Introduction to Nucleic Acids Sequencing Technologies

November 4, 2024

Professor Qasim Ayub

Director Monash University Malaysia Genomics Platform (MUMGP)

Deputy Head of School (Research)

qasim.ayub@monash.edu

wellcome connecting science

MONASH University | MALAYSIA

Genomics Platform

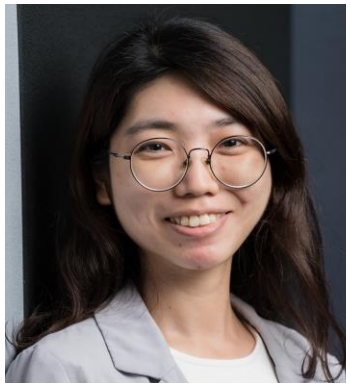
bsi ISO 9001 Quality Management Systems CERTIFIED

The Malaysian Team
School of Science

Dr. Wee Wei Yee
Senior Lecturer

Prof. Qasim Ayub
Director, Monash University Malaysia Genomics Platform (MUMGP)
Deputy Head of School (Research)

Dr. Aswini Leela Loganathan
Research Fellow



Ms. Lim Shu Yong
Assistant Manager

Dr. Hong Leong Cheah
Research Fellow

Ms. Ling Fong Yoke
Senior Technical Officer (Research)

Health and Safety

Your RESPONSIBILITIES towards Occupational Health and Safety (OHS).

- Protect own health and safety together with the health and safety of any other persons.
- Comply with health and safety direction given by School of Science staff.
- Report health and safety related hazards and incidents immediately to our staff.

General Laboratory Rules

- **FOOD AND BEVERAGES ARE NOT ALLOWED** in the laboratory.
- **BE AWARE** of any special health or safety hazards posed by the chemicals with which you are working.
- **CLEAN YOUR WORK AREA AND YOUR HANDS** after completing laboratory work.
- During **EMERGENCY EVACUATION**, follow our staff's instructions.

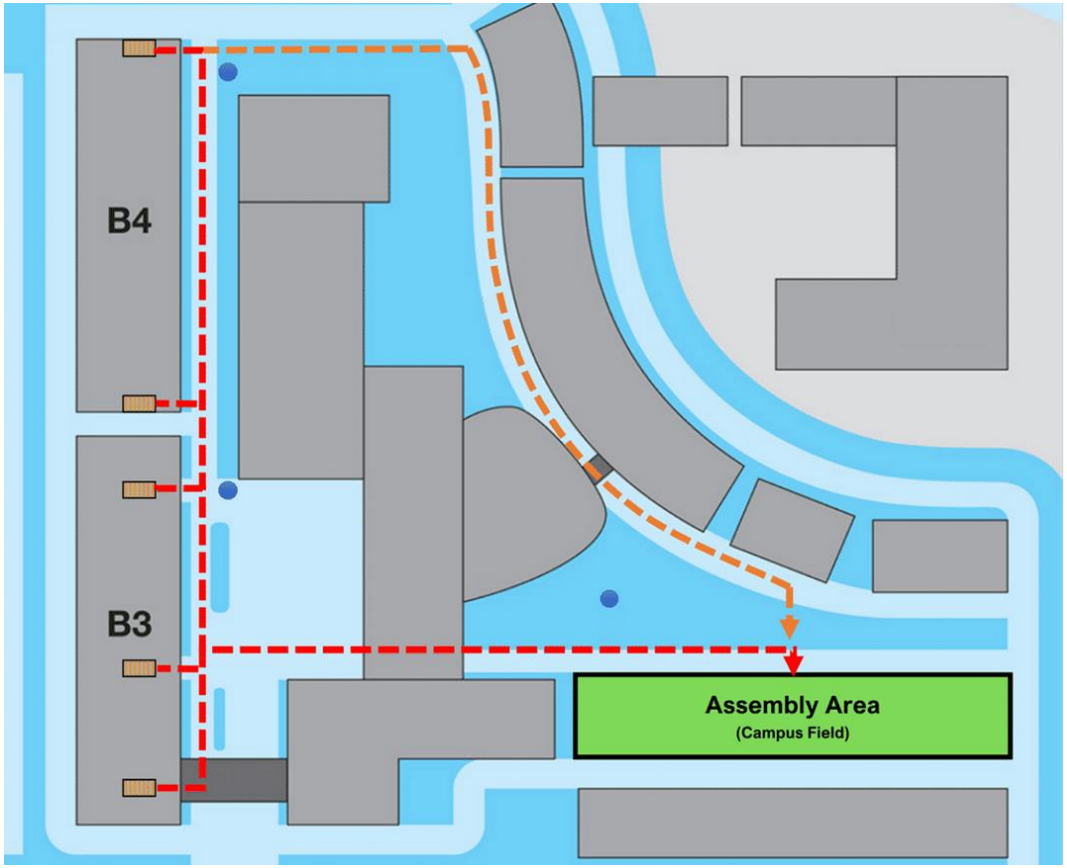
PPE in teaching & research laboratories

Evacuation Routes

MONASH University

School of Science Main Building

- **Building 3**
- **Building 4**



Legend	
	Main Emergency Route
	Alternative Emergency Route
	Evacuation Controller Position

On Campus Emergency Assembly Area

Follow instructions of Incident Commander, queue up to do headcount reporting.

Learning Outcomes

- Understand various “Next Generation Sequencing - NGS” or massively parallel, high-throughput sequencing technologies.
- Explain differences between short- and long-read sequencing.
- Outline of advantages and disadvantages of each sequencing platform.
- Evaluate sequencing data.
- Design your own genomic experiments.

Outline

- Introduction and historical background.
- “Next generation” or massively parallel, high-throughput sequencing technologies.
- Outline of advantages and disadvantages.
- Future developments.

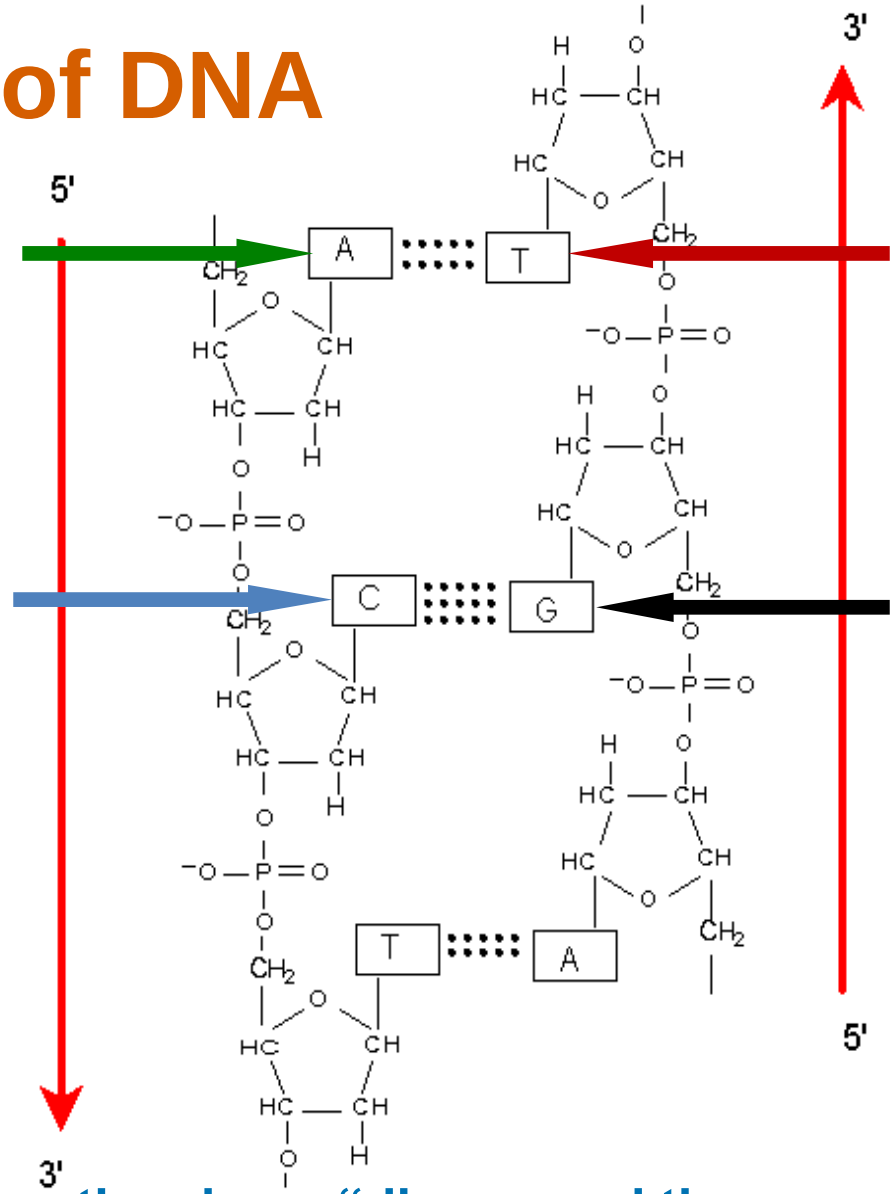
What These Technologies Do?

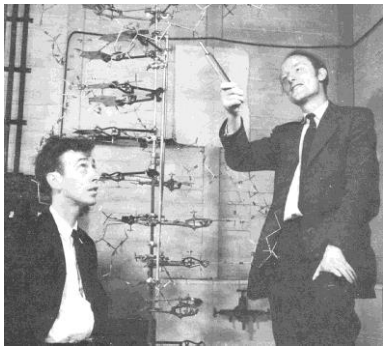
- They characterise the order of the nucleic acids (chemical bases) in the DNA or RNA molecules that make up genomes of organisms.
- Useful to study the genomes or total genetic make up of an individual, organism or species. In almost all species DNA constitutes the hereditary genetic material in the genome.
- The genome provides a set of instructions or recipe for making the enzymes, cells, tissues, organs and body.
- It includes direct analysis of nucleic acids and indirectly of proteins.

Determining the Sequence

AGGAAAGAGCACCCAGGACTGTCATATGGAAGAAAGACAGGACTGCAACTCACCCTTCACAAAATGAGGACCAGACACAGCT
GATGGTATGAGTTGATGCAGGTGTGTGGAGCCTCAACATCCTGCTCCCCTCCTACTACACATGGTTAAGGCCTGTTGCTCTGTC
TCCAGGTTCACACTCTCTGCACTACCTCTTCATGGGTGCCTCAGAGCAGGACCTTGGTCTTTCCTTGTTTGAAGCTTTGGGCTA
CGTGGATGACCAGCTGTTCGTGTTCTATGATCATGAGAGTCGCCGTGTGGAGCCCCGAACTCCATGGGTTTCCAGTAGAATTT
CAAGCCAGATGTGGCTGCAGCTGAGTCAGAGTCTGAAAGGGTGGGATCACATGTTCACTGTTGACTTCTGGACTATTATGGAA
AATCACAACCACAGCAAGGGTATGTGGAGAGGGGGCCTCACCTTCCTGAGGTTGTCAGAGCTTTTCATCTTTTCATGCATCTTG
AAGGAAACAGCTGGAAGTCTGAGGTCTTGTGGGAGCAGGGAAGAGGGAAGGAATTTGCTTCCTGAGATCATTTGGTCCTTGG
GATGGTGGAAATAGGGACCTATTCCTTTGGTTGCAGTTAACAAGGCTGGGGATTTTTCCAGAGTCCCACACCCTGCAGGTCAT
CCTGGGCTGTGAAATGCAAGAAGACAACAGTACCGAGGGCTACTGGAAGTACGGGTATGATGGGCAGGACCACCTTGAATTCT
GCCCTGACACACTGGATTGGAGAGCAGCAGAACCCAGGGCCTGGCCCACCAAGCTGGAGTGGGAAAGGCACAAGATTCGGG
CCAGGCAGAACAGGGCCTACCTGGAGAGGGACTGCCCTGCACAGCTGCAGCAGTTGCTGGAGCTGGGGAGAGGTGTTTTG
GACCAACAAGGTATGGTGGAAACACACTTCTGCCCCTATACTCTAGTGGCAGAGTGGAGGAGGTTGCAGGGCACGGAATCCC
TGGTTGGAGTTTCAGAGGTGGCTGAGGCTGTGTGCCTCTCCAAATTCTGGGAAGGGACTTTCTCAATCCTAGAGTCTCTACCT
TATAATTGAGATGTATGAGACAGCCACAAGTCATGGGTTTAATTTCTTTTCTCCATGCATATGGCTCAAAGGGAAGTGTCTATGGC
CCTTGCTTTTTATTTAACCAATAATCTTTTGTATATTTATACCTGTTAAAAATTCAGAAATGTCAAGGCCGGGCACGGTGGCTCACC
CCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGTCACAAGGTCAGGAGTTTGAGACCAGCCTGACCAACATGGTGA
AACCCGTCTCTAAAAAAATACAAAAATTAGCTGGTCACAGTCATGCGCACCTGTAGTCCCAGCTAATTGGAAGGCTGAGGCAGG
AGCATCGCTTGAACCTGGGAAGCGGAAGTTGCACTGAGCCAAGATCGCGCCACTGCACTCCAGCCTAGGCAGCAGAGTGAG
ACTCCATCTTAAAAAAAAAAAAAAAAAAAAAAAGAGAATTCAGAGATCTCAGCTATCATATGAATACCAGGACAAAATATCAAGTG
AGGCCACTTATCAGAGTAGAAGAATCCTTTAGGTTAAAAGTTTCTTTCATAGAACATAGCAATAATCACTGAAGCTACCTATCTTAC
AAGTCCGCTTCTTATAACAATGCCTCCTAGGTTGACCCAGGTGAAACTGACCATCTGTATTCAATCATTTTCAATGCACATAAAGG
GCAATTTTATCTATCAGAACAAAGAACATGGGTAACAGATATGTATATTTACATGTGAGGAGAACAAGCTGATCTGACTGCTCTCC
AAGTGACACTGTGTTAGAGTCCAATCTTAGGACACAAAATGGTGTCTCTCCTGTAGCTTGTTTTTTTCTGAAAAGGGTATTTCCT
TCCTCCAACCTATAGAAGGAAGTGAAAGTTCCAGTCTTCCTGGCAAGGGTAAACAGATCCCCTCTCCTCATCCTTCCTCTTTCC
TGTCAAGTGCATGCCAAGTAGGAGAGTATAAGGCATACTGGGAGATTAGAAATAATTACTGTACCTTAACCCTGAGTTTGCGTAG
CTATCACTCACCAATTATGCATTTCTACCCCCTGAACATCTGTGGTGTAGGGAAAAGAGAATCAGAAAGAAGCCAGCTCATACAG
AGTCCAAGGGTCTTTTGGGATATTGGGTTATGATCACTGGGGTGTCATTGAAGGATCCTAAGAAAGGAGGACCACGATCTCCCT
TATATGGTGAATGTGTTGTTAAGAAGTTAGATGAGAGGTGAGGAGACCAGTTAGAAAGCCAATAAGCATTTCCAGATGAGAGATA
ATGGTTCTTGAAATCCAATAGTGCCCAGGTCTAAATTGAGATGGGTGAATGAGGAAAATAAGGAAGAGAGAAGAGGCAAGATG
GTGCCTAGGTTTGTGATGCCTCTTTCCTGGGTCTCTTGTCTCCACAGGAGGAGCCATGGGGCACTACGTCTTAGCTGAACGTG
AGTGACACGCAGCCTGCAGACTCACTGTGGGAAGGAGACAAAACTAGAGACTCAAAGAGGGAGTGCATTTATGAGCTCTTCAT

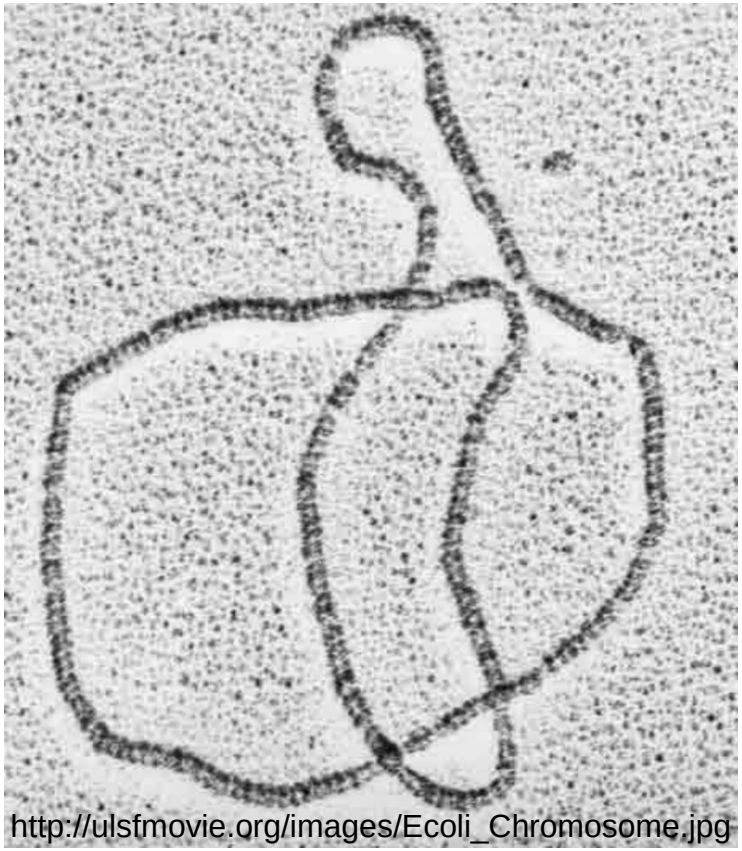
Structure of DNA





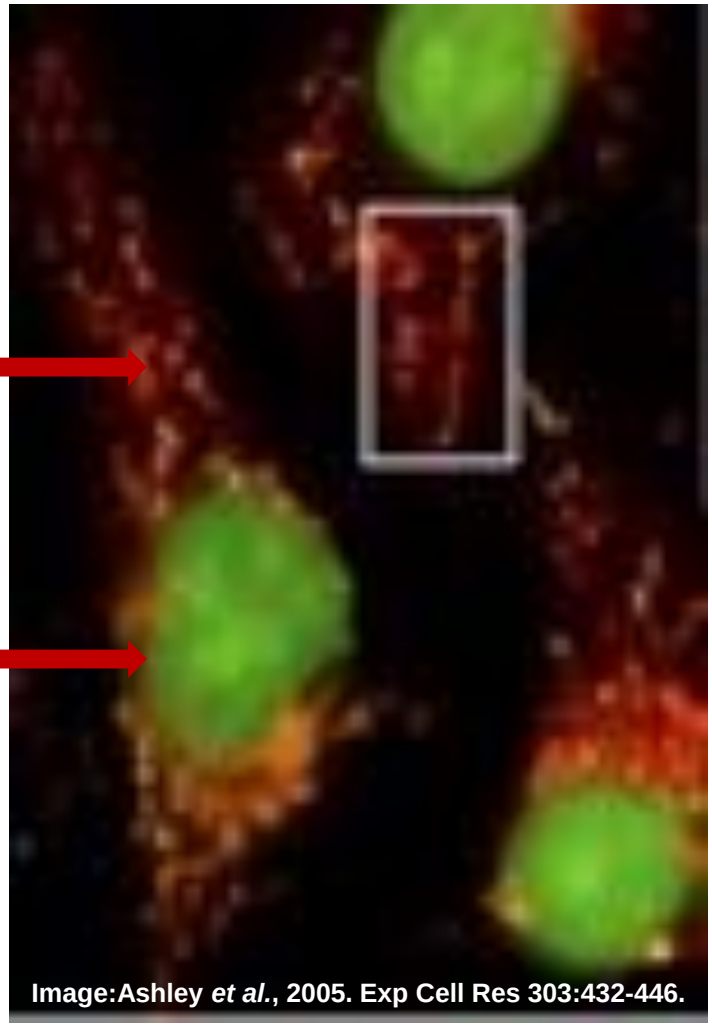
Announce they have "discovered the secret of life"
The Eagle, Benet Street on 28th February 1953.

Where is the DNA Found?



http://ulsfmovie.org/images/Ecoli_Chromosome.jpg

E. coli chromosome

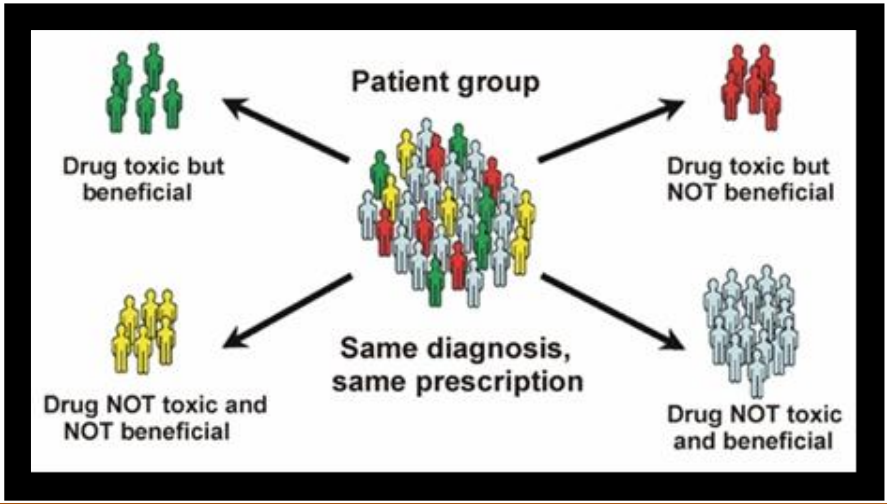


Human ECV60 cell line

Why Sequence DNA?

- Cataloging biodiversity.
- Understanding biology and evolution of organism or species.
- Marker assisted breeding.
- Promise of Personalized Medicine.
  - Genetic diagnosis and accurate disease prediction.
  - Genotype-phenotype association studies.
  - Disease monitoring.
  - Personalized treatment.
  - Forensic identification.
  - Population genomics,
  - Gene environment interactions.

https://www.earthbiogenome.org/



History of Sequencing Genomes

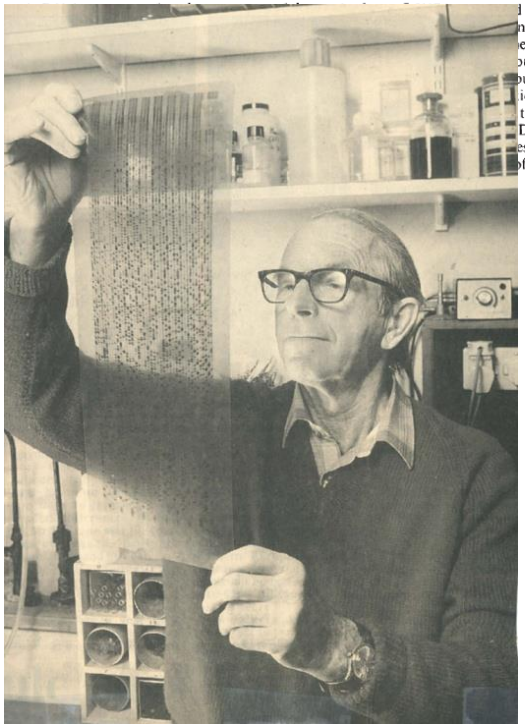
Nature Vol. 265 February 24 1977

articles

Nucleotide sequence of bacteriophage ΦX174 DNA

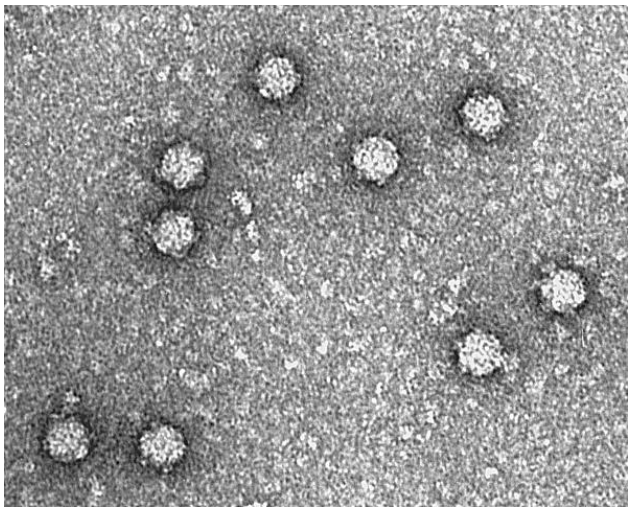
F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown†, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III‡, P. M. Slocombe§ & M. Smith‖

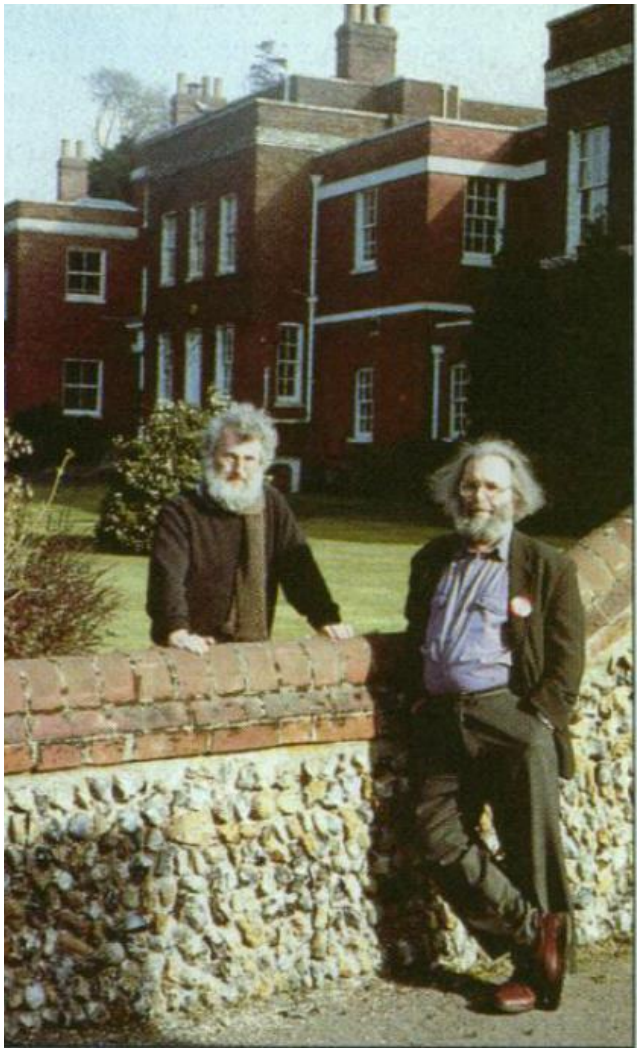
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK



DNA of ΦX has the same sequence as the mRNA and, in conditions, will bind ribosomes so that a protected ent can be isolated and sequenced. Only one major site ound. By comparison with the amino acid sequence data it ound that this ribosome binding site sequence coded for the ion of the gene *G* protein[15] (positions 2,362–2,413).

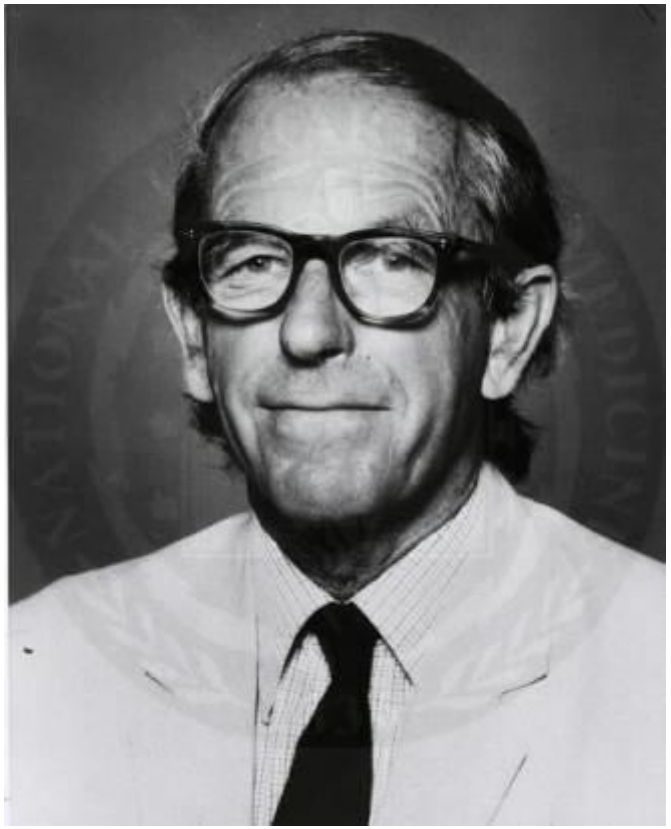
this stage sequencing techniques using primed synthesis DNA polymerase were being developed[16] and Schott[17] esised a decanucleotide with a sequence complementary to f the ribosome binding site. This was used to prime into





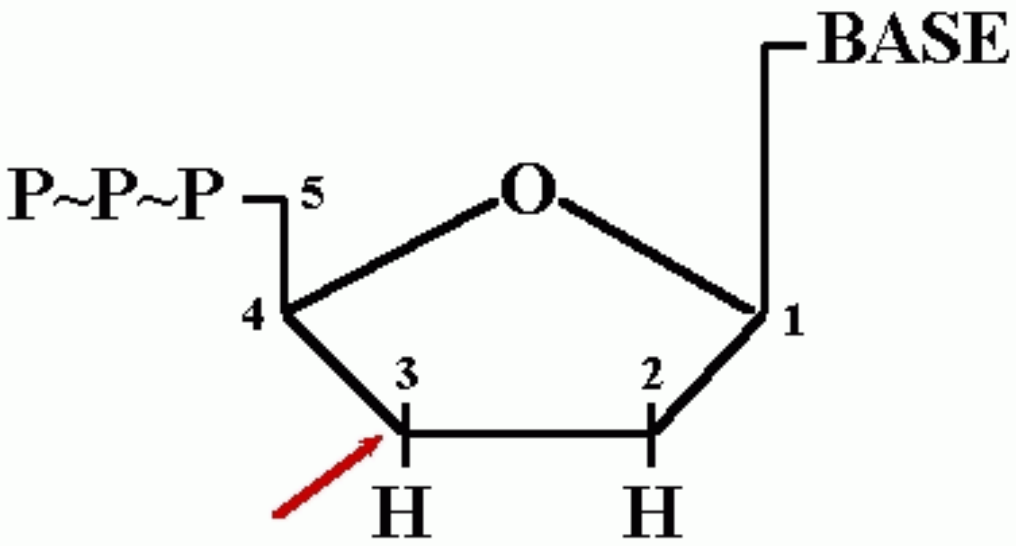
Frederick Sanger

- Developed the DNA sequencing by chain termination method.
- First Nobel Prize 1958:
  - Complete amino acid sequence of insulin.
- Second Nobel Prize 1980:
  - DNA sequencing.



Uses Dideoxy Nucleotides

- Lack an -OH group at the 3' carbon position.
- Cannot add another nucleoside at that position, hence preventing further DNA synthesis.



All Possible Terminations

DNA Polymerase reads the template strand and synthesizes a new second strand to match:

IF 5% of the T nucleotides are actually dideoxy T, then each strand will terminate when it gets a ddT on its growing end:

5' - TACGCGGTAACGGTATGTTCGACCGTTTAGCTACCGAT•

5' - TACGCGGTAACGGTATGTTCGACCGTTTAGCT•

5' - TACGCGGTAACGGTATGTTCGACCGTTT•

5' - TACGCGGTAACGGTATGTTCGACCGTT•

5' - TACGCGGTAACGGTATGTTCGACCGT•

5' - TACGCGGTAACGGTATGTT•

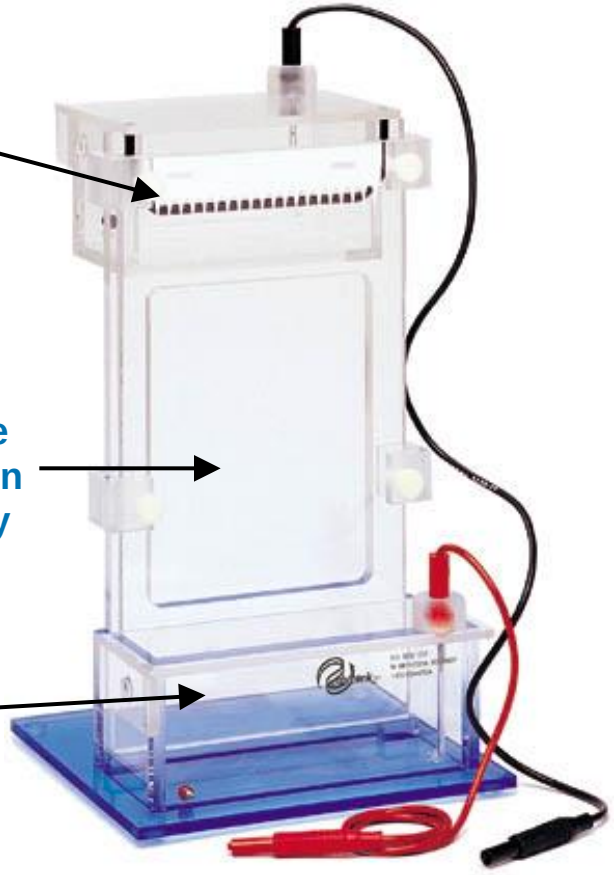
5' - TACGCGGTAACGGTATGT•

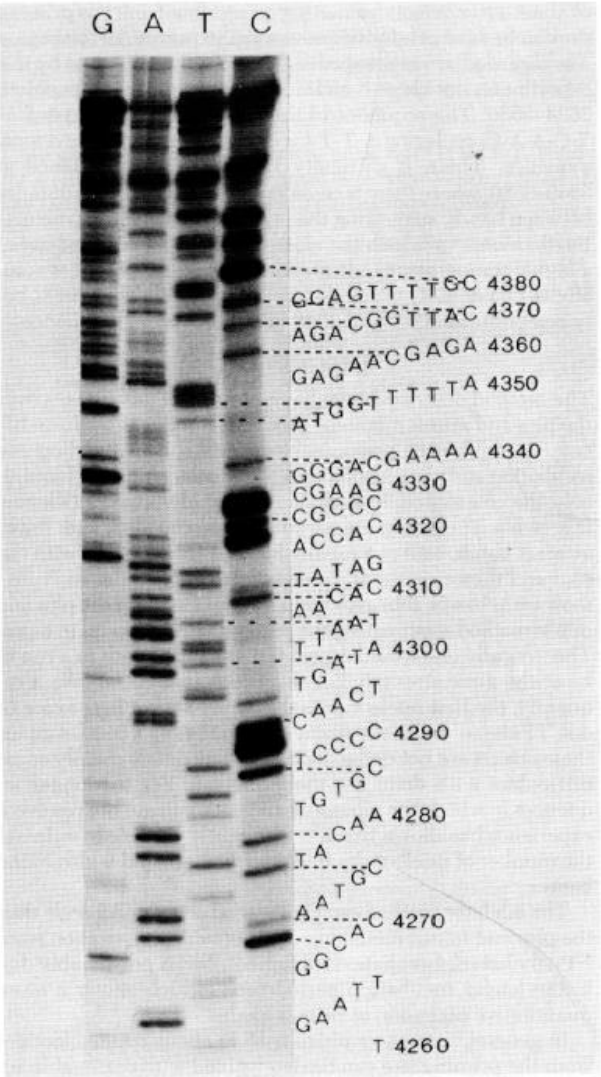
5' - TACGCGGTAACGGTAT•

5' - TACGCGGTAACGGT•

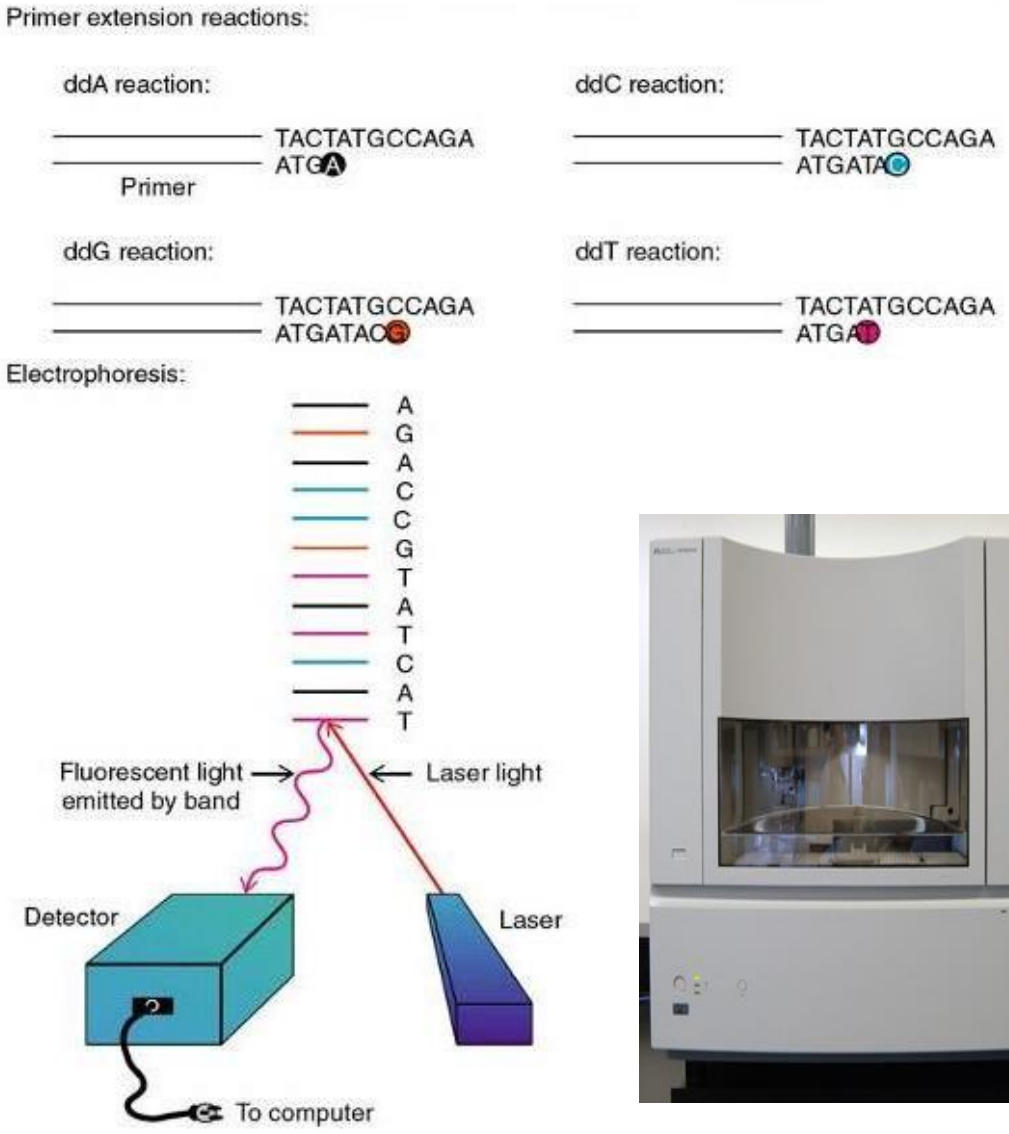
5' - TACGCGT•

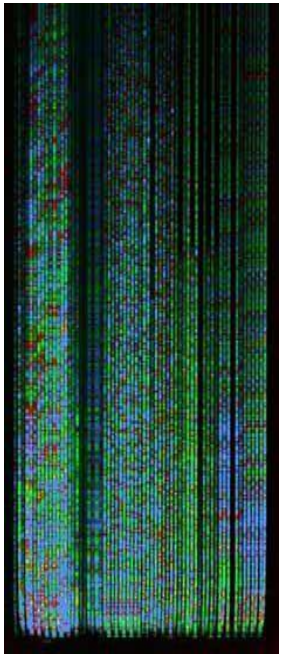
How it was done from 1980 – 1990s?



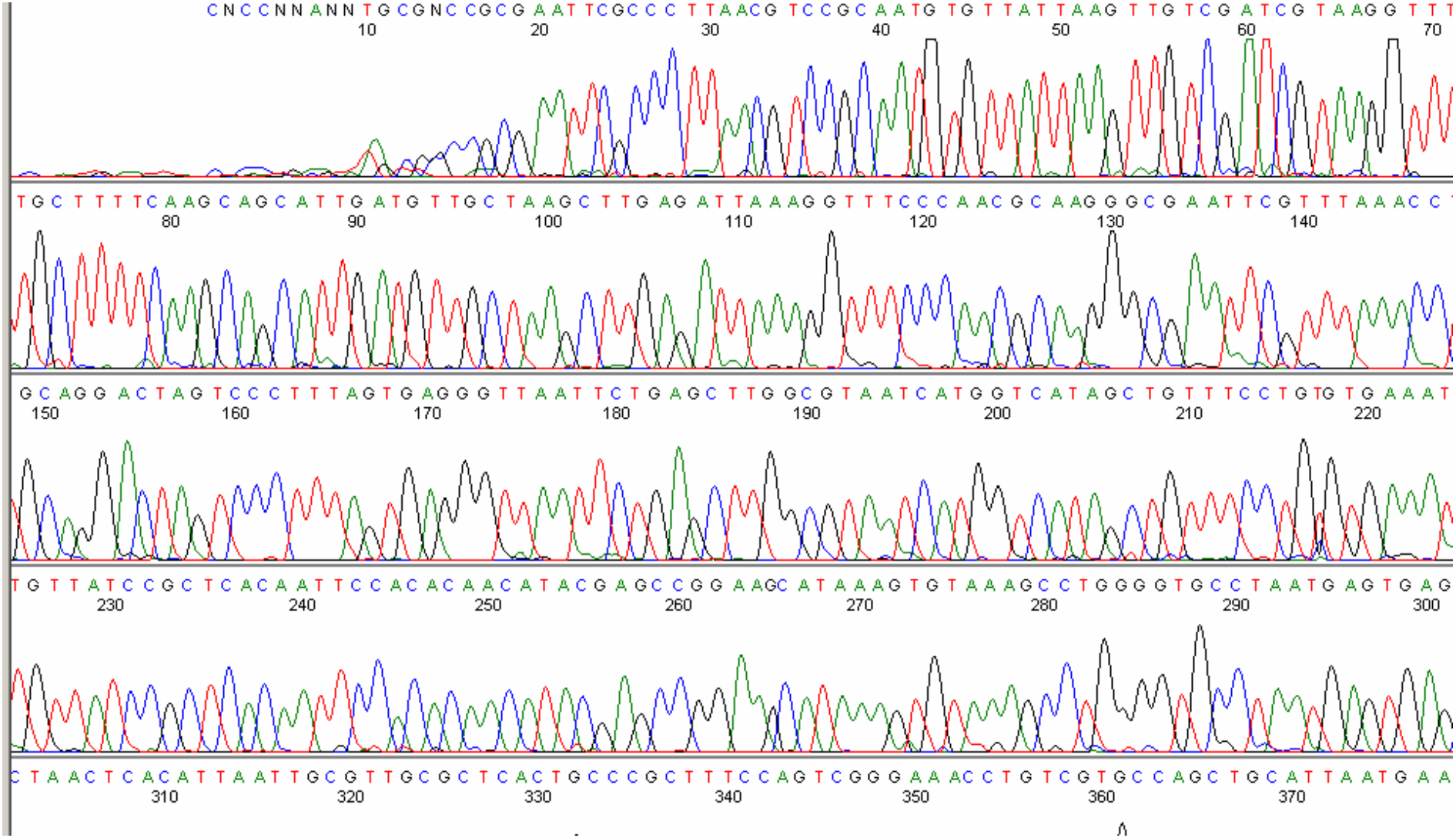


Development of Fluorescent Terminators





Electropherograms DNA Sequence Files

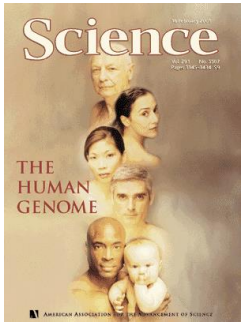


Dideoxy or Sanger Sequencing

https://www.yourgenome.org/video/dna-sequencing

Sequenced Genomes

➢ 1995	*Haemophilus influenzae*	1.8 Mb
➢ 1996	*Saccharomyces cerevisiae*	12.0 Mb
➢ 1998	*Caenorhabditis elegans*	100.0 Mb
➢ 1999	*Drosophila melanogaster*	125.0 Mb
➢ 2000	*Arabidopsis thaliana*	115.0 Mb
➢ 2001	*Homo sapiens* (draft)	3.0 Gb
➢ 2002	Mus musculus	2.6 Gb
➢ 2004	*Homo sapiens* (“finished”)	3.2 Gb

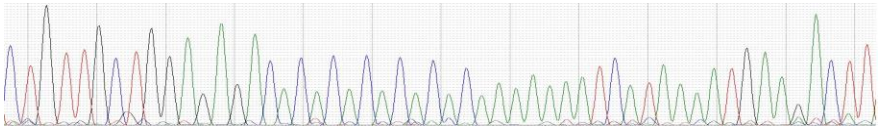


Next Generation Sequencing (NGS)

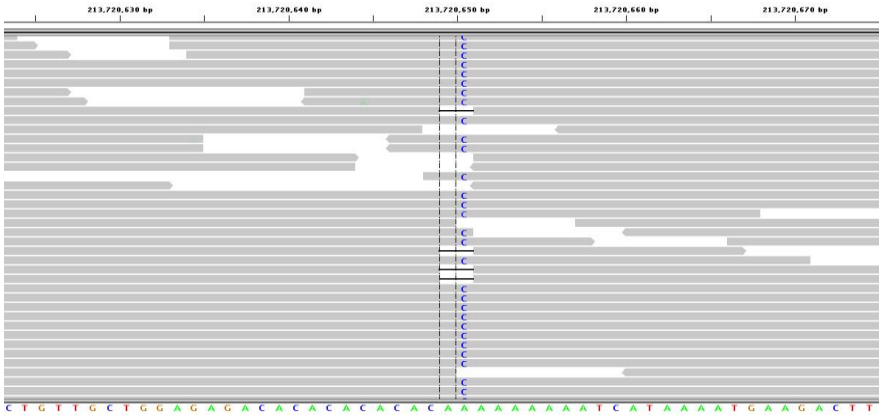
- From 2004 – present.
- Is massively parallel high-throughput sequencing.
- Not limited to few reactions per run.

Sequencing Revolution

2008

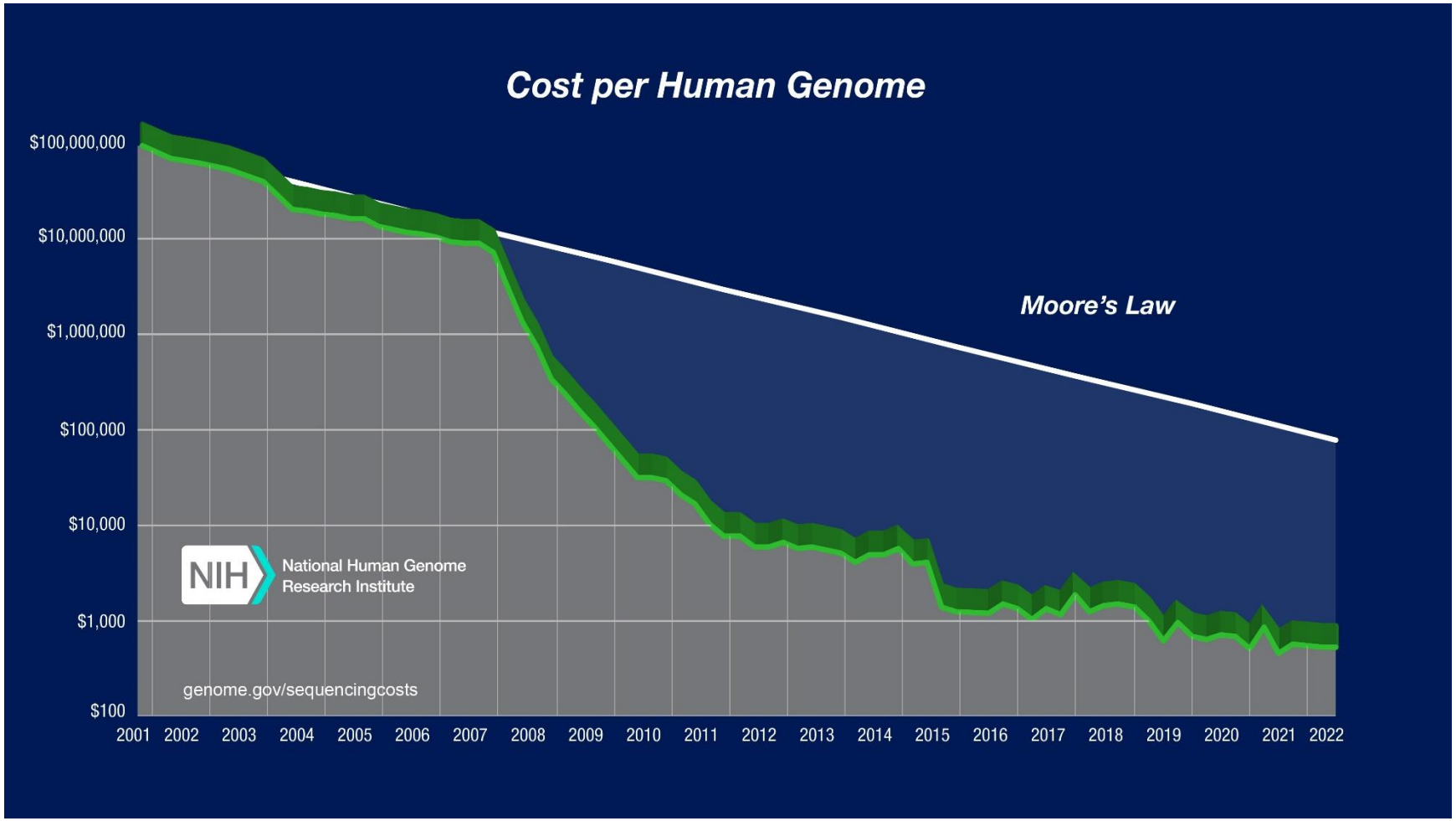


2017



The Ogilvie Building

Plummeting Sequencing Costs



The National Human Genome Research Institute (NHGRI)

https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

NGS Platforms

- **454 Sequencing** 2005
  - 454
  - GS FLX+/ GS Junior

- **Illumina (Solexa)** 2006

http://www.illumina.com/

  - HiSeq/MiSeq/NovaSeq/NextSeq Genome Analyzers

- **ThermoFisher Scientific (Invitrogen Life Technologies)**

https://www.thermofisher.com/au/en/home/life-science/sequencing/next-generation-sequencing.html

  - SOLiD 2007
  - Ion Torrent /PGM 2011
  - Ion Proton™ Sequencer 2012

Other NGS Platforms

- **Helicos Heliscope™ 2009**

- **MGI Tech (BGI – Complete Genomics)**

- **Pacific Biosciences 2010**

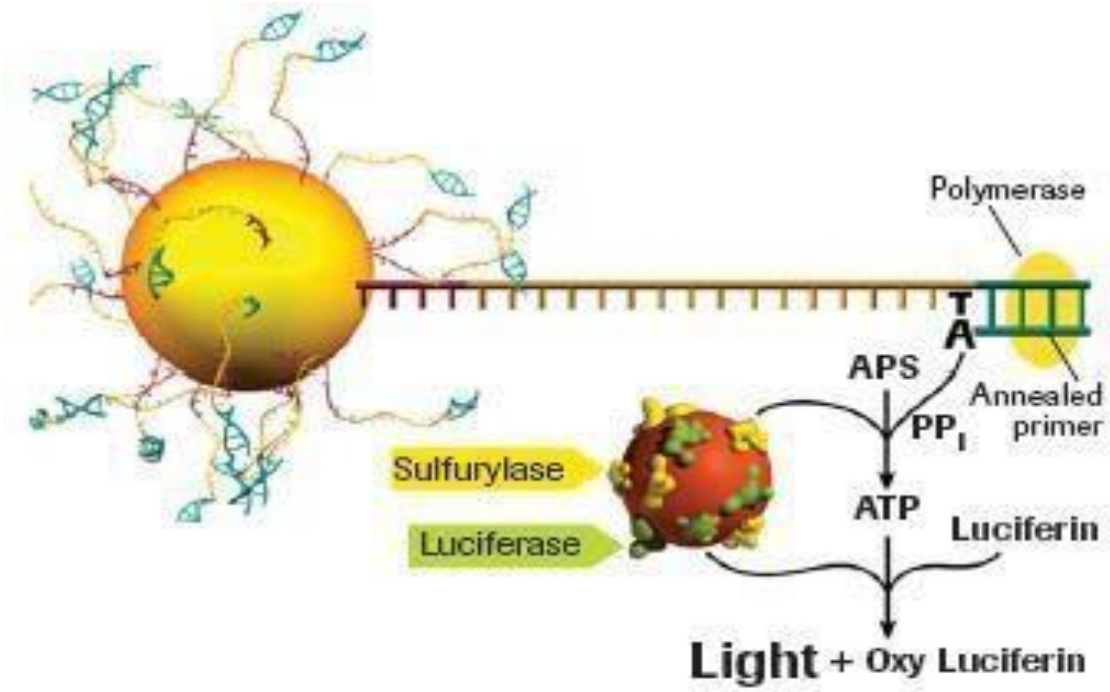
  http://www.pacificbiosciences.com/

  - **PACBIO *RS.***
  - **Sequel System.**
  - **Revlo**
  - **Onso**

- **Oxford Nanopore Technologies**

454

- First generation NGS (2005).
- First massively parallel sequencer.
- Bought by Roche in 2007. Now discontinued.
- Based on pyrosequencing of bead-bound DNA in microwells. A chemical cascade converts luciferin to



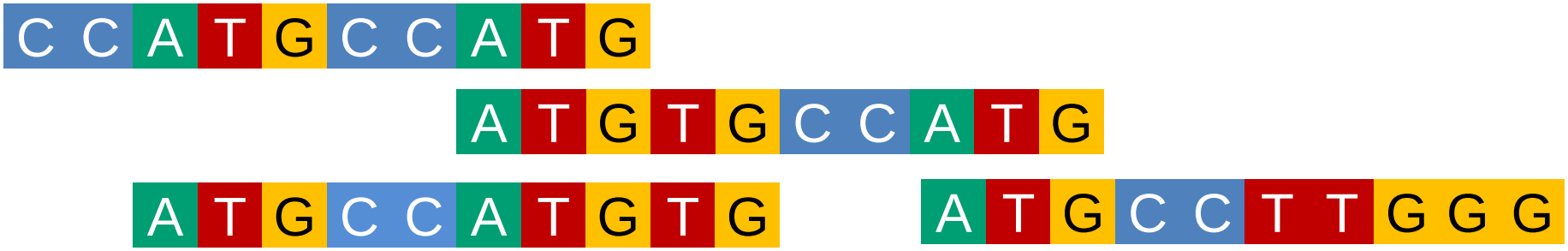
Massively Parallel Sequencing

- Sequencing technologies and instruments have rapidly expanded over the past two decades.
- Many short read and a few long read options are available.
- Each has its own advantages and disadvantages.

Short Read Technologies < 1 kb	Long Read Technologies > 10 kb
Illumina	PacBio
MGI Tech Co. Ltd.	Oxford Nanopore Technologies
Element Biosciences	MGI Tech. Co. Ltd.
GeneMind	
PacBio Onso	
ThermoFisher Scientific	
Ultima Genomics	

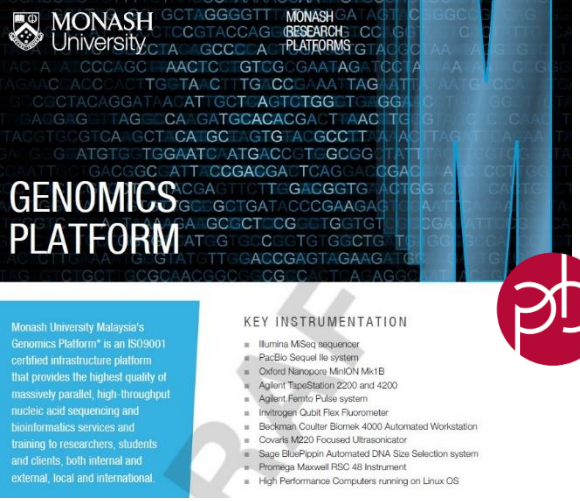
Short vs Long Read Sequencing

36 - 600 bp



> 10 kb ≤ 2 Mb

MUMGP



PACBIO®

illumina®

@genomicsMUMGF

- Monash University Malaysia Genomics Platform is an ISO 9001:2015 certified infrastructure platform under the School of Science.
- It hosts three of the currently used sequencing technologies.
- Vision is to use modern technology, big data analytics and innovative solutions for conservation, food security and good health in alignment with Monash Impact 2030 and United Nations sponsored Sustainable Development Goals.

LinkedIn

Find us on www.facebook.com/mumgf

Sequencing Strategies

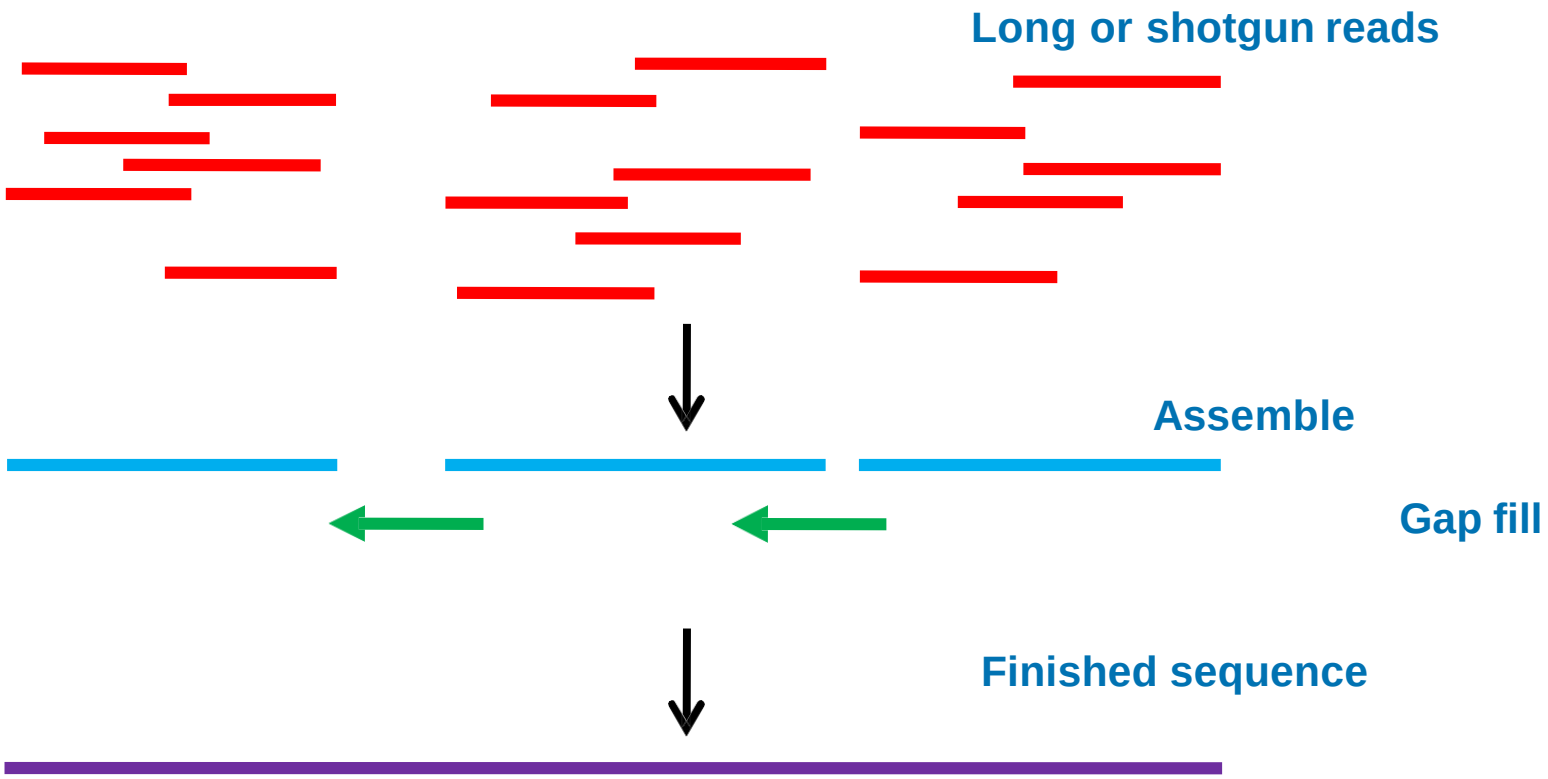
Gene Panels	Whole Exome Sequencing	Whole Genome Sequencing
Covers <1% of genome.	Covers 1 - 2% of genome	Covers 85 - 100% of genome
Requires prior knowledge.	Requires prior knowledge	
Target enrichment required.	Target enrichment required	No enrichment required
Low input DNA.	Low input DNA	Low to high input and HMW DNA
Inexpensive – costly (US$ <100)	Expensive (US$100 -200)	Expensive (US$100 - 4000)
Multiplex numerous samples	Multiplex numerous samples	Multiplex fewer samples
Little computational support	Greater computational support	Massive computational support

- Transcriptomics – RNA-Sequencing.
- Metagenomics – 16S rRNA gene resequencing.
- Epigenomics.

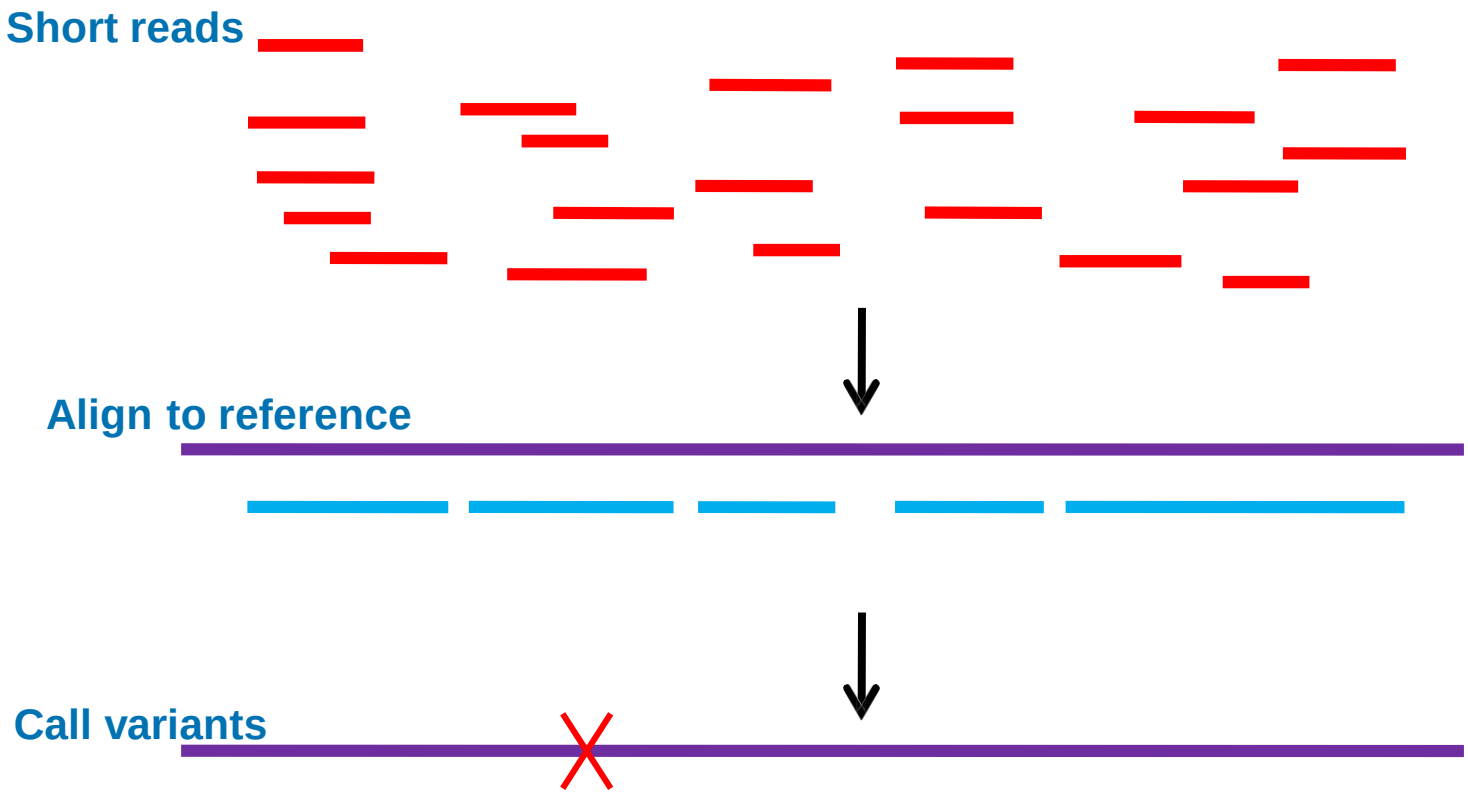
Next Generation Sequencing Steps

- Extract and/or fragment DNA.
- Prepare DNA fragment library.
- Sequence fragments 36 – 4 Mbp.
- Assemble fragments:
  - Map fragments to reference sequence.
  - *De-novo* assembly.
- Call and annotate DNA variants.

De-novo Sequencing

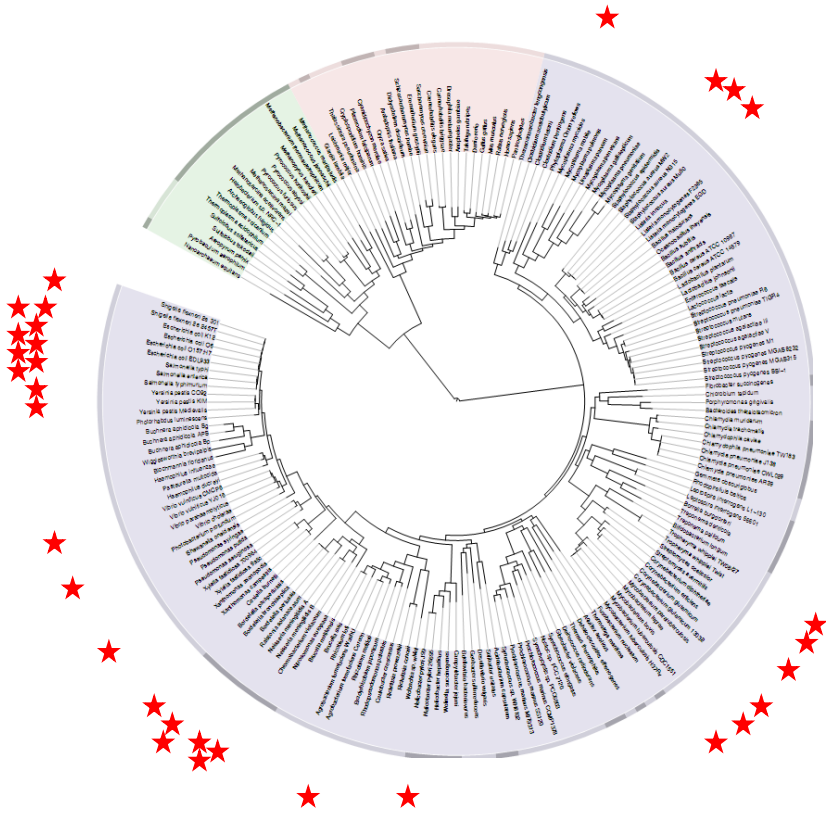


Re-sequencing



What Matters Most?

- Data quality.
- Read length.
- Throughput.



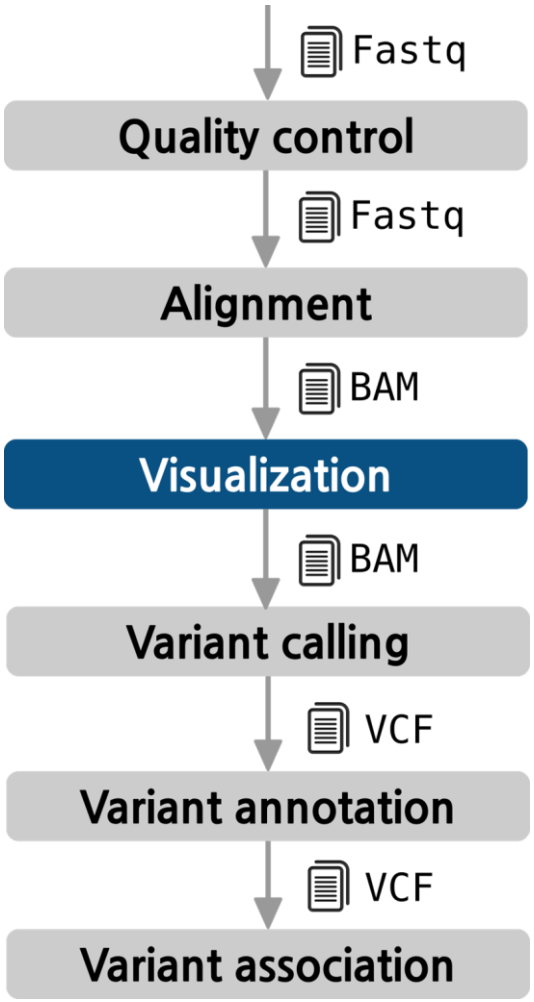
Sequencing Datasets Issues

- Exceedingly large datasets.
- Require complex computational tools.
- Integrative experiments and reproducibility.
- Accessibility and transparency.
- Ethical concerns.

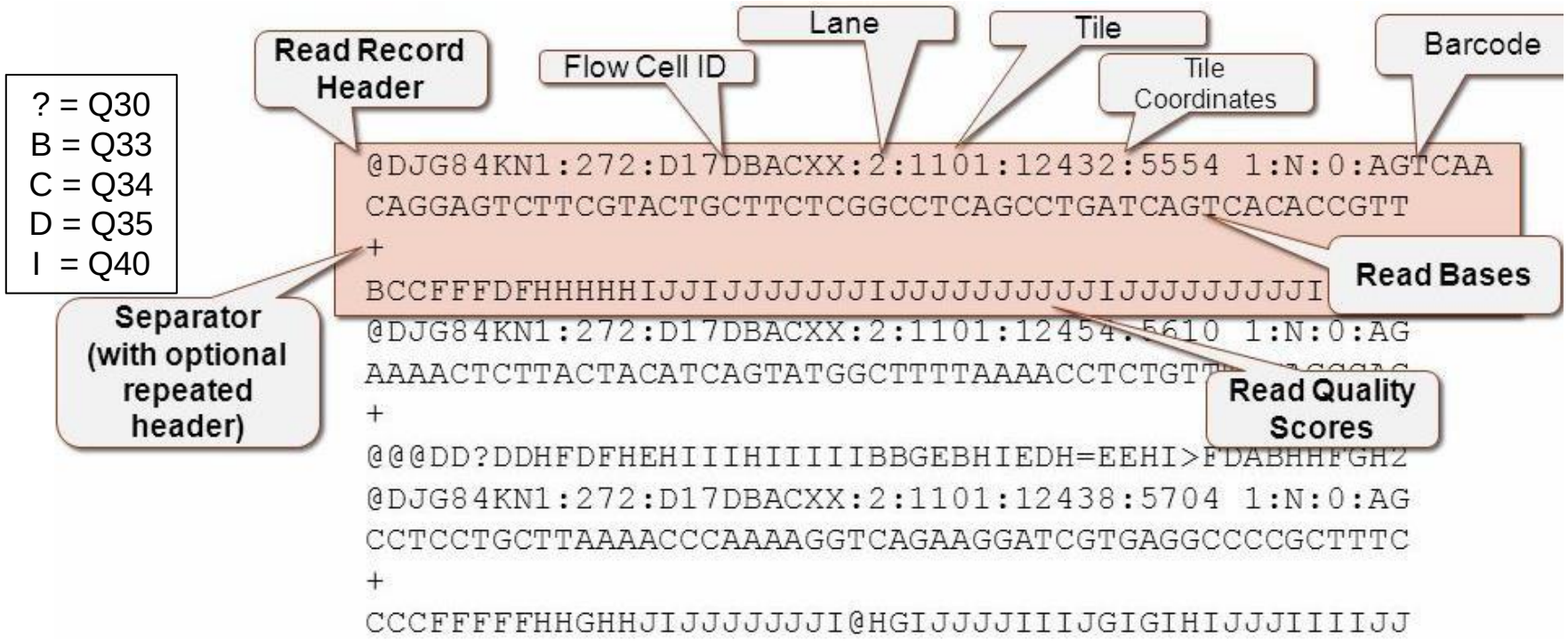
DNA Extraction and Fragmentation

- Separation or breaking DNA strands into pieces.
- DNA Fragmented by:
  - Shearing.
  - Nebulization.
  - Sonication.
  - Enzymatic digestion.
  - Transposon mediated fragmentation.

Typical Sequence Analysis Pipeline



The FASTQ File Format



Probability of Incorrect Base Calls

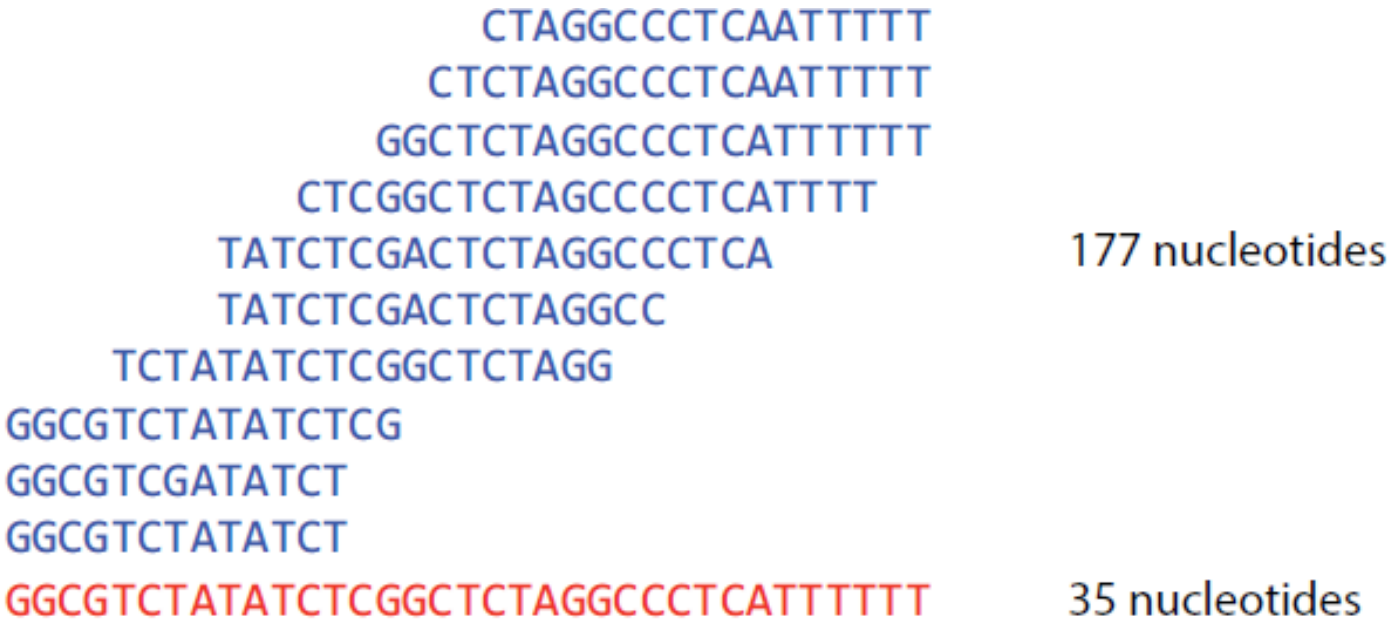
- Assess or measure accuracy of base calling.
- Sequence quality Q is reported on a log scale.
- Defined as a property related to the base calling error probabilities (P): $Q = -10 \log_{10}(P)$

Phred Quality Score (ASCII QS)	Probability of Incorrect Base Call	Call Accuracy (%)
Q10 (+)	1 in 10 bases	90
Q20 (5)	1 in 100 bases	99
Q30 (?)	1 in 1,000 bases	99.9
Q40 (I)	1 in 10,000 bases	99.99
Q50 (S)	1 in 100,000 bases	99.999

- **Q30 means that virtually all bases in a read are called correctly.**

Sequencing Coverage

- Coverage usually refers to average number of reads covering a genome or genomic region. Some people also refer to this as depth.



Average coverage = 177 / 35 ≈ 7x

Estimating Sequencing Runs

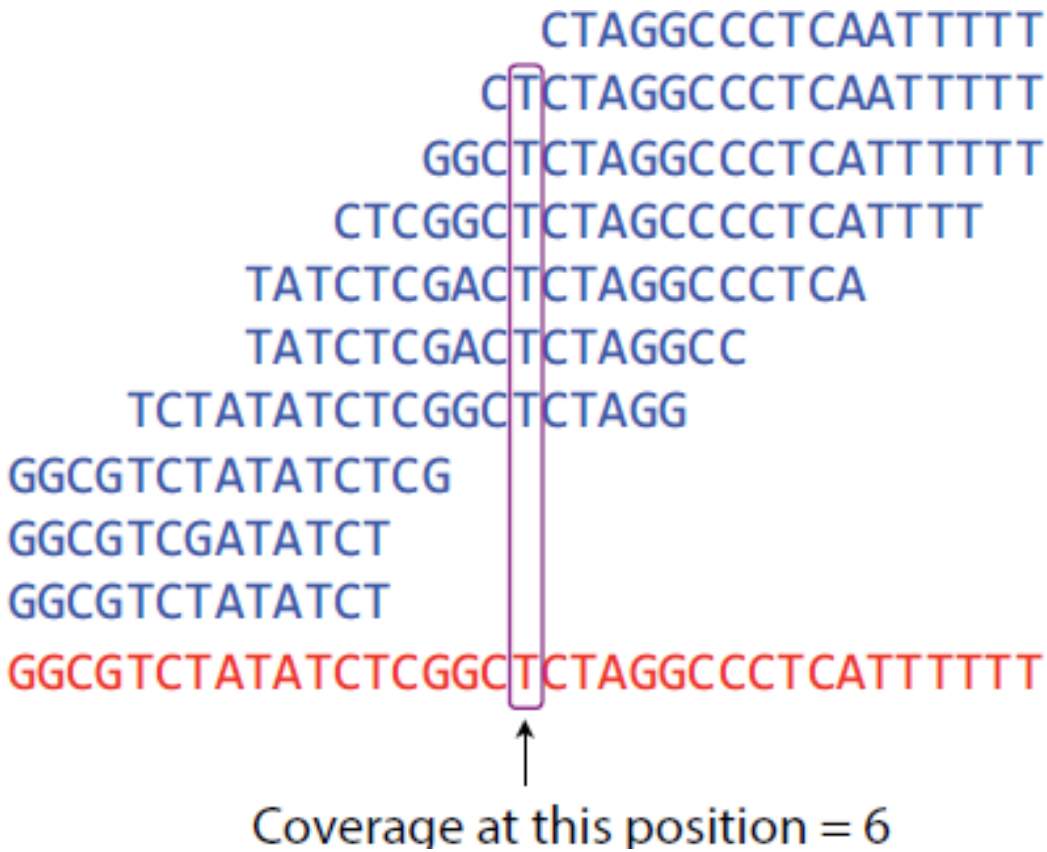
- The Lander/Waterman equation is a method for computing coverage. Lander and Waterman (1988) Genomics 2:231-239.
- The general equation is: C = LN / G
  - C stands for coverage.
  - G is the haploid genome length.
  - L is the read length.
  - N is the number of reads.
  - If we take one lane of single read human sequence with v3 chemistry, we obtain:

$$C = (100\ bp)*(200\times10^6)/(3\times10\ bp) = 6.3$$

https://support.illumina.com/downloads/sequencing_coverage_calculator.html

Sequencing Depth

- Depth usually refers to the number of reads covering a particular position in the genome.



Short Read Sequencing

Instruments

Illumina Instruments

BGISEQ-500

GeneMind

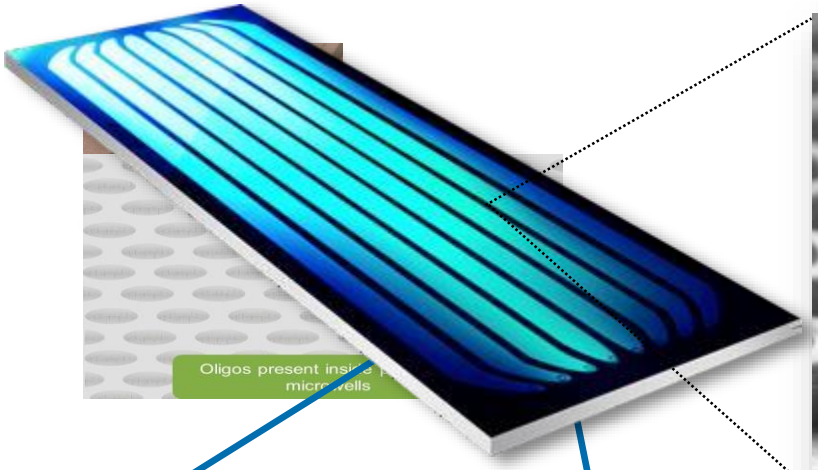
DNBSEQ Sequencers

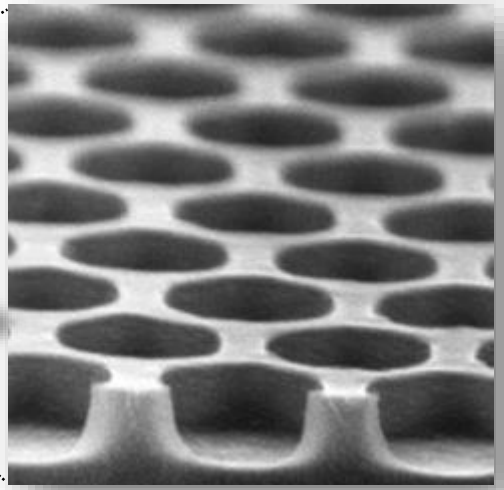
PacBio Onso

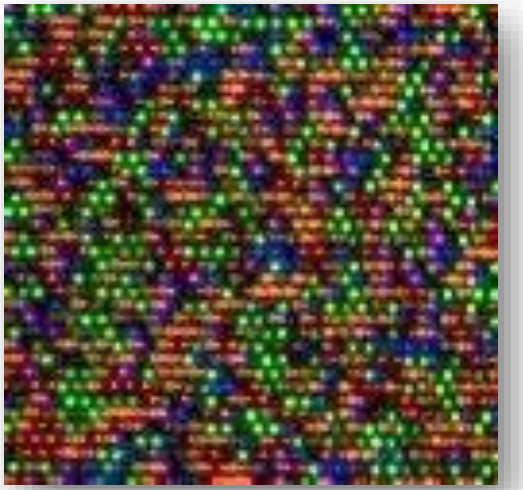
Element Biosciences AVITI24™

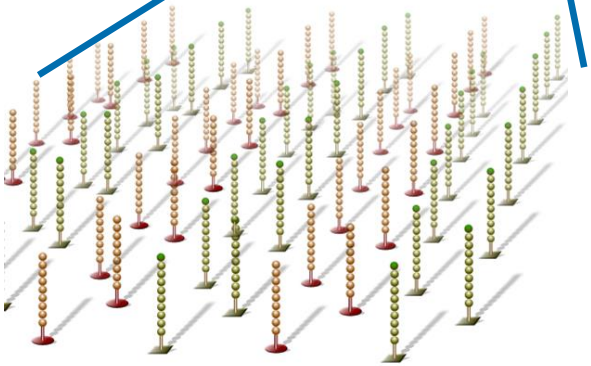
Ion Torrent's PGM

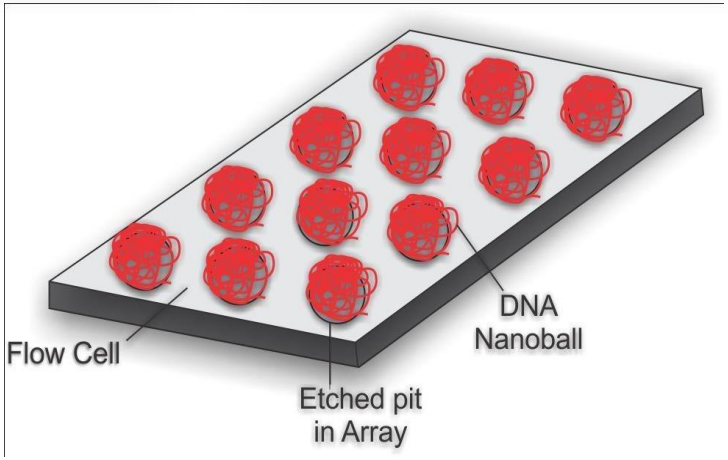
Short-Read Sequencing Flowcells







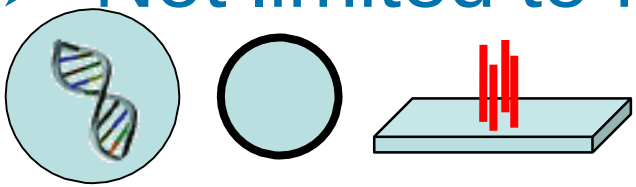


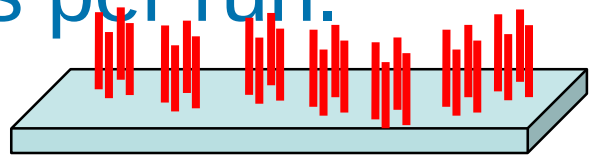


- Single or double flowcell instruments.
- Suitable for all applications including exomes.
- Patterned flowcell (aka “PFCT”).

Flow Cells Enable Parallel Sequencing

- From 2004 – present.
- Conduct high-throughput massively parallel sequencing.
- Not limited to few reactions per run.





1 feature.
1 template.

1 chip, thousands or millions of features.
Output Mb-Tb.

Solexa - Illumina Sequencers

- Launched their Genome Analyzer in 2006.
- Spinout from Cambridge University, set up at Great Chesterford in 2000.
- Genome Analyzer; 1Gb/run.
- Acquired by Illumina in 2007.
- Short read sequencing.
- Highly accurate (0.1-0.2% error).
- Market leader.
- Cost US$6-$50/Gb.

Illumina Sequencers

- The current market leader.
- ~70% of sequencing market.

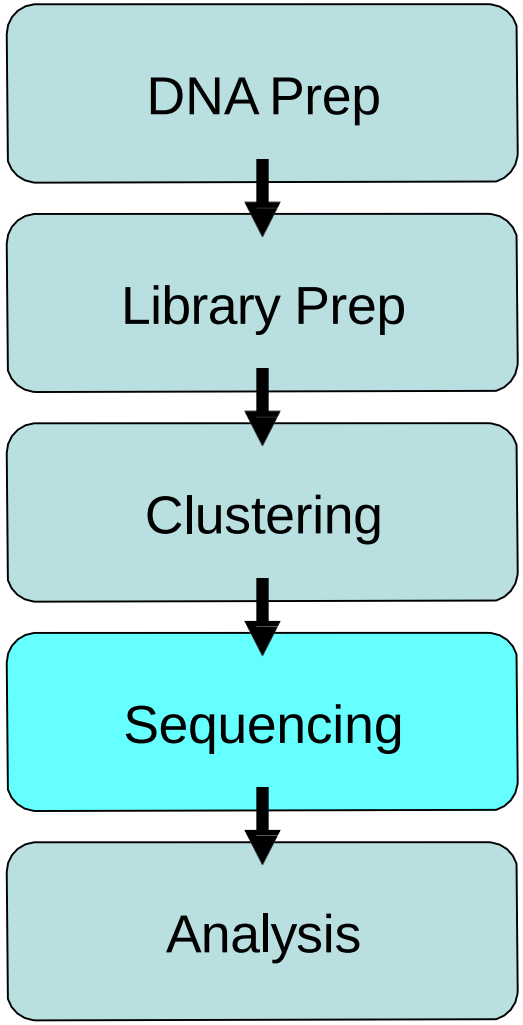
Scale	
Benchtop	**Production**
iSeq	NextSeq
MiniSeq	HiSeq Series
MiSeq	HiSeq X Series
NextSeq1000 NextSeq2000	NovaSeq6000

The Current Market Leader

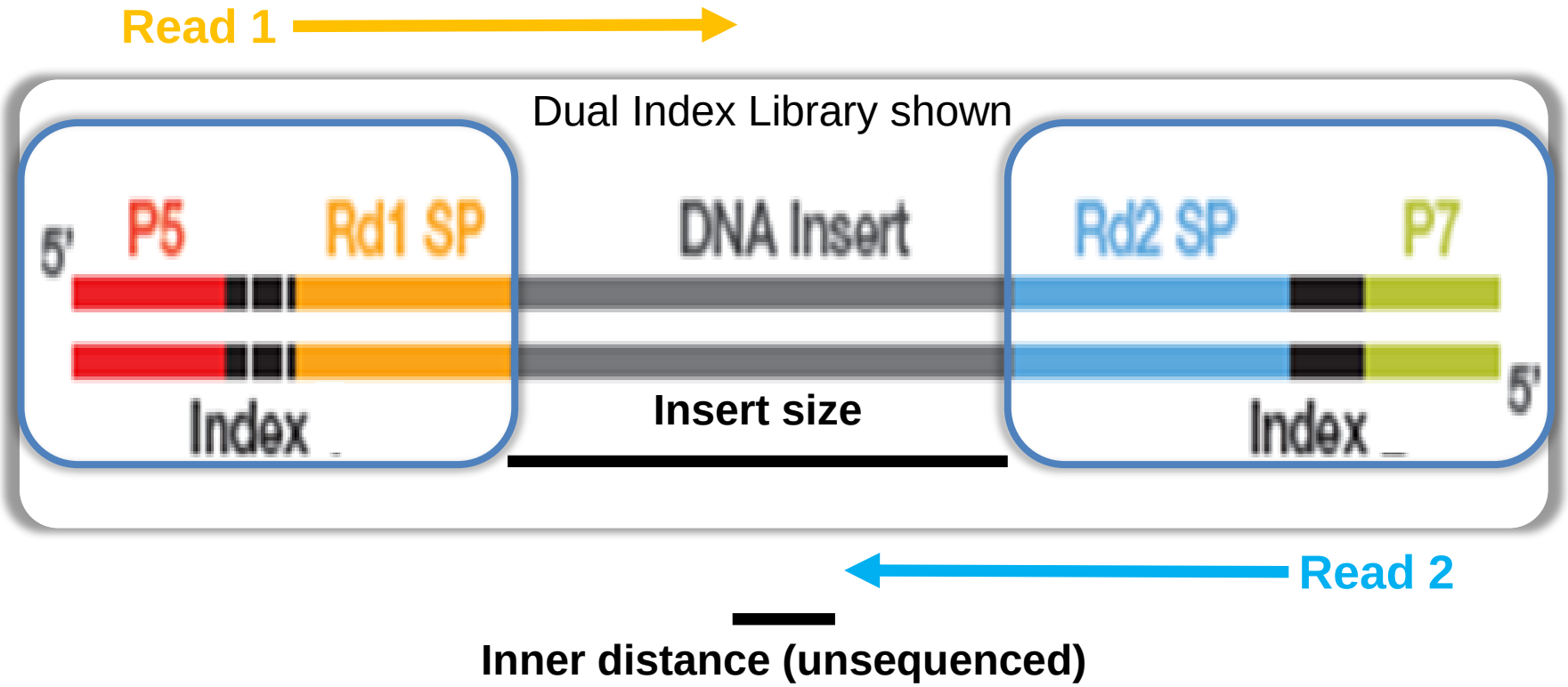
Platform (Run Time)	Read Length	Maximum Output (Gbp/run)	Cost/ Human Genome (US$)
MiniSeq (1 day)	2 X 150 bp	7.5 Gb	NA
MiSeq (1-3 days)	2 X 300 bp	15.0 Gb	NA
NextSeq (1-2 days)	2 X 150 bp	120 Gb	
HiSeq (1-6 days)	2 X 150 bp	1500 Gb	2,500
HiSeq X (<3 days)	2 X 150 bp	1800 Gb	1000
NovaSeq 6000 (1-2 days)	2 X 150 bp	6000 Gb	
Capillary sequencing	700-1000 bp	0.6 Gb	3,000,000,000 10,000,000

Illumina Workflow

- DNA Preparation.
- Library preparation.
- Cluster generation.
- Sequencing.
- Data analysis.

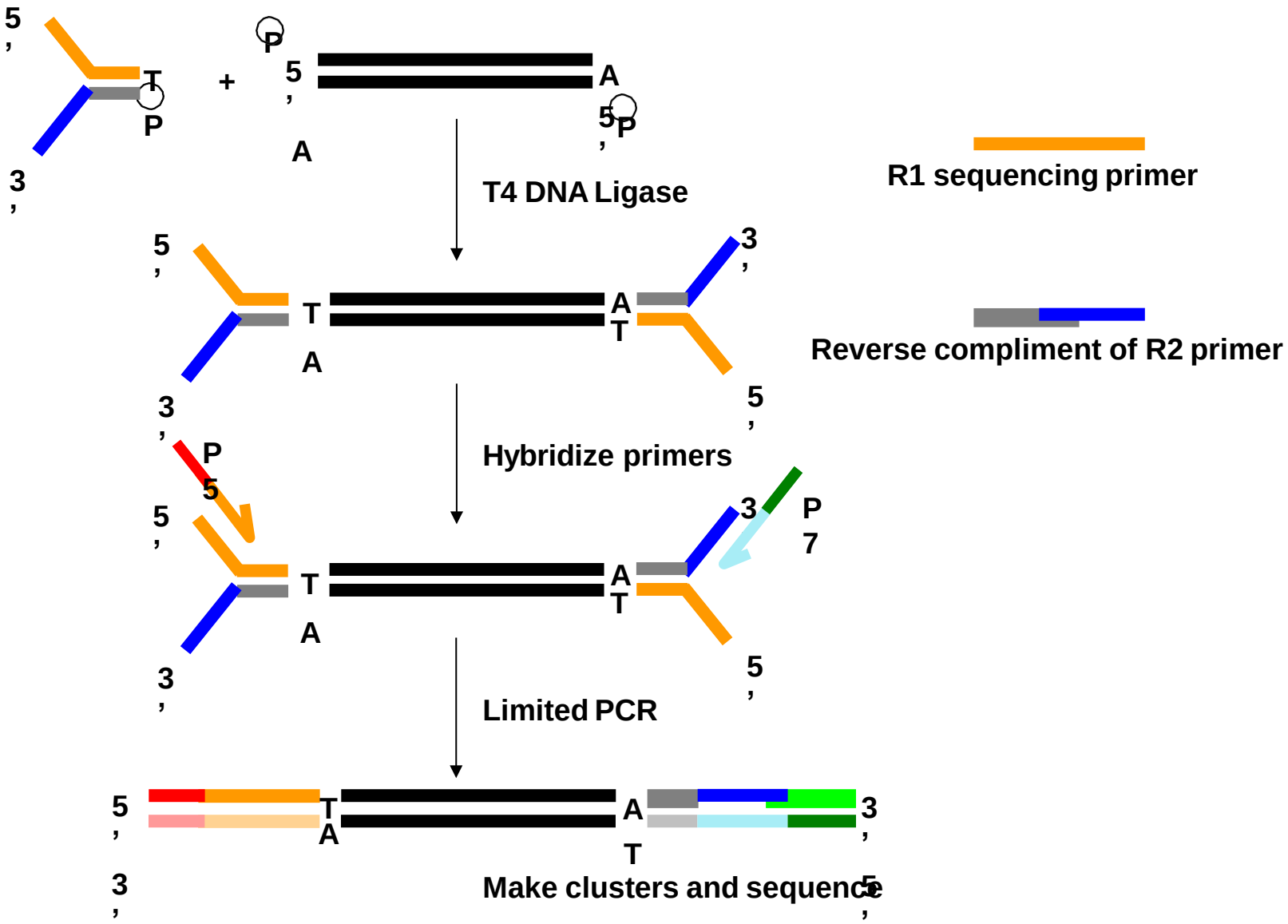


Library Preparation

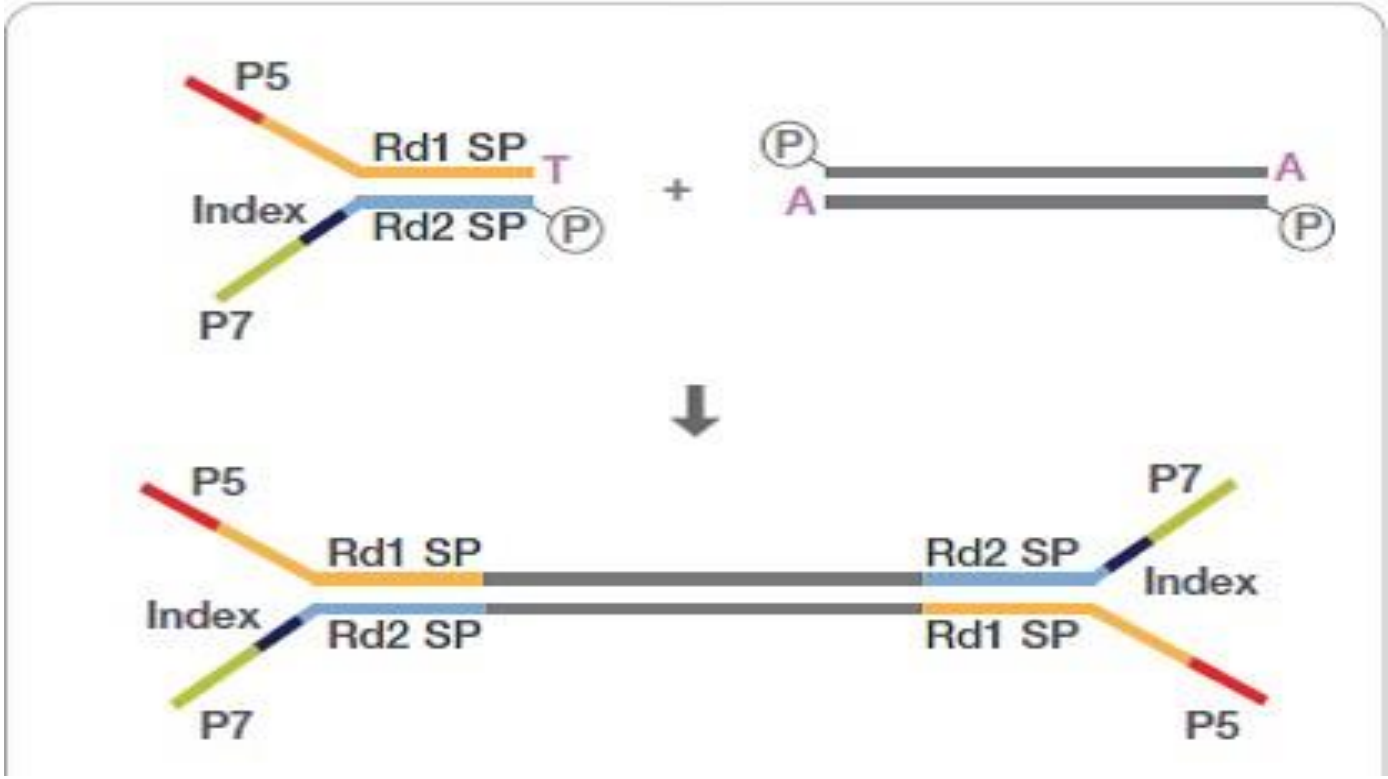


- **The aim of library preparation is to obtain nucleic acid fragments with adapters attached on both ends.**

Illumina Paired End Library Preparation

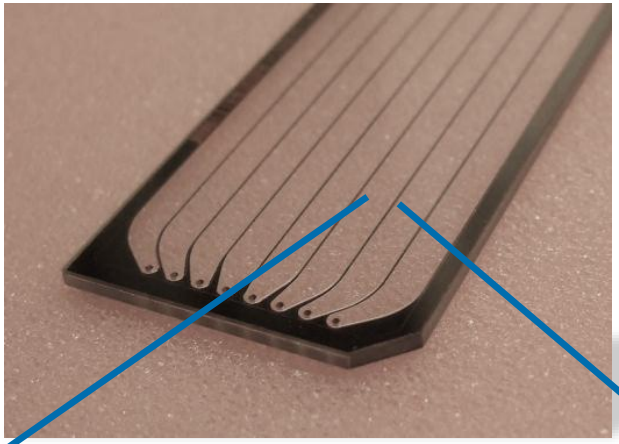


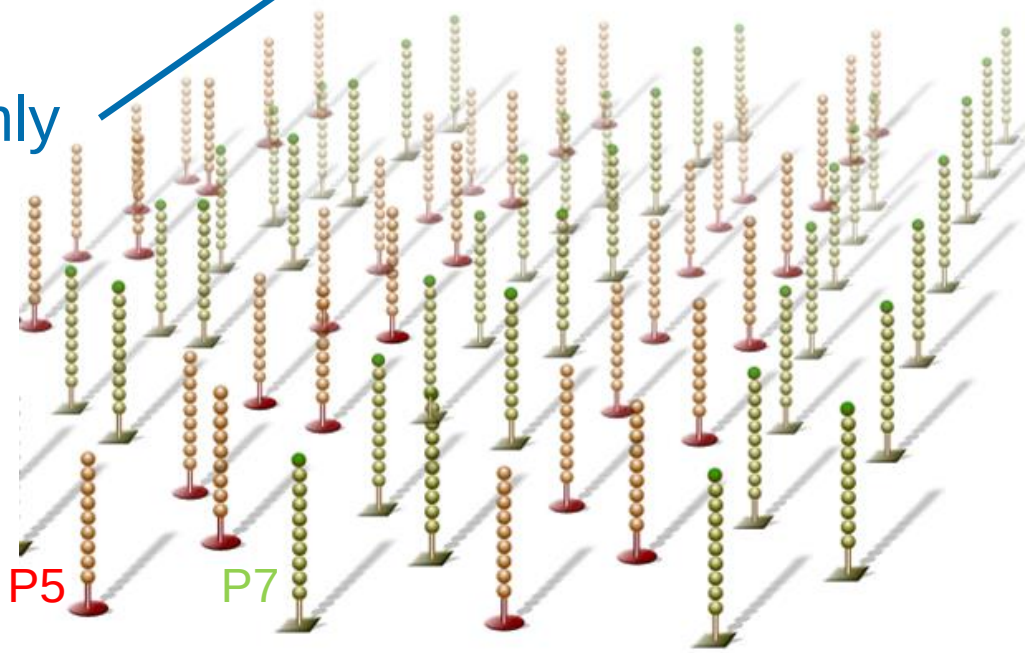
Illumina Truseq Library Prep



Illumina Flow Cell

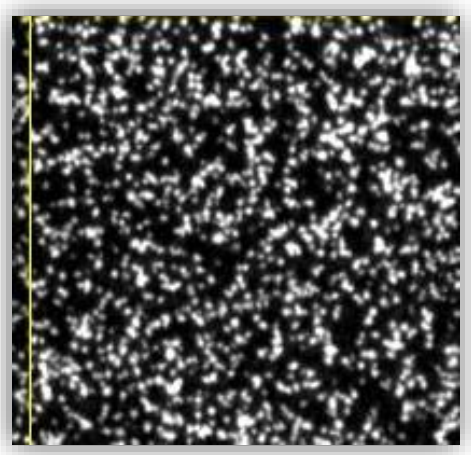
- A flow cell is a thick glass slide with channels or lanes.
- Cluster generation occurs on a flow cell.
- Each lane is randomly coated with a lawn of oligos that are complementary to library P5 and P7 adapters .

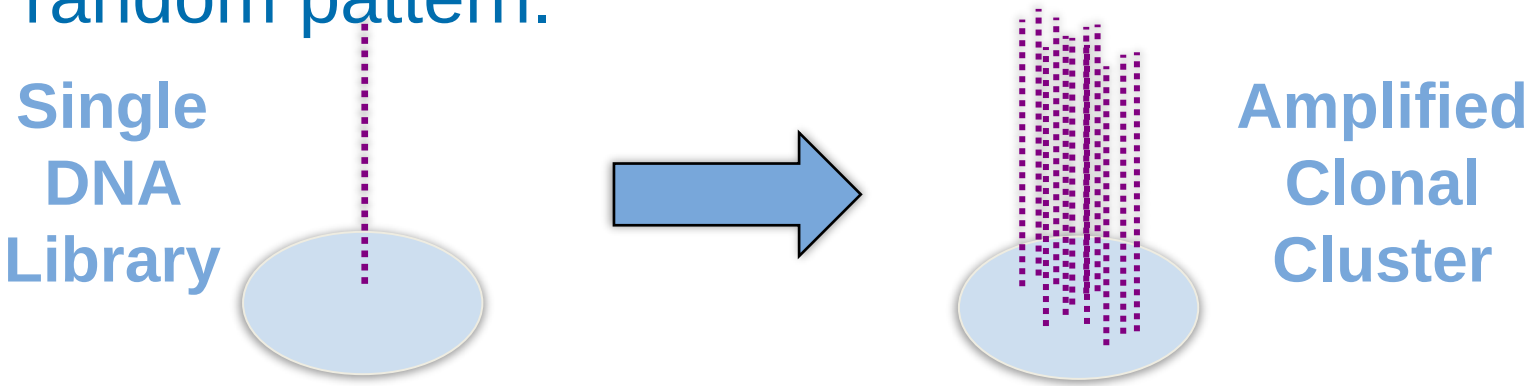




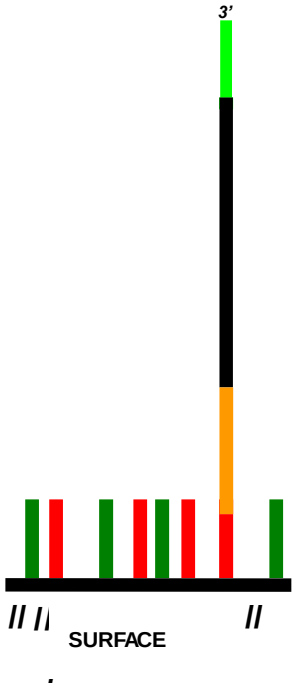
What is a Cluster?

- Clusters are bright spots on an image.
- Each cluster represents thousands of copies of the same DNA strand in a 1–2 micron spot.
- Around 800,000 clusters/mm^2 are bound to the flow cell in a random pattern.

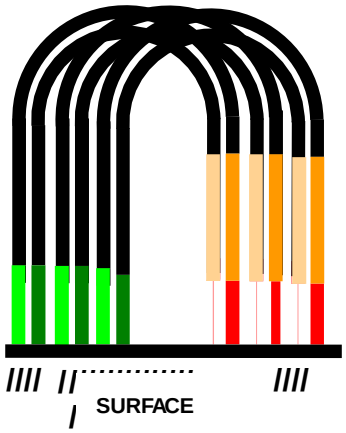




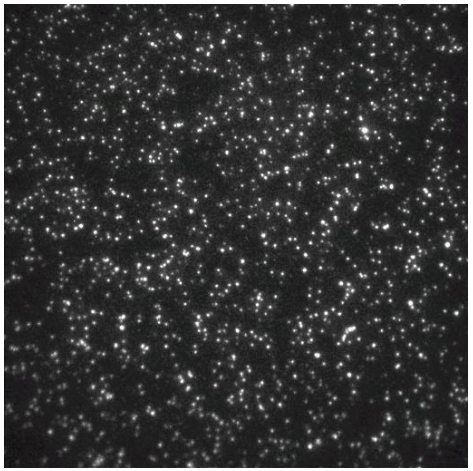
Cluster Amplification



Single-molecule array



Cluster ~1000 molecules

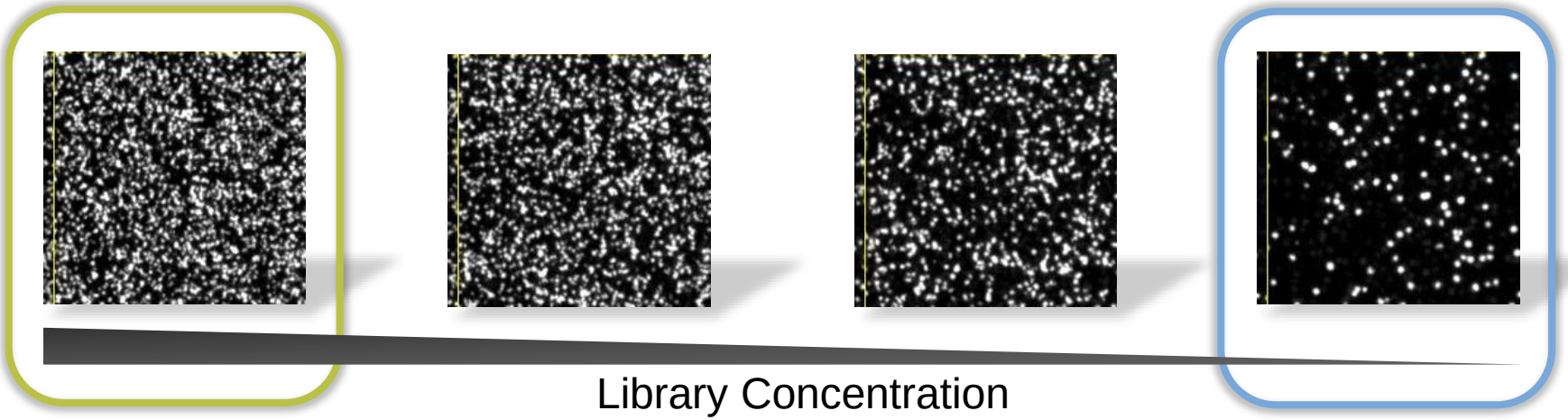


1.5 Billion clusters on a single glass chip

The cBOT

Maximize Data Quality and Quantity

Optimized flow cell clustering determines data quality and overall data yield



Overclustering can result in:

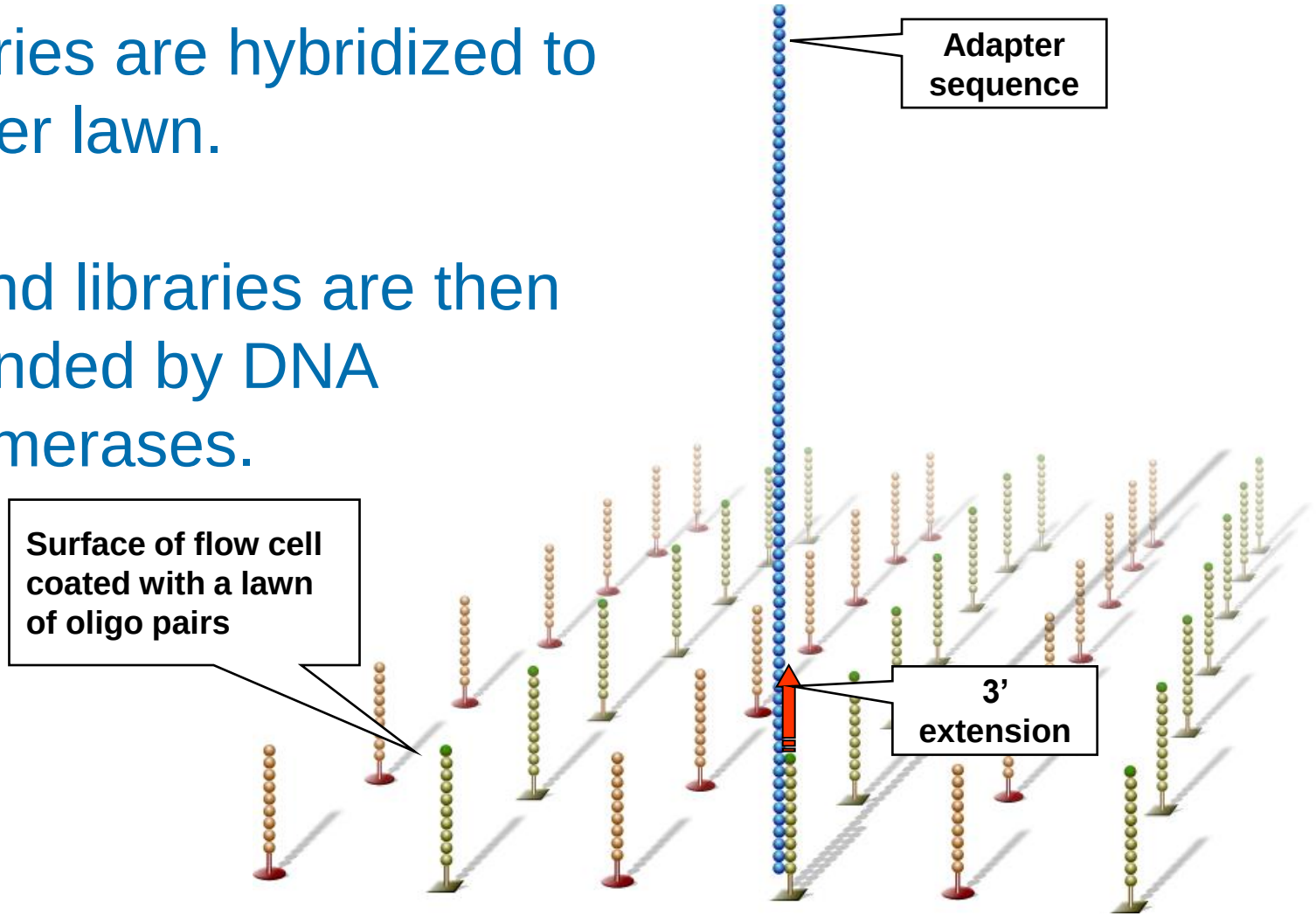
- Loss of data quality and data output
- Loss of focus
- Reduced base calls and Q30 scores

Underclusterring can result in:

- Loss of time and money
- Loss of focus
- Complete run failure

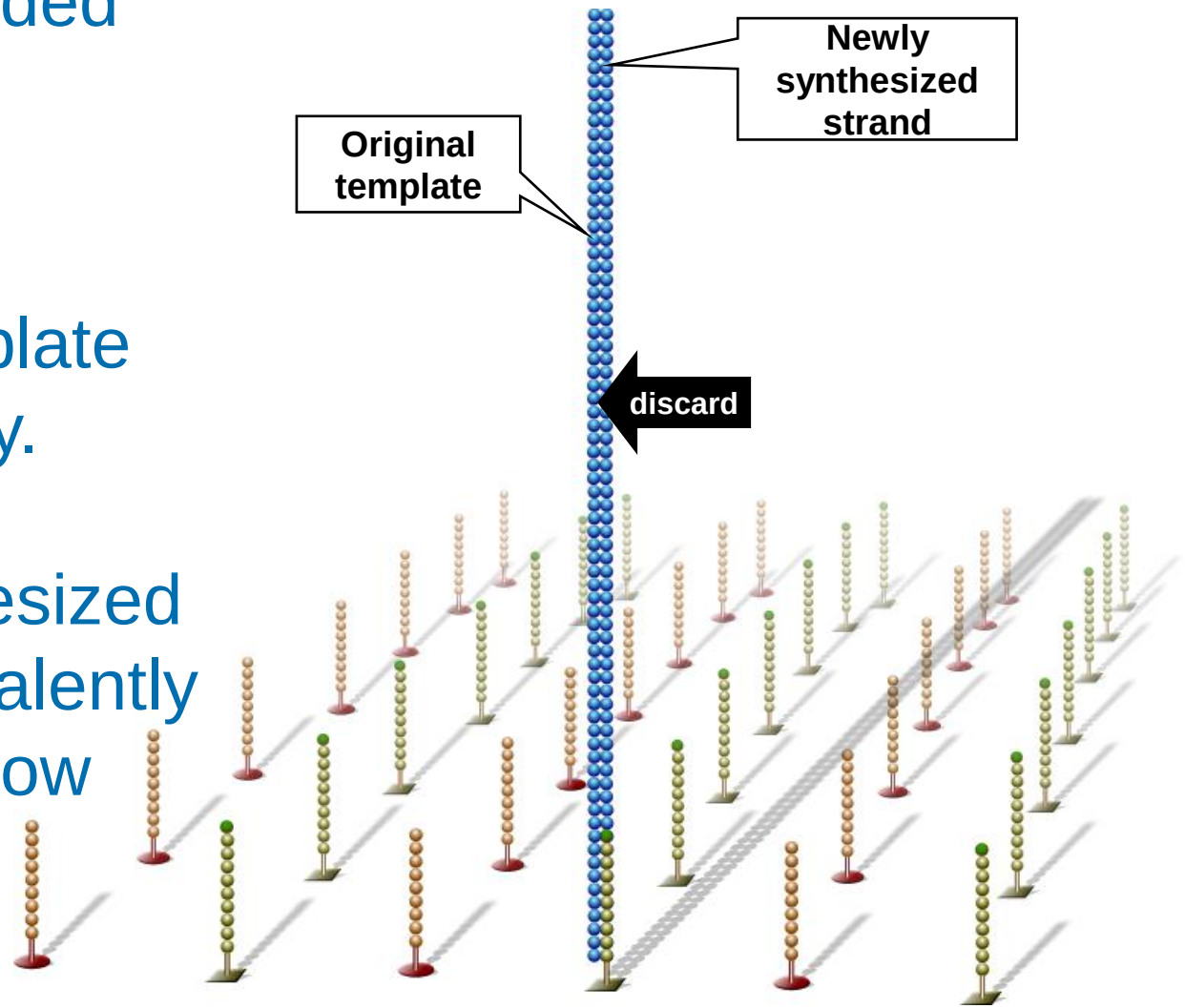
Fragment Hybridization & Extension

- Single stranded DNA libraries are hybridized to primer lawn.
- Bound libraries are then extended by DNA polymerases.



Denaturation

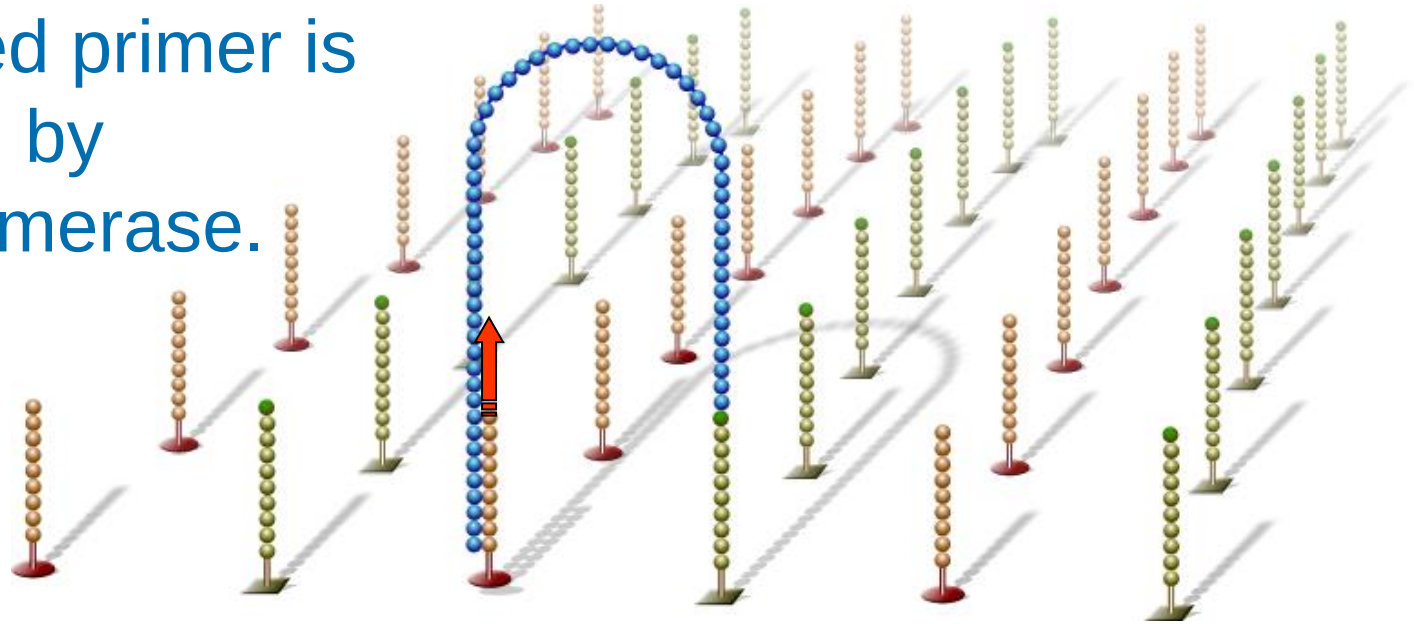
- Double-stranded molecule is denatured.
- Original template washed away.
- Newly synthesized strand is covalently attached to flow cell surface.



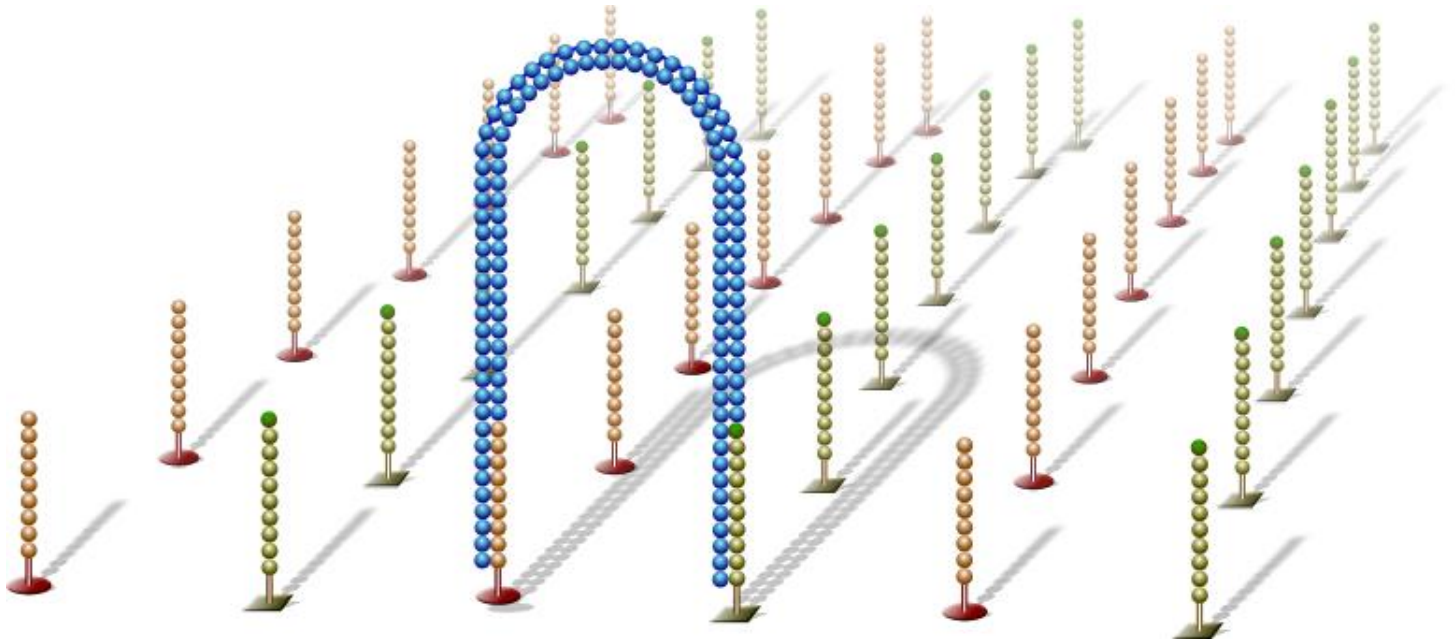
Bridge Amplification

- Single-stranded molecule flips over and forms a bridge by hybridizing to adjacent, complementary primer.

- Hybridized primer is extended by DNA polymerase.

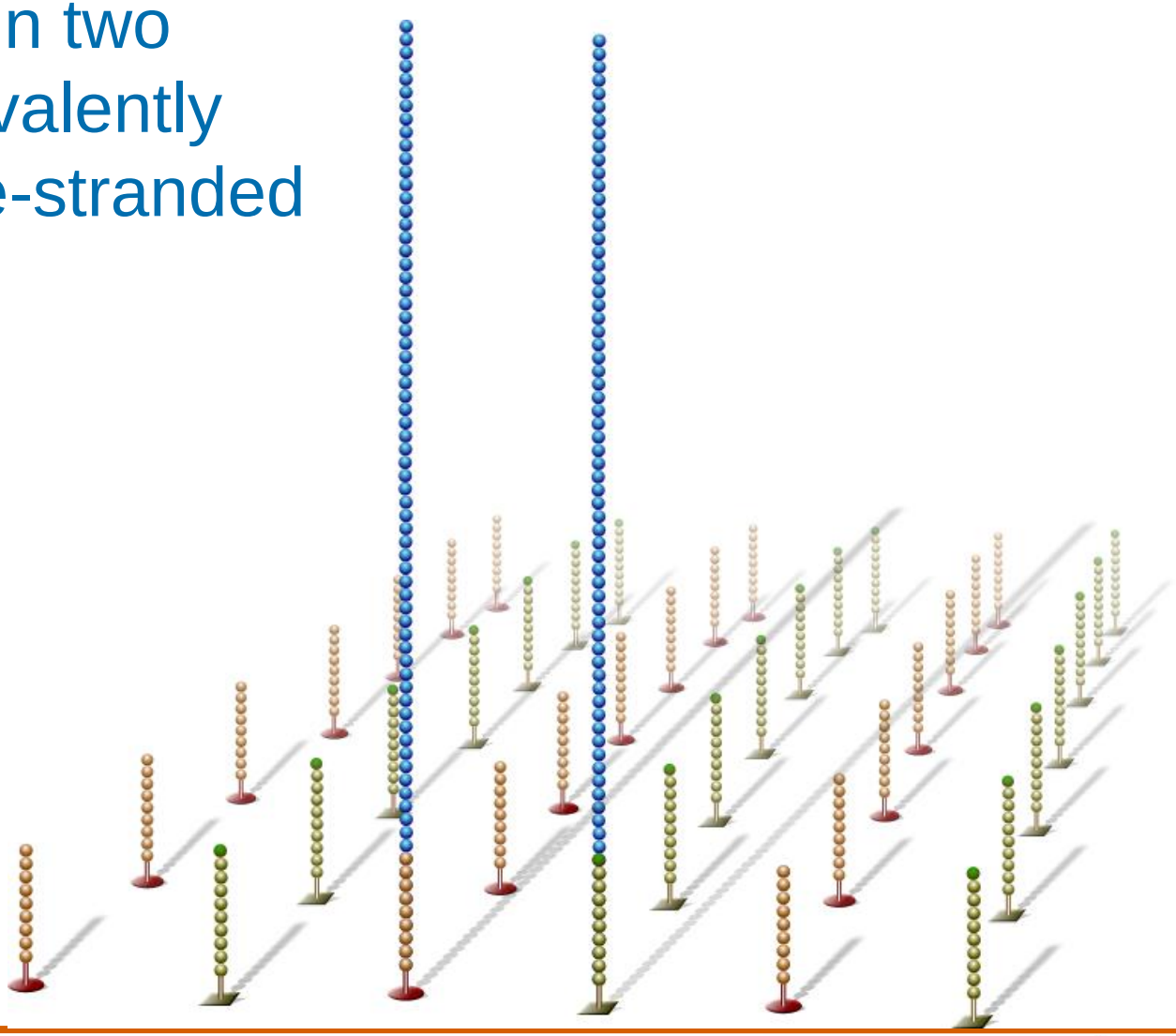


Formation of Double Stranded Bridge



Denaturation of Double-Stranded Bridge

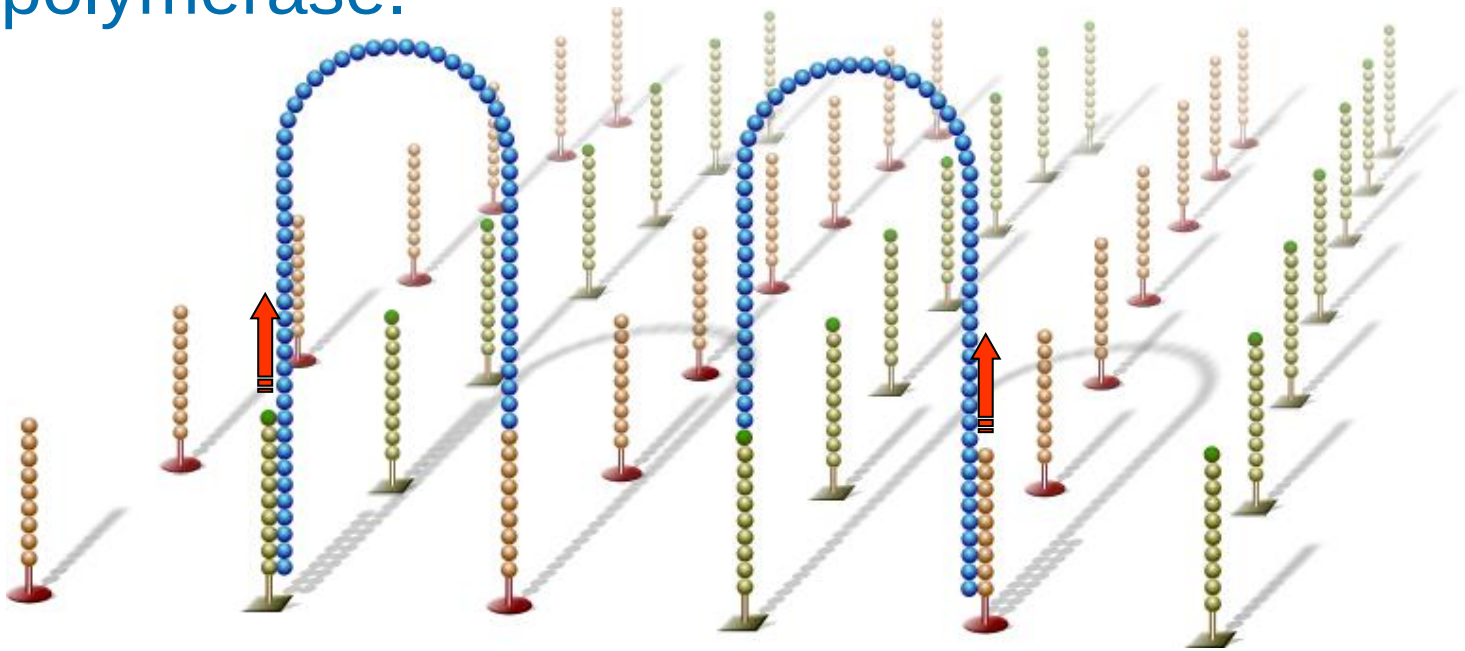
- This results in two copies of covalently bound single-stranded templates.



Continuation of Bridge Amplification

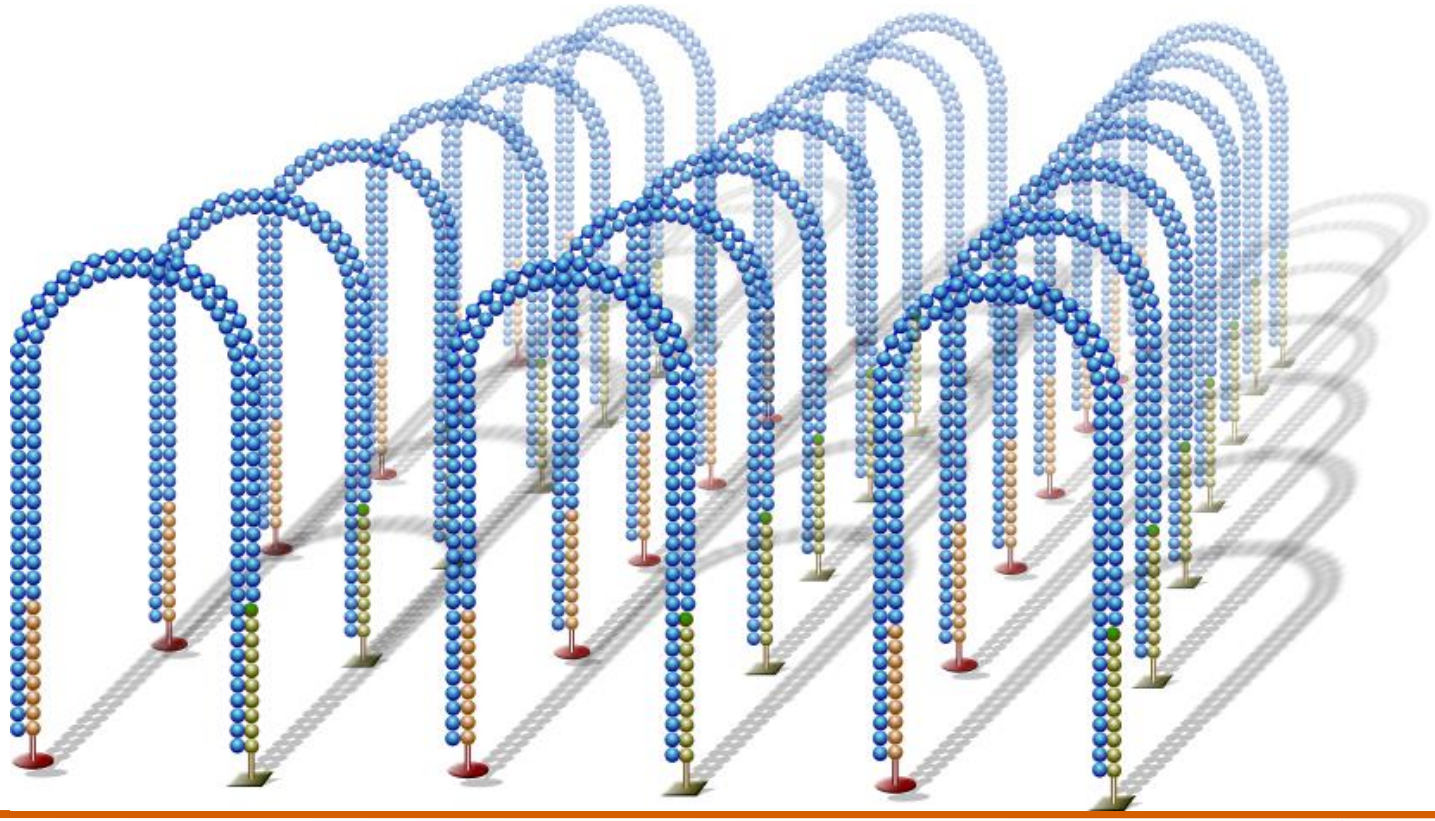
- Single-stranded molecules flip over to hybridize to adjacent primers.

- Hybridized primer is extended by DNA polymerase.



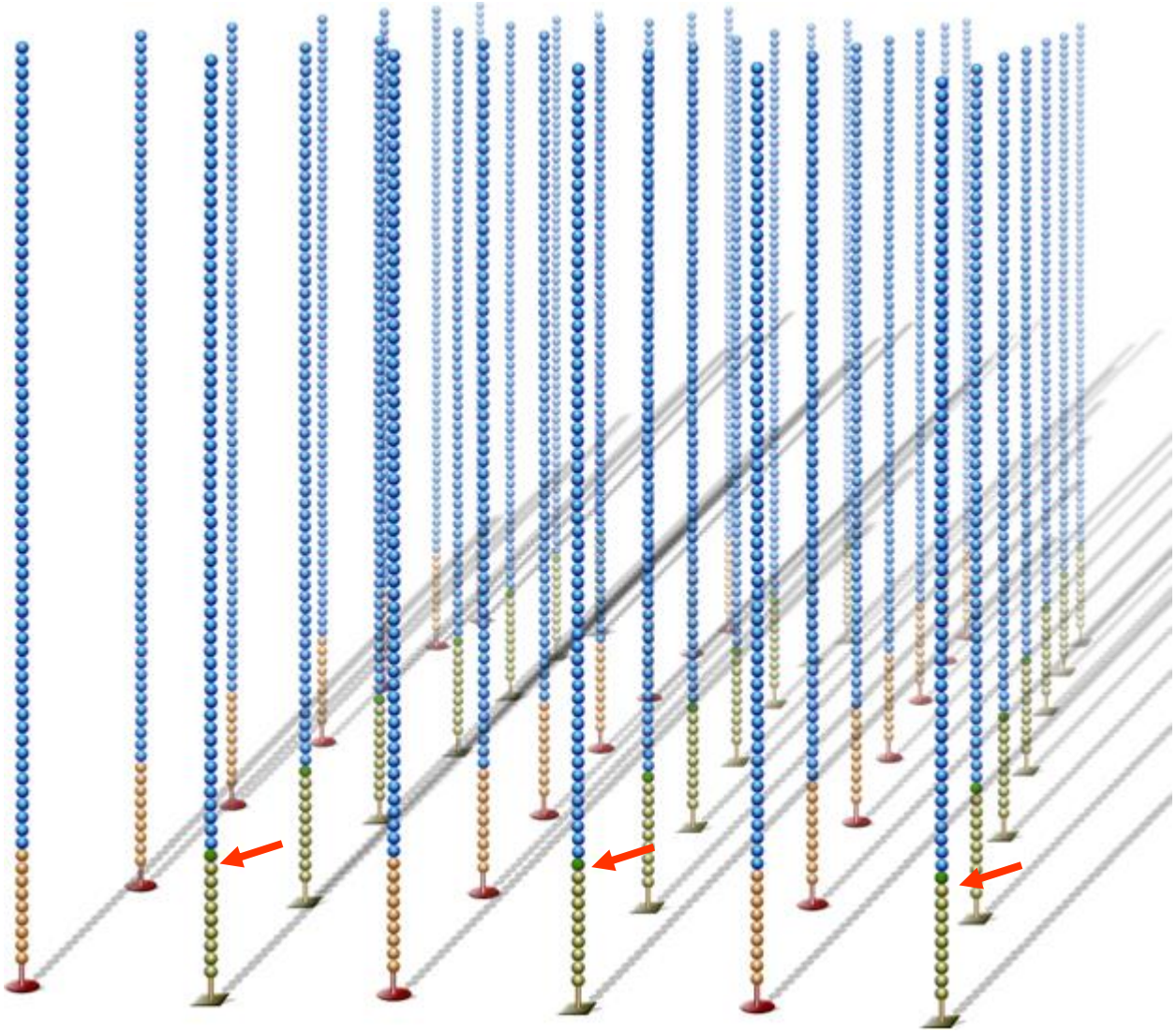
Formation of Multiple Clusters

- Bridge amplification cycle is repeated until multiple bridges are formed.



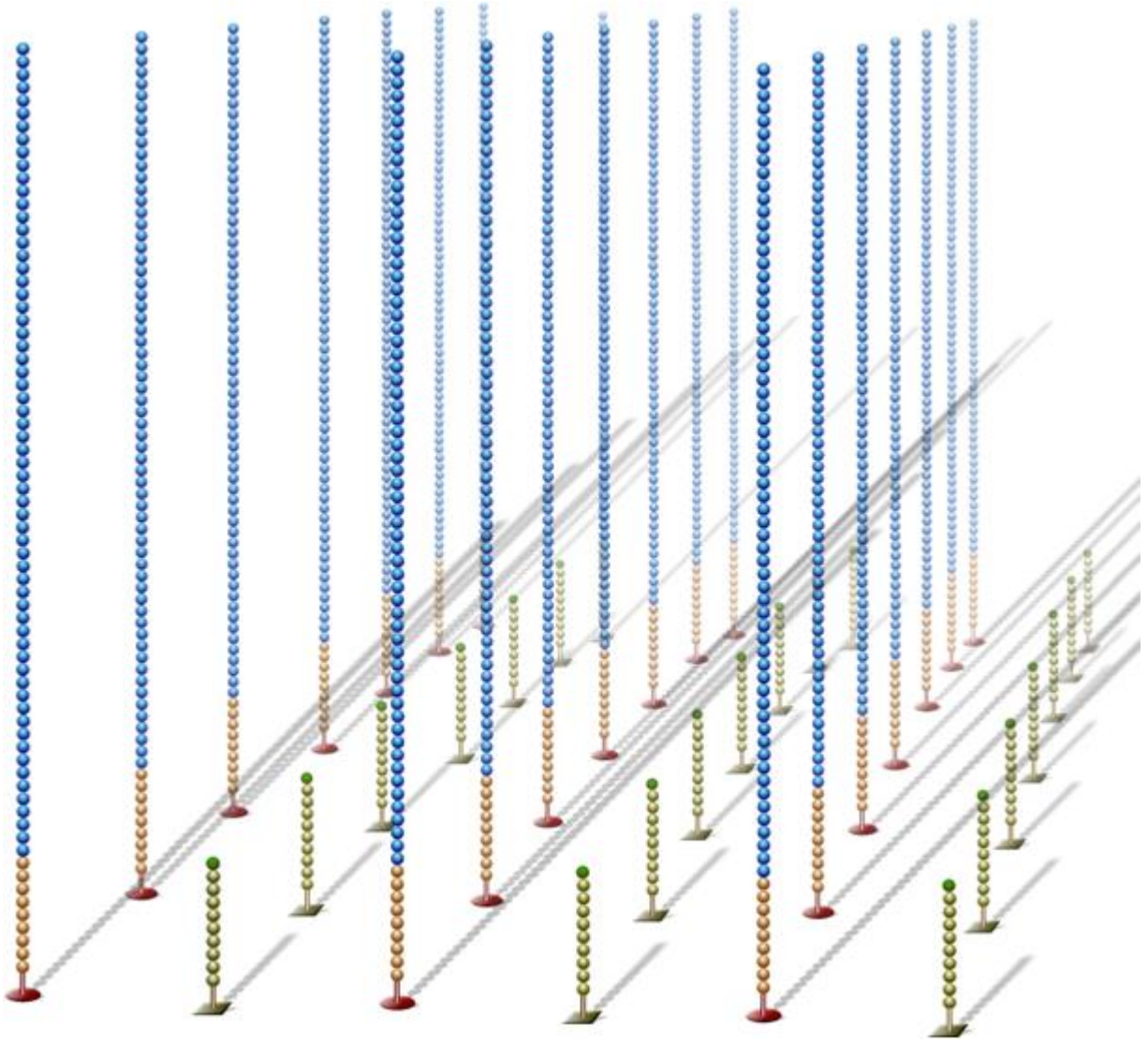
Denaturation and Linearization

- dsDNA bridges are denatured.



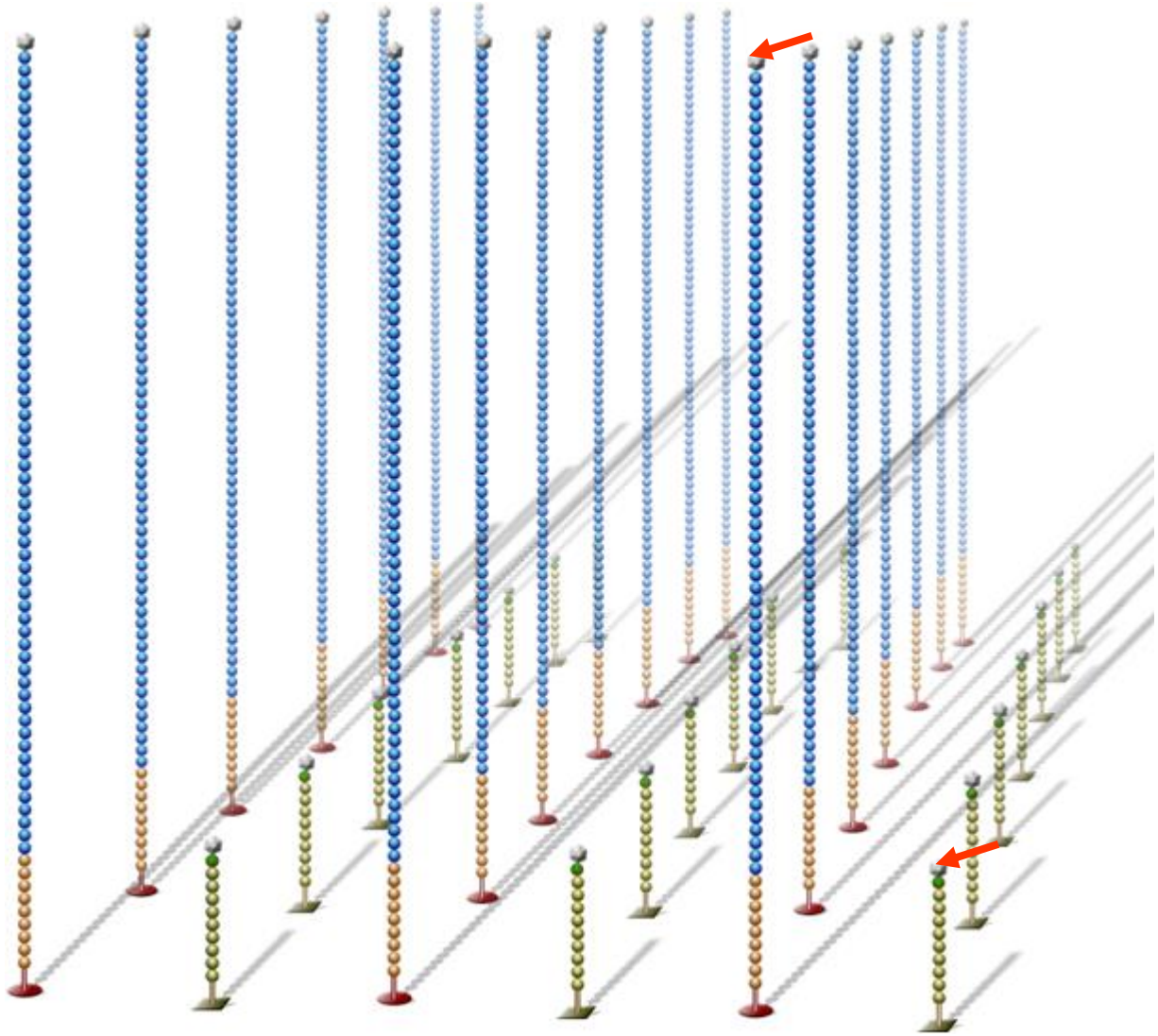
Reverse Strand Cleavage

- Reverse strands are cleaved and washed away, leaving a cluster with forward strands only.



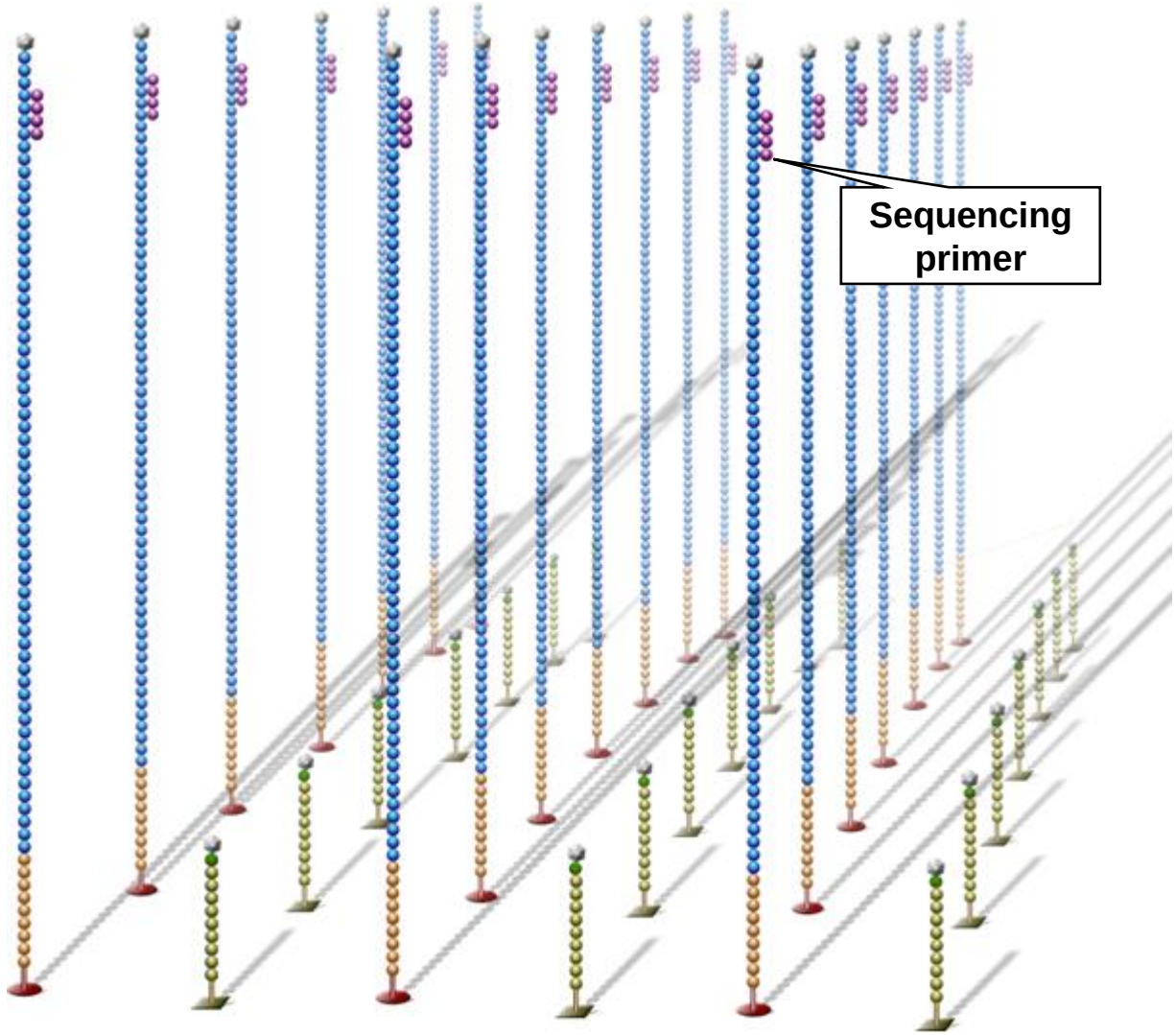
Blocking

- Free 3’ ends are blocked to prevent unwanted DNA priming.



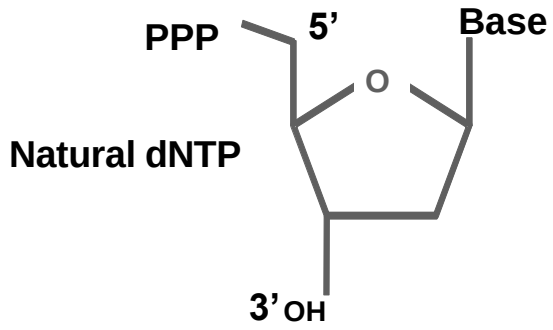
Read 1 Primer Hybridization

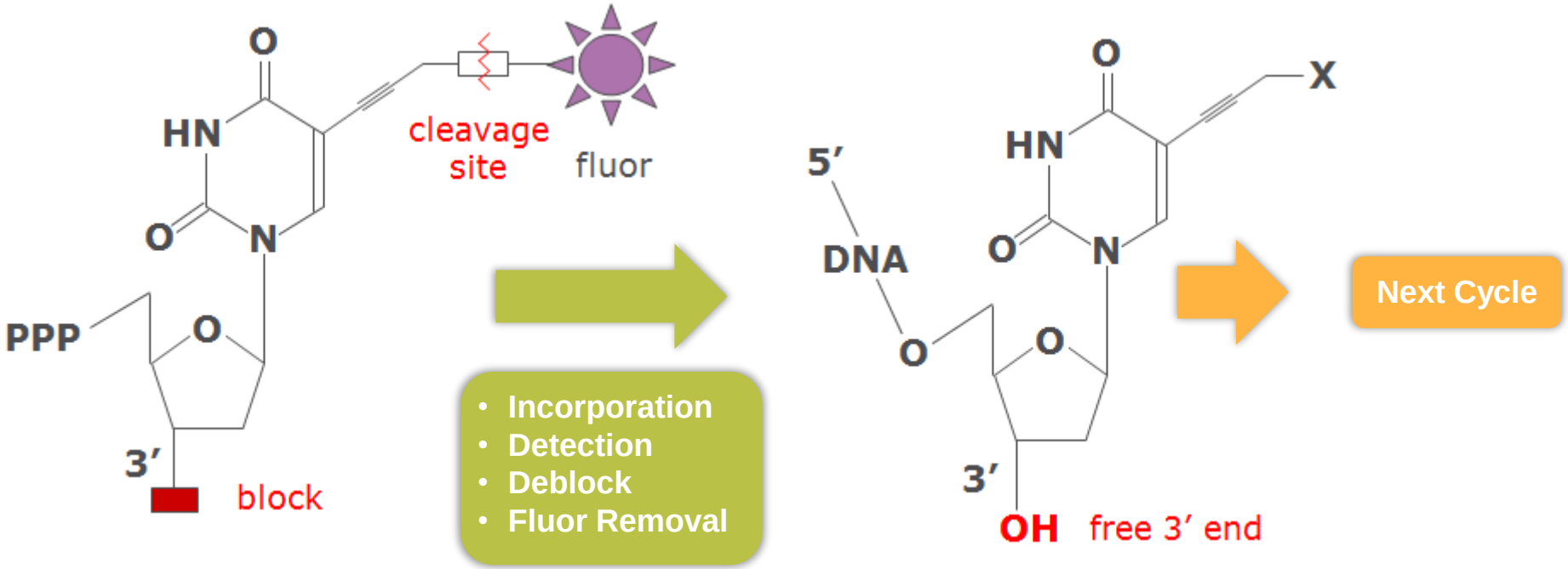
- Sequencing primer is hybridized to adapter sequence.



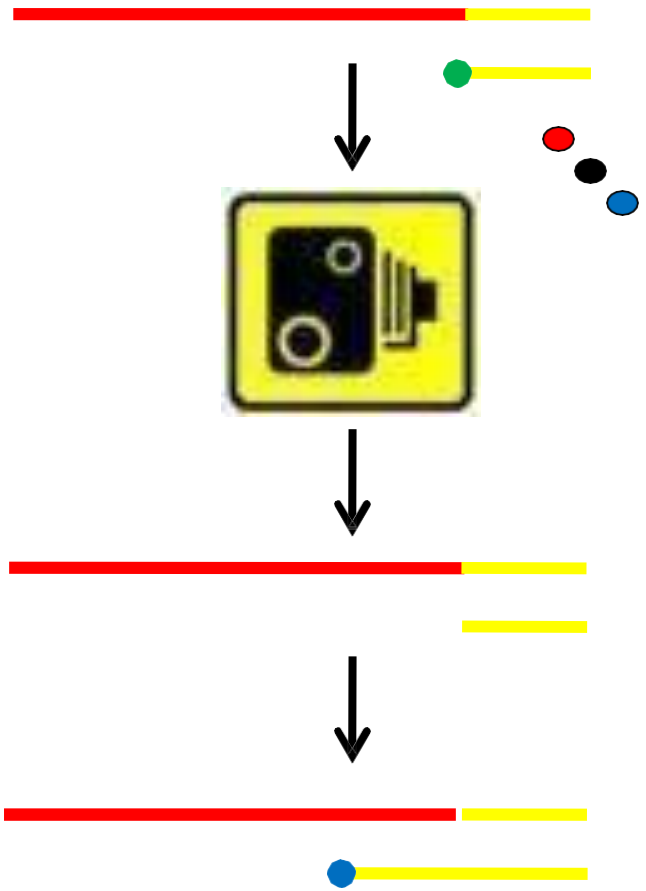
Illumina Modified Nucleotides

- Reversible terminator chemistry.
- All 4 nucleotides in one reaction.
- Higher accuracy.
- No problems with homopolymer repeats.



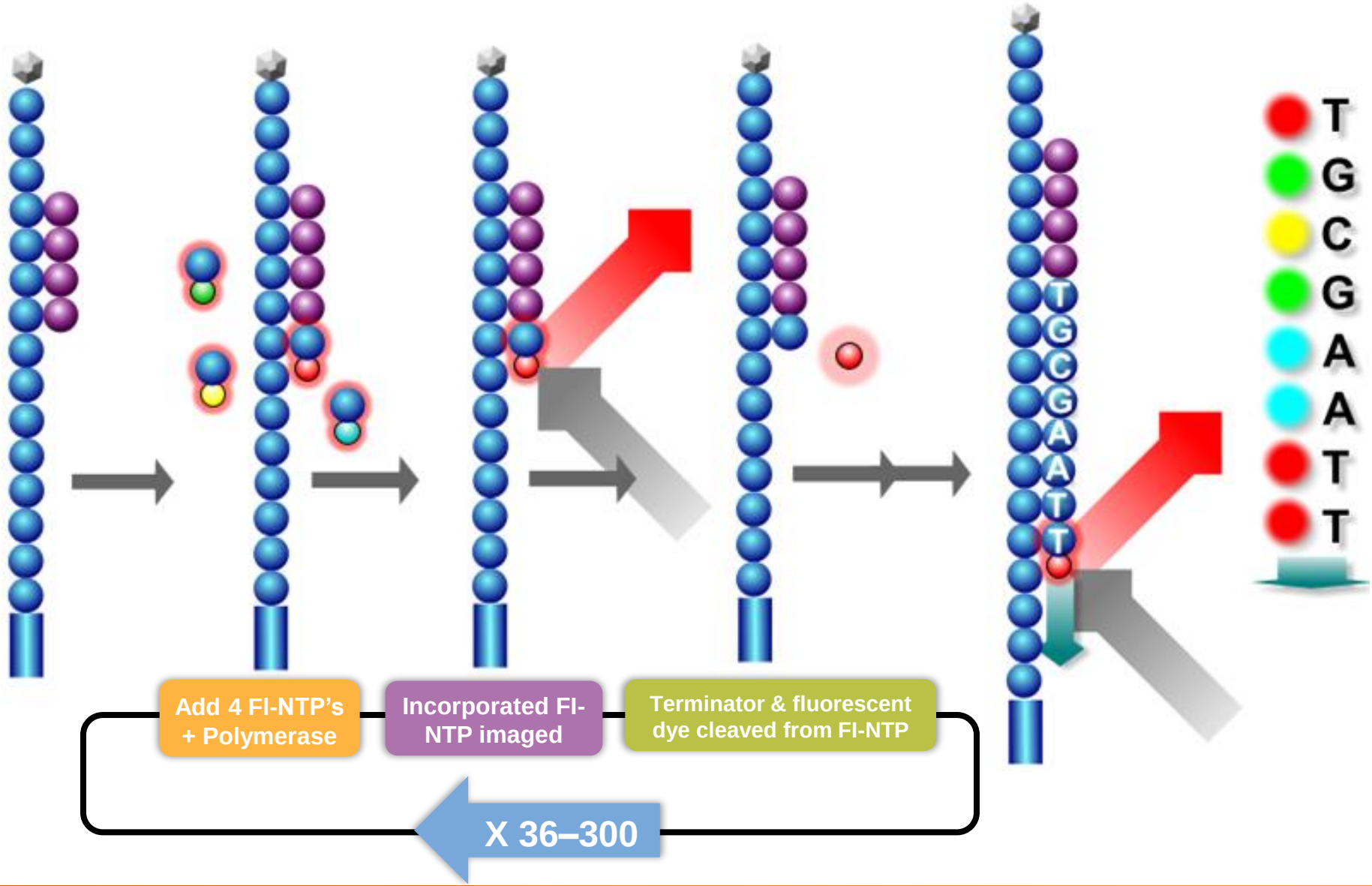


Iluumina Sequencing by Synthesis



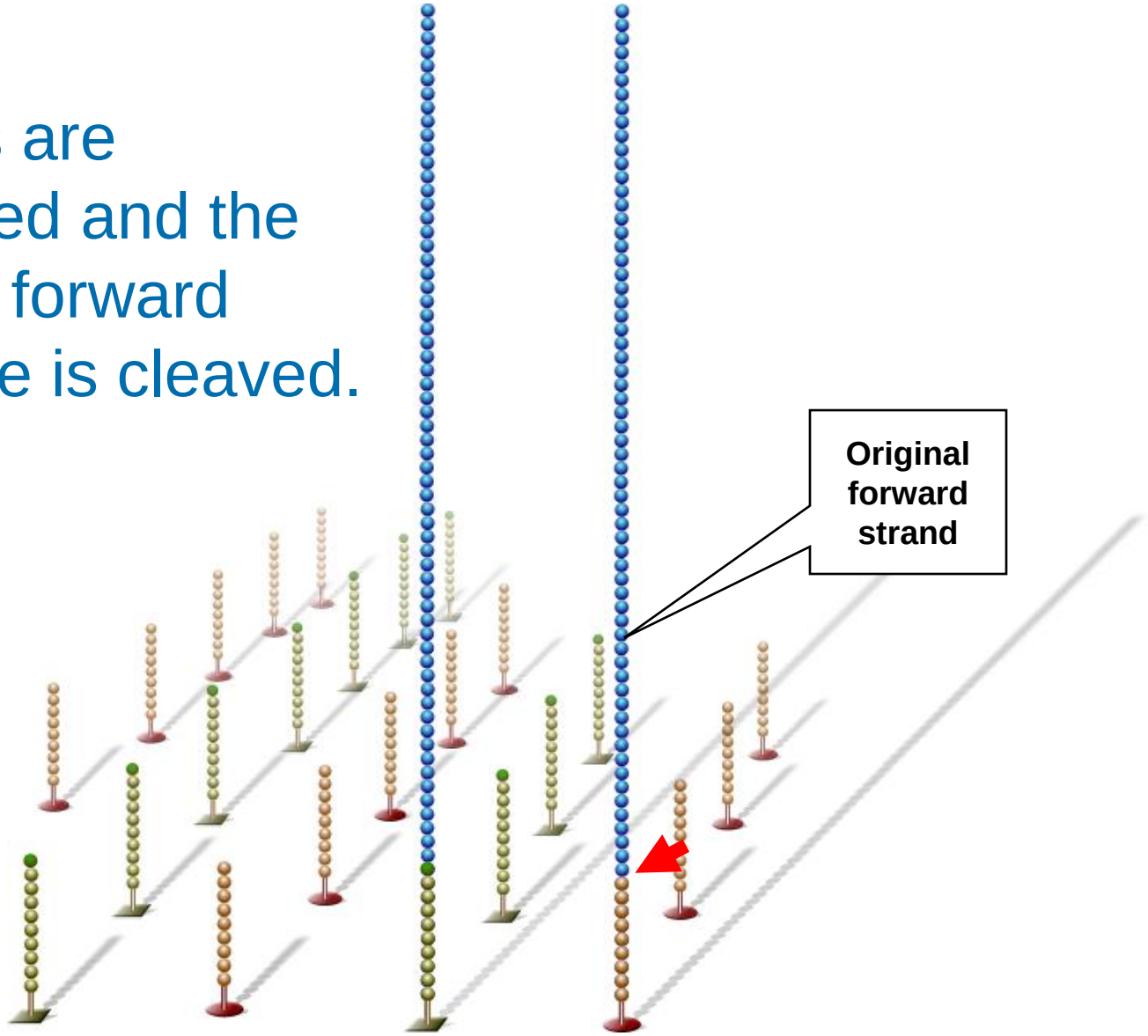
- Extend by 1 base.
- Image.
- Reverse termination.
- Repeat.

Sequencing By Synthesis (SBS)



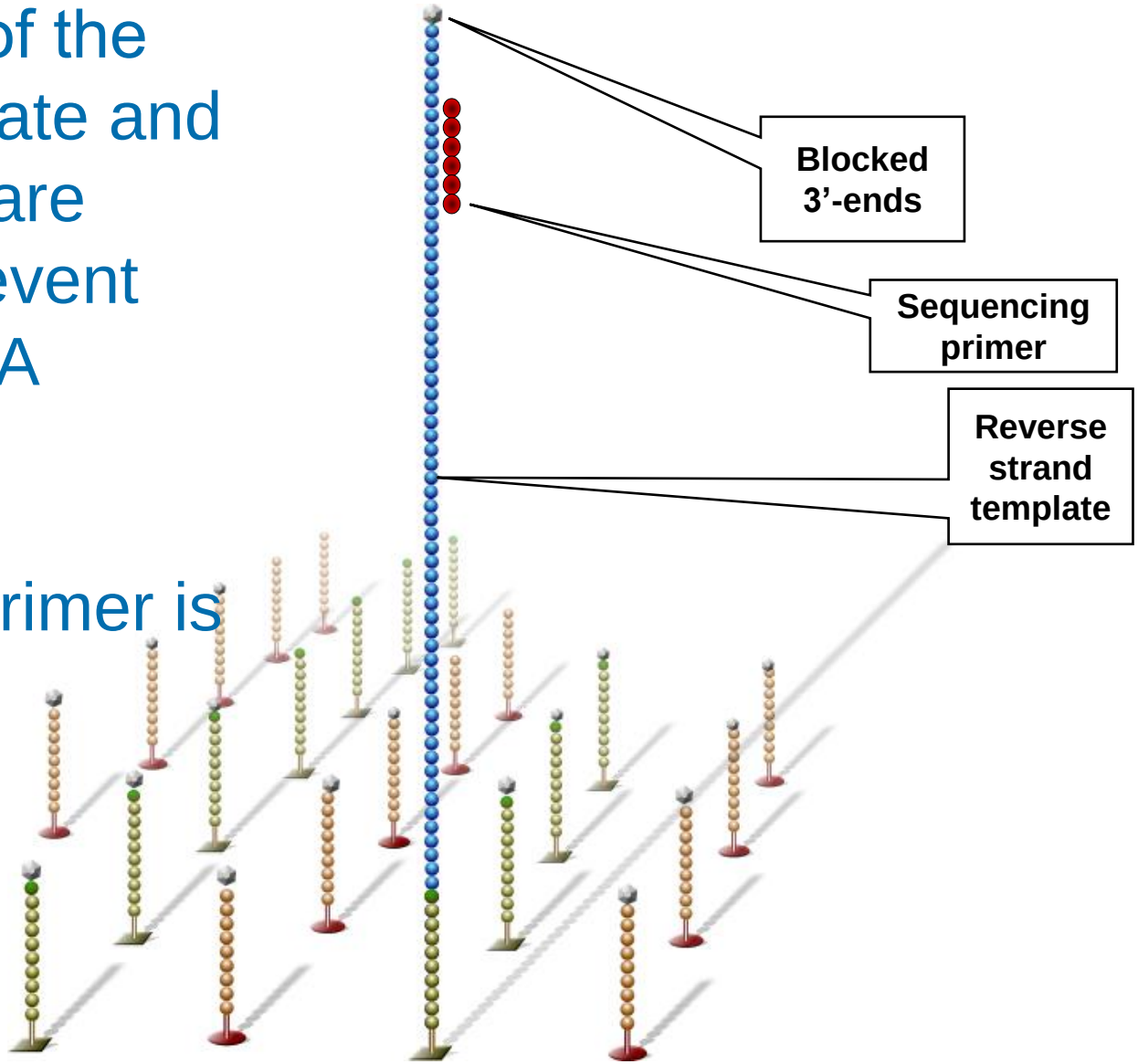
Forward Strand Cleavage

- Bridges are linearized and the original forward template is cleaved.

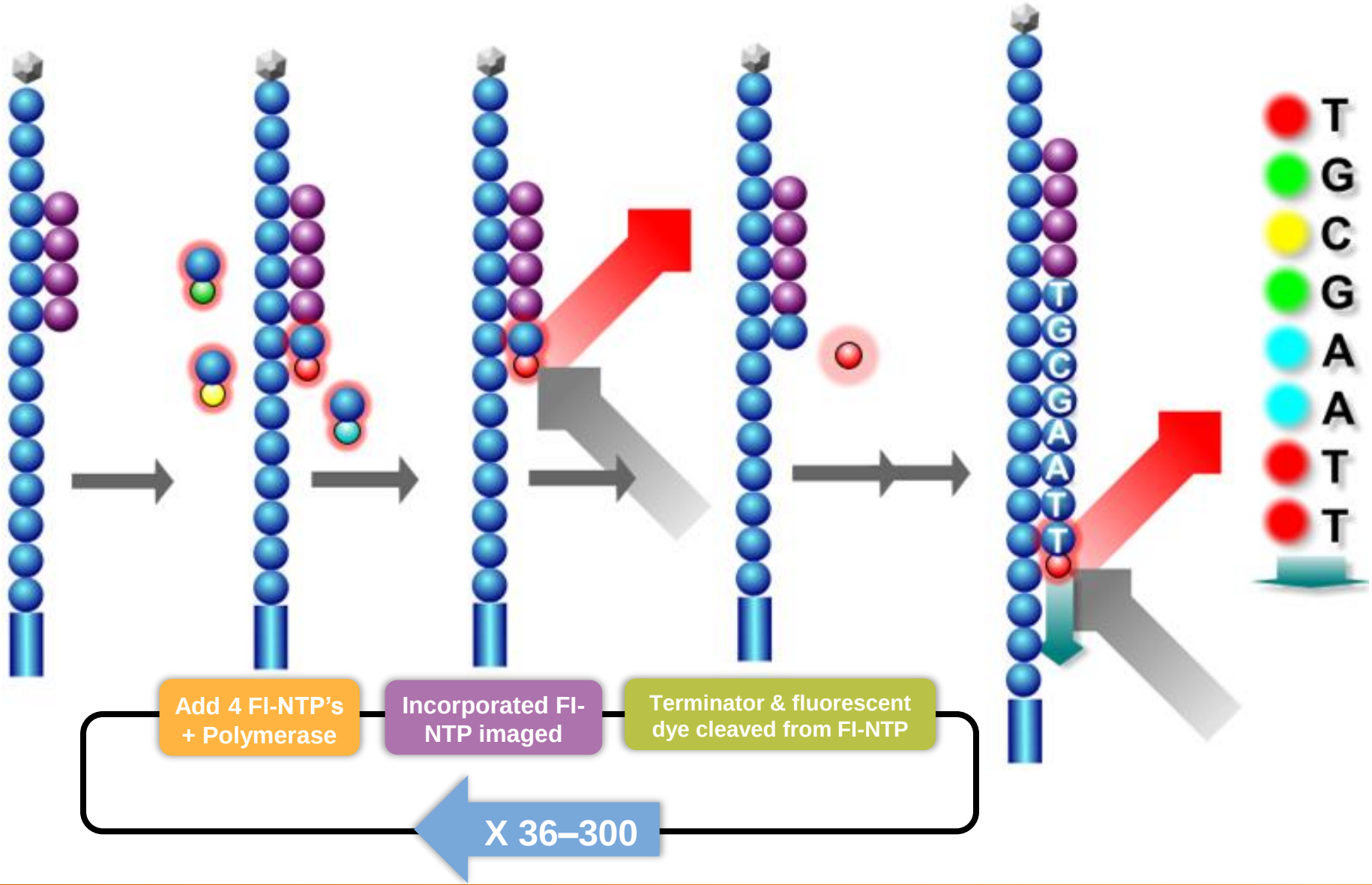


Read 2 Primer Hybridization

- Free 3’ ends of the reverse template and lawn primers are blocked to prevent unwanted DNA priming.
- Sequencing primer is hybridized to adapter sequence.

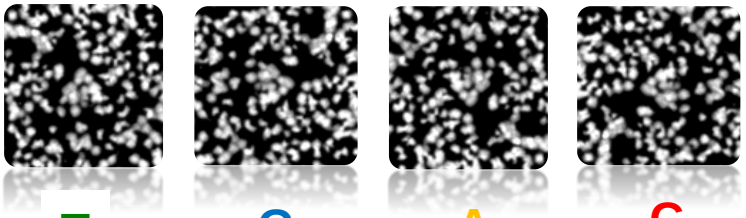


Sequencing By Synthesis (SBS)



Illumina Sequence Capture

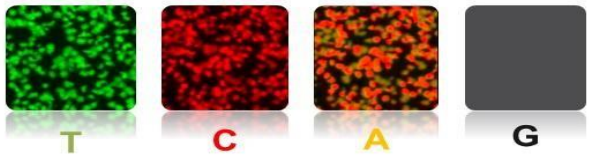
- 4 channel sequencing uses a different fluorescently labeled nucleotide for each base.
- 4 images are required to complete sequencing. Used by GA, HiSeq and MiSeq systems.



Green Laser identifies G and T
Red Laser identifies A and C

- 2 images are required to complete sequencing in NovaSeq, NextSeq and MinSeq systems.

- **Used by NovaSeq, NextSeq and MiniSeq**

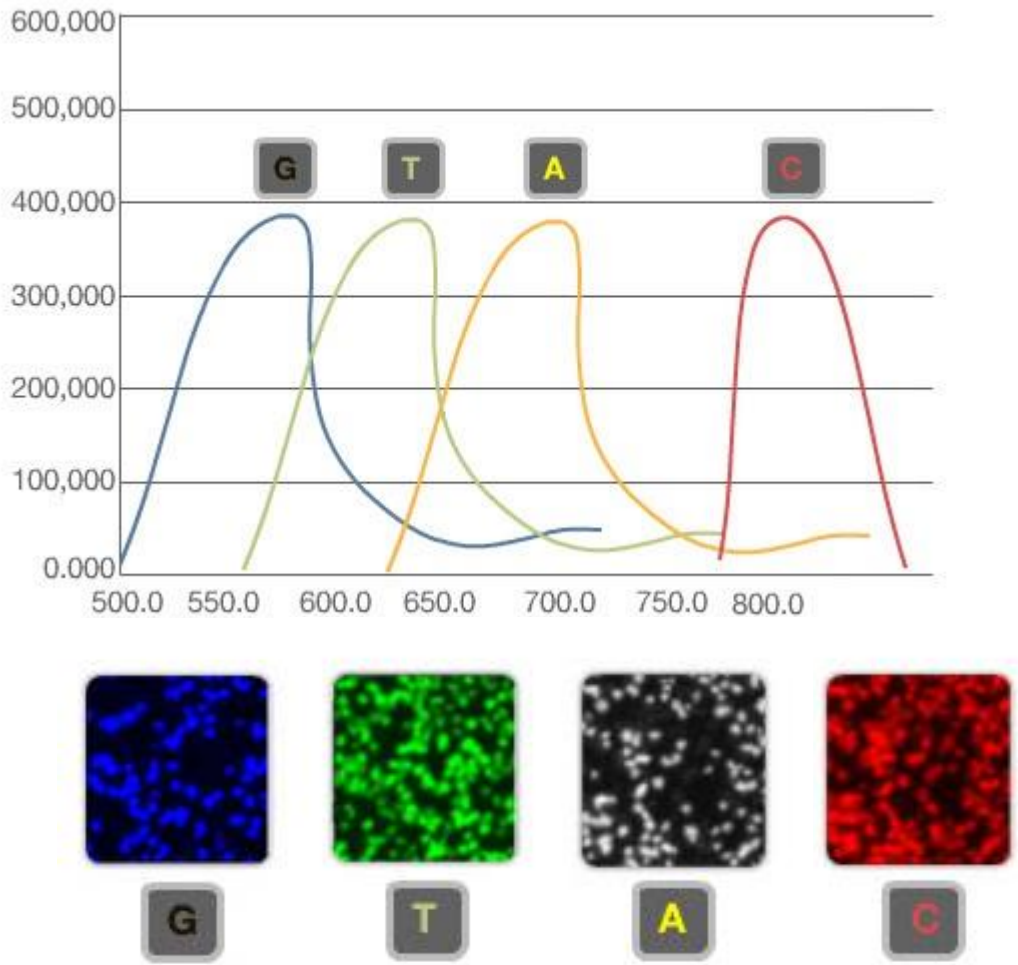


Channel	Green	Red	Green and Red	Dark - Neither

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

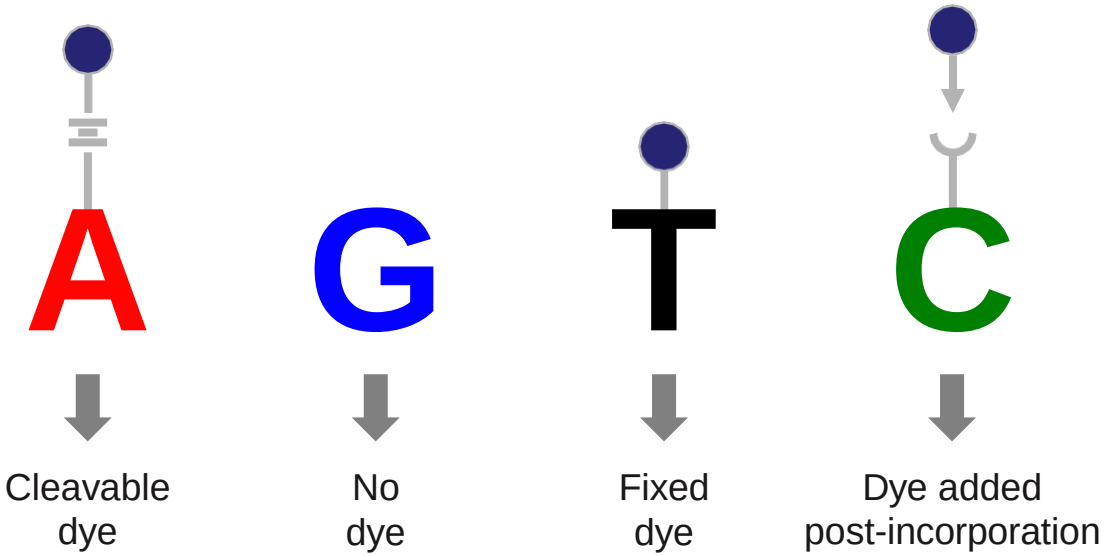
4-Channel SBS Chemistry

- Each of the four DNA bases emit an intensity of a unique wavelength
- Four images are collected during each cycle, each cluster appears in only one of four images.



iSeq 100 1-Dye Chemistry

What's Different?

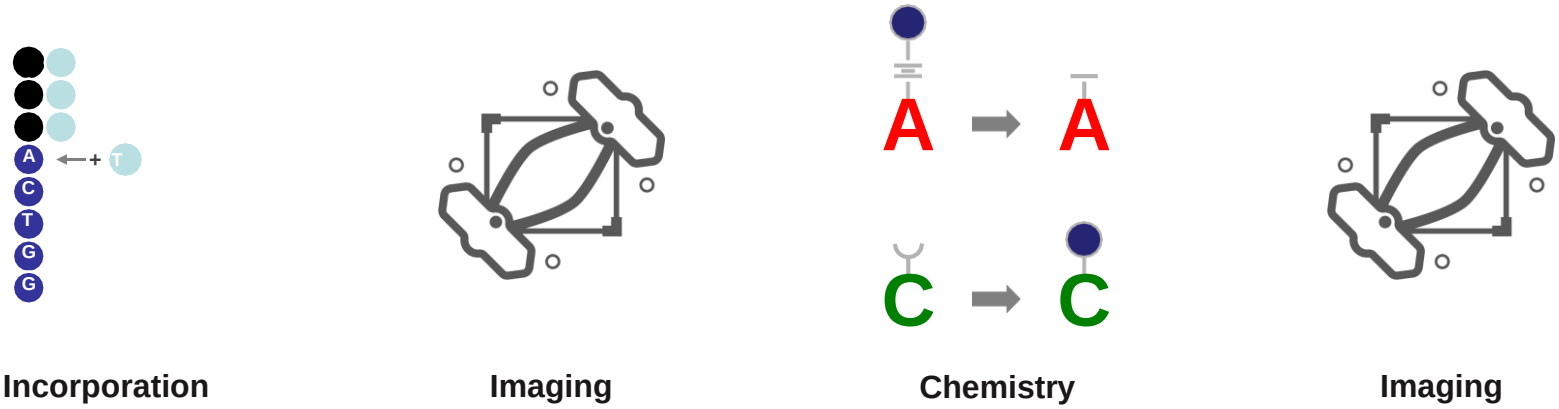


Nucleotides are labeled with a single dye, with the exception of the **G nucleotide**

iSeq 100 Sequencing

Sequencing by Synthesis

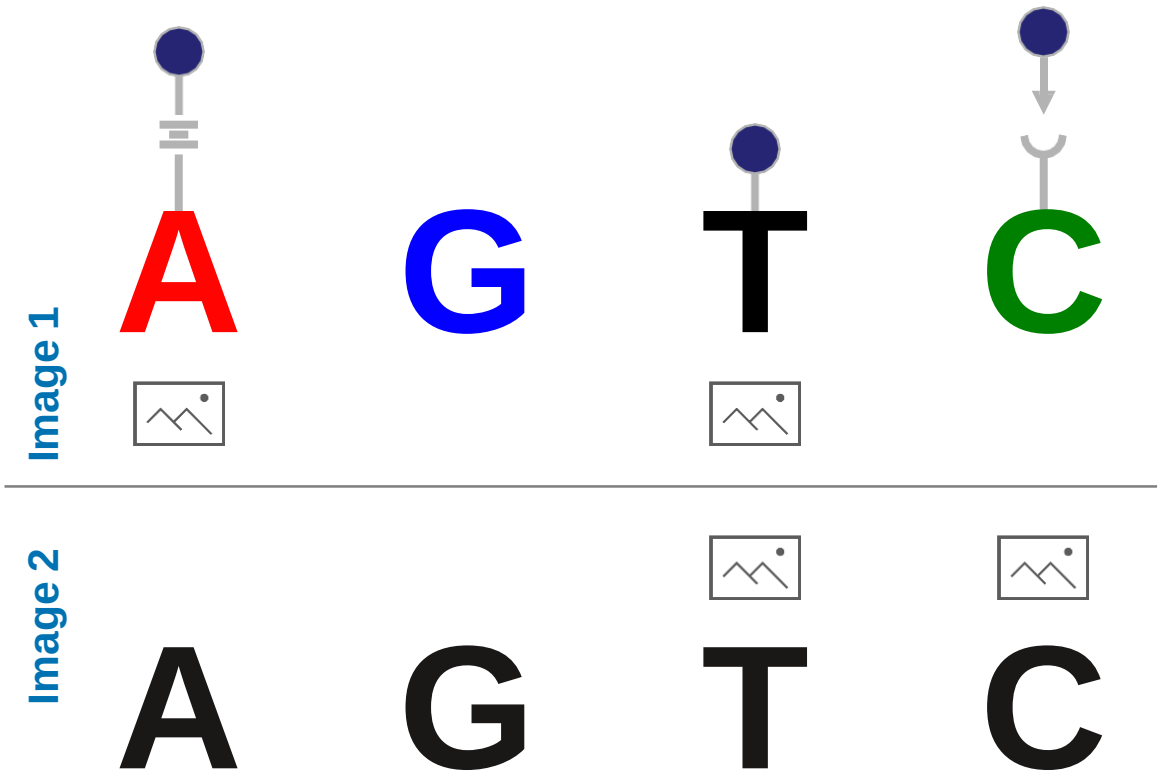
Sequencing Cycle



Incorporation **Imaging** **Chemistry** **Imaging**

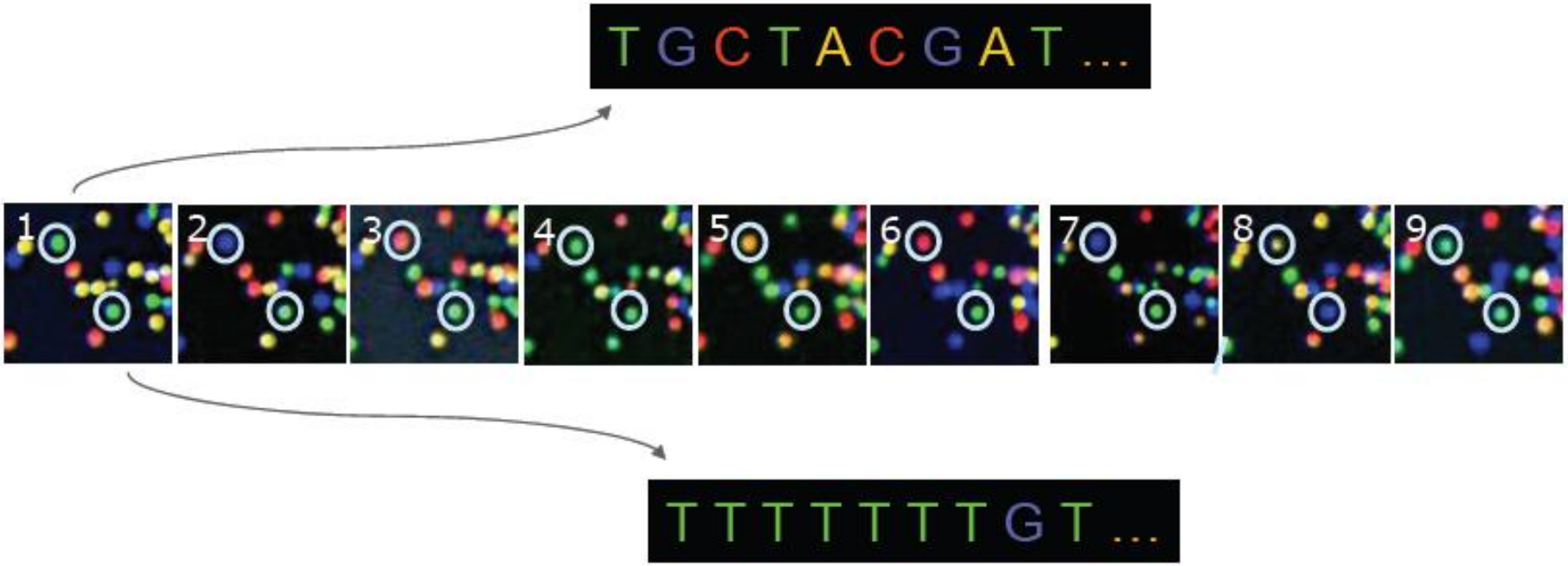
- An intermediate chemistry step, which **removes the dye from the A nucleotide** and **adds a dye to the C nucleotide**, separates the two images.

iSeq 100 Imaging



Using the two images, the iSeq™ 100 innovative data processing approach uniquely determines **which nucleotide was added to the growing template strand**

Illumina Base Calling



Each base of a cluster is read off from sequential images

Primary Data Analysis Workflow

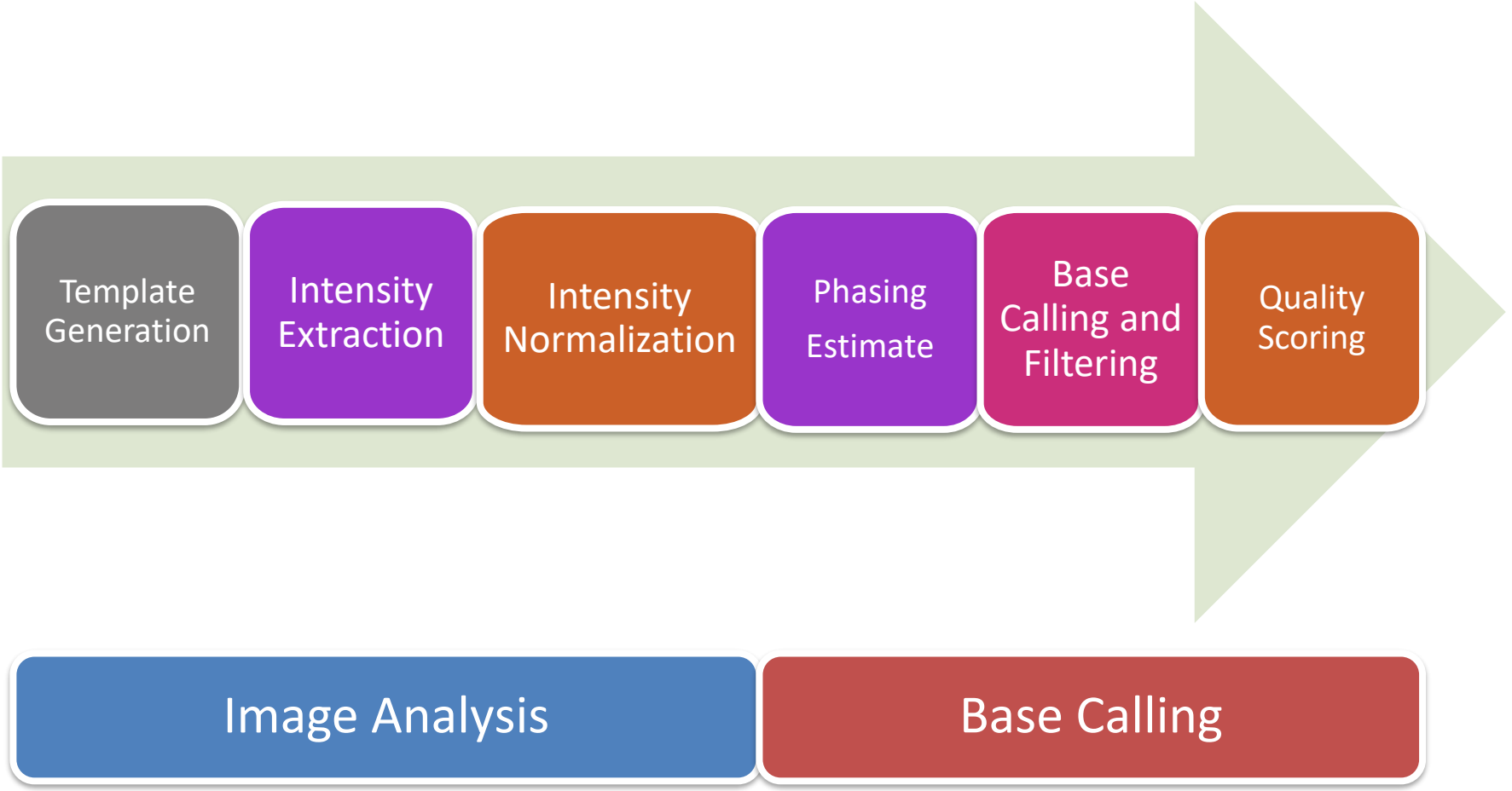
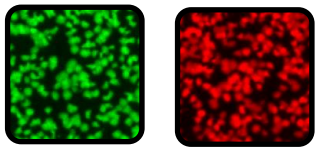


Image Analysis Overview



Images

RunInfo.xml

RunInfo.xml

Image Analysis

Template Generation

Extract Intensities

Normalize Intensities

Locate clusters

Correct Signal

Cluster Locations

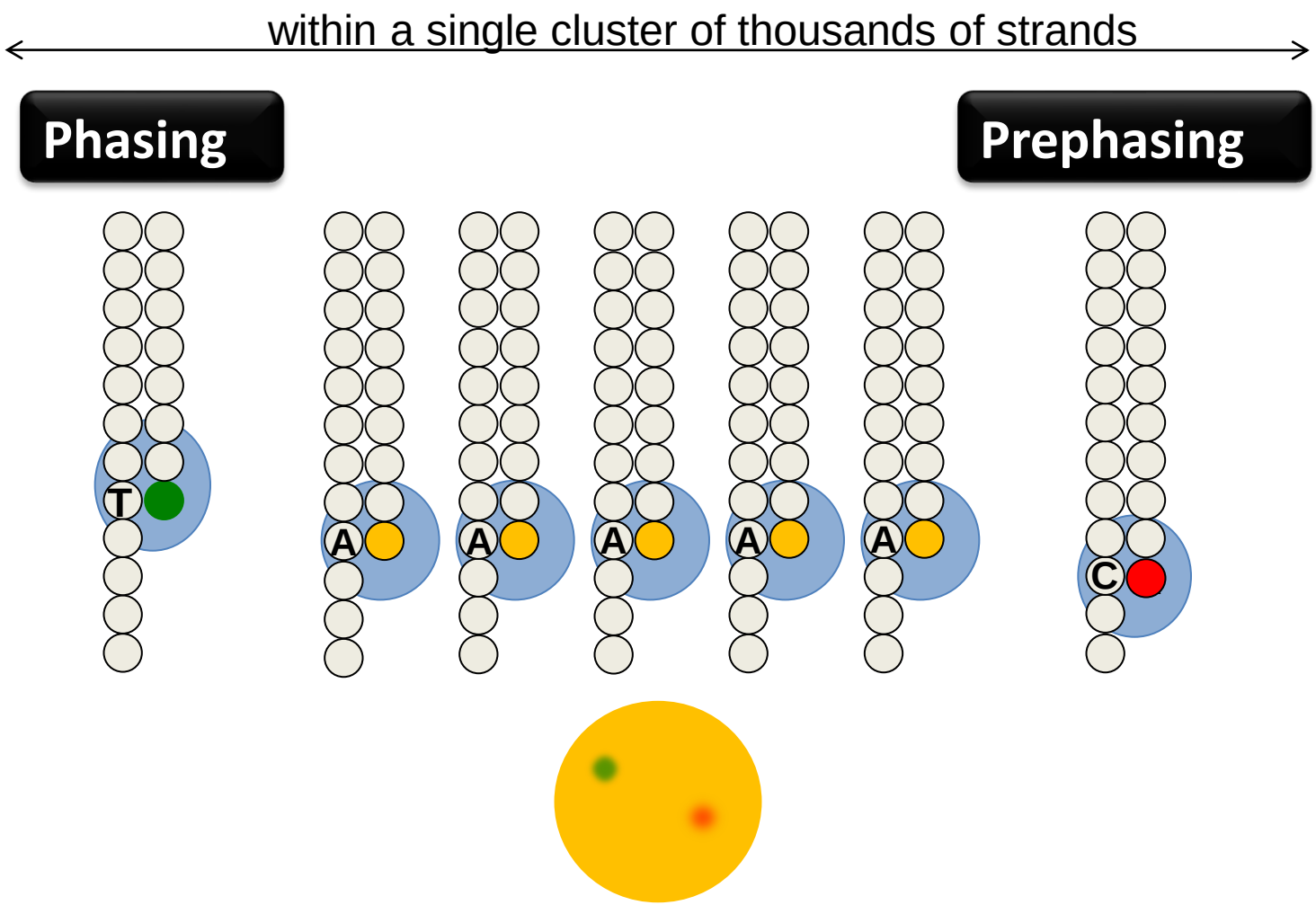
.clocs

.cif

Cluster Intensities

Base Calling Input and Output

Base Calling Phasing Correction

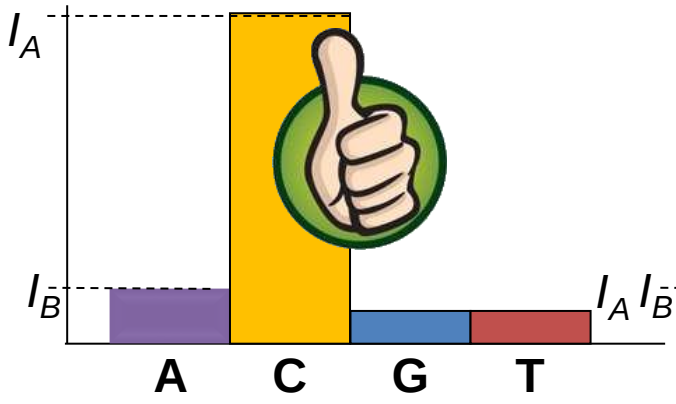


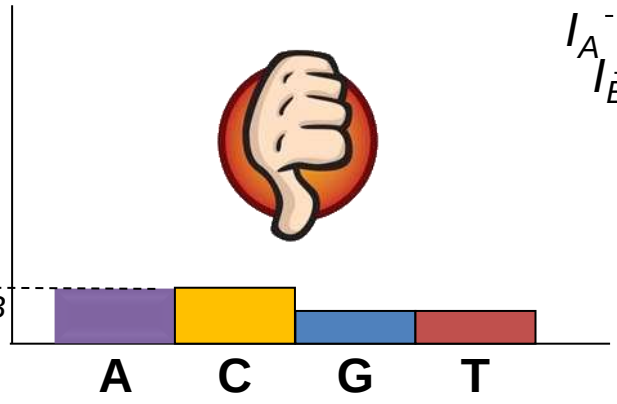
Clusters Passing Filter

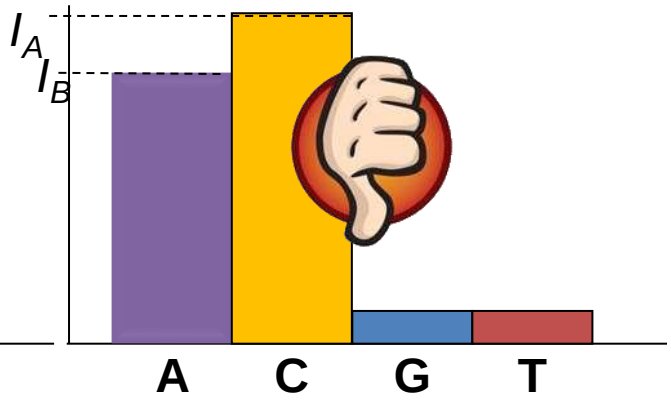
Pass filter is:

$$C = \frac{I_A}{I_A + I_B}$$

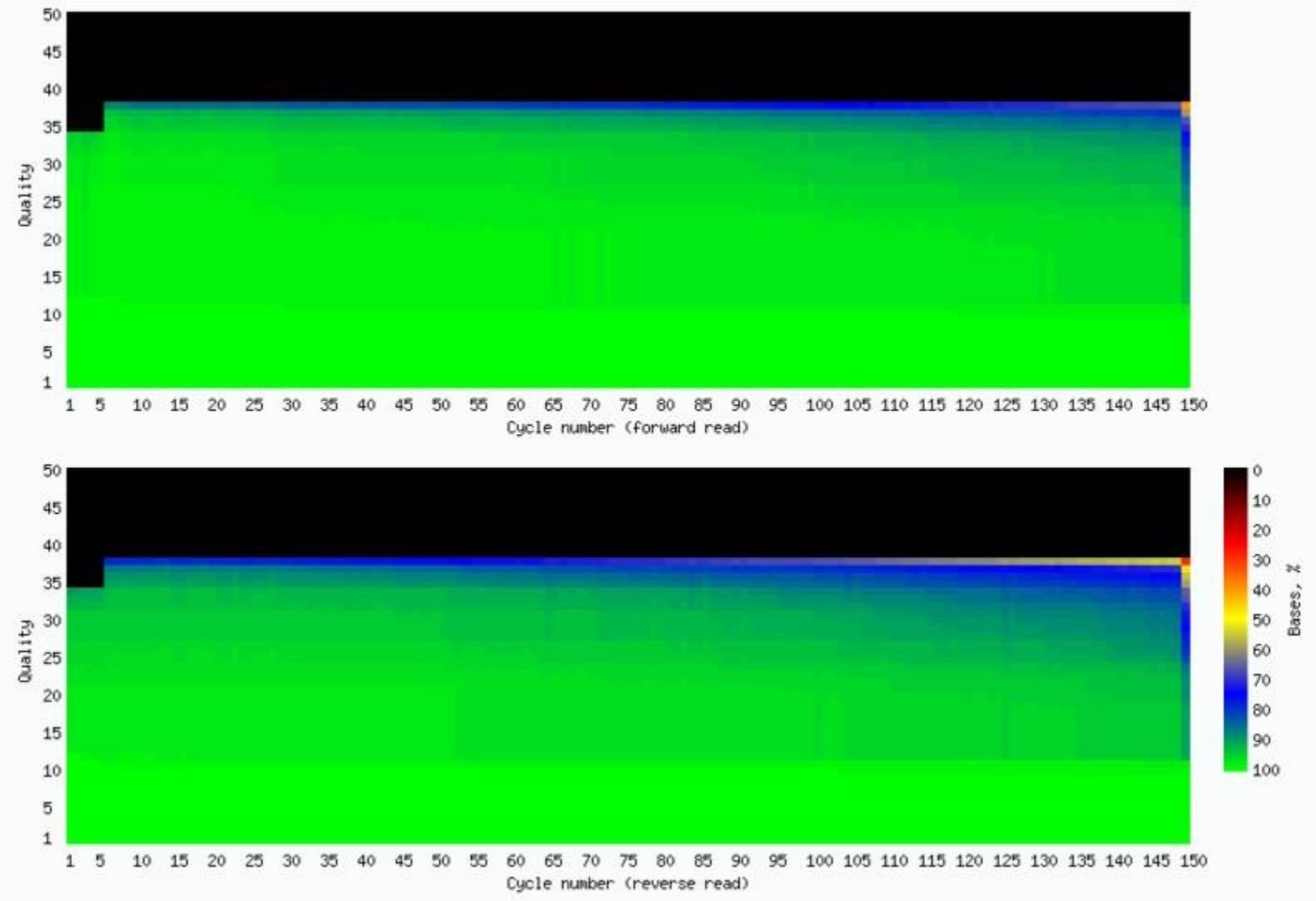
- The highest intensity over the sum of the 2 highest intensities
- Calculated for each cluster over the first 25 bases of the sequence
- A filter to remove overlapping and low-intensity clusters



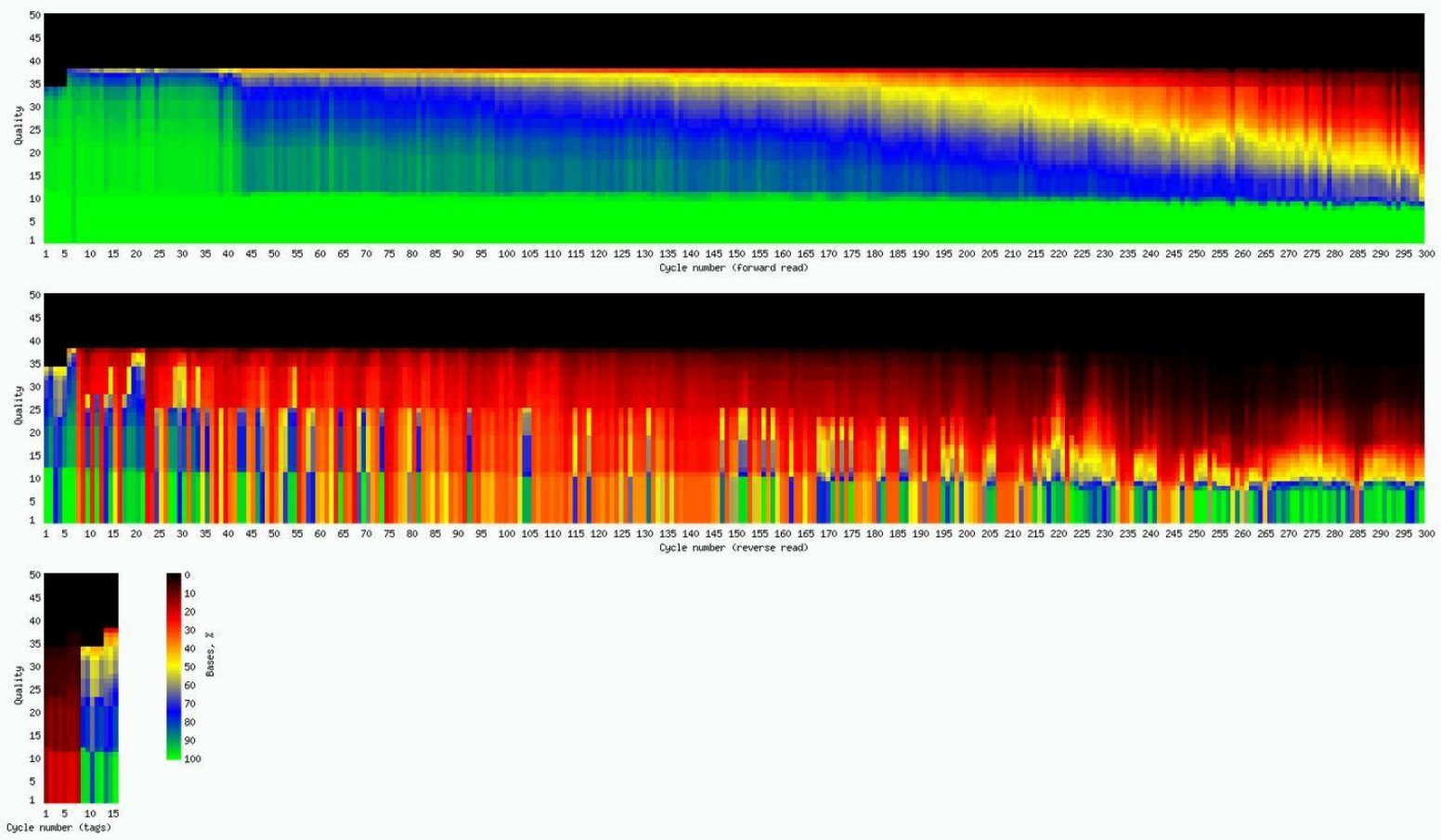




Error Limits Read Length



Low Quality 16S Run



Phred Quality Scores

- **Assess/measure accuracy of base calling.**
- **Defined as a property related to the base calling error probabilities (P):**

$$Q = -10 \log_{10}(P)$$

Quality Scores	• Estimate the probability of an error in base calling based on a quality model
Quality model	• Includes quality predictors of single bases, neighboring bases and reads
Reported	• After Clusters passing filter calculation has completed cycle 25

Probability of Incorrect Base Calls

- Assess or measure accuracy of base calling.
- Sequence quality Q is reported on a log scale.
- Defined as a property related to the base calling error probabilities (P): $Q = -10 \log_{10}(P)$

Phred Quality Score (ASCII QS)	Probability of Incorrect Base Call	Call Accuracy (%)
Q10 (+)	1 in 10 bases	90
Q20 (5)	1 in 100 bases	99
Q30 (?)	1 in 1,000 bases	99.9
Q40 (I)	1 in 10,000 bases	99.99
Q50 (S)	1 in 100,000 bases	99.999

- **Q30 means that virtually all bases in a read are called correctly.**

Illumina Platform Applications

Platform	iSeq 100	MiniSeq	MiSeq I100 i100 Plus	NextSeq	HiSeq	NovaSeq
Large Genomes				●	●	●
Small Genomes	●	●	●	●	●	●
Exome Sequencing				●	●	●
Targeted Resequencing	●	●	●	●	●	●
Transcriptome Sequencing				●	●	●
Gene Expression Profiling				●	●	●
miRNAs	●	●	●	●	●	●
DNA-Protein Interactions			●	●	●	●
Methylation Sequencing				●	●	●
16S Metagenomic sequencing		●	●	●	●	●

Limitations of Illumina Technology

- Some systematic errors:
  - Difficult to spot rare variants (<1%)
  - See Duplex seq by Scmitt *et al.*
- Low complexity templates.
  - Add complex library to 30%, phase ensure variation at start of read.
- Sequencing short fragments doesn't give any long range information.
- Index Hopping

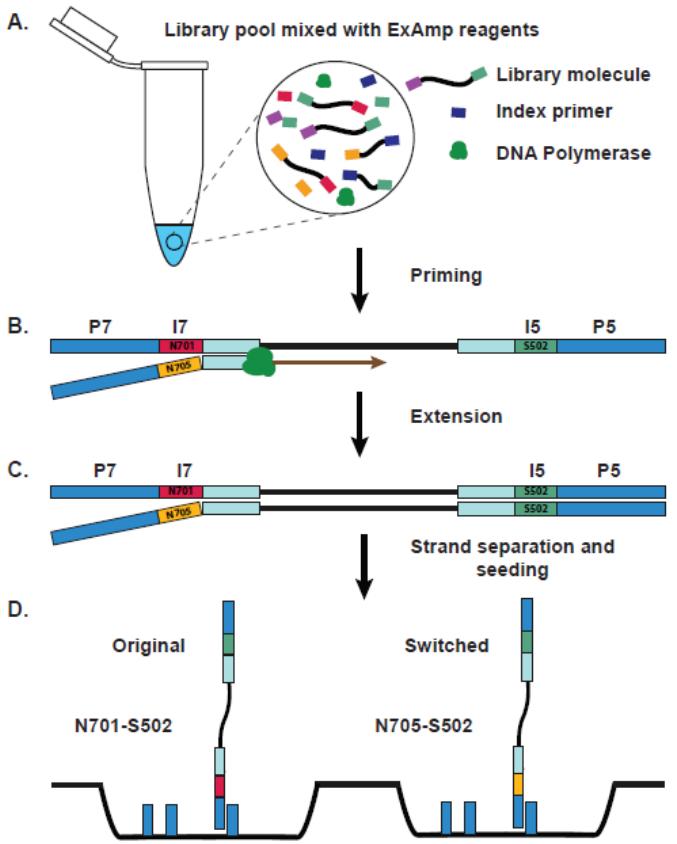


Figure 1

See Sinha et al BioRXIV 2017. http://biorxiv.org/content/early/2017/04/09/125724

GeneMind Platform Applications

Platform	FASTASeq300	GenoLabM	SurfSeq 5000
Output	7.5 -75 Gb	18 -300	50 Gb – 2.2 TB
Large Genomes			●
Small Genomes	●	●	●
Exome Sequencing		●	●
Targeted Resequencing	●	●	●
Transcriptome Sequencing		●	●
Gene Expression Profiling		●	●
miRNAs	●	●	●
DNA-Protein Interactions		●	●
Methylation Sequencing		●	●
16S Metagenomic sequencing	●	●	●

MGI Tech - Complete Genomics

Combinatorial Probe-Anchor Synthesis (cPAS) and DNA nanoballs sequencing.

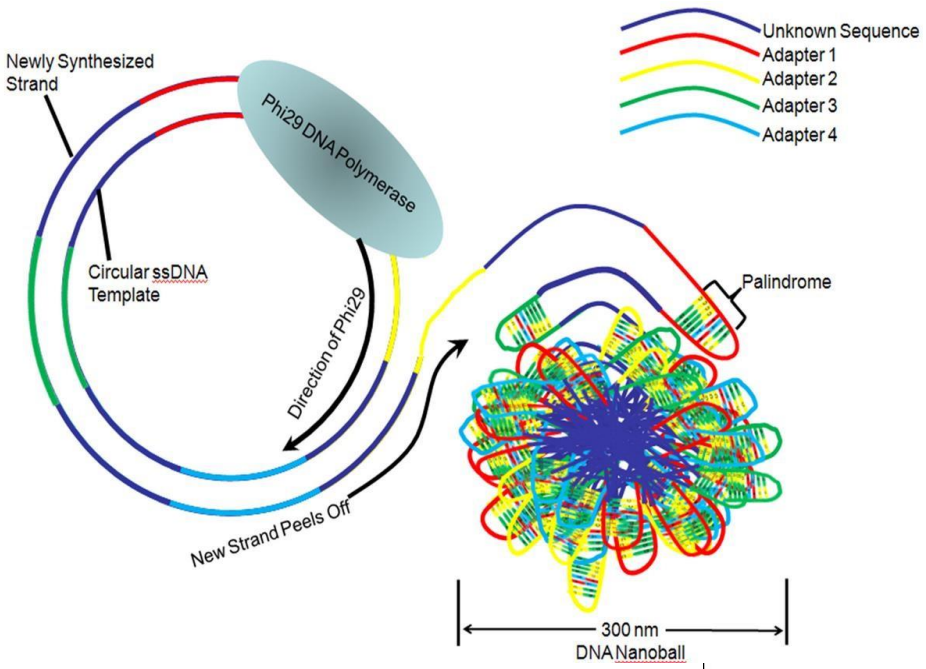
Drmanac et al. (2010) Science 327:78-81.

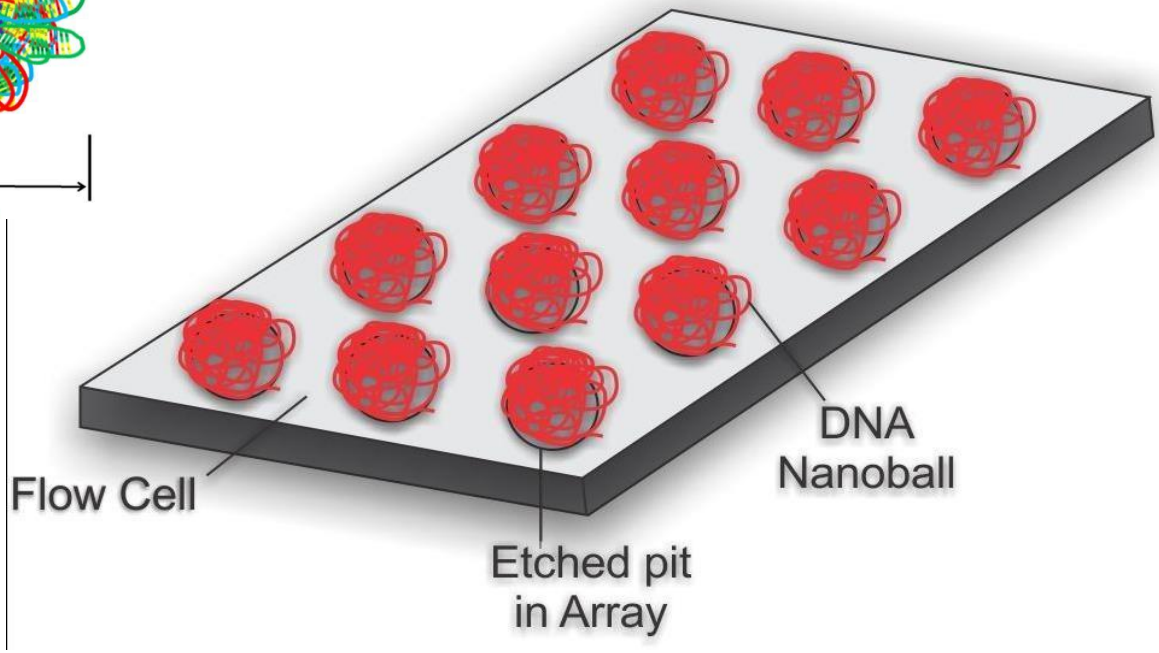
- Acquired Complete Genomics.
- Making sequencers for the Chinese market.
- Short read sequencers.
- Accurate

MGI Tech Instruments

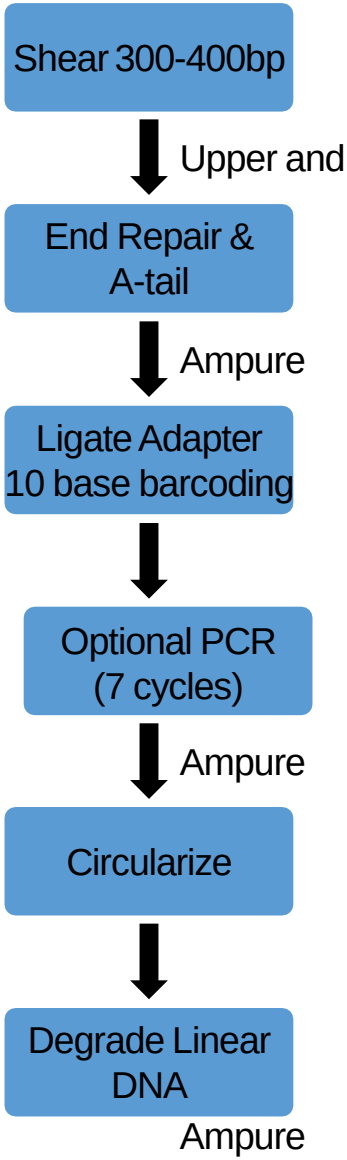
Platform	BGISEQ50	BGISEQ500	DNBSEQ T/G/E	MGISEQ T7
Large Genomes			●	●
Small Genomes	●	●	●	●
Exome Sequencing			●	●
Targeted Resequencing	●	●	●	●
Transcriptome Sequencing			●	●
Gene Expression Profiling			●	●
miRNAs	●	●	●	●
DNA-Protein Interactions			●	●
Methylation Sequencing			●	●
16S Metagenomic sequencing		●	●	●

DNA Nanoball Arrays on Flowcell

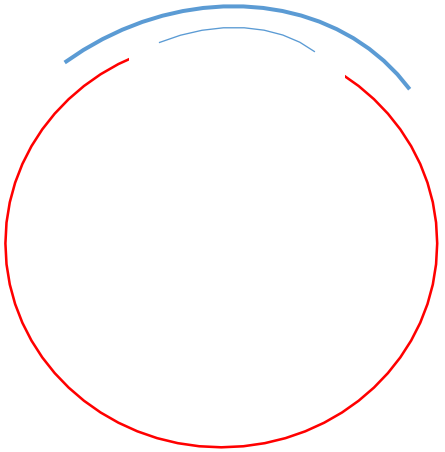




Basic MGI Library Preparation Schema



PCR free requires 1 µg DNA



DNA Nanoball Generation

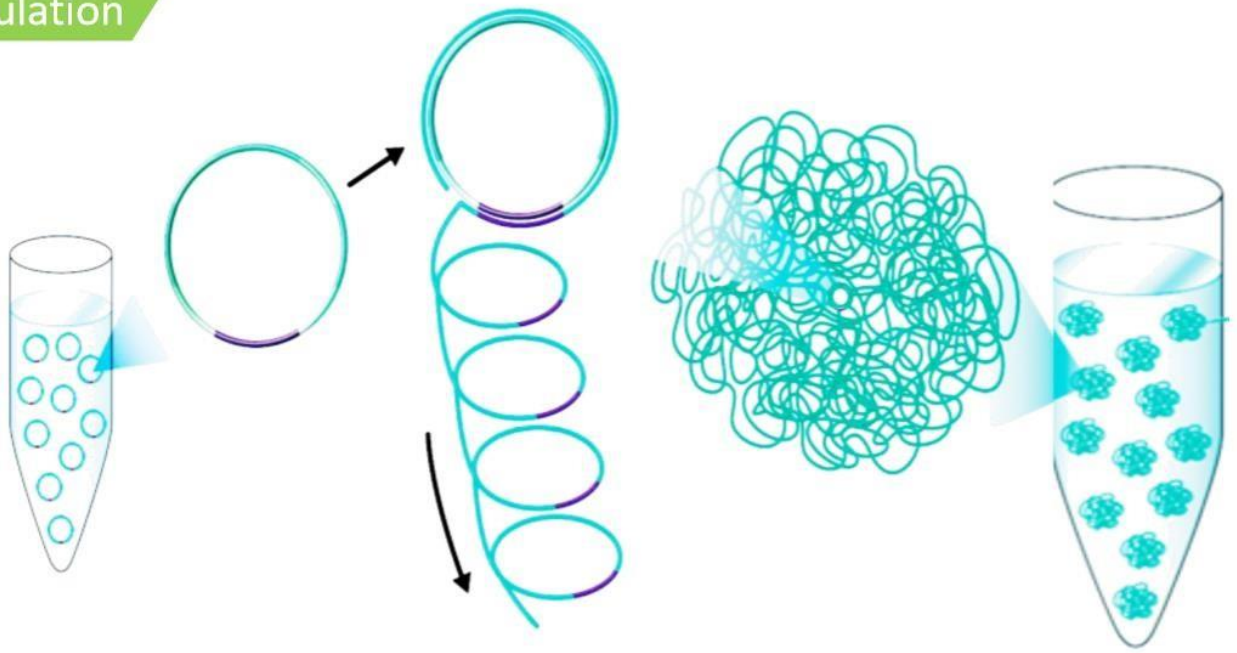
DNB Generation

Low amplification bias

No amplification error accumulation

Linear Amplification

Make DNA Nanoballs (DNB)



RCA

From the ssCirDNA library to DNBs.

From 1 copy to 300-500 copies .

DNA Sequencing

- DNBs loaded on silicon wafer with photomasking to create patterned array of binding sites.
- Primers anchored to DNBs.
- SBS using fluorescent reversible terminators.
- 2 or 4 colour chemistry.
- After sequencing use MDA to produce DNB from second strand allowing paired end sequencing.
- Reported <1% run failure and error rates.

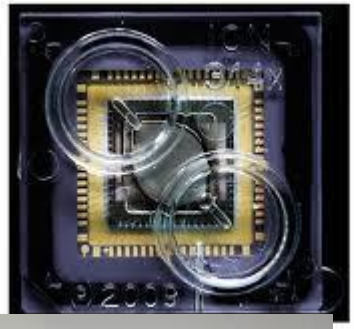
Advantages of MGI Sequencers

- Instruments cheaper than Illumina.
- Sequencing costs cheaper.
- Consistent yields.
- No index hopping.
- No fundamental bias.
- No fragment size dependent representation bias.
- Quality similar to Illumina
- 0.1 - 0.2% raw error

Ion Torrent

Semiconductor sequencing

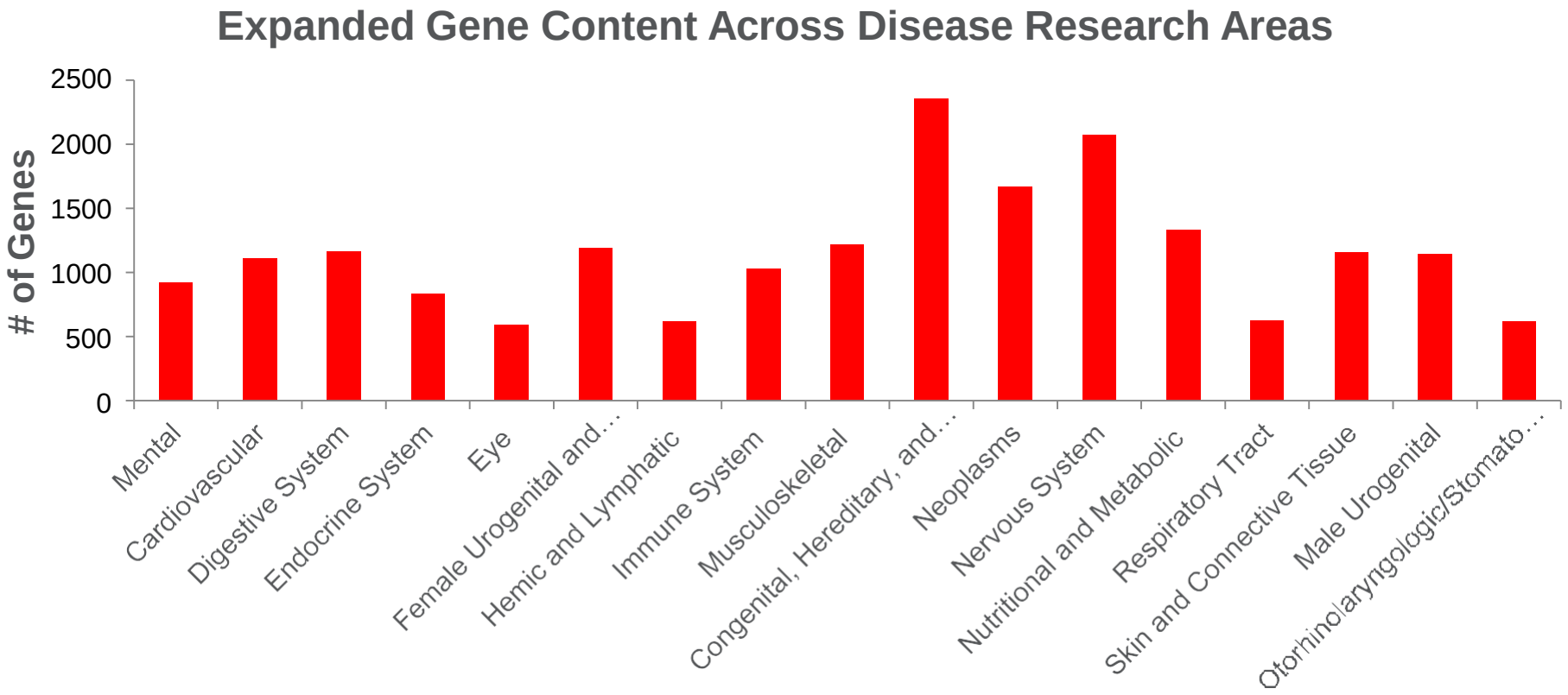
- Similar to 454 but detects H+ released as a base is added.
- Prone to errors near homopolymers.
- Not good for whole genome sequencing, but useful for targeted sequencing as run times are short.
- Used in a lot of clinical settings for disease panel sequencing.



Ion Torrent's PGM

Ion AmpliSeq On-Demand Panels

- Now ~5,000 pre-designed, pre-tested gene panels are available.



Average of 1154 genes per major UMLS disease research area category

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

Ion GeneStudio S5 Series

- Flexible instrument portfolio cconfigurable to your needs.

Ion GeneStudio™ S5	Ion GeneStudio™ S5 **Plus**	Ion GeneStudio™ S5 **Prime**
	New 	New 
Fast.	**Flexible.**	**Powerful**

Ion 510™ Chip	**Ion 520™ Chip**	**Ion 530™ Chip**	**Ion 540™ Chip**	New **Ion 550™ Chip**
				
2–3 M reads Up to 400 bp	3–6 M reads Up to 600 bp	15–20 M reads Up to 600 bp	60–80 M reads Up to 200 bp	100–130 M reads Up to 200 bp

For Research Use Only. Not for use in diagnostic procedures. * Throughputs based on 200bp sequencing

Ion GeneStudio S5 Series Comparison

- Output and turn-around time to meet your laboratory's peak volume needs.

	Ion GeneStudio™ S5	Ion GeneStudio™ S5 Plus	Ion GeneStudio™ S5 Prime
			
Speed*	19 hrs	10 hrs	6.5 hrs
Output (max/day):	15 Gb/80 M	30 Gb/160 M	50 Gb/260 M
Chips (max/day)**:**	1 x 540	**2 x 540** or 1 x 550	2 x 550

* Based off 540 chip – sequencing (2.5 hours) and analysis (varies) time

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

Ion GeneStudio Enables Quick Results

- ~45 minutes hands-on time from DNA to data with only 2 pipetting steps per sample.

Automated Ion AmpliSeq™ library construction		Automated template preparation		Sequencing	
Pipet sample and primer pools into Ion Chef™ cartridge for library prep	**Load cartridge** onto Ion Chef System	**Pipet** library into Ion Chef cartridge for templating and chip loading	**Load cartridge** onto Ion Chef System	**Load reagents** onto Ion GeneStudio™ S5 System	**Transfer chip** to Ion GeneStudio S5 System for sequencing
15 min hands-on time		15 min hands-on time		15 min hands-on time	

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

Ion GeneStudio Costs

List prices 2018

- Ion GeneStudio S5™ System A38194:
  - Runs 510, 520, 530 and 540 chip. **50,528 GBP.**
- Ion GeneStudio S5™ Plus System A38195:
  - Runs 510, 520, 530, 540 and 550 chip. **104,942 GBP.**
- Ion GeneStudio S5™ Prime System A38196:
  - Runs 510, 520, 530, 540 and 550 chip. **132,150 GBP.**
- Ion Chef(TM) System 4484177 **45,240 GBP.**
- All instruments include 12 months warranty.
- Promo: Trade-in any current NGS or CE instrument for up to 50% discount.

ThermoFisher SCIENTIFIC

Genexus

- US$299,000.
- 12-15 M reads in 14 - 30 hours from sample to insight.
- Oncomine knowledgebase.

Genexus System—Tomorrow's Specimen-to-Report NGS Workflow

Genexus Software

- FFPE tissue
- Frozen tissue
- Bone marrow
- Whole blood
- PBL
- Urine
- Saliva

Nucleic acid purification and quantitation*

Ion Torrent™ Genexus™
Purification System (Available 2020).

Up to **32** FFPE tissue samples with DNA **OR** RNA only input

2 hr TAT

Library preparation to variant interpretation | **Report***

Ion Torrent™ Genexus™
Integrated Sequencer (Available Nov 2019)

Ion Torrent™ GX5™ Chip: 12–15M reads/lane

14 hours for a single-lane run (approx. 24 to 30 hours for full chip)
Up to 32 Samples per run

* Specimen-to-report workflow available after Ion Torrent Genexus Purification System launches in 2020.
The content provided herein may relate to products that have not been officially released and is subject to change without notice.

ThermoFisher SCIENTIFIC

Oncomine Precision Assay

Oncomine Precision Assay on Ion Torrent Genexus System

Maximizes your ability to detect relevant variants

Curated pan-cancer content

- Mutations, CNVs, and fusion variant types across 50 key genes
- Tumor suppressors, drivers, and resistance variants

Tissue and plasma samples

- One test, one workflow, multiple sample types
- Maximizes the number of tumors that can be profiled

Molecular tagging

- Enhanced low-level variant detection
- Key for liquid biopsy testing

FusionSync™ Detection Technology

- Sensitive and specific—targeted isoform designs
- Novel fusion detection

The content provided herein may relate to products that have not been officially released and is subject to change without notice.

ThermoFisher SCIENTIFIC

Qiagen Genereader

- Sequencing by synthesis using reversible terminators developed by Intelligent Biosystems.
- Low output (1Gb). Optimised for use with Qiagen's targeted sequencing panels.
- End to end system with clinical reporting.

GenapSys

The GenapSys™ Sequencing Platform

Sequencing Without Compromise

Exceptional Accuracy
Highly accurate data validated by numerous applications in experienced third party labs

Unrivaled Scalability
A range of outputs tuned to your sample throughput needs

Amazing Affordability
Modest run and instrument pricing enable operational flexibility

GenapSys Sequencer

GenapSys Sequencing Prep System

A novel and robust Next-Generation Sequencing technology is here, with advantages for labs of all types. The compact GenapSys™ Sequencer combines electronic data detection, CMOS chip technology (Complementary Metal Oxide Semiconductor), and proven sequencing by synthesis (SBS) chemistry to deliver high accuracy data. The sequencer with the 16M chip generates 1.2 - 2.0 Gb of data per run and delivers the high resolution and analytical sensitivity needed for detection of rare variants and transcripts.

- Fast, cost-effective runs that match your sample throughput needs and applications
- Perform runs on your own schedule, without having to wait for additional sample batching
- Control the sequencing process from beginning to end for higher confidence in sample integrity and data analysis results
- Affordable list price puts sequencing within the reach of virtually any lab

Element Biosciences

US$ 2/Gb Sequencing

- Powered by Avidity Sequencing™
- AVITI chemistry harnesses the power of ultra-stable avidities to deliver high-quality data more cost-effectively than ever.
- Avidity Sequencing employs rolling circle amplification (RCA) to minimize common amplification errors.
- No PCR error propagation.
- Negligible index hopping.
- Low duplication rates.

Element AVITI™ FIT
Empowering More Researchers with Flexible Genomic Solutions

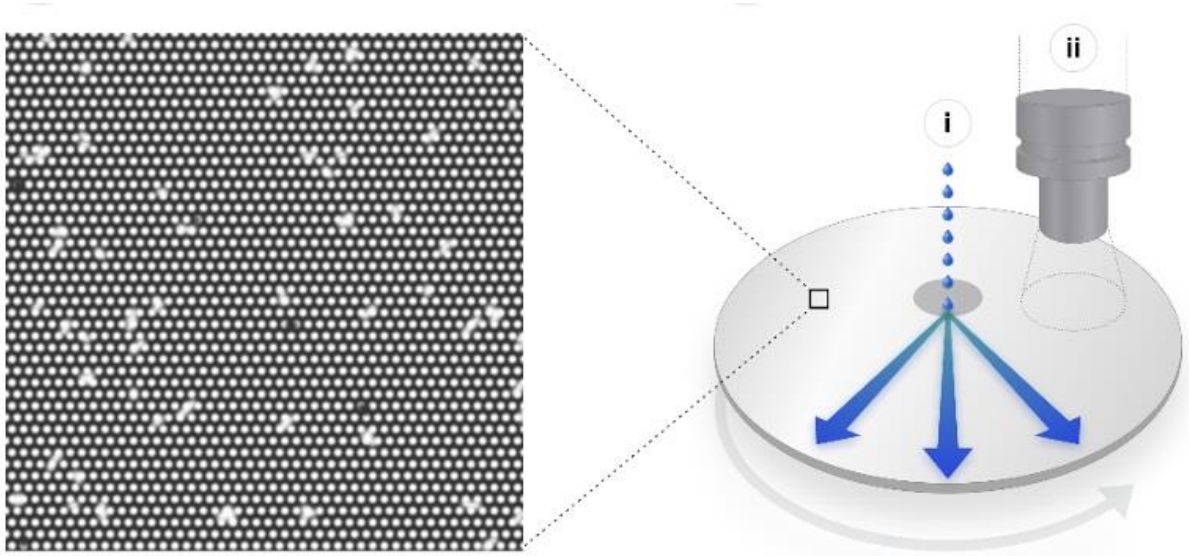
Compatibility			2 x 75 Kit		2 x 150 Kit		2 x 300	
Flow Cell	**AVITI**	**AVITI LT**	**Read Count***	**Q30**	**Read Count**	**Q30**	**Read Count**	**Q30**
Low Output	✓	✓			250 million			
Medium Output	✓	✓	500 million	> 90%	500 million	> 90%	100 million	> 85%
High Output	✓		1 billion	> 90%	1 billion	> 90%	300 million	> 85%

Aviti Systems

Ultima UG100™ Sequencer

US$ 1/Gb Sequencing

- Ultima's revolutionary new sequencing architecture replaces complex and expensive flow cells with a staple of the electronics industry, a 200 mm open silicon wafer.



UG100

PacBio Onso

- PacBio's Onso uses sequencing by binding with improved chemistry to obtain Q40 reads.

PacBio Onso

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

Long Read Sequencing

Long-Read Instruments

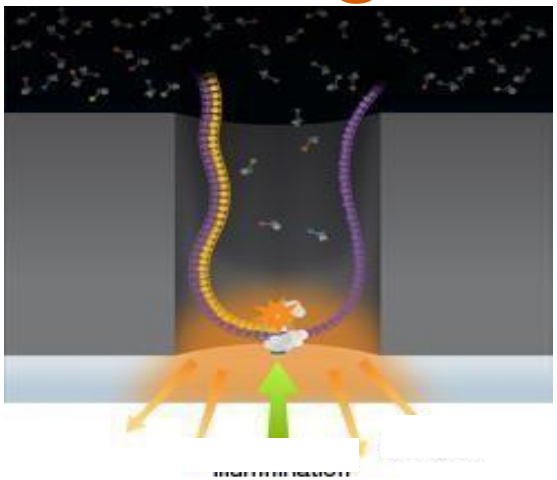
RSII : 1800 lbs. and ~11 feet long !

https://youtu.be/WMZmG00uhwU

PacBio Sequencers

Platform	Onso	Sequel I	Sequel II	Revlo		
Output (Gb)	80 – 150	30	30			400 Gb
Large Genomes			●	●	●	●
Small Genomes	●	●	●	●	●	●
Exome Sequencing			●	●	●	●
Targeted Re-sequencing	●	●	●	●	●	●
Transcriptome Sequencing		●	●	●	●	●
Gene Expression Profiling		●	●	●	●	●
DNA-Protein Interactions		●	●	●	●	●
Methylation Sequencing		●	●	●	●	●
16S Metagenomic sequencing		●	●	●	●	●

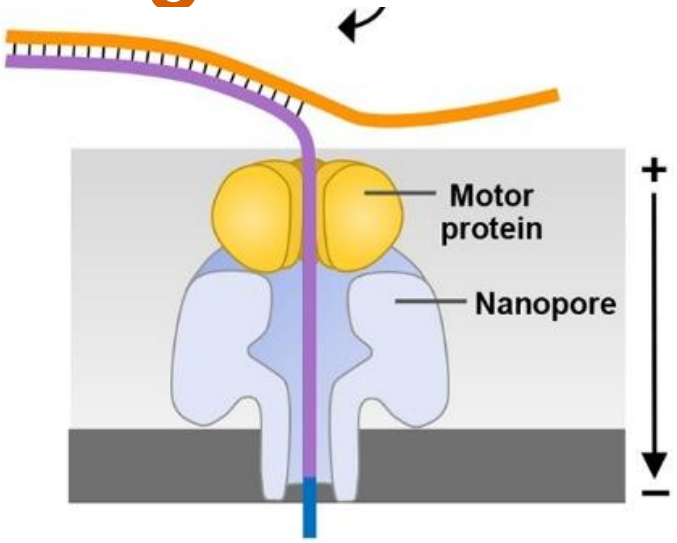
Long-Read Sequencing Flowcells



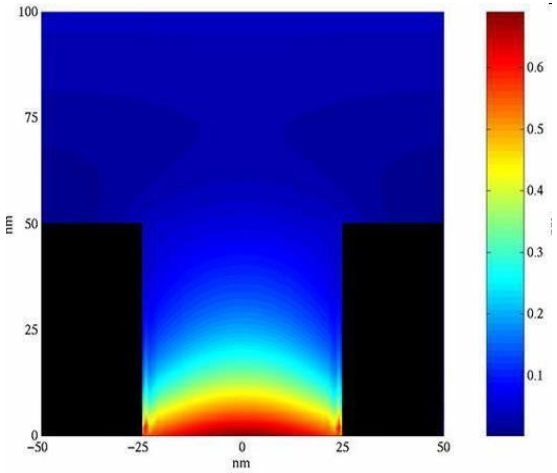
Individual ZMW

Single polymerase molecule bound to SMRT bell loaded in a 20 nm chamber, termed zero-mode waveguide (ZMW).

- Laser light illuminates and penetrates through the lower 20 - 30 nm of each ZMWs.
- Detection volume is 20 zeptoliter (10^{-21} liters).
- Records incorporation in real time - ~2 bases per second.



Nanopores through which nucleic acids pass.

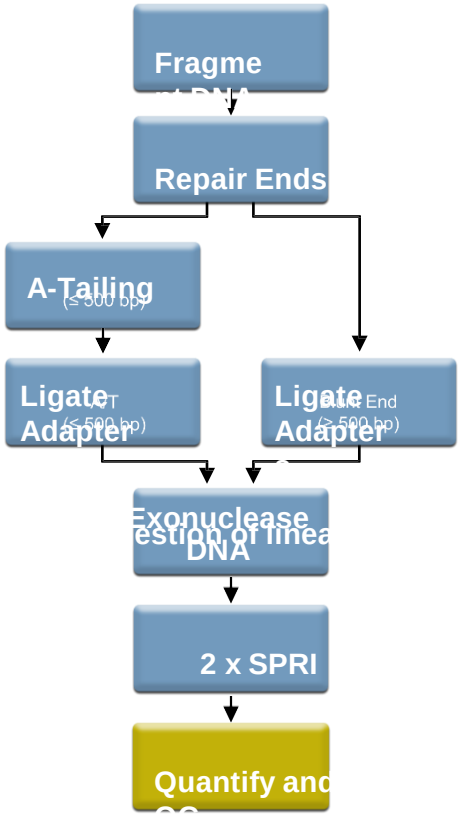


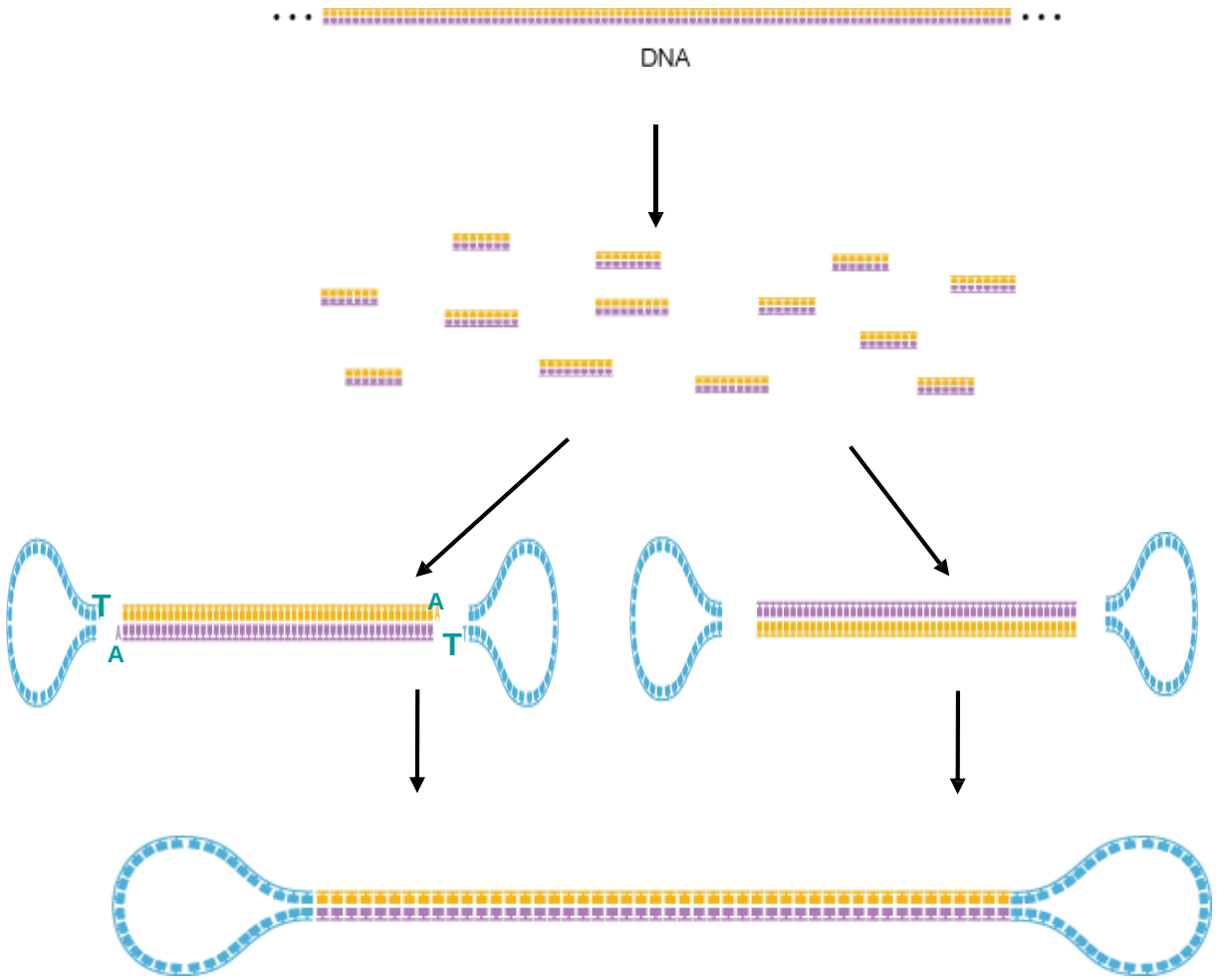
Laser light illuminates the ZMW

PacBio

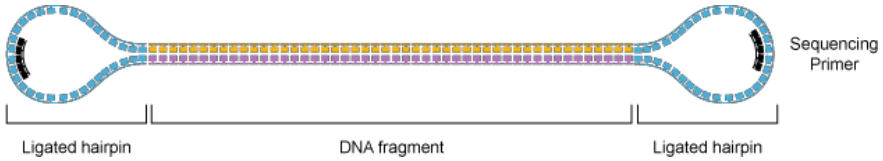
- Long Read sequencer.
- Single Molecule Real Time (SMRT).
- Current technology of choice for *de-novo* sequencing projects.

PacBio Library Prep

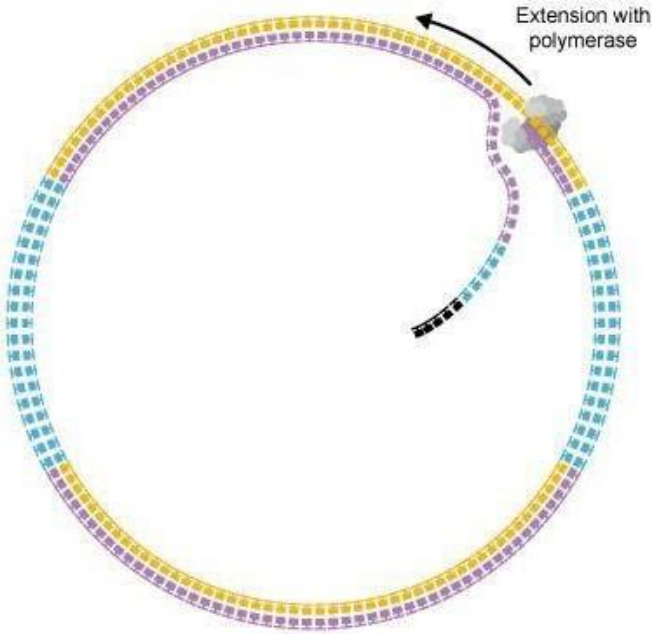




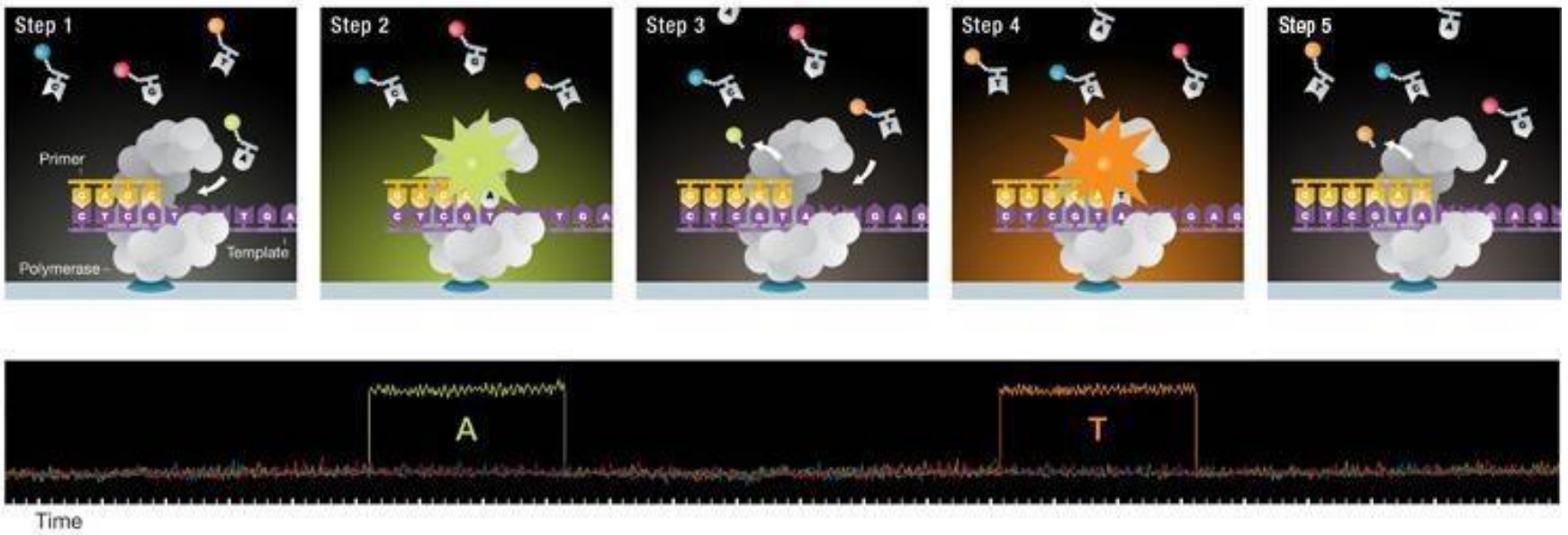
PacBio Template Preparation

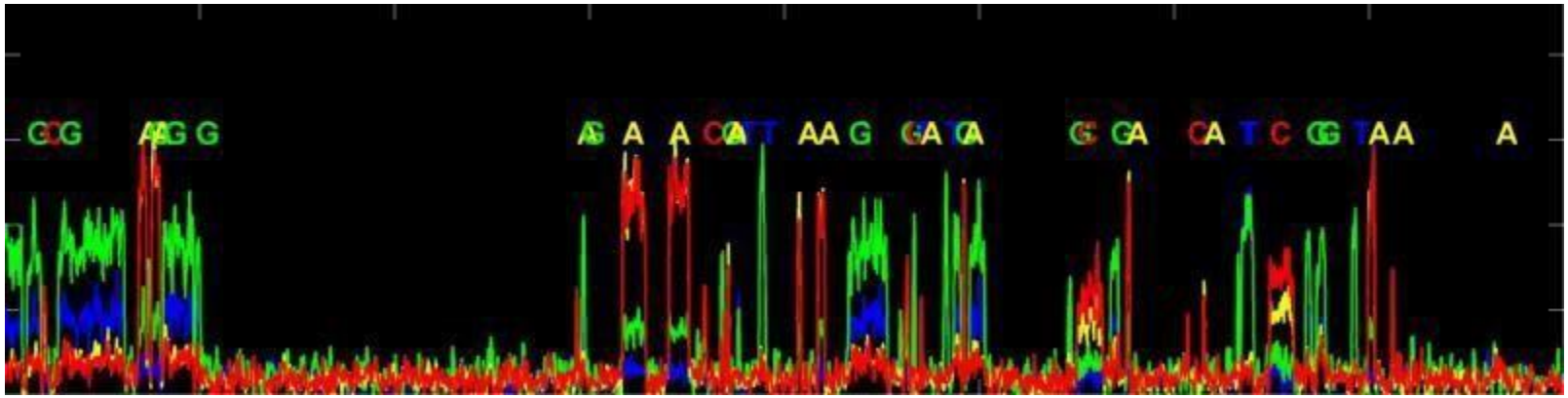


1. Anneal primer.
2. Bind polymerase.
3. Sequence.

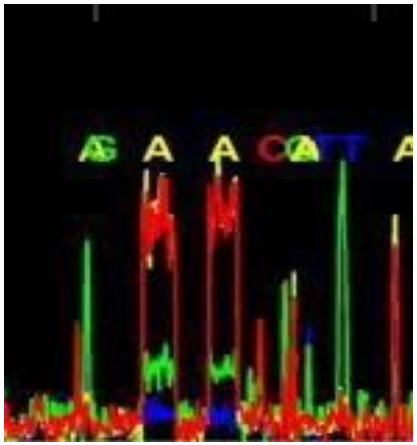


PacBio Sequencing Cycles



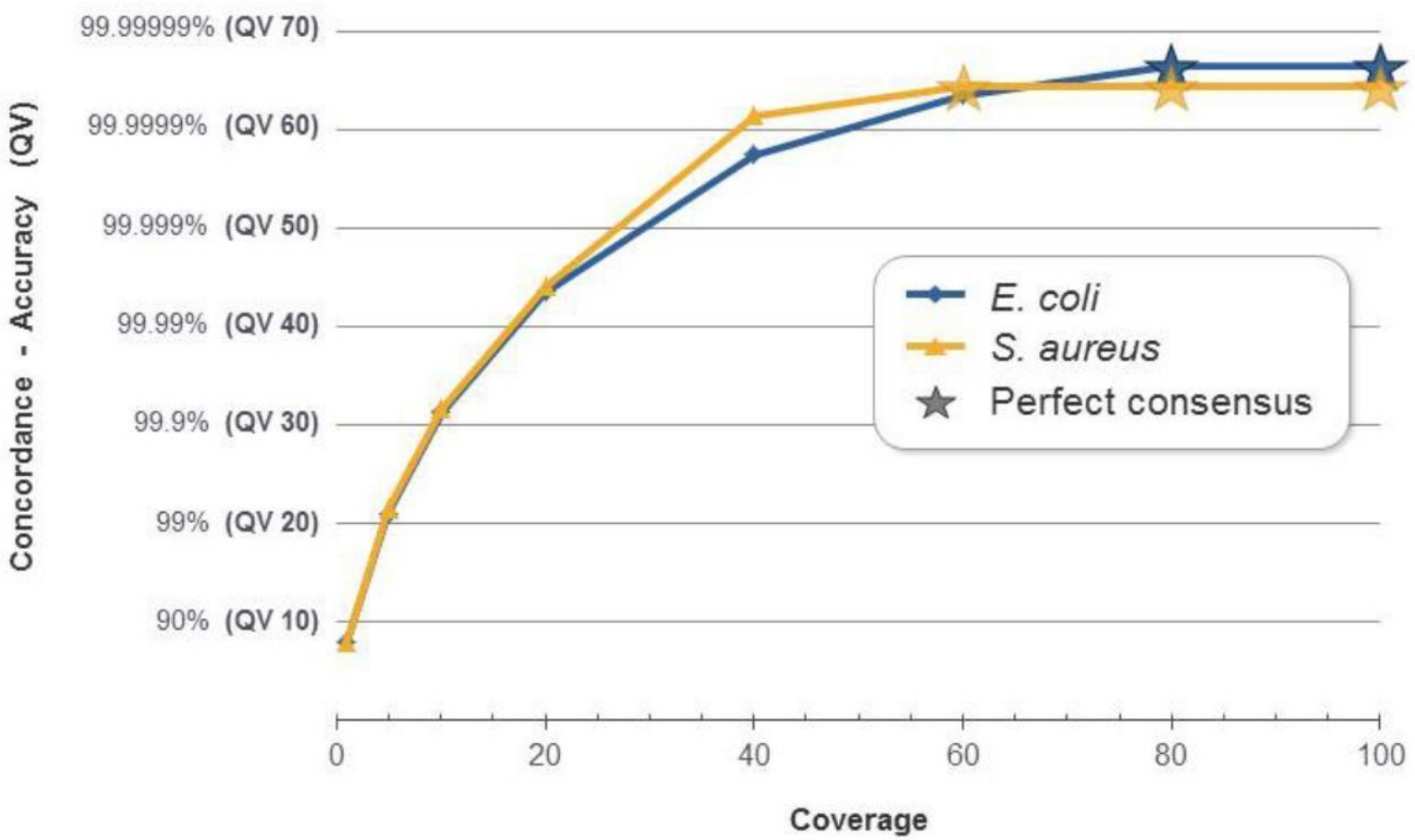


PacBio Sequencing



- Some bases added very quickly and missed.
- Some wrong bases flirt with active site and go away.
- SMRT cell has 1 million ZMWs.
- US$500 each.

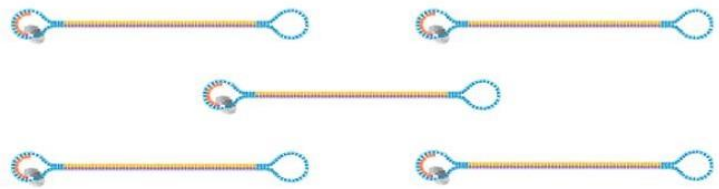
PacBio Accuracy

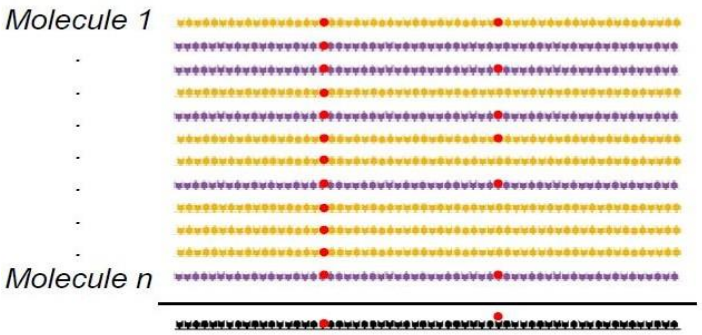


PacBio Sequencing Modes

UNTIL NOW: 2 MODES OF SMRT SEQUENCING

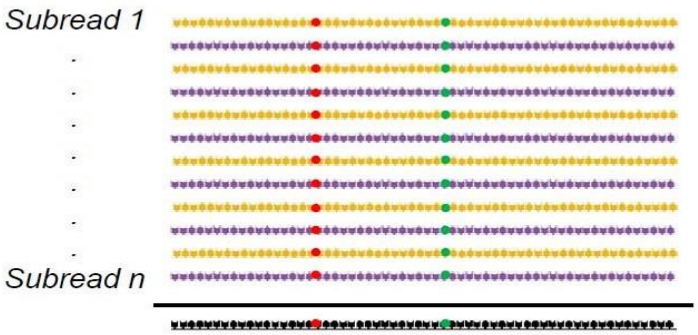
- Long-insert Genome Sequencing:





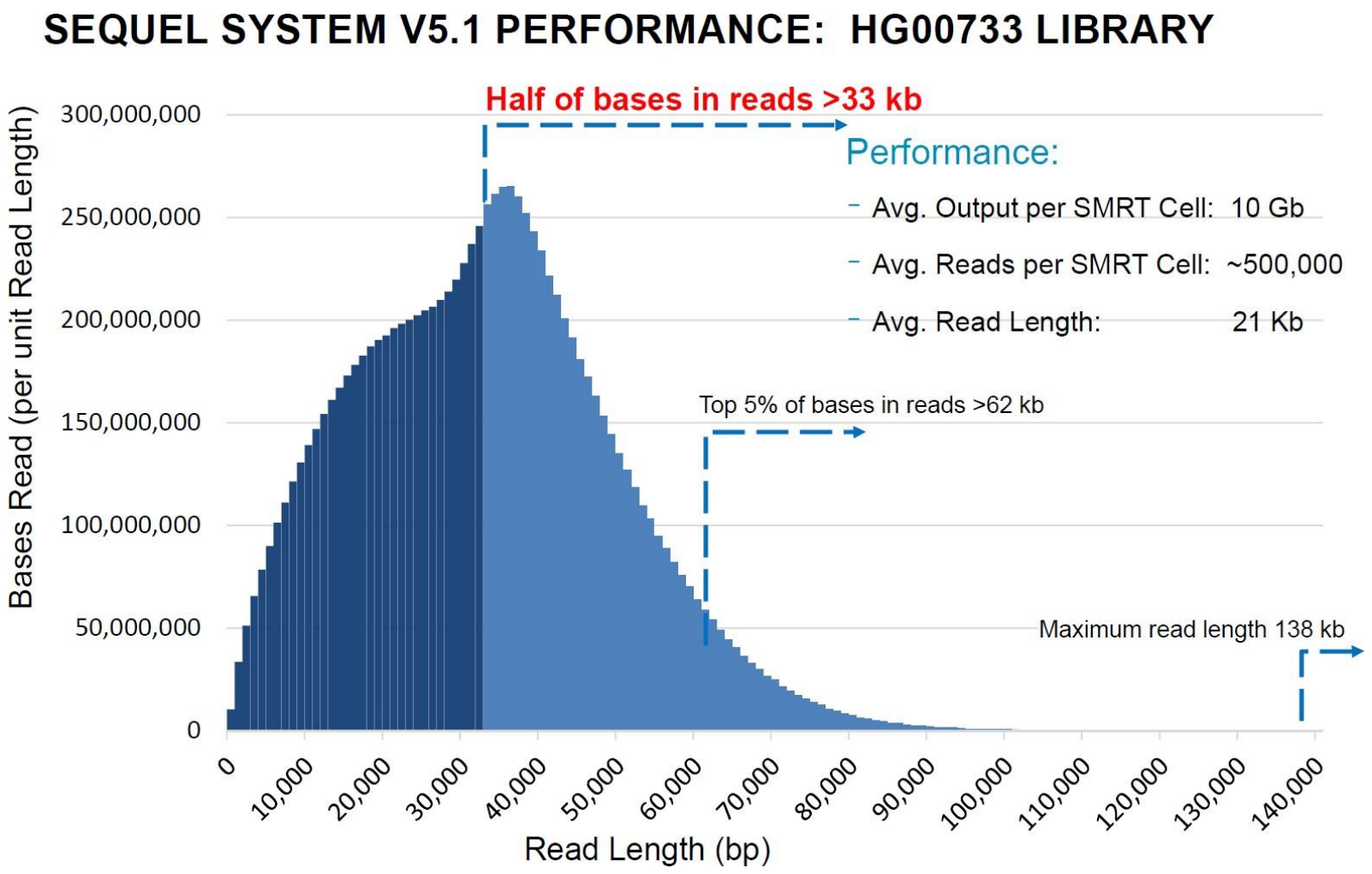
Genotyping, SV detection, *de novo* assembly

- Circular Consensus Sequencing (CCS):



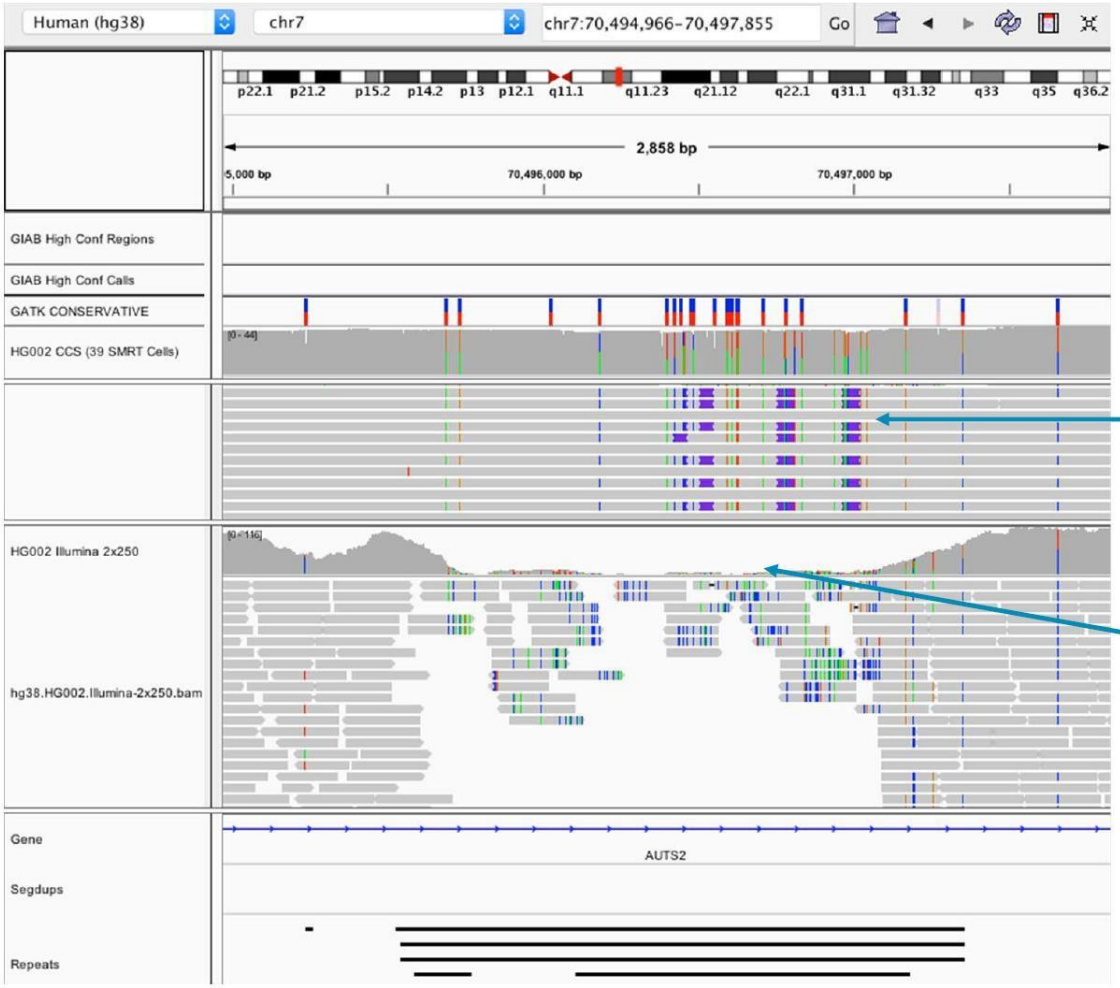
Amplicons, Minor variant detection, Metagenomics

PacBio Sequencing Performance



PacBio Detecting Structural Variation

INTRONIC INSERTION IN *AUTS2* (AUTISM)

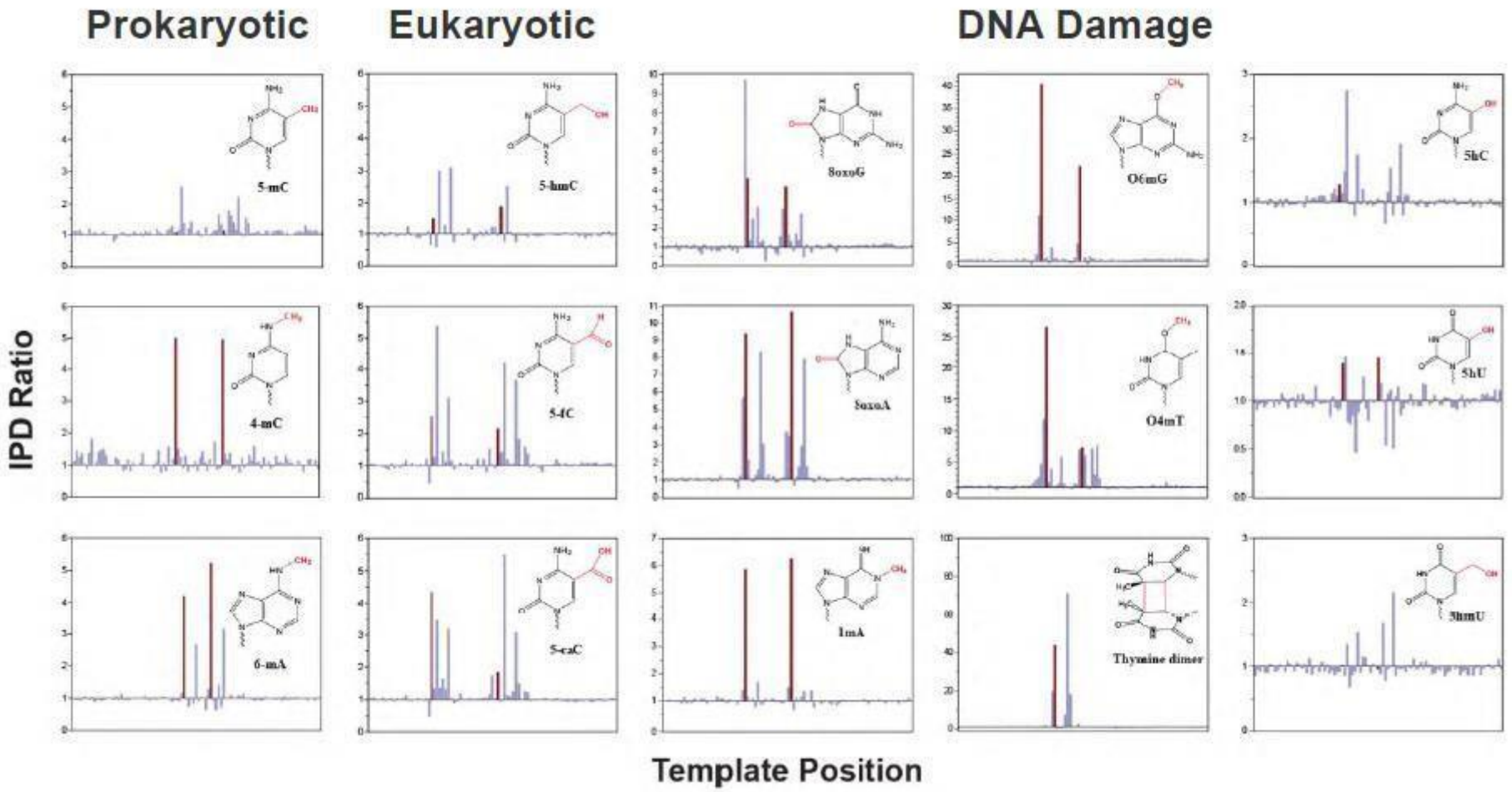


PacBio reads show insertions

Illumina reads do not map

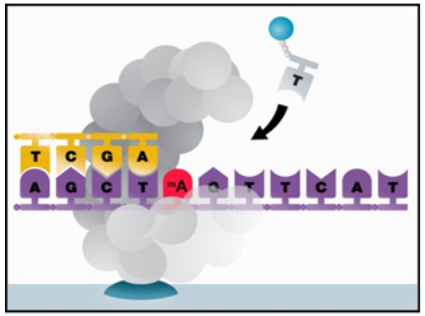
Tandem repeat

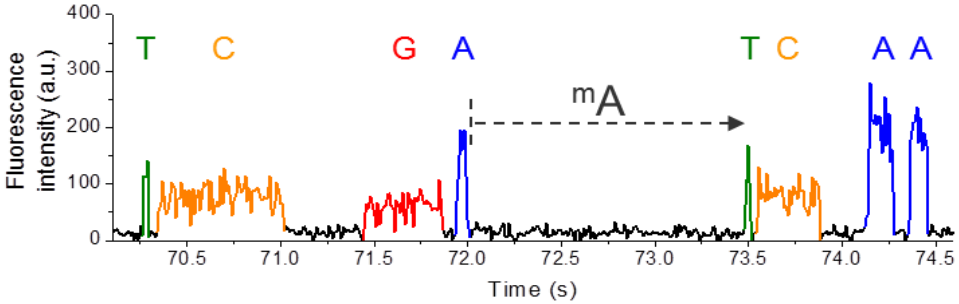
PacBio Can Detect Base Modifications

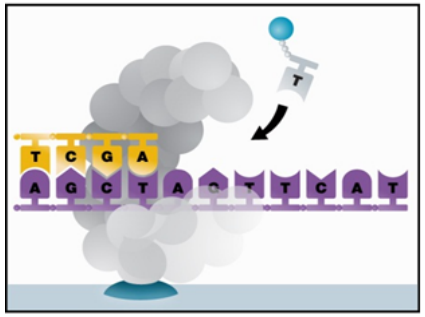


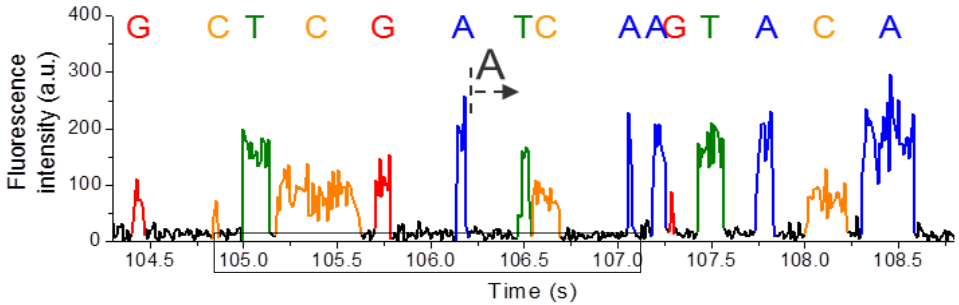
Detection of DNA Base Modifications

Example: N^6-methyladenine



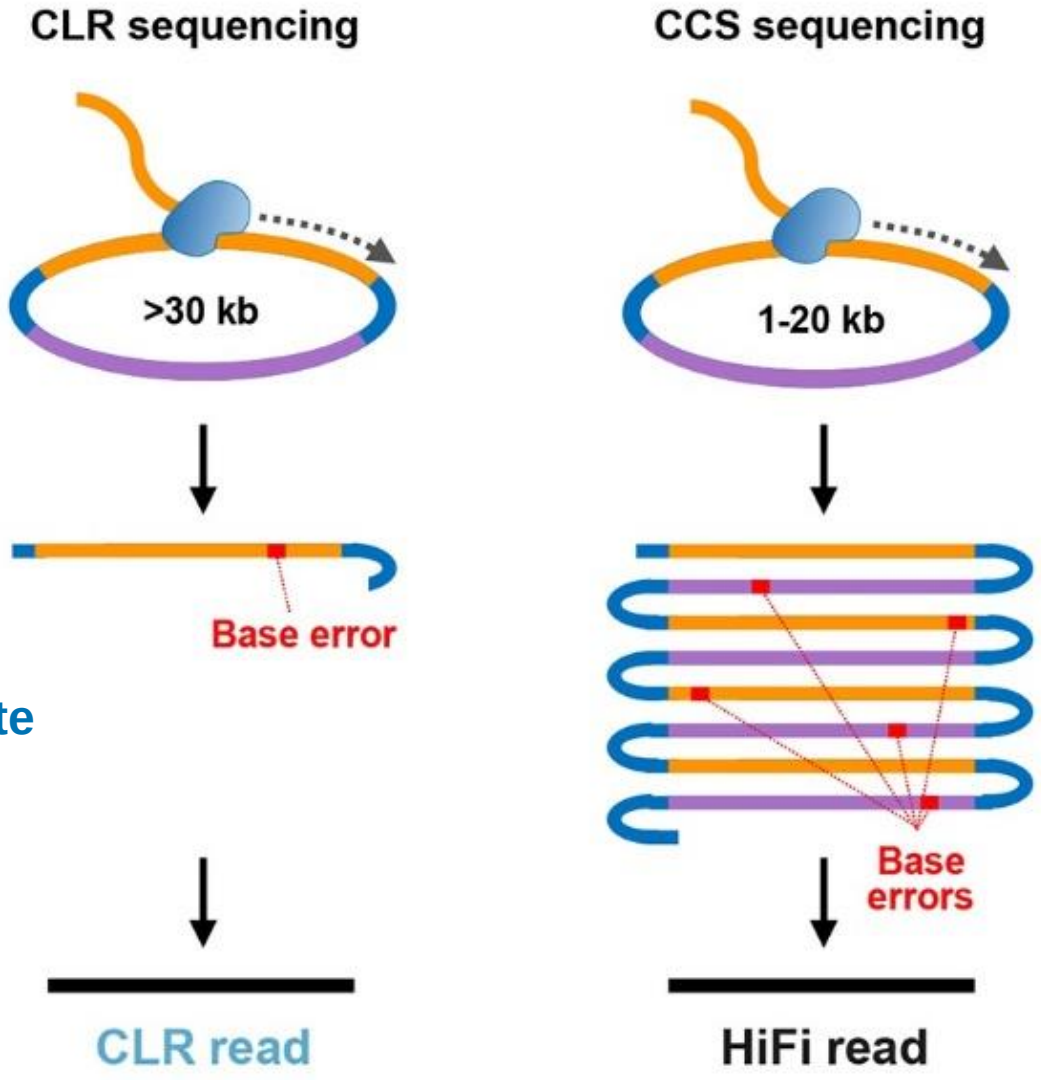






- SMRT Sequencing uses kinetic information from each nucleotide addition to call bases.
- This same information can be used to distinguish modified and native bases by comparing results of SMRT Sequencing to an *in silico* kinetic reference for incorporation dynamics without modifications.

CLR and Hi-Fi Reads



8 - 15% error rate

<1% error rate

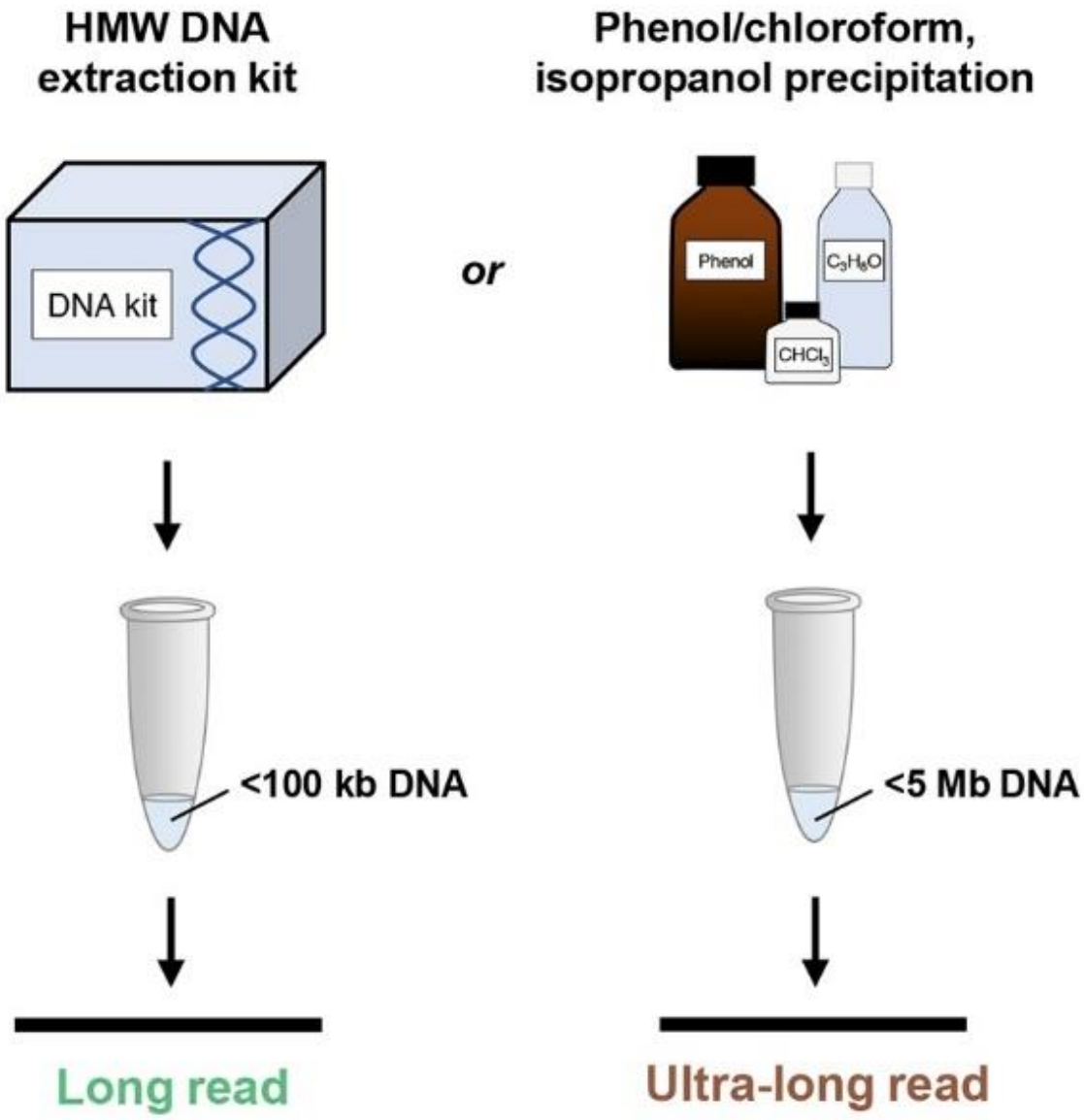
Oxford Nanopore Technology

- Another long read sequencing platform.
- Sequences single molecules of DNA as travel through pore.
- Reads typically 5 – 50 kb but some much longer (up to 4 Mb have been achieved).
- Fast moving technology that may soon be cheaper and have a higher throughput than Illumina.
- Zero capital cost price structures available.

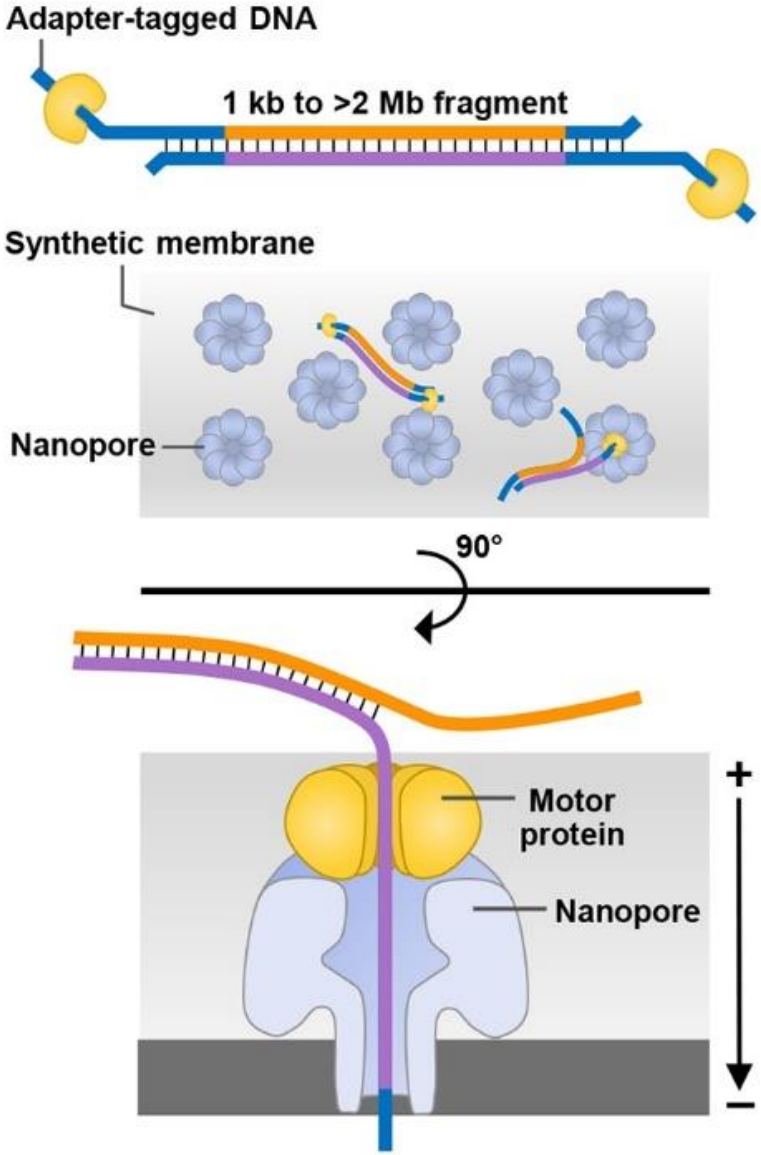
ONT Sequencers

Platform	Flongle	MinION	GridION	PromethION		
				P2	24	48
Output	1 – 2 Gb	15 -35 Gb	6 TB	100 – 200 Gb /cell		
Large Genomes			●	●	●	●
Small Genomes	●	●	●	●	●	●
Exome Sequencing			●	●	●	●
Targeted Re-sequencing	●	●	●	●	●	●
Transcriptome Sequencing		●	●	●	●	●
Gene Expression Profiling		●	●	●	●	●
DNA-Protein Interactions		●	●	●	●	●
Methylation Sequencing		●	●	●	●	●
16S Metagenomic sequencing		●	●	●	●	●

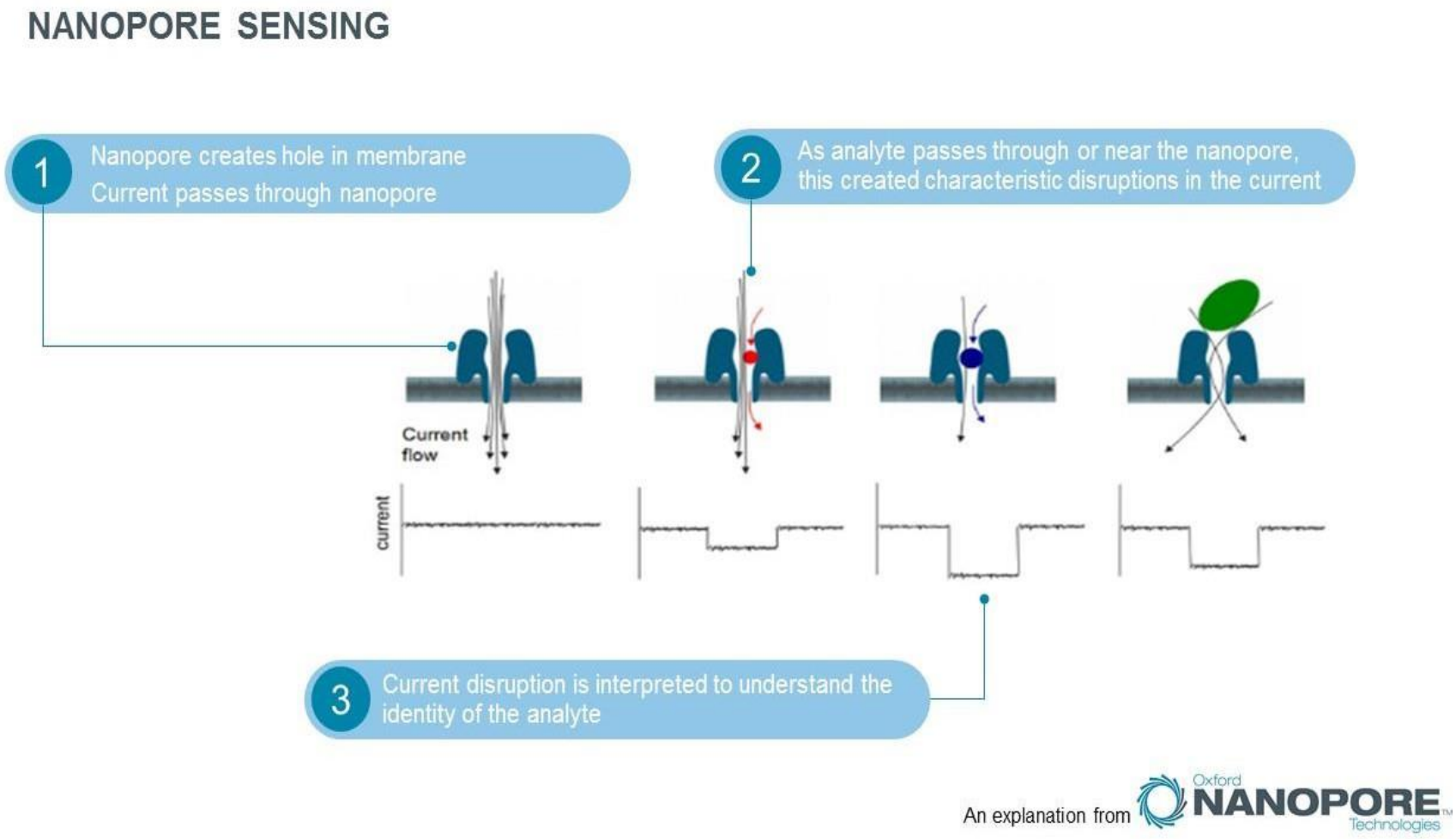
Ultra-long Reads



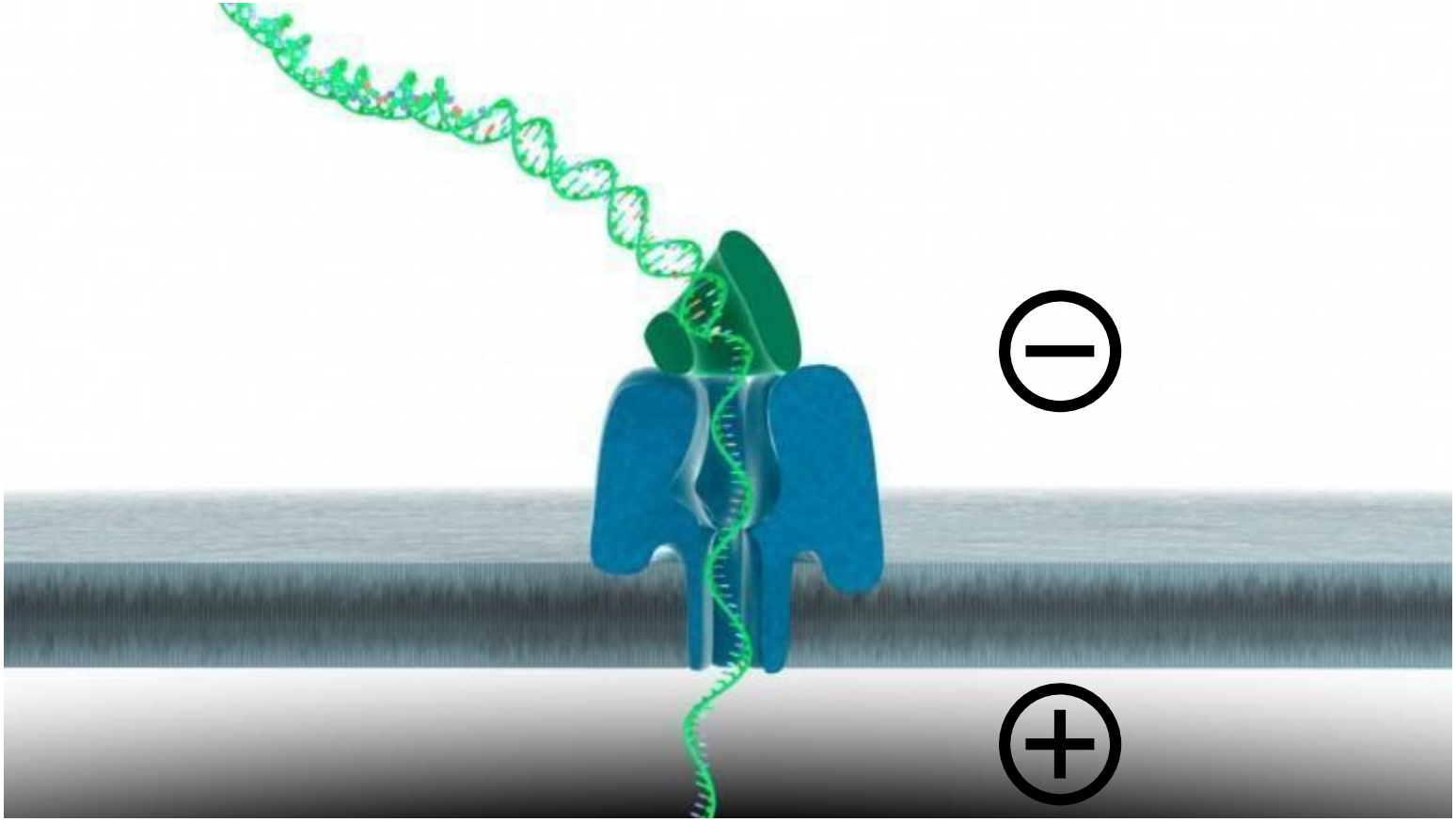
The Protein Nanopore



Nanopore Sensing

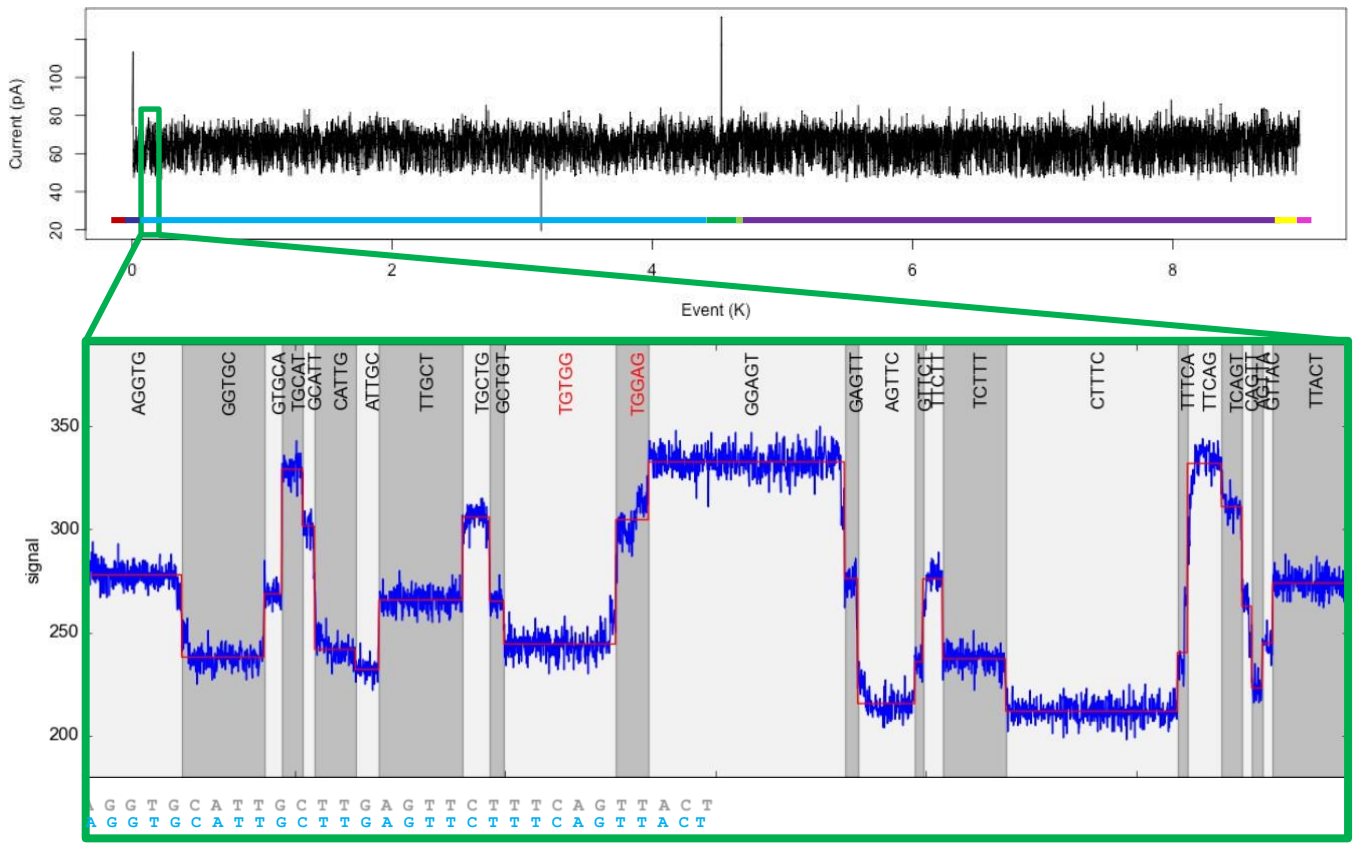


ONT



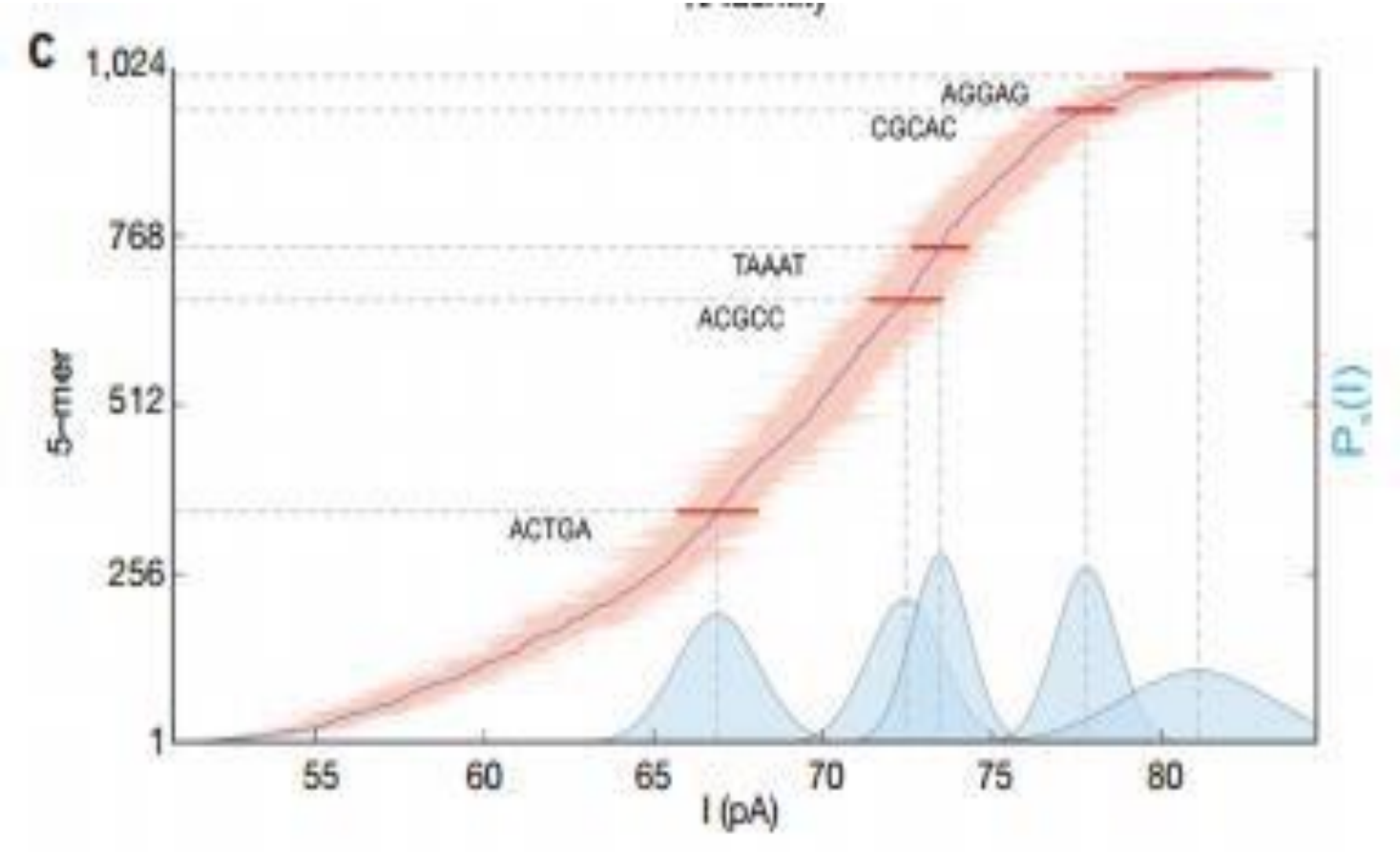
5-base words = 1024 current levels

ONT: The Squiggles



Slide courtesy of David Buck. WTCGH

5mer Current Signals

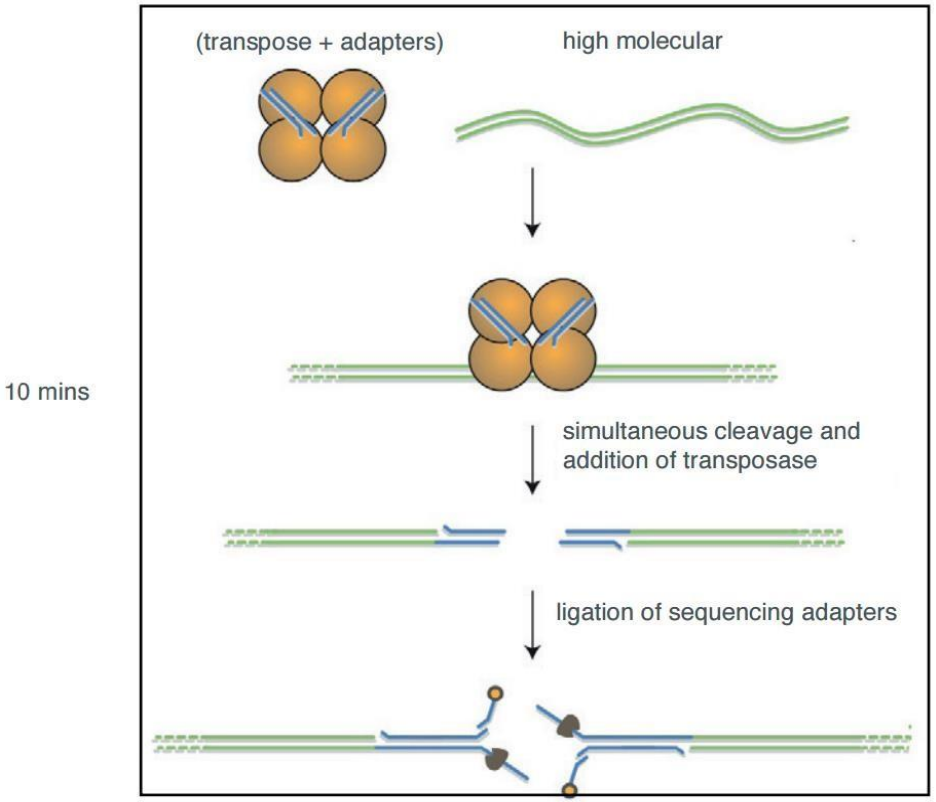


From Szalay & Golovchenko, Nat. Biotech (2015).

ONT Library Preparation

RAPID SEQUECING KIT

A two-step, 10 minute protocol



Starting material will be fragmented; recommended starting size >30 kb for genomic DNA

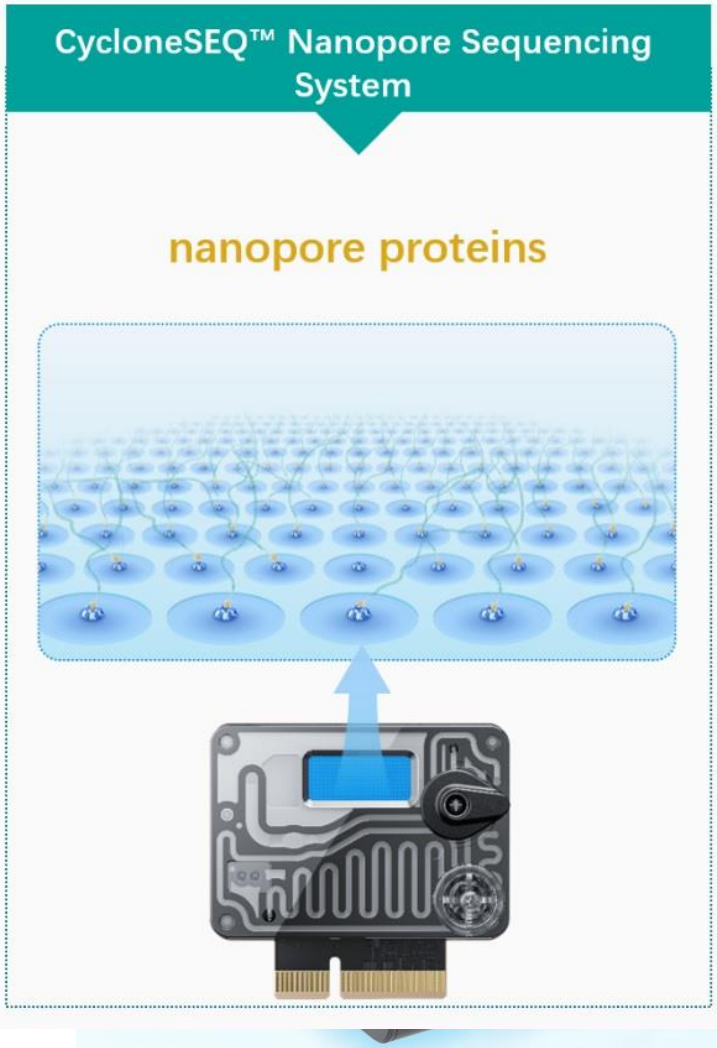
ONT Data Quality

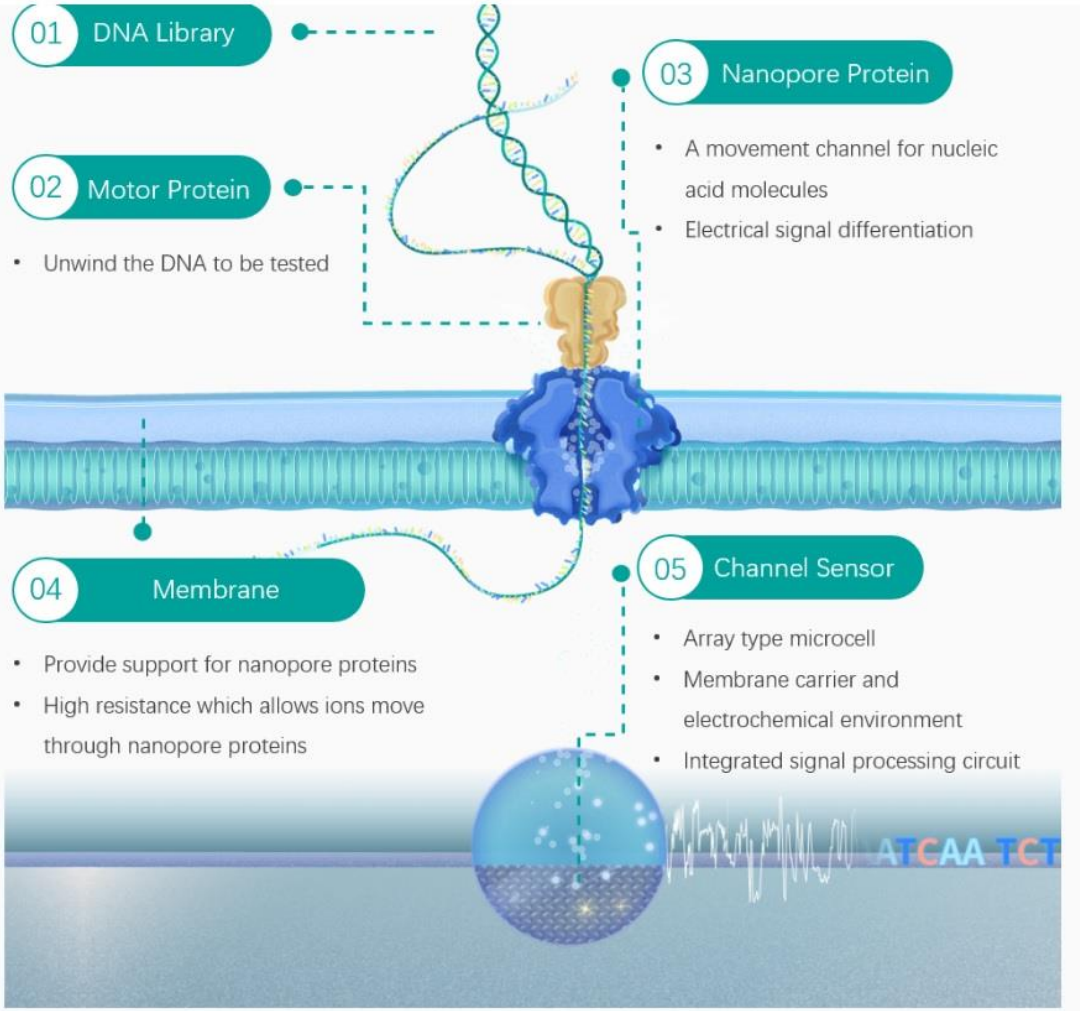
- Low error rates 2 - 5%.
- Read lengths as long as template. Record 2 Mb, but shorter fragments give higher yield.

ONT Applications

- Super-long reads (>1Mb).
- “Run until done".
- Select reads.
- Mobile sequencing.
- Direct methylation and base modification detection.
- Direct RNA sequencing.
- Cas9 mediated enrichment.

MGI Tech CycloneSEQ-WT02





Future Sequencers

- Roswell Technologies.
- Base4.
- Quantapore.
- Gnubio.
- Genia.

Future Sequencers

- Roswell Biotechnologies Inc. and Imec developing molecular electronics biosensor chips (https://www.roswellbiotech.com/).
- Base4 are developing a micro-droplet-based sequencing system (https://www.cambridgemedical.com/sequencing-a-genome-base-by-base-by-base-by-base/).
- Quantapore Inc. are developing another nanopore based platform (https://quantapore.com/).
- Geniais Roche's proprietary nanopore based sequencing platform.
- ZS Genetics, Inc. and Hitachi High Technologies America, Inc. (HTA) are developing a DNA sequencer based on annular dark-field scanning transmission electron microscopy (ADF-STEM) by incorporating 'ZSG labels', modified nucleotides with heavy atoms or other compounds to make them visible.

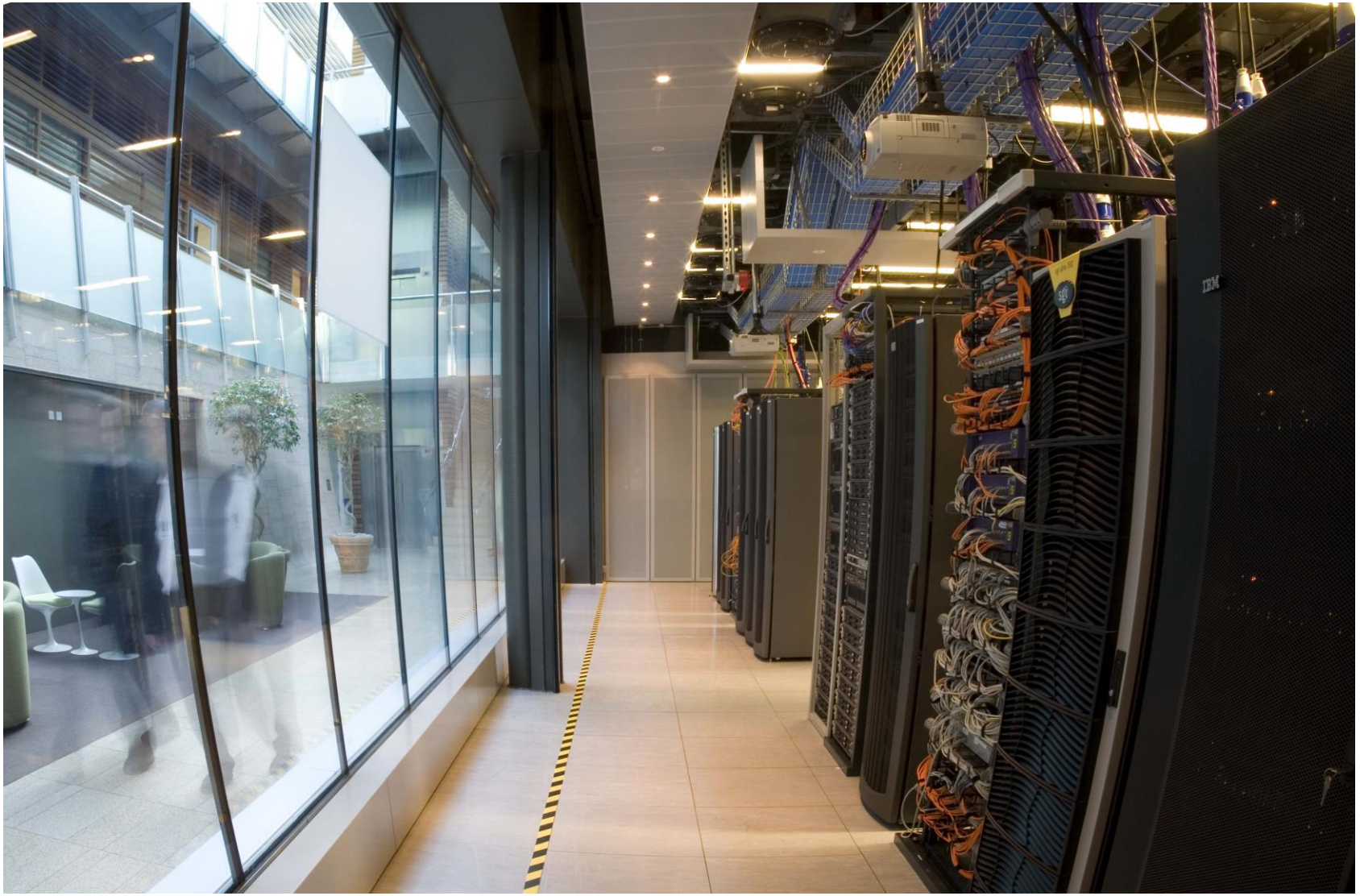
Other Emerging Technologies

http://allseq.com/kb-category/emerging-technologies/

- Gnubioa droplet based sequencing platform under Bio-Rad.
- Halcyon Molecular has developed an advanced DNA manipulation technology that enables ultra-low-cost and ultra-fast sequencing by transmission electron microscopy (TEM). This method is termed IMPRNT for Individual Molecule Placement Rapid Nano Transfer.

Storage Issues

75 Pb vs 1 Pb



References

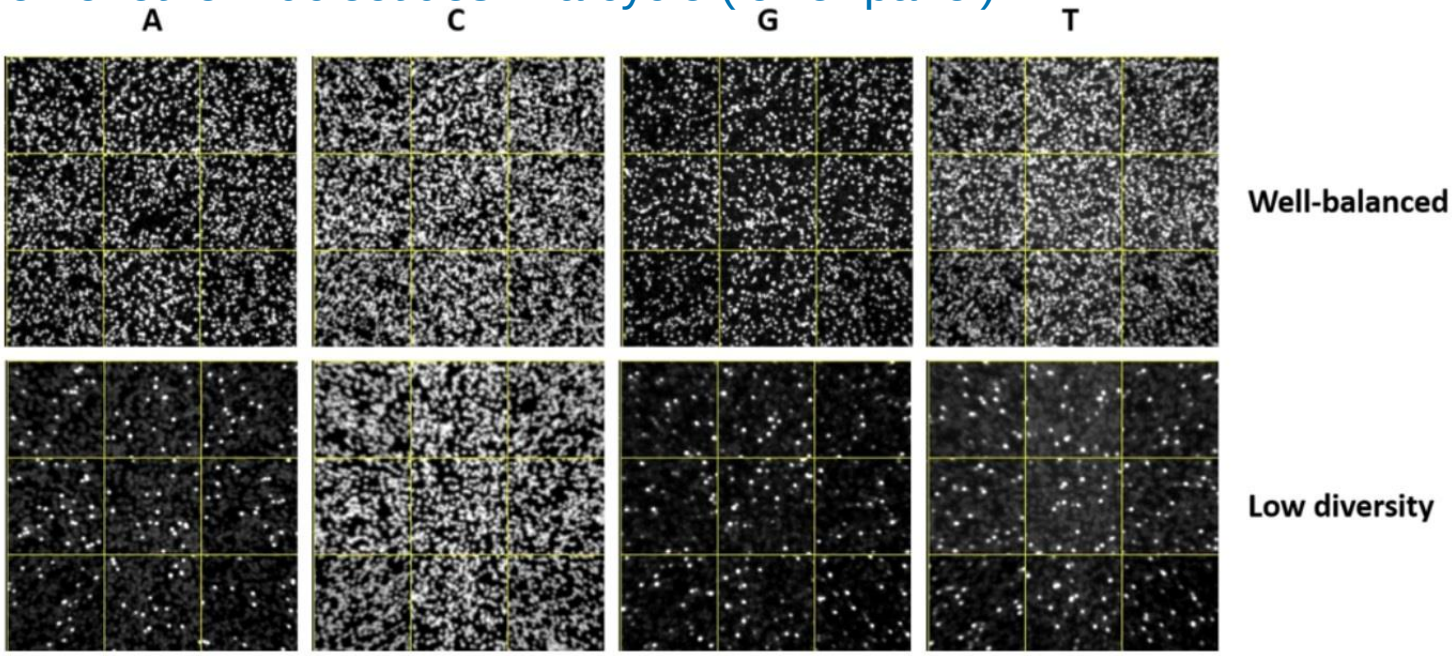
- Gawad C *et al.* 2016. Single-cell genome sequencing: current state of the science. **Nature Review Genetics** 17:175-188.
- Ke R *et al.* 2016. Fourth generation of next-generation sequencing technologies: Promise and Consequences. **Human Mutation** 37:1363-1367.
- Mardis ER. 2008. Next-generation DNA sequencing methods. **Annual Review of Genomics and Human Genetics** 9:387-402.
- Shendure and Ji. 2008. Next-generation DNA sequencing. **Nature Biotechnology** 26:1135-1145.
- Wheeler DA *et al*. 2008. The complete genome of an individual by massively parallel DNA sequencing. **Nature** 452:872-876.
- Wong KM *et al*. 2011. Unraveling the genetics of cancer: Genome sequencing and beyond. **Annual Review of Genomics and Human Genetics** 12:407-430.

Questions

qasim.ayub@monash.edu

Nucleotide Diversity in Sequencing

- Nucleotide diversity refers to the relative proportion of nucleotides A, C, G and T present in every cycle of the run.
- Well-balanced or high diversity libraries have roughly equal proportions of all four nucleotides in each cycle throughout the sequencing run (upper panel).
- Low diversity libraries have a high proportion of certain nucleotides and a low proportion of other nucleotides in a cycle (lower panel).



Solution for Low Base Diversity

- Sequencing libraries with low base diversity or unbalanced nucleotide composition can negatively impact cluster mapping (on non-patterned flow cells) and template registration (on non-patterned flow cells) along with data quality and data output.

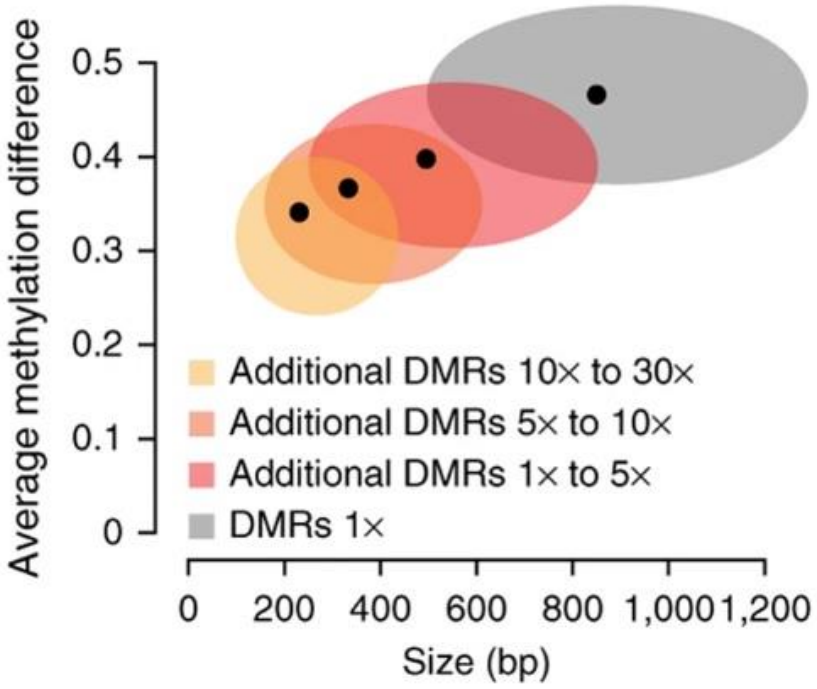
- Illumina recommends using a PhiX Control v3 Library, as an ideal sequencing control (typically with ≥ 1% spike-in) for run quality monitoring, cluster generation, sequencing, and alignment. This is a concentrated Illumina library (10 nM in 10 µl) that has an average size of 500 bp and consists of balanced base composition at ~45% GC and ~55% AT.

- At 5% or higher spike-in it provides the balanced fluorescent signals that low diversity sample libraries lack during each sequencing cycle.

- This, in turn, assists with cluster mapping and template registration and improves overall run performance.

What Coverage is Appropriate for Epigenomics?

- ENCODE recommends 5 -15X per replicate and at least two replicates for Differentially Methylated Regions (DMRs).



Ziller *et. al.* 2015. Nature Methods 12:230–232.