Introduction to Illumina NGS technology

Ester Feldmesser

12.11.19

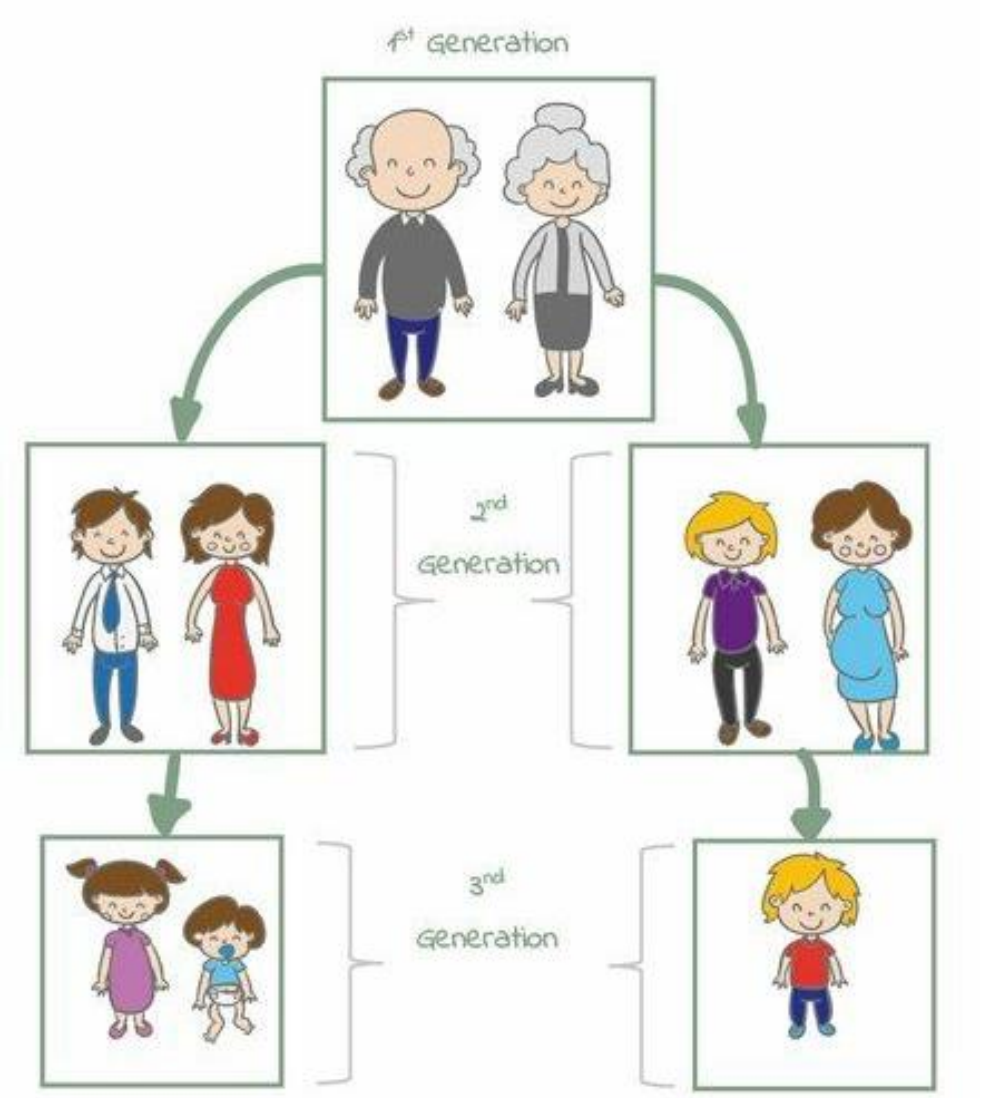
An Introduction to deep-sequencing analysis for biologists

What is sequencing?

DNA sequencing is the process of determining the sequence of nucleotide bases (As, Ts, Cs, and Gs) in a piece of DNA. Today, with the right equipment and materials, sequencing a short piece of DNA is relatively straightforward.

Why do we sequence?

Generations of sequencing



Sanger sequencing

Illumina sequencing
Short read sequencing

Single molecule sequencing
Long read sequencing

Ingredients for Sanger sequencing

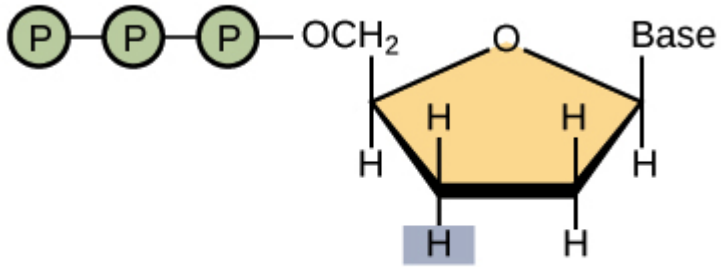
Sanger sequencing involves making many copies of a target DNA region.
New molecules of DNA are **synthesized**.
Its ingredients are similar to those needed for DNA replication in an organism.

Ingredients for Sanger sequencing

Sanger sequencing involves making many copies of a target DNA region.
New molecules of DNA are **synthesized**.
Its ingredients are similar to those needed for DNA replication in an organism.

They include:

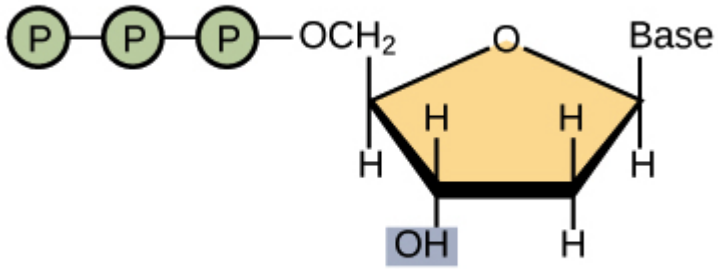
- A DNA polymerase enzyme
- A **primer**
- The four DNA nucleotides (dATP, dTTP, dCTP, dGTP)
- The template DNA to be sequenced



Dideoxynucleotide (ddNTP)

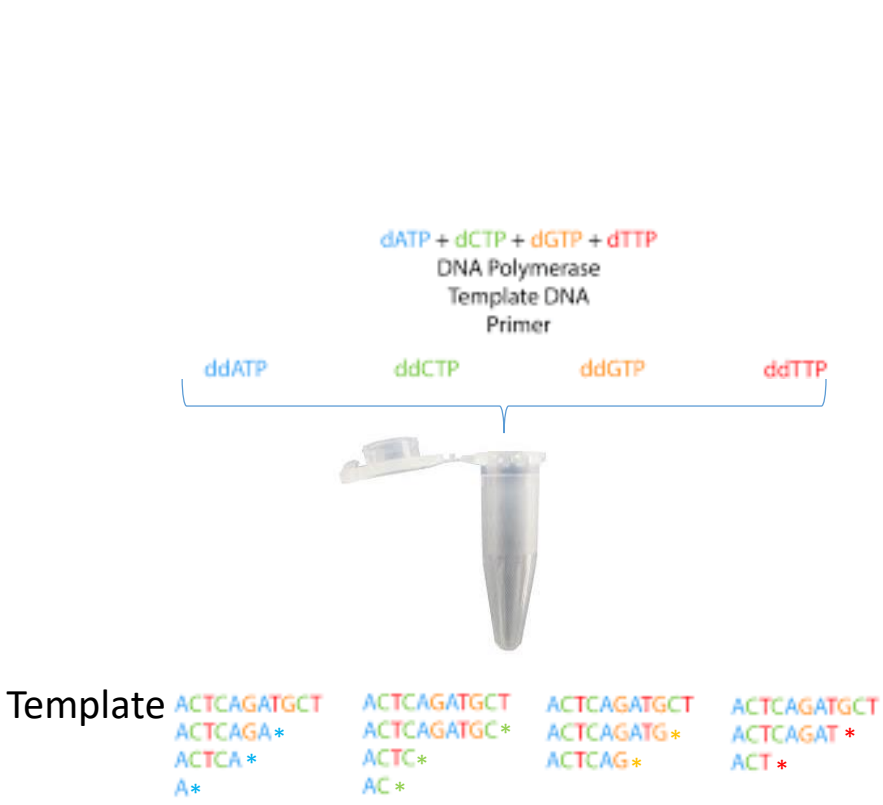
A unique ingredient:

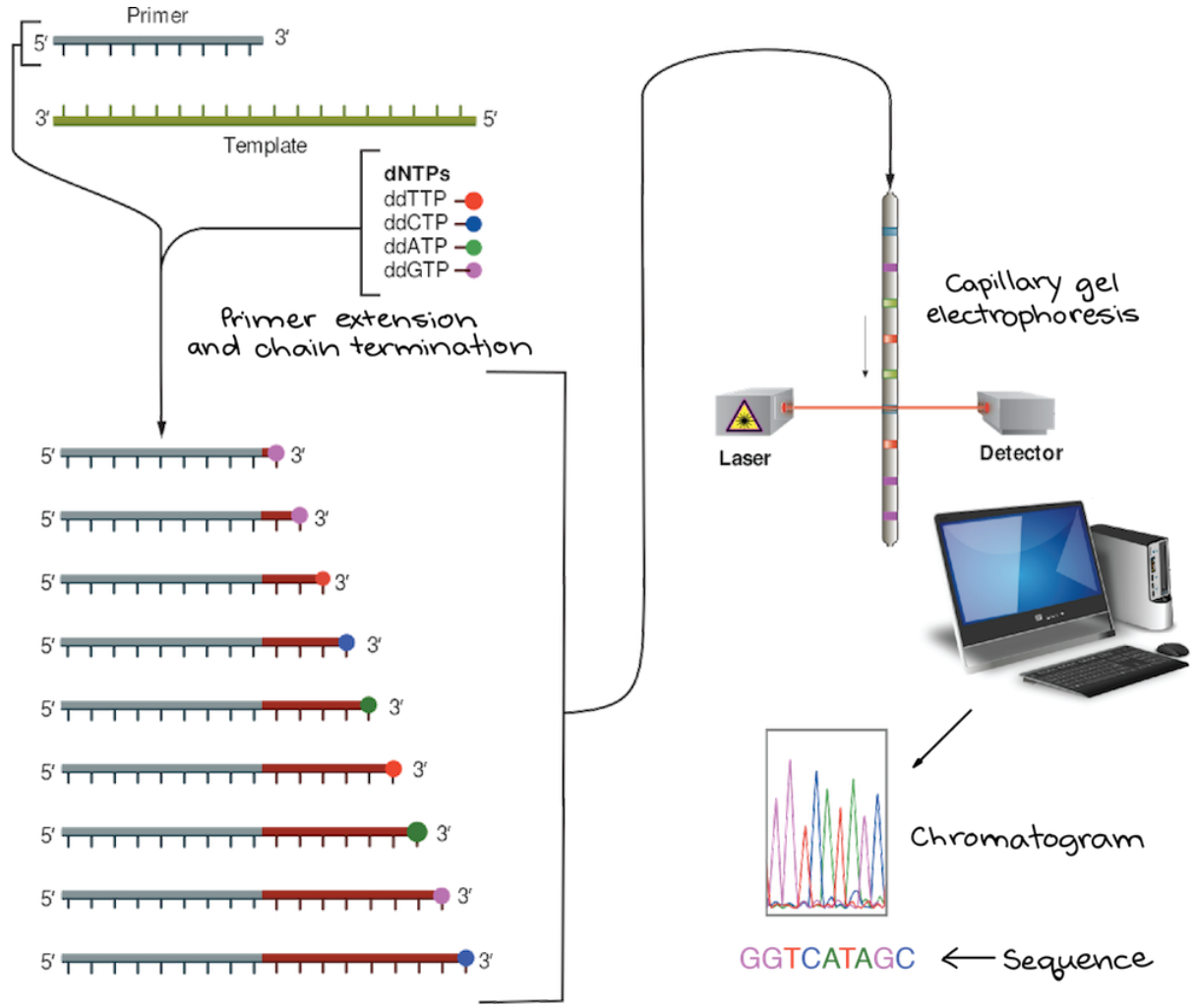
- Dideoxy, or **chain-terminating**, versions of all four nucleotides (ddATP, ddTTP, ddCTP, ddGTP), each labeled with a different color of dye



Deoxynucleotide (dNTP)

Sanger sequencing





https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-biotechnology/a/dna-sequencing

How to scale up to millions of sequences?

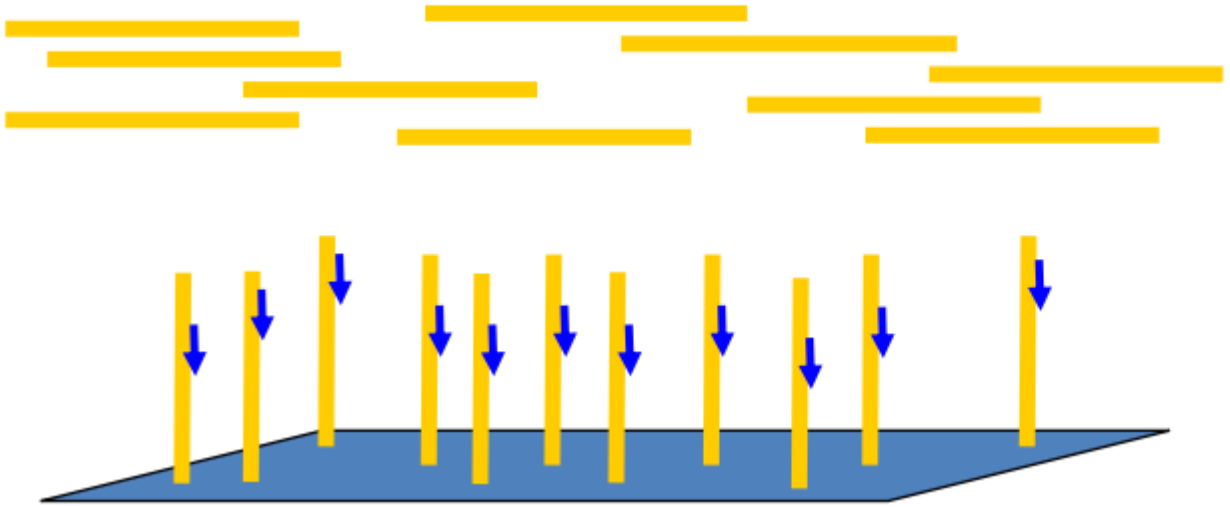
- Problems

- Ideas how to do it

Hint: Do not use ddNTPs. We do not want to run gels or capillaries.

Requirements for Illumina NGS

1. Keep track of each of sequence independently by determining a physical location for each sequence: attachment to a solid surface.



Complex DNA sample

↓

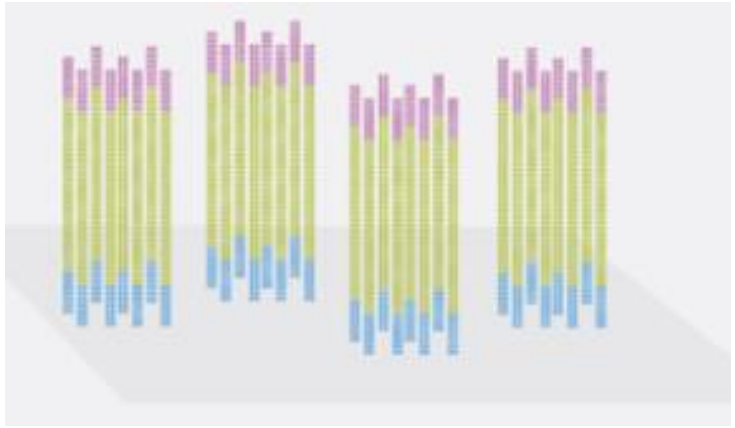
Attachment to solid surface

↓

Parallel Sequencing of all DNA fragments

2. Need to detect the base that we are adding like in Sanger.

Each independent DNA molecule will attach to the solid surface and generate a cluster of identical sequence – similar to bacterial colonies



Illumina Sequencing Workflow

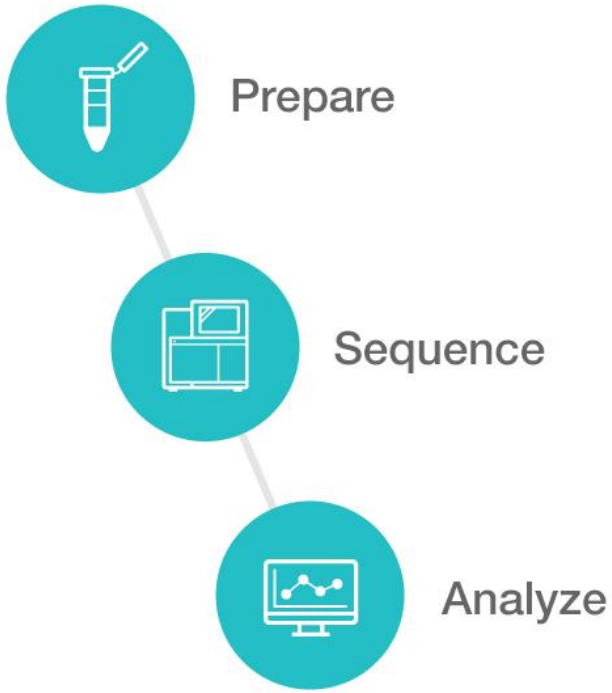
Biological experiment ➡ Designed to answer a biological question

Library preparation ➡ Preparation of DNA/RNA to be loaded on Flow Cell

Cluster generation ➡ Attachment of sample to solid surface and amplification

Sequencing ➡ Simultaneous sequencing of complex library

Data analysis ➡ Align or assemble reads, quantify and more…

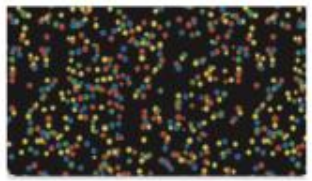


Illumina Workflow

Sample preparation

Cluster generation

Sequencing by synthesis



Data analysis

Sample (library) Preparation

1. DNA shearing

- Physical shearing, random, 200-500bp fragments

2. End repair

- Blunt ends using polymerase

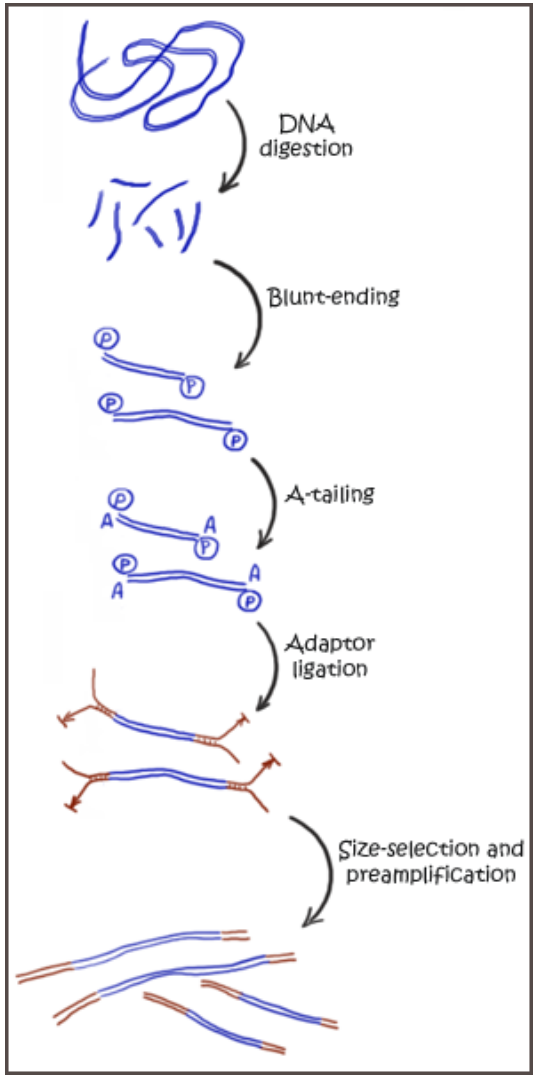
3. A-tailing

- Add an A base on the 3' a overhang to allow primer ligation and prevent self-ligation

4. Adaptor ligation

- To allow pre-amplification, attachment to Flow-Cell and sequencing

5. Pre-amplification



From A. Soldatov, MPI

The library consists of millions of different DNA molecules.

Library in depth

- Structure of Illumina libraries
- Single read SR or Paired end (PE)
- Single Index (barcode) or Dual Index
- Other library types

Illumina libraries

Single End Sequencing

Paired End Sequencing

The indices have additional primers

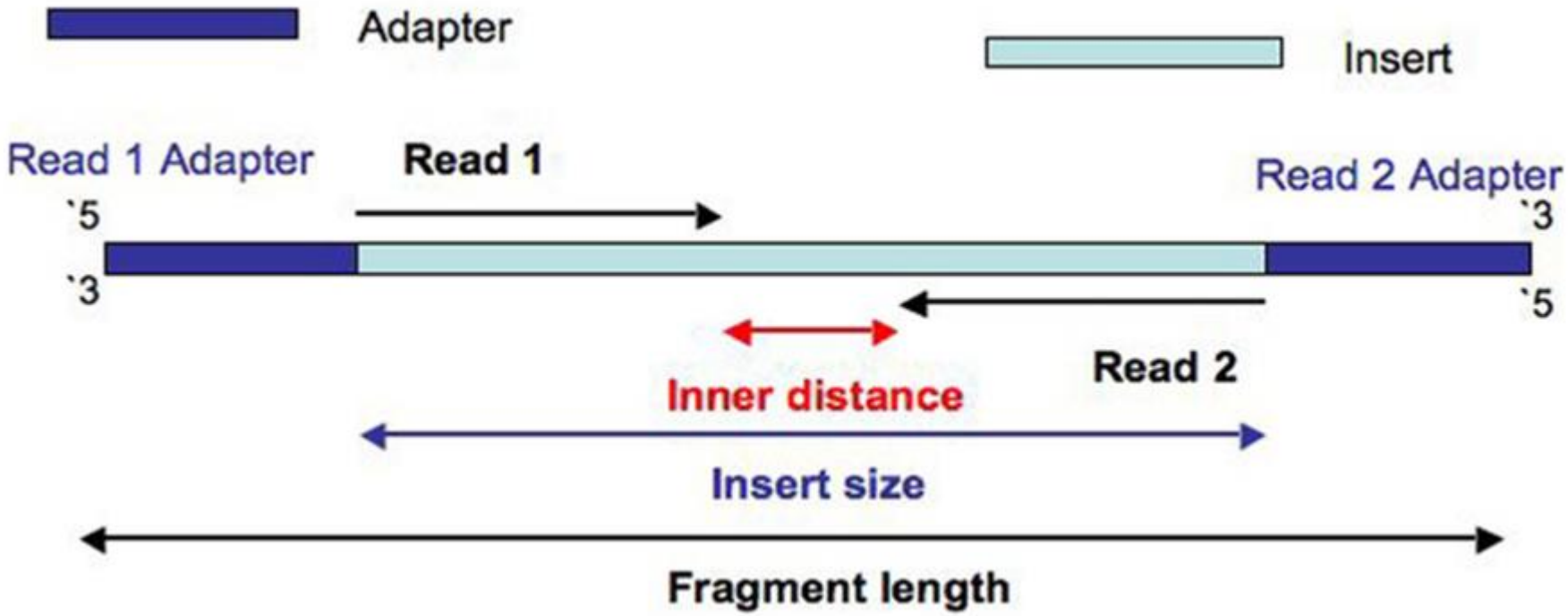
Illumina libraries

Single End Sequencing

Paired End Sequencing

The indices have additional primers

Some terminology…

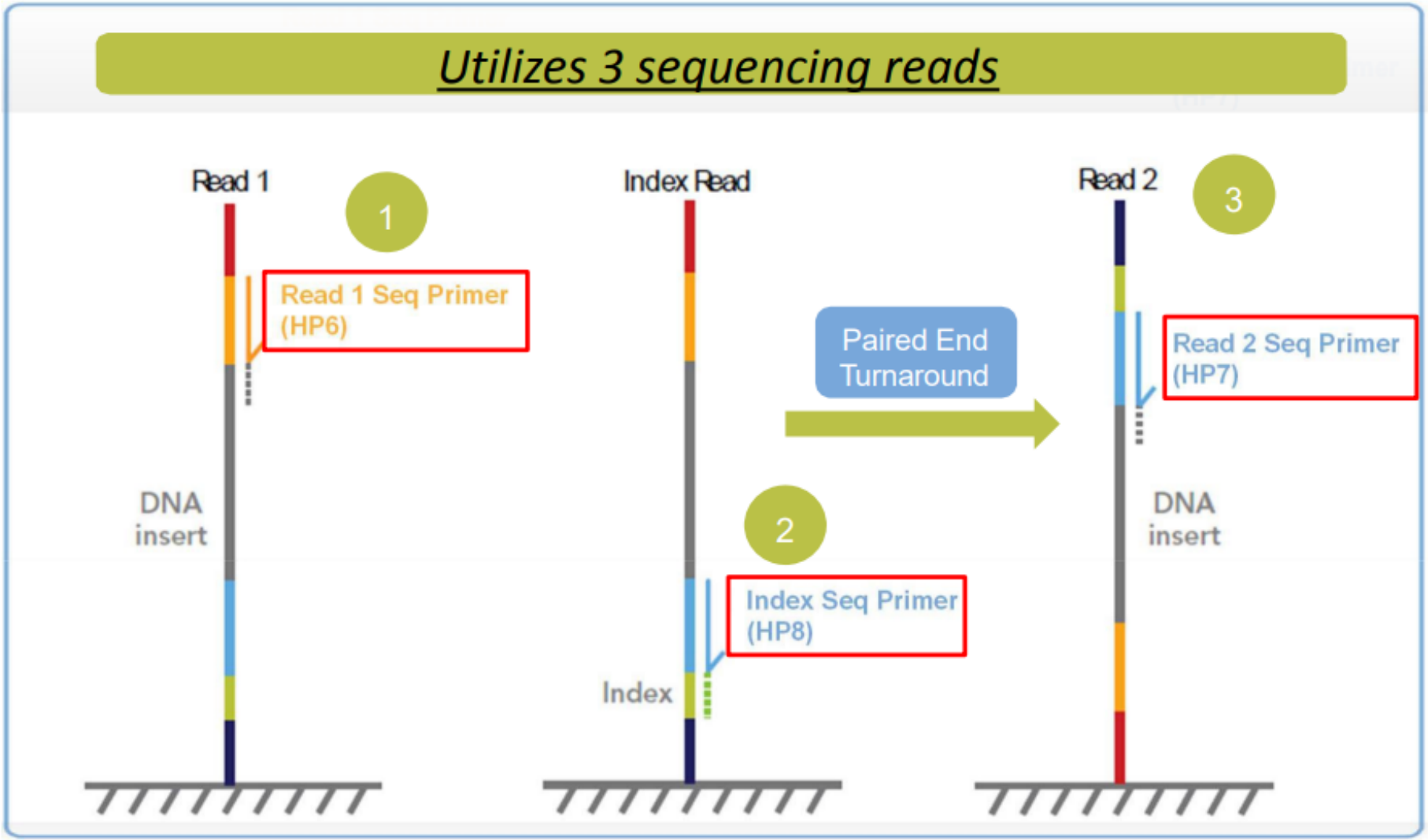


Paired End Sequencing

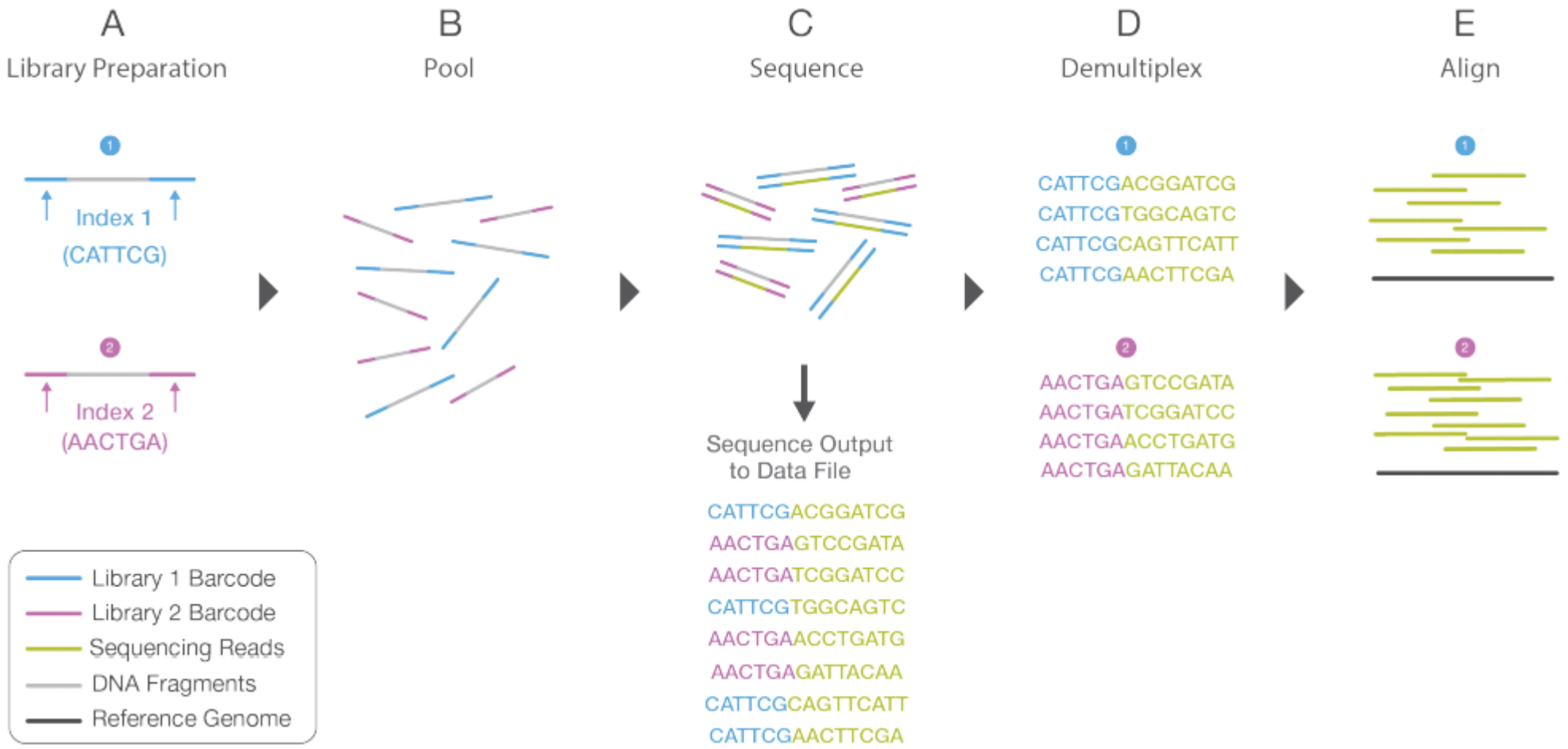
Reference	This is really the best way to do sequencing
Single-reads	This is
…	is really
…	really the
…	the best
…	sequencing
Paired-reads	This is (----100 characters-------) sequencing

Assembly and mapping become easier

Paired End Sequencing of Single-indexed libraries

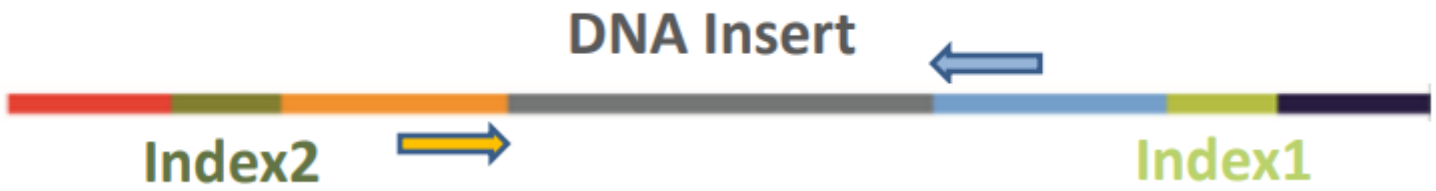


Library Multiplexing Overview



(A) Unique index sequences are added to two different libraries 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.

Sequencing Paired End Libraries with Dual Index Read



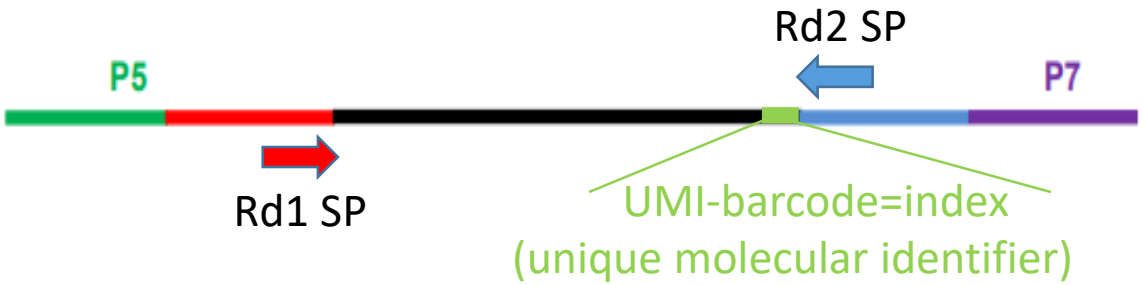
Utilizes 4 Sequencing reads:

- Read 1
- Index Read 1 (i7)
- Index Read 2 (i5)
- Read2

Other types of libraries

- Non-compatible Illumina libraries have indices in a different location than the standard Illumina.
- They cannot be sequenced together because the required read definition is different.
- The automatic demultiplexing pipeline cannot deal at the same time (in one run) with libraries that have indices in different locations. It can deal with each library type in separate runs.

Mars-seq

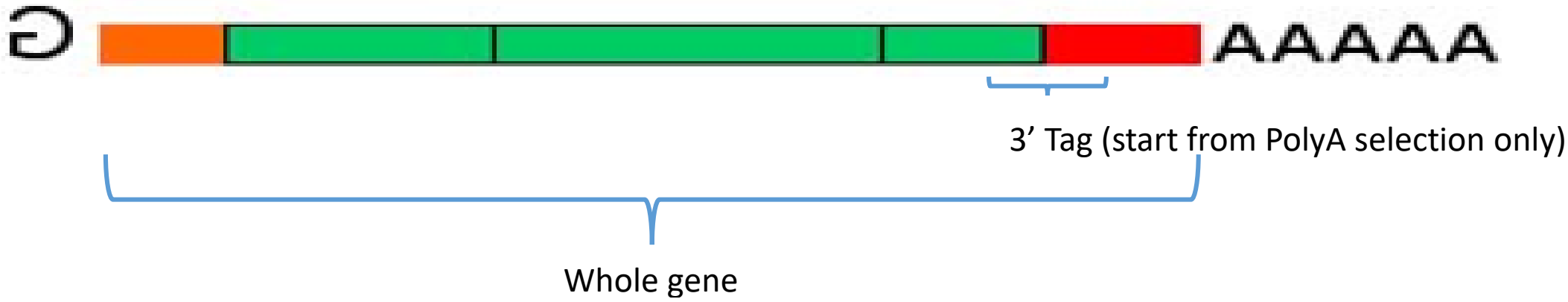


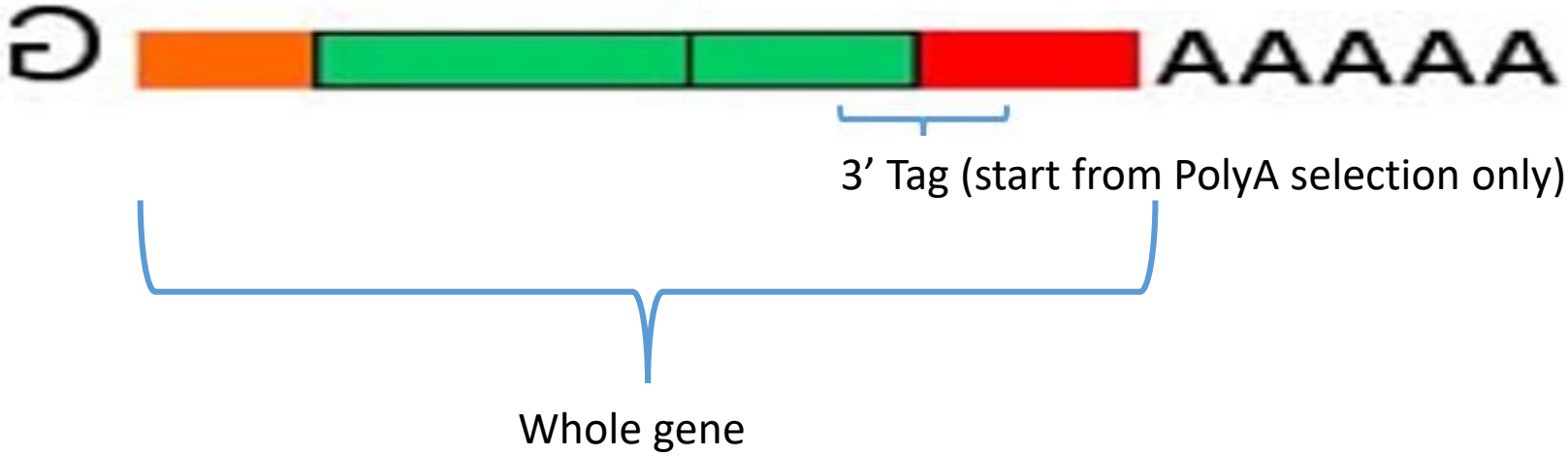
- Library generation for 3’ RNA seq
- Developed in the lab of Ido Amit
- Low input material (1 ng of RNA)
- It is cheap

What do you sequence from each primer?

Rd1 SP – a tag near the 3’ of your gene of interest.
Rd2 SP – the barcode and the UMI

Whole gene versus tag





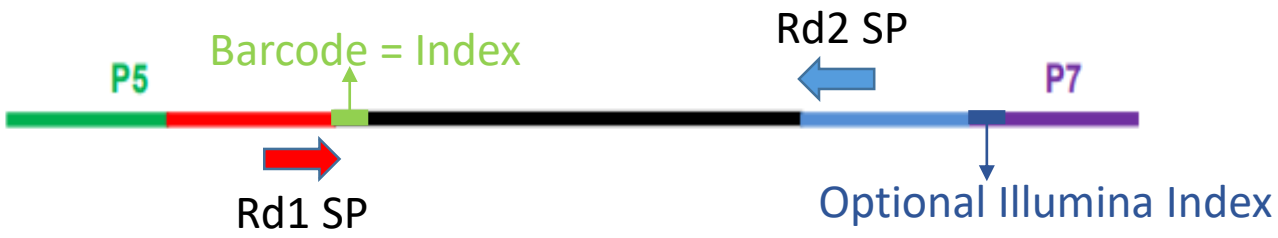
Dual sequence library

A highly multiplexed and sensitive RNA-seq protocol for simultaneous analysis of host and pathogen transcriptomes

Roi Avraham, Nathan Haseley, Amy Fan, Zohar Bloom-Ackermann, Jonathan Livny & Deborah T Hung

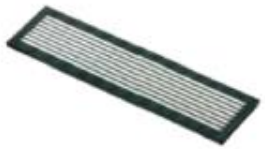
Nature Protocols **11**, 1477–1491 (2016) | Download Citation

Goal: To simultaneously characterize the bacterial and host expression programs during infection



Rd1 SP – the barcode and read1 of your gene of interest
Rd2 SP – read1 of your gene of interest
Index Read 1 (i7) – if you use the optional Illumina Index

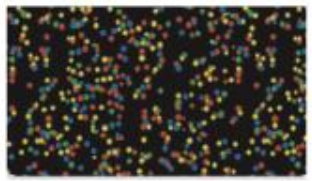
Illumina Workflow



Sample preparation

Cluster generation

Sequencing by synthesis



Data analysis

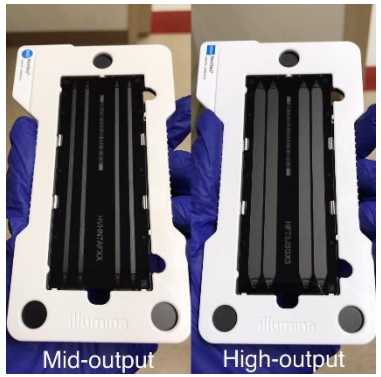
Attachment to a flow cell

NextSeq

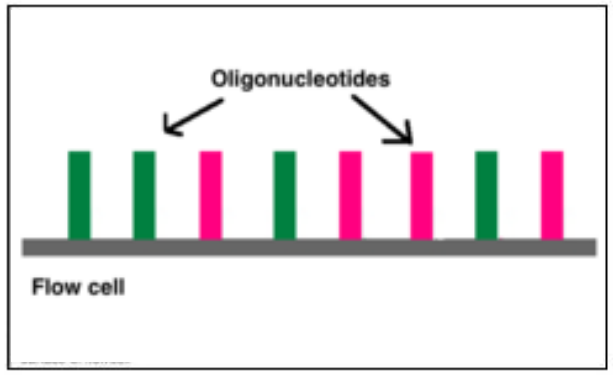
A flow cell is a thick glass slide with channels or lanes

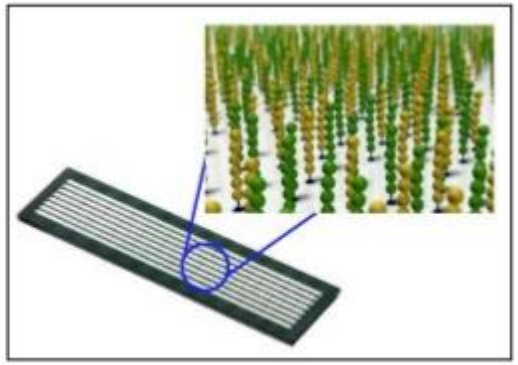
NovaSeq

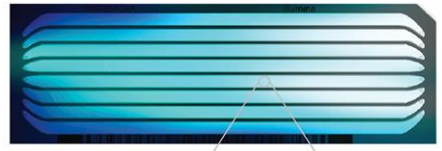
- Oligonucleotides attached to flow cell hybridize to the adaptors
- Individual DNA library fragments are immobilized



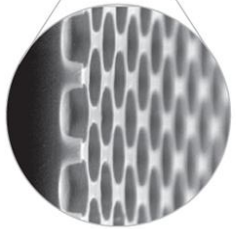
Random



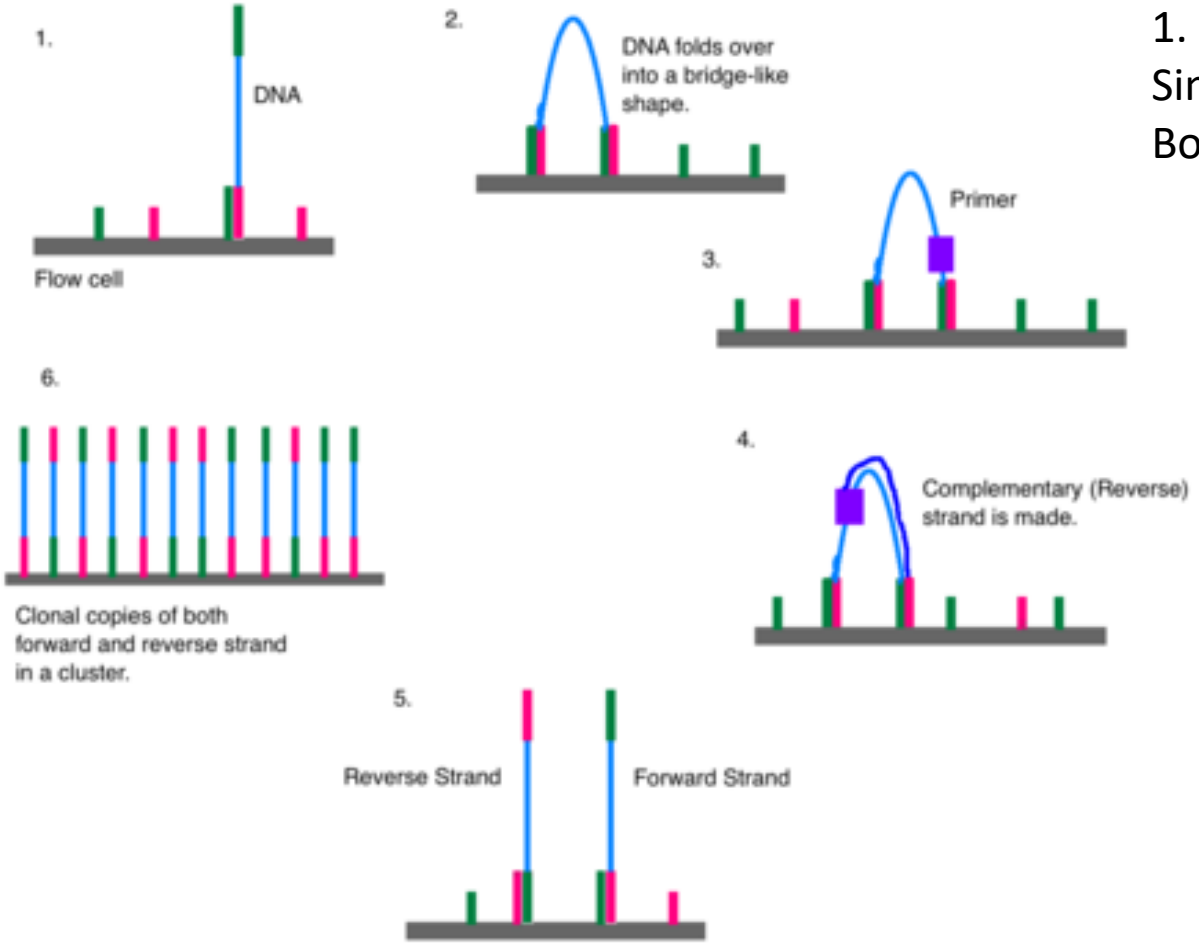




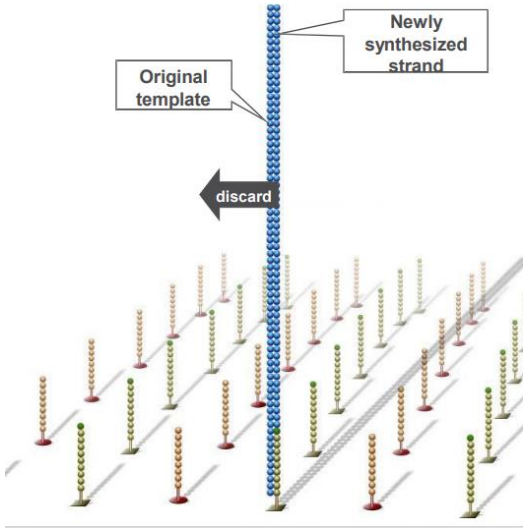
Patterned



Cluster Generation (random flow cell)



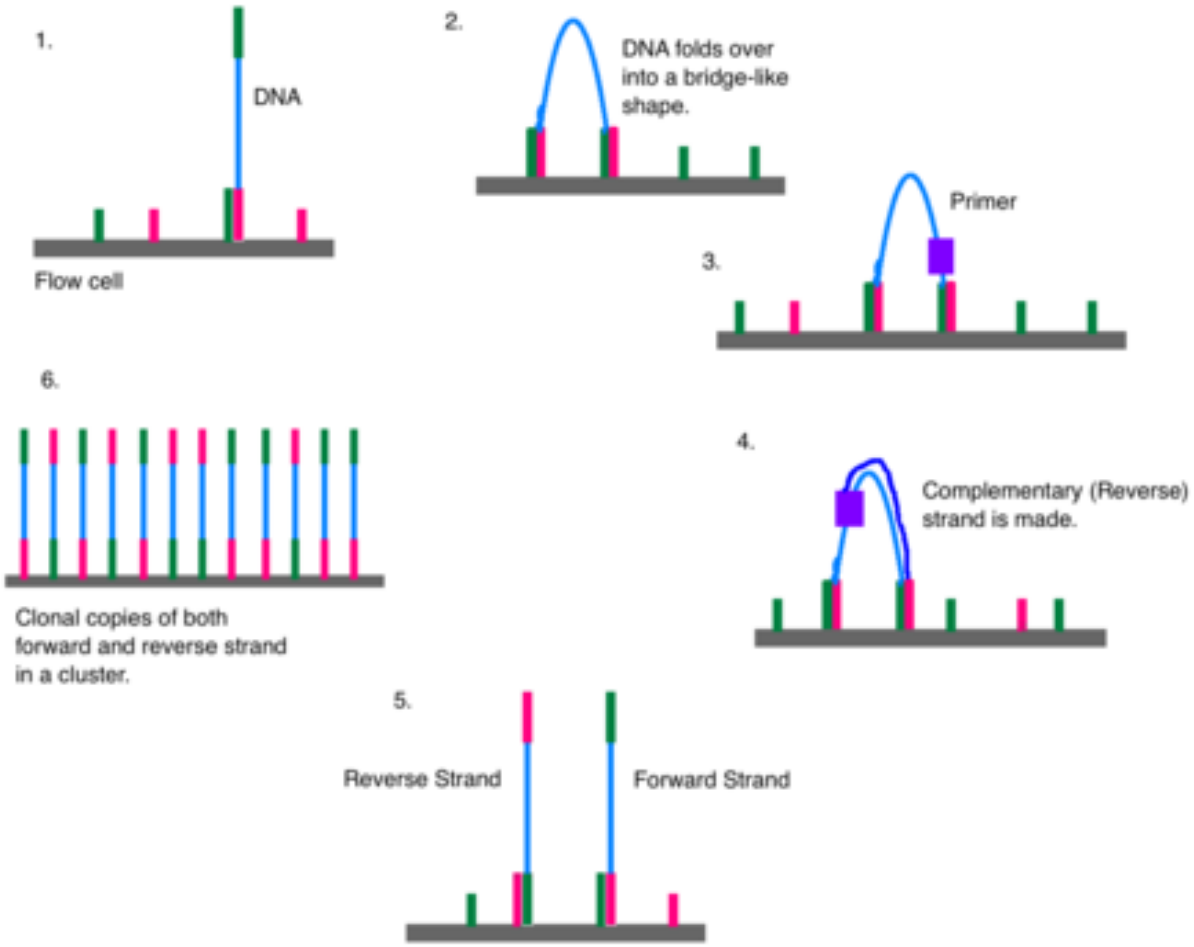
1.
Single strand DNA libraries are hybridized to primer lawn
Bound libraries are then extended by polymerases



Original template washed away
Newly synthesized strand is covalently attached to flow cell surface

https://en.wikipedia.org/wiki/Illumina_dye_sequencing

Cluster Generation



2-6. Bridge Amplification

Single-stranded molecules flip over to hybridize to adjacent primers

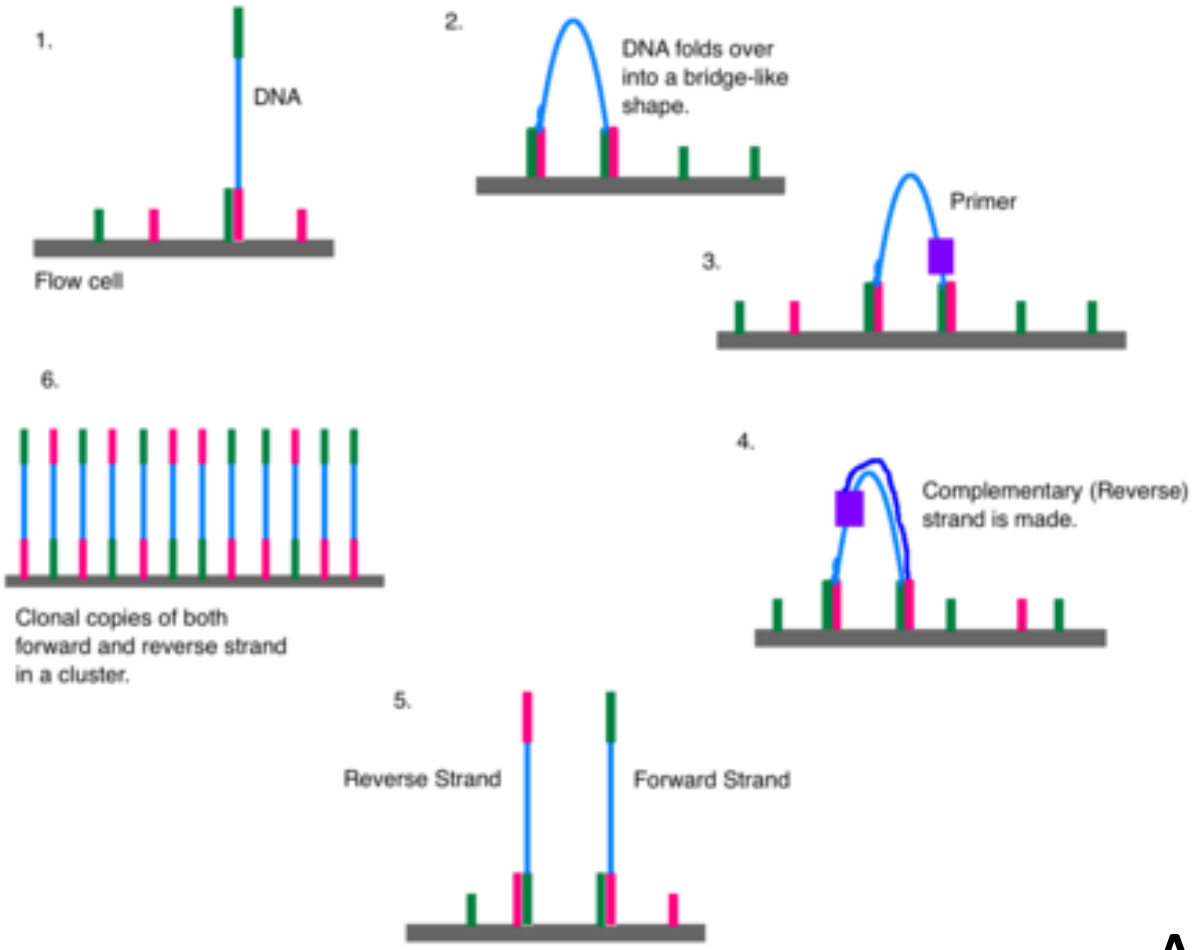
Hybridized primer is extended by polymerase

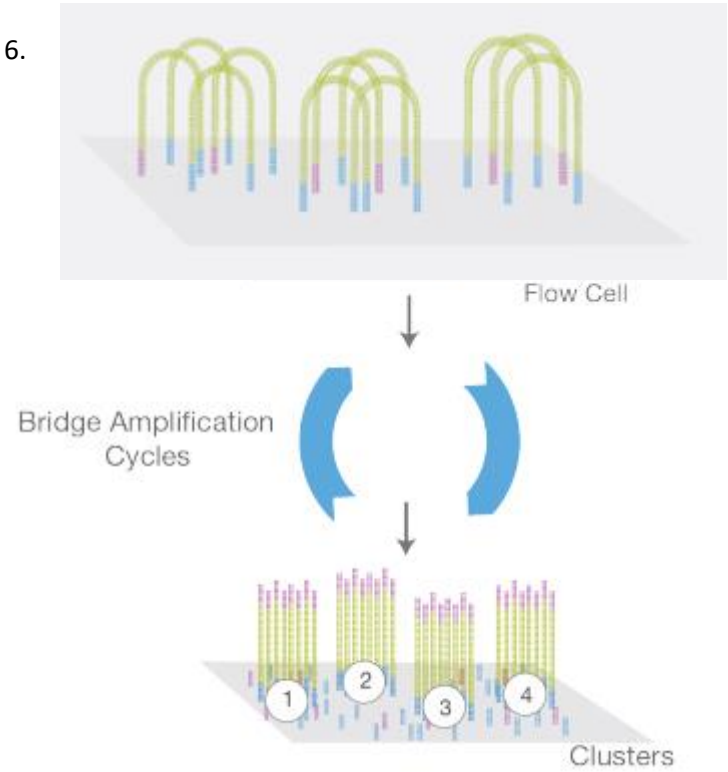
Bridge amplification cycle is repeated until multiple bridges are formed

dsDNA bridges are denatured

https://en.wikipedia.org/wiki/Illumina_dye_sequencing

Cluster Generation





Bridge amplification cycle is repeated until multiple bridges are formed

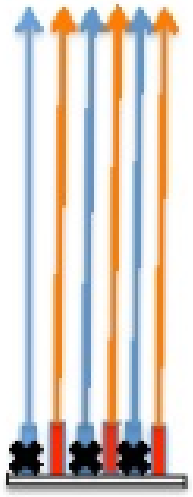
Clusters remain distinct – do not mix, so that the software can distinguish each cluster

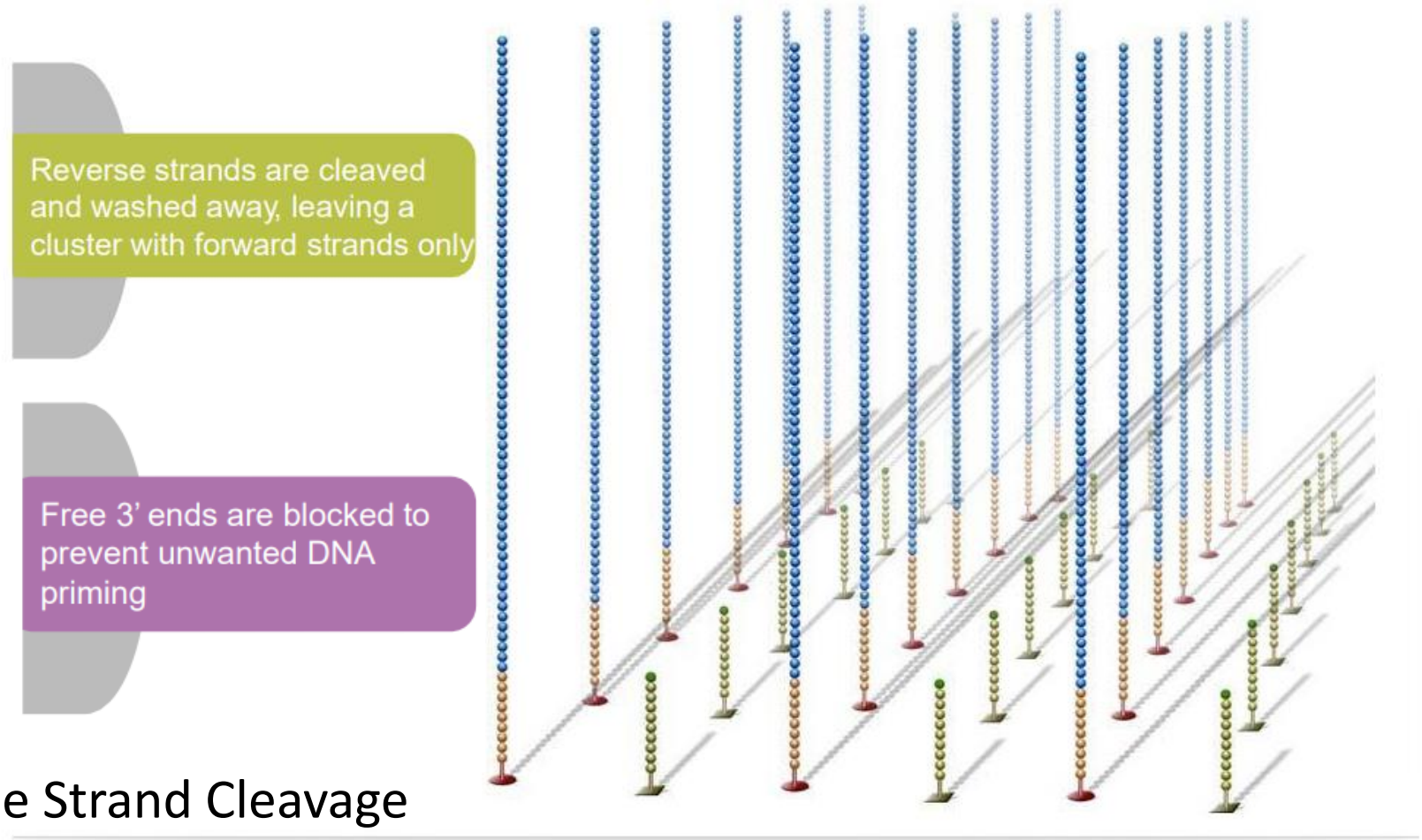
Are we ready for sequencing?

https://en.wikipedia.org/wiki/Illumina_dye_sequencing

Cluster Generation: Just before Sequencing

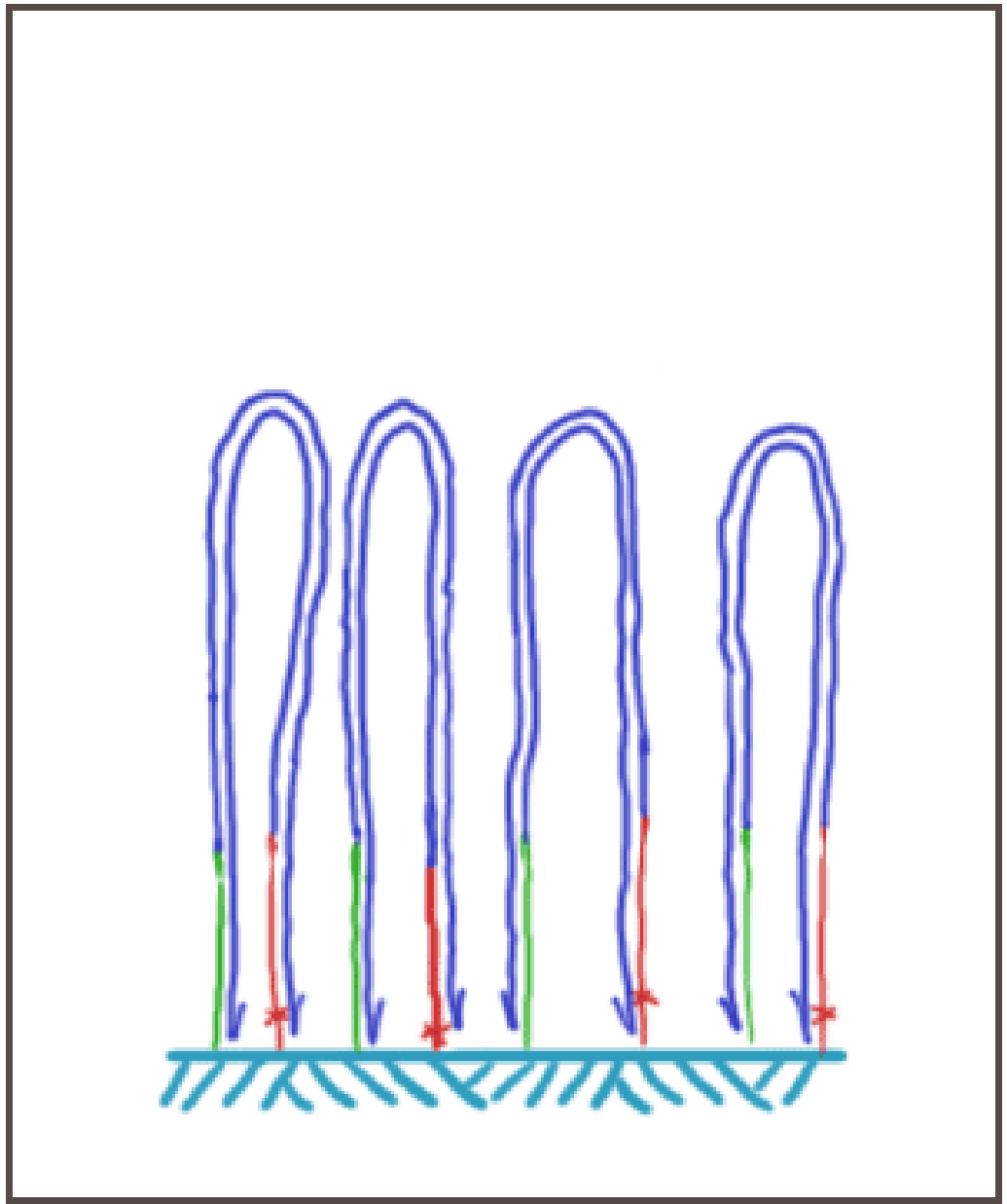
Each cluster has forward and reverse strand of the same molecule.





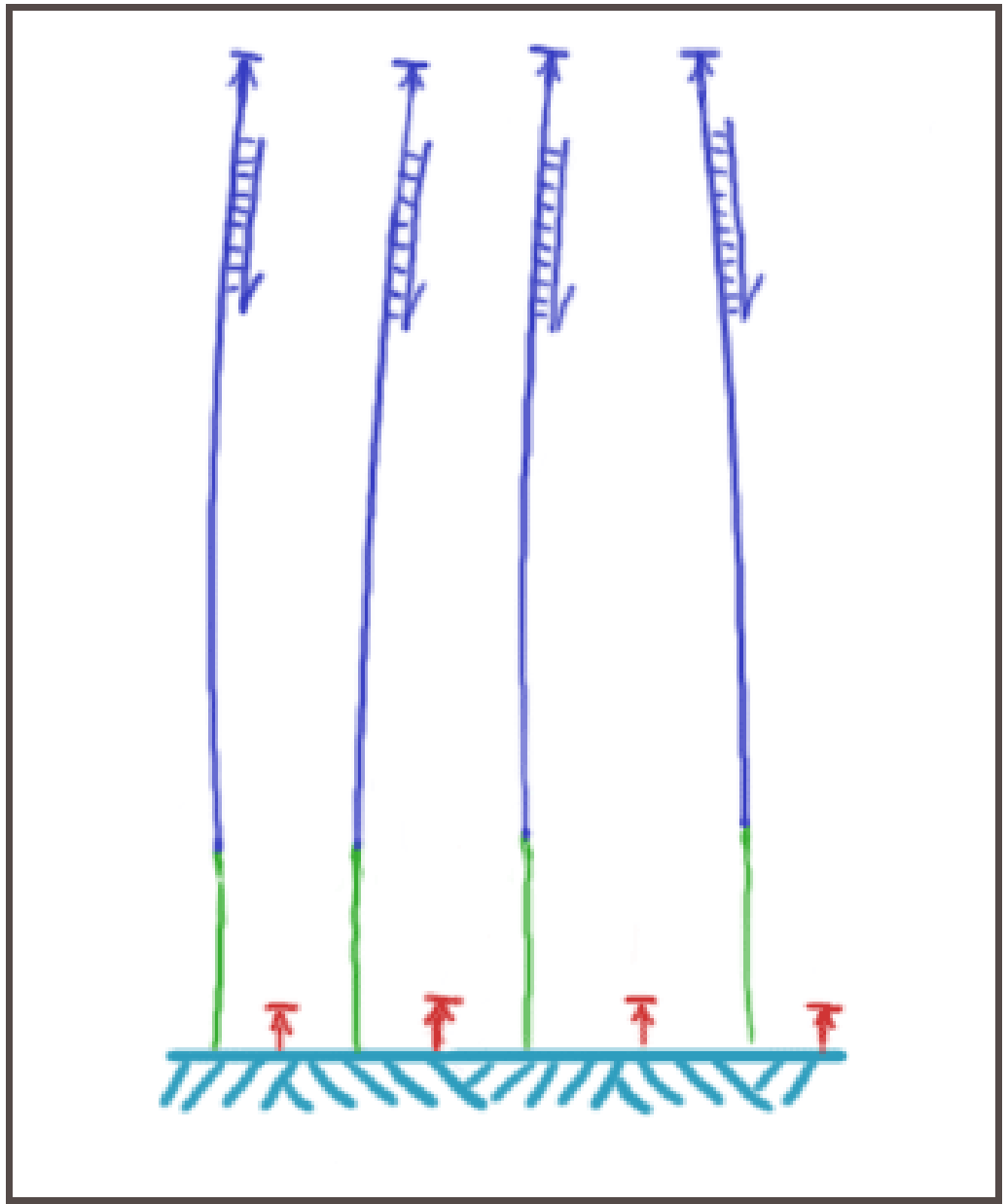
Reverse Strand Cleavage

Bridge Amplification



From A. Soldatov, MPI

Cluster Processing



From A. Soldatov, MPI

Insert size in the library

What will happen during cluster generation if the inserts in the library are very long?

Too big clusters get mixed

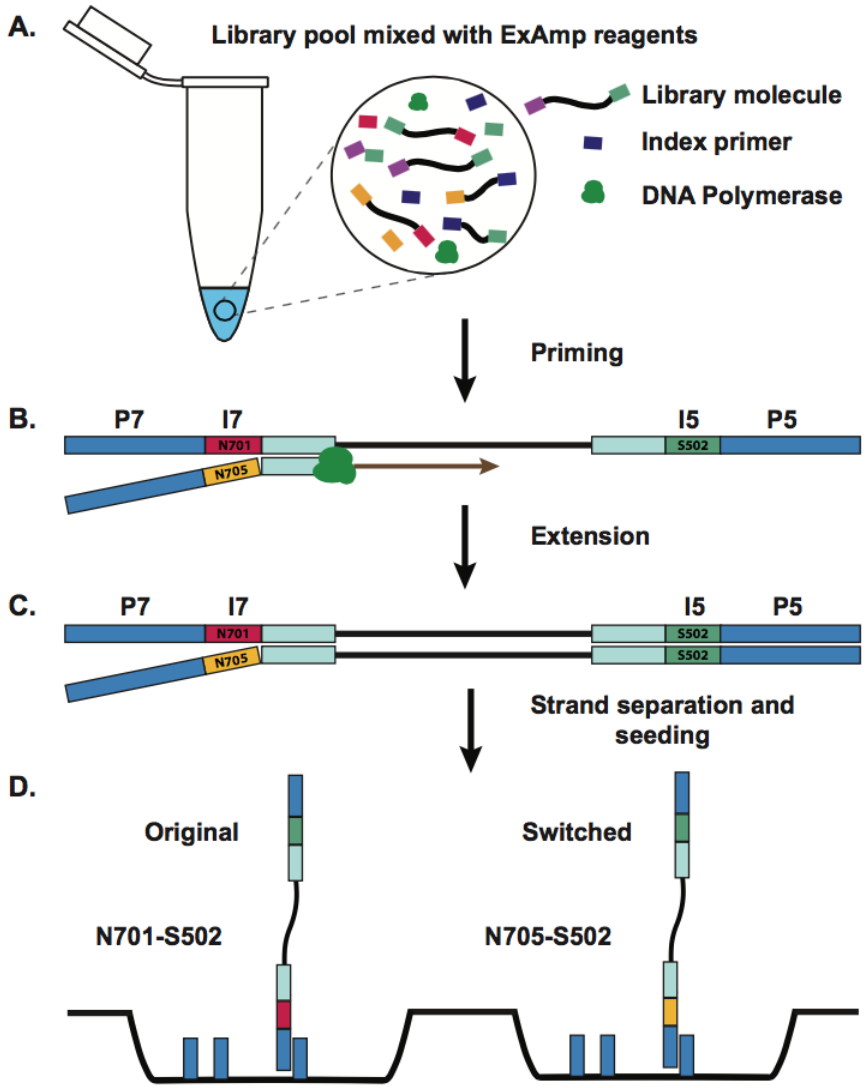
What will happen during cluster generation if the inserts in the library are very short?

Library too short to bridge for amplification
Sequencing of the adapter

Cluster Generation (patterned flow cell)

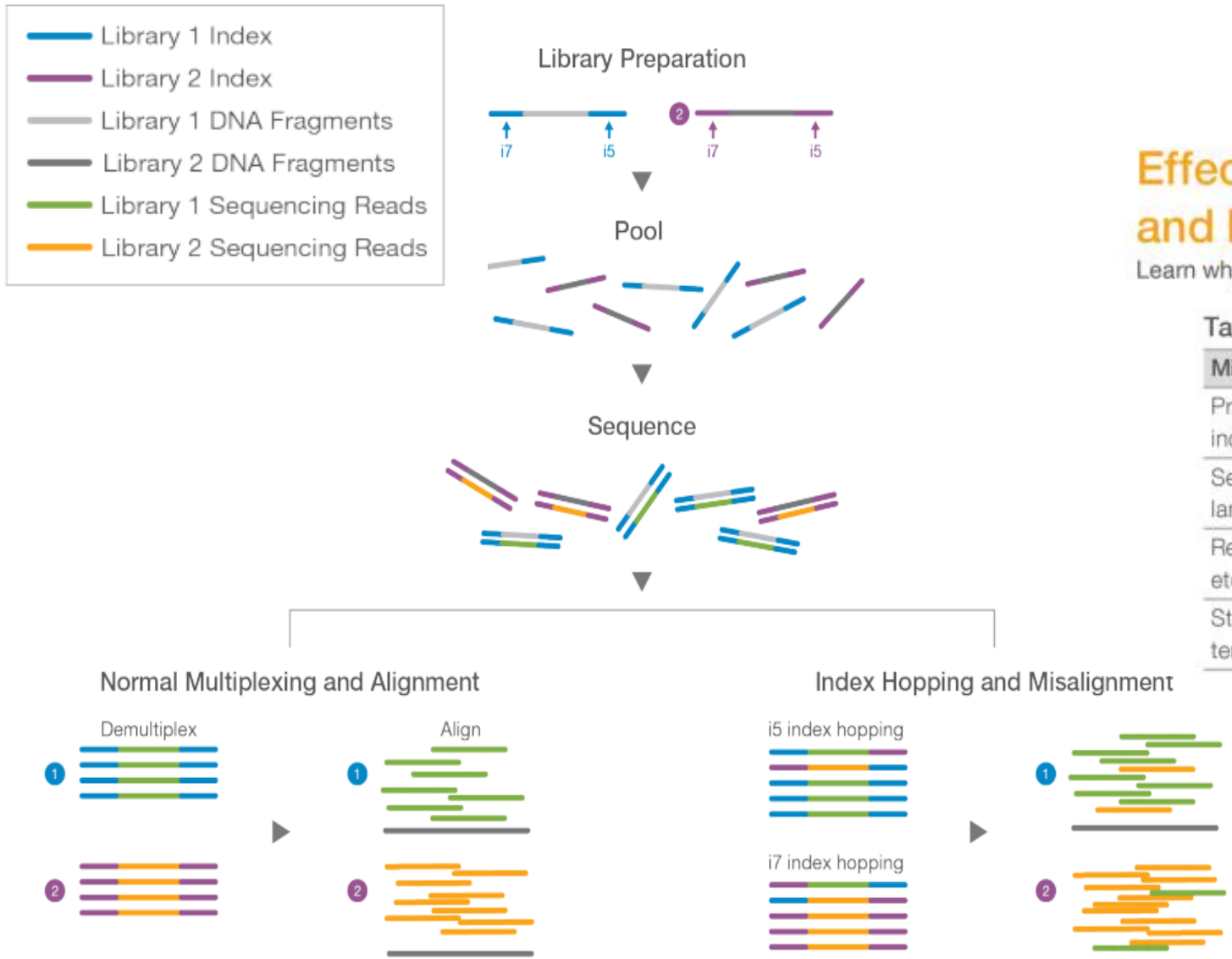
Exclusion Amplification

By enabling simultaneous seeding (landing of the DNA strand in the nanowell) and amplification, exclusion amplification promotes monoclonal cluster generation within the nanowells. This improvement allows the number of monoclonal clusters available for sequencing on a patterned flow cell to exceed the Poisson statistical limit, significantly increasing data output.



http://enseqlopedia.com/2017/04/index-swapping-illumina-examp-clustering/
https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf

Index-swapping with @illumina ExAmp clustering



Illumina's recommendations

Effects of Index Misassignment on Multiplexing and Downstream Analysis

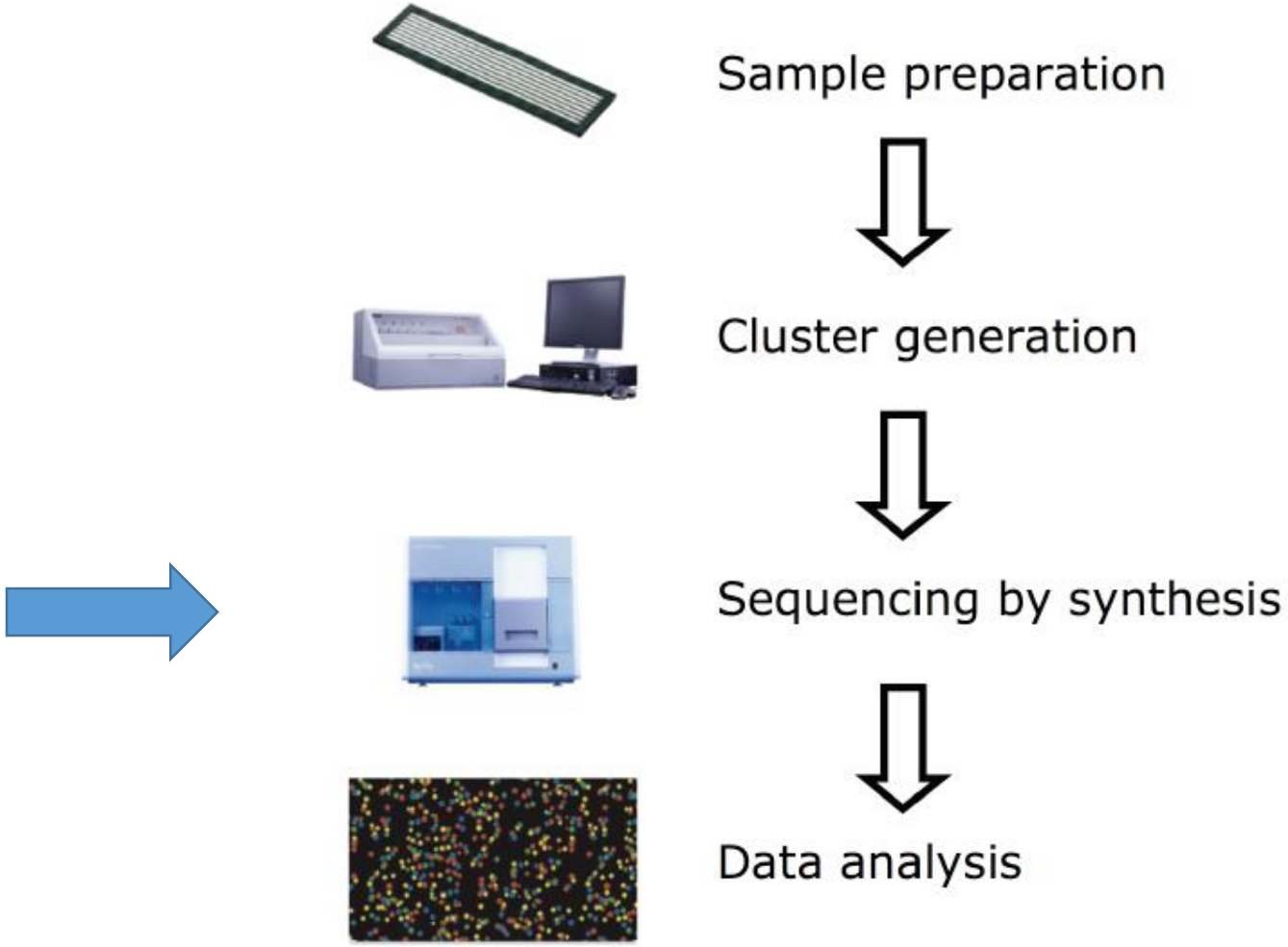
Learn why it happens and best practices to reduce the impact of index hopping.

Table 1: Best Practices for Reducing Index Hopping

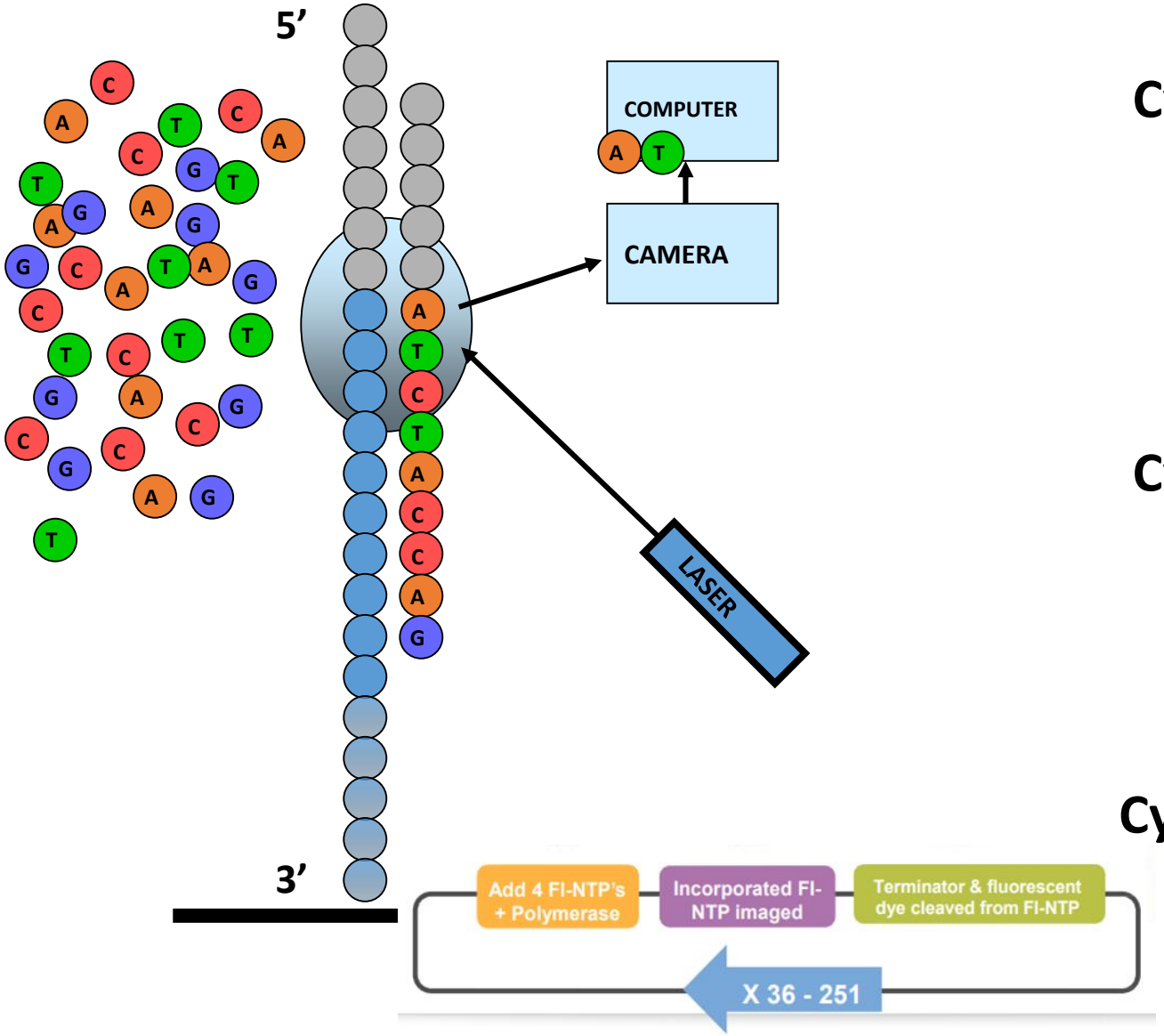
Mitigation/Recommendation	Benefit/Outcome
Prepare dual indexed libraries with unique indexes[a]	Converts index hopped reads to undetermined
Sequence one 30x human genome per lane[b]	Avoids pooling and index hopping
Remove adapters (cleanup, spin columns, etc)[c]	Reduces levels of index hopping
Store prepared libraries at recommended temperature of –20° C[c]	Reduces levels of index hopping

https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf?linkId=36607862

Illumina Workflow



Illumina Sequencing By Synthesis (SBS)



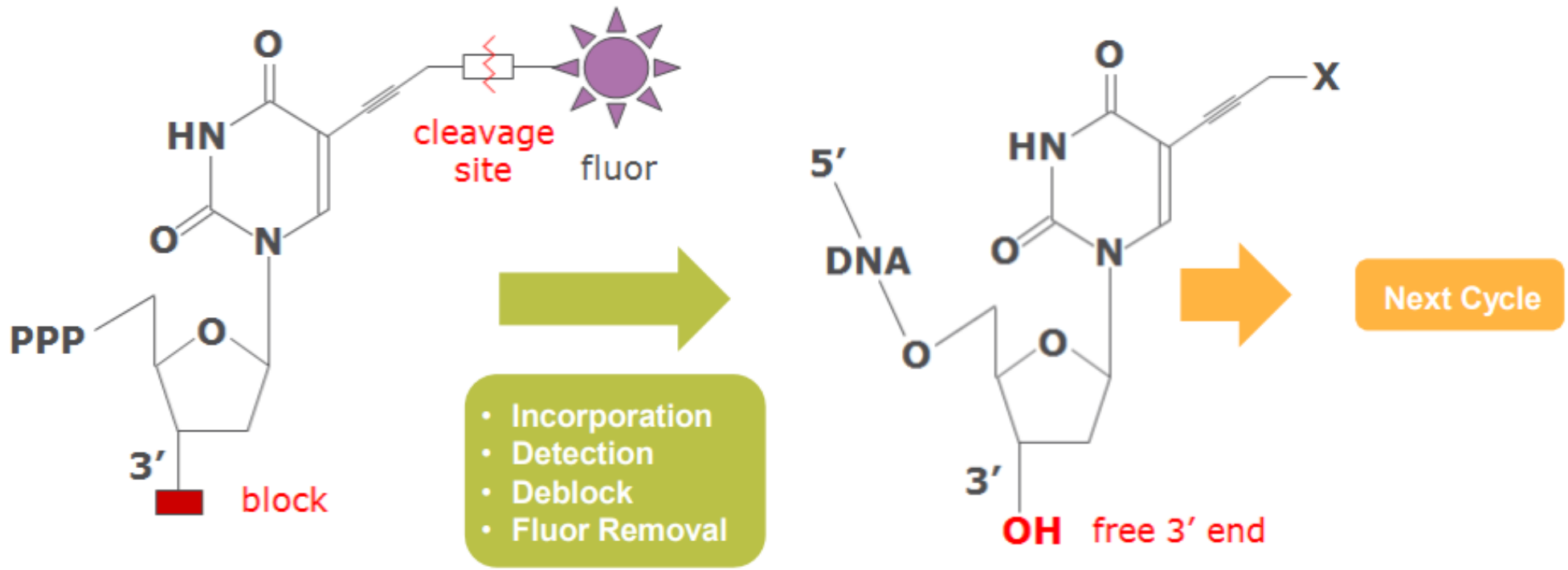
Cycle 1:
Primer hybridization
Add sequencing reagents
First base incorporated
Remove unincorporated bases
Detect signal
Cleave dye and terminator

Cycle 2:
Add sequencing reagents
Second base incorporated
Remove unincorporated bases
Detect signal
Cleave dye and terminator

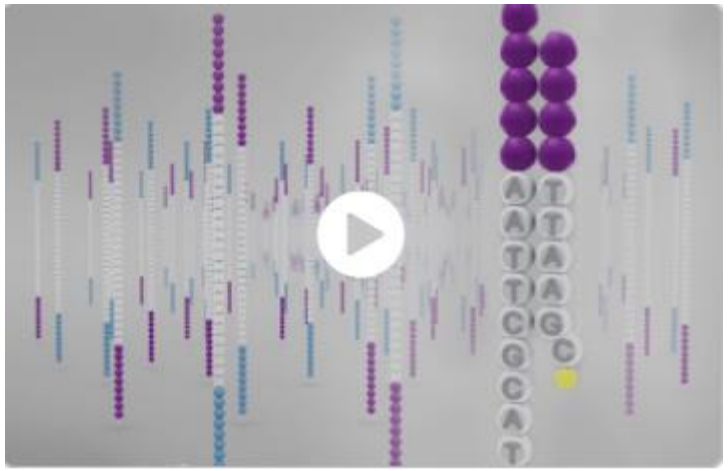
Cycle 3-n: ……

Reversible Terminator Chemistry

- All 4 labeled nucleotides in 1 reaction
- Higher accuracy
- No problems with homopolymer repeats

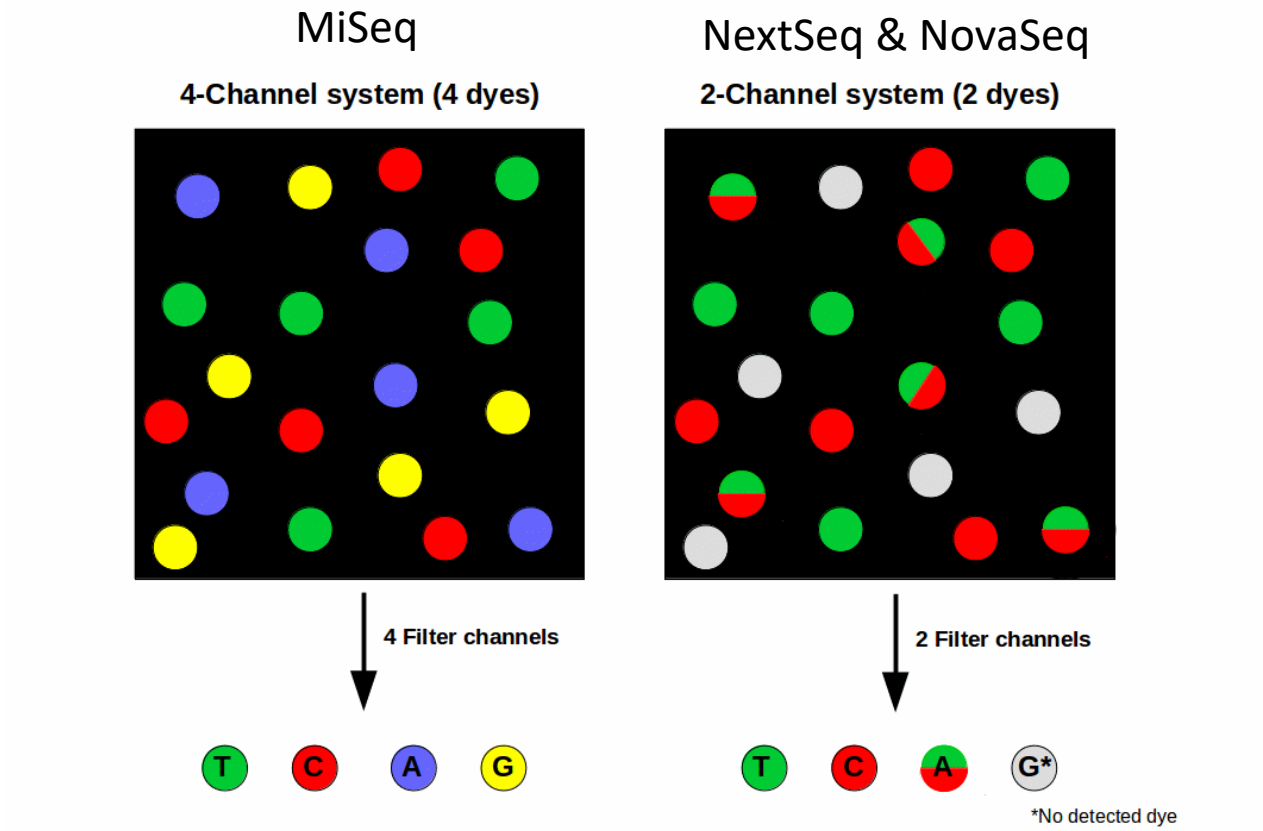


How Does Illumina NGS Work?



https://emea.illumina.com/science/technology/next-generation-sequencing.html

New instruments use 2 dyes



Is the 2-color better or worse?

Pros

Faster sequencing process (scanning)

Cheaper optical instrumentation

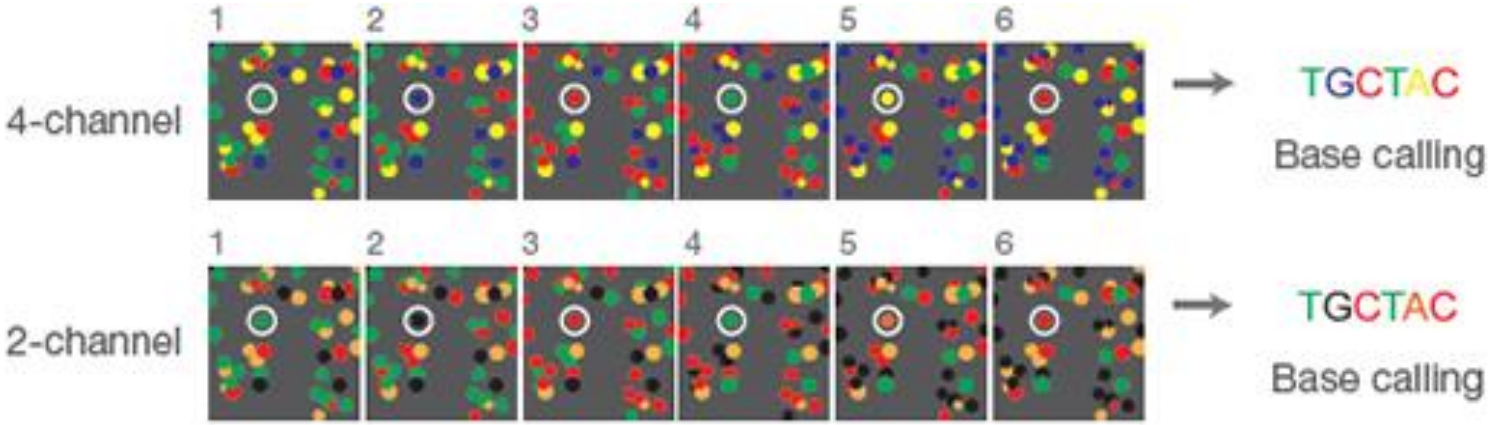
Cons

Distortion of base qualities

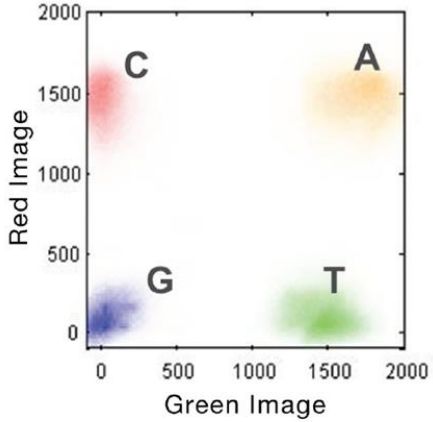
Less accuracy of the sequences.

https://www.ecseq.com/support/ngs/do_you_have_two_colors_or_four_colors_in_Illumina

Imaging



Each cluster or read has a specific physical location



Library = Sample

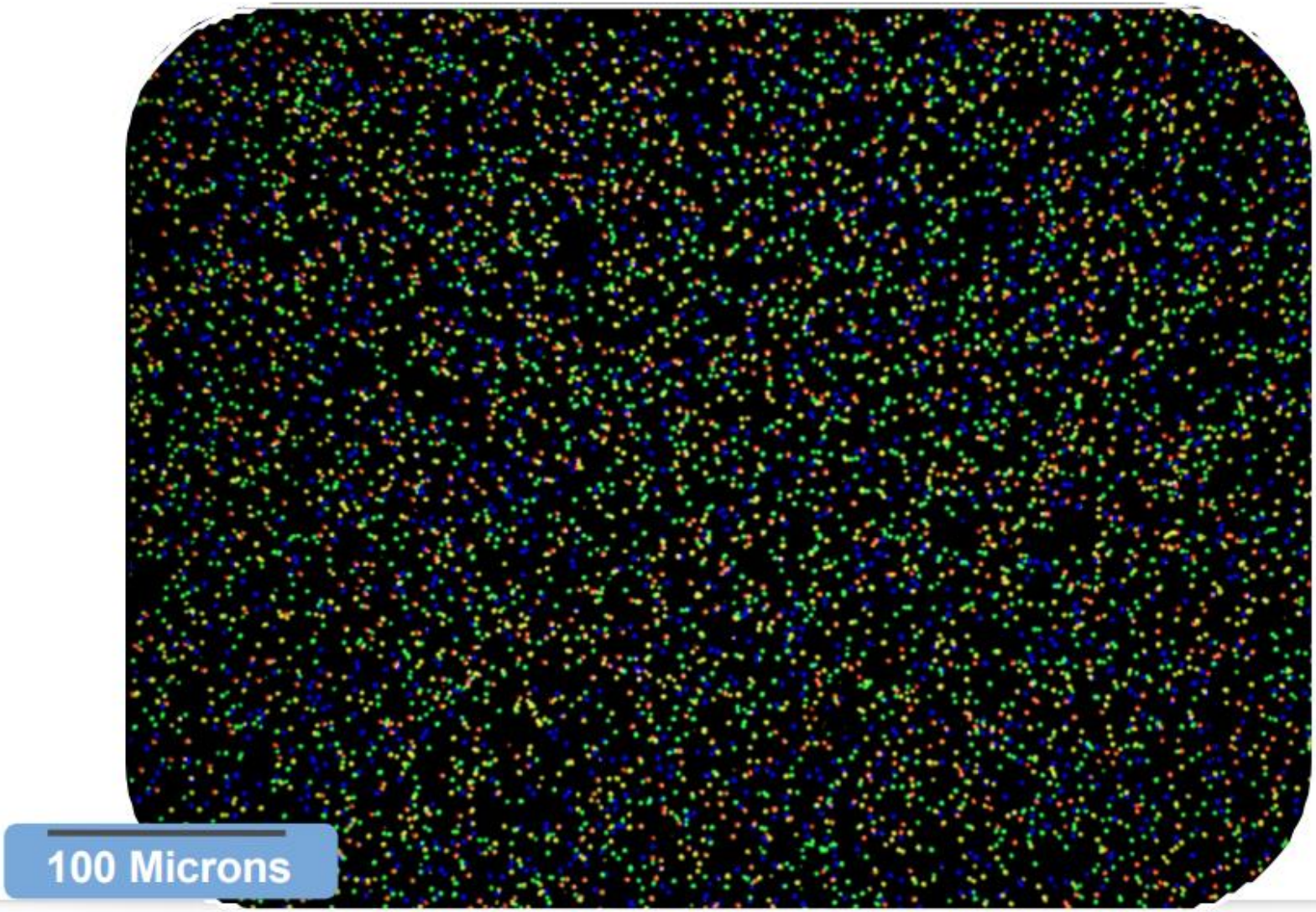
Sequences = Reads = Clusters

Cycles = Bases

Index = Barcode

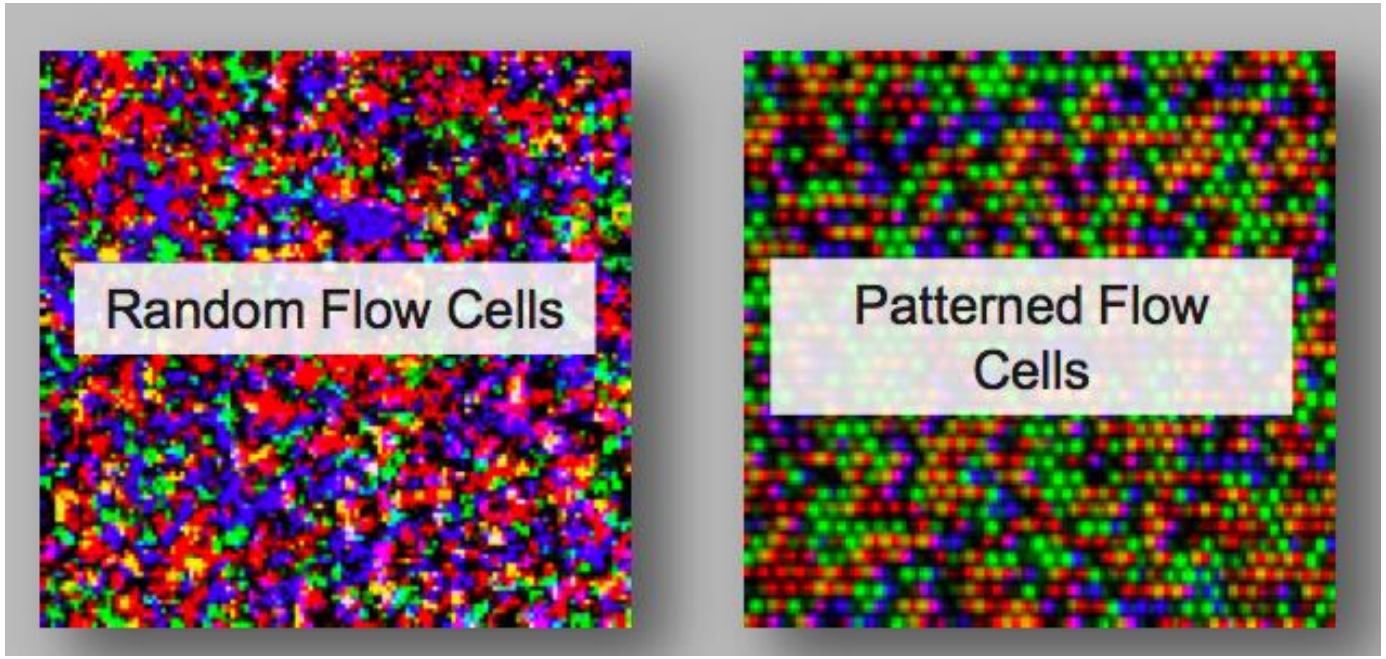
Clusters (of DNA molecules sequenced):

Cluster Intensities collected following every base addition



illumina®

Oops: Patterned Flow Cells



- Single sequence per well
  - Higher density, more data
- Different side effects
  - Index hopping
  - Duplicate reads

GIGO or you always get data…..

Garbage in

Garbage out

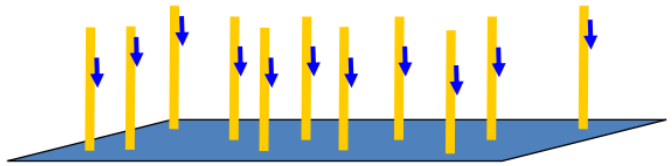
Bad Sample

Bad Library

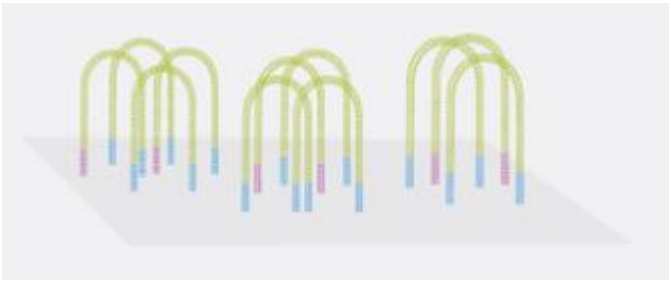
Bad Sequencing Data

Summary

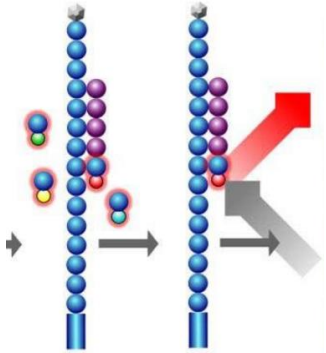
- Attach to solid phase.
  Track each sequence according to its physical location.



- Amplify in a limited area to get a spot with one molecule type (bridge amplification, exclusion amplification). Amplification is needed for detection.



- Short reads for optimal clustering and because of limitations of the chemistry



- Sequence by synthesis

- Detection after each base synthesis

- Barcodes for multiplexing
  You need to know where the barcode is located