Lecture 7

Next-generation sequencing technologies

Reading assignments

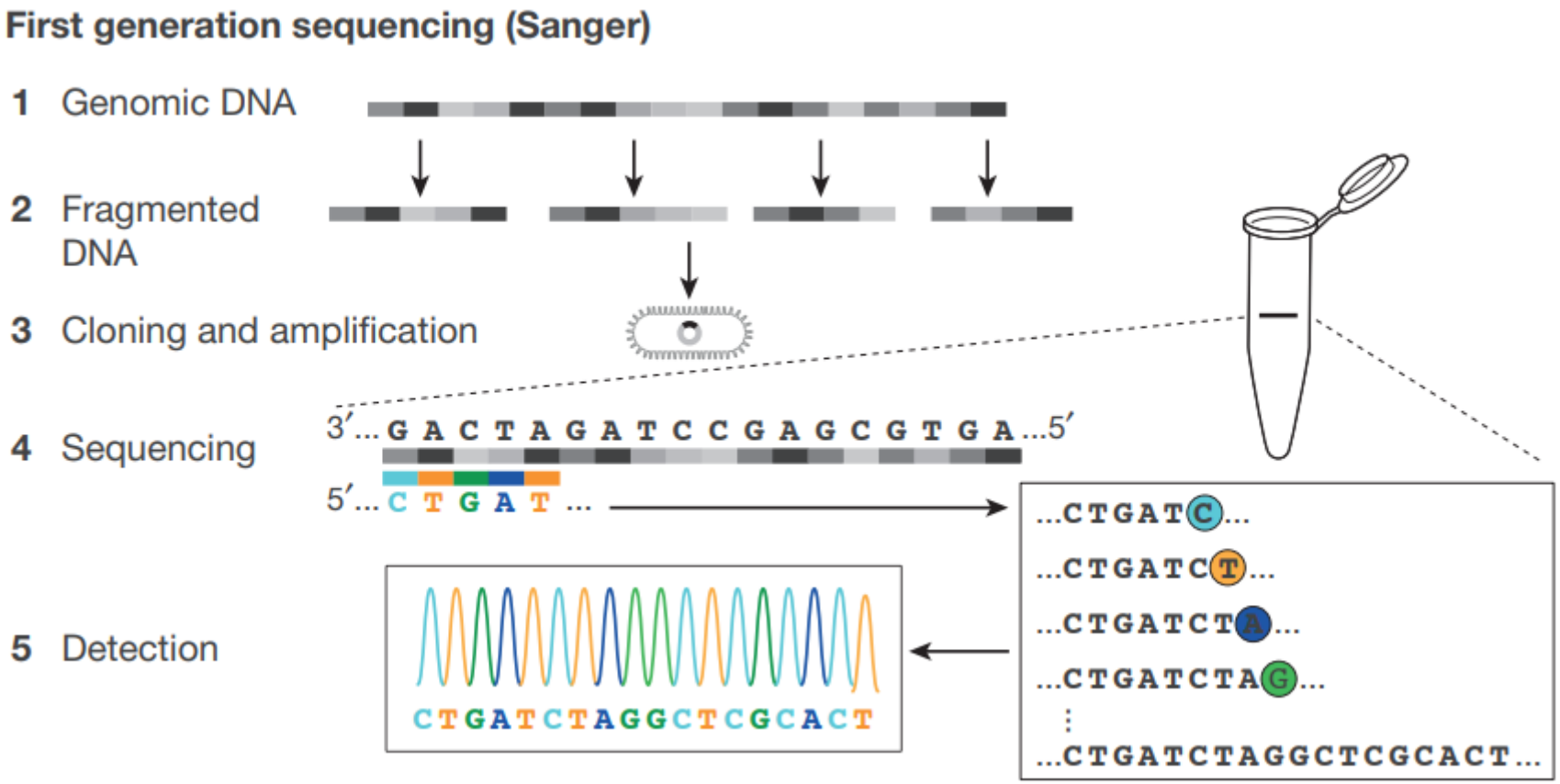
- Coming of age: ten years of next-generation sequencing technologies (https://www.ncbi.nlm.nih.gov/pubmed/27184599)

- DNA sequencing at 40: past, present and future (https://www.ncbi.nlm.nih.gov/pubmed/29019985)

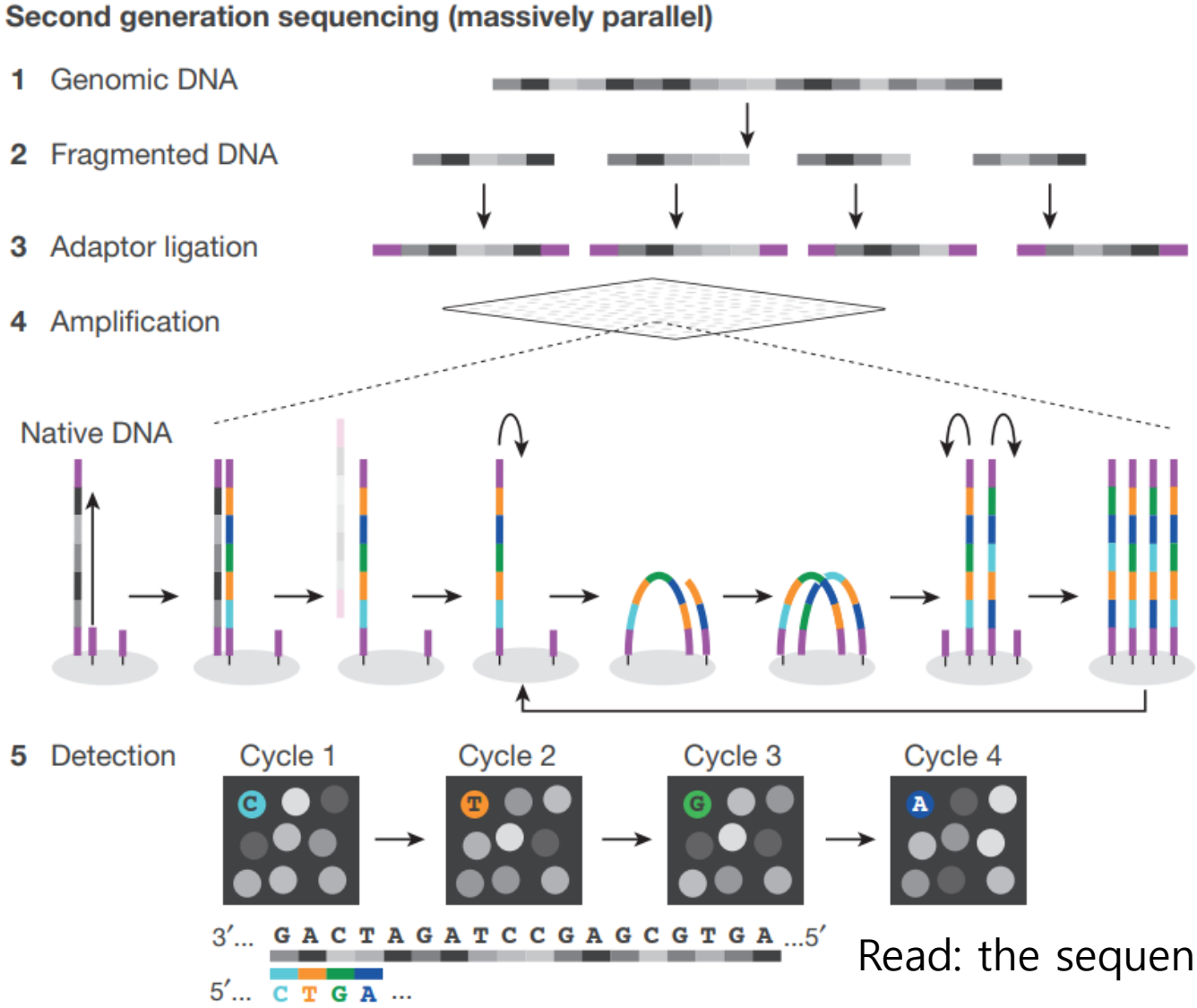
History of sequencing technologies

1953: Sequencing of insulin protein[2]
1965: Sequencing of alanine tRNA[4]
1968: Sequencing of cohesive ends of phage lambda DNA[6]
1977: Maxam–Gilbert sequencing[9]
1977: Sanger sequencing[8]
1981: Messing's M13 phage vector[12]
1986–1987: Fluorescent detection in electrophoretic sequencing[14,15,17]
1987: Sequenase[18]
1988: Early example of sequencing by stepwise dNTP incorporation[139]
1990: Paired-end sequencing[23]
1992: Bodipy dyes[140]
1993: *In vitro* RNA colonies[37]
1996: Pyrosequencing[44]
1999: *In vitro* DNA colonies in gels[38]
2000: Massively parallel signature sequencing by ligation[47]
2003: Emulsion PCR to generate *in vitro* DNA colonies on beads[42]
2003: Single-molecule massively parallel sequencing-by-synthesis[33,34]
2003: Zero-mode waveguides for single-molecule analysis[57]
2003: Sequencing by synthesis of *in vitro* DNA colonies in gels[49]
2005: Four-colour reversible terminators[51–53]
2005: Sequencing by ligation of *in vitro* DNA colonies on beads[41]
2007: Large-scale targeted sequence capture[93–96]
2010: Direct detection of DNA methylation during single-molecule sequencing[65]
2010: Single-base resolution electron tunnelling through a solid-state detector[141]
2011: Semiconductor sequencing by proton detection[142]
2012: Reduction to practice of nanopore sequencing[143,144]
2012: Single-stranded library preparation method for ancient DNA[145]

First generation sequencing



Second (Next) generation sequencing

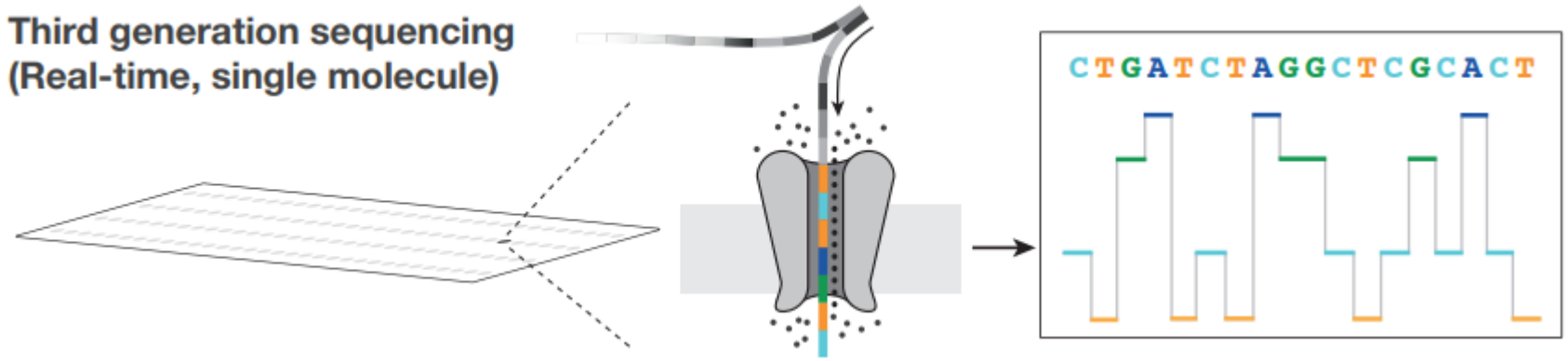


Template: A DNA fragment to be sequenced.

Cluster: A group of identical copies of a DNA template in close proximity.

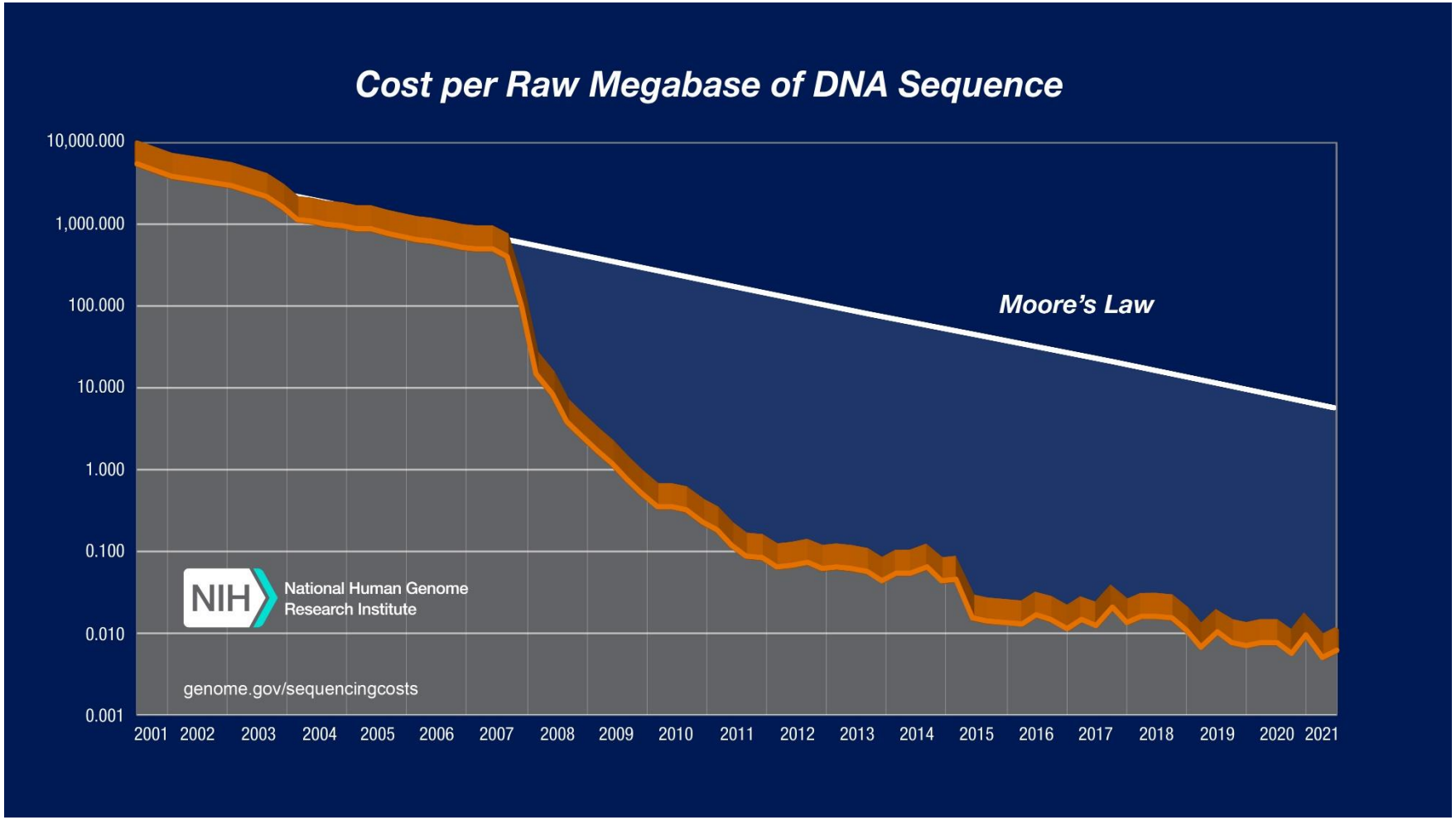
Read: the sequence of bases from a single molecule of DNA

Third generation sequencing



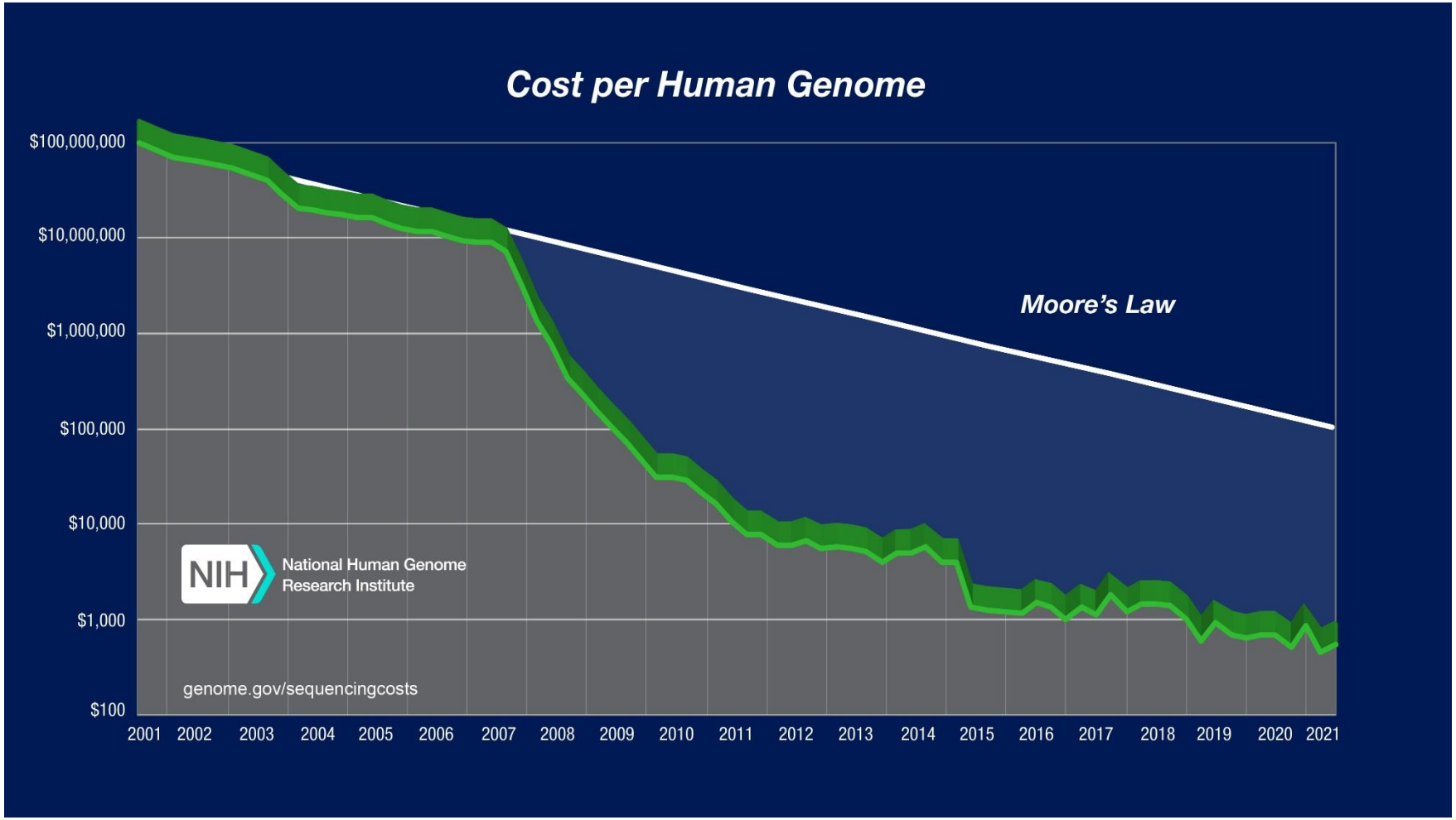
Read lengths of PacBio and ONT sequencing: hundreds of kilobases

Sequencing cost per Mbp



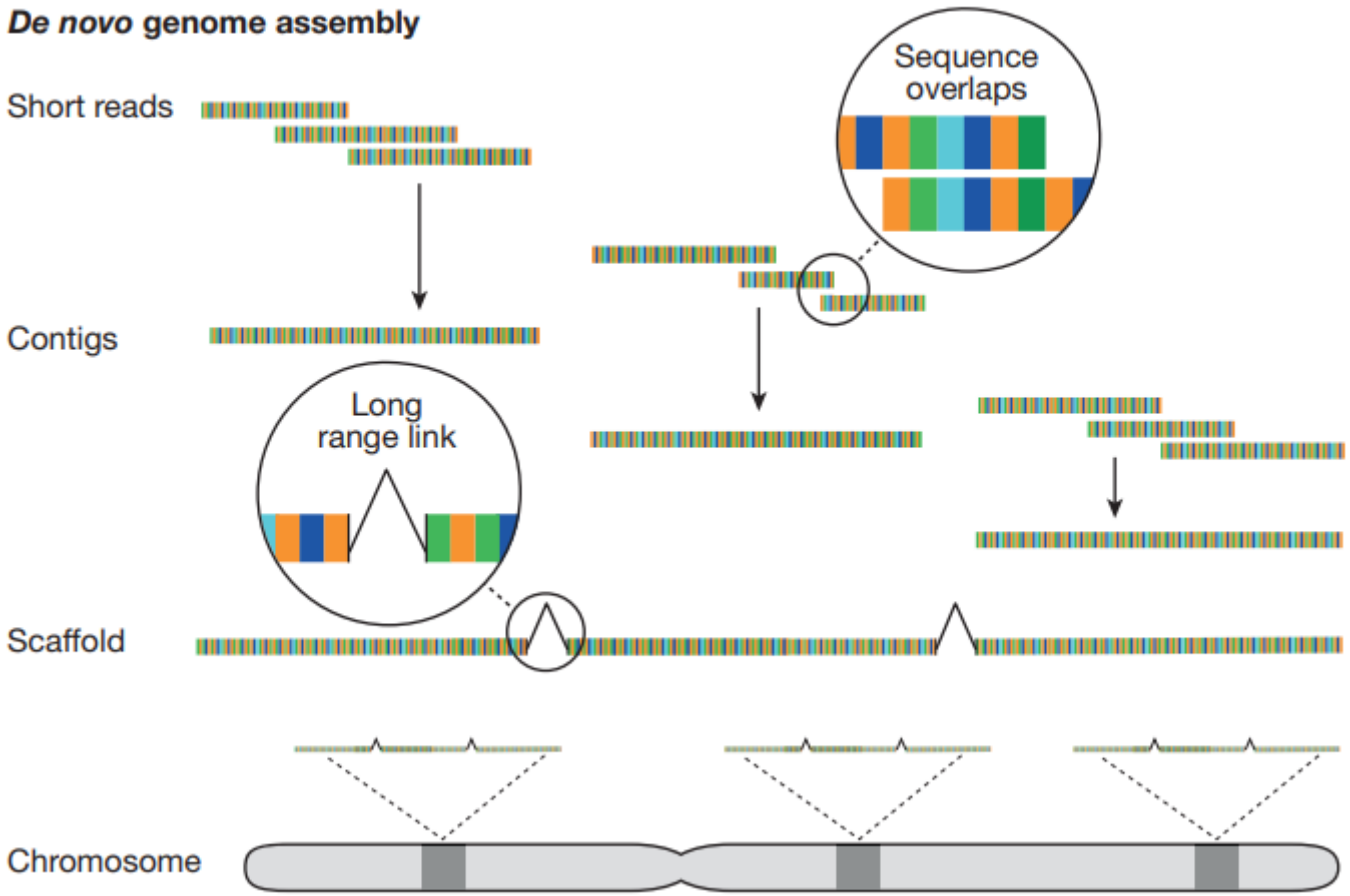
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Sequencing cost per human genome



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

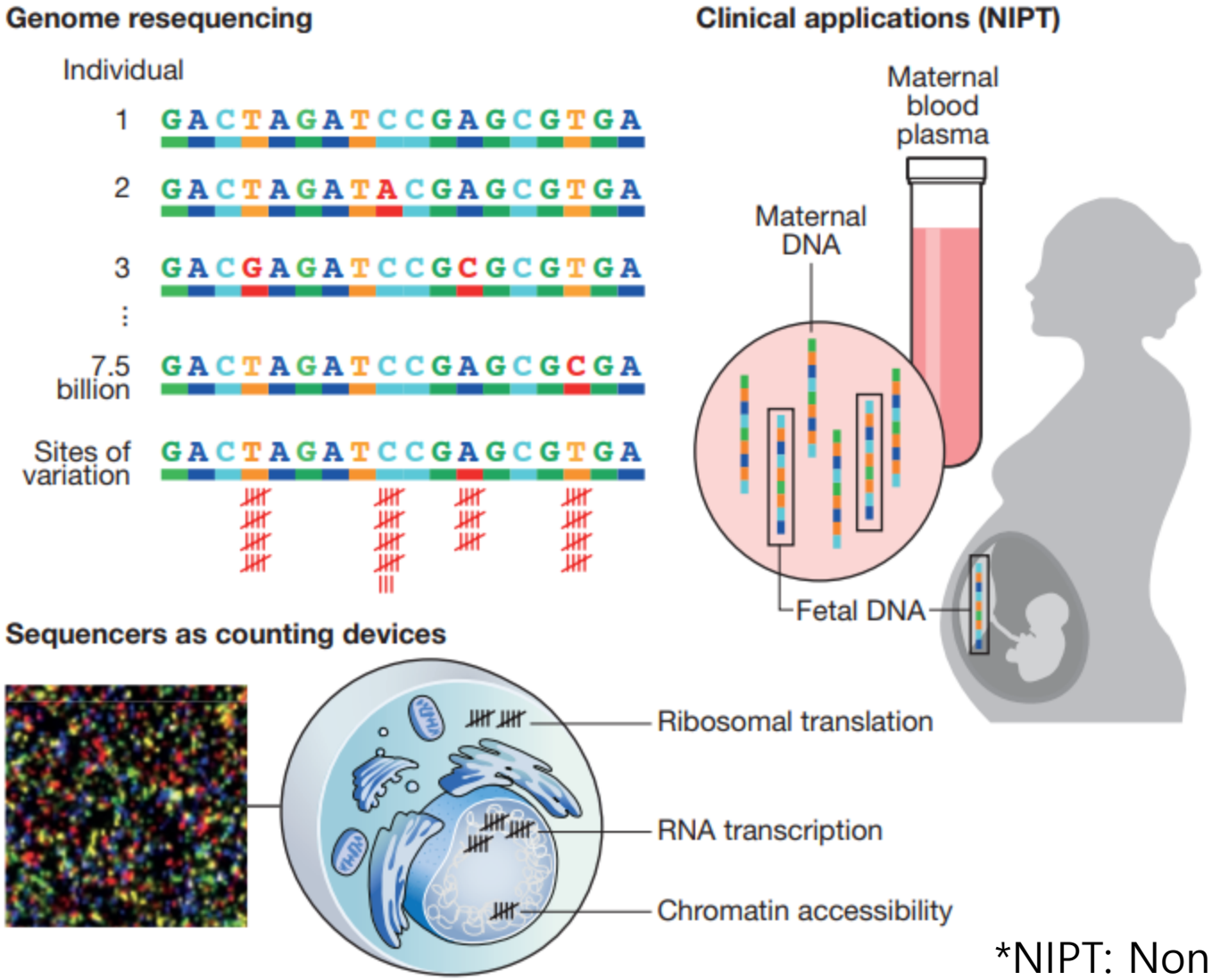
NGS applications



Genome milestones

1977: Bacteriophage ΦX174 (ref. 72)
1982: Bacteriophage lambda[13]
1995: *Haemophilus influenzae*[26]
1996: *Saccharomyces cerevisiae*[27]
1998: *Caenorhabditis elegans*[28]
2000: *Drosophila melanogaster*[32]
2000: *Arabidopsis thaliana*[146]
2001: *Homo sapiens*[29–31]
2002: *Mus musculus*[147]
2004: *Rattus norvegicus*[148]
2005: *Pan troglodytes*[149]
2005: *Oryza sativa*[150]
2007: *Cyanidioschyzon merolae*[126]
2009: *Zea mays*[151]
2010: Neanderthal[88]
2012: Denisovan[145]
2013: The HeLa cell line[152,153]
2013: *Danio rerio*[154]
2017: *Xenopus laevis*[155]

NGS applications



*NIPT: Non-invasive prenatal testing

Application milestones

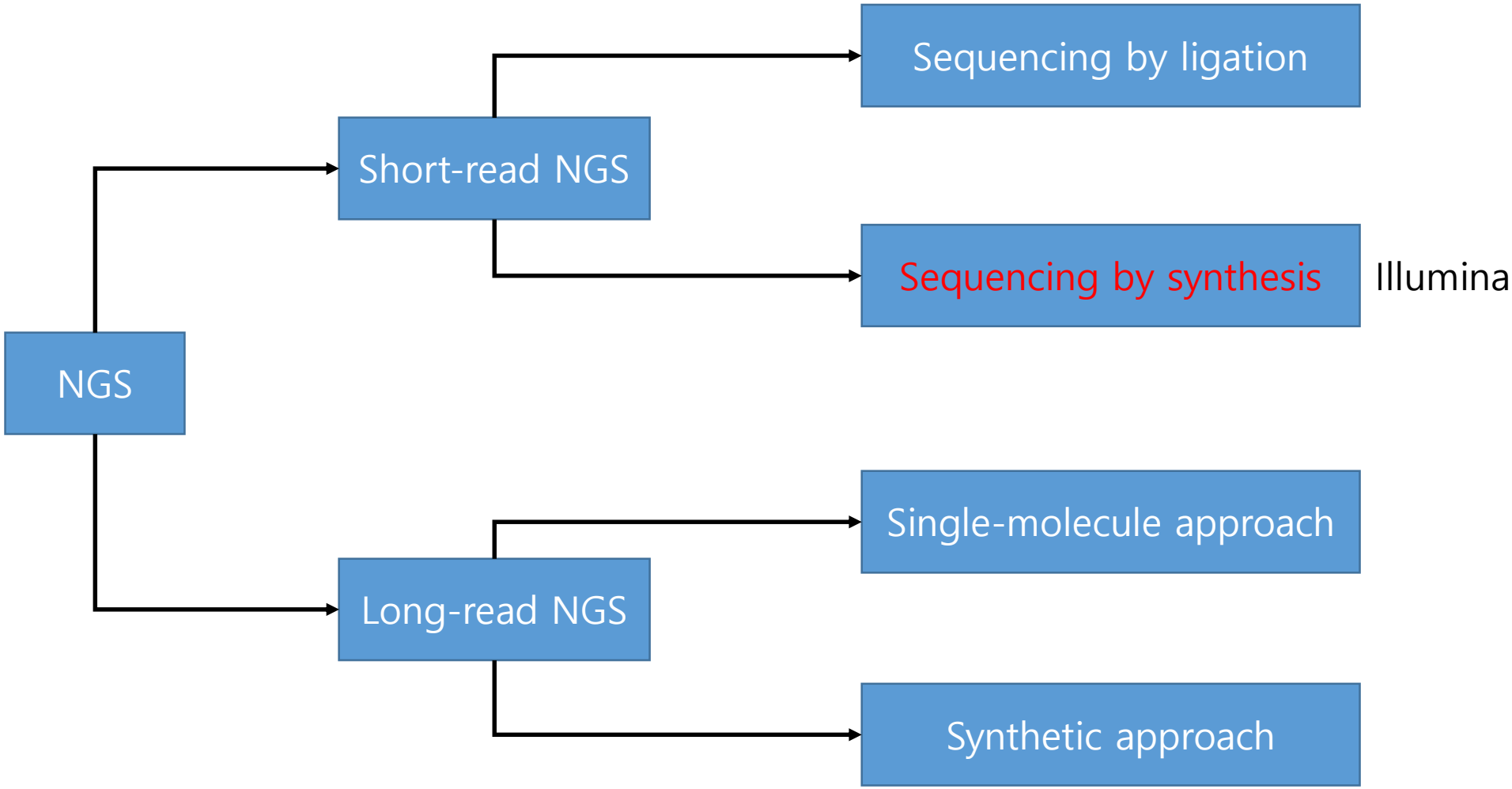
1977: Genome sequencing[72]
1982: Shotgun sequencing[13]
1983, 1991: Expressed sequence tags[107,108]
1995: Serial analysis of gene expression[109]
1998: Large-scale human SNP discovery[168]
2004: Metagenome assembly[122]
2005: Bacterial genome resequencing with NGS[40,41]
2007: Chromatin immunoprecipitation followed by sequencing (ChIP–seq) using NGS[117]
2007–2008: Human genome and cancer genome resequencing using NGS[55,90–92]
2008: RNA-seq using NGS[110–114]
2008: Chromatin accessibility using NGS[118]
2009: Exome resequencing using NGS[97]
2009: Ribosome profiling using NGS[119]
2010: Completion of Phase I of the 1000 Genomes Project[98]
2010: *De novo* assembly of a large genome from short reads[169]
2011: Haplotype-resolved human genome resequencing using NGS[170,171]
2016: Human genome *de novo* assembly with PacBio[172]
2017: Human genome *de novo* assembly with nanopore[64]

Computational milestones

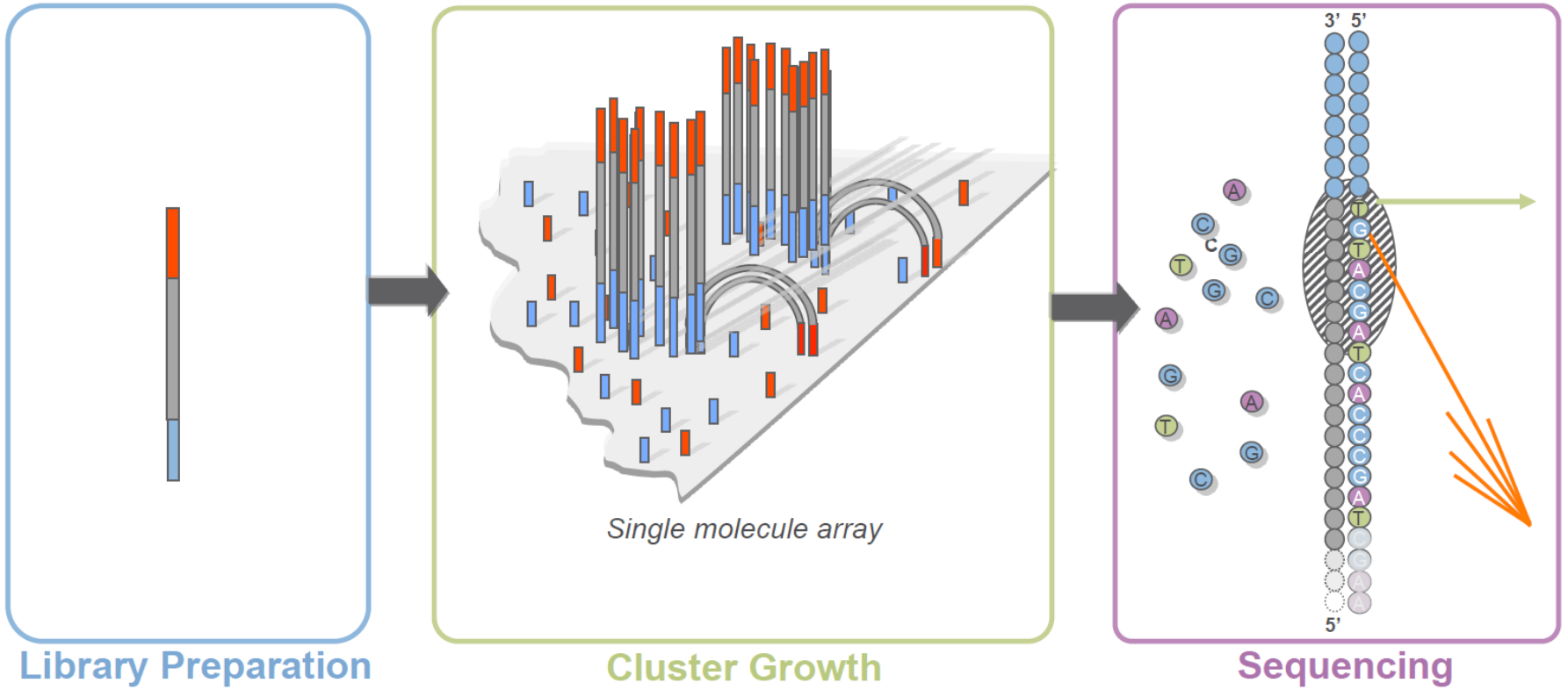
* 1981: Smith–Waterman[156]
* 1982: GenBank (https://www.ncbi.nlm.nih.gov/genbank/statistics/)
1990: BLAST[16]
1995: TIGR assembler[24]
1996: RepeatMasker
1997: GENSCAN[157]
* 1998: phred, phrap, consed[22]
2000: Celera assembler[25]
* 2001: Bioconductor
2001: EULER[74]
2002: BLAT[158]
* 2002: UCSC Genome Browser[159]
* 2002: Ensembl[160]

2005: Galaxy[161]
* 2007: NCBI Short Read Archive
2008: ALLPATHS[162]
2008: Velvet[75]
2009: Bowtie[83]
* 2009: BWA[82]
* 2009: SAMtools[84]
2009: BreakDancer[163]
2009: Pindel[164]
2009: TopHat[115]
2010: SOAPdenovo[165]
2010: GATK[85]
2010: Cufflinks[116]
* 2011: Integrated Genomics Viewer[166]
2013: HGAP/Quiver[167]
2017: Canu[81]

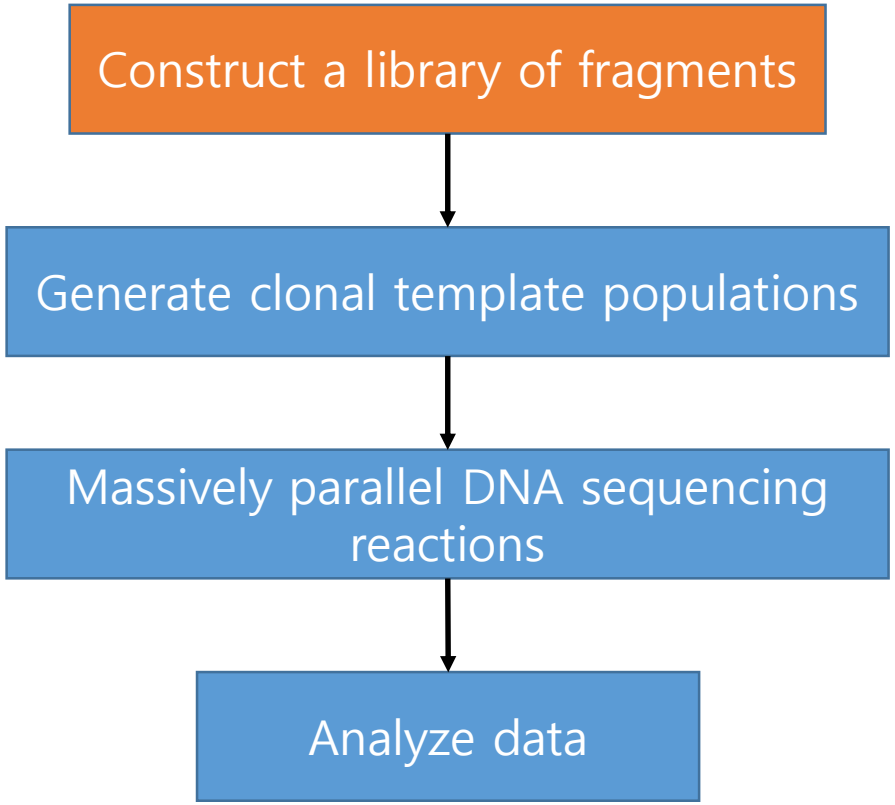
Overview


NGS
Short-read NGS
Long-read NGS
Sequencing by ligation
Sequencing by synthesis
Single-molecule approach
Synthetic approach
Illumina

Illumina sequencing workflow



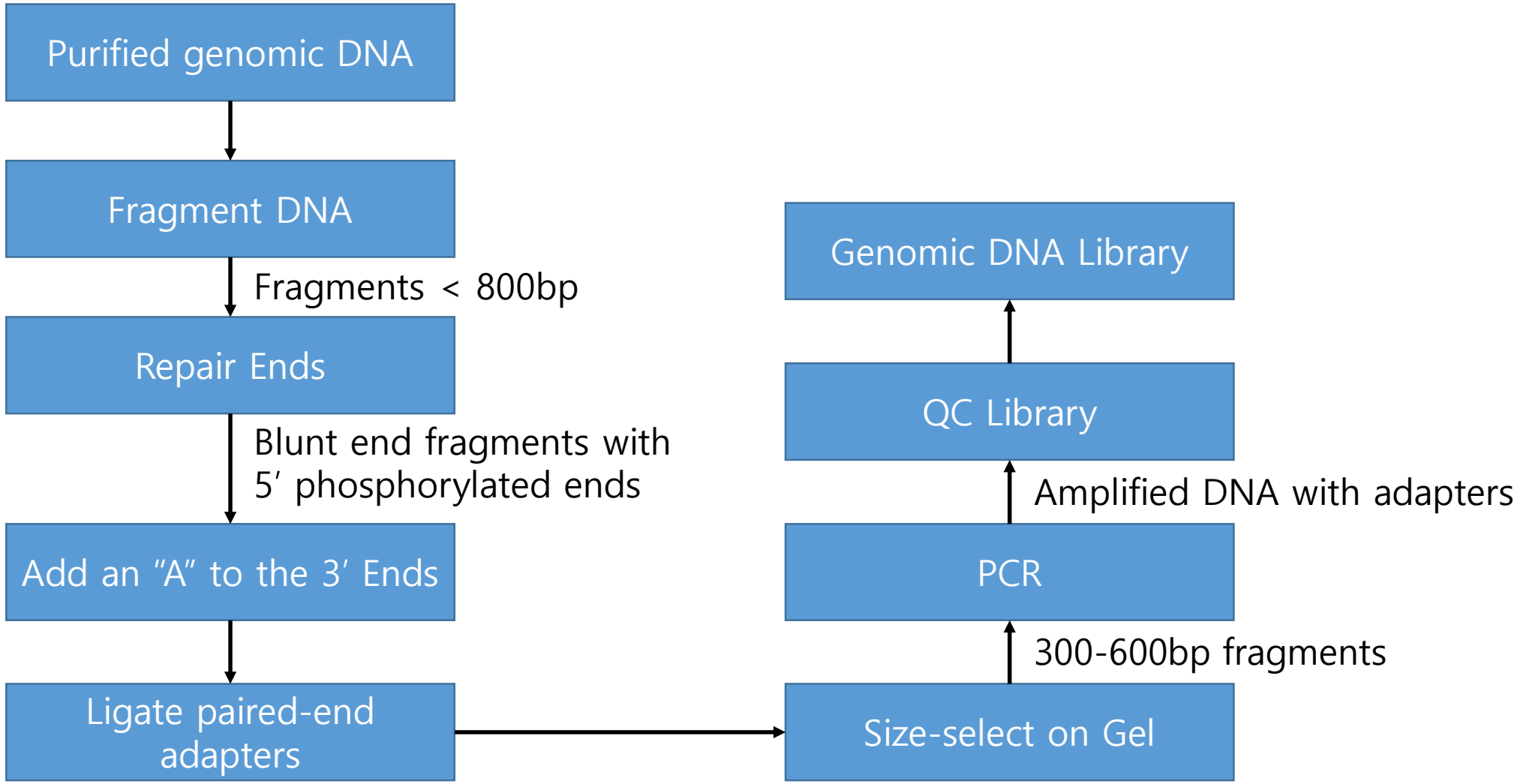
General principles of short-read NGS



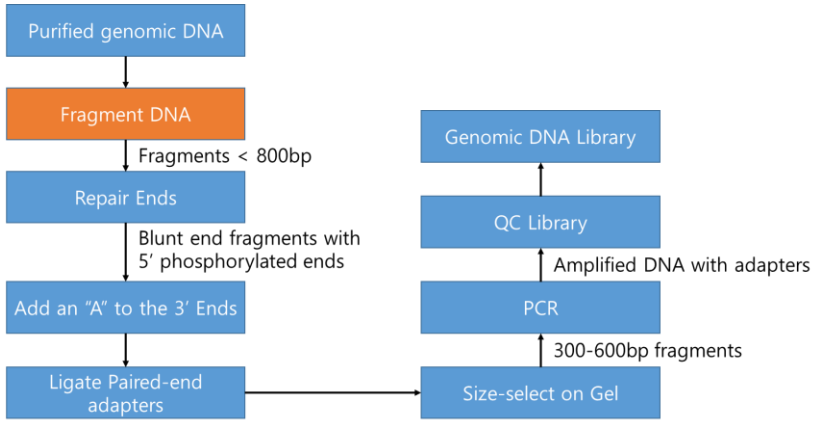
Library preparation

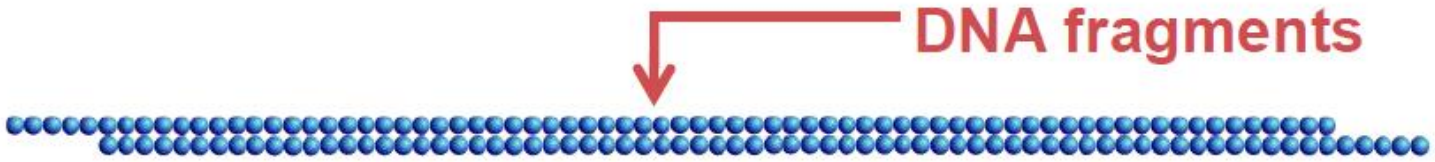
- Prepares sample nucleic acids for sequencing
  - Fragmentation: the process of breaking large DNA fragments into smaller fragments, achieved mechanically, by sonication or enzymatically.
  - Generates double-stranded DNA flanked by Illumina adapters
  - Generates the same general template structure, but variables include
    - Insert size
    - Adapter type
    - Index for multiplexing

Library preparation: Overview



Library preparation: Fragmentation





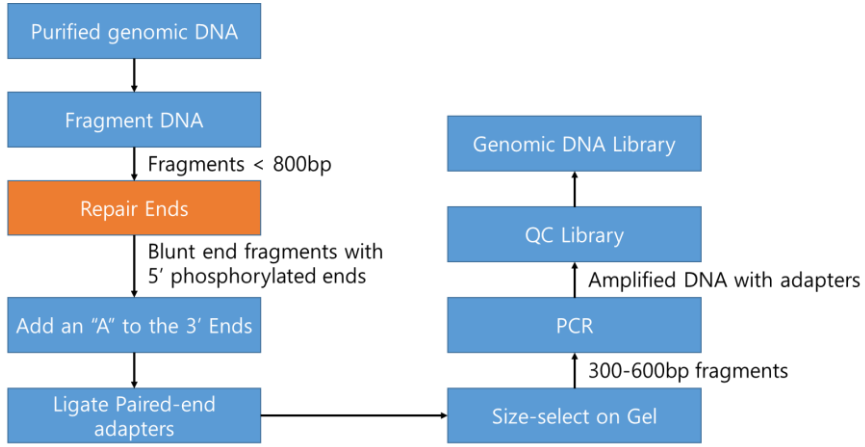
Library preparation: Fragmentation

The size of the target DNA fragments in the final library is a key parameter for NGS library construction.

Optimal library size is impacted by

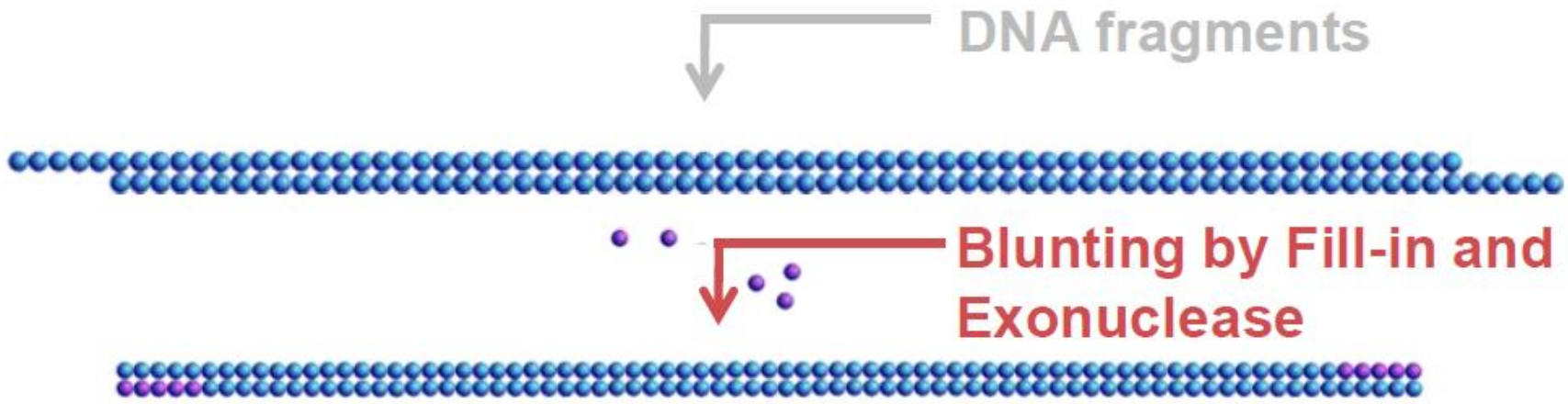
1. the process of cluster generation: Short products amplify more efficiently than longer products. Longer library inserts generate larger, more diffuse clusters than short inserts.
2. the sequencing application: For example, 2×100 PE for exome sequencing since more than 80% of human exomes are under 200bp.

Library preparation: Repair Ends

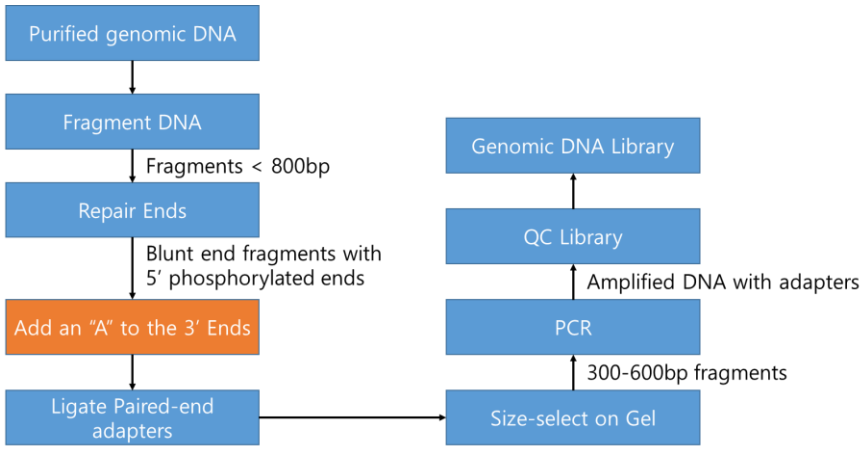


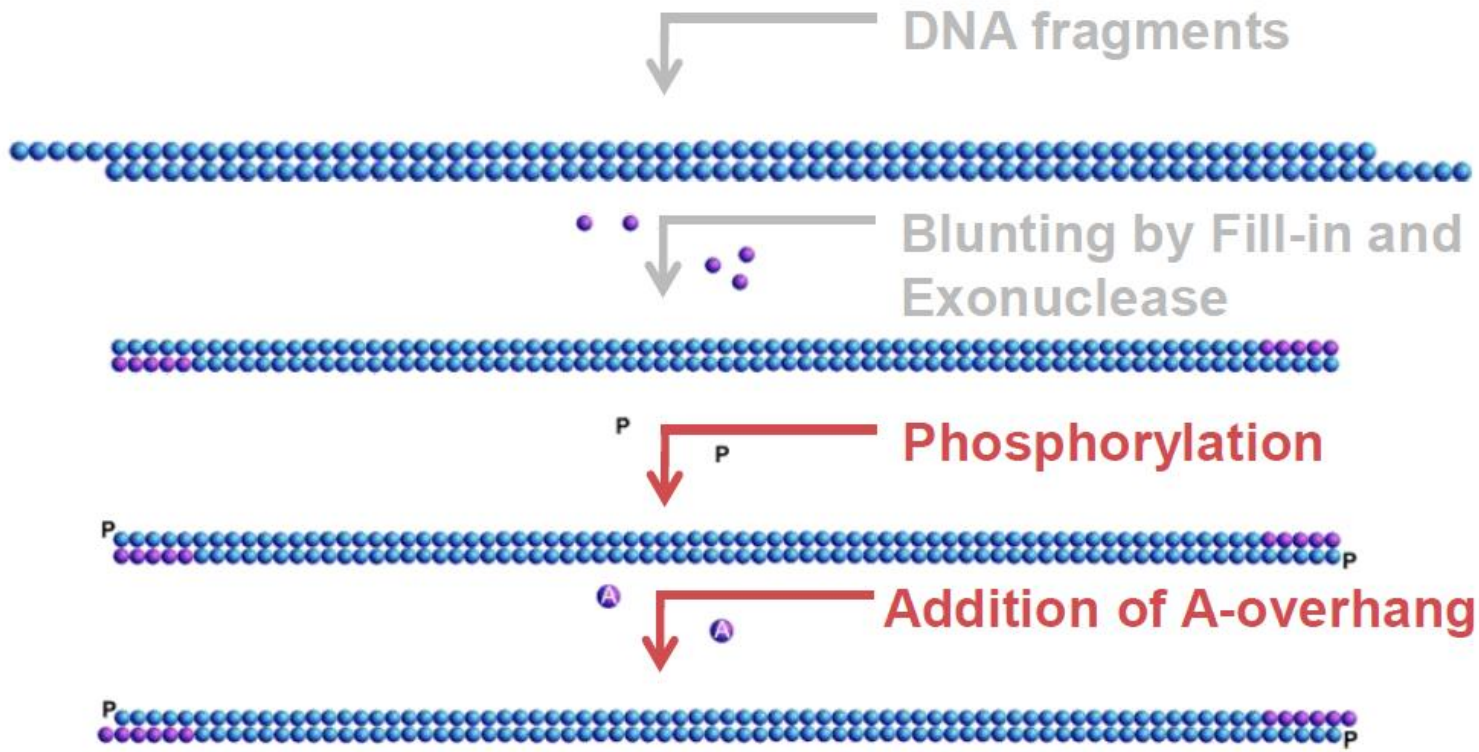
Sticky ends

Blunt ends



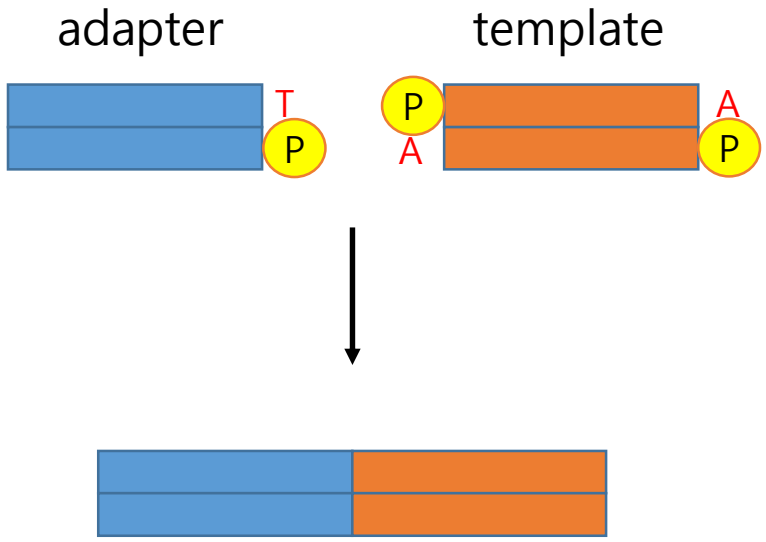
Library preparation: A-tailing



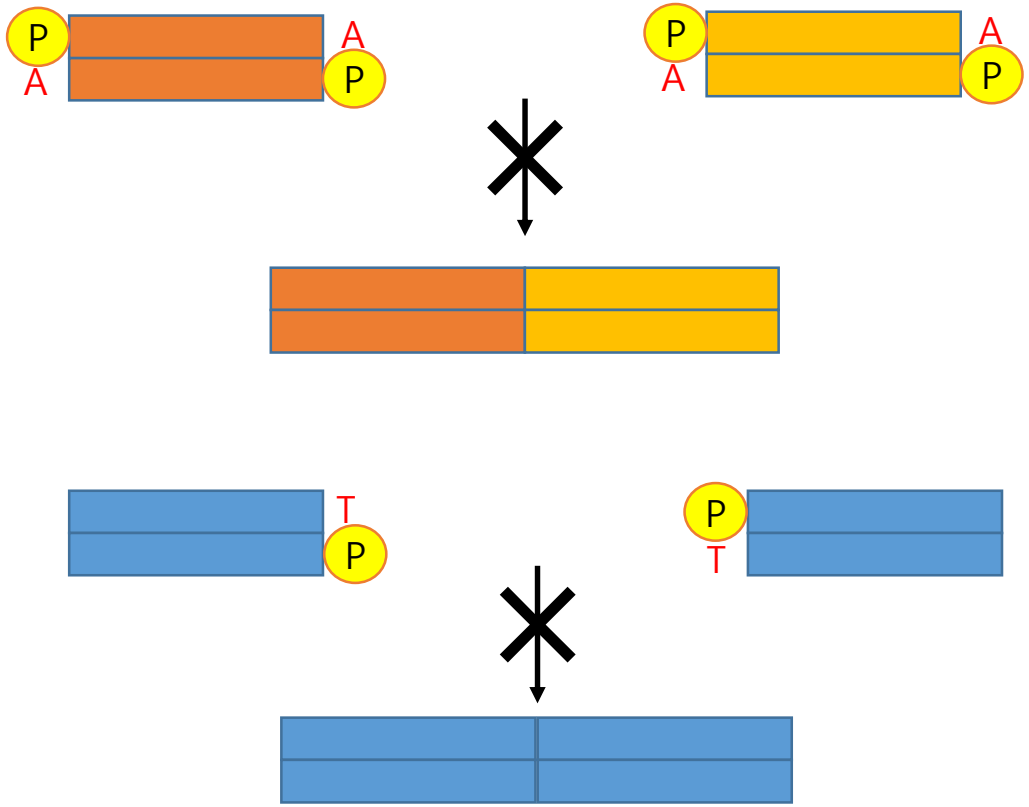


Library preparation: A-tailing

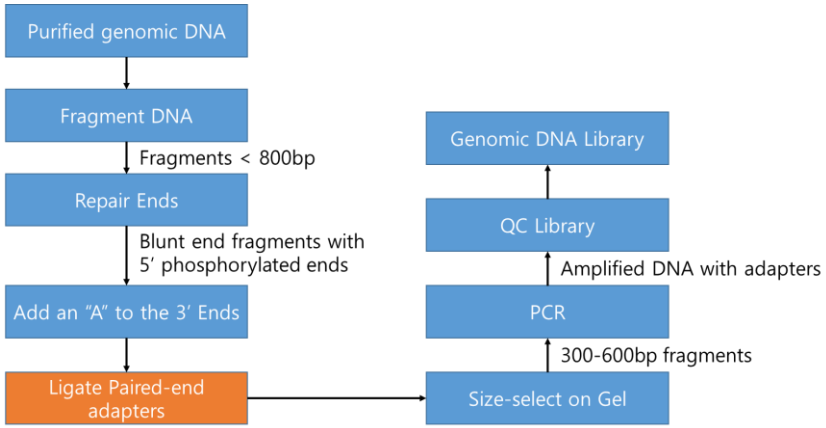
To facilitate ligation to sequencing adapter

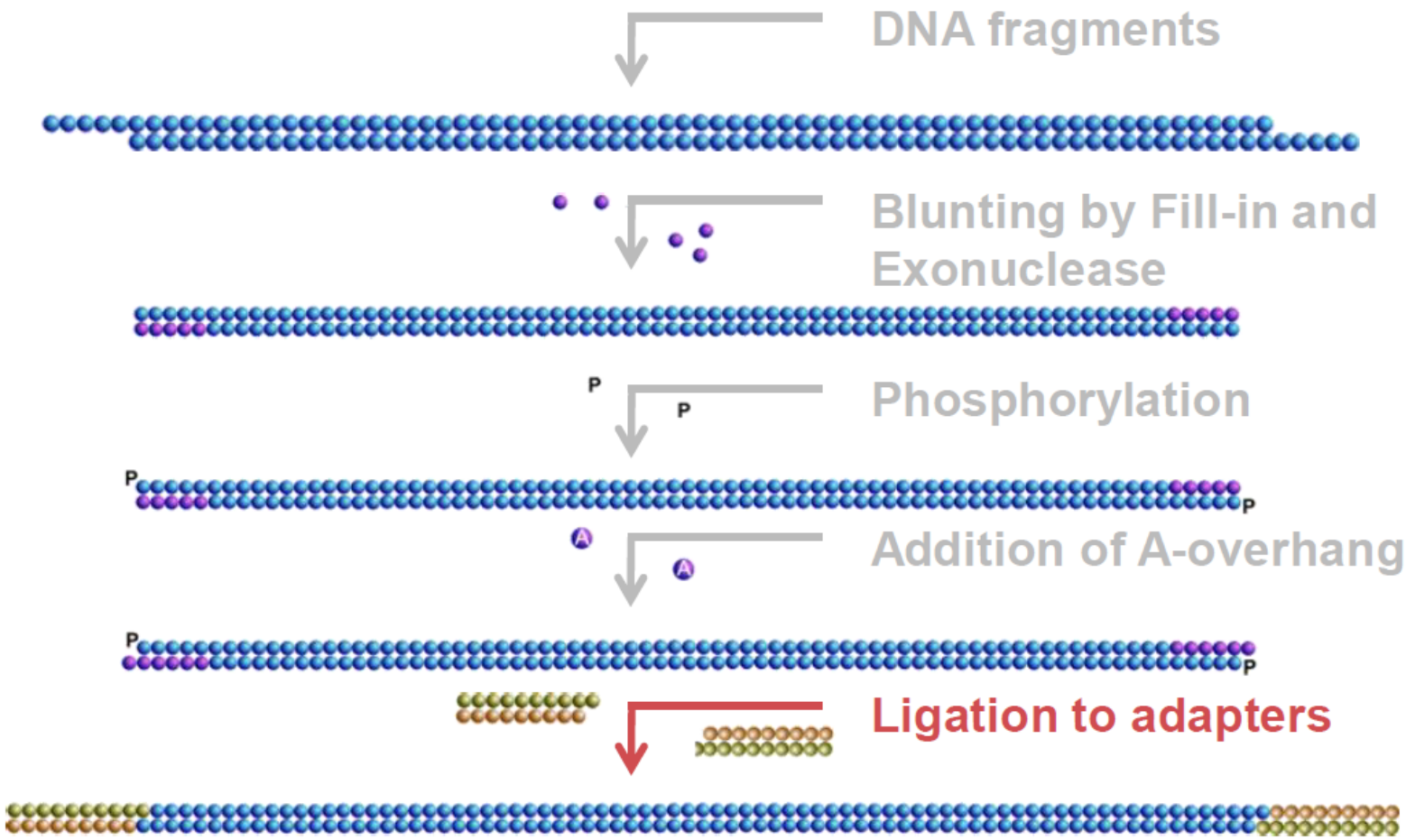


To prevent self-ligation between blunt ended template molecules (concatermers), or between adapters (adapter dimers)

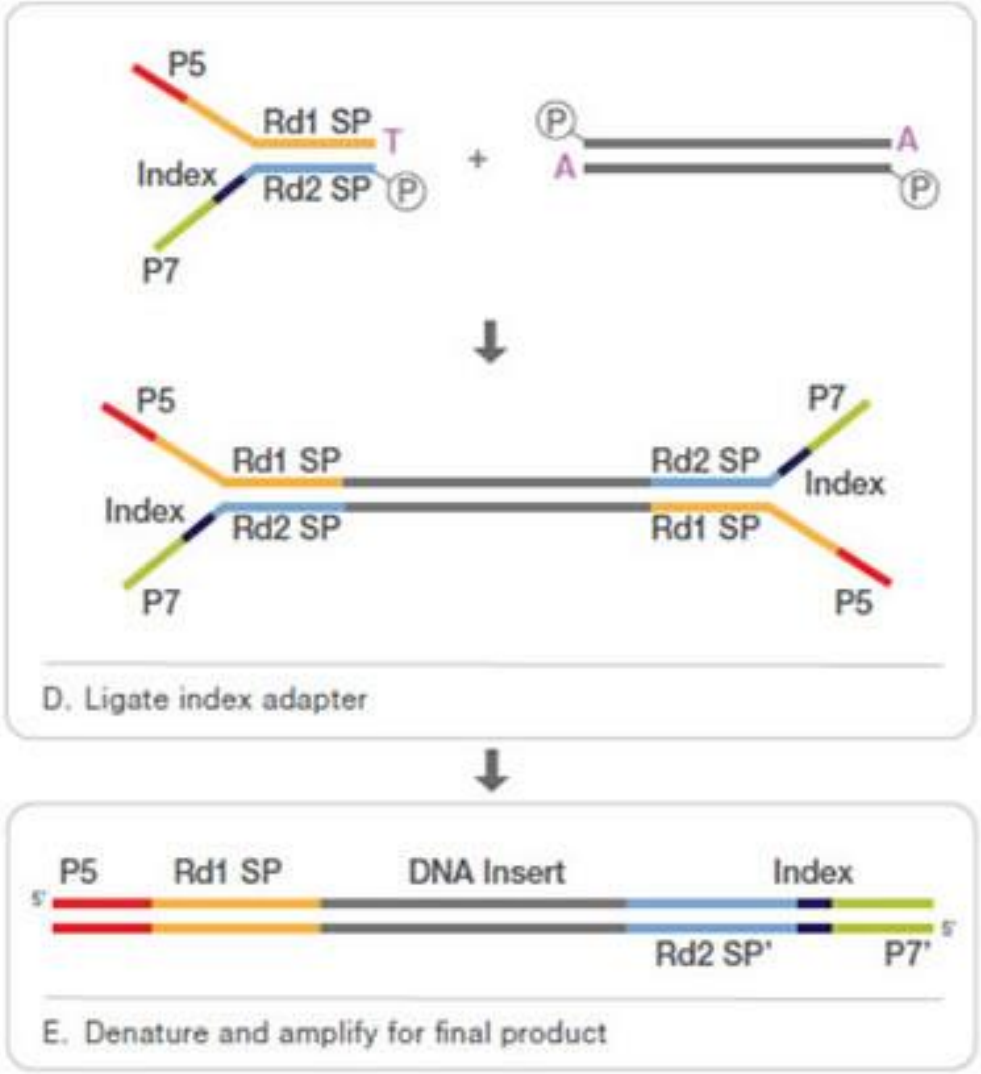


Library preparation: Adapter ligation



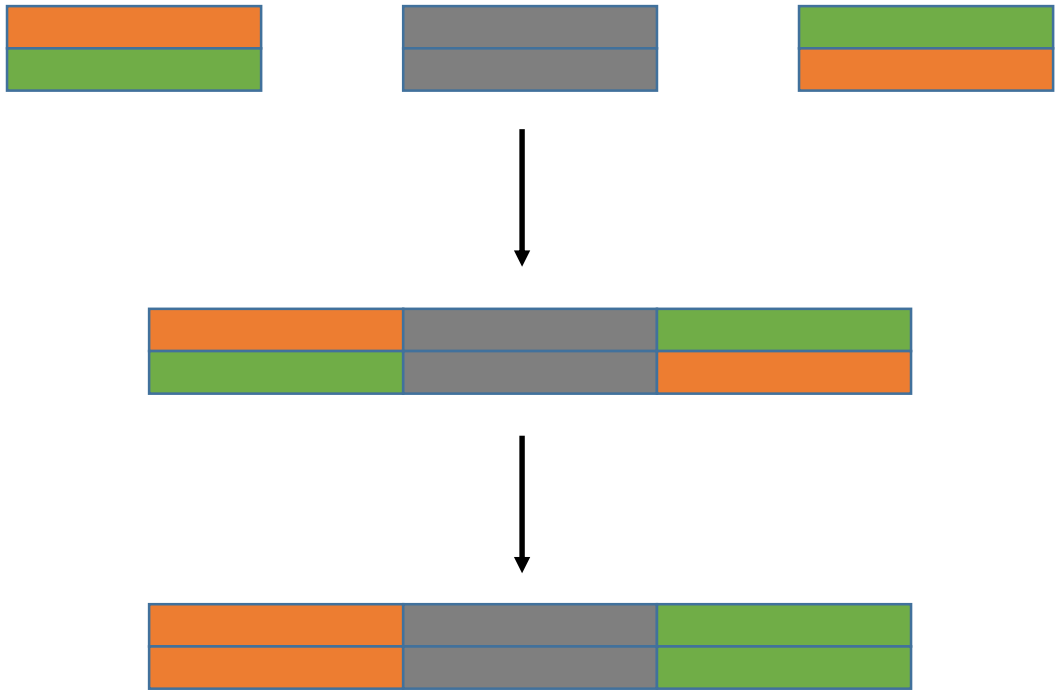


Library preparation: Y-shaped adaptors

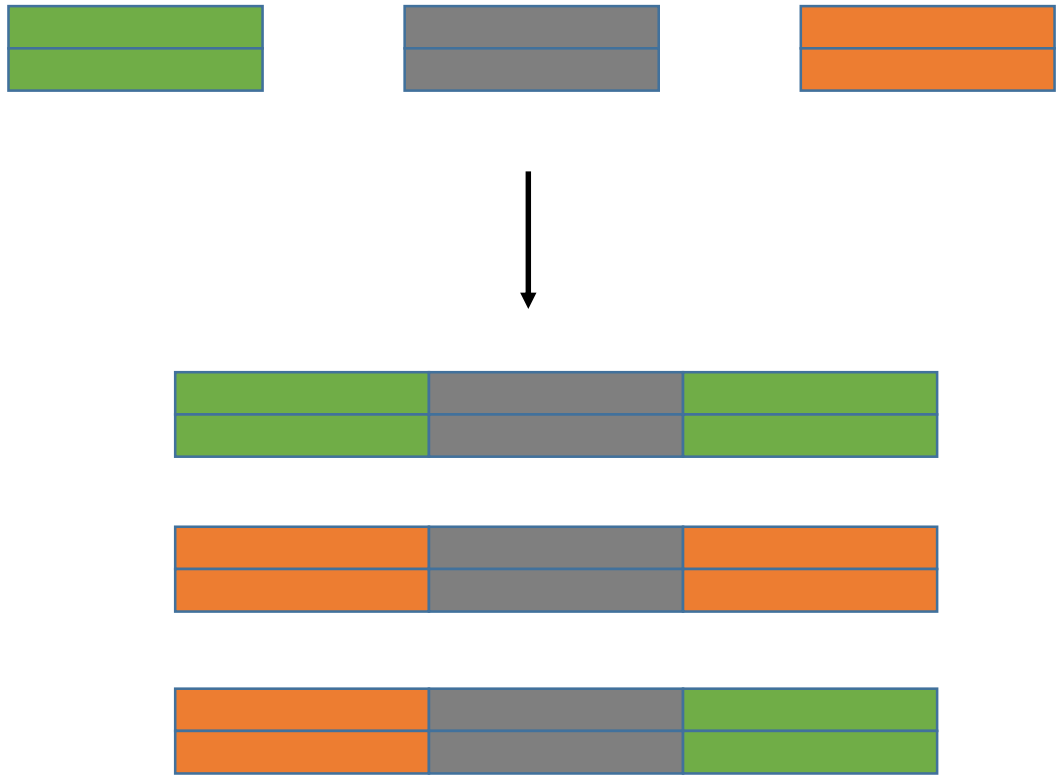


Library preparation: Y-shaped adapters

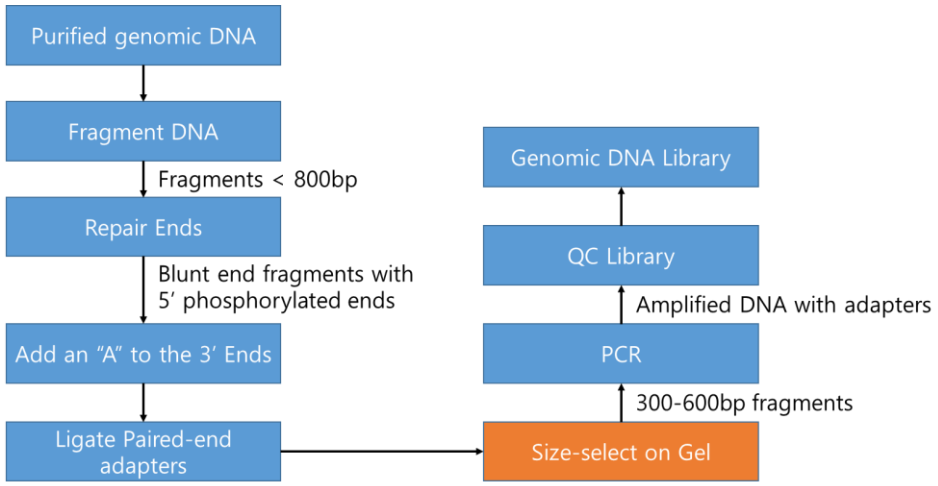
Y-shaped adapters

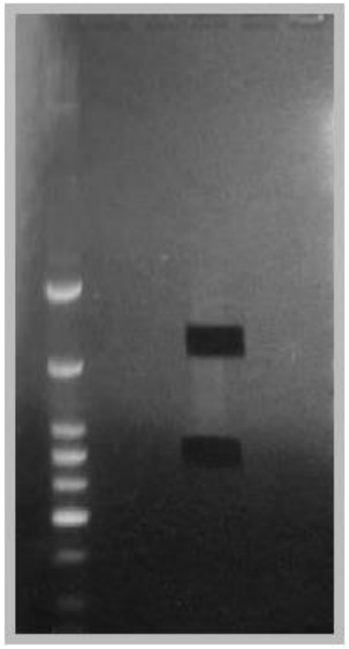


Non Y-shaped adapters



Library preparation: Size-select on Gel

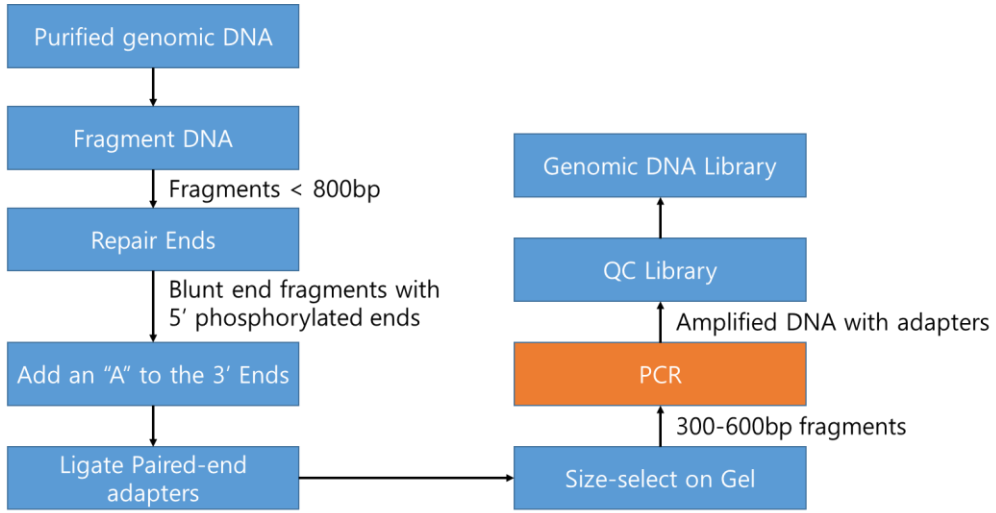




600bp area excised

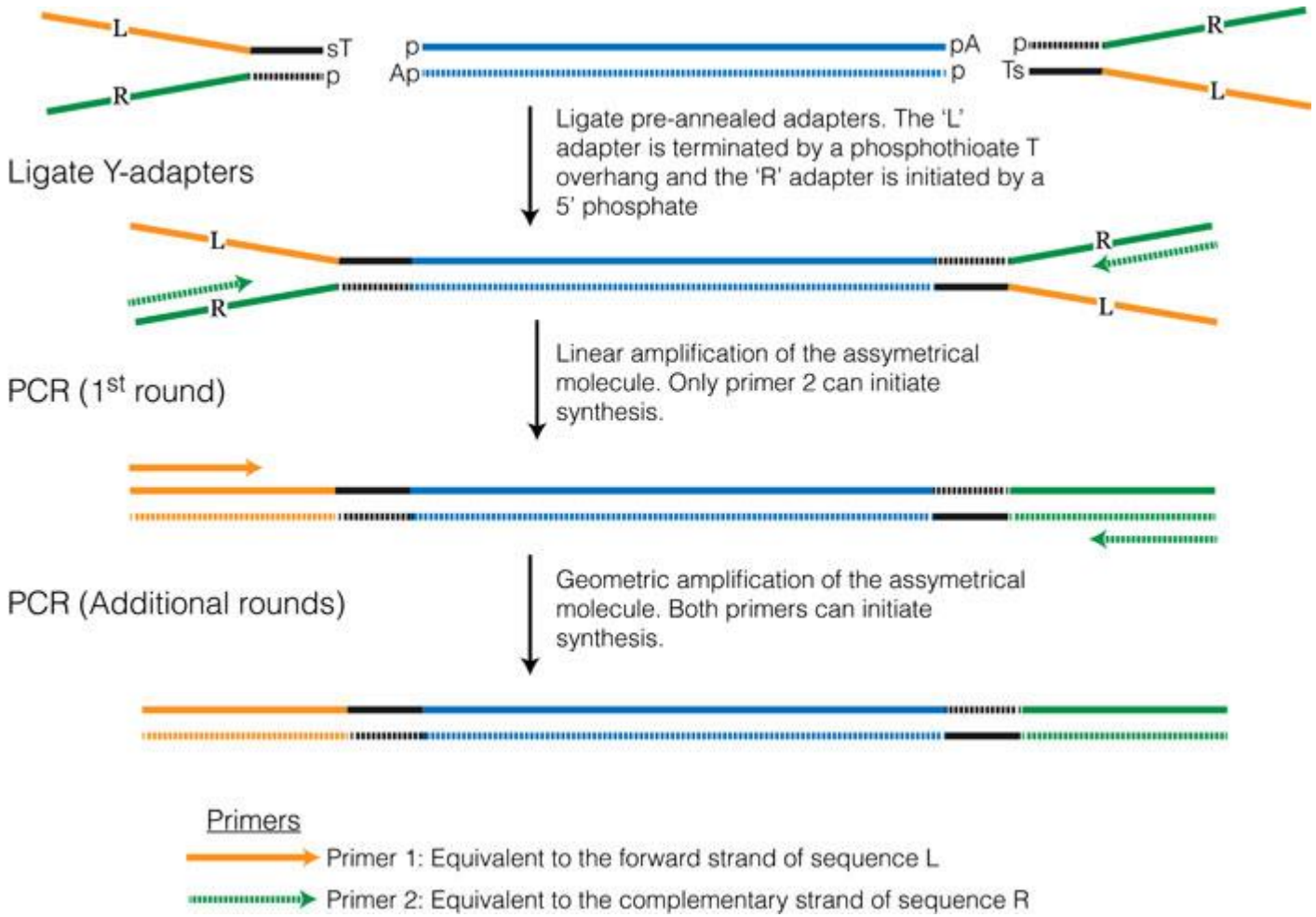
300bp area excised

Library preparation: PCR

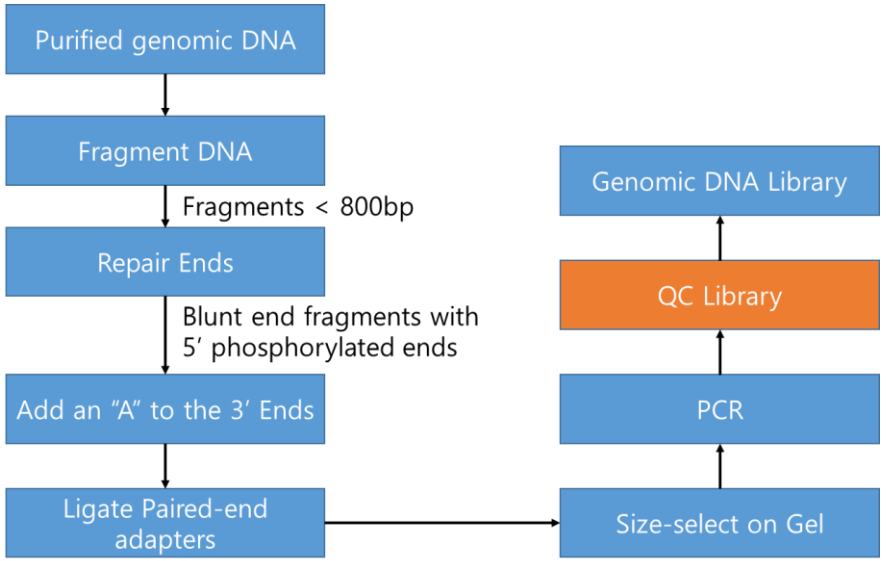


- Selectively enrich DNA fragments with adapters on both ends
- Amplify the amount of DNA in the library

Library preparation: PCR

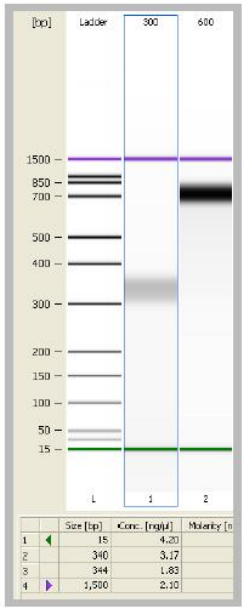


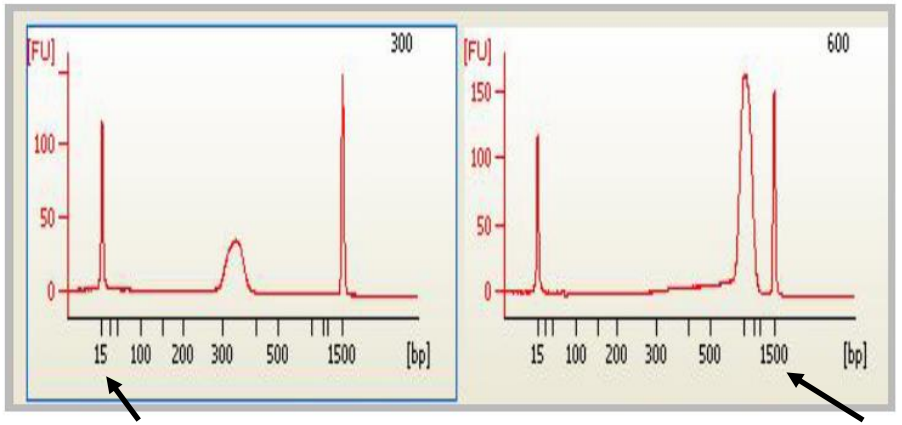
Library preparation: QC Library



QC by Agilent Bioanalyzer: gives size confirmation and visualizes unwanted products

- Agilent Bioanalyzer Results
  - Lane 1: 300 bp library
  - Lane 2: 600 bp library

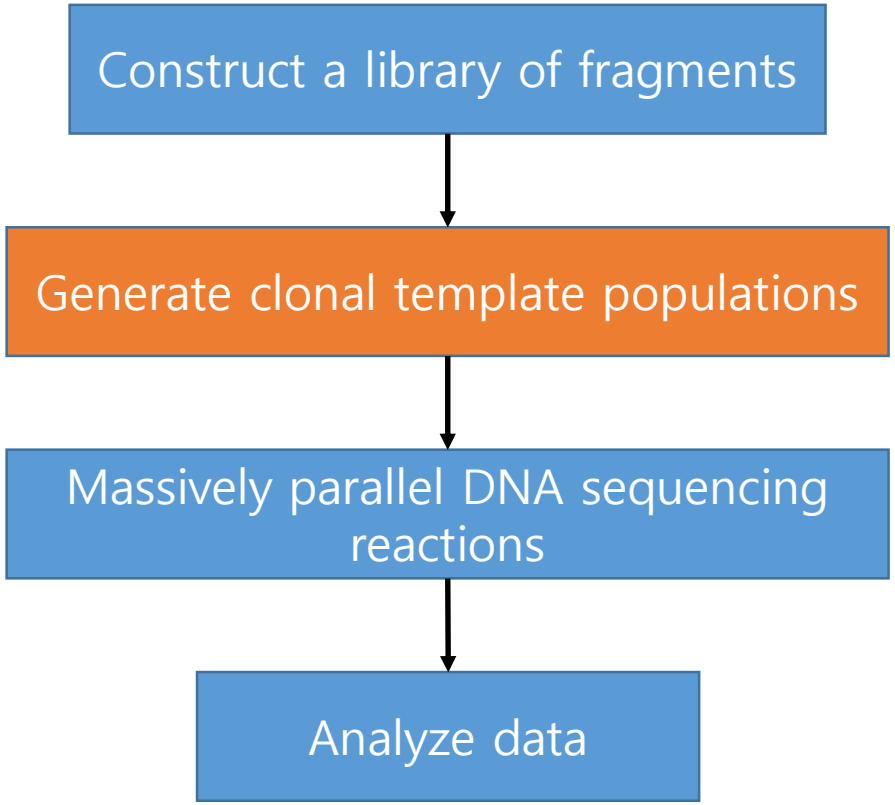




Lower marker
15bp

Upper marker
1500bp

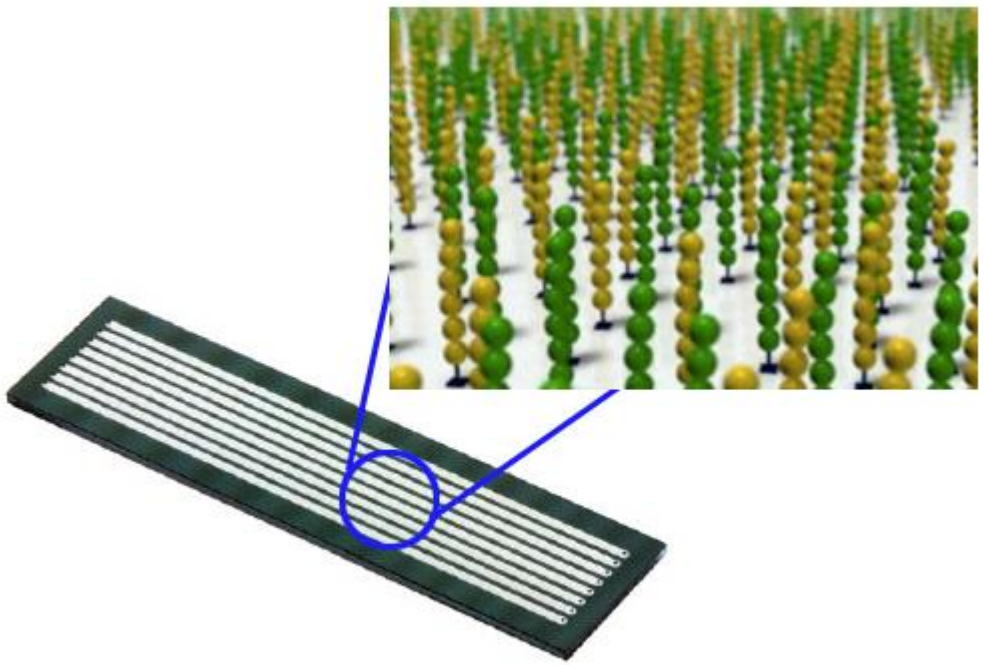
General principles of short-read NGS



Overview

- https://www.youtube.com/watch?v=fCd6B5HRaZ8

Cluster amplification: Flow cells

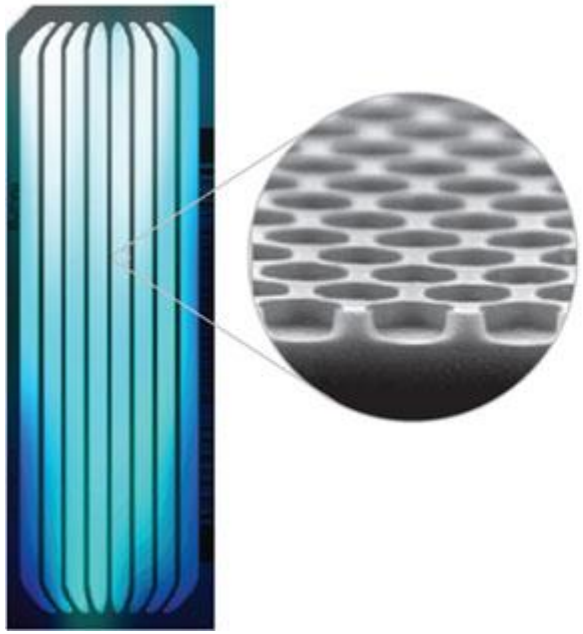


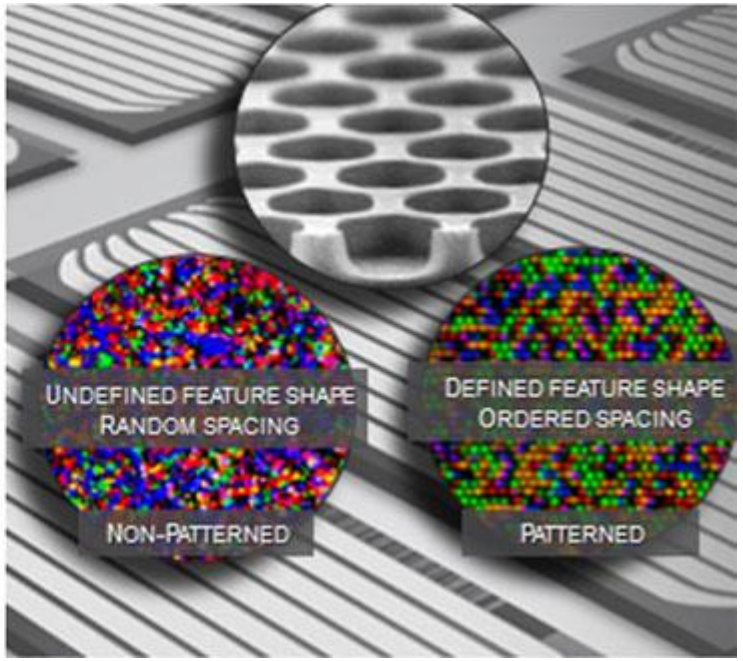
Cluster amplification: Flow cells

- Adapter-ligated library elements hybridize to complementary oligonucleotides on the surface of a flow cell. Each attached library fragment acted as a seed and is amplified to generate a clonal cluster containing thousands of identical fragments.

- Ideally, clusters are of similar size and spaced well apart from each other to achieve accurate resolution during imaging. In reality, DNA clusters are randomly distributed across the flow cell with many clusters in close proximity to neighboring clusters, if the sample is overloaded, making it difficult to discern individual clusters from each others and reducing the amount of information generated during the run.

Cluster amplification: Patterned flow cells



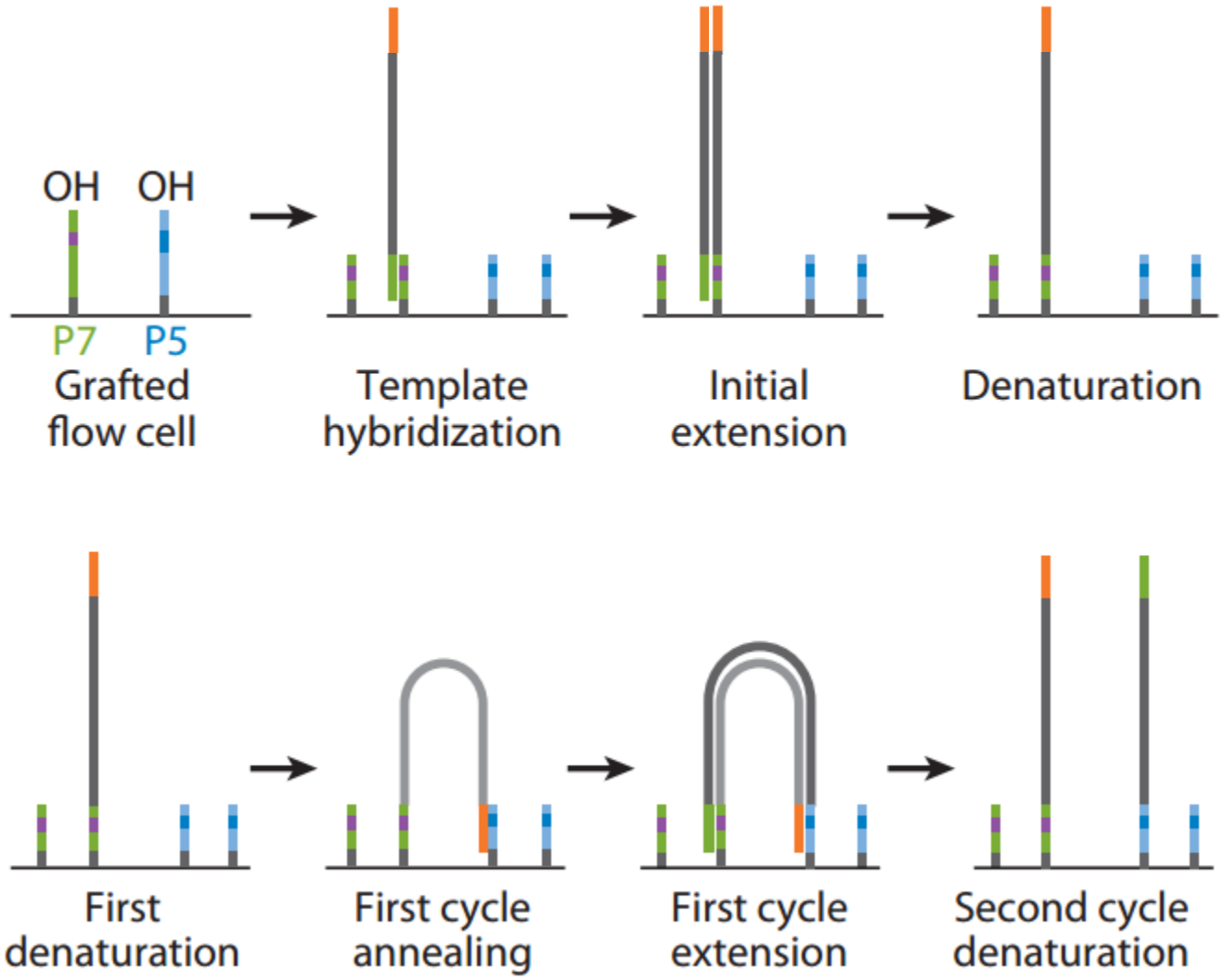


Cluster amplification: Patterned flow cells

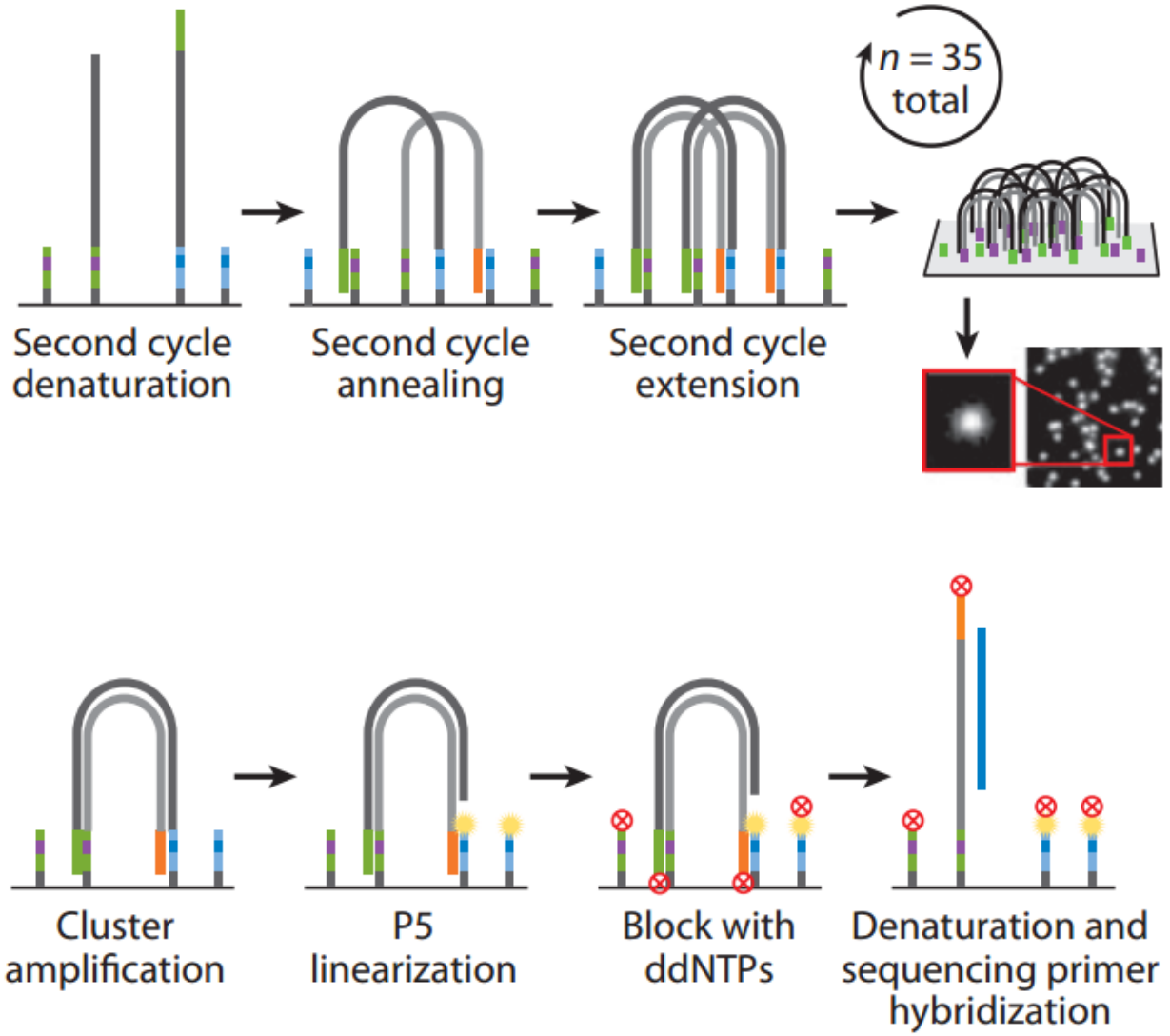
- Patterned flow cell technology provides even cluster spacing and uniform feature size to deliver extremely high cluster densities.

- Clusters can only form in the nanowells, allowing accurate resolution of clusters during imaging.

Cluster amplification

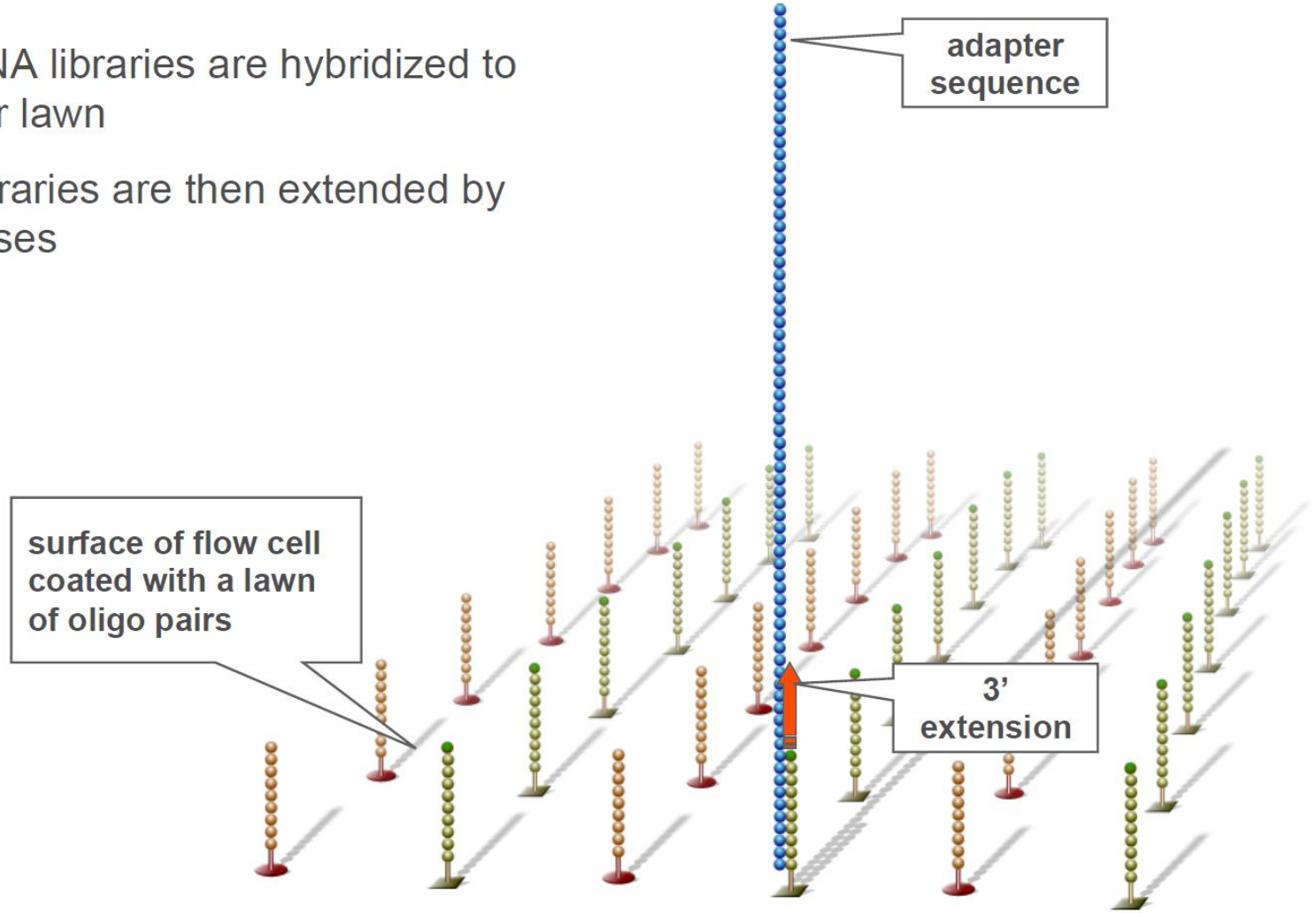


Cluster amplification



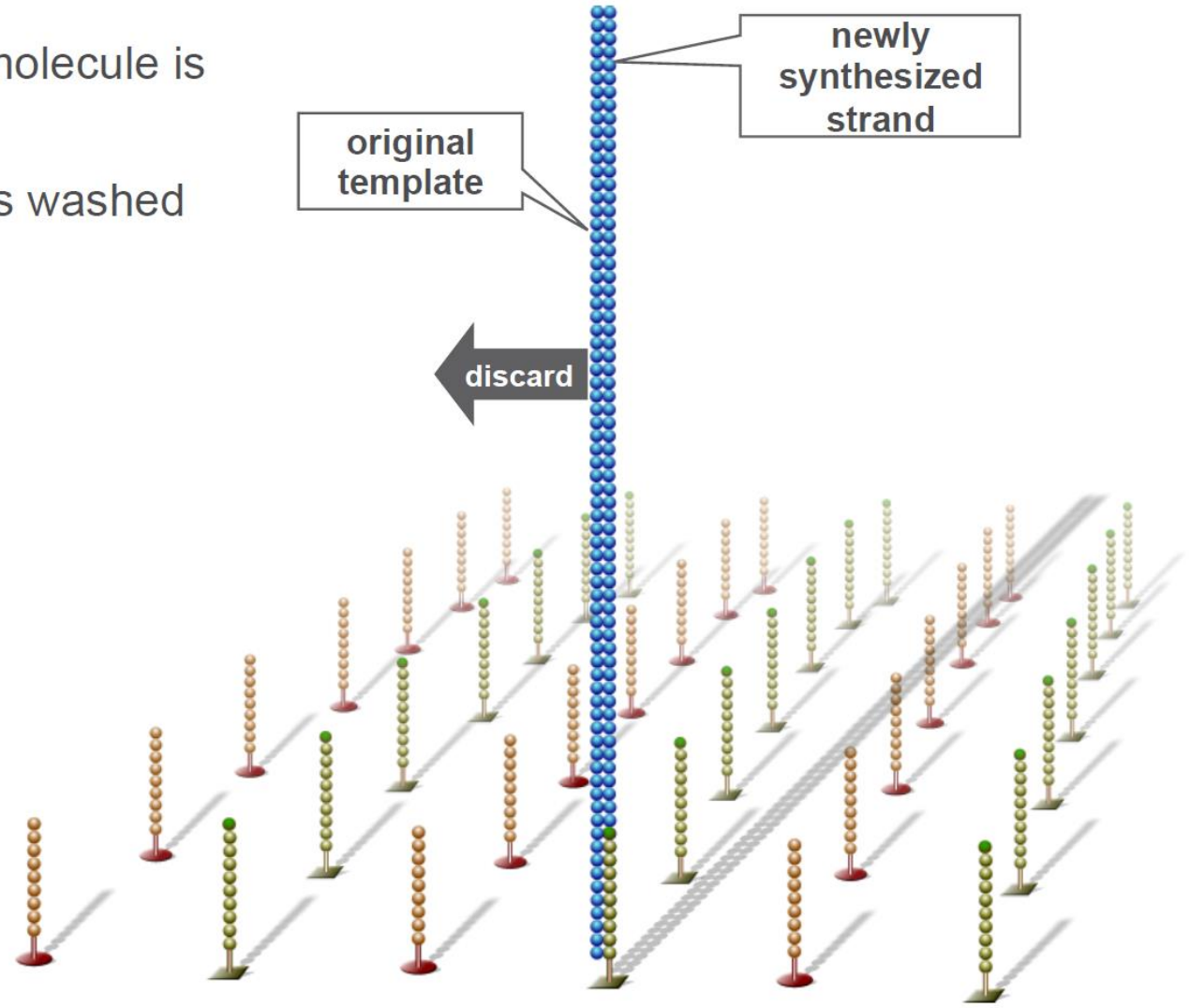
Cluster amplification: Hybridization and extension

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



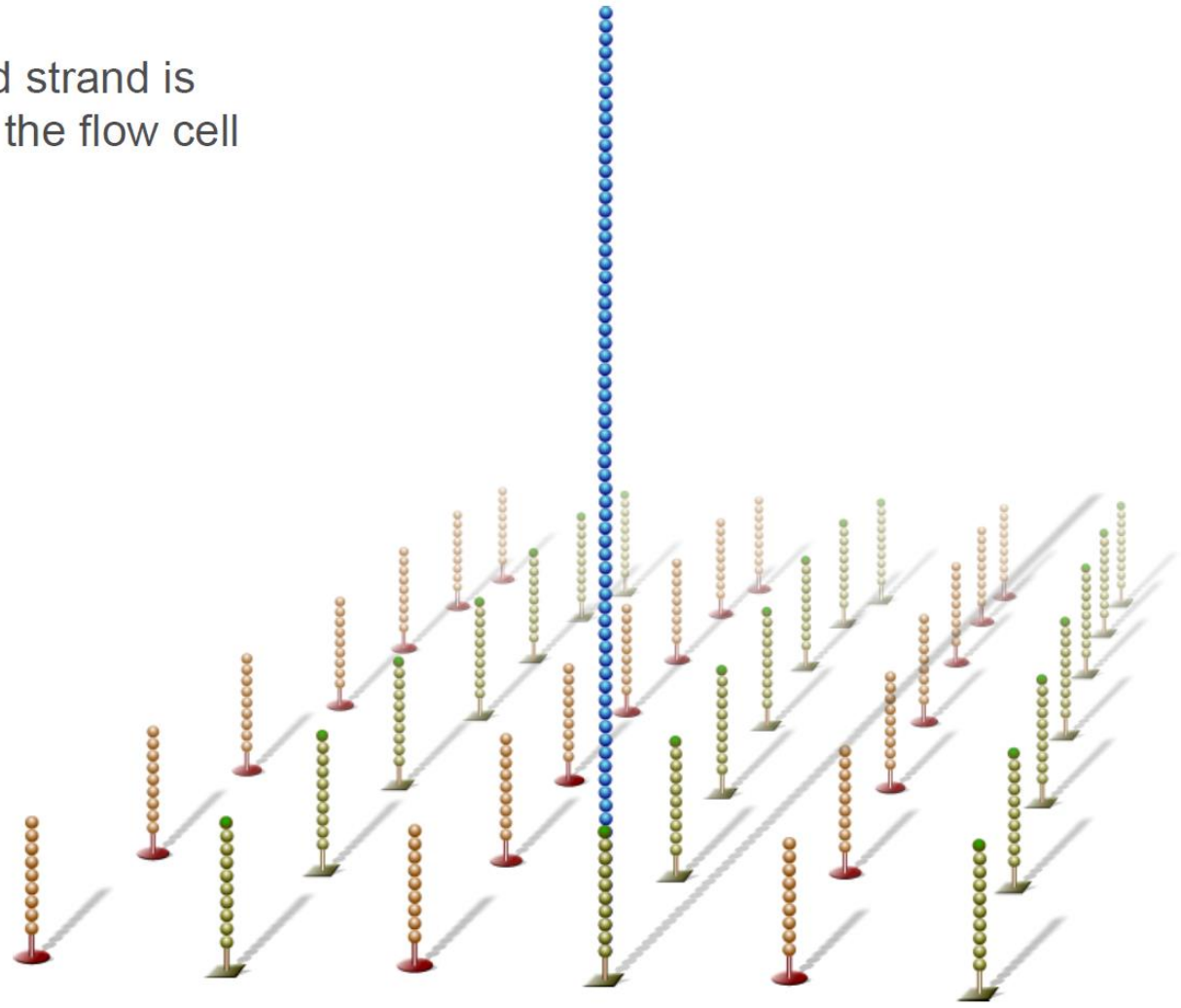
Cluster amplification: Denaturation

- The double-stranded molecule is denatured
- The original template is washed away



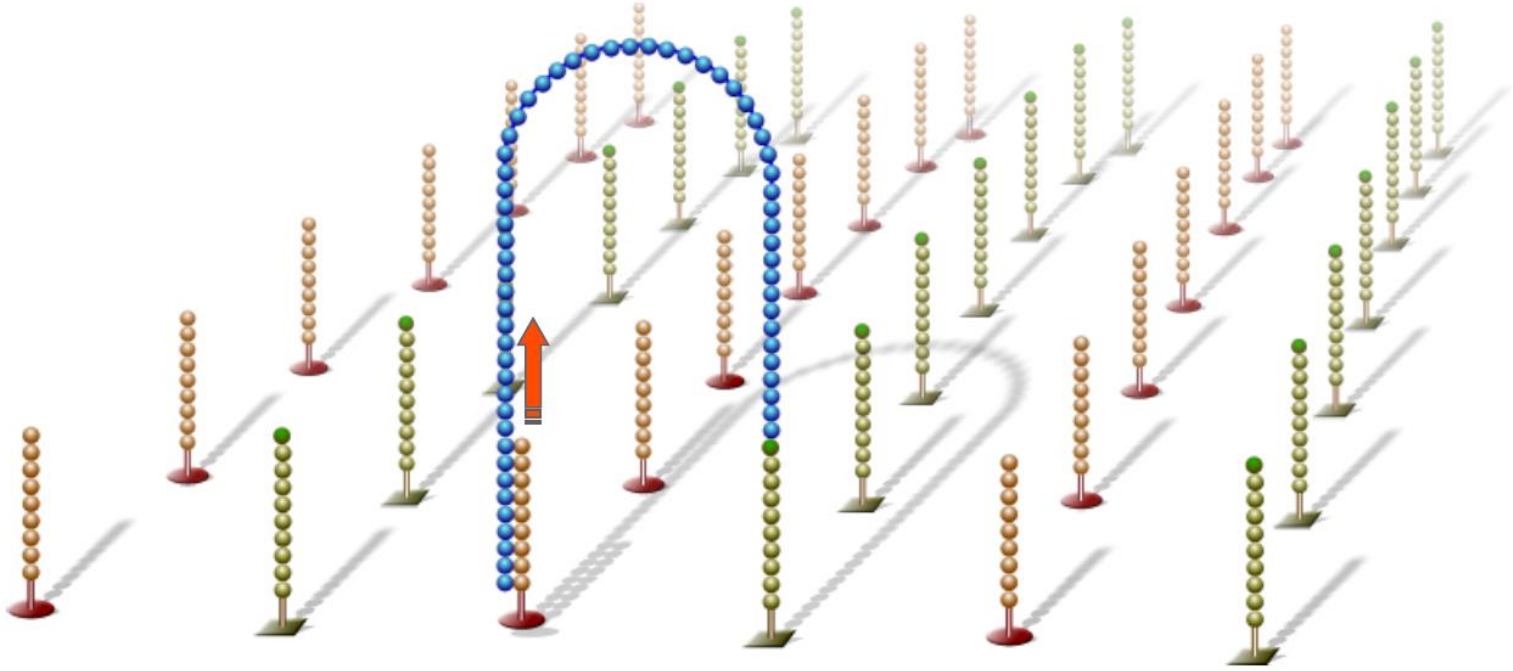
Cluster amplification: Anchor the template to the surface

- The newly synthesized strand is covalently attached to the flow cell surface



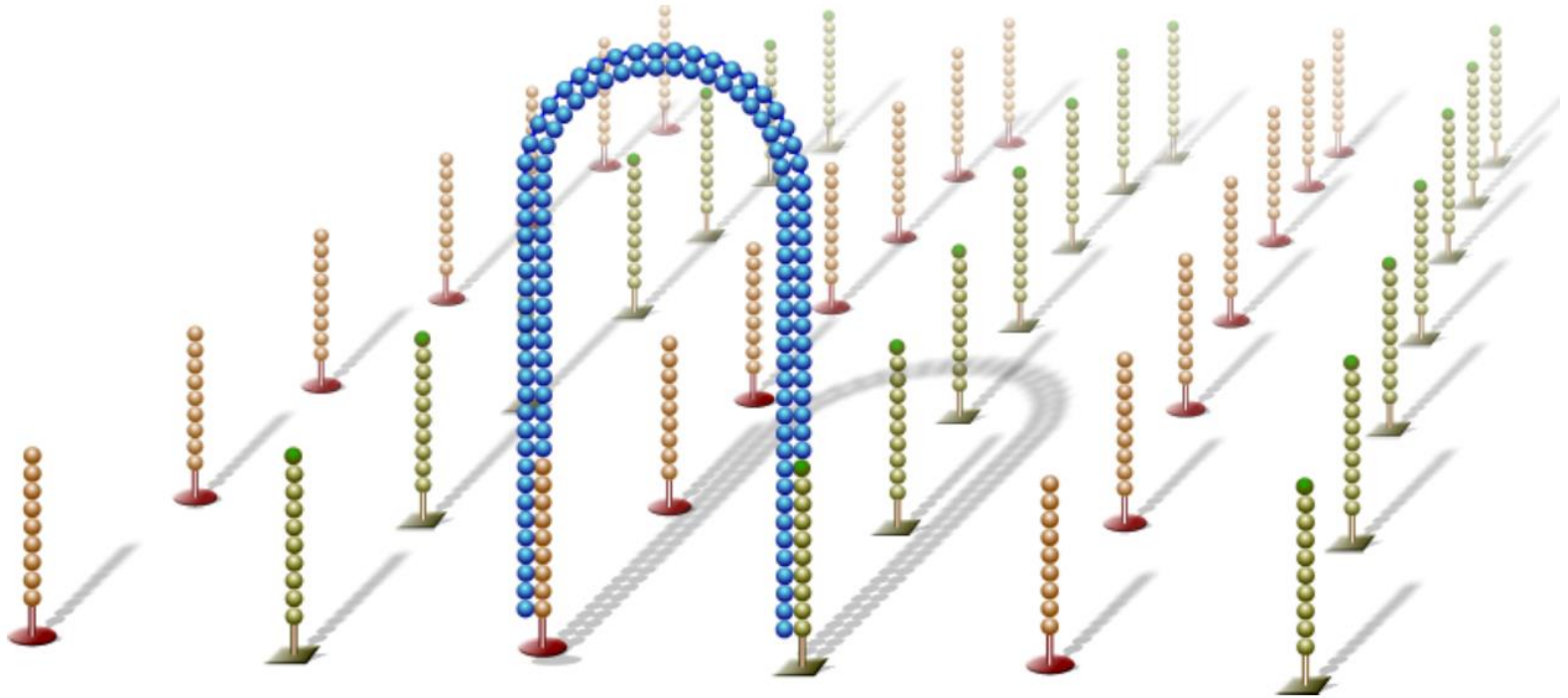
Cluster amplification: Bridge amplification

- The single-stranded molecule flips over and forms a bridge by hybridizing to an adjacent, complementary primer
- The hybridized primer is extended by polymerases



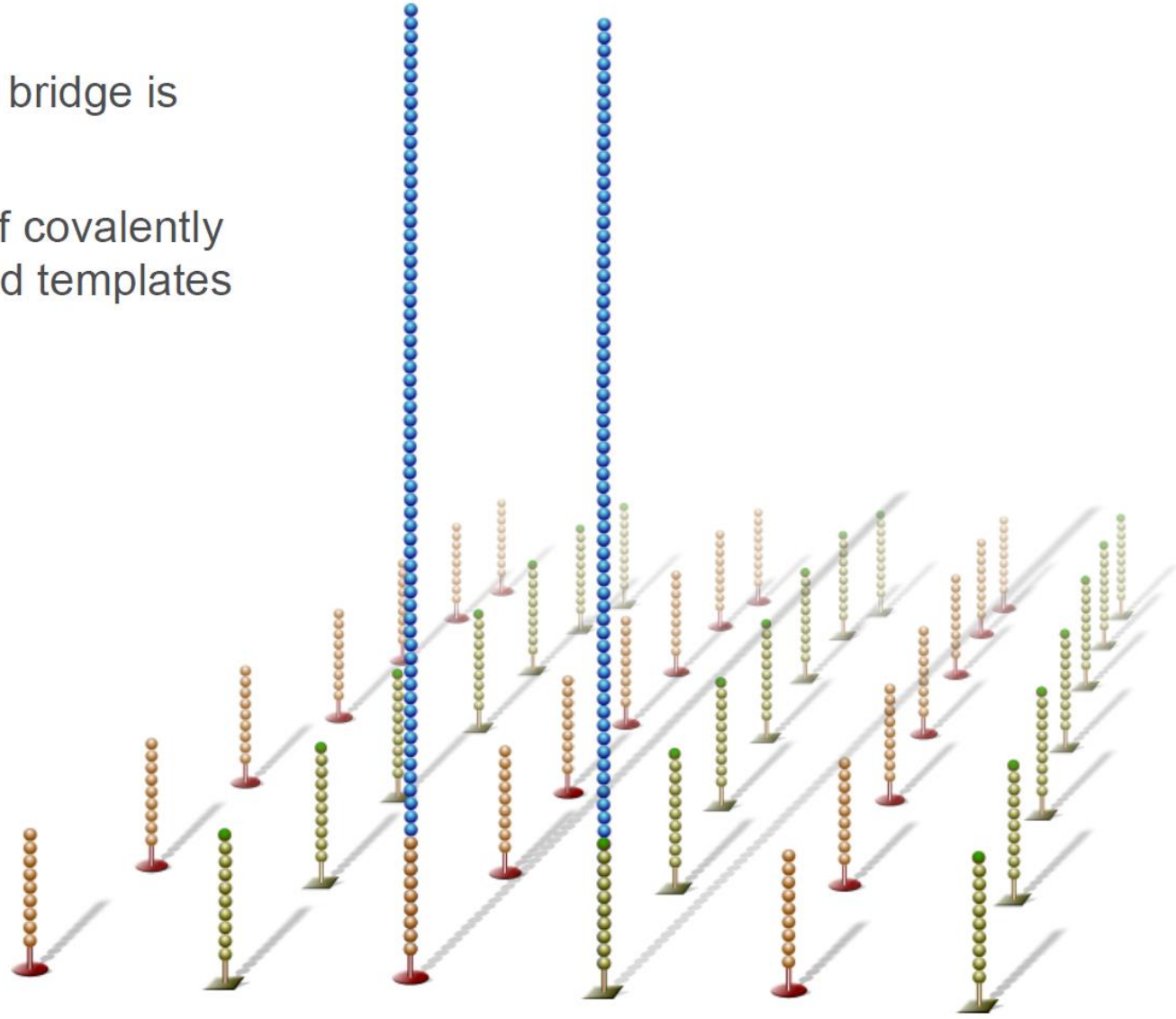
Cluster amplification: Bridge amplification

- A double-stranded bridge is formed



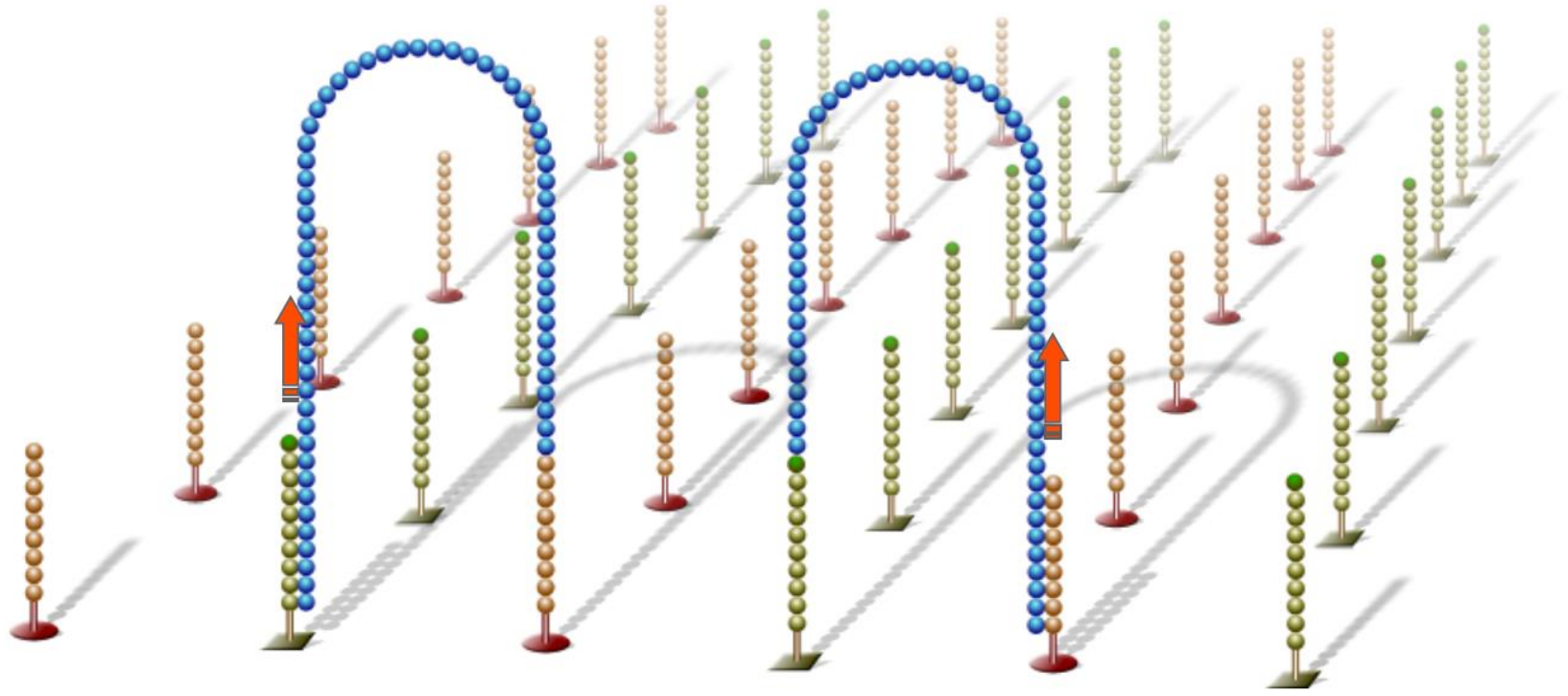
Cluster amplification: Denaturation

- The double-stranded bridge is denatured
- Result: Two copies of covalently bound single-stranded templates



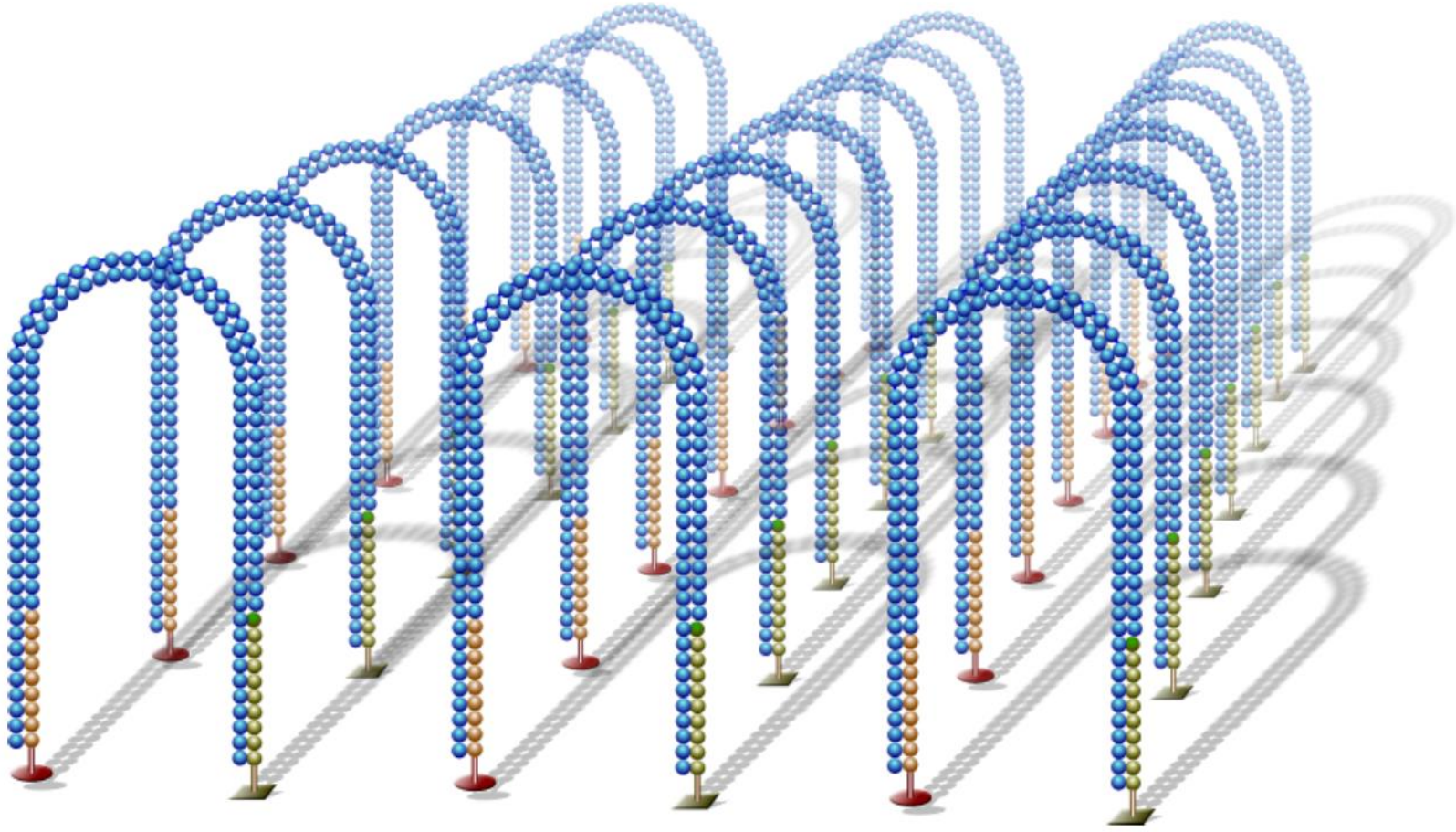
Cluster amplification: Bridge amplification

- Single-stranded molecules flip over to hybridize to adjacent primers
- Hybridized primer is extended by polymerase

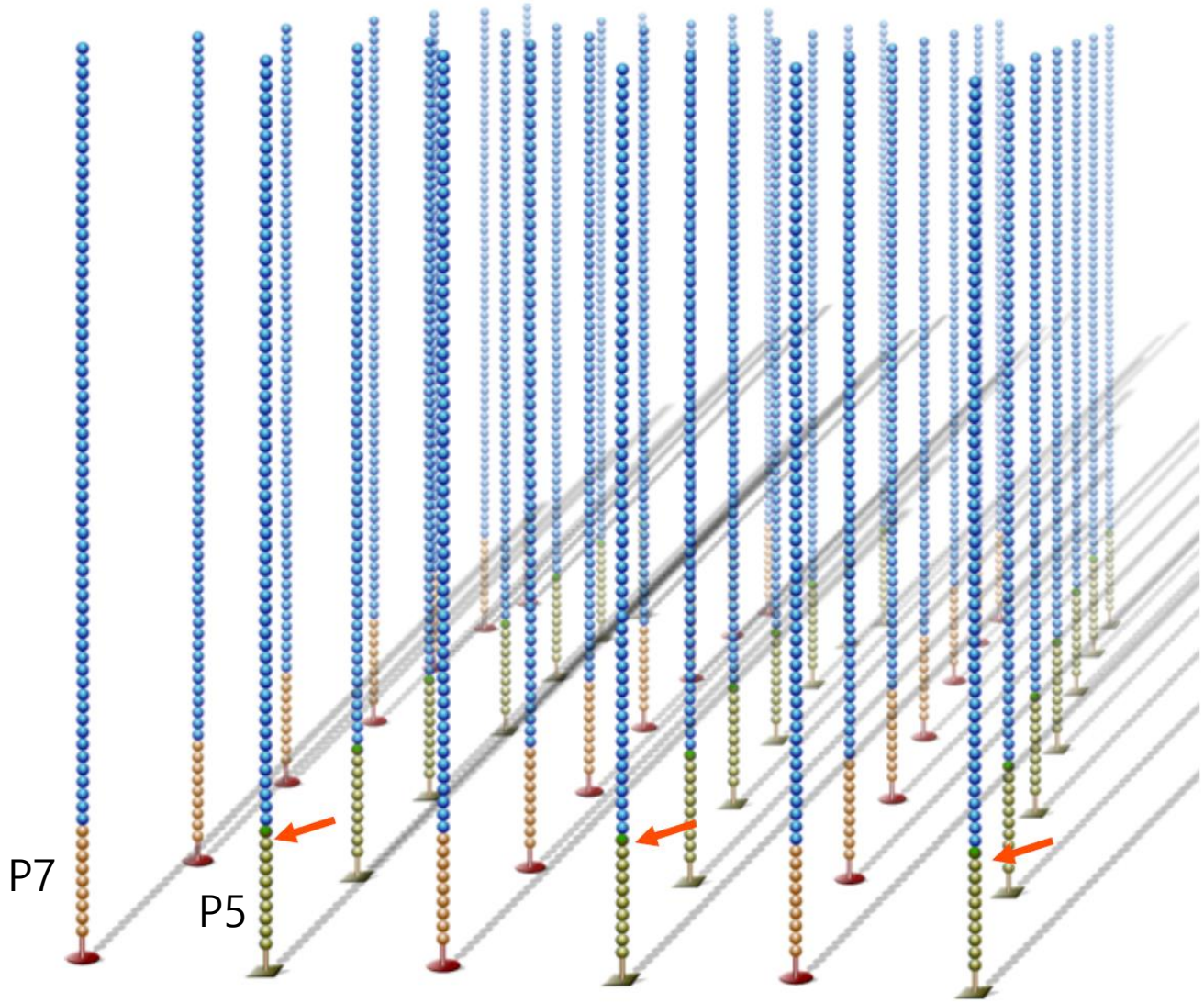


Cluster amplification: Bridge amplification

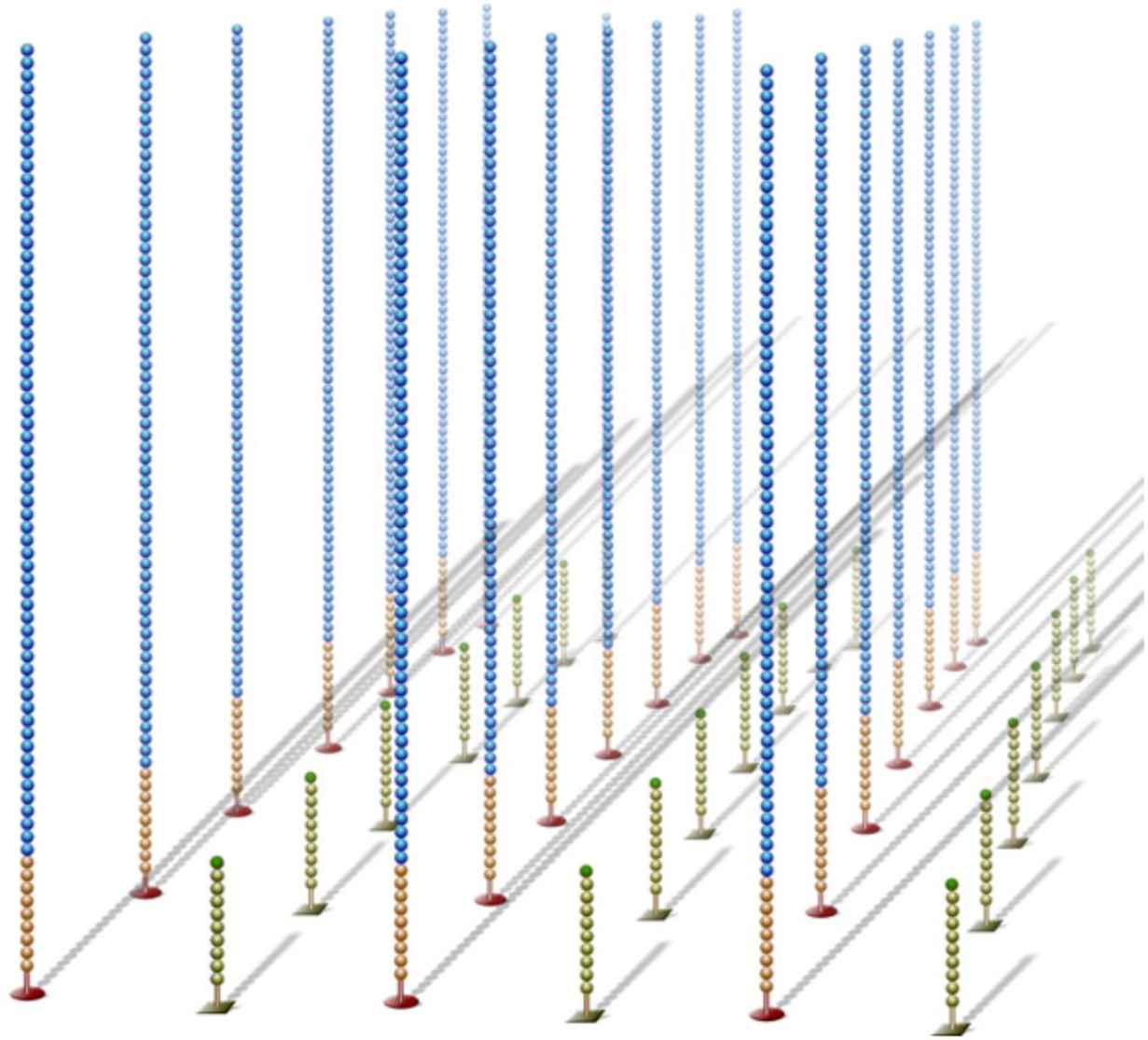
- Bridge amplification cycle repeated until multiple bridges are formed



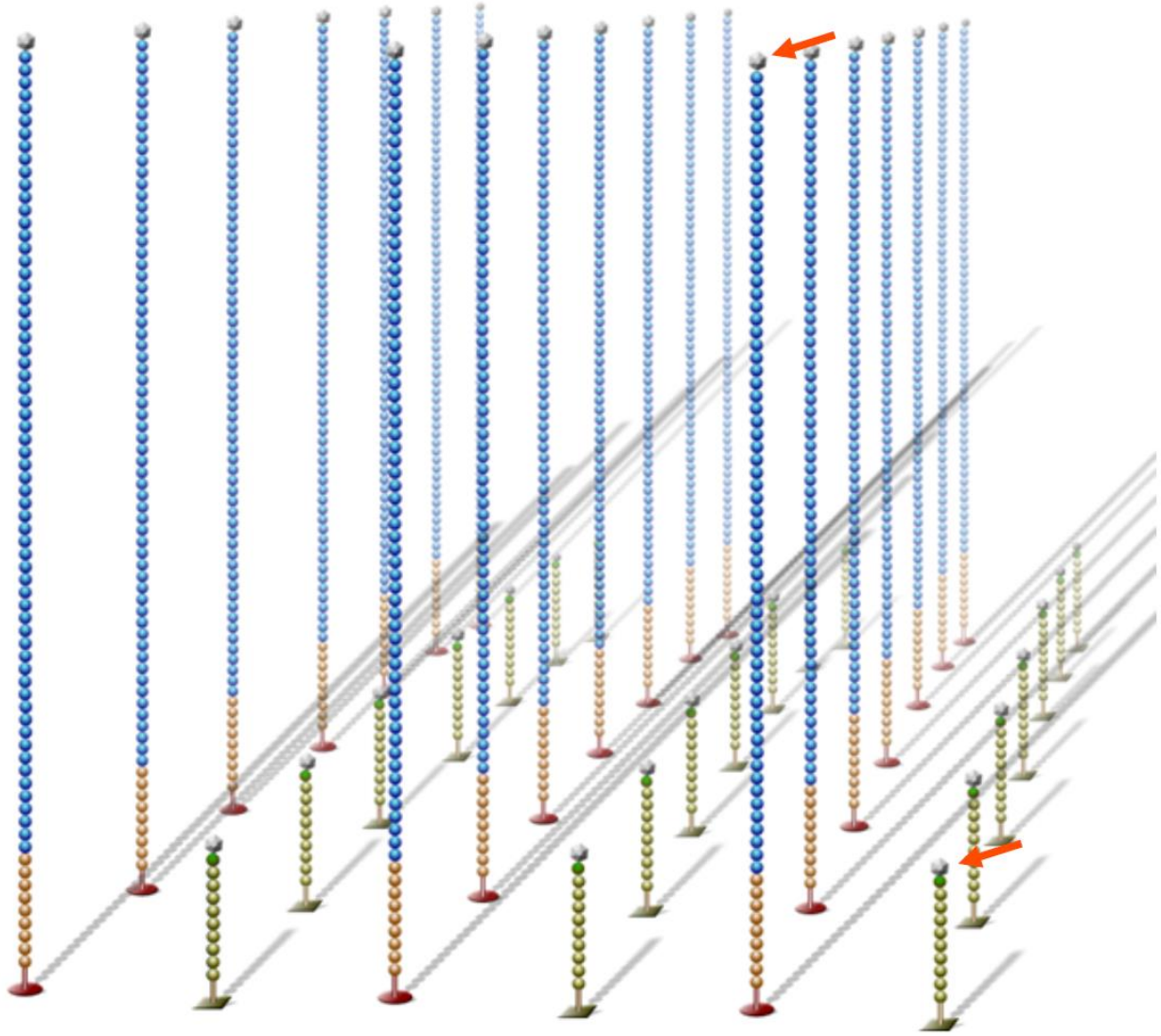
Cluster amplification: P5 Linearization



Cluster amplification: P5 Linearization

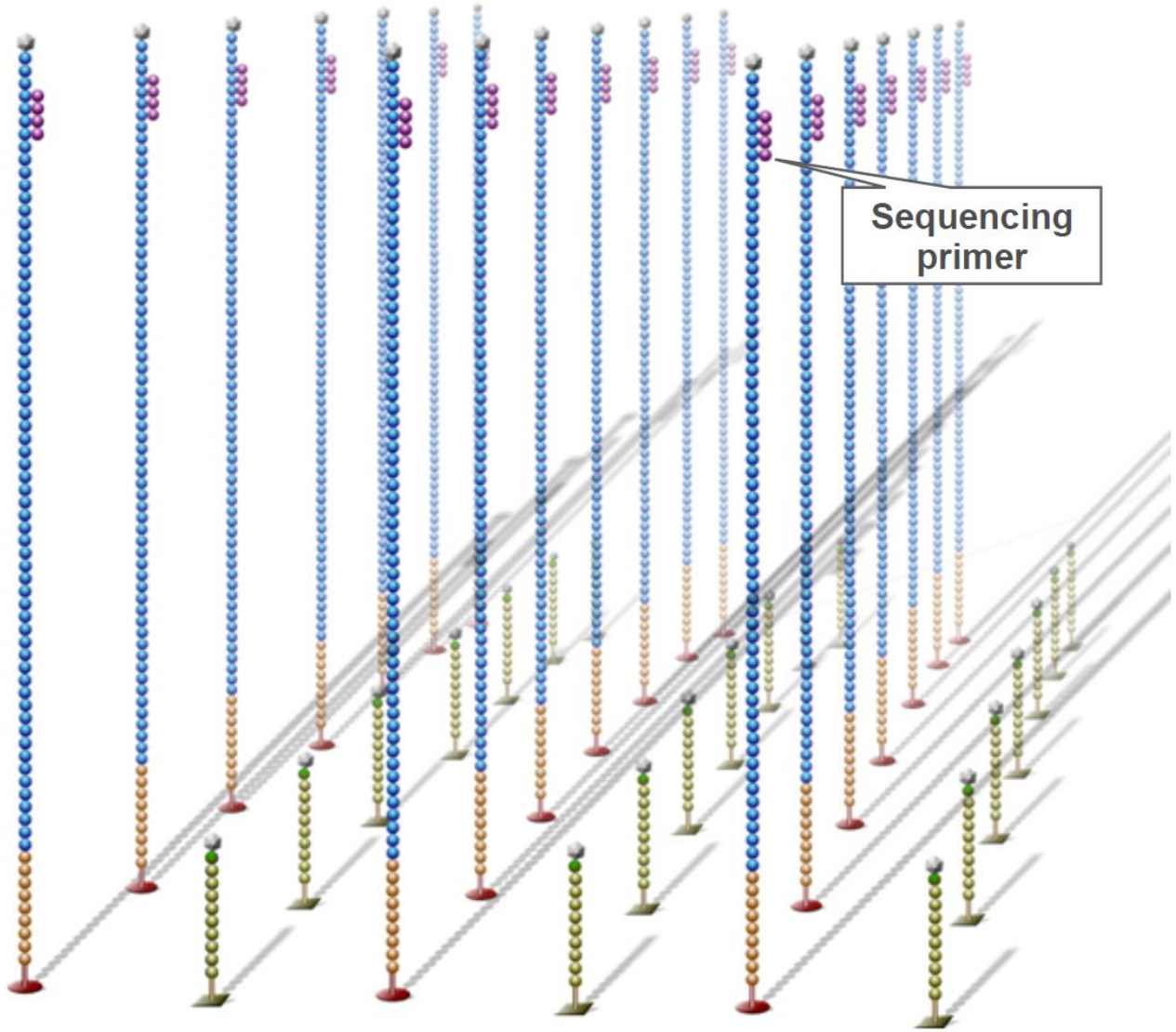


Cluster amplification: Blocking

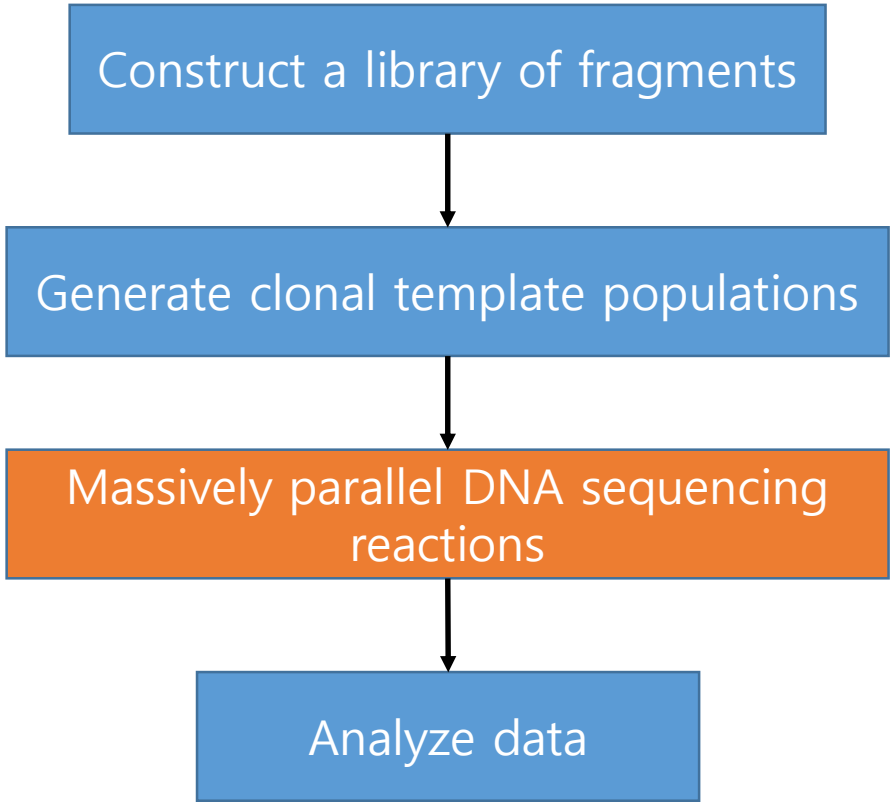


Cluster amplification: Read1 sequencing

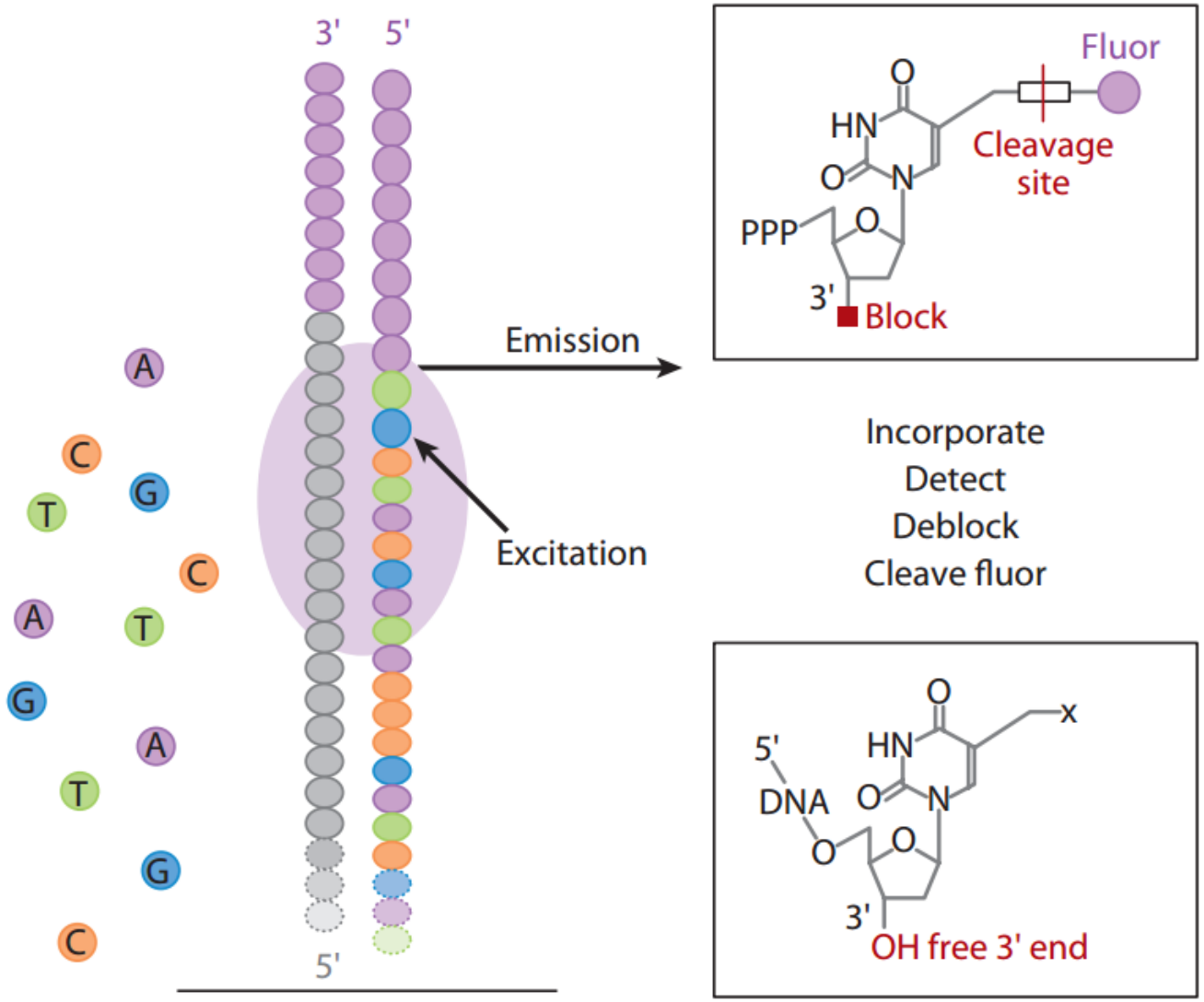
- A sequencing primer is introduced to the flow cell and hybridized to the adapter sequence annealing site



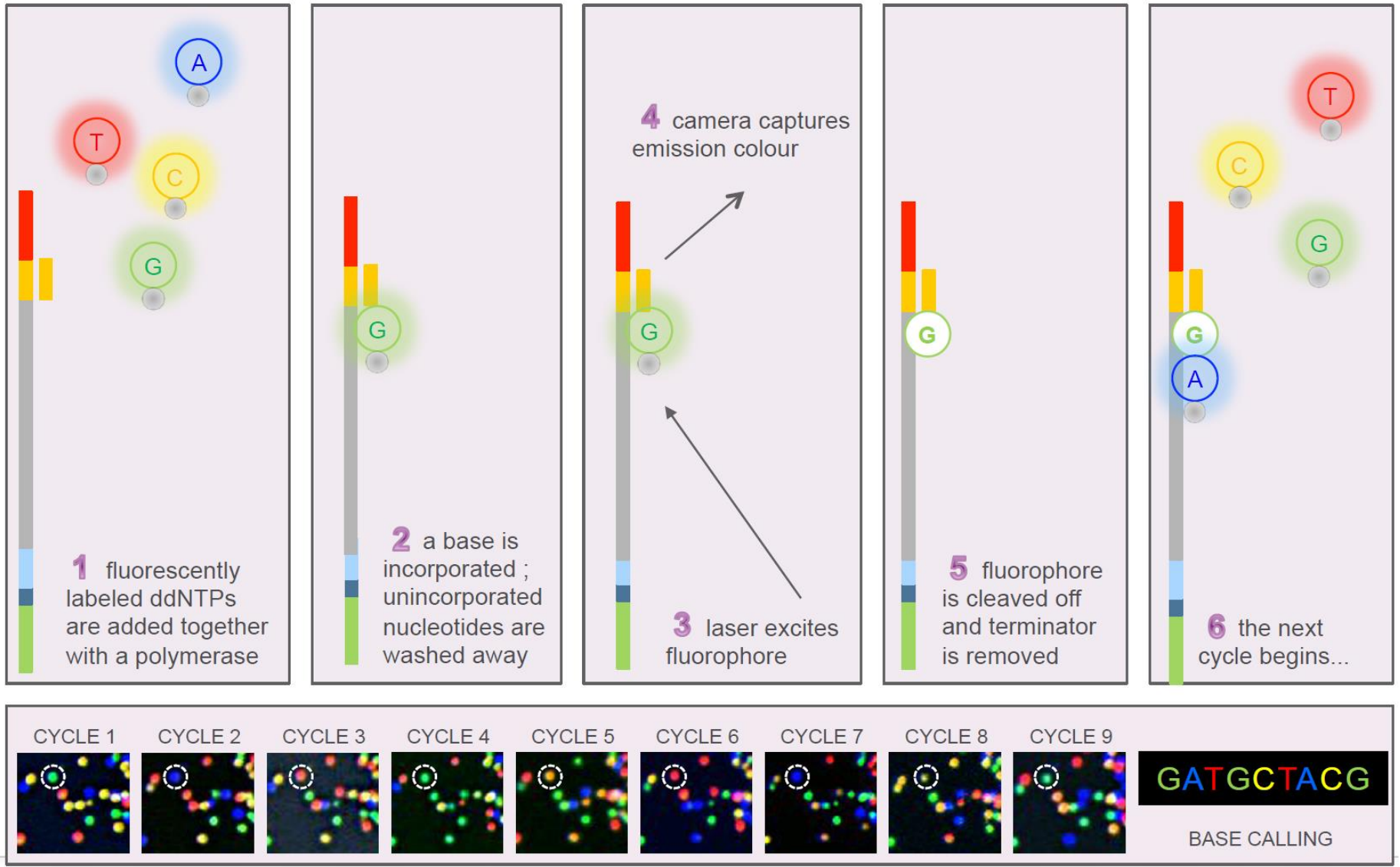
General principles of short-read NGS



Sequencing by synthesis

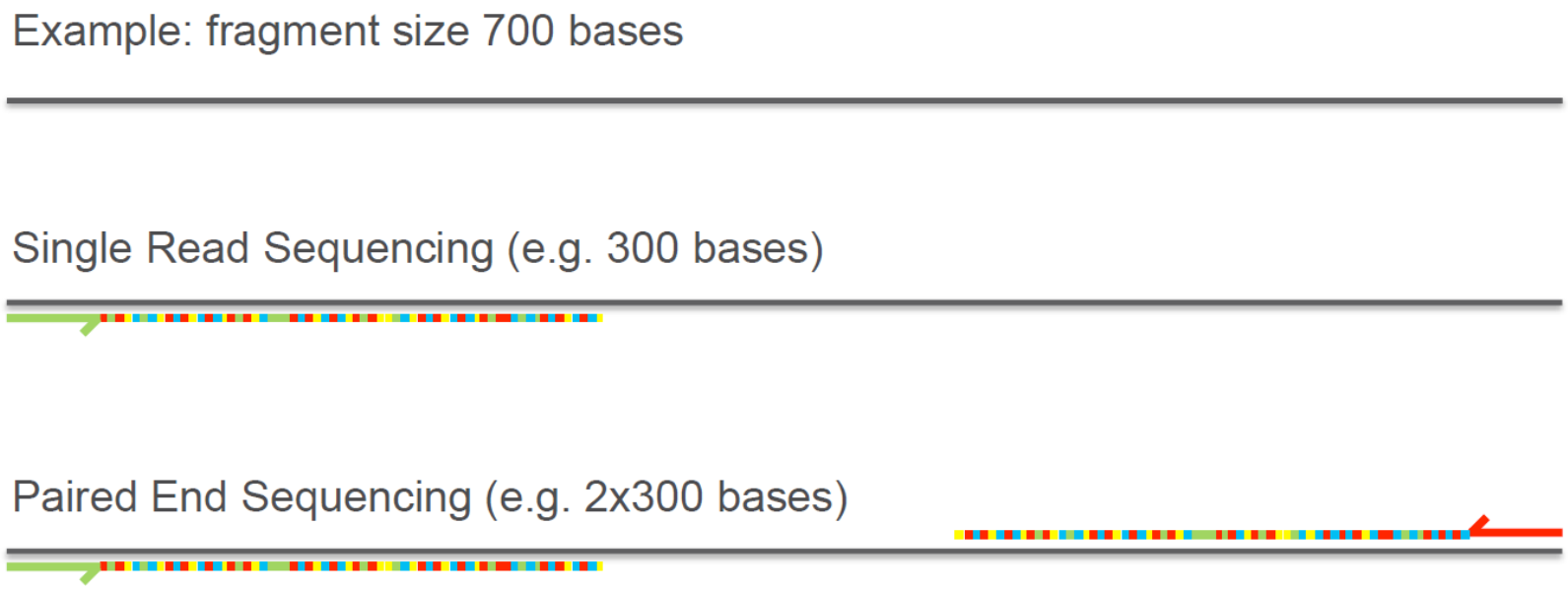


Sequencing by synthesis



Single read, paired-end and read lengths

- Program the system to sequence a specific number of bases (1-600 bases)
- Sequence the strands from both directions to achieve a total of e.g. 600 bases (2×300 bases)



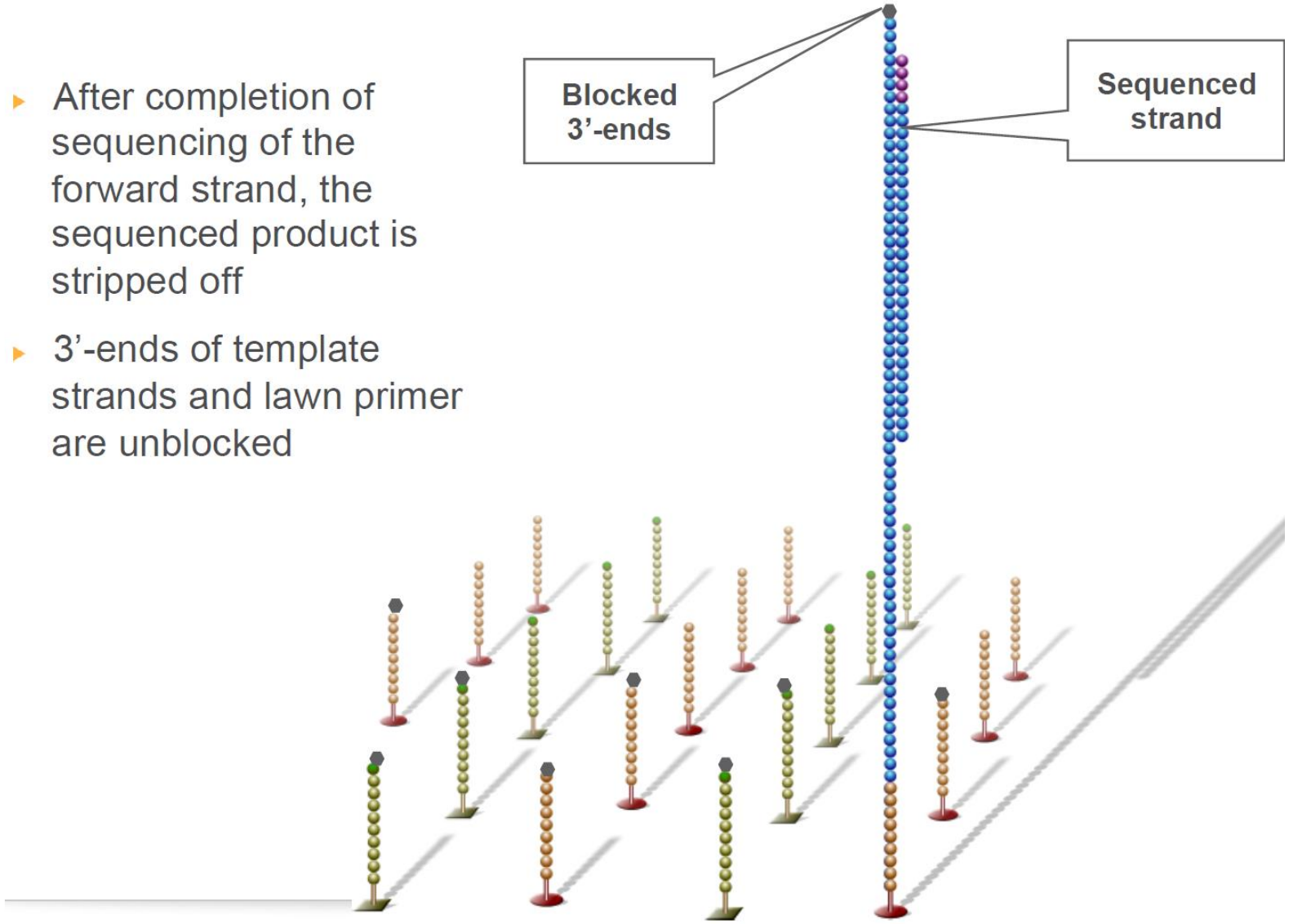
Paired-end sequencing

Longer read lengths improve 1) the overall length of contiguous sequence that can be assembled, and 2) the certainty of short read alignments.

Several next-generation sequencers have offered increases in read length over time. Another improvement has resulted from paired-end sequencing, producing sequence data from both ends of each library fragment.

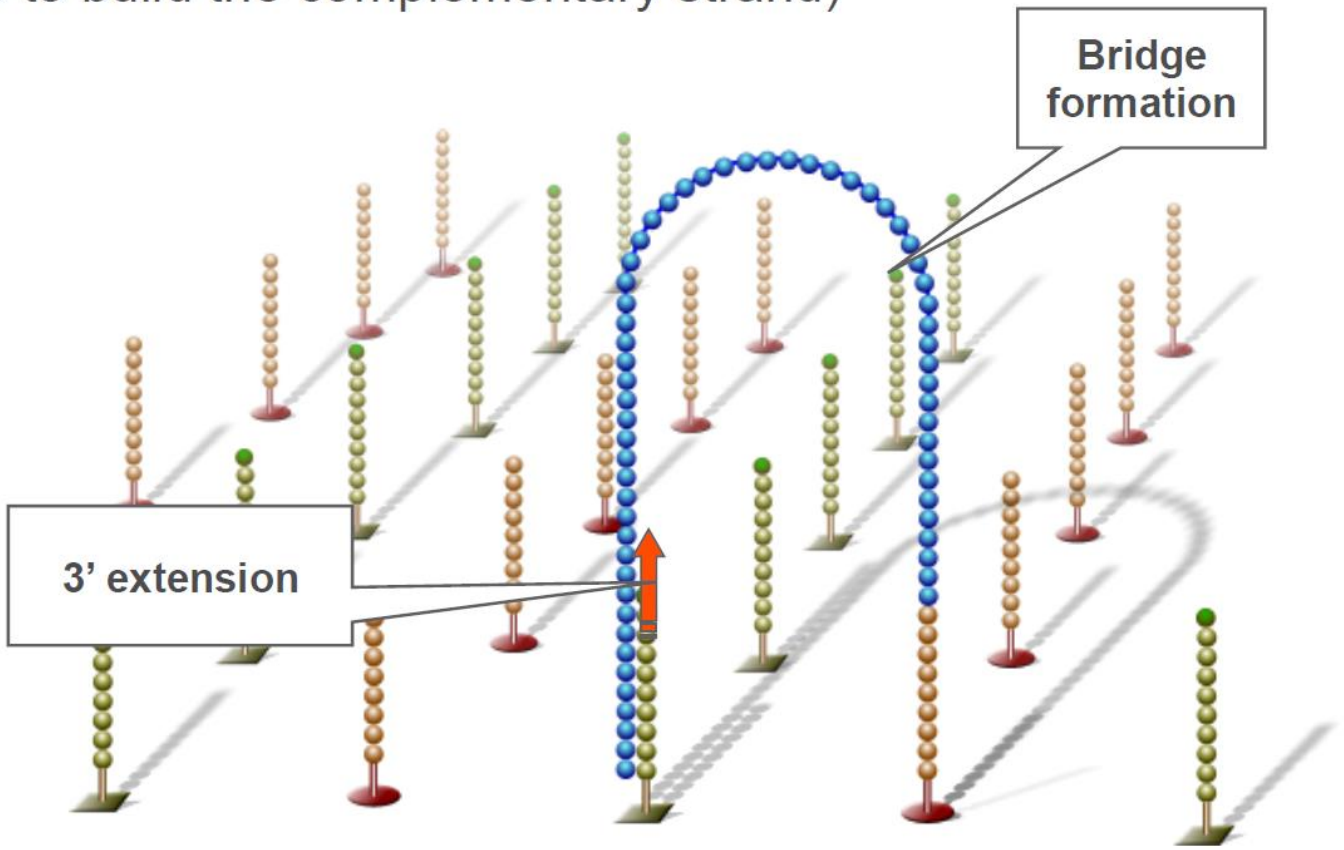
Paired-end sequencing

- After completion of sequencing of the forward strand, the sequenced product is stripped off
- 3’-ends of template strands and lawn primer are unblocked

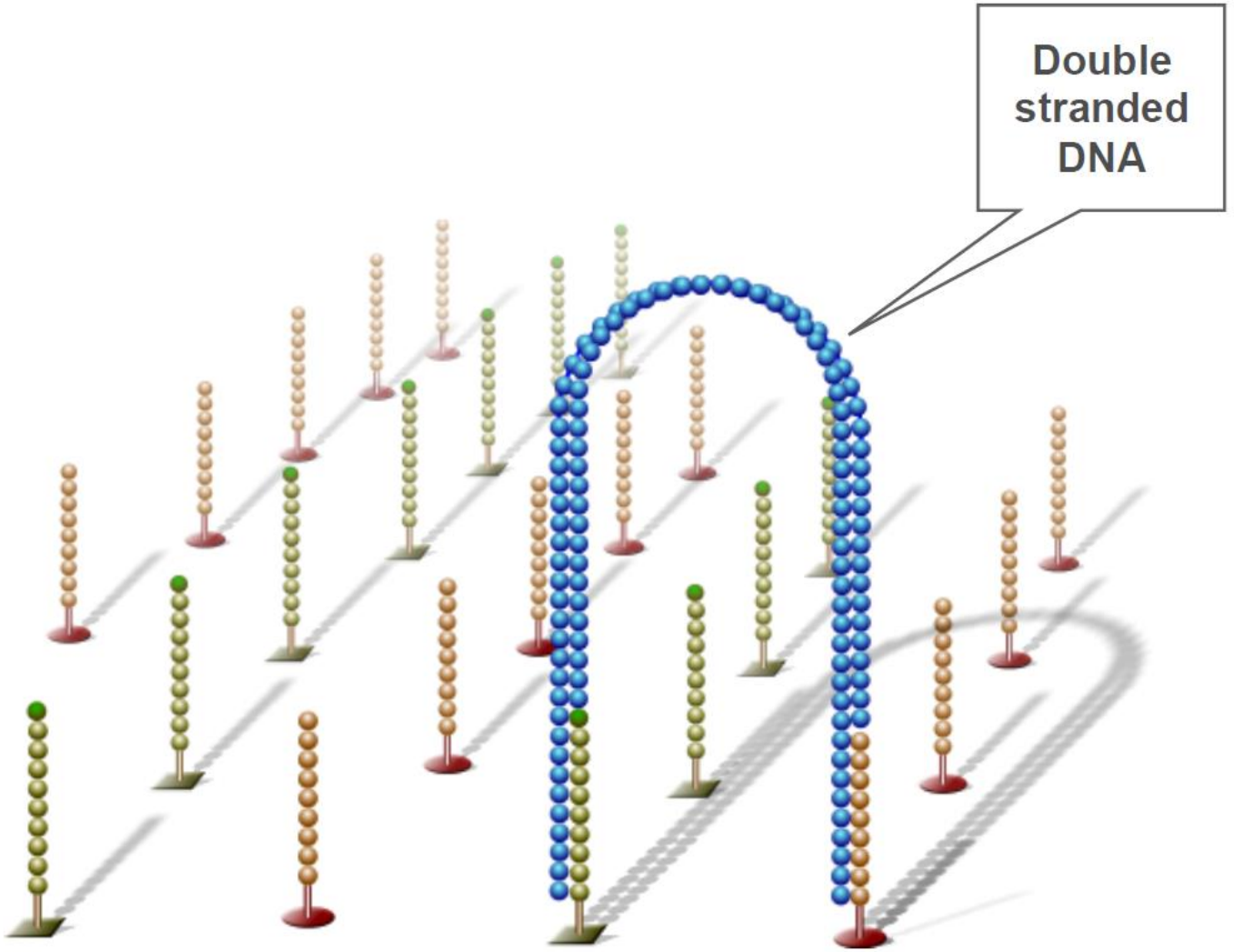


Paired-end sequencing

- Single-stranded template loops over to form a bridge by hybridizing with a lawn primer
- 3'-ends of lawn primer is extended (double-stranded stretch allows the polymerase to build the complementary strand)

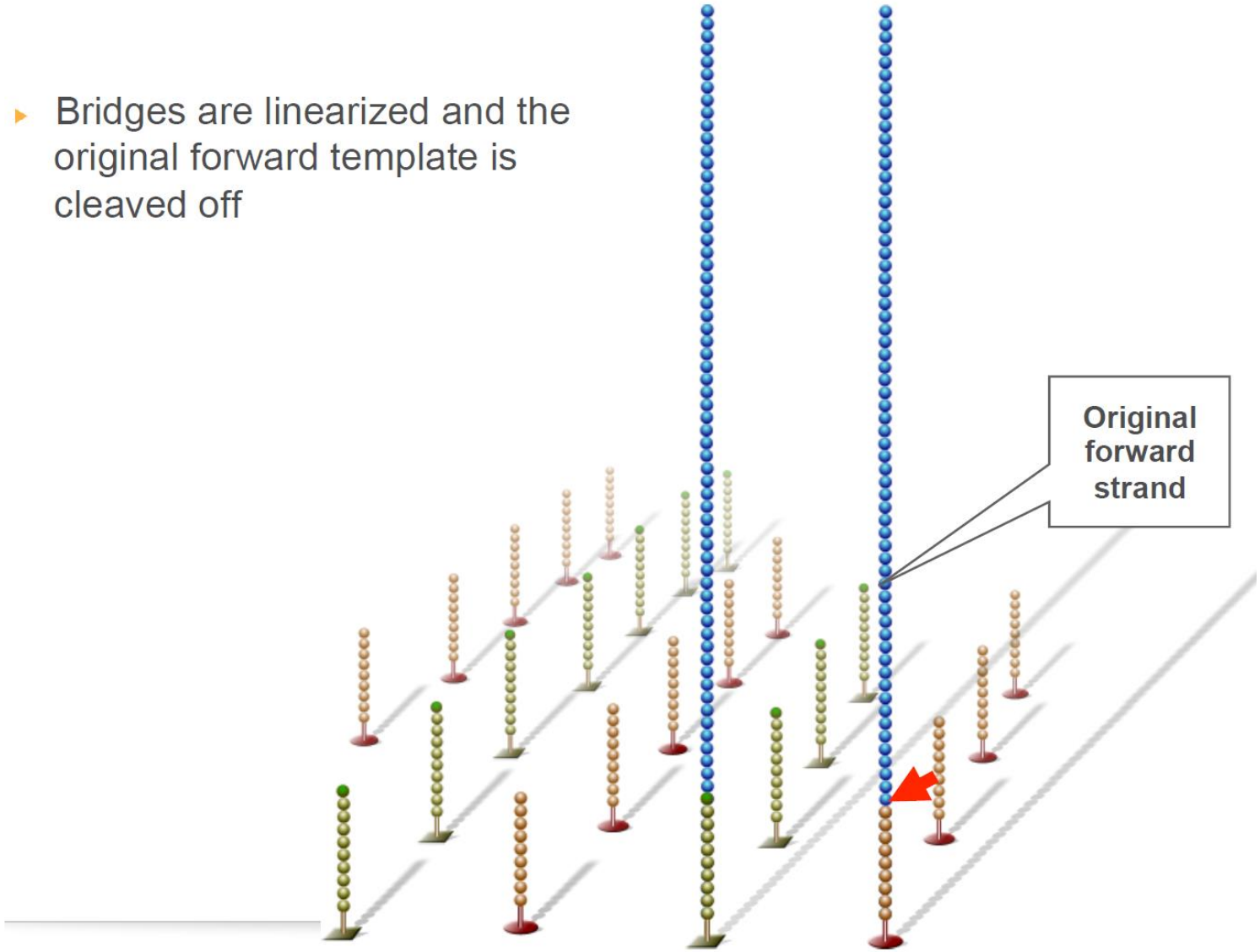


Paired-end sequencing



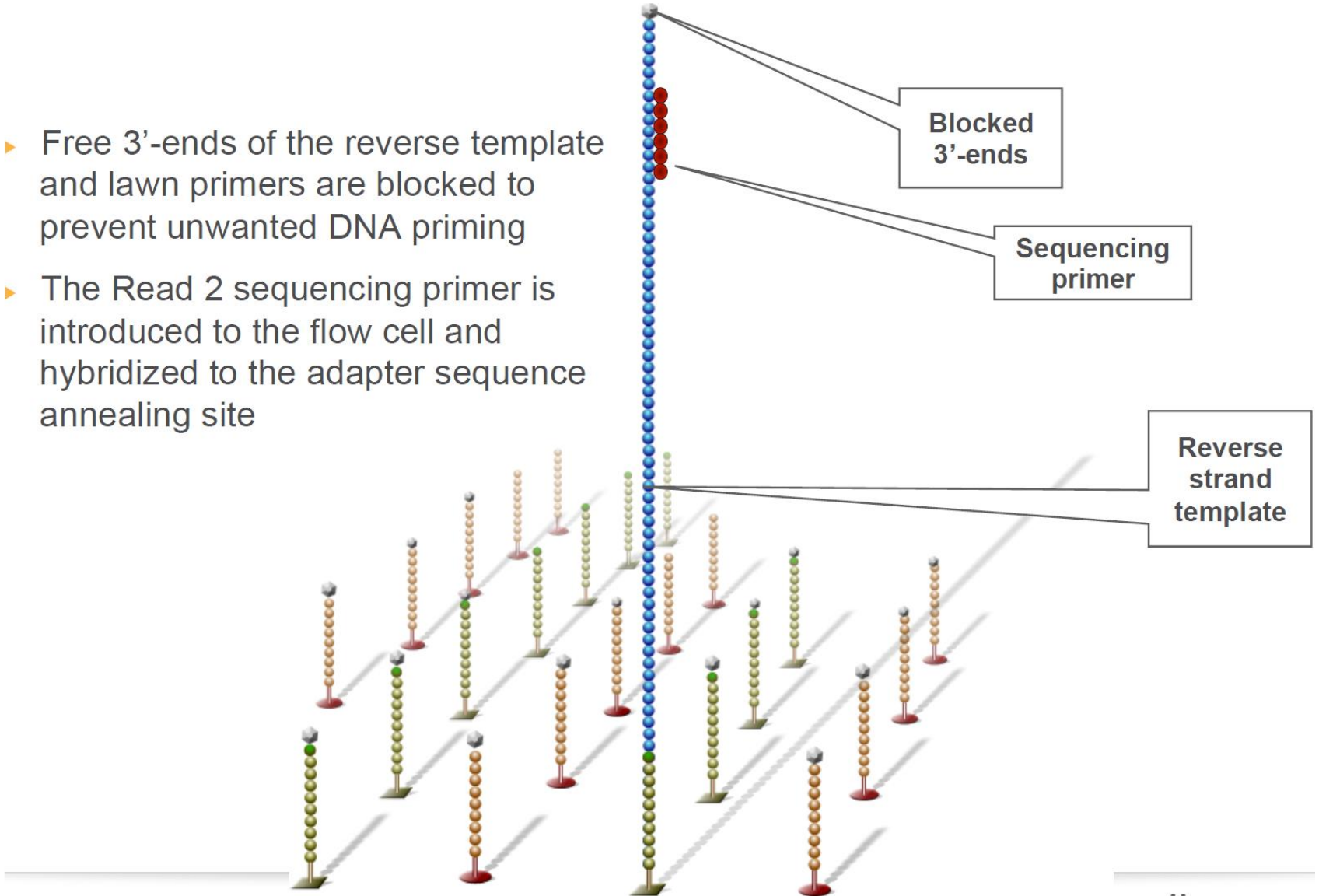
Paired-end sequencing: P7 linearization

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



Paired-end sequencing

- Free 3’-ends of the reverse template and lawn primers are blocked to prevent unwanted DNA priming
- The Read 2 sequencing primer is introduced to the flow cell and hybridized to the adapter sequence annealing site



Illumina platforms: Benchtop sequencers

Popular Applications & Methods	iSeq 100	MiniSeq	MiSeq Series	NextSeq 550 Series	NextSeq 1000 & 2000
	Key Application	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)					
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●	●
Exome & Large Panel Sequencing (enrichment-based)				●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)				●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)				●	●
Targeted Gene Expression Profiling	●	●	●	●	●
miRNA & Small RNA Analysis	●	●	●	●	●
DNA-Protein Interaction Analysis (ChIP-Seq)			●	●	●
Methylation Sequencing				●	●
16S Metagenomic Sequencing		●	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)				●	●
Cell-Free Sequencing & Liquid Biopsy Analysis				●	●

Benchtop Sequencer Sheds Light on Ebola Outbreak

Local scientists use the iSeq 100 Sequencing System to analyze transmission patterns and trace the origin of an Ebola outbreak in the Democratic Republic of the Congo.

Read Article ›

Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11–48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb*
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.1 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

https://www.illumina.com/systems/sequencing-platforms.html

Illumina platforms: Production-scale sequencers

NextSeq 550 Series

NextSeq 1000 & 2000

NovaSeq 6000

Popular Applications & Methods	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)			●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome & Large Panel Sequencing (enrichment-based)	●	●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	●	●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	●	●	●
Chromatin Analysis (ATAC-Seq, ChIP-Seq)	●	●	●
Methylation Sequencing	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)	●	●	●
Cell-Free Sequencing & Liquid Biopsy Analysis	●	●	●

Optimized NGS Sample Tracking and Workflows

See how a Laboratory Information Management System (LIMS) enabled this large genomics lab to standardize lab procedures and cope with increasing sample volumes from diverse clients.

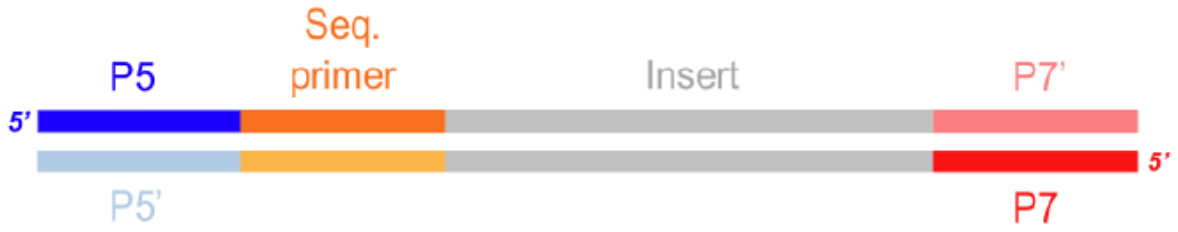
Read Case Study ›

Run Time	12–30 hours	11-48 hours	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	330 Gb*	6000 Gb
Maximum Reads Per Run	400 million	1.1 billion*	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 x 250**

https://www.illumina.com/systems/sequencing-platforms.html

Choosing a library type

- Single read library
  - Unidirectional sequencing
  - Compatible with only single-read flow cells
  - Applications: ChIP-seq, mRNA-seq for quantification, low-coverage resequencing



Choosing a library type

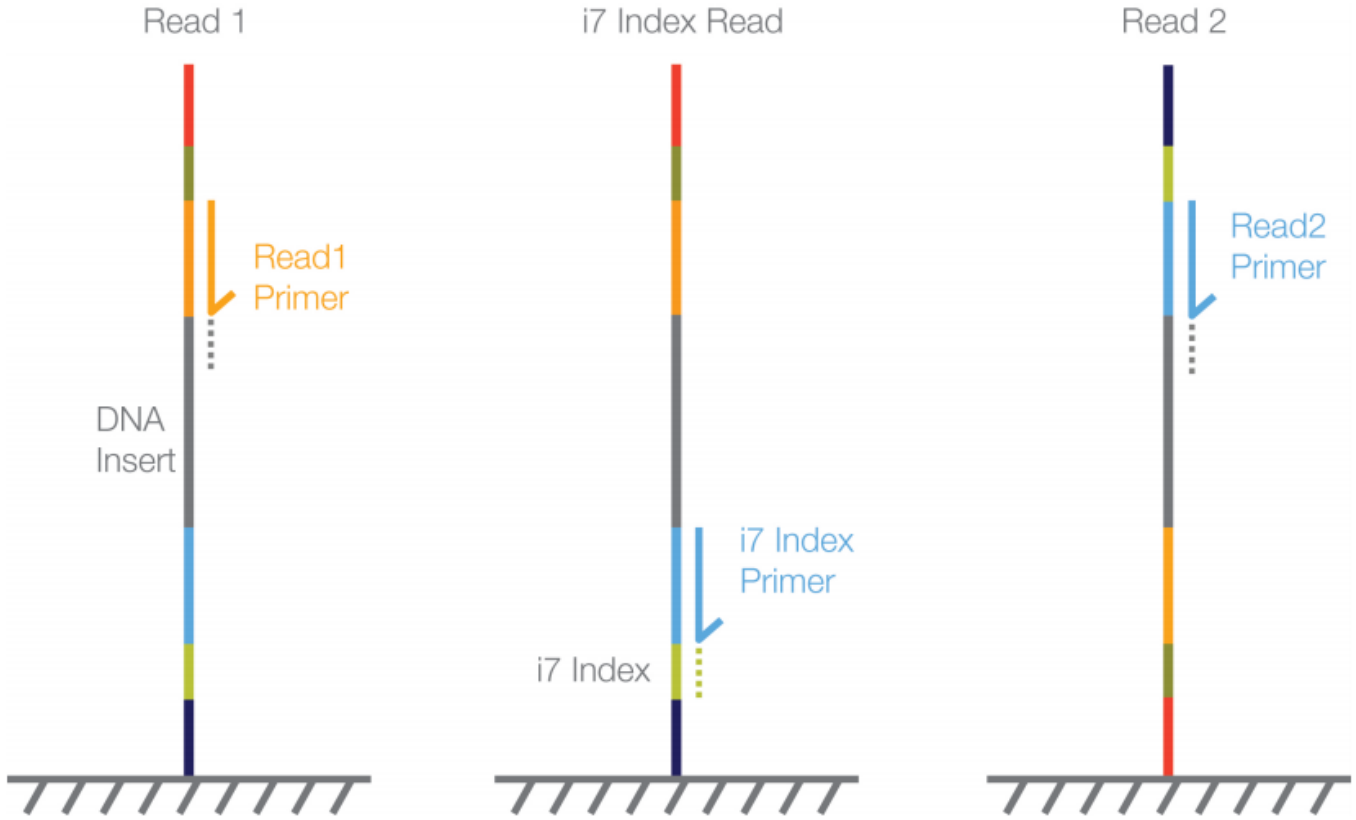
- Paired end library
  - Uni or Bidirectional sequencing
  - Compatible with both single-read and paired-end flow cells
  - Applications: the most common library type, de novo assembly, structural variants detection, high-coverage resequencing

Choosing a library type

- Indexed libraries
  - Uni or bidirectional sequencing
  - Allows multiple libraries per lane
  - Single-indexed libraries: adds up to 48 unique 6-base index 1 (i7) se quences to generate up to 48 uniquely tagged libraries.
  - Dual-indexed libraries: adds up to 24 unique 8-base index 1 (i7) sequences and up to 16 unique 8-base index 2 (i5) sequences to generate up to 384 uniquely tagged libraries.

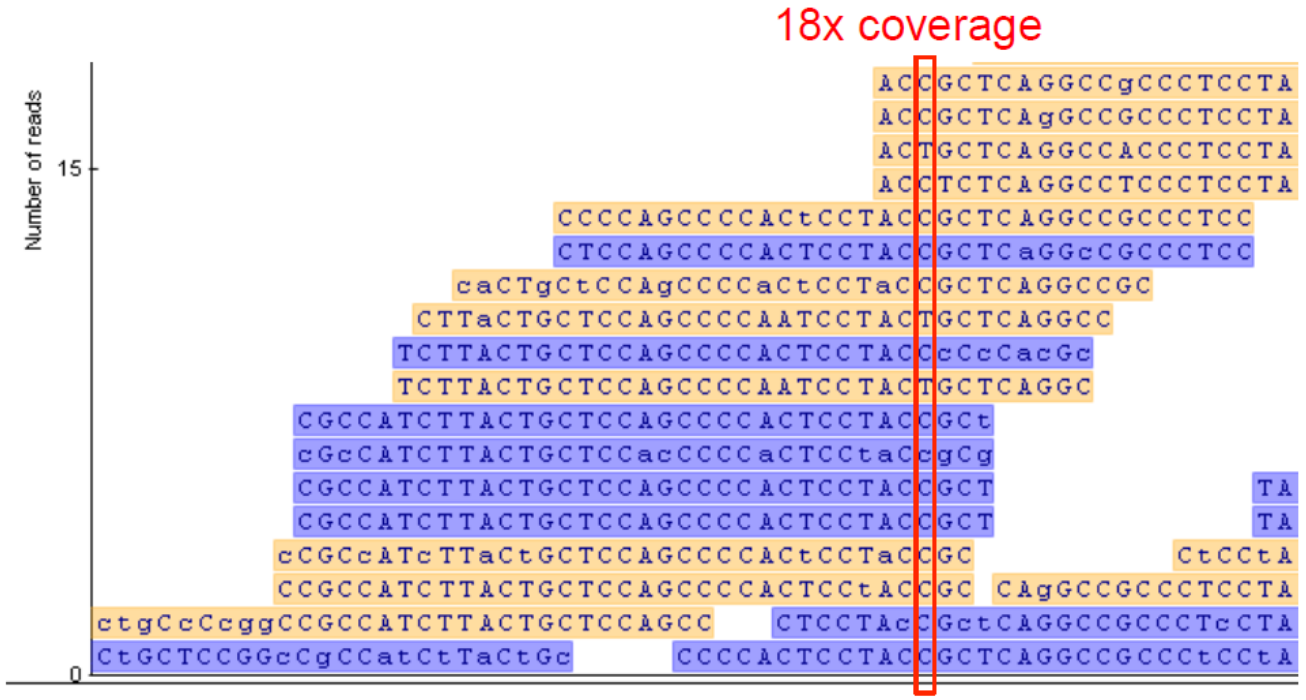
Single-indexed sequencing

The single-indexed sequencing workflow applies to all Illumina sequencing platforms.



Reads and coverage

- The number of reads for a specific region is denoted “depth” or “coverage”



partly overlapping sequencing reads result from the multiple templates being sequenced across the flow cell