# Lesson 1

Introduction to NGS

# Genomics - The Study of Whole Genomes

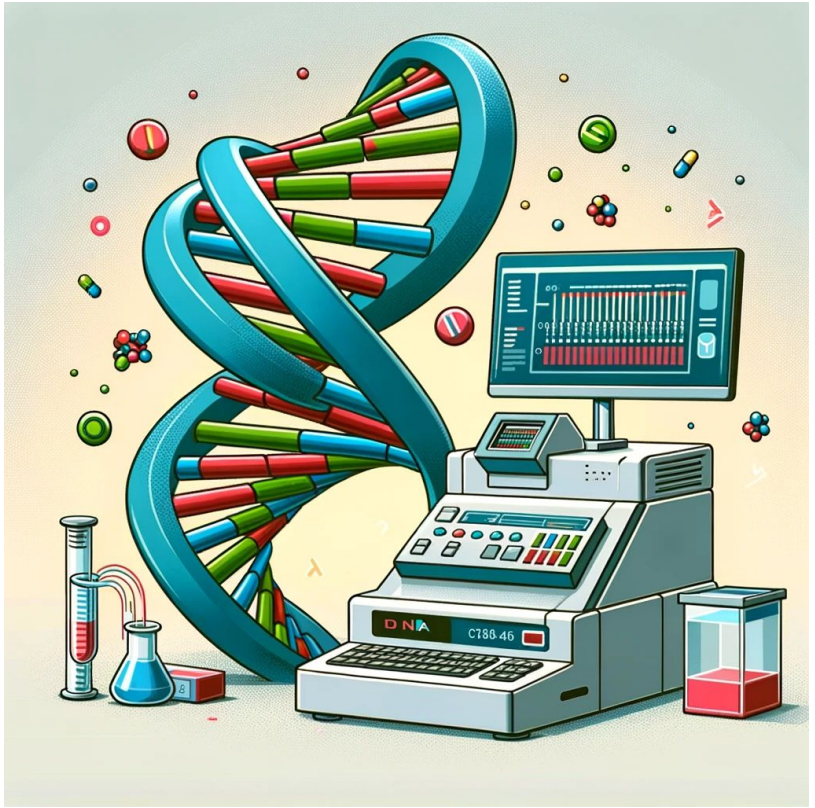
- Structure
- Function
- Evolution

Omics:

DNA - Genomics

RNA - Transcriptiomics

Protein - Proteomics



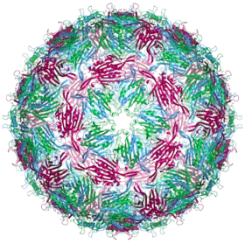
# Sequencing - HOW?

# By the end of this lesson you will...

- Be familiar with the history of nucleic acid sequencing
- Know some of the common uses of NGS
- Understand how the Illumina sequencing technology works
- Be able to use some basic concepts and terminology related to genomics and NGS

# The first sequencing efforts

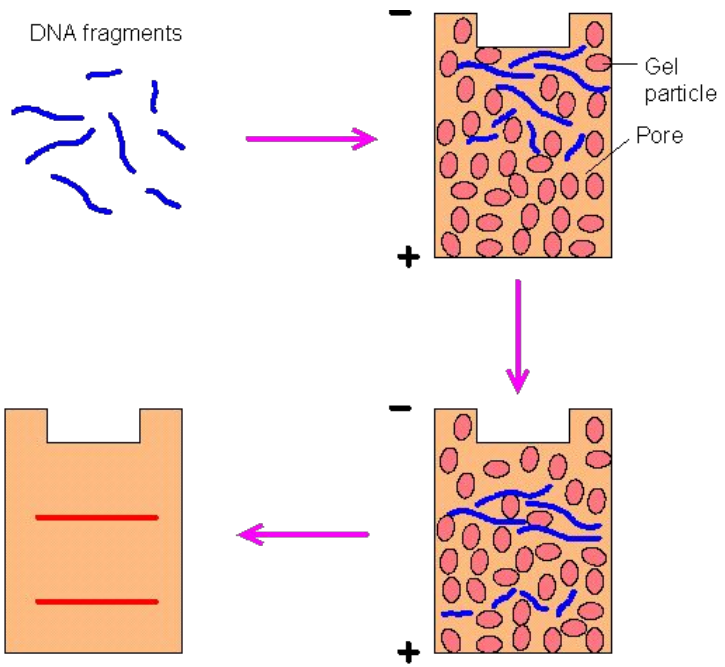
- Highly abundant ssRNA - rRNA, tRNA, phages
- 1965 - first full sequence of yeast tRNA[1]
- 1972 - first protein-coding sequence[2]
- 1976 - first full (RNA) phage genome[3] - 3,569 nucleotides
- Main technology - 2D fractionation of radioactive nucleotides[4]




1. Holley, Robert W., et al. "Structure of a ribonucleic acid." *Science* (1965): 1462-1465.
2. Jou, W. Min, et al. "Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein." Nature 237.5350 (1972): 82.
3. Fiers, Walter, et al. "Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene." Nature 260.5551 (1976): 500.
4. Sanger, F., G. G. Brownlee, and B. G. Barrell. "A two-dimensional fractionation procedure for radioactive nucleotides." Journal of molecular biology 13.2 (1965): 373-IN4.

# 1970’s - 1st generation sequencing

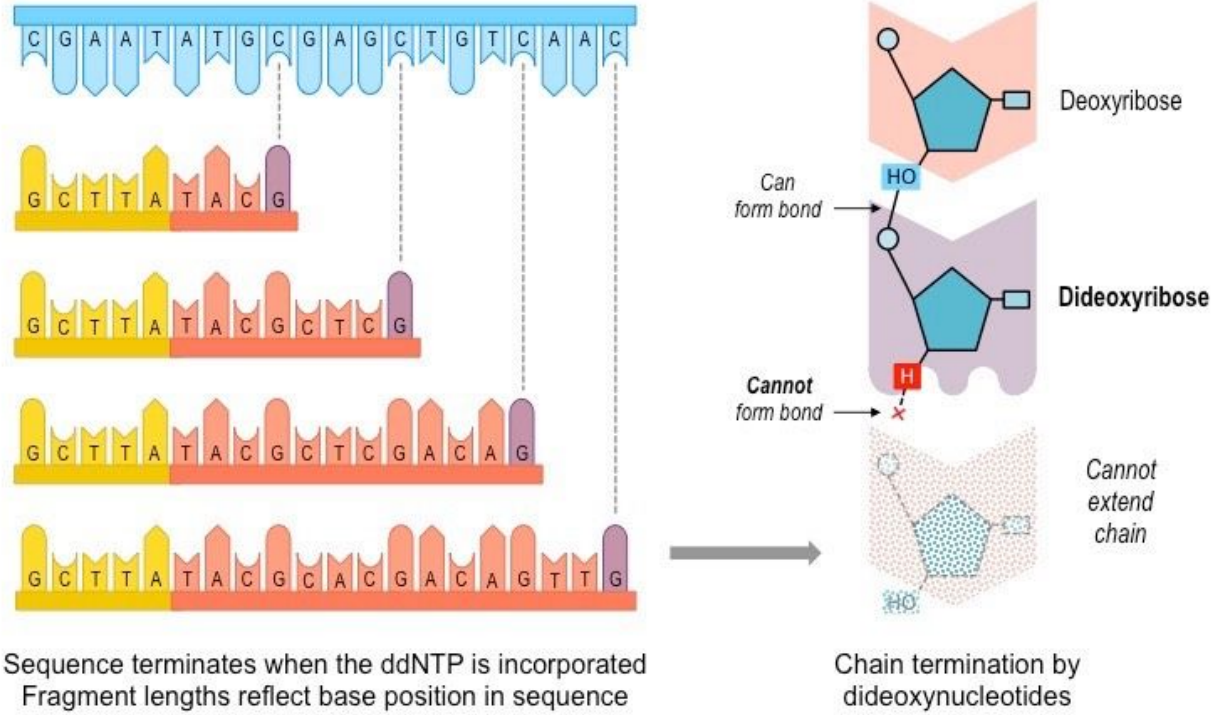
- Chain termination method[1]
- Chemical cleavage method[2]
- Both used polyacrylamide gels
- First DNA phages sequenced





1. Sanger, Frederick, Steven Nicklen, and Alan R. Coulson. "DNA sequencing with chain-terminating inhibitors." Proceedings of the national academy of sciences 74.12 (1977): 5463-5467.
2. Maxam, Allan M., and Walter Gilbert. "A new method for sequencing DNA." Proceedings of the National Academy of Sciences 74.2 (1977): 560-564.

# The chain termination method - “Sanger sequencing”

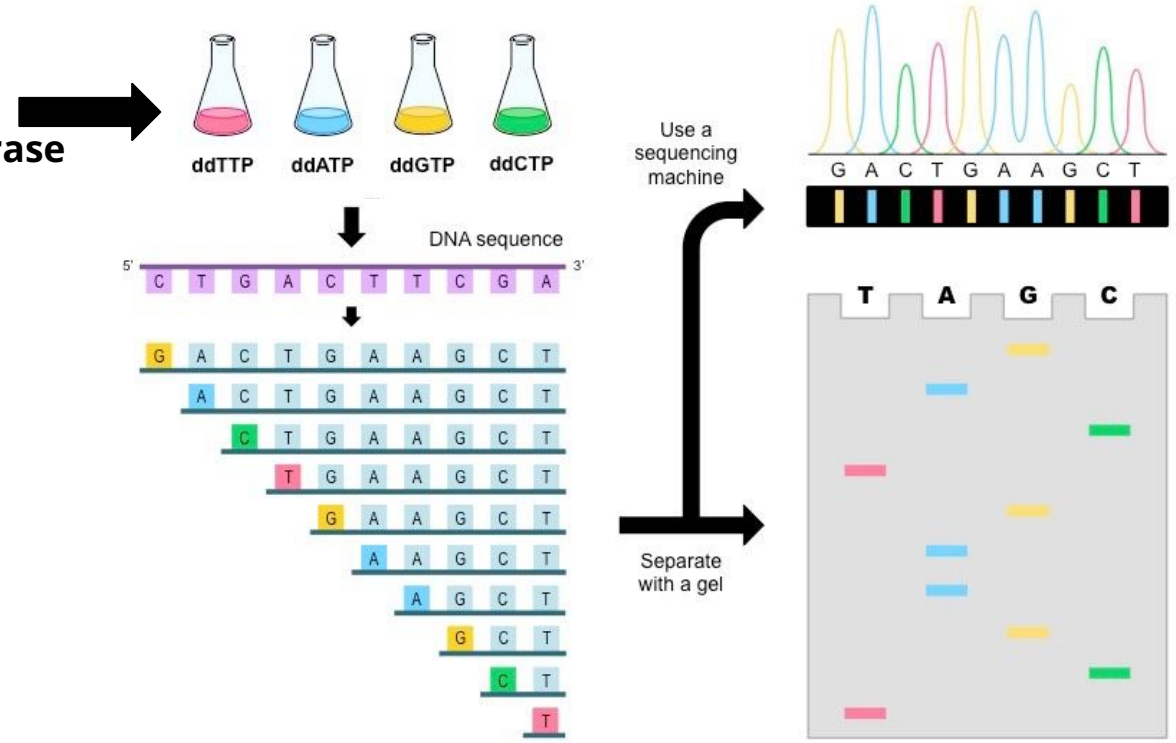



Sequence terminates when the ddNTP is incorporated
Fragment lengths reflect base position in sequence

Chain termination by dideoxynucleotides

# The chain termination method - “Sanger sequencing”

**Target DNA**
**Primer**
**DNA Polymerase**
**4 dNTPs**



# The chain termination method - “Sanger sequencing”

- **SBS - sequencing by synthesis**
- Produces reads 500-1,000 bp long
- First commercial machines
- Still used today!
- Can sequence ~70k bp/hour
- Cost: ~$500/1Mbp
- Used for various whole genome projects
- **Shotgun sequencing**

# The human genome project

**Reminder**: human genome size is ~3Gb

- Started 1990
- First genome draft - 2000
- Project completion - 2003
- Largest ever biological collaborative project
- 20 sequencing centers around the world
- Entirely based on Sanger sequencing
- Estimated cost: $5 billion

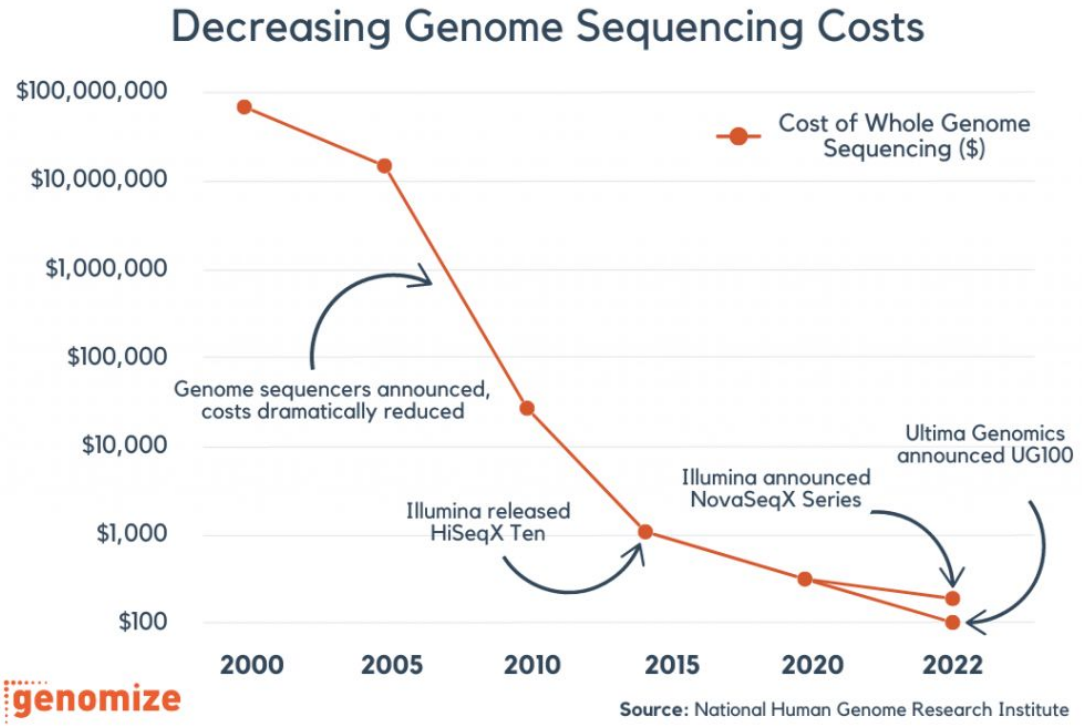
# 2000s - 2nd generation sequencing

- Pyrosequencing[1] - “454 sequencing”
- Illumina (Solexa) method
- Allow massively parallel sequencing
- Produce short reads - usually 50-250 bp
- **Deep sequencing** - each genomic region is sequenced multiple times


1. Ronaghi, Mostafa, Mathias Uhlén, and Pål Nyrén. "A sequencing method based on real-time pyrophosphate." Science 281.5375 (1998): 363-365.
2. Canard, Bruno, and Robert S. Sarfati. "DNA polymerase fluorescent substrates with reversible 3′-tags." Gene 148.1 (1994): 1-6.

# Illumina technologies - standard for NGS

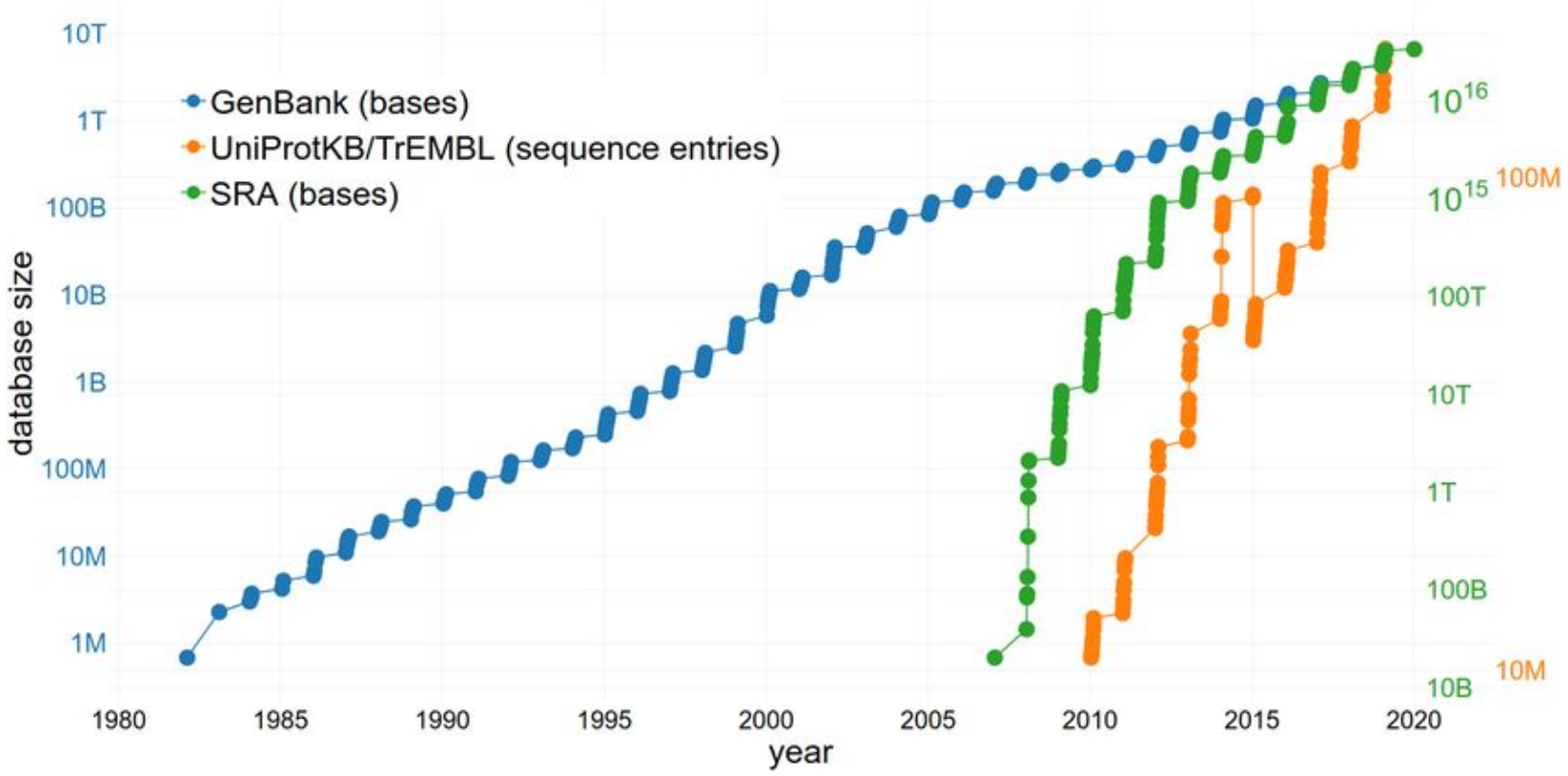
- By far the most popular sequencing technology
- Up to 120 Gb/hour
- Significantly reduced sequencing costs
- Allowed the sequencing of numerous species and samples






How Decreasing Genome Sequencing Costs Effects Clinical Genetics and Precision Medicine

# Number of bases deposited in public databases




DOI:10.7717/peerj.9954

# Some sequencing projects from 2022

nature communications

Explore content | About the journal | Publish with us

nature > nature communications > articles > article

Article | Open Access | Published: 04 March 2022

**Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil**

nature genetics

Explore content | About the journal | Publish with us

nature > nature genetics > articles > article

Article | Open Access | Published: 22 September 2022

**Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics**

ZR ZOOLOGICAL RESEARCH www.zoores.ac.cn

Zool Res. 2022 Jan 18; 43(1): 78–80. PMCID: PMC8743251
doi: 10.24272/j.issn.2095-8137.2021.266 PMID: 34877831

Whole-genome resequencing infers genomic basis of giant phenotype in Siamese fighting fish (*Betta splendens*)

# So... what can we do with NGS?

Detect genomic variation

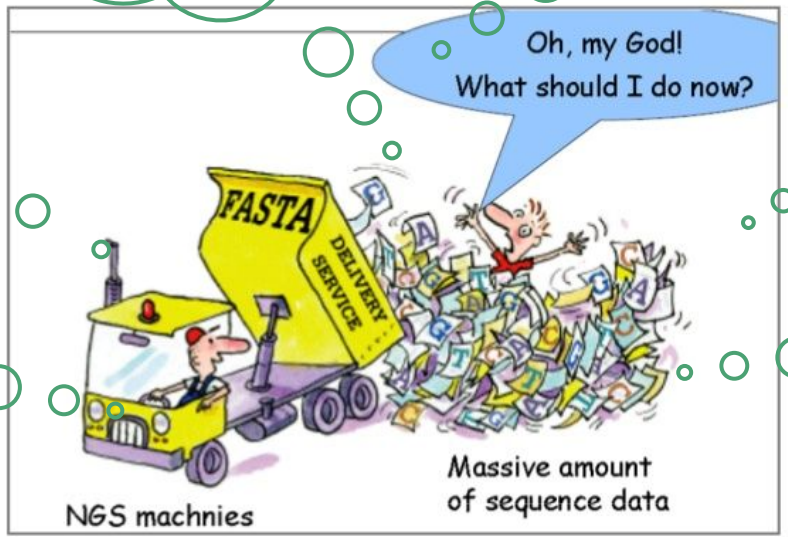
Profile gene expression

Acquire new genomes

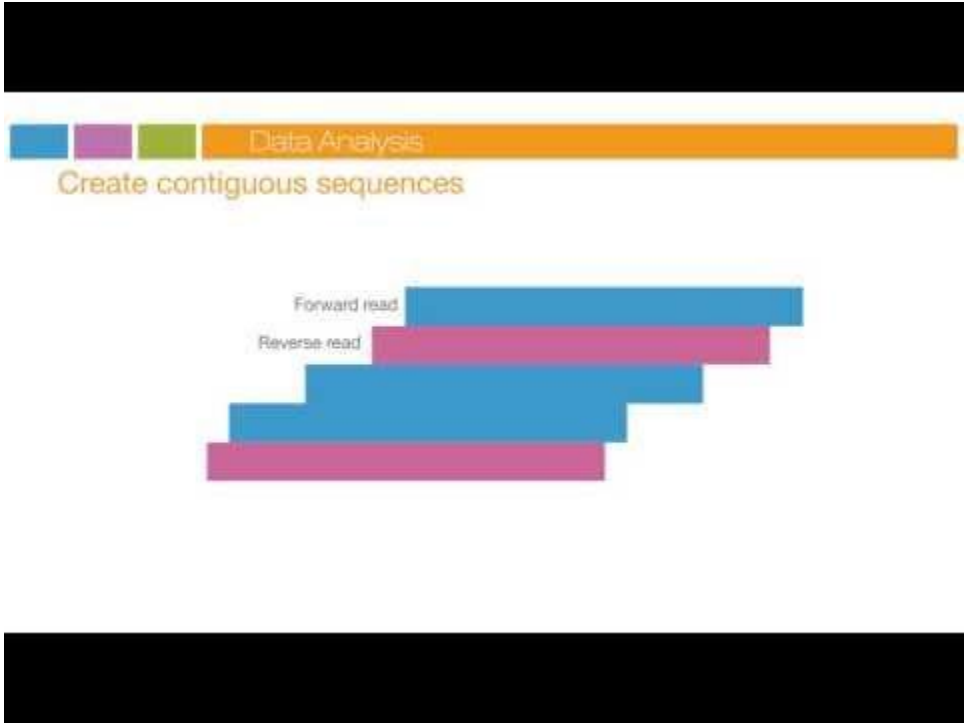
Genotype

Characterize environments

More exotic stuff...



# The Illumina sequencing method

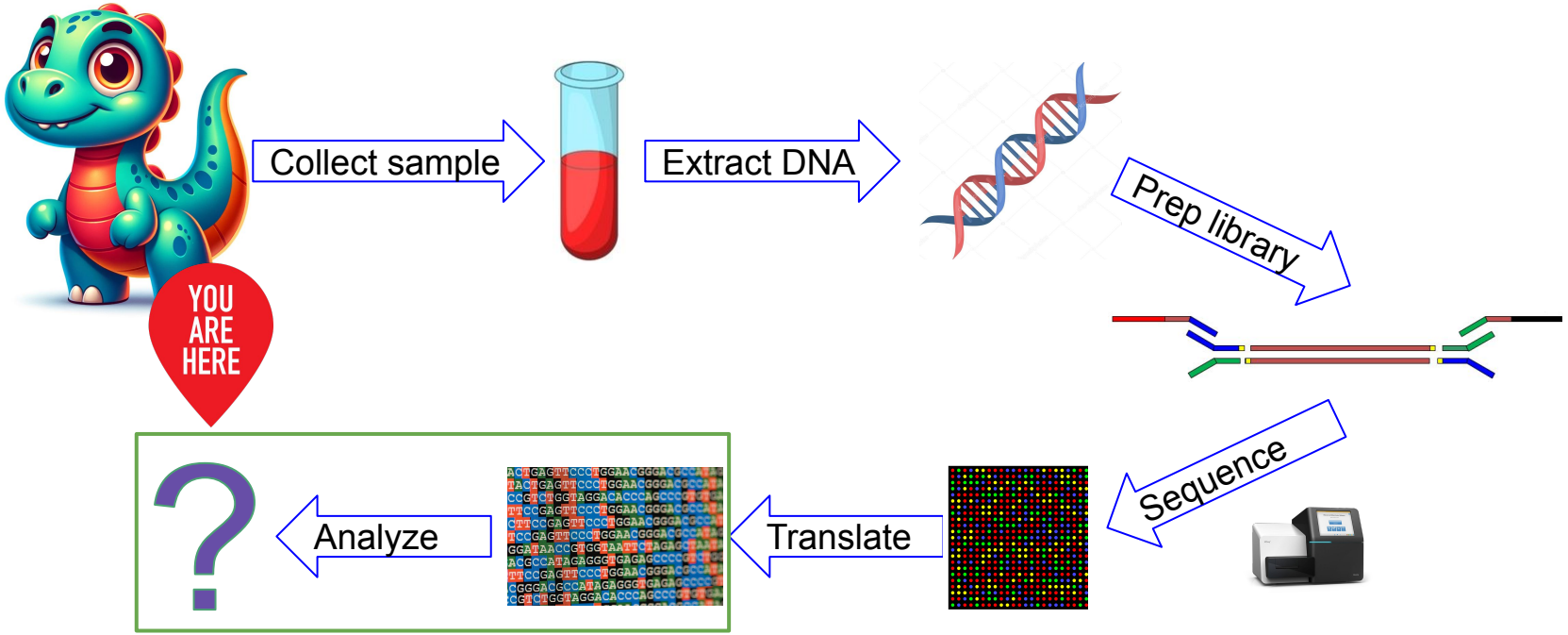


# Illumina NovaSeqX


# Illumina sequencing instruments

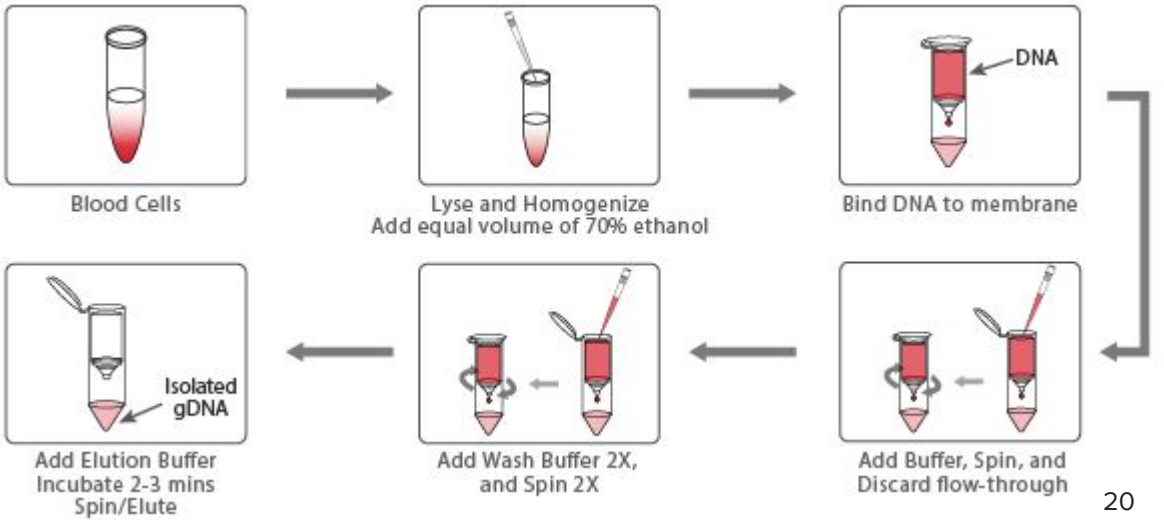
| | iSeq | MiniSeq | MiSeq | NextSeq | HiSeq 4000 | HiSeq X | NovaSeq |
|---|---|---|---|---|---|---|---|
| **Run Time** | 9–17.5 hrs | 4–24 hours | 4–55 hours | 12–30 hours | < 1–3.5 days | < 3 days | 13-44 hours |
| **Maximum Output** | 1.2 Gb | 7.5 Gb | 15 Gb | 120 Gb | 1500 Gb | 1800 Gb | 6000 Gb |
| **Maximum Reads Per Run** | 4 million | 25 million | 25 million | 400 million | 5 billion | 6 billion | 20 billion |
| **Maximum Read Length** | 2 × 150 bp | 2 × 150 bp | 2 × 300 bp | 2 × 150 bp | 2 × 150 bp | 2 × 150 bp | 2 x 250 |

# NGS - from organism to sequence



# DNA/RNA extraction

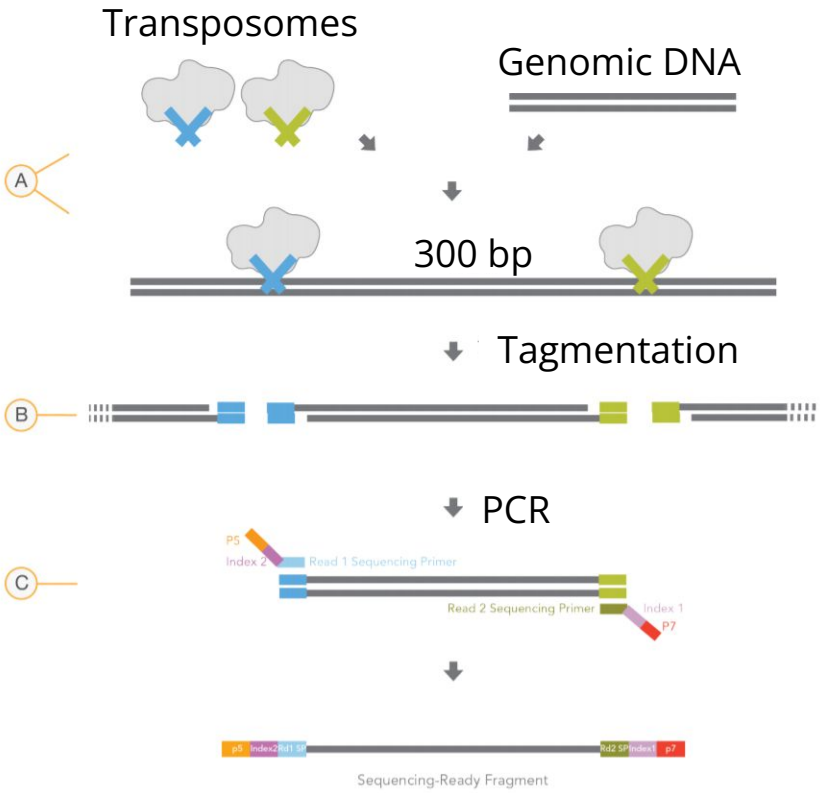
- Protocol depends on organism and tissue
- Required DNA amount depends on application
- DNA should be as intact as possible
  - Fragment sizes
  - Single strand breaks
- May be challenging!



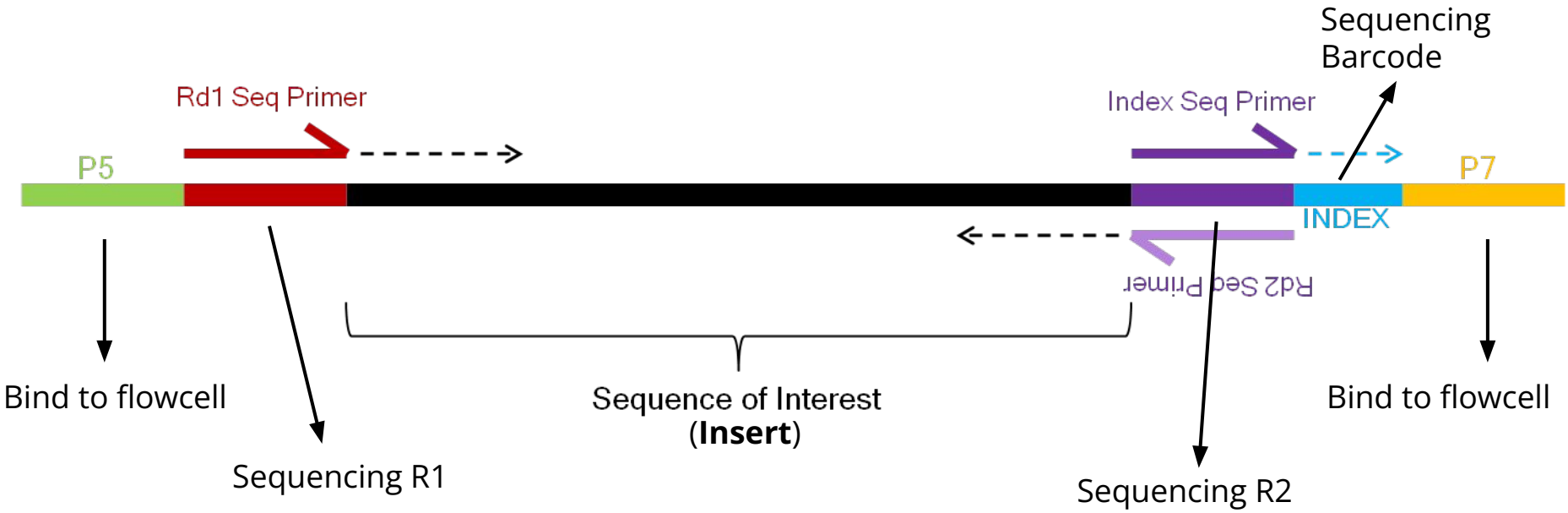
# Library prep

- Many protocols exist
- Main goal - fragment and add adapters
- **Single-end** or **paired-end**?
- Whole genome sequencing (**WGS**) or **targeted** sequencing?
- **Insert size** is determined

# Basic WGS library prep example



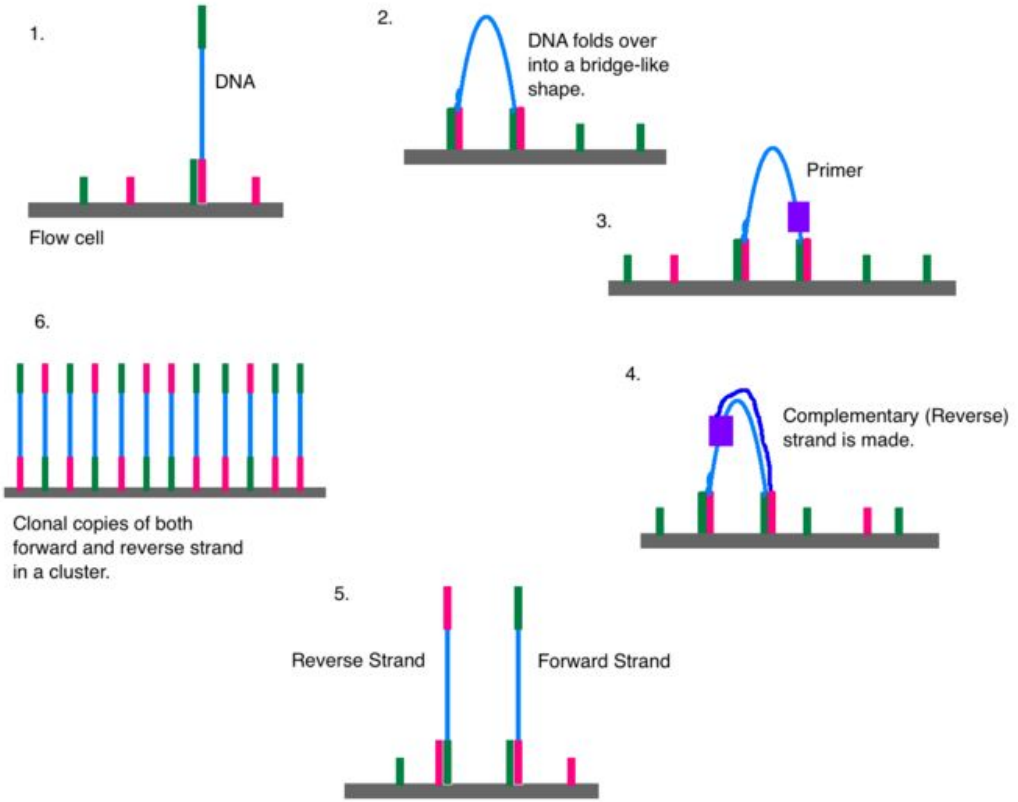
# Library prep - what do we get?



# And finally... Sequencing

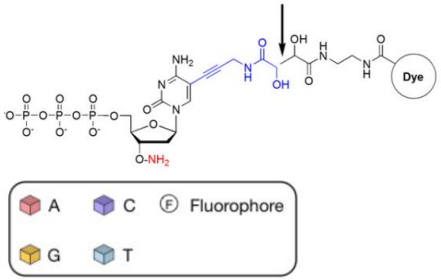
STEP 1 - Cluster amplification
**No sequencing**

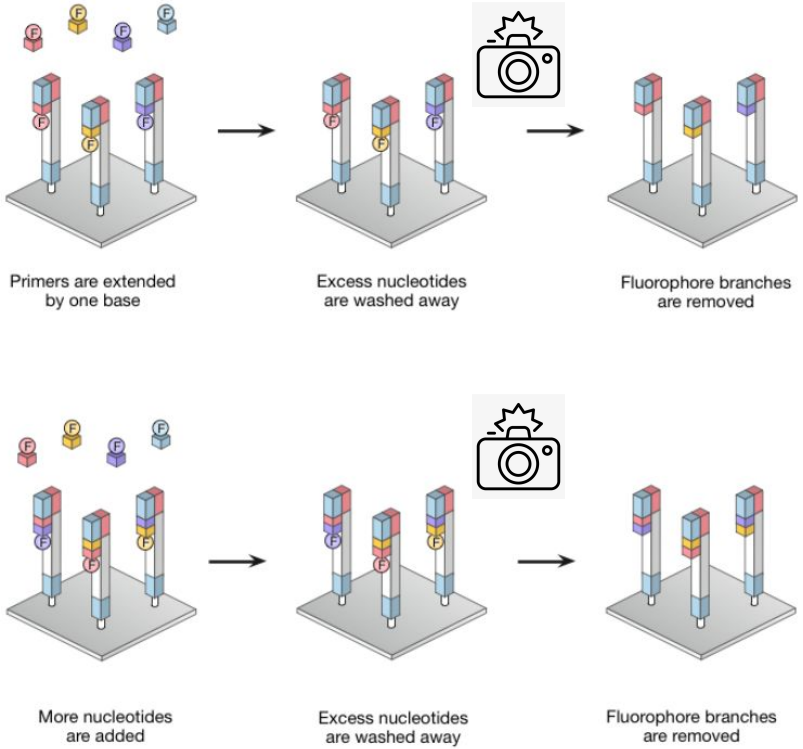
**Do we use tagged nucleotides in cluster amplification?**



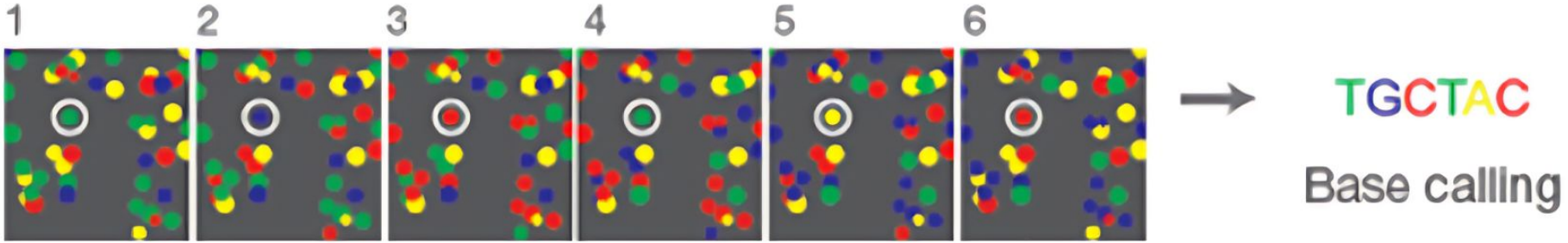
# And finally... Sequencing

STEP 2 - Sequencing



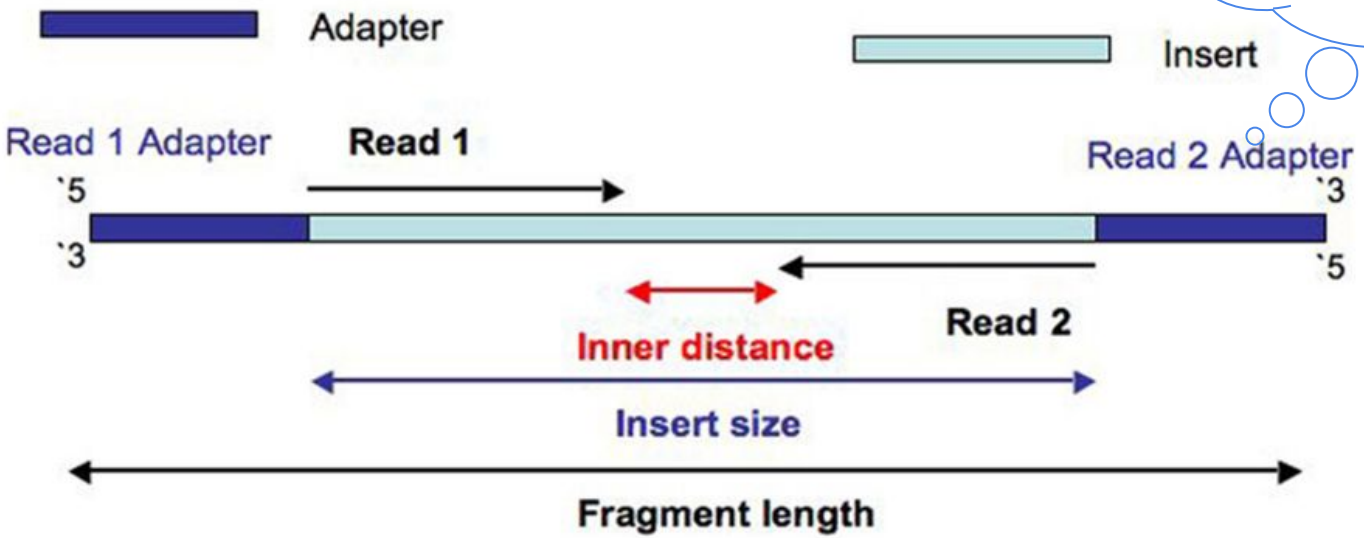



# Output of the sequencing machine




- Each flowcell cluster results in a read or read pair

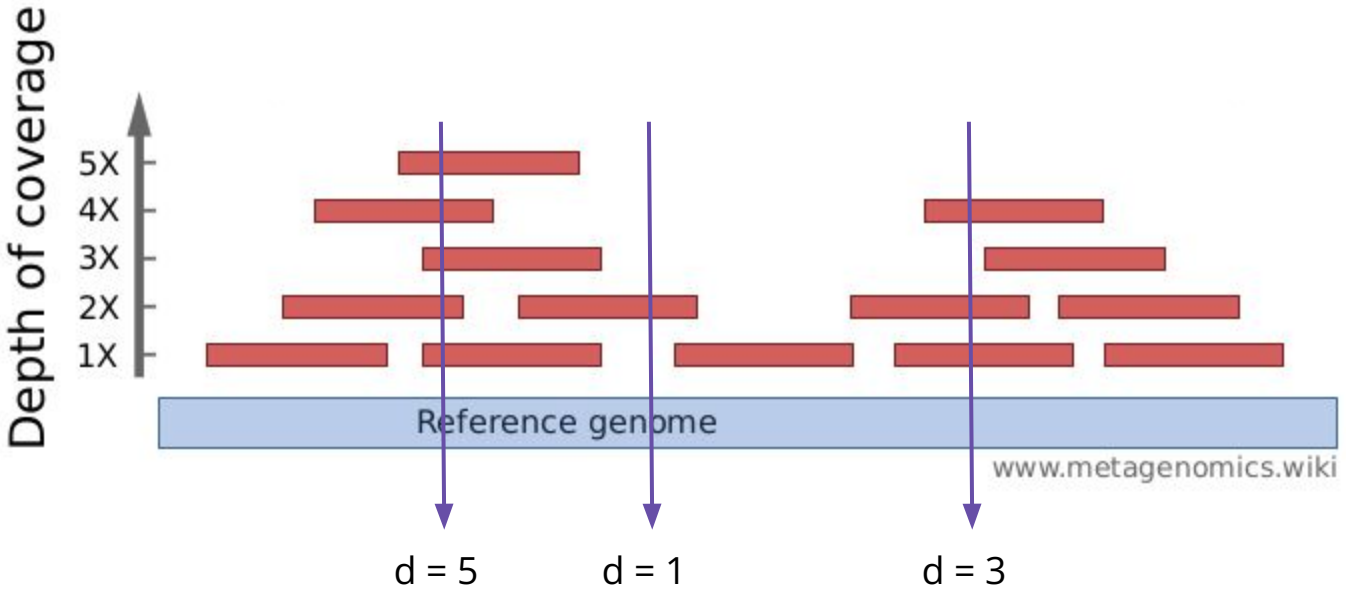
# Some terminology




Are Read1 and Read2 reverse-complement of each other?

Read length - fixed - typically 50 < L < 250
Insert size - distributed - limitation - up to ~700 bp

# Sequencing depth



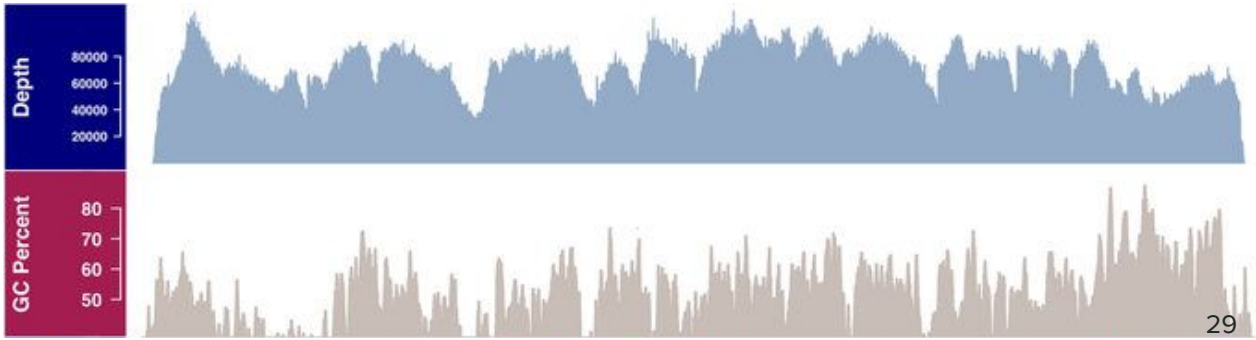
# More about depth

- Not uniform!
  - GC content
  - Low complexity regions
- Consider desired depth when designing the experiment

$$D = \frac{L * N}{G}$$

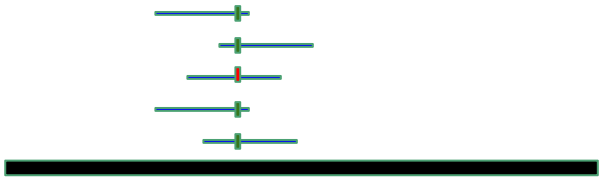
Notation: 5×, 10×, 100×...

D - **average** depth
L - read or pair length
N - number of reads
G - genome size



# Sequencing errors

- Error rate estimated as 1-3/1000 [1]
- ~6.5% of reads contain sequencing errors
- Higher depth can help us filter/correct errors
- Causes:
  - Color cross talk
  - Clusters cross talk
  - Out-of-sync base incorporation (phasing)



1. Pfeiffer, Franziska, et al. "Systematic evaluation of error rates and causes in short samples in next-generation sequencing." Scientific reports 8.1 (2018): 10950.

| | Sanger sequencing | Illumina NGS |
|---|---|---|
| Basic method | | |
| Throughput and cost | | |
| # of molecules sequenced at once | | |
| Read length | | |
| Paired reads | | |
| Sequencing error rate | | |
| **When should we use** | | |

| | Sanger sequencing | Illumina NGS |
|---|---|---|
| Basic method | SBS (chain termination) | SBS (reversible chain termination) |
| Throughput and cost | 70 kb/hr ; $500/Mb | 10 Gb/hr ; $0.01/Mb |
| # of molecules sequenced at once | 1 | millions-billions |
| Read length | 500-1000 | 50-250 |
| Paired reads | No | Optional |
| Sequencing error rate | 0.001% | 0.1% |
| **When should we use** | Sequence one or few sequences | WGS, many genes, RNA-seq |

# 2010s - 3rd generation sequencing

- Single molecule sequencing - no need for amplification
- Long read technologies:
  - PacBio SMRT sequencing - 10-20 kb read length
  - Oxfort Nanopore Technologies X-ION - 10 kb - 1 Mb long
- Linked reads - 10X genomics
- Various technical and computational challenges

# 2020s - 2nd generation of 2nd generation sequencing


BGI

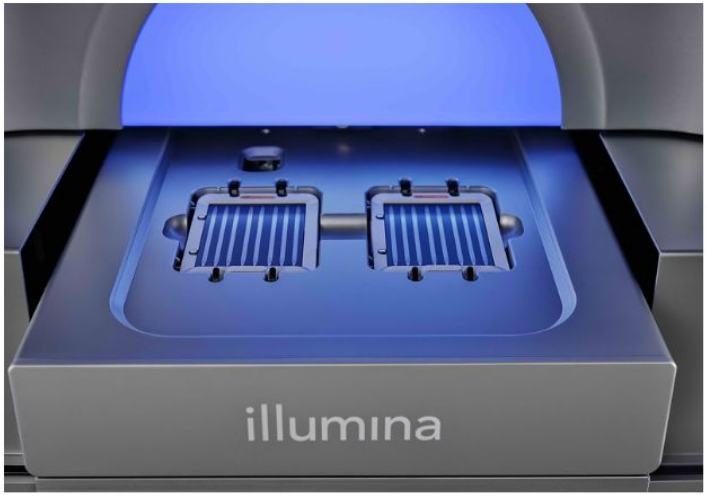

Element Biosciences

ULTIMA GENOMICS

Markets
Prognosis

# Illumina Aims to Push Genetics Beyond the Lab With $200 Genome

- New machine pushes company toward goal of $100 genome sequence
- Could be boon for gene-targeting drugs, Regeneron exec says



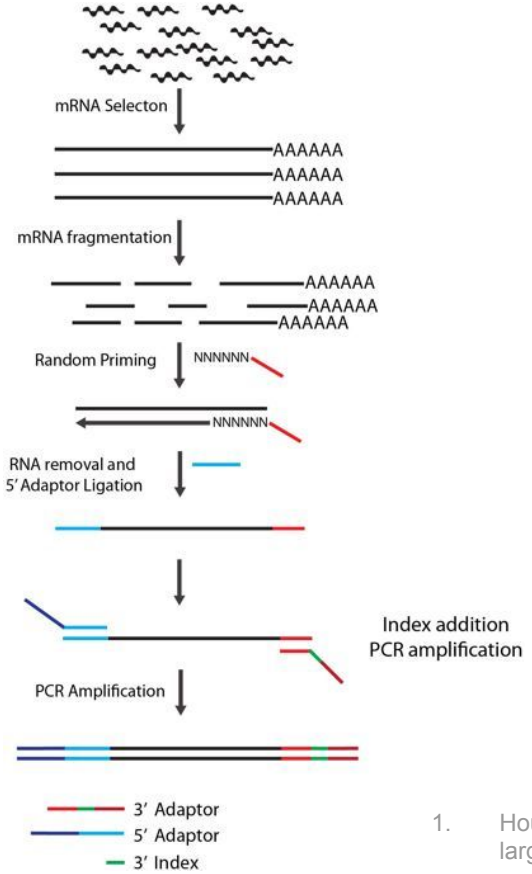

Illumina's new NovaSeq X series genome sequencers *Source: Illumina*

By Angelica Peebles
September 29, 2022 at 7:15 PM GMT+3 *Updated on September 29, 2022 at 9:11 PM GMT+3*

# Specialized NGS protocols

- Sequence total RNA from a sample
- Sequence only the exome
- Detect protein binding sites
- Represent the genome with fewer reads
- Determine DNA methylation sites

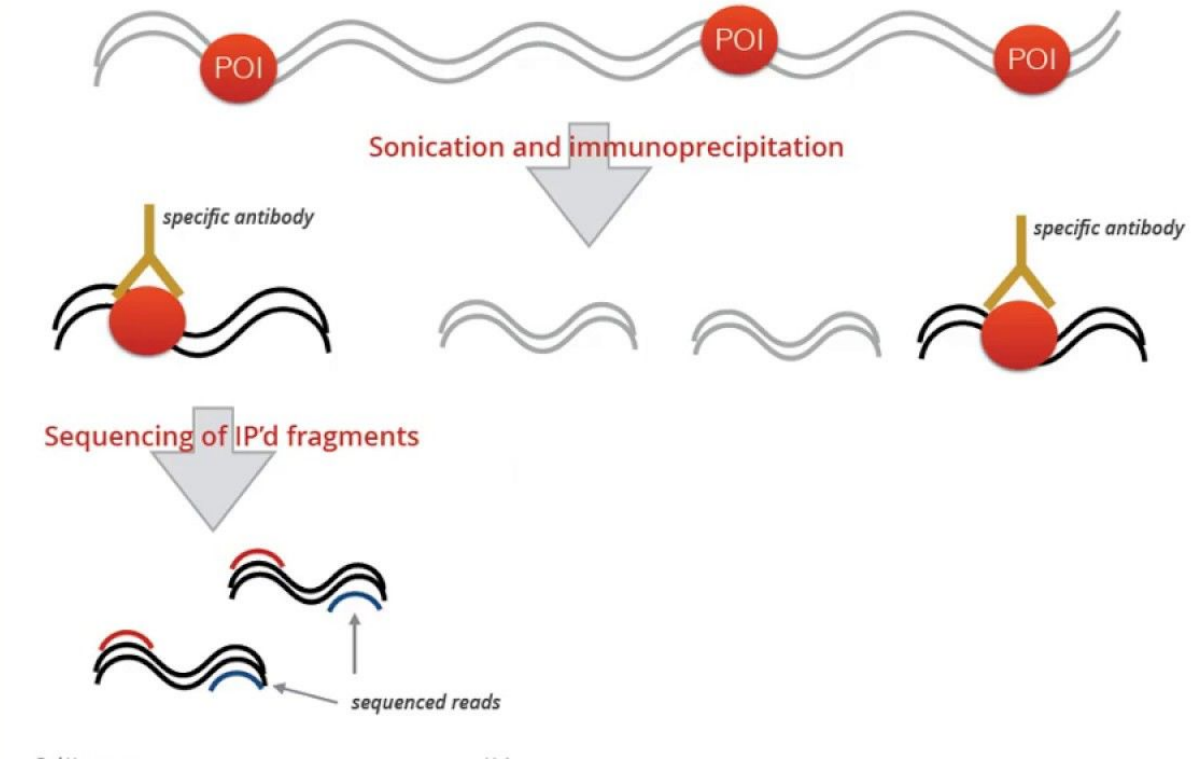
# RNA-Seq




- Must capture mRNA
- Otherwise we mostly get rRNA
- Oligo dT baits or remove rRNA (Ribo-Zero)
- Higher expression = more reads
- Sample of many cells ➜ we see the average!
- Single cell RNA-Seq

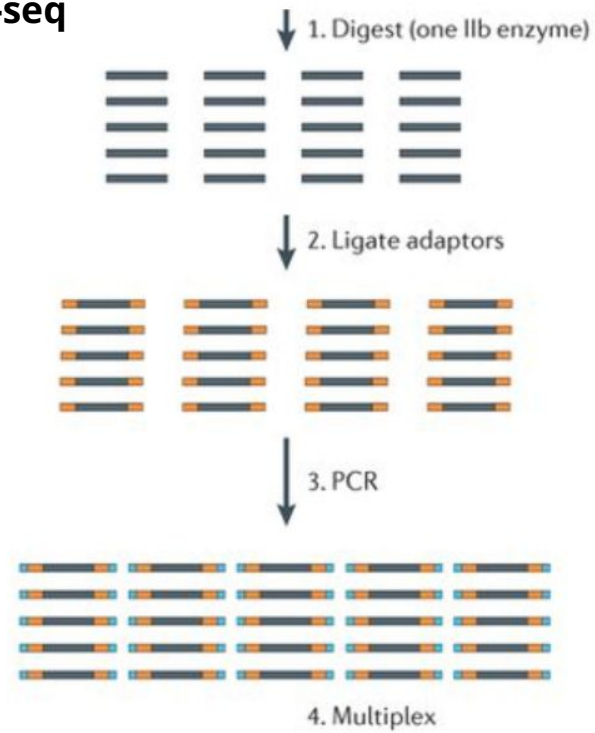
1. Hou, Zhonggang, et al. "A cost-effective RNA sequencing protocol for large-scale gene expression studies." Scientific reports 5 (2015): 9570.

# ChIP-Seq


POI
POI
POI
Sonication and immunoprecipitation
specific antibody
specific antibody
Sequencing of IP'd fragments
sequenced reads

# Reduced-representation sequencing

**RAD-seq**




- Low budget projects
- Limited genomic resources available
- Only look at specific genomic regions

1. Andrews, Kimberly R., et al. "Harnessing the power of RADseq for ecological and evolutionary genomics." Nature Reviews Genetics 17.2 (2016): 81.

# Current challenges of genomics and NGS

- Genomic and NGS data are already all over the place
- Generating new data is relatively easy and cheap
- How do we:
  - Store and share?
  - Analyze effectively?
  - Integrate genomic data from various sources?
  - Extract biological insights?

# Conclusion questions

- Define the terms:
  - Read length
  - Paired end
  - Insert size
  - Sequencing depth
- What are the major steps of an NGS experiment?
- The wheat genome is ~17 Gb in size. How many read pairs with read length 150 do we need to sequence it at x100 depth?