Illumina's sequencing by synthesis

Analysis of Next-Generation Sequencing Data

Friederike Dündar

Applied Bioinformatics Core

Slides at https://bit.ly/2CUdS9z[1]

January 22, 2019

[1] http://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2018/

DNA Sequencing Overview & Recap

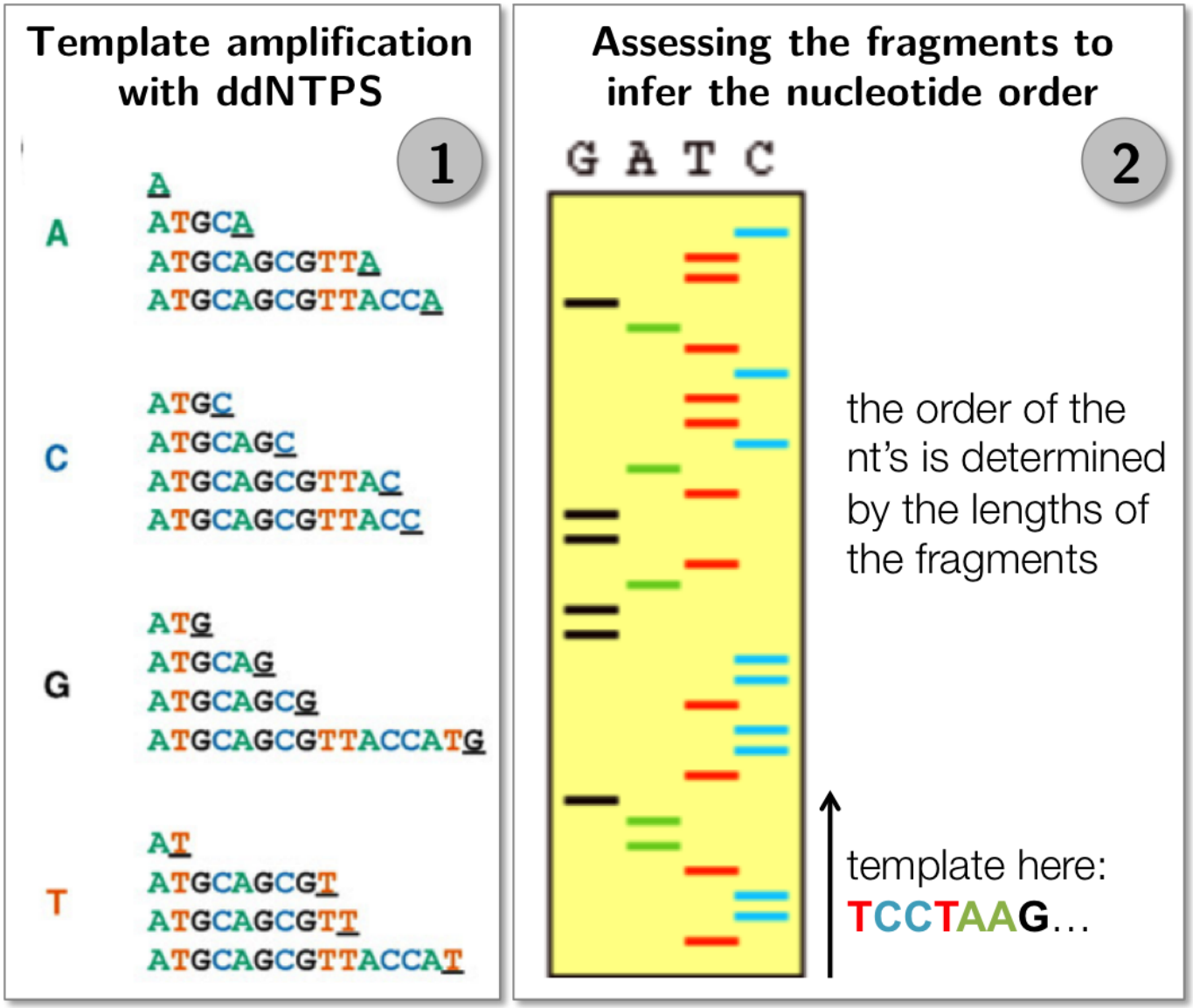
Three Generations of DNA Sequencing

- 1st: **Sanger sequencing** [Sanger et al., 1977]
  - Cost per Mb: **USD 2,400**
  - Read length: 800 bp
  - Run time: 3 hrs
- 2nd: **Next-generation** or **high-throughput** sequencing [Illumina]
  - Cost per Mb: (less than) **USD 0.07**
  - Read length: 50-150 bp
  - Run time: 10 days
- 3rd: **Single-molecule** and/or **long-read** sequencing [PacBio]
  - Cost per Mb: **USD 0.13-0.6**
  - Read length: 1.4 kb
  - Run time: 0.5-2h

Ease-of-use and through-put have been dramatically increased at the cost of (some) accuracy.

1st Generation: Sanger Sequencing

- based on **chain termination** using ddNTPs [Sanger et al., 1977]
- single fragment sequenced at a time (=< 1,000 bp length)



Three Generations of DNA Sequencing

- 1st: **Sanger sequencing** [Sanger et al., 1977]
  - Cost per Mb: **USD 2,400**
  - Read length: 800 bp
  - Run time: 3 hrs
- 2nd: **Next-generation** or **high-throughput** sequencing [Illumina]
  - Cost per Mb: (less than) **USD 0.07**
  - Read length: 50-150 bp
  - Run time: 10 days
- 3rd: **Single-molecule** and/or **long-read** sequencing [PacBio]
  - Cost per Mb: **USD 0.13-0.6**
  - Read length: 1.4 kb
  - Run time: 0.5-2h

Ease-of-use and through-put have been dramatically increased at the cost of (some) accuracy.

Three Generations of DNA Sequencing

Details of first, second, and third generation sequencing technologies with respect to their cost per megabase, instrument cost, read length, and accuracy

Platform	Company	Cost per megabase (USD)	Cost per instrument (USD)	Read-length (bp)	Run time	Throughput	Raw accuracy
First generation							
Maxam-Gilbert	NA	–	–	_	2h	Low	–
Sanger	Applied Biosystems	2400	95,000	800	3h	Low	99.9999%
Second generation							
GS FLX	454 Life Sciences, Roche	~60.0	500,000	700	24 h	High	99.9%
SOLiD	Life Technologies	~0.13	495,000	35	8–14 days	Very high	99.94%
Genome Analyzer	Solexa, Illumina	~0.07	690,000	36	10 days	Very high	>98.5%
Polonator	Dover	~1.00	155,000	13	8–10 days	High	99.7%
HeliScope	Helicos Biosciences	~1.00	1,350,000	30	7 days	High	>99%
Third generation							
Ion Torrent	DNA Electronics Ltd.	1.00	80,000	200–400	3 h	Moderate	99.2%
CGA	BGI	~0.5–1.00	1200,000	10	6 h	Very high	99.99 %
Pacific Bio RS	Pacific biosciences	0.13–0.6	695,000	1400	0.5–2 h	Moderate	88.0%
Oxford Nanopore	Oxford technologies	Not yet calculated	750,000	Up to 4Tb	Upto 48h	Very high	99.99%

Table from Keith [2017]

Main steps of DNA sequencing experiments

TEMPLATE PREPARATION

1. **Obtaining the molecules of interest**: DNA, RNA, nucleotide-protein complexes
2. **Library preparation**: fragmentation and ligation of sequencing adapters
3. **Amplification**

SEQUENCING

Sequencing by Synthesis

read length

single-end vs. paired-end

number of reads

BIOINFORMATICS

Base calling

Alignment
Identifying loci of the sequenced fragments

Additional processing

Interpretation

Template preparation

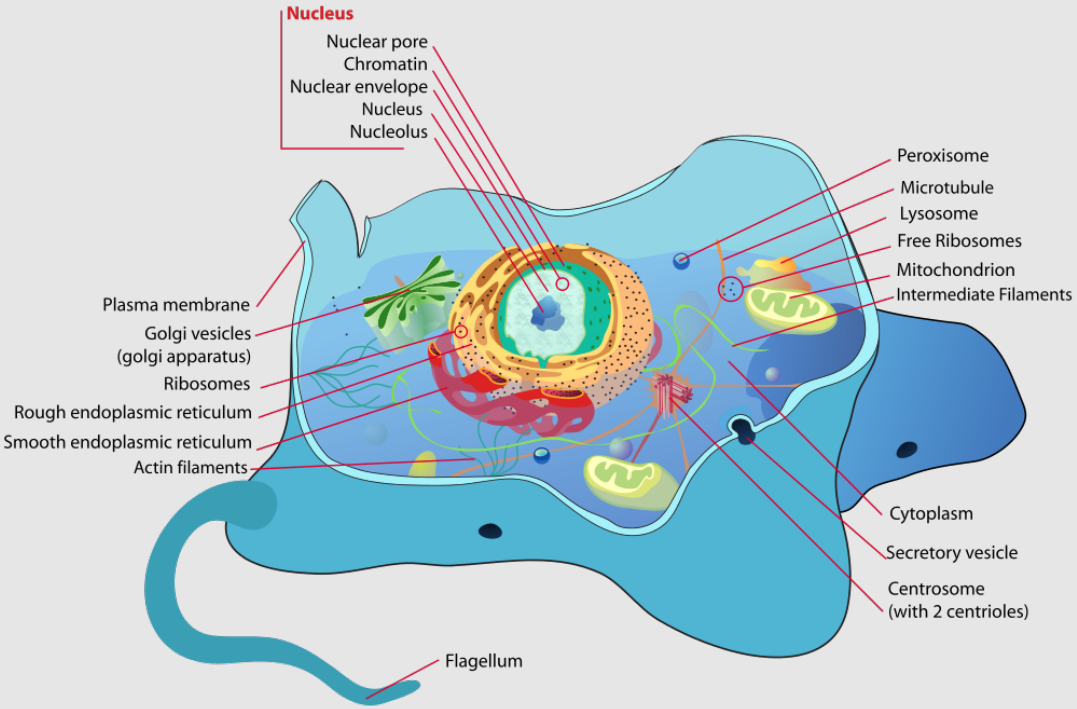
Template preparation

1. Nucleic acid **extraction**
2. **Library preparation**: adding adapters for sequencing
3. **Clonal amplification**: making sure the signal is going to be strong enough

Template preparation

1. DNA/RNA extraction

Nucleic acids must be purified out of a mix of all sorts of organic and inorganic molecules.



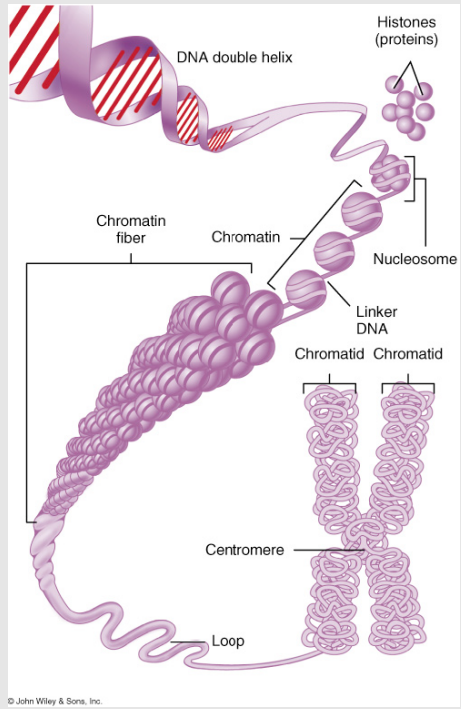
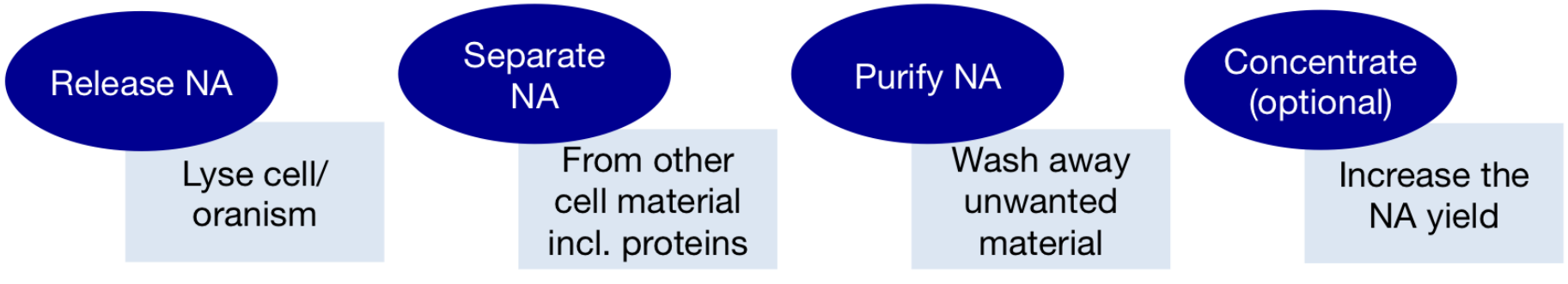


Fig. from: https://en.wikipedia.org/wiki/Eukaryote

1. DNA/RNA extraction

Basic steps

Goal: Little or **no degradation** and complete profiling of the **entire length** of each DNA or RNA molecule.



1. DNA/RNA extraction

Lysis

- Lysis = release of nucleic acids (NA) from cells/nuclei using
  - salt solutions, detergents, lytic enzymes
  - physical forces: mechanical force, heat, freezing
- different cells (bacteria, plant cells, mammalian tissues. . .) have very **different optimal lysis properties** (see Thatcher [2015]!)

1. DNA/RNA extraction

Lysis

- Lysis = release of nucleic acids (NA) from cells/nuclei using
  - salt solutions, detergents, lytic enzymes
  - physical forces: mechanical force, heat, freezing
- different cells (bacteria, plant cells, mammalian tissues. . .) have very **different optimal lysis properties** (see Thatcher [2015]!)

1. DNA/RNA extraction

Lysis

- Lysis = release of nucleic acids (NA) from cells/nuclei using
  - salt solutions, detergents, lytic enzymes
  - physical forces: mechanical force, heat, freezing
- different cells (bacteria, plant cells, mammalian tissues. . .) have very **different optimal lysis properties** (see Thatcher [2015]!)

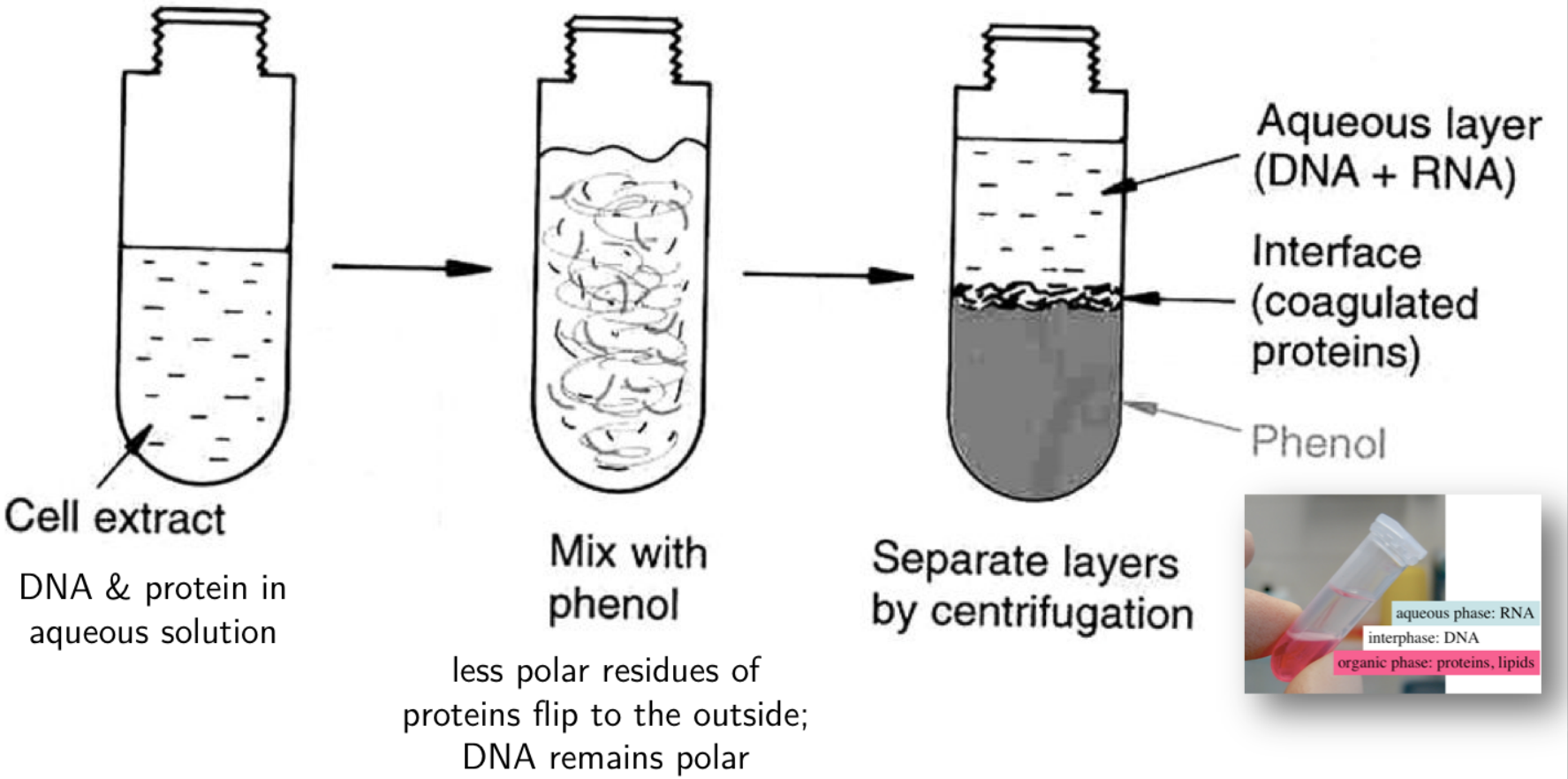
1. DNA/RNA extraction

Lysis

- Lysis = release of nucleic acids (NA) from cells/nuclei using
  - salt solutions, detergents, lytic enzymes
  - physical forces: mechanical force, heat, freezing
- different cells (bacteria, plant cells, mammalian tissues. . .) have very **different optimal lysis properties** (see Thatcher [2015]!)

1. DNA/RNA Extraction

Separate NA: Liquid-liquid extraction (Phenol-Chloroform)



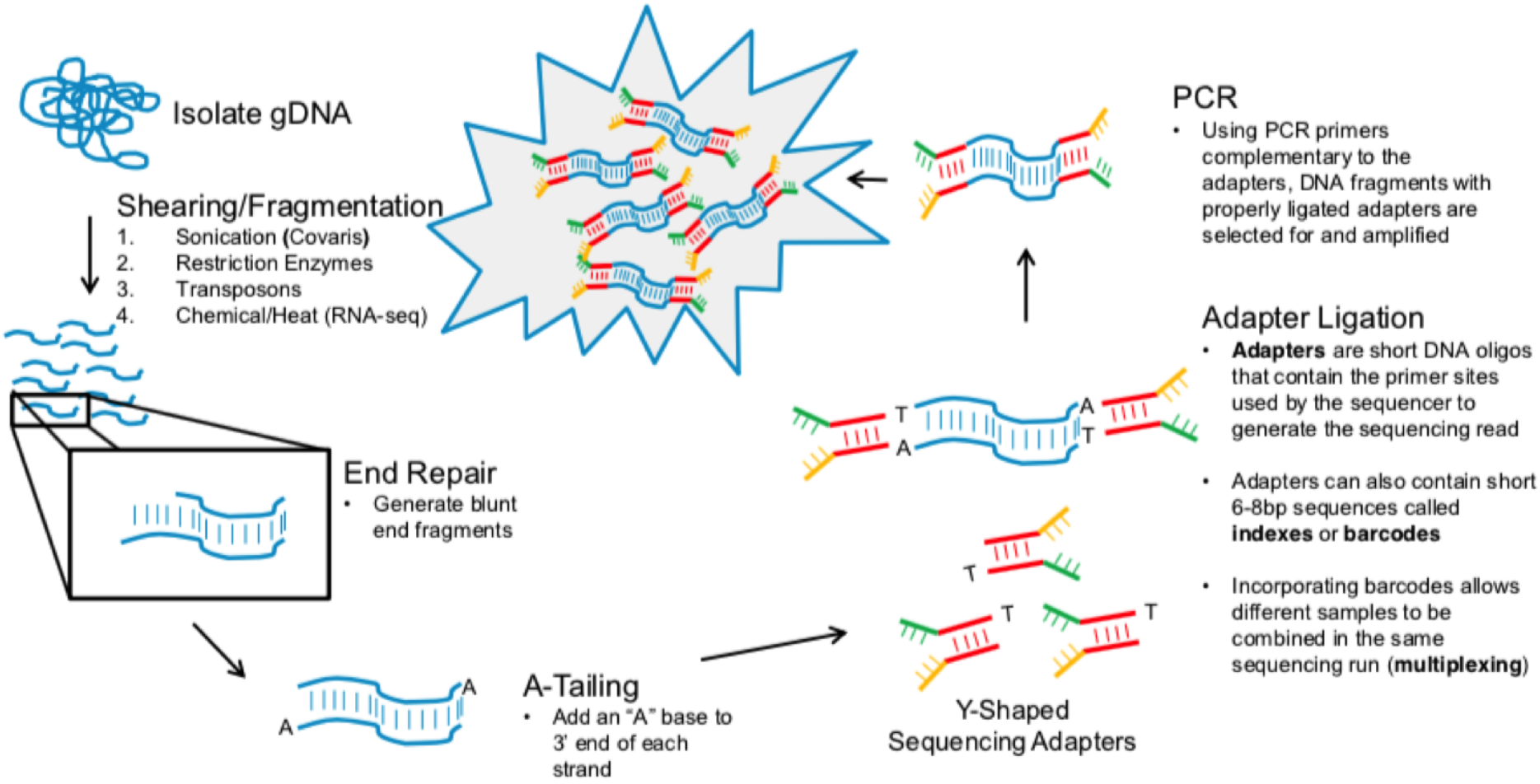
http://slideplayer.com/slide/10173005/34/images/28/Genomic+DNA+prep:+removing+proteins+and+RNA.jpg

1. DNA/RNA extraction

Separate NA: Solid-phase DNA extraction

- liquid-liquid extraction relies on toxic chemicals and is difficult to automate/standardize
- solid phase extraction is based on
  - silica molecules (e.g. within a column or as magnetic silica-based beads) that will
  - bind the nucleic acids
  - in the presence of a chaotropic buffer [a]
- non-DNA components are washed away, before releasing the DNA from the solid adsorber

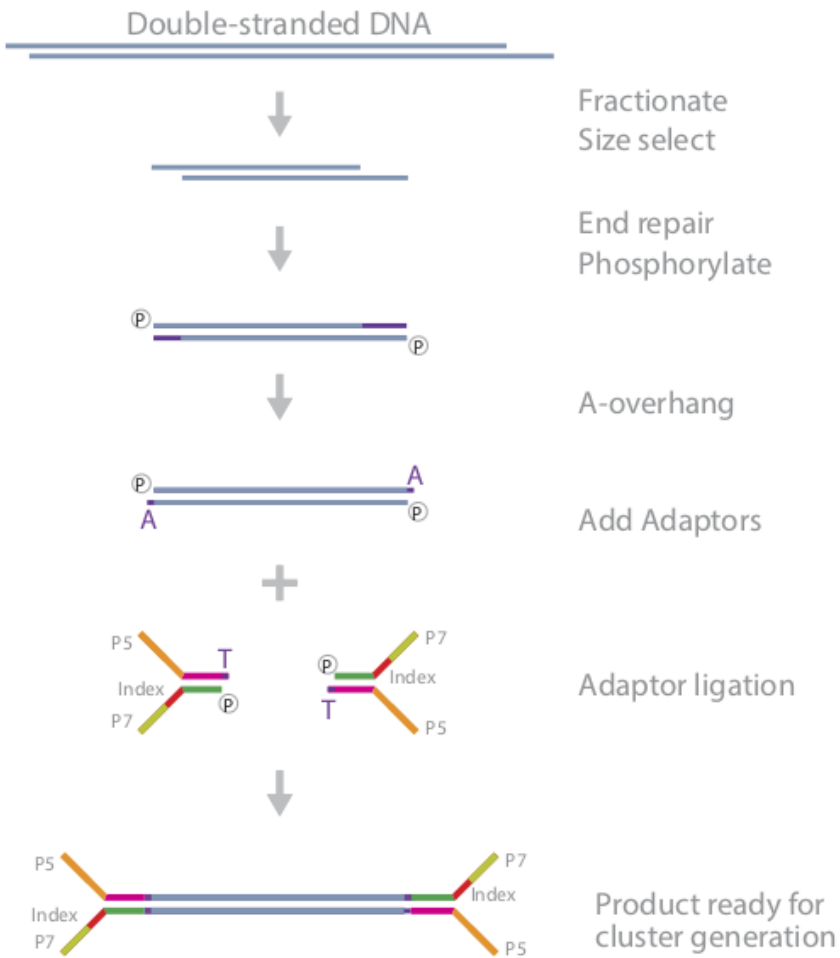
2. Library preparation: getting the NA molecules ready for the sequencer



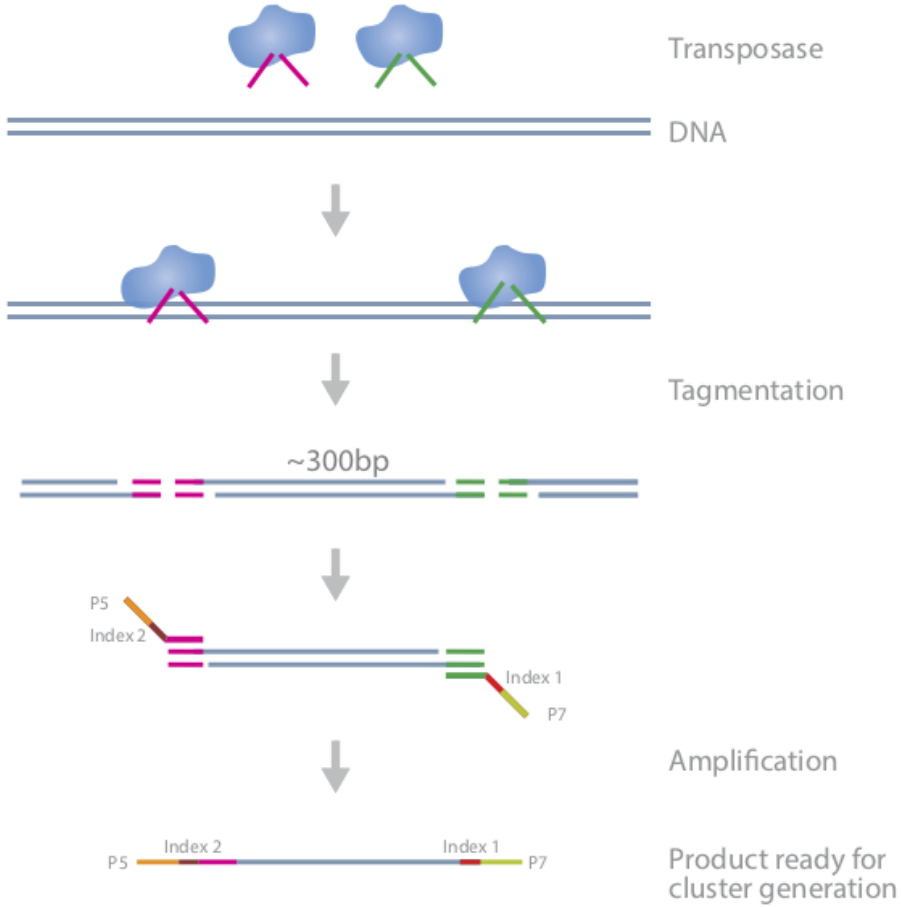
https://www.agilent.com/cs/library/eseminars/public/Next%20Generation%20Sequencing%20101.pdf

2. Library preparation

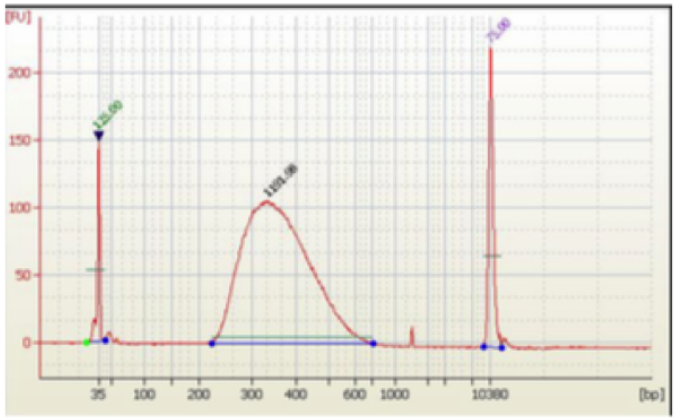
TruSeq Library Prep Protocol



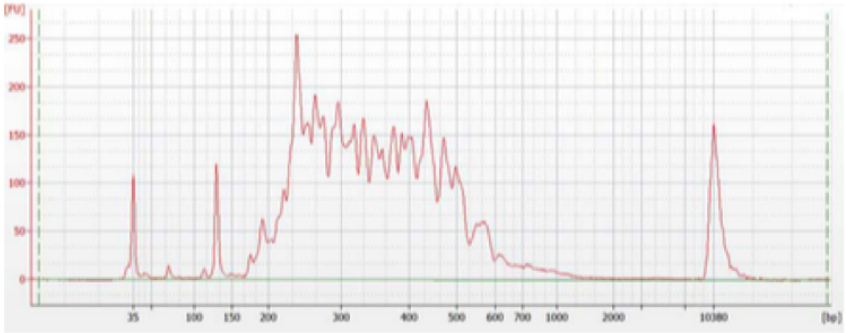
Nextera Library Prep Protocol



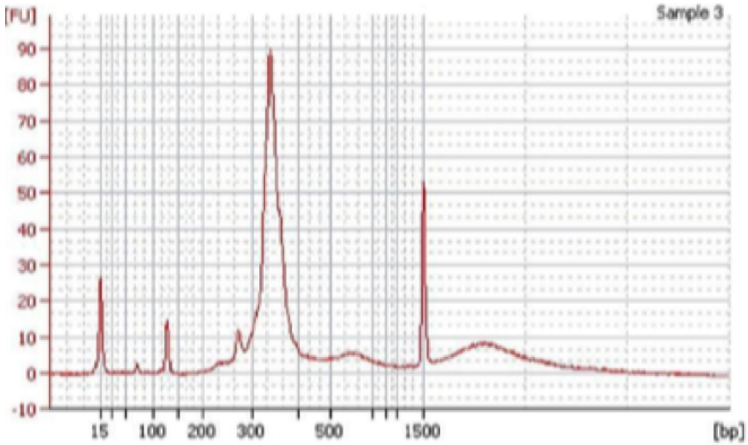
Different library preparations yield different distributions of PCR fragment sizes



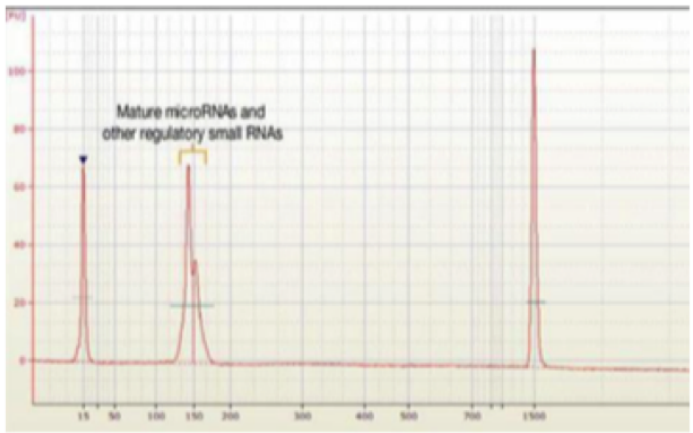
Agilent SureSelect Library Prep



Agilent Haloplex Library Prep



TruSeq Custom Amplicon Library
(adapted from Illumina protocol)



TruSeq Small RNA Library Prep
(adapted from Illumina Protocol)

https://www.agilent.com/cs/library/eseminars/public/Next%20Generation%20Sequencing%20101.pdf

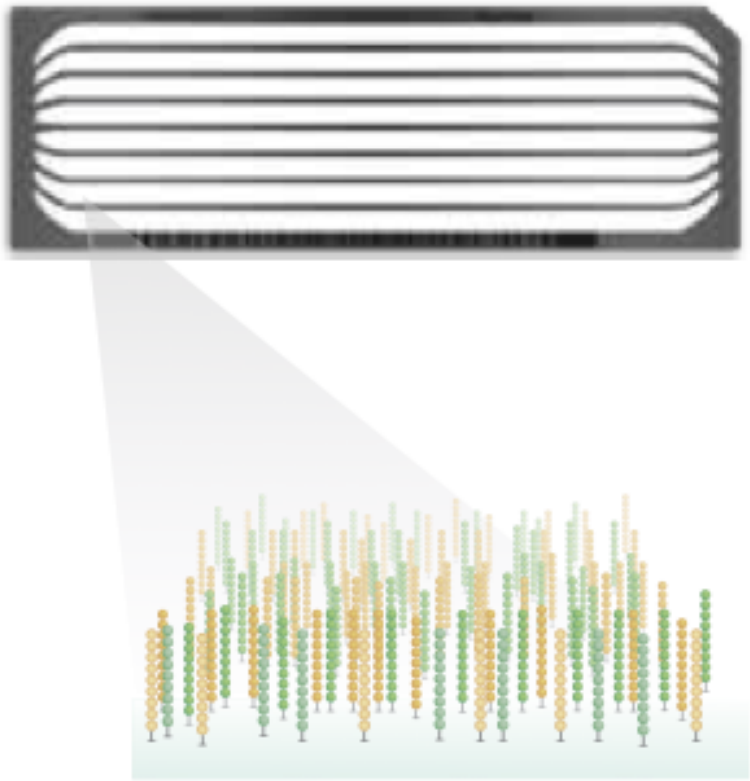
What to consider before choosing a library preparation

1. Sample type
   - High quality DNA? Easy to extract?
   - How much?
2. Experiment goal
   - RNA-seq, ChIP-seq, variant identification, . . . ?
3. Beware of excess PCR cycles!

Library preps all come with their own advantages and disadvantages! Know what to look for during and talk to other people (in your lab, the sequencing facility, online. . .)!

3. Clonal amplification = cluster generation

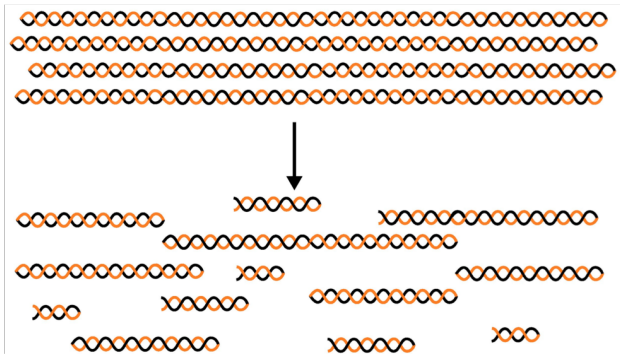
Flowcell



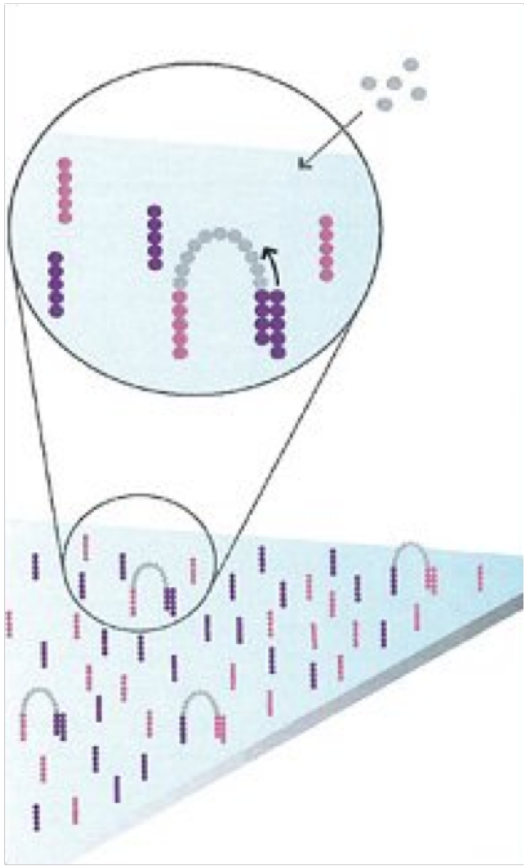
To generate strong signals during sequencing, every fragment is "cloned", yielding physically separate clusters of DNA fragments with identical sequences.

Ideally, the fragments represent the full genome.

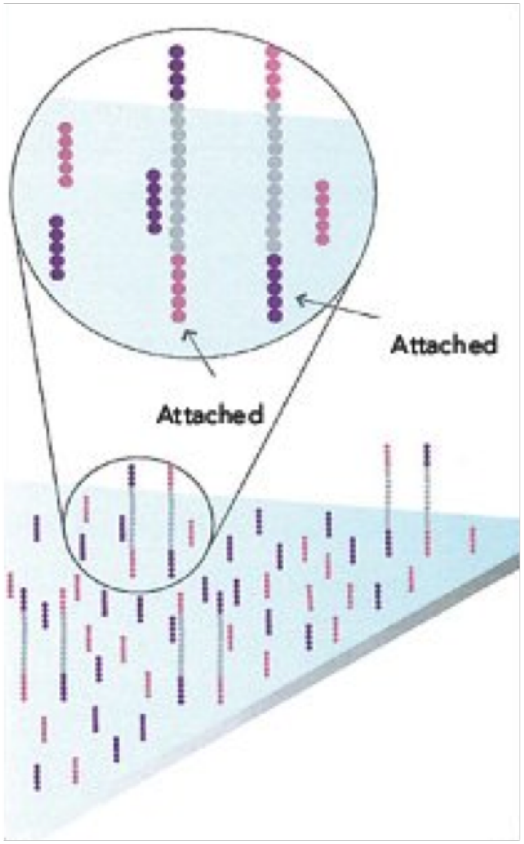
Clusters



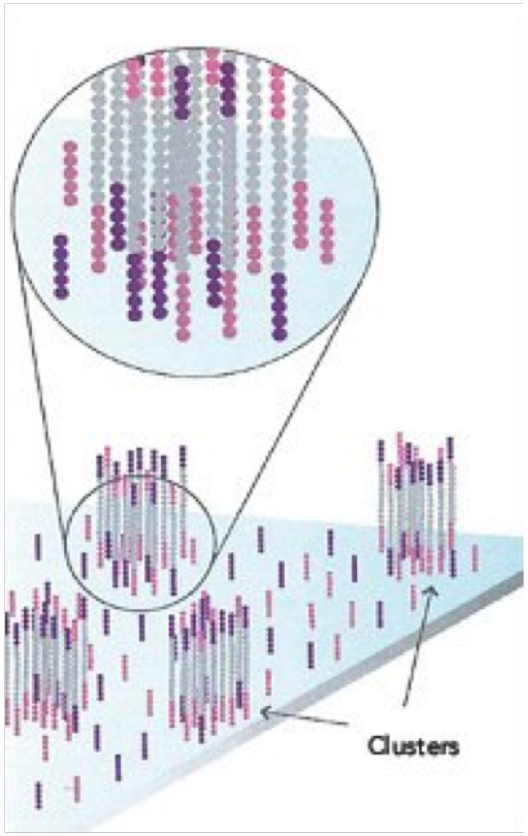
3. Clonal amplification = cluster generation



bridge amplification



denaturation



cluster generation

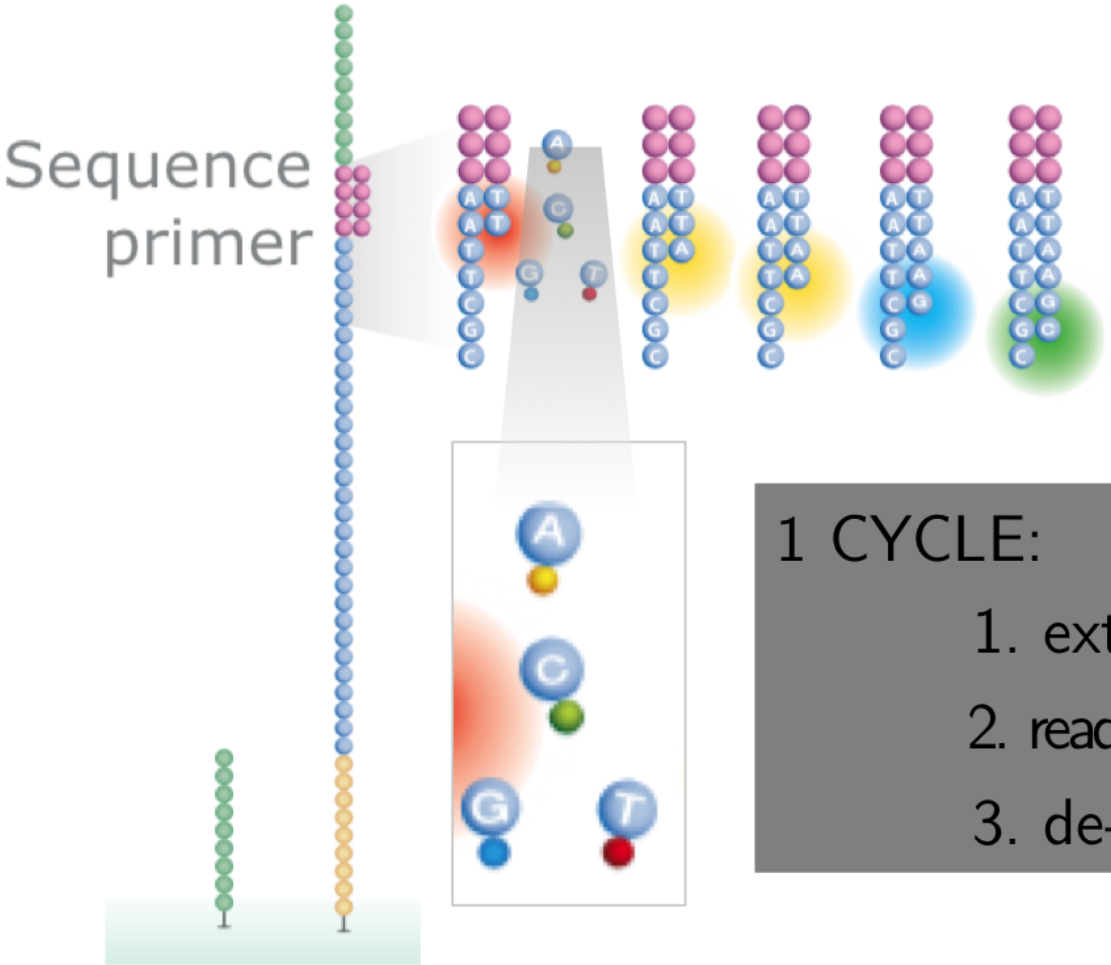
removal of complementary strands
→ identical fragment copies remain

http://informatics.fas.harvard.edu/test-tutorial-page/

Sequencing-by-synthesis

Identifying the order of the nucleotides for every fragment

Illumina's sequencing is based on **fluorophore-labelled dNTPs** with **reversible** terminator elements that will become incorporated and excited by a laser one at a time.

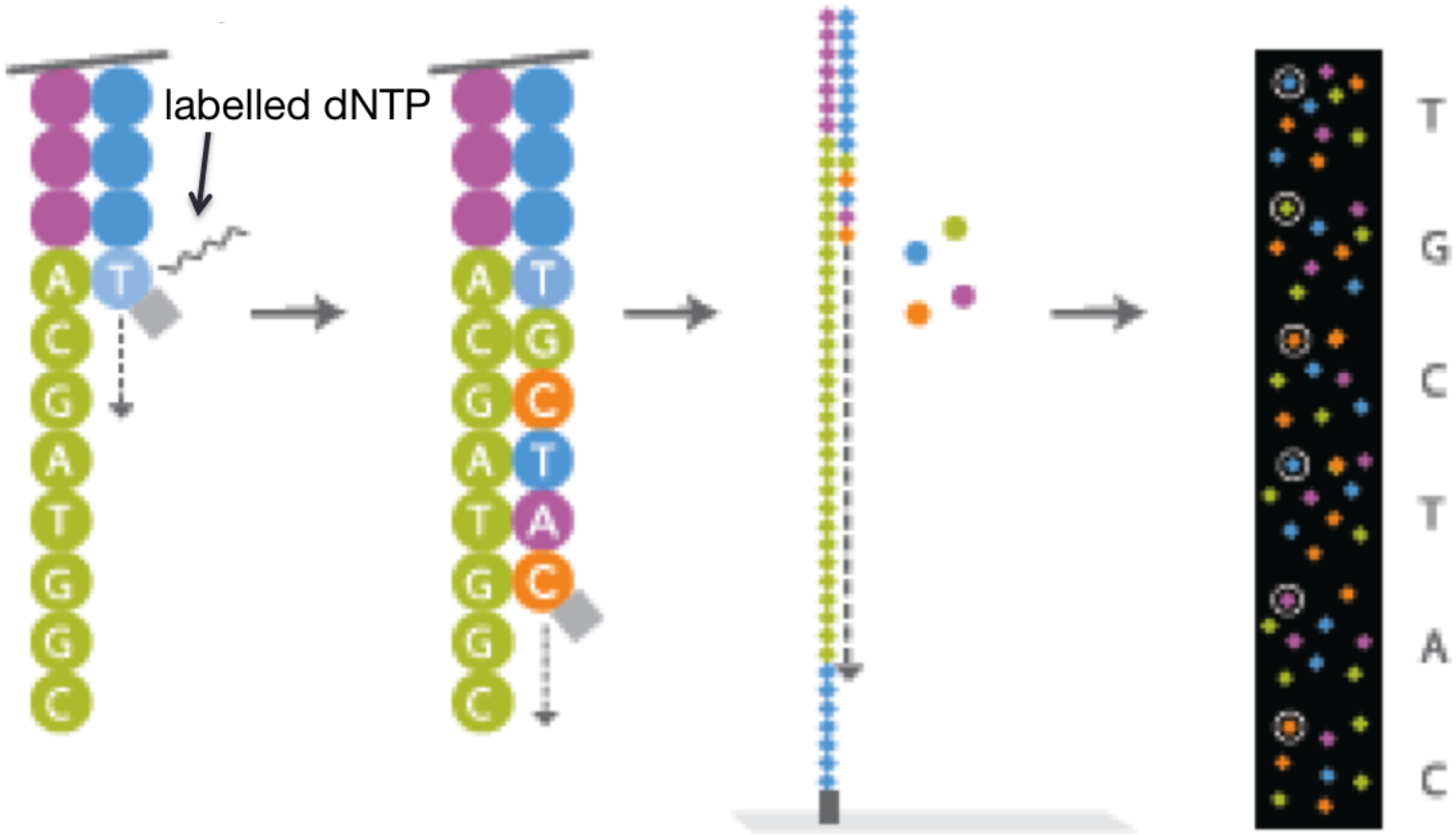


1 CYCLE:

1. extend prev. base
2. read (excite & capture)
3. de-block

The number of cycles determines the read length

50-150 cycle repetitions = 50-150 bp read length



The actual raw data of Illumina sequencing are **images**, but nowadays Illumina will return the **base calls**, i.e. text files of As, Cs, Ts, Gs.

References

See the website

https://bit.ly/2CUdS9

Clinical Chemistry 61:1
89-99 (2015)

Reviews

DNA/RNA Preparation for Molecular Detection

Stephanie A. Thatcher[1*]

Managing Sequence Data

Christopher O'Sullivan, Benjamin Busby, and Ilene Karsch Mizrachi

Abstract

Nucleotide and protein sequences are the foundation for all bioinformatics tools and resources. Researchers can analyze these sequences to discover genes or predict the function of their products. The INSDC (International Nucleotide Sequence Database—DDBJ/ENA/GenBank + SRA) is an international, centralized primary sequence resource that is freely available on the Internet. This database contains all publicly available nucleotide and derived protein sequences. This chapter discusses the structure and history of the nucleotide sequence database resources built at NCBI, provides information on how to submit sequences to the databases, and explains how to access the sequence data.

References

Jonathan M. Keith, editor. *Bioinformatics - Volume I: Data, Sequence Analysis, and Evolution*. Humana Press, methods in edition, 2017.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 1977. doi: 10.1073/pnas.74.12.5463.

Stephanie A. Thatcher. DNA/RNA preparation for molecular detection. *Clinical Chemistry*, 2015. doi: 10.1373/clinchem.2014.221374.