BEST PRACTICES FOR RNA-SEQ DATA ANALYSIS

COURSE OBJECTIVES

Students will gain an understanding of:

- Experimental design
- Quality control
- Theoretical principles of RNA-Seq data analysis process
- How to identify differentially regulated genes
- How to identify alternative splicing events
- How to identify gene fusion events
- Multiple testing correction
- Biological interpretation of RNA-seq data

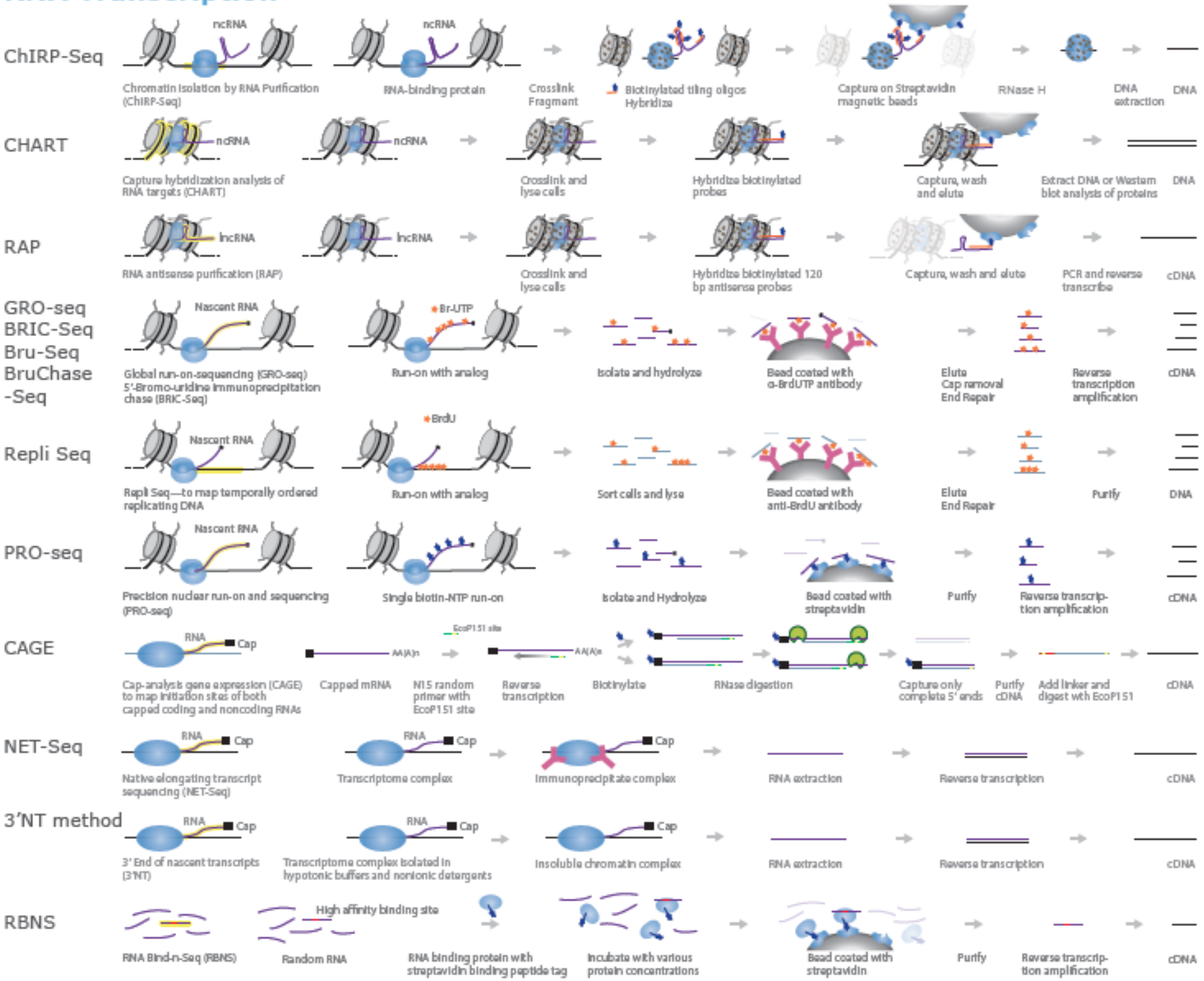
LEARNING OUTCOMES

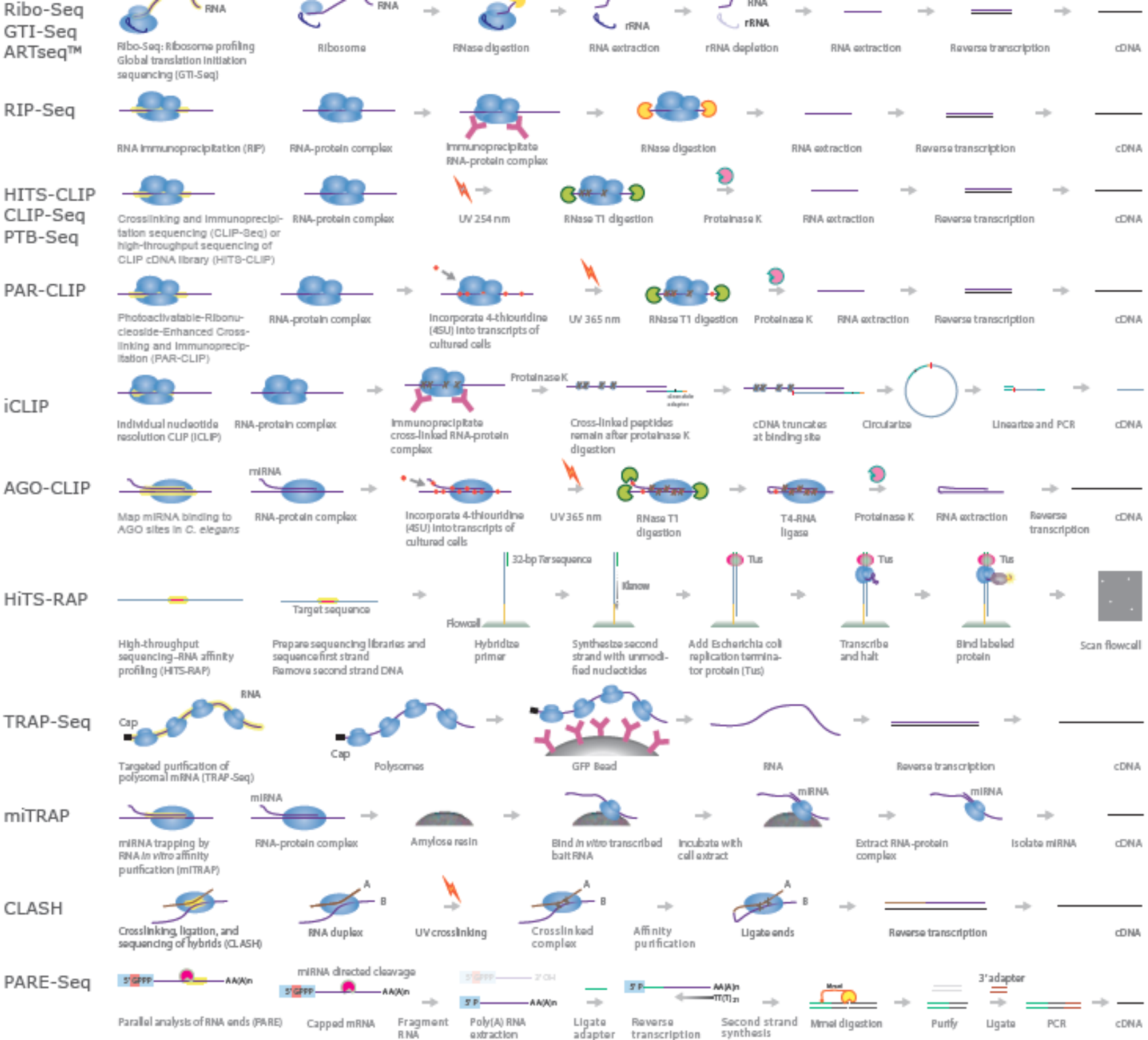
After this course participants should be able to:

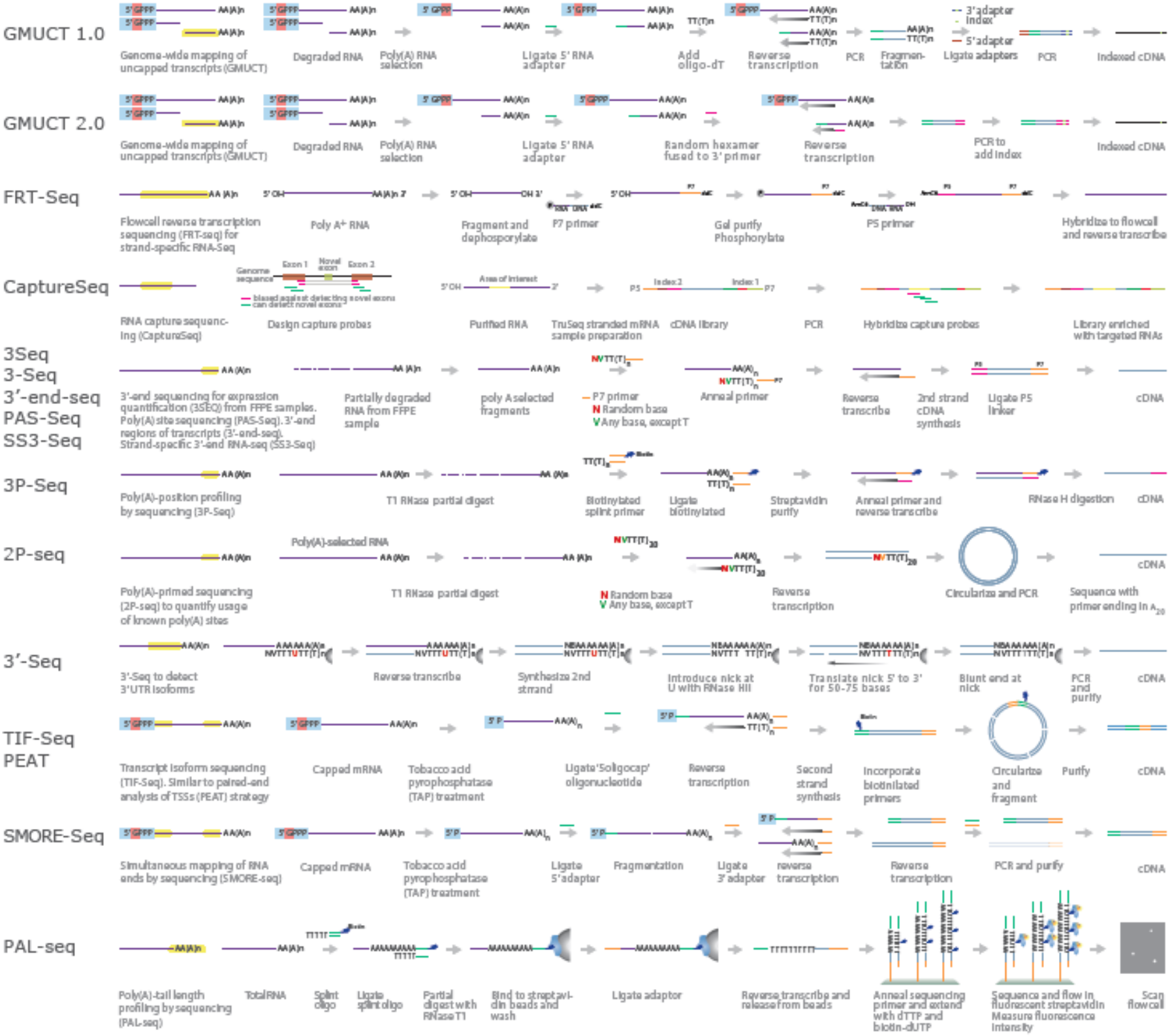
- Understand the importance of RNA-Seq experimental design
- Assess the quality of data
- Preprocess data
- Align data to a reference genome/transcriptome
- Perform a complete analysis of RNA-Seq data
- Estimate known gene and transcript expression
- Discover alternative splicing events and novel isoforms
- Discover fusion gene events
- Perform differential expression analysis
- Perform functional enrichment analysis
- Summarize and interpret the RNA-seq analysis results


RNA Transcription
RNA Low-Level Detection
RNA Modifications
RNA Structure
Rearrangements and Markers
DNA Low-Level Detection
Methylation
DNA-Protein Interactions
Protein-Protein Interactions

RNA Transcription




Ribo-Seq
GTI-Seq
ARTseq™
RNA
Ribo-Seq: Ribosome profiling
Global translation initiation
sequencing (GTI-Seq)
Ribosome
RNase digestion
RNA extraction
rRNA
rRNA depletion
RNA extraction
Reverse transcription
cDNA
RIP-Seq
RNA immunoprecipitation (RIP)
RNA-protein complex
Immunoprecipitate
RNA-protein complex
RNase digestion
RNA extraction
Reverse transcription
cDNA
HITS-CLIP
CLIP-Seq
PTB-Seq
Crosslinking and immunoprecipi-
tation sequencing (CLIP-Seq) or
high-throughput sequencing of
CLIP cDNA library (HITS-CLIP)
RNA-protein complex
UV 254 nm
RNase T1 digestion
Proteinase K
RNA extraction
Reverse transcription
cDNA
PAR-CLIP
Photoactivatable-Ribonu-
cleoside-Enhanced Cross-
linking and Immunoprecip-
itation (PAR-CLIP)
RNA-protein complex
Incorporate 4-thiouridine
(4SU) into transcripts of
cultured cells
UV 365 nm
RNase T1 digestion
Proteinase K
RNA extraction
Reverse transcription
cDNA
iCLIP
Individual nucleotide
resolution CLIP (iCLIP)
RNA-protein complex
Immunoprecipitate
cross-linked RNA-protein
complex
Proteinase K
Cross-linked peptides
remain after proteinase K
digestion
cDNA truncates
at binding site
Circularize
Linearize and PCR
cDNA
AGO-CLIP
Map miRNA binding to
AGO sites in C. elegans
miRNA
RNA-protein complex
Incorporate 4-thiouridine
(4SU) into transcripts of
cultured cells
UV 365 nm
RNase T1
digestion
T4-RNA
ligase
Proteinase K
RNA extraction
Reverse
transcription
cDNA
HiTS-RAP
High-throughput
sequencing–RNA affinity
profiling (HITS-RAP)
Target sequence
Prepare sequencing libraries and
sequence first strand
Remove second strand DNA
32-bp Tersequence
Flowcell
Hybridize
primer
Klenow
Synthesize second
strand with unmodi-
fied nucleotides
Tus
Add Escherichia coli
replication termina-
tor protein (Tus)
Tus
Transcribe
and halt
Tus
Bind labeled
protein
Scan flowcell
TRAP-Seq
RNA
Cap
Targeted purification of
polysomal mRNA (TRAP-Seq)
Cap
Polysomes
GFP Bead
RNA
Reverse transcription
cDNA
miTRAP
miRNA trapping by
RNA in vitro affinity
purification (miTRAP)
miRNA
RNA-protein complex
Amylose resin
Bind in vitro transcribed
bait RNA
Incubate with
cell extract
miRNA
Extract RNA-protein
complex
miRNA
Isolate miRNA
cDNA
CLASH
Crosslinking, ligation, and
sequencing of hybrids (CLASH)
A
B
RNA duplex
UV crosslinking
A
B
Crosslinked
complex
Affinity
purification
A
B
Ligate ends
Reverse transcription
cDNA
PARE-Seq
5'GPPP
AA(A)n
Parallel analysis of RNA ends (PARE)
miRNA directed cleavage
5'GPPP
AA(A)n
Capped mRNA
Fragment
RNA
5'P
AA(A)n
Poly(A) RNA
extraction
Ligate
adapter
5'P
AA(A)n
TT(T)21
Reverse
transcription
Second strand
synthesis
MmeI
MmeI digestion
Purify
3'adapter
Ligate
PCR
cDNA


GMUCT 1.0
Genome-wide mapping of uncapped transcripts (GMUCT)
Degraded RNA
Poly(A) RNA selection
Ligate 5' RNA adapter
Add oligo-dT
Reverse transcription
PCR
Fragmentation
3' adapter
Index
5' adapter
Ligate adapters
PCR
Indexed cDNA
GMUCT 2.0
Genome-wide mapping of uncapped transcripts (GMUCT)
Degraded RNA
Poly(A) RNA selection
Ligate 5' RNA adapter
Random hexamer fused to 3' primer
Reverse transcription
PCR to add index
Indexed cDNA
FRT-Seq
Flowcell reverse transcription sequencing (FRT-seq) for strand-specific RNA-Seq
Poly A+ RNA
Fragment and dephosporylate
P7 primer
Gel purify Phosphorylate
P5 primer
Hybridize to flowcell and reverse transcribe
CaptureSeq
RNA capture sequencing (CaptureSeq)
Design capture probes
Purified RNA
TruSeq stranded mRNA sample preparation
cDNA library
PCR
Hybridize capture probes
Library enriched with targeted RNAs
3Seq
3-Seq
3'-end-seq
PAS-Seq
SS3-Seq
3'-end sequencing for expression quantification (3SEQ) from FFPE samples. Poly(A) site sequencing (PAS-Seq). 3'-end regions of transcripts (3'-end-seq). Strand-specific 3'-end RNA-seq (SS3-Seq)
Partially degraded RNA from FFPE sample
poly A selected fragments
P7 primer
N Random base
V Any base, except T
Anneal primer
Reverse transcribe
2nd strand cDNA synthesis
Ligate P5 linker
cDNA
3P-Seq
Poly(A)-position profiling by sequencing (3P-Seq)
T1 RNase partial digest
Biotinylated splint primer
Ligate biotinylated
Streptavidin purify
Anneal primer and reverse transcribe
RNase H digestion
cDNA
2P-seq
Poly(A)-primed sequencing (2P-seq) to quantify usage of known poly(A) sites
Poly(A)-selected RNA
T1 RNase partial digest
N Random base
V Any base, except T
Reverse transcription
Circularize and PCR
Sequence with primer ending in A20
cDNA
3'-Seq
3'-Seq to detect 3'UTR isoforms
Reverse transcribe
Synthesize 2nd strand
Introduce nick at U with RNase HII
Translate nick 5' to 3' for 50-75 bases
Blunt end at nick
PCR and purify
cDNA
TIF-Seq
PEAT
Transcript isoform sequencing (TIF-Seq). Similar to paired-end analysis of TSSs (PEAT) strategy
Capped mRNA
Tobacco acid pyrophosphatase (TAP) treatment
Ligate 'Soligocap' oligonucleotide
Reverse transcription
Second strand synthesis
Incorporate biotinilated primers
Circularize and fragment
Purify
cDNA
SMORE-Seq
Simultaneous mapping of RNA ends by sequencing (SMORE-seq)
Capped mRNA
Tobacco acid pyrophosphatase (TAP) treatment
Ligate 5' adapter
Fragmentation
Ligate 3' adapter
reverse transcription
Reverse transcription
PCR and purify
cDNA
PAL-seq
Poly(A)-tail length profiling by sequencing (PAL-seq)
Total RNA
Splint oligo
Ligate splint oligo
Partial digest with RNase T1
Bind to streptavidin beads and wash
Ligate adaptor
Reverse transcribe and release from beads
Anneal sequencing primer and extend with dTTP and biotin-dUTP
Sequence and flow in fluorescent streptavidin Measure fluorescence intensity
Scan flowcell

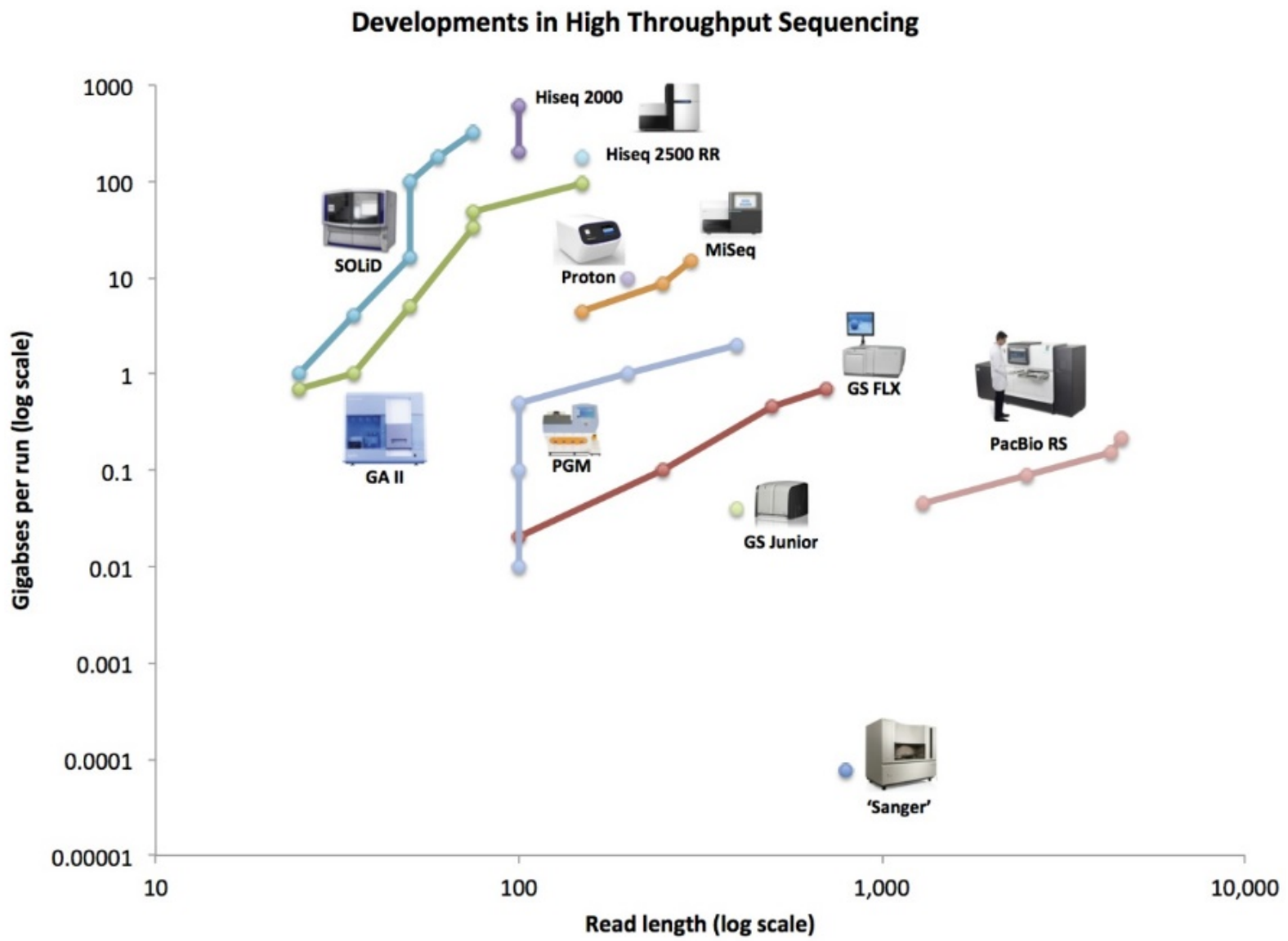
NEXT GENERATION SEQUENCING

In this poster, 108 different methods are described:

- Transcription: 33 variants + 11 for lowly expressed RNAs
- RNA modifications: 4 variants
- RNA structure: 9 variants
- Geomic DNA: 4 variants
- Genomic structural variants: 6 variants
- DNA/RNA methylation: 22 variants
- Protein/DNA interaction: 17 variants

And many more have been described in the scientific literature

NEXT GENERATION SEQUENCING

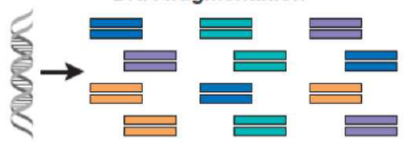


NEXT GENERATION SEQUENCING

Sanger Sequencing

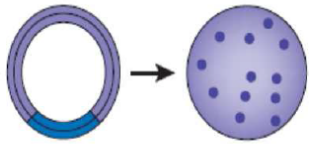
a

DNA fragmentation



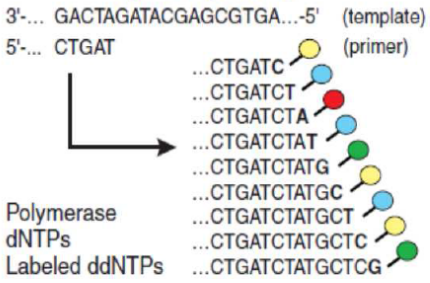
DNA is fragmented

In vivo cloning and amplification



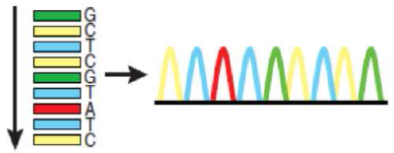
Cloned to a plasmid vector

Cycle sequencing



Cyclic sequencing reaction

Electrophorsesis
(1 read/capillary)



Separation by electrophoresis
Readout with fluorescent tags

NGS: Massive Parallel Sequencing

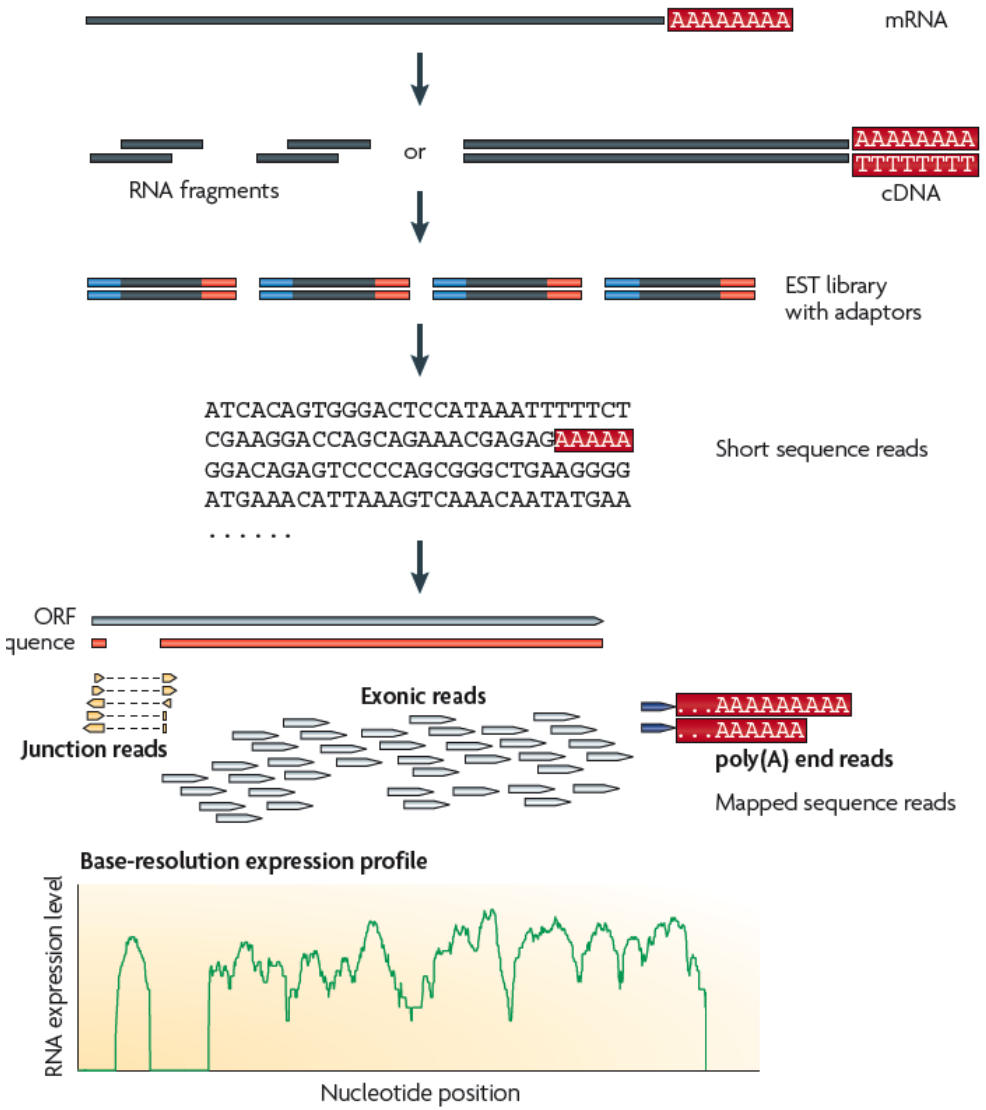
DNA is fragmented

Adaptors ligated to fragments (**Library construction**)

Clonal amplification of fragments on a solid surface (**Bridge PCR** or **Emulsion PCR**)

Direct step-by-step detection of each nucleotide base incorporated during the sequencing reaction

RNA-SEQ



- High-throughput sequencing techniques can be applied to RNAs extracted from a sample, converted to cDNA and fragmented.
- It‘s not currently possible to sequence all RNAs from a sample, but only a (large) subset of them.
- The assumption is that this subset is representative of the population.
- The more a gene is transcribed, the more of its RNAs are present in the subset, and the more will be sequenced.
- Hence, the number of sequencing reads from a given gene is related to its expression level.

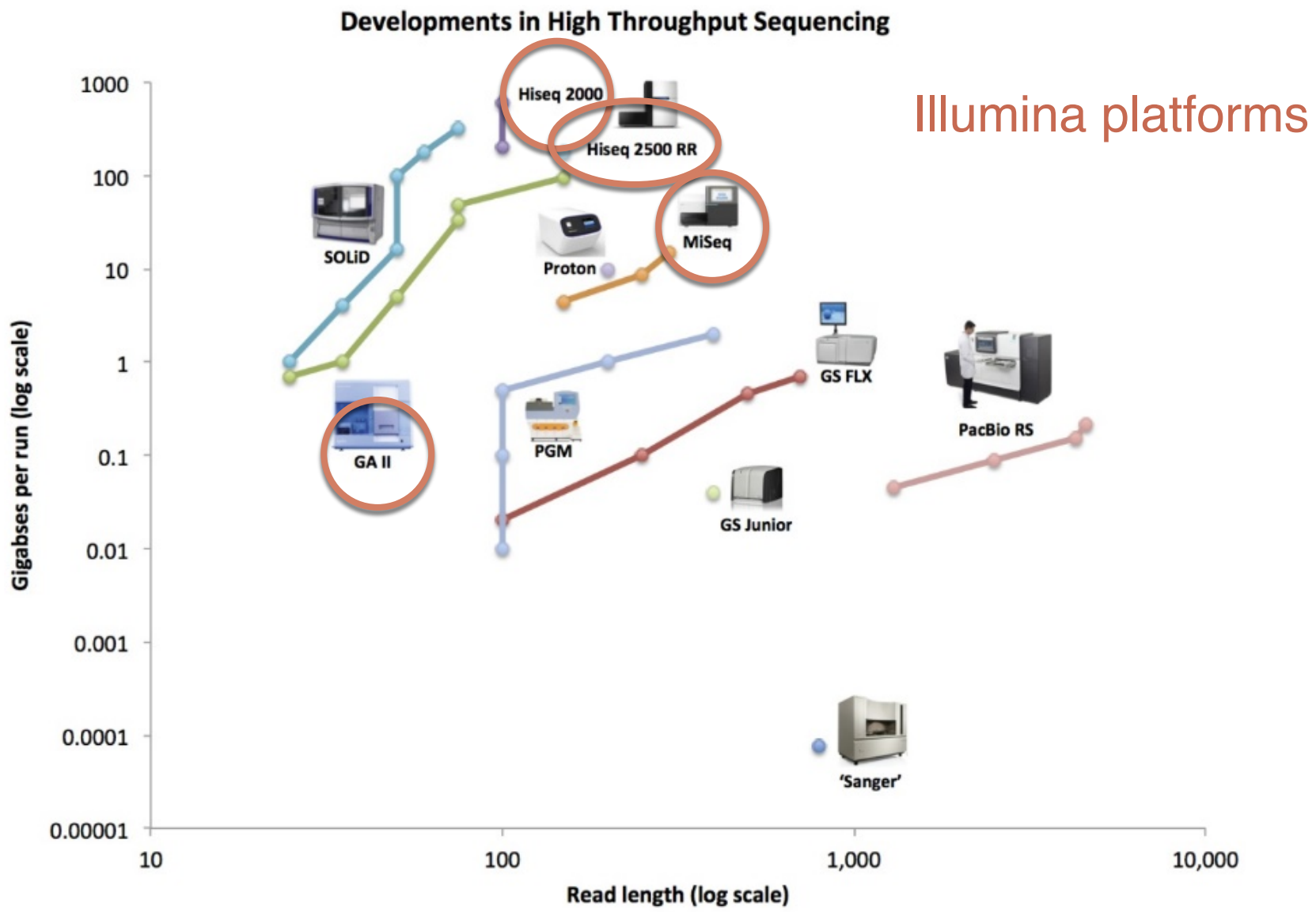
RNA-SEQ

Applications:

- Gene expression levels, differential expression, identification and expression estimate of different RNA sub-classes (nuclear, cytoplasmic, poly-adenylated, small, associated to ribosomes, transcribed from a specific strand)
- Splicing variants expression levels, identification of novel splicing variants
- Identification of novel genes
- Identification of gene fusion events
- Identification of SNPs and small insertions/deletions
- Trans-splicing
- Circular RNAs
- Protein-RNA interactions
- RNA editing and methylation

Martin and Wang Nat. Rev. Genet. (2011)

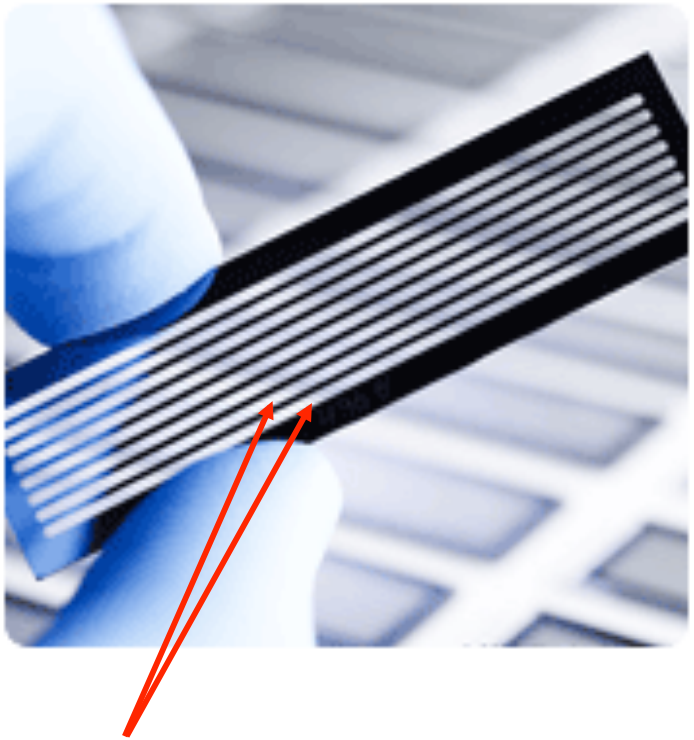
RNA-SEQ



Illumina platforms

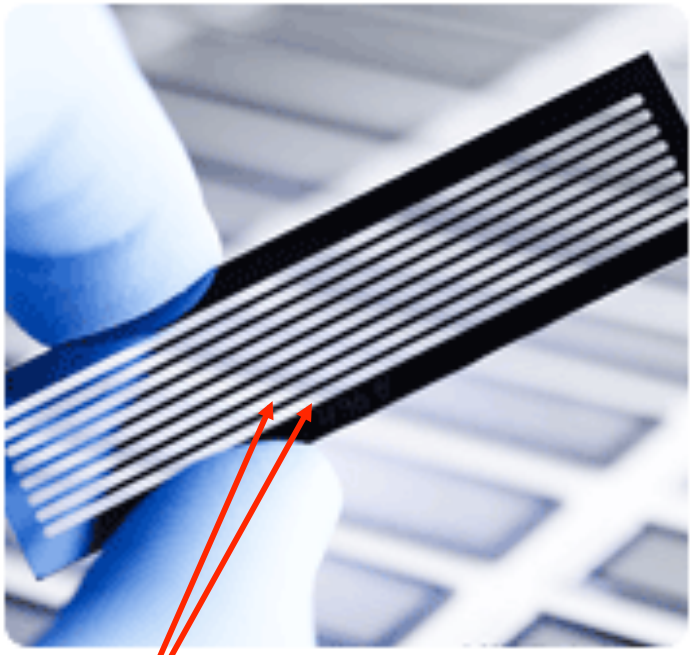
ILLUMINA SEQUENCING

Illumina Genome Analyzer Flow cell

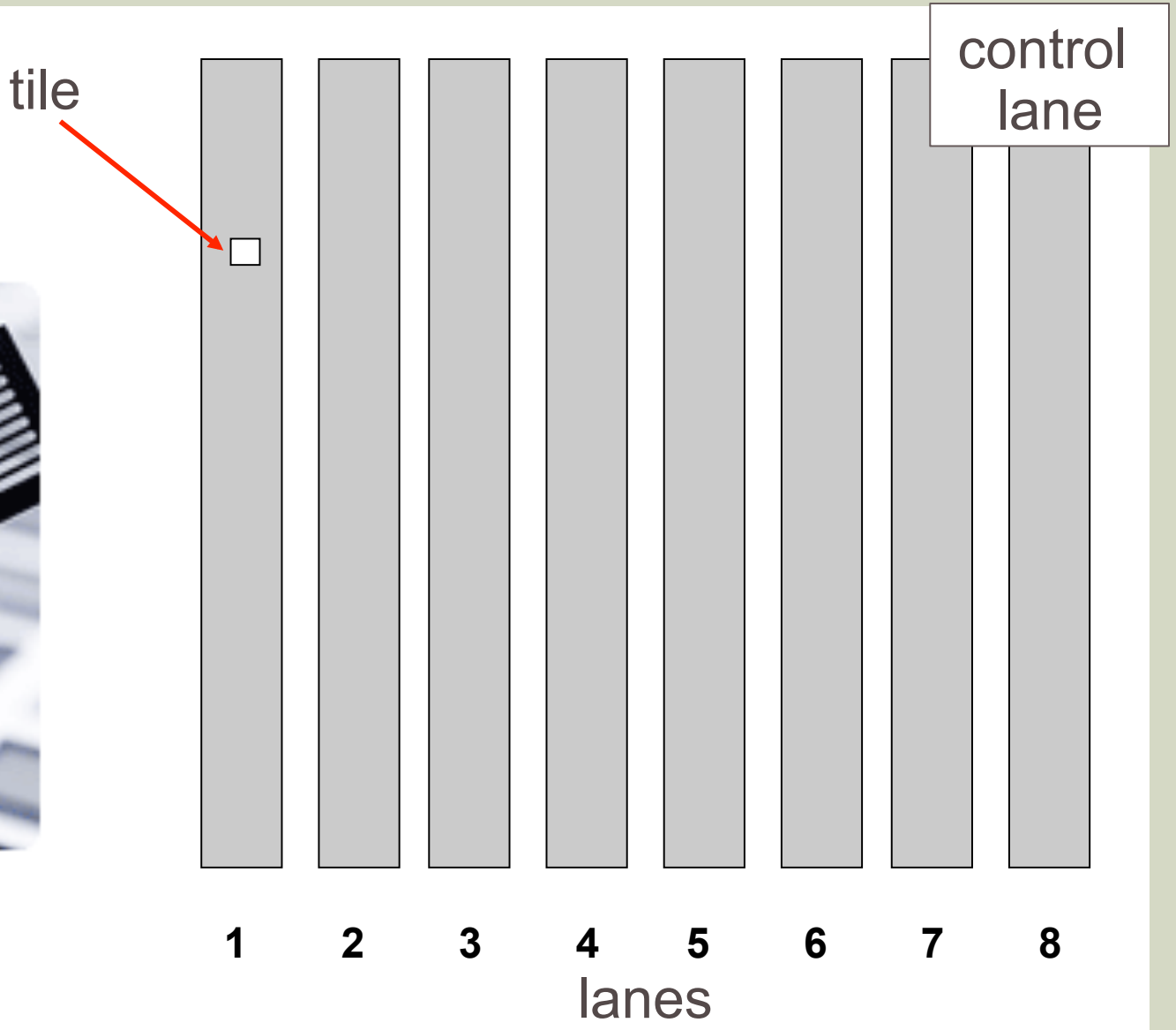


- Divided into 8 channels (lanes)
- Each channel can be loaded with multiple samples, each identified by a different tag (multiplexing)
- Input: 0.1–1.0 µg;
- Output: Hundreds of millions (or more) of reads (clusters) for flow cell
- Each cluster contains ~1,000 copies of the same template

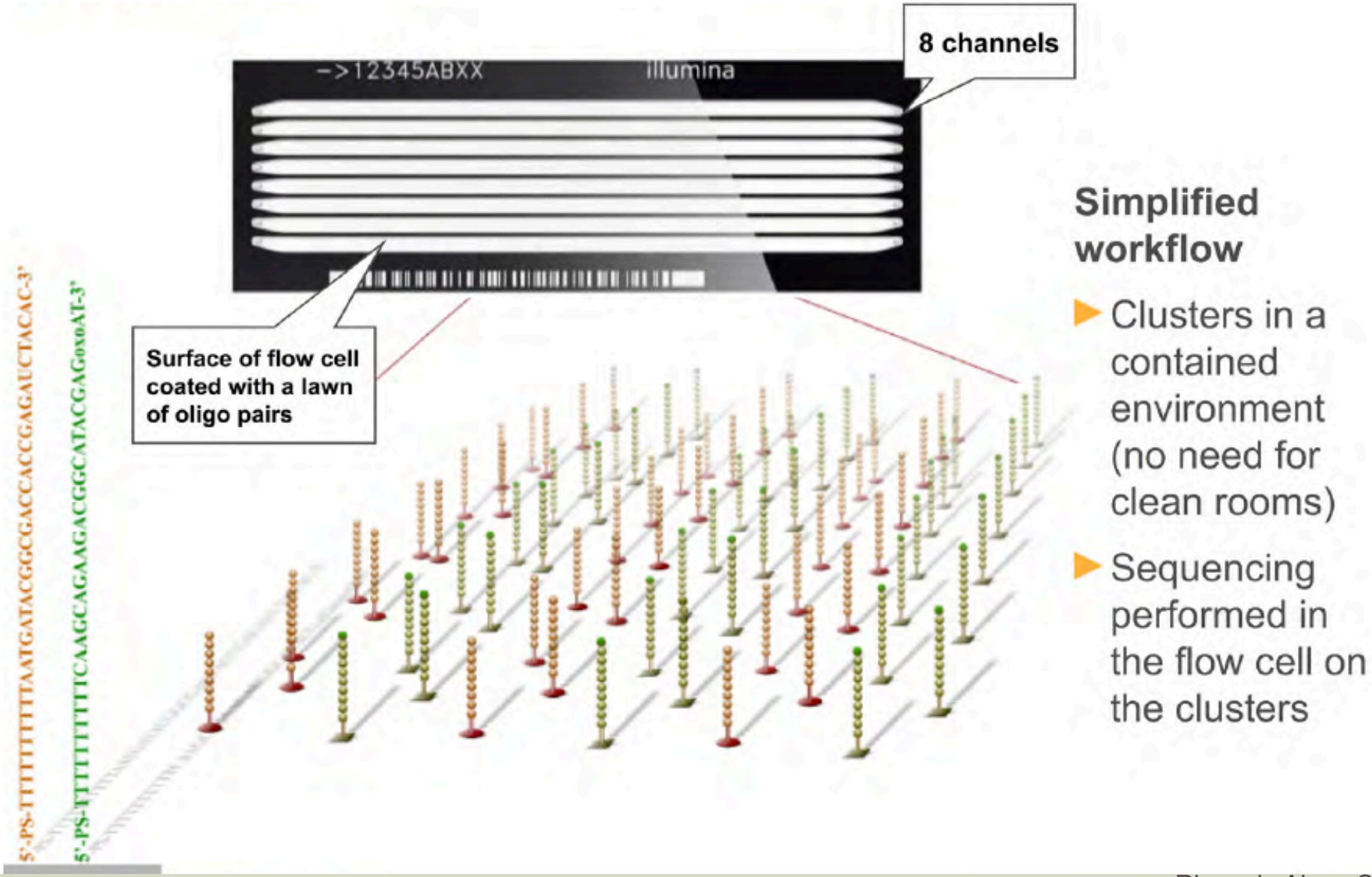
ILLUMINA SEQUENCING



lanes



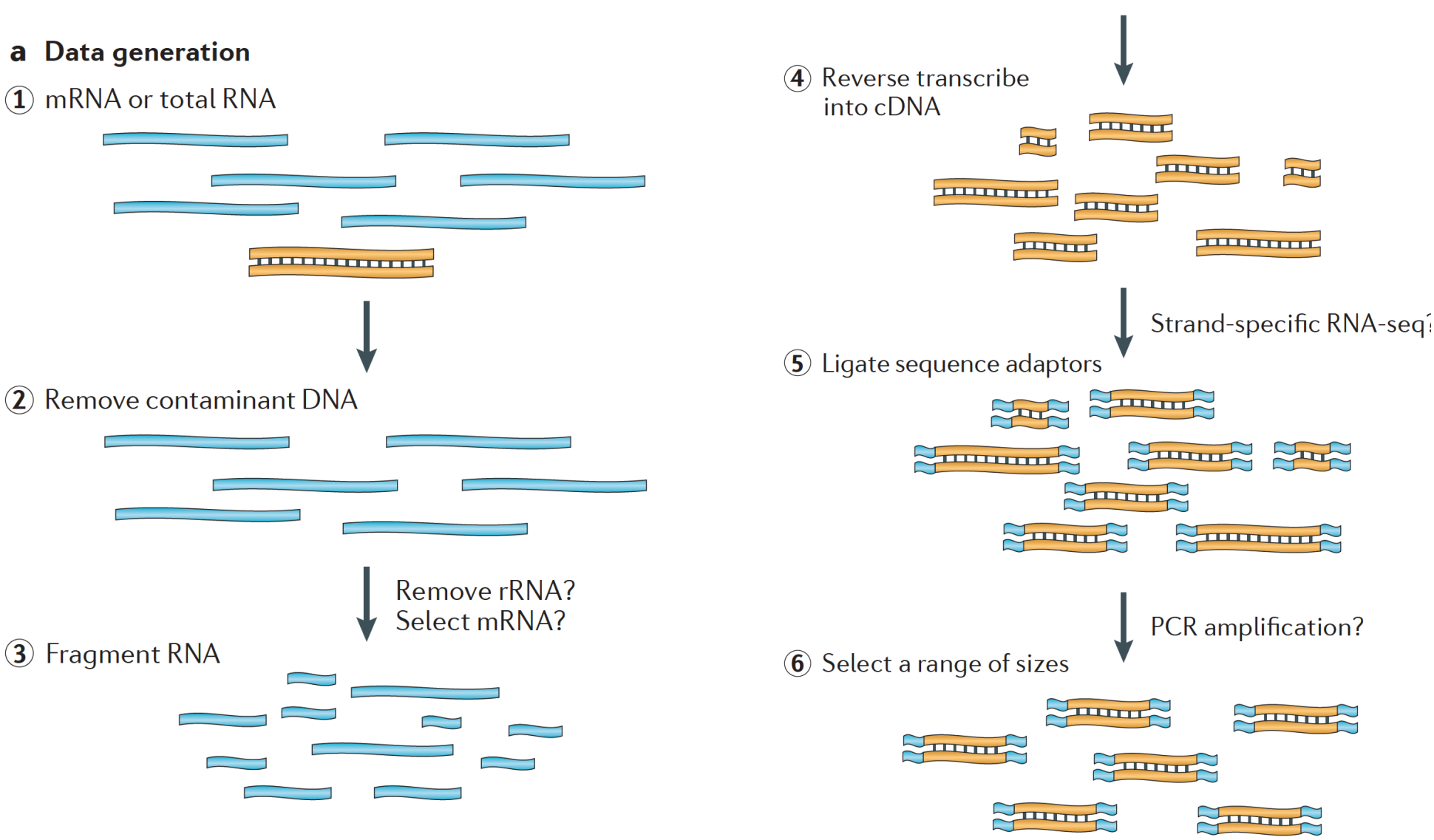
ILLUMINA SEQUENCING



Simplified workflow

- Clusters in a contained environment (no need for clean rooms)
- Sequencing performed in the flow cell on the clusters

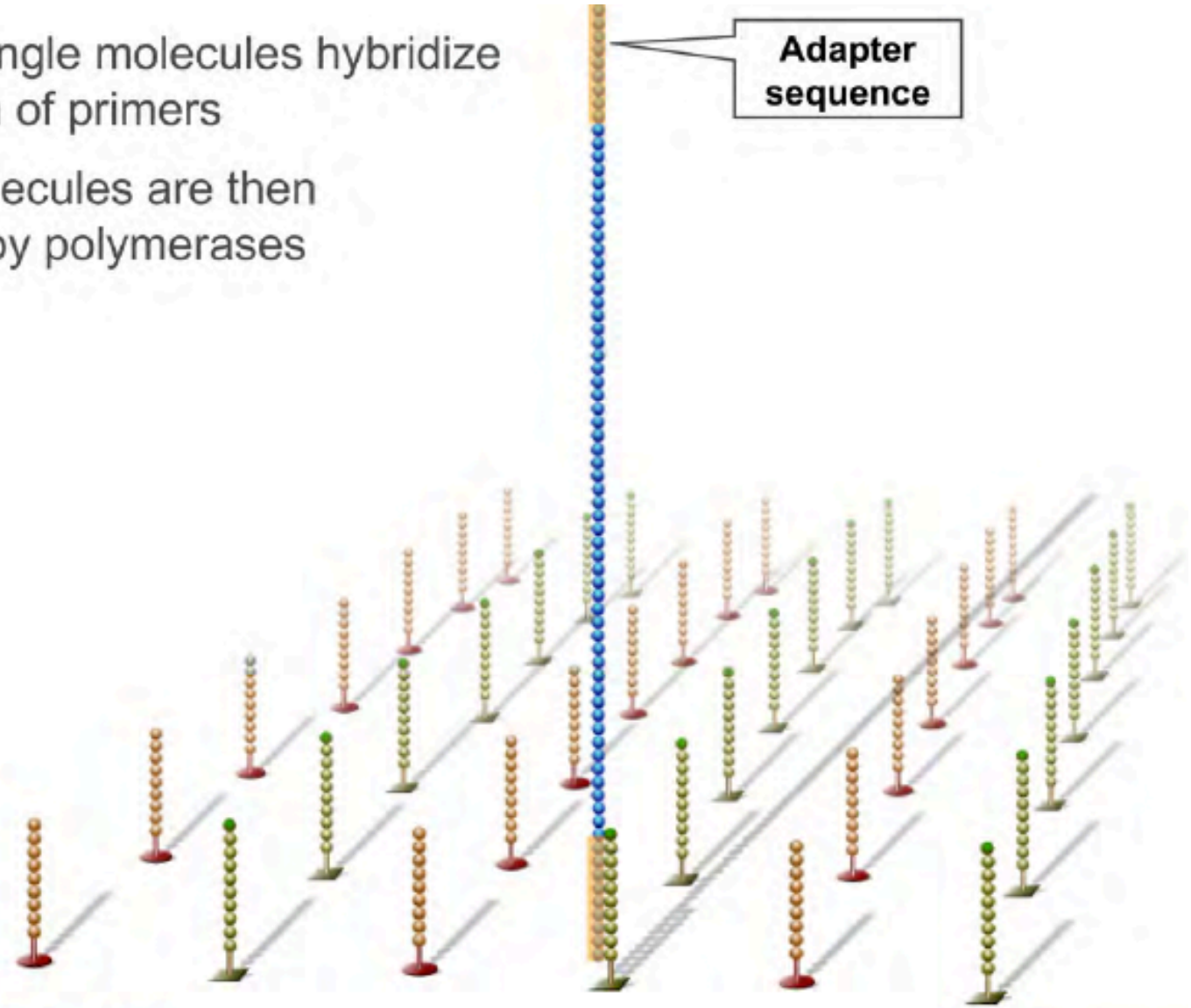
ILLUMINA SEQUENCING



Martin and Wang Nat. Rev. Genet. (2011)

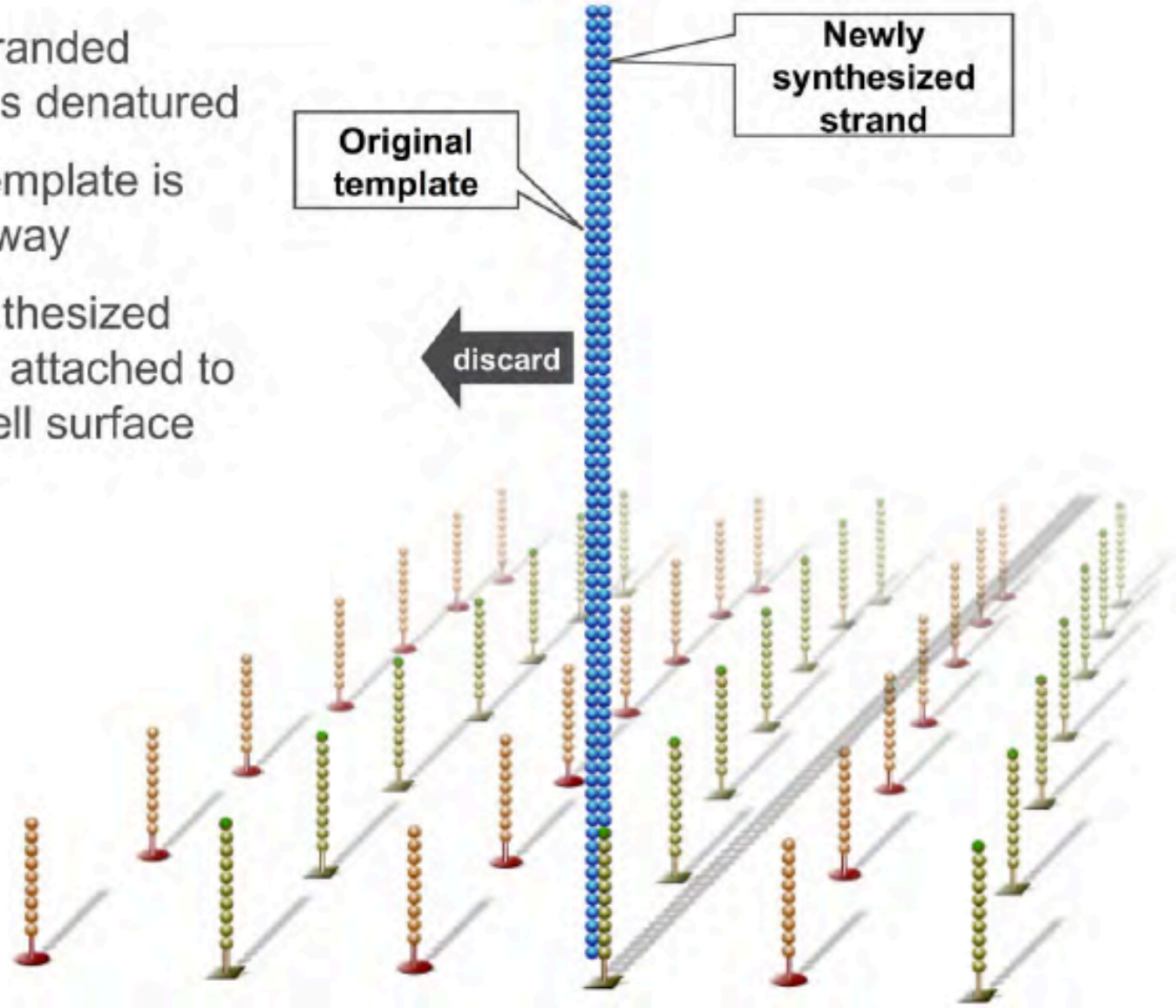
ILLUMINA SEQUENCING

- > 100 M single molecules hybridize to the lawn of primers
- Bound molecules are then extended by polymerases

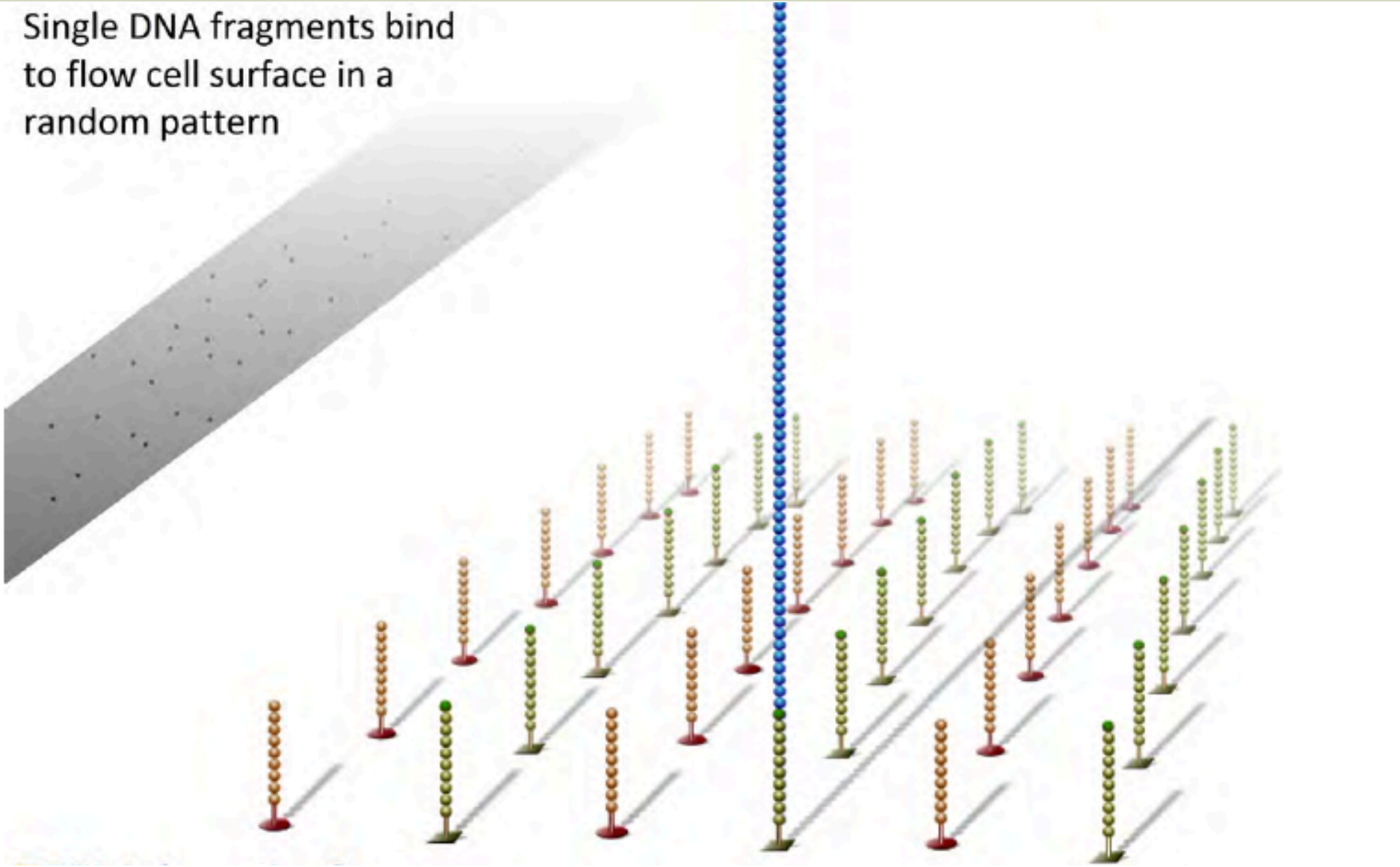


ILLUMINA SEQUENCING

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

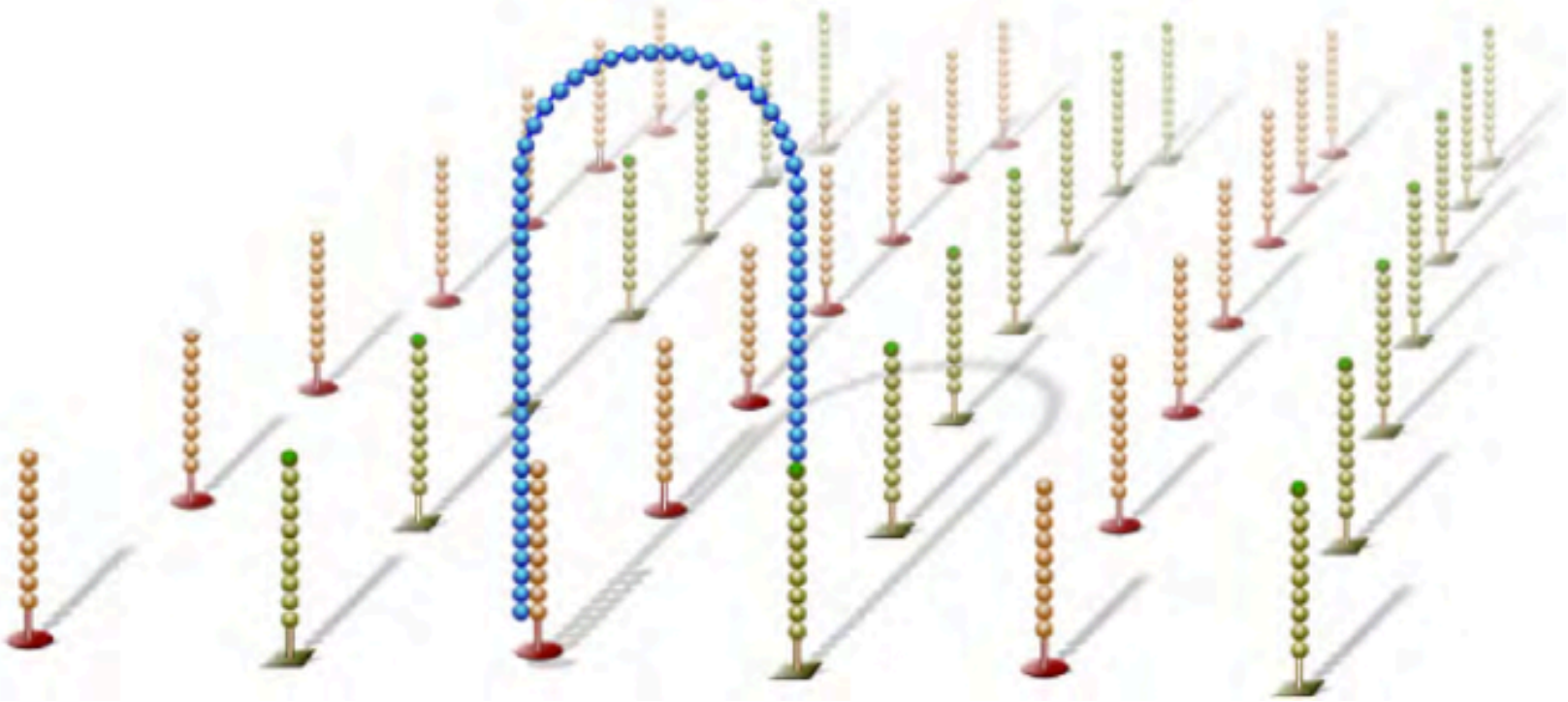


ILLUMINA SEQUENCING



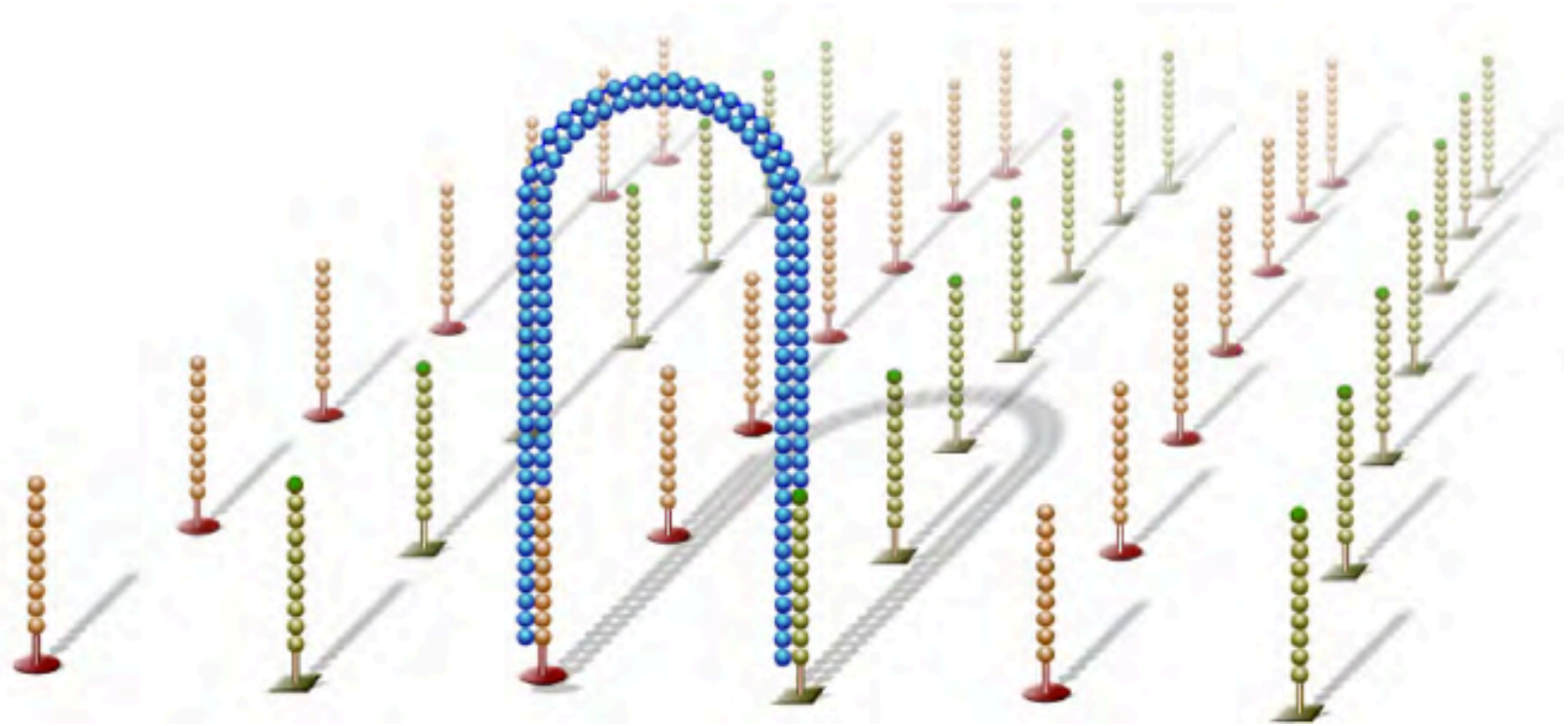
ILLUMINA SEQUENCING

- Single-strand flips over to hybridize to adjacent primers to form a bridge
- Hybridized primer is extended by polymerases



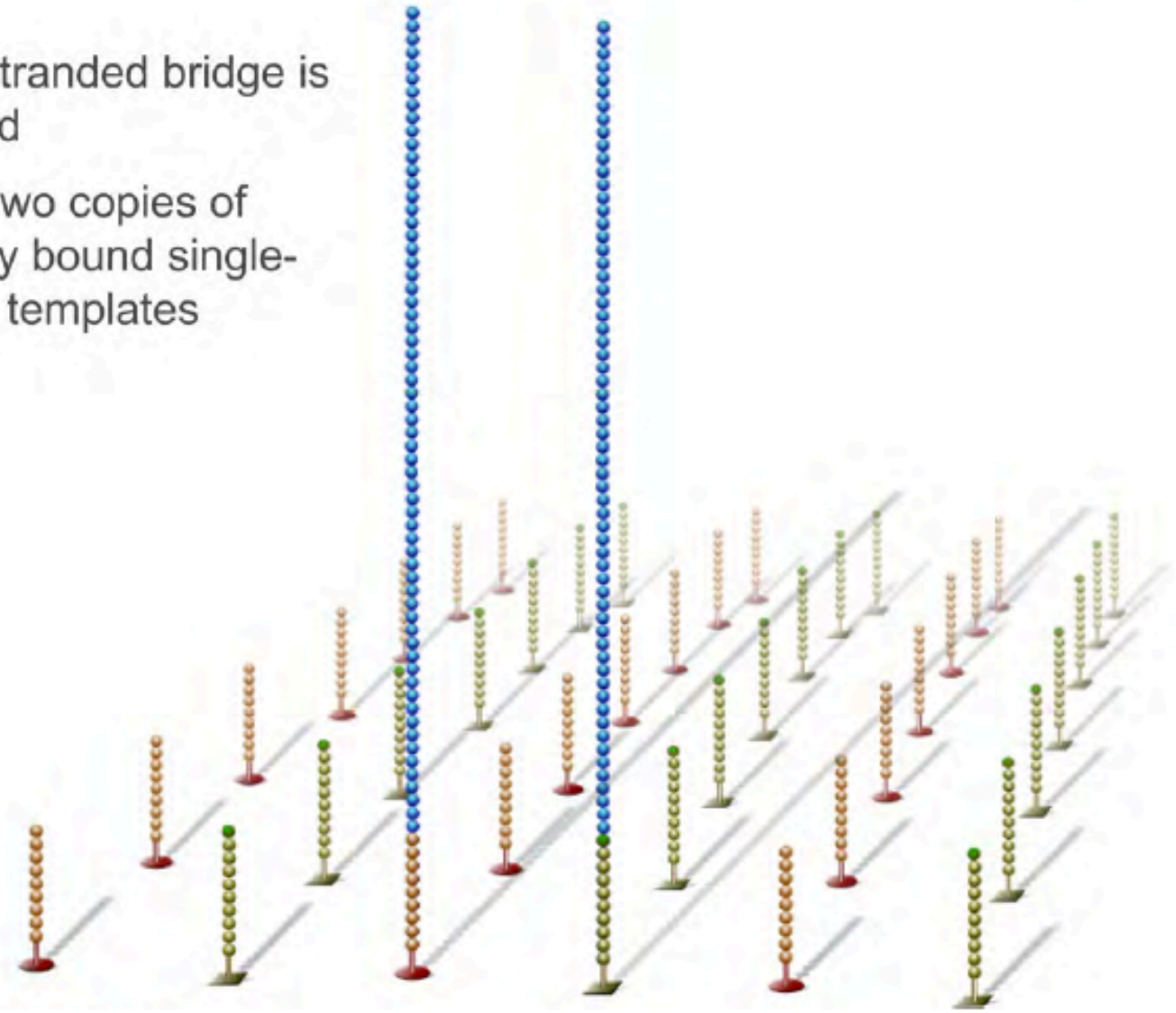
ILLUMINA SEQUENCING

- Double-stranded bridge is formed



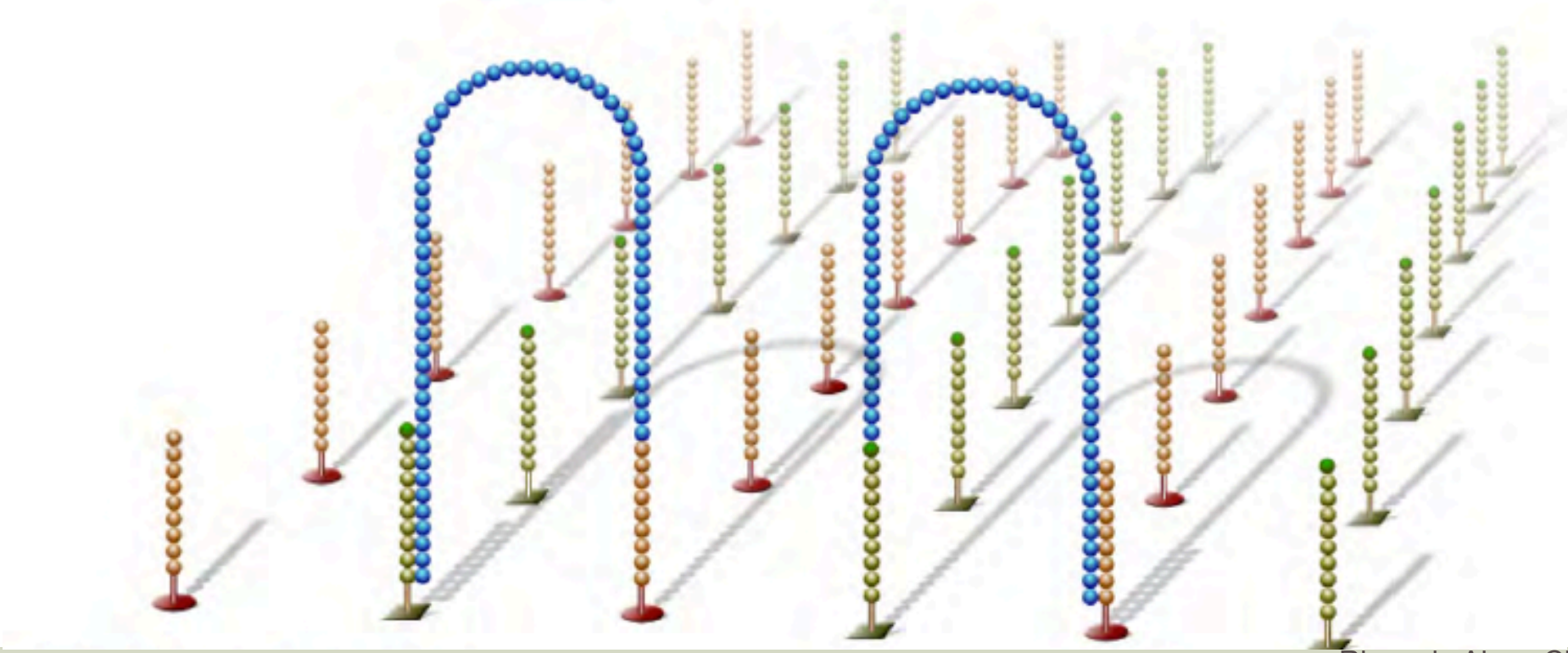
ILLUMINA SEQUENCING

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



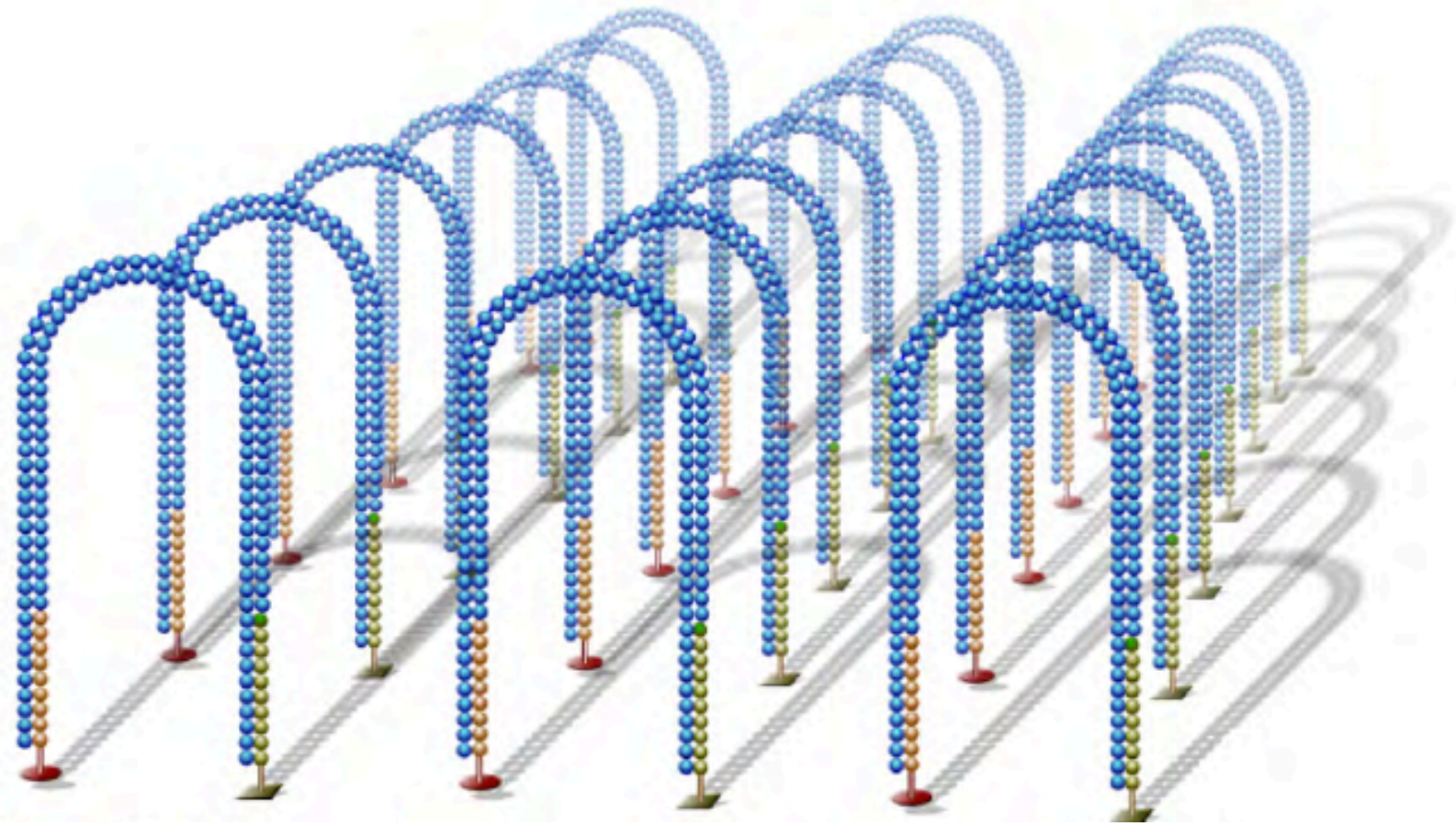
ILLUMINA SEQUENCING

- Single-strands flip over to hybridize to adjacent primers to form bridges
- Hybridized primer is extended by polymerase

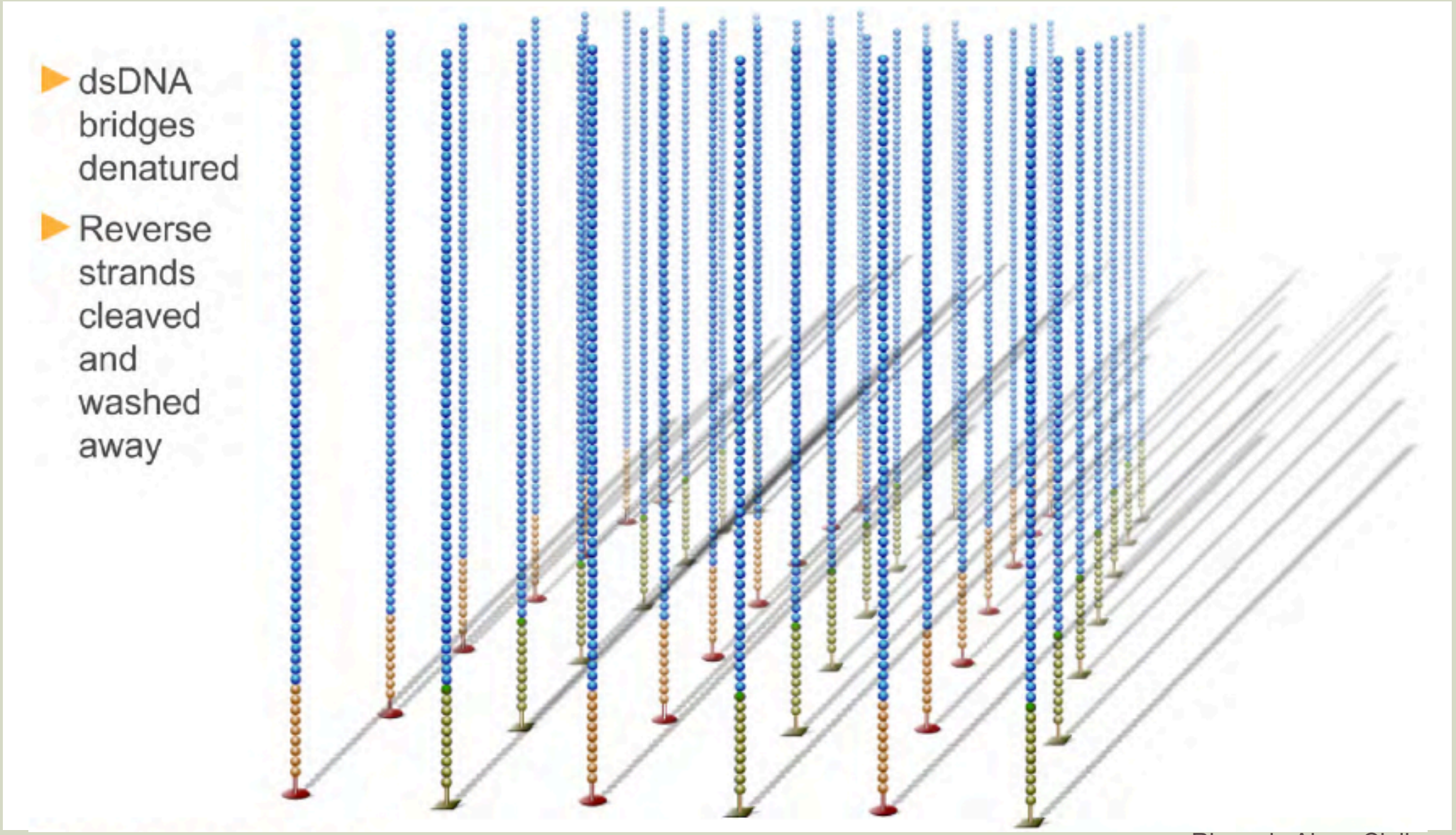


ILLUMINA SEQUENCING

- Bridge amplification cycle repeated until multiple bridges are formed

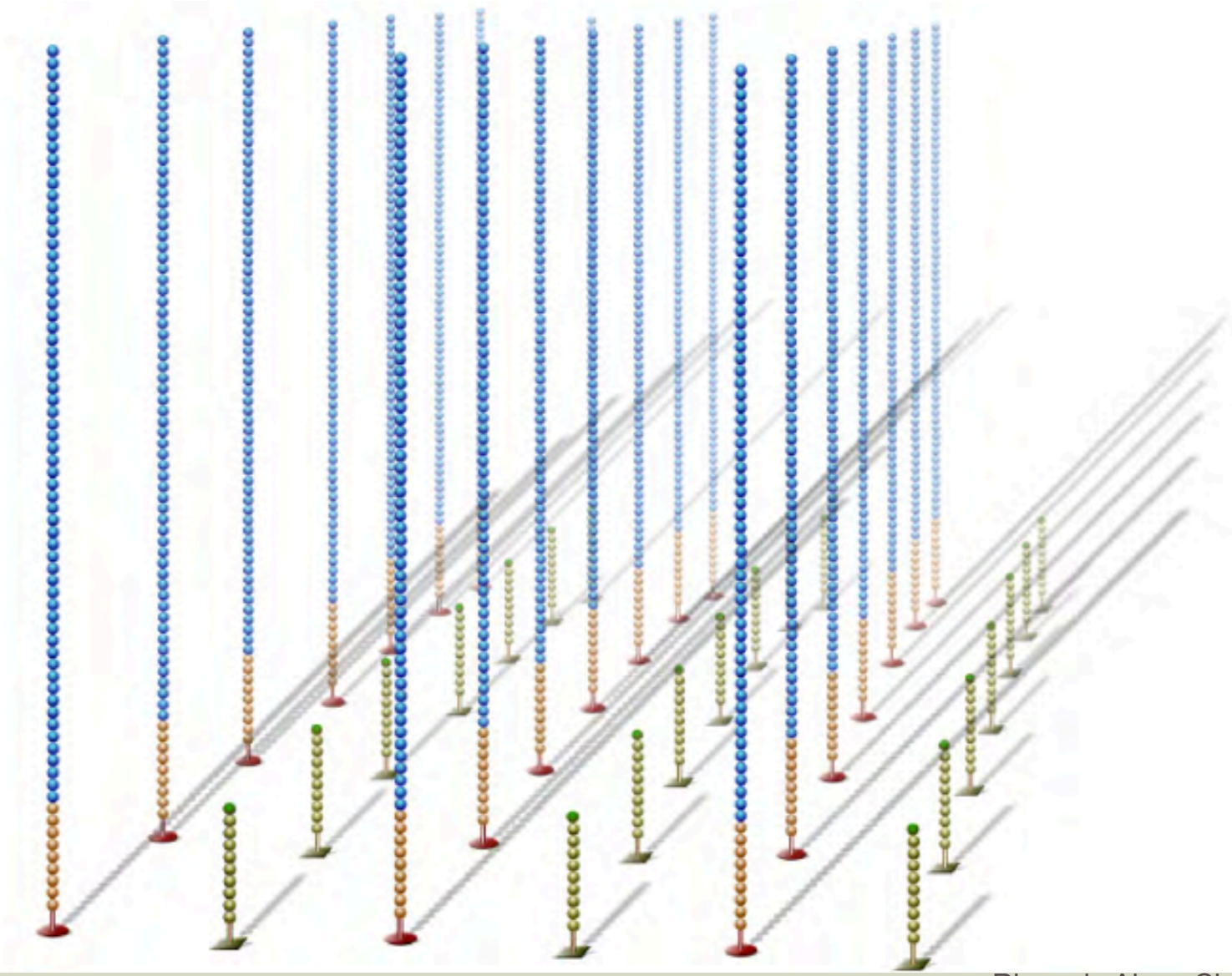


ILLUMINA SEQUENCING



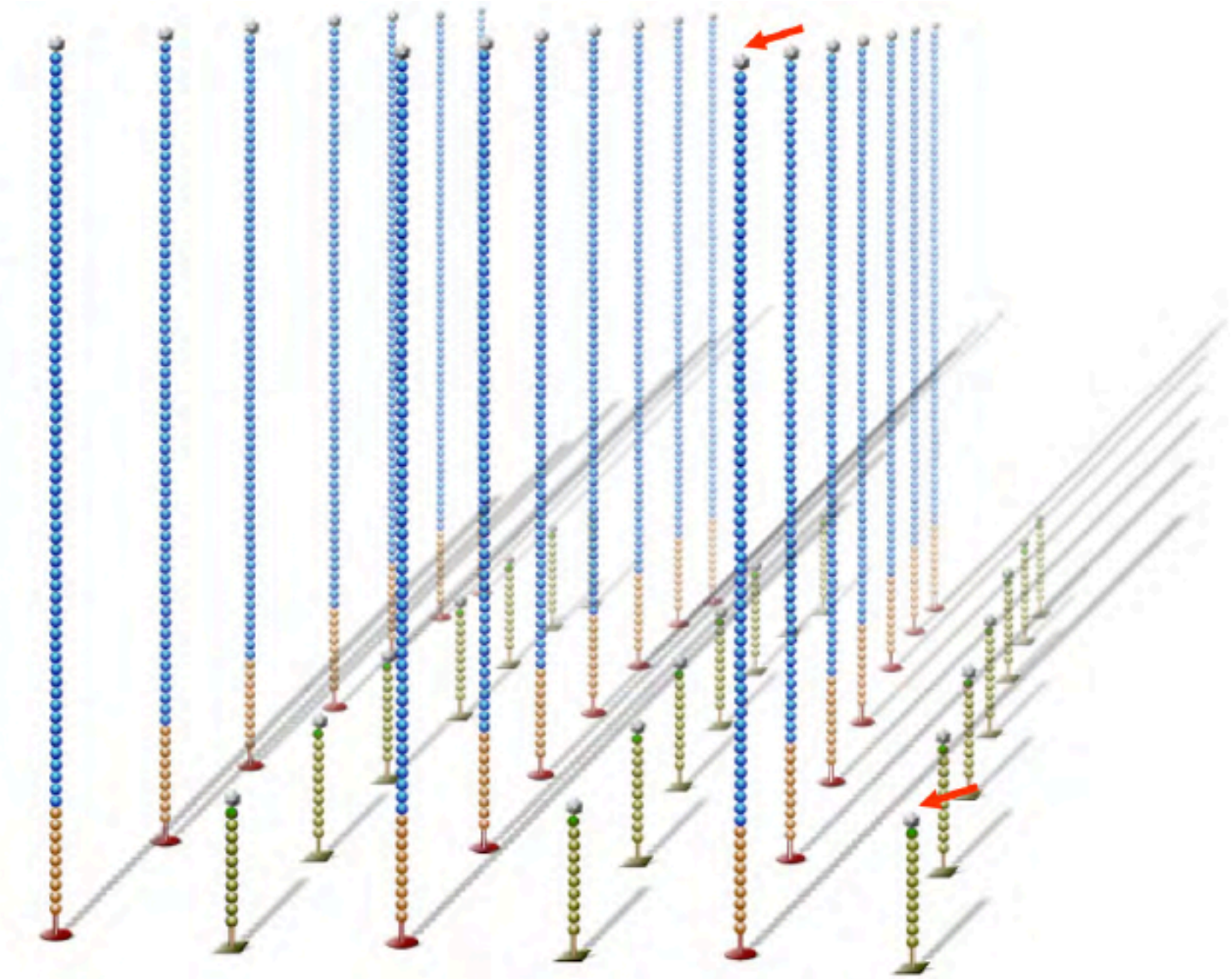
ILLUMINA SEQUENCING

- …leaving a cluster with forward strands only



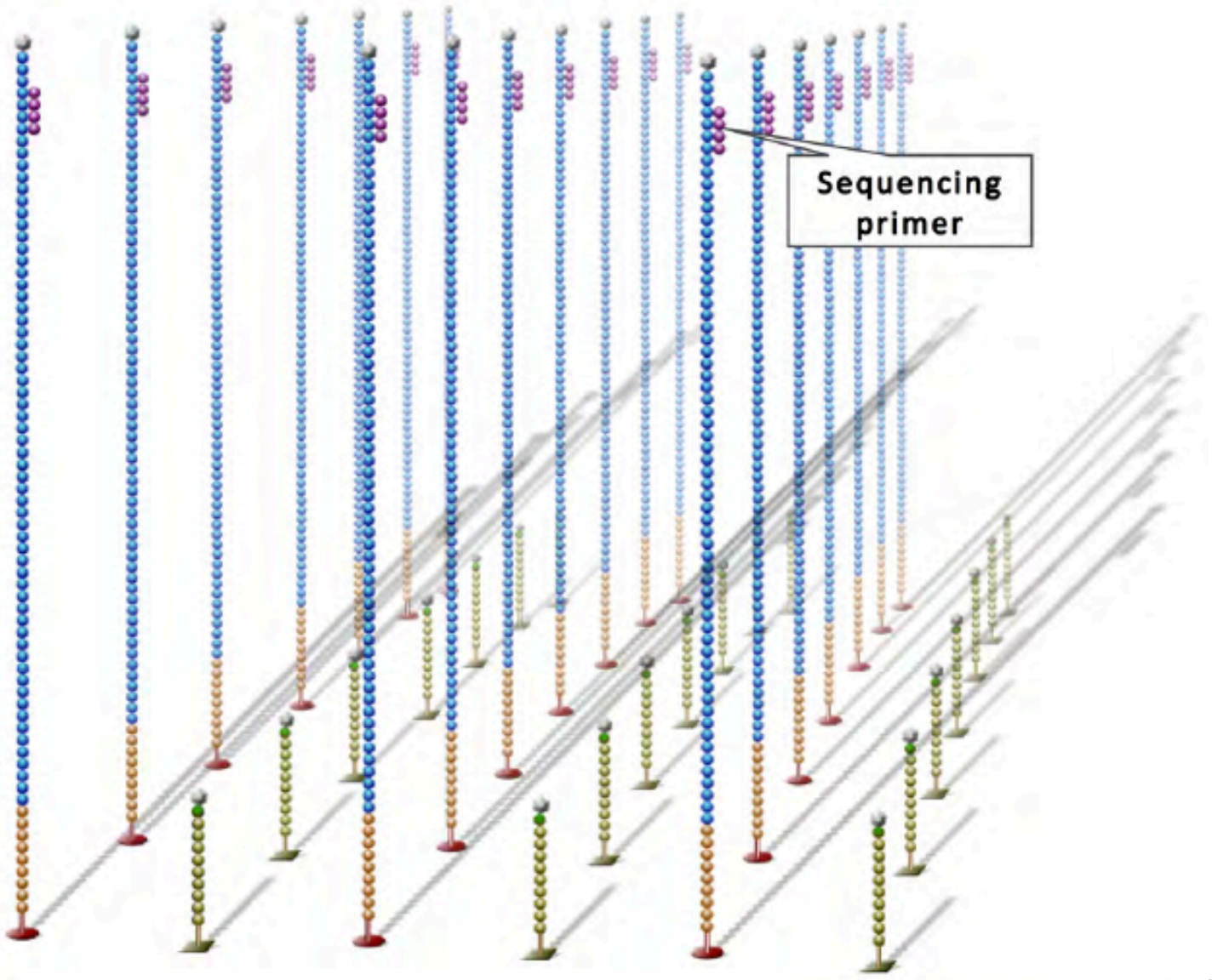
ILLUMINA SEQUENCING

- Free 3' ends are blocked to prevent unwanted DNA priming



ILLUMINA SEQUENCING

- Sequencing primer is hybridized to adapter sequence



ILLUMINA SEQUENCING

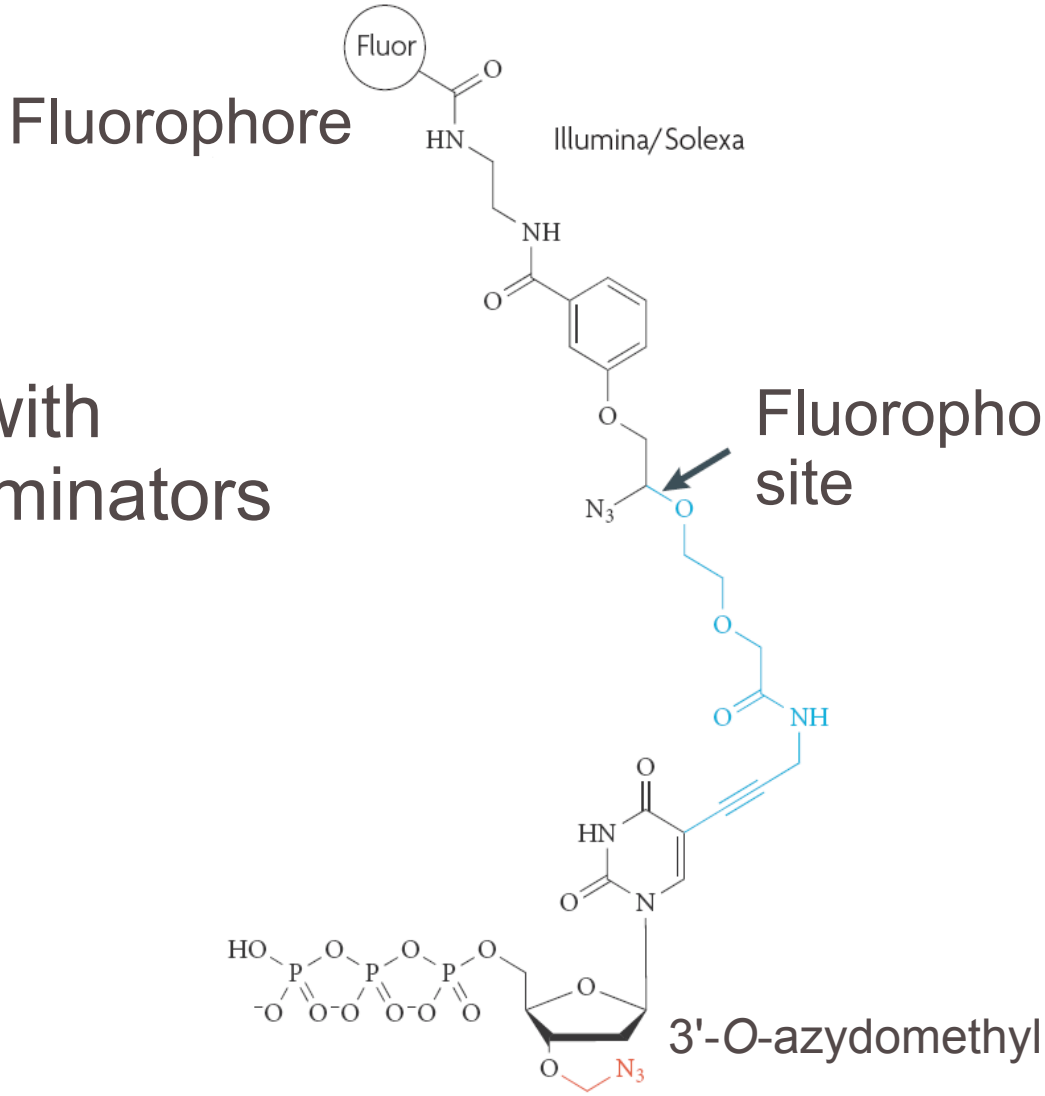
A DNA replication enzyme is added to the matrix, along with 4 modified nucleotides. These nucleotides are chemically modified in a way so that each one would:

- Emit light at different wavelengths when excited by a laser
- Terminate the replication, in order that the complementary strands in each cluster are extended by only one nucleotide, which identity can be detected by its emission wavelength

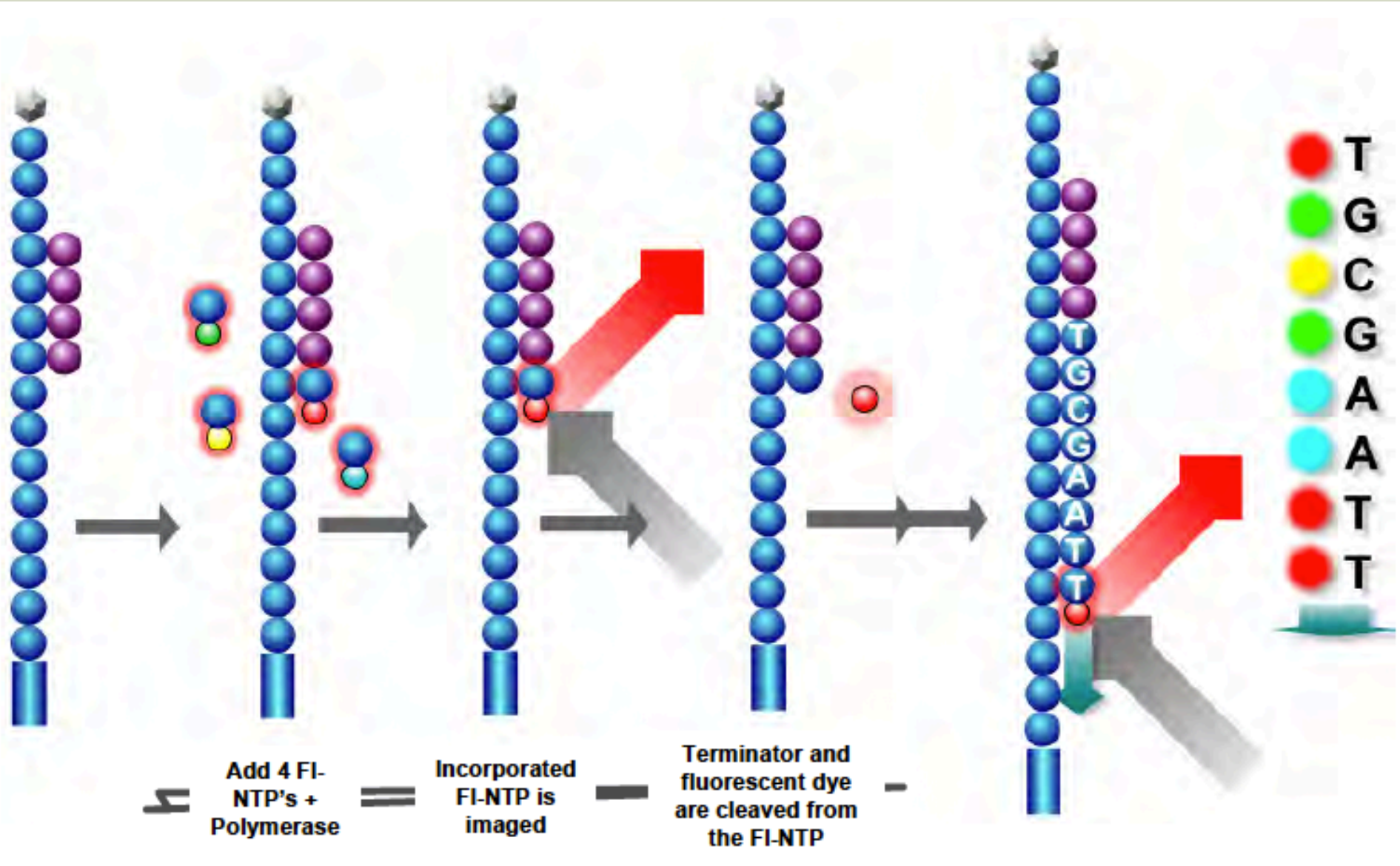
Both the fluorophore and the replication extension block can be removed after the image capture

ILLUMINA SEQUENCING

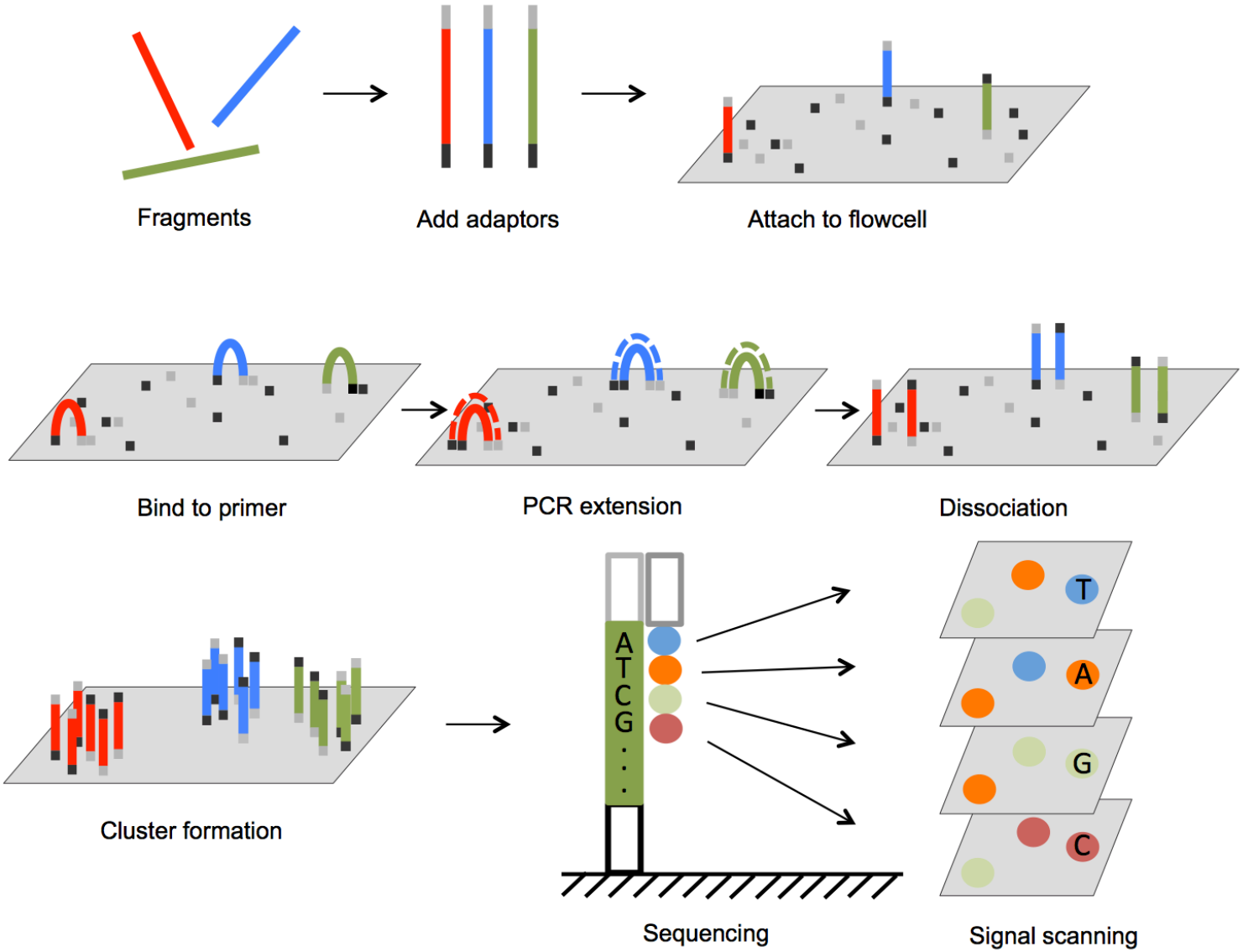
Sequencing with reversible terminators



ILLUMINA SEQUENCING

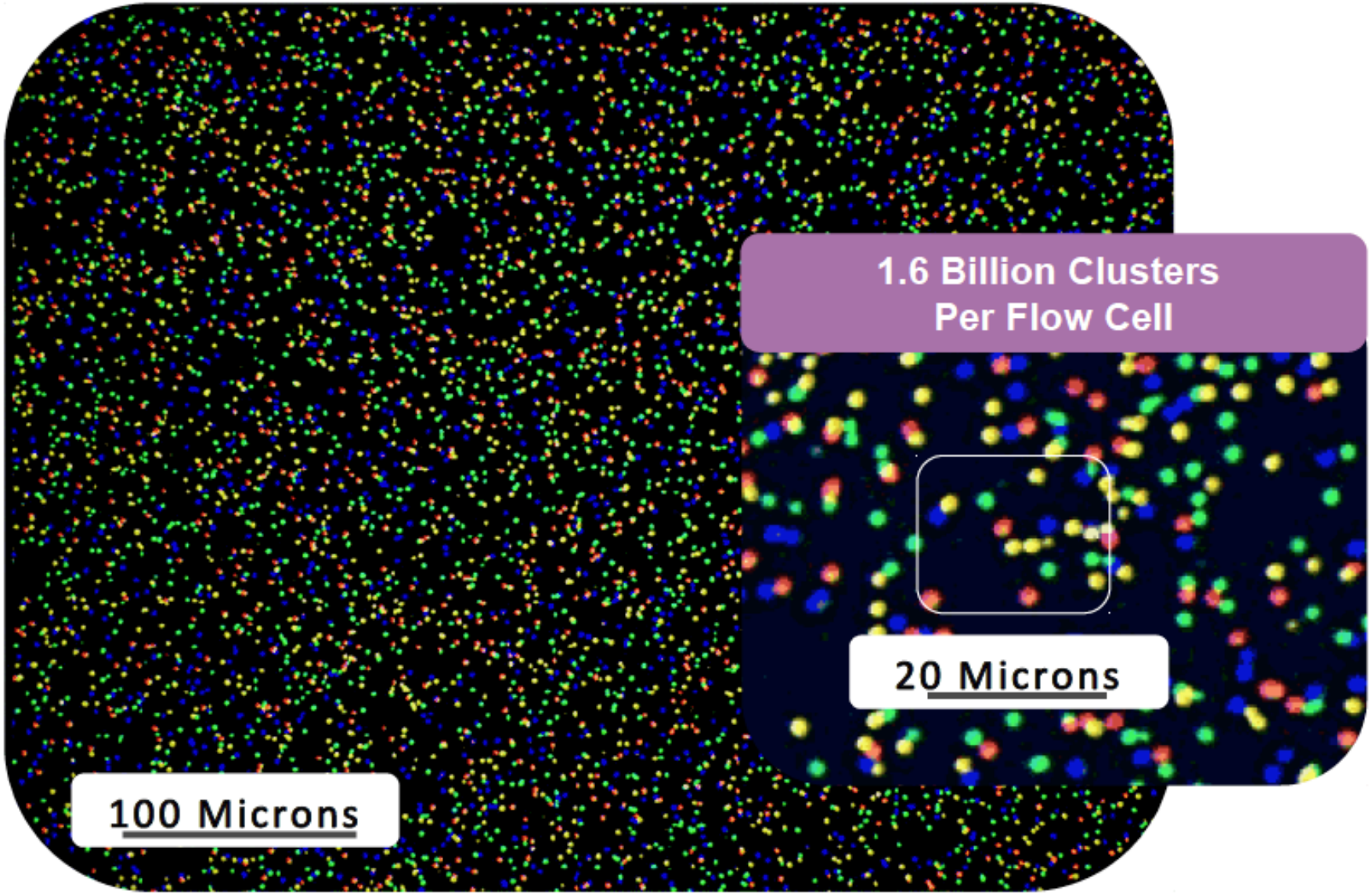


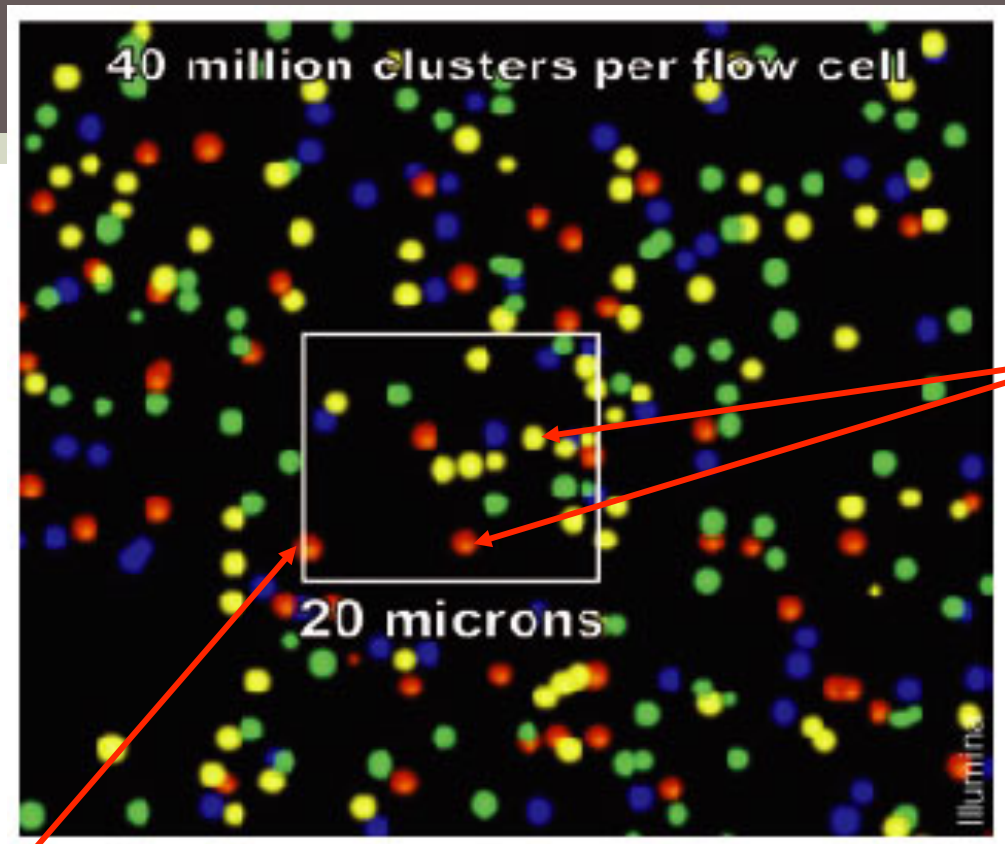
ILLUMINA SEQUENCING



Nature Reviews Microbiology 10, 599-606 (2012)

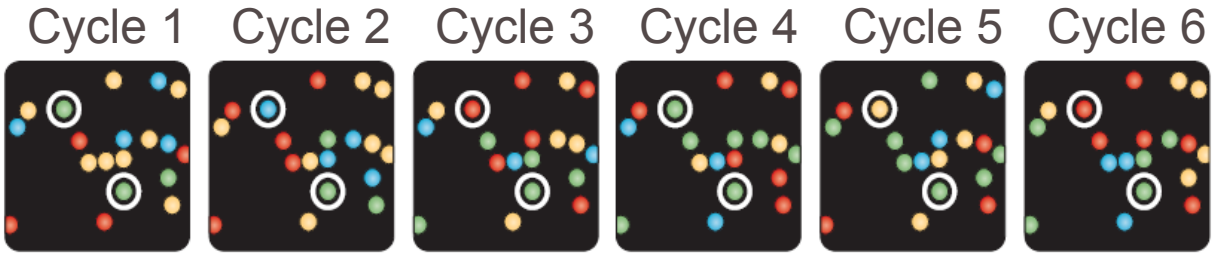
ILLUMINA SEQUENCING

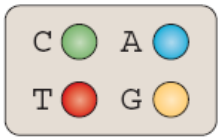




sequence clusters

tile





Top: CATCGT
Bottom: CCCCCC

ILLUMINA SEQUENCING

- Several different Illumina platforms are available on the market, differing by the ampount of data that can be obtained, from the MiniSeq (8 Gb, 25M read, 2x150 bp) to the HiSeq (1800 Gb, 6G reads, 2X150 bp)

ILLUMINA SEQUENCING

- Illumina NovaSeq (6000 Gb, 20G read, 2x150 bp)

ILLUMINA SEQUENCING

- Illumina reads have fixed length, usually between 50 and 300 bp, depending on the number of sequencing cycles.
- The most common sequencing error is base substitution, the overall sequencing accuracy is quite high, having error rate less than 0.1%. Insertions and deletions are rare
- Nevertheless, substitutions rate is not uniform, and it increases towards the read end
- Some sequencing errors are dependent on base composition: substitutions are more frequent at high GC content, while insertions and deletions tend to increase at homopolymeric sites or at inverted repeats

ILLUMINA SEQUENCING

Reads.What do they look like?

```
@SEQUENCE1
GCCCGGCGGGTTCATGCTGAAGAAAGGCGAAGTGTTCGGTTGGGCGGC
+
fffffffefe^eeceedffdcd^dXecffbeed`Reebe`db\]XWSS
```

This is a part of the standard FASTQ format showing a single read, reporting each sequenced nucleotide and an estimation of the probability that this nucleotide is called incorrectly

ILLUMINA SEQUENCING

```
@SEQUENCE1
GCCCGGCGGGTTCATGCTGAAGAAAGGCGAAGTGTTCGGTTGGGCGGC
+
fffffffefe^eeceedffdcd^dXecffbeed`Reebe`db\]XWSS
```

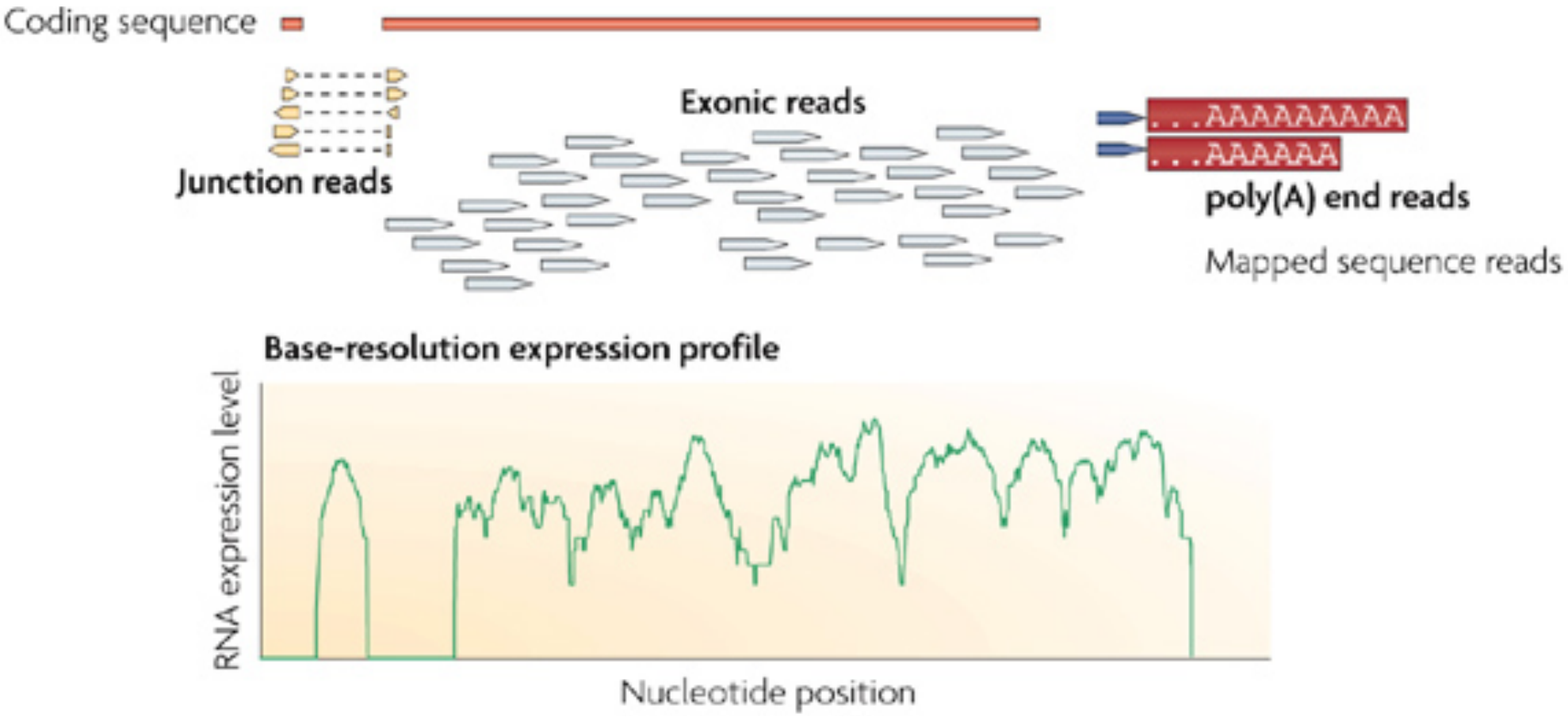
A FASTQ file employs 4 rows for each read:

- The first row starts with @ followed by the unique read identifier
- The second row contains the sequence.
- The third row starts with a + symbol and it might contain the read unique identifier
- The fourth row contains the quality scores for each nucleotide in the read sequence, encoded as decimal conversion of the ASCII code (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange) of the corresponding character (es.]=93,f=102).

RNA-SEQ


genome
transcripts
aligned reads
junction reads
exonic reads

RNA-SEQ


Coding sequence
Exonic reads
Junction reads
. . .AAAAAAAAA
. . .AAAAAA
poly(A) end reads
Mapped sequence reads
Base-resolution expression profile
RNA expression level
Nucleotide position

EXPERIMENTAL DESIGN

Things to consider beforehand:

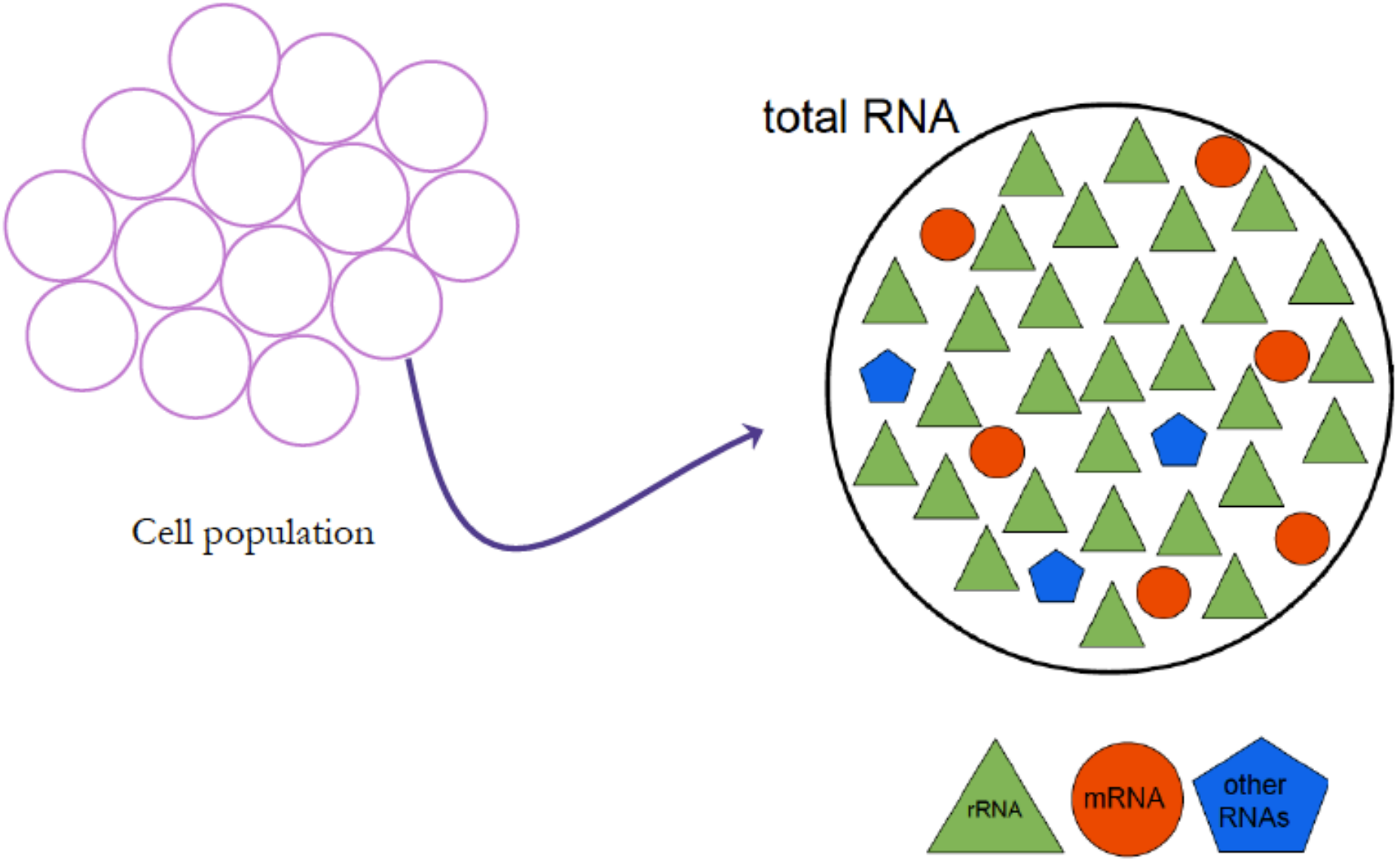
- Library type and preparation
- Sequencing depth
- Number of replicates
- Control and avoidance of biases

EXPERIMENTAL DESIGN

RNA extraction protocol:

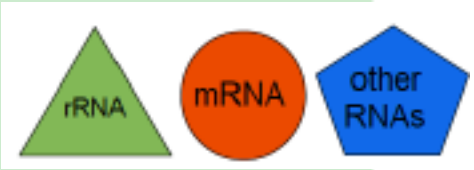
- rRNA removal
- RNA integrity check
- Size selection

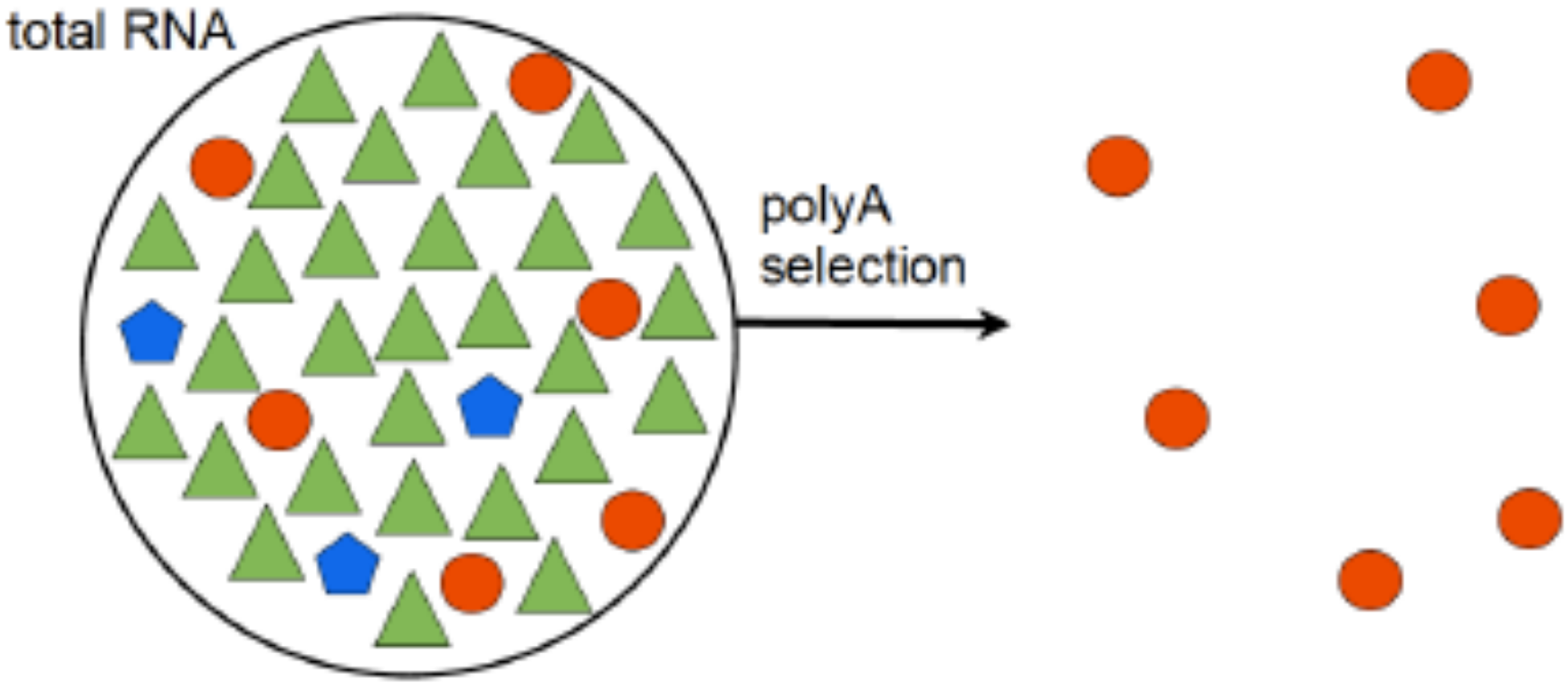
EXPERIMENTAL DESIGN



EXPERIMENTAL DESIGN

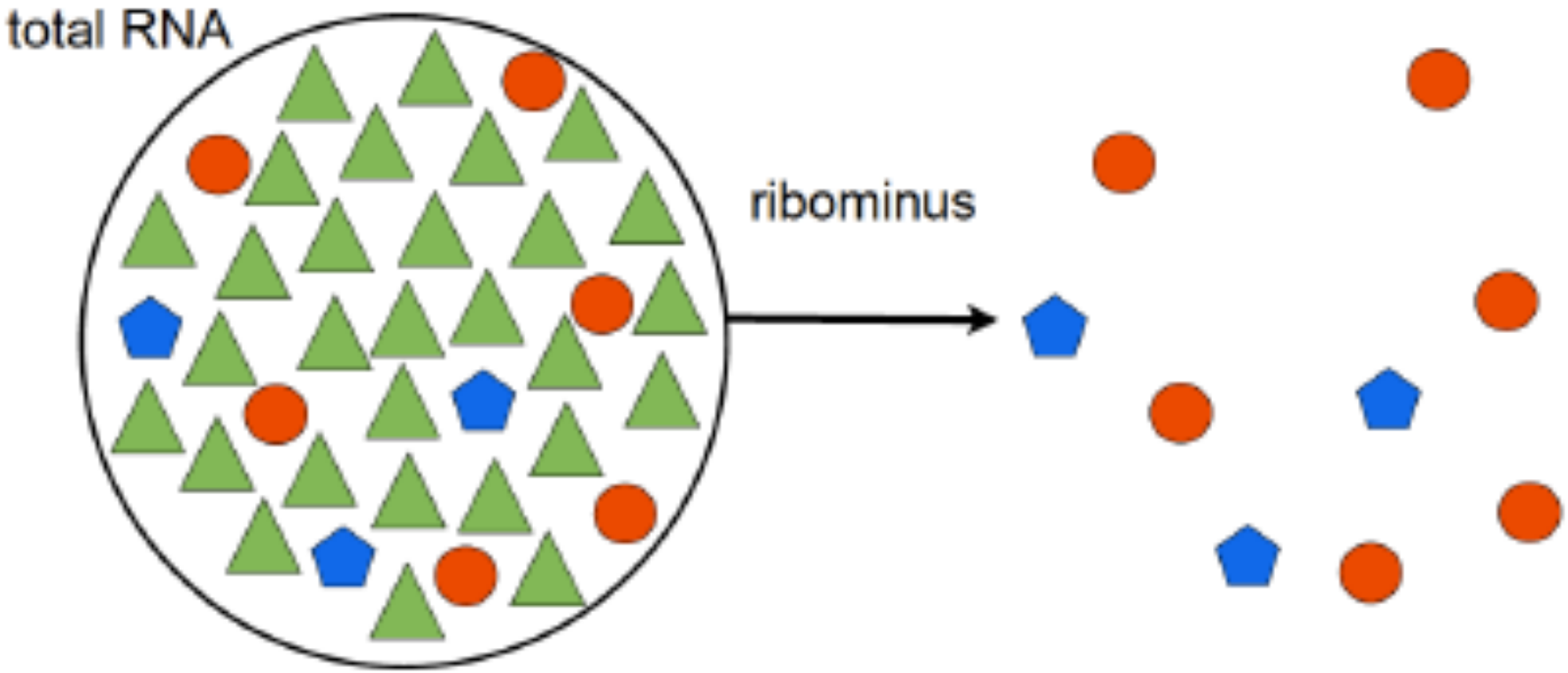
poly-$(A)^+$ selection





EXPERIMENTAL DESIGN

ribominus selection

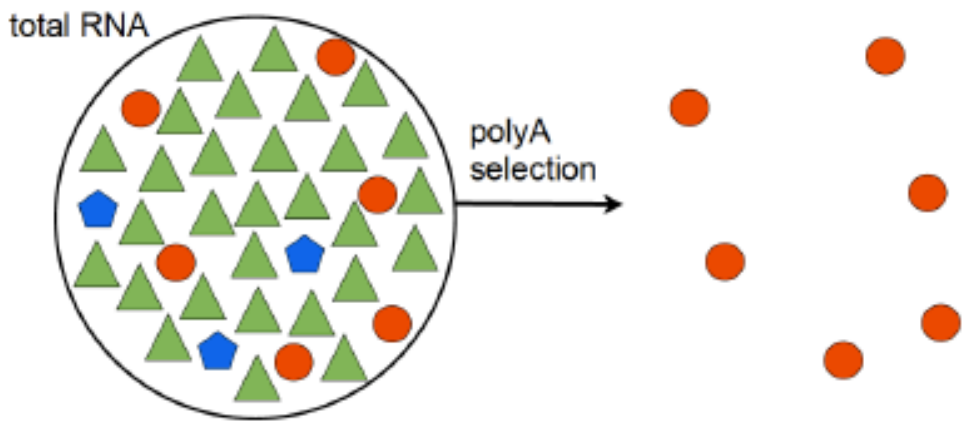


EXPERIMENTAL DESIGN

- Two main strategies were developed for rRNA depletion from an RNA extract:
  - Subtractive hybridization using rRNA specific probes
  - Selective degradation of rRNAs (and of all the other 5'monophosphate RNAs) using exonucleases. mRNAs are protected from degradation by the 5'cap, or, in prokaryotes, by a 5' triphosphate
- Yet, highly expressed genes with short half-life, and partially degraded mRNAs could be lost following exonuclease treatment
- Additionally, RNAs having highly stable secondary structures might resist degradation by impairing exonuclease processing

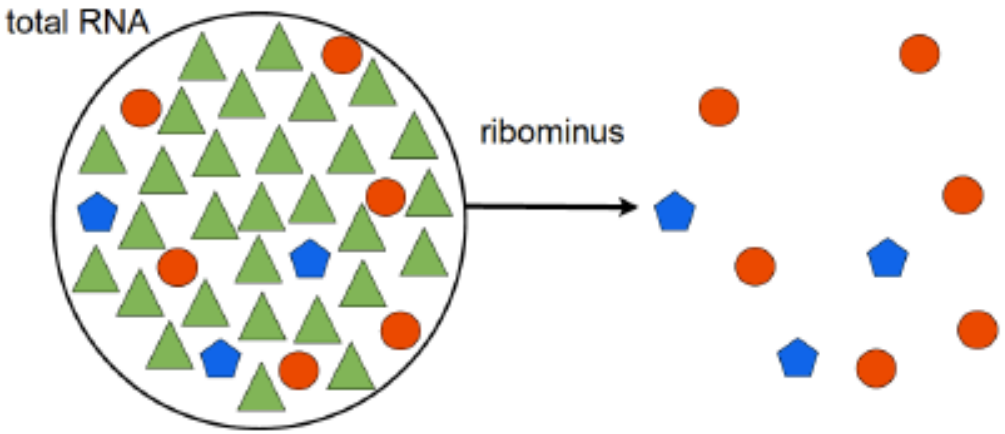
EXPERIMENTAL DESIGN

poly-(A)$^+$ selection



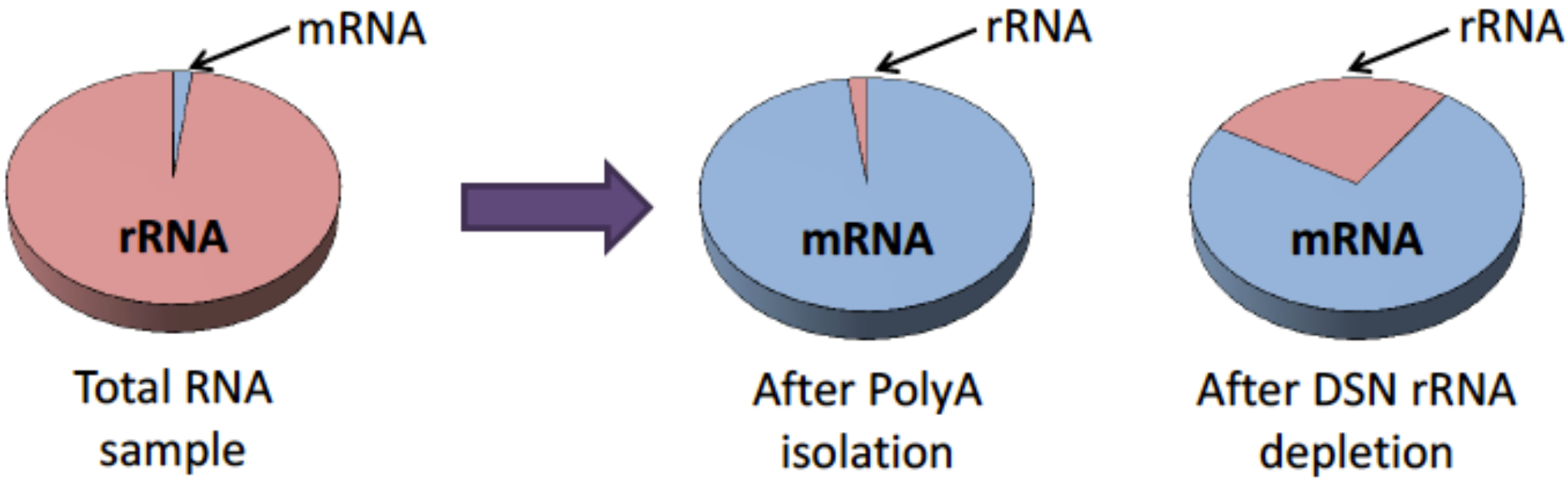
- Most commonly used
- Many non-coding RNAs will be missed
- Requires more RNA
- Requires non-degraded RNAs

ribominus selection



- More complex and expensive
- Provides “complete” transcriptome
- Requires less RNA
- Less affected by RNA degradation

EXPERIMENTAL DESIGN



rRNA depletion is generally less effective in removing all rRNA from the extract compared to poly(A) selection

EXPERIMENTAL DESIGN

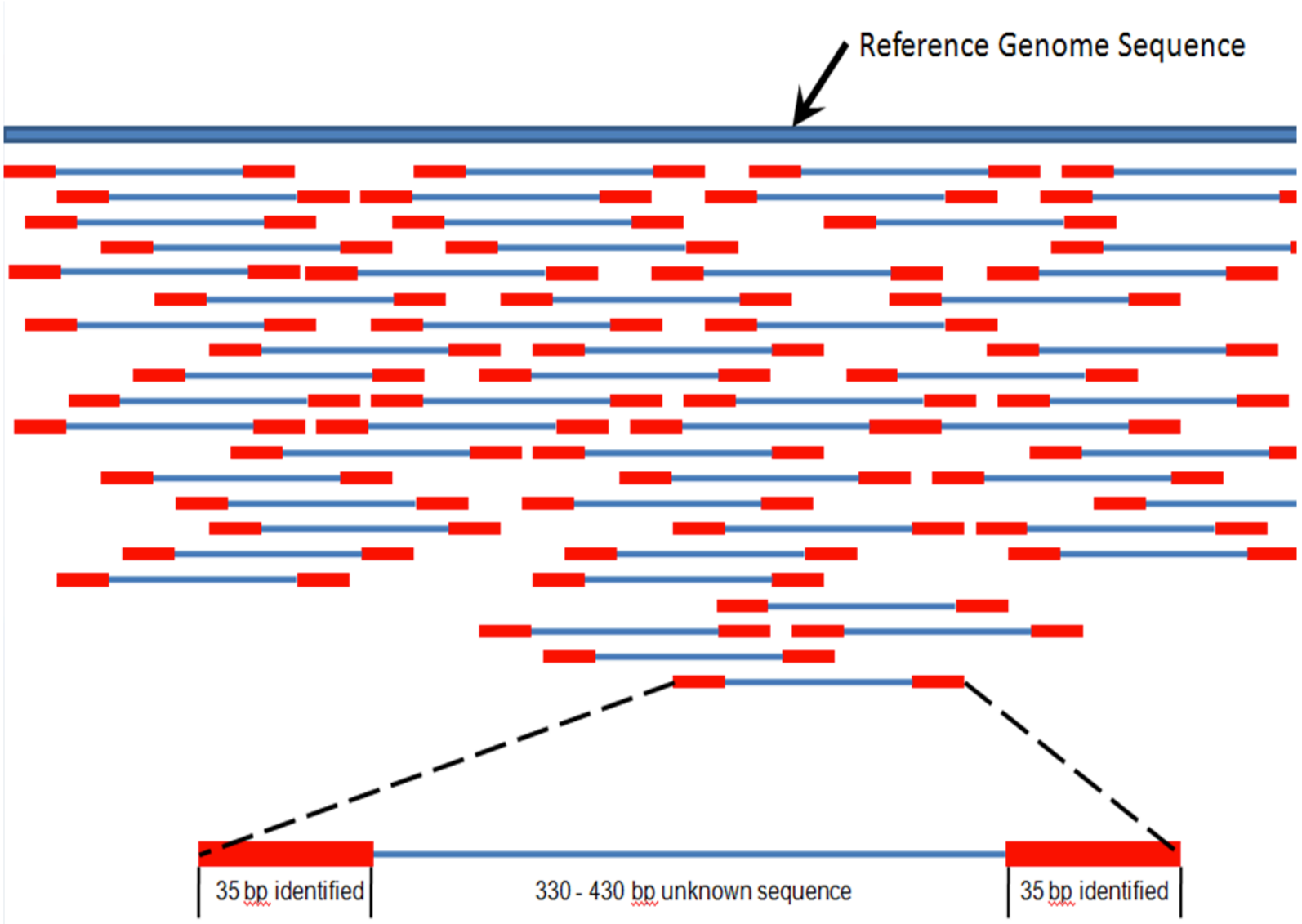
- RNA molecules are easily degraded, hence it is usual practice to check for RNA integrity during library preparation
- RIN (RNA integrity number) is a measure of RNA degradation based on gel electrophoretic runs
- For RNA selection methods based on poly(A) capture, a degraded RNA sample would enrich the library with fragments corresponding to the 3'end of the transcripts, thus creating a possible bias in expression estimates and alternative splicing investigation
- RNA depletion methods are less affected by RNA degradation

EXPERIMENTAL DESIGN

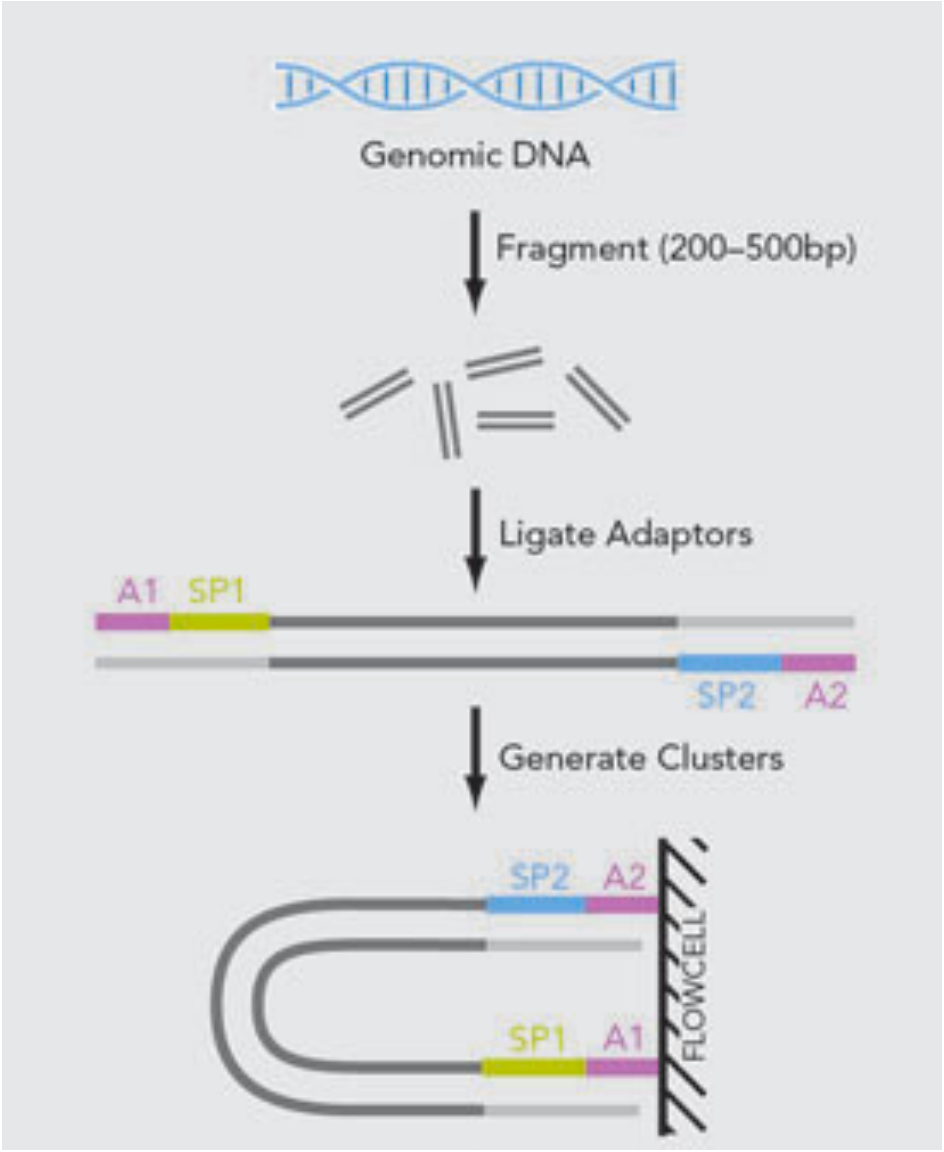
Given transcriptome fragments:

- Only one end of each fragment can be sequenced (single end sequencing)
  - Simpler, less expensive, can be the better choice for gene expression studies in well annotated genomes
- Both ends of each fragment can be sequenced (paired end sequencing)
  - The paired reads generated from a fragment usually do not cover the entire fragment length, but they can still be useful to obtain long-range connection and better characterize alternative splicing events

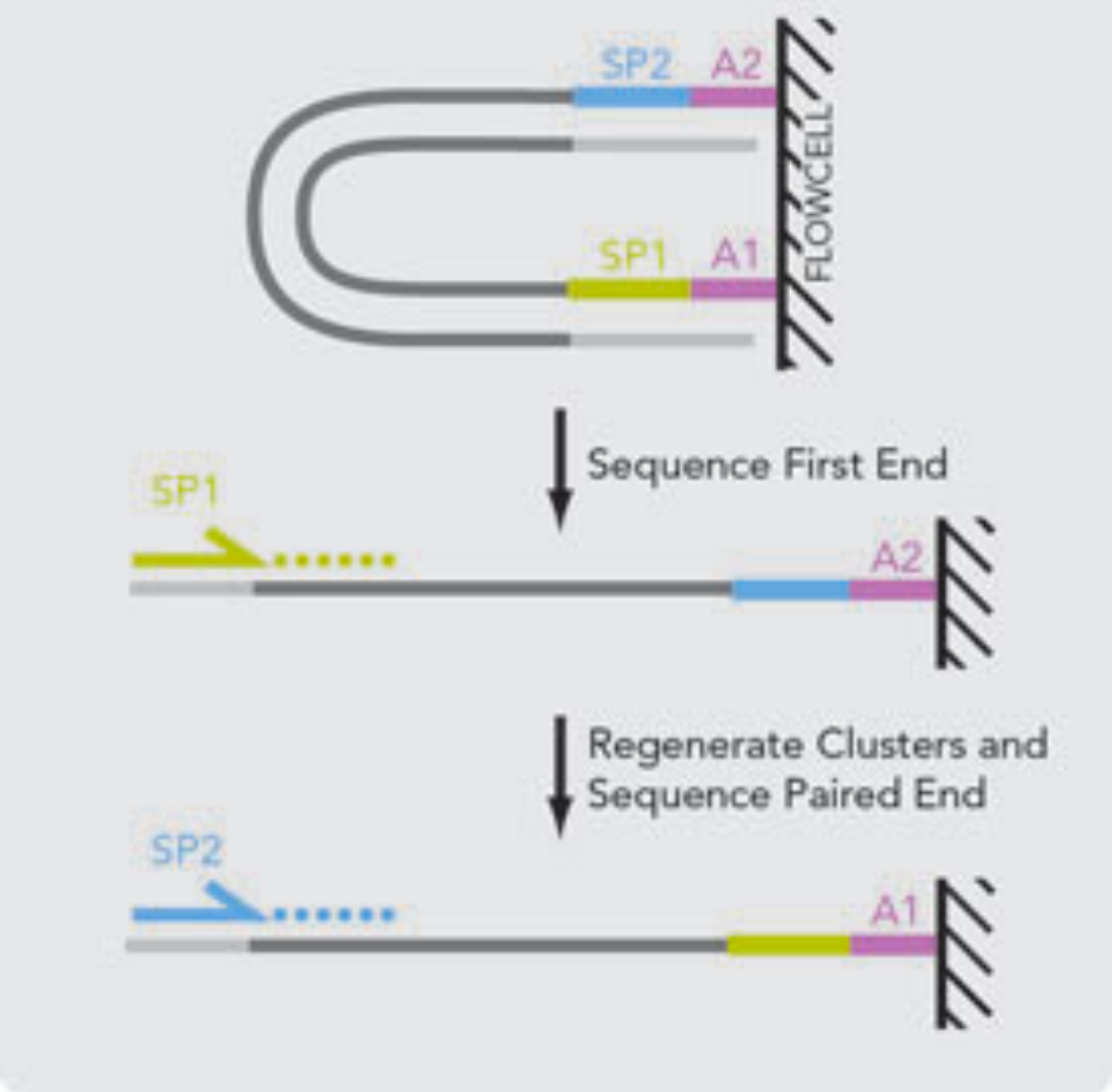
EXPERIMENTAL DESIGN



EXPERIMENTAL DESIGN

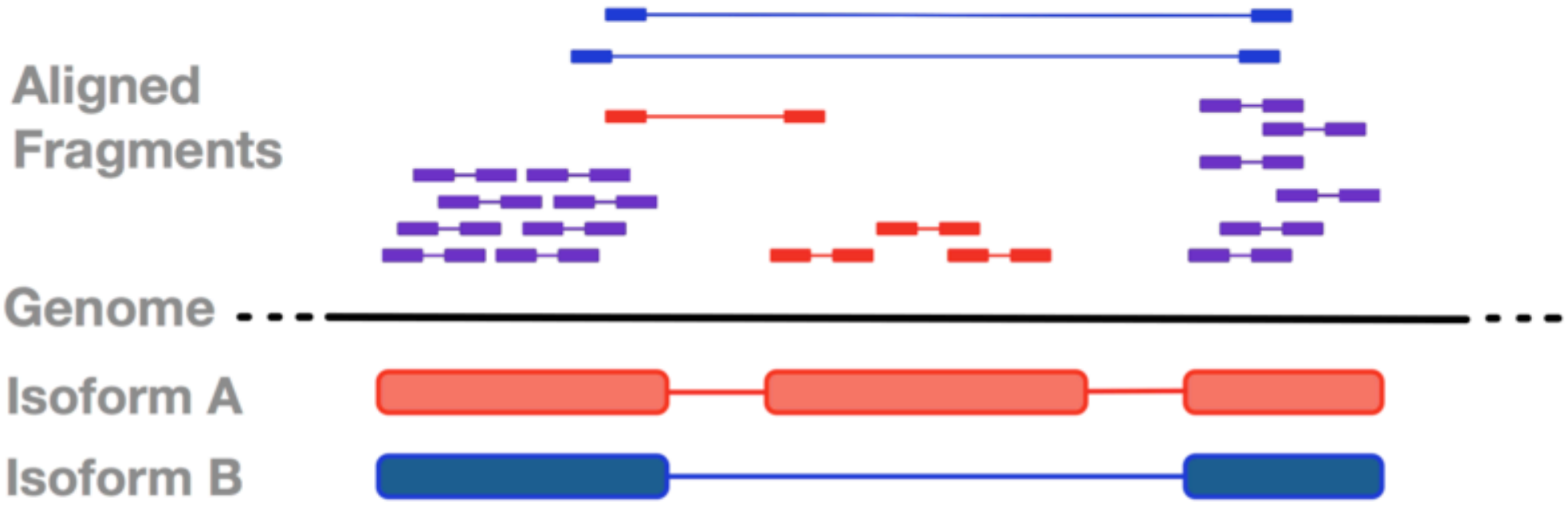


EXPERIMENTAL DESIGN

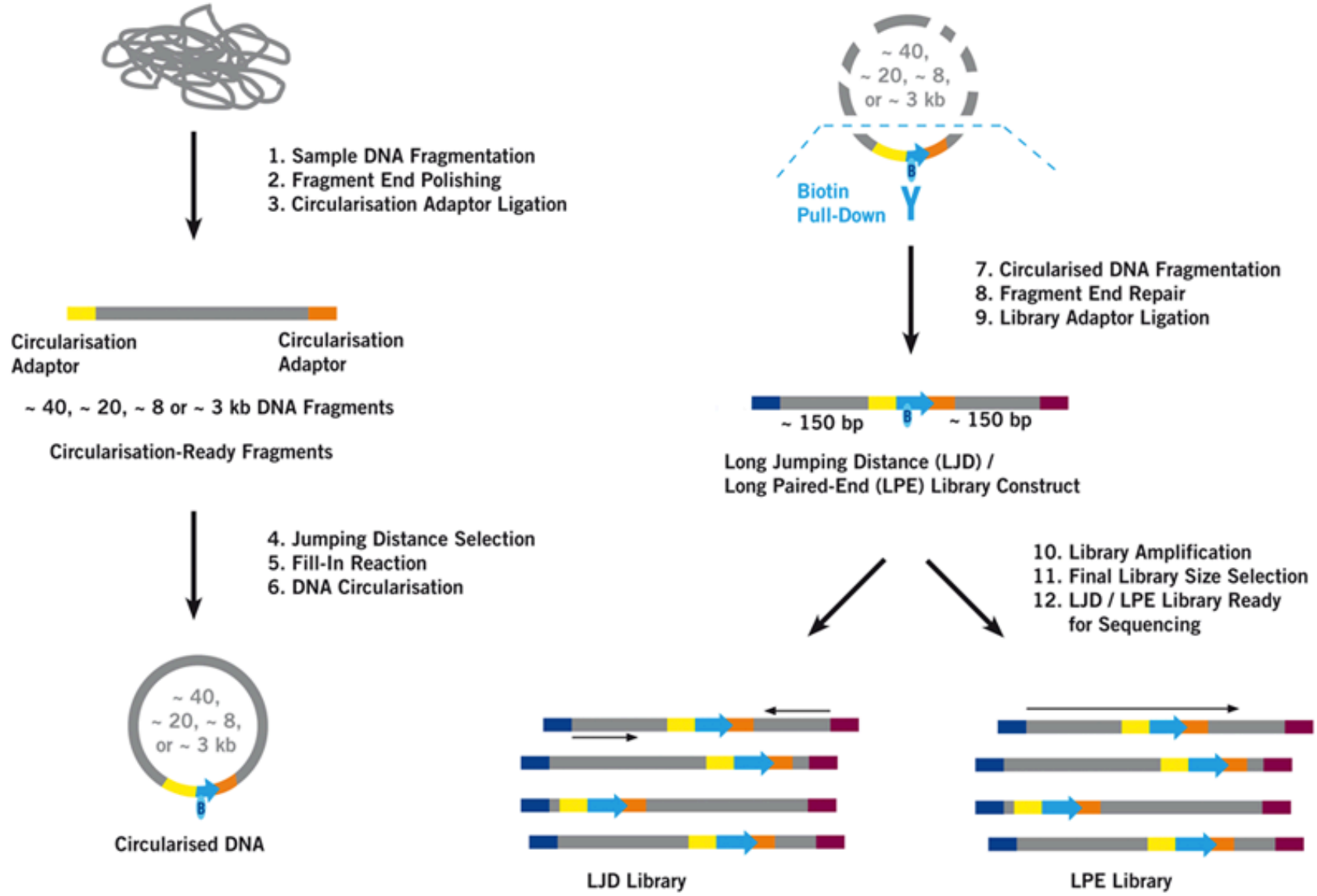


EXPERIMENTAL DESIGN

Paired end reads facilitate in the read mapping, in solving repetitive regions and in allowing a better disentangling of the expression of splicing variants



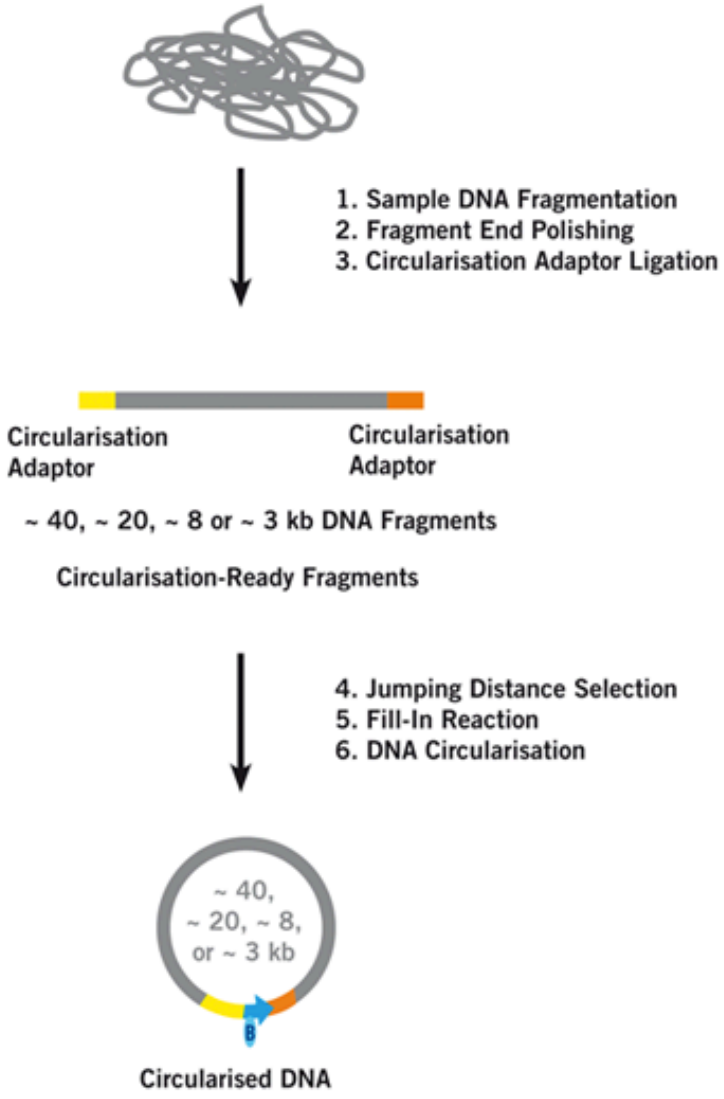
EXPERIMENTAL DESIGN



https://www.eurofinsgenomics.eu/

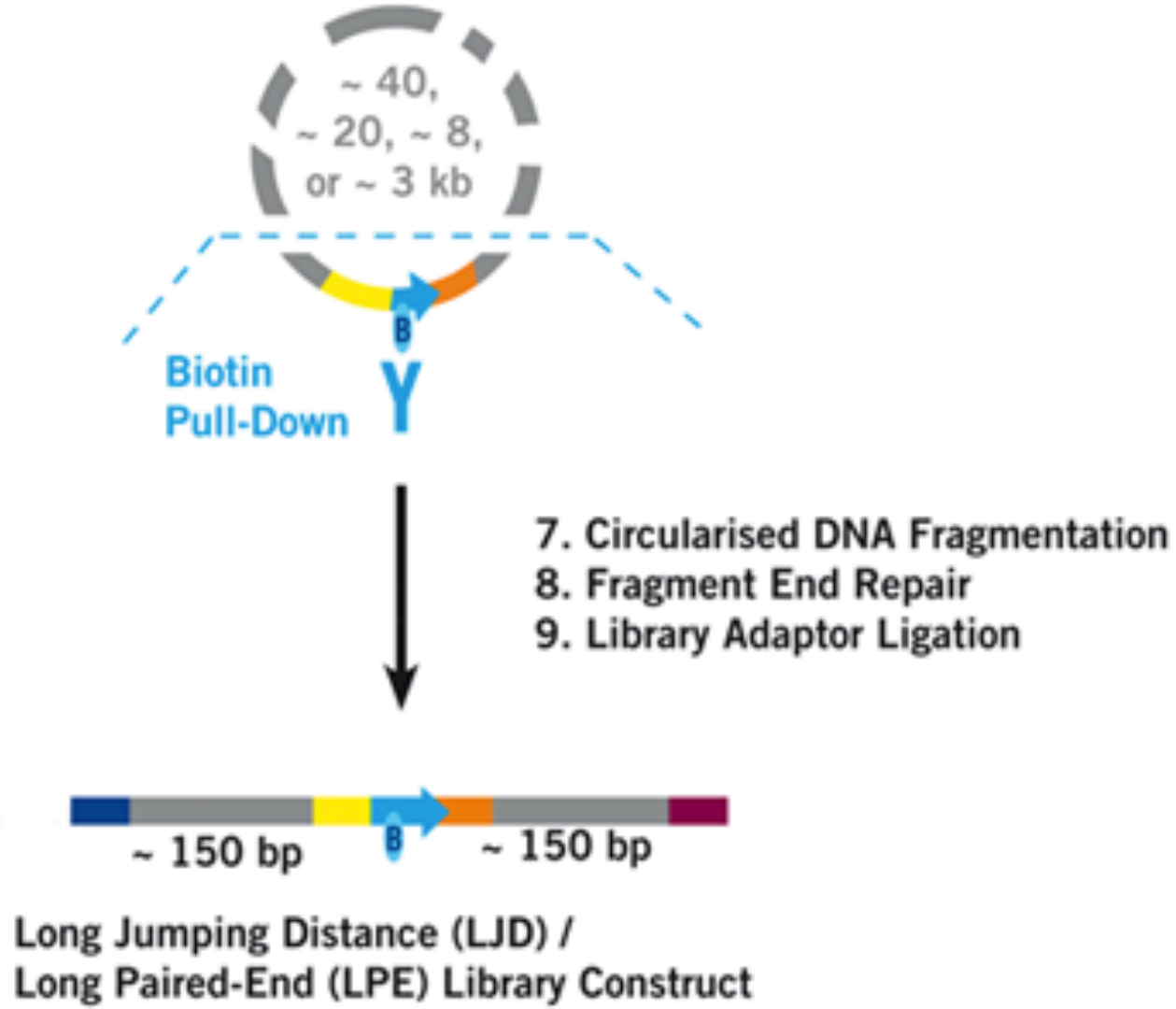
EXPERIMENTAL DESIGN

- Large genomic fragments cannot be sequences as paired end reads
- To determine the sequences at the extremities of large genomic fragments (up to tens of kbp), these are circularized in order to bringin proximity the fragment ends. The junction is marked (e.g. with biotin)
- This circular RNA is fragmented, and the marked fragments are rescued and sequenced as paired ends



https://www.eurofinsgenomics.eu/

EXPERIMENTAL DESIGN



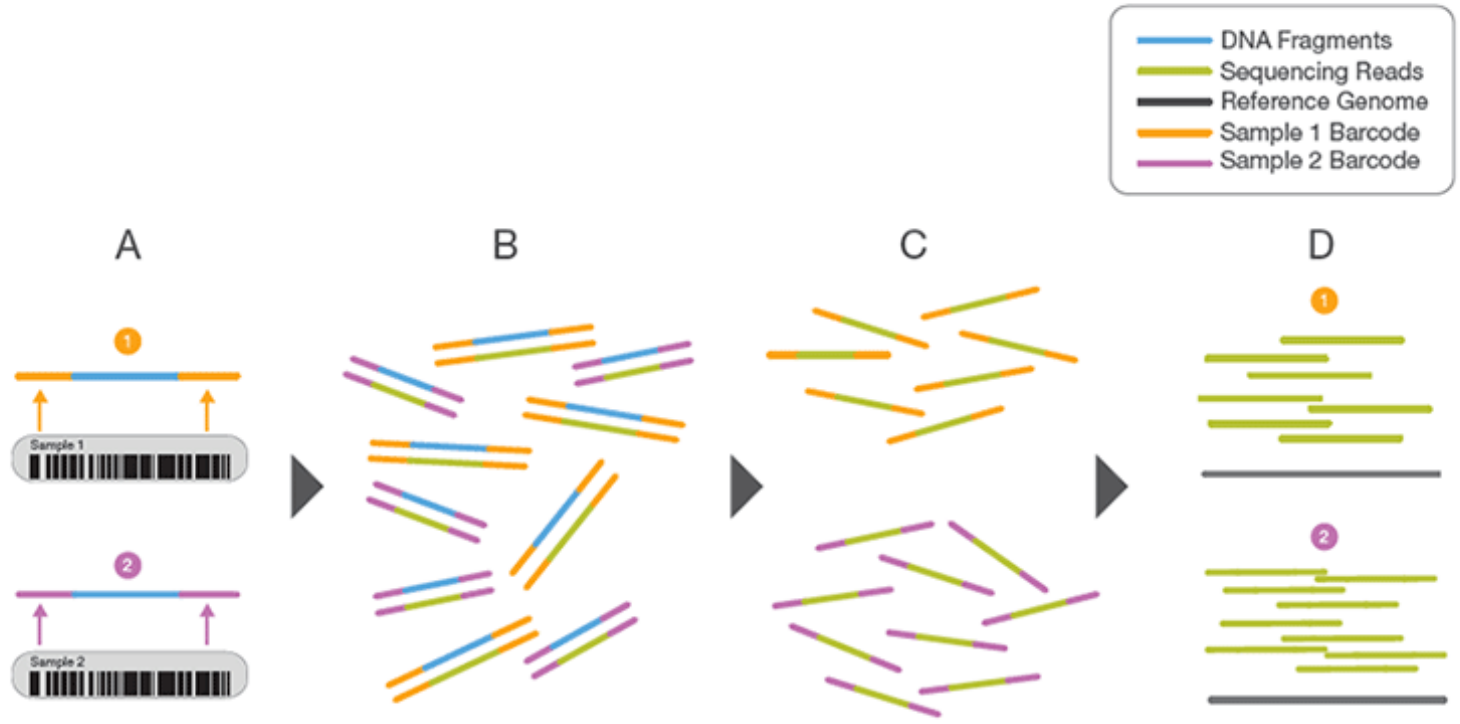
https://www.eurofinsgenomics.eu/

EXPERIMENTAL DESIGN

- It is possible to sequence multiple samples together in the same flow cell lane
- This is due to reduce costs, and minimizing biases due to lane effects. It is generally denoted as multiplexing or barcoding
- To distinguish the sample of origin for each read sequenced in a lane, short sequences (barcodes) are used, which are attached to the fragments together with the adapters. These barcodes have a different sequence for each sample
- After the read sequences are determined, the first nucleotides of their sequence allow the separation of the read pool into the different samples

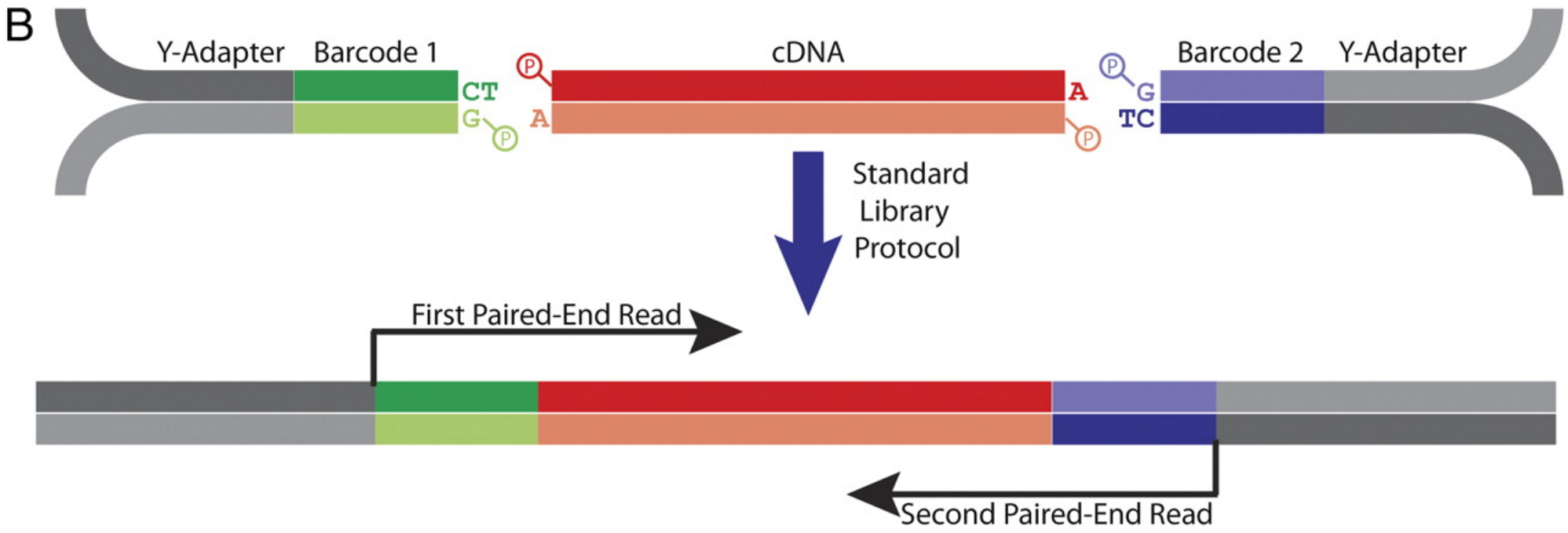
EXPERIMENTAL DESIGN

Figure 2: Conceptual Overview of Sample Multiplexing



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

EXPERIMENTAL DESIGN



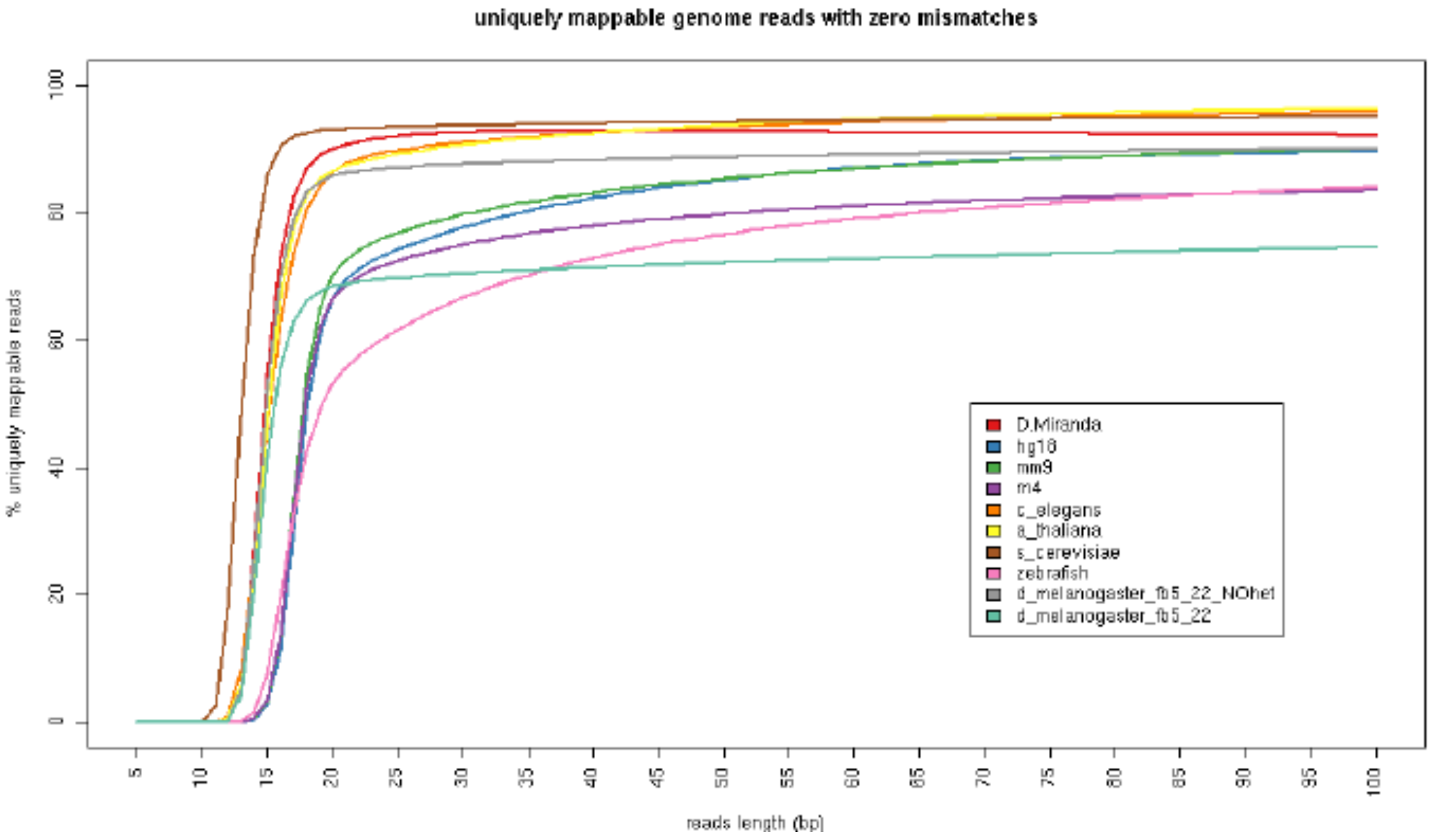
Illumina TruSeq Barcodes

ATCACG CGATGT TTAGGC TGACCA ACATGT GCCAAT
CAGATC ACTTGA GATCAG TAGCTT GGCTAG CTTGTA

EXPERIMENTAL DESIGN

- Read length affects the mapping of the read sequence onto a reference genome/transcriptome
- The shorter the read, the more likely is that its sequence would be similar by chance to the wrong genomic region
- For Illumina platforms, read length is fixed and depends on the employed platform
- Over a certain length, which depends on the genomic composition, higher lengths might not be beneficial anymore

EXPERIMENTAL DESIGN



From Francesco Ferrari

EXPERIMENTAL DESIGN

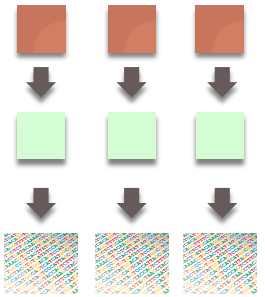
- Sequencing depth (also called library size) is the number of sequenced reads for a sample
- Higher sequencing depth generates more informational reads, which increases the statistical power to detect differentially expressed genes
- Nevertheless, very high depths not necessarily lead to any novel insight, reaching saturation (i.e. the expression estimates do not become more accurate)
- There is also evidence that too large sequencing depths can result in the detection of transcriptional noise and off-target transcripts

EXPERIMENTAL DESIGN

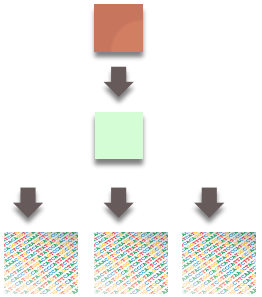
- The number of replicates that should be included in a RNA-seq experiments depends on the amount of technical variability and the biological variability of the system under study
- Technical variability is generally very low, and can be minimized by controlling batch effects and randomizing samples (by multiplexing, running on more lanes, etc)
- Biological variability is instead often very large, requiring the preparation and analysis of different biologically equivalent samples

EXPERIMENTAL DESIGN

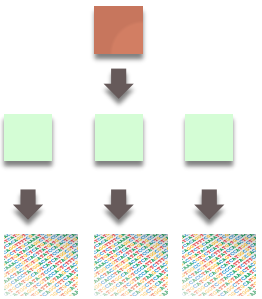
- Biological replicates: different cell cultures are used, from which different libraries are prepared in parallel



- Technical replicates: one culture is used, one library is prepared then divided into aliquots



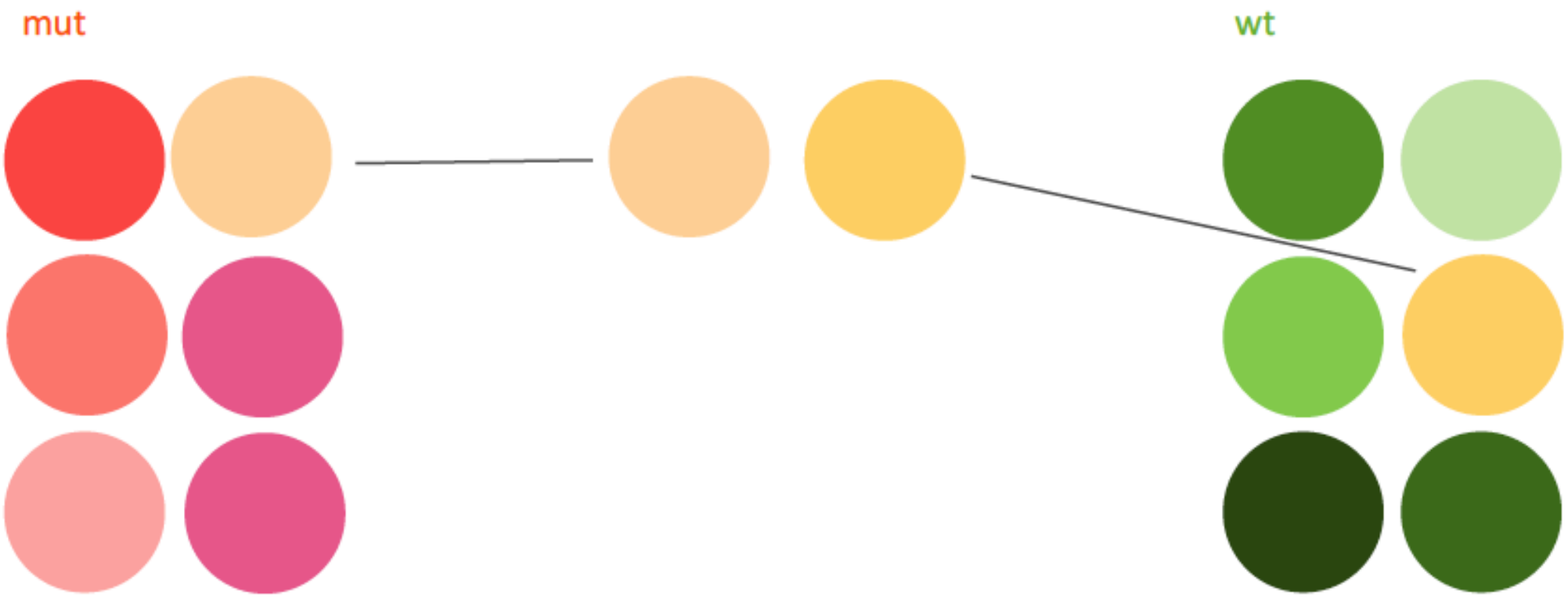
- Intermediate strategies: only one culture is employed, then divided in more cultures processed independently



EXPERIMENTAL DESIGN

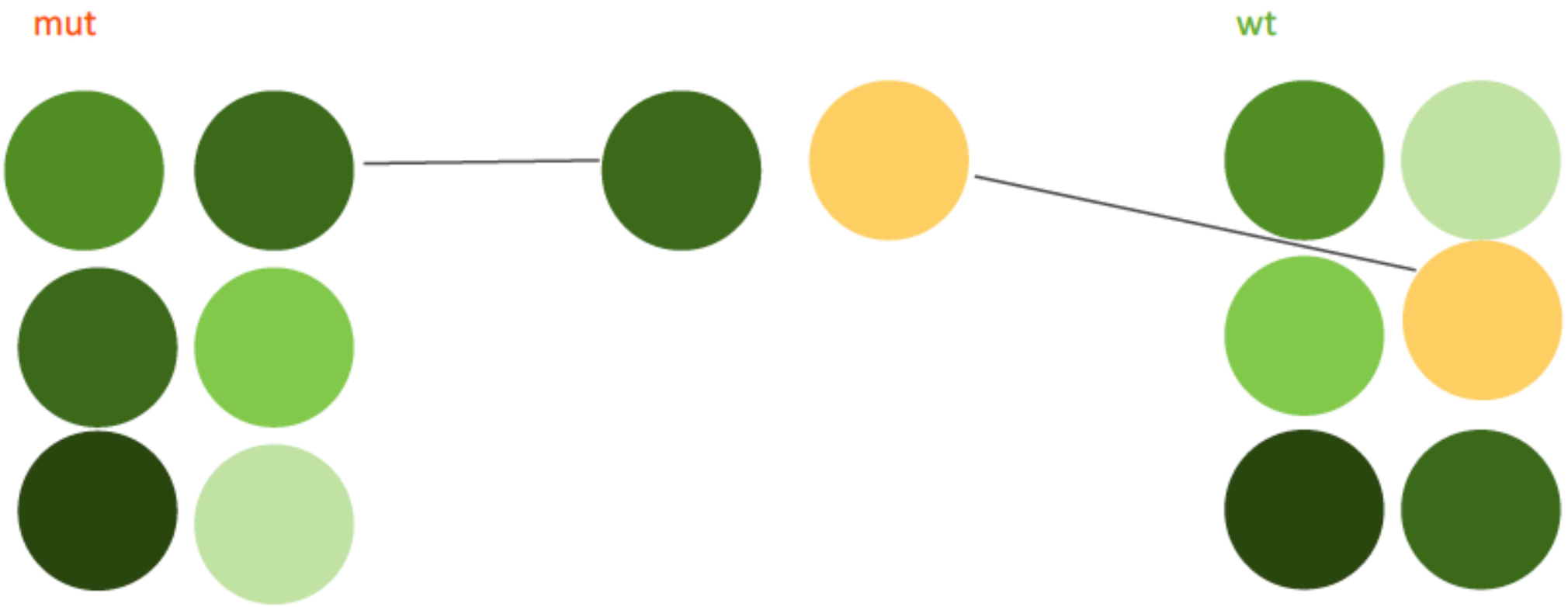
Biological heterogeneity

EXPERIMENTAL DESIGN



Here, groups differ, but individual replicates from each group can be very similar. If only these two replicates are analyzed and compared, the two groups would incorrectly appear similar

EXPERIMENTAL DESIGN

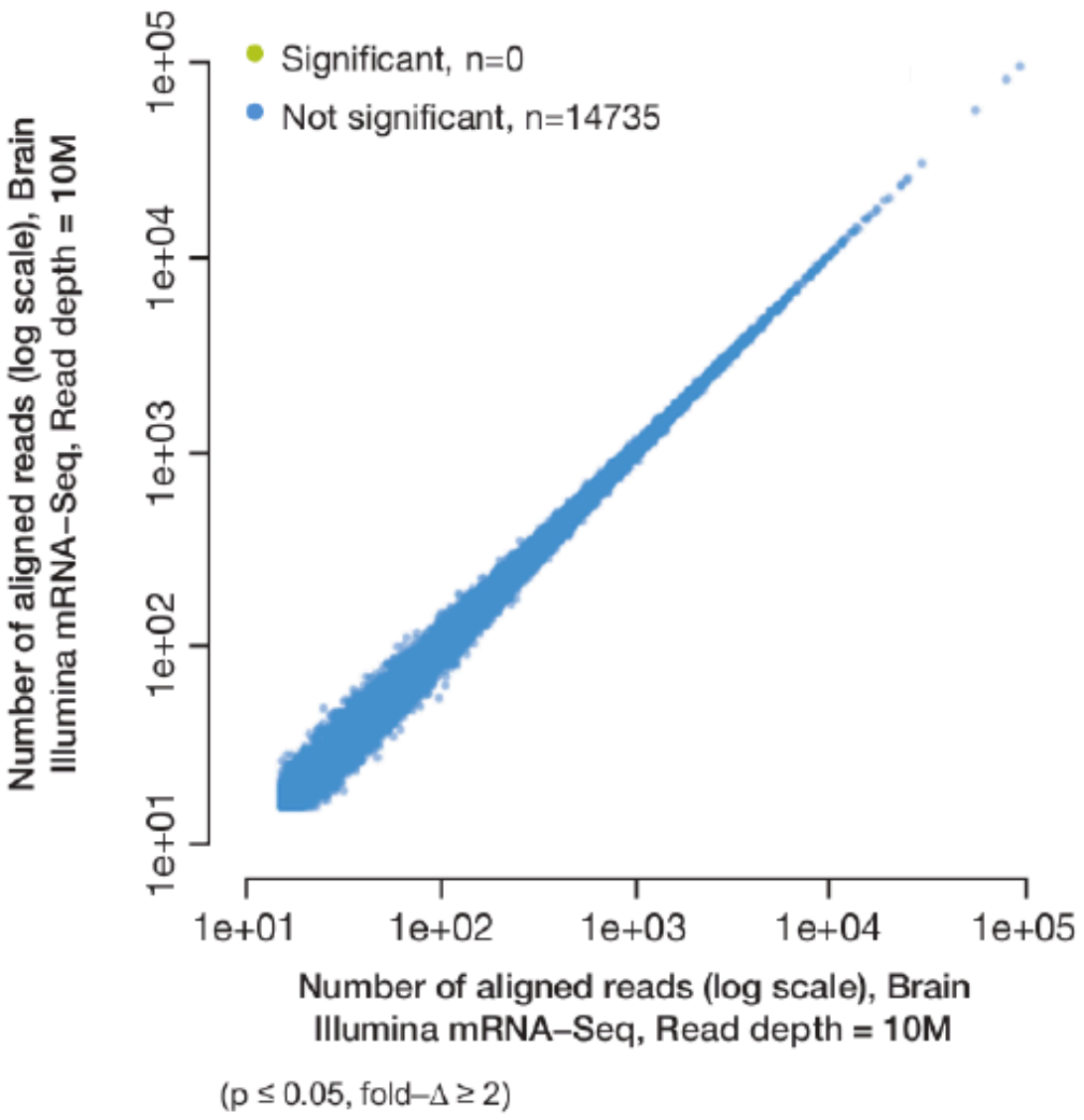


Here, groups are similar, but outlying observation from the group on the right could incorrectly lead to noteworthy differences in an unreplicated experiment

EXPERIMENTAL DESIGN

- Single biological replicates may not be representative of a whole group representing a biological condition
- Single replicates also strongly limit the statistical power to test for differential expression
- Some analysis software can estimate within-group variability under some assumptions, but results can be misleading
- Three biological replicates per condition is accepted as a good compromise between statistical power and cost

EXPERIMENTAL DESIGN



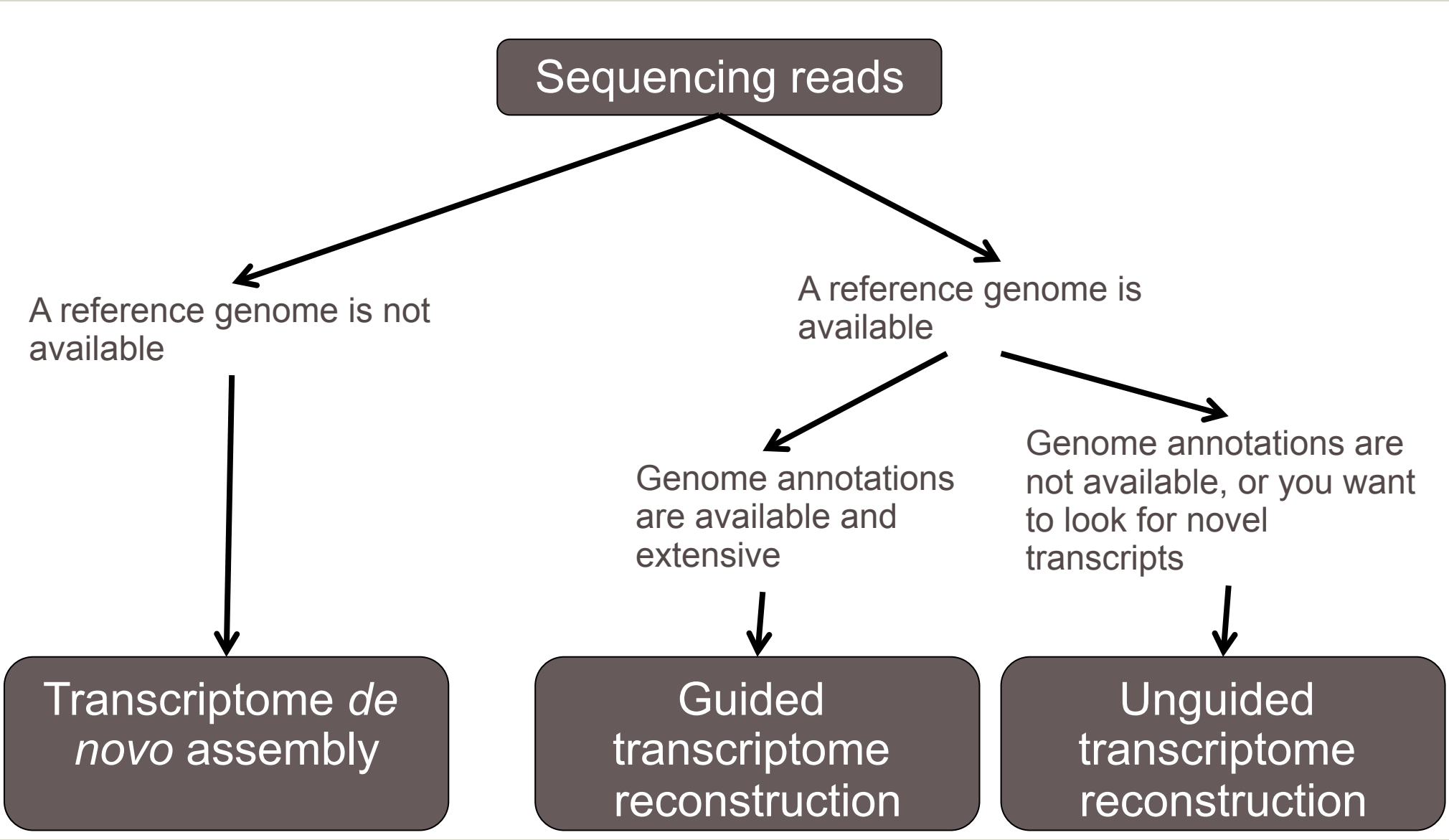
- Technical replicates show very high degrees of correlation, therefore are not employed
- Remember: in RNA-seq, biological variability >> technical variability
- Also remember: technical replicates cannot substitute biological replicates

EXPERIMENTAL DESIGN

Library construction and sequencing decisions:

Project Goals:	*De novo* Assembly of transcriptome	Refine gene model	Differential Gene Expression	Identification of structural variants
Library Type:	PE, Mated PE	PE, SE	PE	PE, Mated PE
Sequencing Depth:	Extensive (> 50 X)	Extensive	Moderate (10 X ~ 30 X)	Extensive

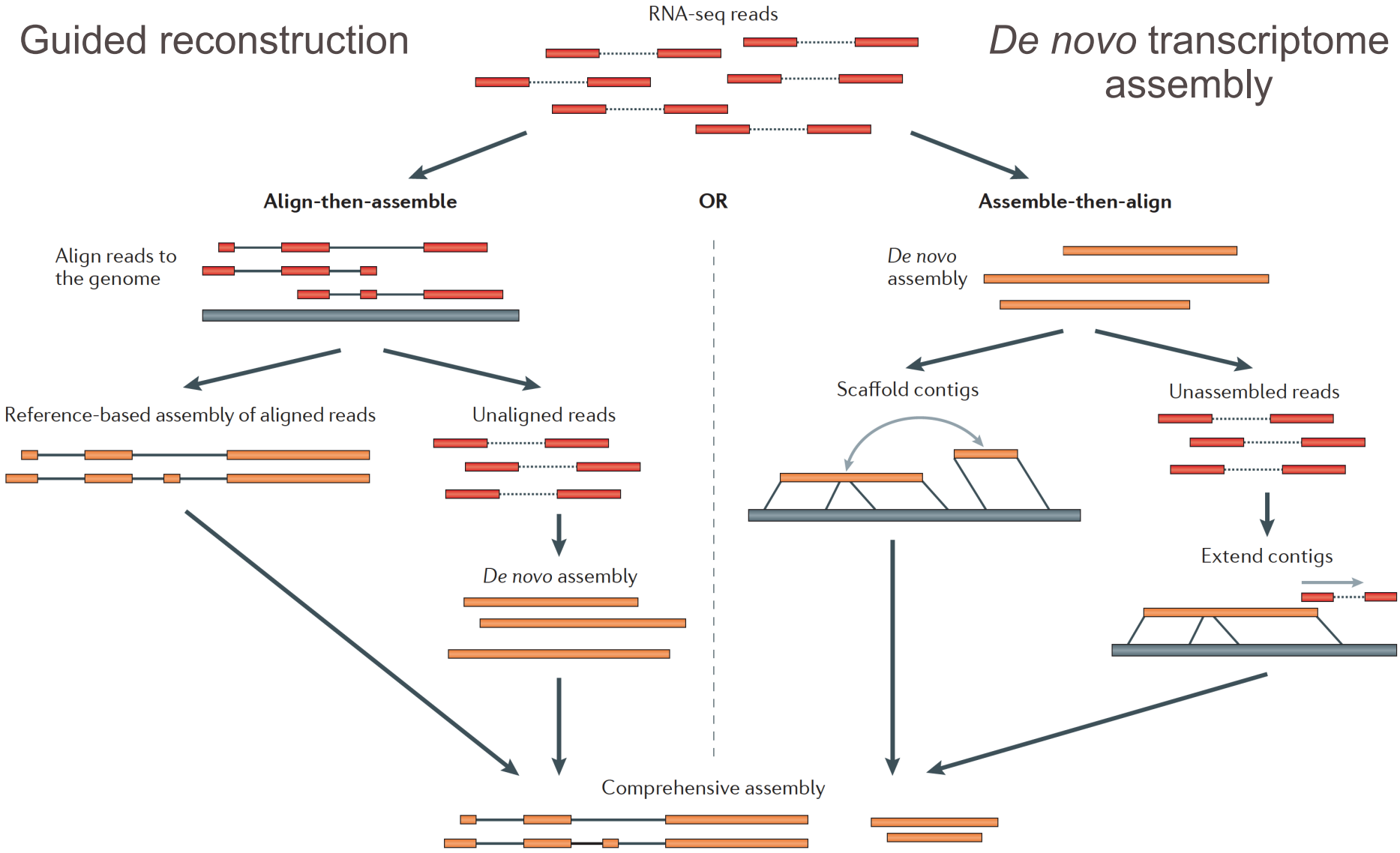
TRANSCRIPTOME RECONSTRUCTION



TRANSCRIPTOME RECONSTRUCTION

- *De novo assembly* of the transcriptome:
  - Reads are compared between each other, looking for shared segments that suggest provenience from the same transcript. Starting from these overlaps, reads are assembled into contiguos sequences that should correspond to while RNAs.
- Guided reconstruction of the transcriptome:
  - Reads are aligned onto the genomic sequence. Gene boundaries, their structure and known splicing variants are used to estimate gene and splicing variants expression.
- Unguided reconstruction of the transcriptome:
  - Reads are aligned onto the genomic sequence. Gene boundaries, their structure and splicing variants are not known, or are not used, or are used only as a general reference. The algorithms try to infer gene structure based on read clusters on the genome, aided by spliced reads at exon-exon junctions and, when available, on paired-end reads

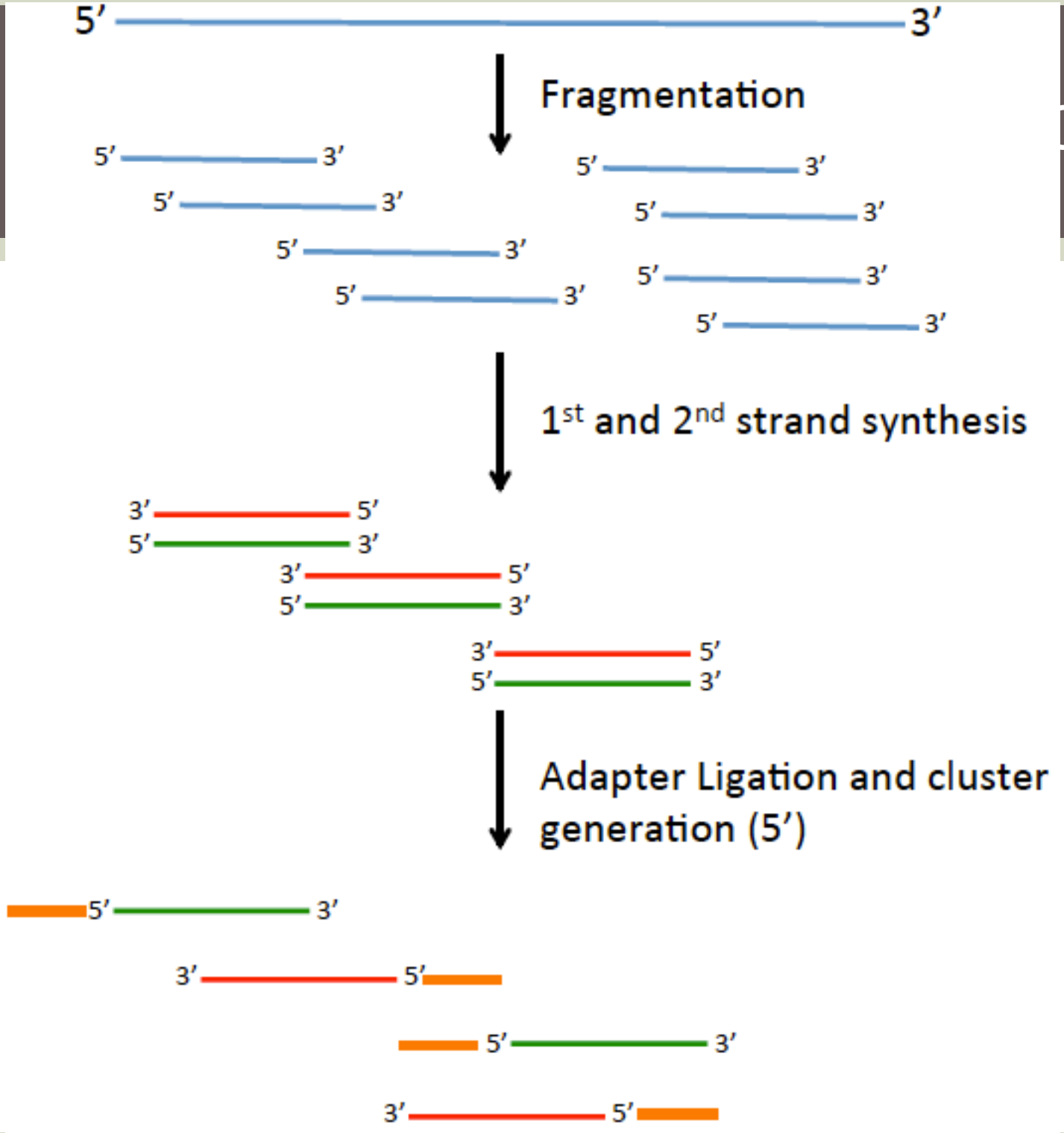
TRANSCRIPTOME RECONSTRUCTION



TRANSCRIPTOME RECONSTRUCTION

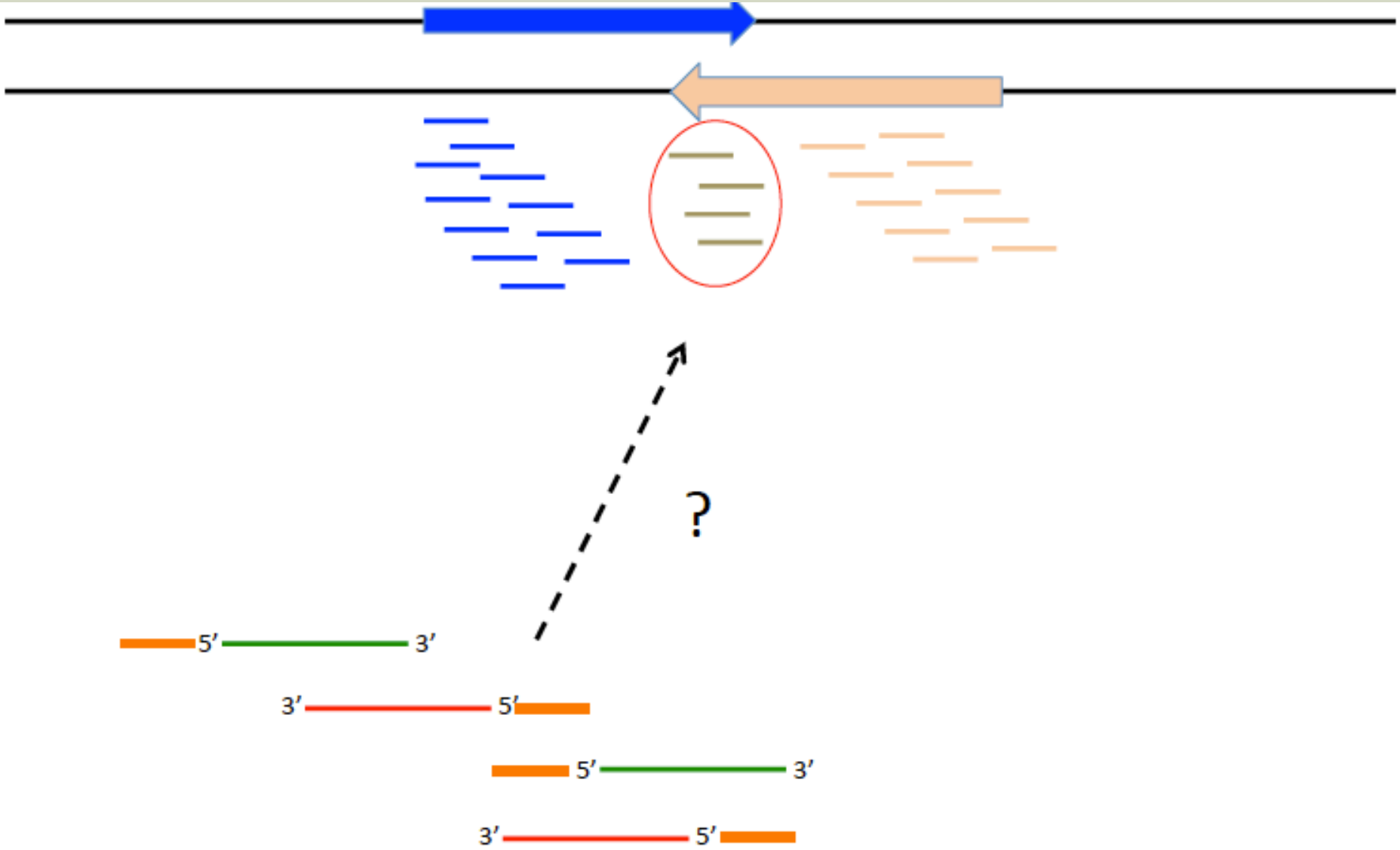
- The three approaches are not mutually exclusive:
  - If you have reliable reference genome and annotations, you always starts fro a guided transcriptome reconstruction
  - If annotations are incomplete, you can try also an unguided reconstruction, to see whether new genes can be found
  - If the reference genome is incomplete or not well tested, you can try a *de novo* assembly to identify genes hosted in genome loci that are absent from the current assembly, or where the assembly is wrong

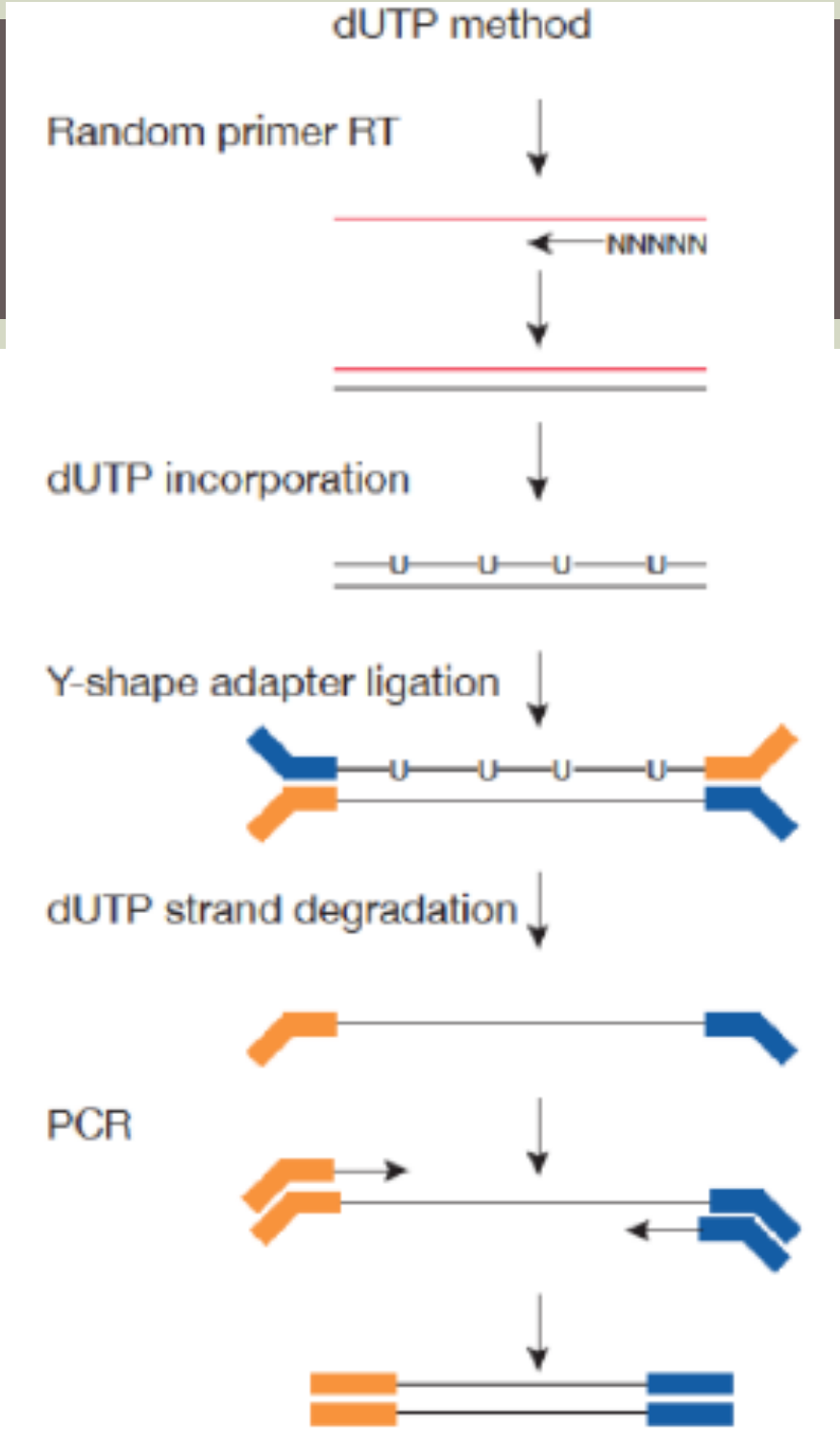
GN



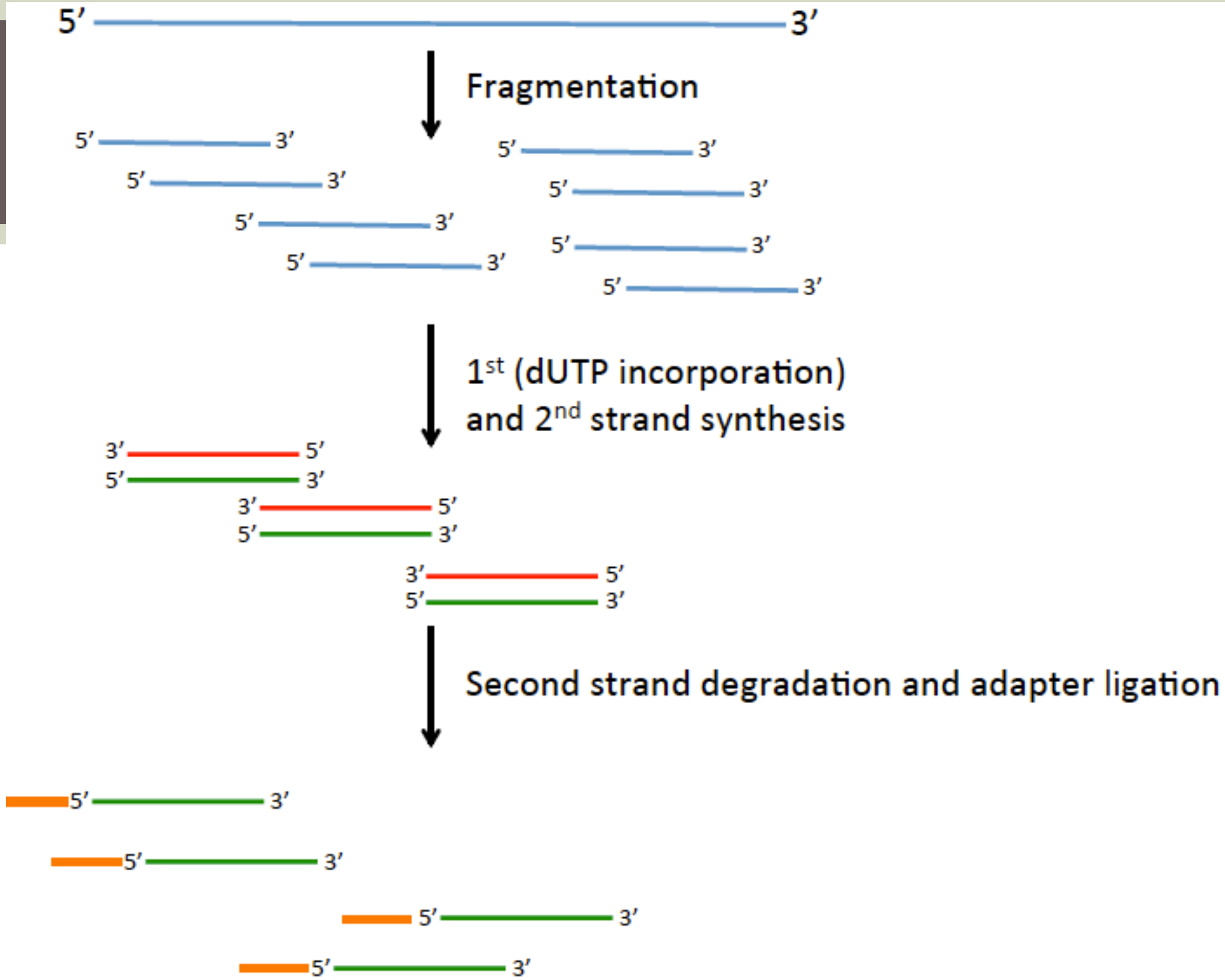
The obtained reads are randomly coming from either the forward or reverse strand

EXPERIMENTAL DESIGN

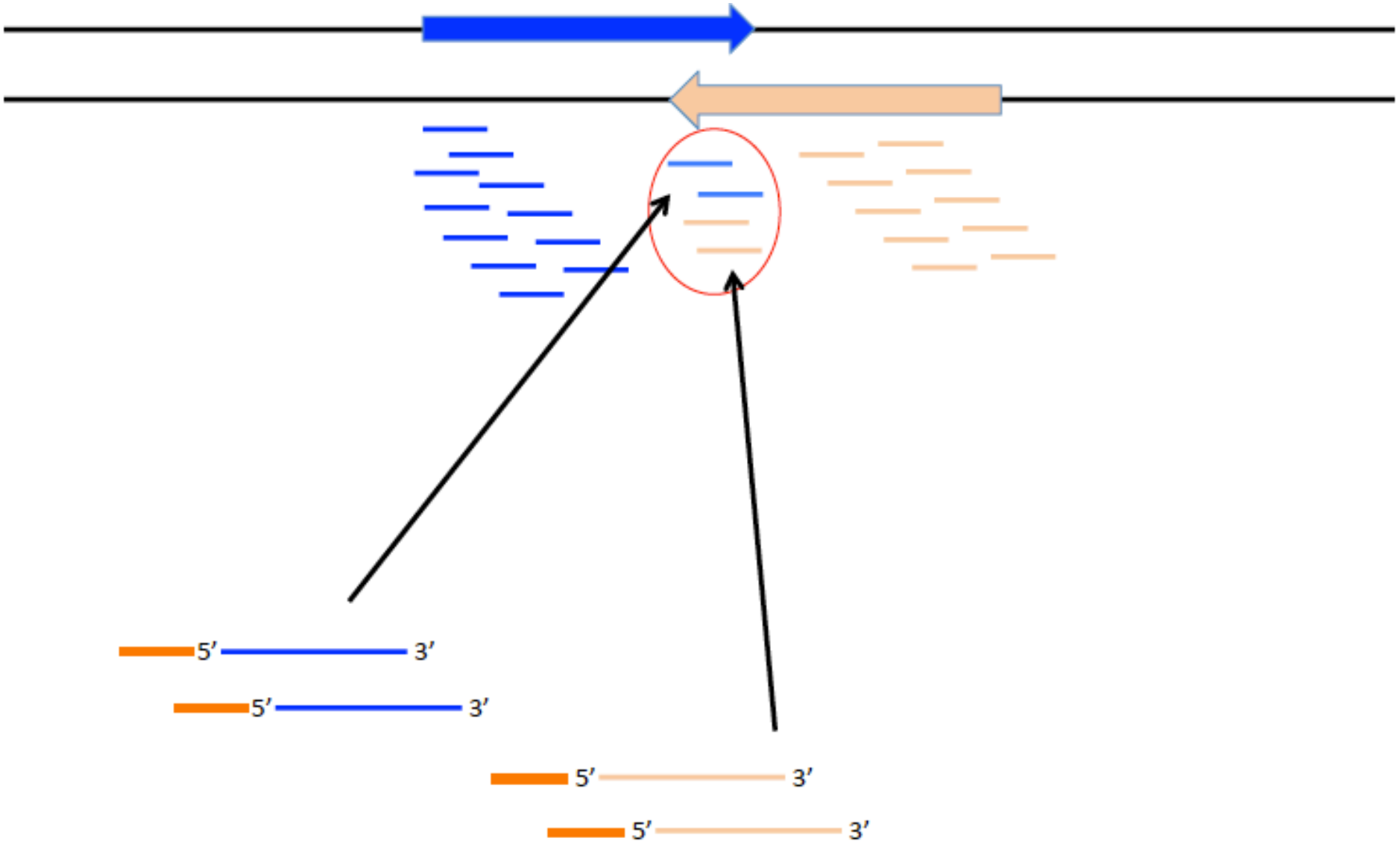




- With this method, during retro-transcription the strand identical to the RNA sequence is marked and destroyed, after adapter ligation.
- This allow the new synthesis of the complementary strand, and the adapter used as primer for sequencing will be on the transcription specific strand


5′
3′
Fragmentation
1st (dUTP incorporation)
and 2nd strand synthesis
Second strand degradation and adapter ligation

EXPERIMENTAL DESIGN



FUTURE DEVELOPMENTS

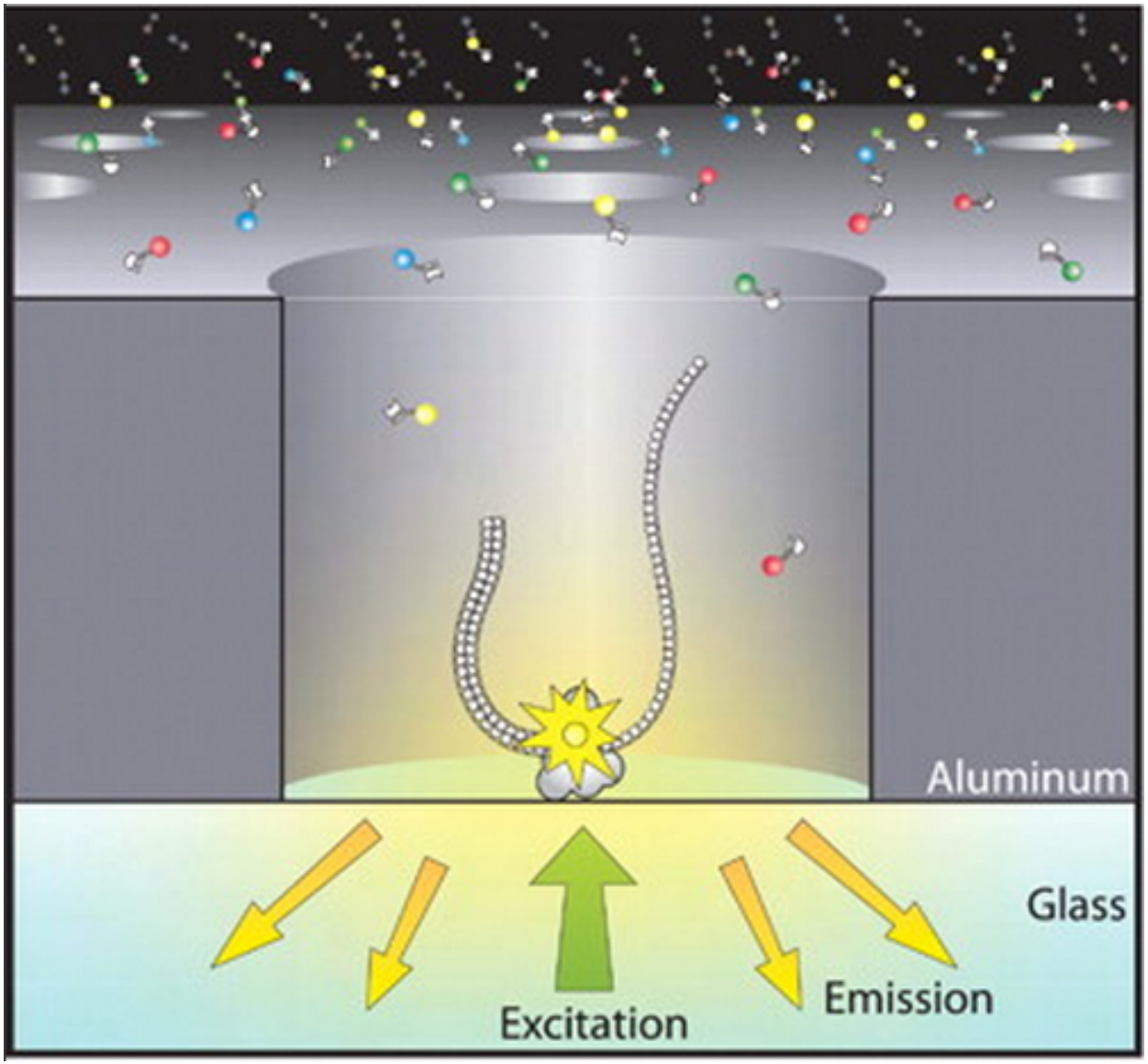
Illumina sequencing, as well as other sequencing platforms, is limited by some factors:

- The need for amplification of each fragment, which can introduce biases
- The short length of the reads, that impairs the reconstruction and expression estimate of splicing variants

PACIFIC BIOSCIENCES

- The PacBio RS platform, from Pacific Biosciences of California, Inc (PacBio), is the first example of single molecule sequencing technology, called SMRT, *single-molecule real-time sequencing*
- It can produce very long reads (up to 30 kbp), therefore allowing the sequencing of full-length cDNAs

PACIFIC BIOSCIENCES



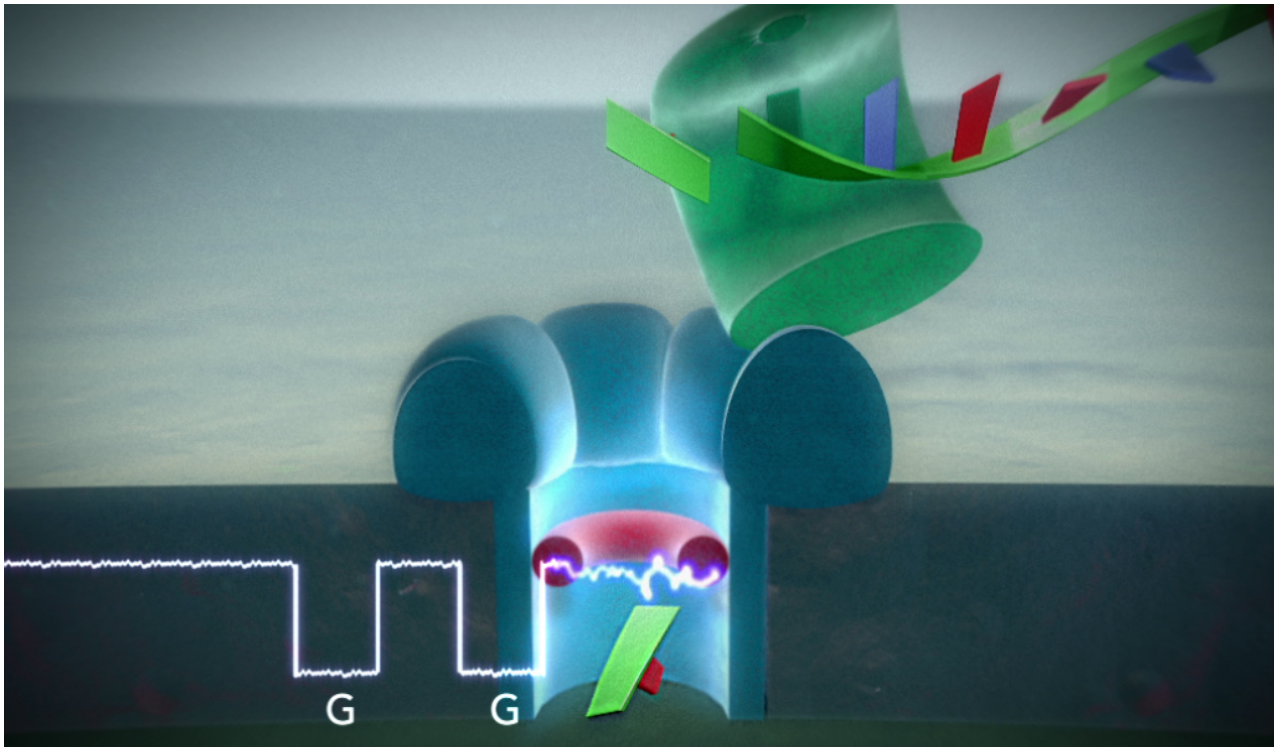
PACIFIC BIOSCIENCES

The PacBio still has relevant limits:

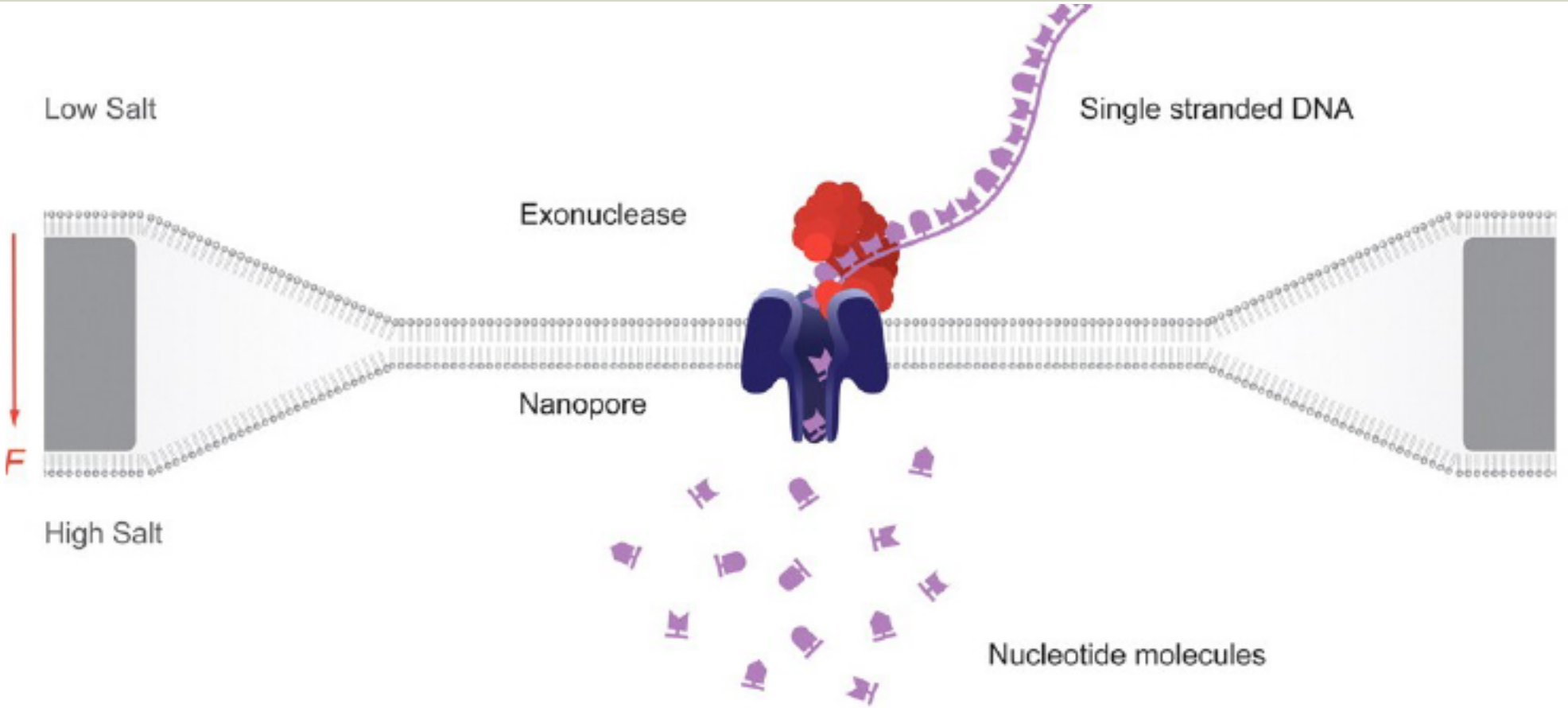
- It is expensive and not easy to use and maintain
- Error rate is quite large, around 19%, mostly insertions and deletions. To minimize this, DNAs are circularized and sequenced many times, in order to correct errors
- The throughput is relatively low, thus not allowing the expression estimate but only the transcript identification. To overcome this aspect, PacBio transcriptome sequencing has been coupled with Illumina sequencing: the Illumina reads can be mapped on the PacBio sequencing to obtain expression estimates

OXFORD NANOPORE

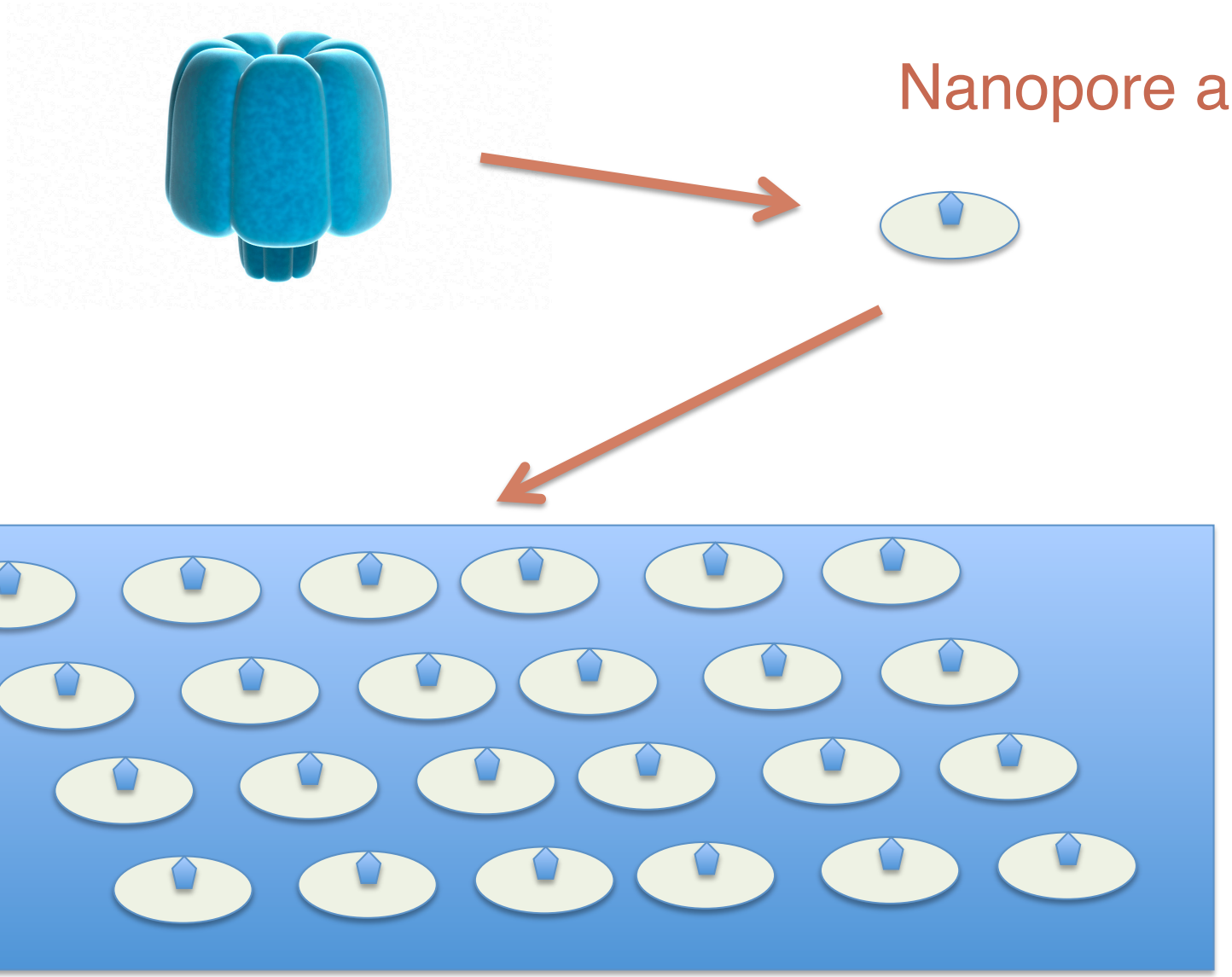
- Another issue is that we cannot sequence RNA directly, but it must always be converted into a cDNA
- The nanopore-based technologies are instead able to sequence full-length RNA, without needing to retro-transcribe or amplify it



OXFORD NANOPORE

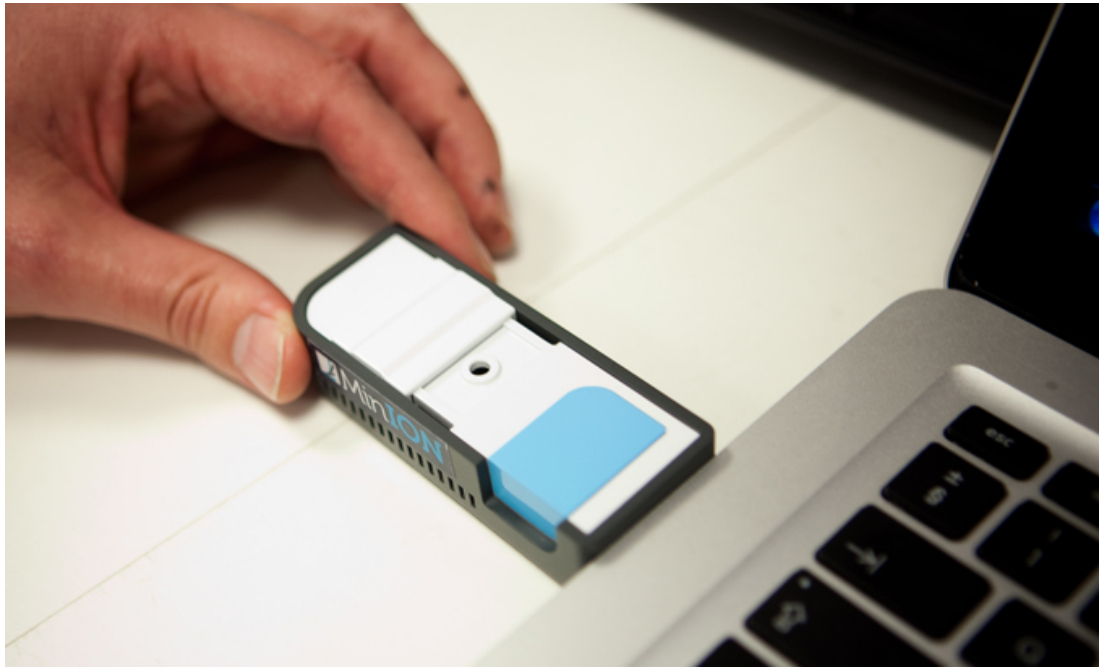


OXFORD NANOPORE



OXFORD NANOPORE

MinION



OXFORD NANOPORE

MinION

- 150Mb for run
- Reads up to 48kb
- 500 pores for cartridge
- $900 for single-usage cartridge

OXFORD NANOPORE

The MinION on the ISS

OXFORD NANOPORE

GridION

- Cartridge array
- From 2000 to 8000 pores

PromethION

- 48 flow cells
- Each flow cell contains 3000 nanopores
- In total, 144000 pores
- Still a prototype