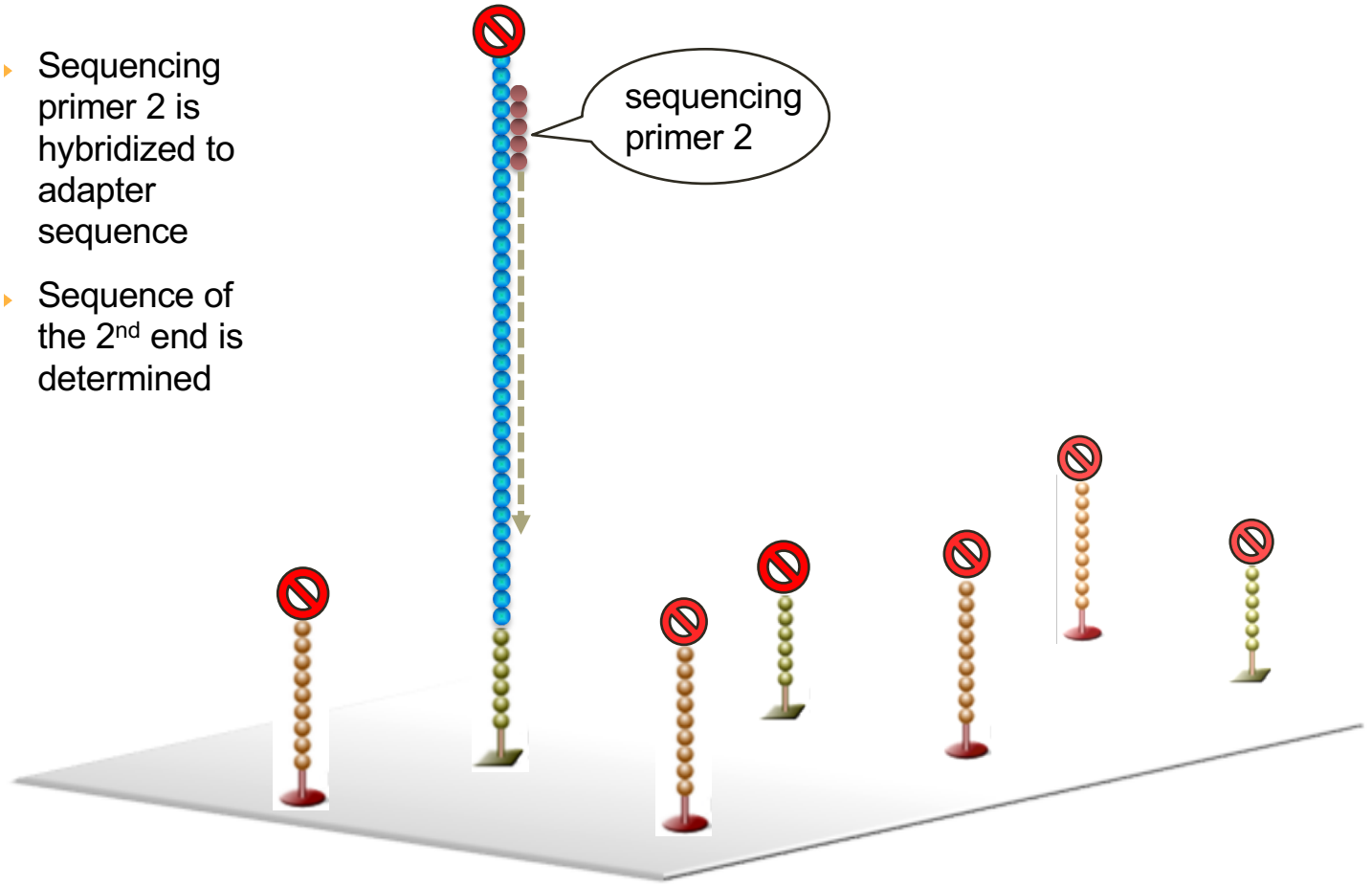
CLEAVAGE AND REMOVAL OF FIRST STRAND

- Clip at P7 adapter
- Remove unbound template.



HYBRIDIZATION OF SEQUENCING PRIMER 2

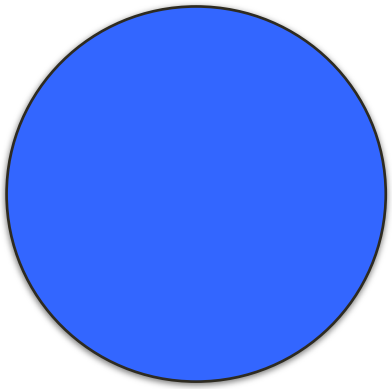
- Sequencing primer 2 is hybridized to adapter sequence
- Sequence of the 2nd end is determined



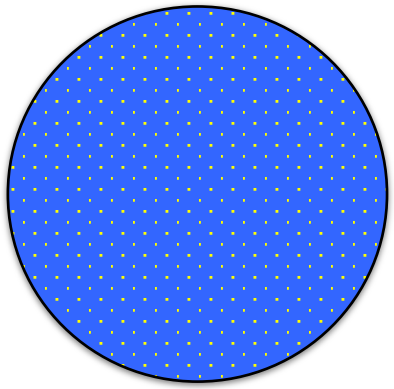
THE METHODS DESCRIBED SO FAR AVERAGE THE SIGNAL OVER MILLIONS OF COPIES OF THE SAME SEQUENCE. WHY IS THIS PROBLEMATIC?

- Errors during PCR amplification render copies not 100% identical. Especially errors at an early stage of the PCR can mimic heterozygous positions.

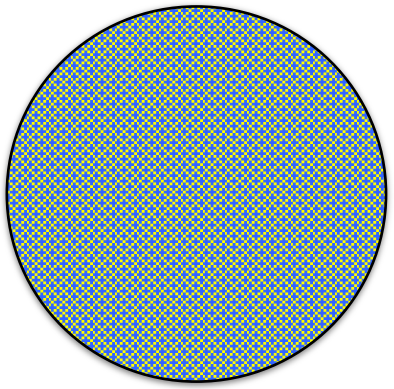
- Not every copy of a pool of millions of sequences will incorporate a base in each cycle. With increasing numbers of cycles the length heterogeneity of the already sequenced fraction will increase and the sequencing will get **out of phase.**



Template: AGACTATTTA
TC**T**



Template: AGACTATTTA
(9x) TCTGA**T**
Template: AGACTATTTA
(1x) TCTG**A**



Template: AGACTATTTA
(5x) TCTGATAAA**T**
Template: AGACTATTTA
(5x) TCTGATAA**A**

SEQUENCE READS ARE STORED TYPICALLY IN FASTQ FORMAT

Sequenzheader (-ID)

Sequenz

Separator[1]

Basenqualitäten

```
@D00689:288:CBUB7ANXX:2:2202:1336:1998#ATTACTCGTATAGCCT/1
CATCCTTCTCAGCTTGCAGGTCGGCGGCGCAGCTGGCGGATGTCTGTTTCTCGGCCTCCAGGT
CGGCGGCGCAGGCTCCAGAGTCNTCCTTTTCTATCTCCAGGTCGCCAGGACACTGCTCCAGC
+
<</<</FFFBBBFBFFFFF</</BF/FFFFBFFFFFB7BFB7///7/BF/F/BFF/7FF/B
B//</B</B/////77////#7<77///7/////7FFB7/7/7/////F/7BFFF<B/
```

1 – Forward

2 – Reverse

```
@D00689:288:CBUB7ANXX:2:2202:1336:1998#ATTACTCGTATAGCCT/2
NNNNNNNNNNNNNNCCNNNNNNNCNGNNNNNNNNNNNNNNNNNNNNNNNNANNAGNNCNNNNN
NNNNNNNNNNNNNTNGNNANNTGGAGNNNNNNAAGGANGNNNNNNNNNNNNNNNNNNNNNNN
+
##############B<#######<#<#########################<##</##/#####
############7#/##/##<</</######///77#7#########################
```

1 Optional kann hier noch einmal die Sequenz-Id wiederholt werden

SEQUENCE READS ARE STORED TYPICALLY IN FASTQ FORMAT

```
@D00689:288:CBUB7ANXX:2:2202:1336:1998#ATTACTCGTATAGCCT/1
CATCCTTCTCAGCTTGCAGGTCGGCGGCGCAGCTGGCGGATGTCTGTTTCTCGGCCTCCAGGT
CGGCGGCGCAGGCTCCAGAGTCNTCCTTTTCTATCTCCAGGTCGCCAGGACACTGCTCCAGC
+
<</<</FFFBBBFBFFFFF</<///BF/FFFFBFFFFFB7BFB7//7/BF/F/BFF/7FF/B
B//<///B</B/////77////#7<77///7/////7FFB7/7//7//////F/7BFFF<B/
```

Header Information:

D00689 the unique instrument name

288 the run id

CBUB7ANXX the flowcell id

2 flowcell lane

2202 tile number within the flowcell lane

1336 'x'-coordinate of the cluster within the tile

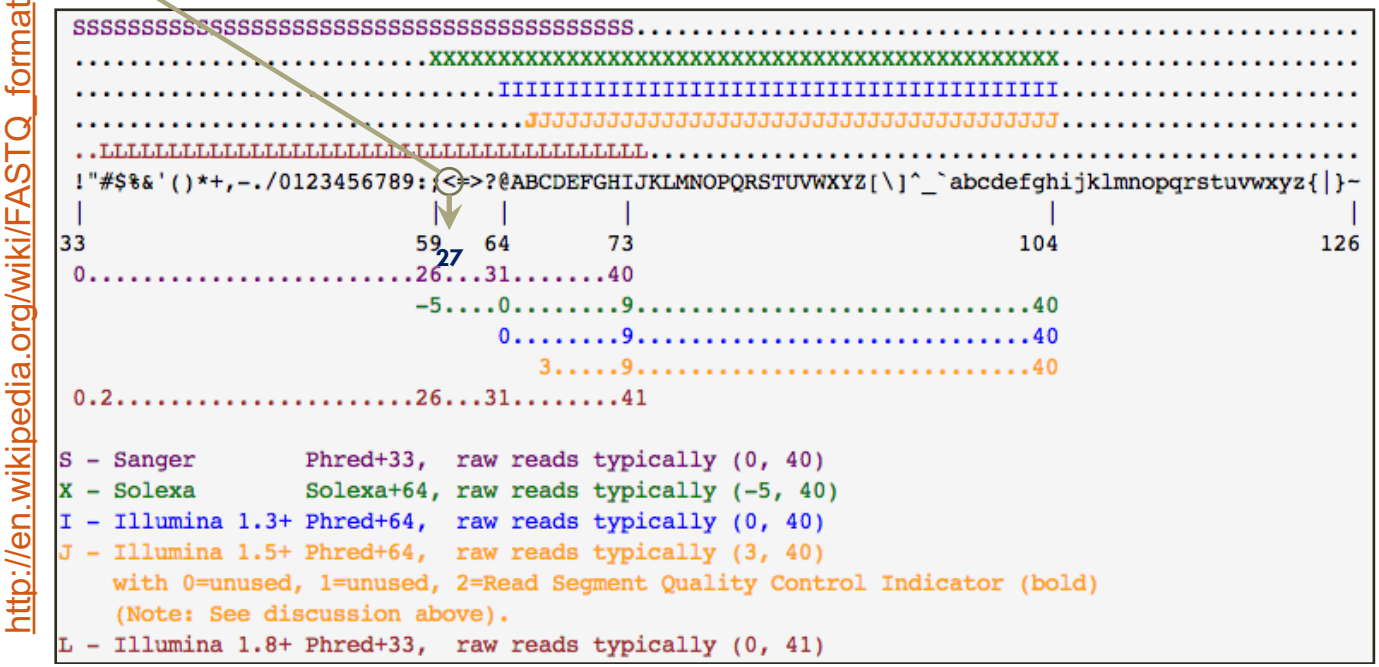
1998 'y'-coordinate of the cluster within the tile

ATTACTCGTATAGCCT index sequence

1 the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*

FILE FORMATS: FASTQ – QUALITY DECODING

```
@D00689:288:CBUB7ANXX:2:2202:1336:1998#ATTACTCGTATAGCCT/1
CATCCTTCTCAGCTTGCAGGTCGGCGGCGCAGCTGGCGGATGTCTGTTTCTCGGCCTCCAGGT
CGGCGGCGCAGGCTCCAGAGTCNTCCTTTTCTATCTCCAGGTCGCCAGGACACTGCTCCAGC
+
<</<</FFFBBBFBFFFFF</<///BF/FFFFBFFFFFB7BFB7///7/BF/F/BFF/7FF/B
B//<///B</B/////77/////#7<77///7/////7FFB7/7//7/////F/7BFFF<B/
```

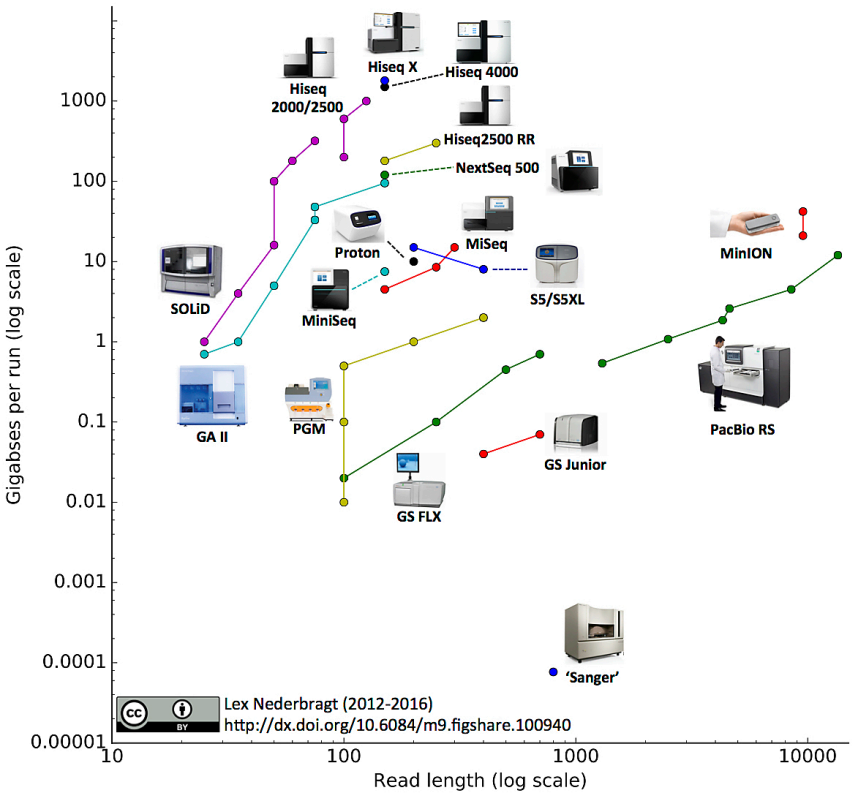


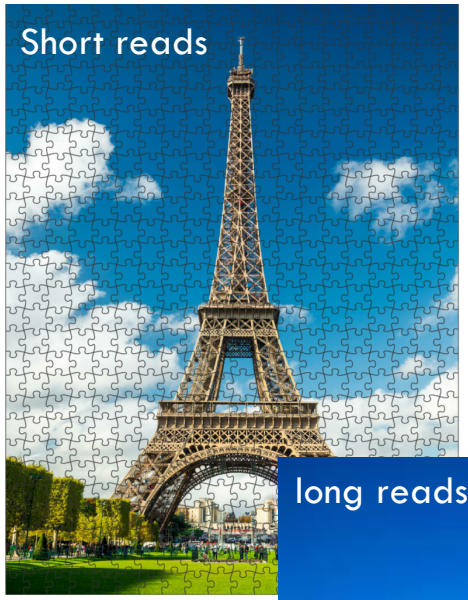
http://en.wikipedia.org/wiki/FASTQ_format

ASCI Value

Base quality values

FROM SHORT TO LONG READ SEQUENCING

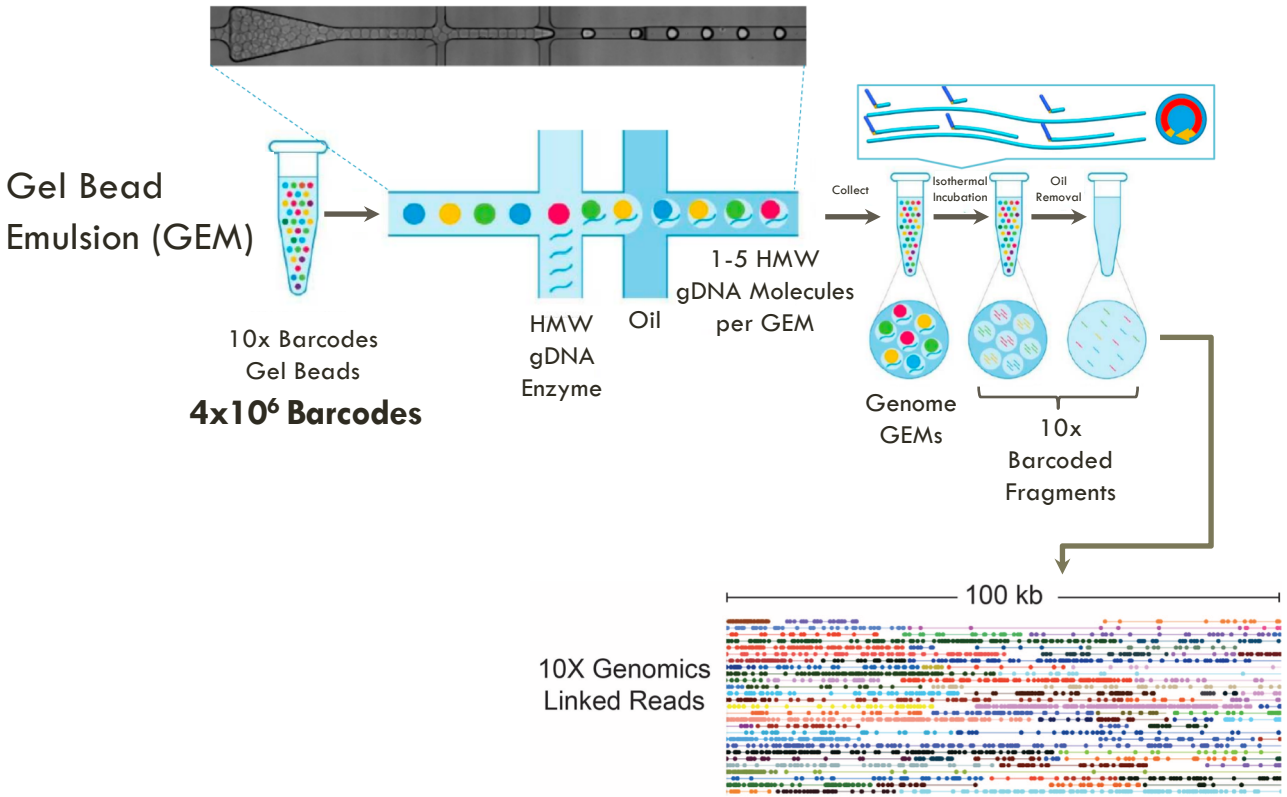




CHROMIUM PLATFORM - 10X GENOMICS

Key points

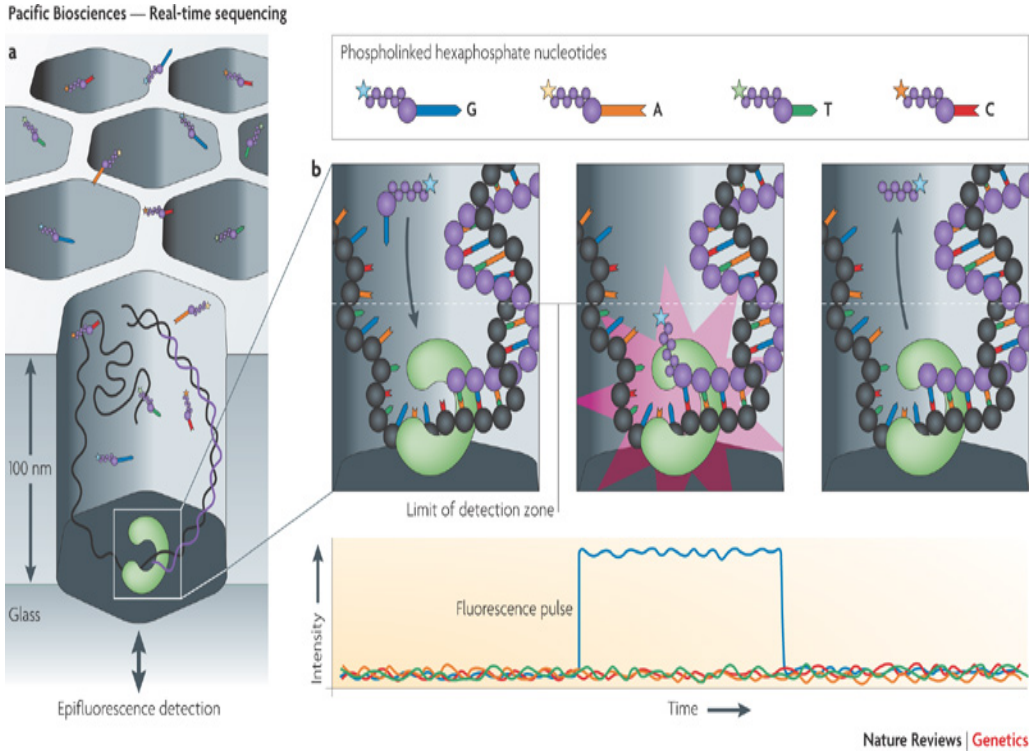
- Library Prep technique for Illumina sequencing
  - High throughput, low sequencing error
- long DNA molecules are partitioned into >1M individual reactions each containing a unique barcode
- Short reads from each partition have the same barcode
- Libraries maintain haplotype and other long-range information
- The resulting datatype is called Linked-Reads.



SINGLE MOLECULE REAL TIME SEQUENCING (SMRT)

Key Points

- Sequencing by synthesis
- Terminator free technology
- fluorescent labeled phosphate chain
- Uses DNA polymerase
- Read length ~ 15 kbp
- Individual reads have a substantial sequencing error (~15%)
- Optimal for repeat resolution and scaffolding



SMRT – LIBRARY PREPARATION

Key Points

•library of overlapping inserts

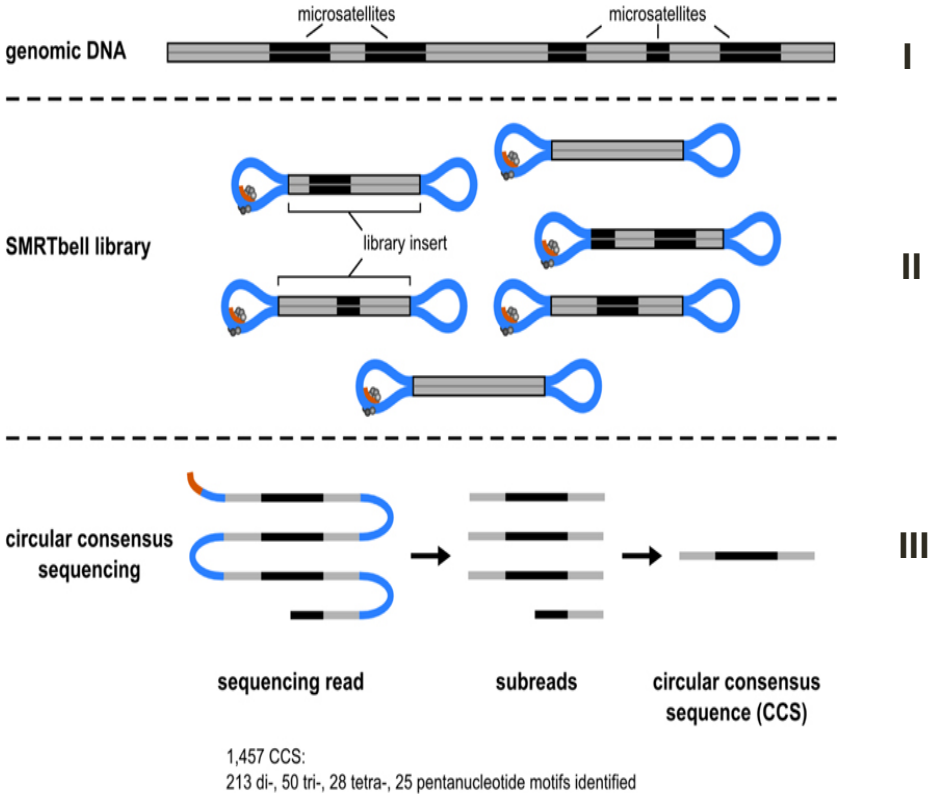
• hairpin adaptors create a circular molecule

• adaptors contain binding site for DNA polymerase

• sequencing results in a long sequencing read

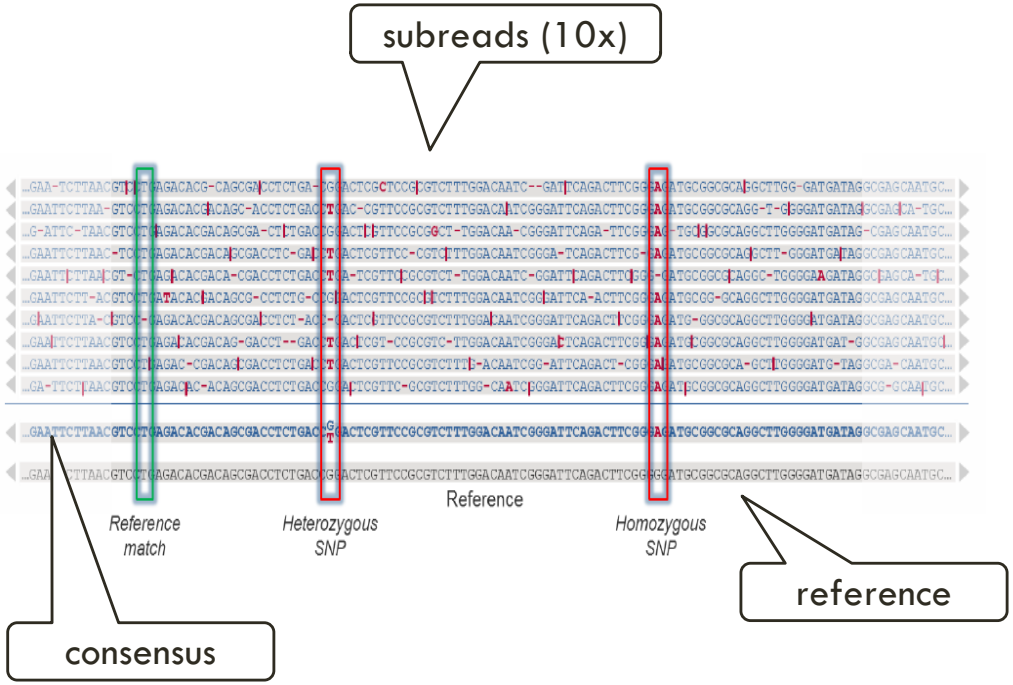
• can generate multiple subreads from one template

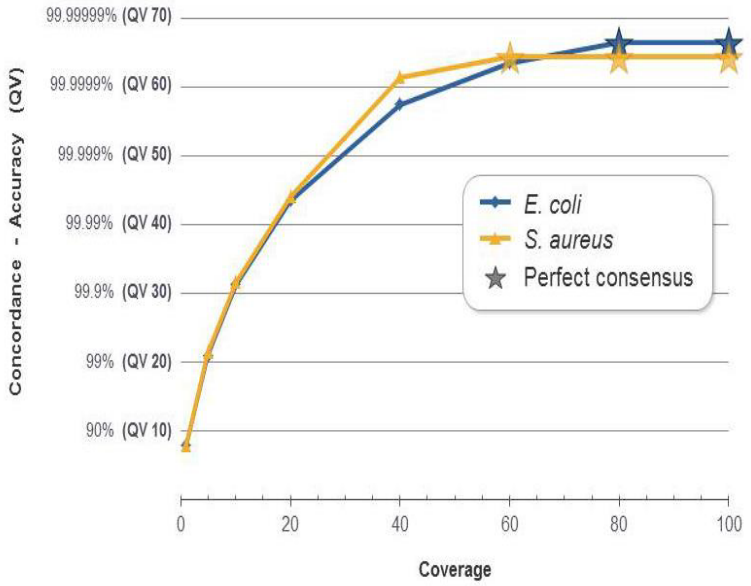
• combine subreads to create circular consensus read



* single read accuracy ~85%

SMRT – CIRCULAR CONSENSUS SEQUENCES INCREASE SEQUENCING ACCURACY



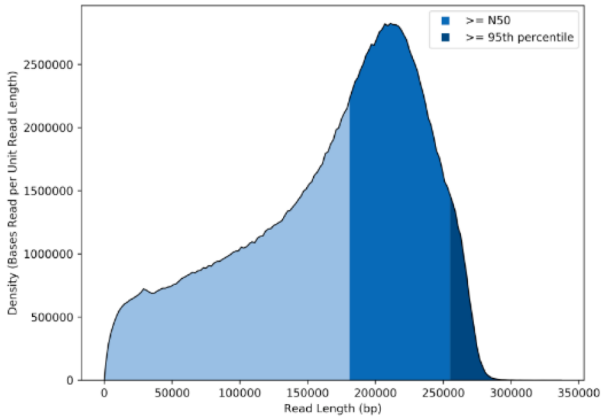


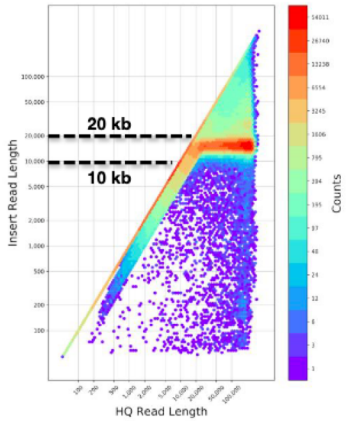
* QV = phred quality values

PACBIO HIFI PROTOCOL USES CCS FOR GENERATING HIGH QUALITY SEQUENCES[1]

B. Primary Sequencing Performance Metrics for a 15 kb HiFi Express 2.0 Library (Sequel II System)

Sample	Name	Status	Movie Time (hours)	Pre-extension Time (hours)	Total Bases (Gb)	Unique Molecular Yield (Gb)	Read Length				Productivity		
							Polymerase		Longest Subread				
							Mean	N50	Mean	N50	P0	P1	P2
1	Frac_4 15 kb HiFi Library	Complete	30	2	392.68	56.88	91960	181775	14514	15649	44.6%	53.3%	2.1%

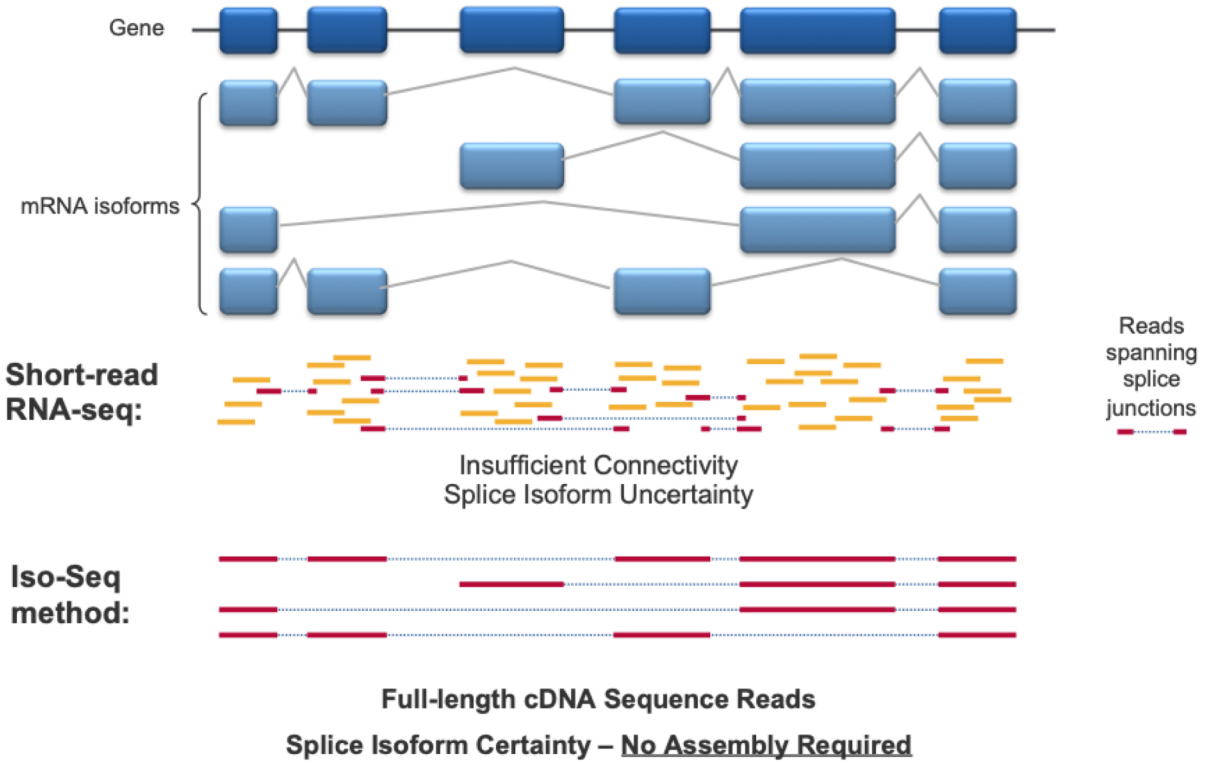




1 Sequencing error < 1%

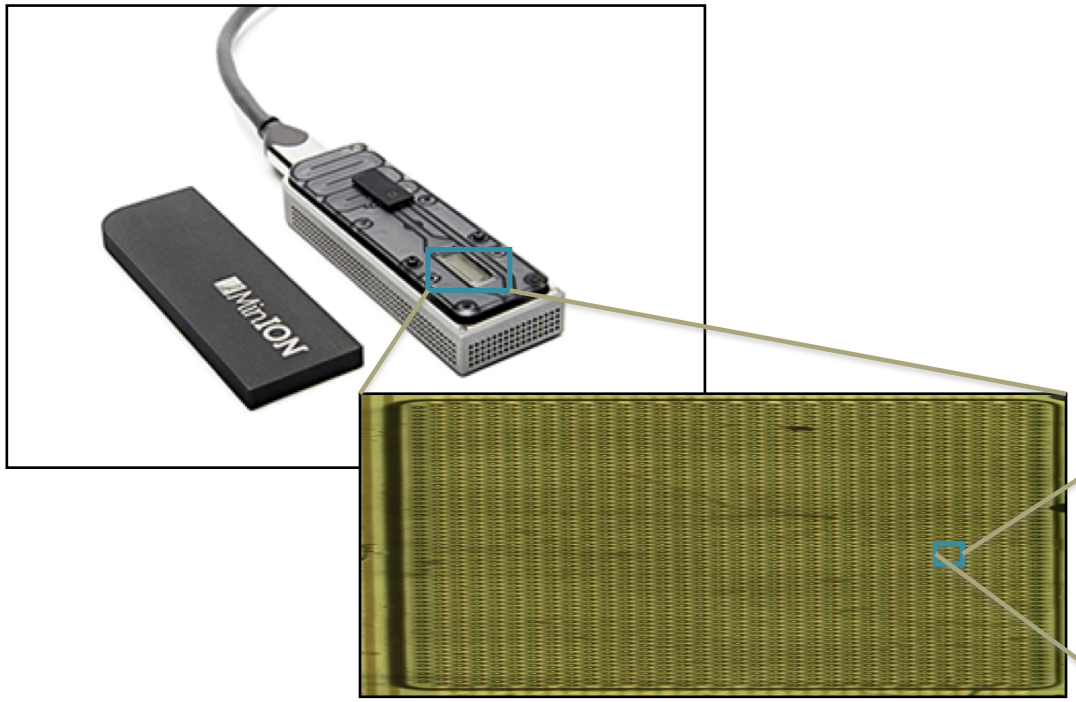
Slide source: https://www.pacb.com/wp-content/uploads/HiFi-Library-Preparation-Using-SMRTbell-Express-TPK-2.0-Customer-Training.pdf

FULL-LENGTH RNA SEQUENCING USING ISO-SEQ

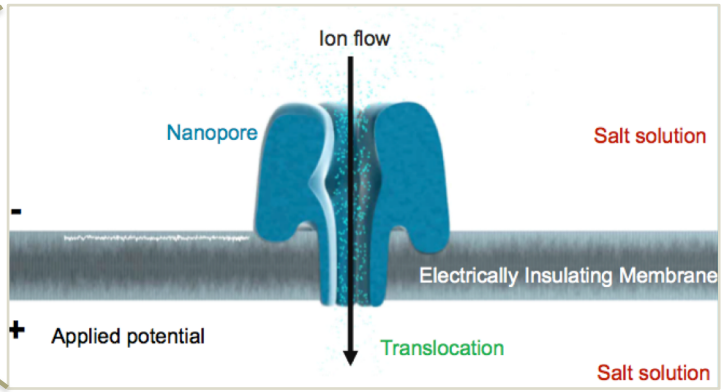


Source: https://www.pacb.com/wp-content/uploads/2018-10-NA-UGM-Iso-Seq-Method.pdf

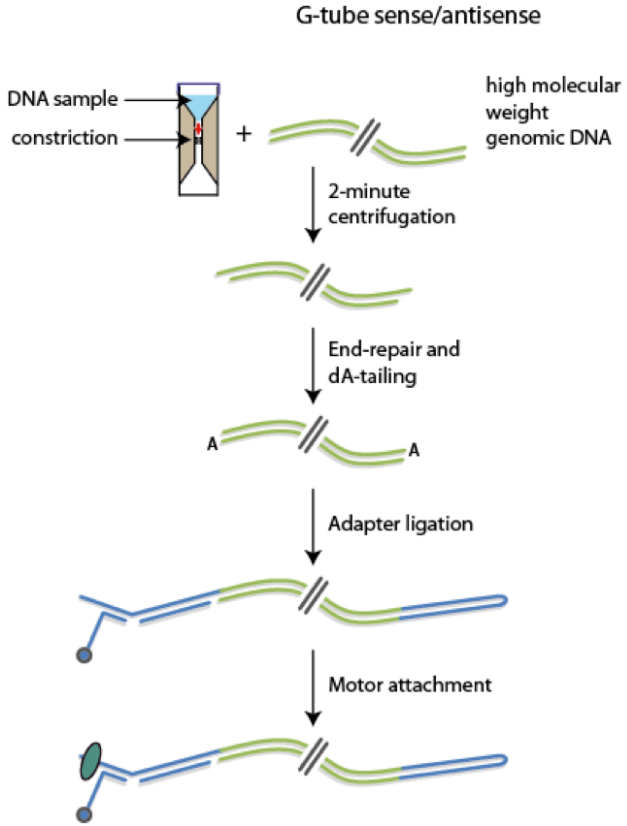
OXFORD NANOPORE SEQUENCING : SINGLE MOLECULE SEQUENCING

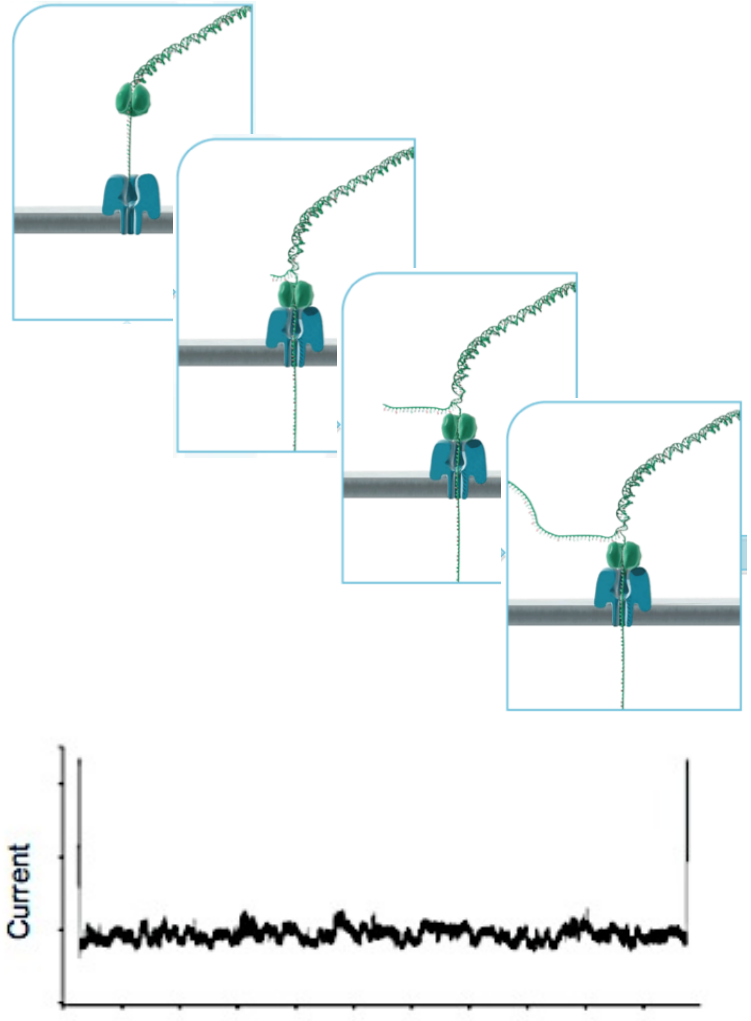


- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow

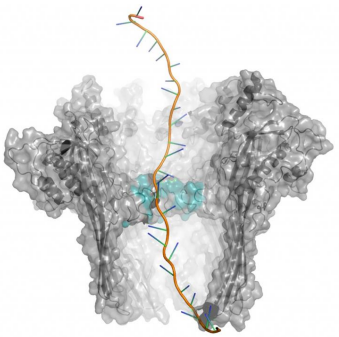


NANOPORE SEQUENCING

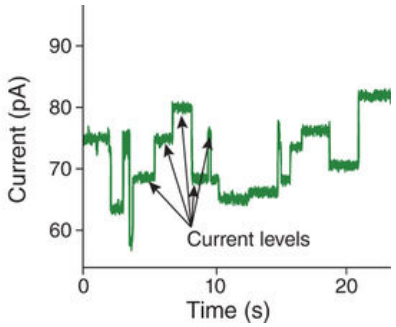




OXFORD NANOPORE SEQUENCING : SINGLE MOLECULE SEQUENCING



Oxford Nanopore Google Hangout March 2016



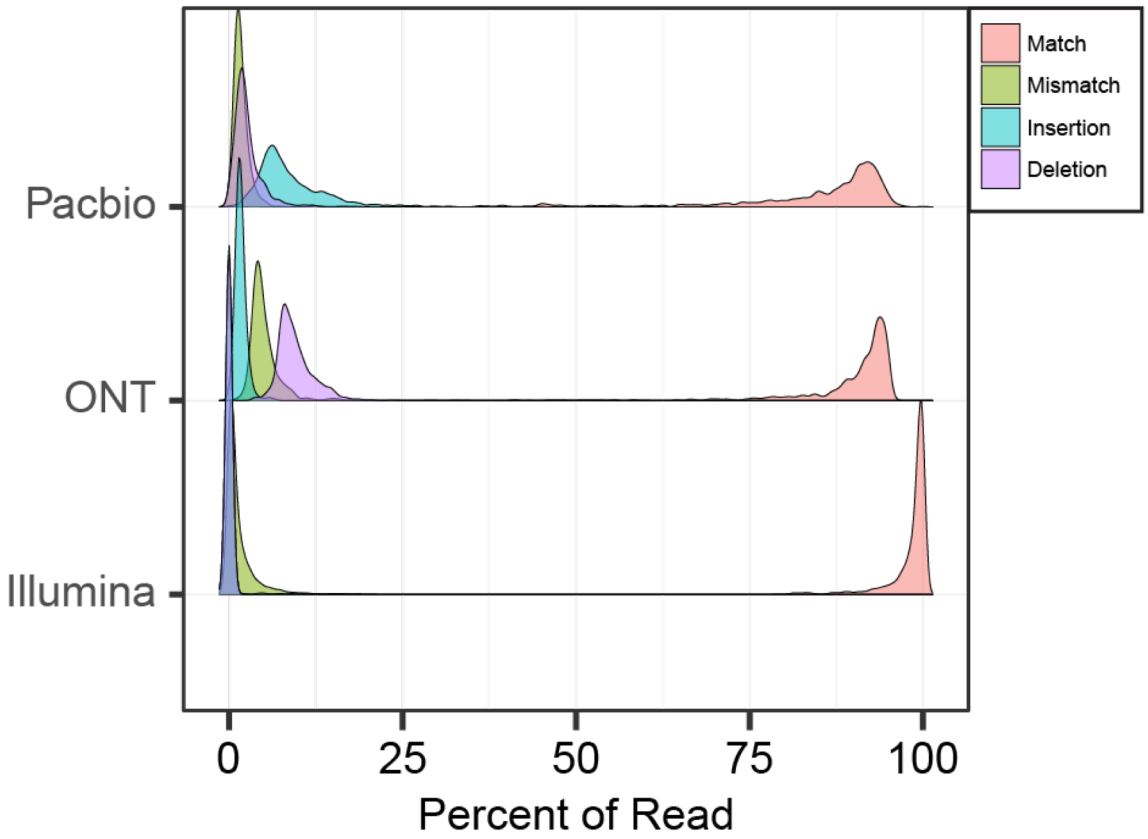
Deamer et al 2016, Nature Biotech

ATCGATCGATA
GTATTAGATAC
GACTAGCGAT
CAG

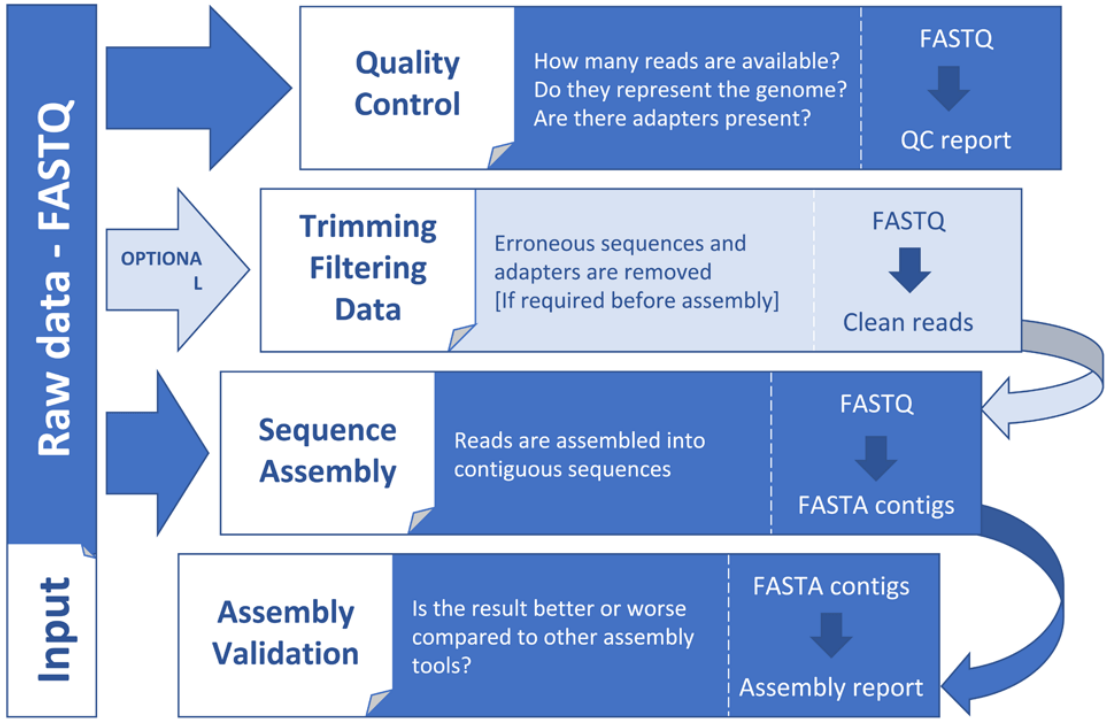
- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow
- Multiple base-pairs at a time (“k-mers”)
- Characteristic current signature is converted to nucleotide sequences
- No theoretical upper limit to sequencing read length, practical limit only in delivering DNA to the pore intact
- Predicted sequencing output 5-10Gb

SEQUENCING ERROR COMPARISON USING A REFERENCE BASED MAPPING OF E. COLI SEQUENCING DATA

- Map Illuminan reads with bowtie2 and minimap2 for Pacbio and ONT alignment, use samtools to compare to reference.
- Illumina reads align almost perfectly, with a per read median of 99.3% correct. Indels almost never occur.
- PacBio reads which have an median of 89.2% of the read correct. Most frequent error type: insertions (7.45% median) with mismatches only 1.5% median % of read.
- ONT reads have a per read median of 92.4% correct, with deletions (9%) and mismatches (4.5%) both at a relatively high median per read.



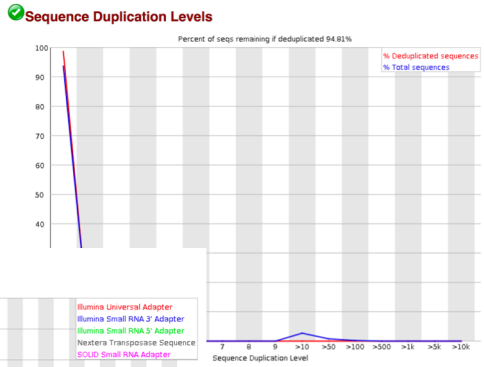
PRE-PROCESSING OF DNA SEQUENCING READS IS THE FIRST STEP OF DATA ANALYSIS

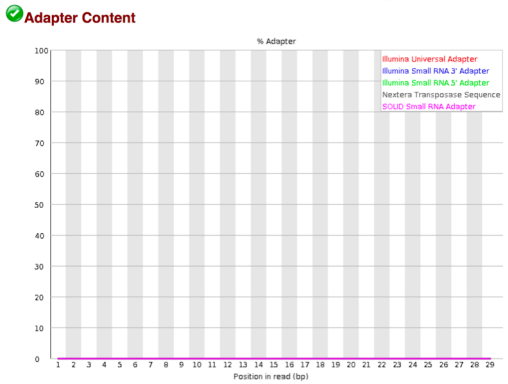


FASTQC AND *CUTADAPT* HELP TO VISUALIZE AND CLEAN HIGH THROUGHPUT DNA SEQUENCING DATA

Sequence file (FastQ Format)

FASTQC



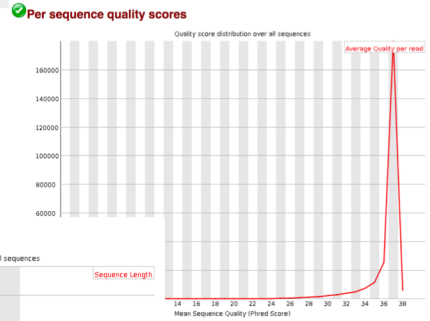


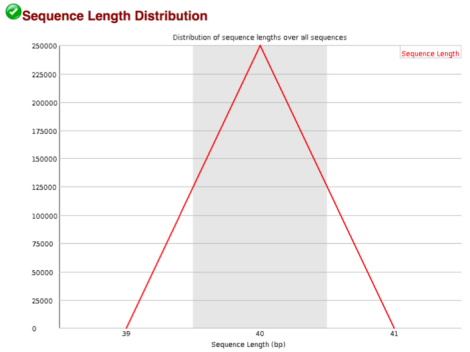
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

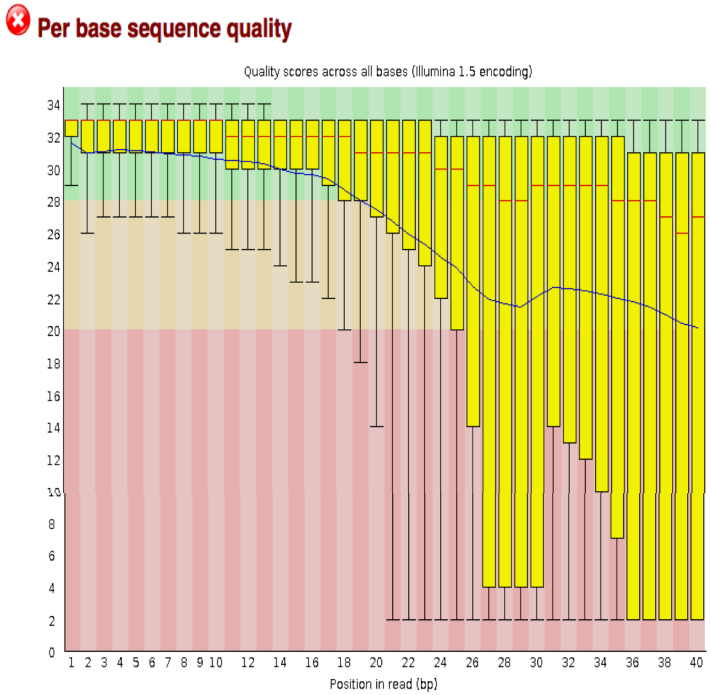
Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45





REMEMBER - SEQUENCING READS ARE NOT ERROR FREE



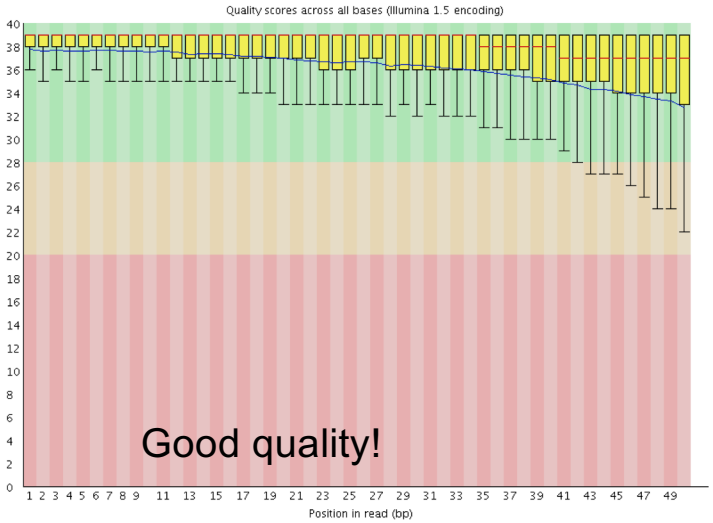
Base quality distribution for an Illumina short read run. The plot shows the typical decay of average base quality towards the end of the reads

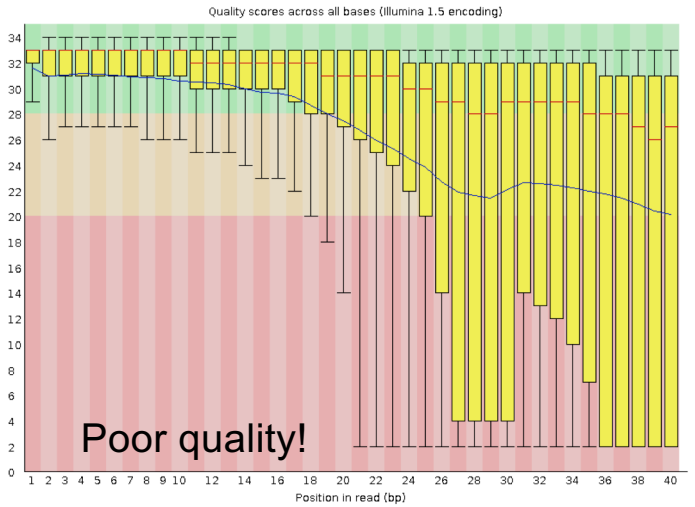
There are two main error sources

- Misinterpretation of the signal by the basecaller. This type of error results in low base qualities
- PCR error during template preparation and/or amplification. This error can result in high quality but wrong base calls[1]

1 remember, each sequencing reaction starts from a single molecule. Early PCR errors will be propagated to (almost) all amplificaiton products

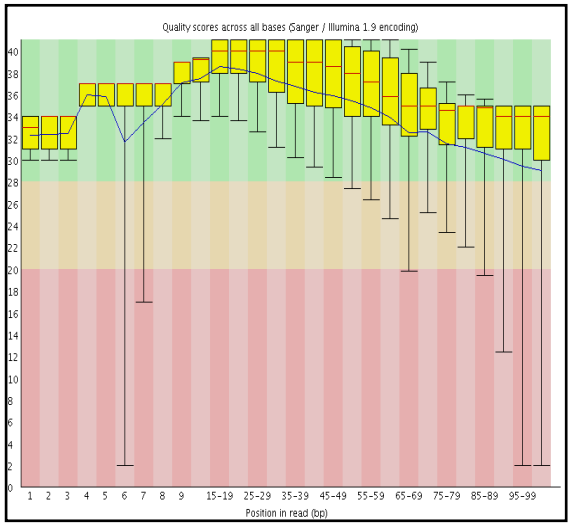
REMEMBER - SEQUENCING READS ARE NOT ERROR FREE



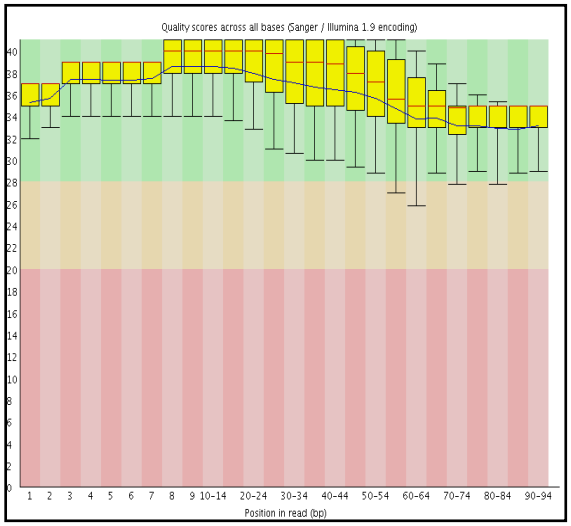


QUALITY TRIMMING USING TRIMMOMATIC

Before quality trimming

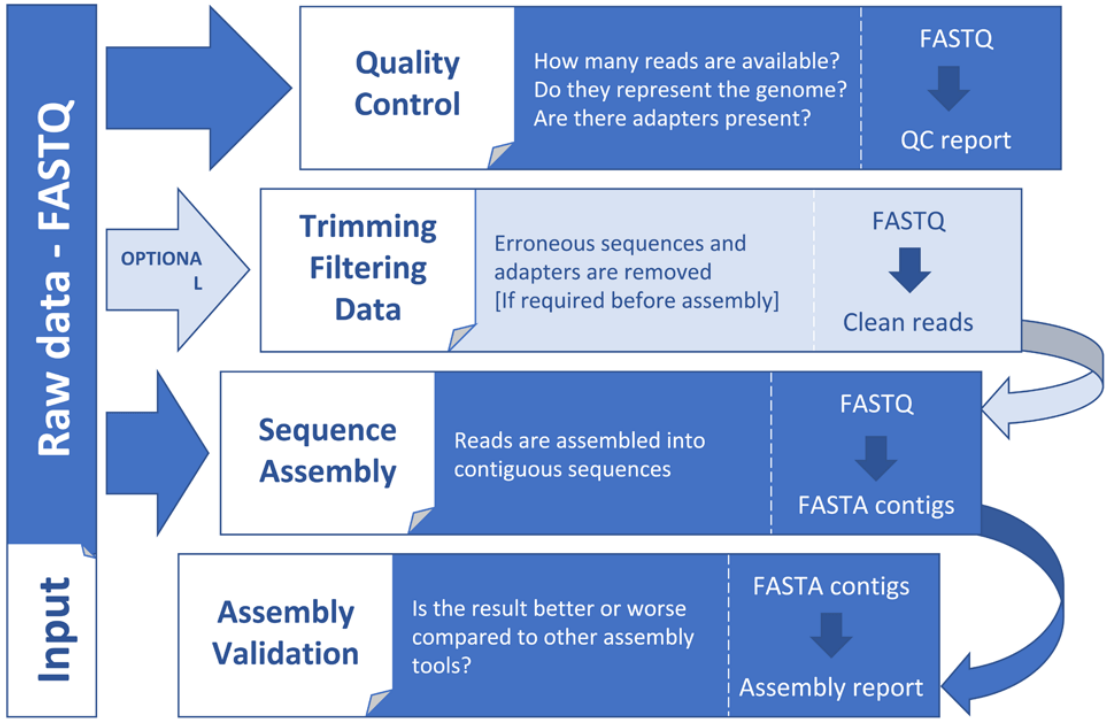


After quality trimming



http://www.usadellab.org/cms/?page=trimmomatic

RECONSTRUCTING THE TEMPLATE SEQUENCE – THE ASSEMBLY OF SEQUENCING READS

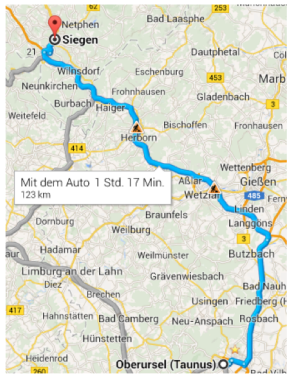


HOW BIG A PROBLEM IS SEQUENCE ASSEMBLY?

Human Chr 8: 146,000,000 bp

Method	Approach	Real-time	Read-length	Bp per run	# of runs for 10x coverage
Illumina	Sequencing by synthesis	Yes	125 bp	1000 Mb	~0.02

150,000,000 Km



In fact, the problem is at least 2 orders of magnitude larger since:

* The entire human genome consists of approx. 3.2 Billion base pairs
* 1-fold coverage is not sufficient. Typically at least 10 x coverage* should be achieved. Thus, we need to sequence 32 Billion base pairs.

*ca 80x required for short read sequencer