Genomics of Human Diseases

2019 Dragon Star Bioinformatics Course

Housekeeping issues

- **Course date: July 29 – August 2, 2019**
- **Theme: Genomics of Human Diseases**
- Morning (9am-12pm with break): lecture session on basic concepts and bioinformatics methods to study the genomics of human diseases
- Afternoon (2pm-4pm): computing exercise to practice software tools discussed in the morning
- Lecture notes and computing exercises in English
- The lecture slides and exercises are shared to students. The computing exercise instructions are also available at GitHub: https://github.com/WGLab/dragonstar2019

Housekeeping issues (continued)

- The afternoon computing exercise session will be split into four rooms, with maximum capacity of 110, 110, 55 and 55 students, respectively.
- We will have teaching assistants in each room to address questions and troubleshoot problems.
- Each student will be assigned a separate IP address to connect to a cloud server by SSH.
  - The username and password are `biouser` and `biouser`, respectively.
  - You should change your password using `passwd` command upon logging into the server.
- Each cloud server is only used by one student, and is not shared with the rest of the class.

Acknowledgements

- This course is co-developed by 王凯(宾夕法尼亚大学), 沈宇锋(哥伦比亚大学) ,李明瑶(宾夕法尼亚大学)
- We thank many people who made this course possible:
  - Course organization：刘小乐，李程，杨恩策，高歌，李川昀，孔雷，刘克胜，罗海涛，刘闯，曹瑶
  - Teaching assistants：王欢，韩雅，孙冬青，郑荣斌，万昌鑫，董鑫，李亭亭，吴佳奇，任云晓，曲素素，刘运泽，郭徨纯，陈钊铭
  - Cloud computing: 赵洋，王猛
  - Preparation of the exercise: 刘乾，方立，Jim Havrilla, Abolfazl Doostparast and other lab members
- We thank 北京大学医学部基础医学院、研究生院、医用理学系计算机教研室、信息通讯中心、教室管理服务中心、保卫处 for facilitating the logistics of the course
- We thank 中软国际CIG云智能集团 for providing the cloud computing platform and technical assistance for computing exercises
- We thank all the IBW conference organization committee members and sponsors

Outline of the five days

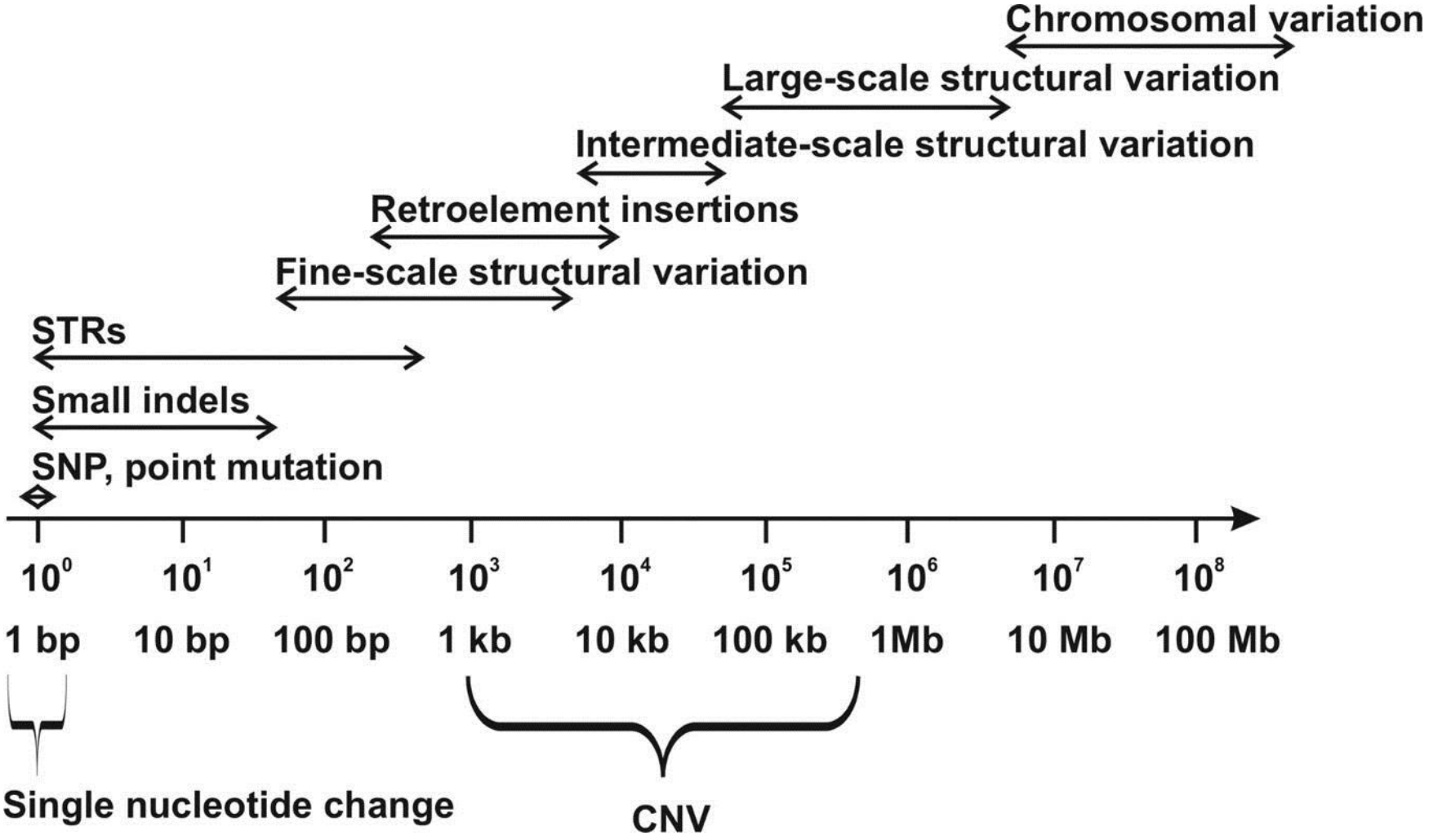
- Day 1: genomic technologies, typical file formats
- Day 2: alignment of sequencing data and genome assembly
- Day 3: detection of structural variants, phenotype-driven variant annotation/interpretation
- Day 4: SNP and sequencing based association studies, rare or de novo variants in human diseases
- Day 5: RNA-Seq and other advanced topics (artificial intelligence, somatic mutation, etc)

For questions and comments before and after class, contact me at kaichop@gmail.com

Genomic technologies in disease studies

2019 Dragon Star Bioinformatics Course (Day 1)

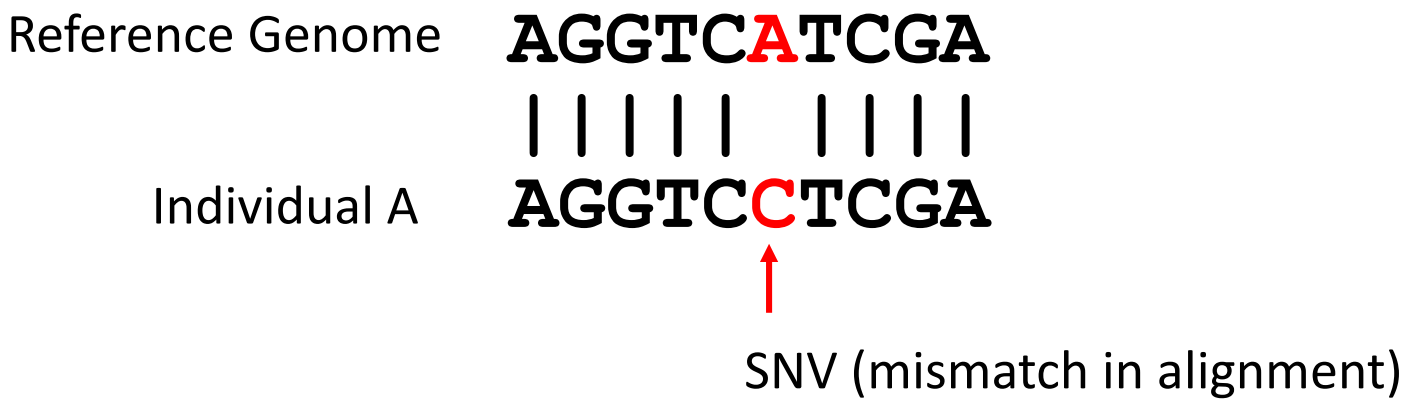
Human Genetic Variation



Pollex et al, Circulation. 2007

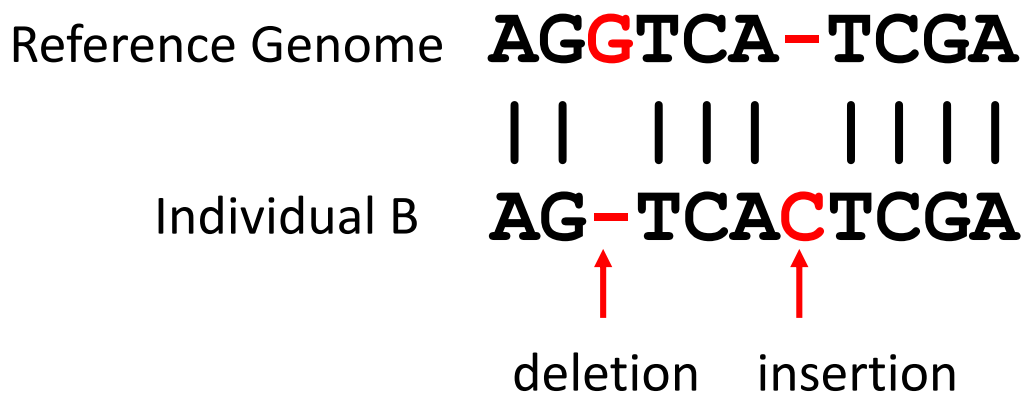
Types of genetic variation

Single Nucleotide Variants (**SNVs**).



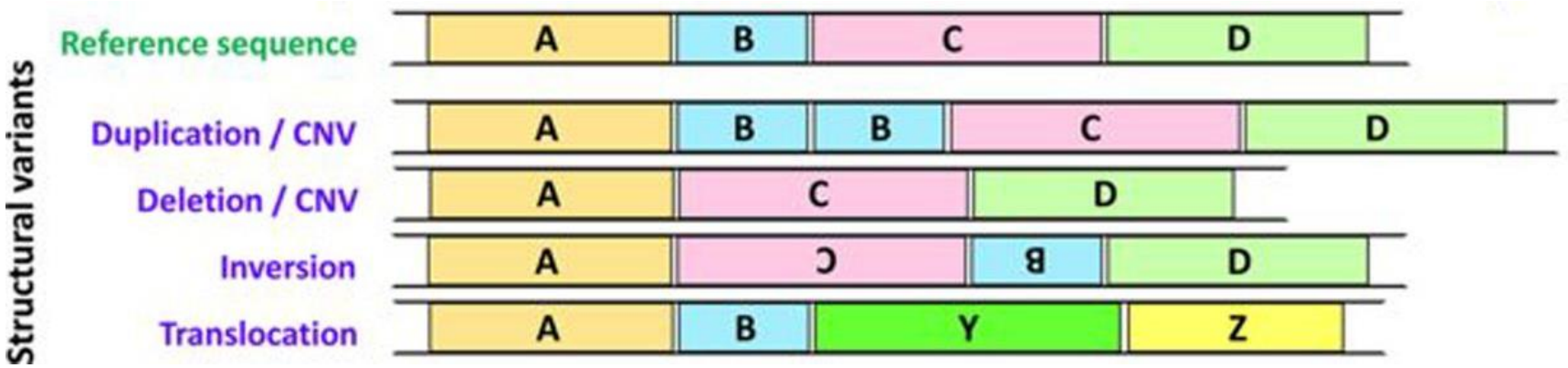
Types of genetic variation

Insertion or deletion (< 50 bp), also known as **Indel**.



Types of genetic variation

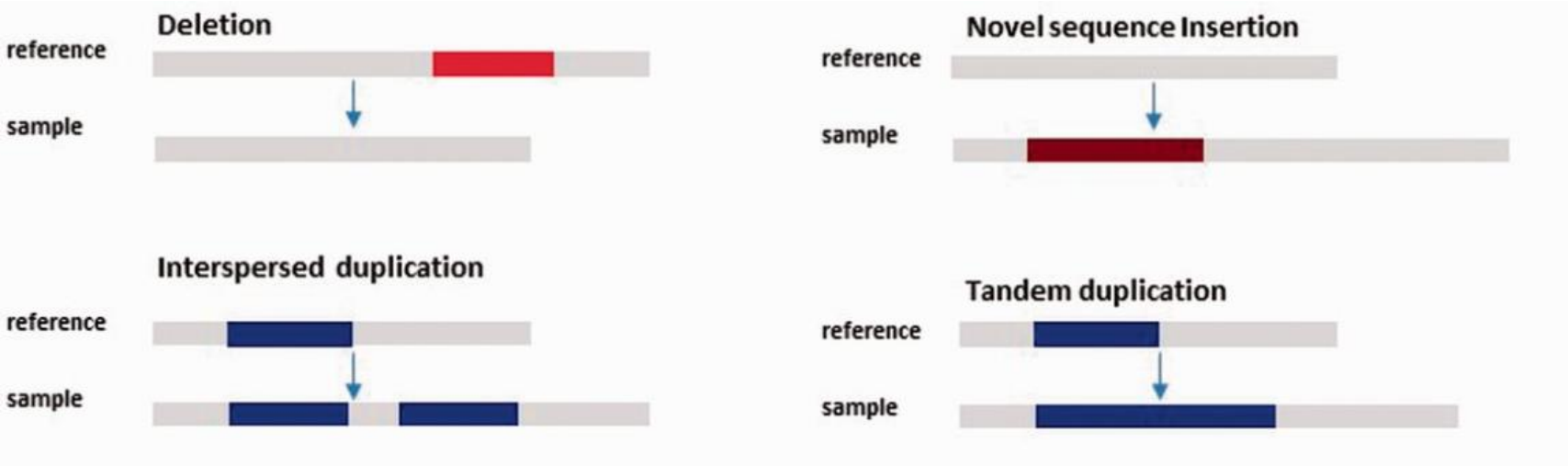
- **Structural Variants** (**SV**): generally defined as a region of DNA that shows a change in
  - Copy number (deletions, insertions and duplications)
  - Orientation (inversions) or
  - Chromosomal location (translocations) between individuals.



Jackson et al, Essays In Biochemistry 2018

Different types of SVs

- SV can be *balanced or unbalanced.*
  - Unbalanced events: deletions/insertions/duplications
  - Chromosomal aneuploidies (such as trisomy 21) are extreme cases of unbalanced SV.

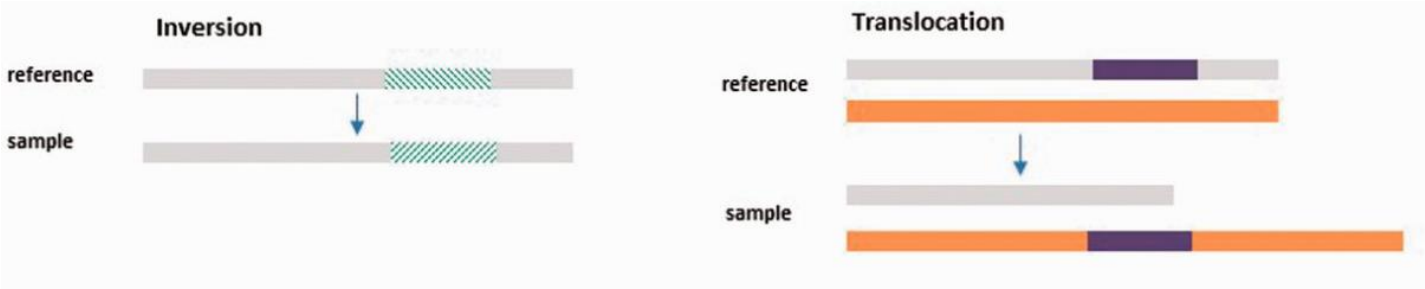


Deletions and duplications are two subtypes of CNVs (**Copy Number Variants**).

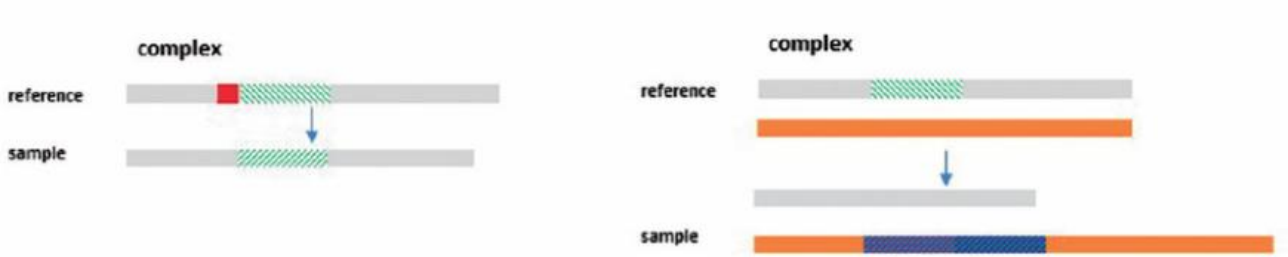
Escaramís G, et al. Briefings in Functional Genomics, 2015

Different types of SVs

- Balanced events do not involve gain or loss of genetic materials
  - Inversions and translocations

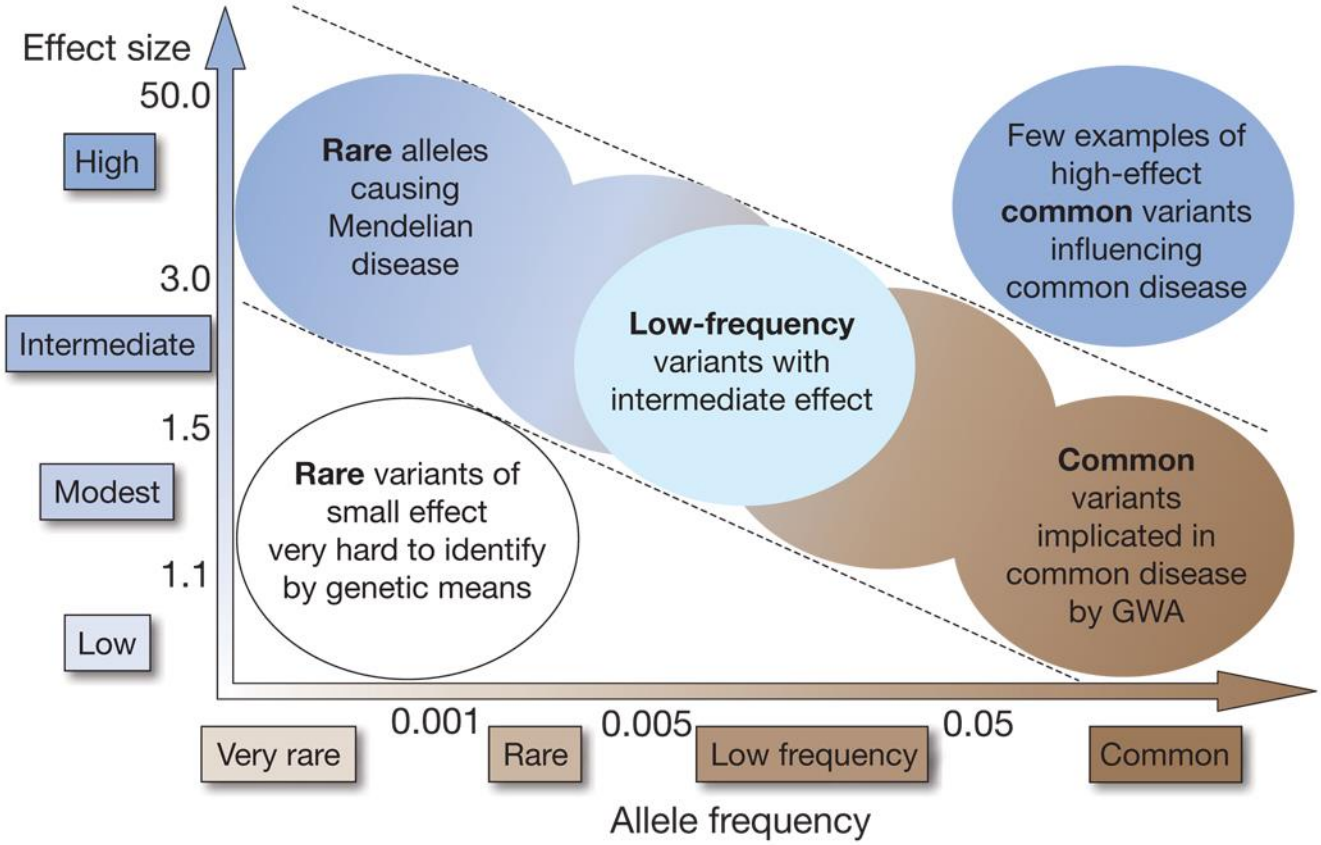


  - Complex SVs (several types together)



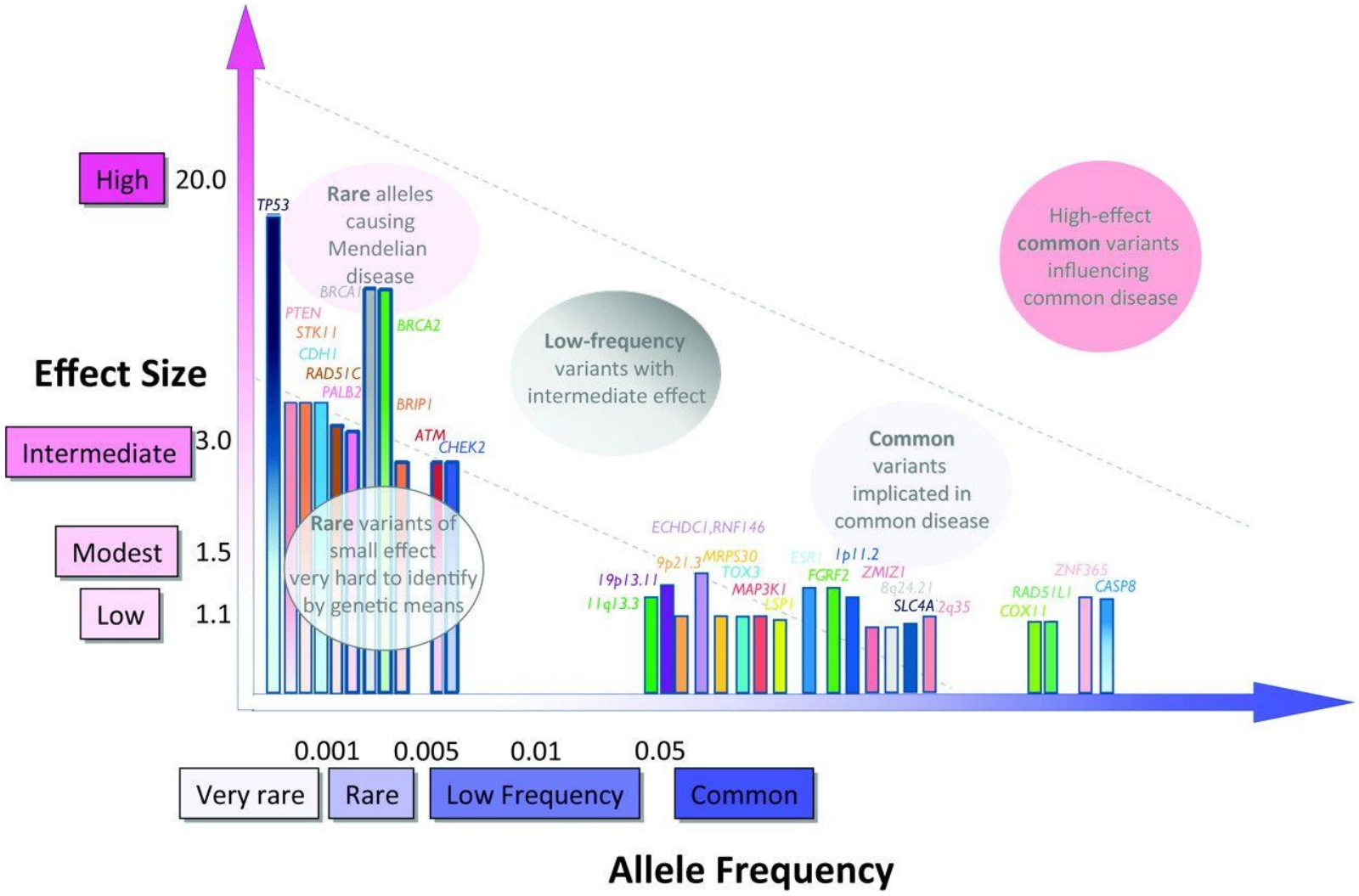
Escaramís G, et al. Briefings in Functional Genomics, 2015

Allele frequency and effect size



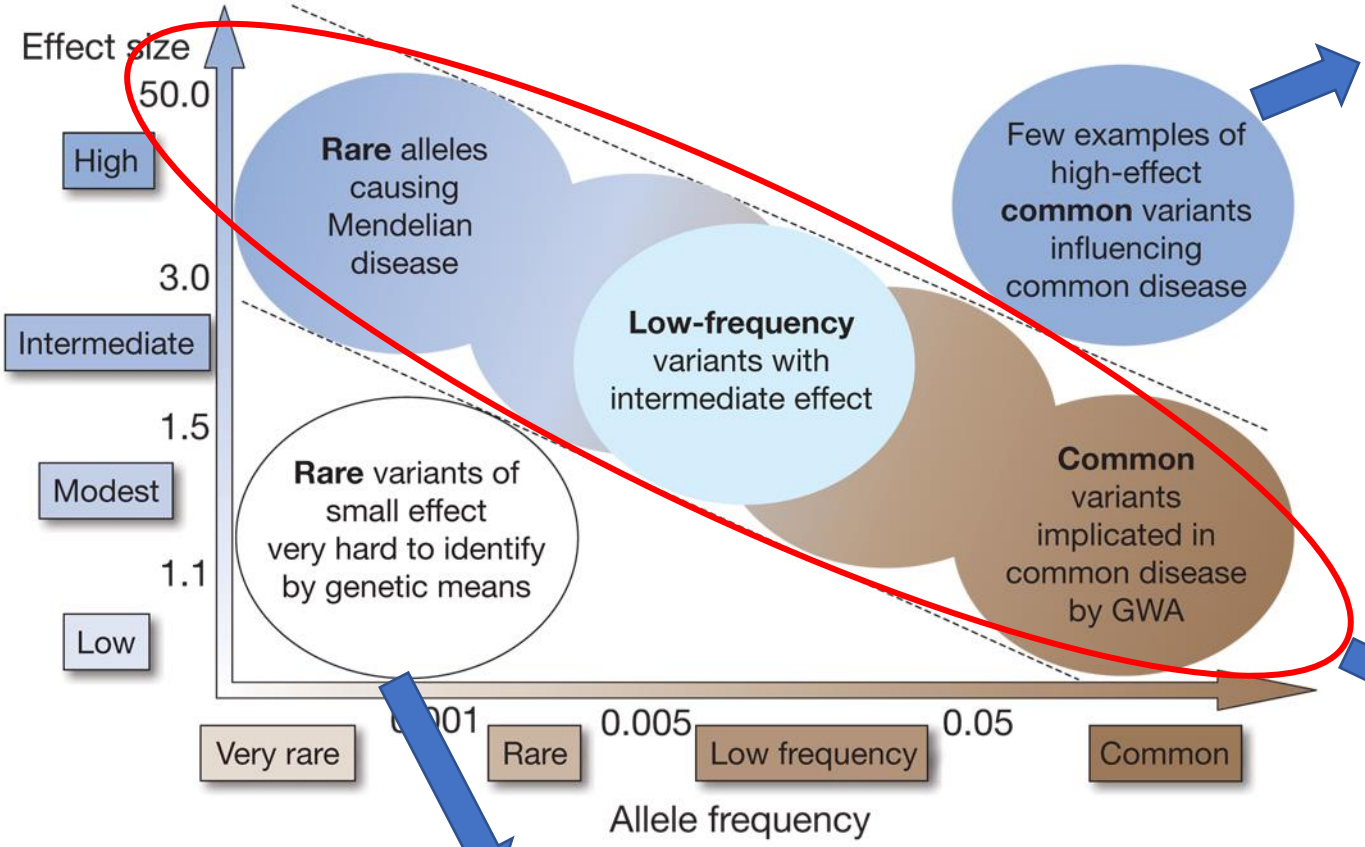
Manolio et al, Nature, 2009

Breast cancer as an example



Hindorff et al, Carcinogenesis, 2011

Variant frequency and effect size



Only a few variants may belong to this category because of purifying selection. Example: APOE4 (a high penetrance variant with modest population contributes to Alzheimer's disease).

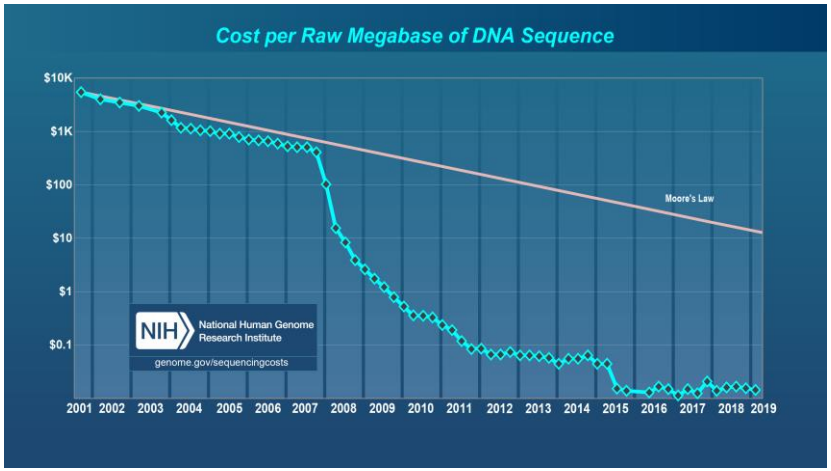
The focus of current human genetic studies

Very hard to detect with GWAS or sequencing, but we usually don't put much attention on them as they contribute to the phenotype so little.

History of DNA Sequencing

Technical milestones

1953: Sequencing of insulin protein[2]
1965: Sequencing of alanine tRNA[4]
1968: Sequencing of cohesive ends of phage lambda DNA[6]
1977: Maxam–Gilbert sequencing[9]
1977: Sanger sequencing[8]
1981: Messing's M13 phage vector[12]
1986–1987: Fluorescent detection in electrophoretic sequencing[14,15,17]
1987: Sequenase[18]
1988: Early example of sequencing by stepwise dNTP incorporation[139]
1990: Paired-end sequencing[23]
1992: Bodipy dyes[140]
1993: *In vitro* RNA colonies[37]
1996: Pyrosequencing[44]
1999: *In vitro* DNA colonies in gels[38]
2000: Massively parallel signature sequencing by ligation[47]
2003: Emulsion PCR to generate *in vitro* DNA colonies on beads[42]
2003: Single-molecule massively parallel sequencing-by-synthesis[33,34]
2003: Zero-mode waveguides for single-molecule analysis[37]
2003: Sequencing by synthesis of *in vitro* DNA colonies in gels[49]
2005: Four-colour reversible terminators[51–53]
2005: Sequencing by ligation of *in vitro* DNA colonies on beads[41]
2007: Large-scale targeted sequence capture[93–96]
2010: Direct detection of DNA methylation during single-molecule sequencing[65]
2010: Single-base resolution electron tunnelling through a solid-state detector[141]
2011: Semiconductor sequencing by proton detection[142]
2012: Reduction to practice of nanopore sequencing[143,144]
2012: Single-stranded library preparation method for ancient DNA[145]



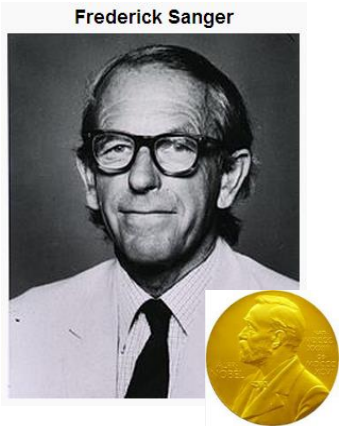
Illumina NovaSeq

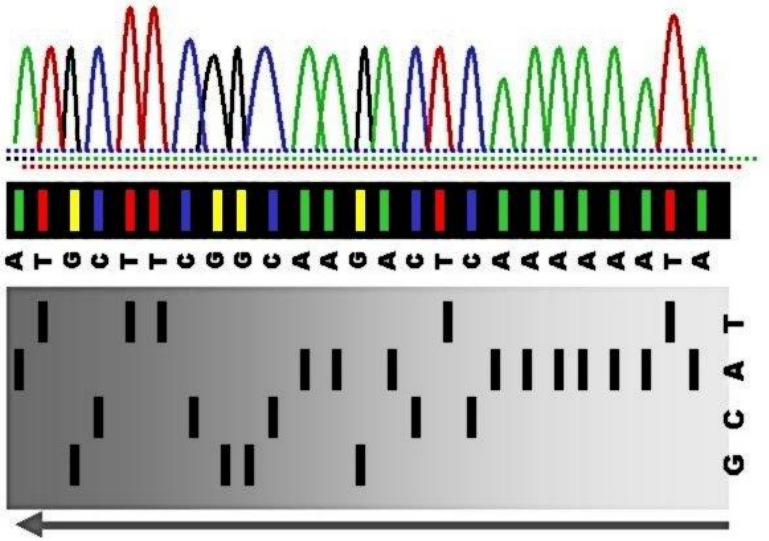
S4	10 B	2000 Gb	3000 Gb
Flow cell type	Single reads*	2 x 100 output	2 x 150 output

Shendure et al, Nature, 2017

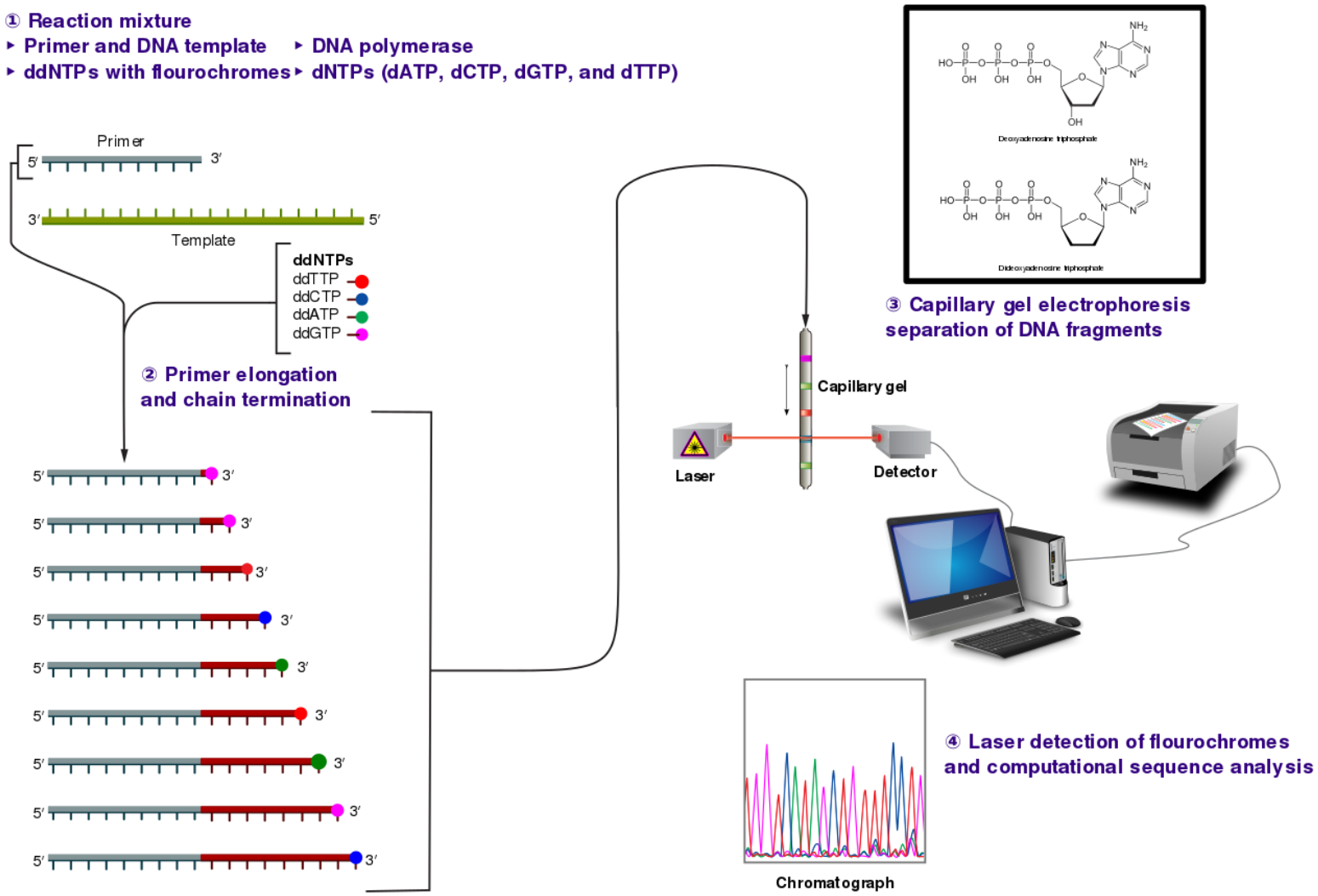
Sanger Sequencing

- Developed by Frederick Sanger and colleagues in 1977
- Up to 1,000 bases
- First human genome draft was based on Sanger sequencing
- Remains in wide use today, for smaller-scale projects and for validation of next-generation sequencing results





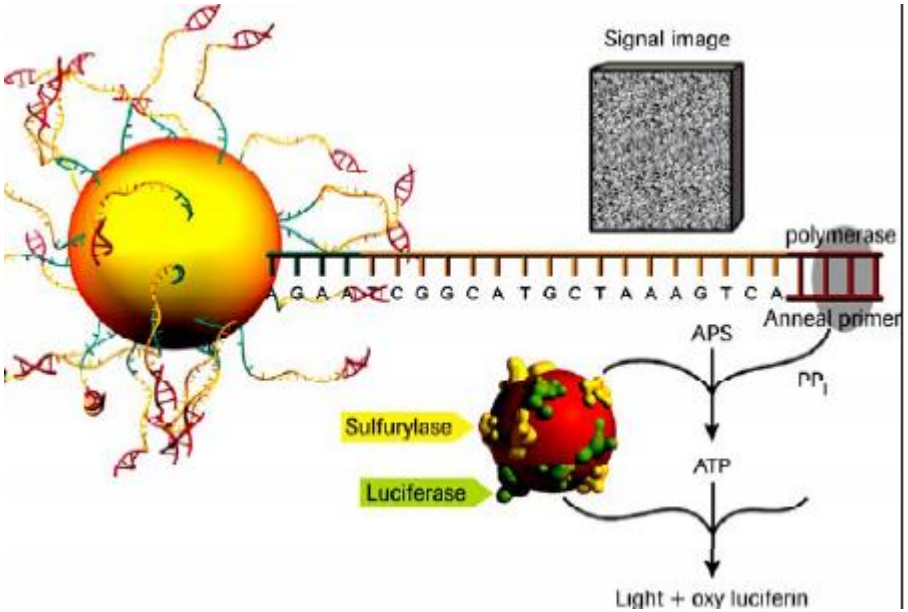
Sanger Sequencing



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

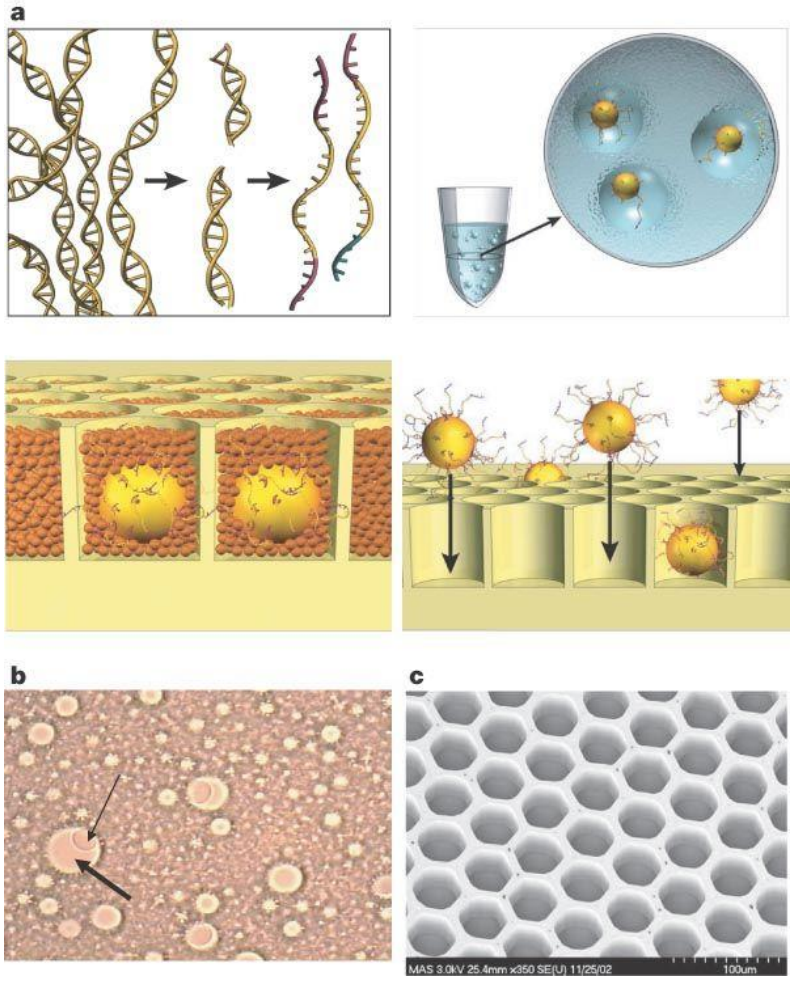
Next-generation sequencing

- Pyrosequencing: incorporation of nucleotide that results in the release of pyrophosphates which fuels the production of light by firefly enzyme luciferase.
- licensed to 454 Life Sciences, where it evolved into the first major successful commercial ‘next-generation sequencing’ (NGS) technology.



http://symposcium.com/2013/05/roche-454-pyrosequencing/

454 Sequencing

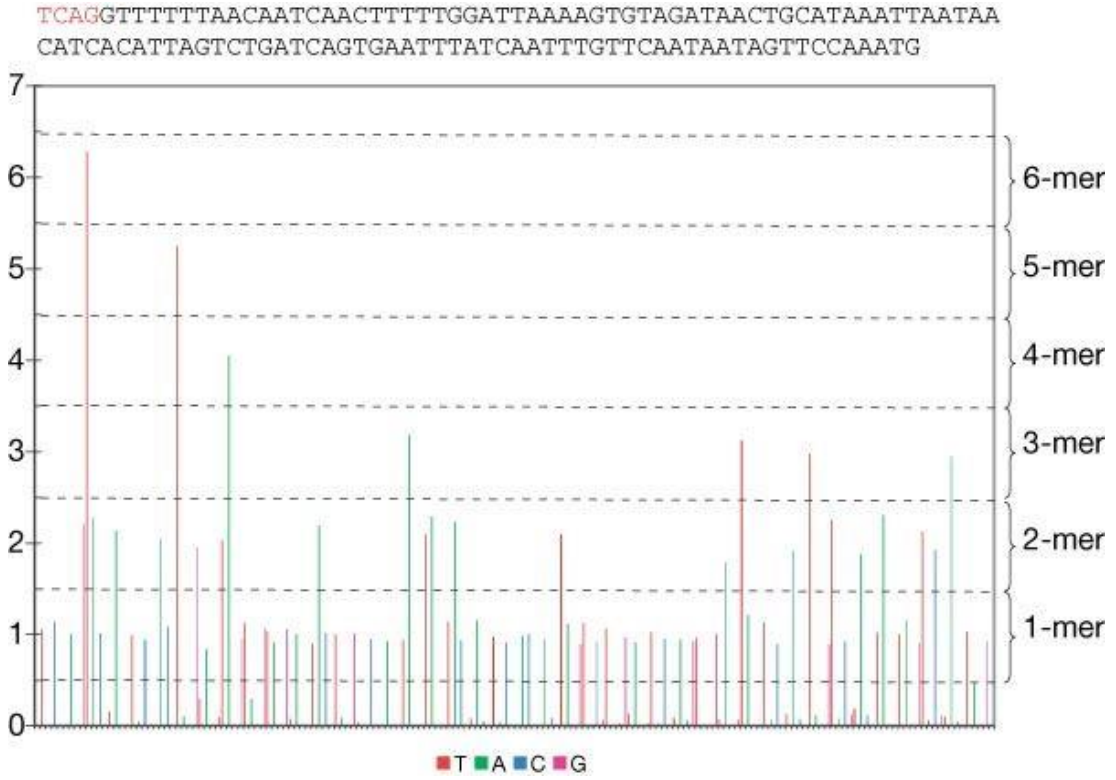


- Genomic DNA is isolated, fragmented, ligated to adapters and separated into single strands.
- Fragments are bound to beads (one fragment per bead), the beads are captured in the droplets; emulsion PCR occurs within each droplet
- Beads carrying single-stranded DNA clones are deposited into wells of a fibre-optic slide
- After the flow of each nucleotide, a wash containing apyrase is used to ensure that nucleotides do not remain in any well before the next nucleotide being introduced.

Margulies et al, Nature, 2005

454 sequencing: base calling

Nucleotide incorporation is detected by the associated release of inorganic pyrophosphate and the generation of photons



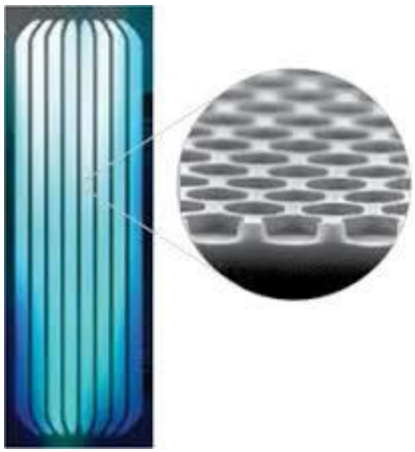
Margulies et al, Nature, 2005

Illumina short-read sequencing

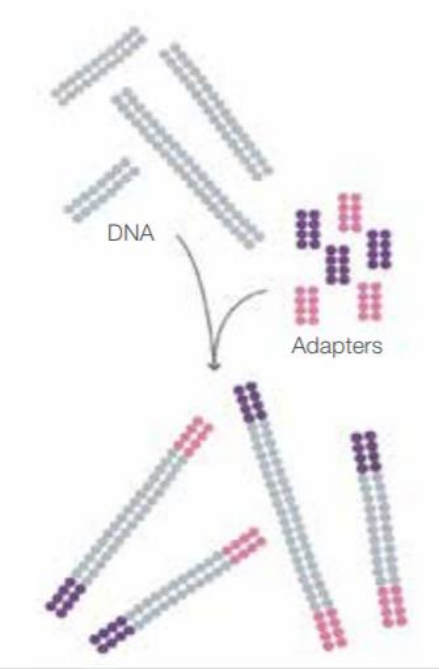
- Illumina sequencing technology, sequencing by synthesis (SBS), is a widely adopted next-generation sequencing (NGS) technology worldwide, responsible for generating more than 90% of the world's sequencing data

How Illumina sequencing works

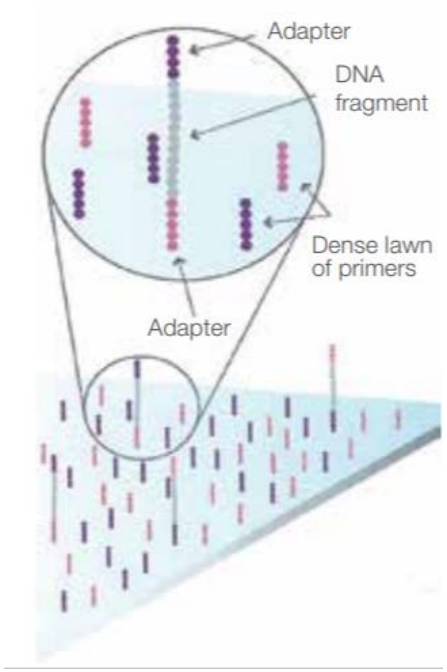
- Cluster generation:
  - Solid-phase amplification creates up to 1,000 identical copies of each single template molecule in close proximity (diameter of 1um or less), using unlabeled nucleotides.
- Sequencing by synthesis (SBS):
  - Four fluorescently labeled nucleotides to sequence the tens of millions of clusters on the flow cell surface in parallel
  - During each sequencing cycle, a single labeled deoxynucleoside triphosphate (dNTP) is added to the nucleic acid chain. The nucleotide label serves as a terminator for polymerization, so after each dNTP incorporation, the fluorescent dye is imaged to identify the base and then enzymatically cleaved to allow incorporation of the next nucleotide.



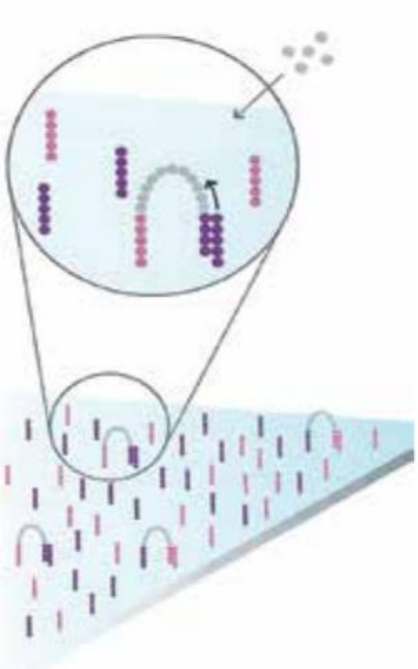
https://www.illumina.com



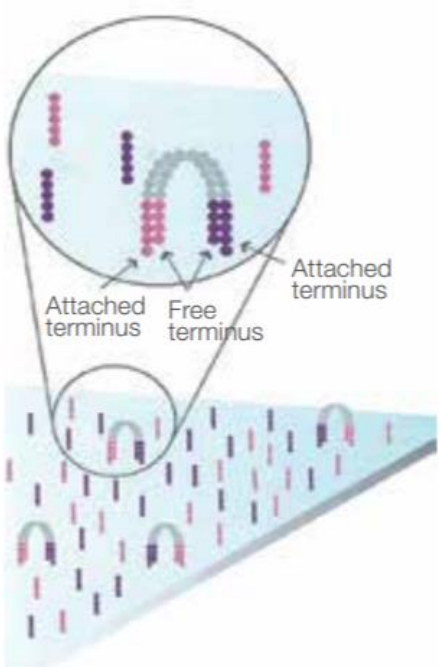
Prepare DNA Sample



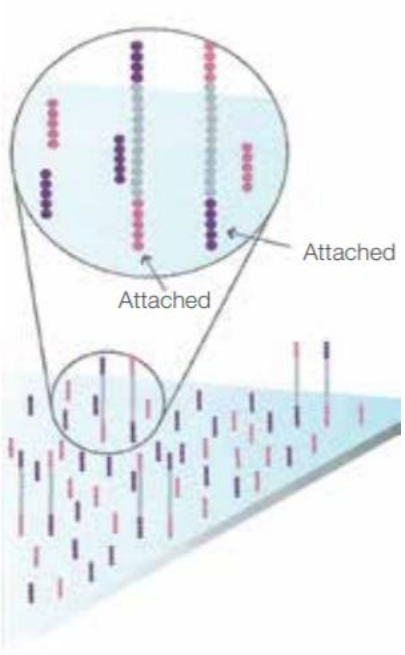
Attach DNA to Surface



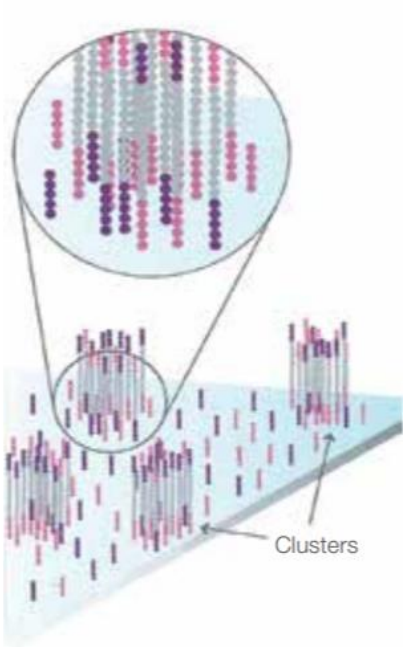
Bridge Amplification



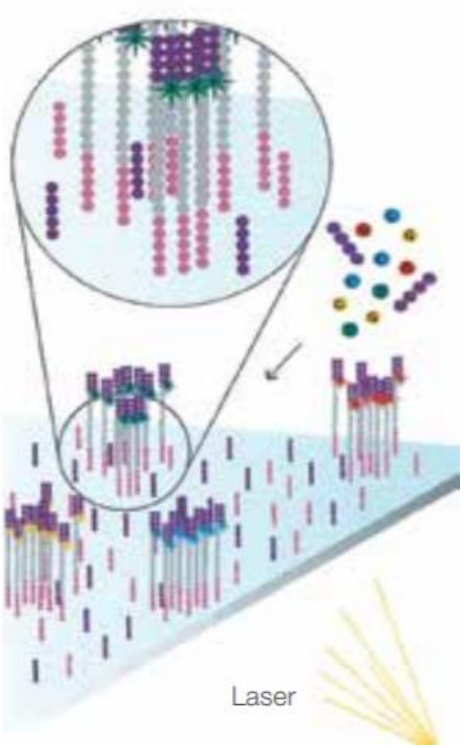
Fragments Become Double Stranded



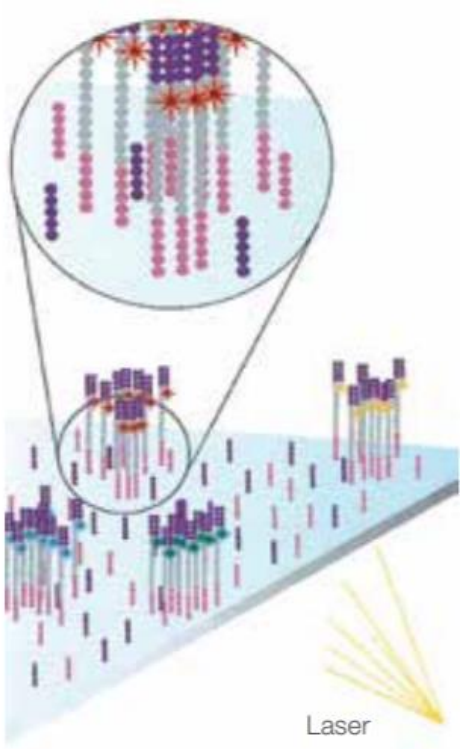
Denature the Double-Stranded Molecules



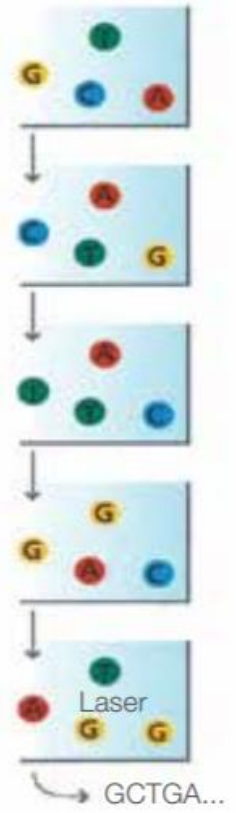
Complete Amplification



Determine First Base



Determine Second Base



Sequencing Over Multiple Cycles

BGISeq and MGISeq

Complete Genomics

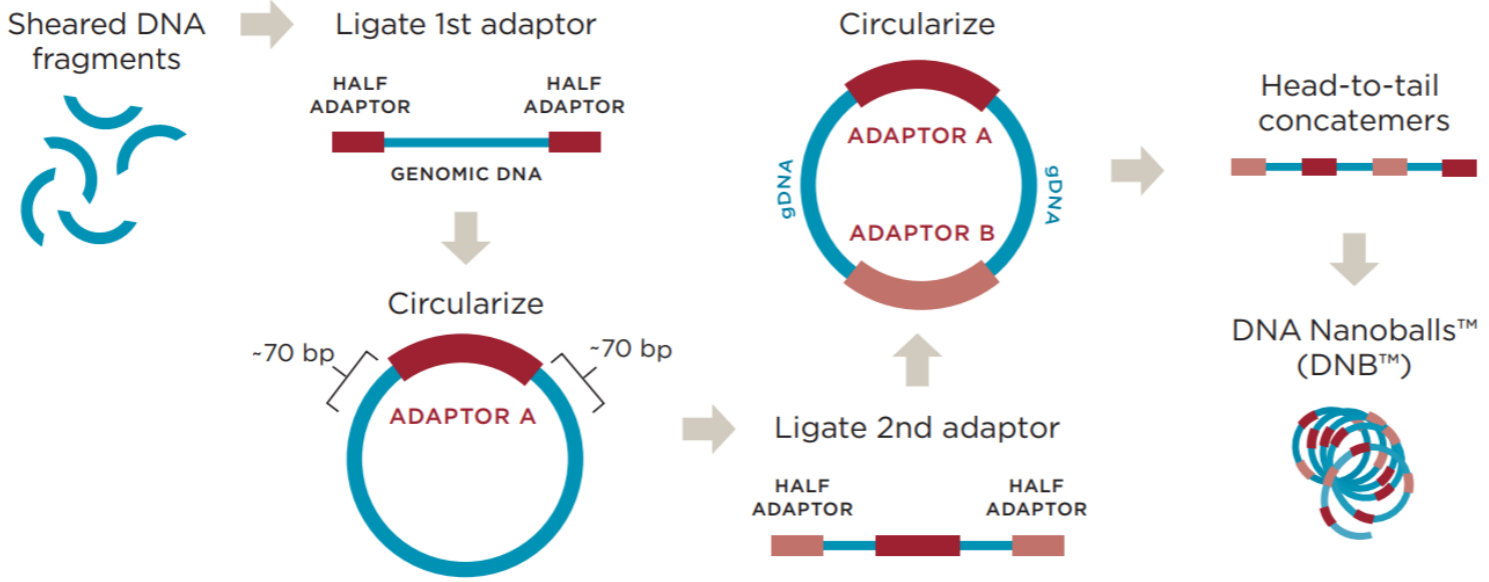
BGISEQ-500

MGISEQ-2000

MGISEQ-T7

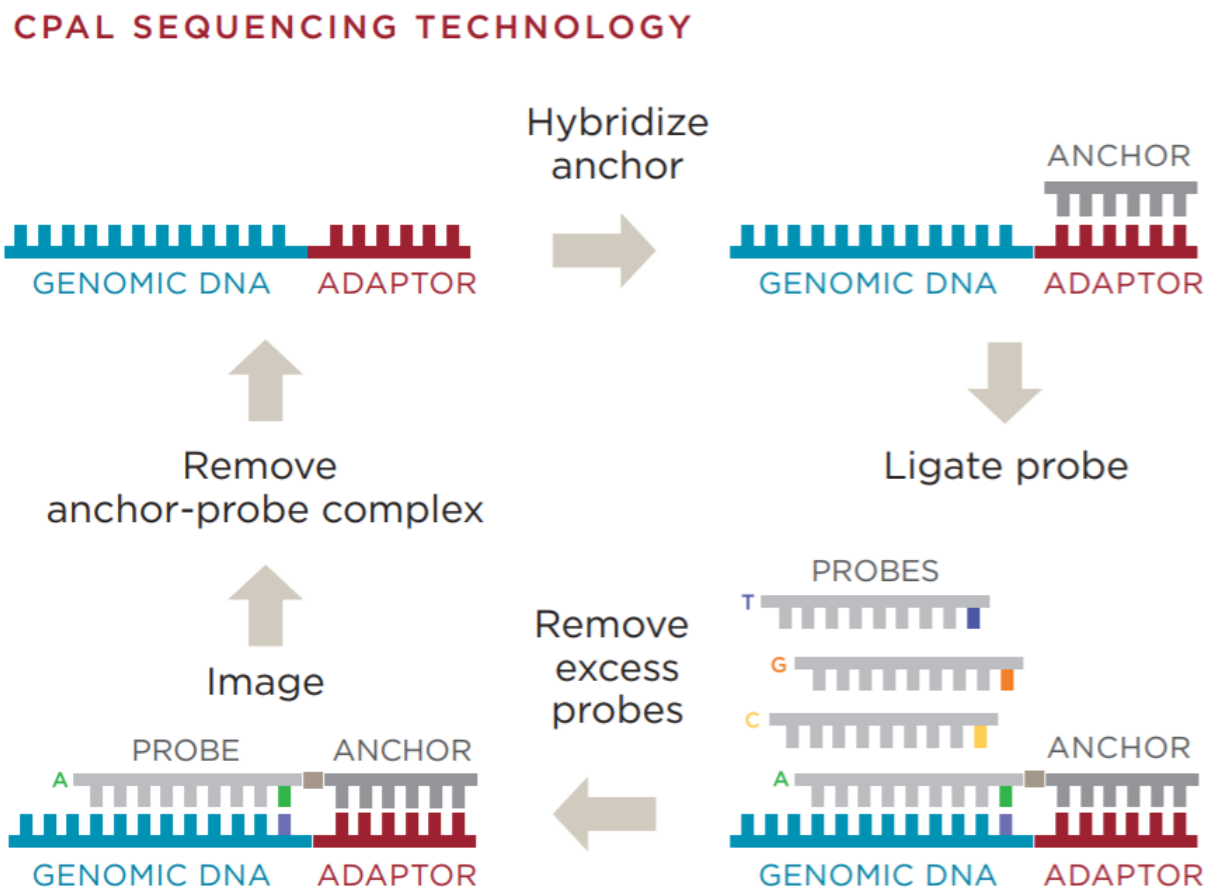
Production of DNA Nanoballs

- The circular DNA molecules in the flow cell library are clonally amplified and modified to produce DNA Nanoballs (DNBs), each containing more than 200 copies of the original template



Drmanac et al, Science, 2010

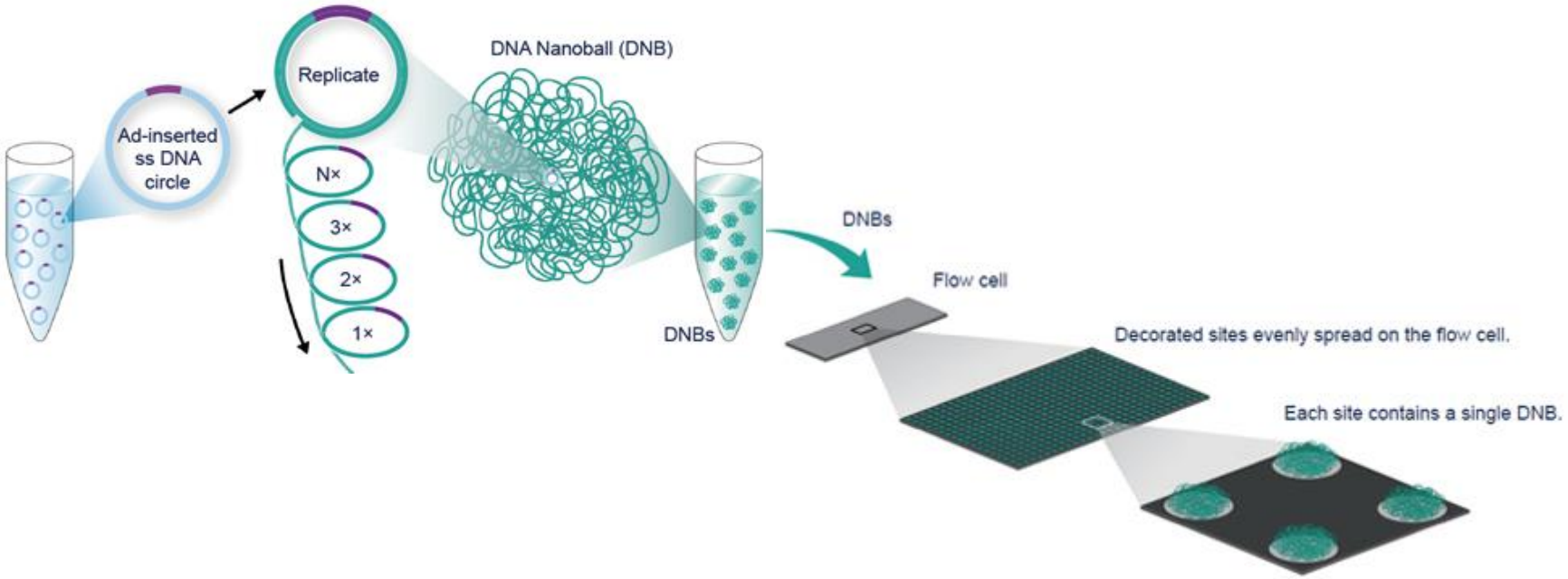
Ligation-based cPAL (Combinatorial Probe-Anchor Ligation) sequencing chemistry



Drmanac et al, Science, 2010

From cPAL (hybridization) to cPAS (synthesis)

- Each cycle: addition of fluorescently labelled terminated dNTPs, cleavage of a terminator, and the detection of the produced fluorescent signal



https://en.mgitech.cn/product/Sequencer.html

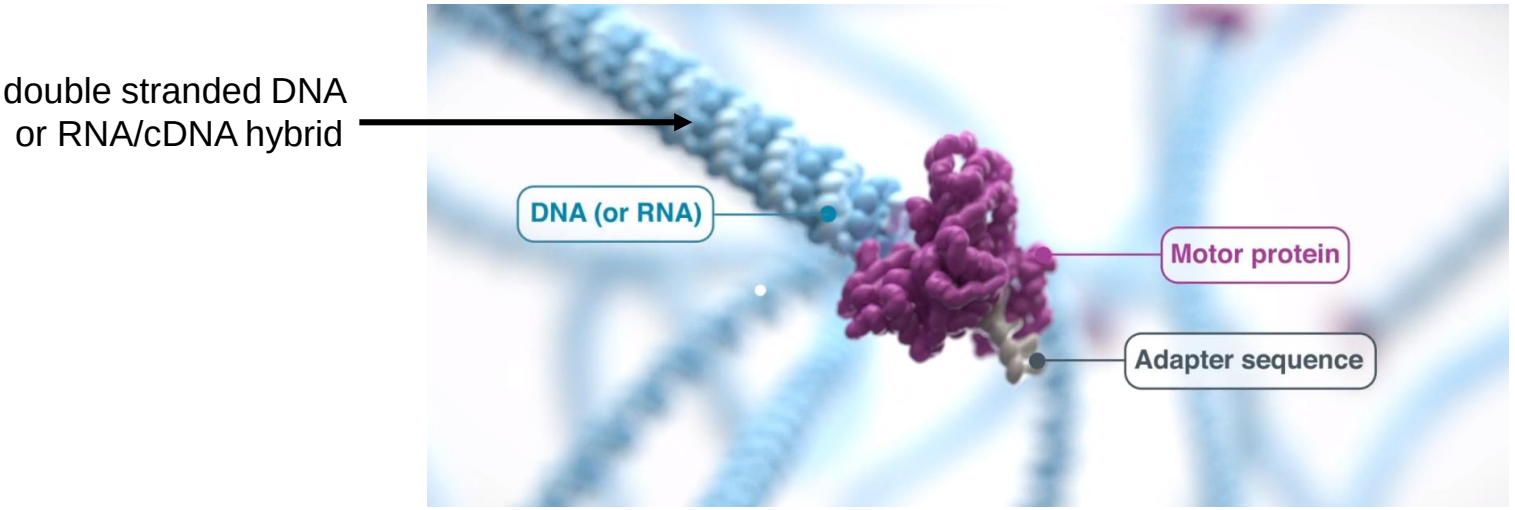
Revolution: single-molecule long-read sequencing

PacBio

Oxford Nanopore

Oxford Nanopore Sequencing

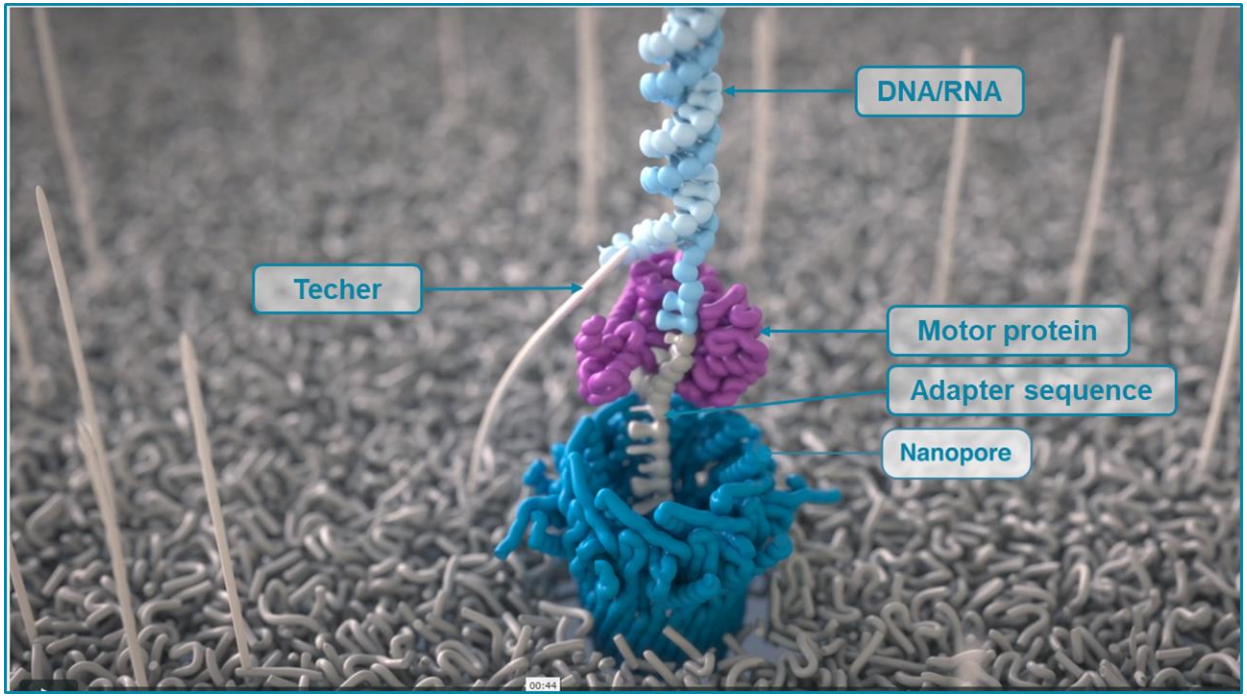
- Oxford Nanopore Sequencing is a real-time, direct DNA/RNA sequencing technology.
- The DNA/RNA is sequenced when it is going though a protein pore.



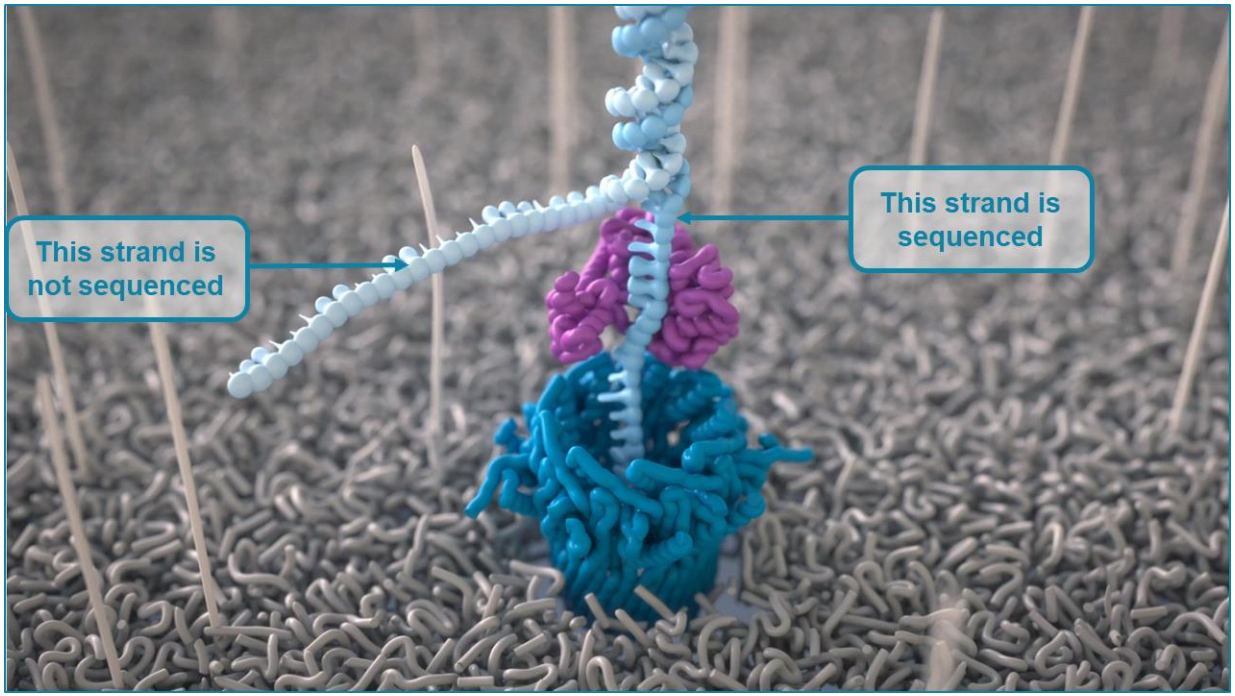
Oxford Nanopore Sequencing

Each well contains a nanopore

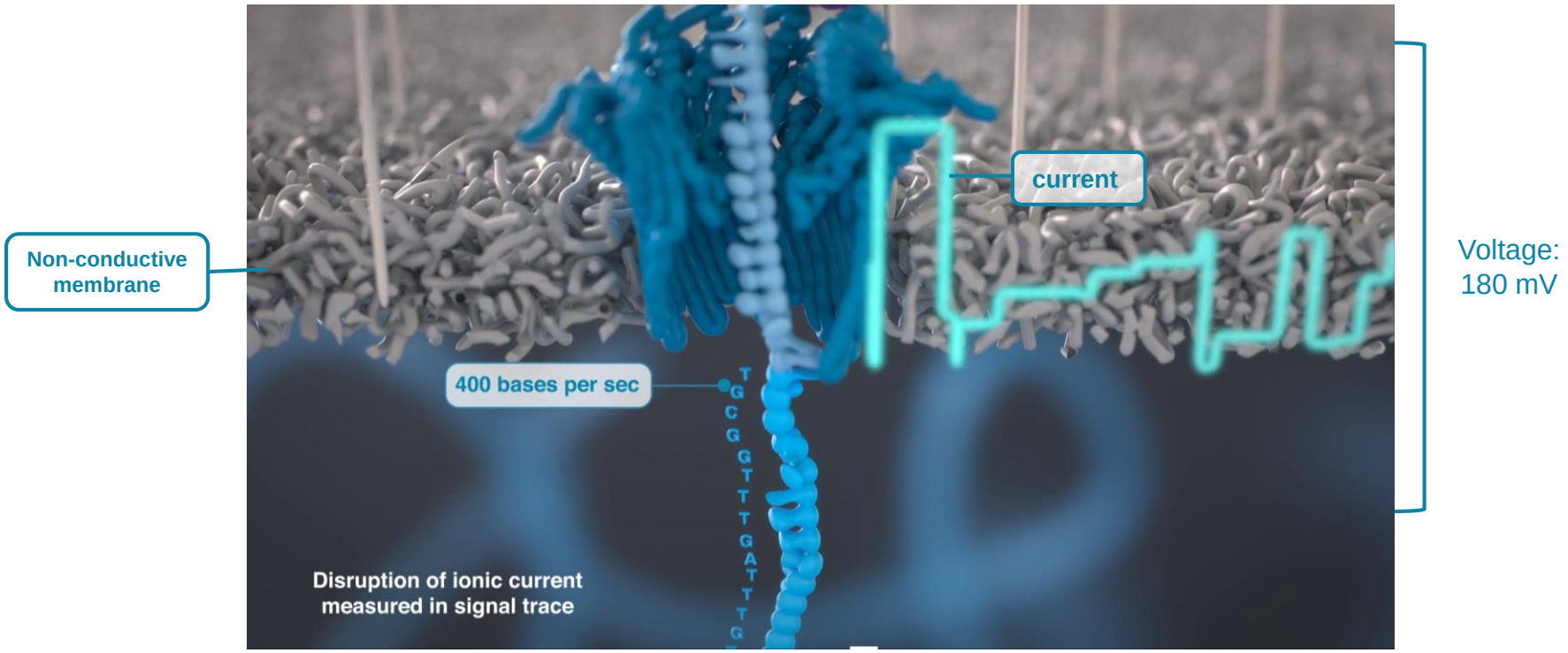
Oxford Nanopore Sequencing



Oxford Nanopore Sequencing



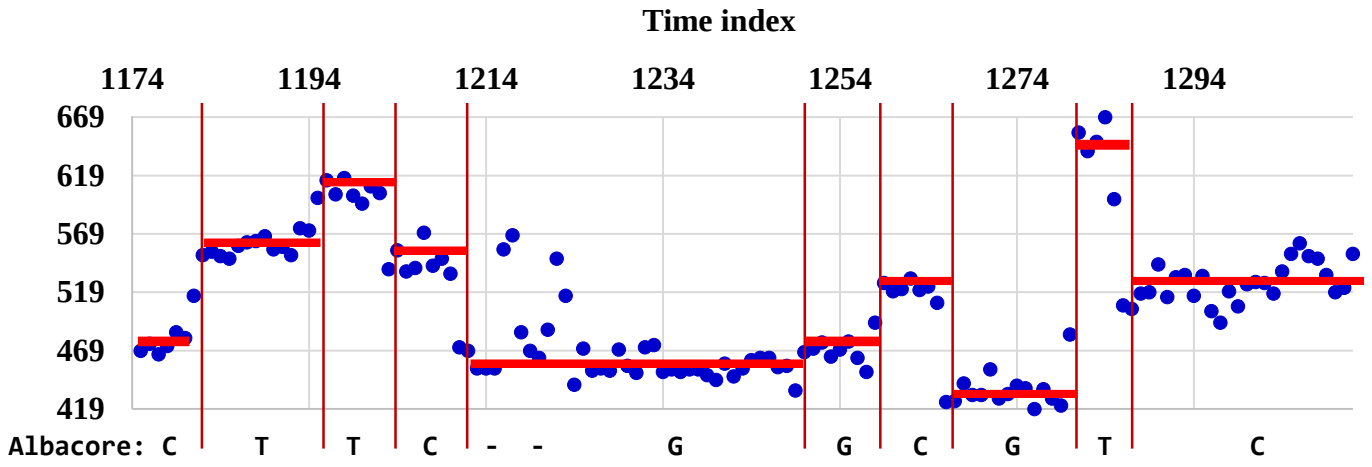
Oxford Nanopore Sequencing



The nucleotides in the DNA/RNA block the ionic current and induce changes of current, which can be measured.

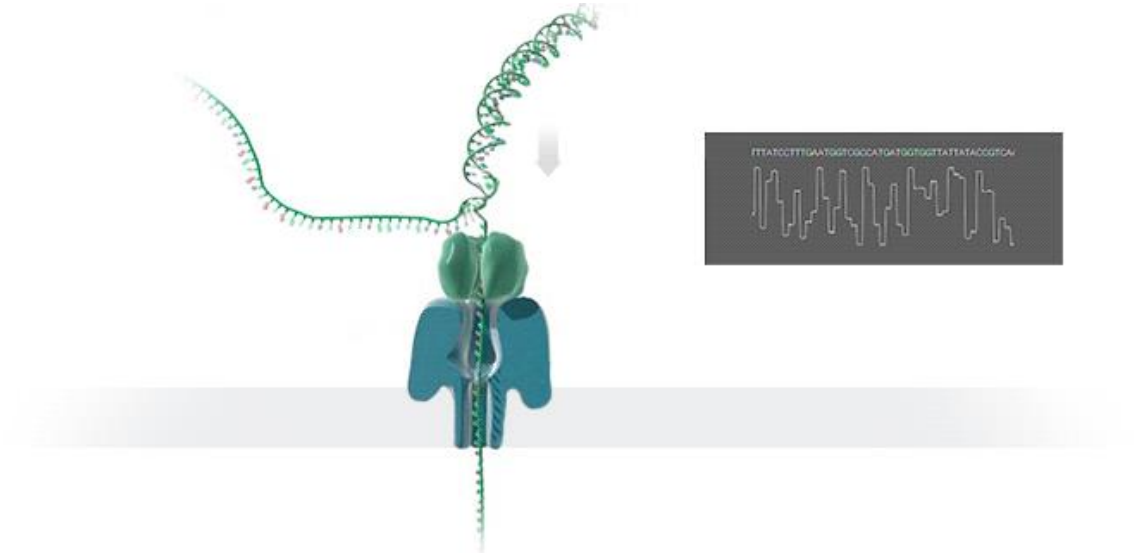
How does the data look like?

- Nanopore sequencing
  - The raw data is electric current (dot)
  - Event detection (red)
  - Base calling: A, C, G, T



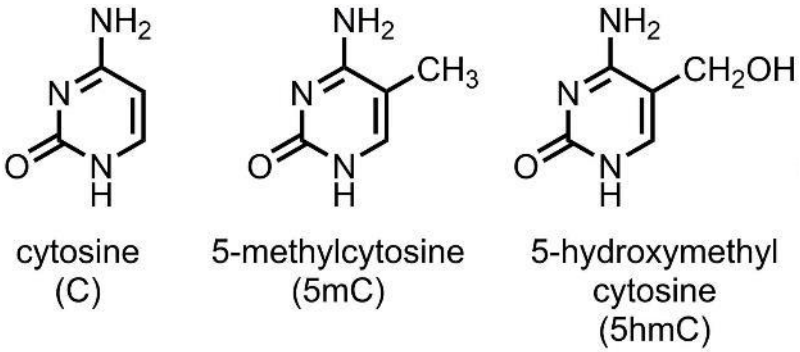
From A/C/G/T to DNA modifications

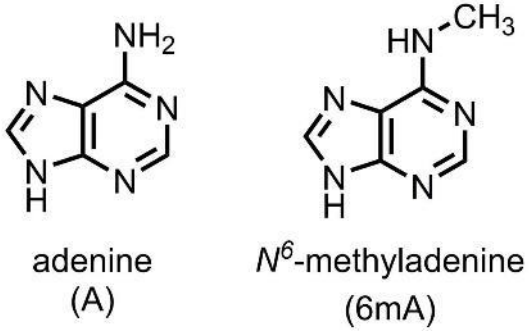
- Change of current when a molecule pass through a tiny hole
- Different types of nucleotides and different modifications of nucleotides would generate different signals
- Currently, homopolymer repeats are an issue

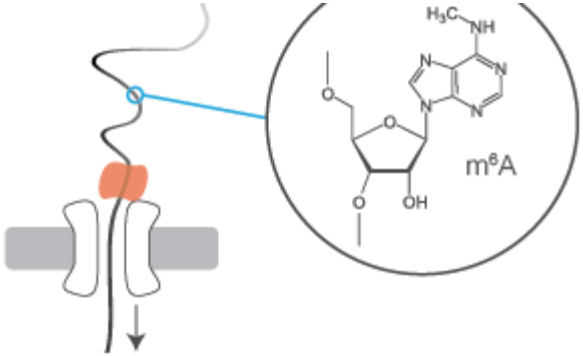


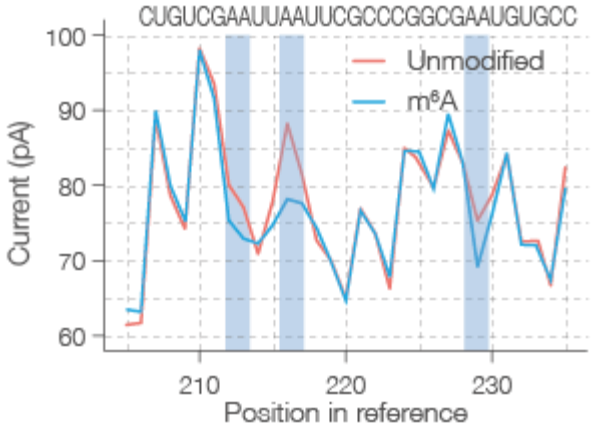
https://www.jove.com/video/55979/nanopore-dna-sequencing-for-metagenomic-soil-analysis

Detect direction of DNA methylations









Shi et al, Front. Genet, 2017

https://nanoporetech.com/resource-centre/tombo-detection-non-standard-nucleotides-using-genome-resolved-raw-nanopore-signal

PacBio Single-molecule real-time (SMRT) sequencing

PacBio RS II

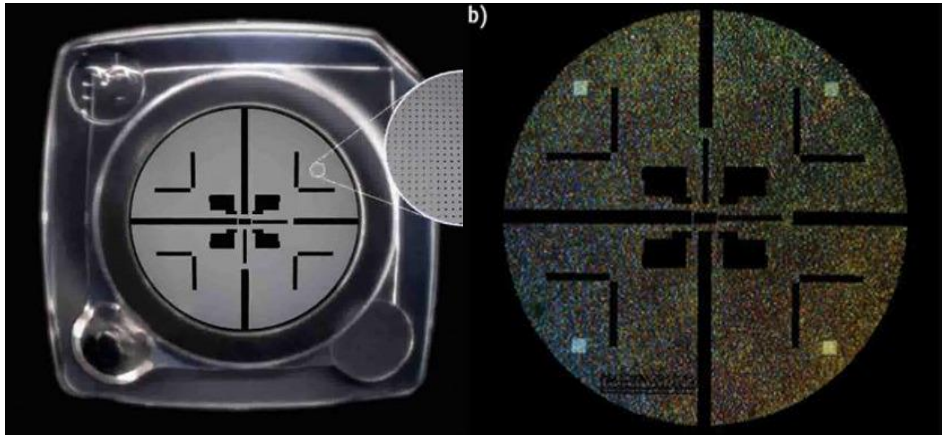
Sequel

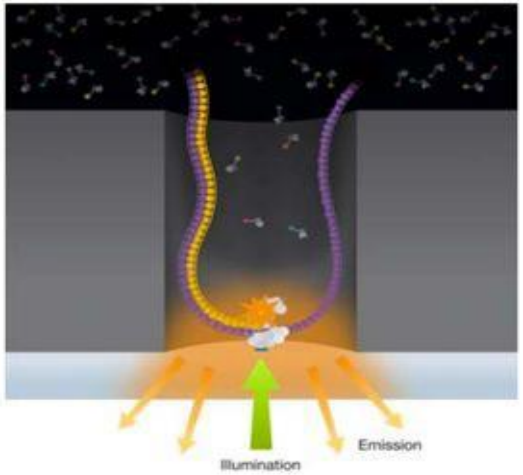
Sequel II

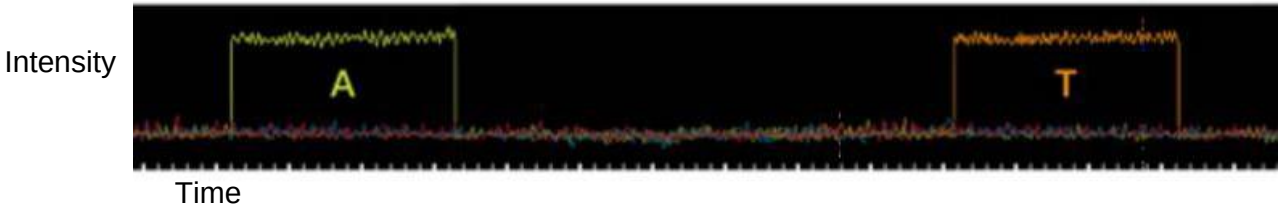
150K/1M/8M zero-mode waveguides (ZMWs)

SMRT sequencing

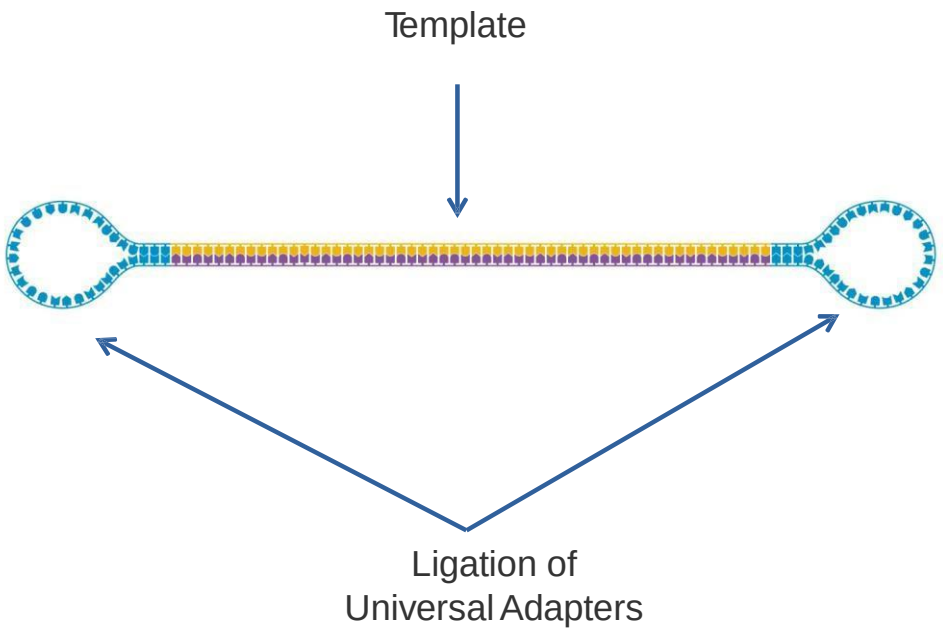
- Imaging of fluorescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).



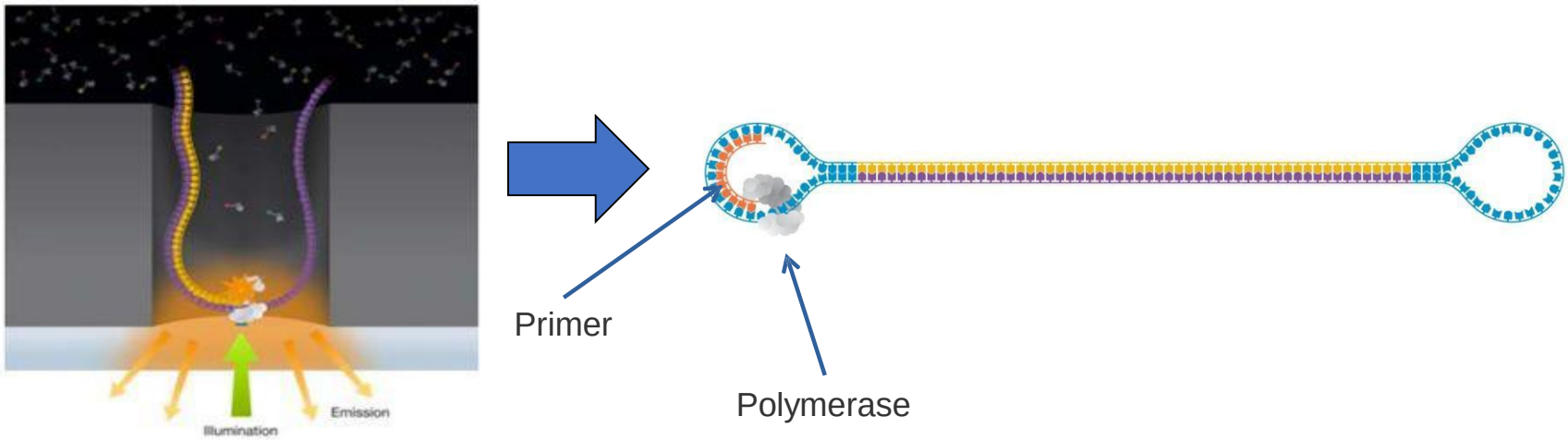




SMRTbell library construction



SMRTbell sequencing

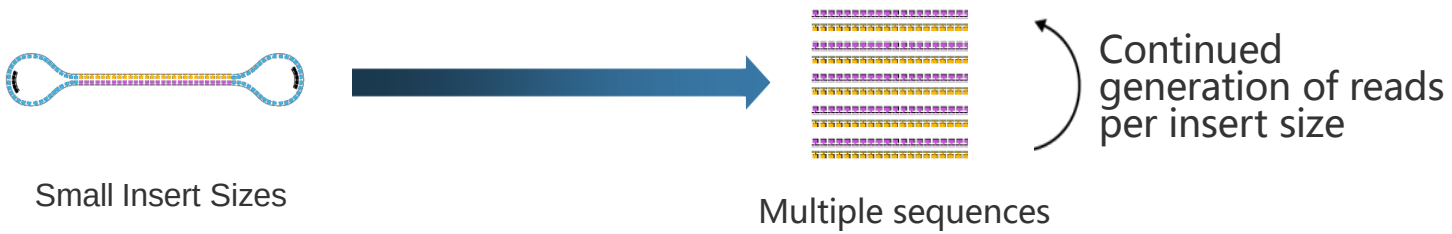


Types of SMRT sequencing reads

Continuous Long Reads (CLR)

Long inserts so that the polymerase can synthesize along a single strand

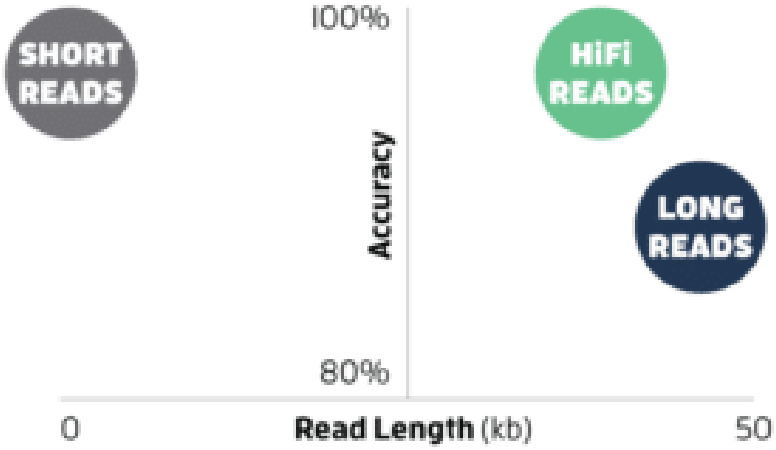
Circular Consensus Sequencing (CCS)



Short inserts, so polymerase can continue around the entire SMRTbell multiple times and generate multiple sub-reads from the same single molecule

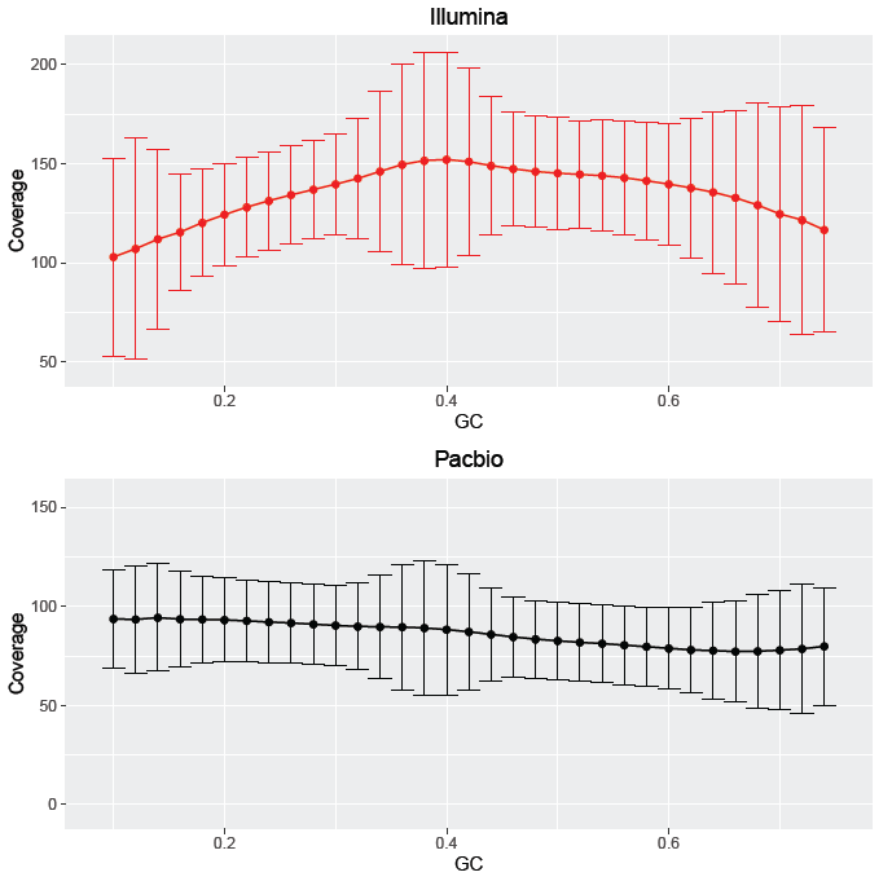
Difference between CLS and CCS (HiFi)

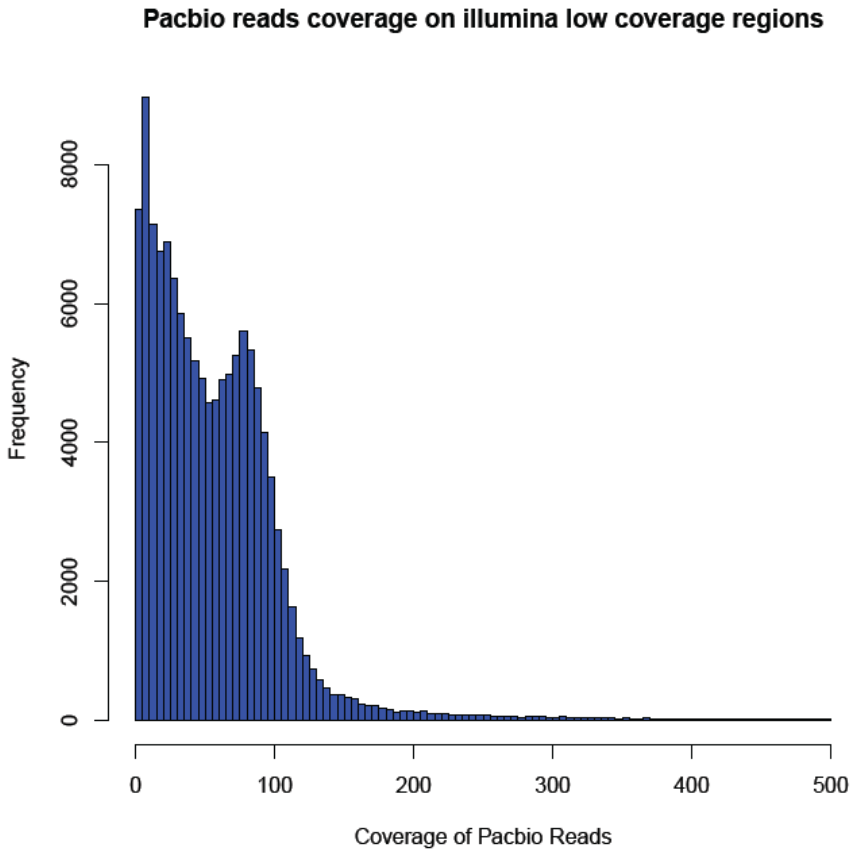
- On Sequel 2, typically ~100 Gb per 8M SMRT cell using long insert/CLR libraries; or ~12-15 Gb >Q30 HiFi CCS per 8M SMRT cell using HiFi libraries.



https://www.pacb.com/products-and-services/sequel-system/latest-system-release/

Impacts of GC on read depth

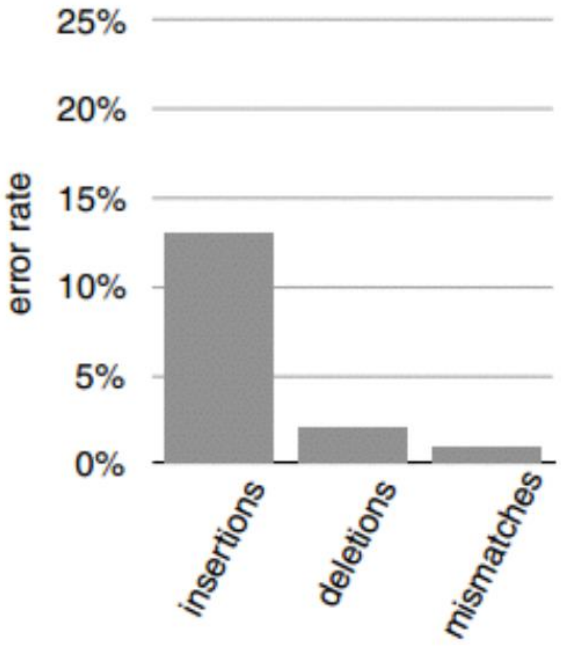


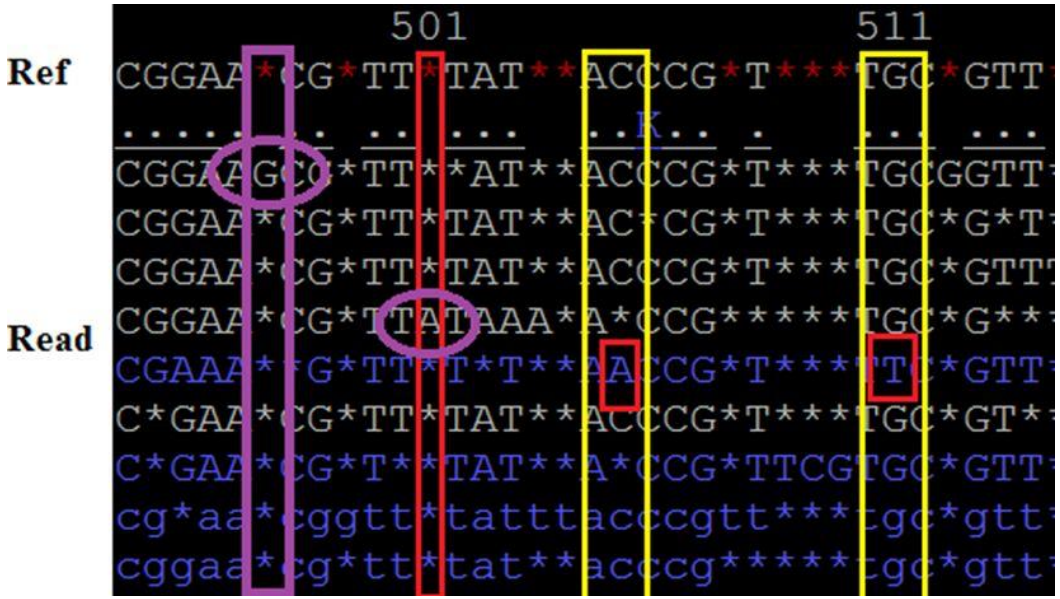


Shi et al, *Nature Communications,* 2016

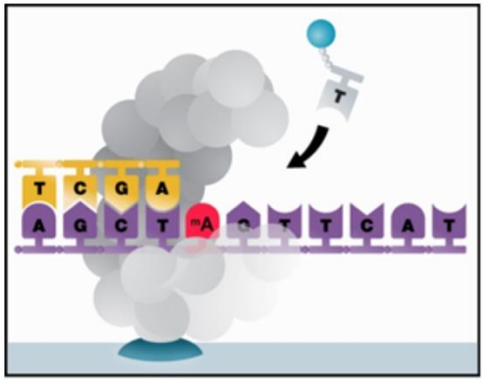
Error profile

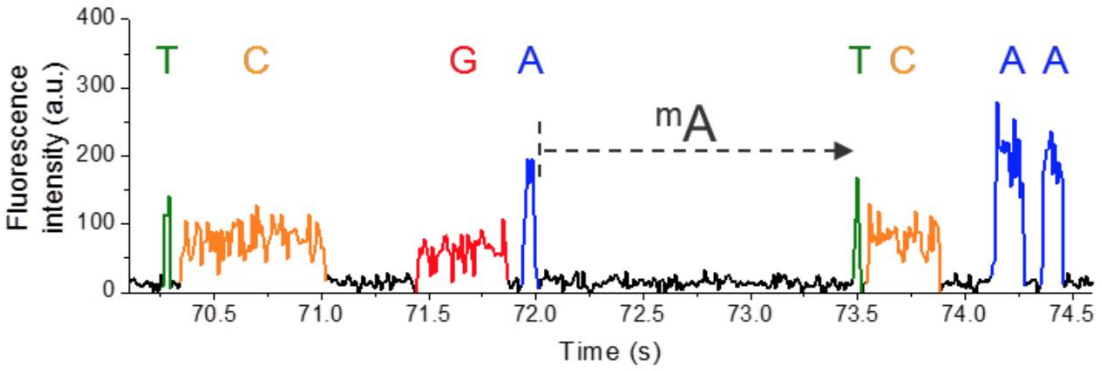
- Insertions tend to be more than deletions and substitutions

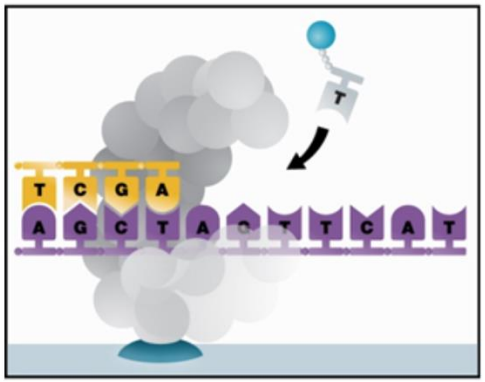


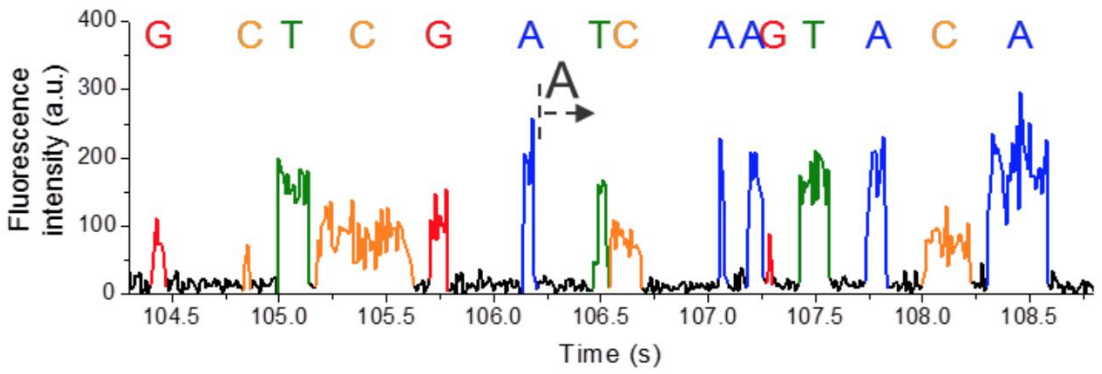


Detection of DNA Base Modifications Using IPD (Interpulse duration ratio)



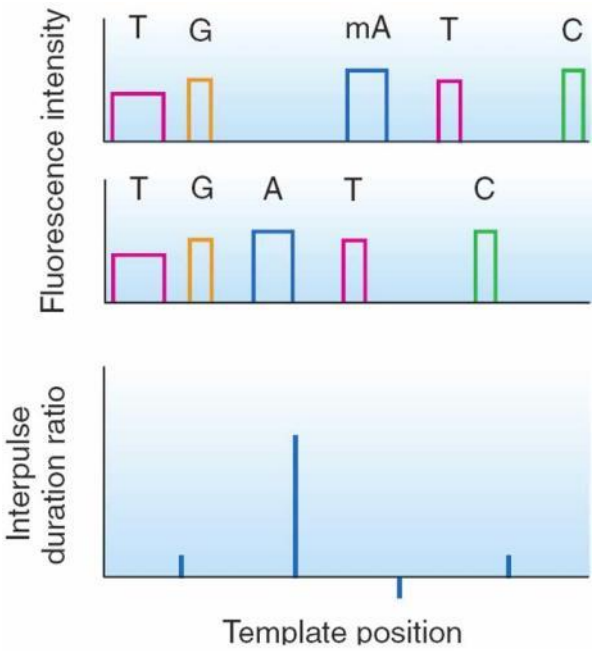


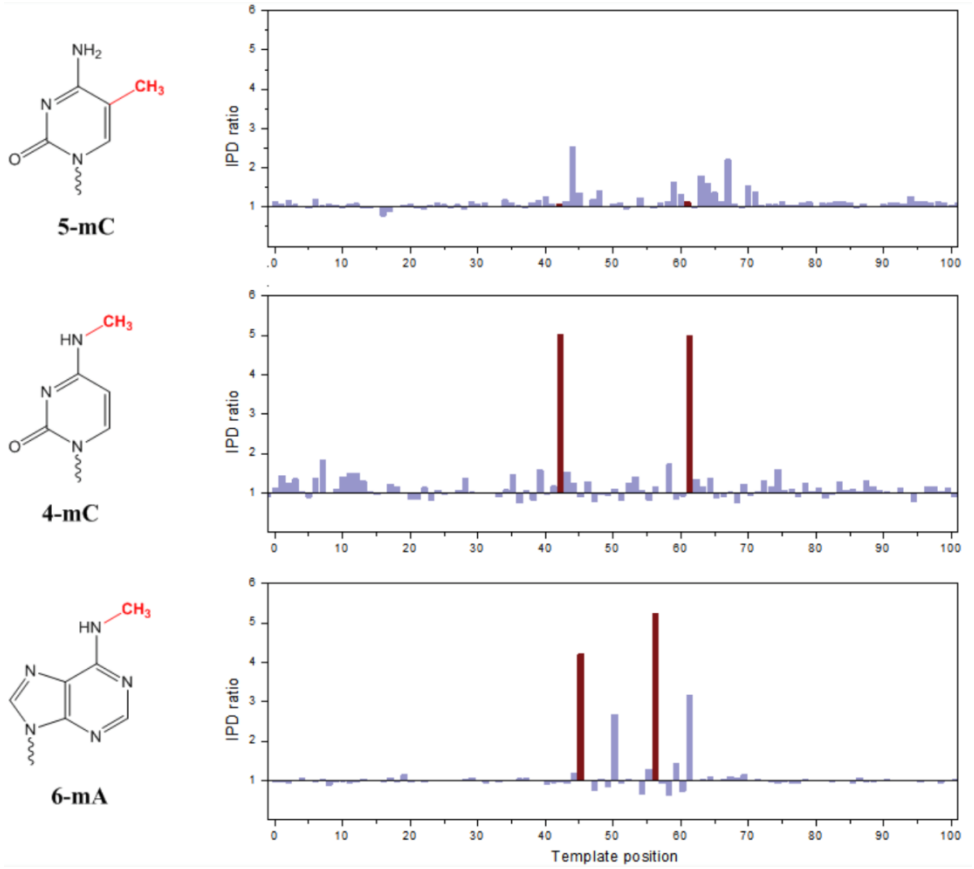




Flusberg et al, *Nature Methods*, 2010

Different modifications have different IPD patterns

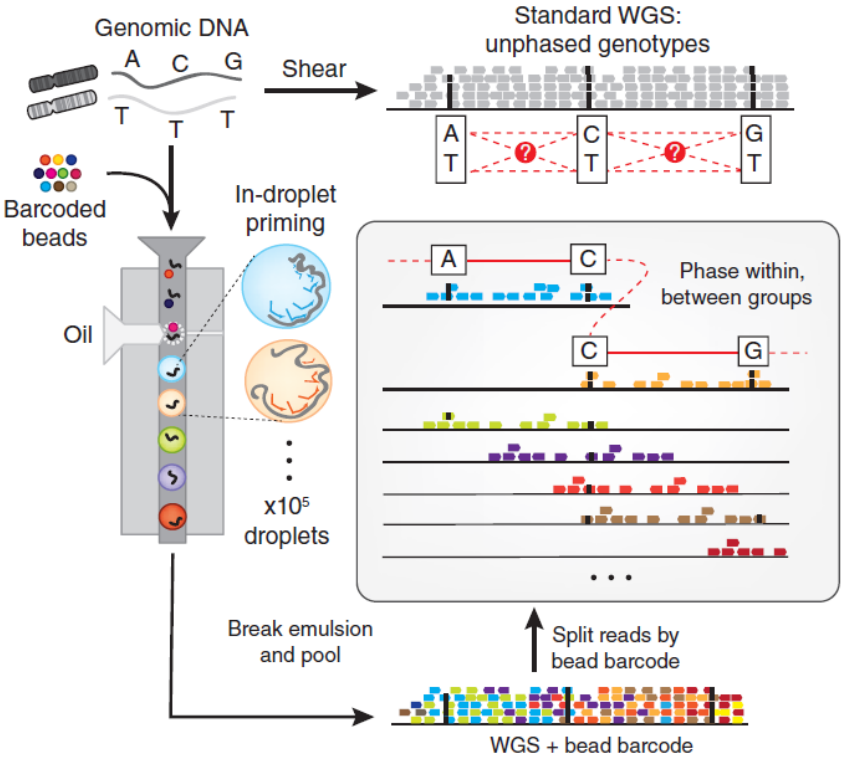




- Coverage needs vary based on the strength of the kinetic signal.
- Kinetic signal strength varies by modification type.

Linked-read Sequencing

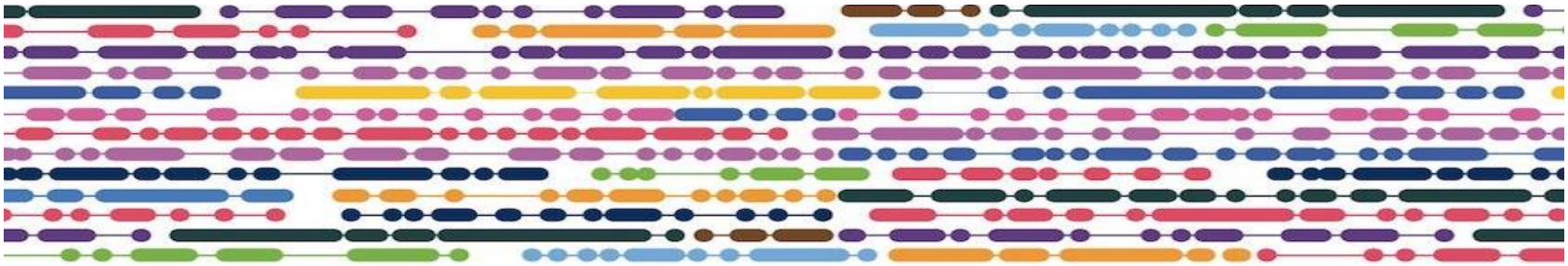
- By adding a unique barcode to every short read generated from an individual molecule, the short reads are linked together.



Zheng et al, *Nature Biotechnology*, 2016

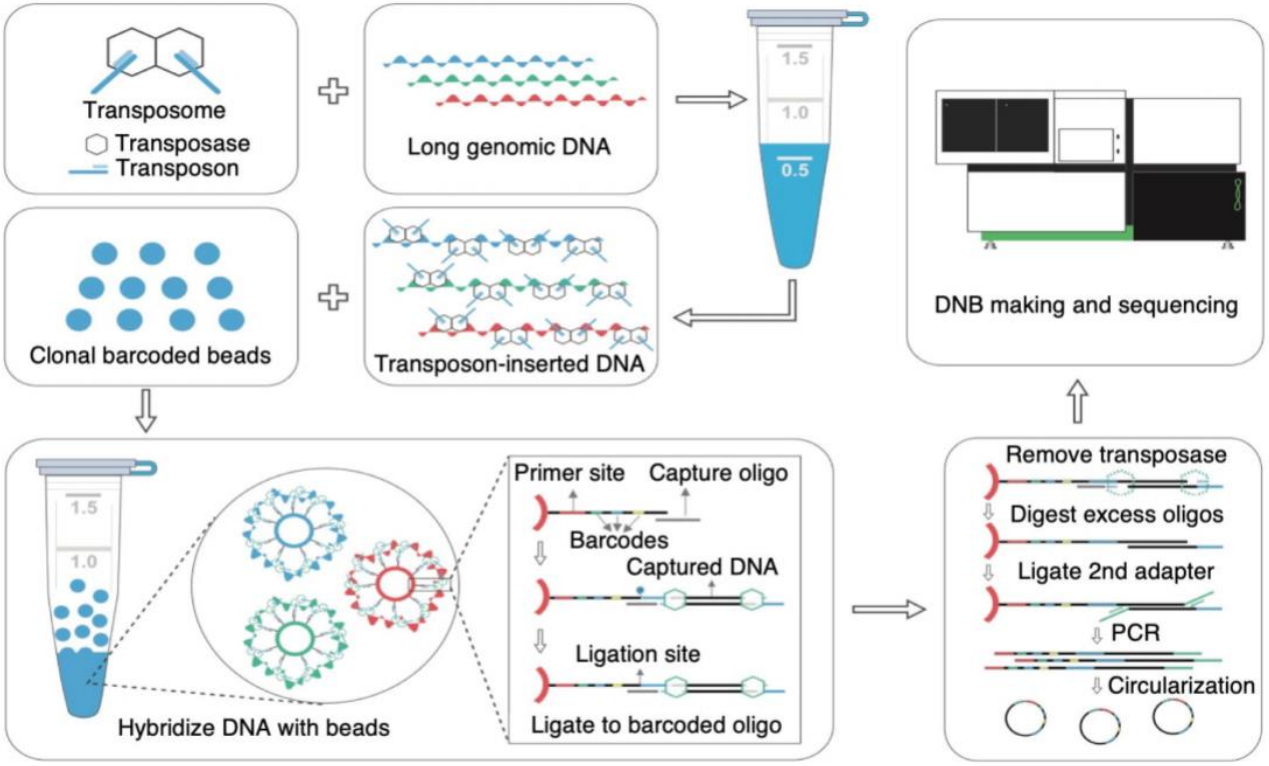
Linked read use molecular barcoding to preserve long-range information

- 1. Generation of long DNA fragments (weighted mean: ~50 kb)
- 2. The long DNA fragments are randomly dispersed into ~1 million droplet partitions with different barcodes; thus, only a small number (~ 10) of DNA fragments are loaded per partition.
- 3. Short read pairs (2 x150 bp) are generated using barcode-containing primers.
- 4. Short reads that contains the same barcode and within a certain distance can be linked together to “reconstruct” the original long DNA fragment.



Other linked-reads technologies

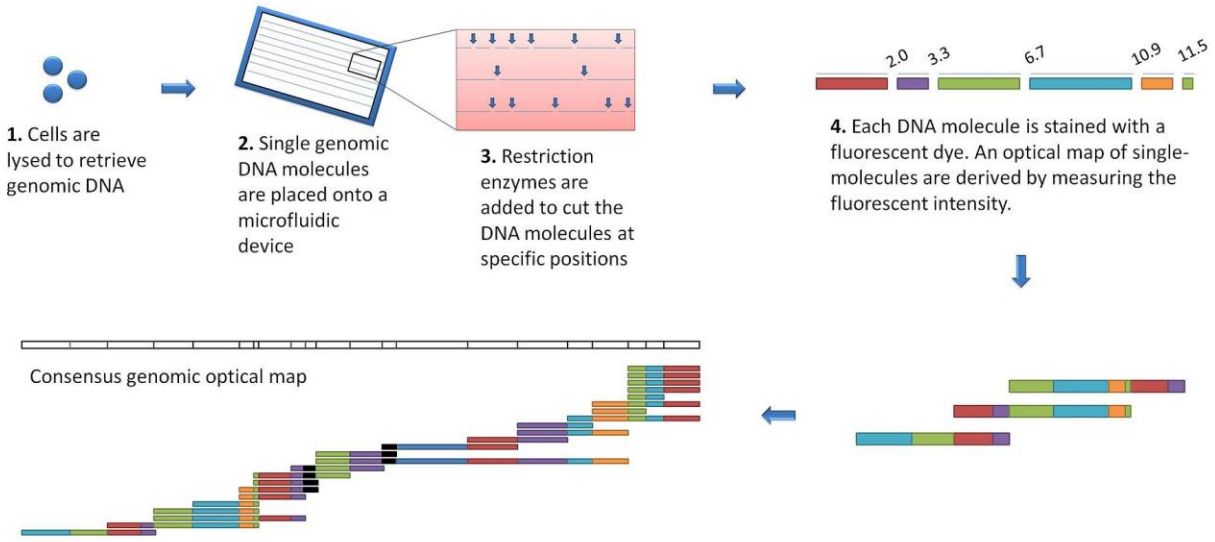
- Single tube long fragment read (stLFR): adding the same barcode sequence to sub-fragments of the original long DNA molecule (DNA cobarcoding).



Wang et al. Genome Research, 2019

Optical mapping

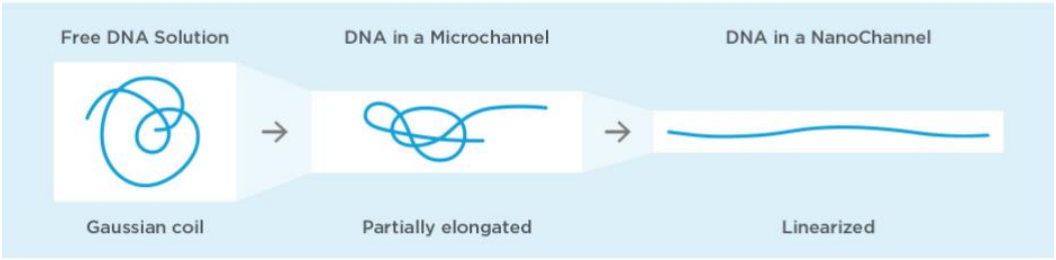
- **Optical mapping** is a technique for constructing ordered, genome-wide, high-resolution restriction maps from single, stained molecules of DNA, called "optical maps".



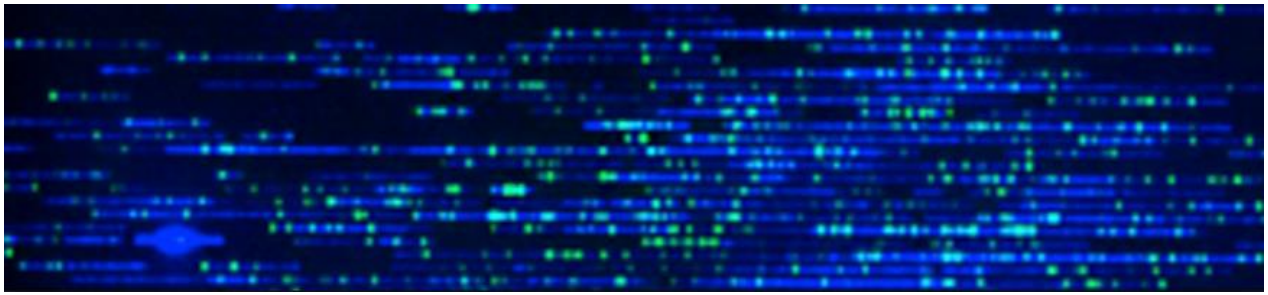
https://en.wikipedia.org/wiki/Optical_mapping

Single-molecule optical mapping (Bionano Genomics)

- Single DNA molecule linearization in Nanochannel.



- Single DNA molecules are labeled with restriction enzymes. The images are scanned and converted to DNA molecules.



https://bionanogenomics.com