Next Generation Sequencing

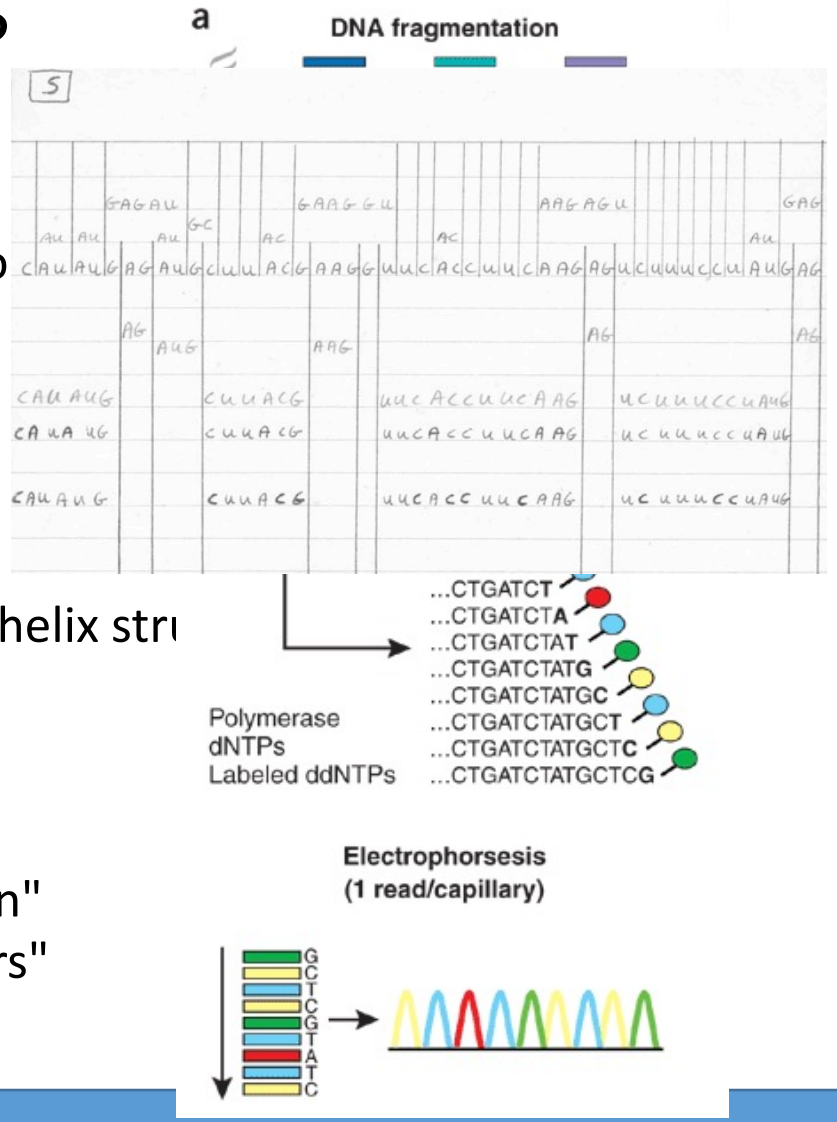
Wouter Bossuyt- BESHG

08/10/2021

Sequencing: a history

Landmarks in DNA sequencing

- 1911:
  - Thomas Hunt Morgan disproves himself and find chromoso
- 1944-1952
  - Avery–MacLeod–McCarty experiment (DNA in bacteria)
  - Hershey–Chase experiment (DNA in phages)
- 1953
  - Rosalind Franklin, Watson, Crick: Discovery of DNA double helix stru
- 1953-1977: the ‘desperate’ era
  - Walter Fiers use RNase digest in competition with Sanger
- 1977
  - A Maxam and W Gilbert "DNA seq by chemical degradation"
  - F Sanger"DNA sequencing with chain-terminating inhibitors"



Landmarks in DNA sequencing

- 1984
  - DNA sequence of the Epstein-Barr virus, 172 kb
- 1987
  - Applied Biosystems - first automated sequencer
- 1991
  - Sequencing of human genome in Venter's lab
- 1996
  - P. Nyrén and M Ronaghi - pyrosequencing
- 2001
  - A draft sequence of the human genome
- 2003
  - human genome completed
- 2004
  - 454 Life Sciences markets first NGS machine

doi:10.1038/nature03959 nature

ARTICLES

Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies[1]*, Michael Egholm[1]*, William E. Altman[1], Said Attiya[1], Joel S. Bader[1], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Kshama B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[3], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

The proliferation of large-scale DNA-sequencing projects in recent years has driven a search for alternative methods to reduce time and cost. Here we describe a scalable, highly parallel sequencing system with raw throughput significantly greater than that of state-of-the-art capillary electrophoresis instruments. The apparatus uses a novel fibre-optic slide of individual wells and is able to sequence 25 million bases, at 99% or better accuracy, in one four-hour run. To achieve an approximately 100-fold increase in throughput over current Sanger sequencing technology, we have developed an emulsion method for DNA amplification and an instrument for sequencing by synthesis using a pyrosequencing protocol optimized for solid support and picolitre-scale volumes. Here we show the utility, throughput, accuracy and robustness of this system by shotgun sequencing and *de novo* assembly of the *Mycoplasma genitalium* genome with 96% coverage at 99.96% accuracy in one run of the machine.

DNA Sequencing – the next generation

- NGS refers to non-Sanger-based high-throughput DNA sequencing technologies.

- NGS technologies constitute various strategies that rely on a combination of
  - Library/template preparation
  - Parallel sequencing

Different technologies

- 454
- Solexa/Illumina
- Ion Torrent/Thermo Fischer
- Nanopore
- Pacbio
- MGI
- …

Roche 454

Solexa/Illumina

Ion torrent

Pacbio

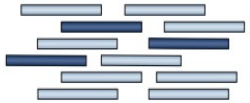
MGI

Nanopore

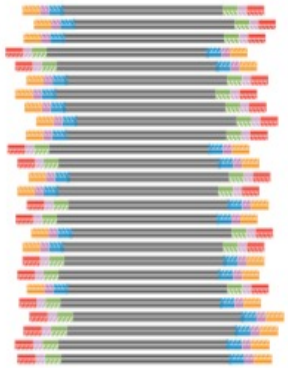
Sequencing workflows

NGS workflow

Genomic DNA



Fragments

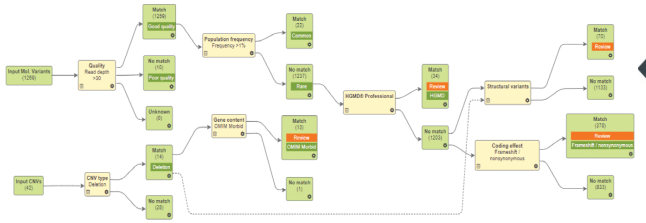


Library

DNA sequencing

AGGTCGTTACGTACGCTAC
GACCTACATCAGTACATAG
GCATGACAAAGCTAGGTGT

Mapping, alignment, variant calling



Variant filtering & interpretation

What do I chose? Long read, short read,…

	Illumina	Nanopore	Pacbio
Read length	35 bp to 600 bp	Anything goes	250bp to 25kb (or 100kb)
Accuracy	High	Medium and improving	high
Capacity	Small to very large	Tiny to large	medium
Biases	Fragment size and GC		DNA modifications
Applications			
WGS	+++	++	+++
RNAseq	+++	+	++
Targeted resequencing	+++	Only large fragments	++
Single cell sequencing	+++		

Illumina (solexa) sequencing

- Illumina MiSeq, NextSeq 500, Nextseq2000, HiSeq4000 & NovaSeq 6000



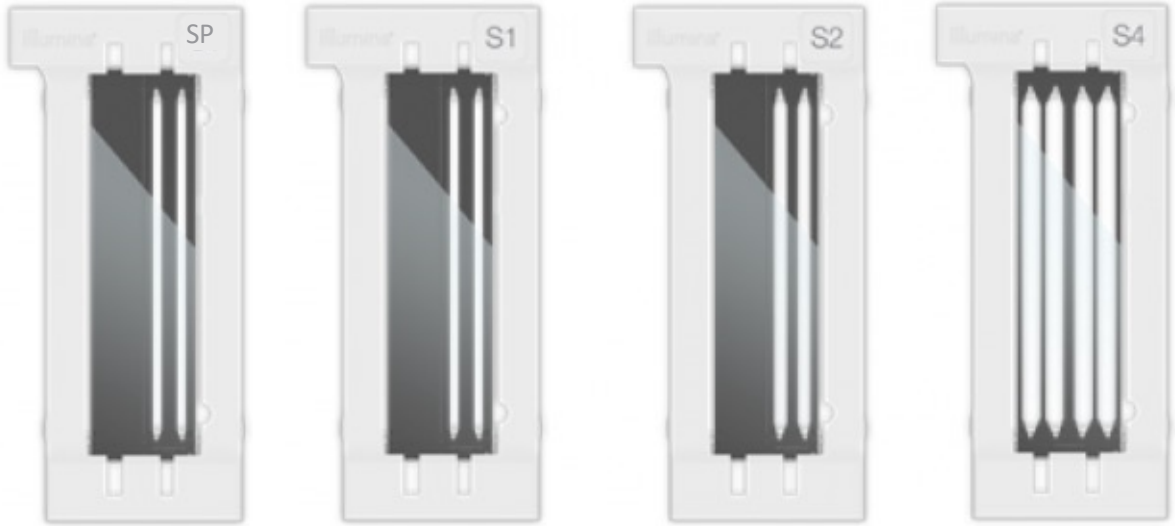
Illumina flowcells

MiSeq

NextSeq500

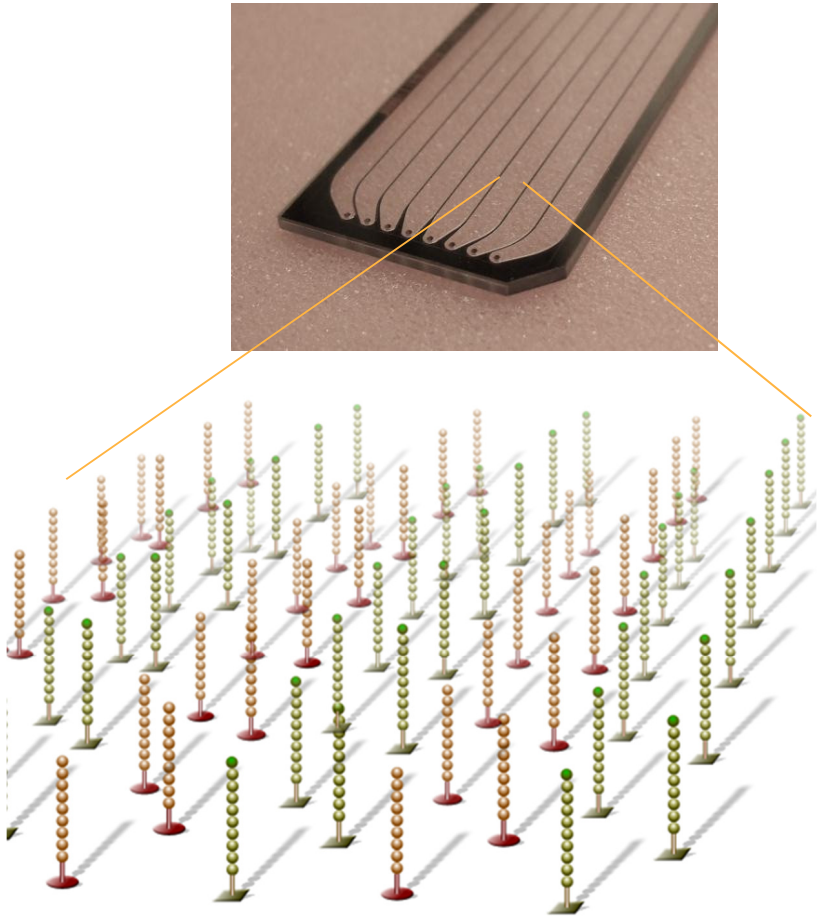
NextSeq2000

HiSeq4000



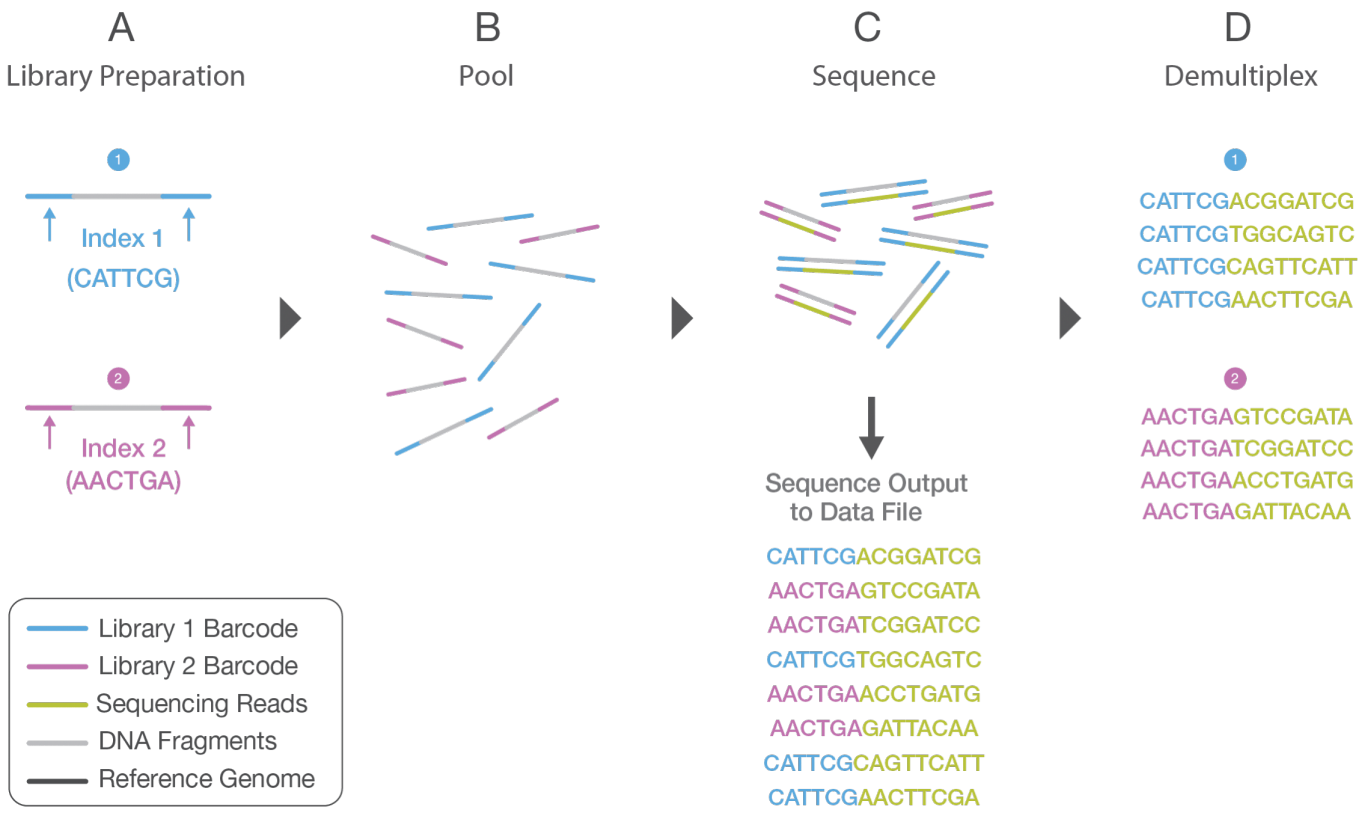
NovaSeq

NGS Library Prep



NGS Library Prep

- Indexing
  - Sample barcodes
  - High diversity necessary
  - Unique dual indexing is top



NGS Library Prep

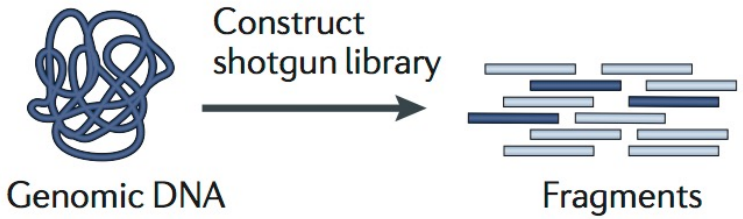
- General DNA library prep
- General RNA library prep
- Targeted library prep approaches

NGS Library Prep

- General DNA library prep
- General RNA library prep
- Targeted library prep approaches

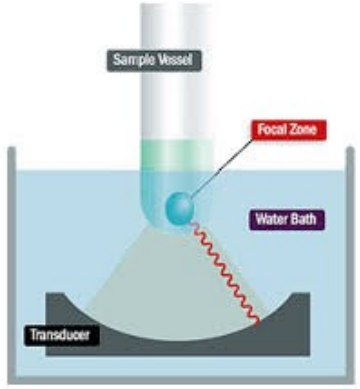
NGS Library prep

- Fragmentation



  - acoustic

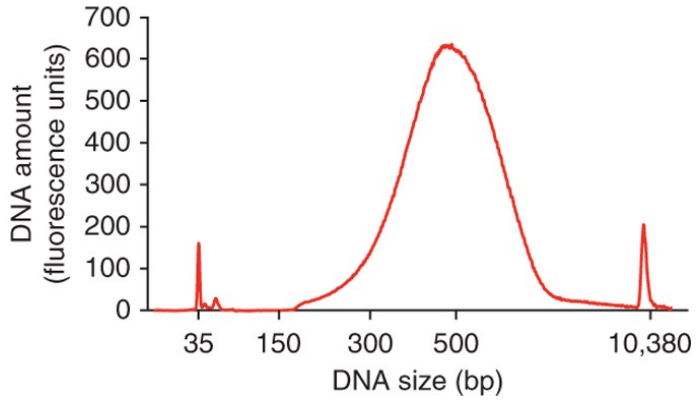
covaris



  - enzymatic

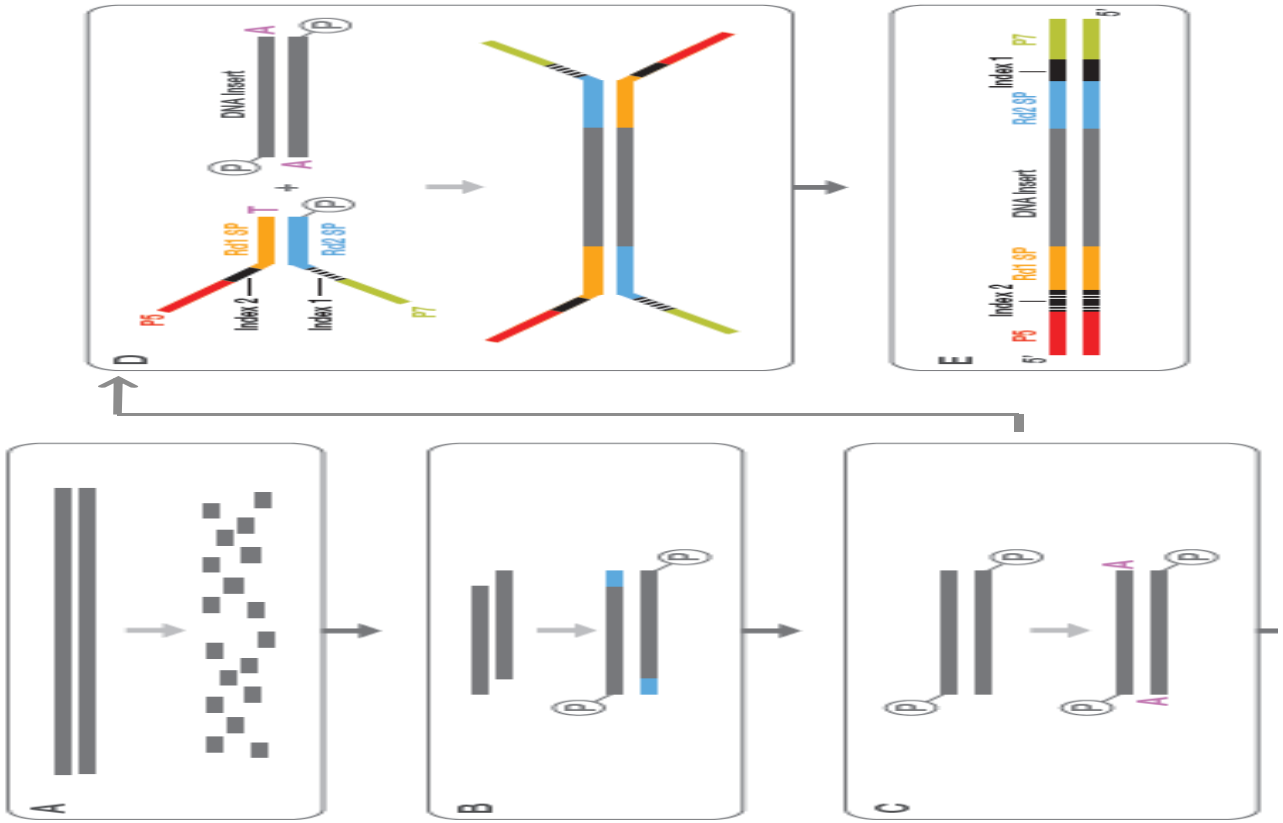
NGS Library prep

- Fragmentation
  - quality check: BioAnalyser / FragmentAnalyser



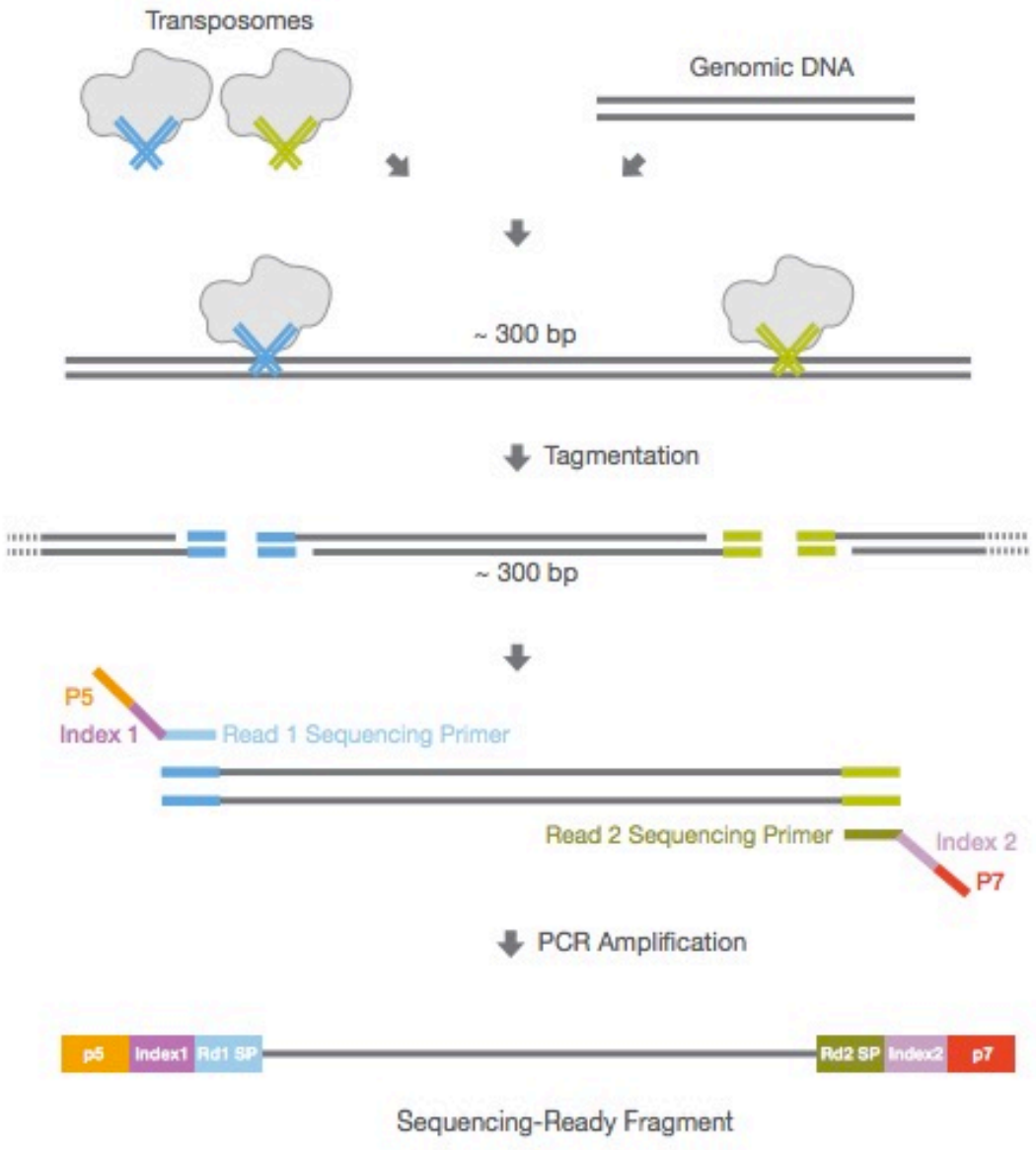
NGS Library prep

- Standard A-tailing & adaptor-ligation (DNA/RNA)



NGS Library Prep

- Illumina Nextera tagmentation
- Transposon-based adapter insertion
- PCR-based indexing



NGS Library Prep

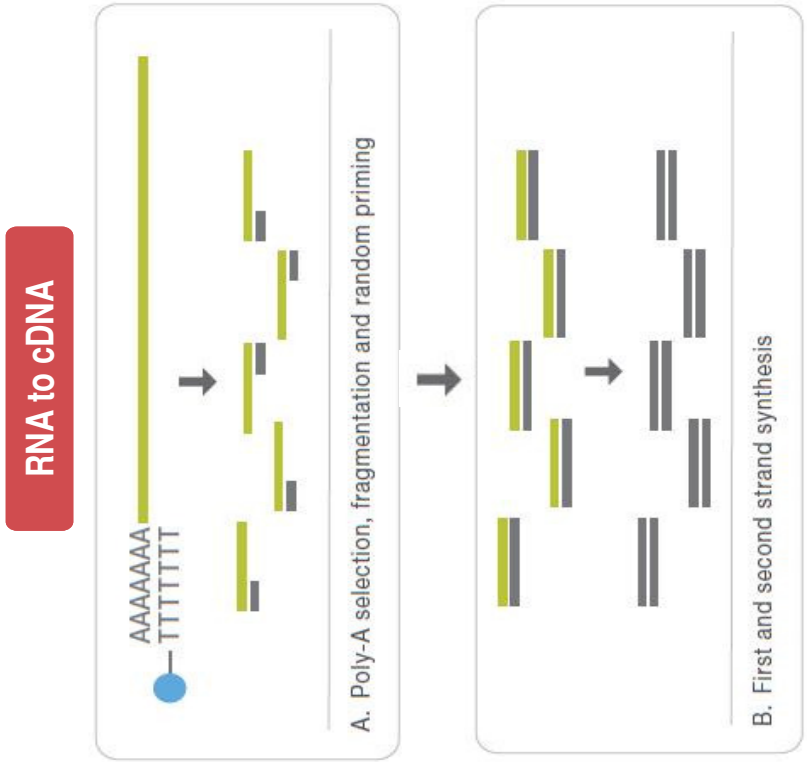
- General DNA library prep
- General RNA library prep
- Targeted library prep approaches

NGS Library Prep

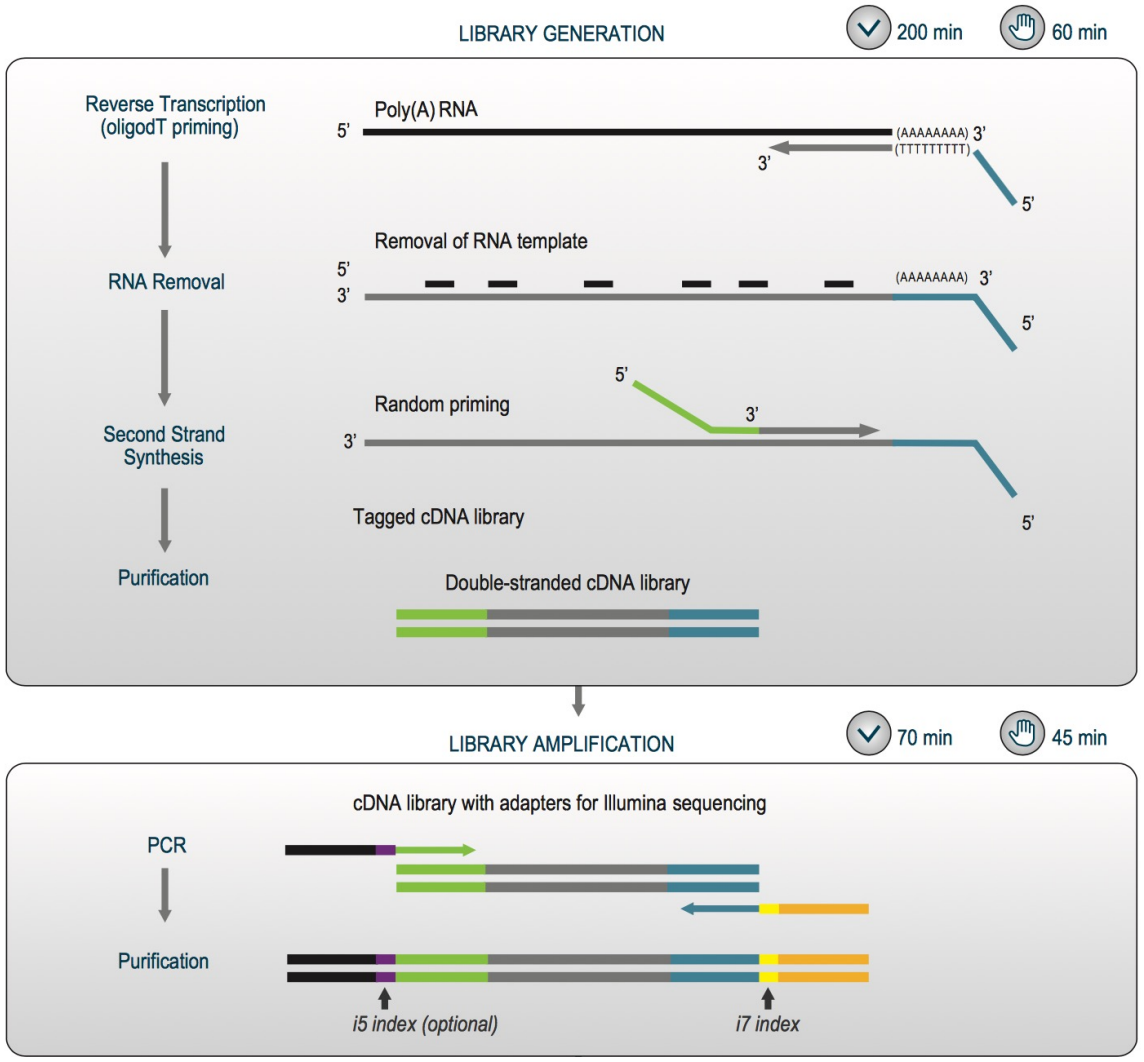
- General DNA library prep
- General RNA library prep
- Targeted library prep approaches

NGS Library Prep

- Illumina TruSeq Stranded mRNA



LexoGen QuantSeq (RNA)



NGS Library Prep

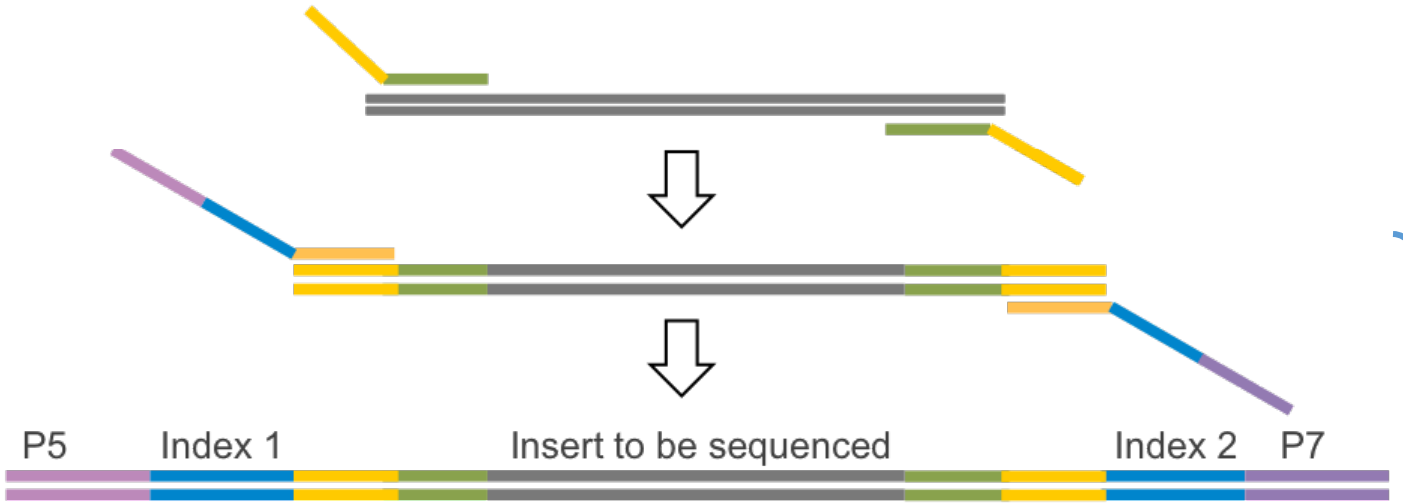
- General DNA library prep
- General RNA library prep
- Targeted library prep approaches

NGS Library Prep

- General DNA library prep
- General RNA library prep
- Targeted library prep approaches
  - amplicon based
  - ligation based
  - enrichment based

NGS Library Prep

- Custom two step PCR

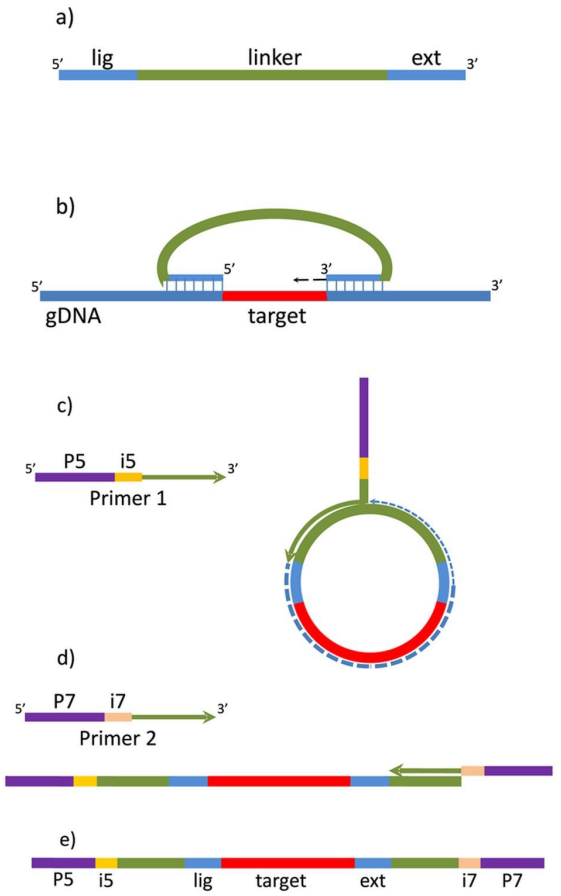


Performed by you

Performed by the GC

NGS Library Prep

- MIPs



Multiplex probes

Hybridization and fill-in reaction

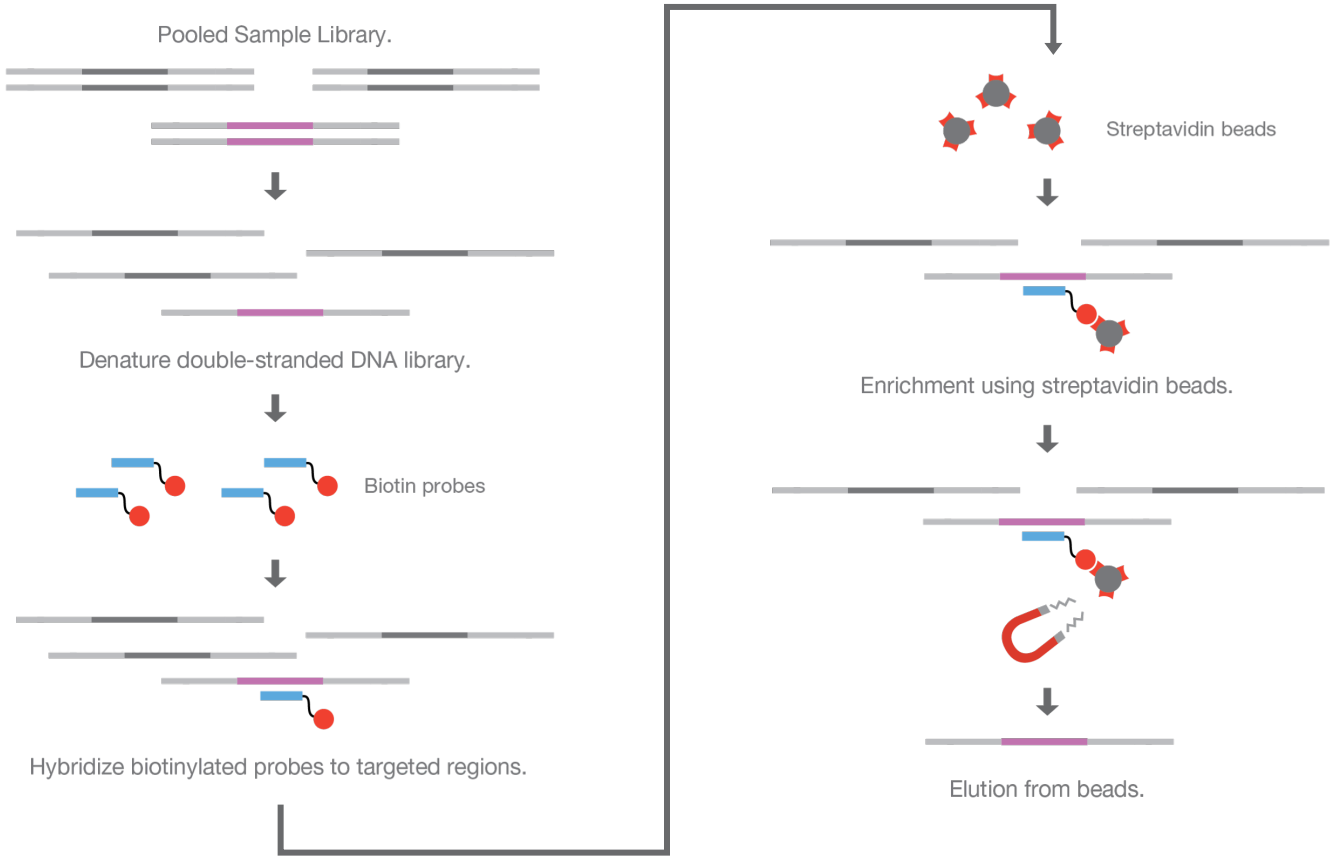
PCR linearization

PCR amplificaiton

Niedzicka *et al.* 2016

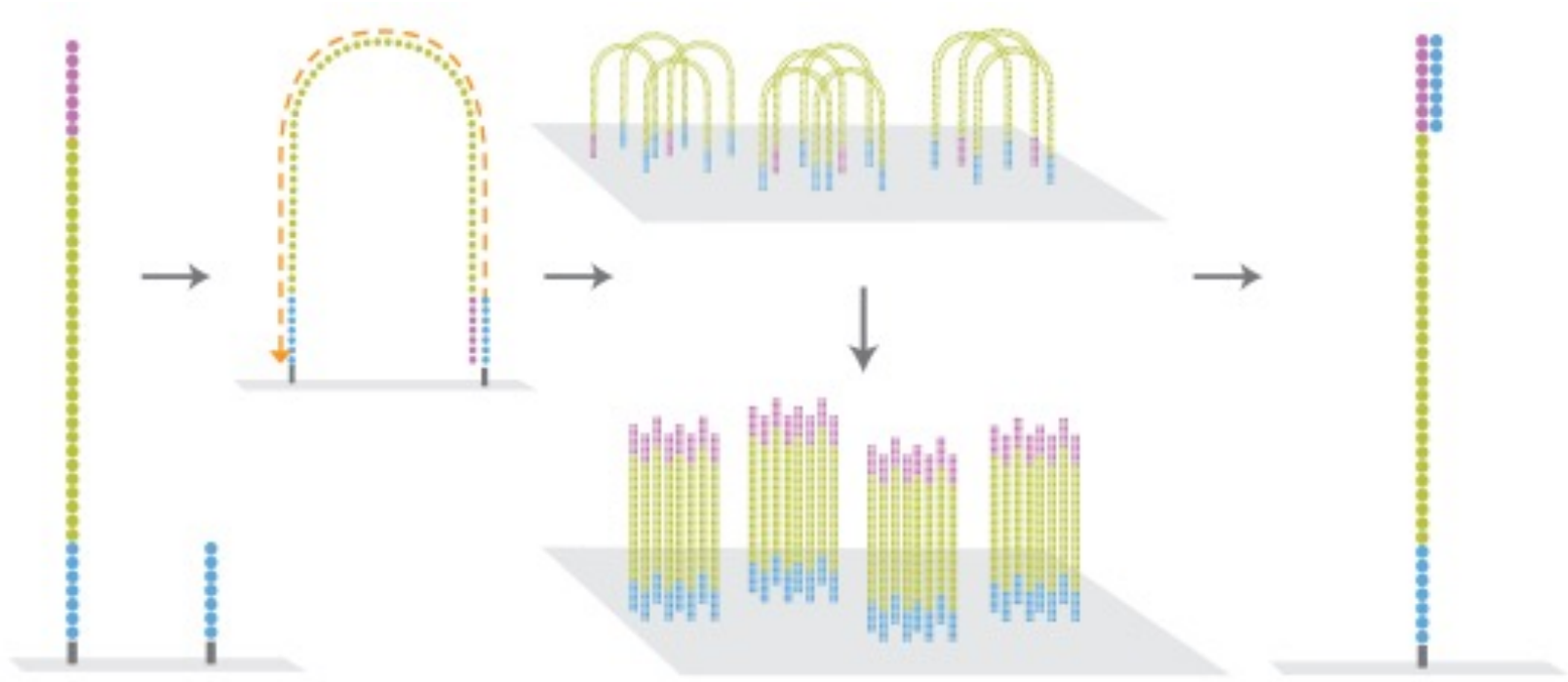
NGS Enrichment

- Sequence capture



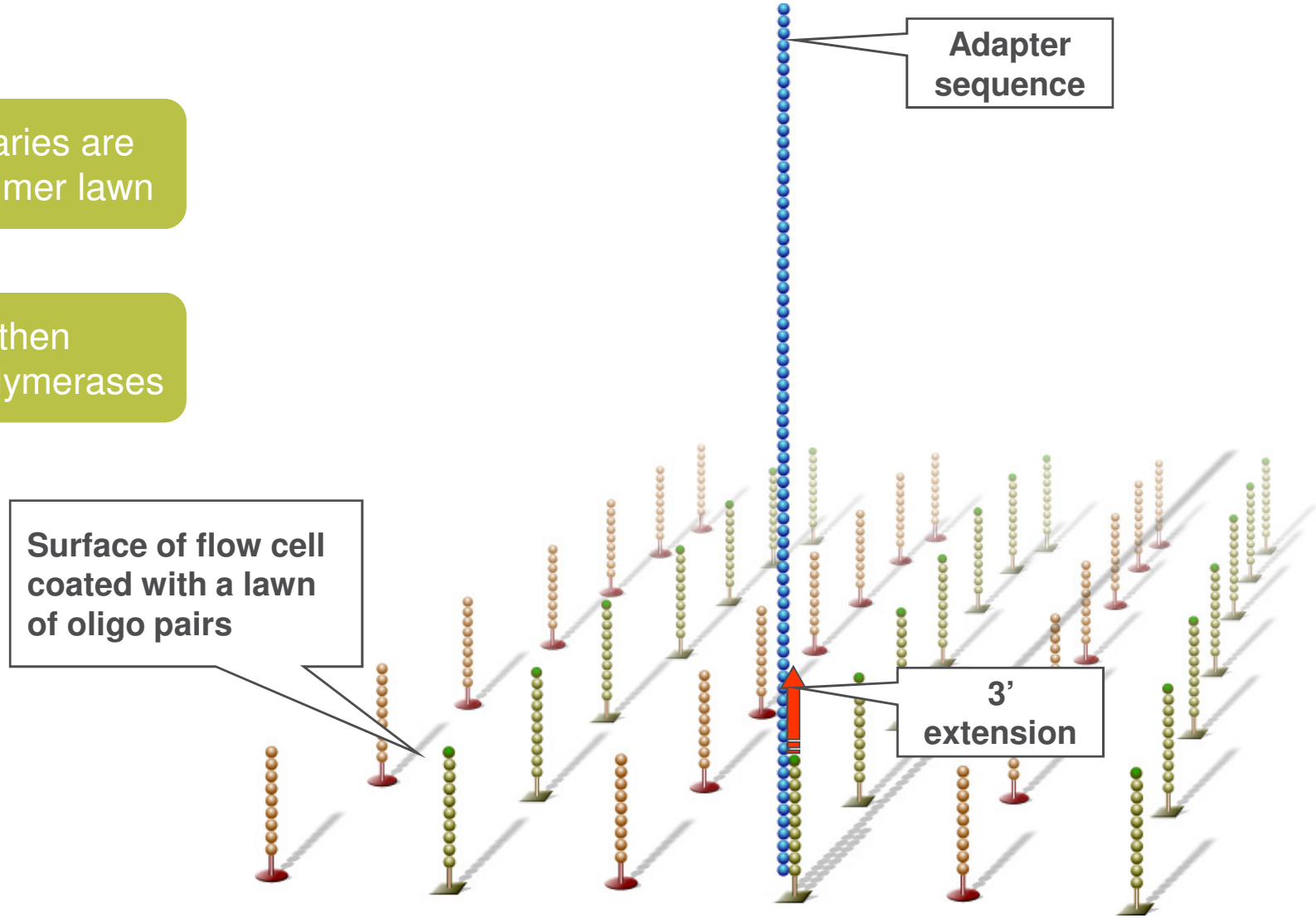
Illumina sequencing

NGS Illumina Clustering

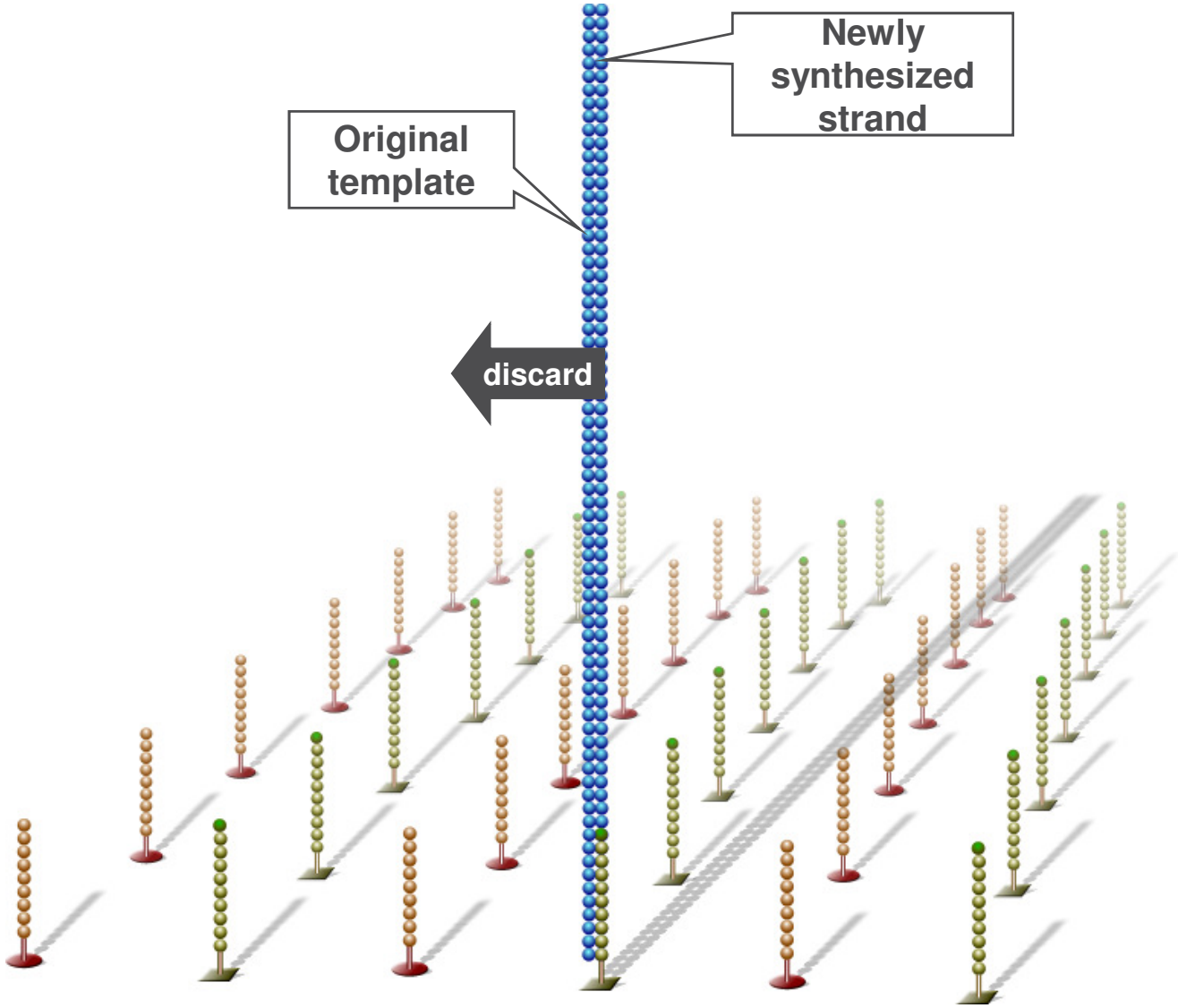


Single DNA libraries are hybridized to primer lawn

Bound libraries then extended by polymerases

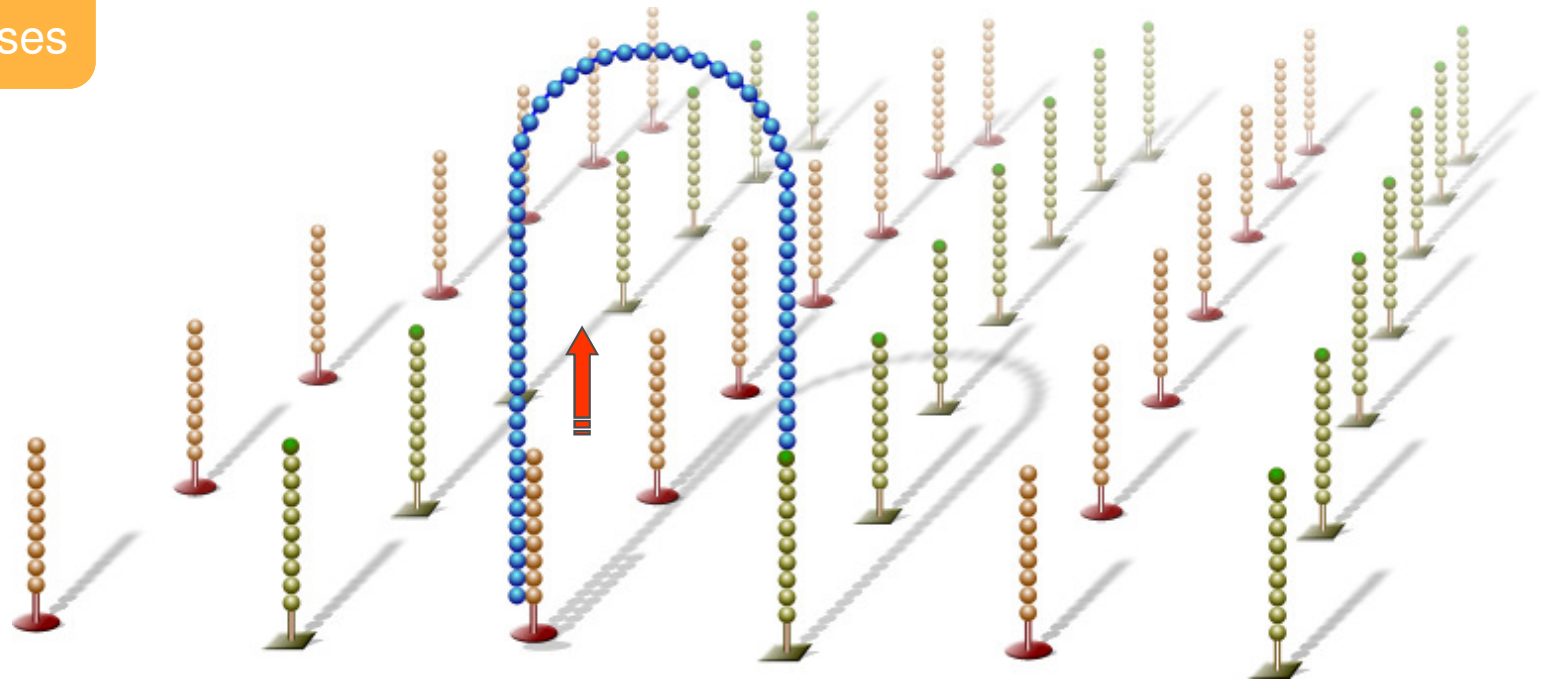


- Double-stranded molecule is denatured
- Original template washed away
- Newly synthesized strand is covalently attached to flow cell surface

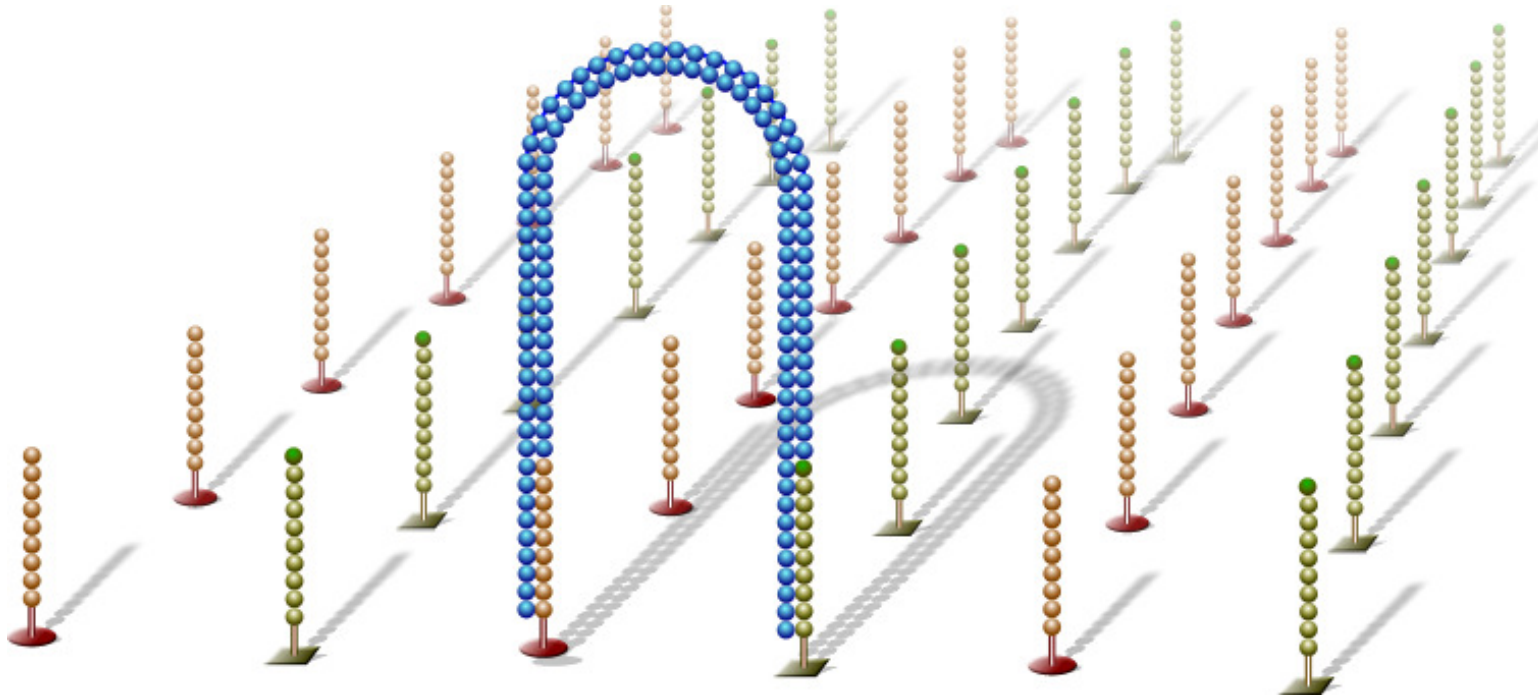


Single-stranded molecule flips over and forms a bridge by hybridizing to adjacent, complementary primer

Hybridized primer is extended by polymerases



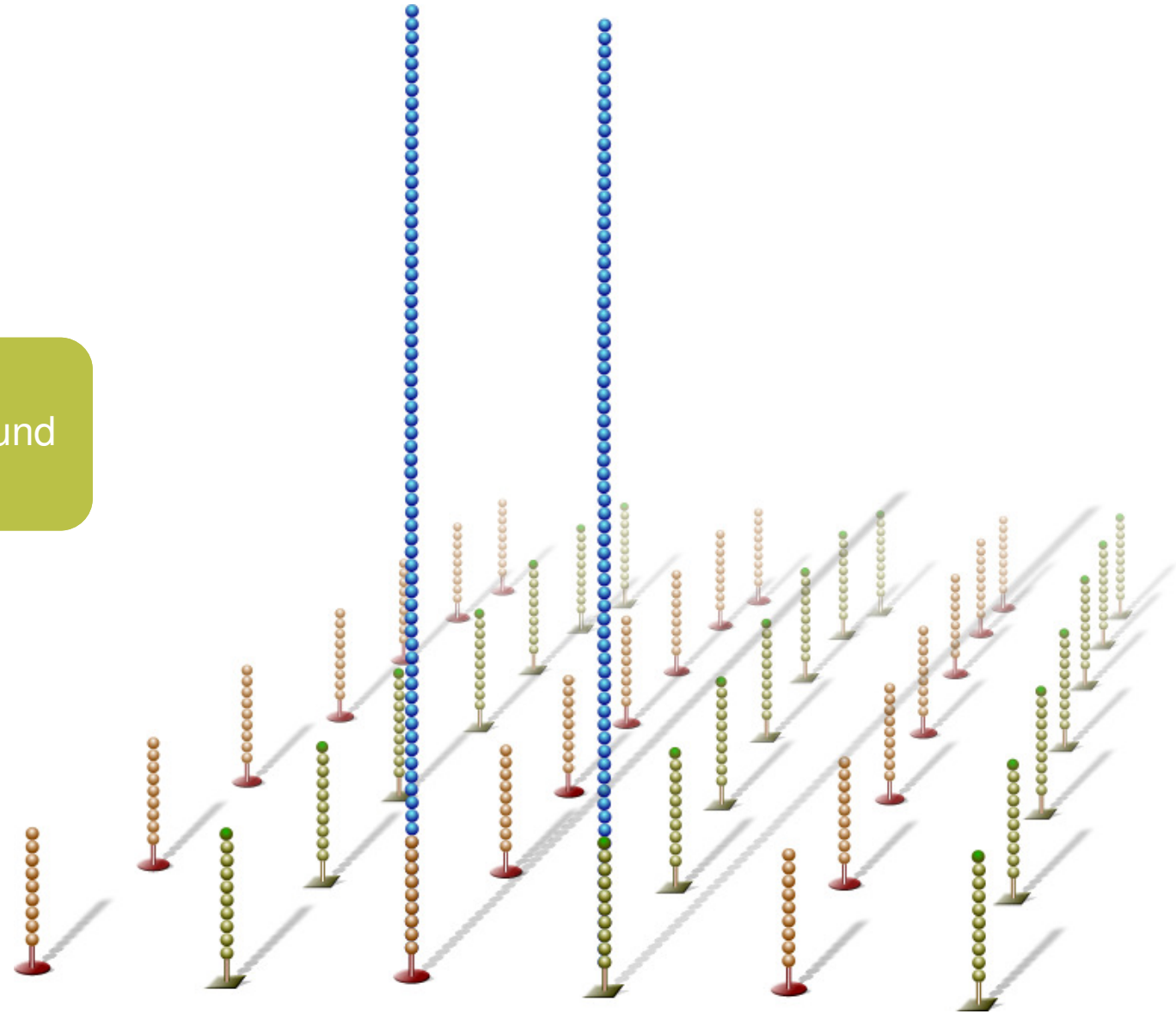
Double-stranded bridge is formed



Sample prep → Clonal Amplification → Parallel sequencing

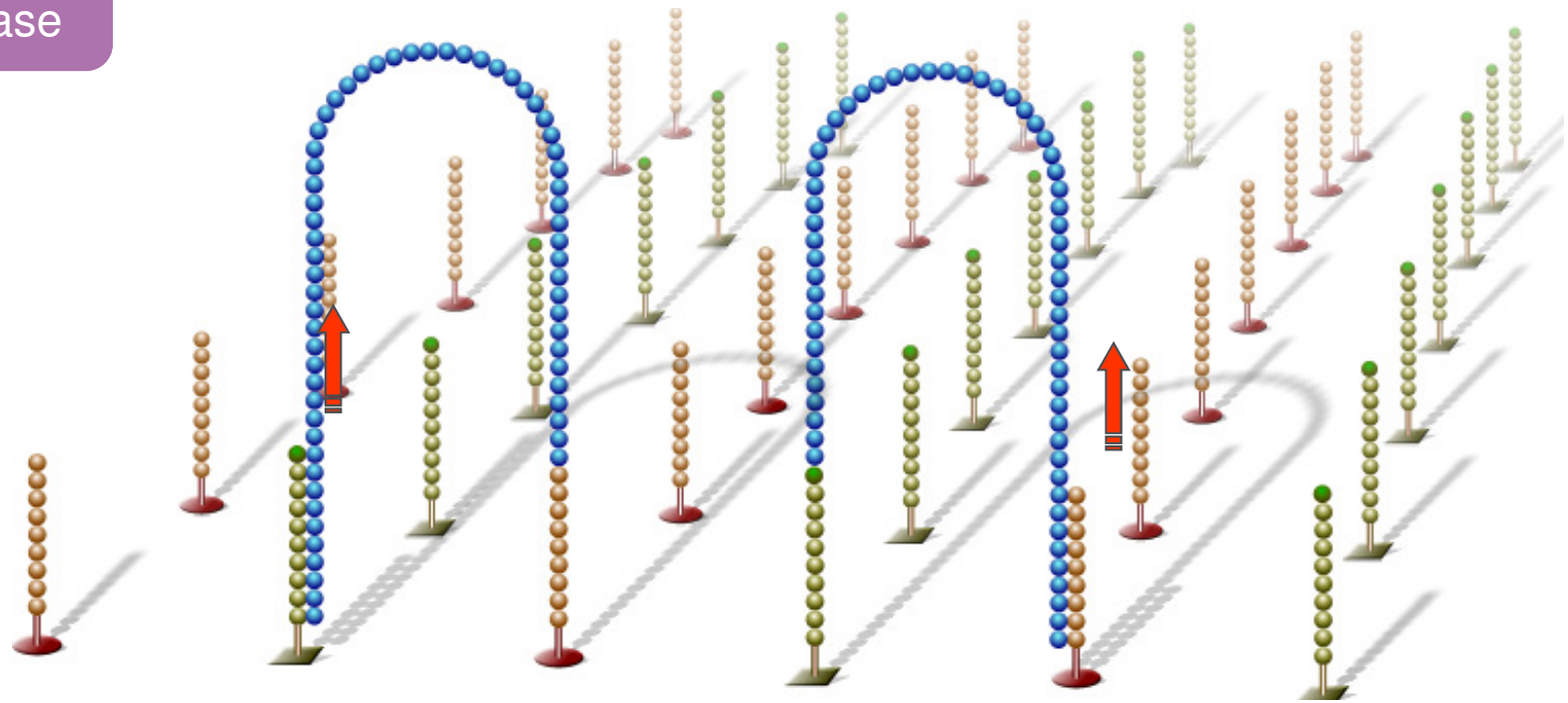
Double-stranded bridge is denatured

Result:
Two copies of covalently bound single-stranded templates

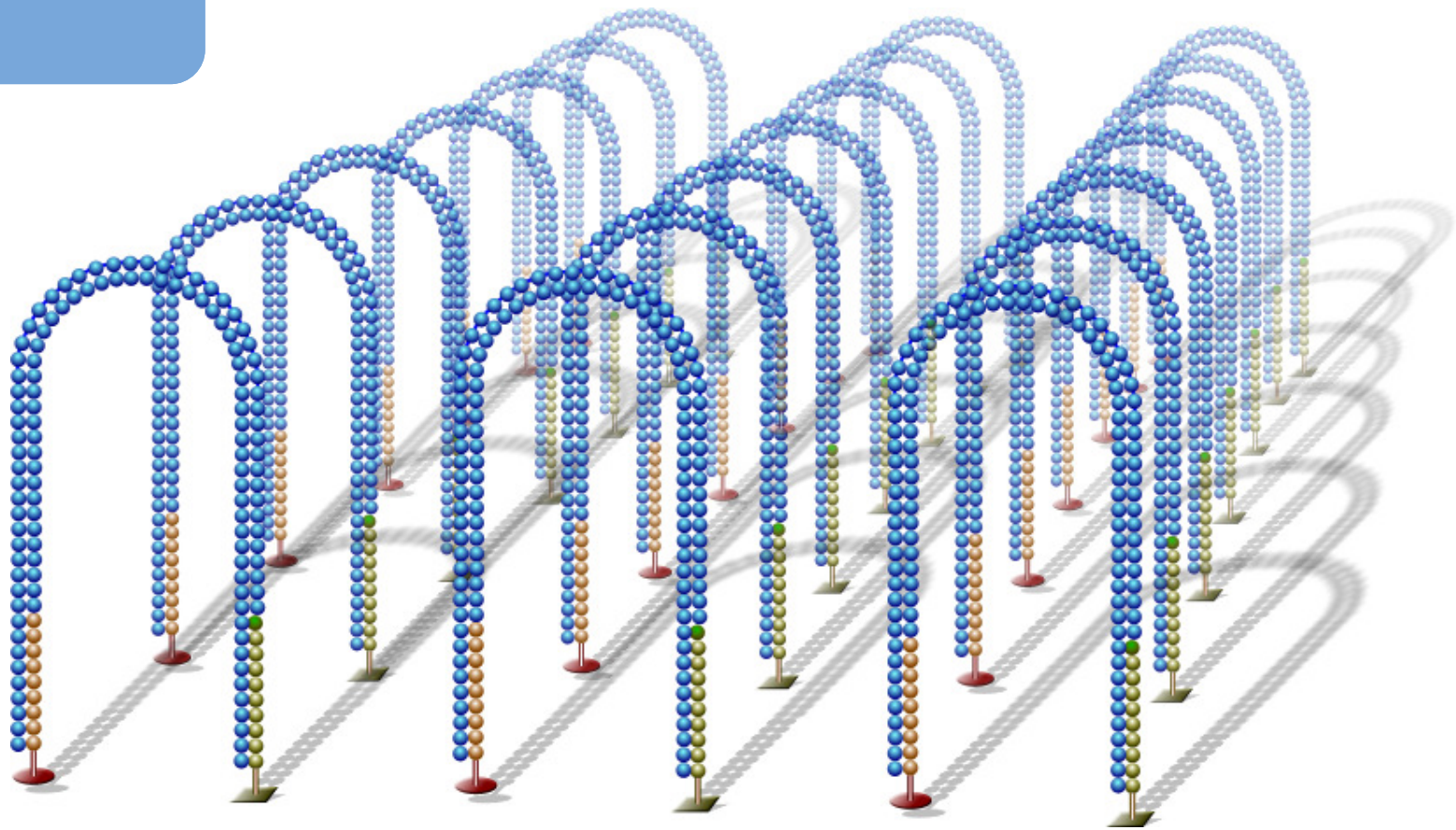


Single-stranded molecules flip over to hybridize to adjacent primers

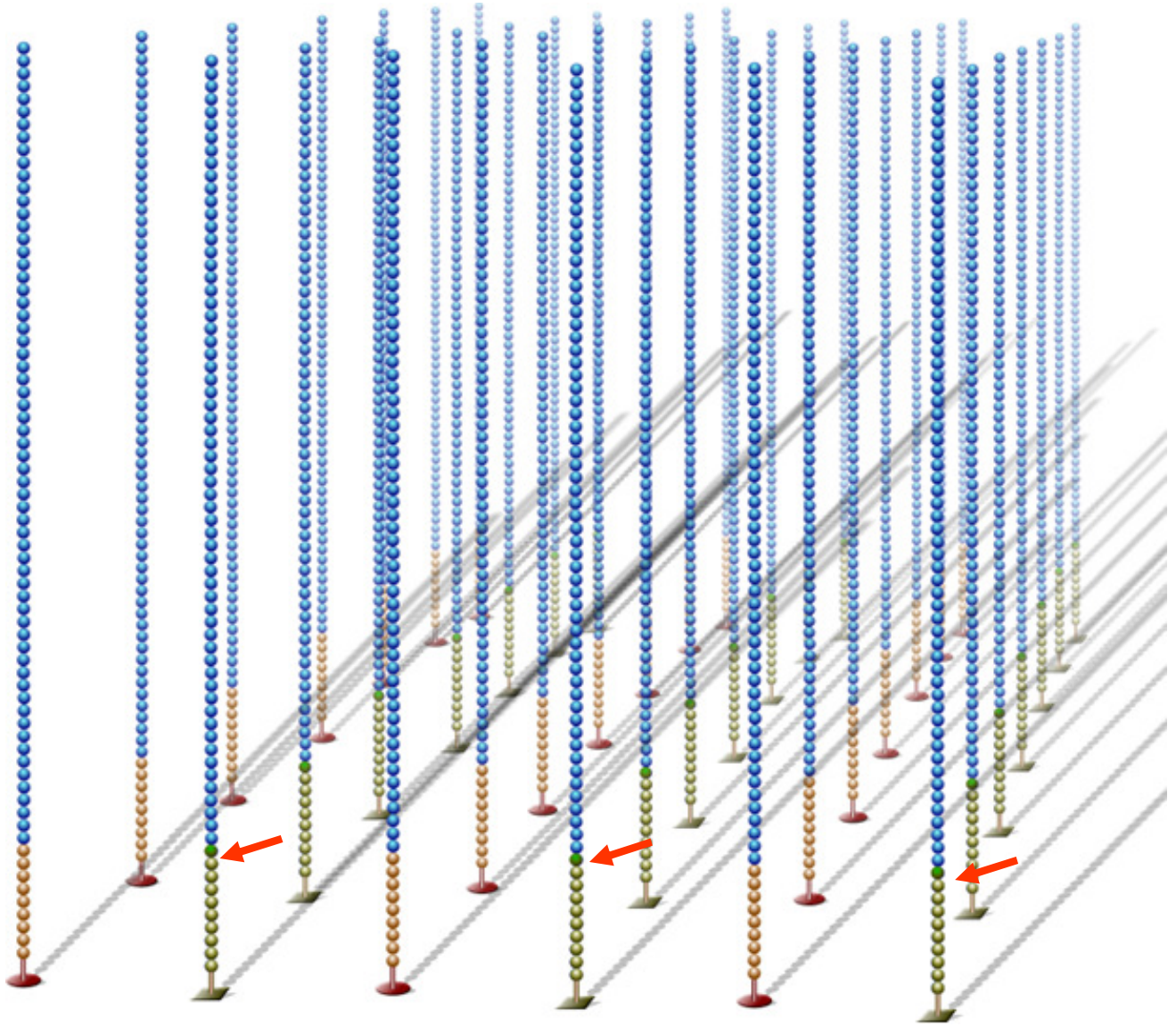
Hybridized primer is extended by polymerase



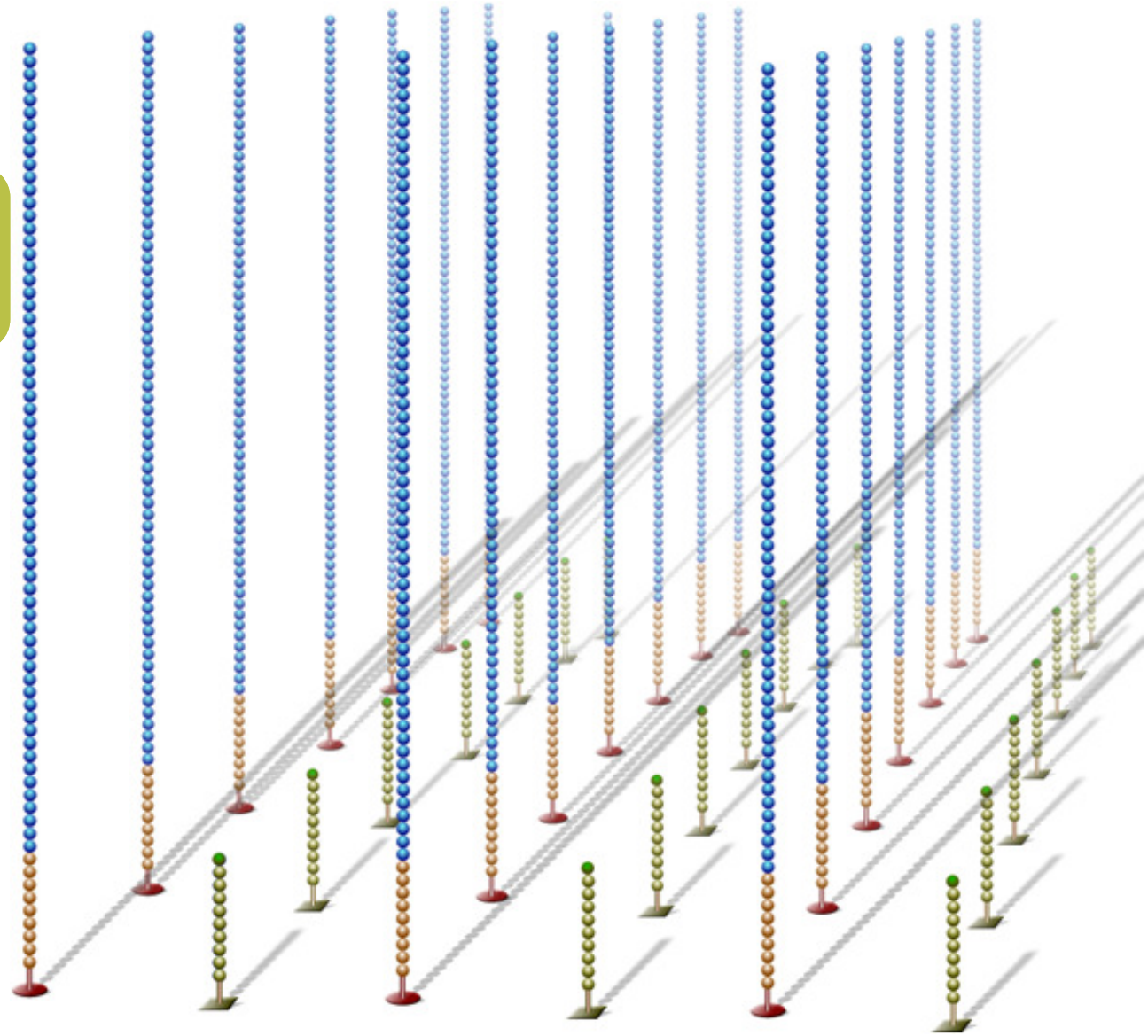
Bridge amplification cycle repeated until multiple bridges are formed



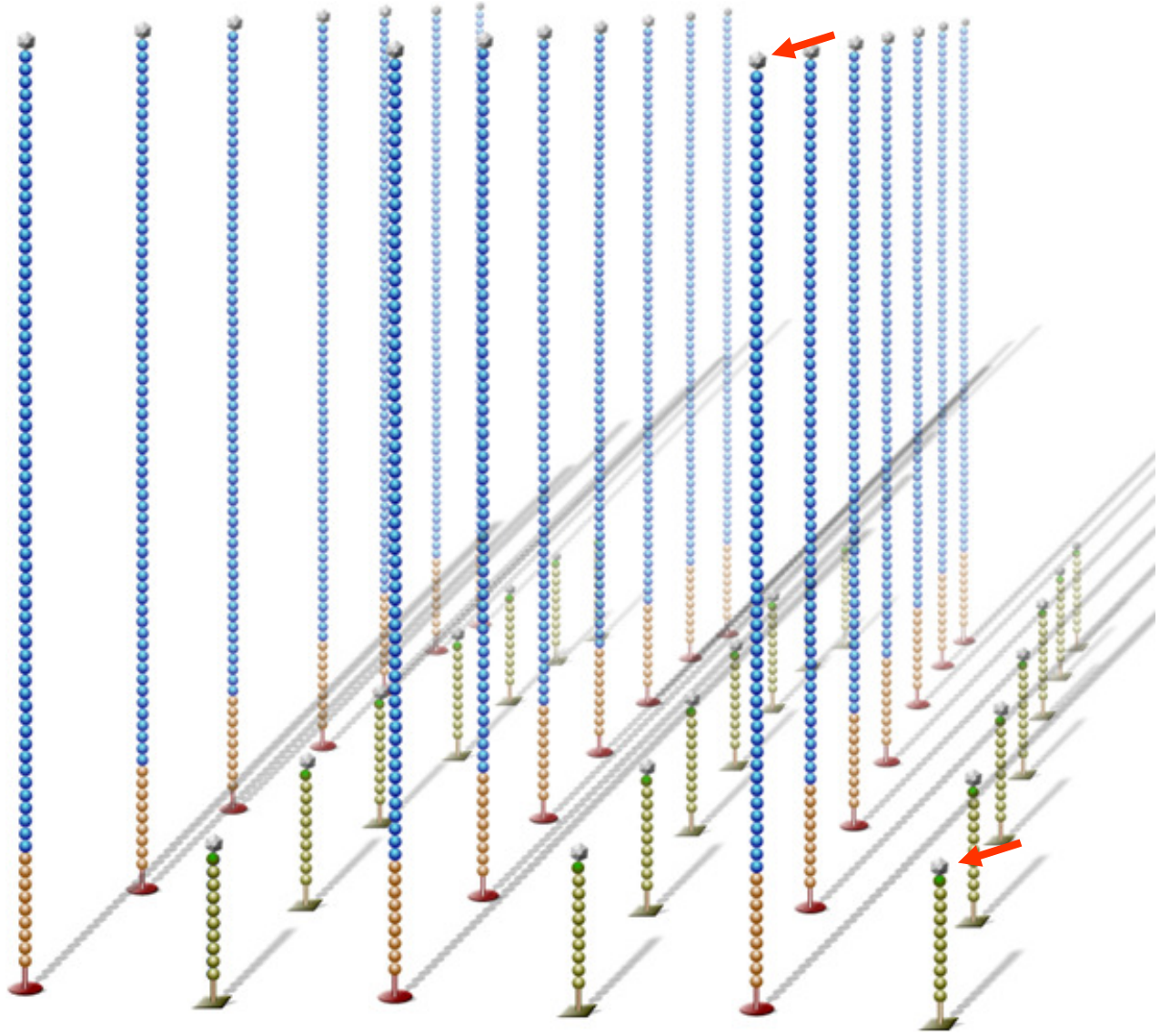
dsDNA bridges are denatured



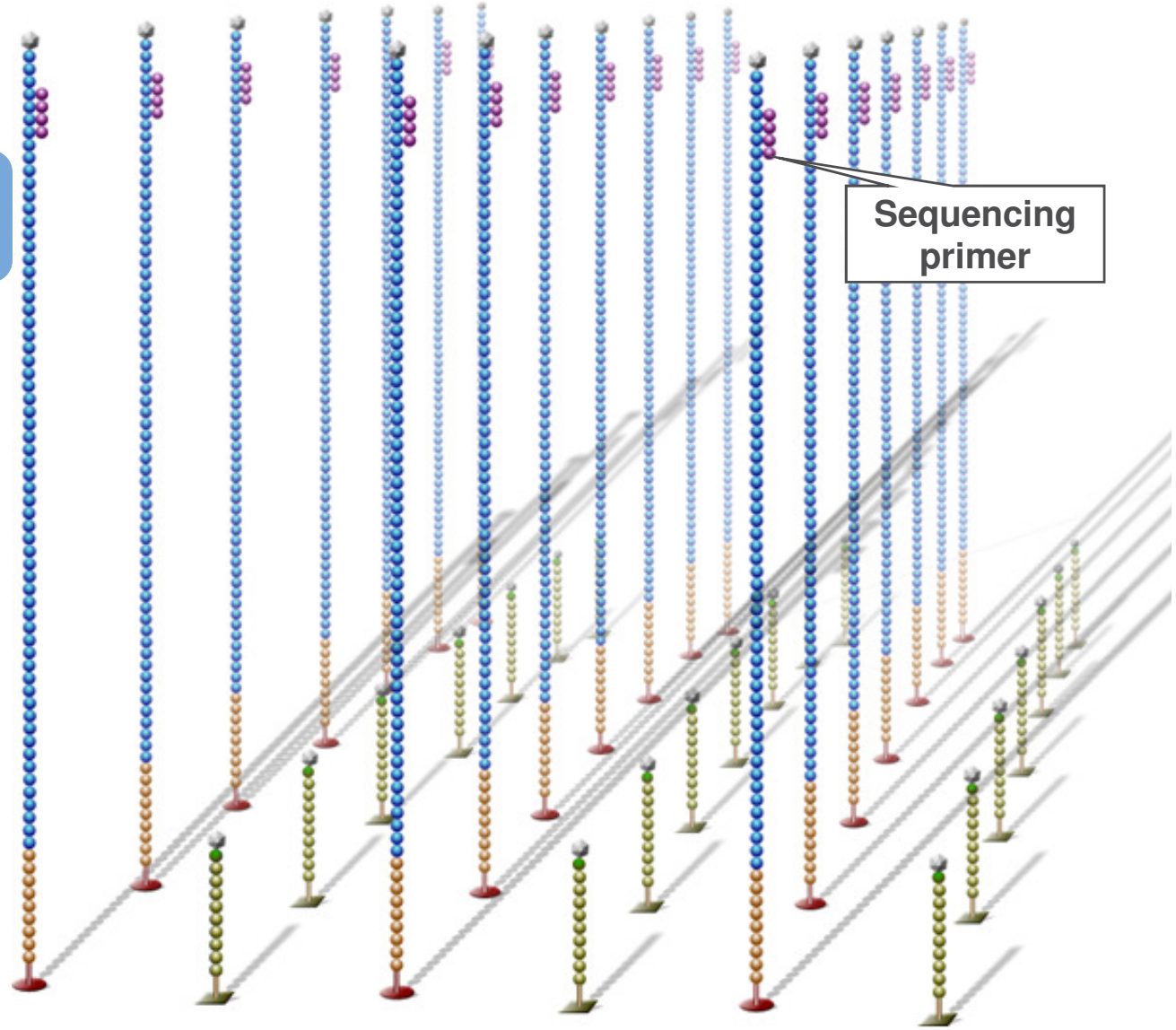
Reverse strands cleaved and washed away, leaving a cluster with forward strands only



Free 3’ ends are blocked to prevent unwanted DNA priming

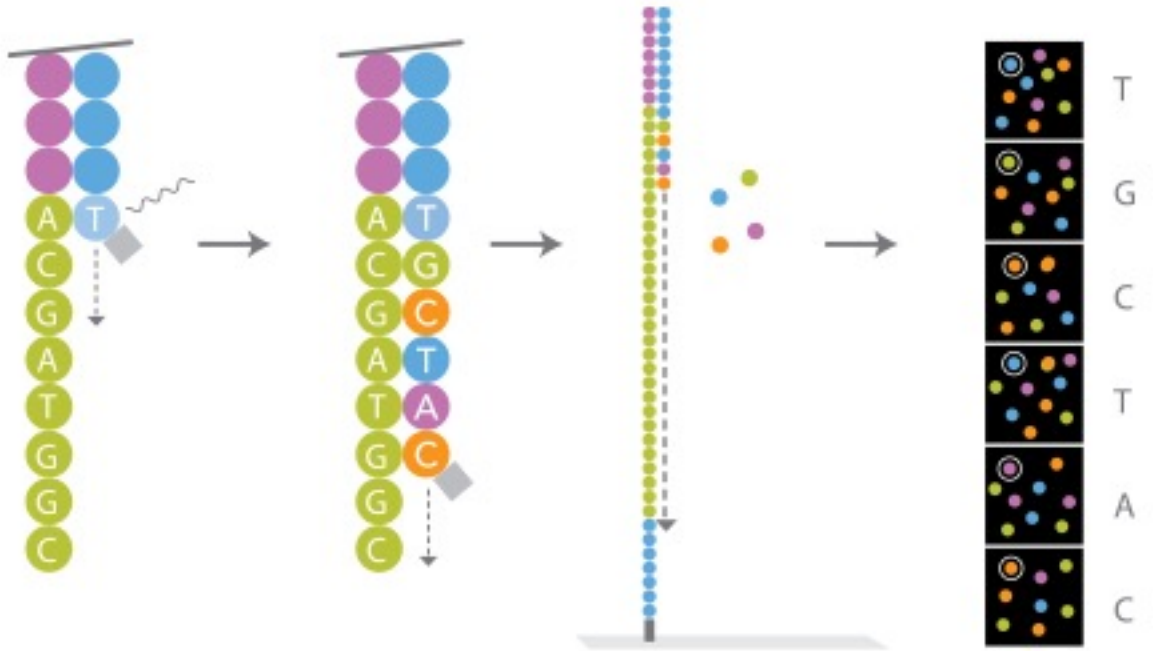


Sequencing primer is hybridized to adapter sequence



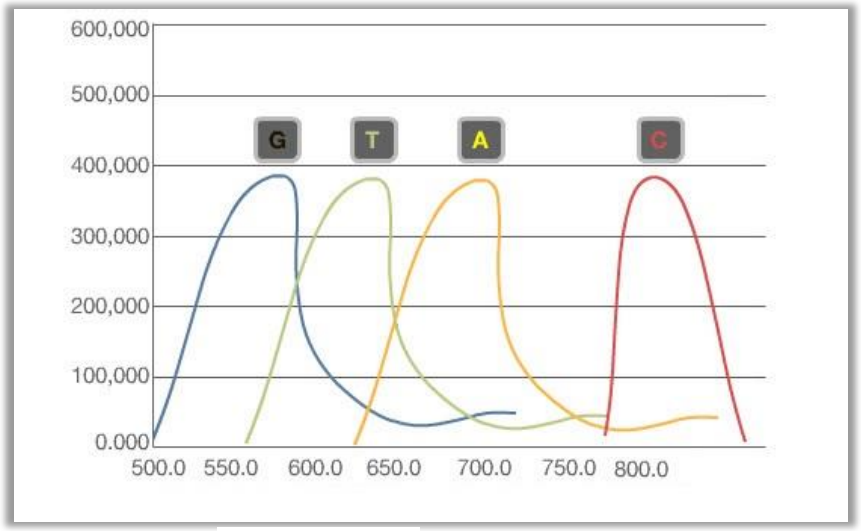
NGS Illumina Sequencing

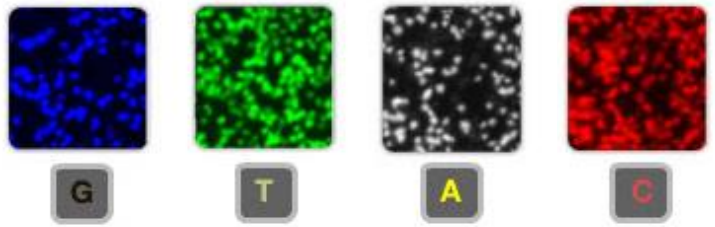
- **S**equencing **B**y **S**ynthesis (SBS)



NGS Illumina Sequencing

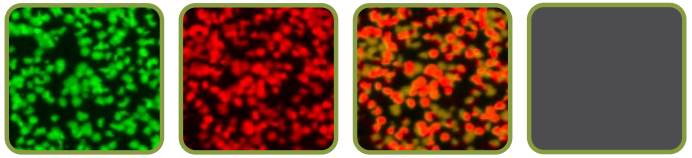
- Imaging MiSeq, HiSeq4000

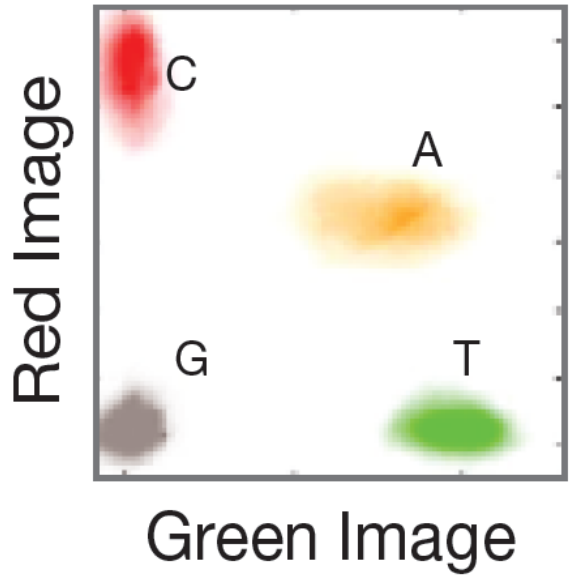




NGS Illumina Sequencing

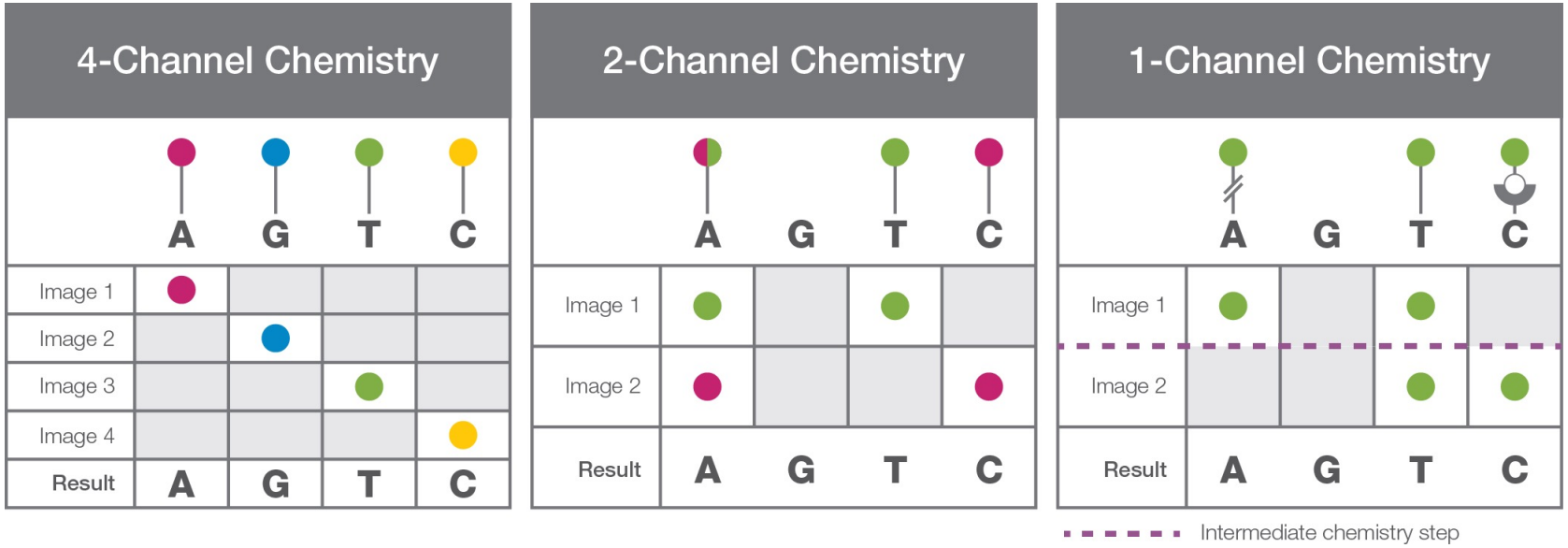
- Imaging NextSeq500, Nextseq2000* NovaSeq





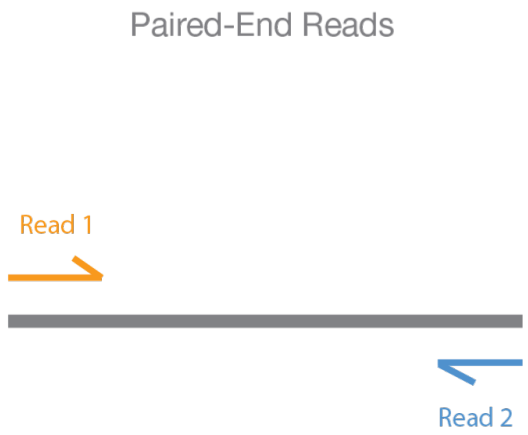
NGS Illumina Sequencing

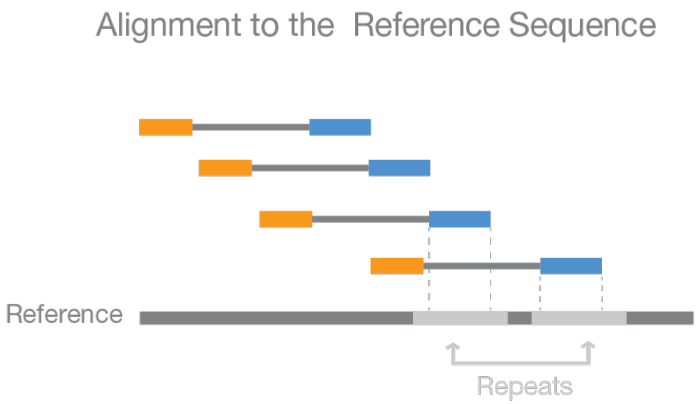
- Different SBS dyes

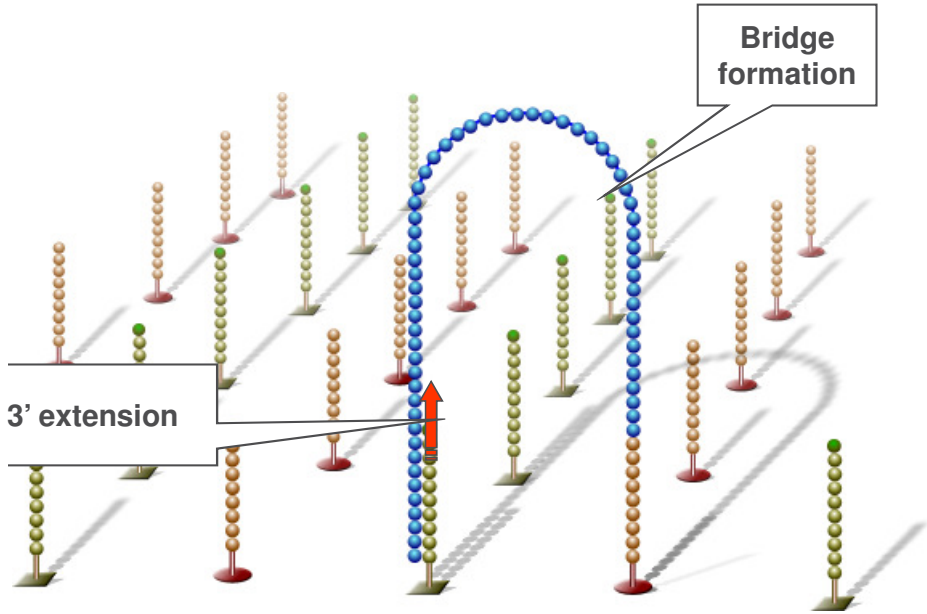


NGS Illumina Sequencing

- Paired-end sequencing







Illumina sequencing QC

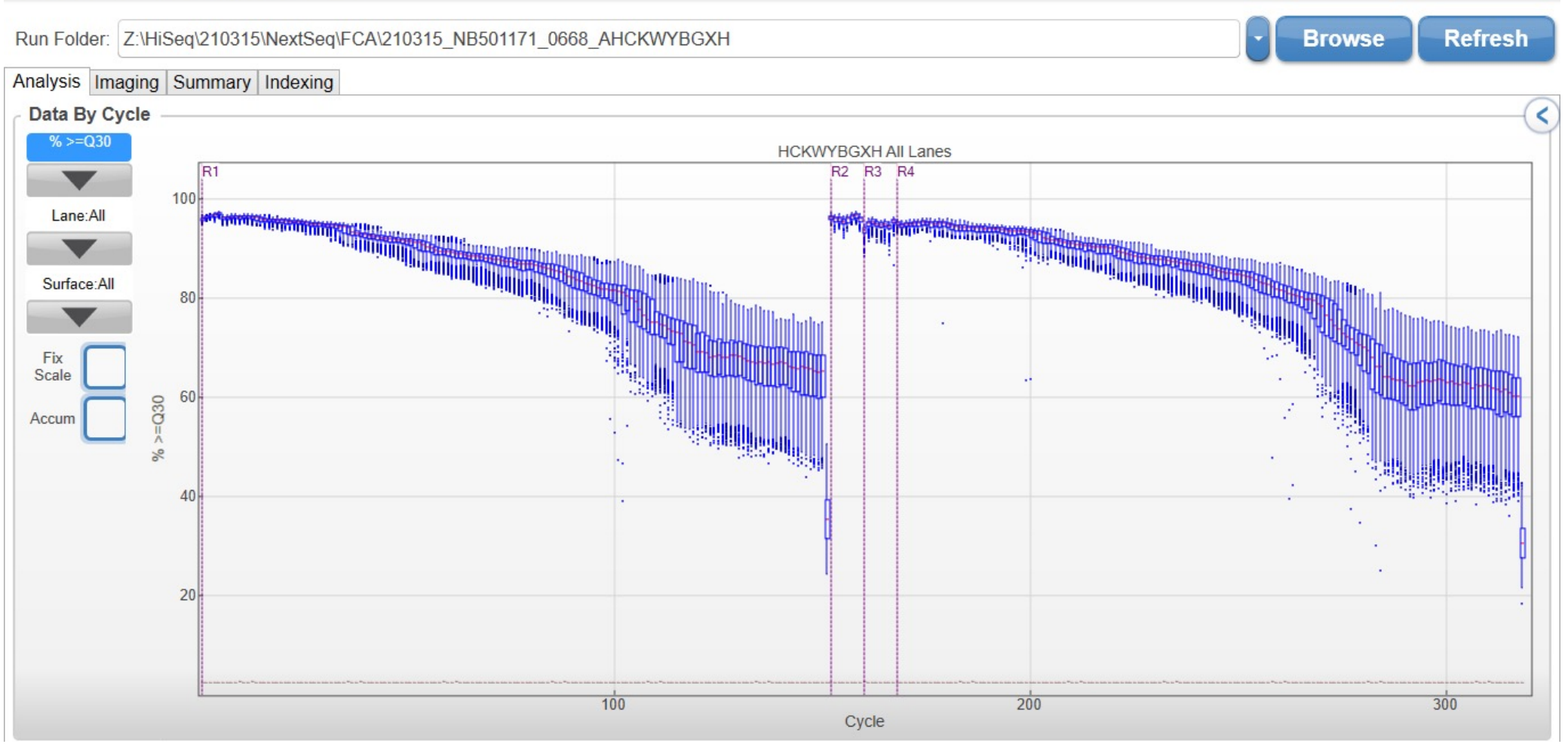
Quality Control

Run Summary

Level	Yield Total (G)	Projected Total Yield (G)	Aligned (%)	Error Rate (%)	Intensity Cycle 1	% >= Q30
Read 1	0,3	0,3	33,59	1,76	76	95,3
Read 2 (I)	0,0	0,0	0,00	0,00	77	64,9
Read 3 (I)	0,0	0,0	0,00	0,00	379	97,2
Read 4	0,3	0,3	33,07	1,79	100	
Total	0,6	0,6	33,33	1,78	158	

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Q30 by cycle


Run Folder: Z:\HiSeq\210315\NextSeq\FCA\210315_NB501171_0668_AHCKWYBGXH
Browse
Refresh
Analysis
Imaging
Summary
Indexing
Data By Cycle
% >=Q30
Lane:All
Surface:All
Fix Scale
Accum
HCKWYBGXH All Lanes
R1
R2
R3
R4
100
80
60
40
20
% >=Q30
100
200
300
Cycle

NGS Illumina Sequencing

- Quality Control

Run Summary

Level	Yield Total (G)	Projected Total Yield (G)	Aligned (%)	Error Rate (%)	Intensity Cycle 1	% >= Q30
Read 1	0,3	0,3	33,59	1,76	76	95,3
Read 2 (I)	0,0	0,0	0,00	0,00	77	64,9
Read 3 (I)	0,0	0,0	0,00	0,00	379	97,2
Read 4	0,3	0,3	33,07	1,79	100	88,3
Total	0,6	0,6	33,33	1,78	158	91,5

Read 1

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	2	877 +/- 6	94,89 +/- 1,62	0,108 / 0,078	1,23	1,16	95,3	0,3	250	33,59 +/- 0,22	1,76 +/- 0,03	0,16 +/- 0,08	0,20 +/- 0,08	0,26 +/- 0,06	76 +/- 8

Read 2 (I)

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	2	877 +/- 6	94,89 +/- 1,62	0,000 / 0,000	1,23	1,16	64,9	0,0	0	0,00 +/- 0,00	0,00 +/- 0,00	0,00 +/- 0,00	0,00 +/- 0,00	0,00 +/- 0,00	77 +/- 0

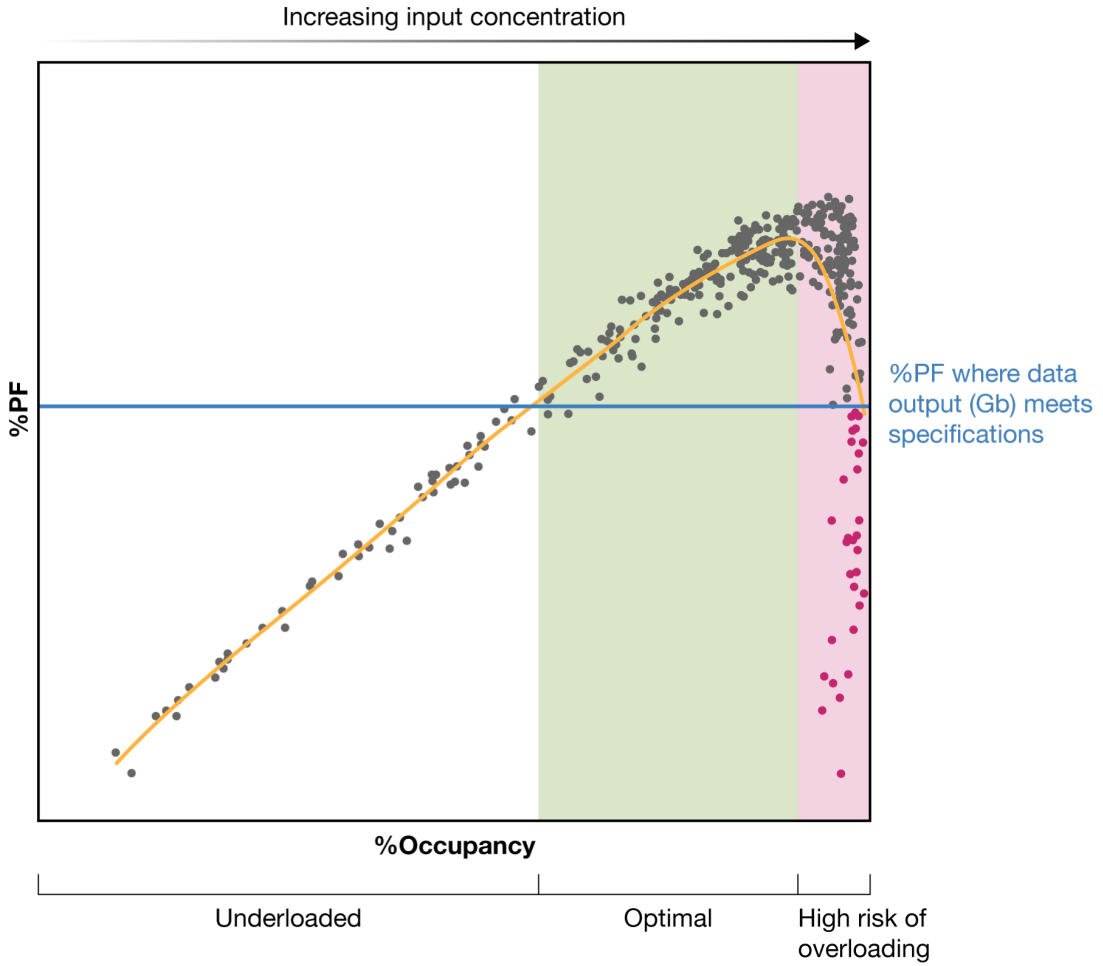
Read 3 (I)

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	2	877 +/- 6	94,89 +/- 1,62	0,000 / 0,000	1,23	1,16	97,2	0,0	0	0,00 +/- 0,00	0,00 +/- 0,00	0,00 +/- 0,00	0,00 +/- 0,00	0,00 +/- 0,00	379 +/- 6

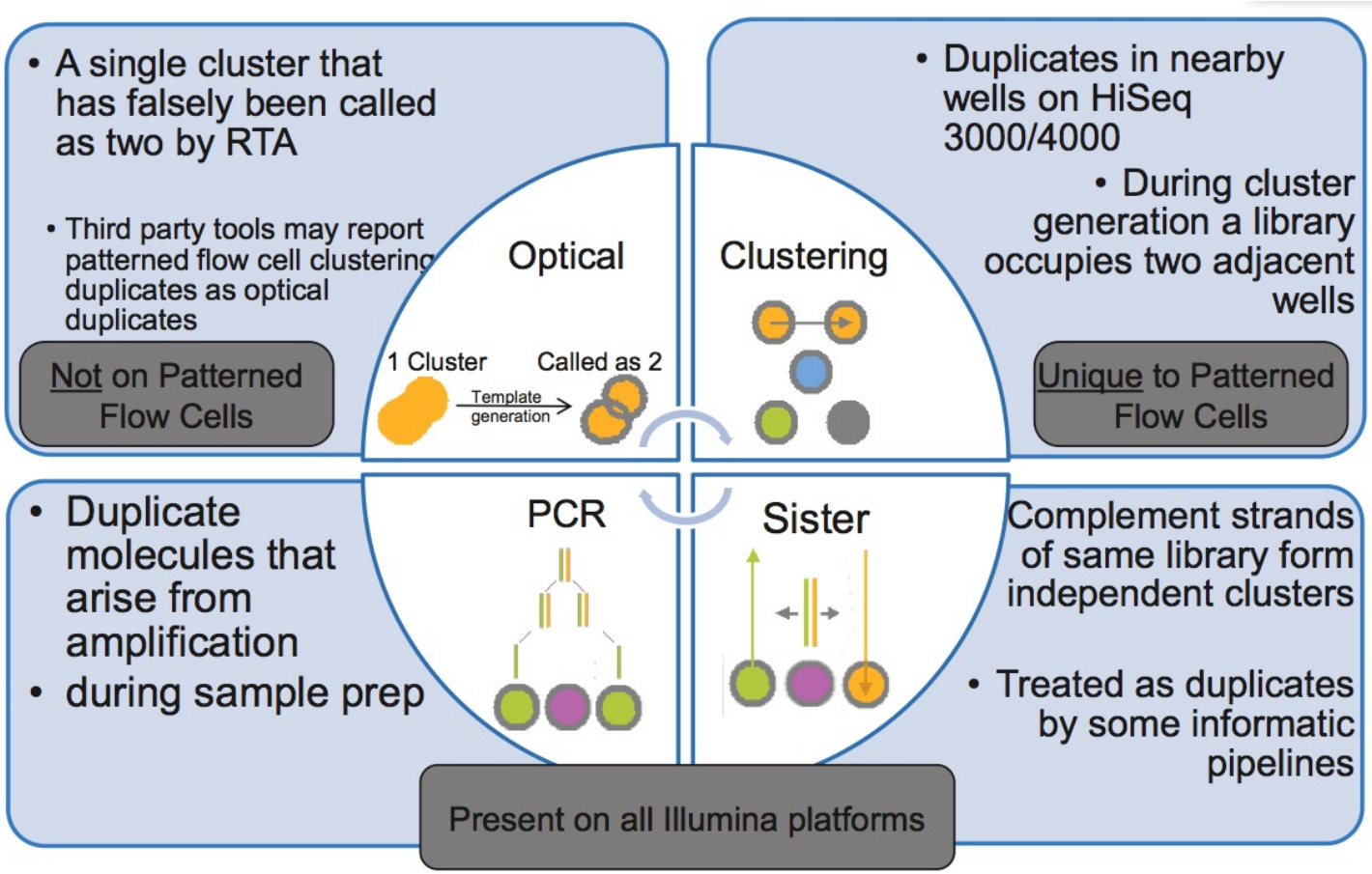
Read 4

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	2	877 +/- 6	94,89 +/- 1,62	0,147 / 0,068	1,23	1,16	88,3	0,3	250	33,07 +/- 0,61	1,79 +/- 0,01	0,20 +/- 0,10	0,25 +/- 0,08	0,30 +/- 0,06	100 +/- 5

Optimal loading



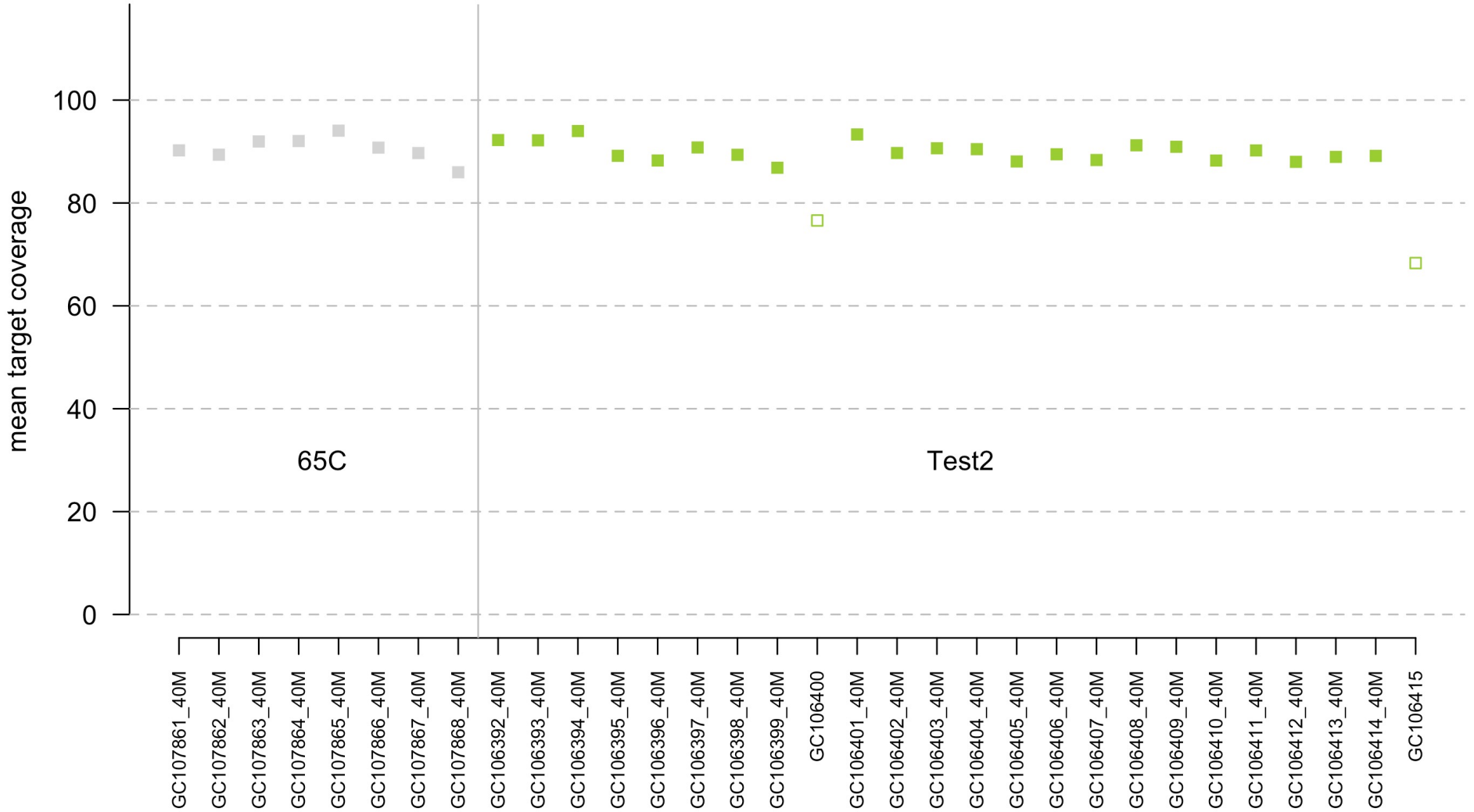
Duplicates


• A single cluster that has falsely been called as two by RTA
• Third party tools may report patterned flow cell clustering duplicates as optical duplicates
Not on Patterned Flow Cells
Optical
1 Cluster
Called as 2
Template generation
Clustering
• Duplicates in nearby wells on HiSeq 3000/4000
• During cluster generation a library occupies two adjacent wells
Unique to Patterned Flow Cells
• Duplicate molecules that arise from amplification
• during sample prep
PCR
Sister
Complement strands of same library form independent clusters
• Treated as duplicates by some informatic pipelines
Present on all Illumina platforms

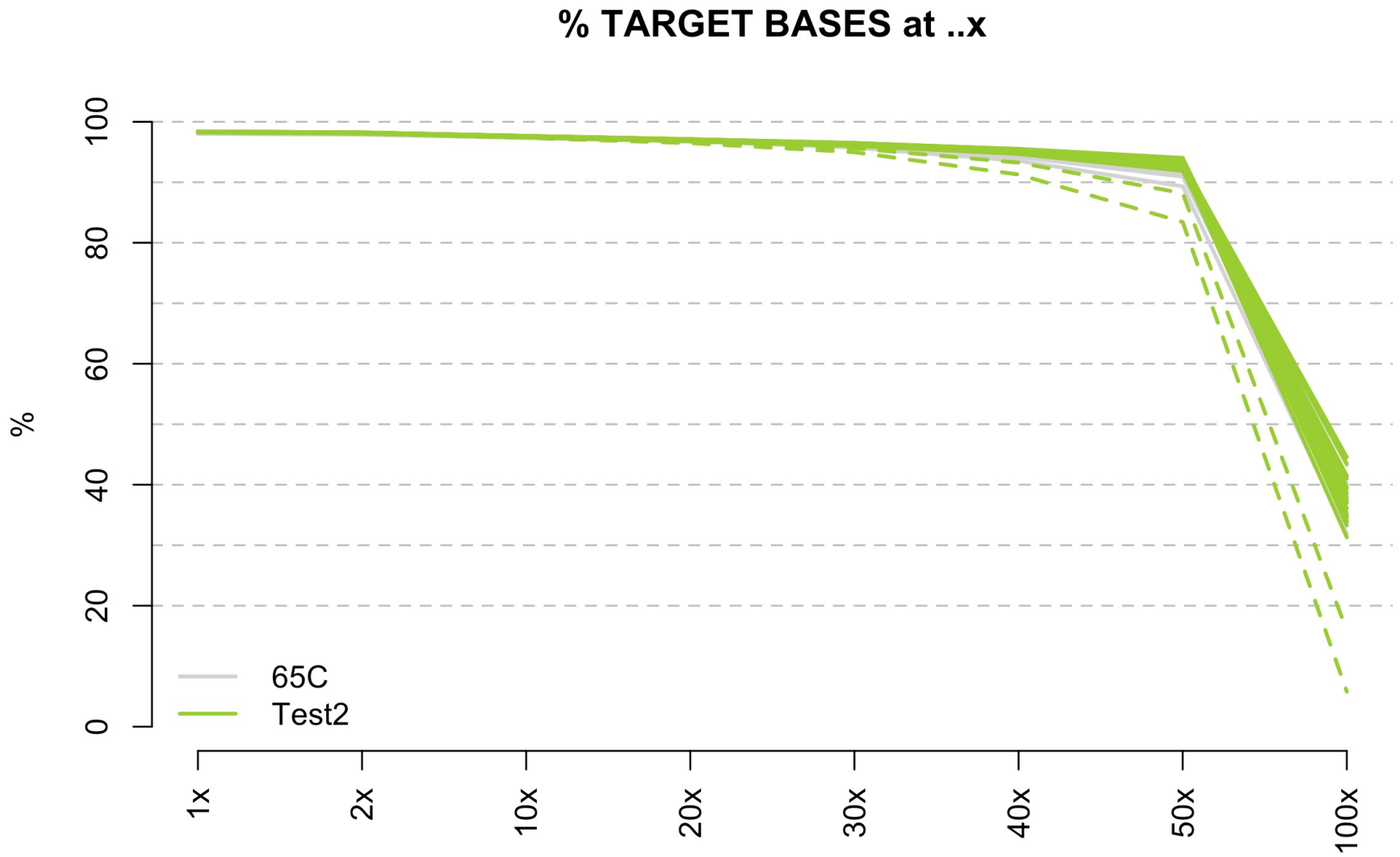
Common loss in Illumina

- Insufficient coverage: suboptimal loading
- Low complexity library
- Overloading
- Repeats
- Duplicates
- Too short fragment size
  - Overlap
  - High % adapter
- Low Q30

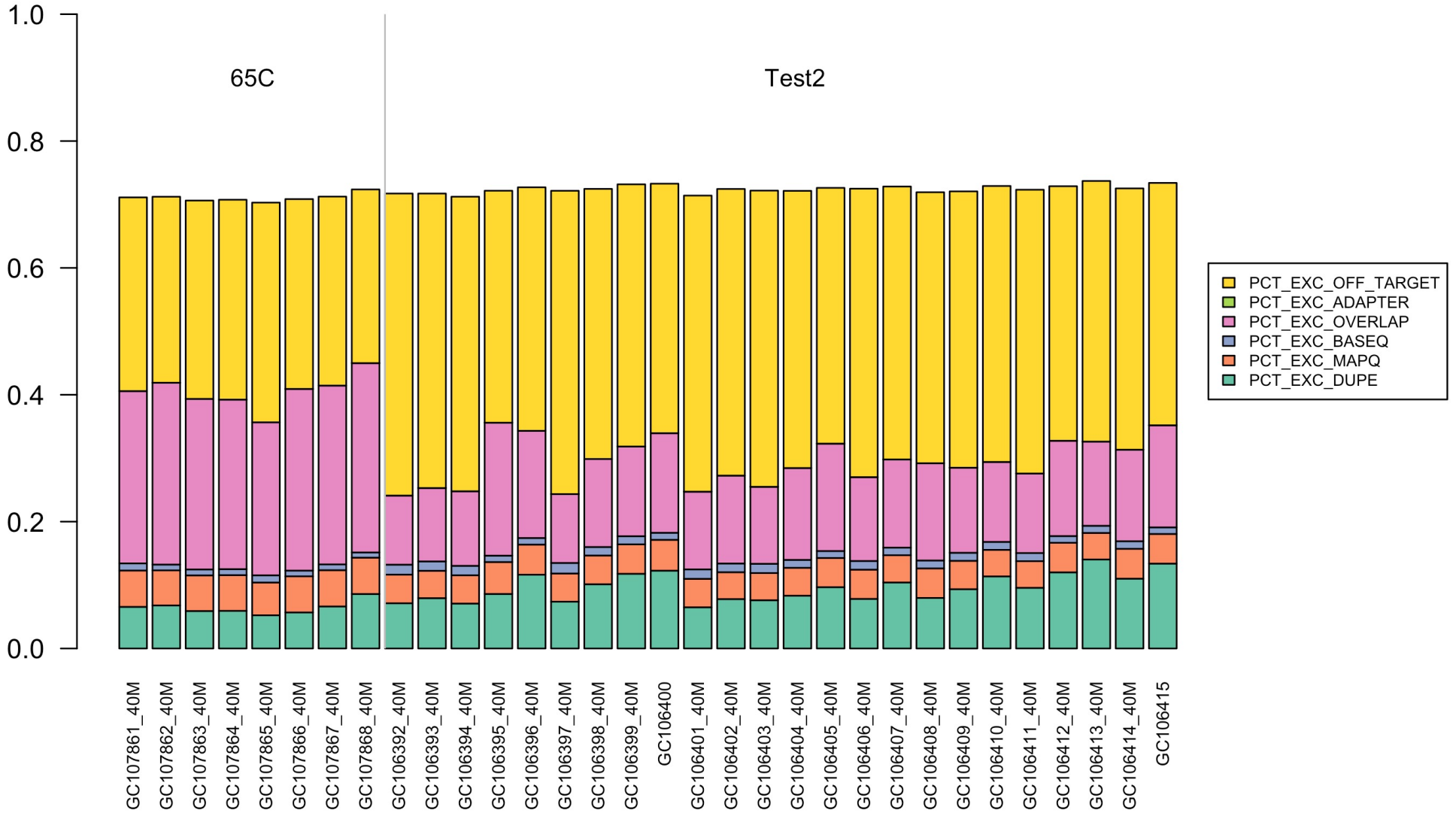
Metrics Picard – coverage technical target



Metrics Picard – uniformity



Metrics Picard – Excluded bases



CORELEUVEN

IGV screenshot of part of TRDN (chr6:123,247,689-123,284,977)

Nanopore sequencing

- MinION
- GridION
- PromethION
- SmidgION

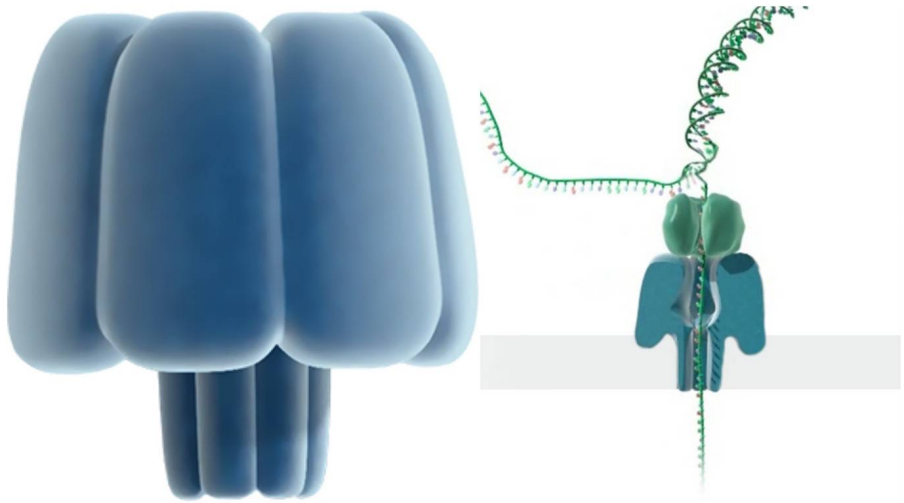
Oxford Nanopore sequencing

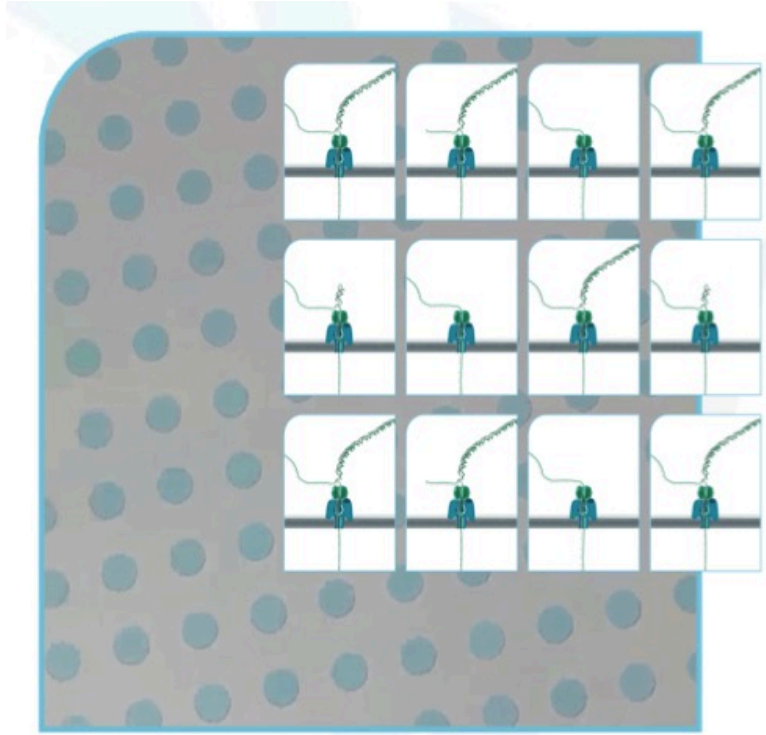
Tether keeps DNA fragment on the membrane leading to a ~20K fold higher DNA concentration close to the pore.

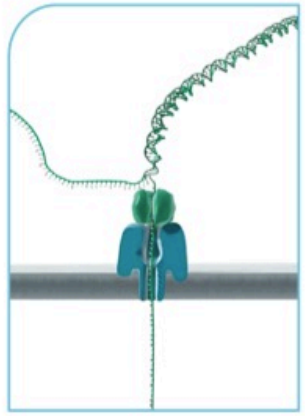
Motor protein unwinds DNA and ratchets it though the pore.

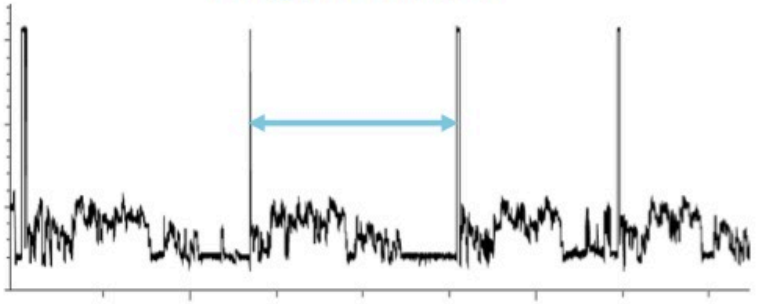
Abasic nucleotides in the hairpin are a recognition point.

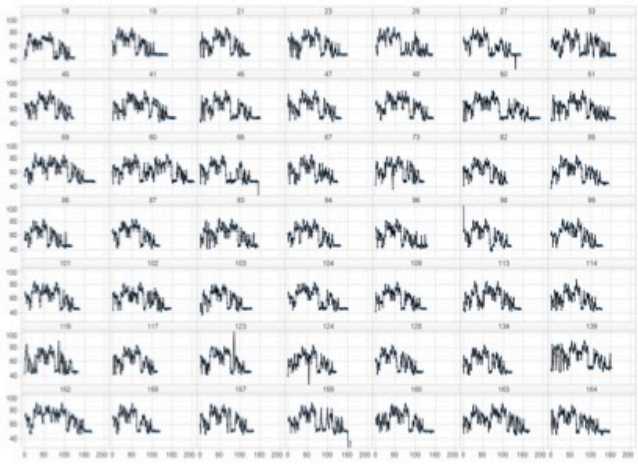
Brake protein prevents the motor protein from zipping through the complement strand.







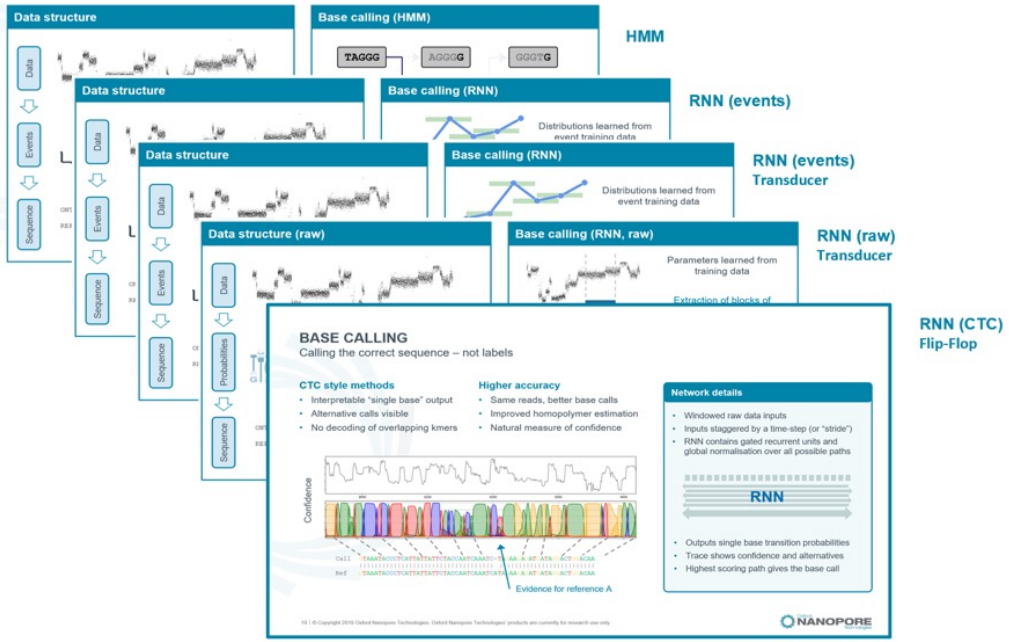




- Data acquired as full length reads – real time
- Data throughput = No. pores x average speed/pore

From squiggles to sequencing

- New basecaller: extract more correct information from squiggles
- Training of base caller for methylation data



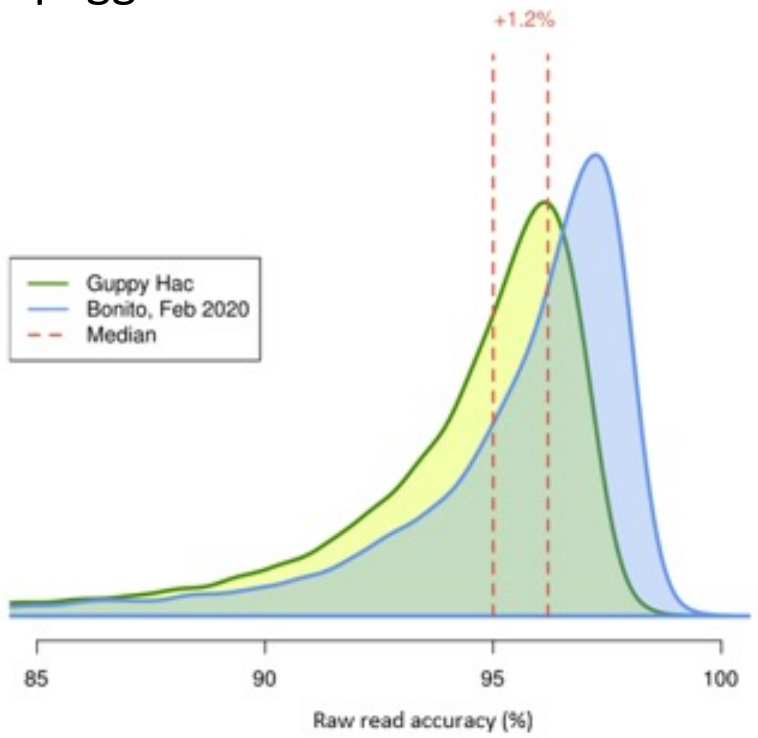
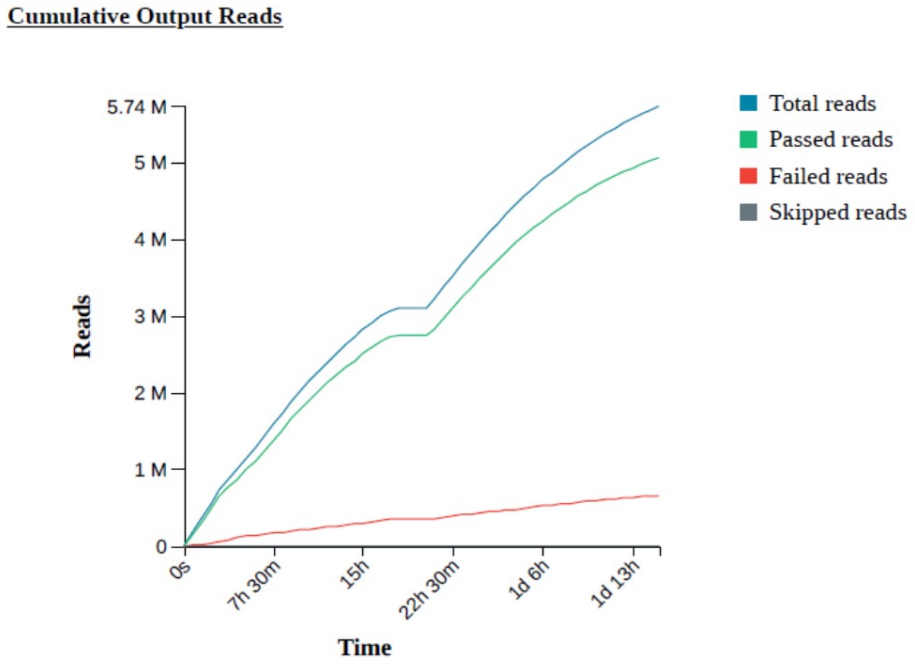
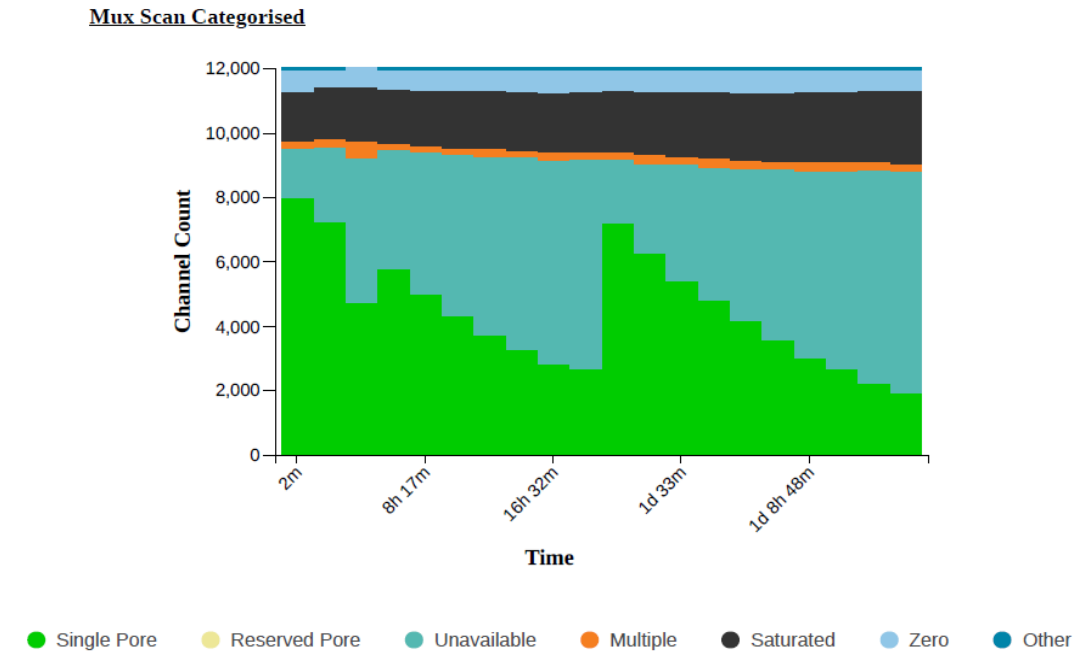


Figure 1: Raw read accuracy of Bonito basecaller on the human reference genome NA12878 against high-accuracy Guppy, currently integrated into MinKNOW onboard nanopore devices.

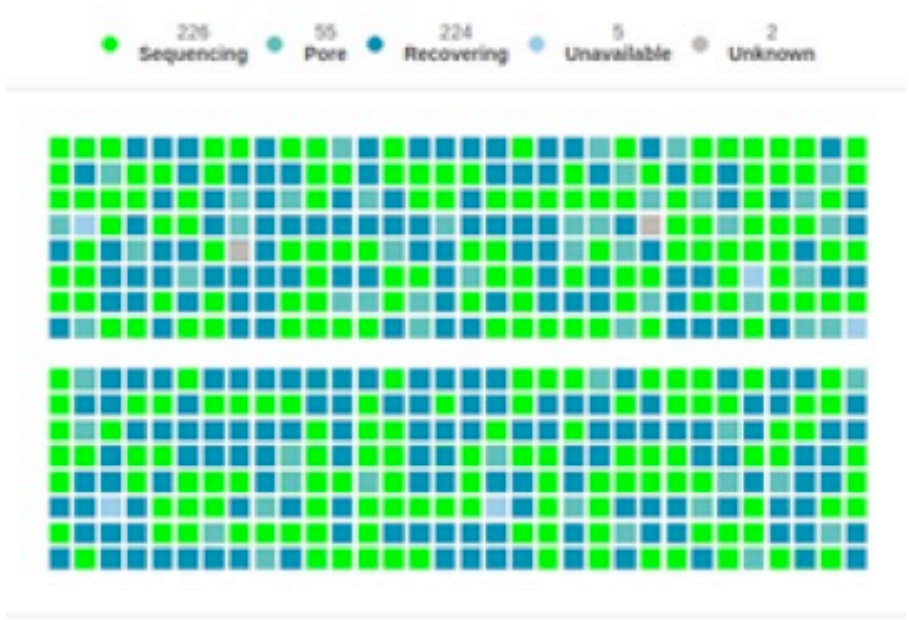
Run QC on Nanopore

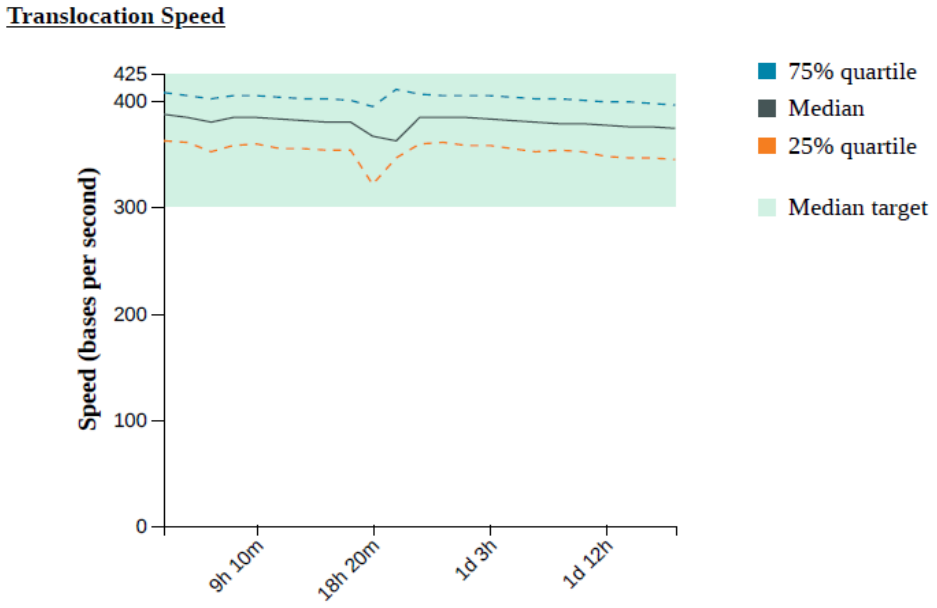
- Every flowcell is different
- Nuclease wash (and refueling) increases output





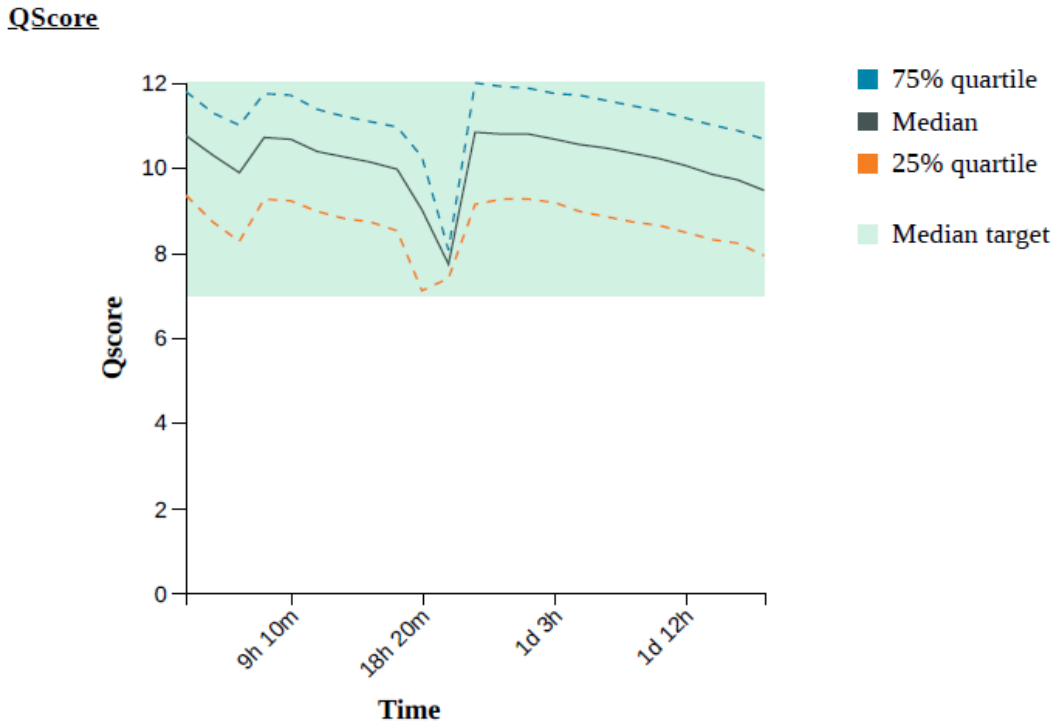
Translocation speed and pore status





Run QC on Nanopore

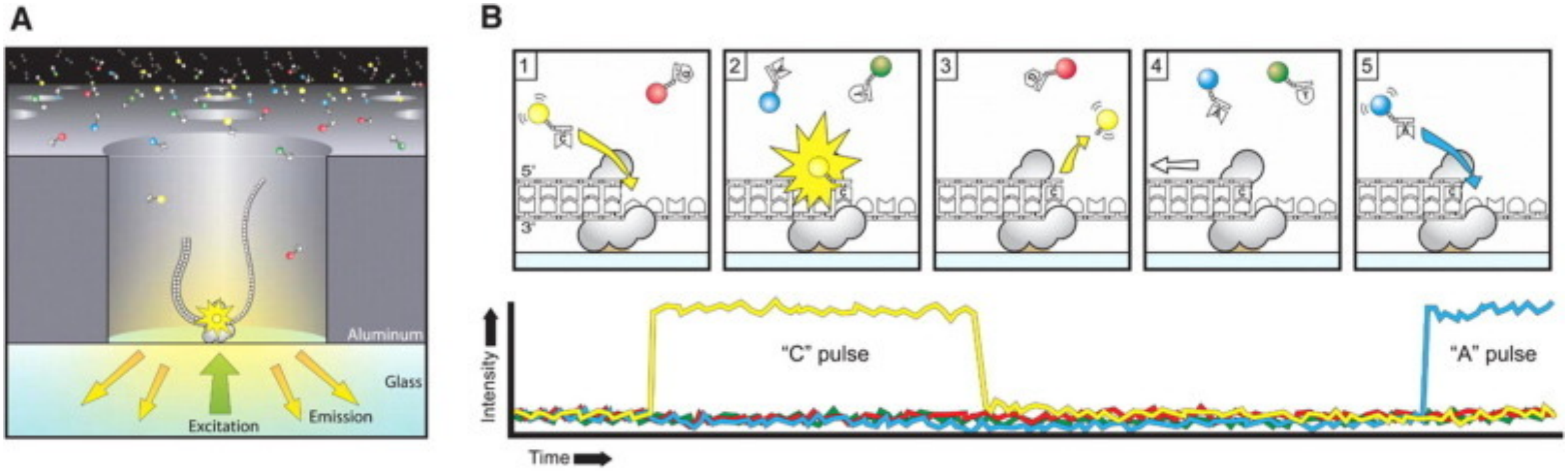
- Q-score need to be stable
- New flowcell chemisty improves Q-score
- Barcode selection → select high quality door



SMRT Sequencing

- Pacific Biosciences
  PacBio Sequel IIe

SMRT® Technology



Template preparation

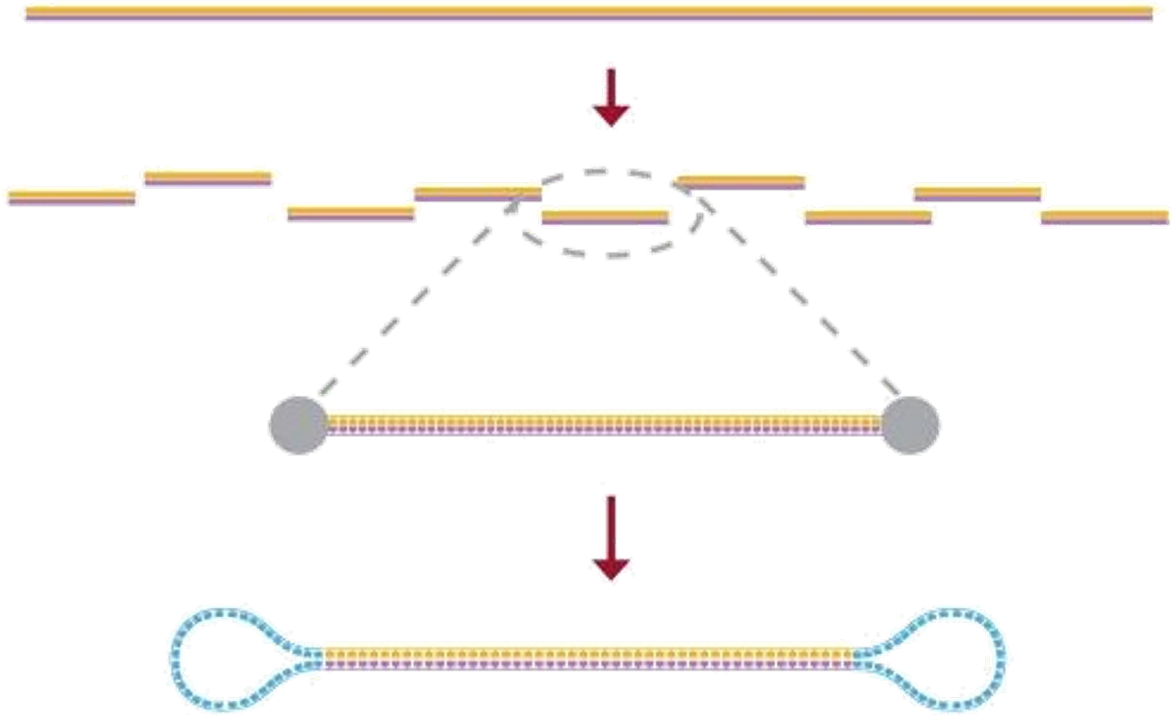
DNA Sample

Fragment DNA

Damage Repair/ End Repair

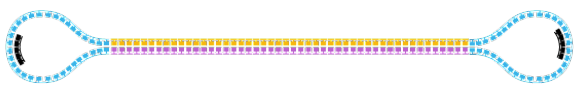
Ligate adapters

Purify DNA



- from 250 bp to 400 kb
- sequences of both forward and reverse strands in the same trace

Universal SMRTbell™ Template

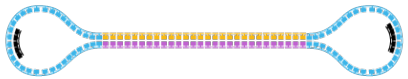


Large Insert Sizes

- Recommended Insert Size: > 3 kb
- Maximum length over 300 kb

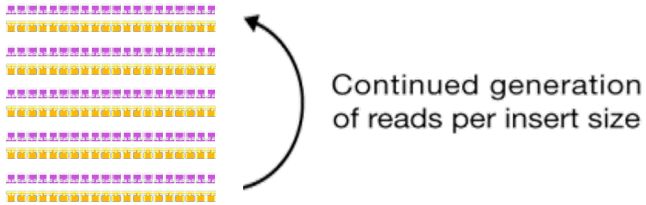
Generates one pass on each molecule sequenced

Circular Consensus Sequencing (CCS)



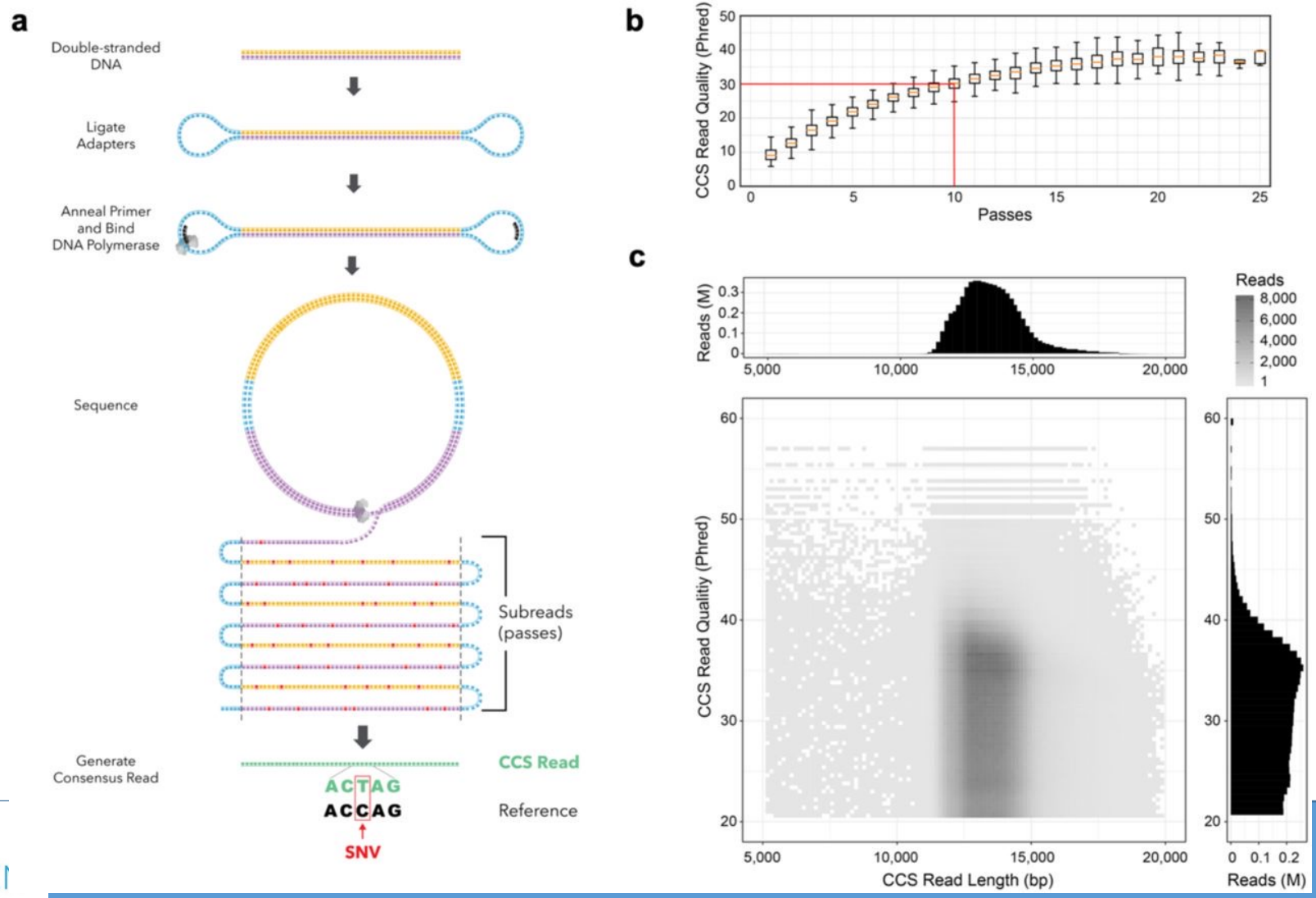
Small Insert Sizes

- Recommended Insert Size: 500 bp-20kb



Generates multiple passes on each molecule sequenced

SMRT® Sequencing Accuracy



Benefits of SMRT® Sequencing

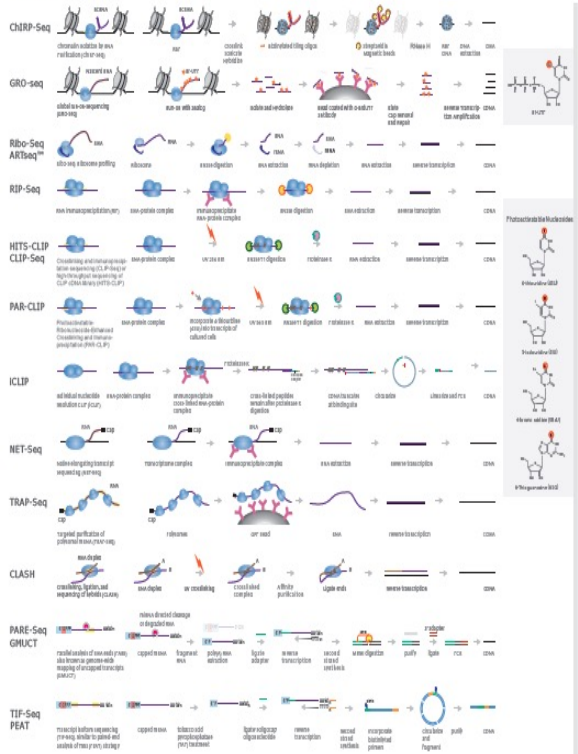
- Produce reads with average lengths of 6000 to 10000, with longest reads over 175,000 base pairs
- Greater than **99.999% (QV 50)** accurate sequencing results
- Sensitivity to detect minor variants at frequency less than 0.1%
- Detect broad spectrum of base modification events in the same sequencing run that reads canonical base sequence
- No amplification bias and least GC bias for improved coverage uniformity


PHASE TWO: INTERPRETATION
SHERMAN Tallahassee Democrat
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

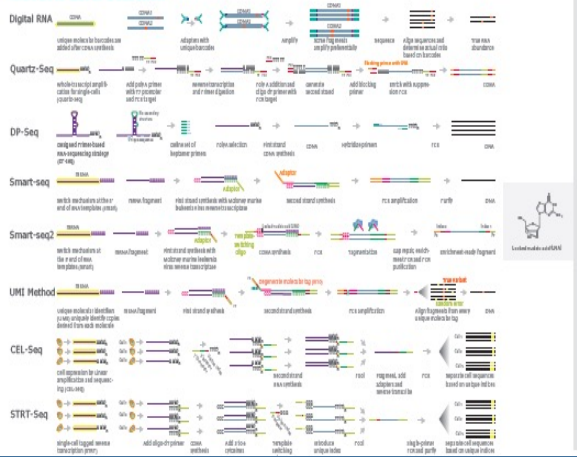
NGS Applications

NGS as a tool for studying genome variation and regulation

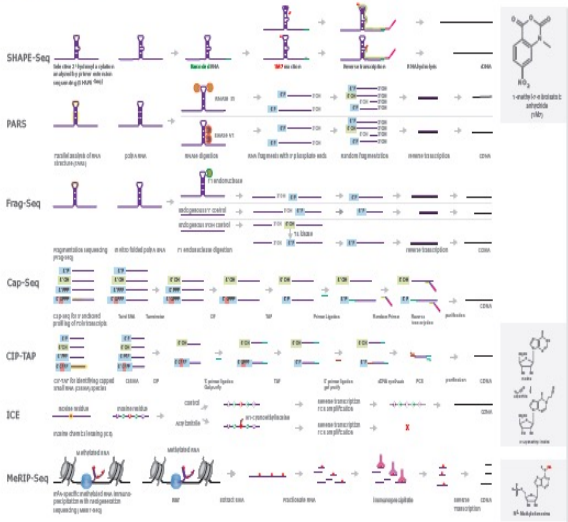
RNA Transcription



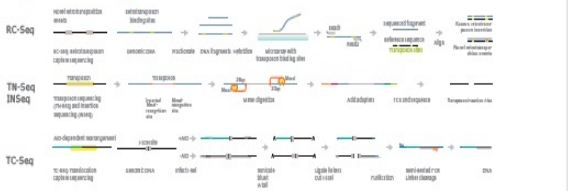
RNA Low-Level Detection



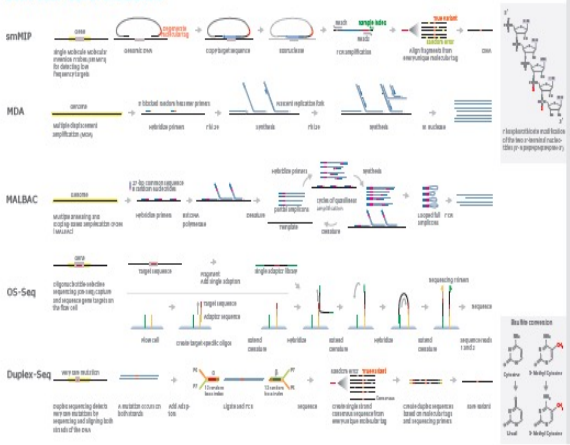
RNA Structure



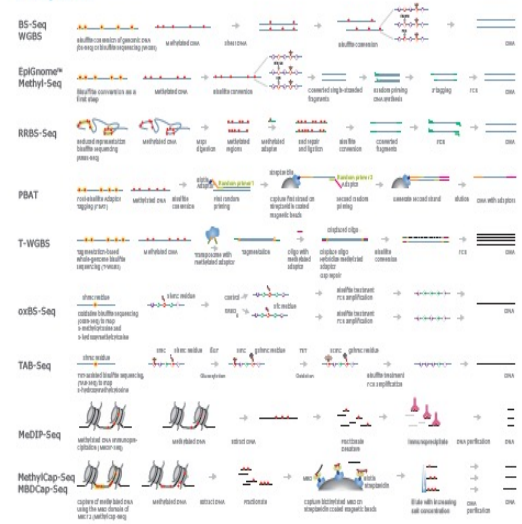
Sequence Rearrangements



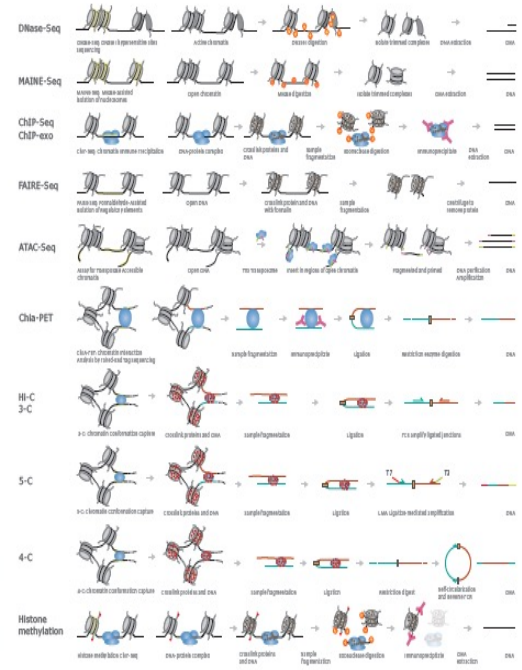
DNA Low-Level Detection



Methylation



DNA-Protein Interactions



DNA

- Targeted resequencing
  - Amplicon
  - MIPs
  - Capture panels
- *de novo* assembly
- Bacterial WGS
- Vertebrate WGS
- Long read sequencing

RNA

- Truseq stranded mRNA
- Lexogen quantseq
- IsoSeq (Pacbio)

Single cell genomics

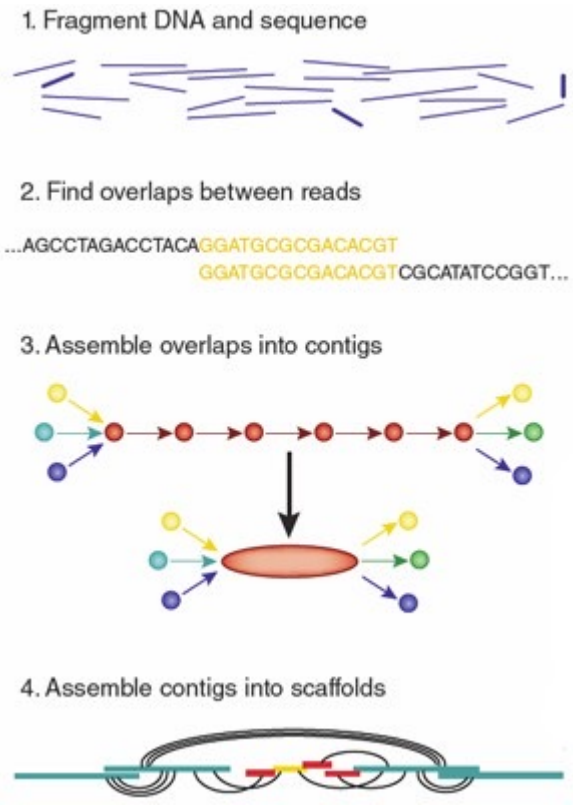
- Various single cell library prep methods for DNA and RNA

Whole genome sequencing

- **Copy number variation analysis**
  - **Sequencing a genome at 0.1-0.3x**
  - **Sequencing a genome at 1-3x**
- **Structural variation analysis**
  - **Sequencing a genome at 5-10x**
- **Whole genome re-sequencing**
  - **Sequencing a genome at >30x**
  - **yeast, fruit fly, bacterial genomes, human…**

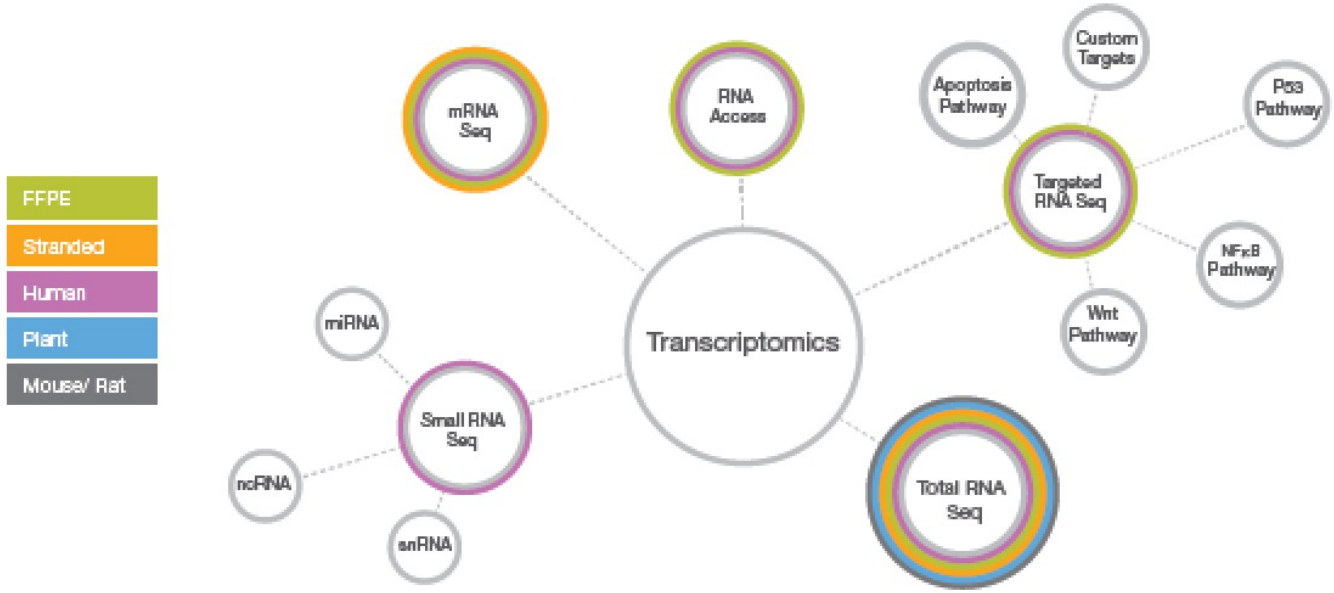
De novo assembly

- Assembling a genome from scratch
- Extremely computationally heavy
- No reference to distinguish variation from artefacts
- Combination of multiple sequencing and optical mapping techniques required



RNA SEQUENCING

- Rapid expression profiling, transcriptome sequencing and small RNA's

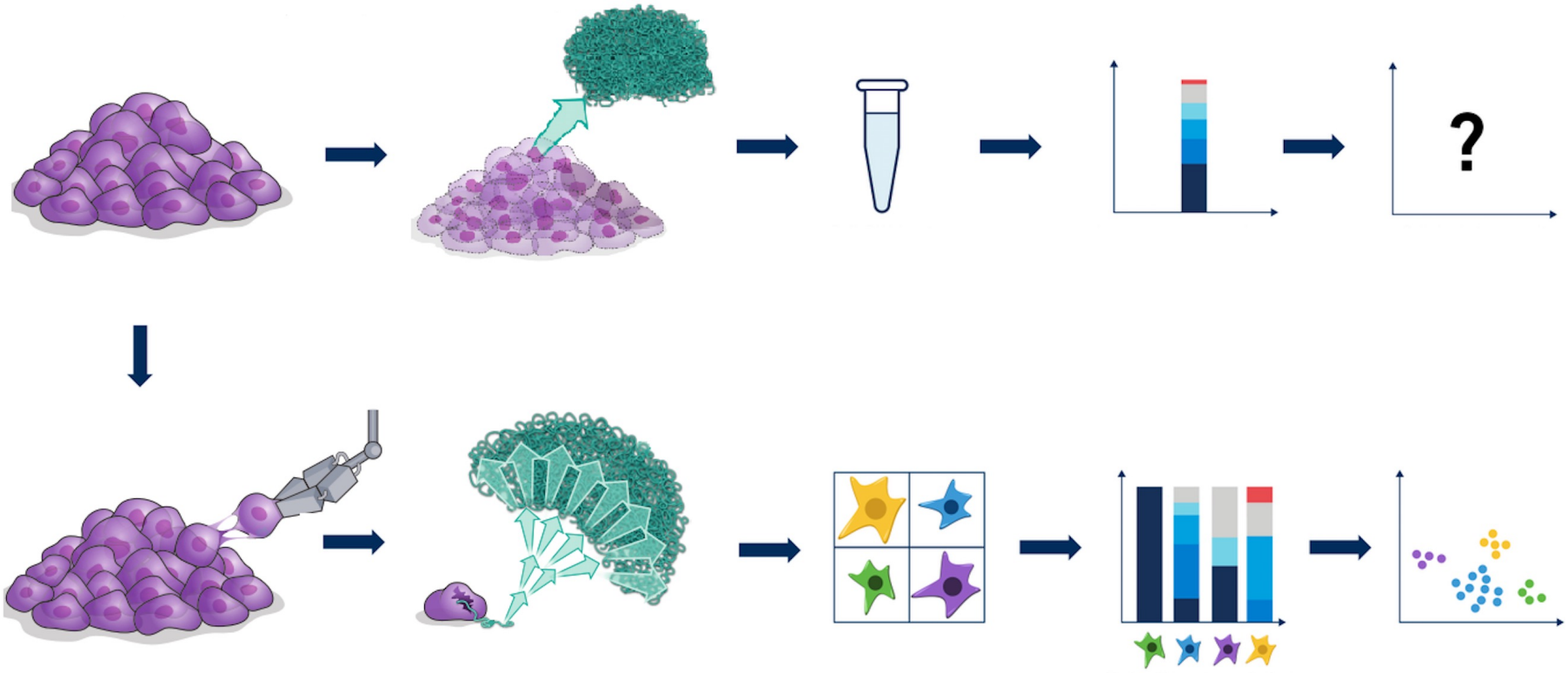


RNAseq library preps

	Differential expression	Whole transcript, fusion, isoforms	Small RNA	Illumina compatible	Low input
Lexogen QuantSeq 3’ mRNA	✓			✓	✓
Lexogen Small RNA seq	✓		✓	✓	✓
Illumina TruSeq stranded mRNA	✓	✓		✓	
Illumina TruSeq stranded total RNA	✓	✓	✓	✓	
IsoSeq	(✓)	✓			
Smart-Seq2	✓	✓		✓	✓✓✓

Single cell RNA-seq as a complementary technique to bulk RNA-seq

Side note



Single cell RNA-seq as a complementary technique to bulk RNA-seq

	RNA type	Transcript targeted	sensitivity	throughput	sequencing	Specific property
SMART SEQ2	mRNA	full transcript	sensitive	low	deep	FACS sorting specific populations
10x genomics 3' RNA seq	mRNA	3' end	medium	High (> 10000)	shallow	Can be combined with surface markers