Sequencing instruments

Bioinformatics for Beginners using the Biostar Handbook

Peter FitzGerald (Genome Analysis Unit, BTEP)
Desiree Tillo (CCR Genomics Core, GAU)

Outline

Next **G**eneration **S**equencing (NGS)

- Overview of sequencing technologies/platforms
- NGS resources at NCI/CCR
- Illumina Technology
- Long Read Technology
- Preparing your samples
- Receiving your data

Next Generation Sequencing

History

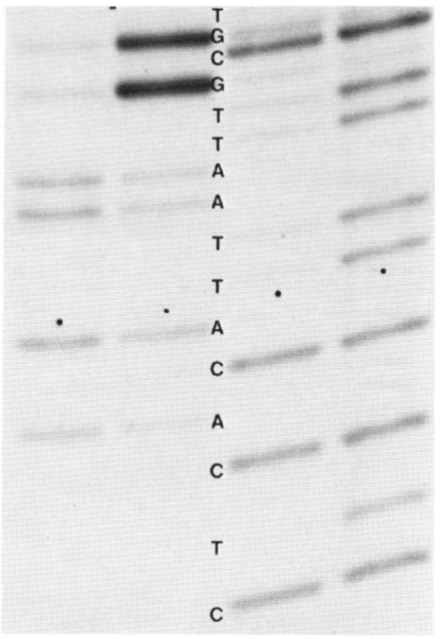
P^{32} radioactivity

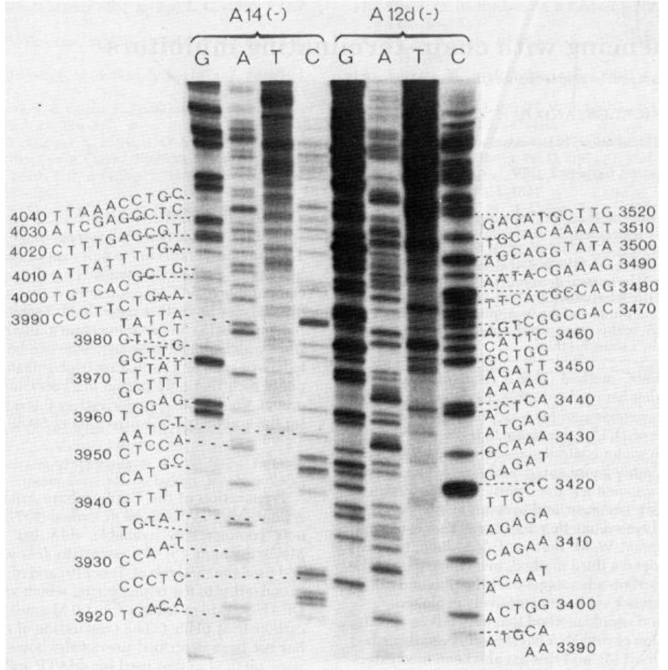
Polyacrylamide gels

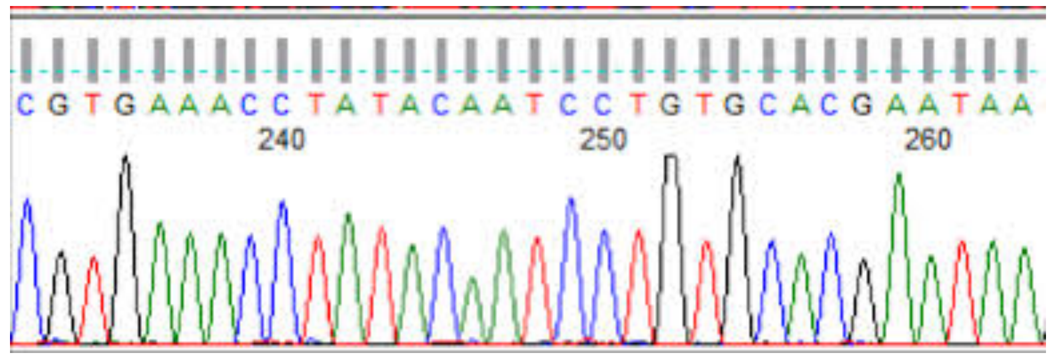
First Generation Sequencing

- Maxam-Gilbert - *A new method for sequencing DNA* (1977)
  Sequencing by degradation

- Sanger Sequencing - DNA sequencing with chain-terminating inhibitors (1977)
  Sequencing by synthesis







Next Generation Sequencing

History

Second/Next Generation Sequencing (NGS) - Massively Parallel, Short Reads

- **Roche** - 454 DNA sequence (2007)
- **ABI** Solid
- **Illumina** (short-read, sequencing by synthesis) is most commonly used platform
- **Ion Torrent**
  Third Generation Sequencing - Long Reads
- **Pacific Biosciences**
- **Oxford Nanopore**

Single Cell Sequencing

- **10X Genomics**
- **Drop Seq**

Where can I get my samples sequenced?

NGS resources at the NCI

- **Office of Science and Technology Resources (OSTR)**
  https://ostr.ccr.cancer.gov

- **Supplemental Technology Award Review System (STARS)**
  https://ostr.ccr.cancer.gov/stars/

- **Collaborative Research Exchange (CREx)**
  https://crex.scientist.com/users/sign_in

Where can I get my samples sequenced?

NGS resources at the NCI

- **NCI Sequencing Facility (SF)** - ATRF, Frederick
- **NCI CCR Genomics Core (GC)** - Bethesda, Bldg 37
- **NCI CCR Single Cell Analysis Facility (SCAF)** - Bethesda, Bldg 37
- **NCI Genomics Technology Laboratory (GTL)** - Frederick
- **NIH Intramural Sequencing Center (NISC)** - Rockville
- ***Commercial (Various)***

Next Generation Sequencing

Sequencing Facility (SF) - ATRF, Frederick

https://ostr.ccr.cancer.gov/resources/sequencing-facility/

- **Illumina Sequencing Technology** (Short Read)
  NovaSeq6000, NextSeq500, Hiseq4000, and MiSeq sequencer
- **PacBio Sequel II Sequencing** (Long Reads)
  Long Read single-molecule real-time (SMRT) technology
- **10X Genomics Chromium System** (Single Cell)
  Single Cell Gene Expression, Single Cell Immune Profiling, Single Cell ATAC (Assay for Transposase Accessible Chromatin) and Single Cell CNV
- **Bionano Genomics**
  Non-sequencing-based genome mapping technology

Highlights:

- Large/production-scale projects (i.e. lots of samples, standardized protocols)
- All NGS applications, incl. whole-genome/exome sequencing, RNA-seq, ChIP-seq, etc
- Primary and secondary analyses for all NGS projects, including initial base-calling, demultiplexing, data quality control, and reference genome alignment of NGS reads.

Next Generation Sequencing

CCR Genomics Core - Bldg 37, Bethesda

https://genomics.ccr.cancer.gov

- **Next-Generation Sequencing** (MiSeq, NextSeq 550) - (Short reads)
- **Sanger Sequencing** (2 -ABI 3500xL and 1-3730 xL DNA sequencers)
- **Digital Gene Expression** (NanoString nCounter System)
- **Digital droplet PCR** (BioRad QX200 ddPCR)
- **Analytical and preparative electrophoresis** (Tapestations 4150 and 4200, Pippin HT)
- **Automation** (2 Agilent Bravo and Mantis liquid handlers)
- Oxford Nanopore MinION (Long reads)   coming…

Highlights:

- Smaller-scale/pilot projects/fast turnaround
- RNA-Seq, ChIP-Seq, targeted panels, ATAC-seq, amplicon sequencing, CRISPR libraries
- Primary analyses for all NGS projects, including initial base-calling, demultiplexing, data quality control.

Next Generation Sequencing

Single Cell Analysis Facility (SCAF) - Bldg 37, Bethesda

https://ostr.ccr.cancer.gov/emerging-technologies/single-cell-analysis/

- Menarini Silicon Biosystems DEPArray system
- 10X Genomics Chromium system

Emerging, technologies will include:

- BD Genomics Rhapsody system
- Akoya Biosciences CODEX protein imaging system

Highlights

- 10X Genomics Single Cell 3' and 5' Whole Transcriptome Profiling & VDJ Sequencing
- Plate-based Single Cell Sequencing (e.g. Smart-Seq2)
- 10x Genomics Single Cell ATAC Sequencing

Next Generation Sequencing

NCI Genomics Technology Laboratory (GTL) - Frederick

https://ostr.ccr.cancer.gov/resources/genomics-laboratory/

- Variety of options for **NGS library preparations**, including substantial project design consultation, Ion Torrent (**CLIA certified)**, and Illumina MiSeq
- **Whole exome capture** using Agilent's SureSelect reagents
- **16s Microbiome Pipeline**: Automated fecal sample extraction, library preparation, normalization, and sequencing on MiSeq for bacterial 16s RNA gene.
- Specific single nucleotide polymorphism (SNP) detection and DNA methylation analysis on the Qiagen Pyromark platform
- Large projects requiring **laboratory automation** are managed using one of several Beckman BioMek FX liquid handling systems.

Highlights

- Whole exome capture library preparation
- Custom assay design

Next Generation Sequencing

NIH Intramural Sequencing Center (NISC) - Rockville

https://nisc.nih.gov/

- **Illumina Sequencing Technology** (Short Read)
  Illumina NovaSeq 6000, NextSeq 550, Illumina MiSeq
- **PacBio Sequel II Sequencing** (Long Reads)
  Long Read single-molecule real-time (SMRT) technology

Highlights:

- Still have an Illumina 2500
- Use Globus for data delivery

Next Generation Sequencing

Experimental Considerations

- **Please** talk to the experts **(Sequencing Core AND Bioinformatician) BEFORE** you do your experiment to ensure proper experimental design.
- For publishable experiments you should have at least 3 biological replicates (absolute minimum), but 4 if possible (optimum minimum - a safety net for failed samples).
- If you are unable to process all your samples together and need to process them in batches, make sure that replicates for each condition are in each batch so that the batch effects can be measured and removed
- Sequence depth and machine requirements estimates can be obtained from the Illumina Sequencing Coverage website *(https://support.illumina.com/downloads/sequencing_coverage_calculator.html)*
- Cost estimates can be previewed at the NCI Sequencing Facility website *(https://ostr.ccr.cancer.gov/resources/sequencing-facility/?target=Pricing)*

Next Generation Sequencing

The Reality

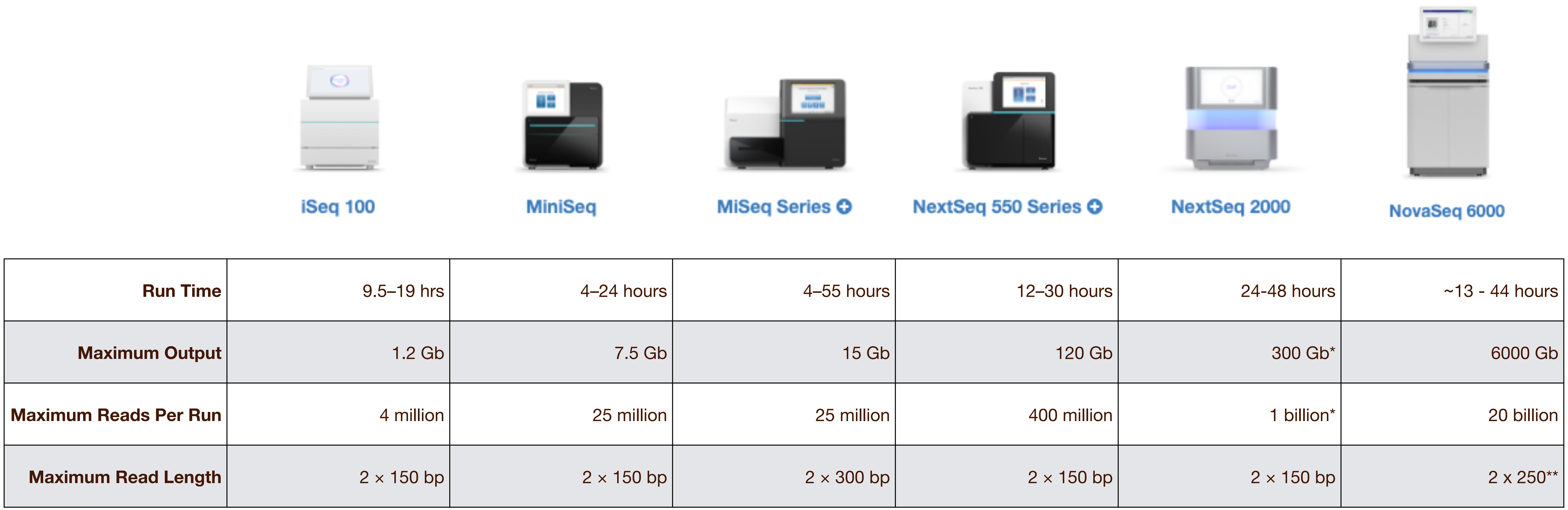
The vast majority of DNA sequencing done today is on Illumina platforms, and most likely this is the type of data you will be dealing with. Also, the techniques and programs for dealing with the other technologies are often specialized and somewhat proprietary

Thus the next section of today’s talk with deal with the specifics of Illumina technology.

Reasons for not using Illumina:

- Long Reads (PacBio, NanoPore)
- RNA Isoforms
- Whole Genomes (Microbial)
- Direct RNA sequencing (NanoPore/PacBio)
- Other Specialized applications

Illumina sequencers



	iSeq 100	MiniSeq	MiSeq Series ⊕	NextSeq 550 Series ⊕	NextSeq 2000	NovaSeq 6000
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	24-48 hours	~13 - 44 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	300 Gb*	6000 Gb
Maximum Reads Per Run	4 million	25 million	25 million	400 million	1 billion*	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 x 250**

https://www.illumina.com/systems/sequencing-platforms.html

Obsolete: Genome Analyzer I/II, HiSeq

How does Illumina sequencing work?

Basic steps:

1. Sample/library preparation
2. Cluster generation/amplification
3. Sequencing

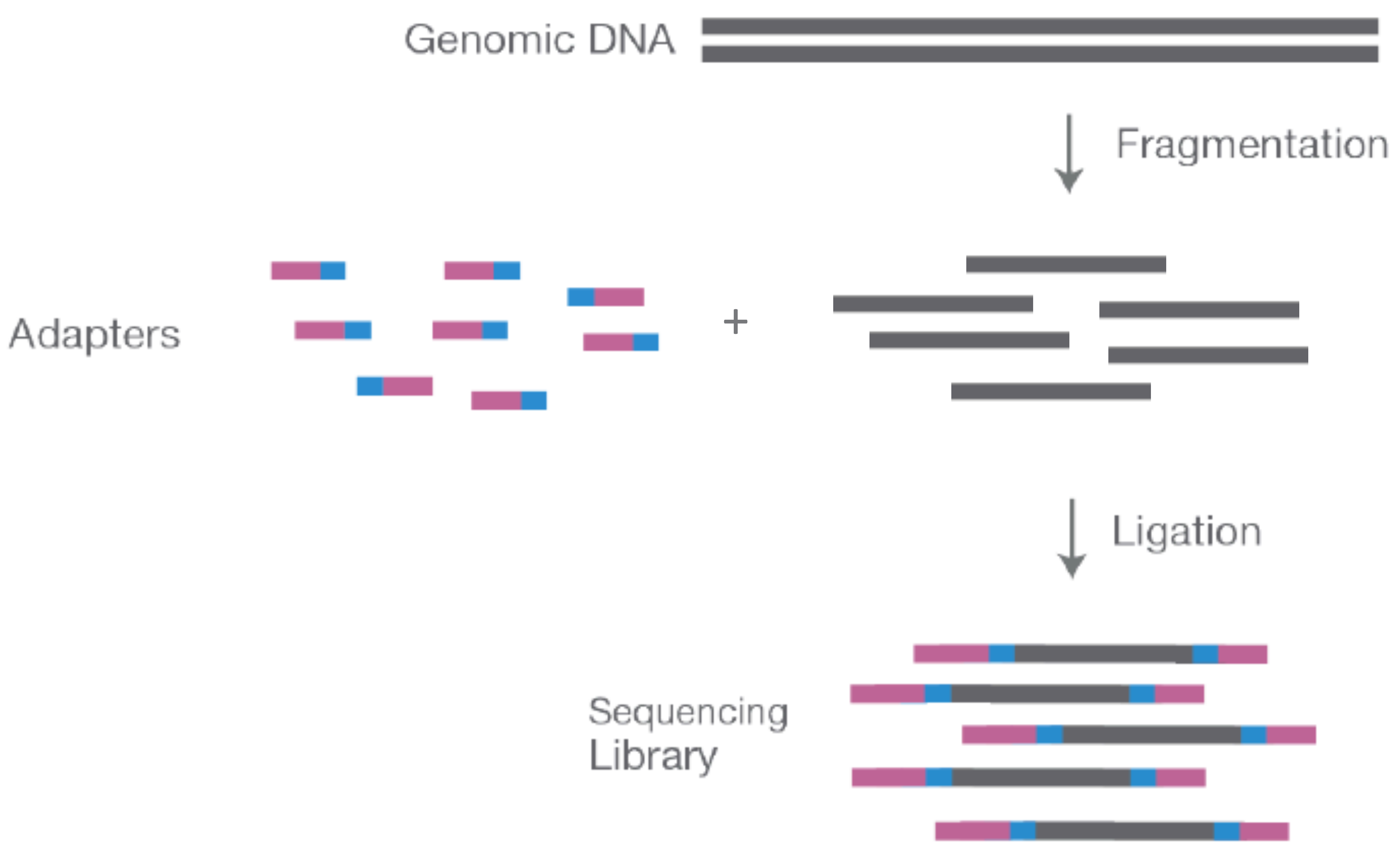
https://www.youtube.com/watch?v=fCd6B5HRaZ8

How does Illumina sequencing work?

Sample preparation

Adapters contain:

1. Platform-specific sequences for library binding to the sequencing instrument (P5, P7)
2. Binding sites for sequencing primers
3. Index sequences (used for multiplexing)



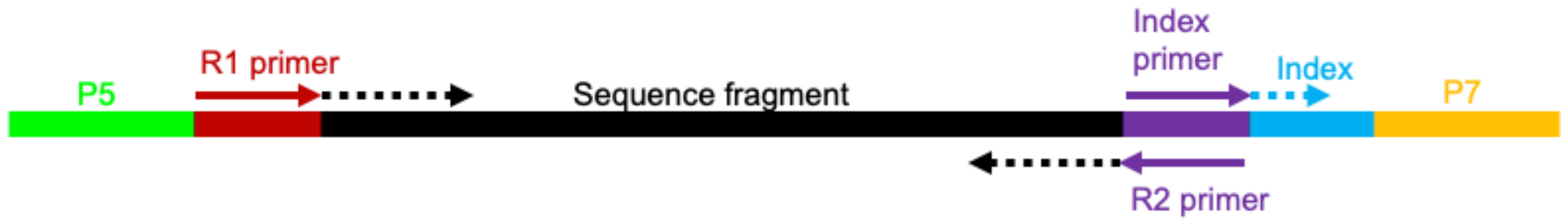
Modified from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

How does Illumina sequencing work?

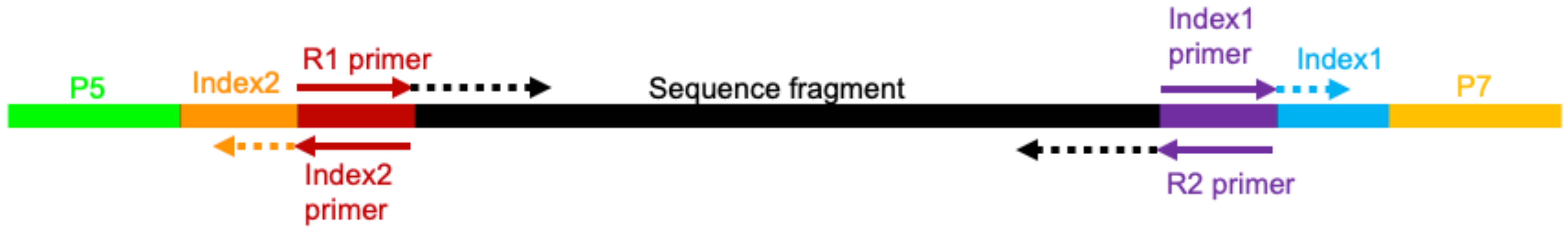
Sample preparation

Structure of typical Illumina libraries:

Single-indexed library



Dual-indexed library



Adapters contain:

1. Platform-specific sequences for library binding to the sequencing instrument (P5, P7)
2. Binding sites for sequencing primers
3. Index sequences (used for multiplexing)

tp://nextgen.mgh.harvard.edu/CustomPrimer.html

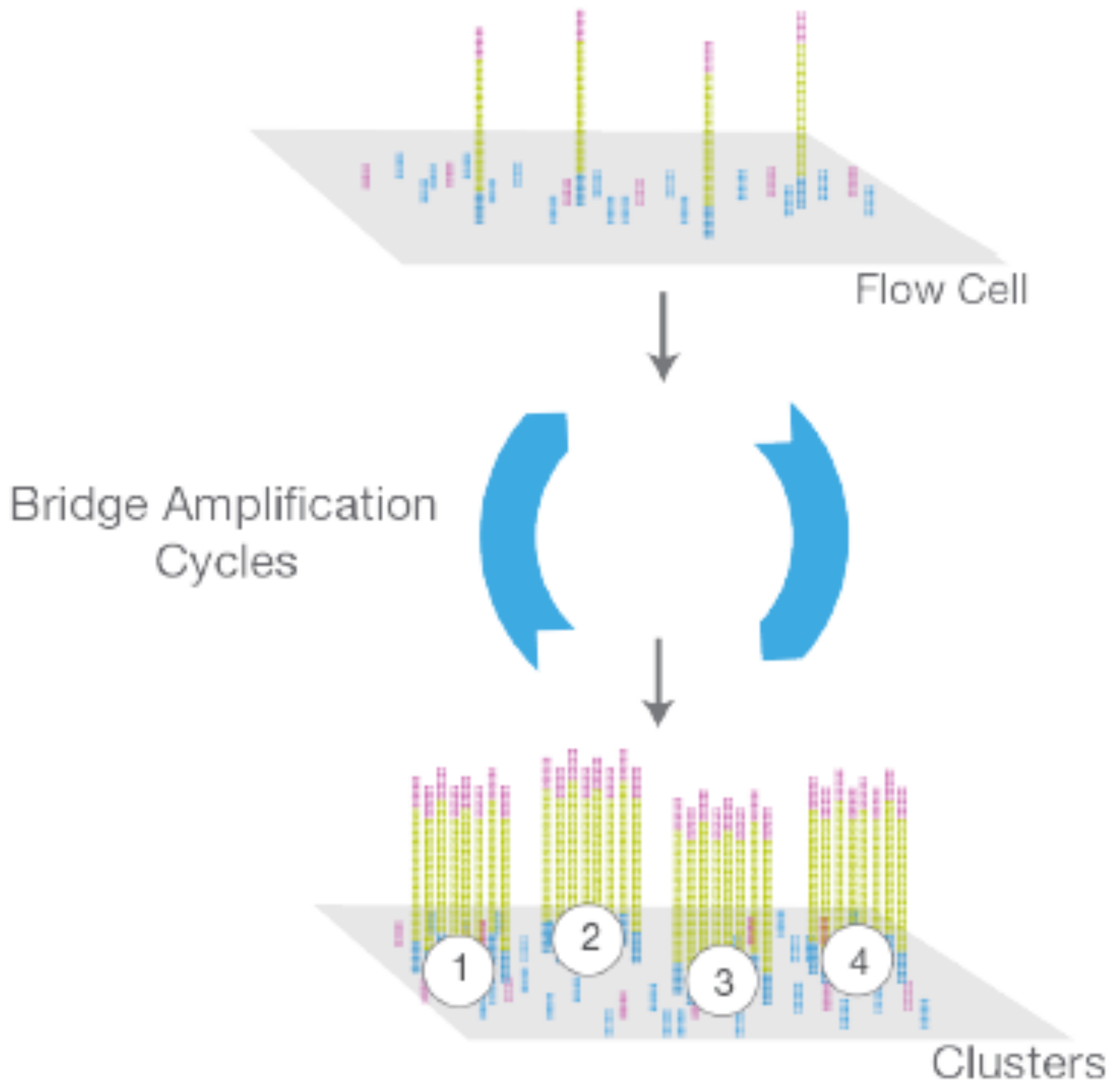
How does Illumina sequencing work?

Cluster generation/amplification

Flow cell surface contains oligonucleotides complementary to library adapter sequence

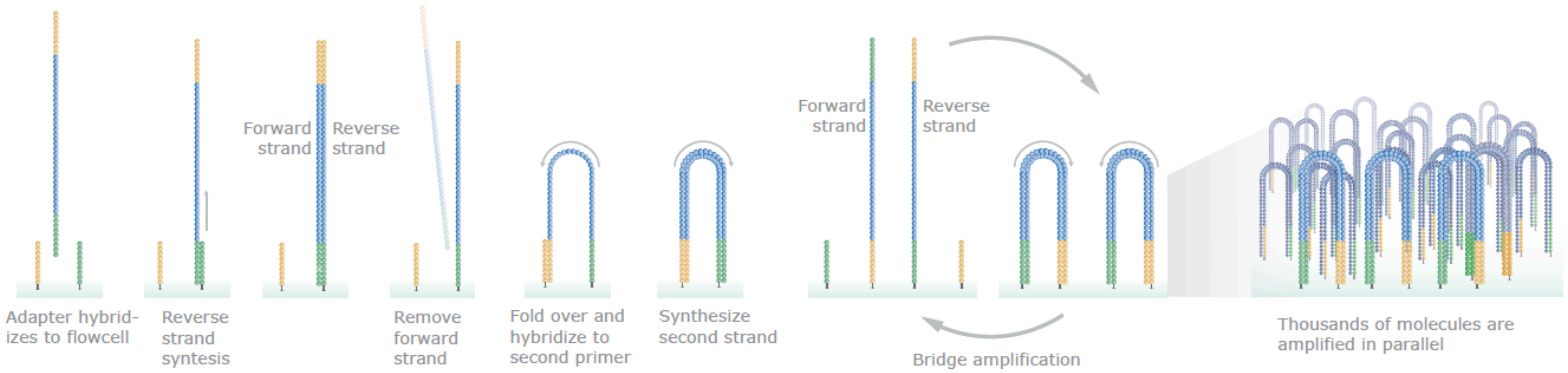
Denatured library loaded onto flow cell, fragments hybridize to the flow cell surface

Bound fragments are amplified via “bridge amplification” to generate clonal clusters containing ~1,000 copies of a single fragment



How does Illumina sequencing work?

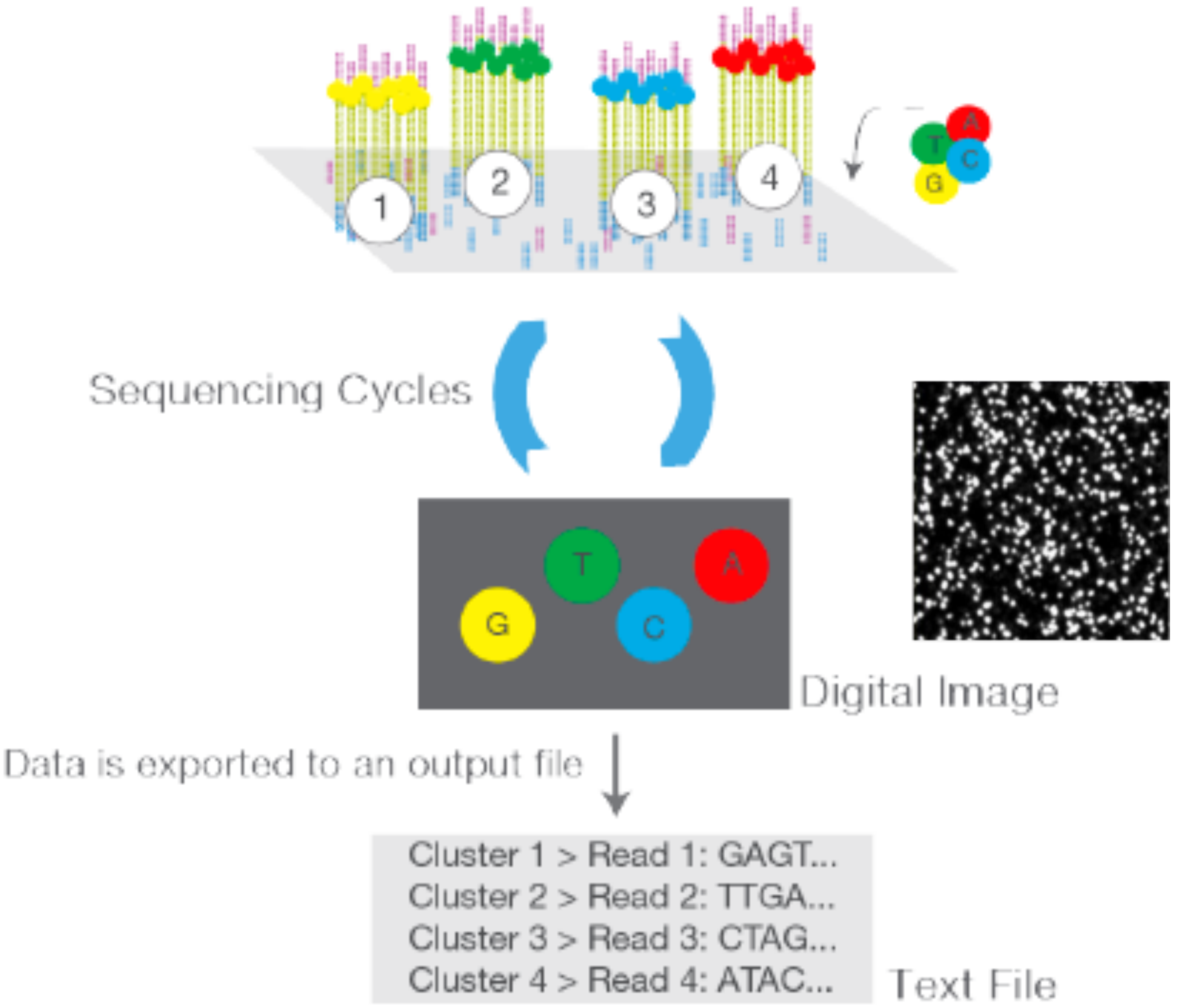
Cluster generation/amplification



https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/for-all-you-seq-dna.pdf

How does Illumina sequencing work?

Sequencing



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

Illumina uses “Sequencing by synthesis” (SBS) chemistry.

Sequencing reagents (sequencing primers, polymerase, fluorescently labeled nucleotides) are added to the flowcell.

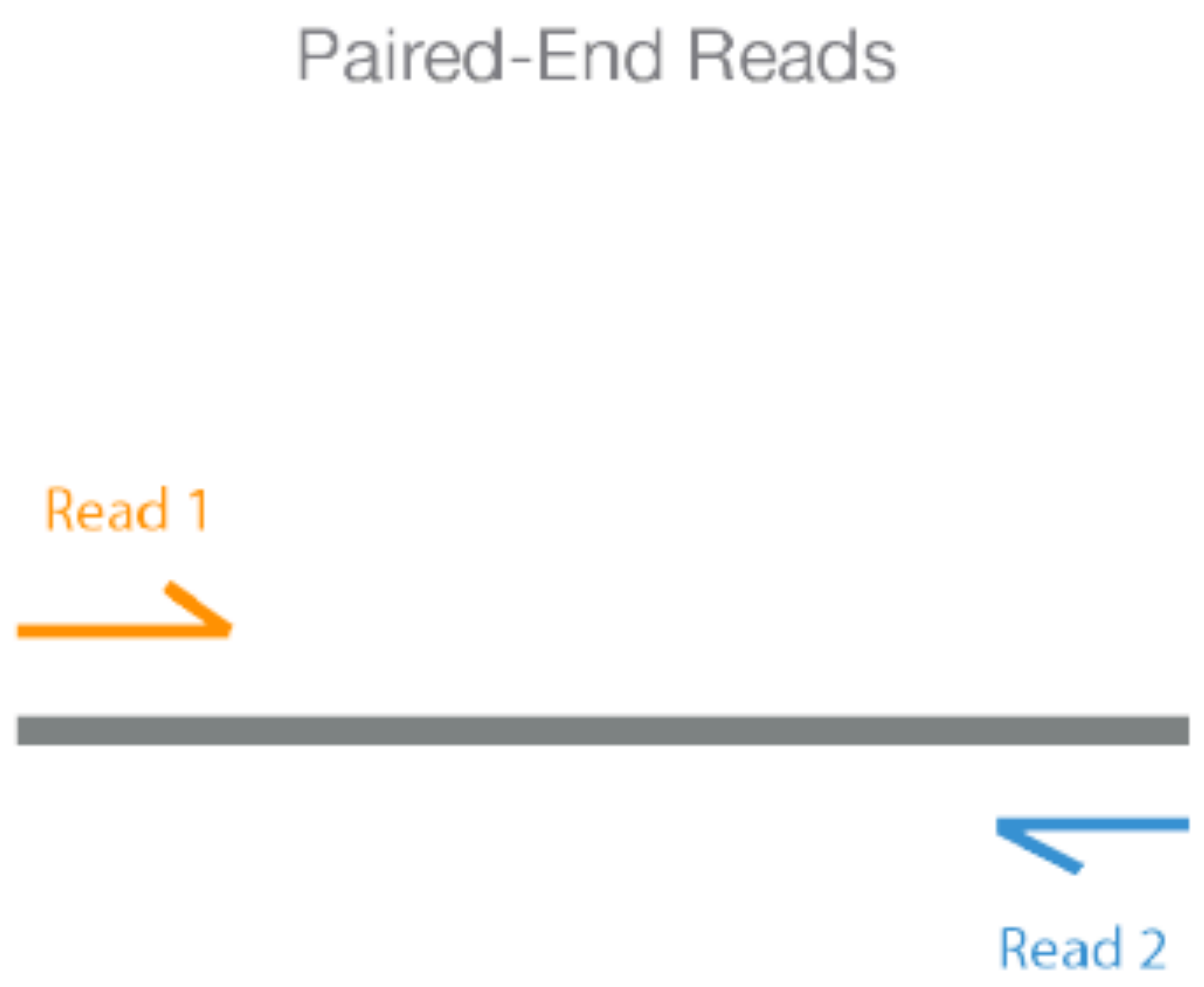
DNA polymerase incorporates a single nucleotide into the DNA template strand. The flow cell is imaged, and the fluorescence at each cluster is recorded. Each nucleotide has a characteristic fluorescence, and the base in each cluster at each cycle can be identified.

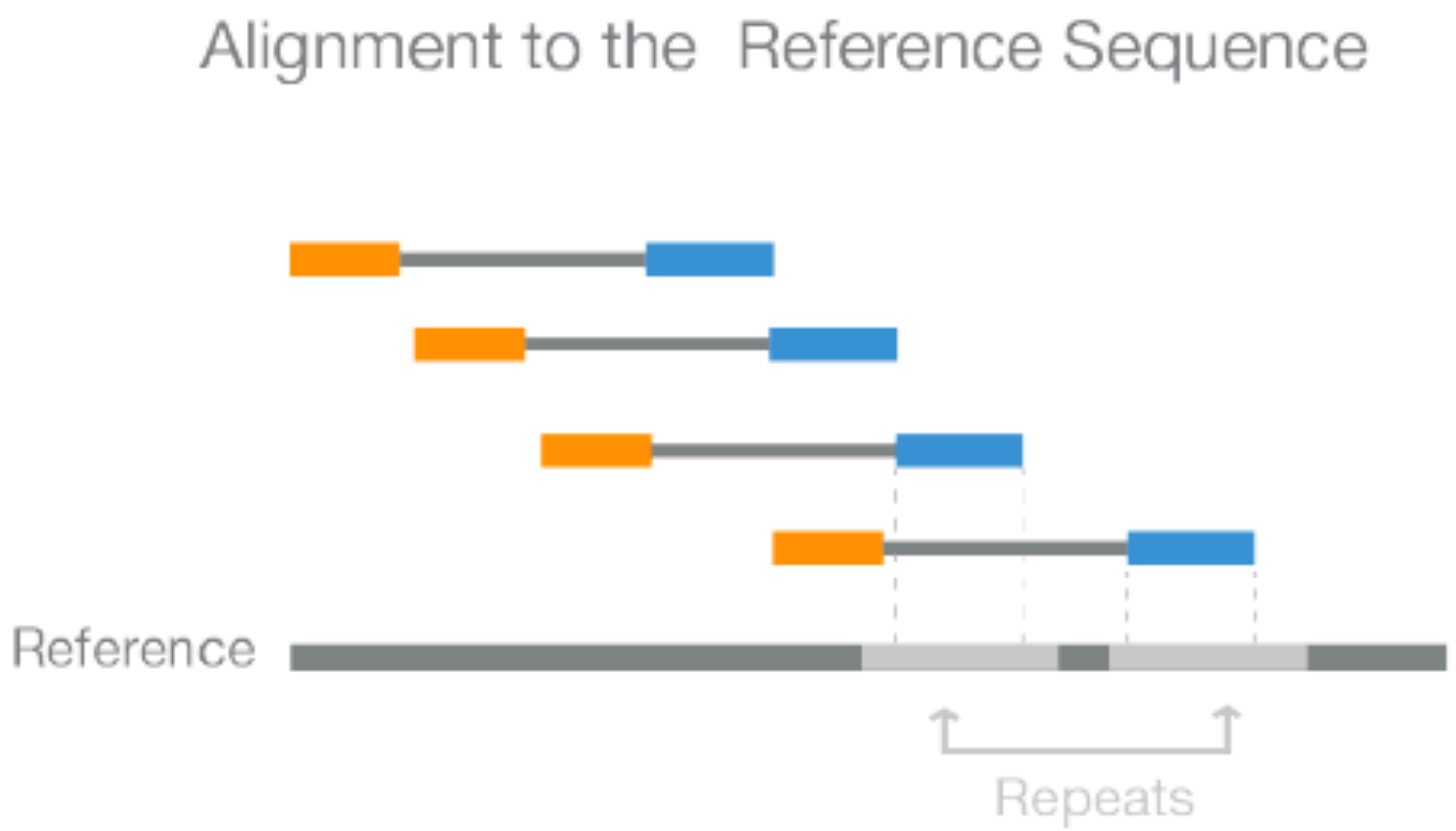
The process is repeated, with the number of cycles determining the length of the read (# cycles = read length)

Modified from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Sequencing details/terminology

Single end vs. paired end





Pros:

- Helps with mapping and assembly
- Sequencing same fragment twice
  - Allows for error estimation and correction
- Great for: variant analysis, applications requiring assembly of DNA sequences (plasmid/whole genome sequencing)
-

Cons

- Is more expensive
- Generates redundant data (not necessary for some applications)
- Takes longer

Sequencing details/terminology

Paired-end reads

- Generally two fastq files - often labelled R1 & R2 (some workflows require this naming)

- Entries within each file must be in the EXACT same order (watch out for trimming)

- No distinction within the files as to which is which

- CAN exist as interleaved files (alternating R1 and R2)

    - Default format in SRA downloads (hence `--split-files` in `fastq-dump`)
    - **AVOID** - most programs cannot process these correctly.

Sequencing details/terminology

Single End Read or Paired End Read R1

FASTQ

```
@M02511:190:000000000-CN9FK:1:1101:10703:1276 1:N:0:5
TCACGACCAGAAAACTGGCCTAACGACGTTTGTTCATTTCCTTCTACTTCT
+
-8ACCGGGGGGGGDFFGFFFFGGG7,@@@DF,,;,,,,;<@6,<6,,6,<,,
```

DNA

```
TCACGACCAGAAAACTGGCCTAACGACGTTTGTTCATTTCCTTCTACTTCT
```

Short fragment

Adaptor

FASTQ

```
@M02511:190:000000000-CN9FK:1:1101:11168:1418 1:N:0:5
CTTATGGAAGCCAAGCATTGGGGATTGAGAAAGAGTAGAAATGCCACAAGC
+
CCCCCGGGGGFGGGGGFGFFCEFGGGFFFGGGG8AF99<,<C<,CFFC<,,
```

DNA

```
CTTATGGAAGCCAAGCATTGGGGATTGAGAAAGAGTAGAAATGCCACAAGC
```

Long fragment

Sequencing details/terminology

Paired End Read R2

Reported

```
@M02511:190:000000000-CN9FK:1:1101:10703:1276 1:N:0:5
TCACGACCAGAAAACTGGCCTAACGACGTTTGTTCATTTCCTTCTACTTCT
+
-8ACCGGGGGGGGDFFGFFFFGGG7,@@@DF,,;,,,,;<@6,<6,,6,<,,
```

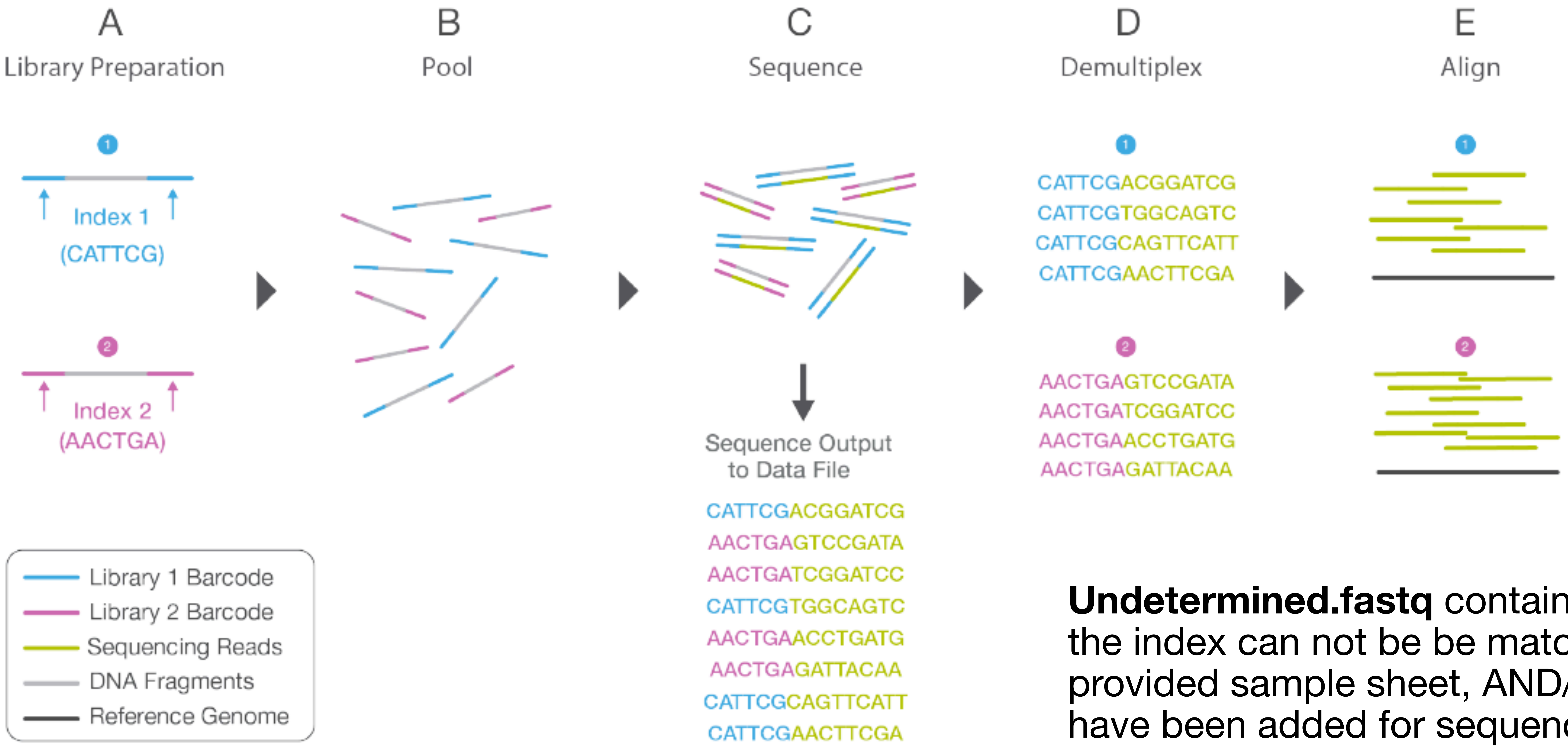
RAW

```
TCGAAGCTTCACGACCAGAAAACTGGCCTAACGACGTTTGTTCATTTCCTTCTACTTCT
```

Index Sequence Adaptor

Sequencing details/terminology

Multiplexing samples



Undetermined.fastq contains reads where the index can not be be matched to those in the provided sample sheet, AND/OR PhiX reads that have been added for sequencing optimization.

Sequencing details/terminology

How many reads do I need?

- Some key terms:
  - Depth: # of useable reads from the sequencing machine
  - Coverage: # times a read covers a known reference (a genome/locus)
- Can estimate coverage for an experiment here: https://support.illumina.com/downloads/sequencing_coverage_calculator.html
- Read depth or coverage varies depending on organism/experiment/application, but for human and mouse samples, here are some recommendations:

RNA-Seq

- **mRNA: 10-20M**, paired-end (PE) reads
  - Your RNA has to be high quality (not degraded, RNA integrity number (RIN) > 8)
- **total RNA (includes long noncoding RNAs): 25-60M** PE reads.
- This is also an option if your RNA is degraded.

Adapted from CCBR Experimental Design Best practices:
https://ccbr.ccr.cancer.gov/project-support/experimental-design-best-practices/

Sequencing details/terminology

How many reads do I need?

ChIP-Seq

- Narrow/punctate binding patterns (e.g. sequence-specific transcription factors): **10-15M** reads
- Broad binding patterns (non-specific binding, histone/chromatin marks): **>30M** reads
- Generally, single-end sequencing (read length=75nt) is recommended, as it is usually most economical.
- If you know your protein binds to repetitive or low-complexity regions, consider longer and/or paired-end reads.

ATAC-seq

- **50M** PE reads (75nt)

Tumor/Normal Variant Calling (Whole exome)

- Mean target depth is **>=100X for tumor**, and **>=50X for germline sample**

Germline Variant Detection

- Mean target depth of **>=50X for exome** and **>=30X for genome** - whole genome sequencing is recommended, rather than exome.

Adapted from CCBR Experimental Design Best practices:
https://ccbr.ccr.cancer.gov/project-support/experimental-design-best-practices/

Long read technologies

Overview

Short-reads can make reconstruction and quantification of original (often longer) molecule difficult.

- Transcript isoform detection and quantification
- Genome assembly, especially for repetitive regions
- Structural variations (copy number, large insertions/deletions) difficult to detect using short reads

Long read technologies (PacBio, Oxford Nanopore)

- single molecule sequencing

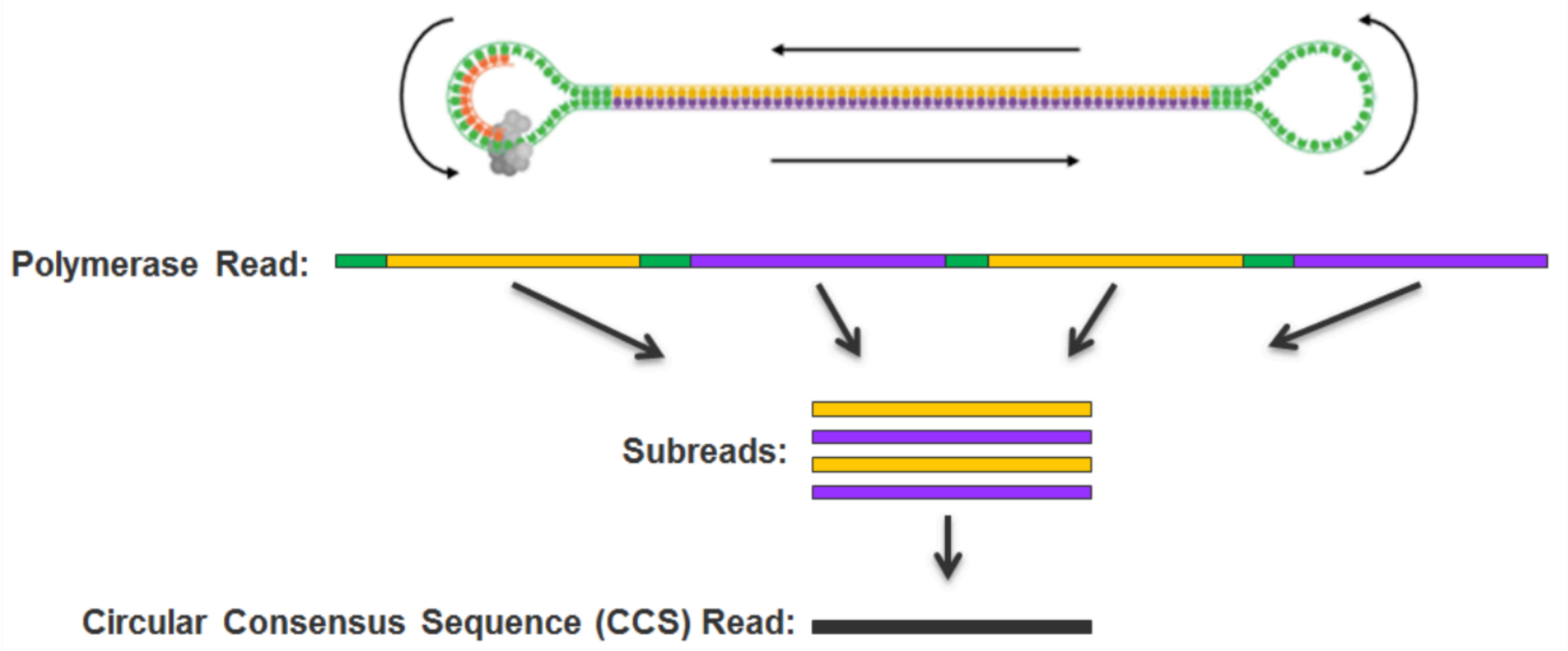
- read length in 10s of Kb in length

- lower accuracy (lower quality scores) than Illumina, but is improving, and not as big of a limitation as some may suggest

Long read technologies

PacBio

- 500bp-50kb inserts
- Flow-cells contain specialized wells (one DNA-molecule/well) - sequencing at single molecule resolution
- Direct detection of modified bases



Each fragment is circularized using **specialized adapters**

Polymerase and **primer** allow incorporation of labelled bases detected in real time

Once polymerase reads the entire fragment, it will loop back and read the fragment again (each pass generates a “subread”)

Aligning and piling up each subread gives a highly accurate consensus sequence (“Circular consensus sequence”)

Long read technologies

Oxford Nanopore

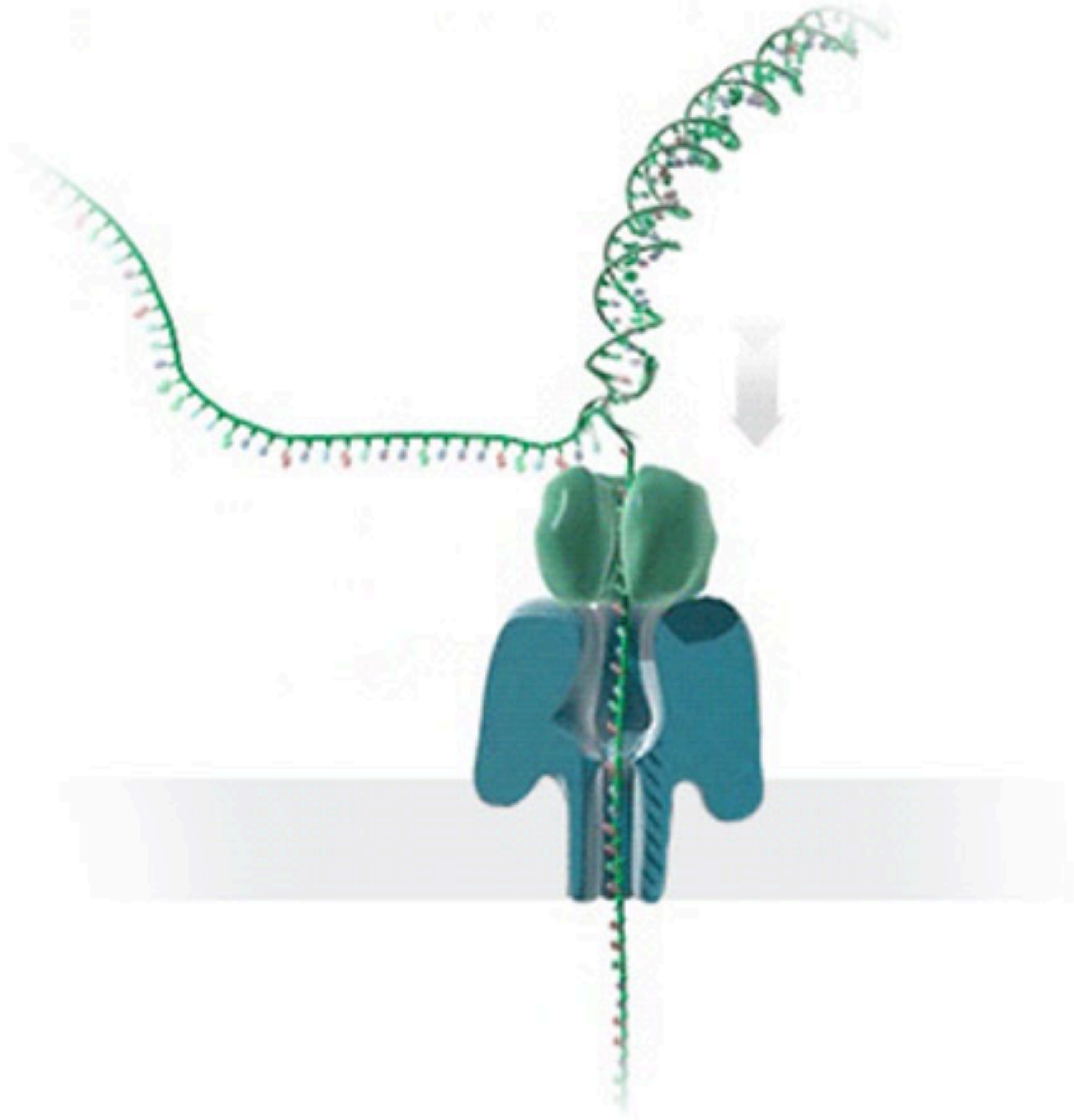
Sequencing through special pores in a membrane

- Membrane has an electric current flowing through it
- Nucleotide identity detected based on change in current

Direct sequencing of nucleic acid (DNA/RNA)

- No cDNA conversion
- Direct detection of base modifications (e.g. 5mC)

Read length is determined by your sample. Longest read detected: >2Mb



For MinION / GridION

Flongle

Adapter to enable small, rapid nanopore sequencing tests, for mobile or desktop sequencers

MinION Mk1B

Your personal nanopore sequencer, putting you in control

MinION Mk1C

Your personal nanopore sequencer including compute and screen, putting you in control

GridION Mk1

Higher-throughput, on demand nanopore sequencing at the desktop, for you or as a service

PromethION 24/48

Ultra-high throughput, on-demand nanopore sequencing, for you or as a service

Sample preparation

Sample QC before sequencing

- Size distribution (200-500bp for Illumina, too long can cause low yields)
- Quantification of sample concentration
  - Avoid overloading of flow cell (too much material leads to overclustering)
- RNA-seq: RNA Integrity Number (RIN), a measure of RNA degradation
- Sequence diversity (determines %PhiX spike-in)
- Beware of using sample names beginning with a number (R HATES it)

Getting your data

Filetypes

- Files delivered depend on platform:
    - FASTQ
    - QC package: read statistics, quality scores
    - PacBio (subreads BAM, consensus reads BAM or FASTQ files)

Can also get even raw (BCL) or secondary analysis files by request

- Alignment (.bam) files
- Variant calls
- Bigwigs

Getting your data

Modes of data delivery

Each of the NCI cores has a different method for distributing the output from sequencing runs. All typically deliver fastq.gz files and a QC report. Some may also deliver alignment data or more. The data is almost always composed of multiple files and is delivered as a single archived file (tar or zip).

Delivery vehicles:

- Globus Data Transfer Utility
- NCI Data Management Environment (DME)
- HTML download via browser, wget or curl.