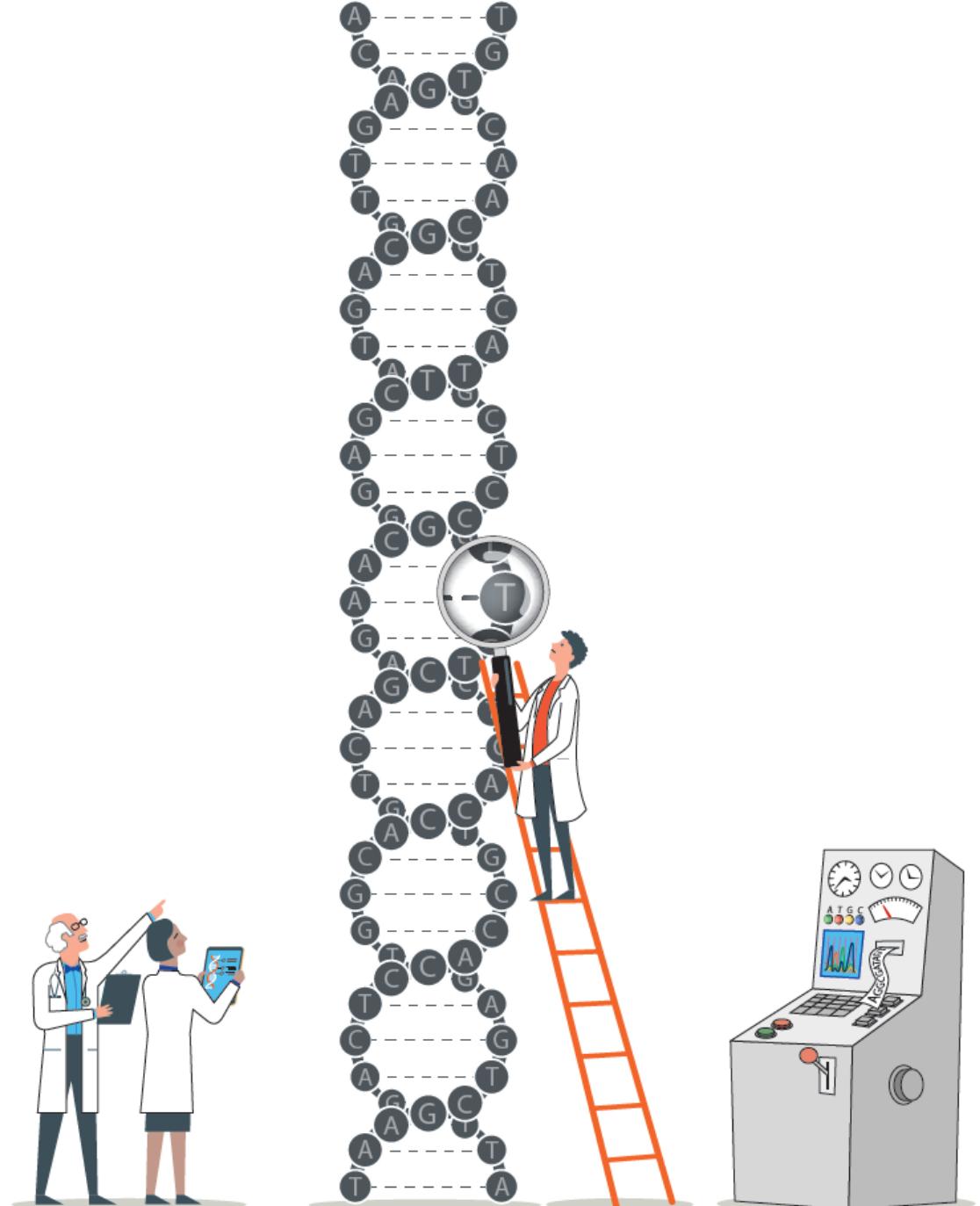


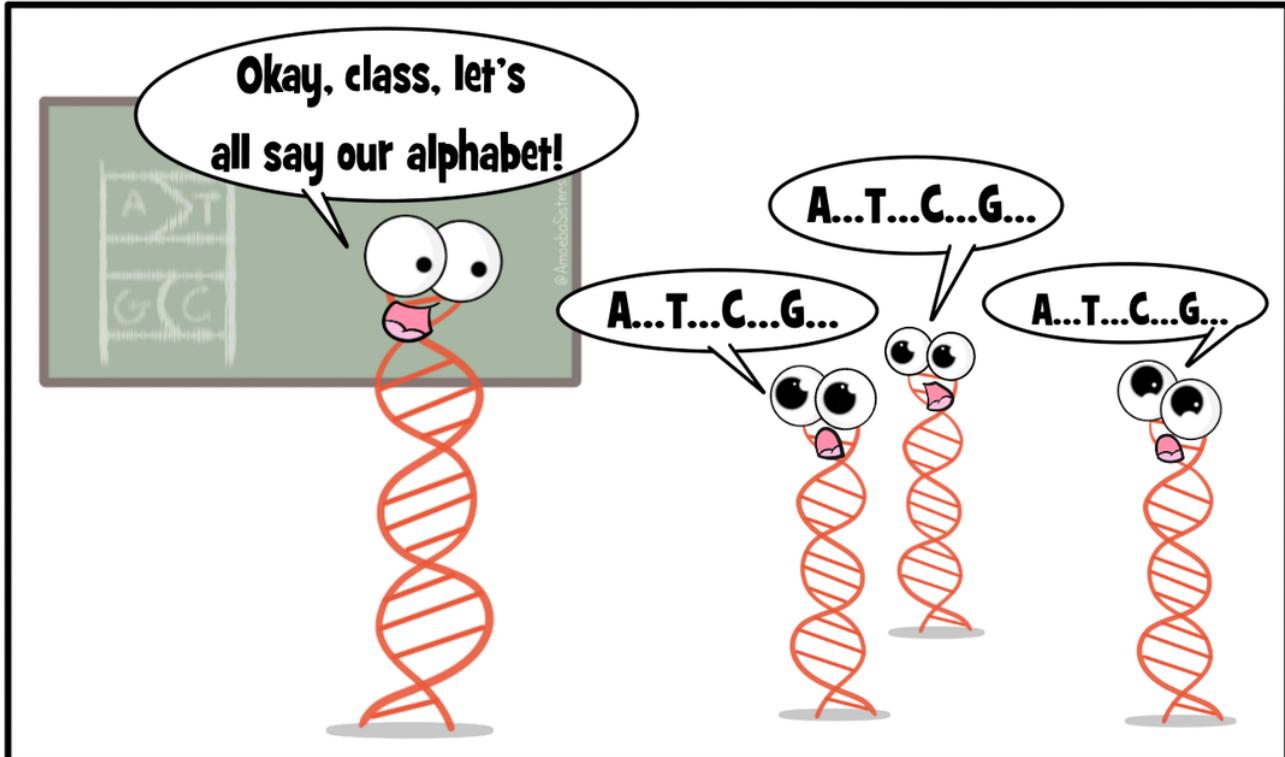
# DNA SEQUENCING TECHNOLOGIES IN 2022

Molly Zeller, MS (zeller2@wisc.edu)  
University of Wisconsin Biotechnology Center  
DNA Sequencing Facility



# AGENDA

## Paramecium Parlor



- **Introduction**
  - Who am I?
- **History of Sequencing**
  - How did we get here?
- **Technologies in 2022**
  - Illumina
  - PacBio
  - Oxford Nanopore
  - Newly Launched
- **Applications**
  - DNA
  - RNA
- **Library Quality Control**
- **Data Quality Control**

I will have QR codes in the presentation so have your phone ready if you would like to scan them!

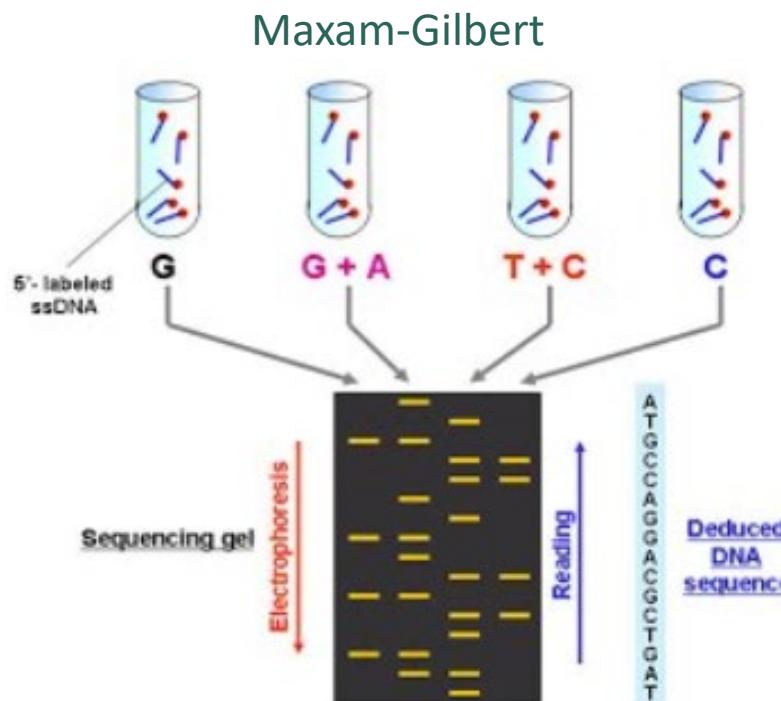
# INTRODUCTION

- UWBC
  - Founded in 1984
  - Located in Madison, Wisconsin
- Many cores in our center
  - DNA Sequencing
  - Gene Expression Center
  - Bioinformatics Resource Center
  - Genome Editing/Animal Models
  - Mass Spectrometry
  - Flow Cytometry



# HISTORY

- First Generation Sequencing!
  - Maxam-Gilbert (chemical method)
  - Sanger dideoxy (chain termination method)



• <https://www.youtube.com/watch?v=cl2s-ZMmcbc>

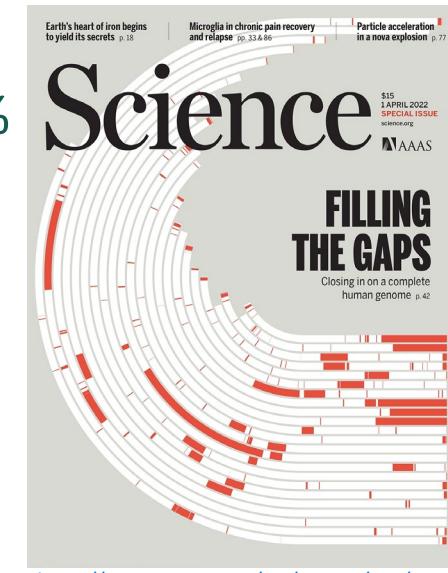


<https://youtu.be/e2G5zx-OJlw>

# HISTORY

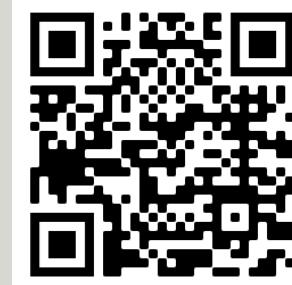
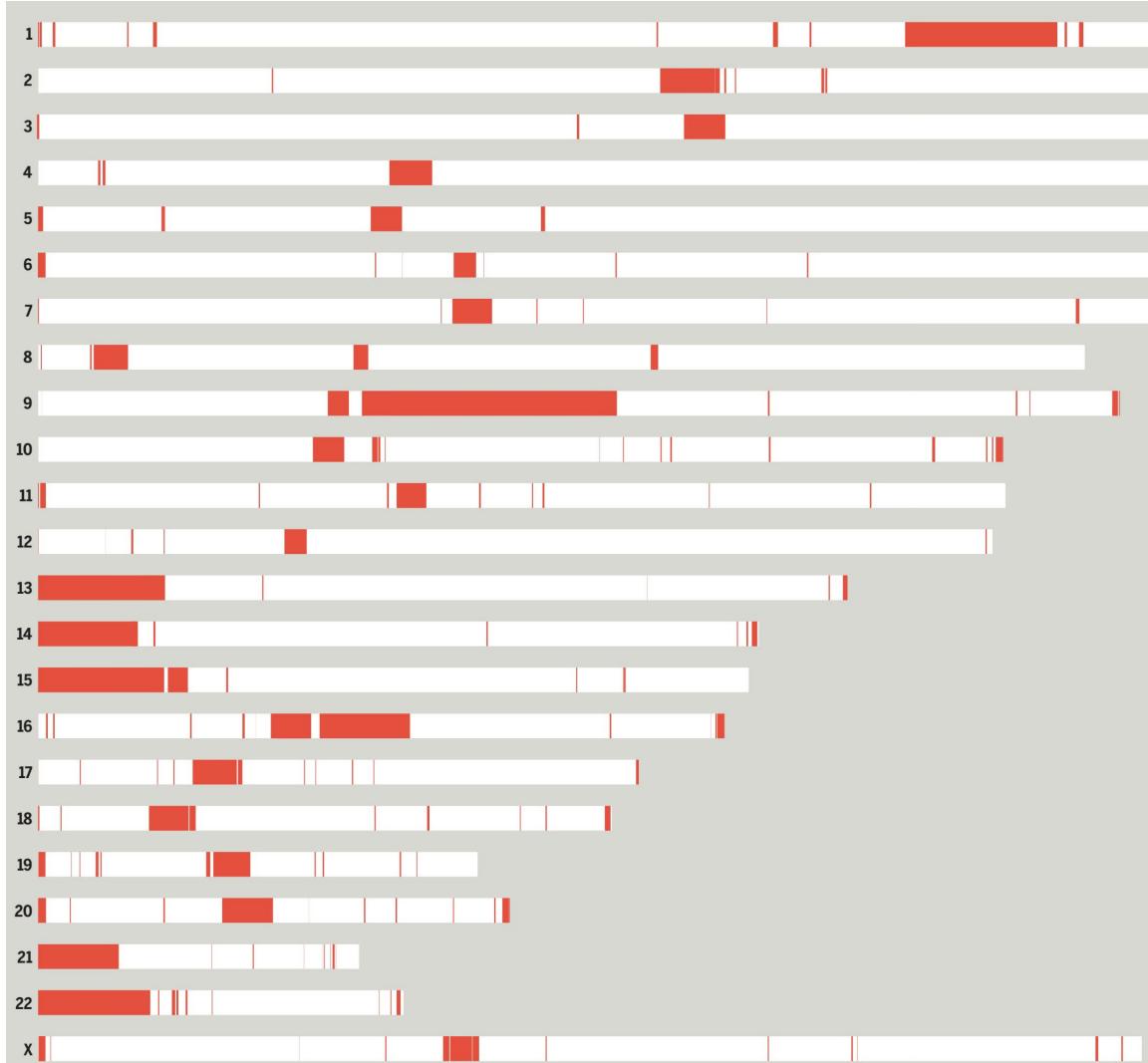


- ✓ 1995 – *Haemophilus influenzae*
- ✓ 1996 – *Saccharomyces cerevisiae*
- ✓ 1997 – *Escherichia coli* K-12 (Blattner, UW Madison)
- ✓ 1998 – *Caenorhabditis elegans*
- ✓ 1999 – Human Chromosome 22
- ✓ 2000 – *Drosophila melanogaster*
- ✓ 2001 – First draft of human genome
- ✓ 2002 – *Mus musculus*, *Plasmodium falciparum*
- ✓ 2003 – Human Genome Project completed (99.99% accuracy)
- ✓ 2004 – *Rattus norvegicus*
- ✓ 2005 – Chimpanzee
- ✓ 2010 – Neanderthal
- ✓ 2013 – *Danio rerio*
- ✓ 2019 – Telomere-to-Telomere Consortium was founded
- ✓ 2022 – Telomere-to-Telomere Consortium provided the first gapless assembly



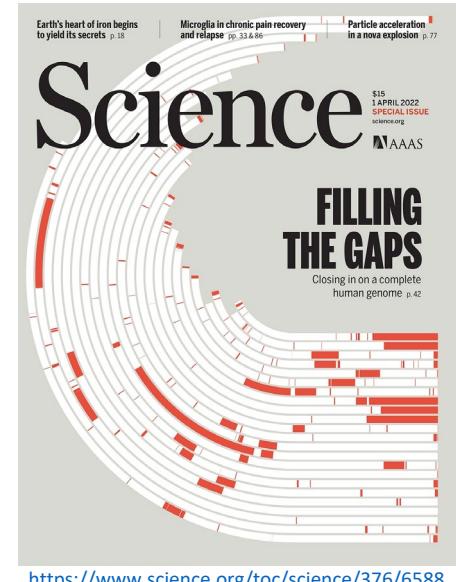
<https://www.science.org/toc/science/376/6588>

# HISTORY



## Abstract

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion-base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

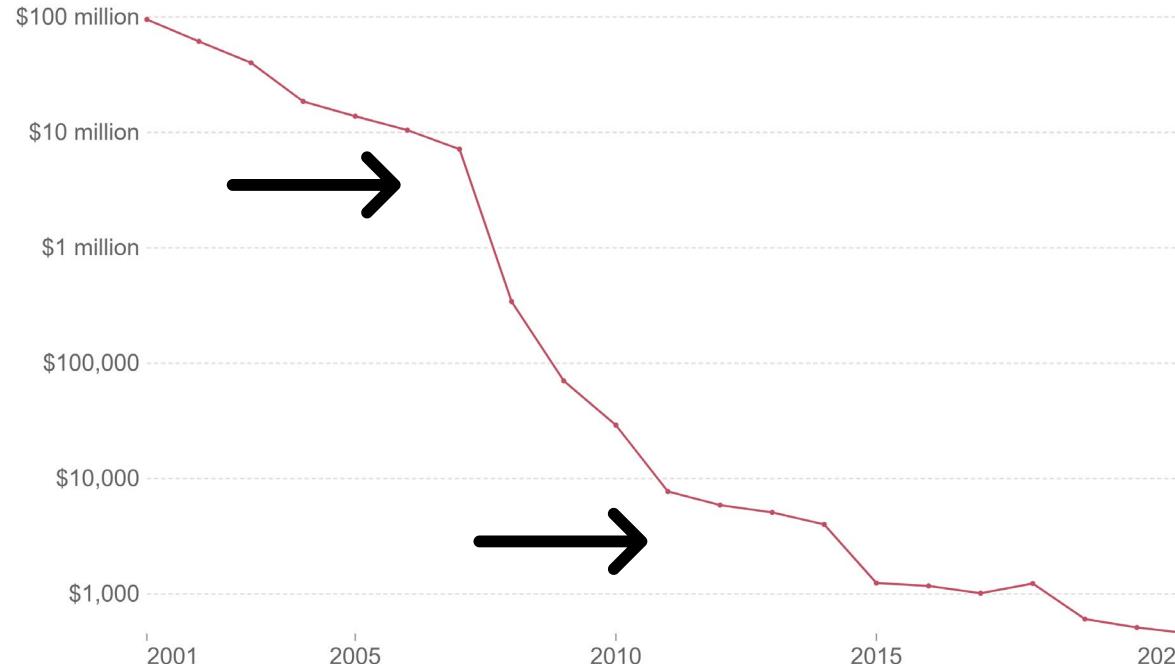


<https://www.science.org/toc/science/376/6588>

# HISTORY

## Cost of sequencing a full human genome

The cost of sequencing the DNA of the complete human genome, measured in US\$. This data is not adjusted for inflation.



Source: National Human Genome Research Institute (2022)

OurWorldInData.org/technological-change • CC BY

Our World  
in Data

## Cost of sequencing a full human genome

The cost of sequencing the DNA of the complete human genome, measured in US\$. This data is not adjusted for inflation.



Source: National Human Genome Research Institute (2022)

OurWorldInData.org/technological-change • CC BY

Our World  
in Data

# TECHNOLOGIES

Illumina



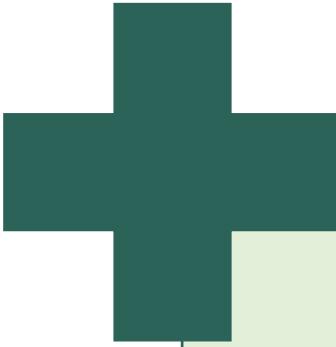
PacBio



Oxford  
Nanopore  
(ONT)



# TECHNOLOGIES



## Short Read Sequencing (<1kb)



- |   |  |
|---|--|
| <ul style="list-style-type: none"><li>- Cost per base is lower than LR</li><li>- Works well for SNP detection</li><li>- Great for gene expression studies or ChIP-seq</li></ul> | <ul style="list-style-type: none"><li>- PCR bias introduced</li><li>- Difficulty sequencing high GC regions</li><li>- Difficulty in resolving substitution, deletions, duplications, etc</li></ul> |
|---|--|

# TECHNOLOGIES

## Long Read Sequencing (>1kb)

- No PCR bias
- Can determine phasing, haplotypes, structural variations, indels, etc
- Aids in *de novo* assembly by spanning low complexity and repetitive regions
- Can determine naturally occurring base modifications such as methylation
- Cost per base is much higher than SR
- Throughput is low
- Need more input DNA
- Analysis tools are still evolving
- Storage AND analysis needs are greater

# TECHNOLOGIES



## Fast facts



Founded in  
**1998**



Number of employees  
**>9,100**



Annual revenue (2021)<sup>1</sup>  
**\$4.5B USD**



Cumulative sequencing  
installed base  
**>20,000**

Headquarters

**San Diego, California, USA**

Countries served

**>140**

President and CEO

**Francis deSouza**

# TECHNOLOGIES

illumina®



<https://youtu.be/fCd6B5HRaZ8>

# TECHNOLOGIES

illumina®



Genomic DNA



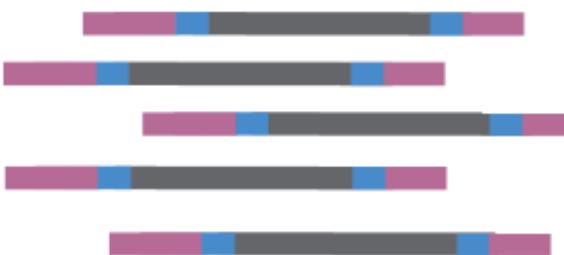
Fragmentation

Adapters



Ligation

Sequencing  
Library



# TECHNOLOGIES



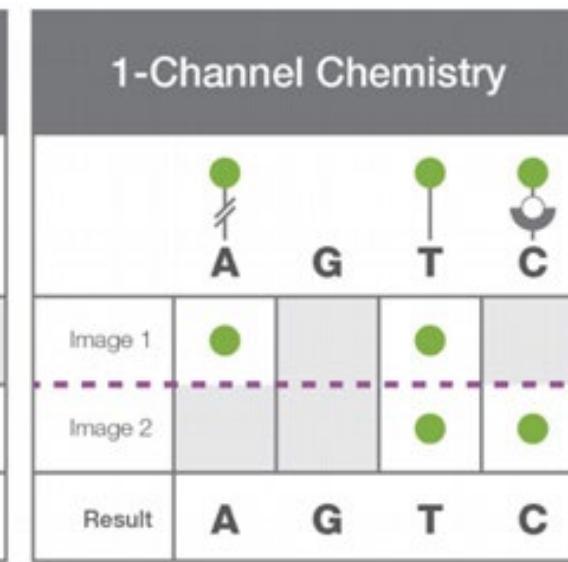
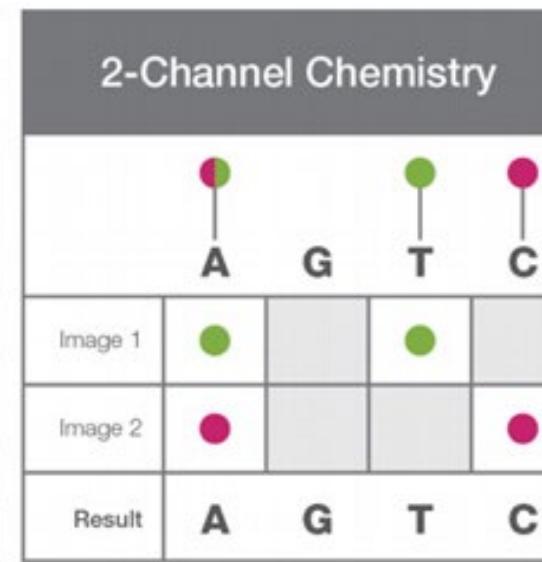
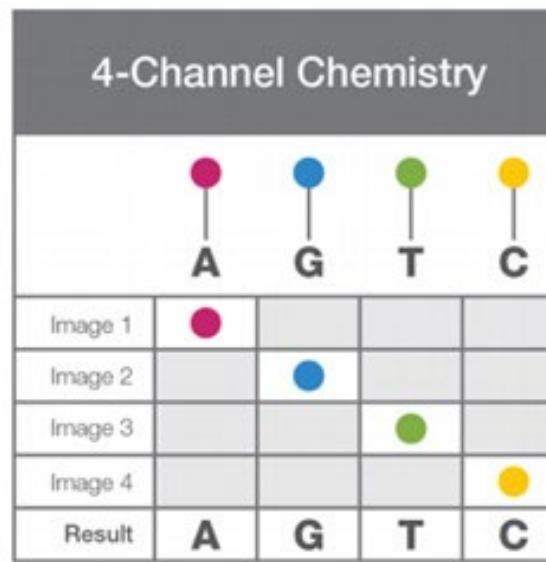
- Illumina MiSeq
  - Has the longest reads at 300bp
  - Great for amplicon sequencing and metagenomics
  - 4 channel chemistry

# illumina®



- Illumina NovaSeq6000

- Can generate up to 10B reads (clusters)
- Great for projects that need a LOT of reads – ex human whole genome
- 2 channel chemistry



# TECHNOLOGIES



NextSeq 1000 & 2000

NovaSeq 6000

NovaSeq X Series

## Popular Applications & Methods

Large Whole-Genome Sequencing (human, plant, animal)



Small Whole-Genome Sequencing (microbe, virus)



Exome & Large Panel Sequencing (enrichment-based)



Targeted Gene Sequencing (amplicon-based, gene panel)



Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)



Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)



Chromatin Analysis (ATAC-Seq, ChIP-Seq)



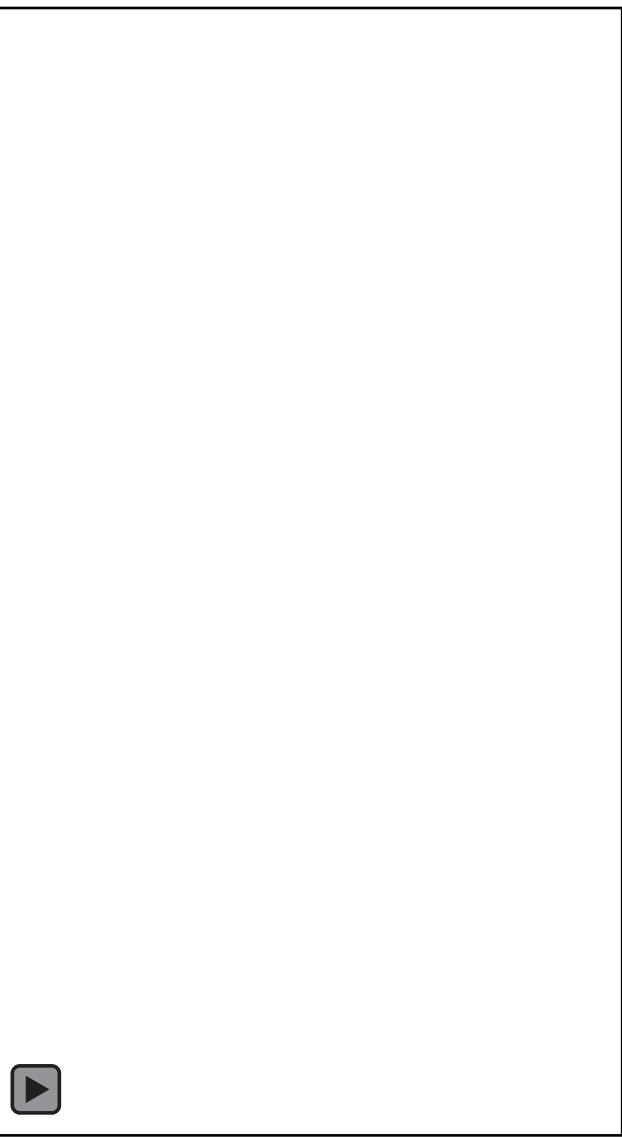
Methylation Sequencing



Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)



Cell-Free Sequencing & Liquid Biopsy Analysis



# TECHNOLOGIES

PacBio

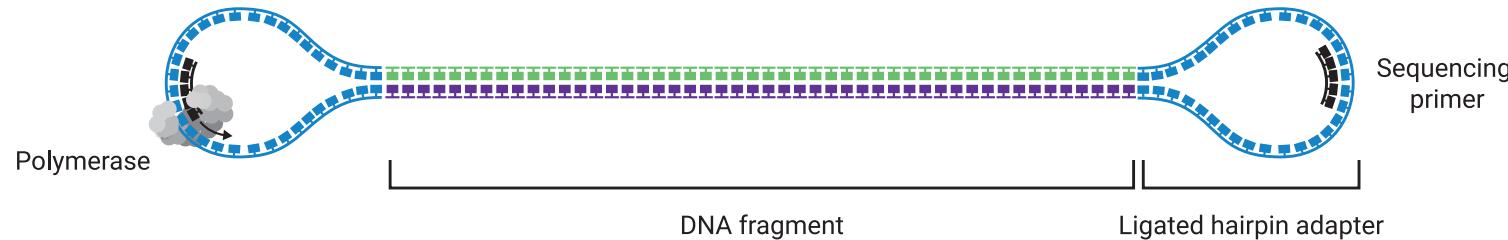


[https://youtu.be/\\_ID8JyAbwEo](https://youtu.be/_ID8JyAbwEo)

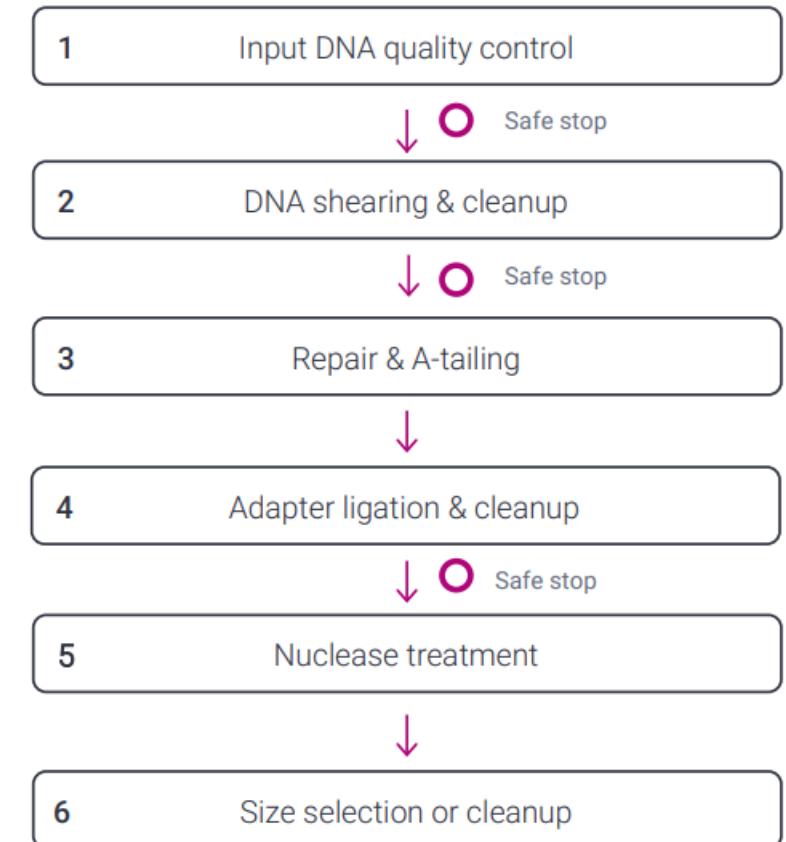
<https://www.pacb.com/blog/steps-of-smrt-sequencing/#:~:text=At%20the%20heart%20of%20SMRT,is%20measured%20in%20real%20time>

# TECHNOLOGIES

# PacBio



- Founded in 2004 (named Nanofluidics, Inc)
- Sequences **native** DNA
- Instruments
  - RS – 2010/11
  - RSII – 2013
  - Sequel – 2015
  - Sequel II – 2019
  - Sequel Ile – 2020

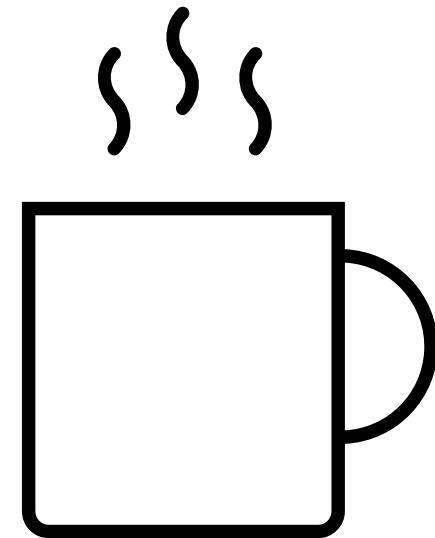
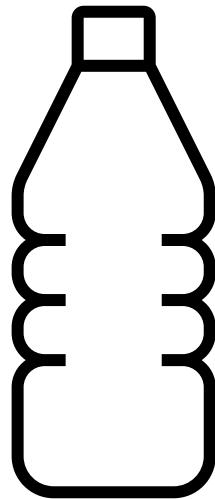


# TECHNOLOGIES



	<b>Sequel II system</b>	<b>Sequel system</b>
<b>Supported SMRT Cell</b>	SMRT Cell 8M	SMRT Cell 1M
<b>Number of HiFi reads &gt;99%* accuracy</b>	Up to 4,000,000	Up to 500,000
<b>Sequencing runtime per SMRT Cell</b>	Up to 30 hrs	Up to 20 hrs
<b>Instrument control software</b>	v10.1	v8.0
<b>SMRT Link</b>	v11.0	v10.2

# Break



# TECHNOLOGIES



<https://www.youtube.com/watch?v=RcP85JHLmnI>

# TECHNOLOGIES



- No limitations in length of molecules that can be sequenced
- Sequences **native** DNA
- Capable of detecting DNA and RNA modifications
- Flongle and MinION are portable!

- Has difficulty sequencing homopolymer regions
- Error rate is higher than other platforms
- Cannot sequence the same strand multiple times

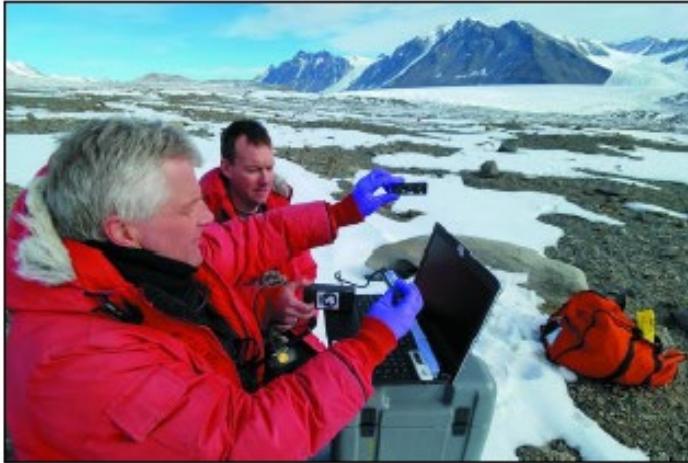
# TECHNOLOGIES



	MinION	GridION	PromethION
Flow cells	1	5	24 (48)
Real-time base calling	Needs MinIT	Yes	Yes
Channels	512	5 x 512	24 (48) x 3,000
Yield per flow cell	15-30 Gb	15-30 Gb	100-180 Gb
Yield per device	15-30 Gb	75-100 Gb	2.4-8.6 Tb



# TECHNOLOGIES



Sequencing microbial DNA in Antarctica



NASA astronaut sequencing bacteria in space



Sequencing Ebola virus in Guinea



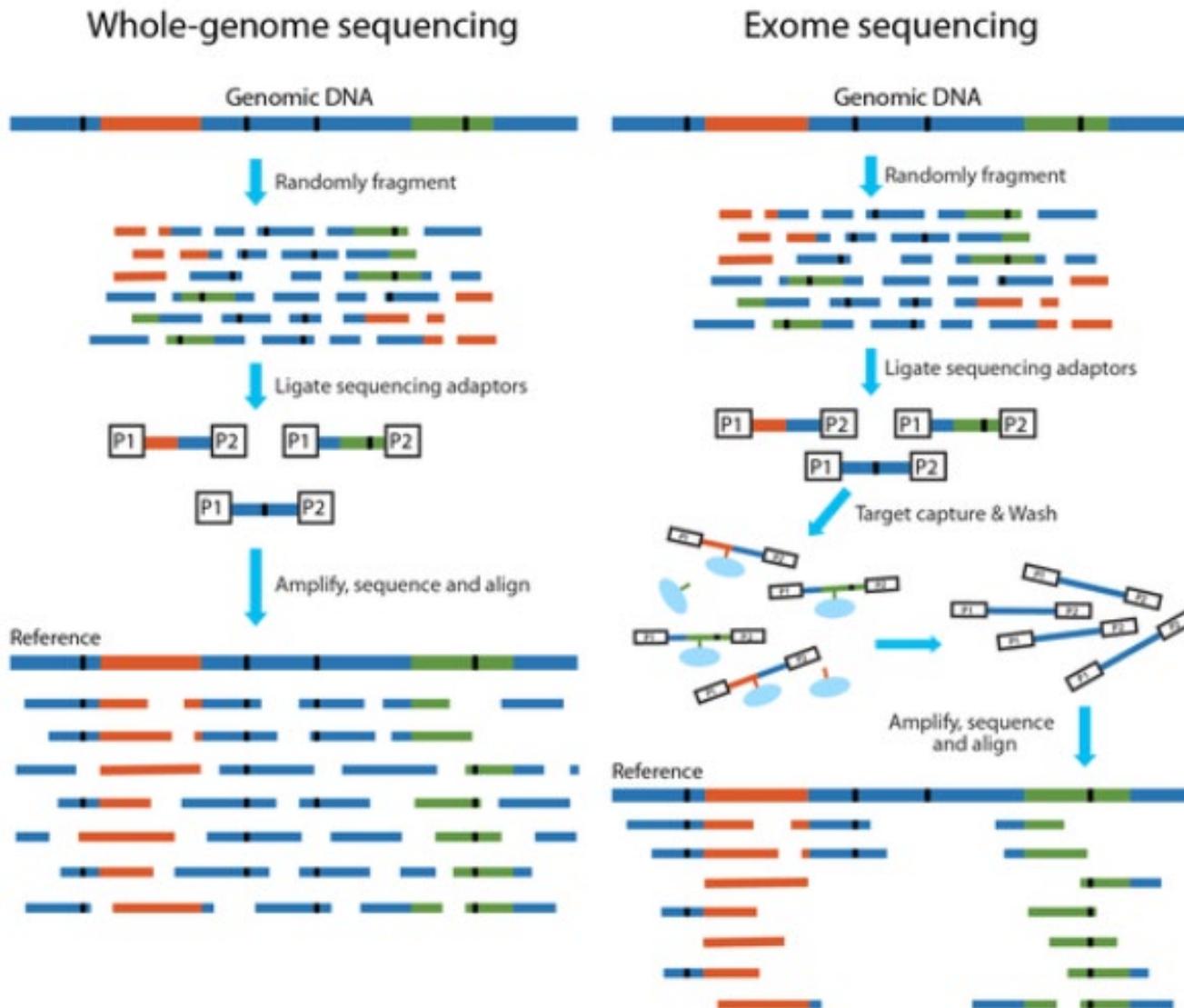
Sequencing Zika virus in Brazil

# TECHNOLOGIES



# APPLICATIONS - DNA

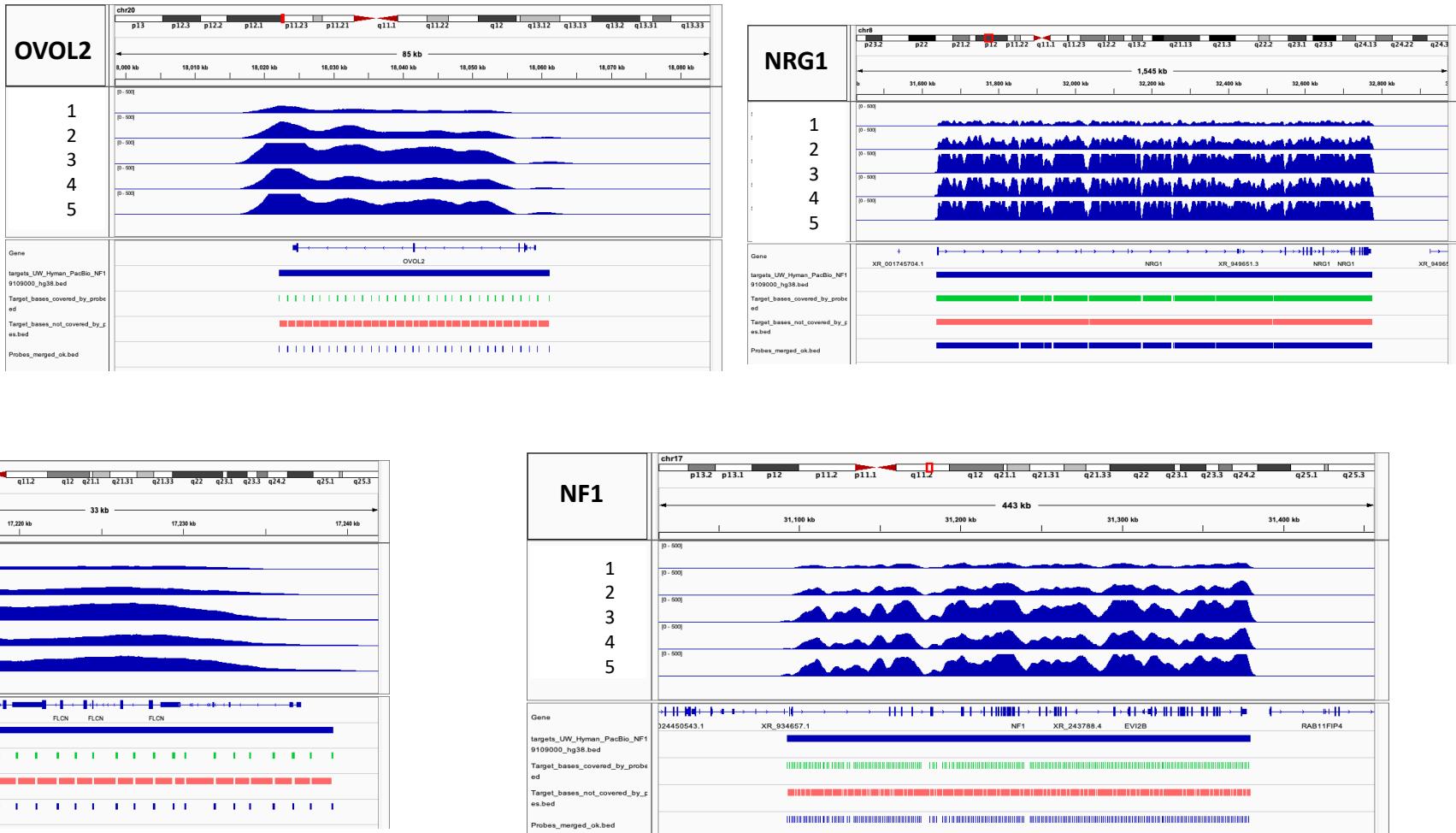
- Whole genome
- Exome
- Targeted (amplicon/gene)
- GBS
- Metagenome



# APPLICATIONS - DNA

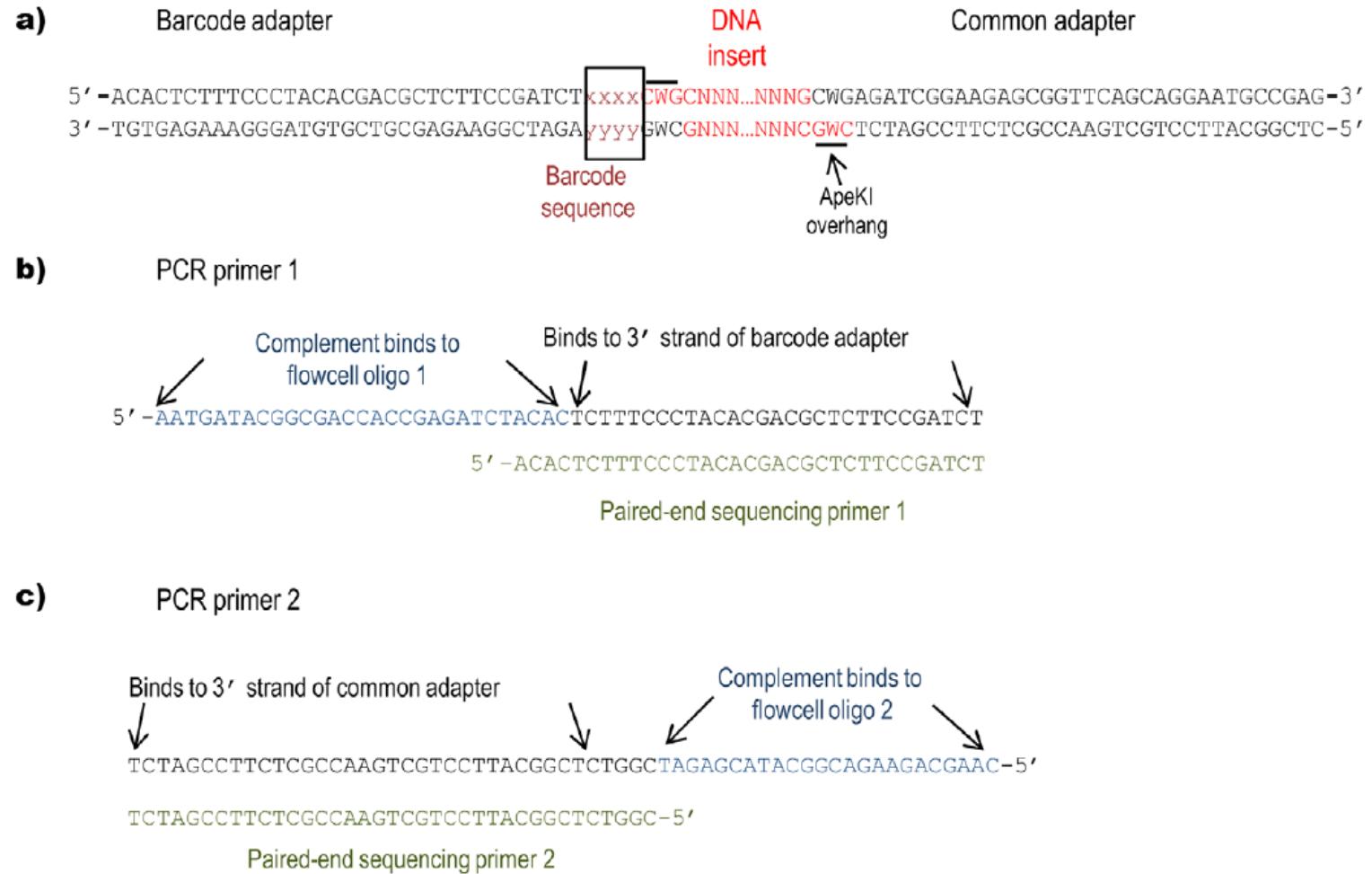


- Whole genome
- Exome
- Targeted (amplicon/gene)
- GBS
- Metagenome



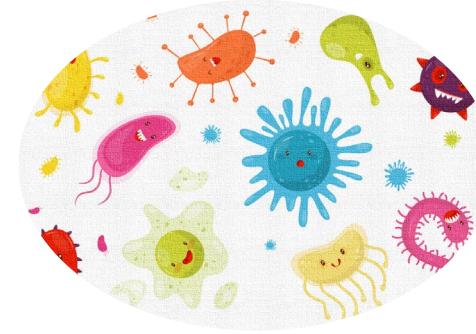
# APPLICATIONS - DNA

- Whole genome
- Exome
- Targeted (amplicon/gene)
- **Genotyping-by-Sequencing (GBS)**
- Metagenome



# APPLICATIONS - DNA

- Whole genome
- Exome
- Targeted (amplicon/gene)
- Genotyping-by-Sequencing (GBS)
- **Metagenome**



Amplicon Metagenomics

Whole Metagenomics

16S, 18S, ITS

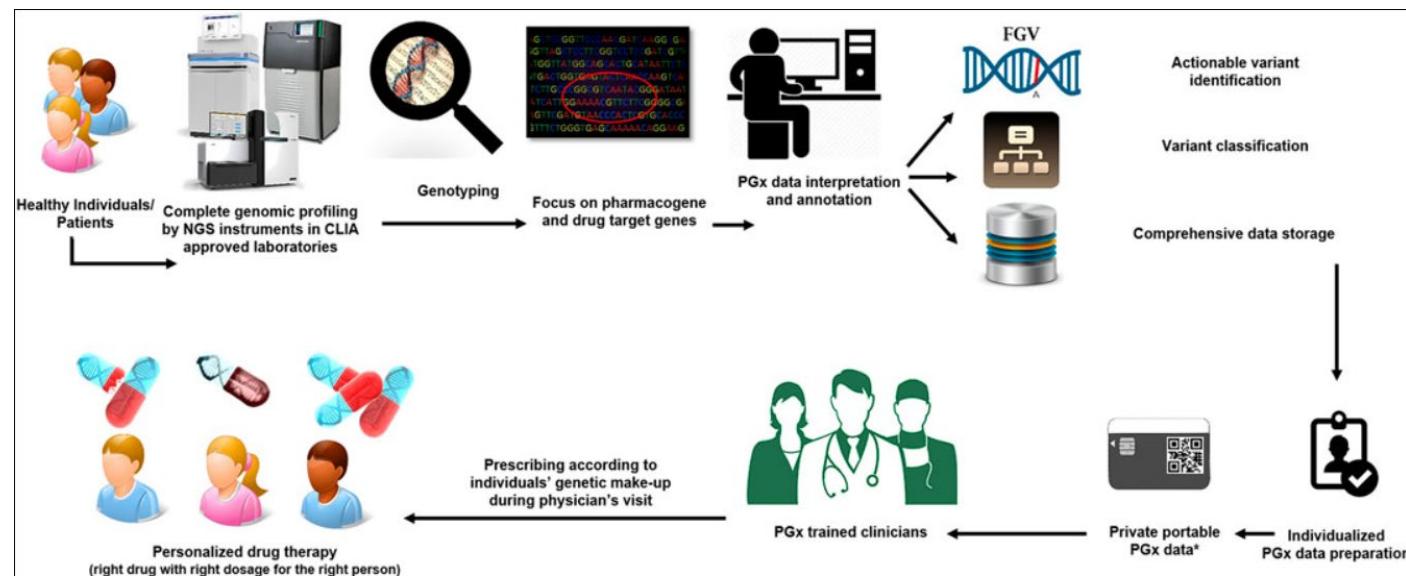
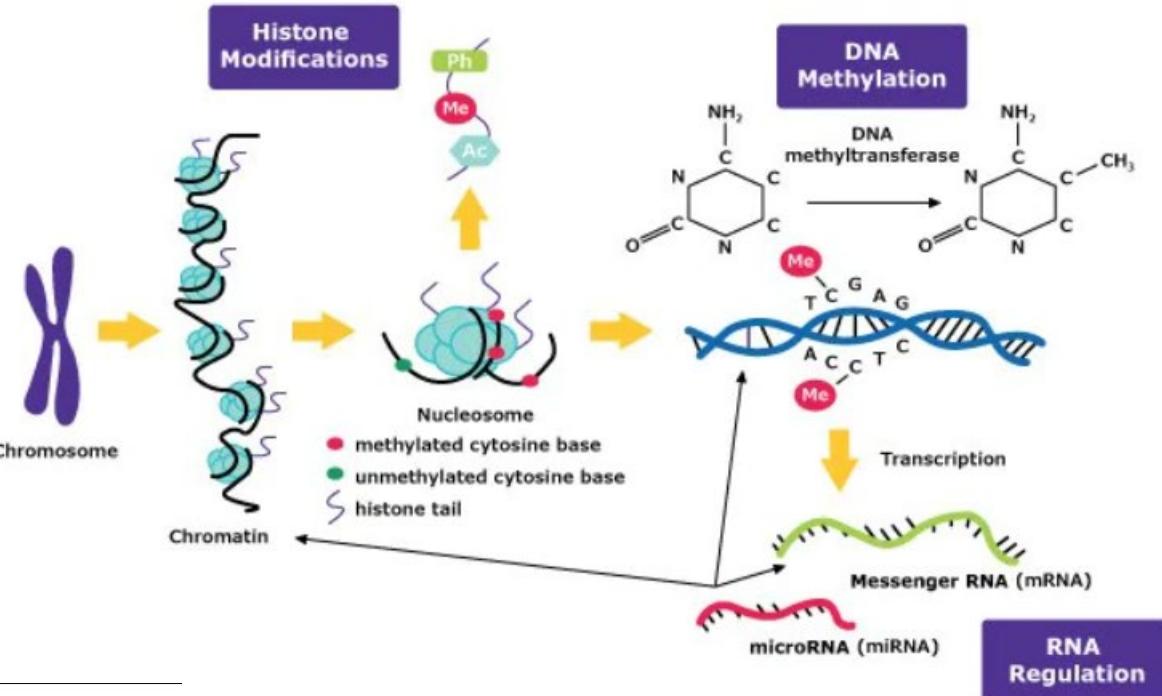
All genes

Community composition

Profiling, functional genomics

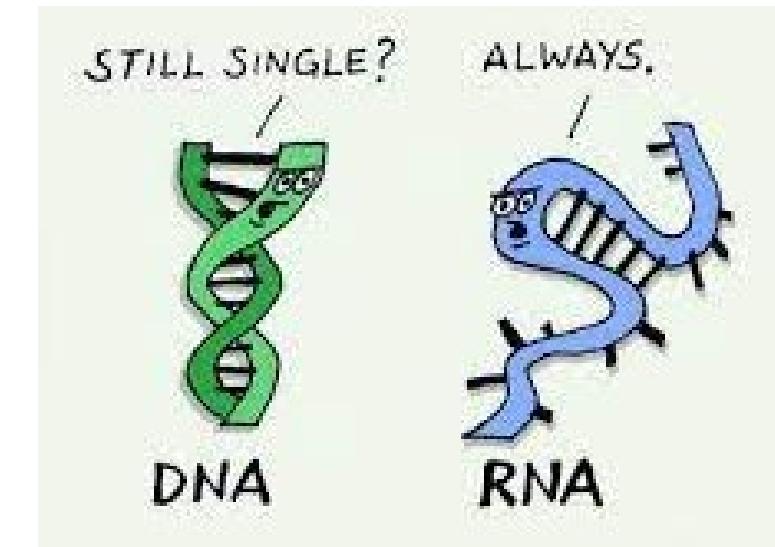
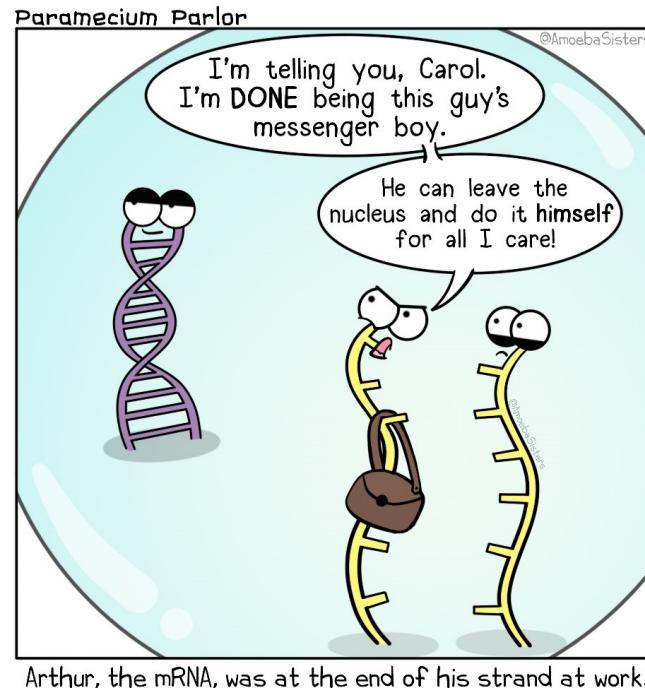
# APPLICATIONS

- Other applications of DNA sequencing
  - Methylation detection
  - Pharmacogenomics
  - Newborn screening
  - So much more!



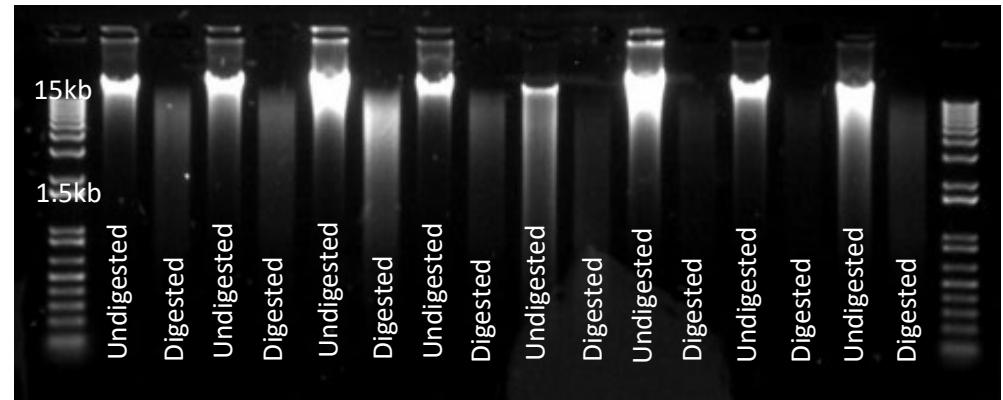
# APPLICATIONS - RNA

- Whole transcriptome
- RNA Exome
- Targeted RNA
- Metatranscriptome



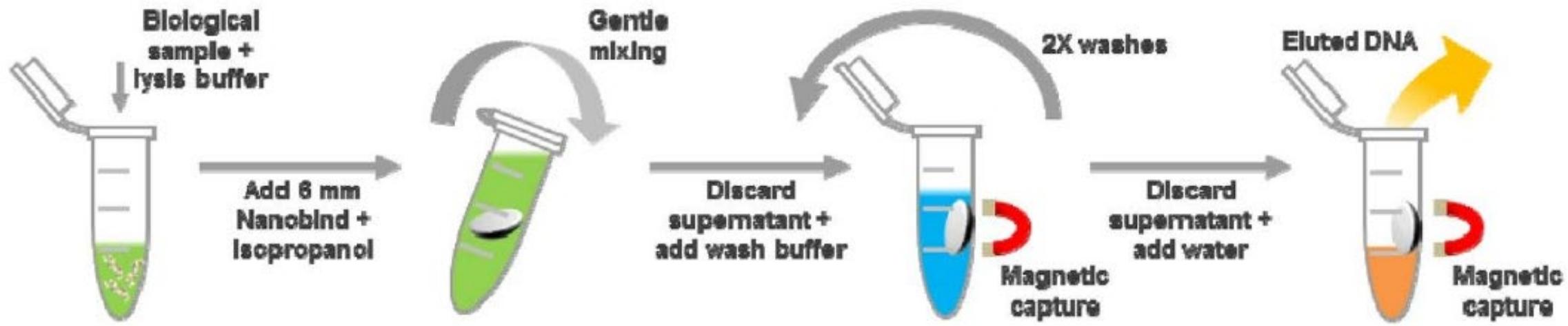
# LIBRARY PREP

- Things to consider for library preparation
  - Quality of DNA
  - Quantity of DNA
  - What kit will be used?



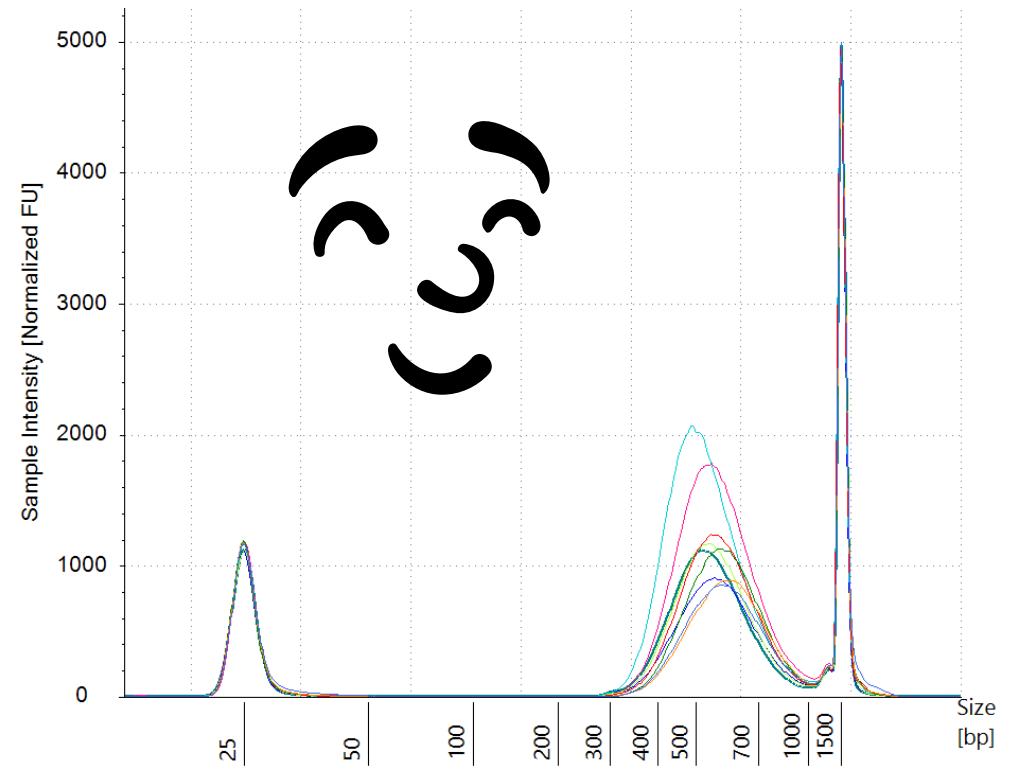
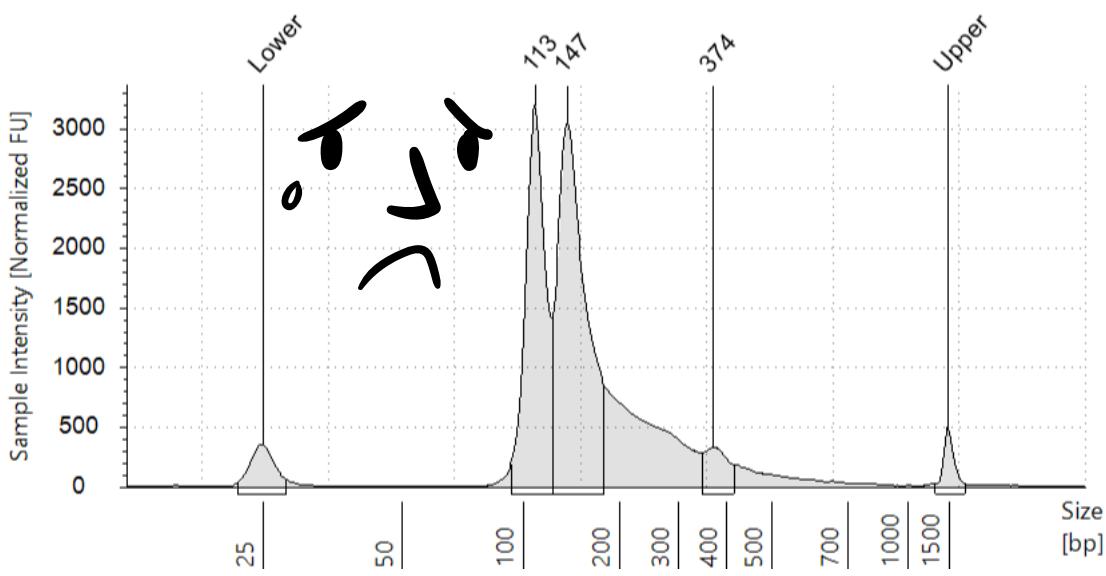
# EXTRACTION

# PacBio



# LIBRARY QC

- Things to consider for library quality control
  - Quality of DNA
  - Quantity of DNA
  - What are the downstream goals?



# PRIMARY ANALYSIS

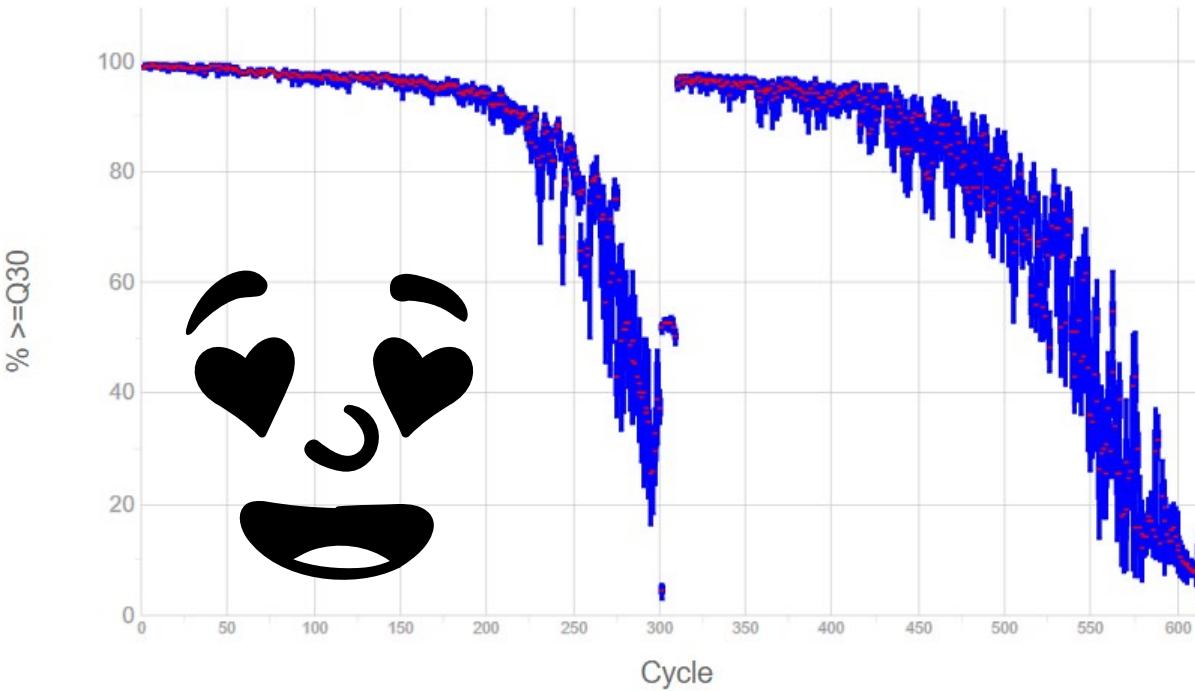
Output Folder (Lab		Sample Name	Index	Reads	% > Q30
Lane	Name)				
1	Zeller	Sample 1	TCTTCAGAGA-CTACGTGACG	99,088	95
1	Zeller	Sample 2	GTCAAGCTCG-CCGACAGCTT	101,727	97
1	Zeller	Sample 3	TAGAGTTGGA-GACGATATGA	134,816	97
1	Zeller	Sample 4	CTGATGATCT-TTGTACTCCA	125,639	97
1	Zeller	Sample 5	ACTAGGTGTT-GTGCACATAA	109,037	97
1	Zeller	Sample 6	CTGTTAGCGG-AGGACAAGTA	108,294	96
1	Zeller	Sample 7	ATCGCACCAA-CCGATTGAG	105,061	96
1	Zeller	Sample 8	CTTACTTGGT-GTAGGAACTT	117,641	97
1	Zeller	Sample 9	TCTCGCCTAG-ATGACCTTGA	109,566	97
1	Zeller	Sample 10	CCTTAATGCG-TACACTACGA	107,212	97

# PRIMARY ANALYSIS

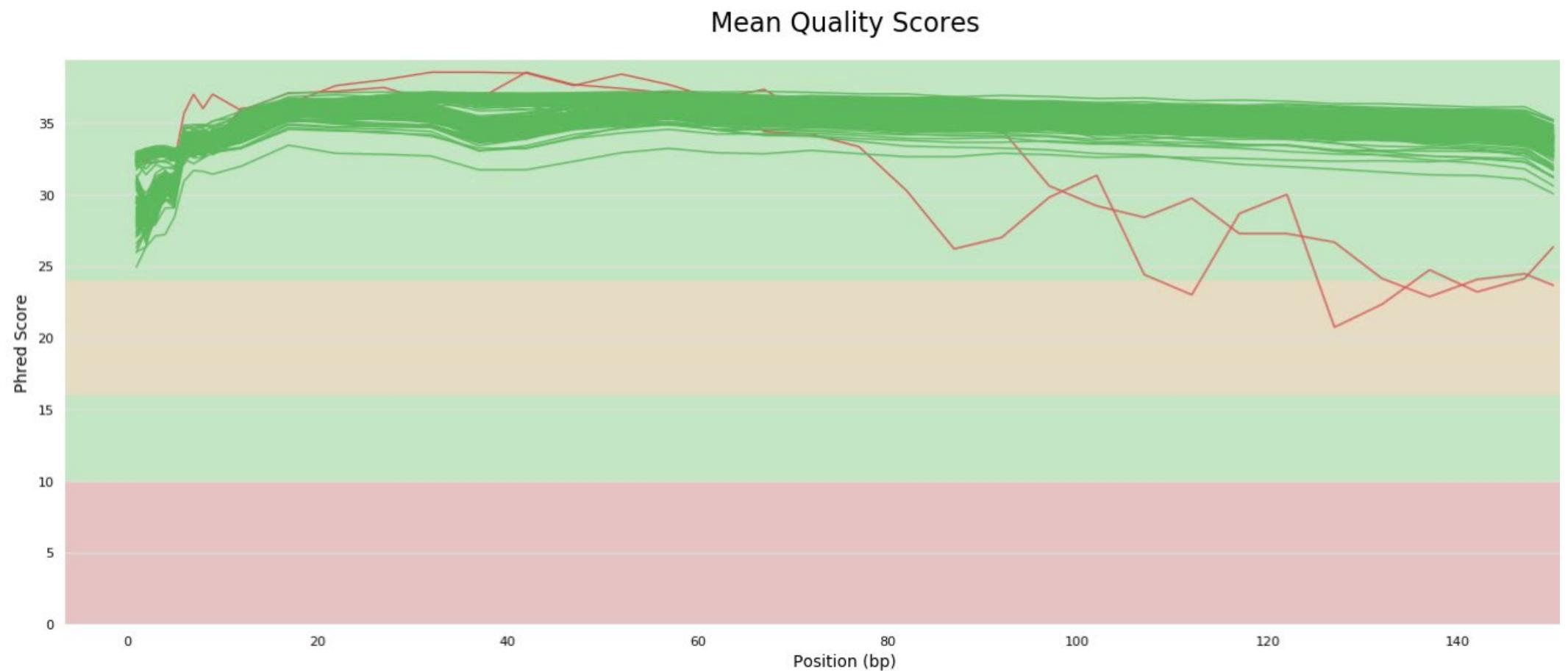
- Q30 score is averaged across whole run not individual cycle

Phred Quality Score	Probability of Incorrect Based Call	Base Call Accuracy	Q-score
---------------------	-------------------------------------	--------------------	---------

10	1 in 10	90%	Q10
20	1 in 100	99%	Q20
30	1 in 1000	99.9%	Q30
40	1 in 10000	99.99%	Q40

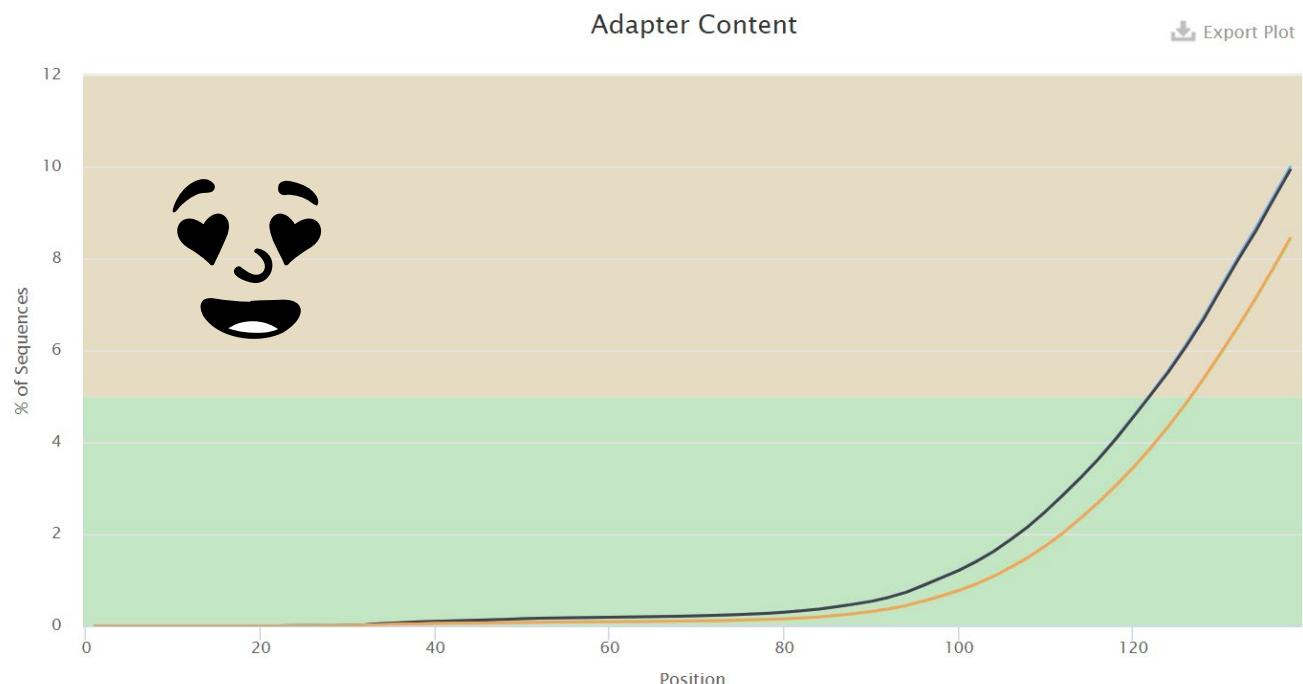
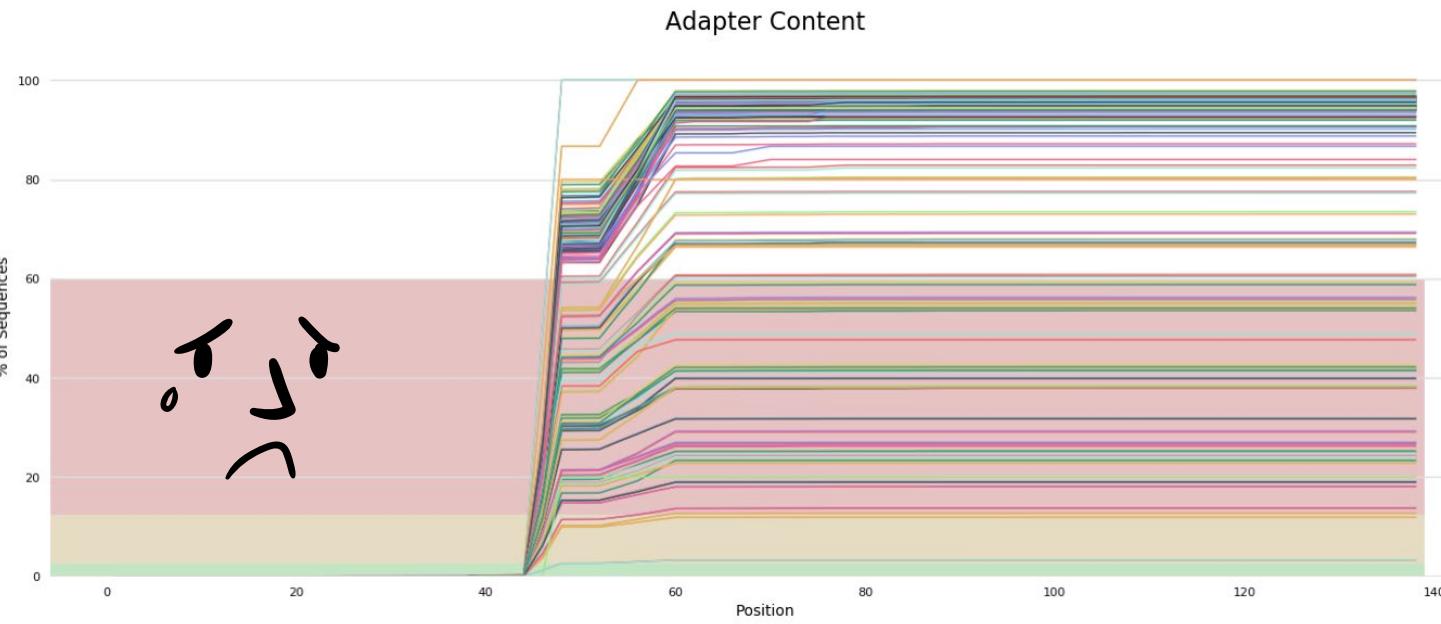


# PRIMARY ANALYSIS

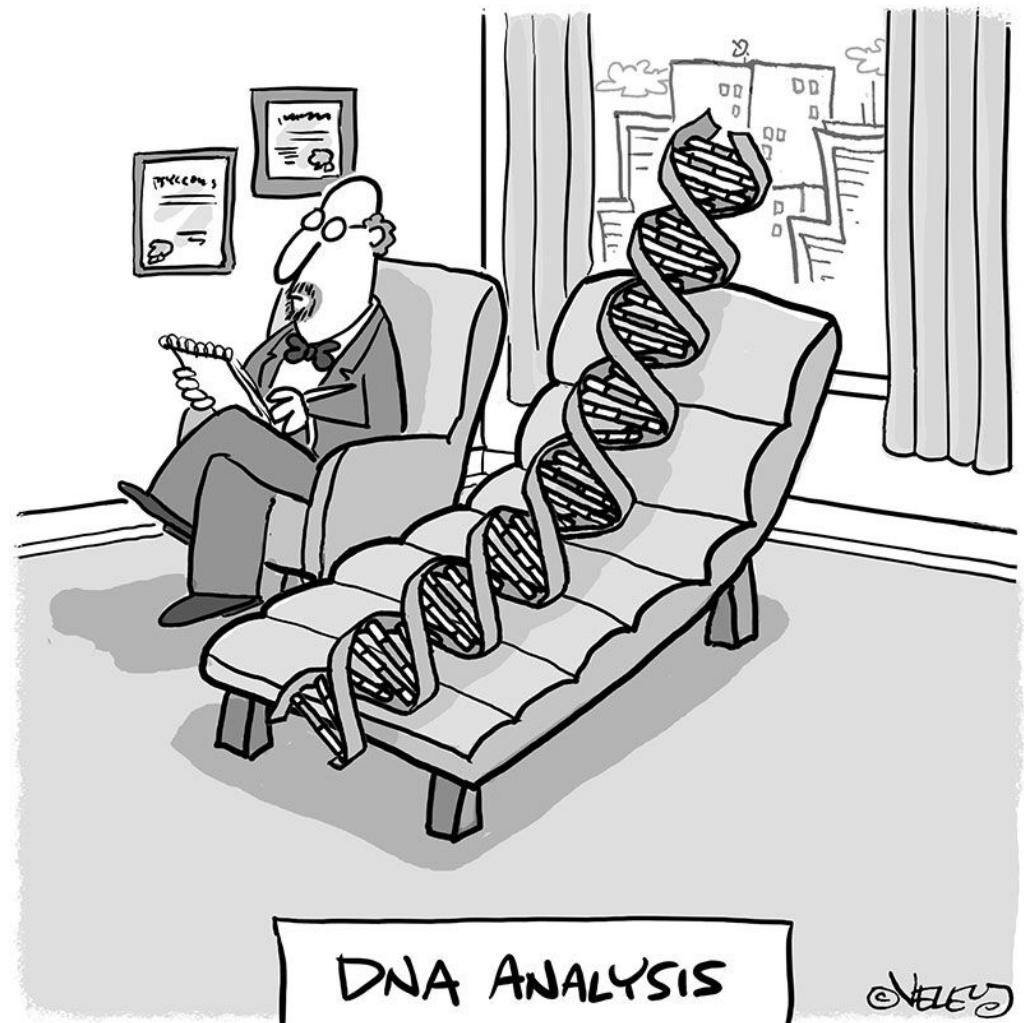


# PRIMARY ANALYSIS

- Ideally no adapter would be present in data
- Short inserts or poor size selection can result in adapter presence



thank  
you



CARTOONSTOCK.COM

CARTOONCOLLECTIONS.COM