Agilent
Trusted Answers

The Next Generation Sequencing: Technologies and Applications

Stefano Gabriele Ph.D.

Market Specialist – Healthcare Diagnostics

DGG-GSD Division

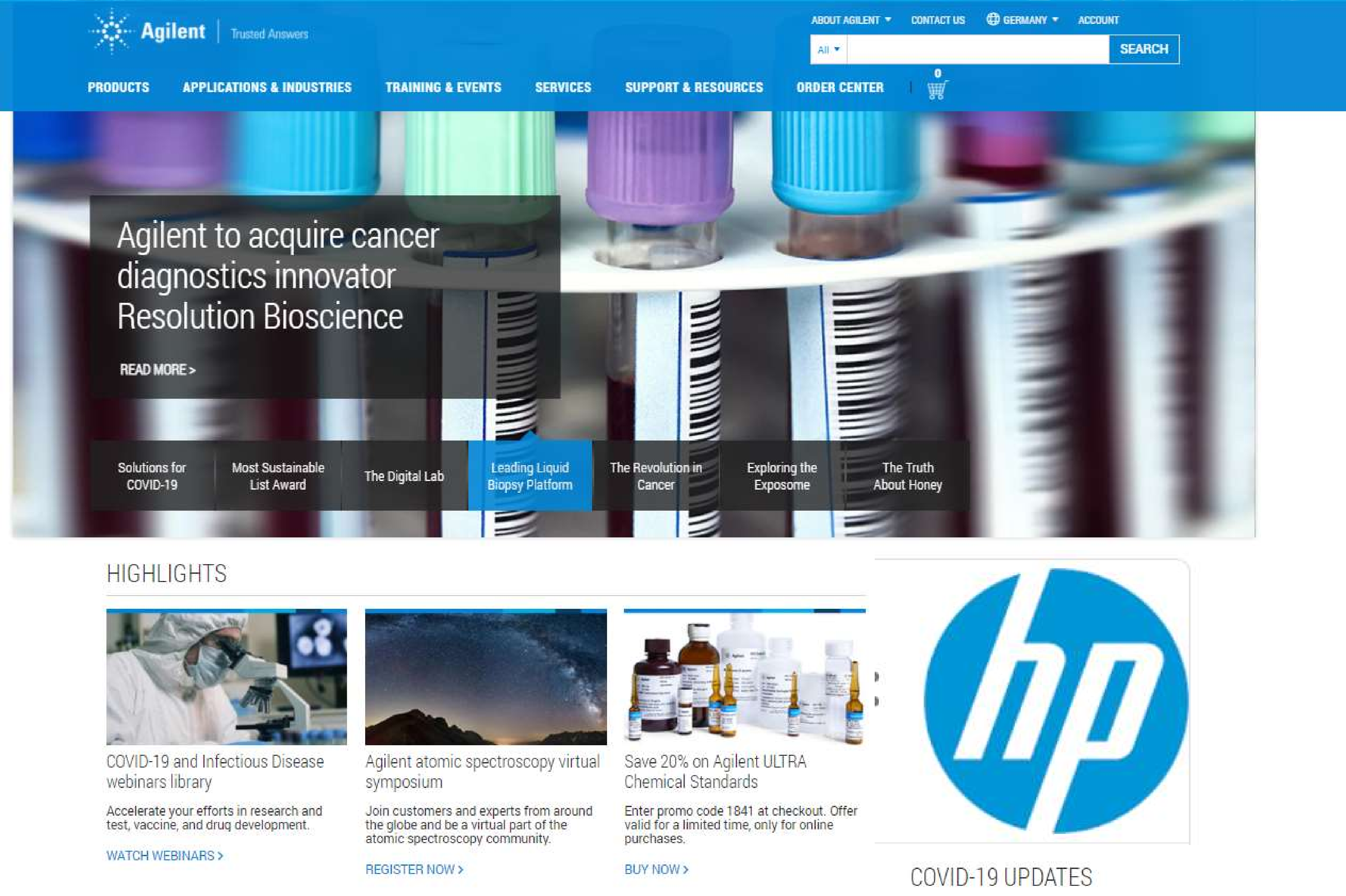
Agilent Tecnhologies

Roma, 17 Aprile 2024

stefano.gabriele@agilent.com

April 22, 2024

AGILENT TECHNOLOGIES SpA



Agilent | Trusted Answers

ABOUT AGILENT ▾ CONTACT US GERMANY ▾ ACCOUNT

All ▾ SEARCH

PRODUCTS APPLICATIONS & INDUSTRIES TRAINING & EVENTS SERVICES SUPPORT & RESOURCES ORDER CENTER | 0

Agilent to acquire cancer diagnostics innovator Resolution Bioscience

READ MORE >

Solutions for COVID-19 | Most Sustainable List Award | The Digital Lab | Leading Liquid Biopsy Platform | The Revolution in Cancer | Exploring the Exposome | The Truth About Honey

HIGHLIGHTS

COVID-19 and Infectious Disease webinars library

Accelerate your efforts in research and test, vaccine, and drug development.

WATCH WEBINARS >

Agilent atomic spectroscopy virtual symposium

Join customers and experts from around the globe and be a virtual part of the atomic spectroscopy community.

REGISTER NOW >

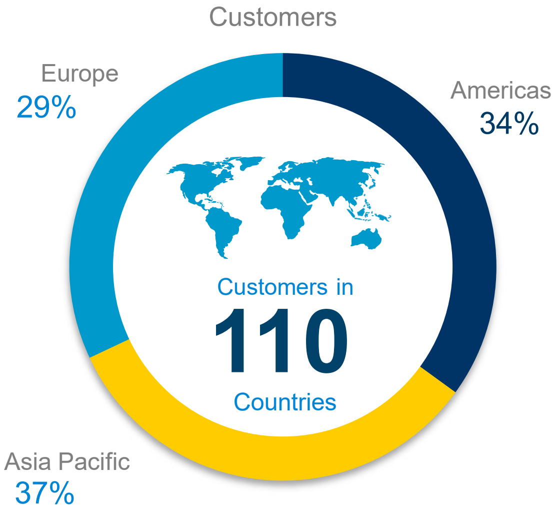
Save 20% on Agilent ULTRA Chemical Standards

Enter promo code 1841 at checkout. Offer valid for a limited time, only for online purchases.

BUY NOW >

COVID-19 UPDATES

Agilent Technologies S.p.A
Who We Are – By the Numbers



265,000
Customer Labs

>16,000
Employees

3,600
Service Engineers

1,320
Remote Support

900
Applications Support

7%
R&D Spend vs. Revenue

1,000,000
Customer Interactions Per Year

#1

Sustainability
DOW JONES SUSTAINABILITY RANKINGS 2019 –
Science Tools and Services
BARRON's 100 MOST SUSTAINABLE 2020

Top 100

Best Places to Work
China, US, Germany

Employer of Women in China
Great Places to work (GPW) Institute

Employer of Women
FORBES 2019

Topics for Today's Presentation

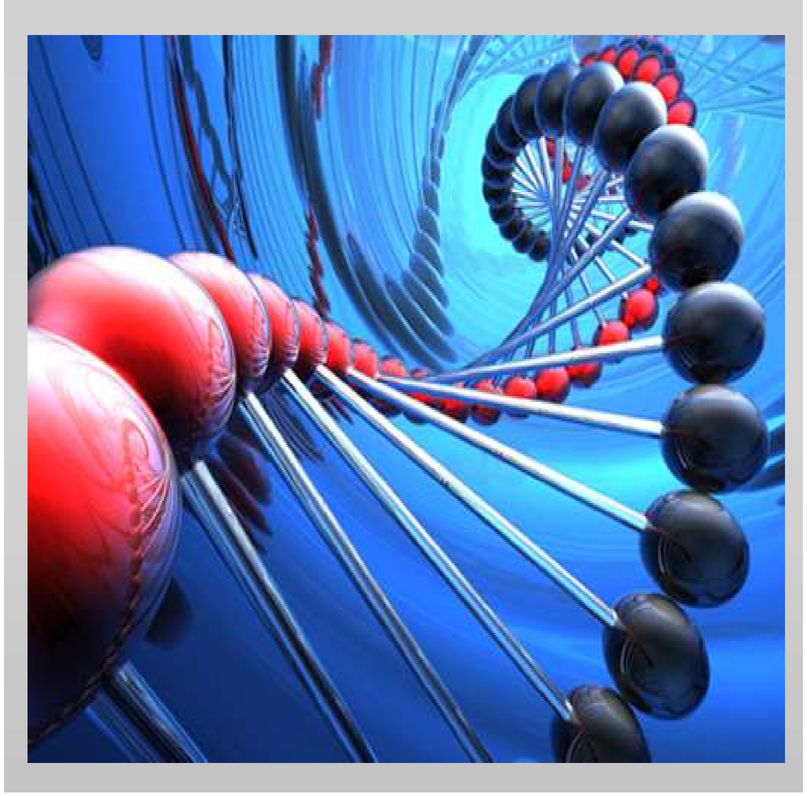
1 What is Next-Gen Sequencing?

2 Sequencers

3 The NGS Library Prep Workflow

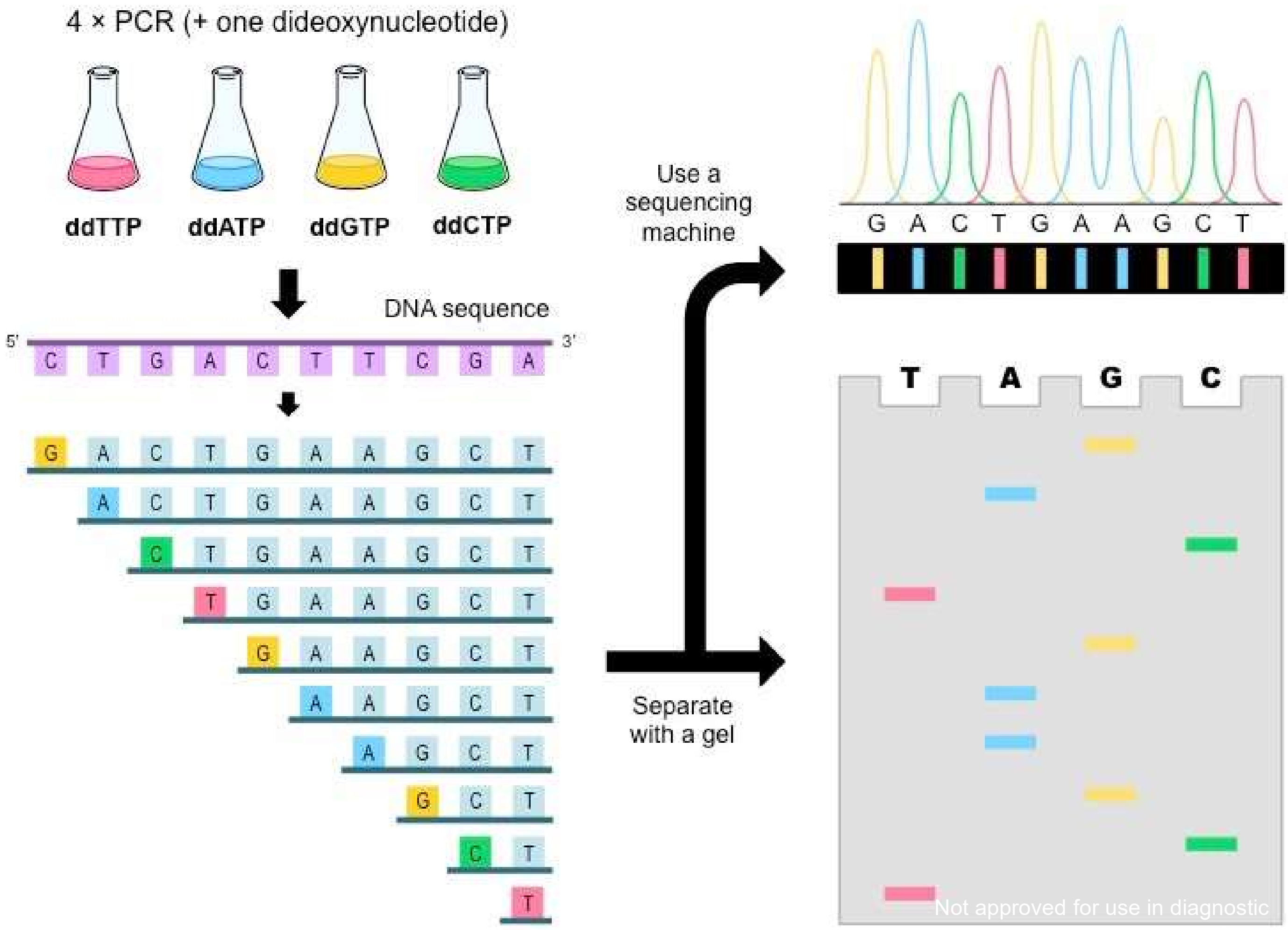
4 QC control

5 Analysis



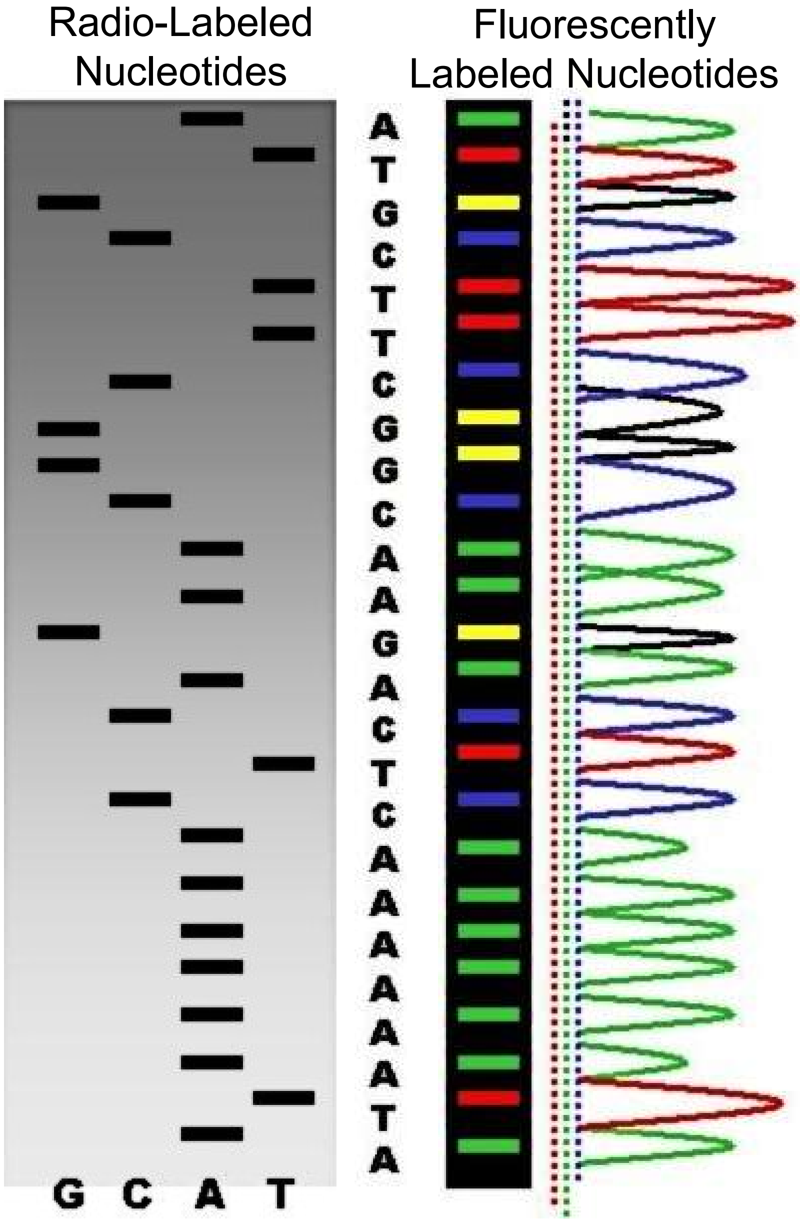
What is Next-Gen Sequencing? A Brief History

- Frederick Sanger (Sanger Sequencing)
  - “First Generation” (circa 1977)
    - Radiolabeled Nucleotides
    - Sequencing Gels



Not approved for use in diagnostic

What is Next-Gen Sequencing? A Brief History



- Frederick Sanger (Sanger Sequencing)
  - "First Generation" (circa 1977)
    - Radiolabeled Nucleotides
    - Sequencing Gels
- Automated Capillary Electrophoresis
  - "Second Generation"
    - ABI 370 generate 500 Kilobases/day
      - Thousands of bases (Kb)
    - ABI 3730 generate 2.8 Megabases/day
      - Millions of bases (Mb)
    - Fluorescence based vs radiolabeling
    - Helped drive the Human Genome Project

The Cornerstone Driving Next-Gen Sequencing Technology

Research: 10 years

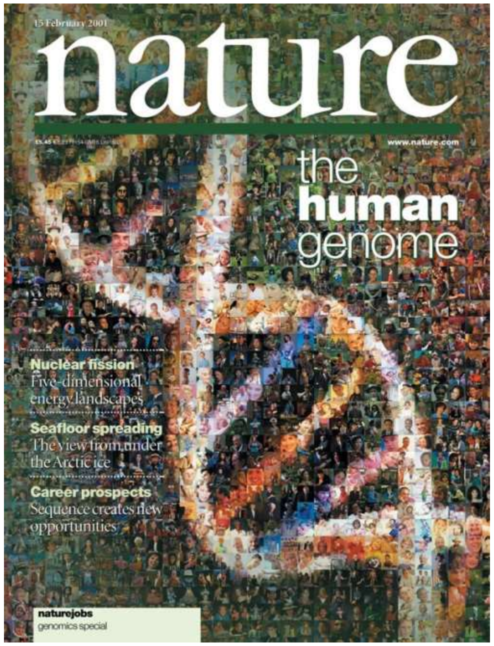
Cost: ~ $3 Billion

Human Genome

3.2Gb

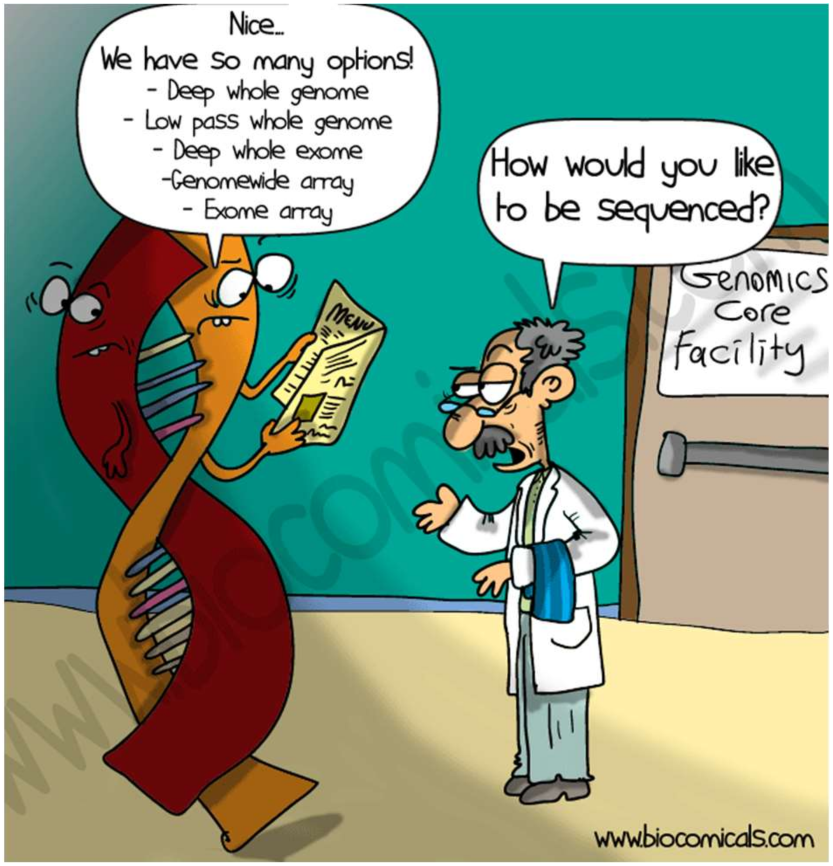
January 15th 2001

Completing The Human Genome…



…Priceless

What is Next-Gen Sequencing?
A Brief History



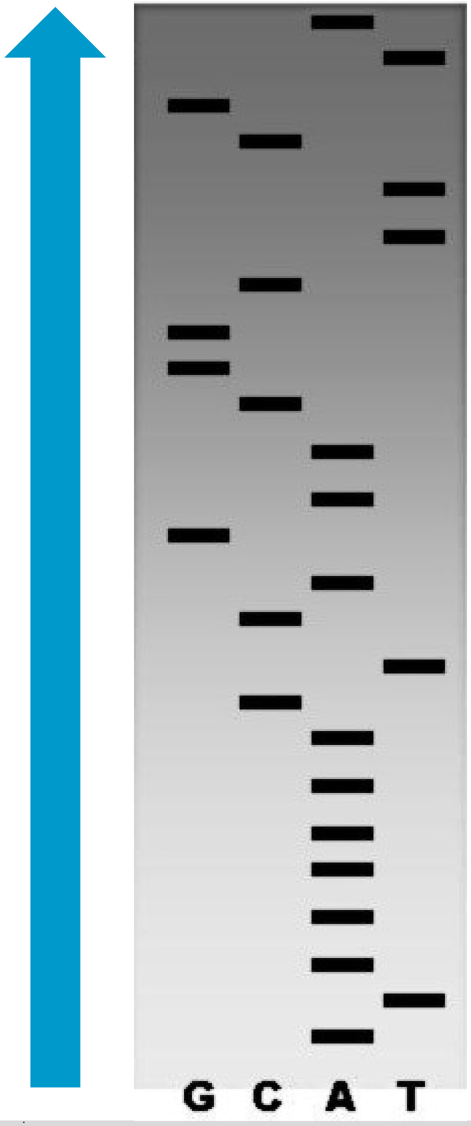
- Massively Parallel Sequencing
  - “Next-Generation Sequencing” (NGS)
    - Does not use Sanger method
    - Different Platforms = Different Chemistries
    - Very High throughput instruments
      - >100 gigabases of DNA sequence/day
- Desktop Sized Sequencing Instruments & Beyond!
  - “Next-Next Generation Sequencing”
    - Scaled down
    - Medium throughput
    - Individual Labs vs Core Facilities
- Some food for thought:
  - What will sequencing be like 5, 10, 15 years from now?

What is Next-Gen Sequencing:
Sanger Sequencing vs Next-Gen Sequencing

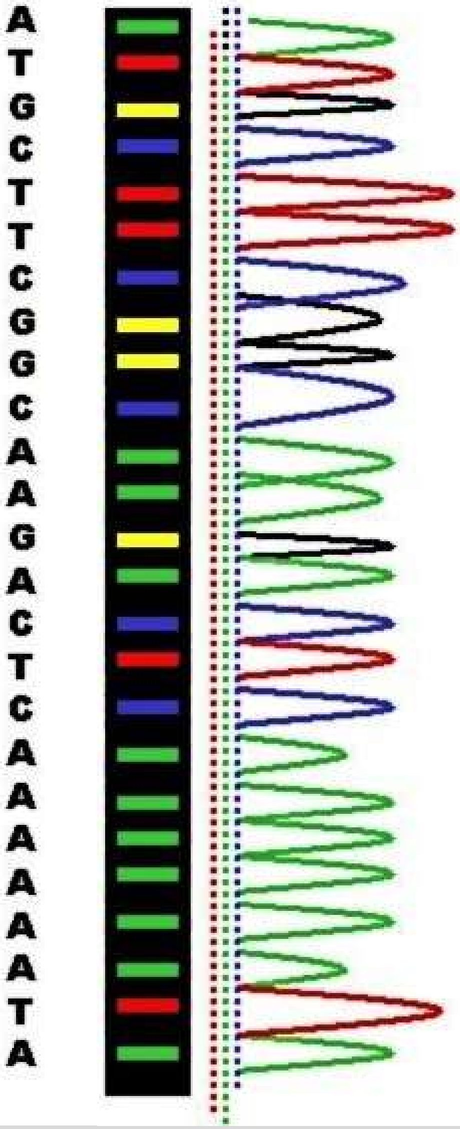
“Single” Read System/Run (i.e. 1 DNA Fragment)

“Multi” Read System/Run (i.e. Thousands of Fragments)

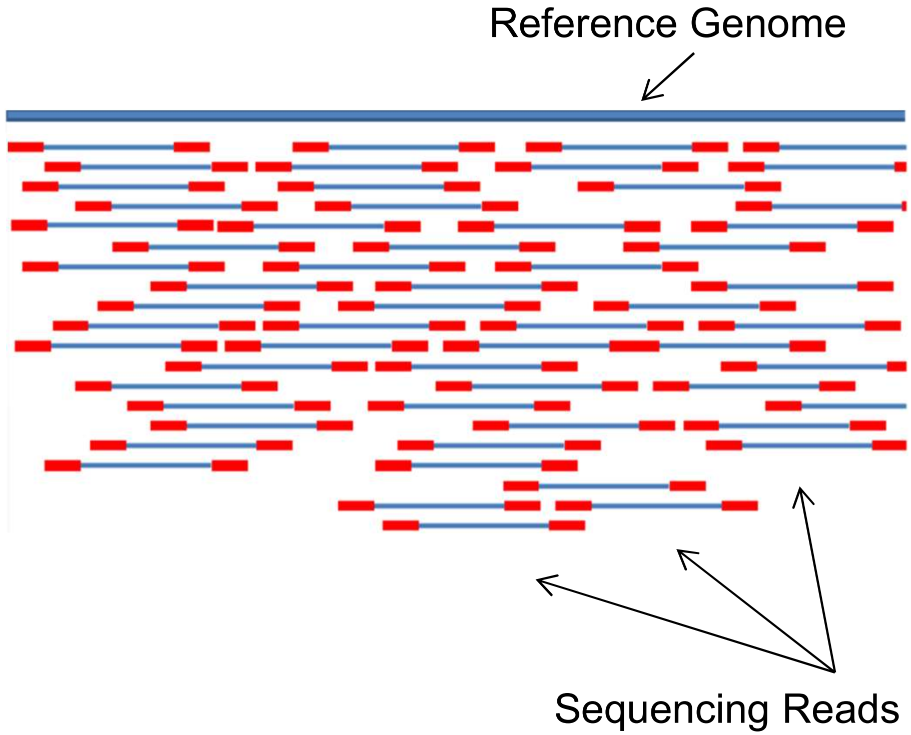
Radio-Labeled Nucleotides



Fluorescently Labeled Nucleotides

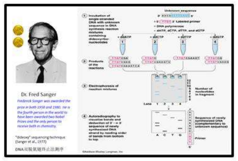


Fluorescently labeled nucleotides of many different DNA fragments being sequenced in parallel



Next-Gen Sequencing Technology Timeline…

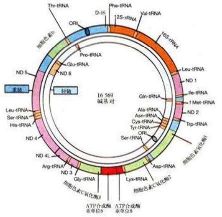
1977



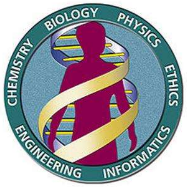
Sanger method

1981

Human mitochondrial genome sequence



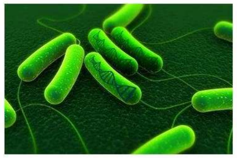
1990



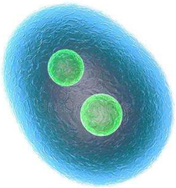
Human Genome Project

1995

Complete cell genome



1996



Complete eukaryotic genome

2001

Complete the Human Genome Project

2005

Second generation sequencer：454 GS20

2007

Second generation sequencer：Genetic Analyzer 2

2008

Second generation sequencer：Proton

2011

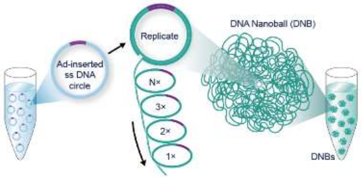
Third generation sequencer：PacBio RS

2014

Nanospace sequencing

2019

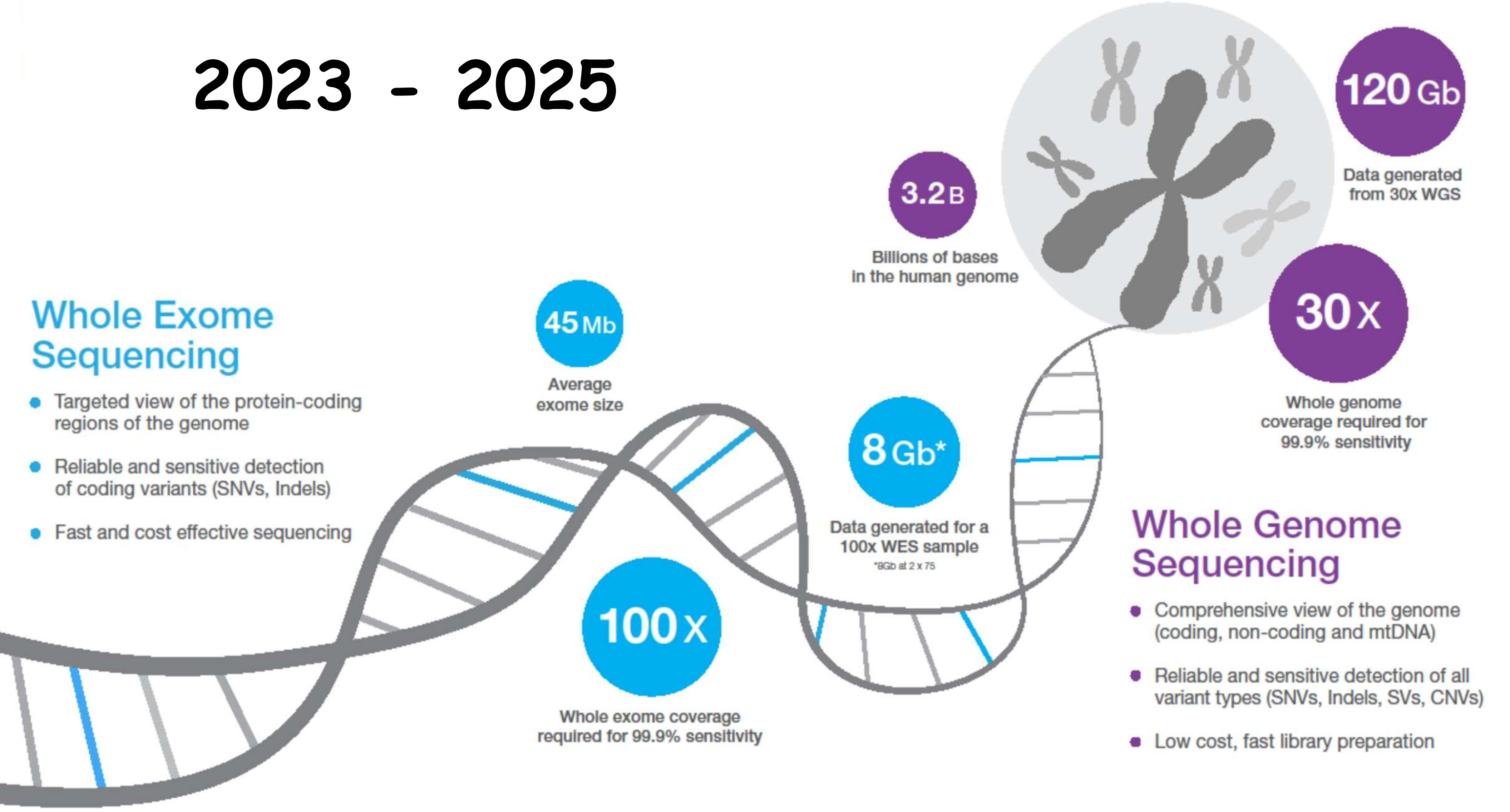
Third generation sequencer: MGI



?

Whole Genome Seq vs Whole Exome Seq

2023 - 2025



Whole Exome Sequencing

- Targeted view of the protein-coding regions of the genome
- Reliable and sensitive detection of coding variants (SNVs, Indels)
- Fast and cost effective sequencing

45 Mb Average exome size

100 x Whole exome coverage required for 99.9% sensitivity

8 Gb* Data generated for a 100x WES sample

*8Gb at 2 x 75

3.2 B Billions of bases in the human genome

120 Gb Data generated from 30x WGS

30 x Whole genome coverage required for 99.9% sensitivity

Whole Genome Sequencing

- Comprehensive view of the genome (coding, non-coding and mtDNA)
- Reliable and sensitive detection of all variant types (SNVs, Indels, SVs, CNVs)
- Low cost, fast library preparation

illumina®

WES = < 300 EUR
WGS = < 100 EUR

Topics for Today's Presentation

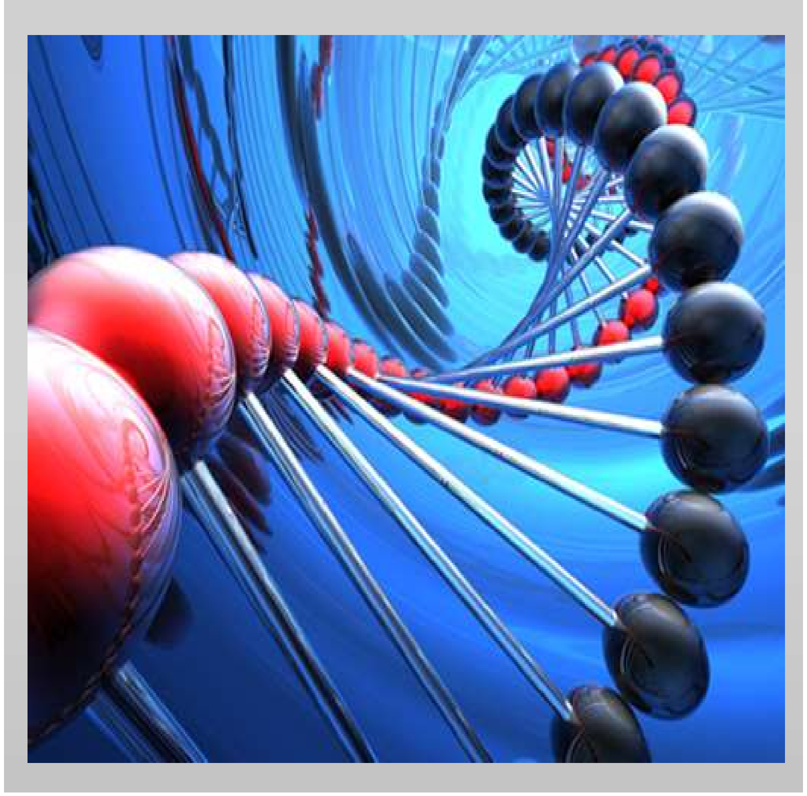
✓ 1 What is Next-Gen Sequencing?

● 2 Sequencers

3 Reviewing NGS Terminology

4 The NGS Library Prep Workflow

5 Analysis



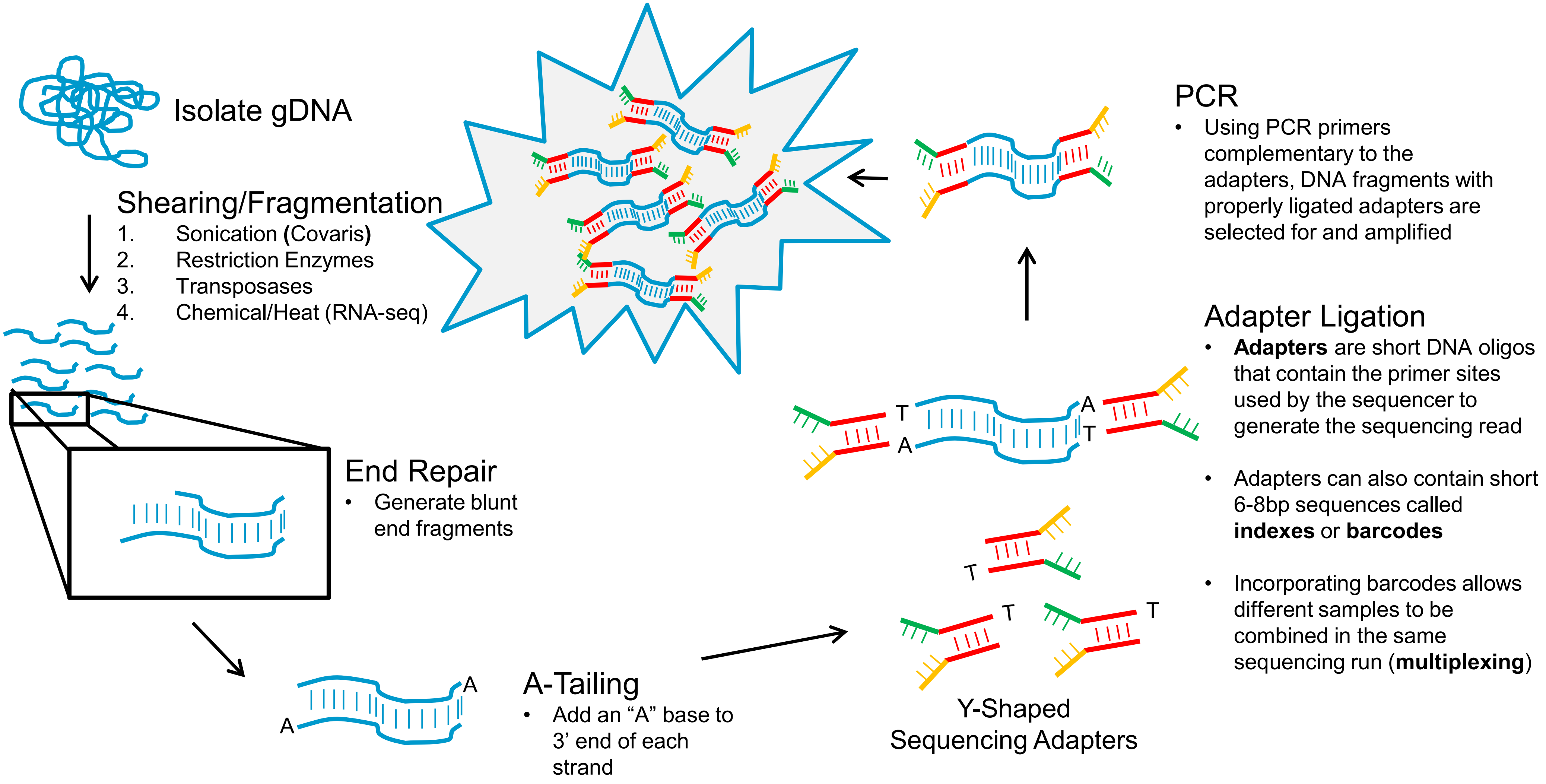
Current technologies and available platforms

- Genome Analyzers (Illumina)
- Ion Torrent (Thermo Fisher)
- Pacific Bioscience
- Oxford Nanopore sequencing

- MGI (BGI) Genome Analyzer
- AVITI Systems (Element Bioscience)
- G4 Sequencer (Singular Genomics)

Learning the NGS Workflow: Generating a Sequencing Library

Library - A collection of DNA or cDNA fragments prepared for sequencing by a performing a series of enzymatic steps. These steps are commonly referred to as the **Library Prep.**



Isolate gDNA

Shearing/Fragmentation

1. Sonication (Covaris)
2. Restriction Enzymes
3. Transposases
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

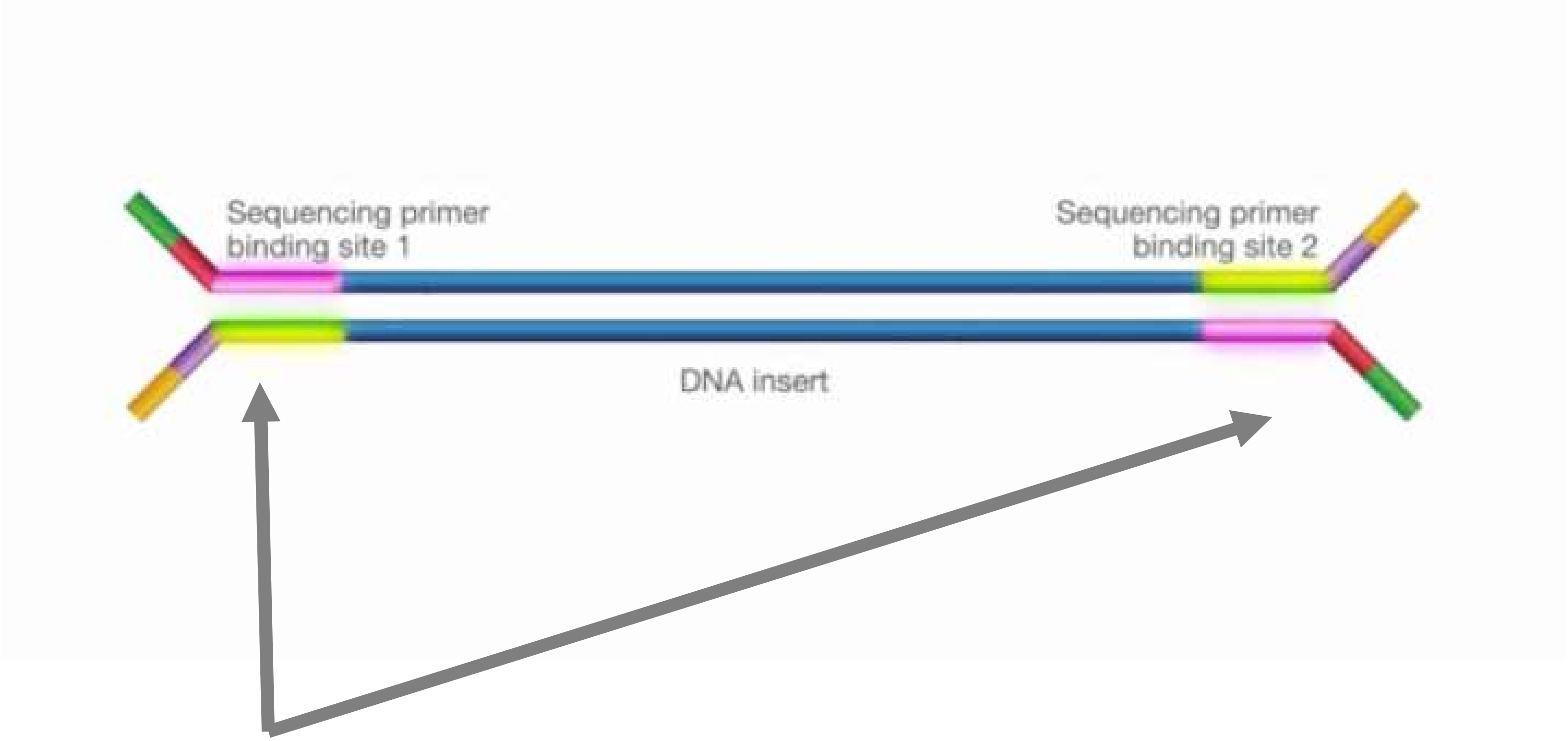
- **Adapters** are short DNA oligos that contain the primer sites used by the sequencer to generate the sequencing read
- Adapters can also contain short 6-8bp sequences called **indexes** or **barcodes**
- Incorporating barcodes allows different samples to be combined in the same sequencing run (**multiplexing**)

PCR

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

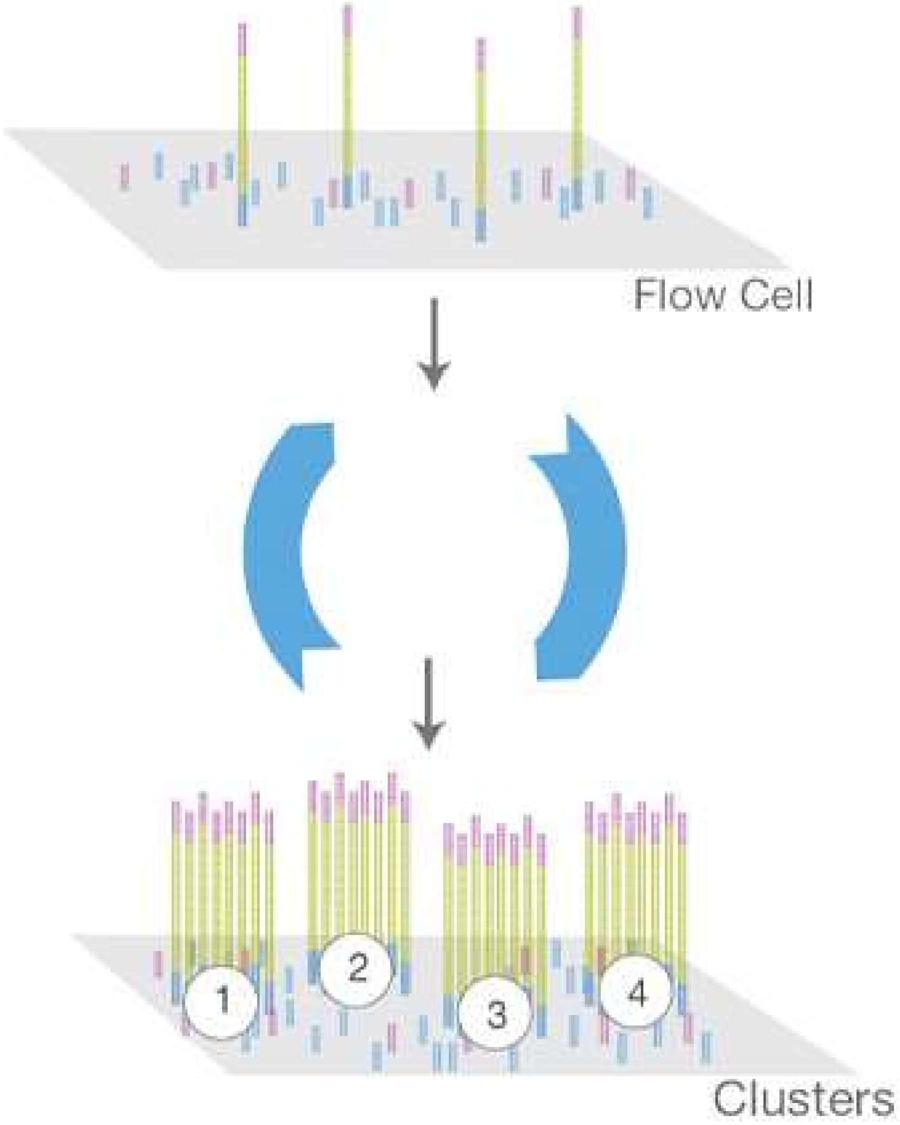
Learning the NGS Workflow: Generating a Sequencing Library

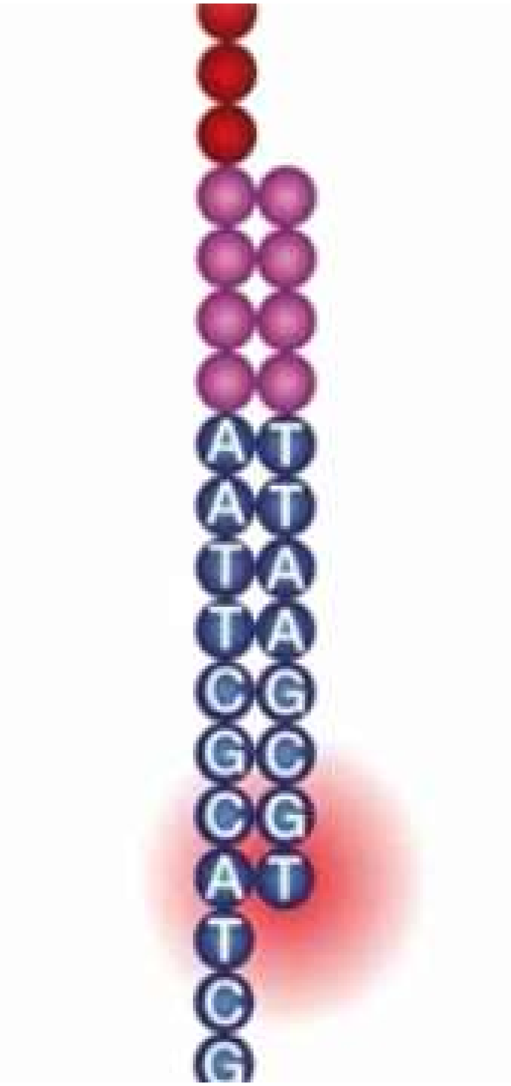
DNA fragments need to be “modified” to meet NGS platforms


Sequencing primer binding site 1
Sequencing primer binding site 2
DNA insert

ADAPTERS

Learning the NGS Workflow: generate a cluster of fragments





Learning the NGS Workflow
Understanding Reads: Types of Reads

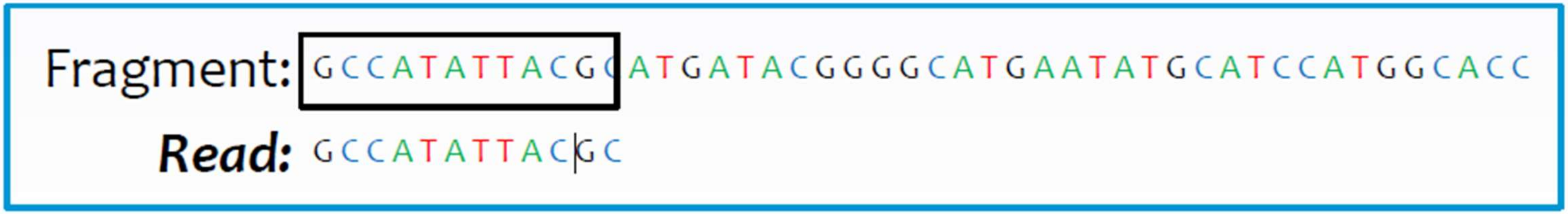


Figure Adapted from Ambry Genetics

- **Single-End Reads**: Provide sequence from <u>one</u> end of a DNA insert
- **Paired-End Reads**: Provide sequence from <u>both</u> ends of a DNA insert.
  - Provides improved alignment of sequencing data
  - Better detection of chromosomal rearrangements: insertions/deletions/translocations and fusions.

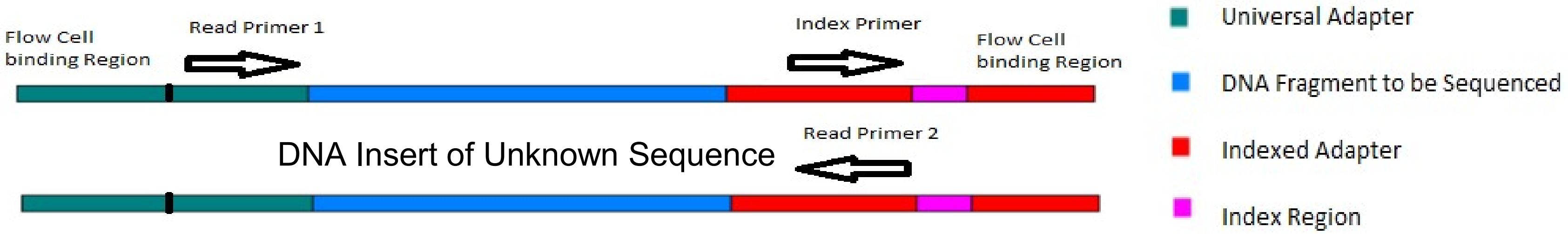


Figure Adapted from: tucf-genomics.tufts.edu

Learning the NGS Workflow
Understanding Reads: Lengths of Reads

Read lengths vary across sequencing platforms:

- Short reads – Illumina, Ion Torrent/Proton,

  <100bp (ex. 1 x 36bp, 2 x 50bp, 1 x 75bp)

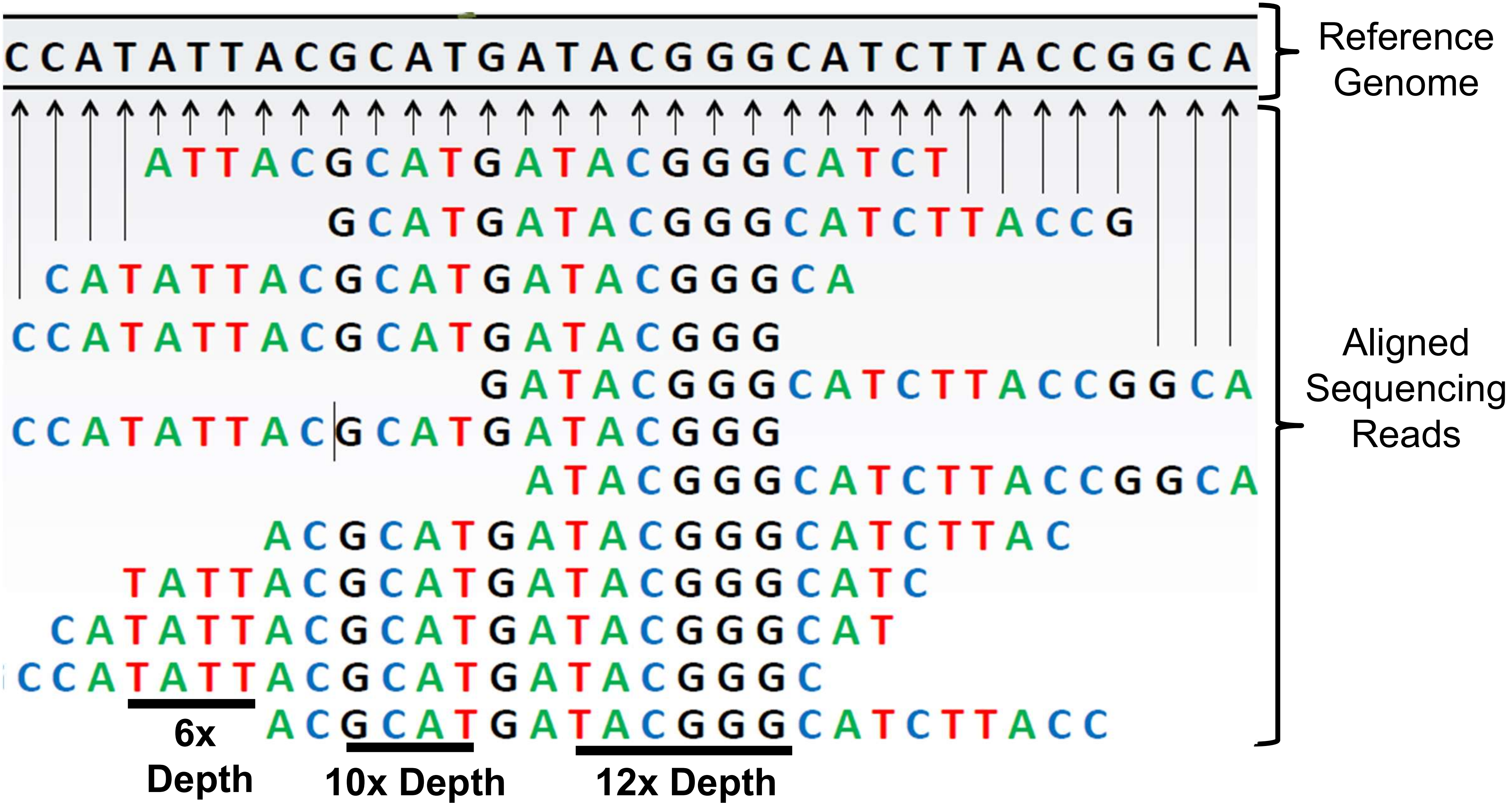
- Medium reads – Illumina, Ion Torrent/Proton, Qiagen, BGI

  >100bp but <1000bp (ex. 2 x 100bp, 2 x 150bp, 1 x 400bp, 1x 600bp)

- Long Reads – Pacific Biosciences (PacBio), Oxford Nanopore

  >1000bp (ex. 1x1000bp, >20,000bp, >300,000bp)

Learning the NGS Workflow
Understanding Reads: Depths of Reads



Reference Genome

CCATATTACGCATGATACGGGCATCTTACCGGCA

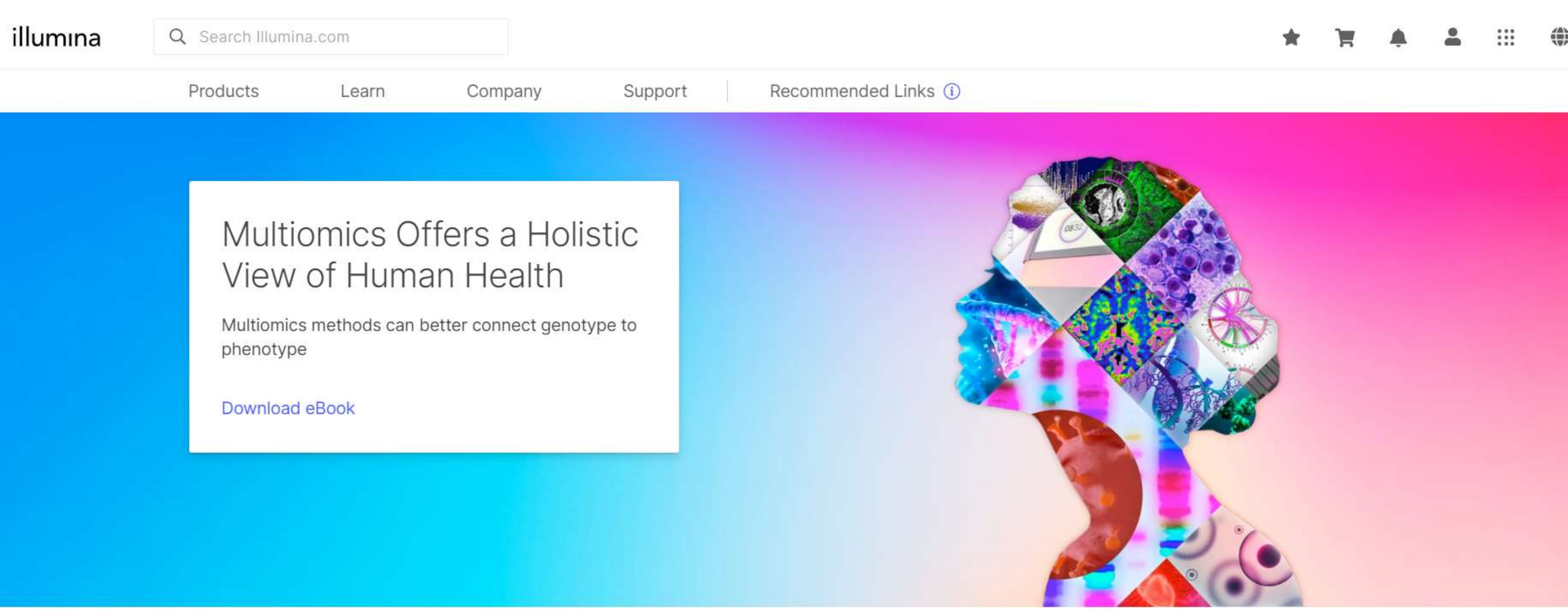
Aligned Sequencing Reads

ATTACGCATGATACGGGCATCT
GCATGATACGGGCATCTTACCG
CATATTACGCATGATACGGGCA
CCATATTACGCATGATACGGG
GATACGGGCATCTTACCGGCA
CCATATTACGCATGATACGGG
ATACGGGCATCTTACCGGCA
ACGCATGATACGGGCATCTTAC
TATTACGCATGATACGGGCATC
CATATTACGCATGATACGGGCAT
CCATATTACGCATGATACGGGC
ACGCATGATACGGGCATCTTACC

Figure Adapted from Ambry Genetics

Next-generation sequencing

1. Illumina Benchtop Sequencer



Multiomics Offers a Holistic View of Human Health

Multiomics methods can better connect genotype to phenotype

Download eBook

Next-generation sequencing

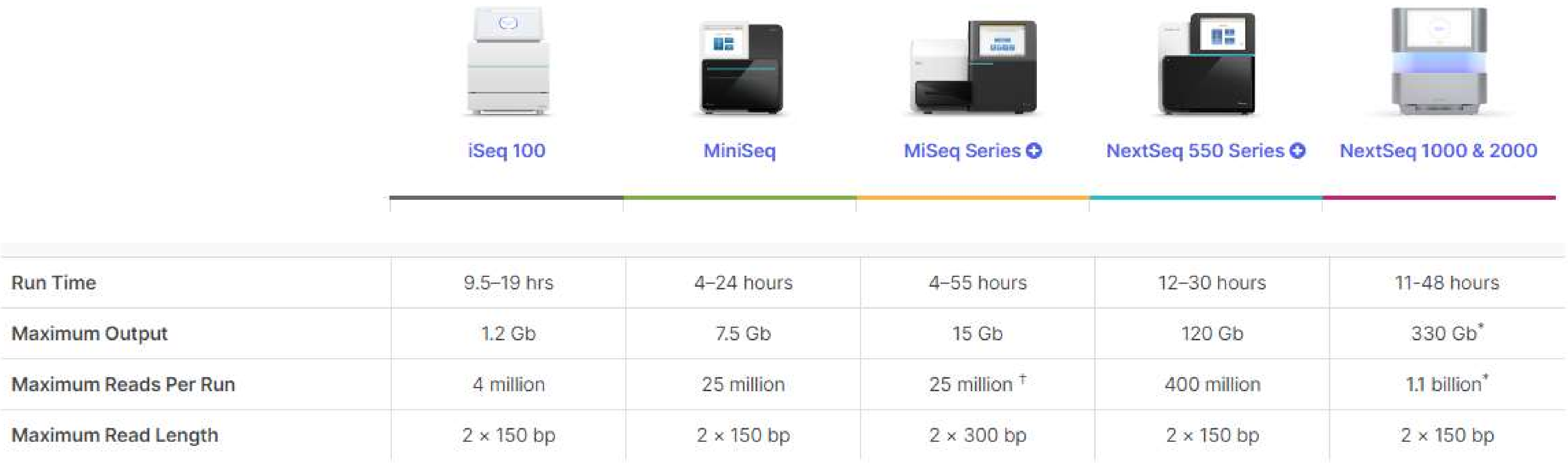
1. Illumina Benchtop Sequencer

illumina®

	MiniSeq	MiSeq	NextSeq 500	HiSeq 4000	NovaSeq
Run Time	24 hours	56 hours	29 hours	3.5 days	40 hours
Read length (pb)	2x 150	2x 300	2x 150	2x 150	2x 150
Read number	$50\ 10^6$	$50\ 10^6$	$800\ 10^6$	$5\ 10^9$	$3.3\ 10^9$
Ouput	7.5 Gb	15 Gb	120 Gb	1,500 Gb	1,000 Gb
Throughput	7 Gb/day	7 Gb/day	100 Gb/day	430 Gb/day	500 Gb/day

Next-generation sequencing

1. Illumina Benchtop Sequencers

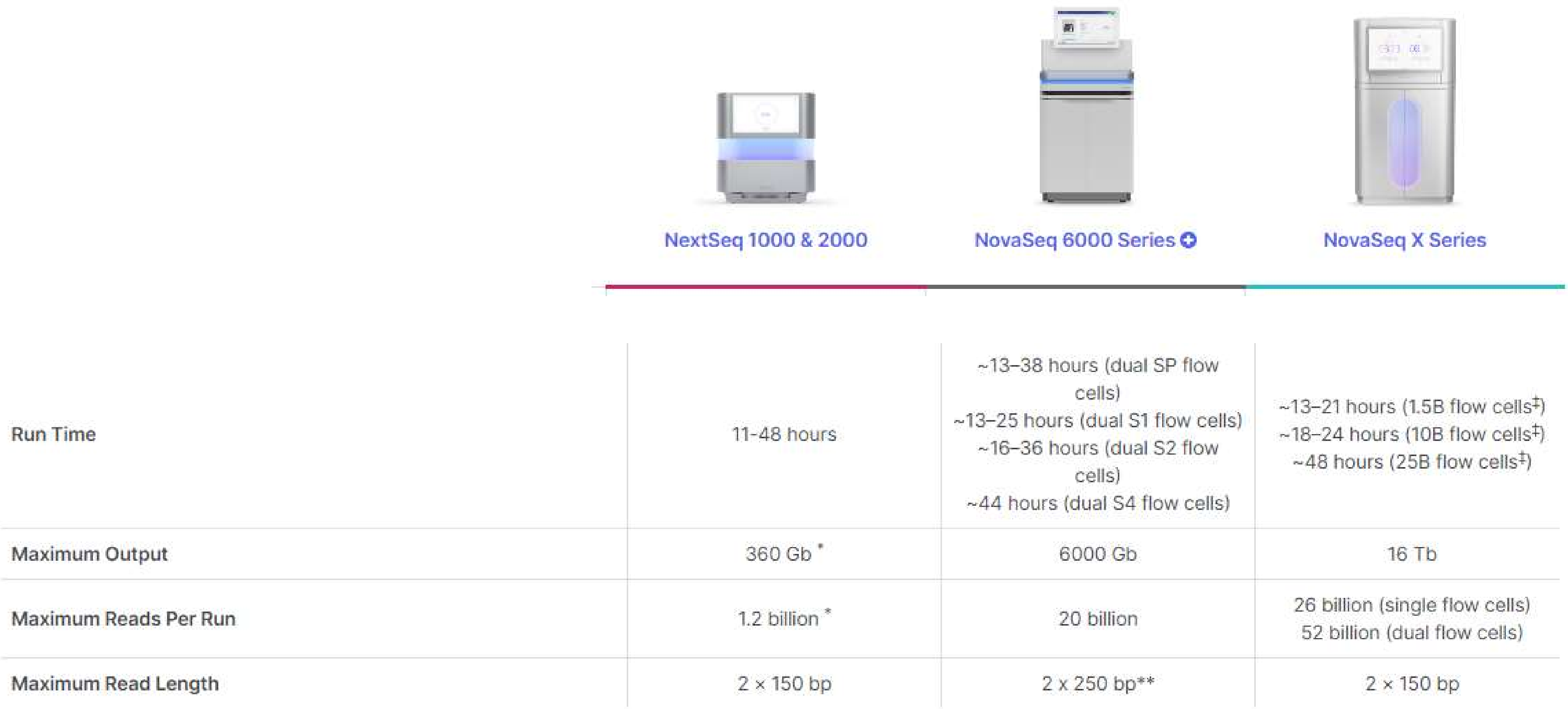


	iSeq 100	MiniSeq	MiSeq Series ⊕	NextSeq 550 Series ⊕	NextSeq 1000 & 2000
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb*
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.1 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

Next-generation sequencing

1. Illumina Production-scale Sequencers

illumina®



	NextSeq 1000 & 2000	NovaSeq 6000 Series ⊕	NovaSeq X Series
Run Time	11-48 hours	~13–38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)	~13–21 hours (1.5B flow cells‡) ~18–24 hours (10B flow cells‡) ~48 hours (25B flow cells‡)
Maximum Output	360 Gb *	6000 Gb	16 Tb
Maximum Reads Per Run	1.2 billion *	20 billion	26 billion (single flow cells) 52 billion (dual flow cells)
Maximum Read Length	2 × 150 bp	2 x 250 bp**	2 × 150 bp

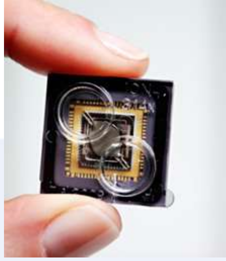
NovaSeq 6000: up to 24 WGS samples at 30x coverage
NovaSeq X Plus: more than 128 genomes per run

Ion Torrent S5 and S5 XL Systems

Thermo Fisher SCIENTIFIC

Ion S5 System

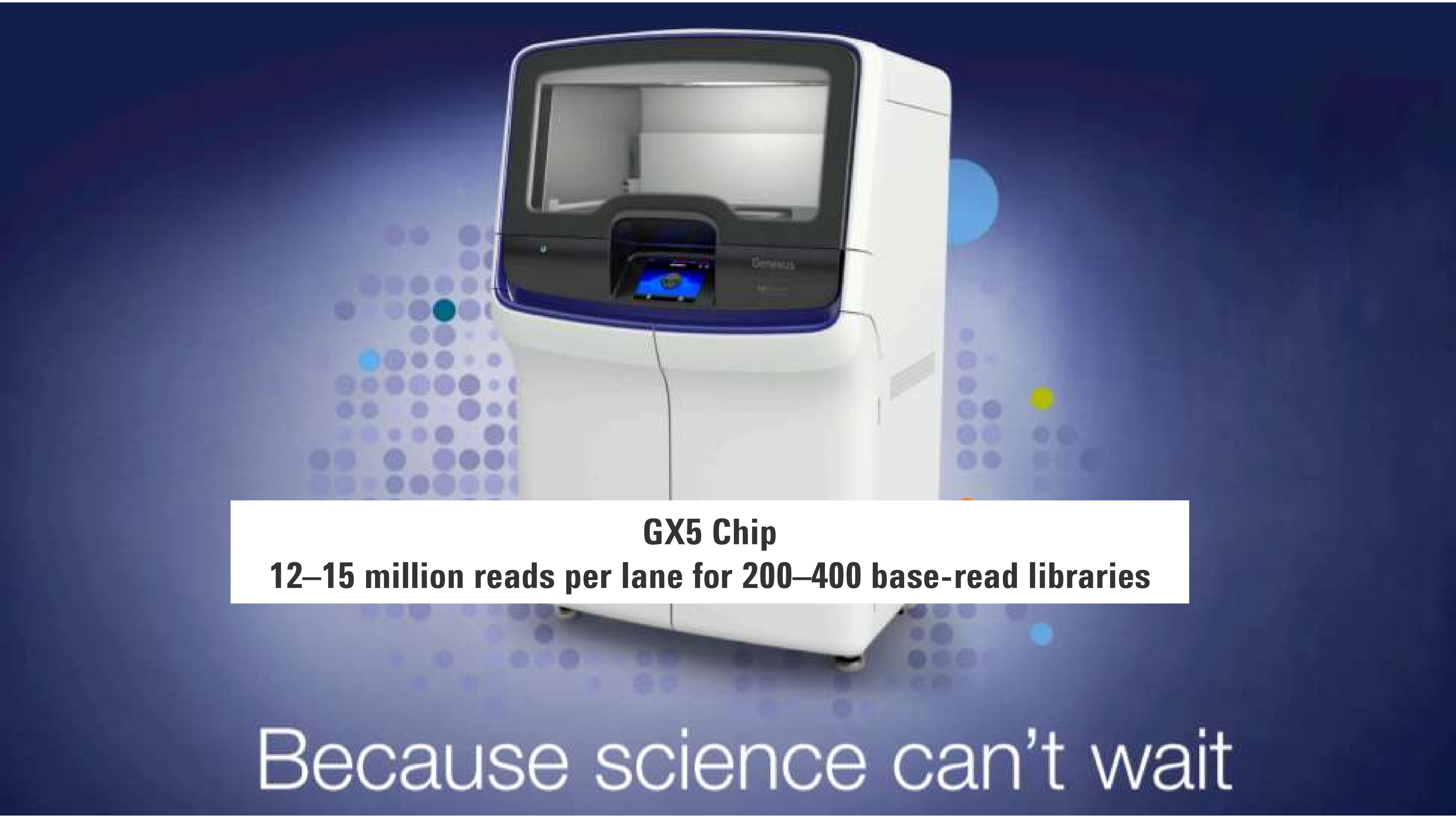
Ion S5 XL System



		Ion 520 Chip	Ion 530 Chip	Ion 540 Chip
Reads		3–5 million	15–20 million	60–80 million
Output*	**200 bp**	0.6–1 Gb	3–4 Gb	10–15 Gb
	400 bp	1.2–2 Gb	6–8 Gb	—
Run times	**200 bp**	2.5 hr	2.5 hr	2.5 hr
	400 bp	4 hr	4 hr	—
Analysis time†	**200 bp**	5 hr	8 hr	16.5 hr
	400 bp	8 hr	17.5 hr	—

Next-generation sequencing

2. Ion Torrent



GX5 Chip

12–15 million reads per lane for 200–400 base-read libraries

PacBio System Throughput

	PacBio RSII	Sequel
Capacity	1-16 SMRT cells / run	1-16 SMRT cells / run
Run Time	30min – 6h	30min – 6h
# of reads	~150.000 / SMRT cell	~1.000.000 / SMRT cell
Read length	Average 4.5 kb	Average 4.5 kb
Output	~675 Mb / SMRT cell 10 Gb / run	4.5 Gb / SMRT cell 72 Gb / run

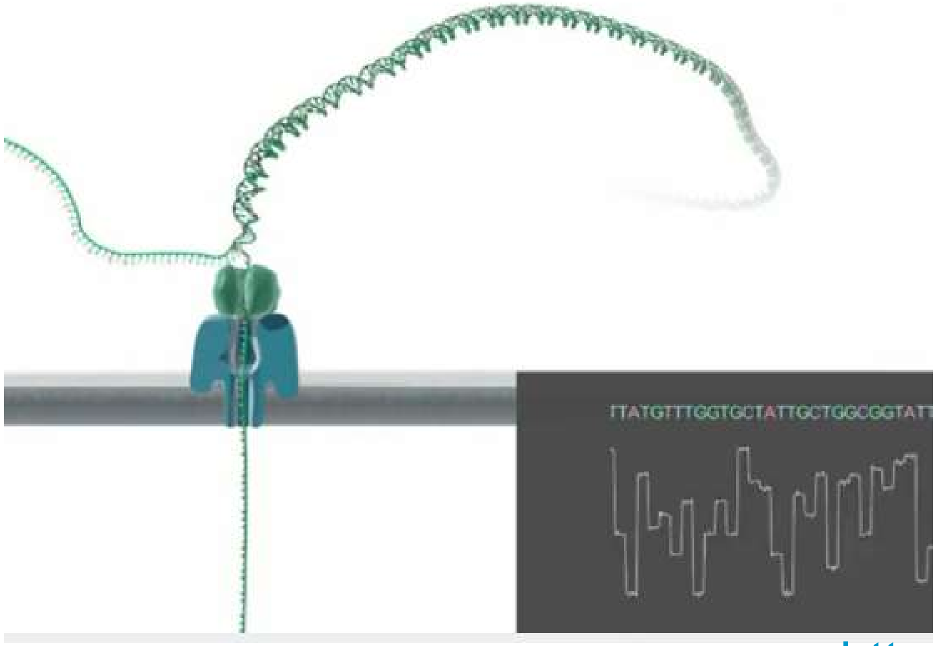
http://www.pacb.com/smrt-science/smrt-sequencing/

Next-generation sequencing

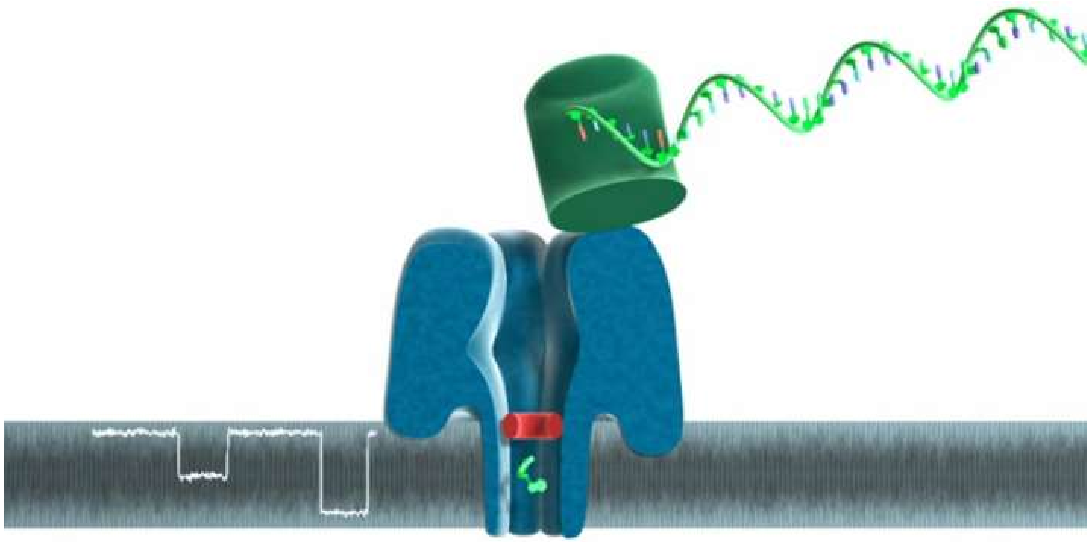
5. Oxford Nanopore Technology

Oxford NANOPORE Technologies

Strand sequencing



Exonuclease sequencing



https://www.youtube.com/watch?v=CE4dW64x3Ts

A single-use cartridge contains arrayed sensors and microfluidics, and inserted in a GridIon instrument

Next-generation sequencing

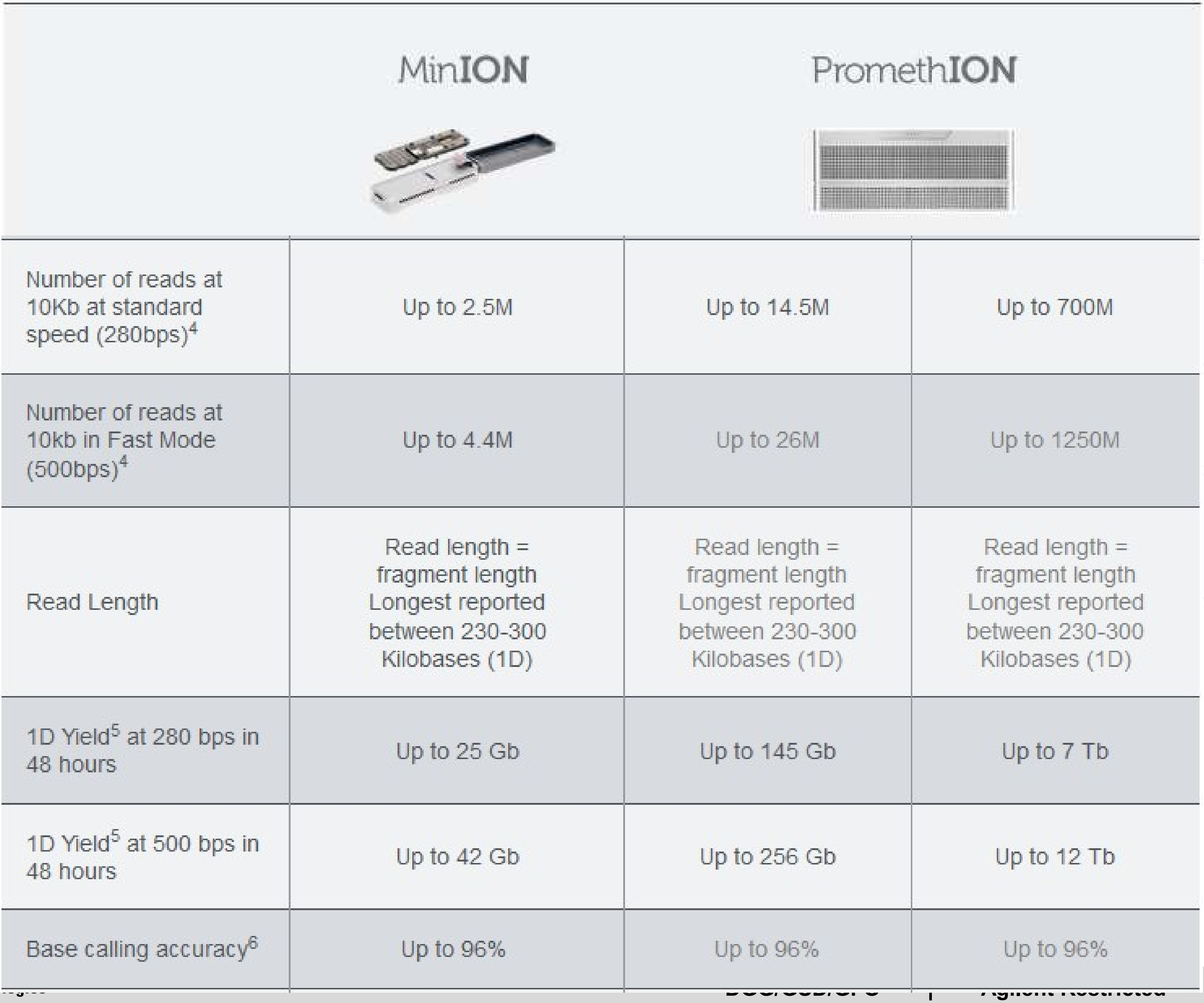
5. Oxford Nanopore Technology

GridIon : scalable

MinIon : USB-sized

Next-generation sequencing

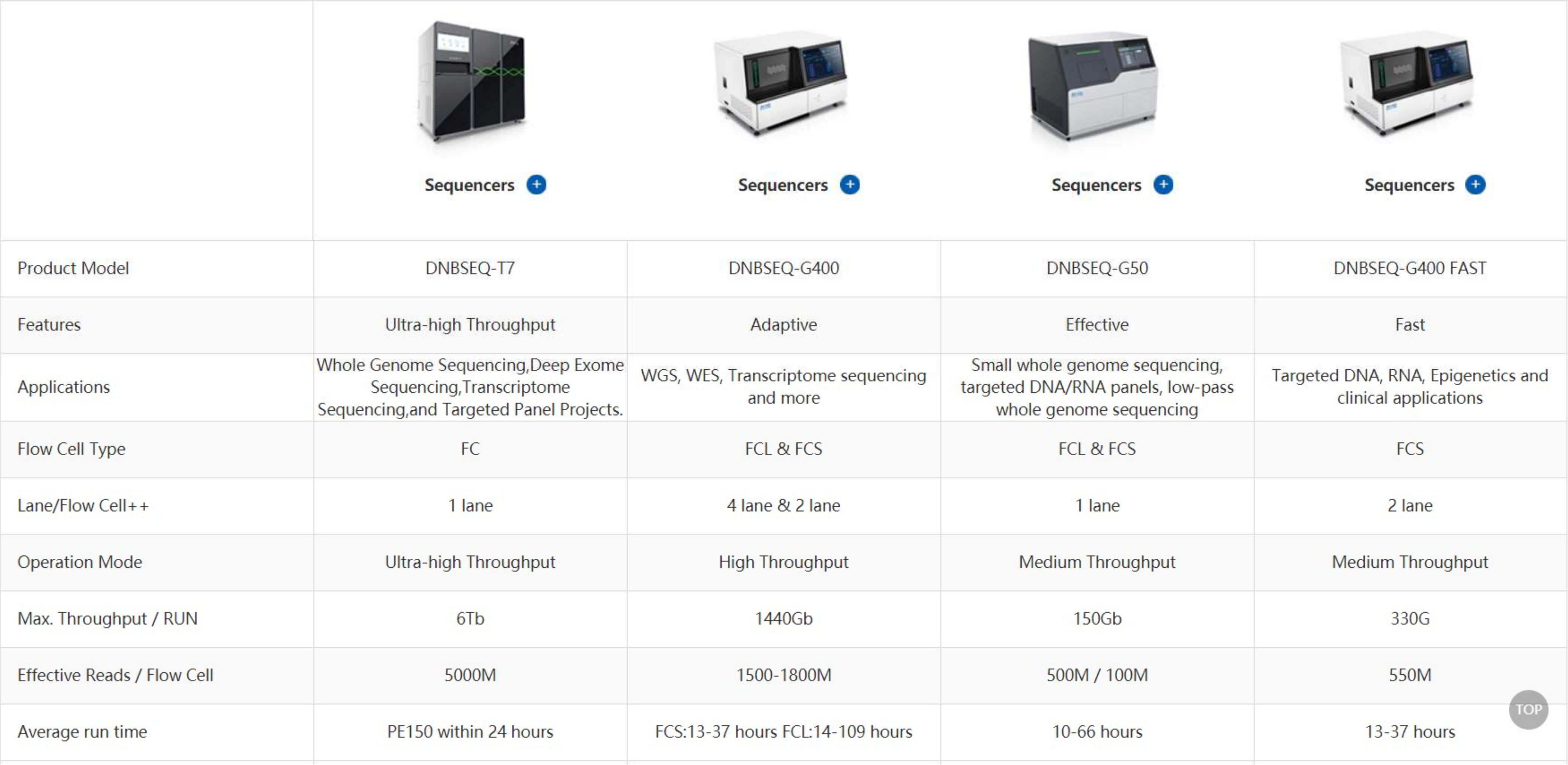
5. Oxford Nanopore Technology



	MinION	PromethION	
Number of reads at 10Kb at standard speed (280bps)[4]	Up to 2.5M	Up to 14.5M	Up to 700M
Number of reads at 10kb in Fast Mode (500bps)[4]	Up to 4.4M	Up to 26M	Up to 1250M
Read Length	Read length = fragment length Longest reported between 230-300 Kilobases (1D)	Read length = fragment length Longest reported between 230-300 Kilobases (1D)	Read length = fragment length Longest reported between 230-300 Kilobases (1D)
1D Yield[5] at 280 bps in 48 hours	Up to 25 Gb	Up to 145 Gb	Up to 7 Tb
1D Yield[5] at 500 bps in 48 hours	Up to 42 Gb	Up to 256 Gb	Up to 12 Tb
Base calling accuracy[6]	Up to 96%	Up to 96%	Up to 96%

Next-generation sequencing

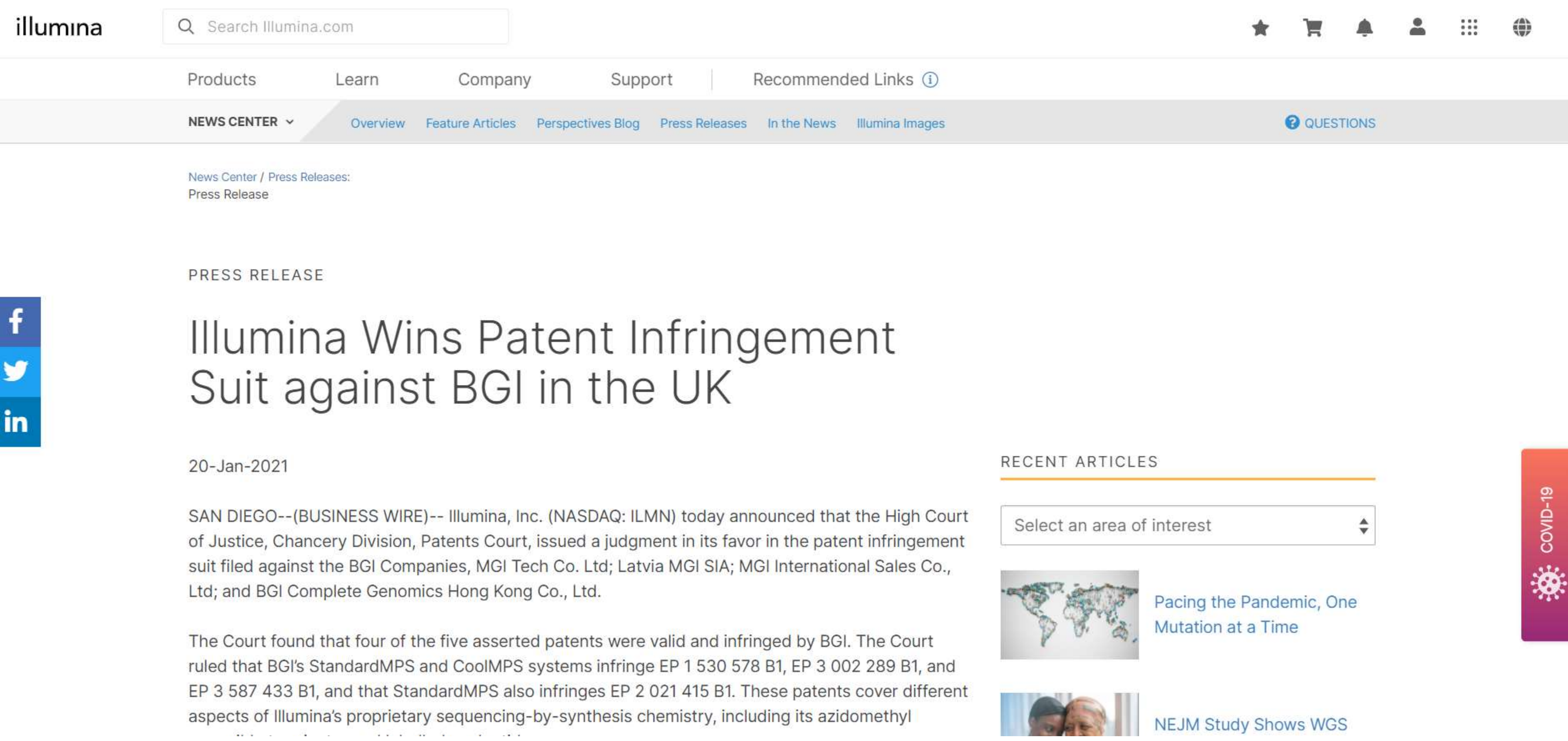
5. Beijing Genomics Institute (BGI)



	Sequencers ⊕	Sequencers ⊕	Sequencers ⊕	Sequencers ⊕
Product Model	DNBSEQ-T7	DNBSEQ-G400	DNBSEQ-G50	DNBSEQ-G400 FAST
Features	Ultra-high Throughput	Adaptive	Effective	Fast
Applications	Whole Genome Sequencing,Deep Exome Sequencing,Transcriptome Sequencing,and Targeted Panel Projects.	WGS, WES, Transcriptome sequencing and more	Small whole genome sequencing, targeted DNA/RNA panels, low-pass whole genome sequencing	Targeted DNA, RNA, Epigenetics and clinical applications
Flow Cell Type	FC	FCL & FCS	FCL & FCS	FCS
Lane/Flow Cell++	1 lane	4 lane & 2 lane	1 lane	2 lane
Operation Mode	Ultra-high Throughput	High Throughput	Medium Throughput	Medium Throughput
Max. Throughput / RUN	6Tb	1440Gb	150Gb	330G
Effective Reads / Flow Cell	5000M	1500-1800M	500M / 100M	550M
Average run time	PE150 within 24 hours	FCS:13-37 hours FCL:14-109 hours	10-66 hours	13-37 hours

Next-generation sequencing

5. Beijing Genomics Institute (BGI)



News Center / Press Releases:
Press Release

PRESS RELEASE

Illumina Wins Patent Infringement Suit against BGI in the UK

20-Jan-2021

SAN DIEGO--(BUSINESS WIRE)-- Illumina, Inc. (NASDAQ: ILMN) today announced that the High Court of Justice, Chancery Division, Patents Court, issued a judgment in its favor in the patent infringement suit filed against the BGI Companies, MGI Tech Co. Ltd; Latvia MGI SIA; MGI International Sales Co., Ltd; and BGI Complete Genomics Hong Kong Co., Ltd.

The Court found that four of the five asserted patents were valid and infringed by BGI. The Court ruled that BGI's StandardMPS and CoolMPS systems infringe EP 1 530 578 B1, EP 3 002 289 B1, and EP 3 587 433 B1, and that StandardMPS also infringes EP 2 021 415 B1. These patents cover different aspects of Illumina's proprietary sequencing-by-synthesis chemistry, including its azidomethyl

RECENT ARTICLES

Select an area of interest

Pacing the Pandemic, One Mutation at a Time

NEJM Study Shows WGS

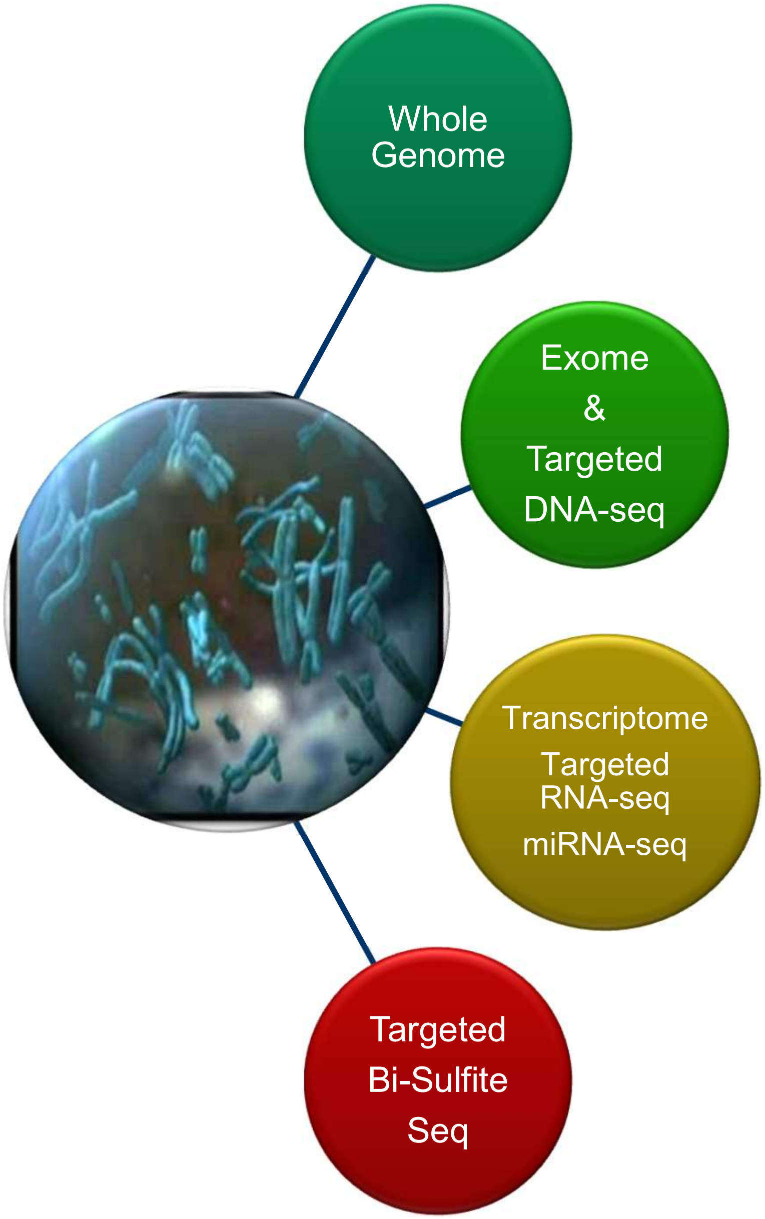
https://emea.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=924a93cb-2ddc-429a-8d4b-984909459305

Next-gen sequencing applications

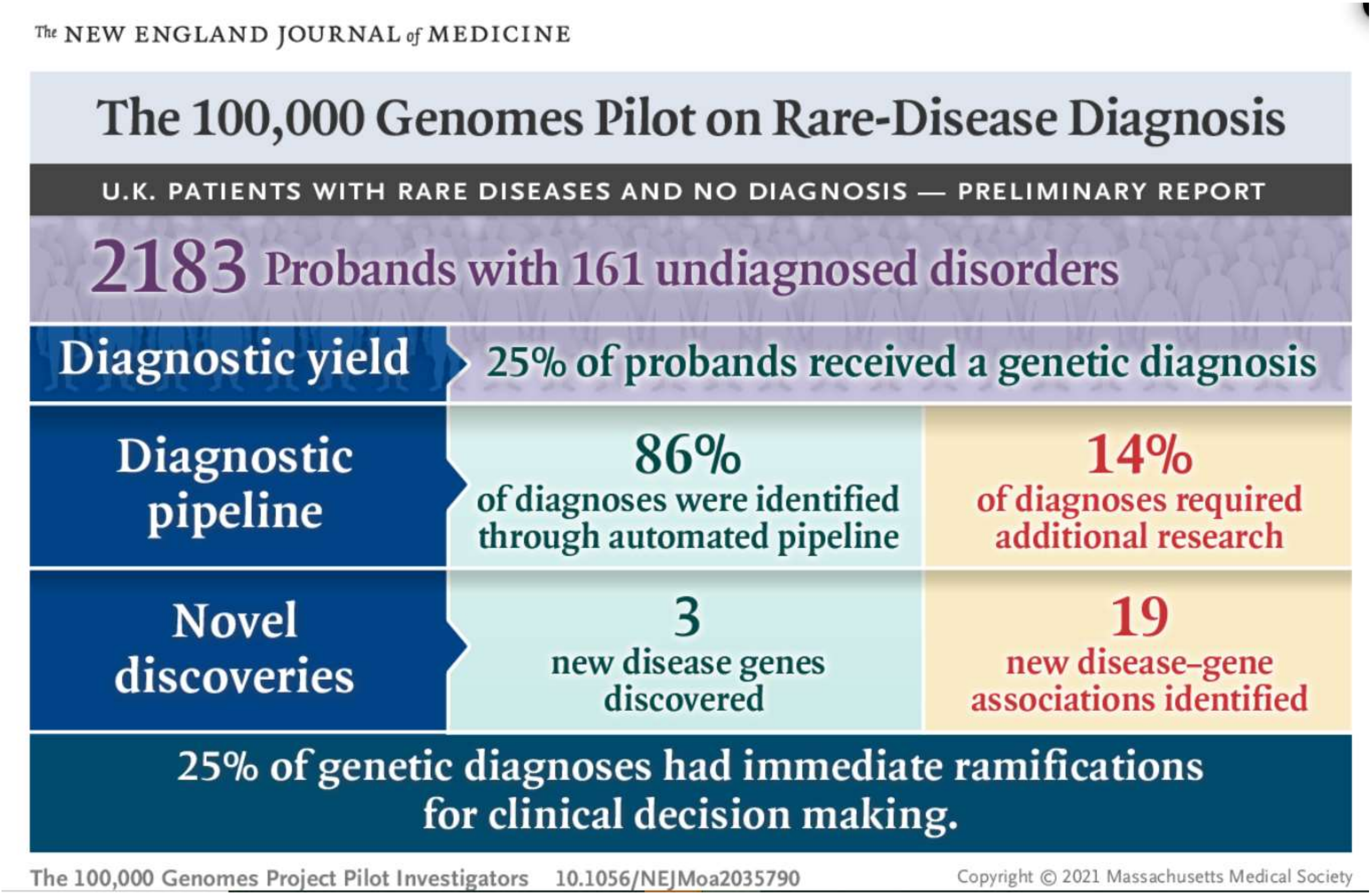
Category	Examples of applications
De-novo genome sequencing	Unknown genomes, Metagenomics (environmental samples)
Genome re-sequencing	Large-scale polymorphism discovery
Targeted resequencing	**Targeted polymorphism and mutation discovery**
Transcriptome (RNA-Seq)	**Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations**
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	**Determining patterns of cytosine methylation in genomic DNA**
Chromatin immunoprecipitation (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions

What can you do using NGS Technology: Applications for Basic and Clinical Research

Types of Variants Detectable using NGS
Large amplifications
Large deletions
Point mutations (SNP)
Insertions/Deletions
Inversions
Translocations
Copy number (CNV)
Fusions/splice variants
Gene expression data
Methylation status



What can you do using NGS Technology: Applications for Basic and Clinical Research



The NEW ENGLAND JOURNAL of MEDICINE

The 100,000 Genomes Pilot on Rare-Disease Diagnosis

U.K. PATIENTS WITH RARE DISEASES AND NO DIAGNOSIS — PRELIMINARY REPORT

2183 Probands with 161 undiagnosed disorders

Diagnostic yield	25% of probands received a genetic diagnosis	
Diagnostic pipeline	**86%** of diagnoses were identified through automated pipeline	**14%** of diagnoses required additional research
Novel discoveries	**3** new disease genes discovered	**19** new disease–gene associations identified

25% of genetic diagnoses had immediate ramifications for clinical decision making.

The 100,000 Genomes Project Pilot Investigators 10.1056/NEJMoa2035790

Copyright © 2021 Massachusetts Medical Society

Topics for Today's Presentation

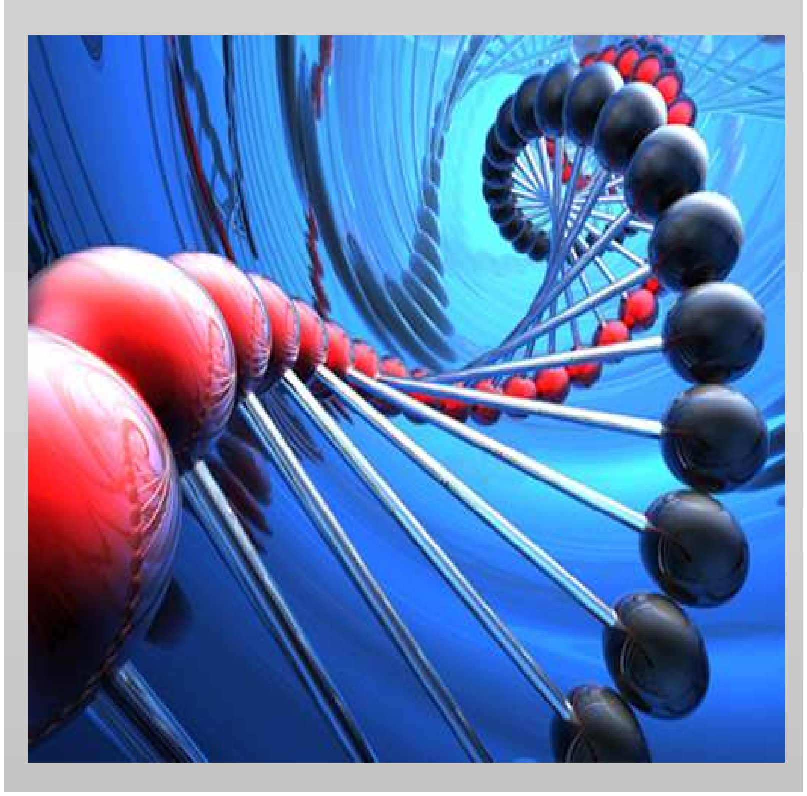
✓ 1 What is Next-Gen Sequencing?

✓ 2 Sequencers

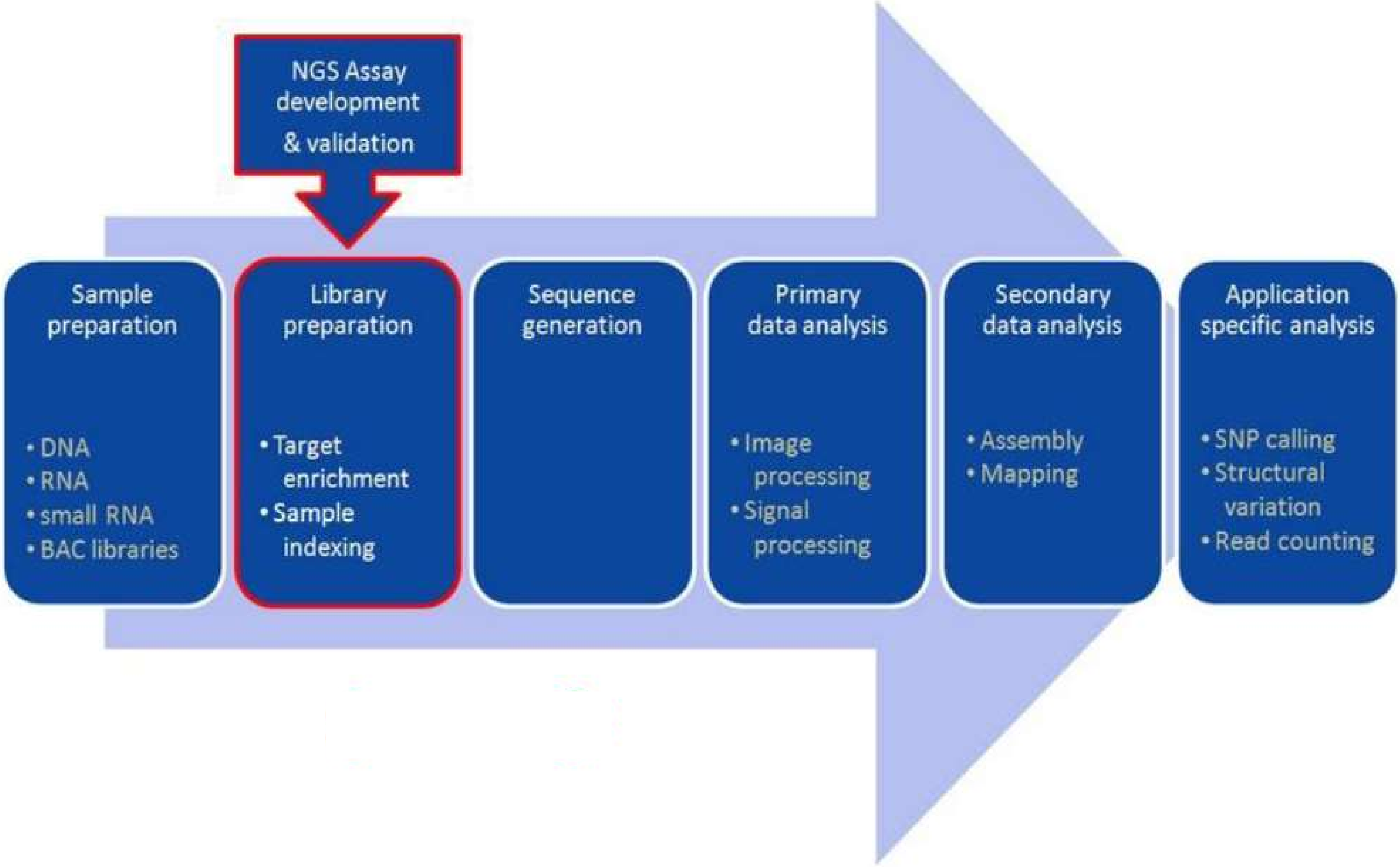
● 3 The NGS Library Prep Workflow

4 QC control

5 Analysis



Overview of the NGS Workflow



NGS Assay development & validation

Sample preparation

- DNA
- RNA
- small RNA
- BAC libraries

Library preparation

- Target enrichment
- Sample indexing

Sequence generation

Primary data analysis

- Image processing
- Signal processing

Secondary data analysis

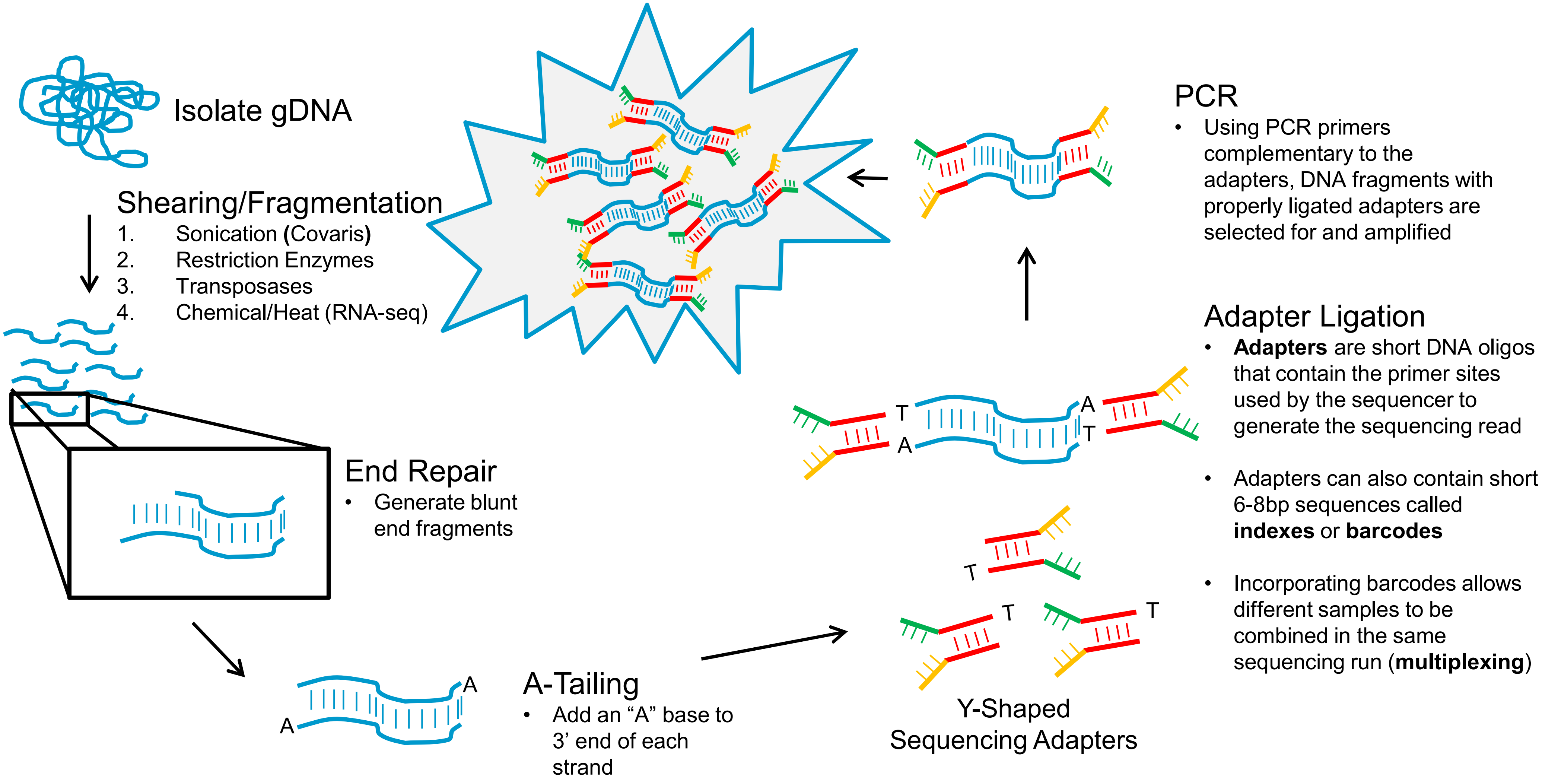
- Assembly
- Mapping

Application specific analysis

- SNP calling
- Structural variation
- Read counting

Learning the NGS Workflow: Generating a Sequencing Library

Library - A collection of DNA or cDNA fragments prepared for sequencing by a performing a series of enzymatic steps. These steps are commonly referred to as the **Library Prep.**



Isolate gDNA

Shearing/Fragmentation

1. Sonication (Covaris)
2. Restriction Enzymes
3. Transposases
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

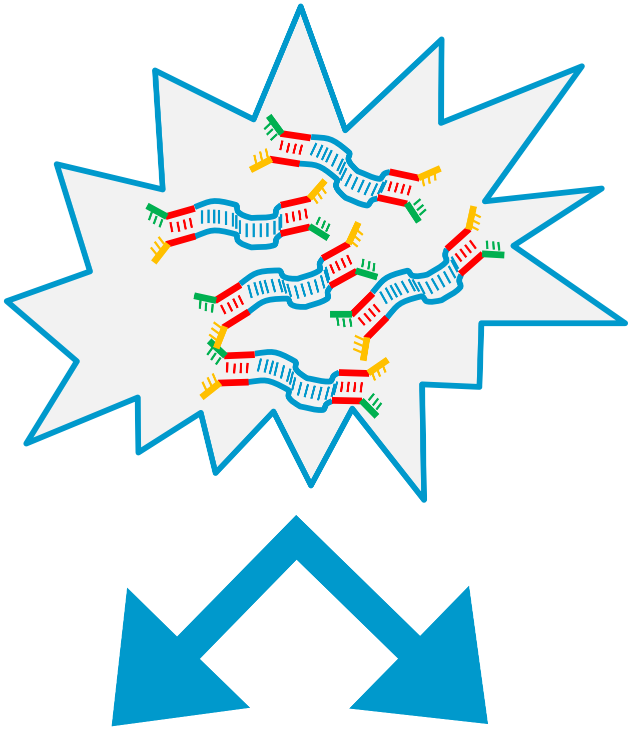
Adapter Ligation

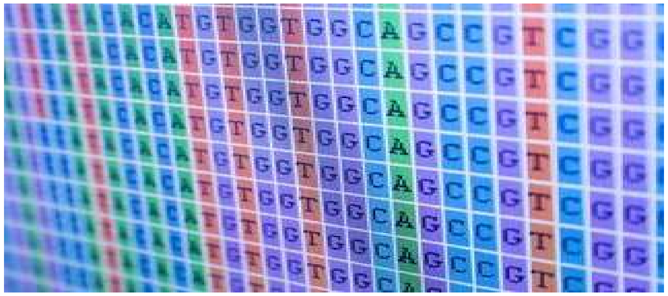
- **Adapters** are short DNA oligos that contain the primer sites used by the sequencer to generate the sequencing read
- Adapters can also contain short 6-8bp sequences called **indexes** or **barcodes**
- Incorporating barcodes allows different samples to be combined in the same sequencing run (**multiplexing**)

PCR

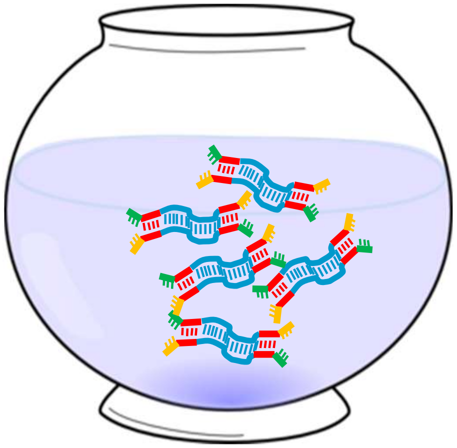
- Using PCR primers complementary to the adapters, DNA fragments with properly ligated adapters are selected for and amplified

So you've made a library….now what?





Sequence It!



Perform Target Enrichment

Target Enrichment: It's just like fishing…

Why perform target enrichment?

1. Sequence only your desired regions of interest (Exons, gene panels, intergenic regions etc...)!

2. Sequence more samples per lane/run (i.e. **Multiplex**)

3. Smaller datasets → Faster time to results

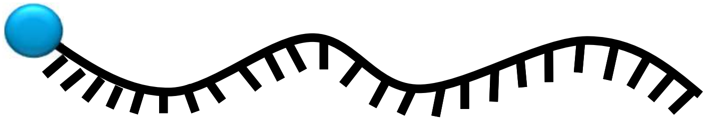
4. Save time and money

5. Increased reliability and accuracy:
   More **Reads** in regions of interest =
   Higher **Depth of Coverage**

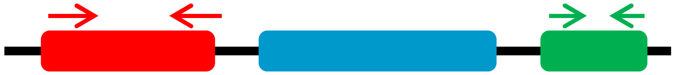
General Methods of Target Enrichment:

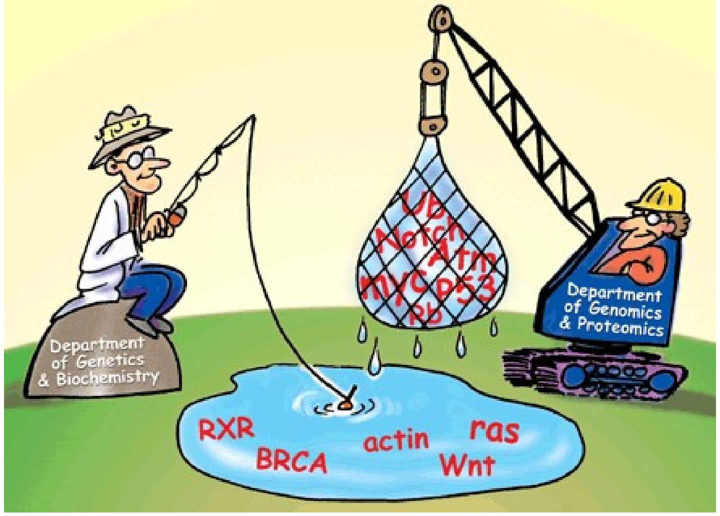
What is the basic concept?

1. Pull out the genes/regions of interest that you care about sequencing
   A. Capture the regions using biotinylated **baits**:
   - **In-solution hybrid capture**



   B. Use primers to selectively amplify the genes/regions you want to sequence:
   - **Amplicon sequencing**




Department of Genetics & Biochemistry
Department of Genomics & Proteomics
Ub
Notch
Atm
MYC
p53
Rb
RXR
BRCA
actin
ras
Wnt

(Adapted from www.sciencemag.org/cgi/content/full/291/5507/1221/F1)

2. Regions that are captured/amplified from initial library (i.e. **pre-capture library**) undergo additional amplification and processing creating a **post-capture library**

3. Off to sequencing!

Simple Target Enrichment Workflow

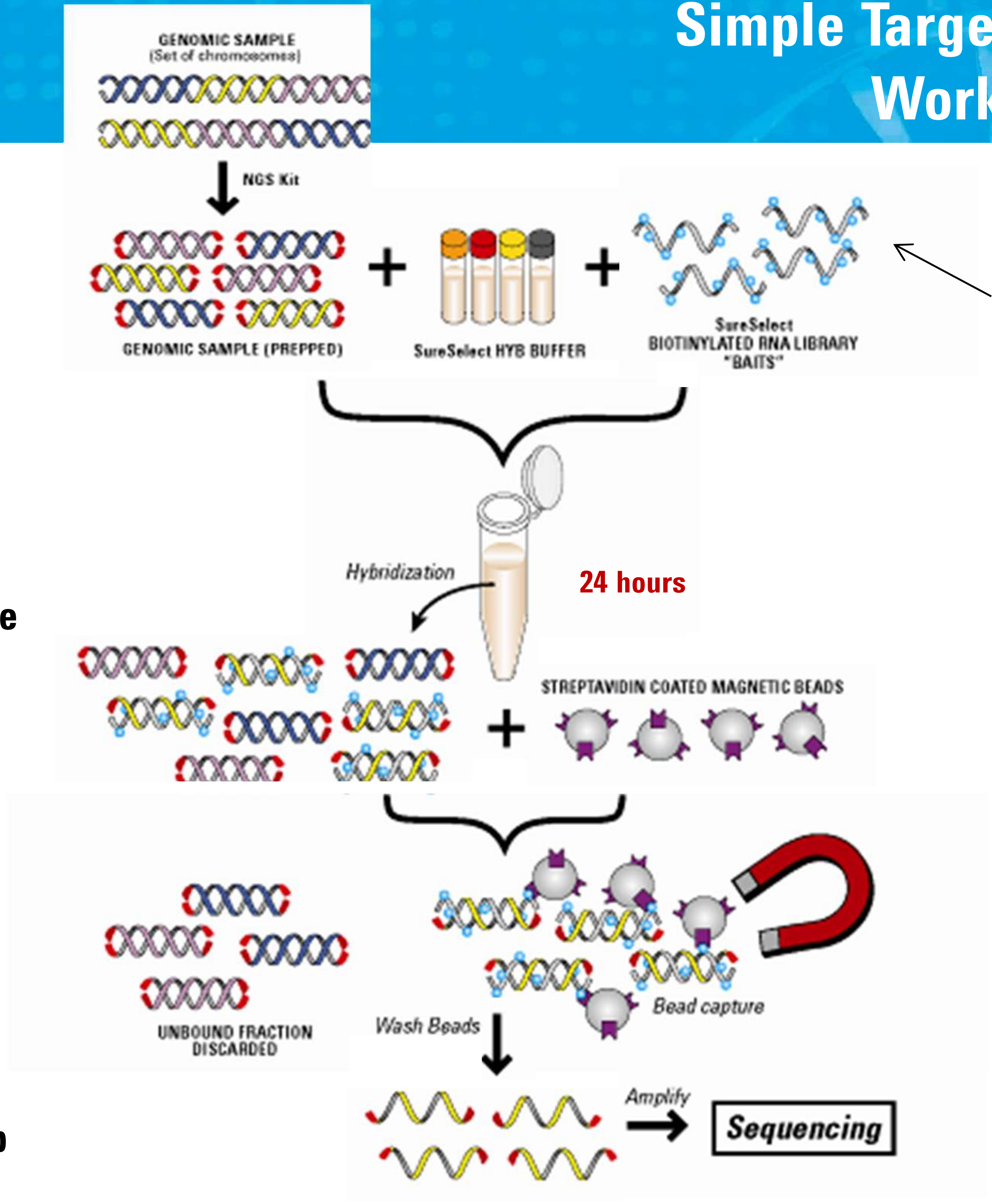
Library Preparation

Now....
200ng + Ion Proton
Protocols!

Hybridization / Capture

Bead Separation

Wash / Elution / Amp



24 hours

Baits:
- cRNA probes
- Long (120bp)
- Biotin labeled
- User-defined

SureSelect Technology

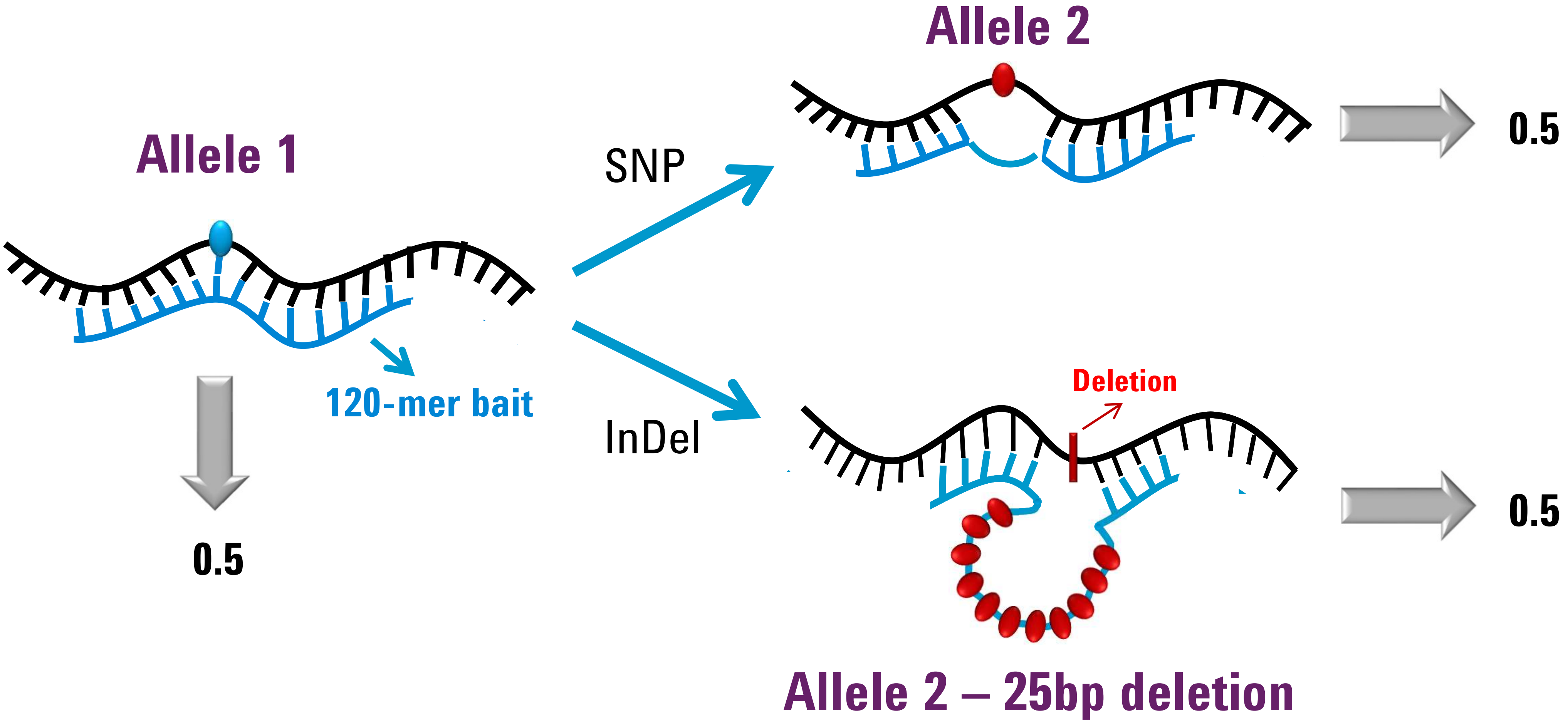
The Best Performance

Core Technology	Benefits
Ultra-Long RNA Baits **(120-mer)** **Binding strength** **RNA:DNA > DNA:DNA**	**Better Sensitivity** Detect more SNP, InDels, CNV, fusions
	Better Workflow Shorter Hybridization
	Better Allelic Balance Equal representation of both alleles

Longer Baits = Better Sensitivity

The Best Performance

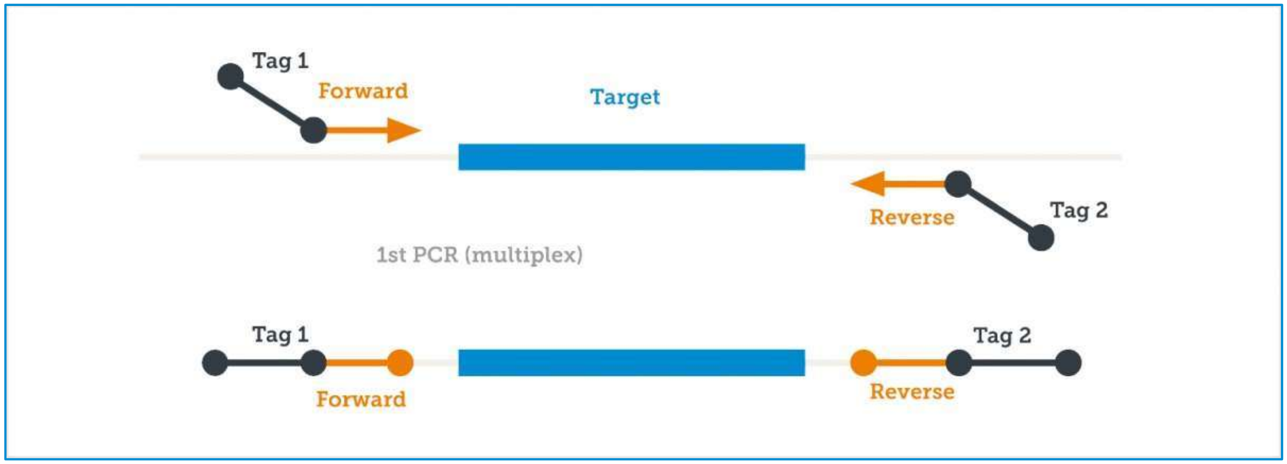
Longer, More Efficient RNA Baits Tolerate Larger Mismatches



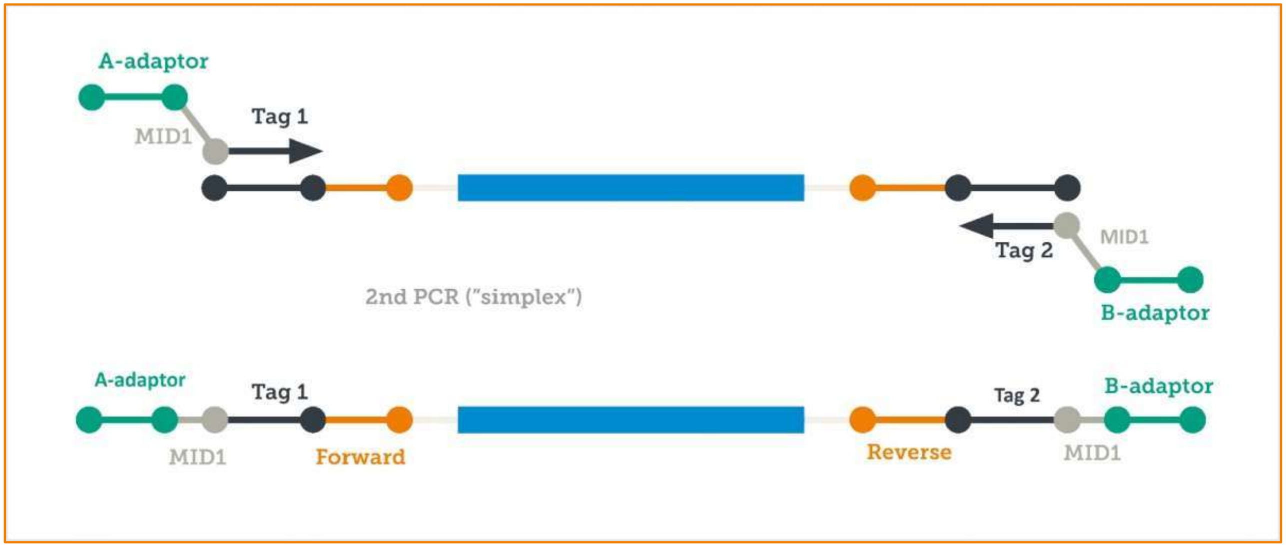
Amplicon Target

Multiplex Amplification of Specific Targets for Resequencing

Step 1: Multiplex PCR For CFTR: 2 PCR reactions per sample; 48 amplicons; 300-450 bp, including **11 control amplicons**

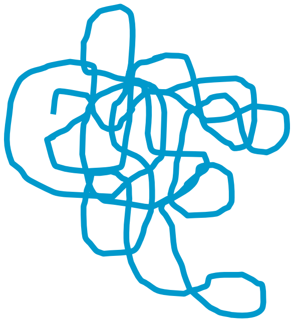


Step 2: Universal PCR for MID and adaptor incorporation



Learning the NGS Workflow: General Comparisons of Target Enrichment Methods

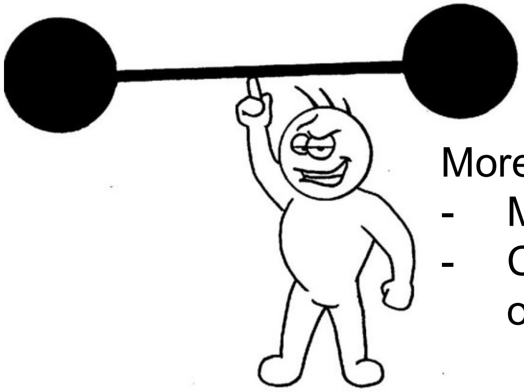
In-Solution Hybridization Capture



gDNA

- Micrograms
- Hundreds of nanograms
- Tens of nanograms ?

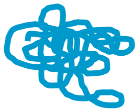
Typically Slower
hyb time range:
3-72hrs



More Robust Data:

- Many unique reads
- Can find large variety of DNA aberrations

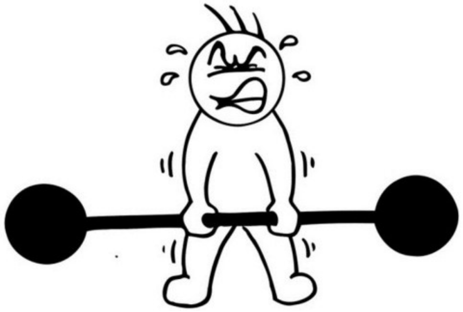
Amplicon Sequencing



gDNA

- Tens of nanograms
- And less…

Typically Faster
(no hyb required)

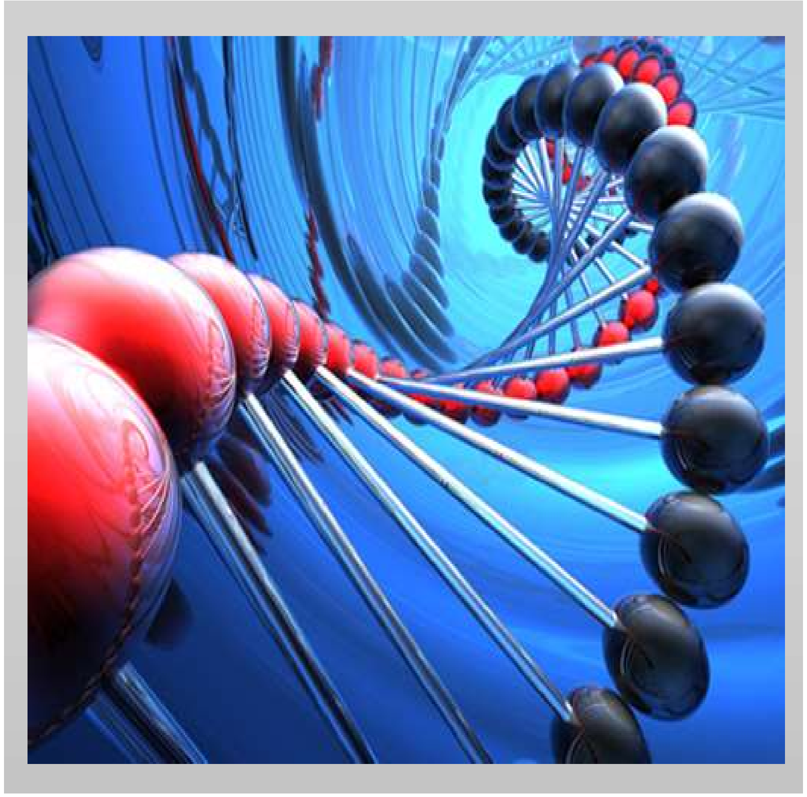


Good but Limited Data:

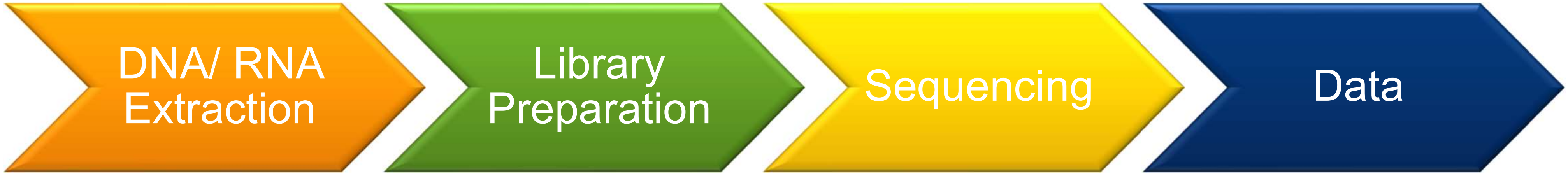
- Few/No unique reads
- Best for small/point mutations

Topics for Today's Presentation

- ✓ 1 What is Next-Gen Sequencing?
- ✓ 2 Sequencers
- ✓ 3 The NGS Library Prep Workflow
- ● 4 QC control
- 5 Analysis



Quality Control of Sequencing Libraries



DNA/ RNA Extraction

Library Preparation

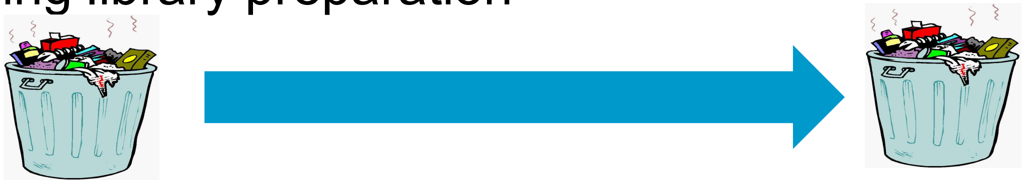
Sequencing

Data

Start with low quality material

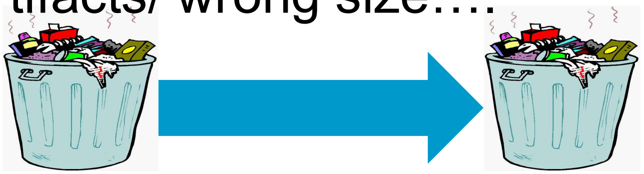
Low quality library & sequencing data

Errors during library preparation



Low/No reads & coverage

Adapter dimers/PCR artifacts/ wrong size….



Low/No reads & coverage

For Research Use Only. Not for use in diagnostic procedures

Microfluidics Product Portfolio

2100 Bioanalyzer System – Electrophoresis in microchannels

- *separation according to mobility (size)*
- *cell counting (pressure driven)*

4200 TapeStation System – ScreenTape Technology

- *Introducing the new TapeStation system*
- Unattended walk away operation with fully automated sample processing for up to 96 samples.

Principle of Electrodriven Flow

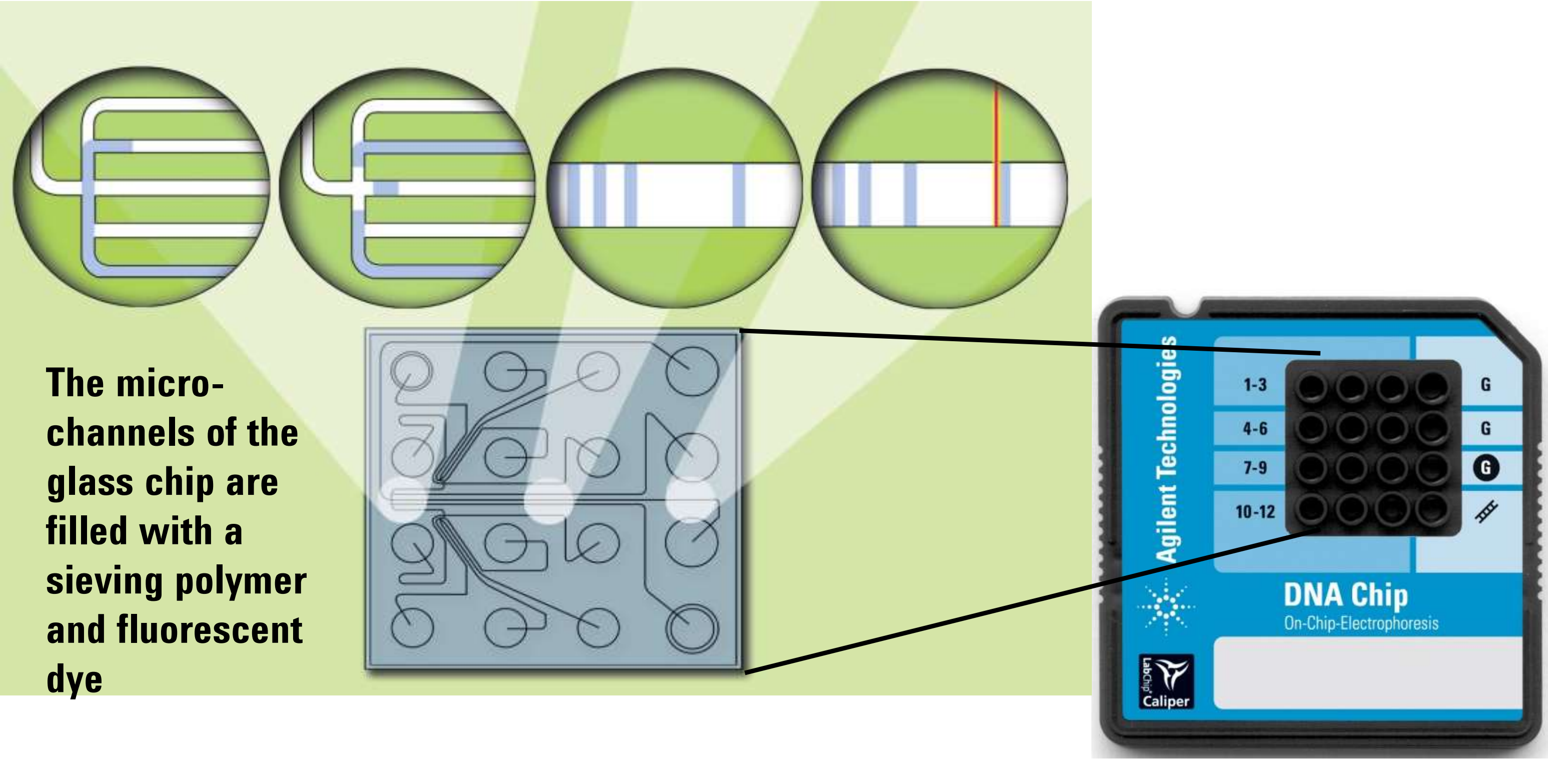
Used for molecular assays (analysis of DNA, RNA and proteins)

The sample moves electrodriven from the sample well through the micro-channels

The sample is electrokinetically injected into the separation channel

Sample components are electrophoretically separated

Components are detected by their fluorescence and translated into gel-like images (bands) and electropherograms (peaks)

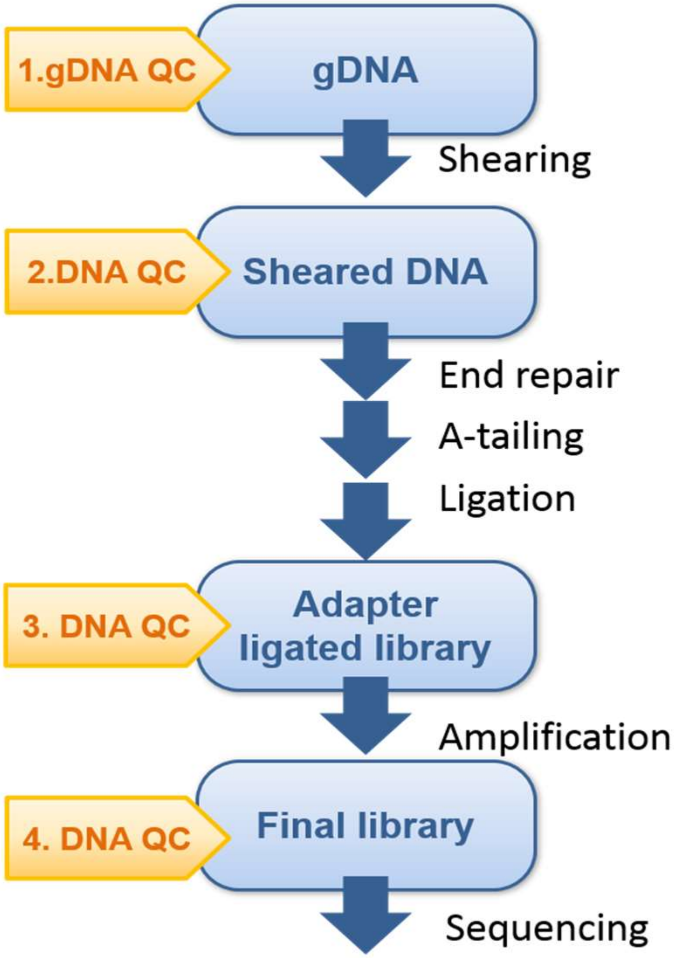


The micro-channels of the glass chip are filled with a sieving polymer and fluorescent dye

For Research Use Only. Not for use in diagnostic procedures.

General steps QC in NGS workflows

All sequencing platforms and library preparation protocols are unique but the general steps are:



1.gDNA QC — gDNA

Shearing

2.DNA QC — Sheared DNA

End repair

A-tailing

Ligation

3. DNA QC — Adapter ligated library

Amplification

4. DNA QC — Final library

Sequencing

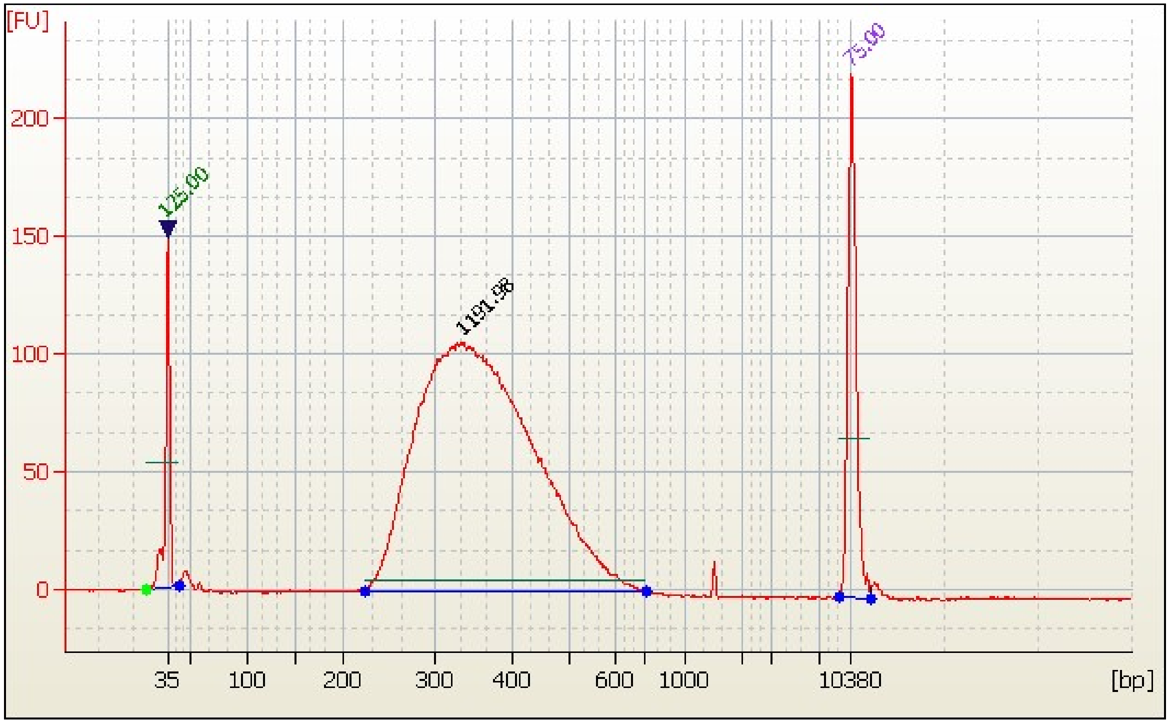
➡ Integrity of starting material

➡ Size distribution after fragmentation

➡ Size distribution of adapter ligated non-amplified library

➡ Quality control and quantification after PCR

Agilent's BioAnalyzer/Tapestation are frequently used for Quality Control of Sequencing Libraries



<u>Electropherogram</u> <u>(i.e. trace) for a Standard Library</u>

<u>Before or After Undergoing Target Enrichment</u>

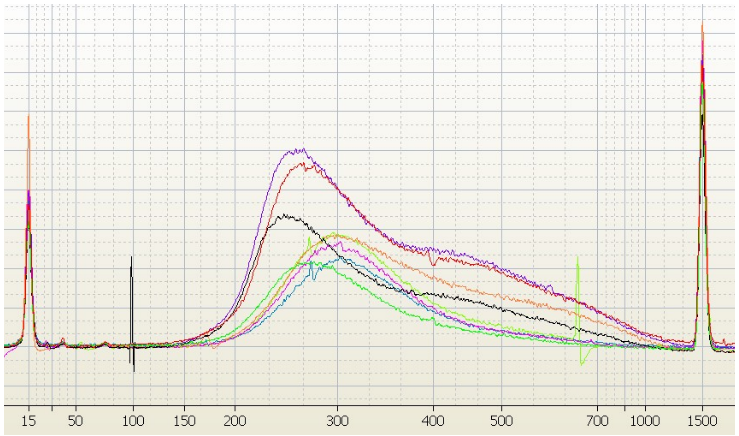
Agilent SureSelect

Illumina TruSeq

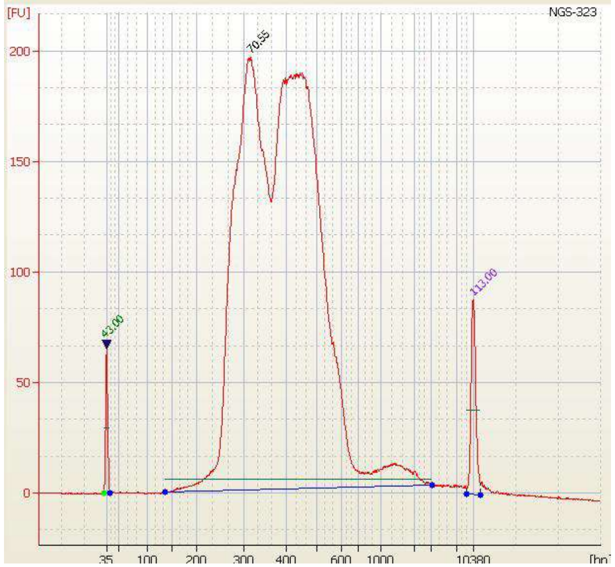
KAPA

NEB

NuGen etc…

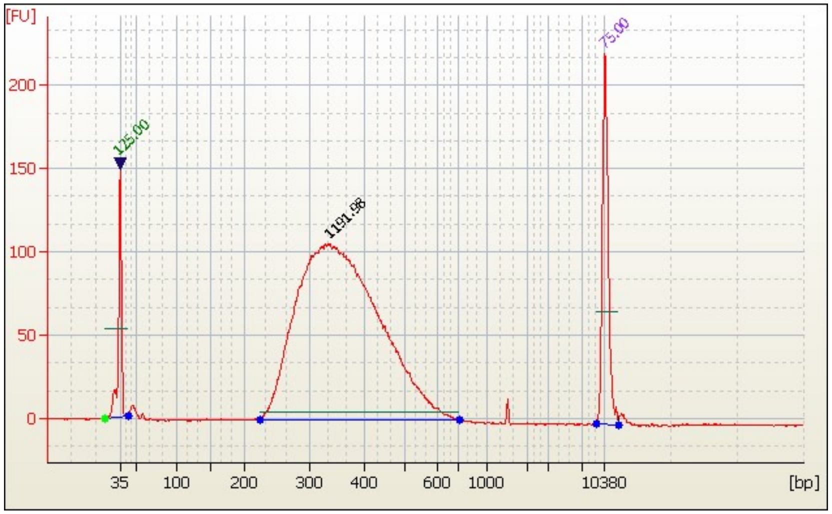


Over-Amplified: Reduce PCR

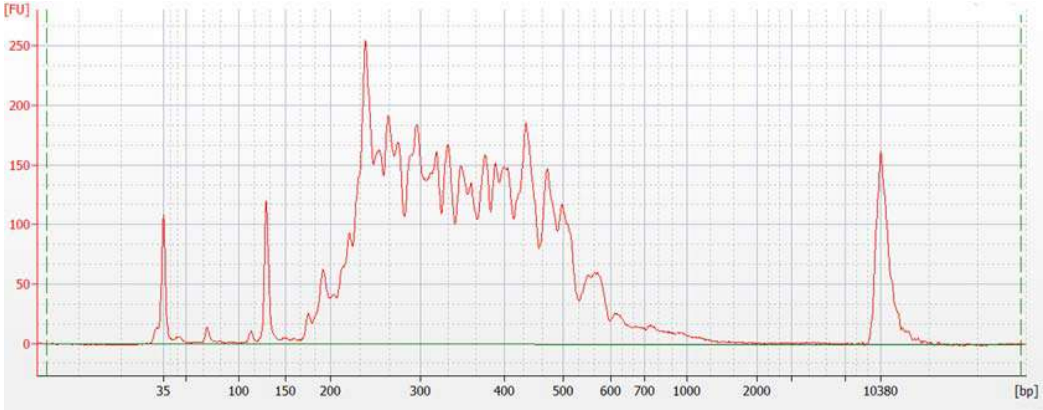


Over-loaded: Dilute and re-run

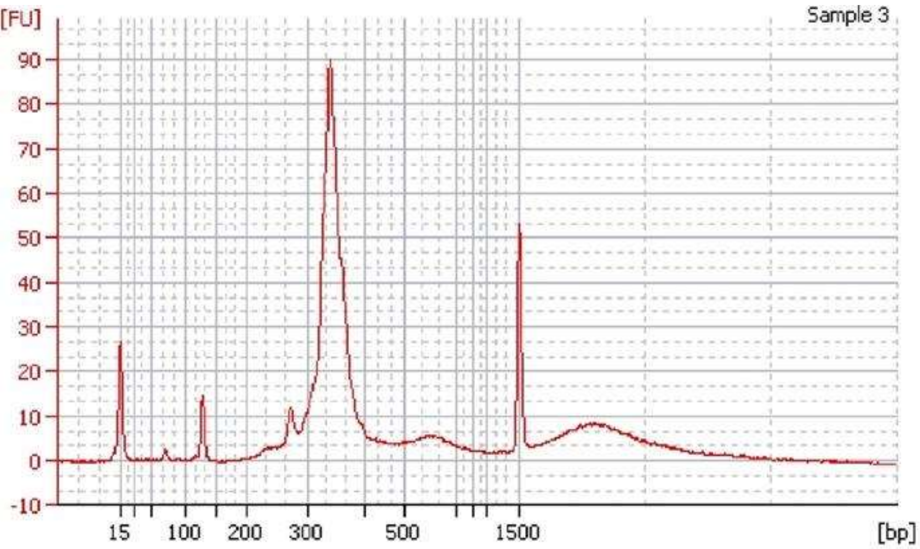
Different Library Preps Generate Different BioAnalyzer Traces



Agilent SureSelect Library Prep

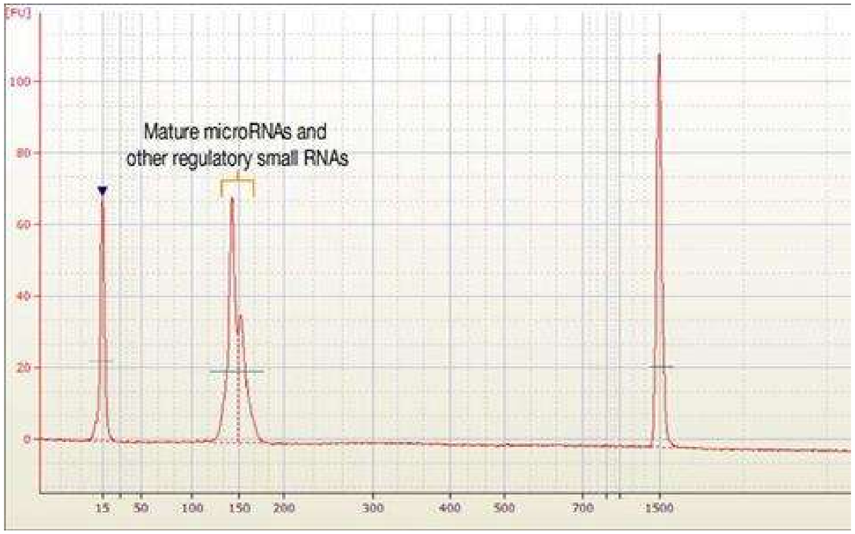


Agilent Haloplex Library Prep



TruSeq Custom Amplicon Library

(adapted from Illumina protocol)



TruSeq Small RNA Library Prep

(adapted from Illumina Protocol)

Four Things to consider beforehand….
Reviewing the NGS Library Prep Workflow

1. What kind of sample am I using and how much do I have?

- High quality gDNA from cells or fresh/frozen tissue?
- Degraded gDNA from Formalin Fixed Parafin Embedded Blocks (**FFPE**)?
- Do you have micrograms, nanograms, picograms

2. What do I want to learn from the samples I prepare?

- Identify single nucleotide polymorphisms/variants (**SNPs/SNVs**)
- Insertions and/or deletions (**InDels**)
- More complex rearrangements: Translocations, Inversions, Copy # Variations (**CNVs**)

Four Things to consider beforehand Reviewing the NGS Library Prep Workflow

3. Set your expectations accordingly

- Poor quality and very low input starting materials may require special handling
- More input required, Whole Genome Amplification
- Results from high quality gDNA ≠ Results from FFPE gDNA

4. Don't be afraid to ask for help!

- While sequencing costs have come down, it's still not cheap!
- Reach out to your sequencing cores, other labs, or vendors for guidance

Topics for Today's Presentation

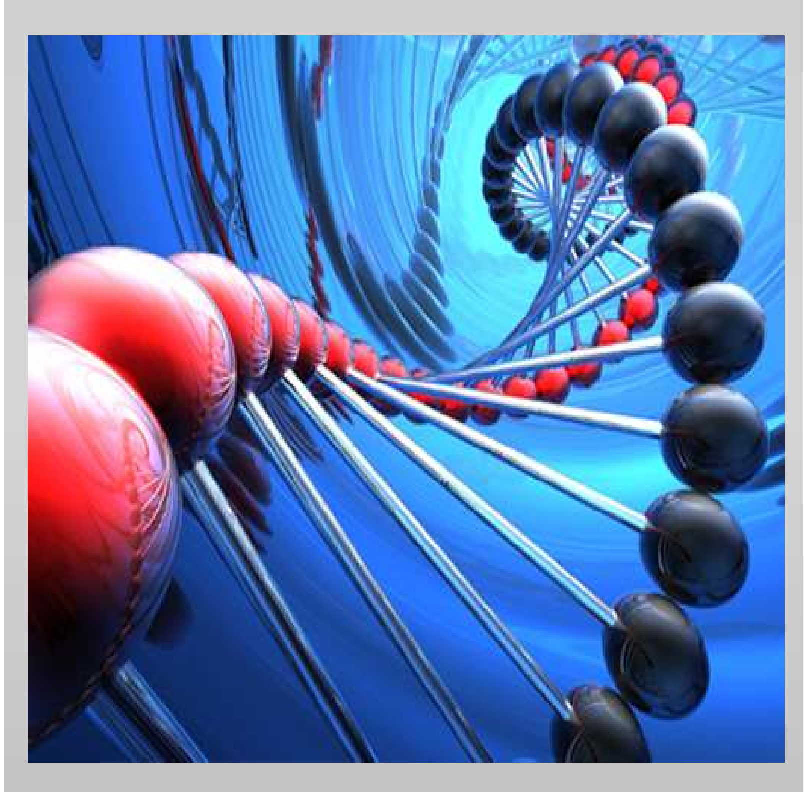
✓ 1 What is Next-Gen Sequencing?

✓ 2 Sequencers

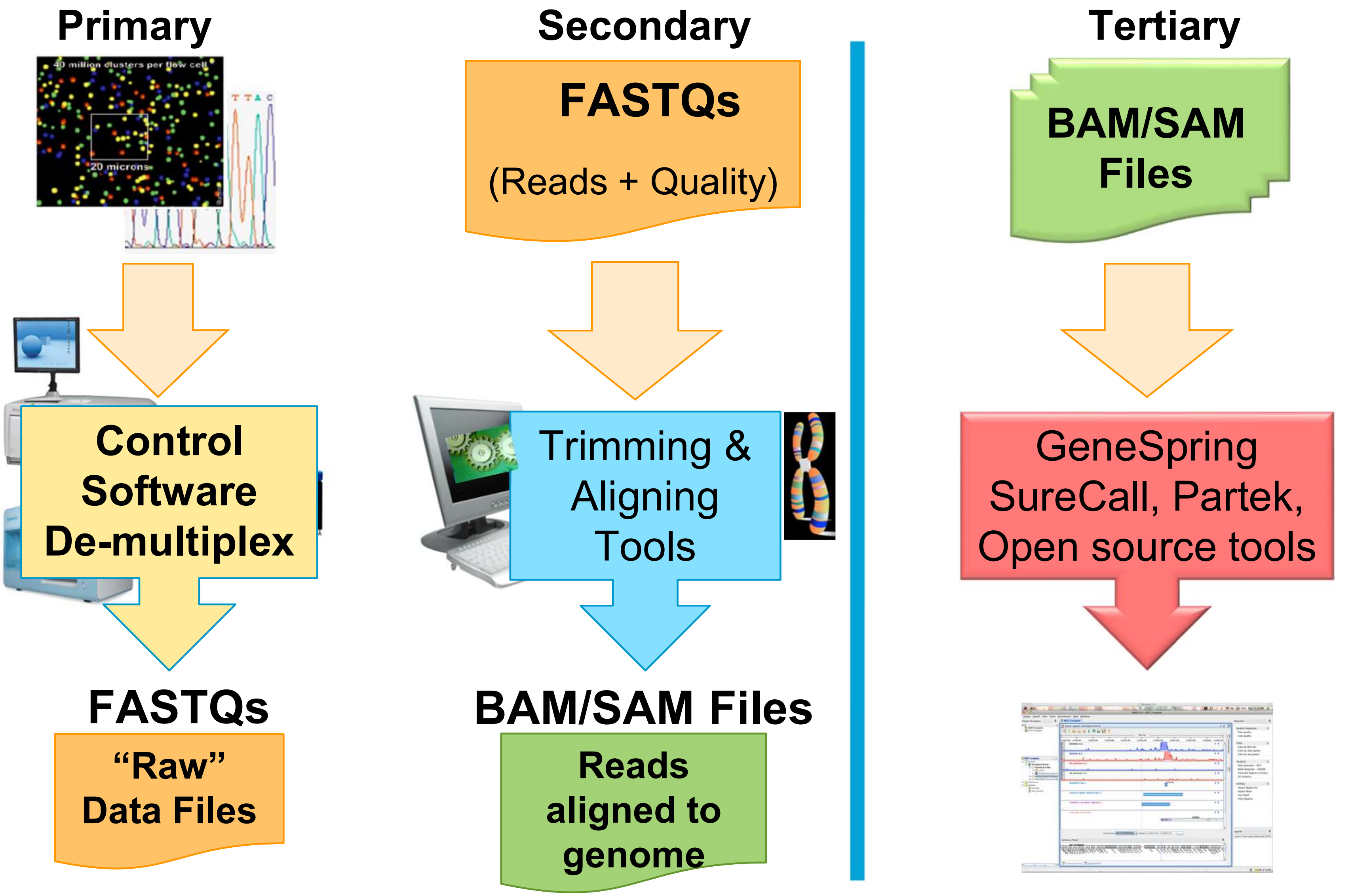
✓ 3 The NGS Library Prep Workflow

✓ 4 QC control

● 5 Analysis



Analysis
What happens after the library is sequenced?



Primary

Control Software De-multiplex

FASTQs

"Raw" Data Files

Secondary

FASTQs

(Reads + Quality)

Trimming & Aligning Tools

BAM/SAM Files

Reads aligned to genome

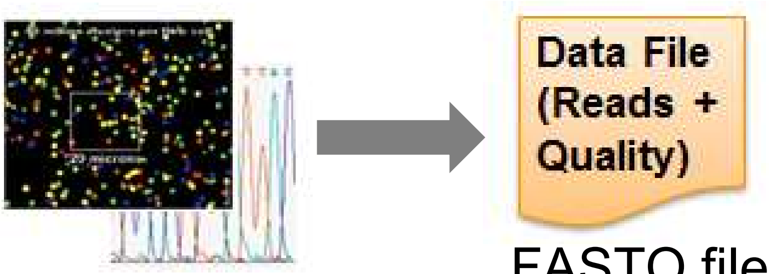
Tertiary

BAM/SAM Files

GeneSpring SureCall, Partek, Open source tools

Analysis
Primary: Clean up the raw data

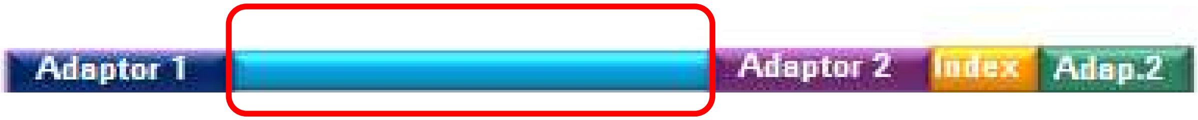
- Sole responsibility of the sequencing platform vendor
- Convert physical signals to base calls, including a quality score per base (quality = confidence in the base call, was it definitely an A, or maybe a T?)



Data File (Reads + Quality)

FASTQ file

- Demultiplex separate reads based on index
- Trim adapters
- Filter out bad reads



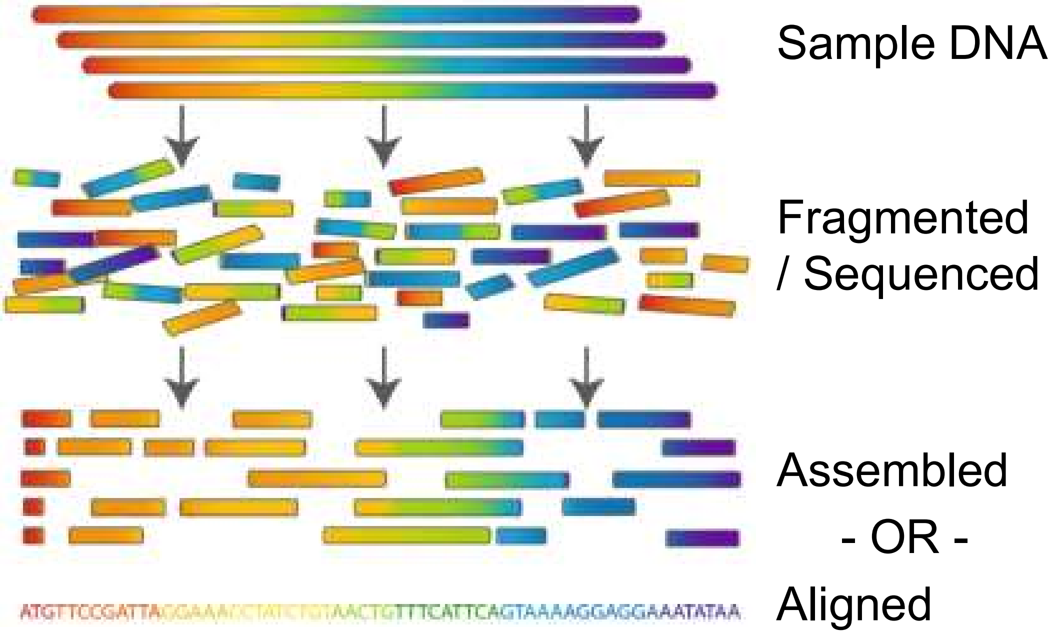
Sequence = A,C,T,G,N + Qual Score/base

Analysis
Secondary: Alignment and Assembly

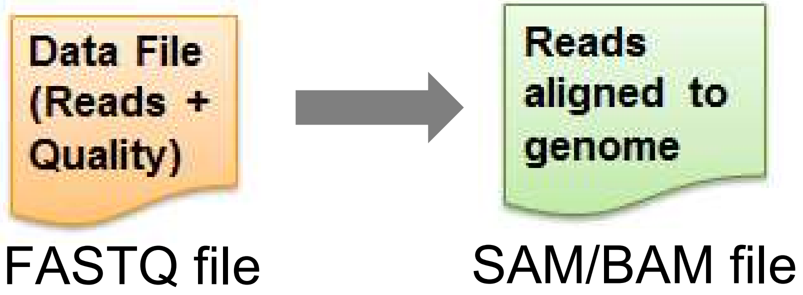
Where do all those library fragments go?

Either align them to a reference genome, or assemble them into contigs based on common overlapping sequences.

Standard output is a SAM/BAM file that stores the location information for each piece (plus a quality score for how well it mapped)

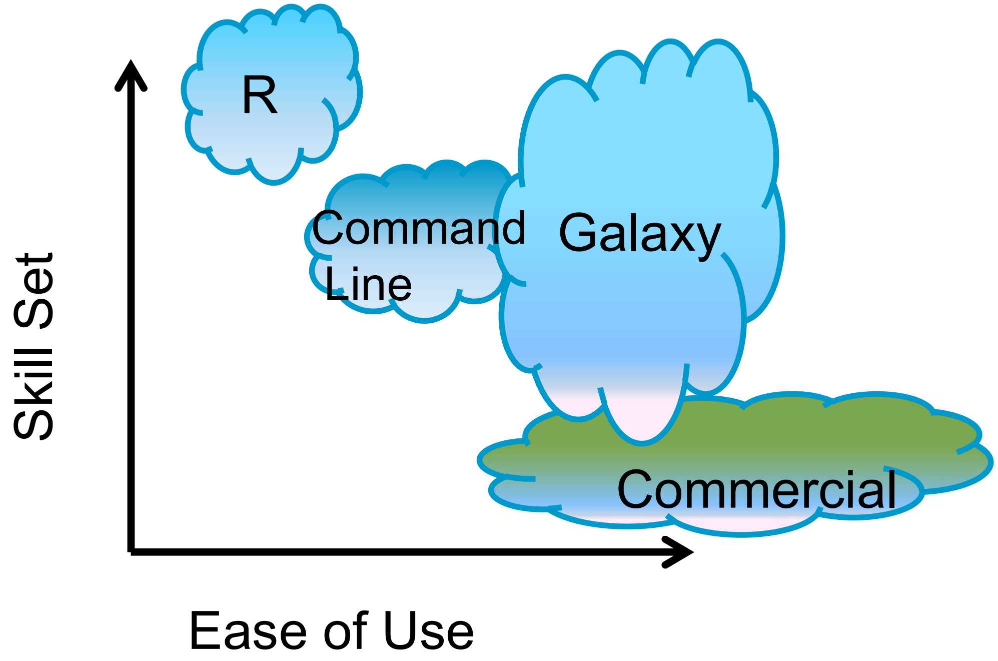


Rosalind.info



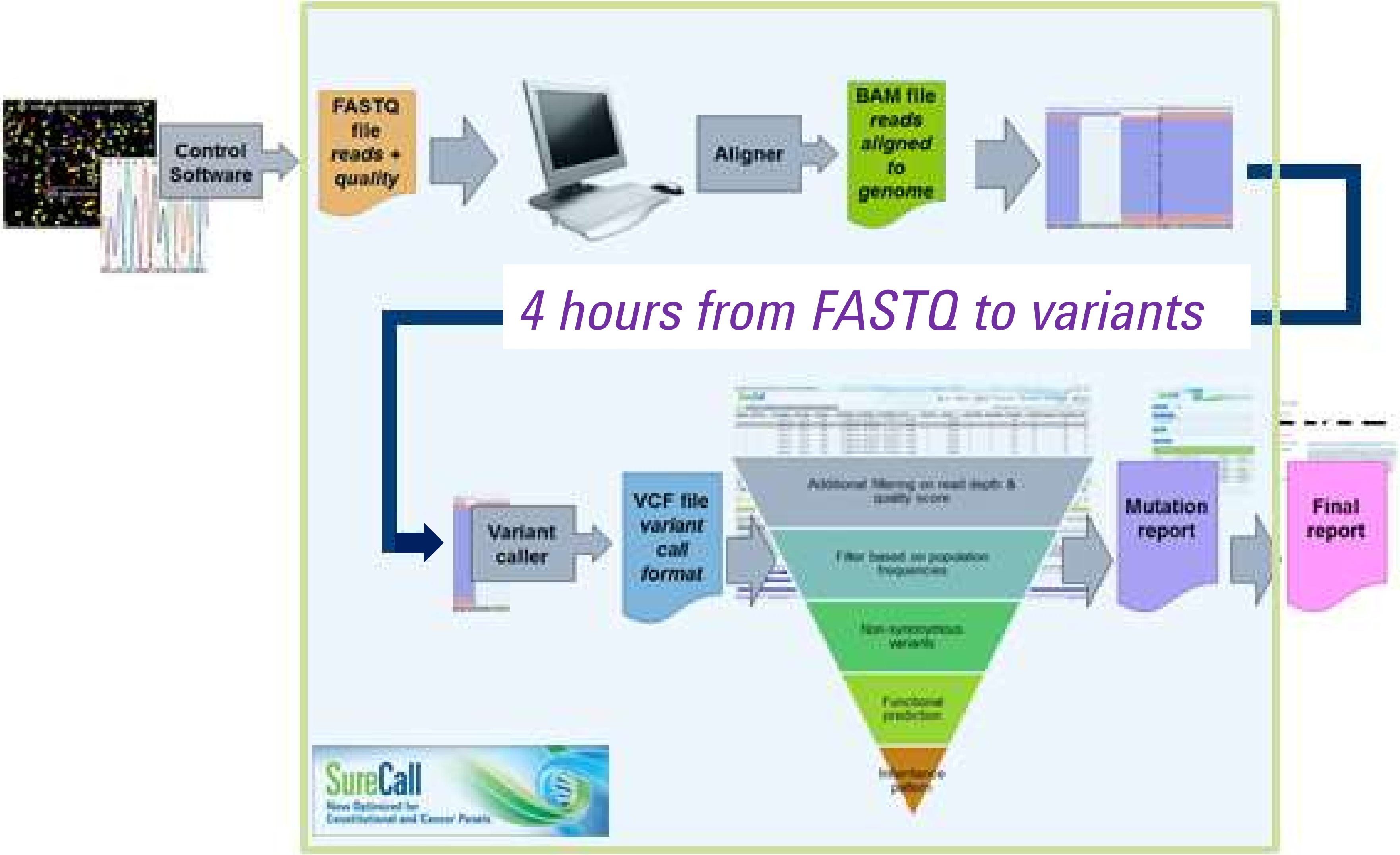
Analysis
Tertiary: How to find the useful information

- Start with the aligned SAM/BAM data file. Analysis from this point will depend on assay type and information you are looking for.
- Freeware and commercial software can help!
  - SureCall (Agilent’s free in house solution)
  - GeneSpring (License-based software Agilent has a collaboration with for multi-omic analysis)
  - Galaxy (web interface for many free NGS tools)



SureCall - SureSelect faster sample to data

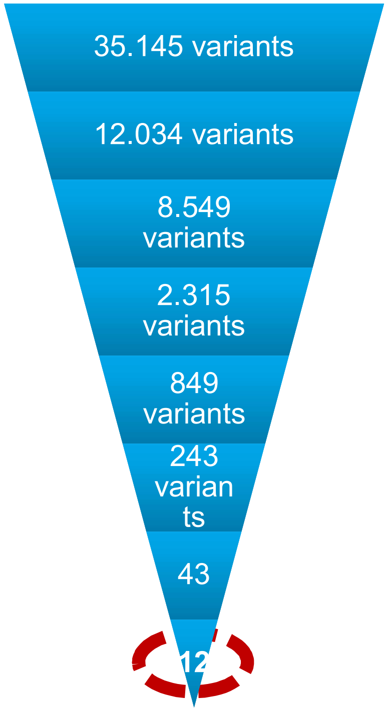
Accelerated sample to answers with SureCall 3.5


Control Software
FASTQ file reads + quality
Aligner
BAM file reads aligned to genome
Variant caller
VCF file variant call format
Additional filtering on read depth & quality score
Filter based on population frequencies
Non-synonymous variants
Functional prediction
Inheritance pattern
Mutation report
Final report
SureCall
Now Optimized for Constitutional and Cancer Panels

4 hours from FASTQ to variants

The interpretation challenge

Which variants are clinically significant?



35.145 variants

12.034 variants

8.549 variants

2.315 variants

849 variants

243 variants

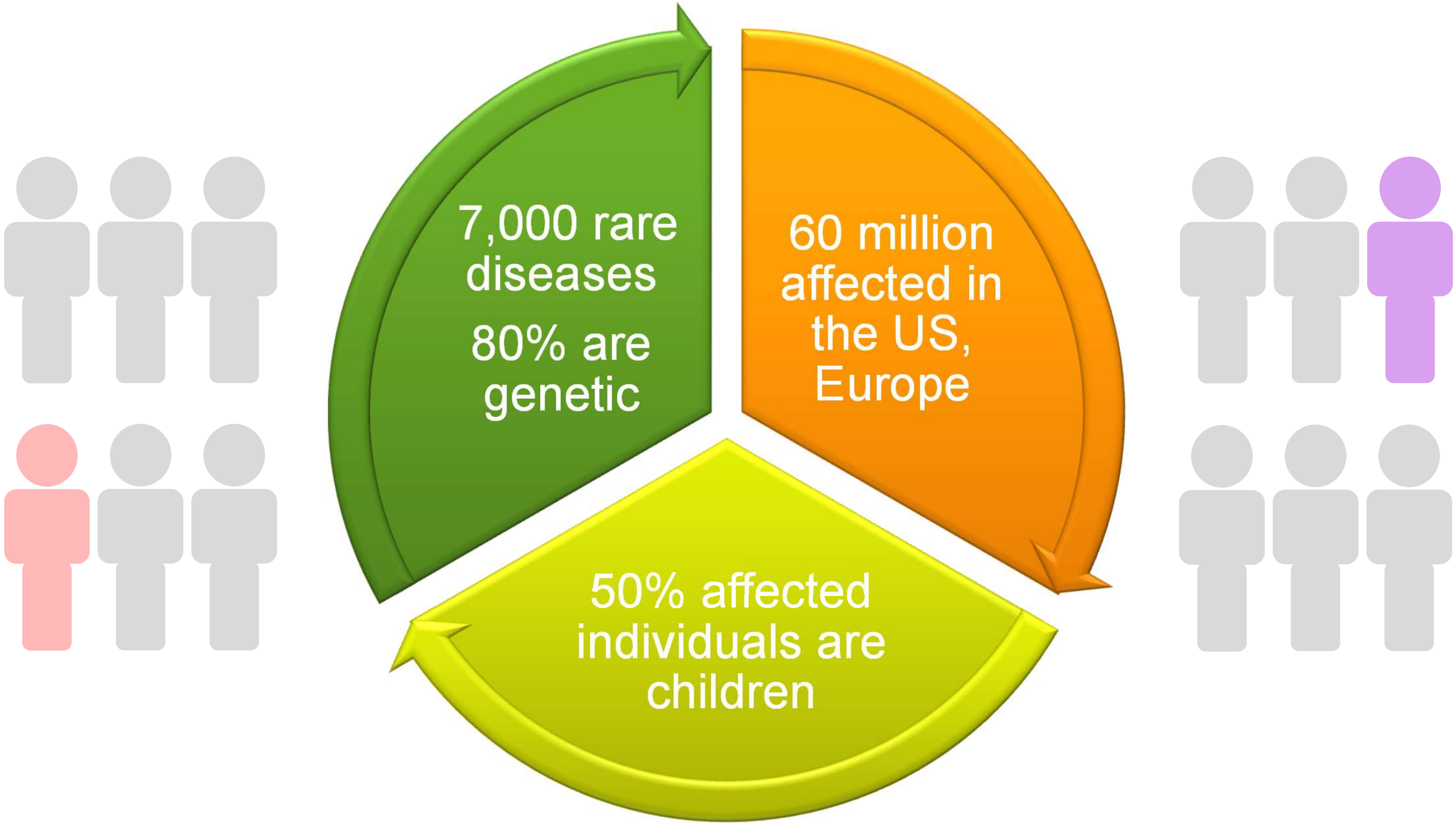
43

12

I need to…

- Discern high quality sequencing results from artefacts and false positives
- Filter out variants that are commonly found in the population
- Prioritize on genes and variants that are linked to that patient's clinical phenotype
- Take into account the whole body of published and community knowledge on variants and their role in disease
- Wade through all my historical findings and previous reports
- Take into account the family history and work through hypotheses on relevant inheritance modes – looking at siblings and parents if available
- Check all public databases on actionable and clinically relevant findings (list)…

Rare diseases affect 350 million people worldwide



7,000 rare diseases

80% are genetic

60 million affected in the US, Europe

50% affected individuals are children

Glossary
Library Prep

1. **Library Preparation (Library Prep)** – The method(s) used to prepare DNA or RNA for next-generation sequencing.

2. **Sequencing Library (Library)** – A collection of DNA or cDNA fragments of a given size range with adapters ligated to each end that can be run through a sequencer. Libraries can be DNA or cDNA (cDNA libraries prepared when performing RNA-seq).

3. **Adapters** – Oligonucleotides of a known sequence that are ligated to each end of a DNA/cDNA fragment (i.e. insert). They provide the primer sites used for sequencing the insert.

4. **Index/Barcode** - Short sequences of typically 6 or more nucleotides that serve as a way to identify/label individual samples when they are sequenced together in a single sequencing lane/chip. Barcodes are typically located within the sequencing adapters.

5. **Multiplexing** – Mixing two or more different samples together such that they can be sequenced in a single sequencing lane or chip. Samples that are to be combined, need to be barcoded/indexed prior to being mixed together.

6. **Library Complexity** – The number of unique DNA fragments contained in a sequencing library.

7. **Electropherogram** – A graphical representation of the size and quantity of a DNA or RNA sample run through a BioAnalyzer, TapeStation or other instrument used for performing quality control.

8. **FFPE DNA/RNA** – Formalin Fixed Parafin Embedded DNA or RNA. When attempting to prepare sequencing libraries from these sample types, modifications are often required to standard library preparation protocols to accommodate the level of DNA/RNA degradation commonly found from samples stored using this technique.

Glossary
Target Enrichment

1. **Target Enrichment (Capture)** – Methods to allow one to isolate and/or increase the frequency of specific genes or other regions of interest from a DNA or cDNA library prior to being sequenced. The regions of interest are retained for sequencing and the remaining material is washed away.

2. **Baits** – Common name given to the oligonucelotide sequences (i.e. probes) that are responsible for identifying and binding to a given region of interest for performing target-enrichment.

3. **In-Solution Capture** – A method of performing target enrichment that requires samples to be hybridized to baits to select and enrich the sample for the desired regions of interest.

4. **Amplicon Sequencing** – A method of performing target enrichment that utilizes one or more pairs of PCR primers to increase the number of copies of the genes or other regions of interest that will ultimately be sequenced.

5. **Gene Panels** – Name frequently given to the selected regions of interest (this can genes or intergenic regions) that will be captured using some form of target-enrichment technology.

6. **Pre-Capture Library** – Common name given to the sequencing library that is created before that library undergoes some form of target-enrichment.

7. **Post-Capture Library** – Common name given to the sequencing library after it has completed some form of target-enrichment.

Glossary
MethylSeq

1. **Epigenetics** – The study of changes in gene expression that are caused by mechanisms that do not effect the underlying DNA sequence. Examples include covalent modification to histones tails and the methylation of DNA.

2. **Epigenetics Writers** – Individual enzymes or protein complexes that facilitate the establishment of covalent modifications to DNA or histones. Examples include DNA methyltransferase and histone methyltransferase.

3. **Epigenetic Readers** - Proteins that identify specific epigenetic marks and either directly bind to or recruit proteins to bind to them in order to modulate gene expression. Examples include methyl CpG binding proteins or members of the Polycomb and Trithorax group proteins.

4. **Epigenetic Erasers** – Proteins that can remove covalent modifications to DNA and histones.

5. **CpGs** – Regions of the genome where cytosines precede guanines along the linear DNA sequence. The "p" in the CpG annotation stands for phosphate which means the cytosine nucleotide occurs 5' of the guanine nucleotide. This nomenclature is used to prevent confusion since cytosines form Watson-Crick base pairing with guanines, which are not sites for DNA methylation.

6. **CpG Islands** – Regions of the genome, typically >500bp, that contain a high density of CpG dinucleotide sequences.

7. **CpG Island Shores** – Term that describing the regions of differentially methylated CpG dinucleotides which occur approximately 2 kb away from annotated CpG islands .

8. **CpG Island Shelves** – Similar to CpG shores, however these regions are found even further from annotated CpG islands in the genome, approximately 4 kb away from annotated CpG islands.

9. **DMRs** – Referring to Differentially Methylated Regions of the genome.

Glossary

Analysis

1. **Assembly** – Process of creating a reference genome or transcriptome from shotgun sequenced data
2. **Alignment** – Assign genomic coordinates to sequences by comparing to a reference genome
3. **Quantification/Mapping** – Assign aligned reads to a particular transcript that overlaps the genomic coordinates
4. **Normalization** – Process of equalizing data between samples and genes so that read counts are comparable
5. **Read** – Base pair information of a given length from a DNA or cDNA fragment contained in a sequencing library. Different sequencing platforms are capable of generating different read lengths.
6. **Single End Read** – The sequence of the DNA is obtained from the 5' end of only one strand of the insert. These reads are typically expressed as 1x "y", where "y" is the length of the read in base pairs (ex. 1x50bp, 1x75bp).
7. **Paired End Read** – The sequence of the DNA is obtained from the 5' ends of both strand of the insert. These reads are typically expressed as 2x "y", where "y" is the length of the read in base pairs (ex. 2x100bp, 2x150bp).
8. **Mate Pair Read** – The sequence of the DNA is obtained similar to paired-end reads, however the size of the DNA insert is often much greater in size (2-10kb in length) and the paired reads originate from a single strand of the DNA insert.
9. **Depth of Coverage** – The number of reads that spans a given DNA sequence of interest. This is commonly expressed in terms of "Yx" where "Y" is the number of reads and "x" is the unit reflecting the depth of coverage metric (i.e. 5x, 10x, 20x, 100x)
10. **Sequencing Depth** – The amount of sequencing a given sample requires to achieve a certain depth of coverage. This is frequently expressed as the number of reads a sample requires (ex. 40 million reads, 80 million reads) or the number of bases of sequencing a sample requires (ex. 4 gigabases, 100 megabases).
11. **Call -** Referring to the identification of a given aberration detected in the sequenced sample when compared to the reference/normal genome.
12. **SNP/SNV** – Referring to a Single Nucleotide Polymorphism or Single Nucleotide Variant detected in a sample.
13. **CNVs** – Referring to Copy Number Variation that is detected in sample.
14. **InDels** – One or more Insertion or Deletion event that is detected in a sample.