Illumina's sequencing by synthesis

Analysis of Next-Generation Sequencing Data

Friederike Dündar

Applied Bioinformatics Core

Slides at https://bit.ly/2T3sjRg[1]

January 21, 2020

Weill Cornell Medicine

[1] https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2020/

1. DNA Sequencing Overview & Recap
2. Template preparation
3. Sequencing-by-synthesis
4. Single and paired-end reads
5. References

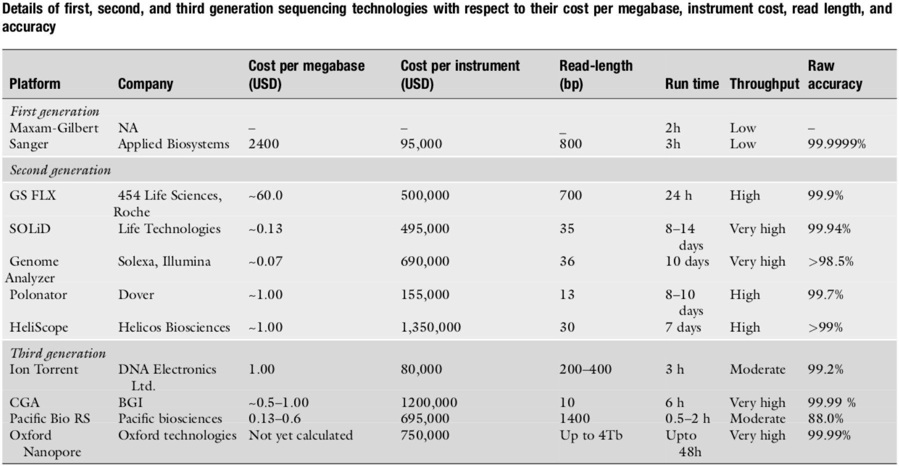
DNA Sequencing Overview & Recap

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.

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%

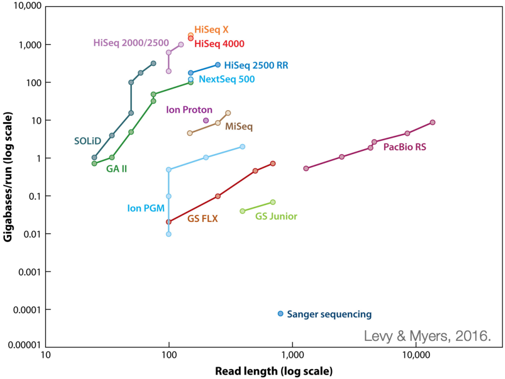
Table from Keith [2017]

NGS = Illumina-based sequencing

In practice, **Illumina's** sequencing platform is by far the most dominant one thanks to its high throughput, constant improvements, and library preparation support (kits).

Since acquiring Solexa in 2006, Illumina has been setting the pace in terms of optimizing yield and costs (e.g. Reuter et al. [2015]).

By mid-2019, PacBio was expected to belong to Illumina, too – on Jan 2, 2020, Illumina stepped away from the deal with a $98M termination fee.



Levy & Myers, 2016.

Main steps of typical NGS experiments

TEMPLATE PREP

Obtaining the molecules of interest: DNA, RNA, nucleotide-protein complexes

⇓

Library preparation: fragmentation and ligation of sequencing adapters

⇓

Amplification

SEQUENCING

Sequencing by **Synthesis**

vs.

Sequencing by **Ligation**

short reads vs. long reads

BIOINFORMATICS

Base calling

⇓

Alignment

Identifying loci of the sequenced fragments

⇓

Additional processing

⇓

Interpretation

Template preparation

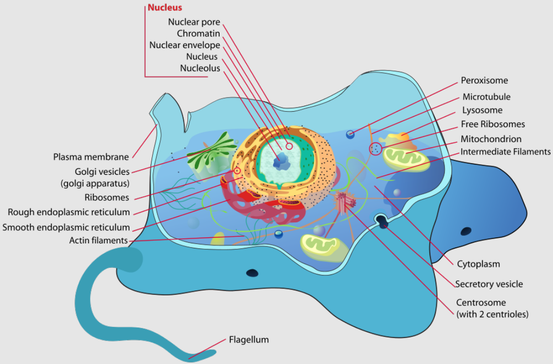
Template preparation

1. Nucleic acid **extraction**
2. **Library preparation** ⇒ 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.



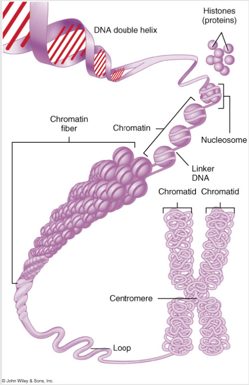
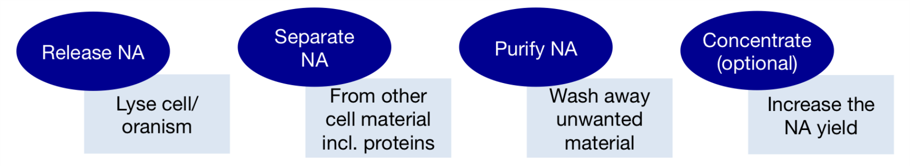


Fig. from: https://en.wikipedia.org/wiki/Eukaryote

1. DNA/RNA extraction

Basic steps

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



Release NA

Lyse cell/ oranism

Separate NA

From other cell material incl. proteins

Purify NA

Wash away unwanted material

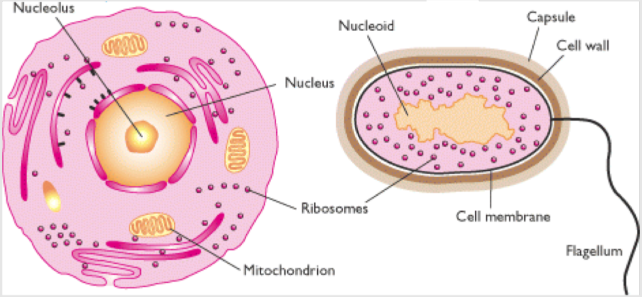
Concentrate (optional)

Increase the NA yield

1. DNA/RNA extraction

Lysis

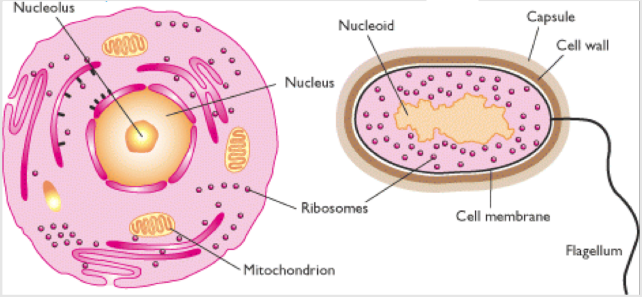
- Lysis = release of nucleic acids (NA) from cells/nuclei (= cell & nucleus destruction) using
  - salt solutions, detergents, lytic enzymes or
  - physical forces: mechanical force, heat, freezing
- different cells (bacteria, plant cells, mammalian tissues. . .) have very **different optimal lysis properties**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



1. DNA/RNA extraction

Lysis

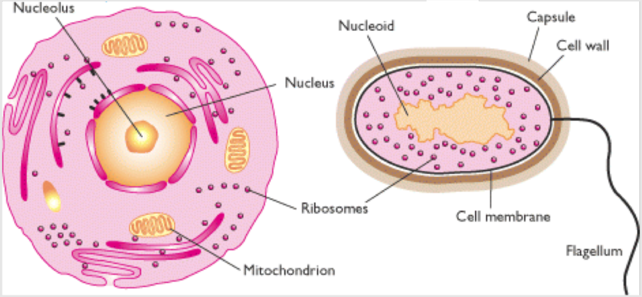
- Lysis = release of nucleic acids (NA) from cells/nuclei (= cell & nucleus destruction) using
  - salt solutions, detergents, lytic enzymes or
  - physical forces: mechanical force, heat, freezing
- different cells (bacteria, plant cells, mammalian tissues. . .) have very **different optimal lysis properties**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



1. DNA/RNA extraction

Lysis

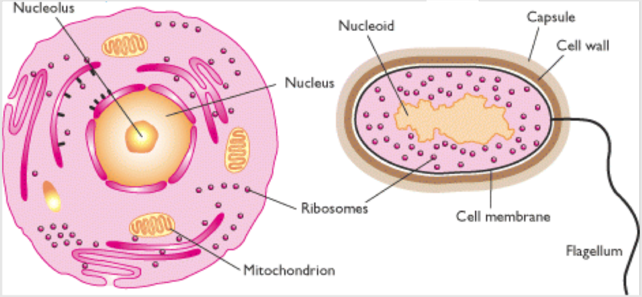
- Lysis = release of nucleic acids (NA) from cells/nuclei (= cell & nucleus destruction) using
  - salt solutions, detergents, lytic enzymes or
  - physical forces: mechanical force, heat, freezing
- different cells (bacteria, plant cells, mammalian tissues. . .) have very **different optimal lysis properties**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



1. DNA/RNA extraction

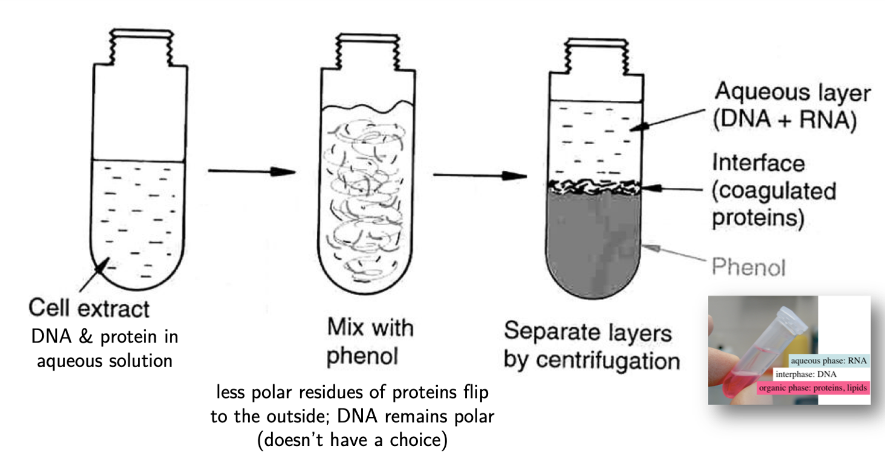
Lysis

- Lysis = release of nucleic acids (NA) from cells/nuclei (= cell & nucleus destruction) using
  - salt solutions, detergents, lytic enzymes or
  - 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)


Cell extract
Mix with phenol
Separate layers by centrifugation
Aqueous layer (DNA + RNA)
Interface (coagulated proteins)
Phenol
aqueous phase: RNA
interphase: DNA
organic phase: proteins, lipids

DNA & protein in aqueous solution

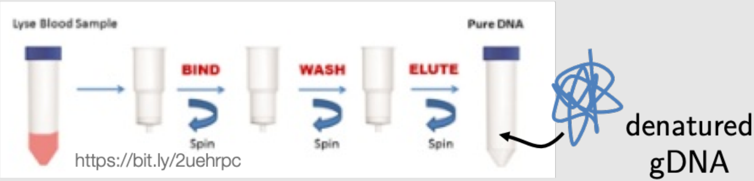
less polar residues of proteins flip to the outside; DNA remains polar (doesn't have a choice)

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

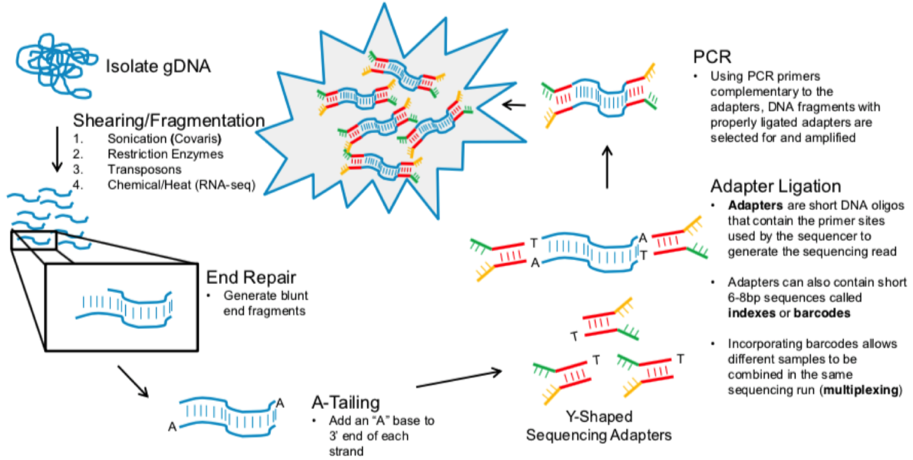
- liquid-liquid extraction relies on toxic chemicals and is difficult to automate/standardize
- solid phase extraction is based on **silica molecules** (e.g. within a column or as magnetic silica-based beads) that will bind the nucleic acids in the presence of a chaotropic buffer [a]
- non-DNA components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



https://bit.ly/2uehrpc

[a] A chaotrope is an ion that disrupts hydrogen bonding, leading to higher protein solubility in water.

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



Isolate gDNA

Shearing/Fragmentation

1. Sonication (Covaris)
2. Restriction Enzymes
3. Transposons
4. Chemical/Heat (RNA-seq)

End Repair

- Generate blunt end fragments

A-Tailing

- Add an "A" base to 3' end of each strand

Y-Shaped Sequencing Adapters

Adapter Ligation

- **Adapters** are short DNA oligos that contain the primer sites used by the sequencer to generate the sequencing read
- Adapters can also contain short 6-8bp sequences called **indexes** or **barcodes**
- Incorporating barcodes allows different samples to be combined in the same sequencing run (**multiplexing**)

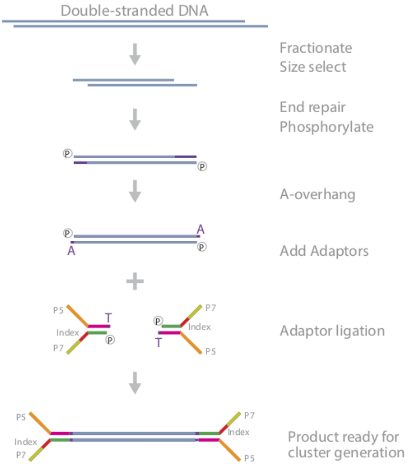
PCR

- Using PCR primers complementary to the adapters, DNA fragments with properly ligated adapters are selected for and amplified

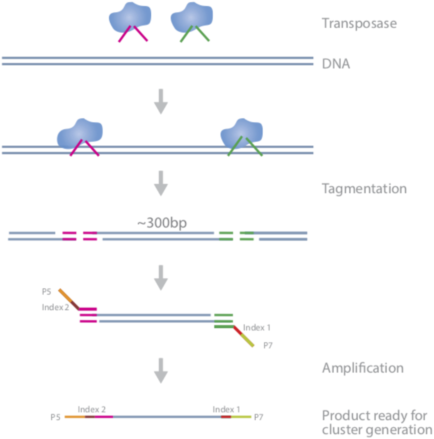
https://www.agilent.com/cs/library/eseminars/public/Next%20Generation%20Sequencing%20101.pdf

2. Library preparation

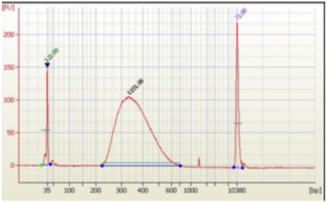
TruSeq Library Prep Protocol



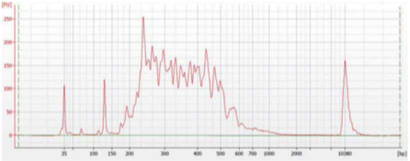
Nextera Library Prep Protocol



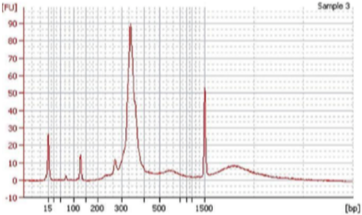
Different library preparations may yield different distributions of PCR fragment sizes – should be suited to the question at hand



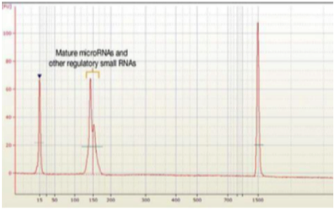
Agilent SureSelect Library Prep



Agilent Haloplex Library Prep



TruSeq Custom Amplicon Library



TruSeq Small RNA Library Prep

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. . .)!

Loading the library onto the **flowcell**

Following library prep, the DNA fragments are floated over the flowcell, which is essentially a glass side full of oligonucleotides that are complementary to the adapters of the library, thereby leading to the physical attachment of the DNA fragments.

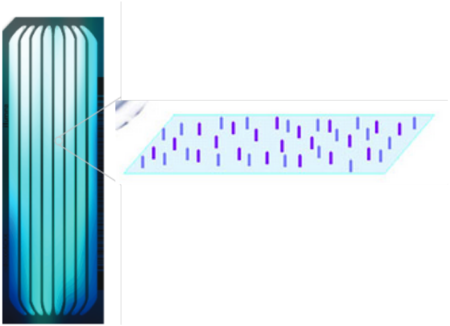
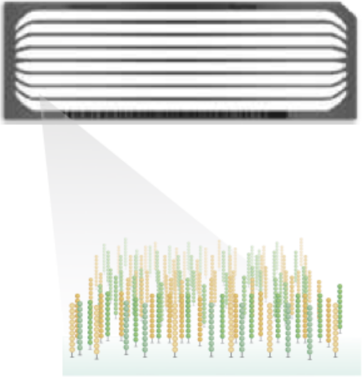


Figure from Illumina Inc [2015]

- 8 microfluidic channels (= "lanes")
- within the channels, the sequencing reaction will happen

3. Clonal amplification = cluster generation

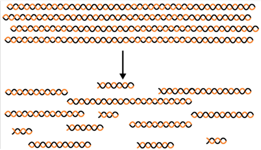
Flowcell



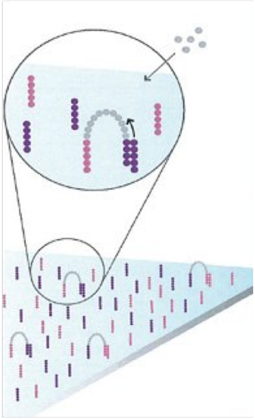
To generate strong signals during sequencing, every fragment is "cloned", yielding physically separate clusters of DNA fragments with identical sequences.

Ideally, the fragments represent the full genome.

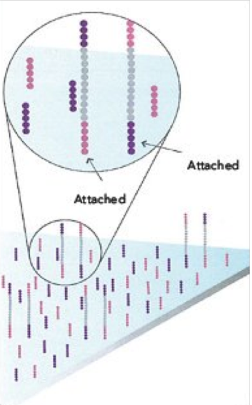
Clusters



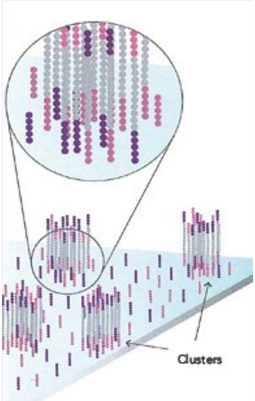
3. Clonal amplification = cluster generation via PCR



bridge amplification



denaturation



cluster generation

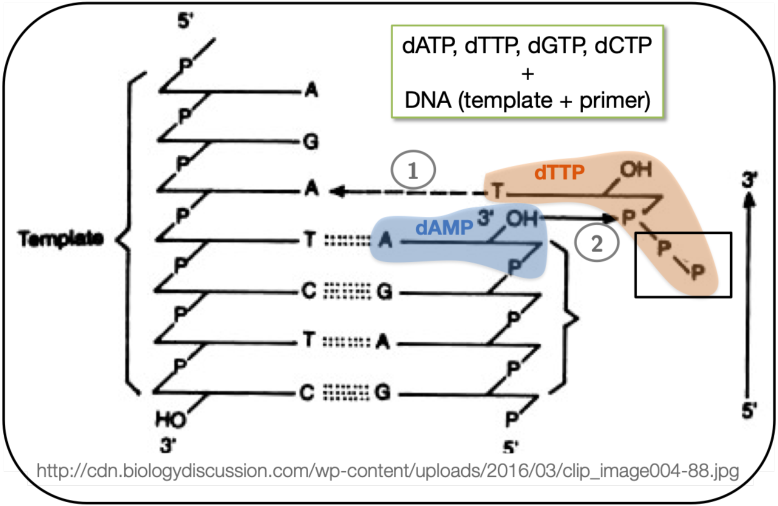
removal of complementary strands
→ identical fragment copies remain

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

Sequencing-by-synthesis

Decoding the DNA: DNA polymerase

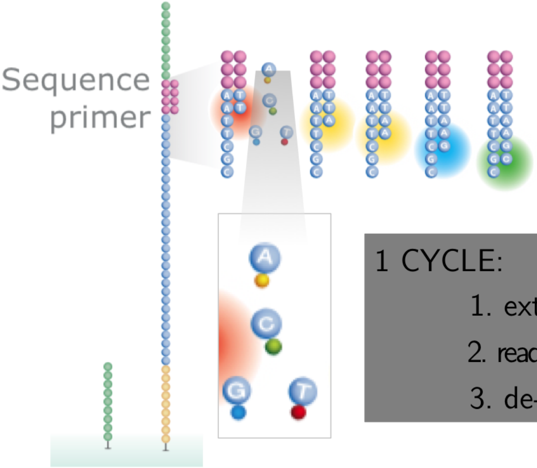
- cannot start DNA synthesis from scratch, always needs **primers**
- relies on the presence of a **template** strand, which is **complemented**



http://cdn.biologydiscussion.com/wp-content/uploads/2016/03/clip_image004-88.jpg

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.

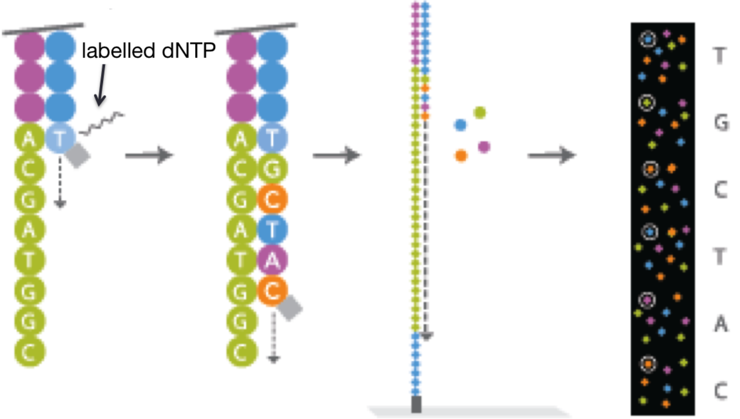


1 CYCLE:

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.

The number of flowcell lanes determines the sequencing depth

Every read represents one cluster on the flowcell.

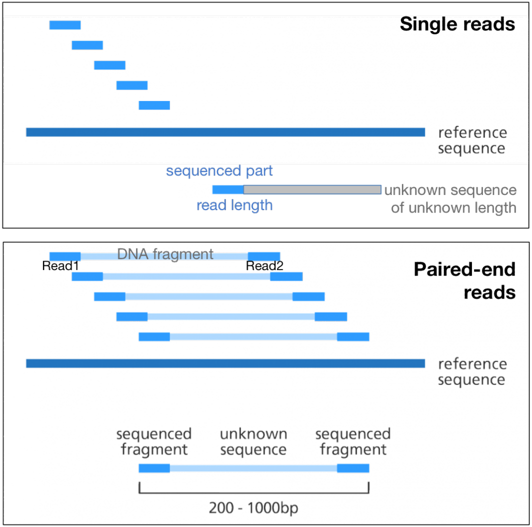
- every cluster = one DNA fragment
- the more clusters one sequences, the more information (= reads) one gets

Machine	Yield per lane
HiSeq4000	400 mio reads
NovaSeq	800-2500 mio reads

Application	Recommended seq. depth
differential gene expression	20 - 50 mio SR, 75 bp
variant calling	30-200x coverage
whole-genome bisulfite sequencing	30x coverage

Single and paired-end reads

Types of reads

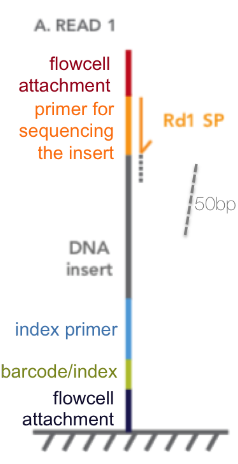


https://www.yourgenome.org/facts/how-do-you-put-a-genome-back-together-after-sequencing

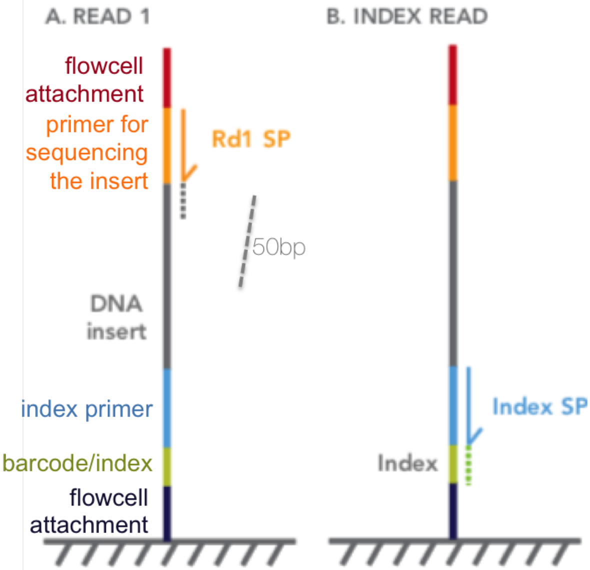
Single reads are the cheaper. Paired-end (PE) reads are helpful for:

- **alignment** along repetitive regions
- chromosomal **rearrangements** and gene fusion detection
- *de novo* genome and transcriptome **assembly**
- precise information about the size of the original fragment (**insert size**)
- PCR duplicate identification

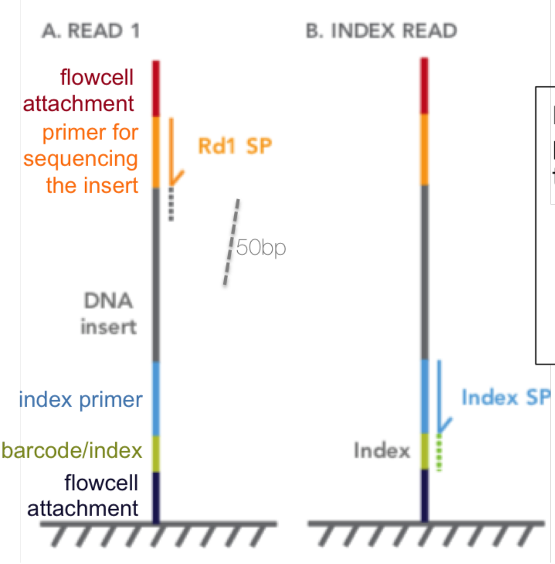
Paired-end read generation

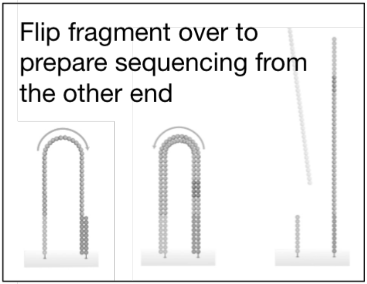


Paired-end read generation

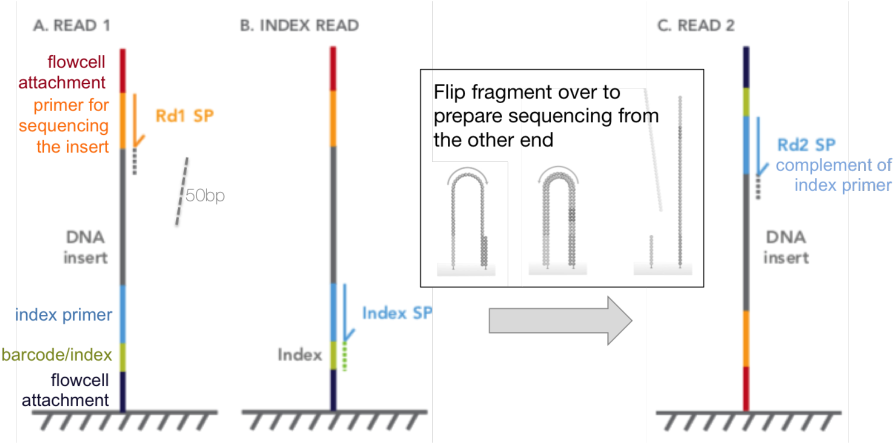


Paired-end read generation




Flip fragment over to prepare sequencing from the other end

Paired-end read generation

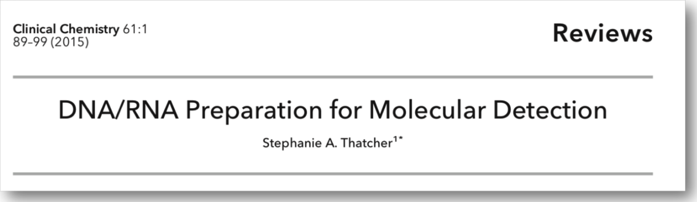


Flip fragment over to prepare sequencing from the other end

References

See the website

https://bit.ly/2T3sjRg

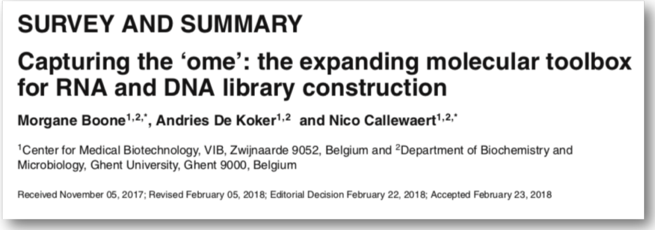


Clinical Chemistry 61:1
89–99 (2015)

Reviews

DNA/RNA Preparation for Molecular Detection

Stephanie A. Thatcher[1*]



SURVEY AND SUMMARY

Capturing the ‘ome’: the expanding molecular toolbox for RNA and DNA library construction

Morgane Boone[1,2,*], Andries De Koker[1,2] and Nico Callewaert[1,2,*]

[1]Center for Medical Biotechnology, VIB, Zwijnaarde 9052, Belgium and [2]Department of Biochemistry and Microbiology, Ghent University, Ghent 9000, Belgium

Received November 05, 2017; Revised February 05, 2018; Editorial Decision February 22, 2018; Accepted February 23, 2018

References

Figure taken from the following publications:
Levy and Myers [2016]

Illumina Inc. Patterned Flow Cell Technology. In *Technical Spotlight: Sequencing*, pages 1–2. 2015. URL https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/patterned-flow-cell-technology-technical-note-770-2015-010.pdf.

Jonathan M. Keith, editor. *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution*, volume 1525. Humana Press, methods in molecular biology edition, 2017.

Shawn E. Levy and Richard M. Myers. Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*, 2016. doi: 10.1146/annurev-genom-083115-022413.

Jason A. Reuter, Damek V. Spacek, and Michael P. Snyder. High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4): 586–597, May 2015. doi: 10.1016/j.molcel.2015.05.004.

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.