September 17, 2021

DNA Sequencing at Scale

Craig Ciesla, Ph.D.

VP – Head of Advanced Platforms and Devices
Illumina Research & Development

illumina®

Agenda

- Illumina Overview
- Introduction to Sequencing by Synthesis
- High throughput DNA sequencing
- Scaling output to meet the needs of DNA-based data storage

Illumina Overview

illumina®

Who We Are

Illumina is an applied genomics technology company making genomics useful for all. We are tirelessly working to create the leading-edge technology that enables clinicians and researchers to not only understand the genome but also fully tap its power.

$3.2 billion (2020)
Annual revenue

>7,800
Number of employees

Francis deSouza
President & CEO

San Diego, CA, USA
Headquarters

1998
Year founded

Making Breakthroughs Possible

At Illumina, we're amidst the most important human health transformation of our time, as sequencing delivers new insights into the genome.

We've made great strides in the field of genomics, and we're just getting started.

From diagnosing disease in critically ill infants to developing new treatments for cancer: Illumina is helping to improve human health by unlocking the power of the genome.

Cost of Sequencing, Per Human Whole Genome

Since 2001, the cost of DNA sequencing has dropped more than 100,000x from $100 million USD per human genome to less than $600 USD today.

1. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcosts; 2. NovaSeq™ 6000 v1.5 Reagent Kit. Data on file.

More than 17,000 Sequencing Systems Installed Around the World

Our Sequencing Instruments

Low-throughput


MiSeq™


MiniSeq


iSeq 100

Mid-throughput


NextSeq 500


NextSeq 550


NextSeq 1000/2000

IVD Instruments


MiSeq™Dx


NextSeq™ 550Dx

High-throughput


NovaSeq™ 6000


HiSeq 2000


HiSeq 2500


HiSeq X Ten


HiSeq 3000/4000

NovaSeq™ 6000

1st sequencer to exceed $1 billion annual revenue

Who We Serve

Markets

- Oncology
- Reproductive Health
- Genetic Disease
- Microbiology
- Agriculture
- Molecular and Cell Biology

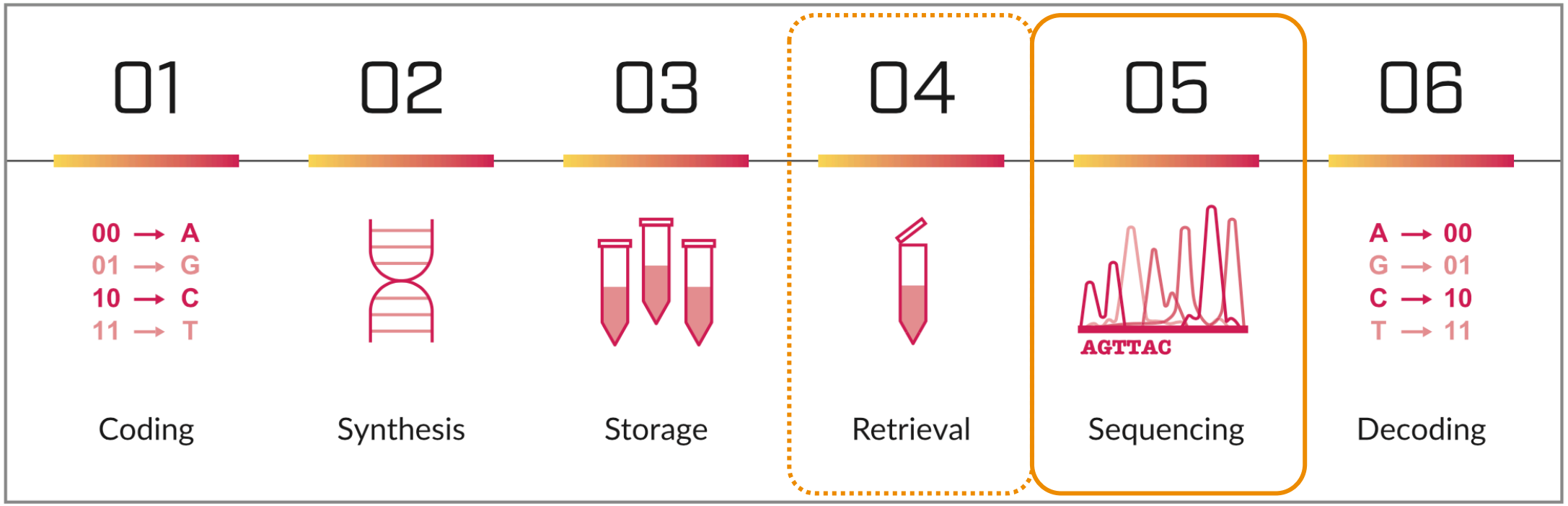
Customers

- Universities and Academic Research Centers
- Pharmaceutical Companies
- Genome Centers
- Biotechnology Companies
- Hospitals
- Consumer Genetics Companies
- Government Agencies

Sequencing by Synthesis (SBS)

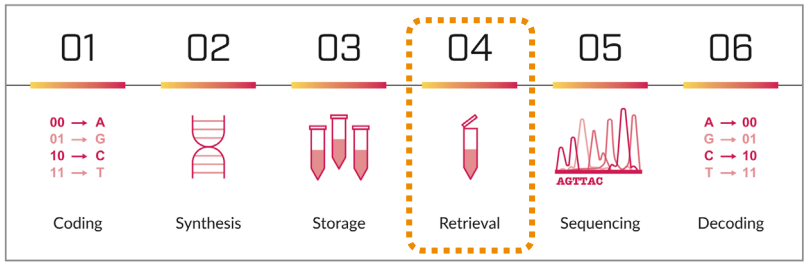
illumina®

Where does Sequencing fit in the Data Storage Pipeline?



* Graphic courtesy of DNA Data Storage Alliance Whitepaper (https://dnastoragealliance.org/publications/)

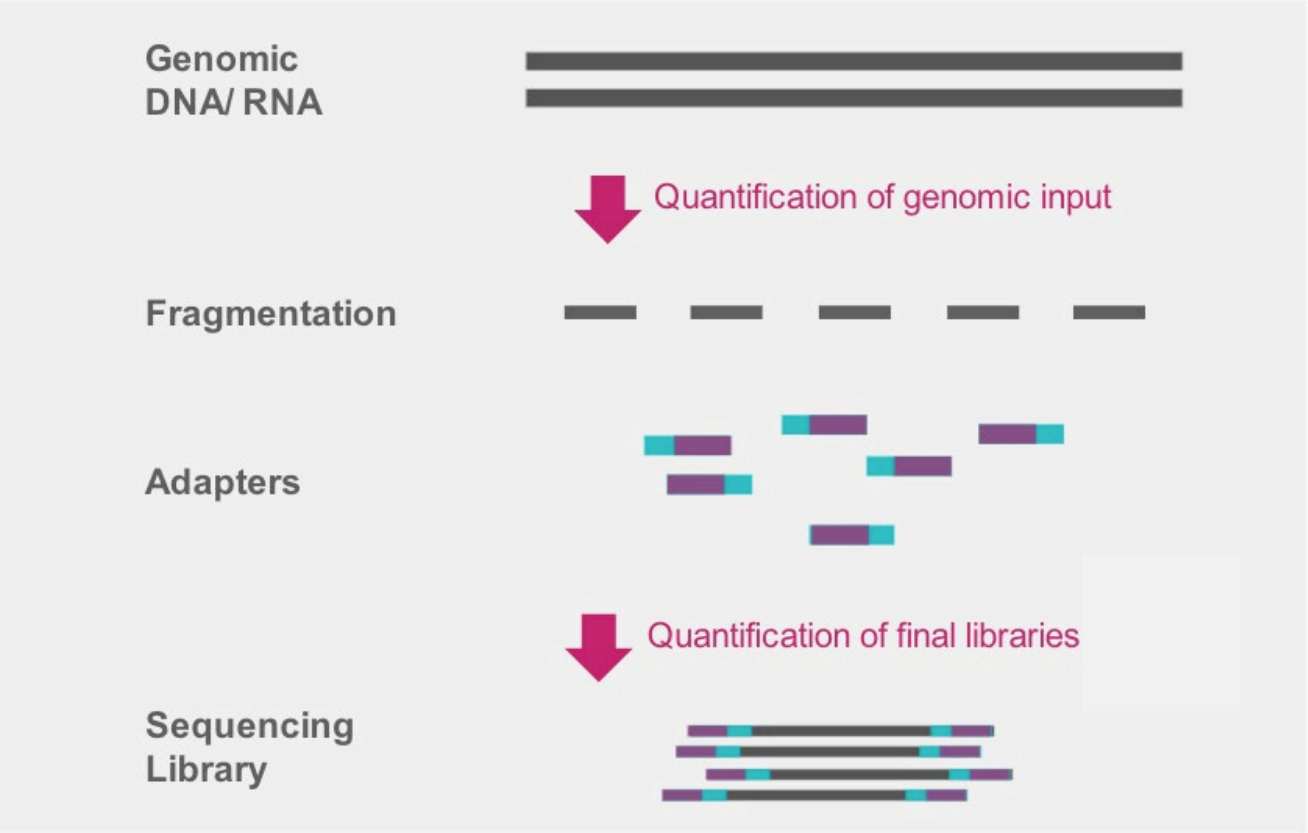
Library Preparation (LP)



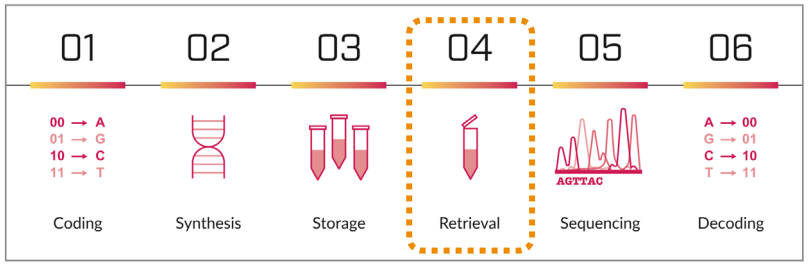
LP is an essential step in the sequencing workflow

Follows DNA (or RNA) extraction from a sample (e.g. blood, saliva)

Fragmentation generates short strands (100s of base pair long) of DNA that are suitable for SBS. Also referred to as DNA inserts.

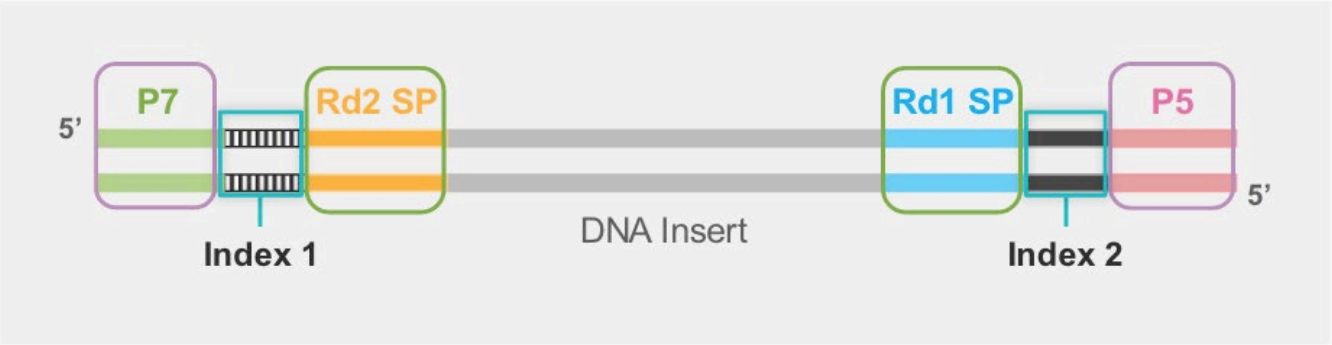


Library Preparation (LP)



LP involves adding three distinct blocks of oligos to both ends of the DNA insert.

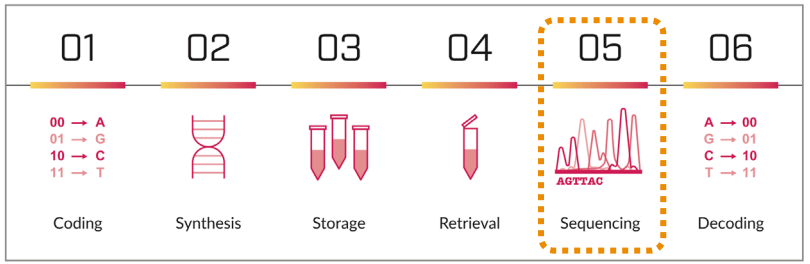
Each serves a critical role in the SBS process.

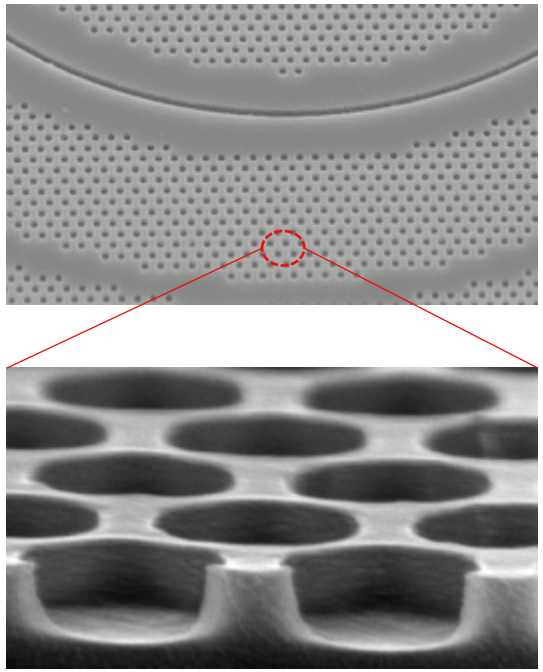


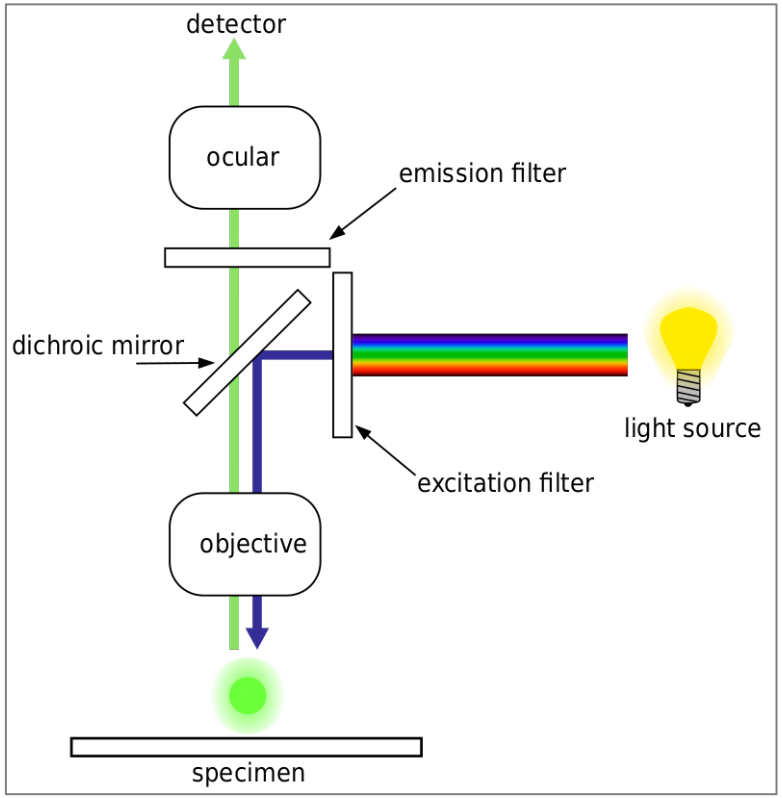
For Clustering	For Multiplexing	For Sequencing
Libraries must have P5 and P7 binding regions, which interact with oligos on the surface of the flow cell.	Libraries must have a unique index or barcode sequence for mixing samples.	Libraries must have binding regions for sequencing primers enable Read 1 and Read 2.

Sequencing

SBS relies on ultra-high throughput fluorescence microscopy to simultaneously measure individual base pairs, across billions of fragments of DNA placed in nanowells

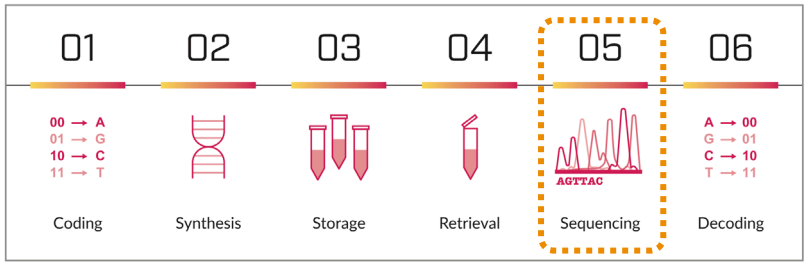






https://en.wikipedia.org/wiki/Fluorescence_microscope

Cluster Generation

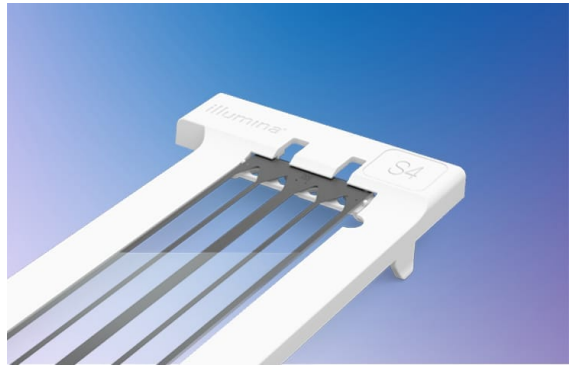


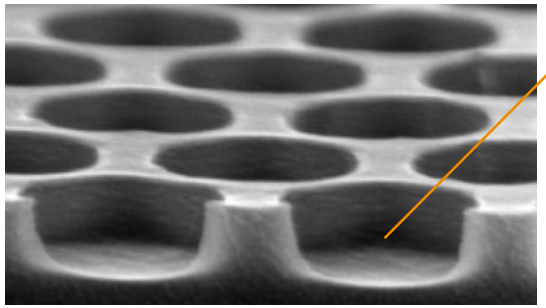
DNA Libraries are flowed into a flowcell to generate a clusters that are used for sequencing.

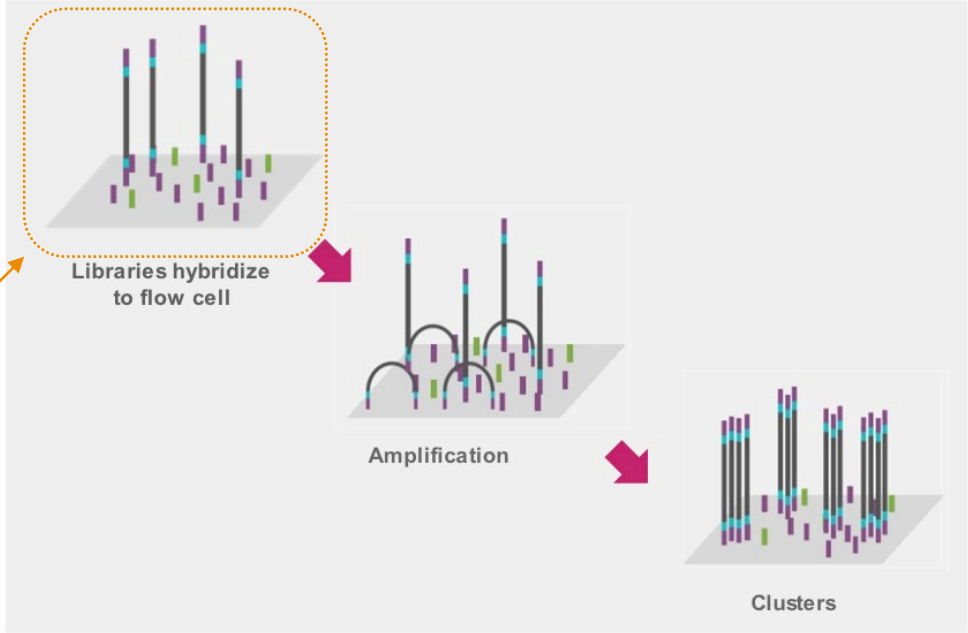
First, libraries hybridize (attach) to the flowcell surface, in the nanowells

Next, each strand is amplified to create a mono-clonal cluster.

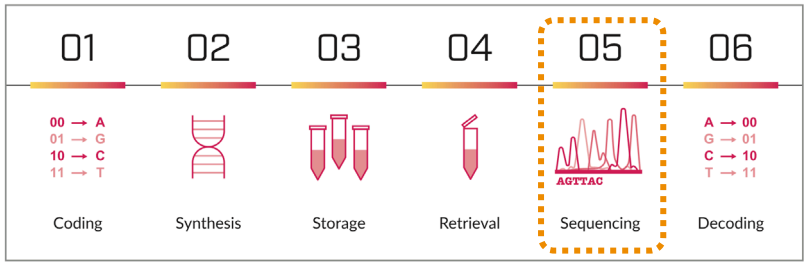
By creating a monoclonal cluster, the signal intensity during fluorescence imaging is increased



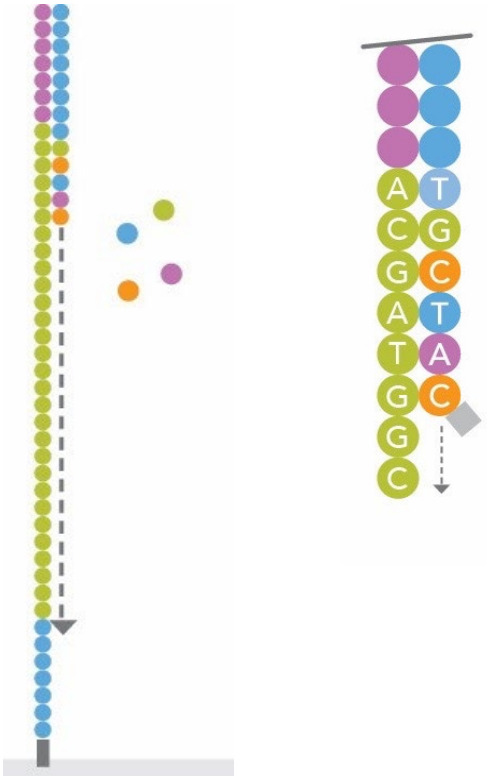




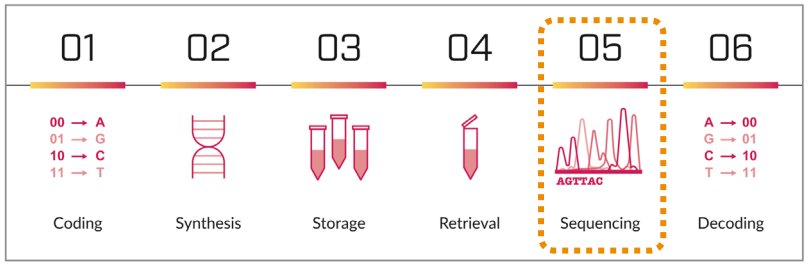
Sequencing

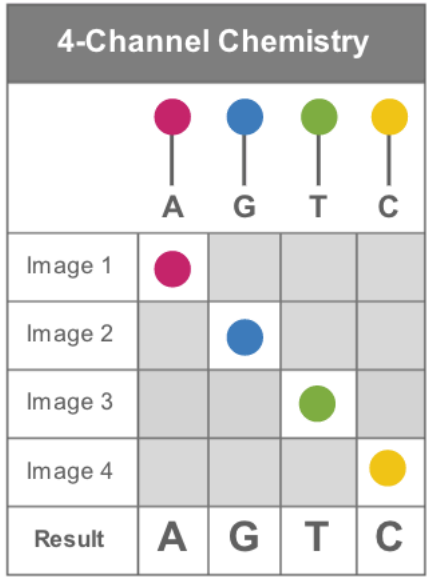


- Polymerase incorporates a nucleotide to create a second DNA strand copy
- The nucleotides are modified to include a fluorescence tag and a block
- The fluorescent tag allows the individual base pairs to be identified (coding scheme in next slide)
- The block prevents the polymerase from incorporating more than one base at a time

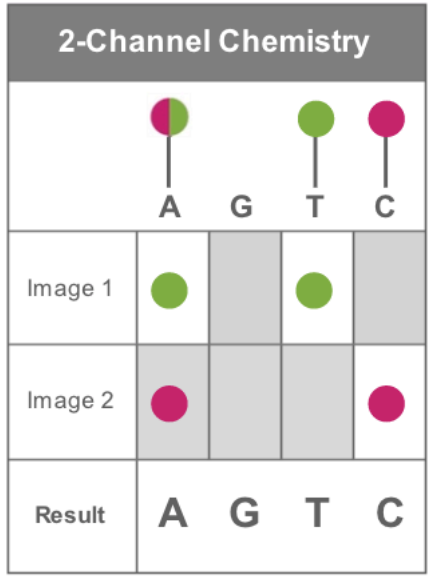


Sequencing Base Pair Coding Scheme

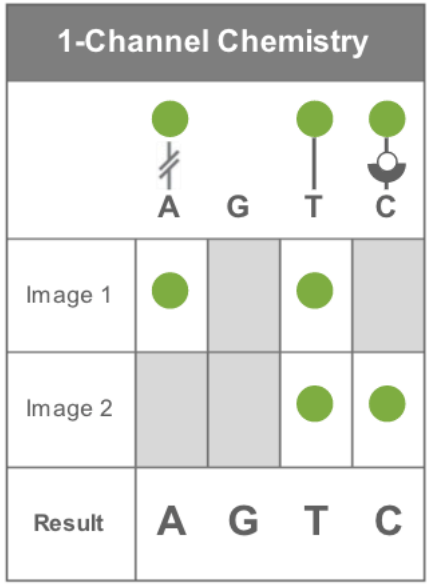




Uses four fluorescent dyes (one for each base), and four images per sequencing cycle.

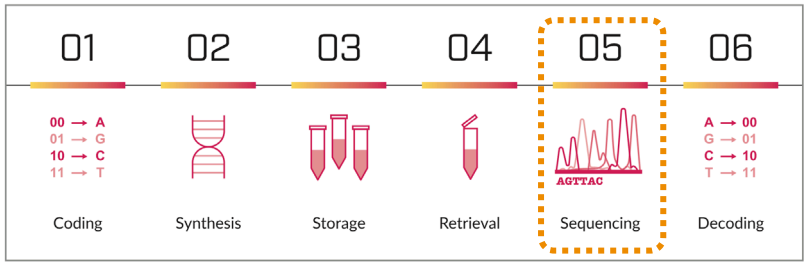


Uses two fluorescent dyes and two images per cycle to determine all four base calls.



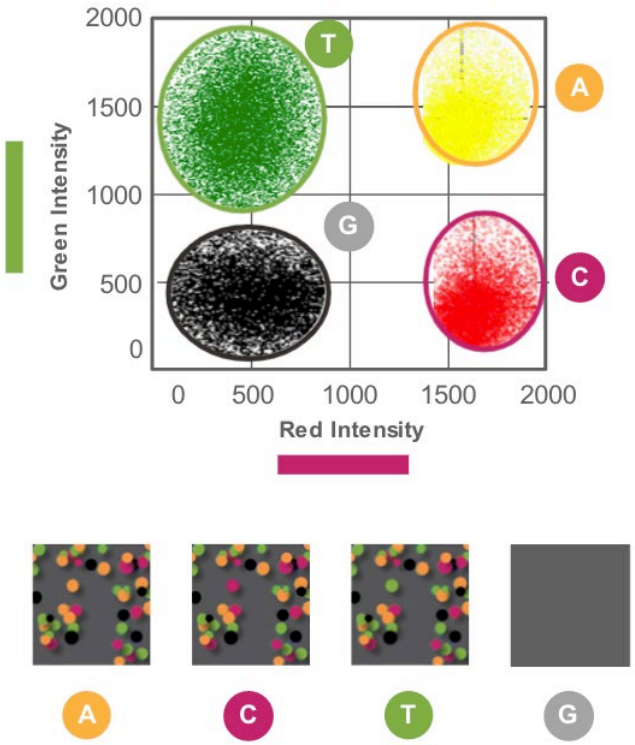
Uses CMOS technology to determine base calls using two images per cycle.

Two-Channel Sequencing



NovaSeq™ 6000 uses two channel red-green chemistry

Nucleotide	Channel
A	
C	
G	
T	

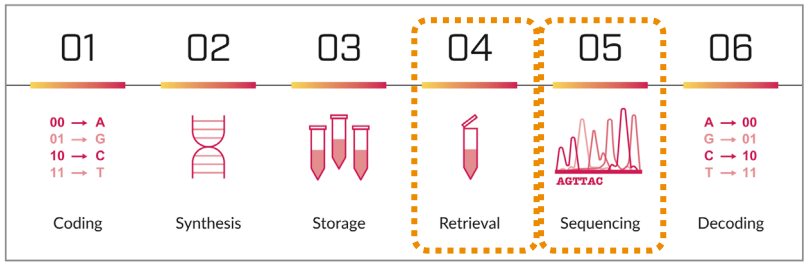


NovaSeq™ 6000

Highest throughput DNA Sequencer

Highest output : S4 Flow cell generates 3T bases in 44 hours
Instrument can run two S4 flow cells simultaneously

LP & Sequencing consumables

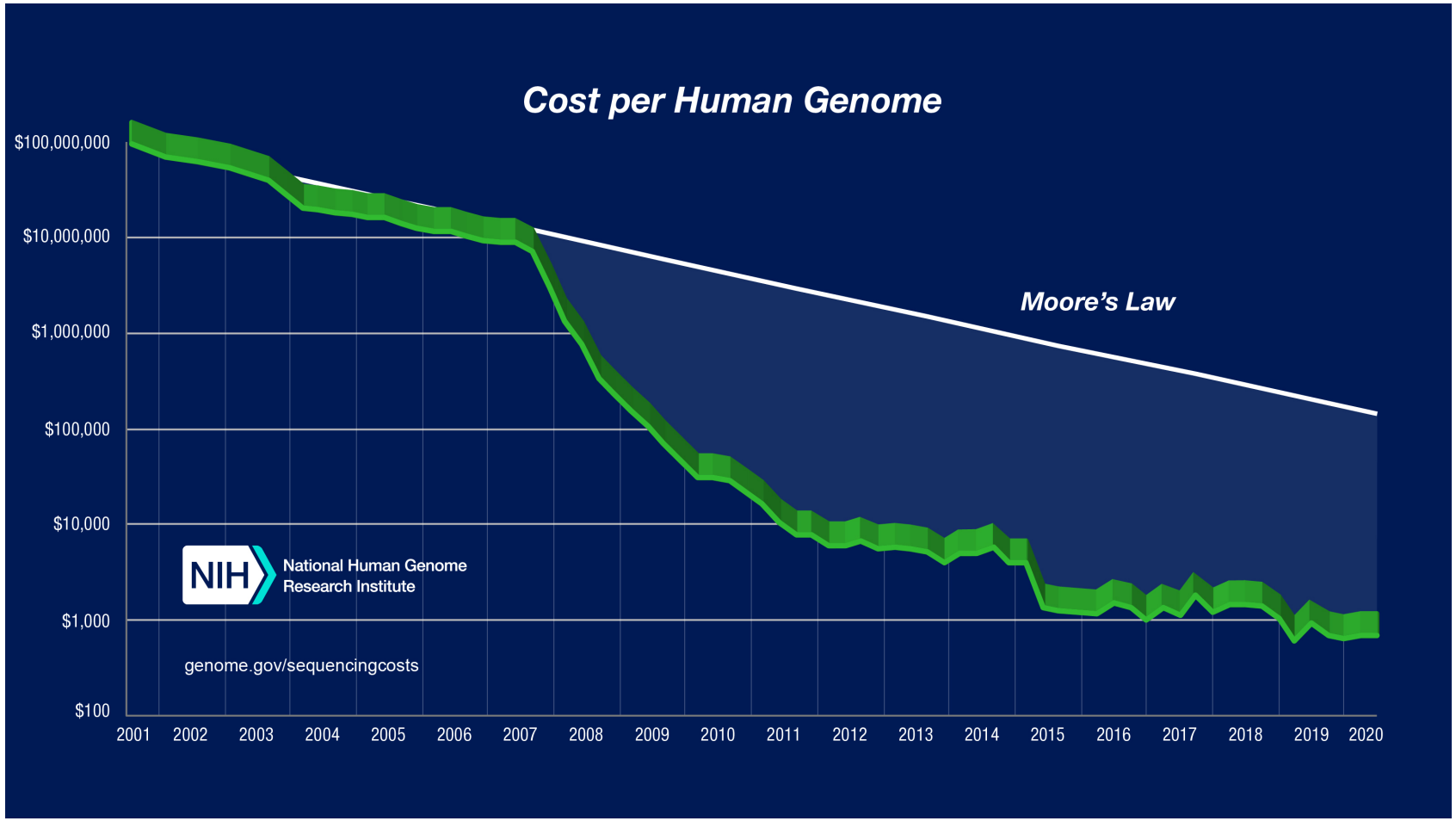


- Both LP and Sequencing require reagents (chemicals).
- Sequencing also requires a flow cell.
- Depending on sample type and quantity, plus the information required, a user will select the appropriate kit.
- Kits are shipped and stored frozen

Sequencing at Scale

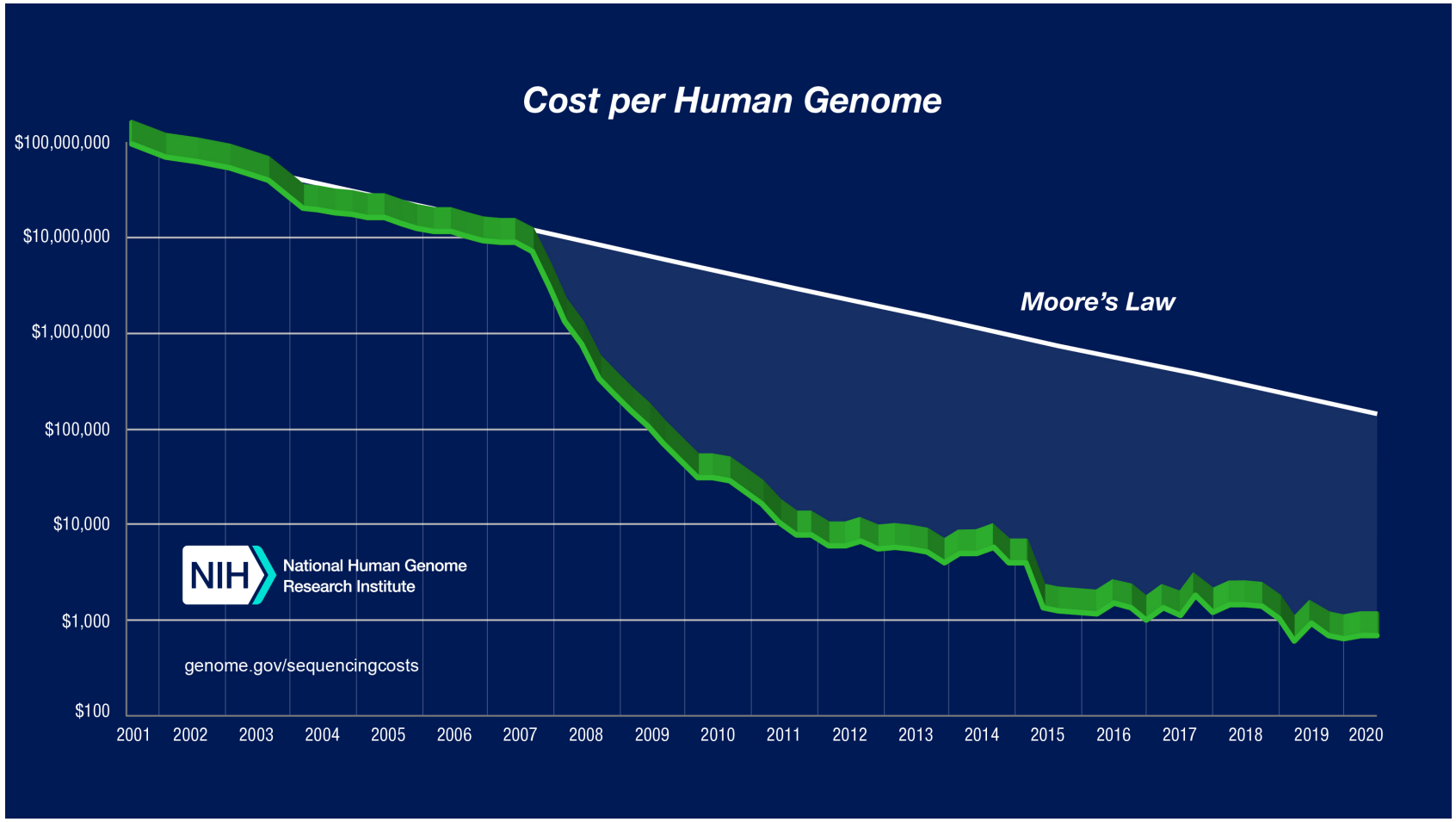
illumina®

Cost of the Human Genome



* https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (downloaded August 2021)

Cost of the Human Genome



NovaSeq 6000 v1.5 today offers a $600- Genome

1 Genome = 120G (~30x coverage) ($5-/Gb)

Plan to deliver then $100- genome ($0.8/Gb)

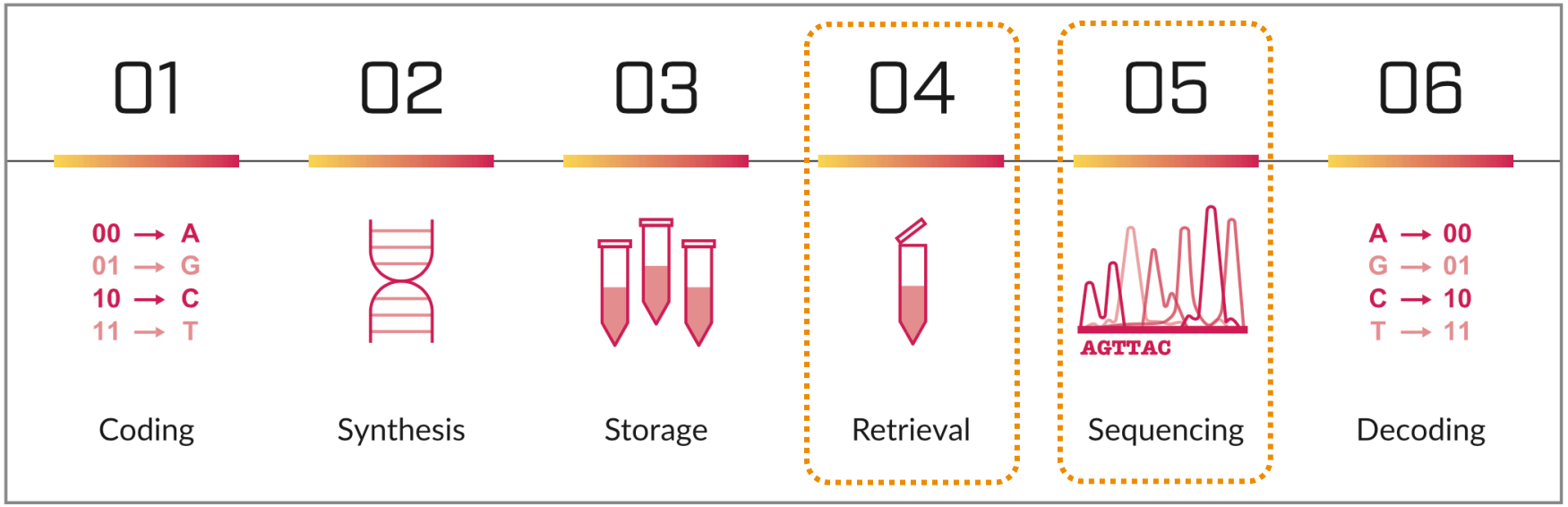
* https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (downloaded August 2021)

Illumina's High Throughput Sequencing lab in Hinxton, UK

Current capacity of ~3900 genomes per month. ~16T bases per day of data being generated*

*Genomes per month averaged over 30 days; assumes 120G per genome

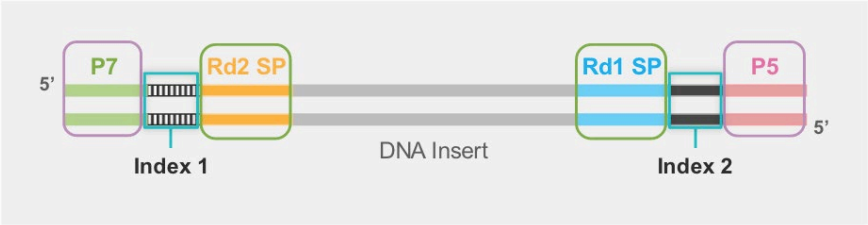
Using DNA for Data Storage presents unique opportunities for Sequencing



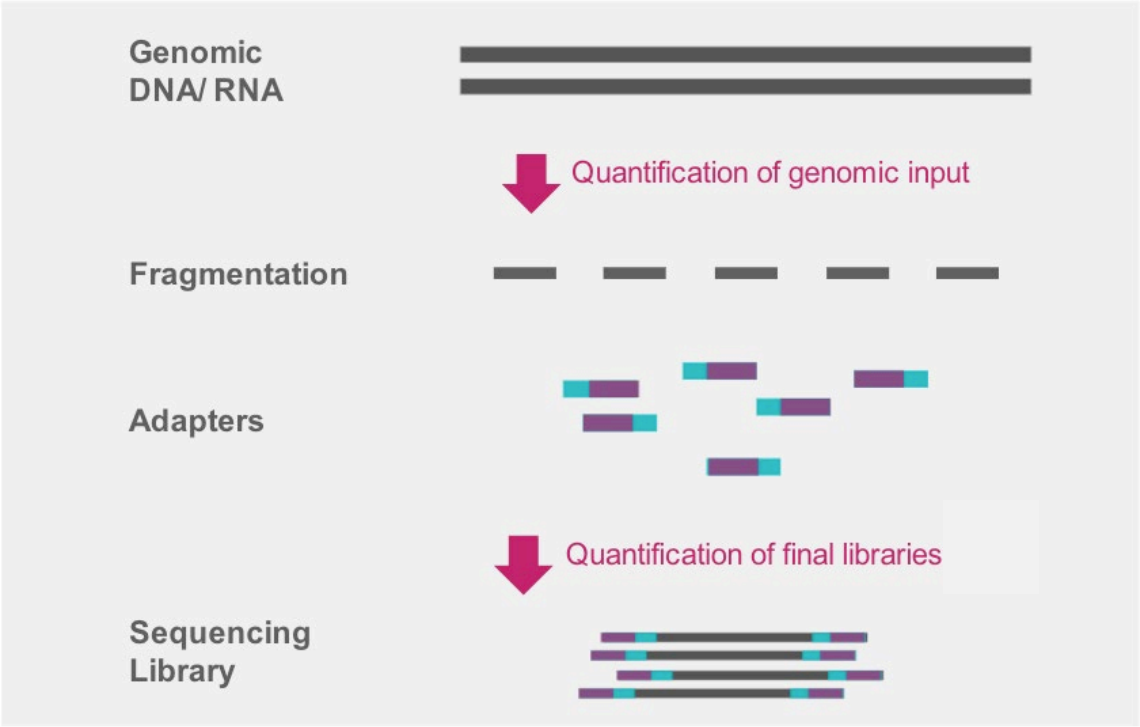
Time-to-data – Simplifying data structure and LP

Data libraries most likely will be written at a length that is readily compatible with SBS – this avoids the fragmentation step.

Having a data library written with primers and indices already incorporated would also save time and cost to retrieve data.

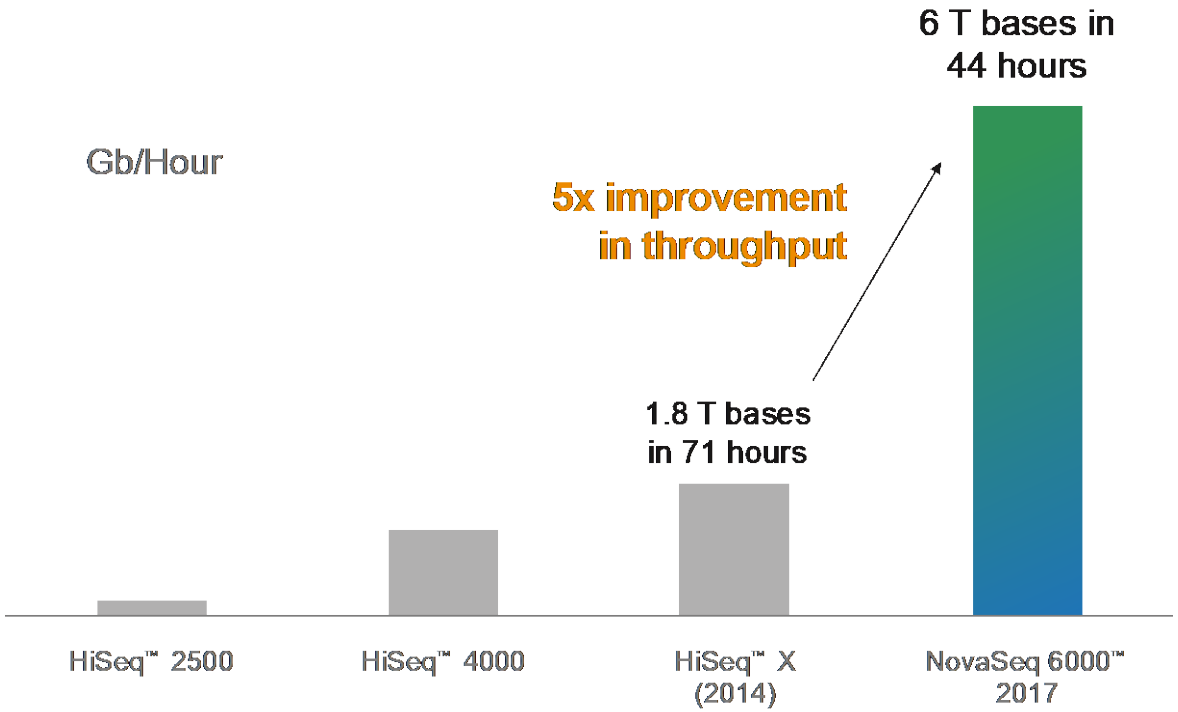


For Clustering	For Multiplexing	For Sequencing
Libraries must have P5 and P7 binding regions, which interact with oligos on the surface of the flow cell.	Libraries must have a unique index or barcode sequence for mixing samples.	Libraries must have binding regions for sequencing primers enable Read 1 and Read 2.



Time-to-data – Increase instrument throughput

Sequencers are configured to run in batches; the larger the batch, the lower the cost per G.



- Higher Density Flow Cell
- Custom Camera and Laser
- New Surface Chemistry & New Dyes
- New Compute Infrastructure

Technologies being developed are expected to offer 2x faster run times and 2x longer read lengths.*

* Illumina at JPM 2021

Reduce cost per Tb - Increase flow cell density

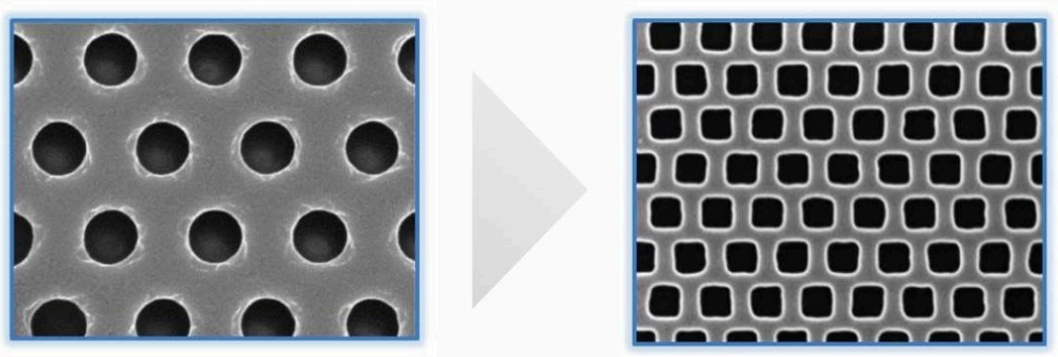
Flow cell fabrication is based on high-volume capable tools from the semiconductor industry

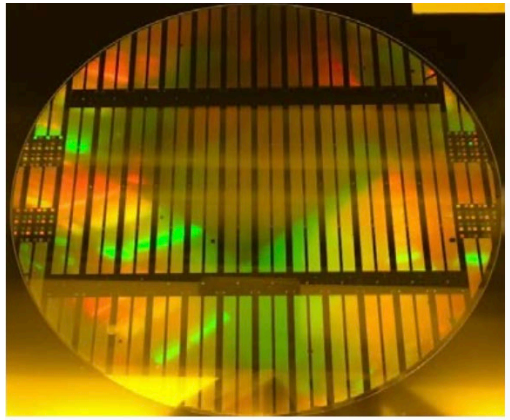
Today, Illumina is scaling production to 300mm diameter glass wafers – doubling the number of flow cells per wafer

Illumina is also increasing flow cell nanowell density by 5x

These innovations are expected to reduce flow cell COGS by up to 90%

Flow cell micrographs of same unit area





Summary

- Sequencing by Synthesis (SBS) based platforms offer the highest throughput and lowest cost sequencing in the market today
- DNA-based data storage has some unique requirements, which provide an opportunity to adapt sequencing to meet the needs of this emerging application
- Changes to the library prep are possible to save time and cost when retrieving data
- With a long history of innovation, Illumina expects to continue to deliver higher-throughput and lower cost-per-base sequencing platforms, to meet the needs of the DNA-based data storage application

Thank you!

Craig Ciesla, Ph.D.

VP – Head of Advanced Platforms and Devices
Illumina Research & Development

illumina®