


DIFFERENT WAYS TO APPROACH BIOLOGICAL QUESTIONS

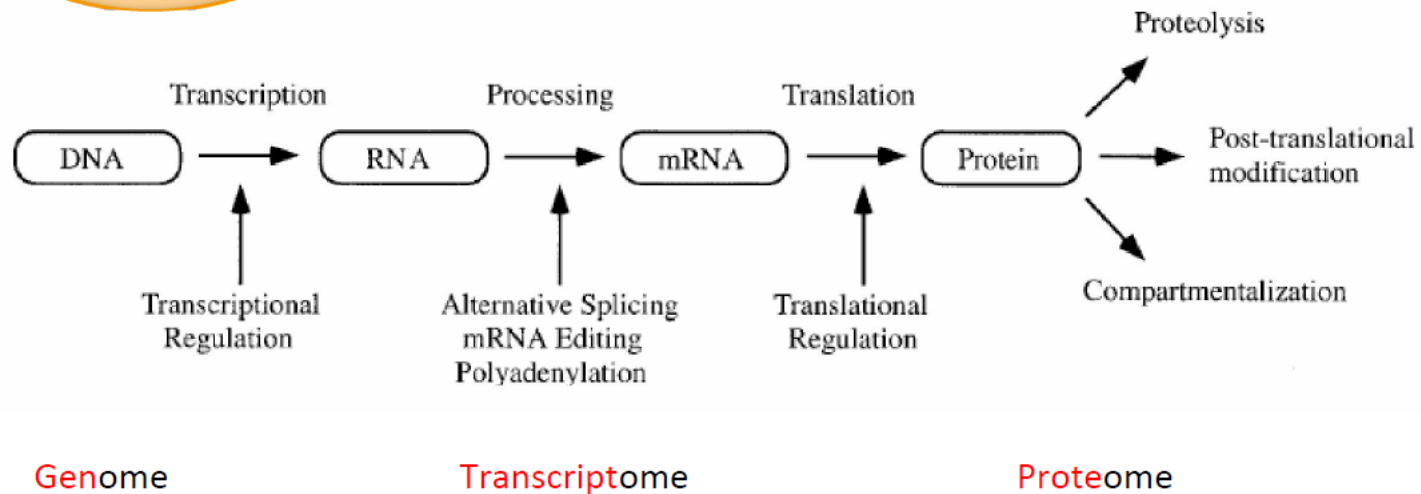
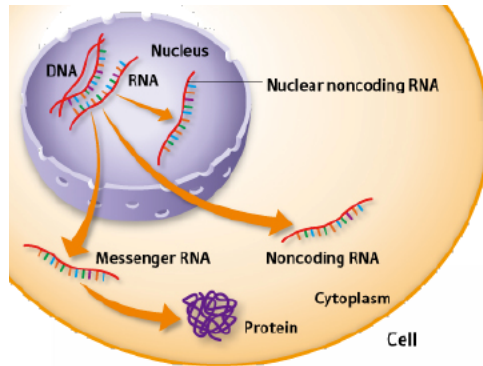
Pre-NGS era



Timeline

- 
- A large, vertical purple arrow pointing downwards, indicating the progression of time from the 1970s to the present.
- '70es - **Recombinant DNA technology**, beginning of gene cloning and sequencing → "*one-gene-at-the-time*"
 - 1975 - "**Chain termination**" **method of DNA sequencing** (Frederick Sanger, Nobel Prize in 1980, shared with Paul Berg and Walter Gilbert)
 - '80es - Polymerase Chain Reaction, or **PCR** (Kary Mullis, Nobel Prize in 1993).
- Cloning of the CFTR gene (results in 1989) → "*Unity is strength*".
 - 1990 - Beginning of the Human Genome Project (**HGP**). Estim. time: 15 yrs
 - 1995 - Completion of the first genome (*Haemophilus influenzae*, prokaryote, 1.8 Mb)
 - 1996 - Completion of the second genome (*Saccharomyces cerevisiae*, eukaryote, 12.2Mb)
 - 2001 - First draft of the human genome
 - 2003 - Completion of the **HGP**: 13 yrs, 3Gb, \$3-billion project
 - today - **Post-Genomic Era**

STEPS THAT ARE ANALYZED BY -OMICS



TRANSCRIPTOME ANALYSIS: WHY?

Issues in the studies on Transcriptome

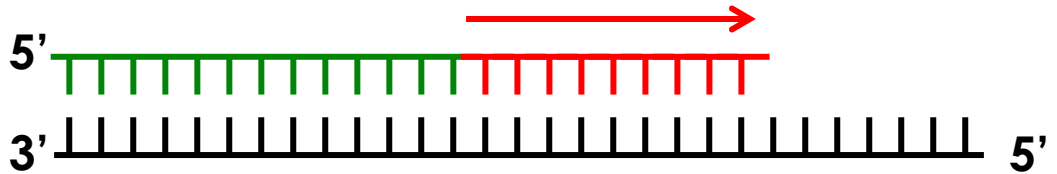
■ The Transcriptome of a cell is a dynamic entity: unlike the Genome, it constantly changes.



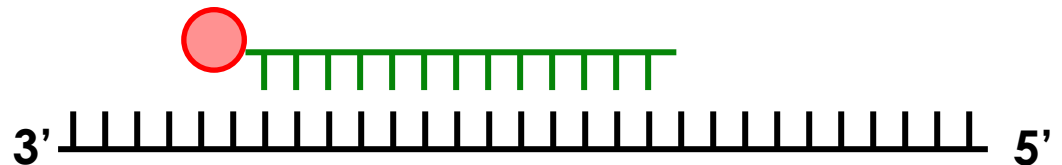
TRANSCRIPTOME ANALYSIS: WHY?

How to detect something that is unknown?

PCR / qPCR /
classic sequencing

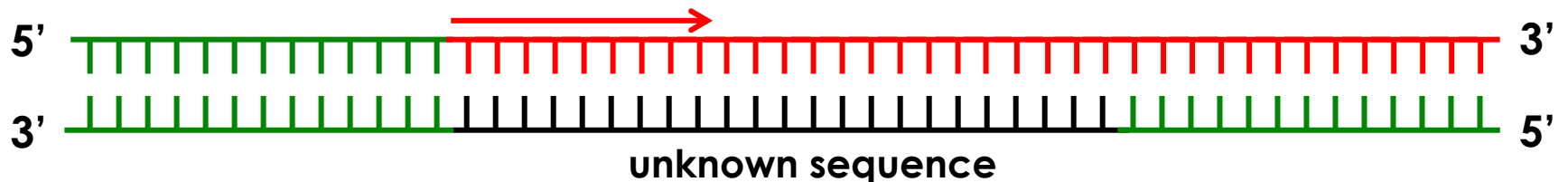


Northern blot /
Southern blot



We need to make detectable something that is not known

Next-Generation Sequencing (NGS)



unknown sequence

SANGER METHOD FOR DNA SEQUENCING

History of Sequencing: Sanger method for DNA sequencing

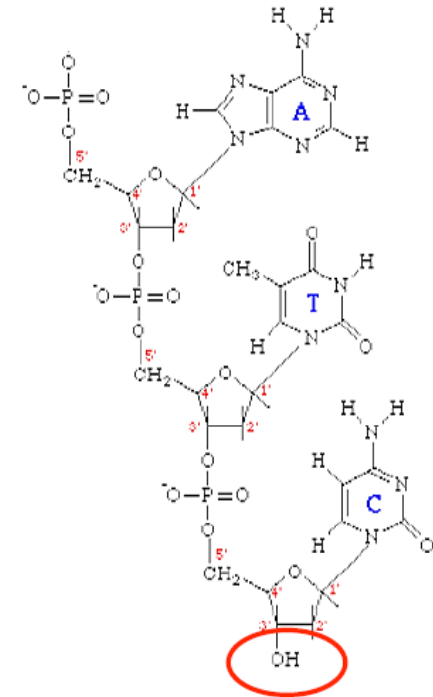
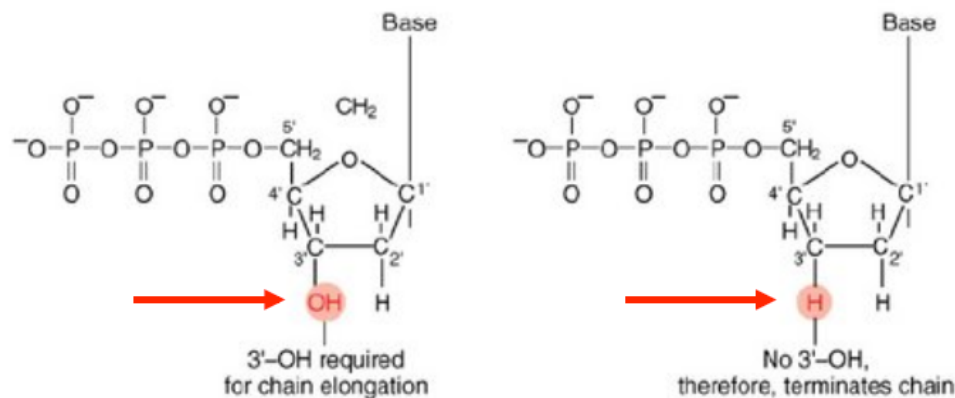
DNA Polymerase can add free nucleotides only to the 3' end of the newly forming strand. This results in elongation of the newly forming strand in a 5'-3' direction. No known DNA polymerase is able to begin a new chain (de novo). DNA polymerase can add a nucleotide only on to a pre-existing 3'-OH group, and, therefore, needs a primer at which it can add the first nucleotide.

DNA Polymerase

5' – TGAGACGAATCGATGCGGACGGATCGATTGATCTGATCGATGCATT
3' – ACTCTGCTTAGCTACGCCTGCCTAGCTAAGCTAGACTAGCTACGTAA – 5'

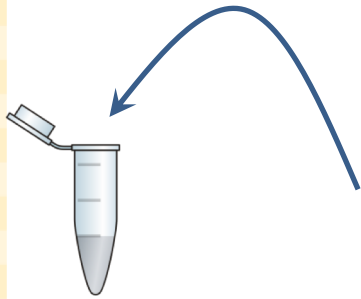
SANGER METHOD FOR DNA SEQUENCING

- **"Sanger Sequencing"** developed by Fred Sanger *et al.* in the mid 1970's
- Uses dideoxynucleotides for "chain termination", generating fragments of different lengths ending in ddATP, ddGTP, ddCTP or ddTTP



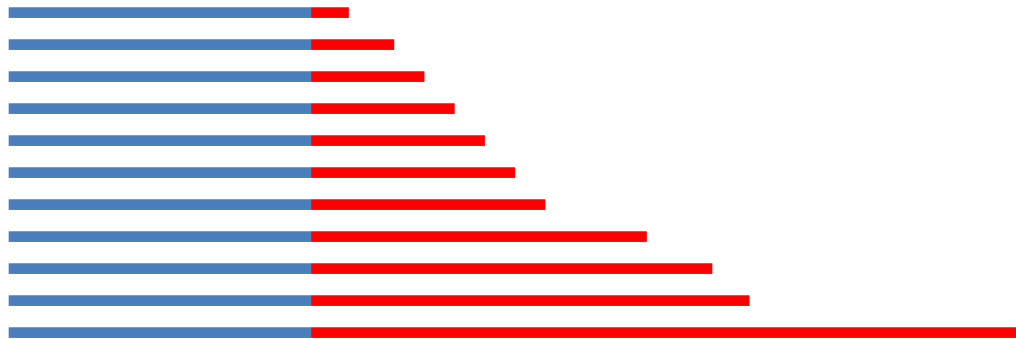
http://openwetware.org/wiki/BE.109:Bio-material_engineering/Sequence_analysis

SANGER METHOD FOR DNA SEQUENCING



- Template DNA
- DNA Polymerase
- Primer
- dATP, dCTP, dGTP, dTTP
- **ddATP** (or ddCTP, ddGTP, ddTTP)

→ **A**T**A**A**A**A**A**CTC**A**G**A**ACGGCTTCG**T**A
GACTGACTGACTATTTTTT**G**AGTCTT**G**CCGAAGCAT



SANGER METHOD FOR DNA SEQUENCING

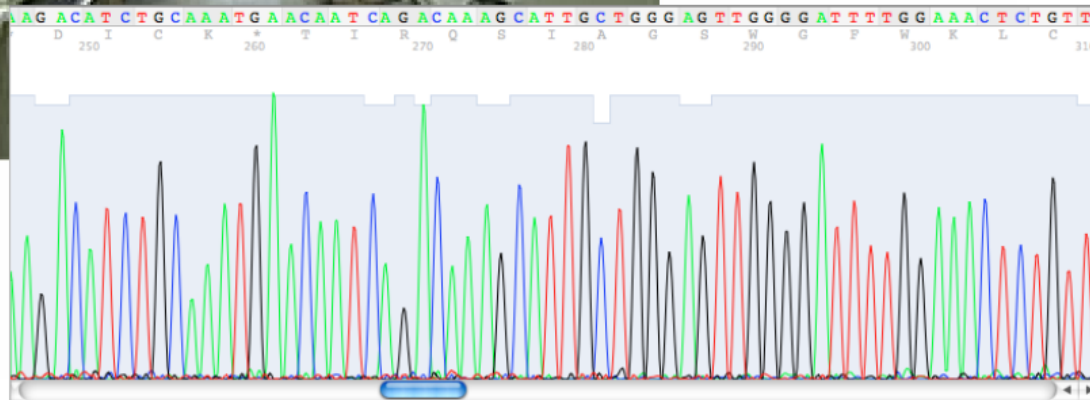


SANGER METHOD FOR DNA SEQUENCING

Automated Sequencing

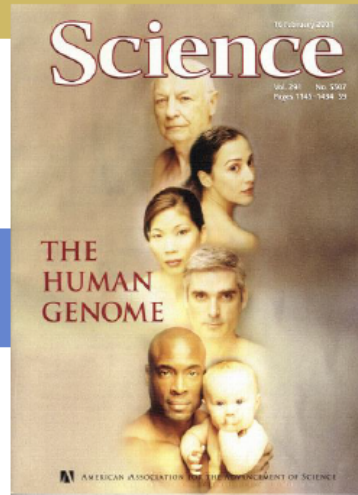


- Sequencing technology was improved in the late 1980s by Leroy Hood who developed fluorescent color labels for the 4 terminator nucleotide bases.
- This allowed all 4 bases to be sequenced in a single reaction and sorted in a single gel lane



HUMAN GENOME PROJECT

Celera Genomics



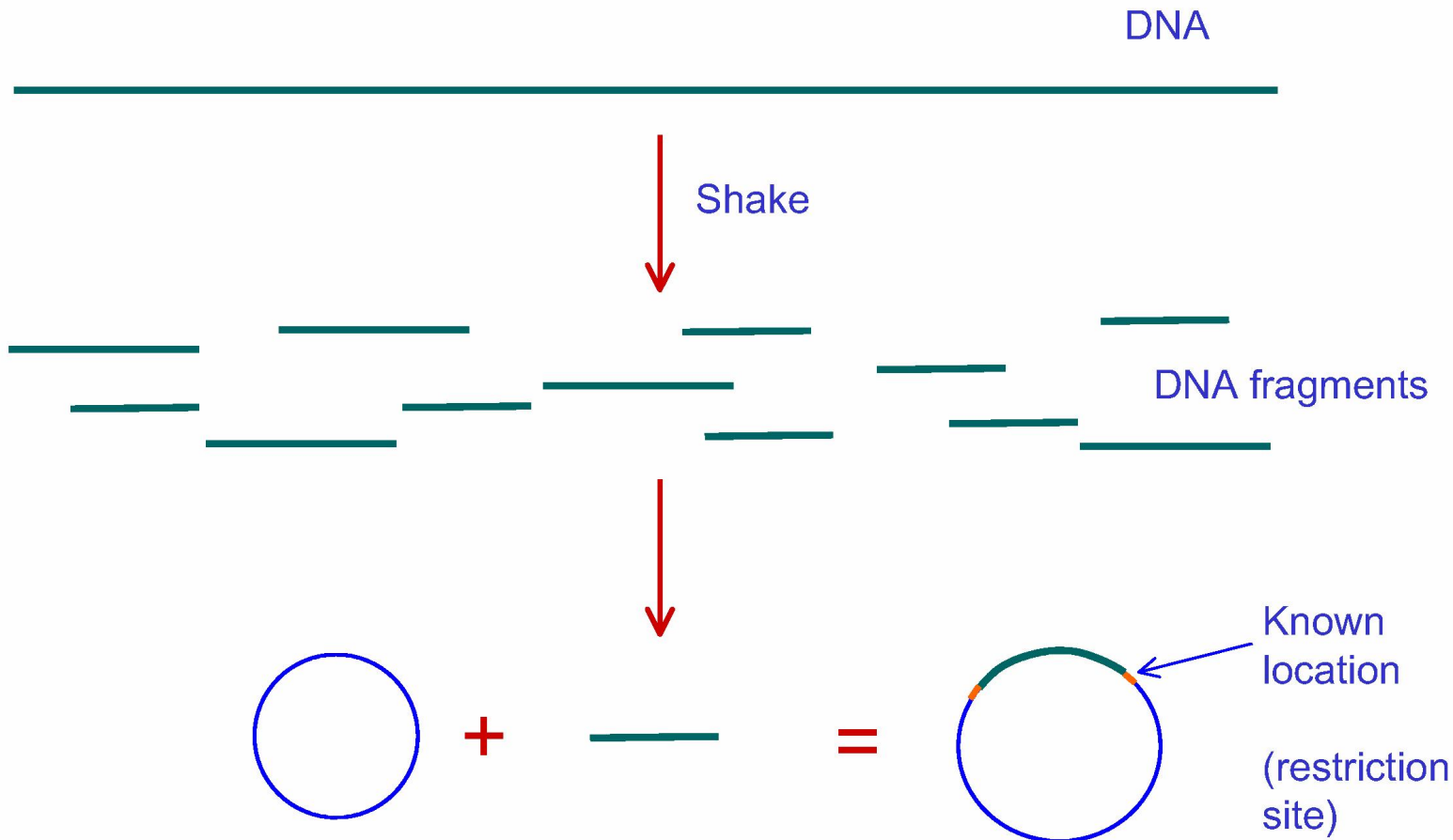
- Private company, founded in 1998
- Beginning of sequencing in 1998
- Patented sequence: **no public access to data**
- Estimated cost: 300 Million \$



International Consortium

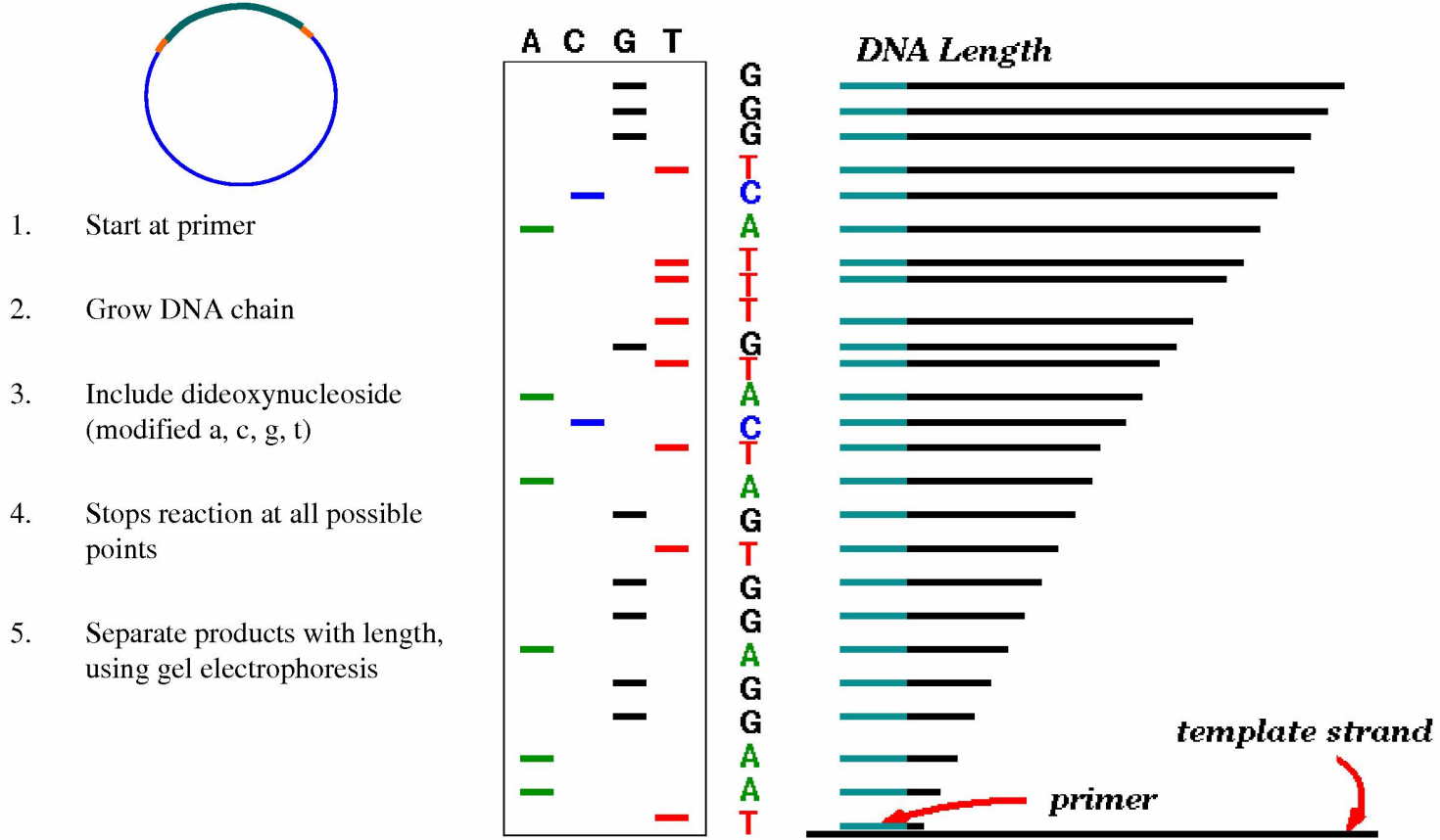
- 20 groups from USA, UK, Japan, France, Germany and China
- Over 1000 scientists involved
- Beginning of sequencing in 1991
- Open-access sequence: **immediate and free release of data**
- Estimated cost: 450 Million – 3 Billion \$

HUMAN GENOME PROJECT: SHOTGUN SEQUENCING



source: robotics.stanford.edu/~serafim/cs262/Spring2003/Slides/Lecture9.ppt


HUMAN GENOME PROJECT: SHOTGUN SEQUENCING



- Can produce DNA fragments 700-900bp long, but it's slow
- Lots of other problems including clone library generation and low-throughput
- The Human Genome Project used Sanger sequencing, completion took over 10 years

SEQUENCING A HUMAN GENOME (3,2 BILLION BP)

Costs and time for sequencing a human genome (3,2 billion bp)

2001	First human genome	13 years	300 million \$
2005	Technology review	6 months	20-30 million \$
2005	454	1 month	900'000 \$ (1X coverage)
2009	Solexa (Illumina)	6 months	50'000 \$ (30X coverage)
2010	Illumina		19'500 \$ (30X coverage)
			
2015	Personalized medicine	1000 \$/genome	

What is it?

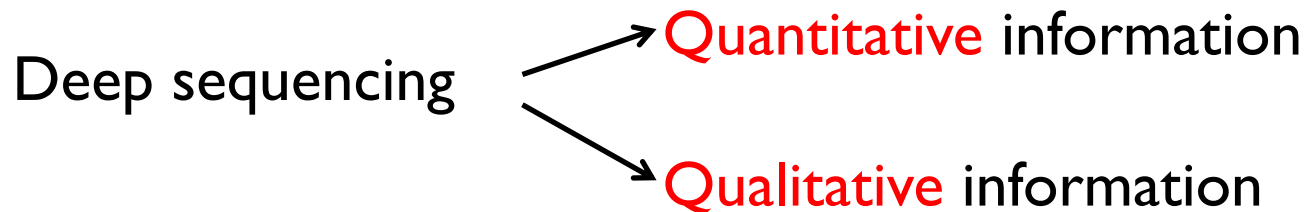
■ Set of new high throughput technologies:

- Allow millions of short DNA sequences from a biological sample to be “read” or sequenced in a rapid manner
- Computational power is then used to assemble or align the “reads” to a reference genome, allowing biologists to make comparisons and interpret various biological phenomena

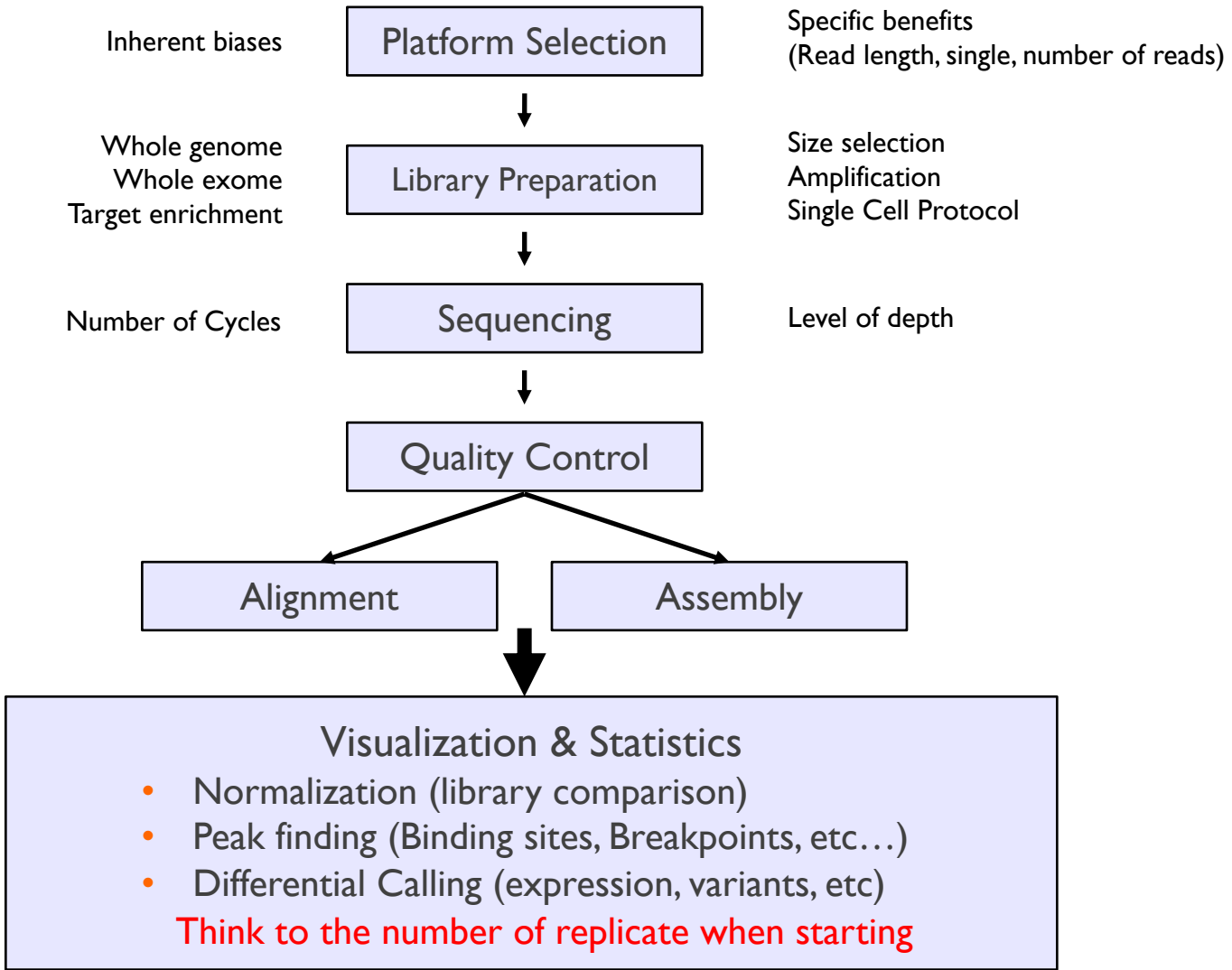
■ Due to high depth of coverage (30-100x), accurate sequencing is obtained much faster and cheaper compared to traditional Sanger/Shotgun sequencing

NEXT GENERATION SEQUENCING

- ◆ Mutation and SNP identification or analysis (genome re-sequencing)
- ◆ Gene/Disease Linkage (genome re-sequencing)
- ◆ Pathogen identification (*de novo* sequence assembly or re-sequencing)
- ◆ Transcriptome analysis (RNAseq)
- ◆ DNA methylation study (medip-seq)
- ◆ Chromatin study (ChIPseq)
- ◆ Transcription factor study (ChIPseq)
- ◆ miRNAs, siRNA, piRNA, tRF, etc... (small RNA seq)
- ◆ Single cell transcriptome analysis

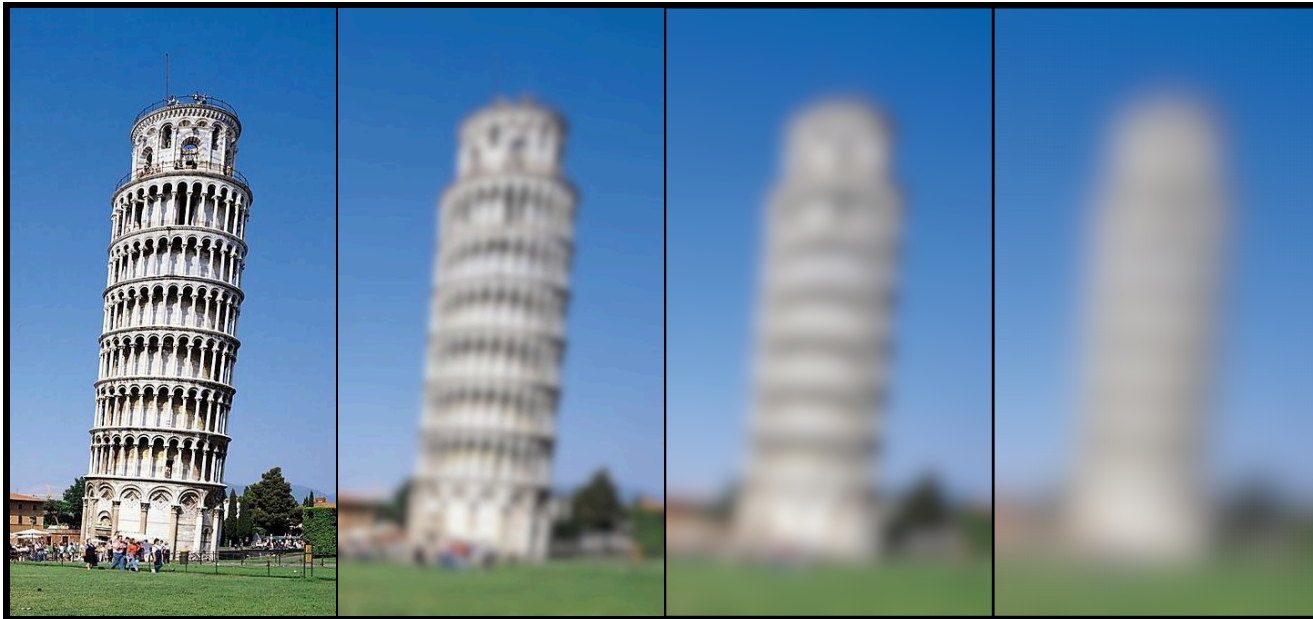


NEXT GENERATION SEQUENCING



NEXT GENERATION SEQUENCING

Sequencing Depth



High resolution



many information

Low resolution



few information

What is RNA-seq?

- RNA-seq is essentially **massively parallel sequencing of RNA** (or, in fact, the corresponding cDNA) and has heralded the second technical revolution in transcriptomics.

- It is **based on next-generation sequencing (NGS) platforms** that were initially developed for high-throughput sequencing of genomic DNA.

- Typically, **all the RNA molecules in a sample are reverse transcribed into cDNA**, and depending on the platform to be used, the **cDNA molecules may (amplification-based sequencing) or may not (single-molecule sequencing (SMS)) be amplified before deep sequencing.**

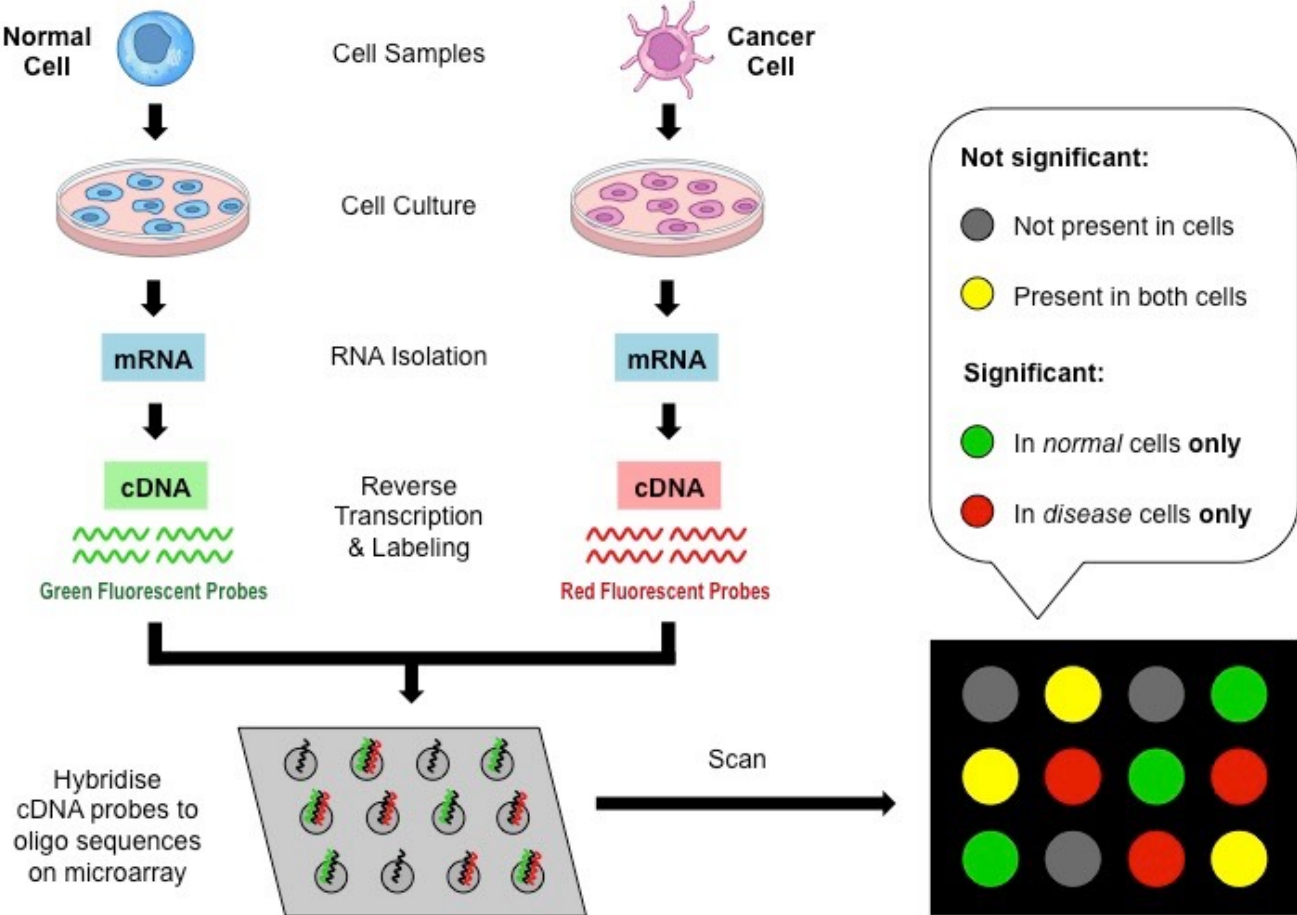
- After the sequencing reaction has taken place, **the obtained sequence stretches (reads) are mapped onto a reference genome** to deduce the structure and/or expression state of any given transcript in the sample.



RNA-Seq provides the ability to look at:

- changes in gene expression
- alternatively spliced transcripts, alternative promoters and polyA sites
- post-transcriptional changes
- gene fusions
- In addition to mRNA transcripts, RNA-Seq can look at different populations of RNA (tRNA, miRNA)
- exon/intron boundaries
- verify or amend previously annotated 5' and 3' gene boundaries.

RNA-Seq VS Microarray (what is it?)



RNA-Seq VS Microarray

RNA-Seq has a wider dynamic range, which depends on the sequencing depth. Microarrays show saturation at high expression levels and loss of signal at low expression levels.

RNA-Seq is more sensitive than microarrays: it is able to identify more genes.

RNA-Seq is able to **identify and quantify novel splicing variants**.

RNA-Seq allows to **identify new SNPs and editing**.

Microarray are cheaper and easier to analyze.

Arrays still have a place for targeted identification of already known common allele variants, making them ideal for regulatory diagnostics.

<https://bioinfomagician.wordpress.com/2014/01/28/rna-seq-vs-microarray-what-is-the-take/comment-page-1/>

Variants of RNA-Sequencing

Traditional RNA-Seq

It allows to quantify the expression of:

- mRNAs and other polyadenilated RNAs (polyA+)
- all RNA species except for rRNAs (RiboMinus, Ribo-Zero).

Small RNA-Seq (Illumina)

Adapters are designed so that they can bind microRNA and other small RNAs which have a 3' hydroxyl group that is the result of the cleavage by Dicer or other RNA processing enzymes.

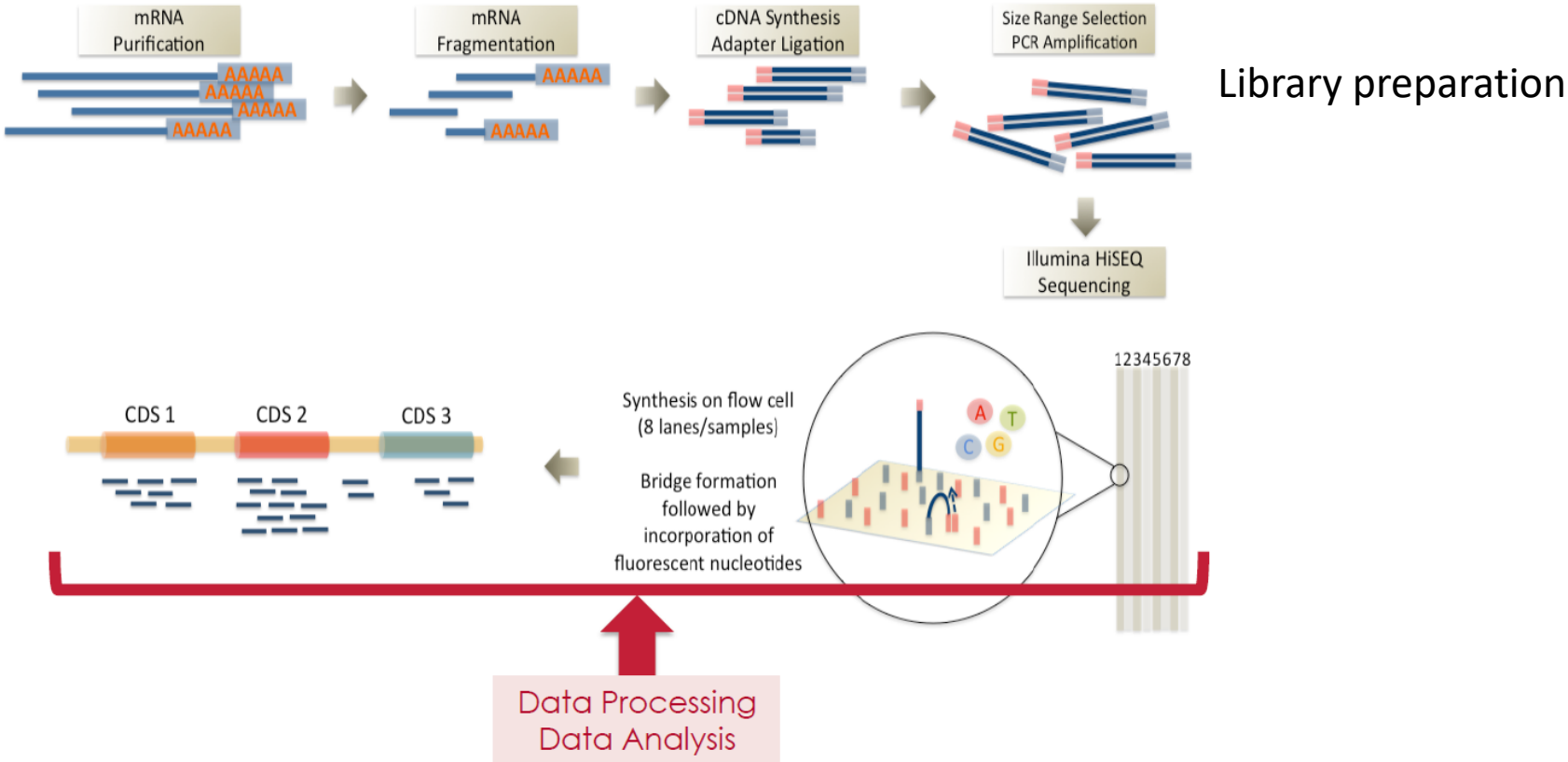
RIP-Seq and CLIP-Seq

All the RNAs which are bound by a protein are sequenced (using a standard protocol), so that they can be identified. CLIP-Seq also allows to find the localization of the binding site.

GRO-Seq

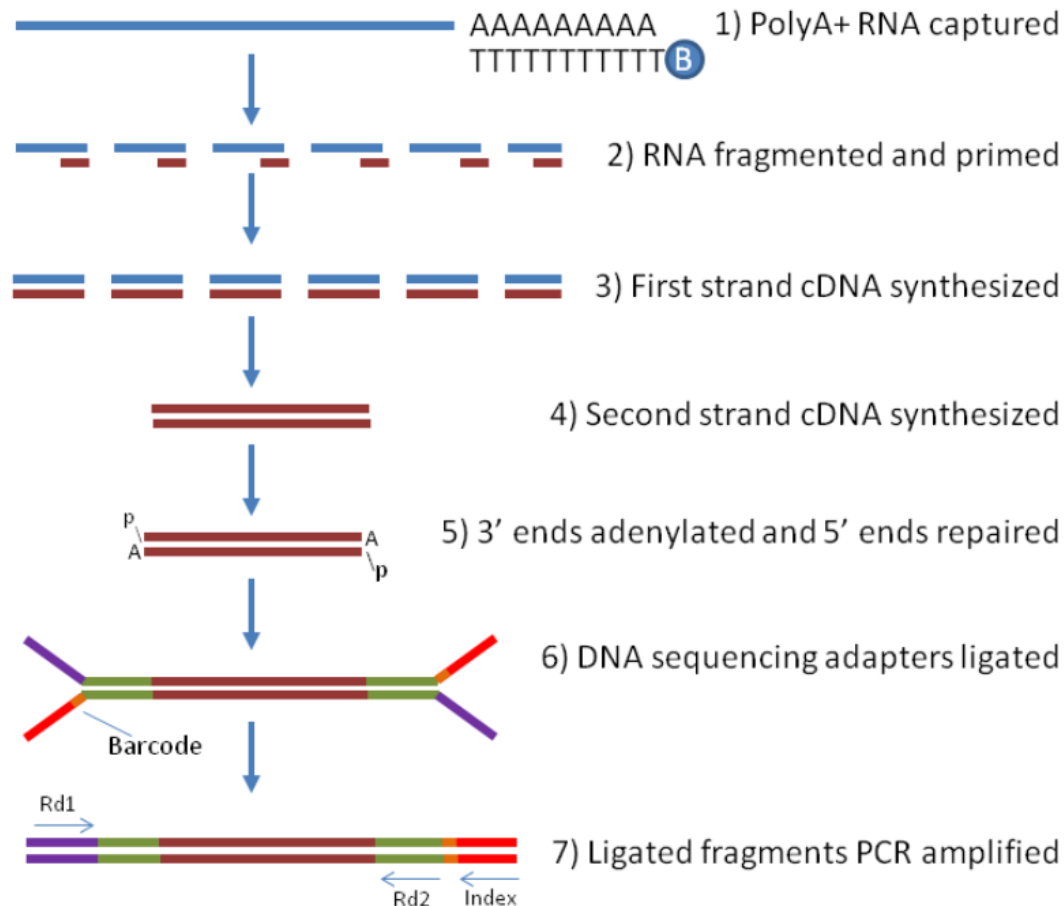
It allows to directly measure nascent RNA production.

The method



RNA-Seq: LIBRARY PREPARATION

Example of library preparation: Illumina Truseq



RNA-Seq: LIBRARY PREPARATION

coding RNAs

mRNA

RNA-seq

non-coding RNAs

large

- rRNA
- Xist
- lincRNA
- Pseudogenes
- circular RNAs

small

- tRNA translation
- snRNAs splicing
- snoRNAs modification
- scRNAs transl. control
- gRNAs editing
- miRNAs transl. control
- siRNAs RNA stability
- rasiRNAs chromatin
- piRNAs genome stability

Small RNA-seq

rRNA + tRNA → ~ 95%

Two ways to isolate long RNA molecules:

■ **1a - Purify and Fragment mRNA**

This process purifies the poly-A containing RNA molecules (mainly mRNA) using poly-T oligo-attached magnetic beads.

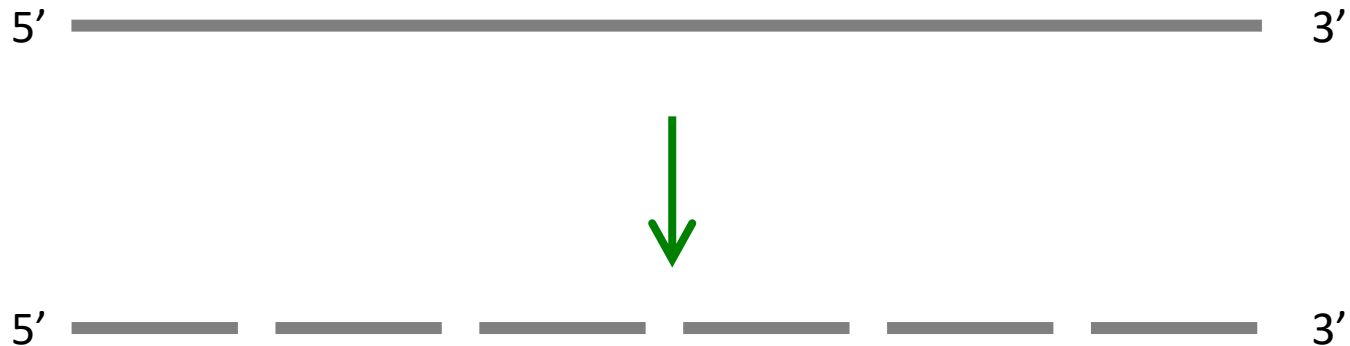
■ **1b - Remove rRNA**

After the ribosomal RNA is depleted, the remaining RNA (not only mRNA) is purified, fragmented and primed for cDNA synthesis. rRNA is removed using a hybridization-based technique.

RNA-Seq: LIBRARY PREPARATION

2 - RNA fragmentation

RNA molecules are fragmented into small pieces using divalent cations under elevated temperature



Range of fragments length: **120-225 bp**

RNA-Seq: LIBRARY PREPARATION

3 - Synthesize First Strand cDNA

This process reverse transcribes the cleaved RNA fragments that were primed with random hexamers into first strand cDNA using reverse transcriptase and random primers.



double stranded
cDNA



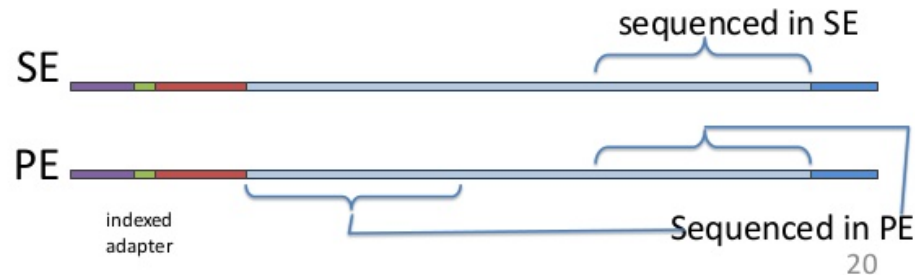
adapter ligation



RNA-Seq: LIBRARY PREPARATION

Single-end VS paired-end sequencing

- **Single-end sequencing (SE)**, involves sequencing of the fragment from only one end.
- **Paired-end sequencing (PE)**, involves sequencing both ends of a fragment, resulting in the production of read pairs. This allows to improve the alignment, to better identify and quantify splicing variants, and to detect rearrangements such as insertions, deletions, and inversions.



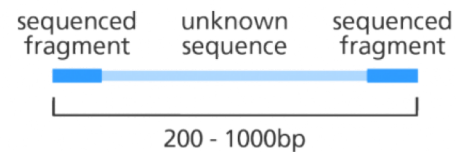
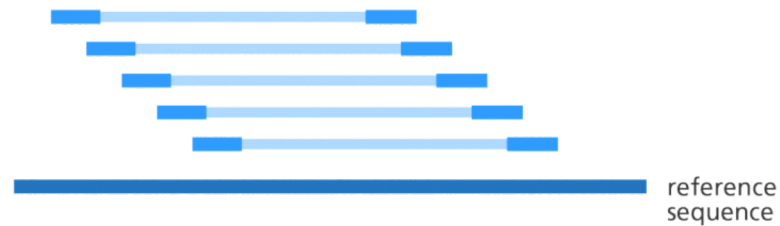
RNA-Seq: LIBRARY PREPARATION

Single-end VS paired-end sequencing

Single-end reads



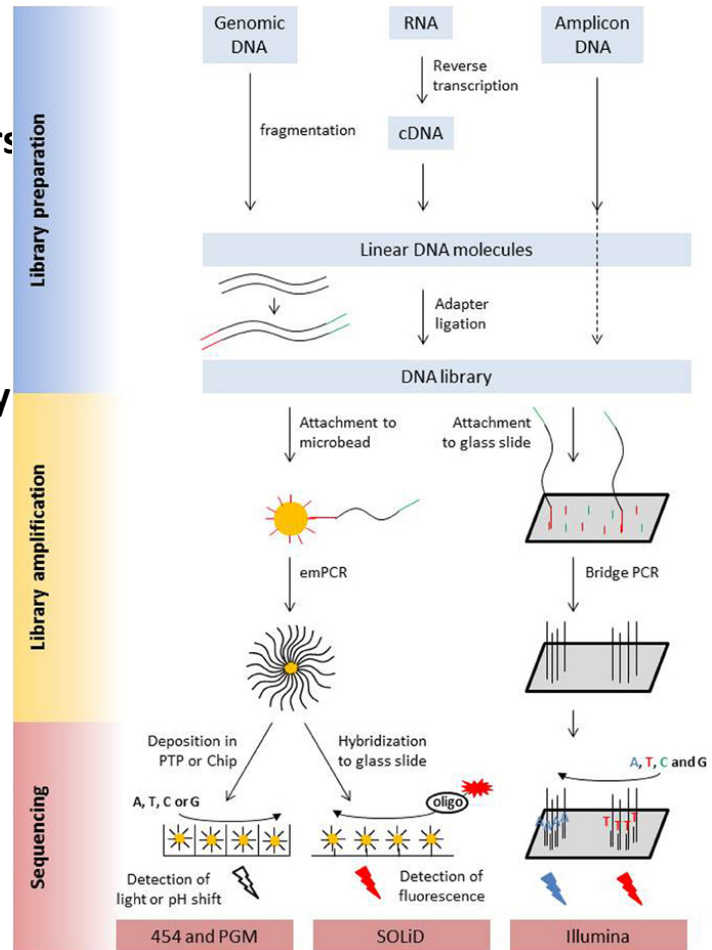
Paired-end reads



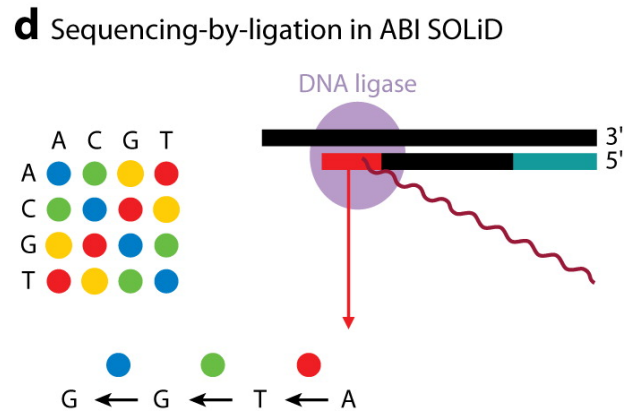
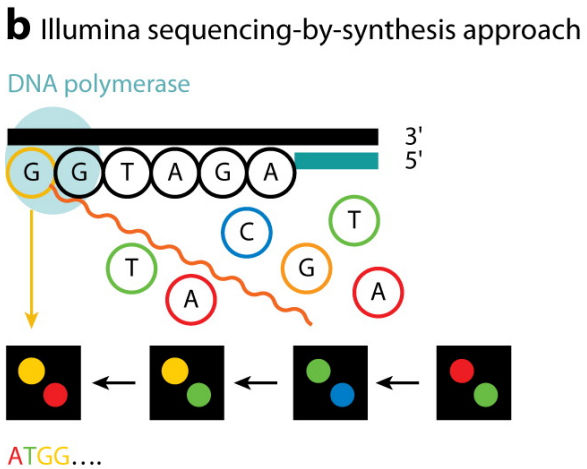
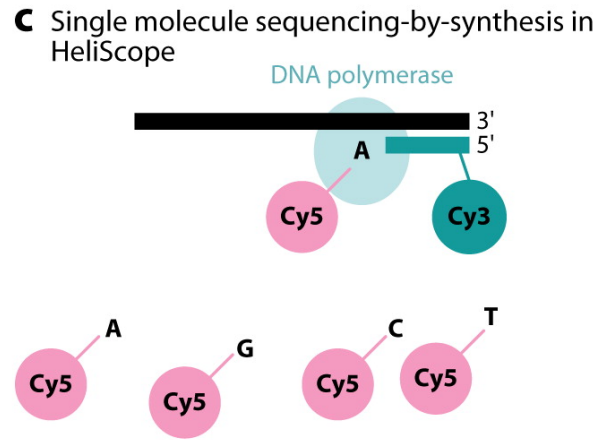
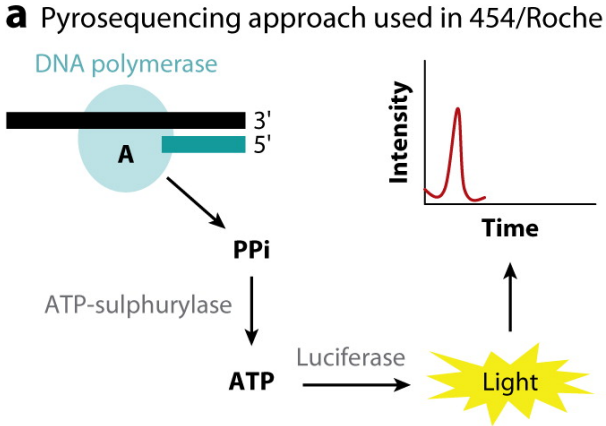
RNA-Seq: SEQUENCING REACTION

Most commonly used sequencing platforms

All different types of starting molecules are converted into doublestranded **DNA molecules that are flanked by adapters**. Adapters are sequencing platform specific and enable the binding of the library molecules to surfaces, either beads or a flow cell, where they are **amplified prior to sequencing**. Clonal amplicons are spatially separated on the glass slides, chips, or picotiterplate. Sequencing is either a **sequencing by ligation** process with fluorescently labeled oligonucleotides of known sequence (SOLiD) or a **sequencing by synthesis** process. During Illumina sequencing, four differently labeled nucleotides are flushed over the flow cell in multiple cycles, depending on the desired read length. During 454 and Ion PGM sequencing unlabeled nucleotides are flushed in a sequential order over the flow cell. Incorporation is detected via a coupled light reaction (454) or the detection of proton release during nucleotide incorporation.

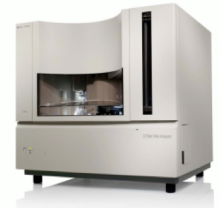


RNA-Seq: SEQUENCING REACTION



AR Morozova O, et al. 2009.
Annu. Rev. Genomics Hum. Genet. 10:135–51

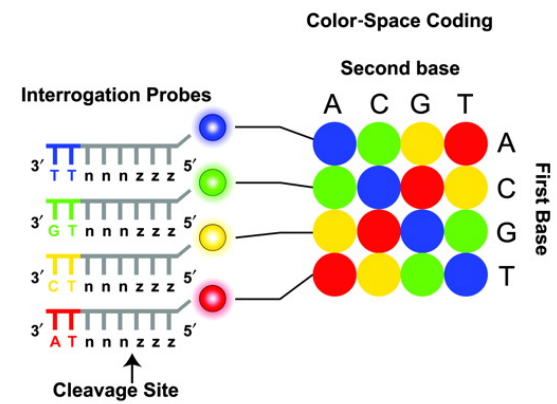
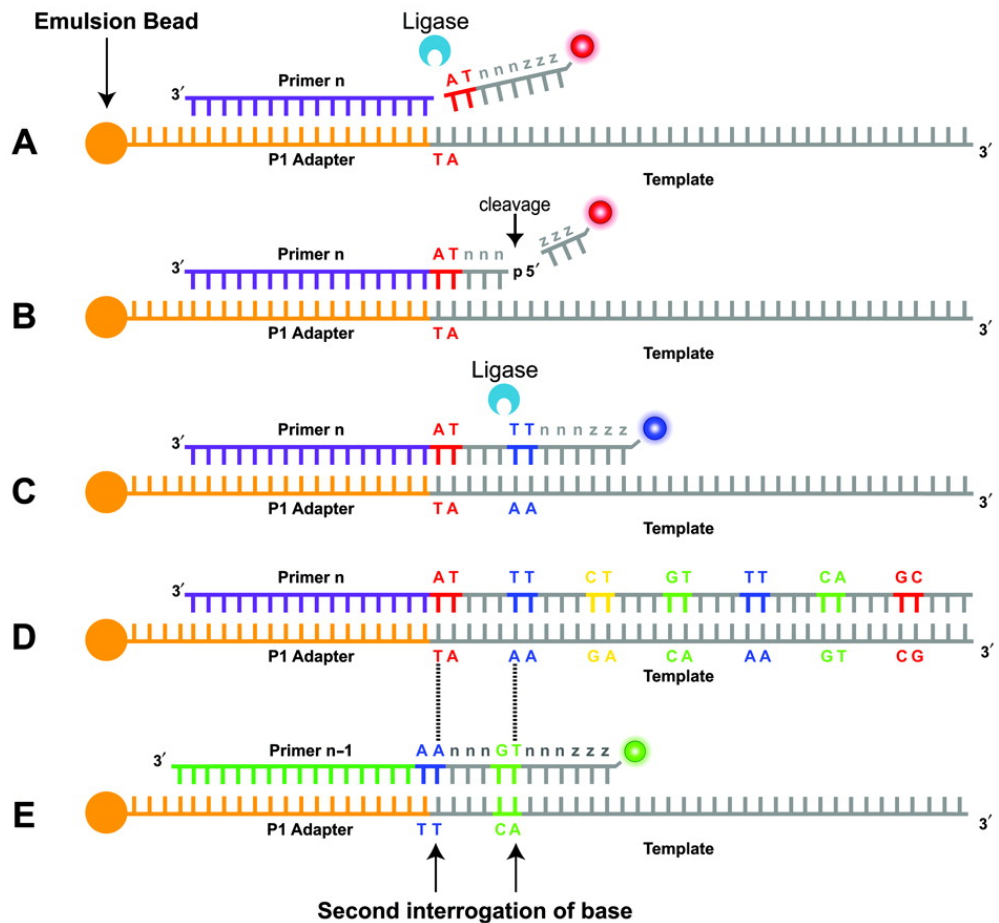
RNA-Seq: SEQUENCING REACTION



Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput
Instrument price	Instrument \$500,000, \$7000 per run	Instrument \$690,000, \$6000/(30x) human genome	Instrument \$495,000, \$15,000/100 Gb	Instrument \$95,000, about \$4 per 800 bp reaction
CPU	2* Intel Xeon X5675	2* Intel Xeon X5560	8* processor 2.0 GHz	Pentium IV 3.0 GHz
Memory	48 GB	48 GB	16 GB	1 GB
Hard disk	1.1 TB	3 TB	10 TB	280 GB
Automation in library preparation	Yes	Yes	Yes	No
Other required device	REM e system	cBot system	EZ beads system	No
Cost/million bases	\$10	\$0.07	\$0.13	\$2400

RNA-Seq: SEQUENCING REACTION

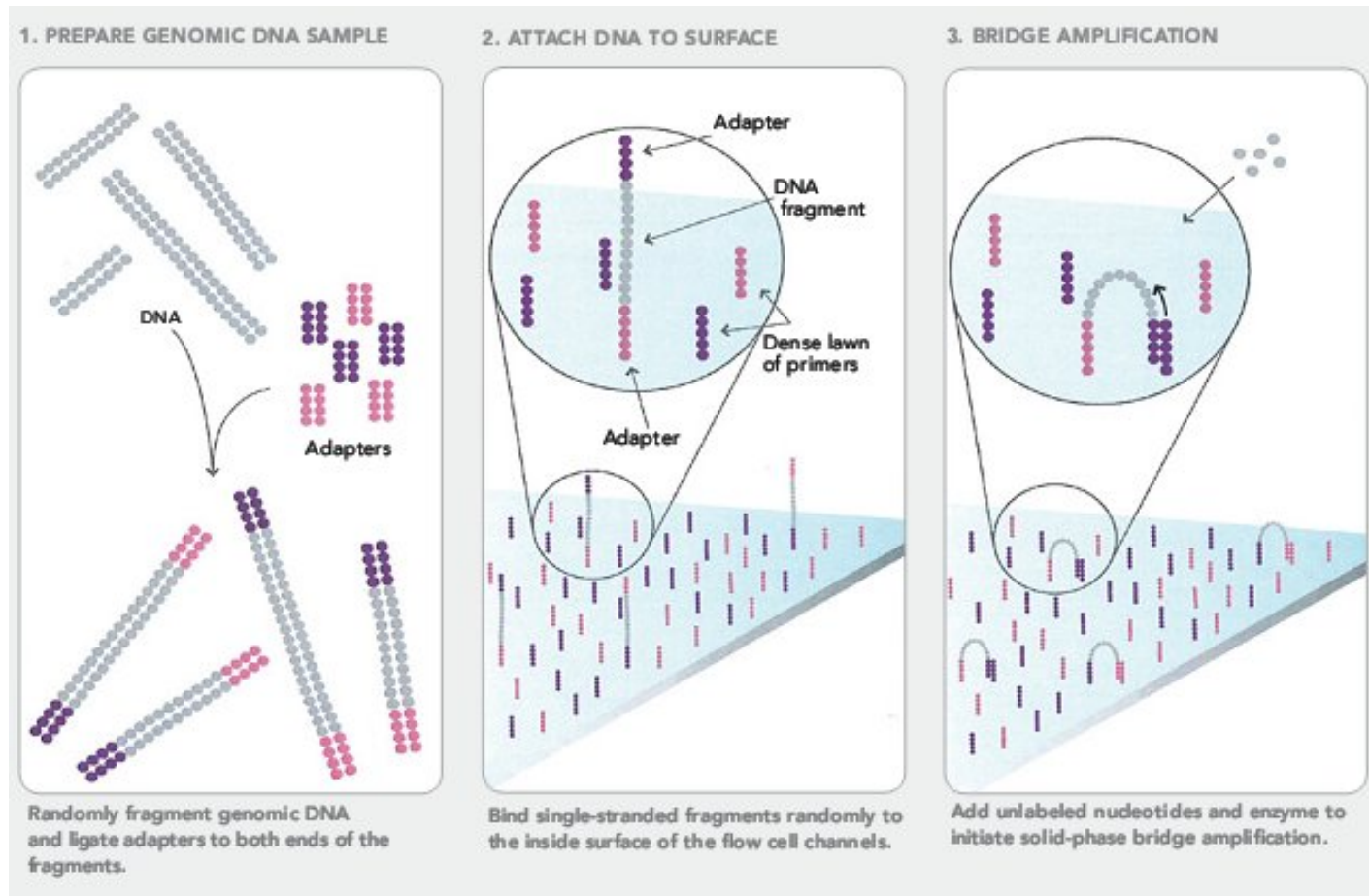
Solid platform: Sequencing by Ligation



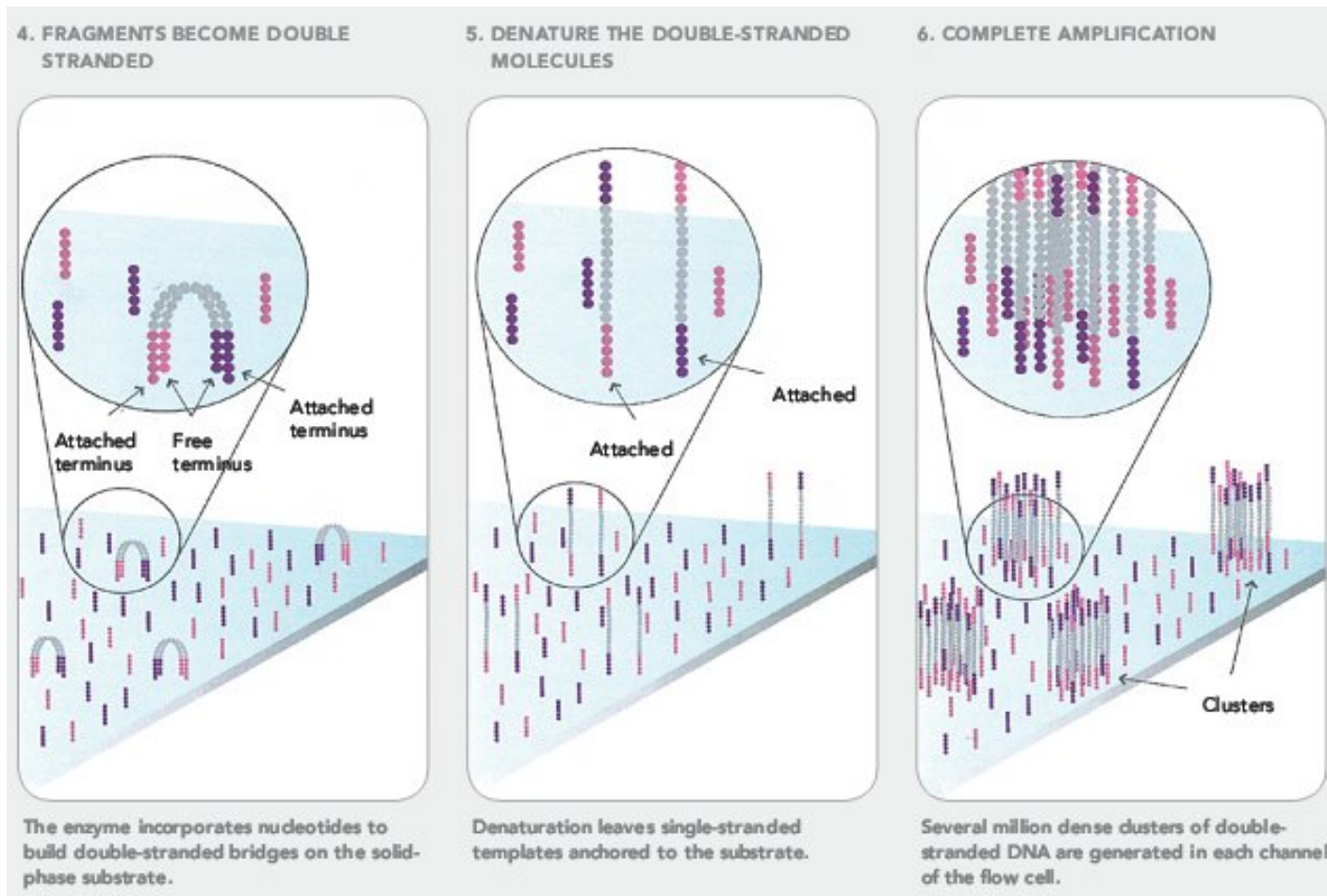
<http://www.clinchem.org/content/55/4/641/F3.expansion>

“In this technology, template bases are interrogated twice.”

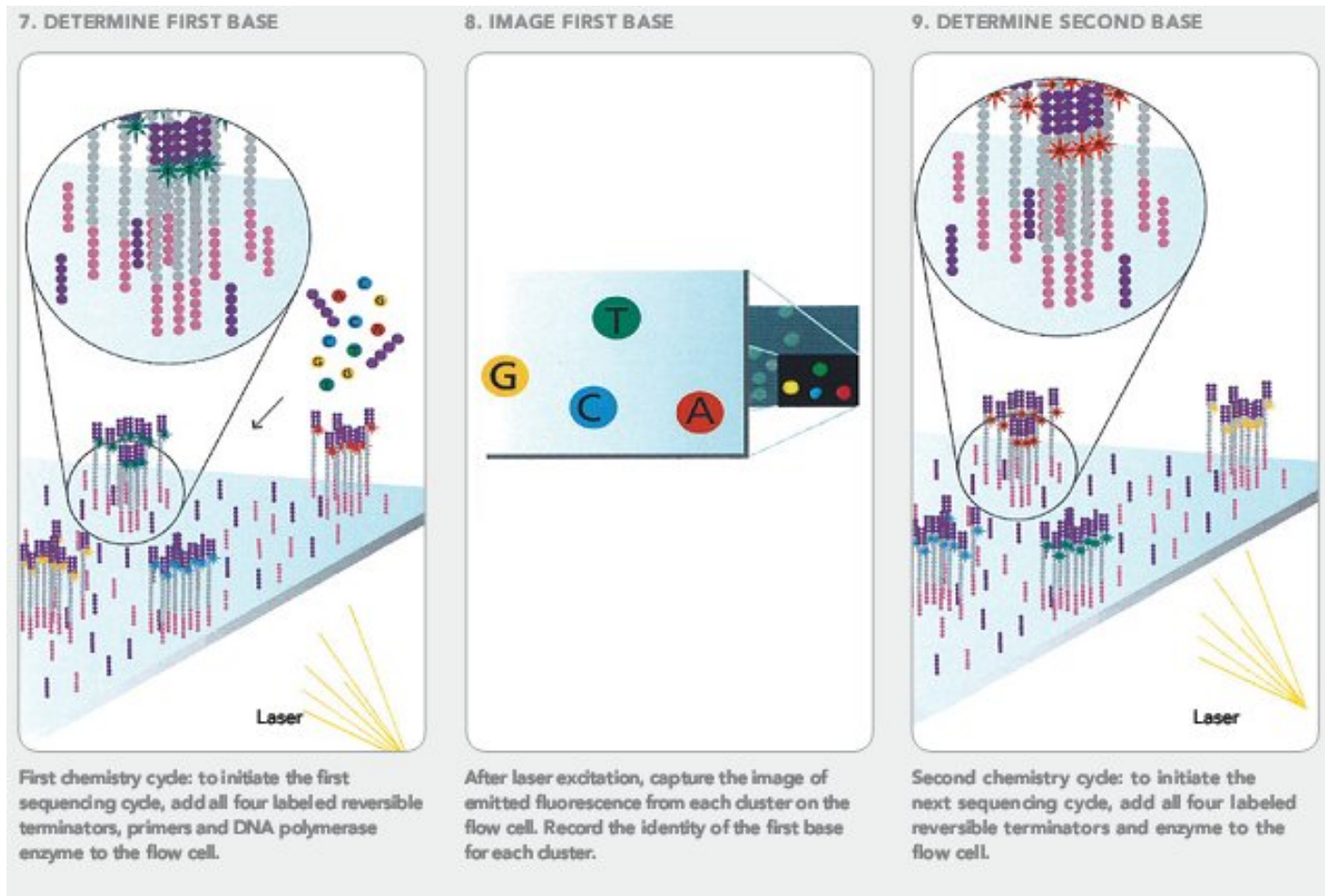
Illumina platform: Sequencing by Synthesis



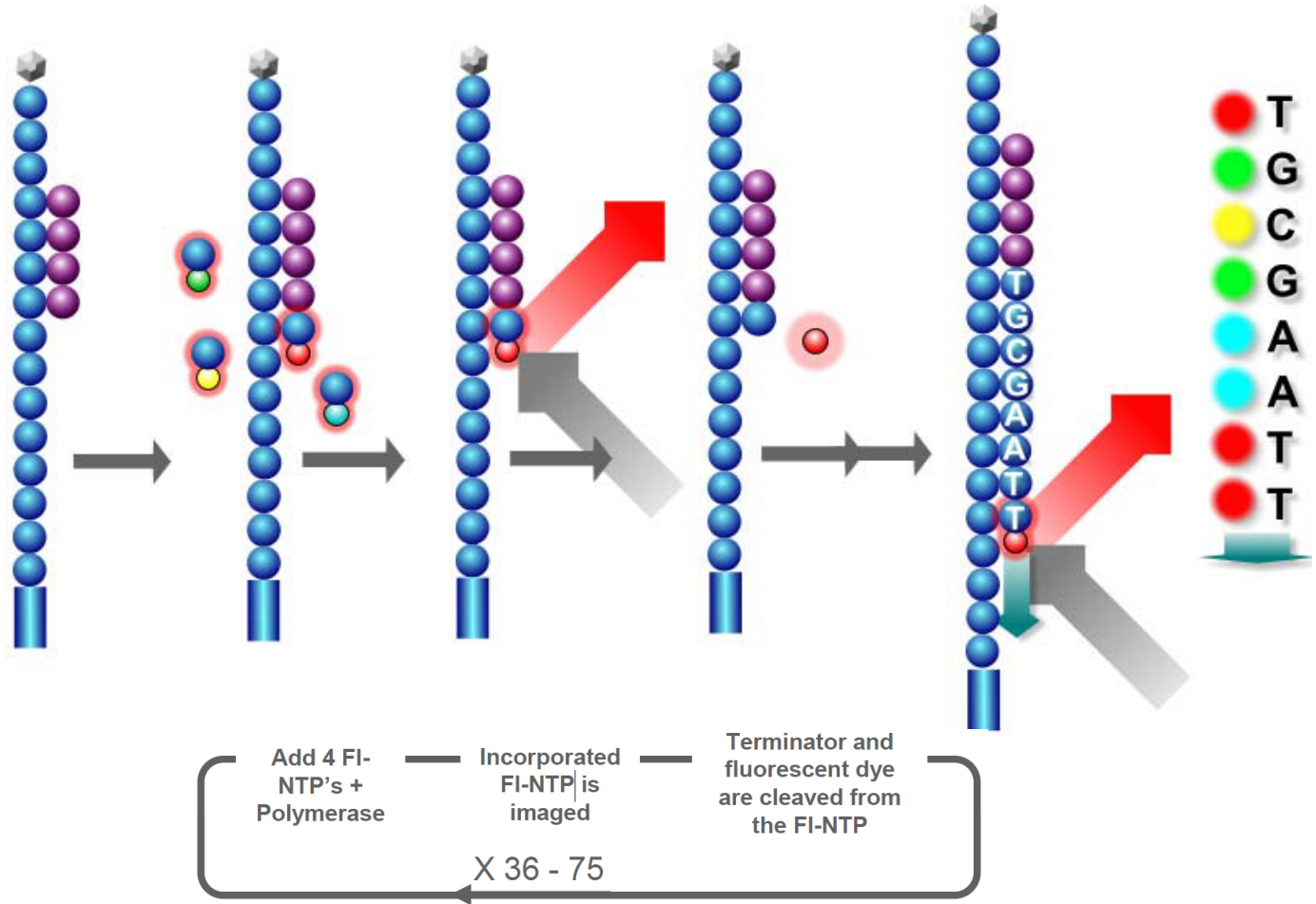
Illumina platform: Sequencing by Synthesis



Illumina platform: Sequencing by Synthesis



RNA-Seq: SEQUENCING REACTION



Illumina platform: Sequencing by Synthesis

