Analysis of Sequencing Data

(Illumina NGS technology)

Marek Mráz
Assistant Professor of Oncology
Group leader at CEITEC MU and Univ. Hospital Brno

Today……

- DNA……rules them all?
- PCR/Sanger
- DNA NGS….principles
- RNA NGS… principles

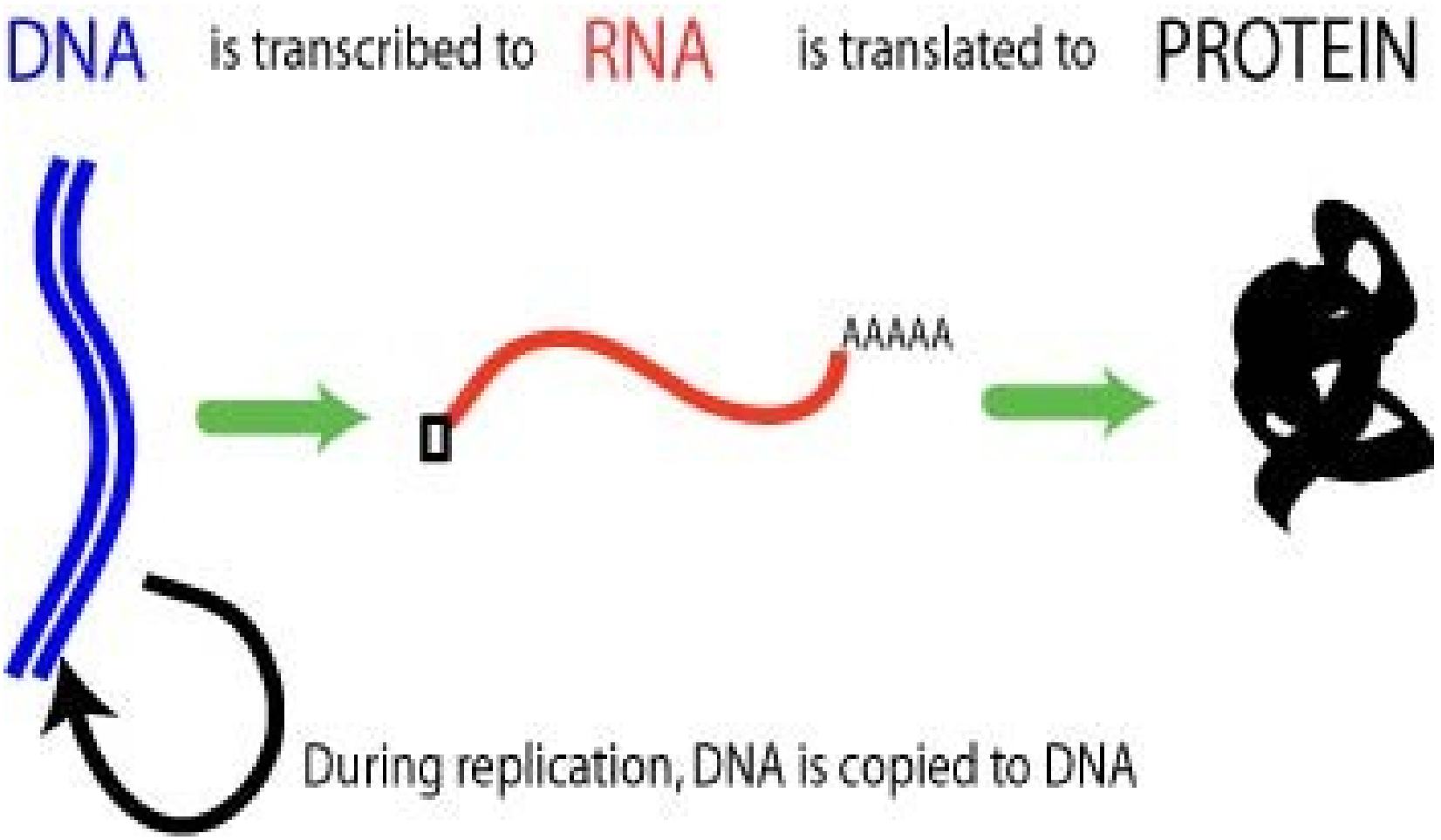
} Illumina platform

- NGS applications in general
- Examples in cancer research

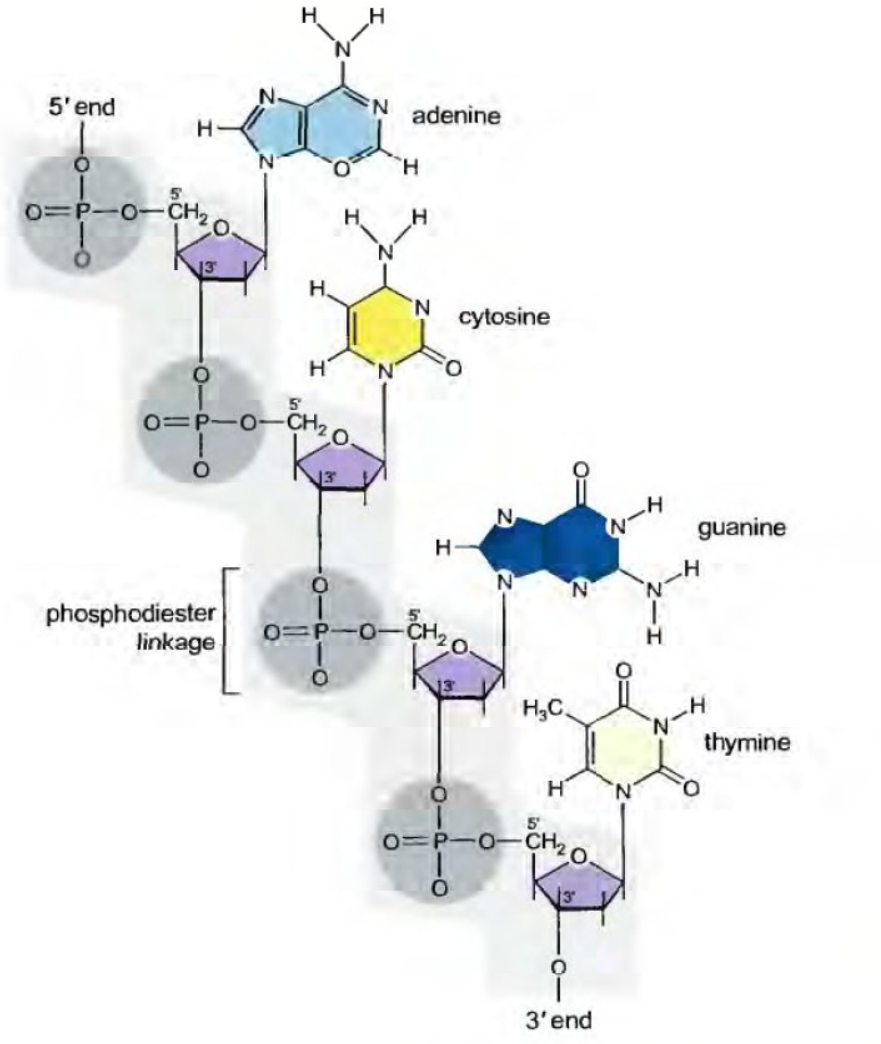
?

- Library
- Adaptor
- Index
- Barcode
- Read
- Flowcell
- Sequencing by synthesis
- T4 Ligase

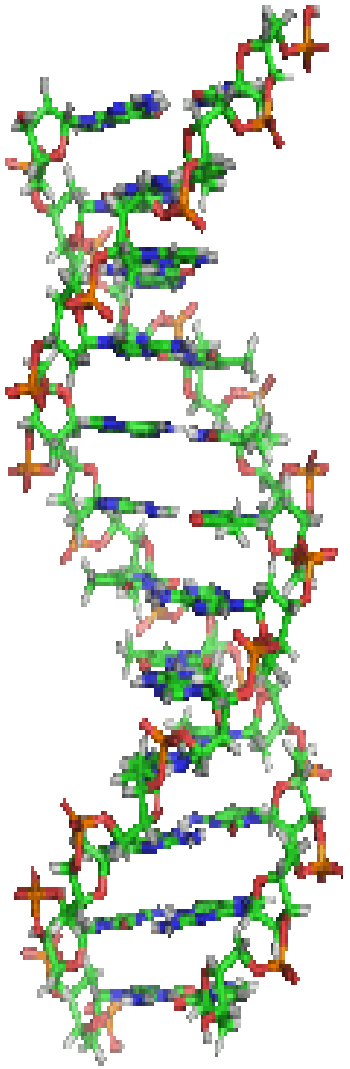
Central Dogma



DNA Has Two Jobs



- It serves as a store of information
- It directs the synthesis of proteins



What are the other types of nucleic acids….?

DNA – nucleus, mitochondria

RNA – mRNA, rRNA, tRNA, snoRNA, miRNA, lncRNA

…… all RNAs can be converted to DNA….we always work/sequence DNA:

DNA or cDNA

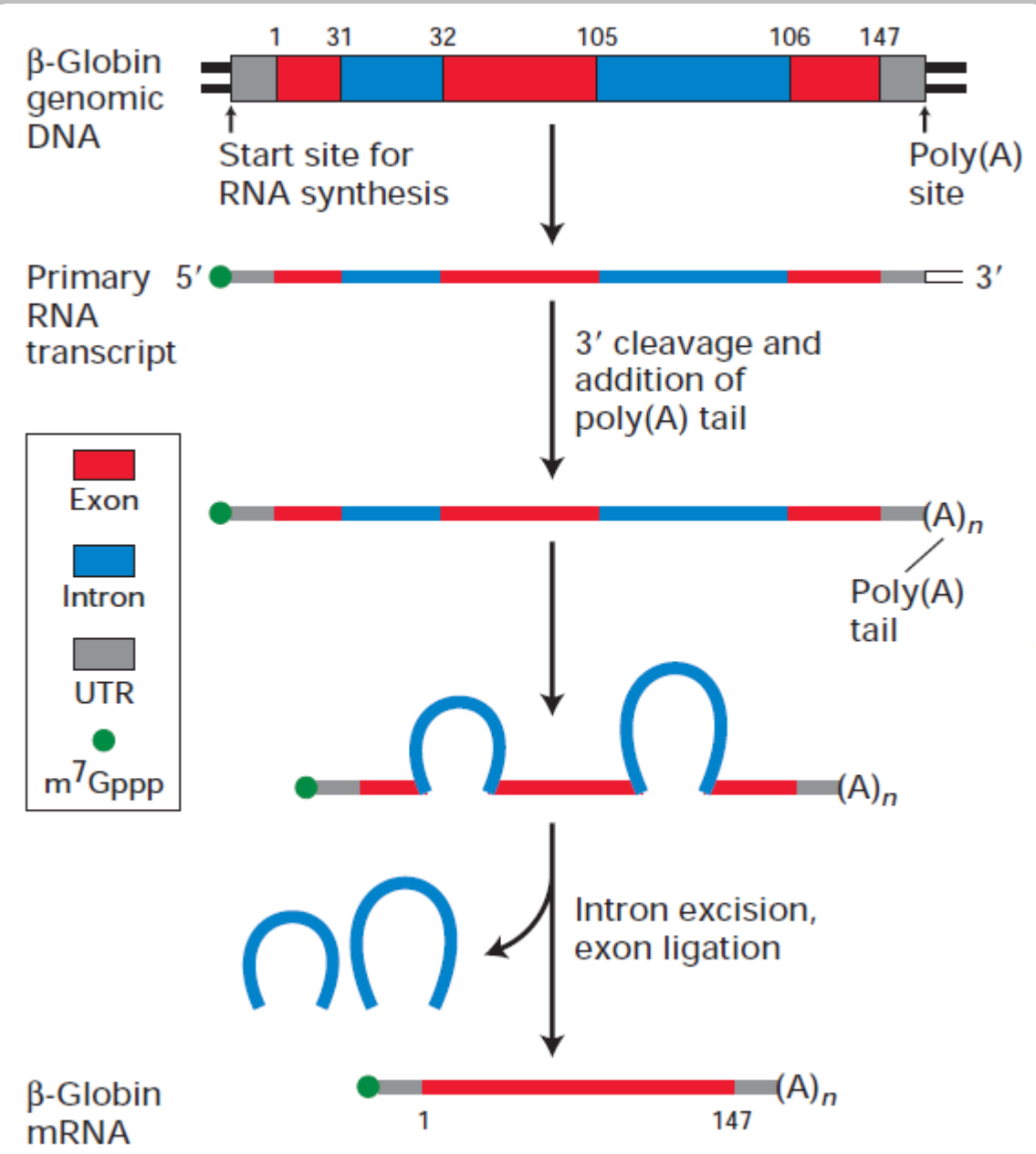
With genomic DNA we are interested in the sequence …..mutations, SNP, CNV, translocations

With RNA we are interested in other things…..
Like?

DNA Sequencing

- You have 3 billion bases
- ~20,000(0) genes

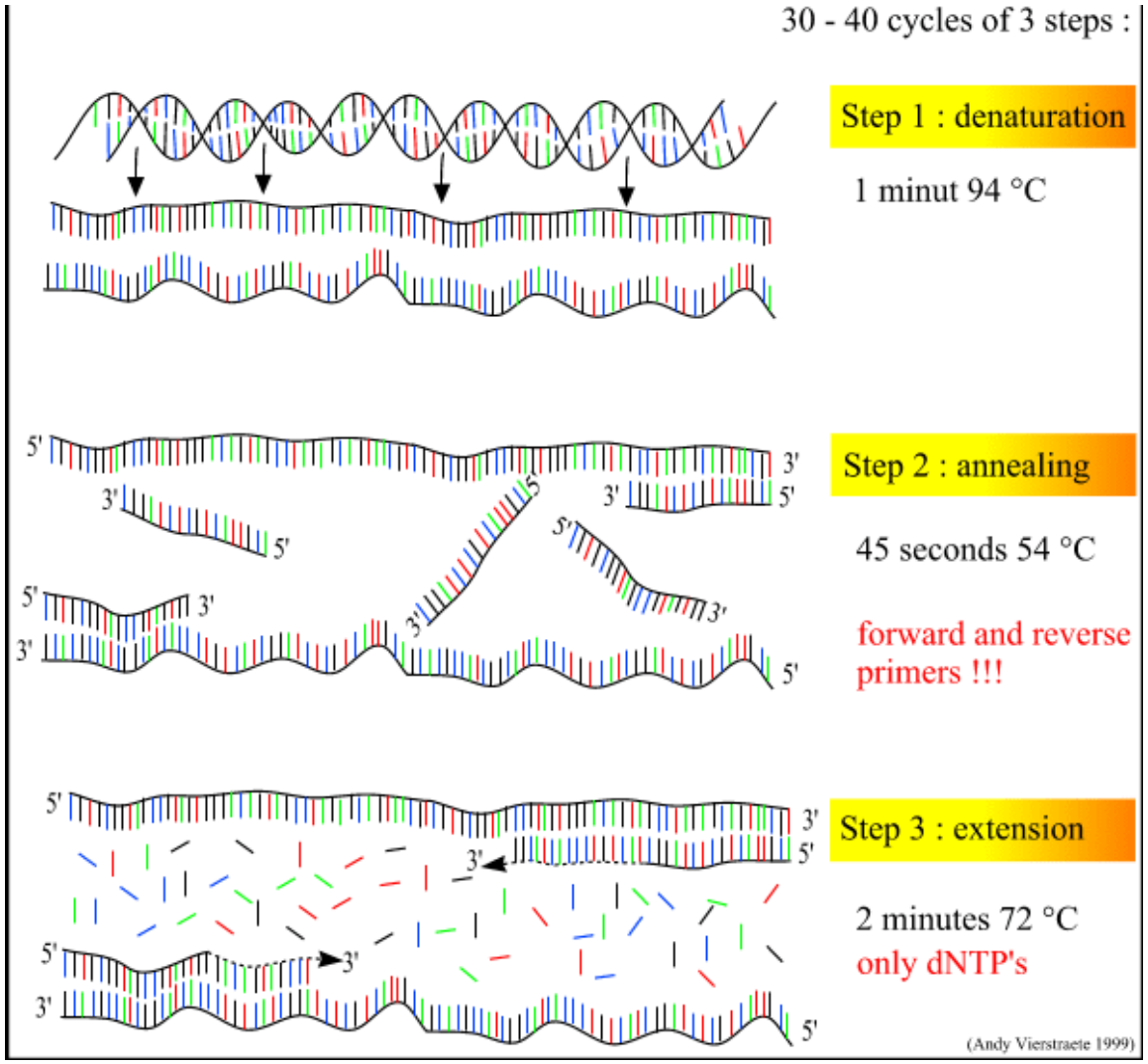
Gene (DNA) gives rise to mRNA

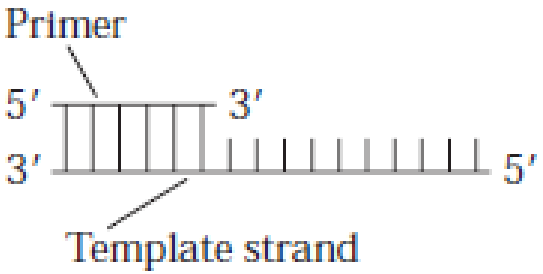


PCR

PCR

- Mix DNA with dNTPs and primer
- Amplify…DNA polymerase

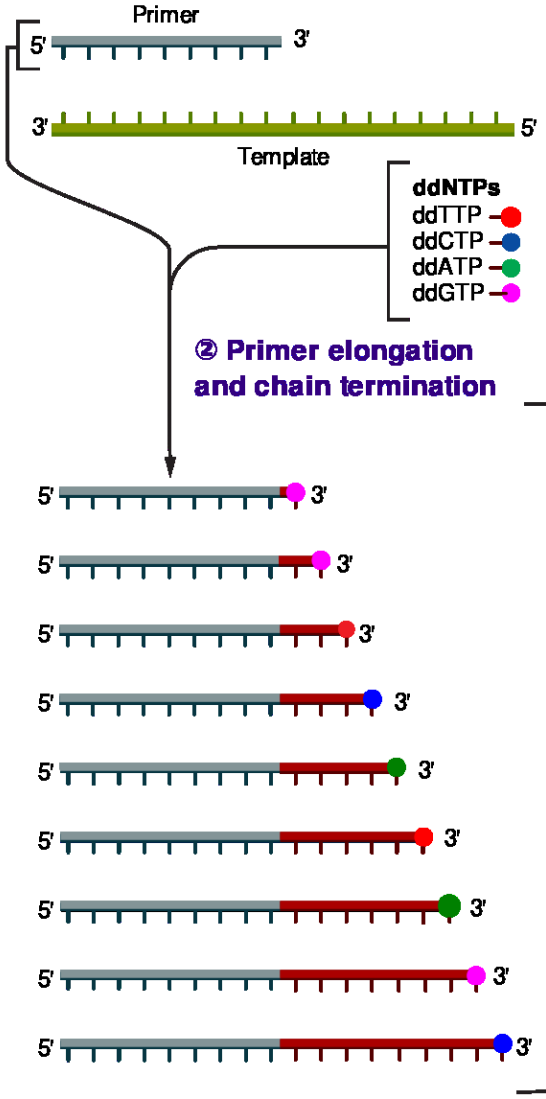


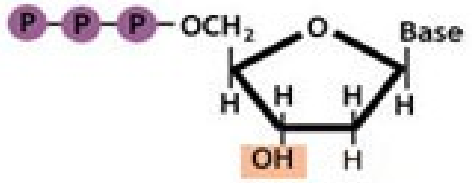


DNA has orientation, need of primer for PCR

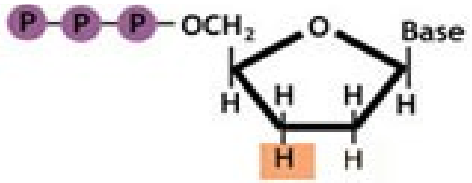
Sanger seq

① Reaction mixture
- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)

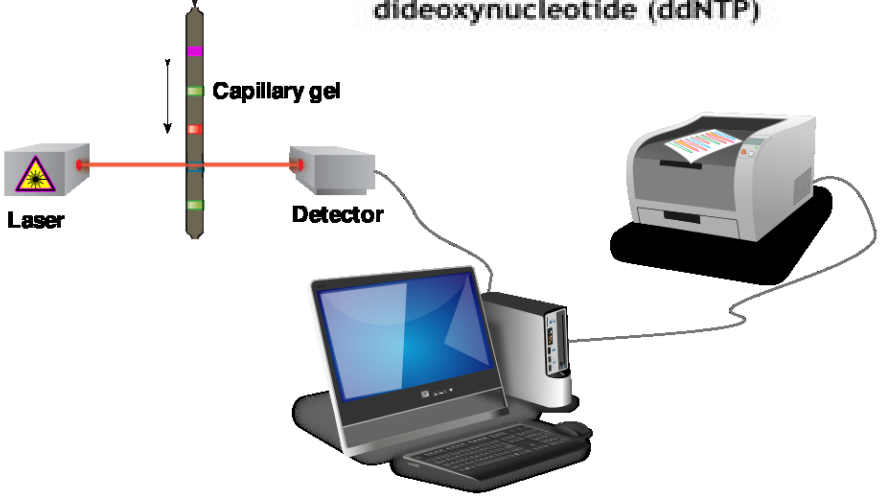


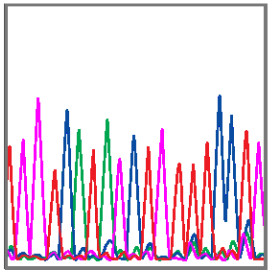


deoxynucleotide (dNTP)



dideoxynucleotide (ddNTP)





Chromatograph

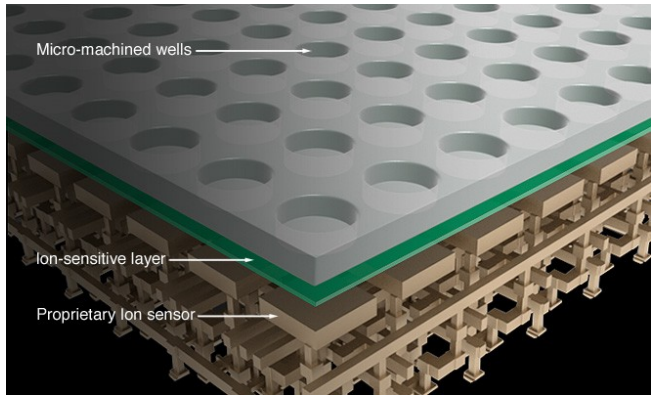
④ Laser detection of flourochromes
and computational sequence analysis

Sanger Sequencing

- Advantages
  - Long reads (~900bps)
  - Suitable for small projects
- Disadvantages
  - Low throughput
  - Expensive (cost per base)

Next Generation Sequencing

- Takes advantage of miniaturization to engage in massively parallel analysis
  - Essentially carrying out millions of sequencing reactions simultaneously in each of 10 million tiny wells/spots
- Sophisticated computer analysis of huge amounts of information allows “assembly" of a given sequence



Massive Parallel Seq workflow

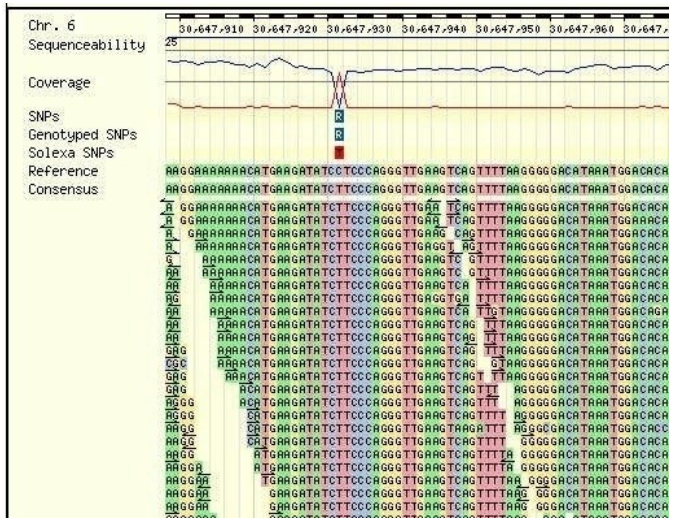
A SAMPLE

1) Library preparation

2) Cluster generation on a flow cell

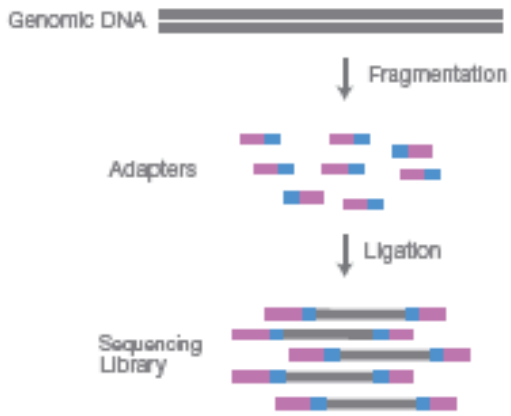
3) Sequencing & imaging

SE, PE reads,
50-250 bases
(miseq)



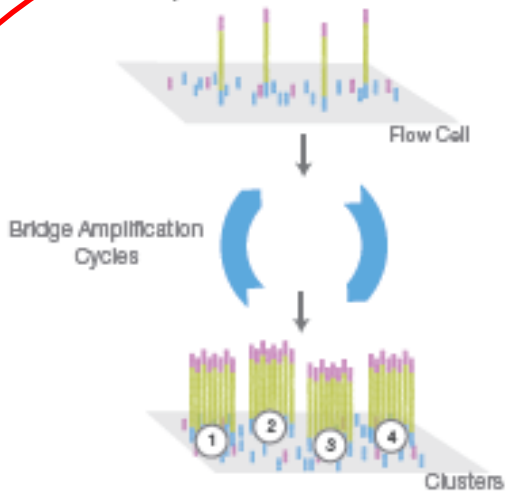
4) Data processing & analysis

A. Library Preparation



NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

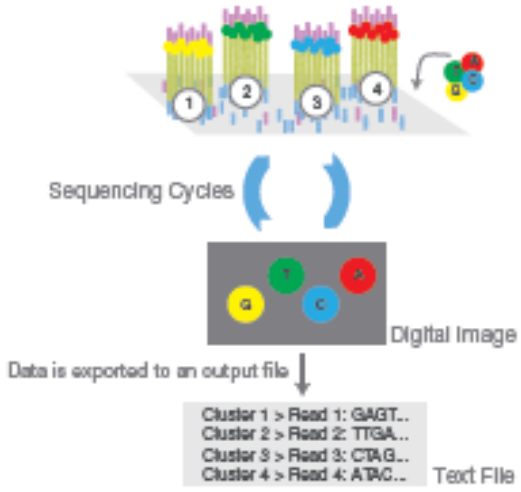
A. Cluster Amplification



Library is loaded into a flow cell and the fragments hybridize to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

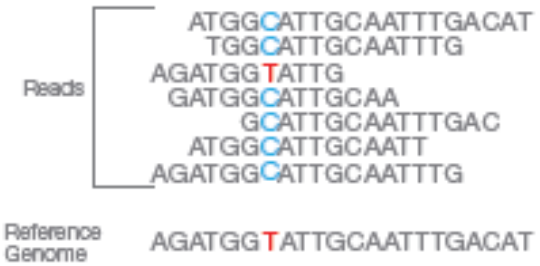
This is the trick

C. Sequencing



Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

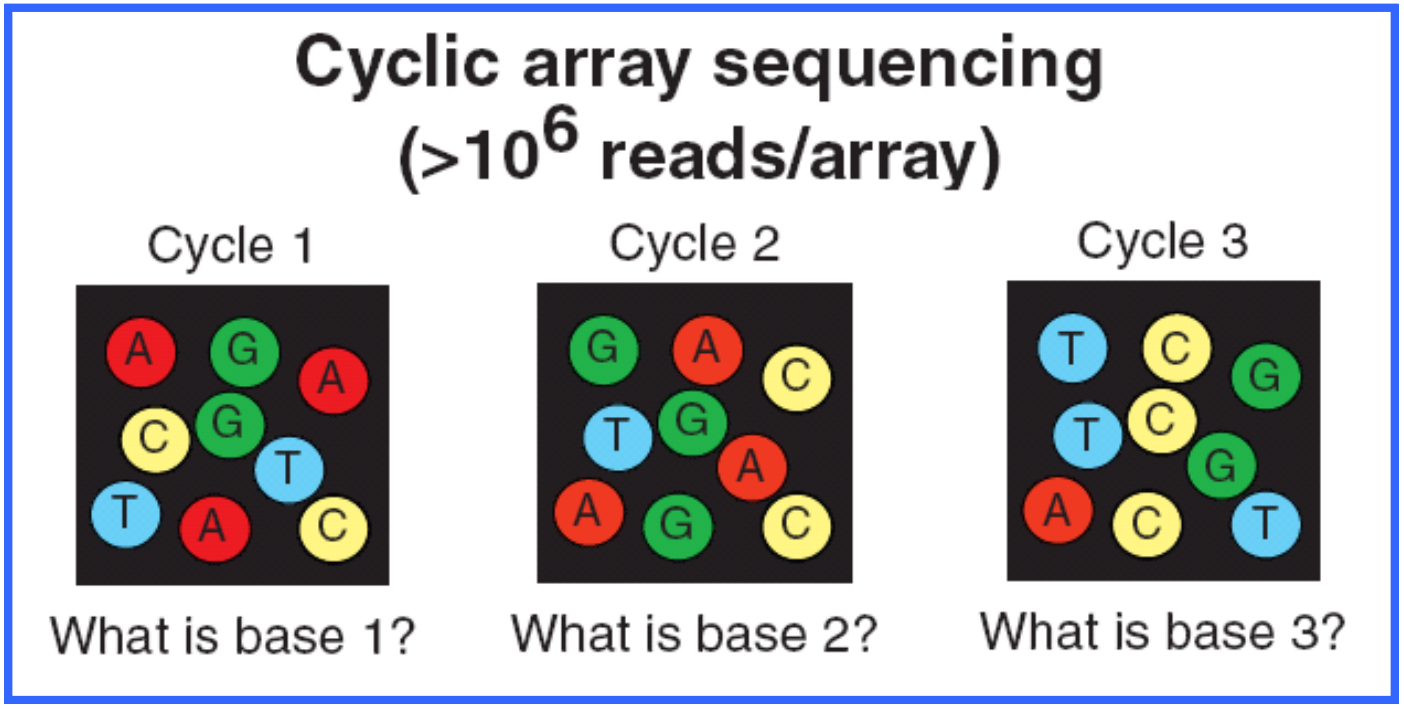
D. Alignment & Data Anaylsis



Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

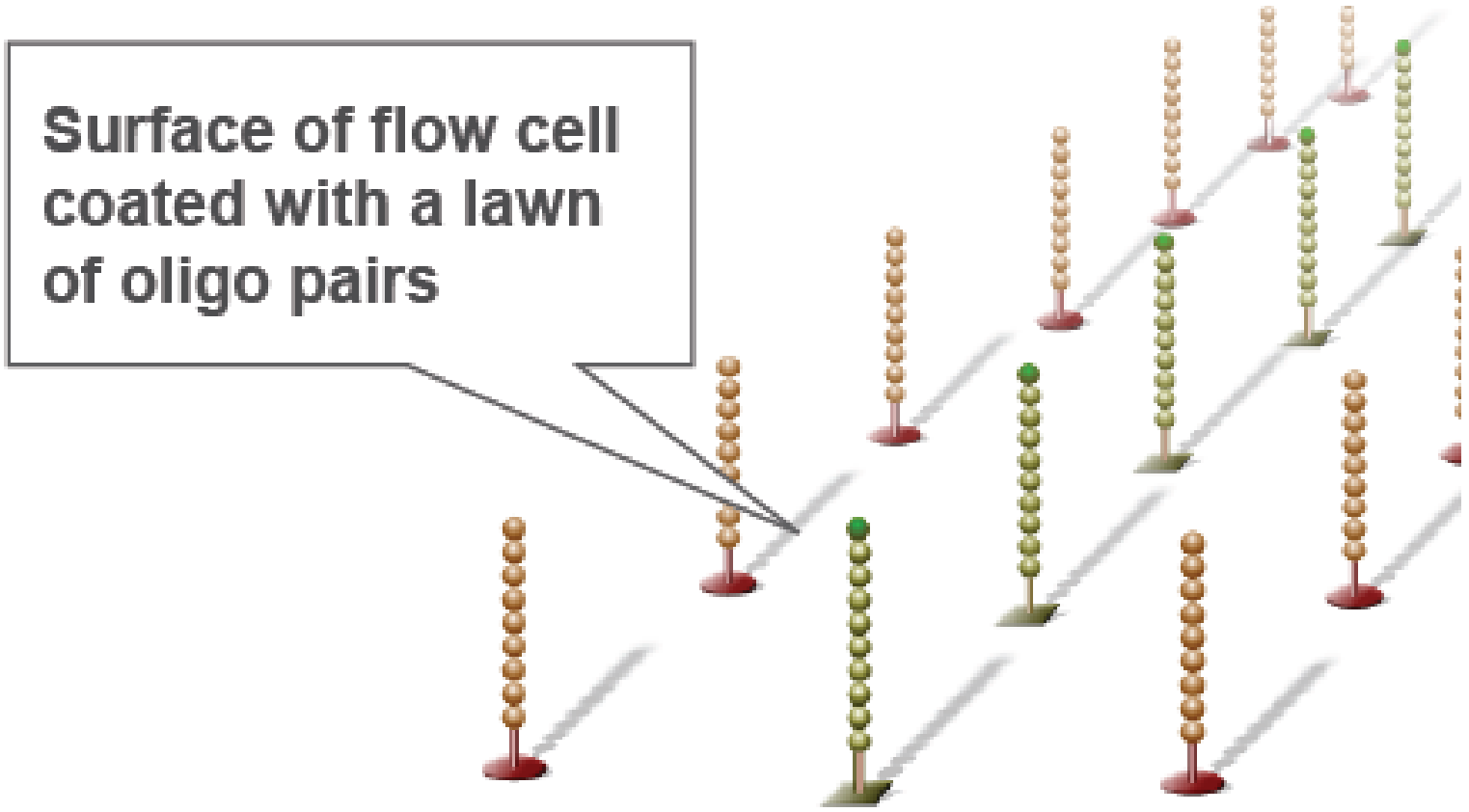
High Parallelism is Achieved in Polony Sequencing

Sanger **Polony**




E
5'
P5
Index 2
Rd1 SP
DNA Insert
Rd2 SP
Index 1
P7
5'

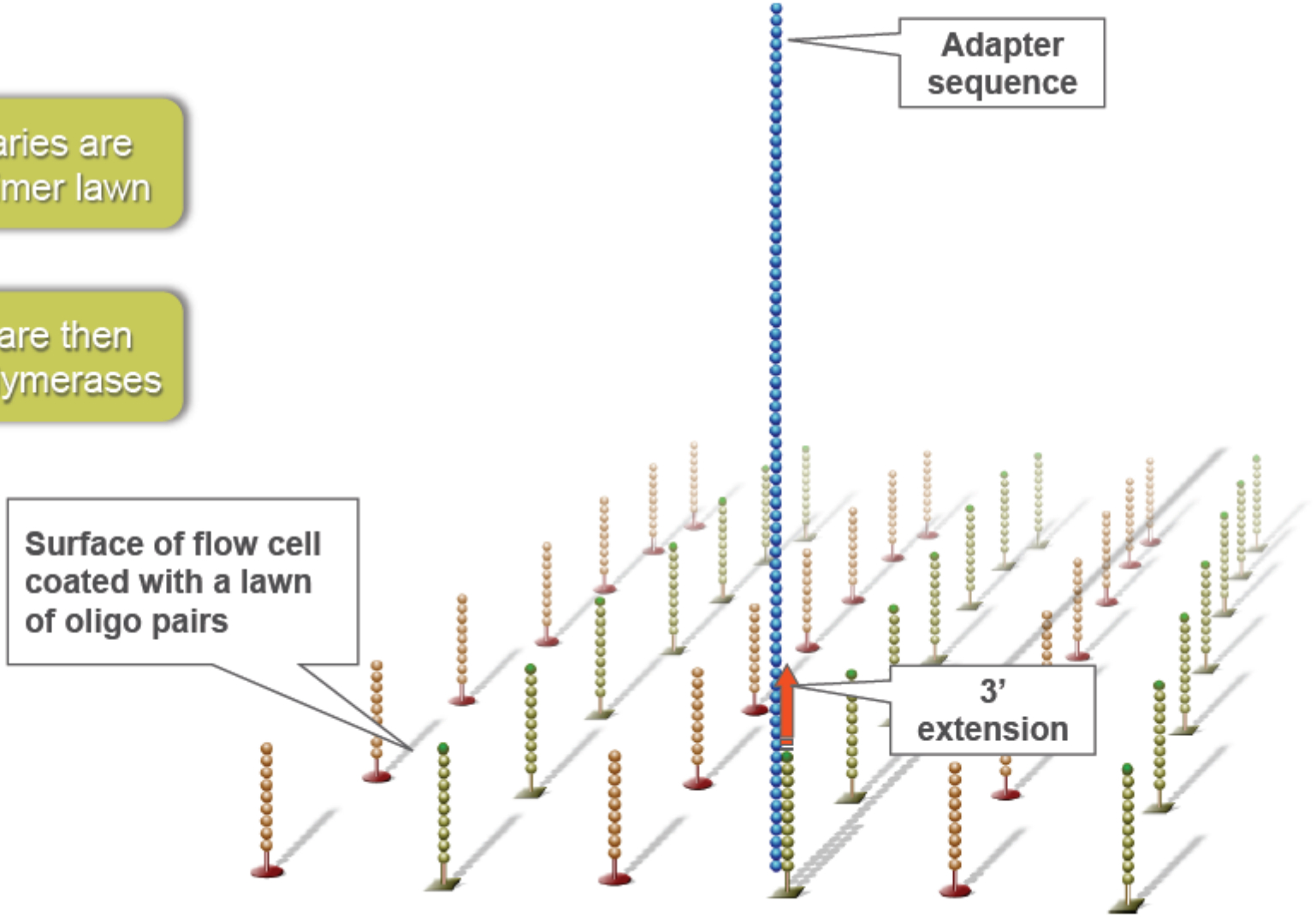
Two PCR primers are attached to the surface of flowcell. One of the primers has a cleavable site



Hybridize Fragment & Extend

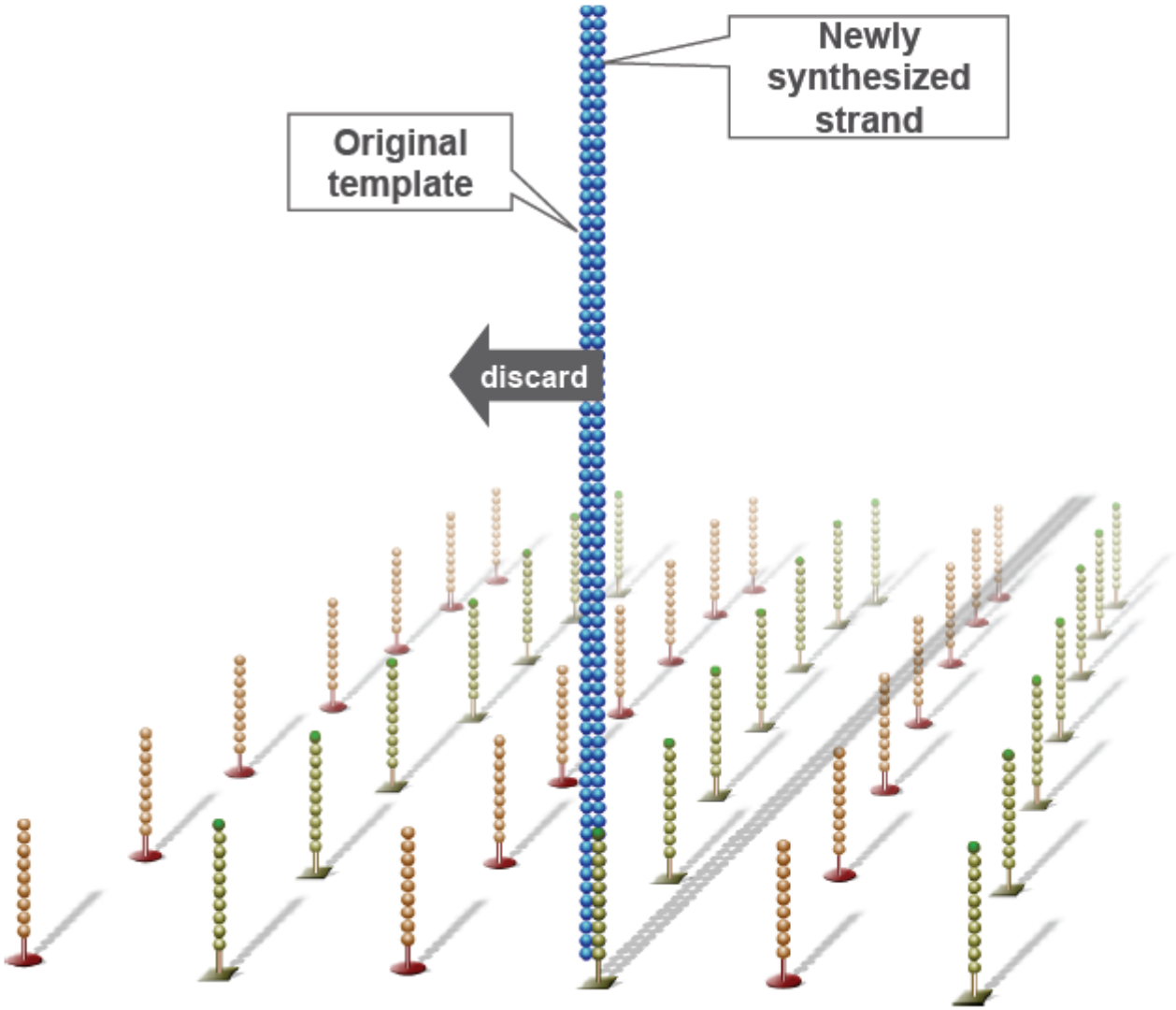
Single DNA libraries are hybridized to primer lawn

Bound libraries are then extended by polymerases

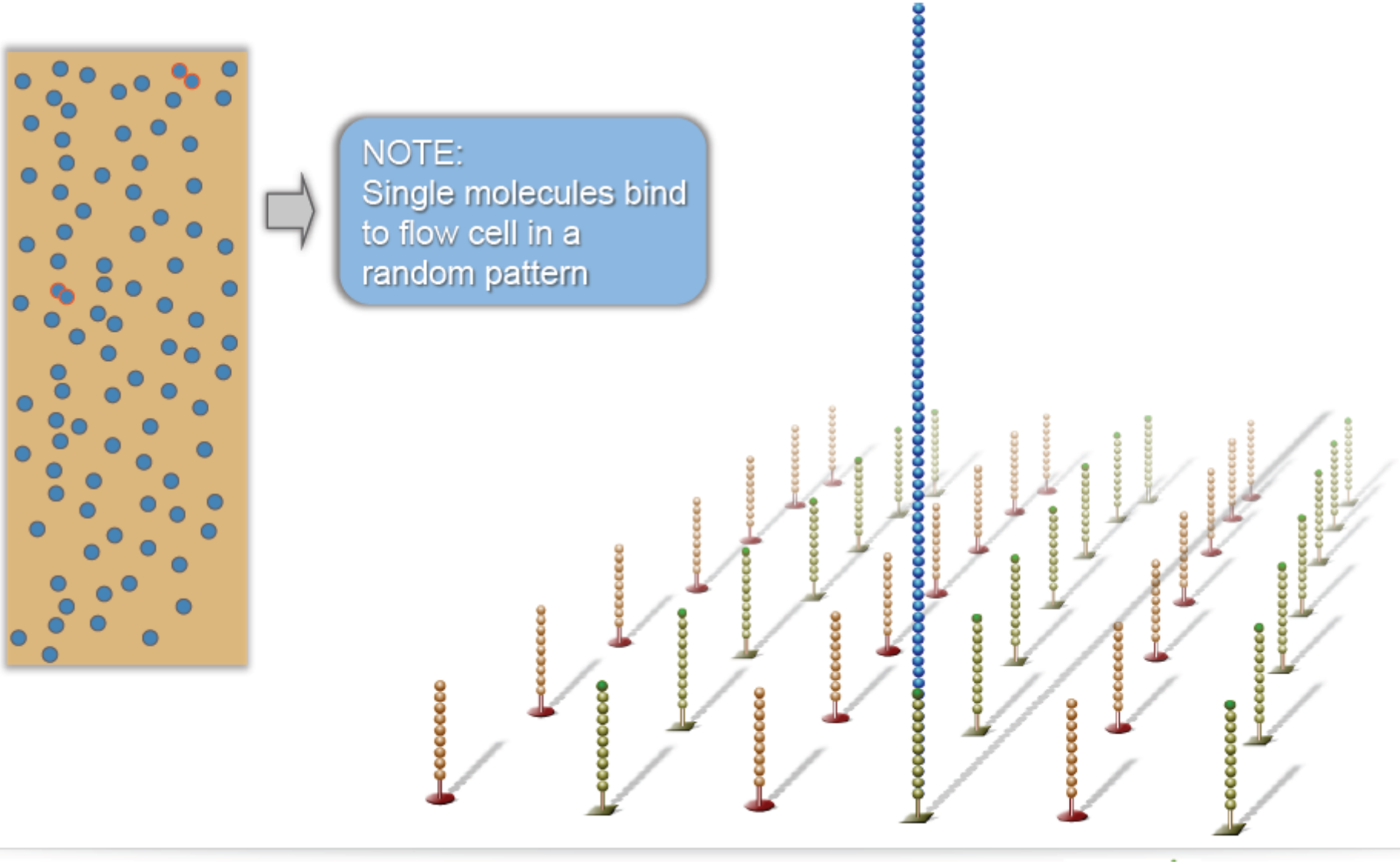


Denature Double-Stranded DNA

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



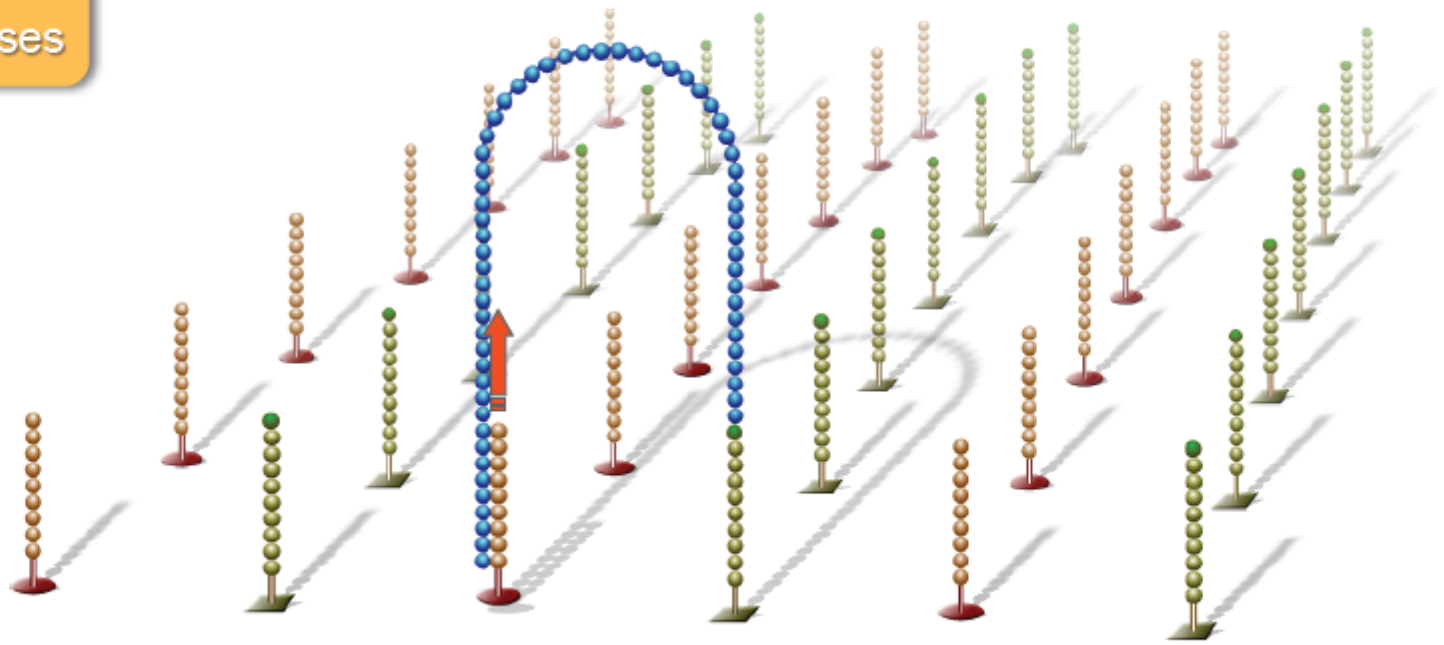
Single-Stranded DNA



Bridge Amplification

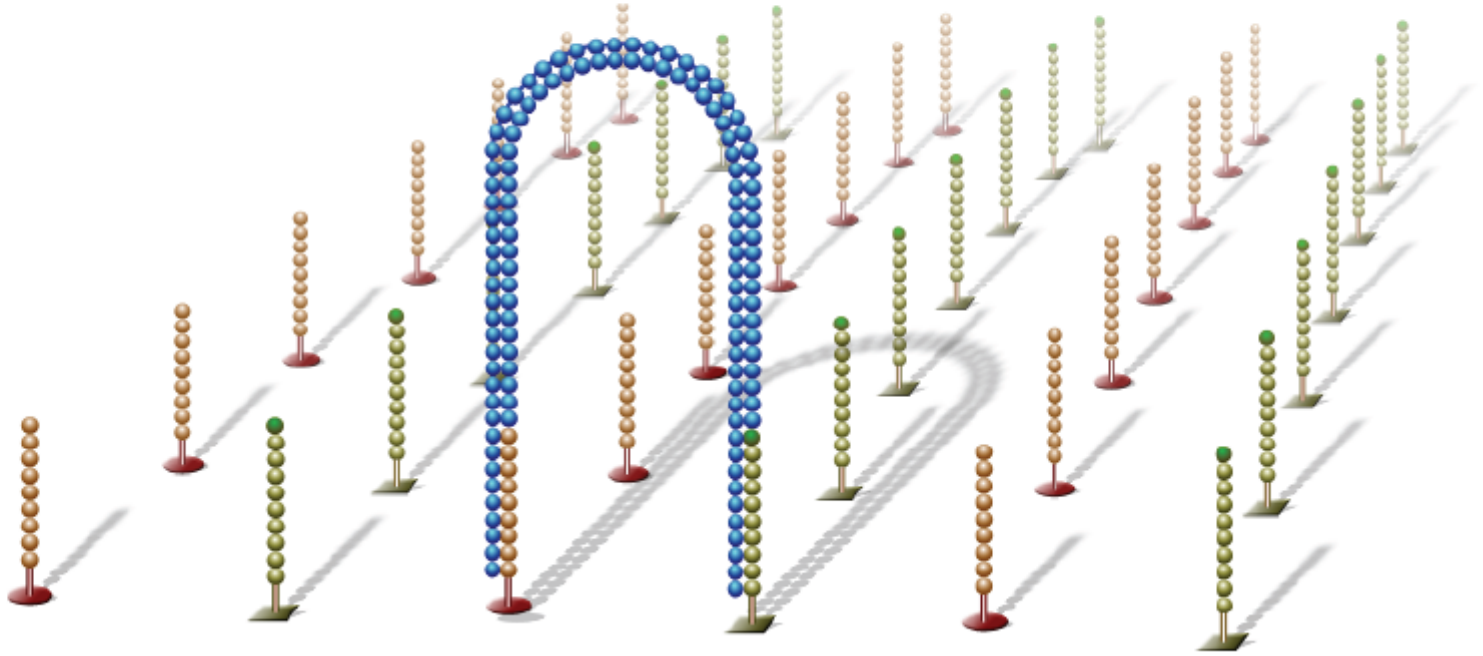
Single-stranded molecule flips over and forms a bridge by hybridizing to adjacent, complementary primer

Hybridized primer is extended by polymerases



Bridge Amplification

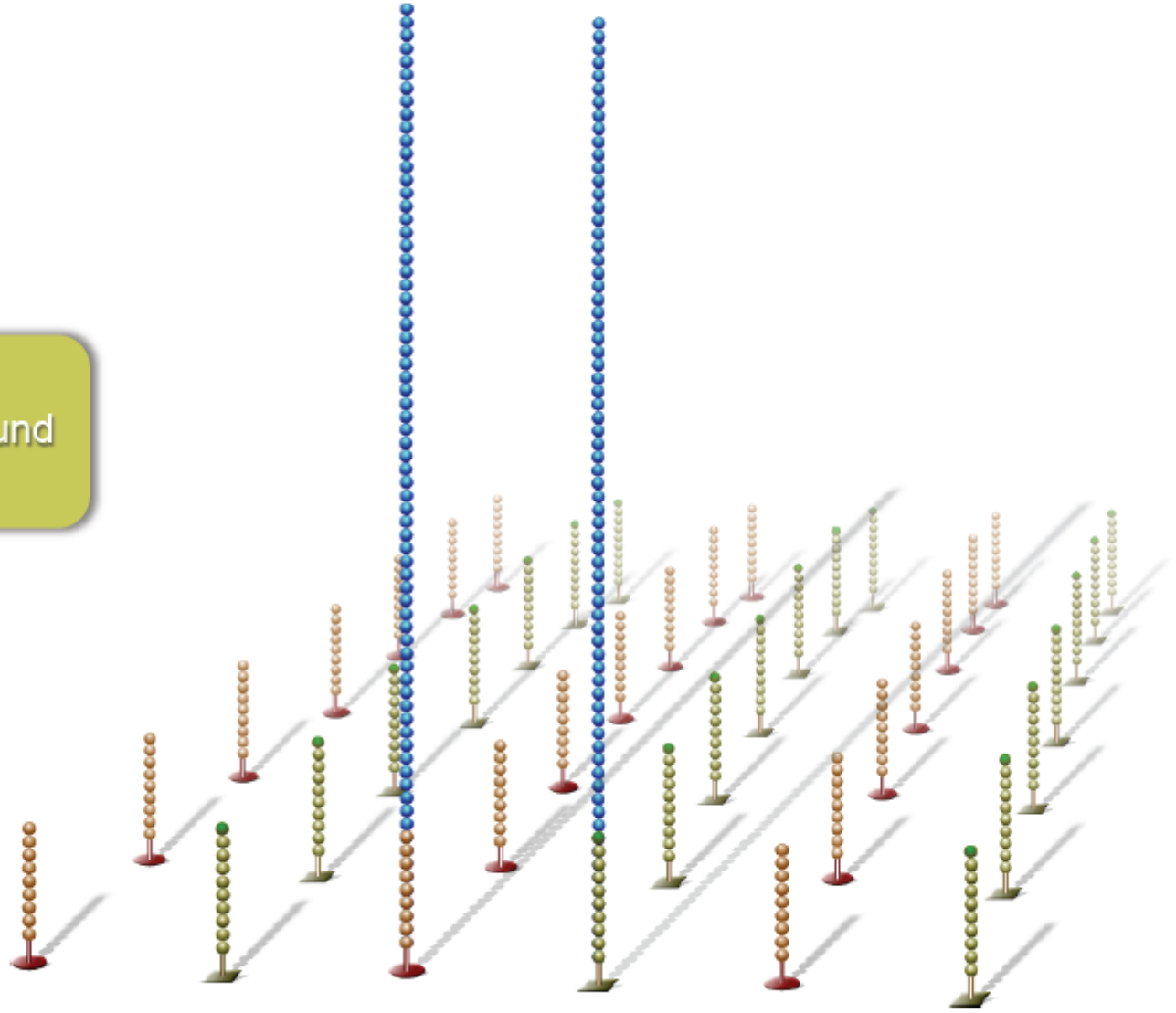
Double-stranded bridge is formed



Denature Double-Stranded Bridge

Double-stranded bridge is denatured

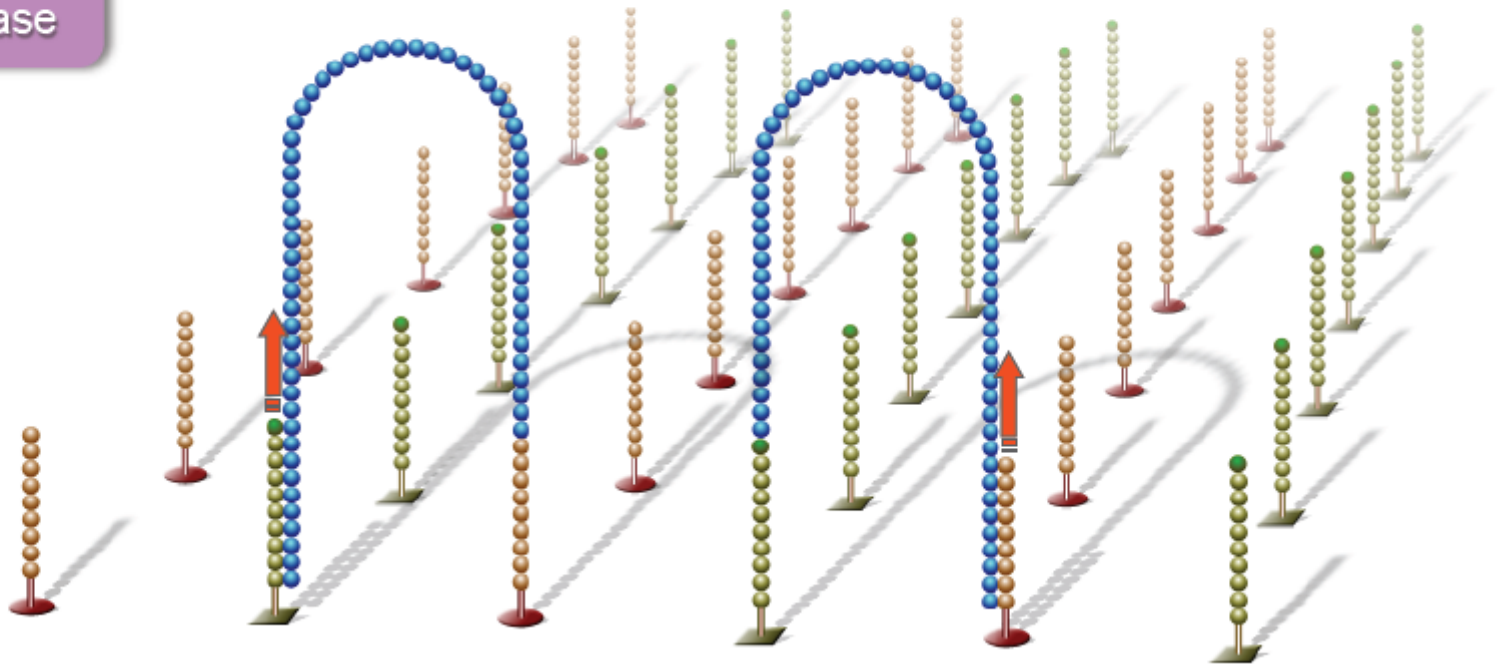
Result:
Two copies of covalently bound single-stranded templates



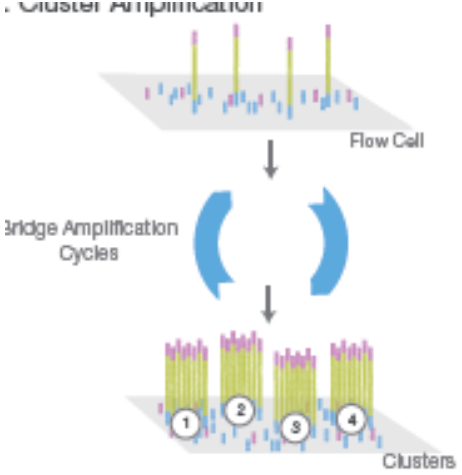
Bridge Amplification

Single-stranded molecules flip over to hybridize to adjacent primers

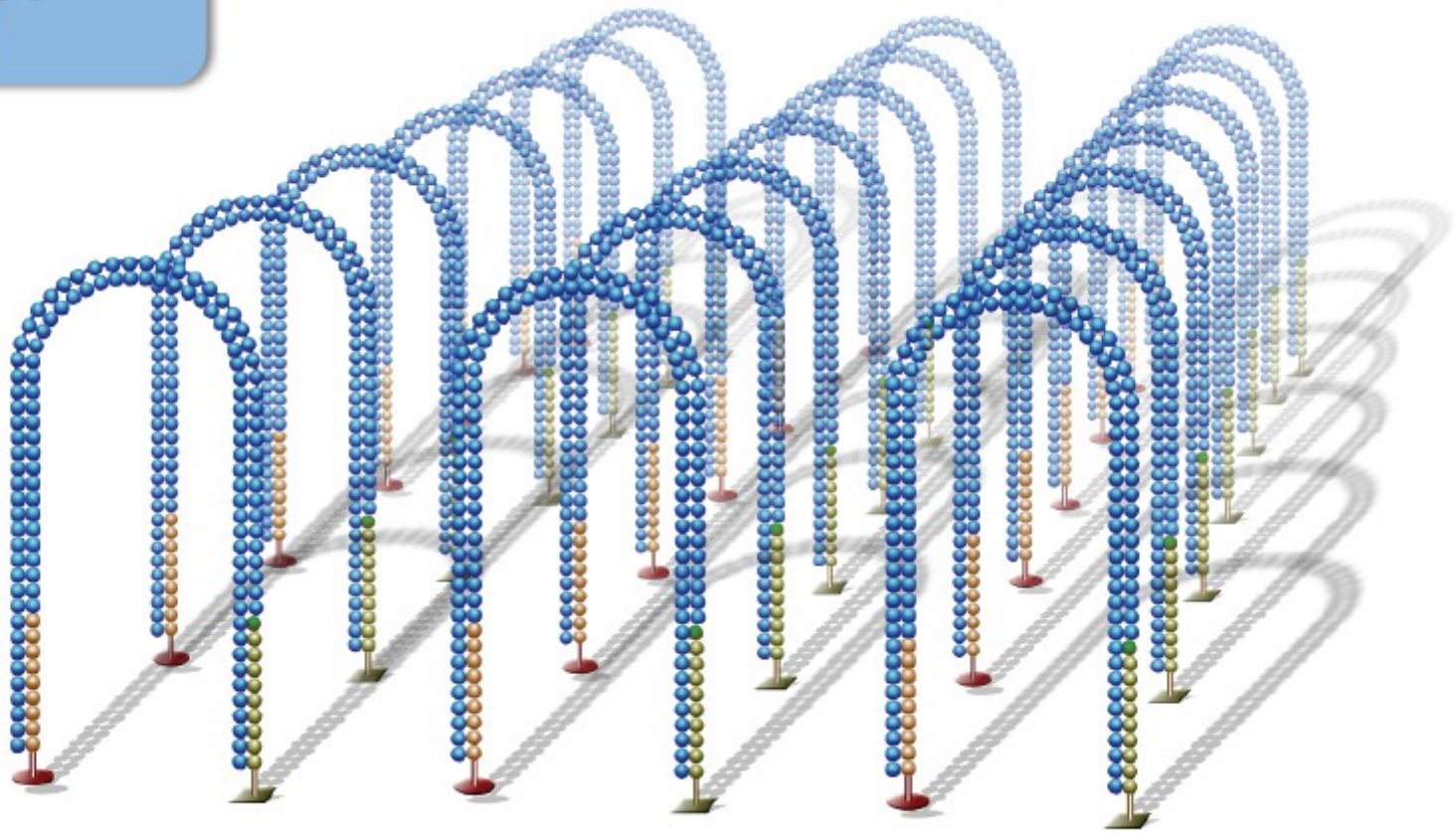
Hybridized primer is extended by polymerase



Bridge Amplification

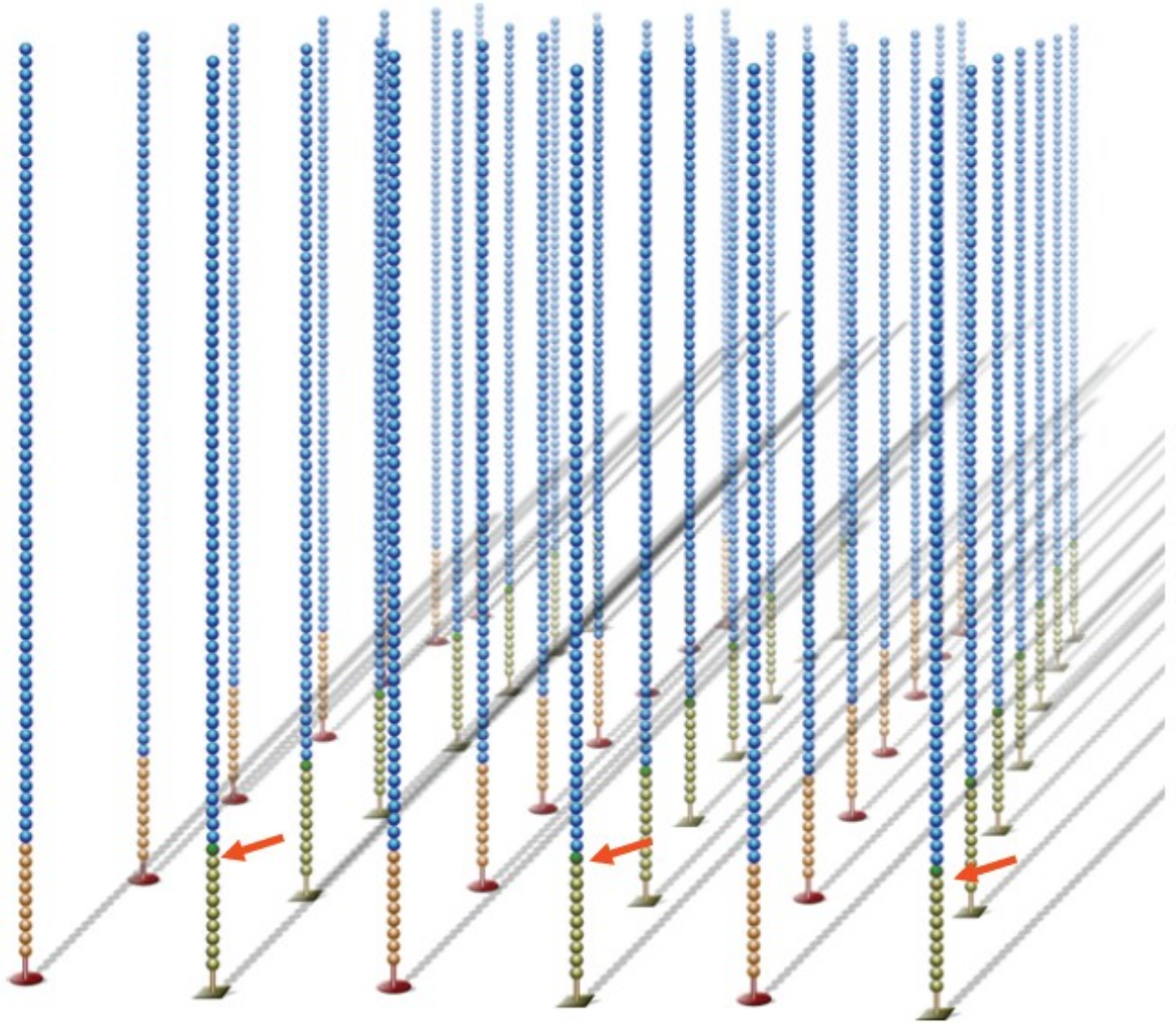


Bridge amplification cycle is repeated until multiple bridges are formed



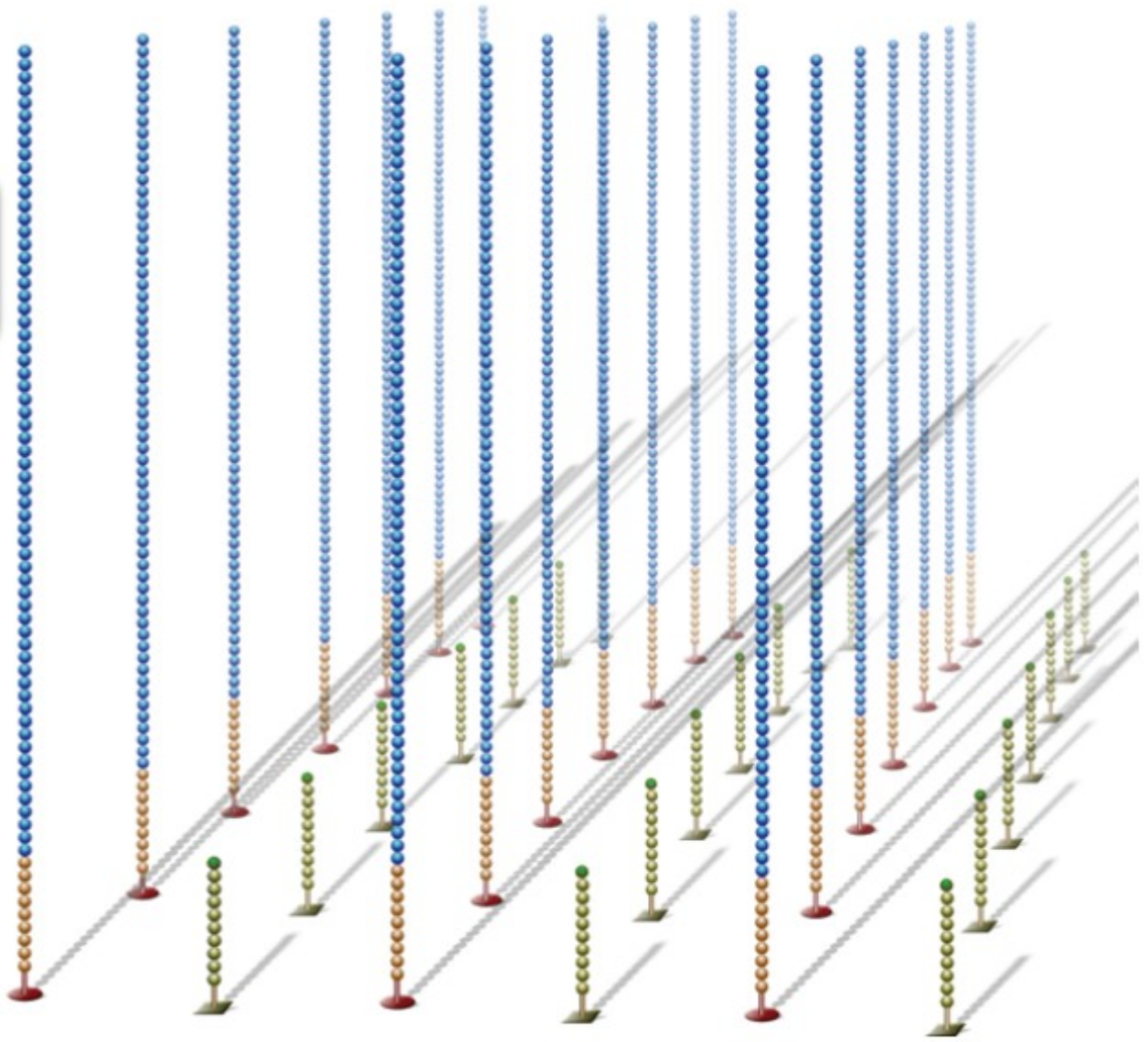
Linearization

dsDNA bridges are denatured



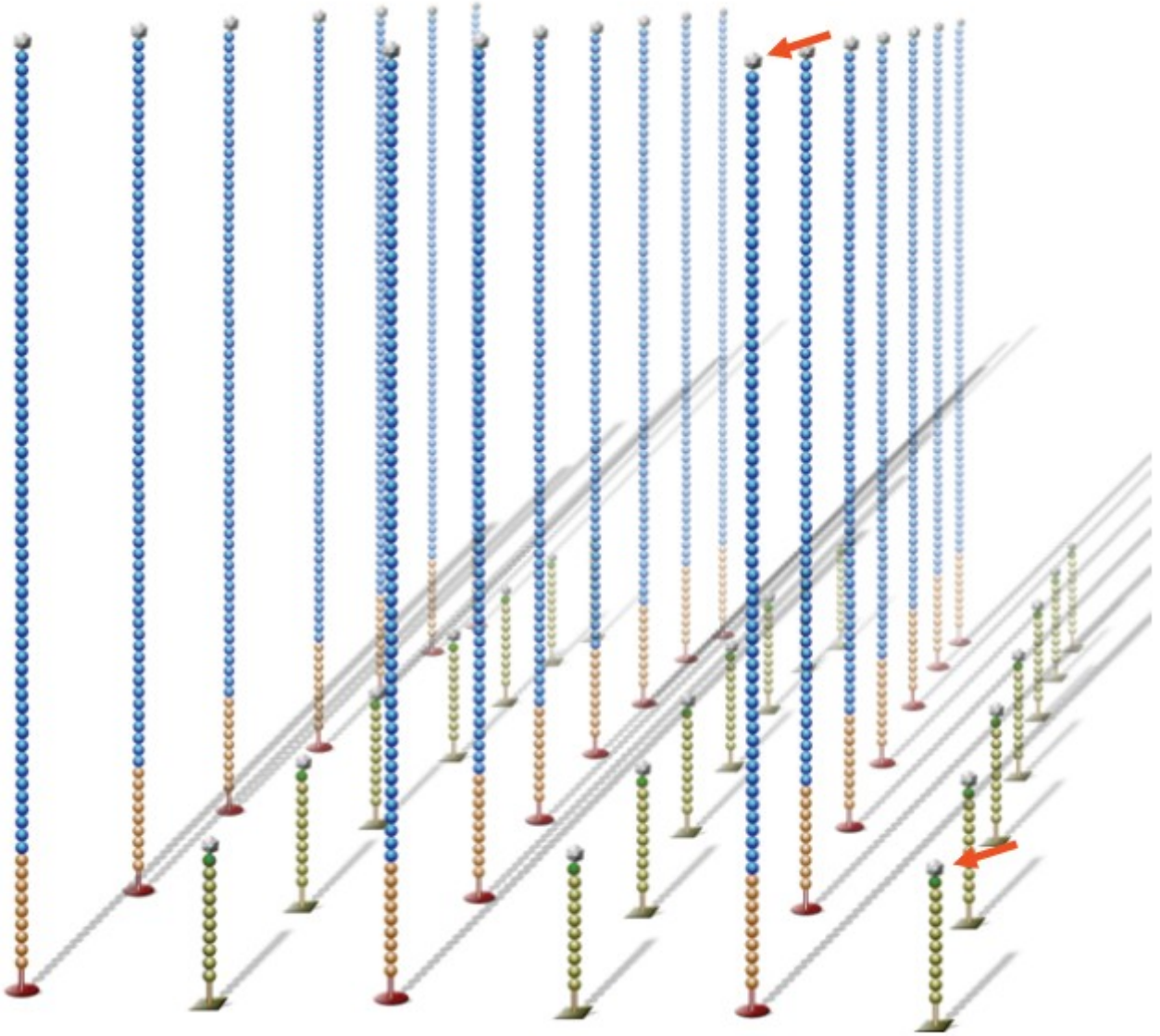
Reverse Strand Cleavage

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



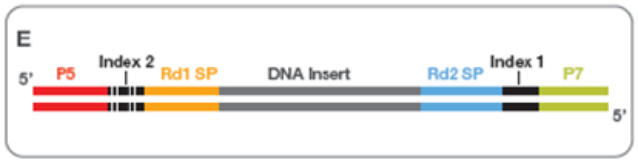
Blocking

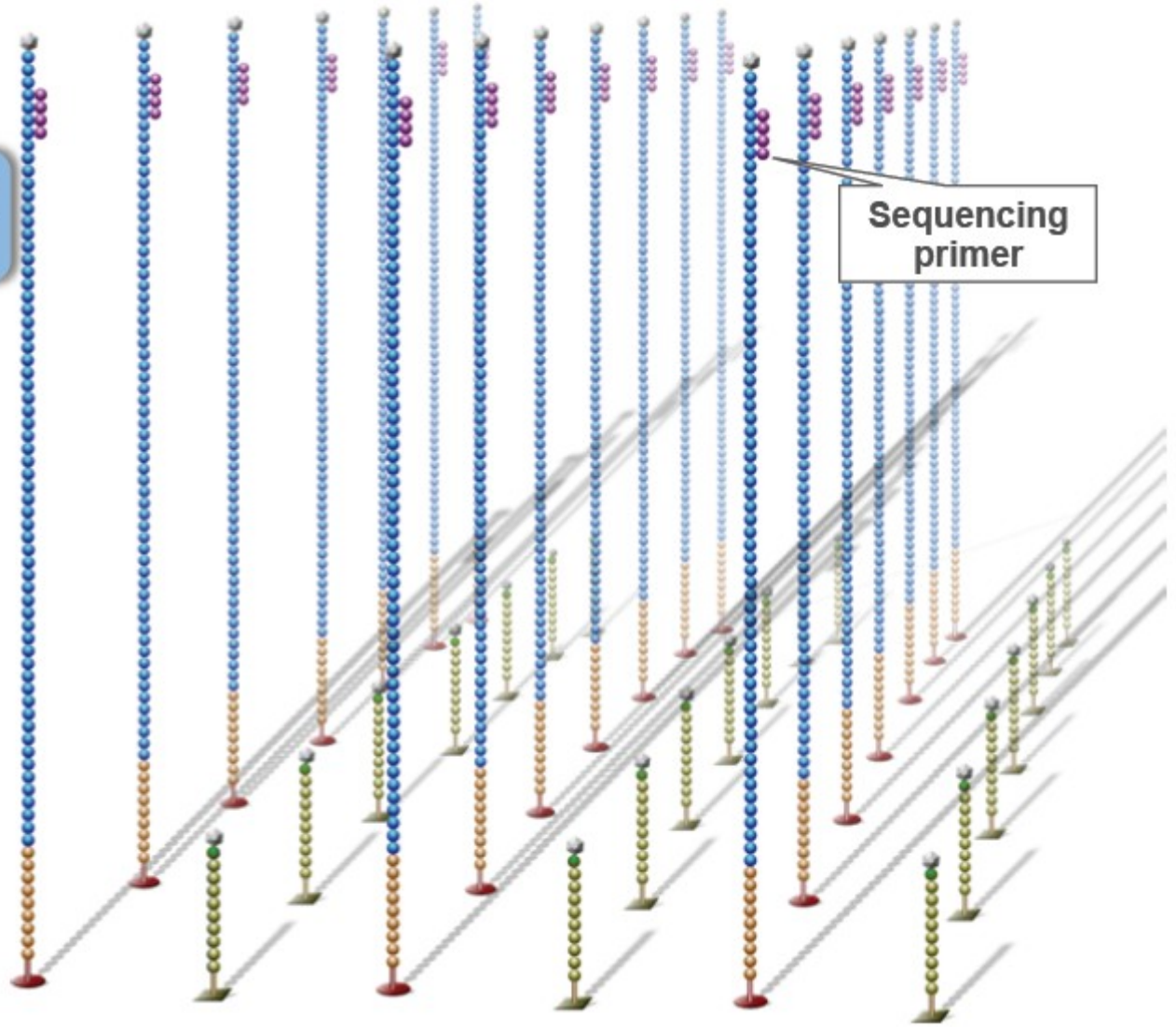
Free 3’ ends are blocked to prevent unwanted DNA priming



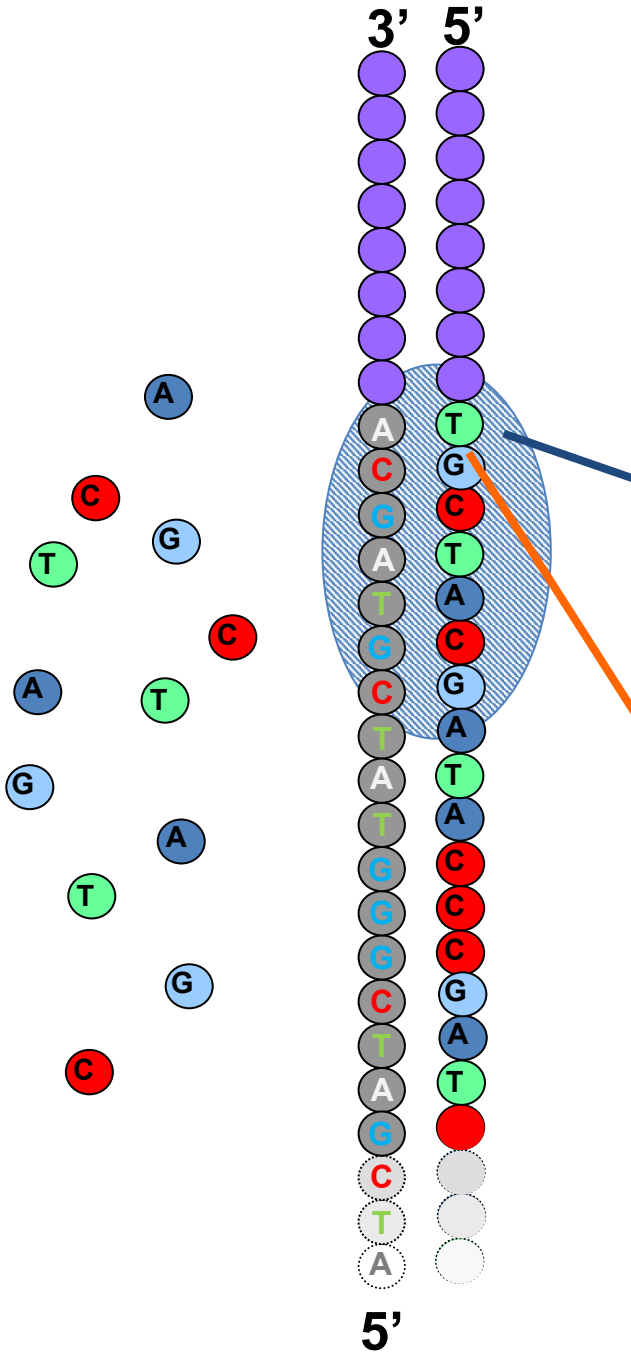
Read 1 Primer Hybridization

Sequencing primer is hybridized to adapter sequence



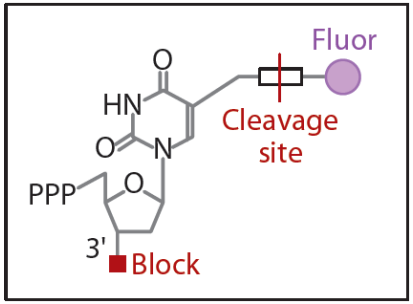


Sequencing by synthesis



Cycle 1: Add sequencing reagents

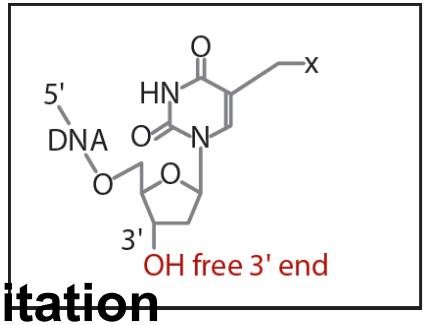
First base incorporated



Emission

Detect signal

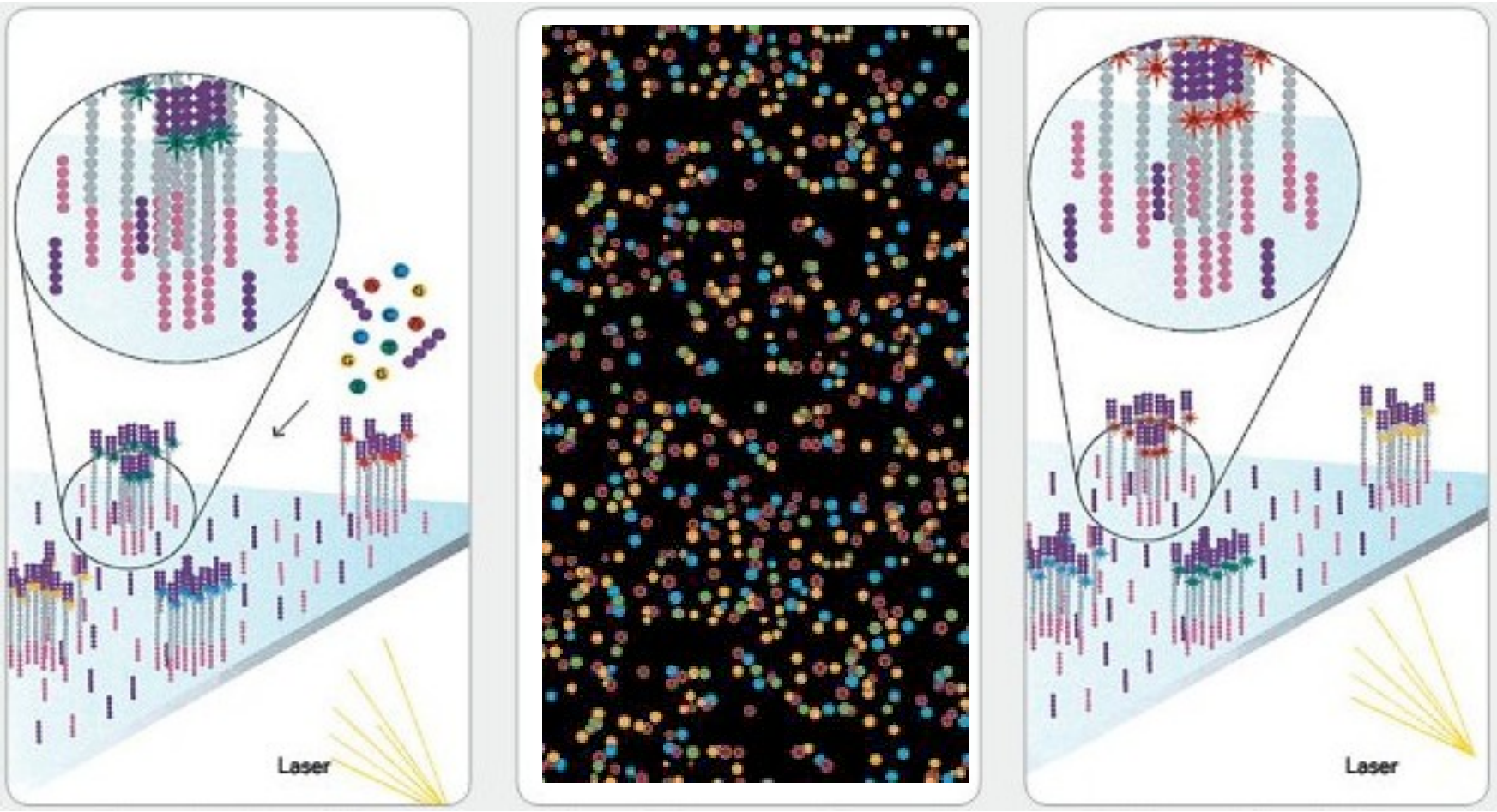
Cleave terminator and dye



Excitation

Cycle 2-n: Add sequencing reagents and repeat

Sequencing by Synthesis - Fluorescently labeled Nucleotides (Illumina)



Complementary strand elongation: DNA Polymerase

Sequencing with Paired Ends

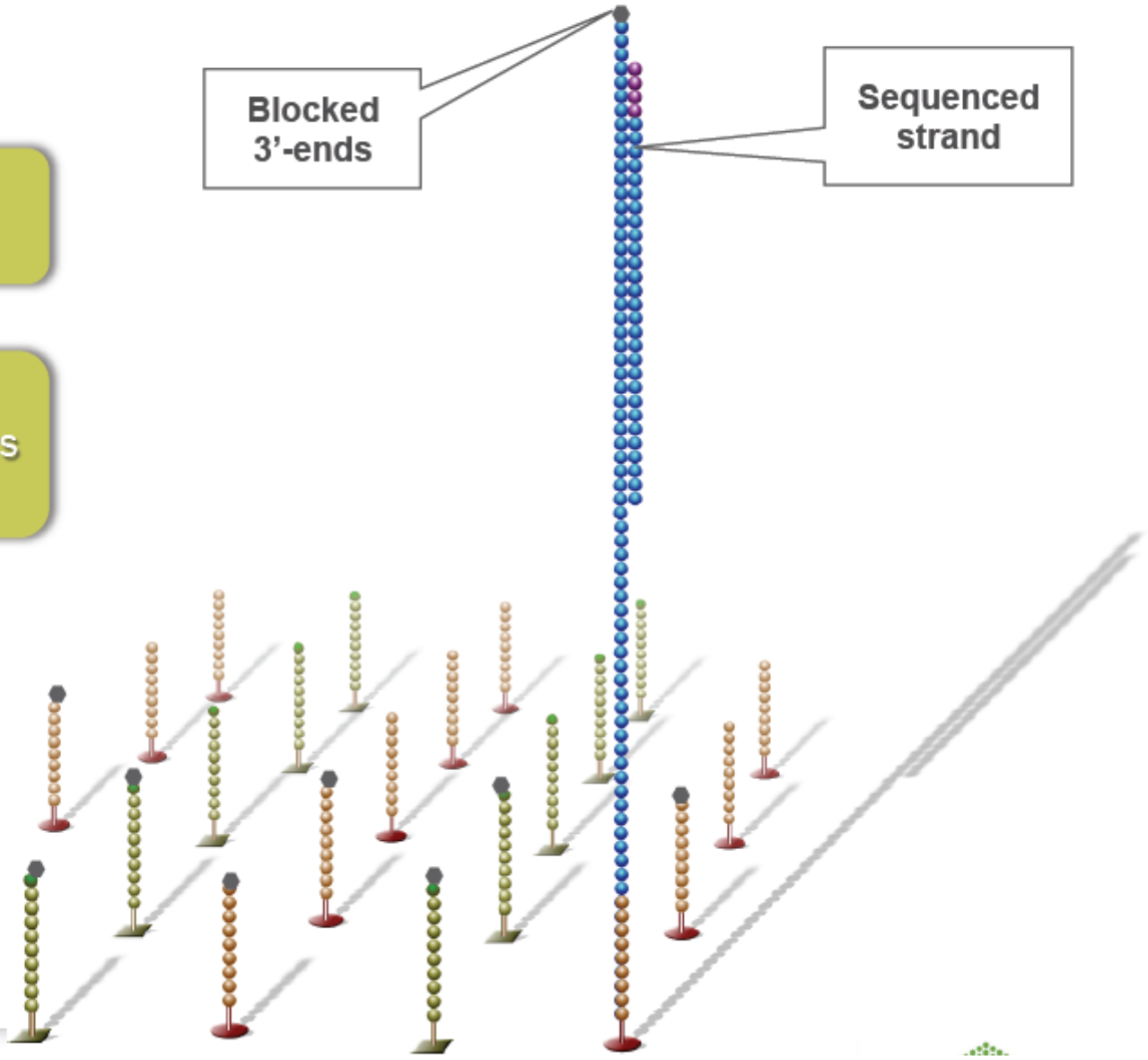
Reference	This is really the best way to do sequencing
Single-reads	This is
…	is really
…	really the
…	the best
…	sequencing
Paired-reads	This is (----100 characters-------) sequencing

Assembly becomes easier!!

Paired End Sequencing

Sequenced strand is stripped off

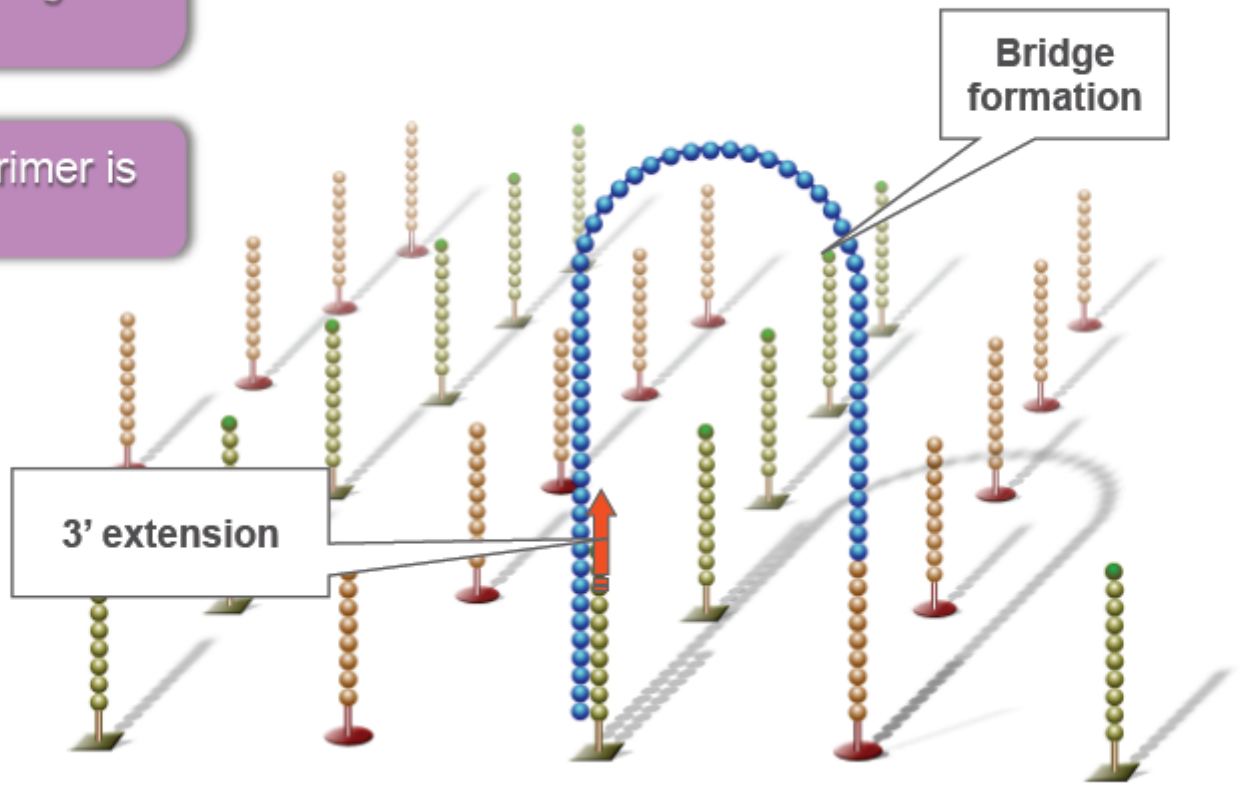
3’-ends of template strands and lawn primers are unblocked



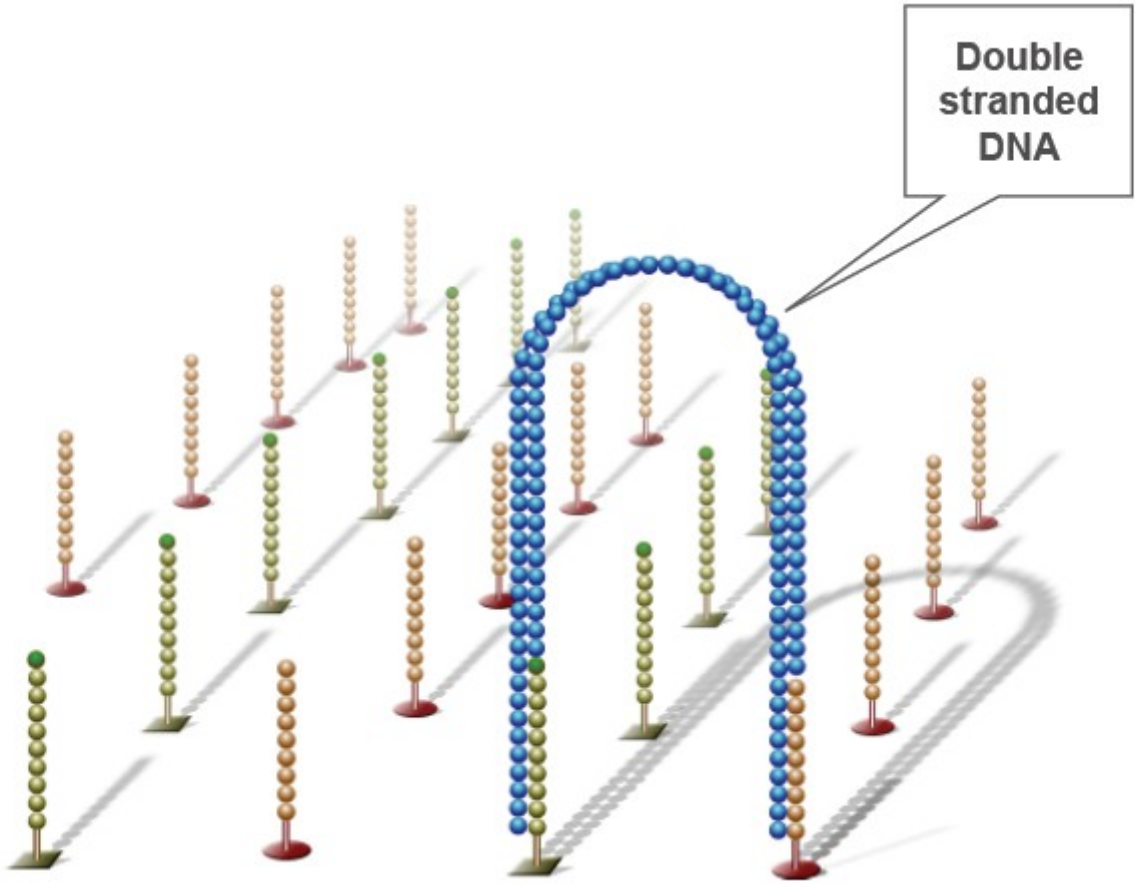
Paired End Sequencing

Single-stranded template loops over to form a bridge by hybridizing with a lawn primer

3’-ends of lawn primer is extended

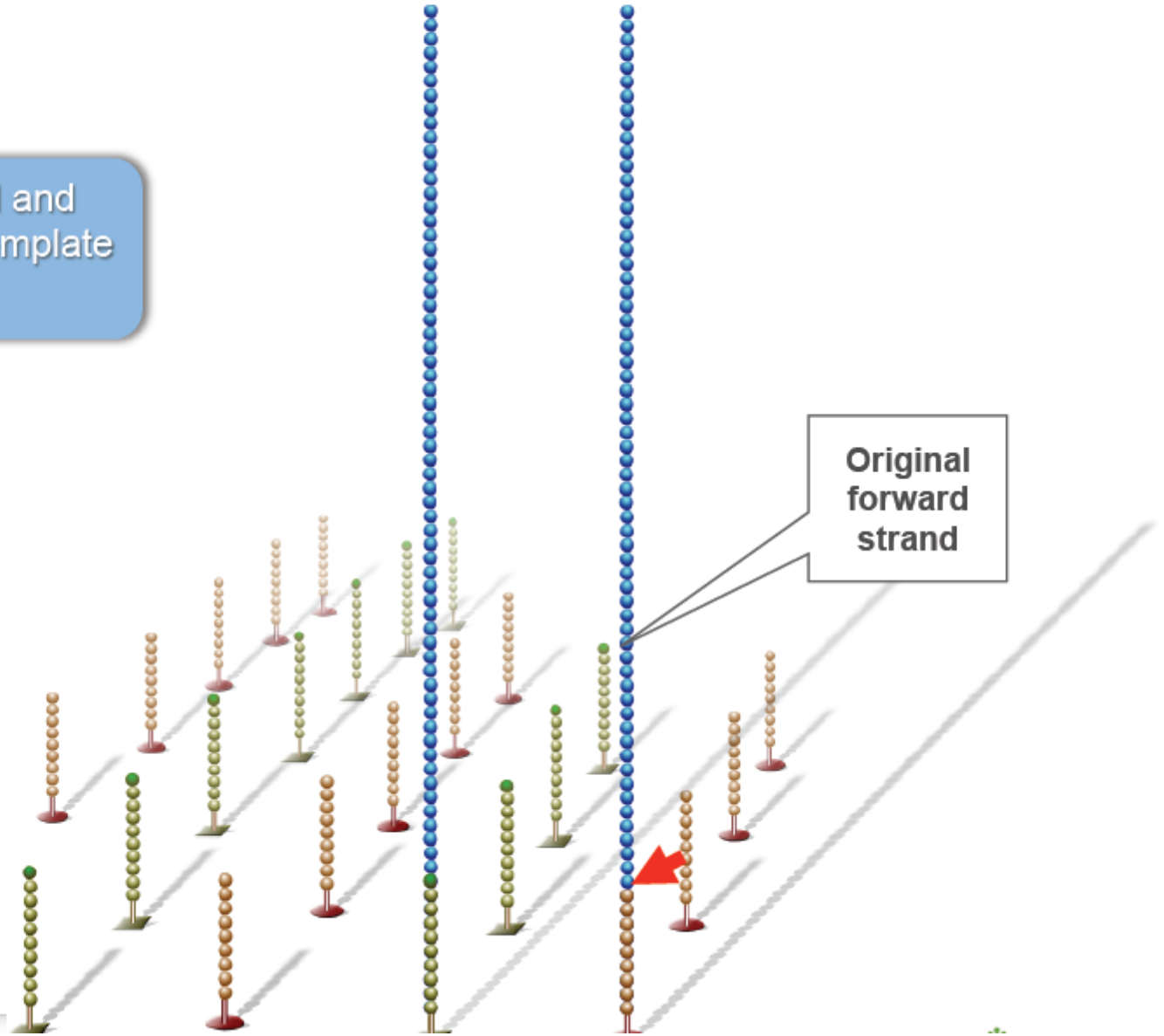


Paired End Sequencing



Paired End Sequencing

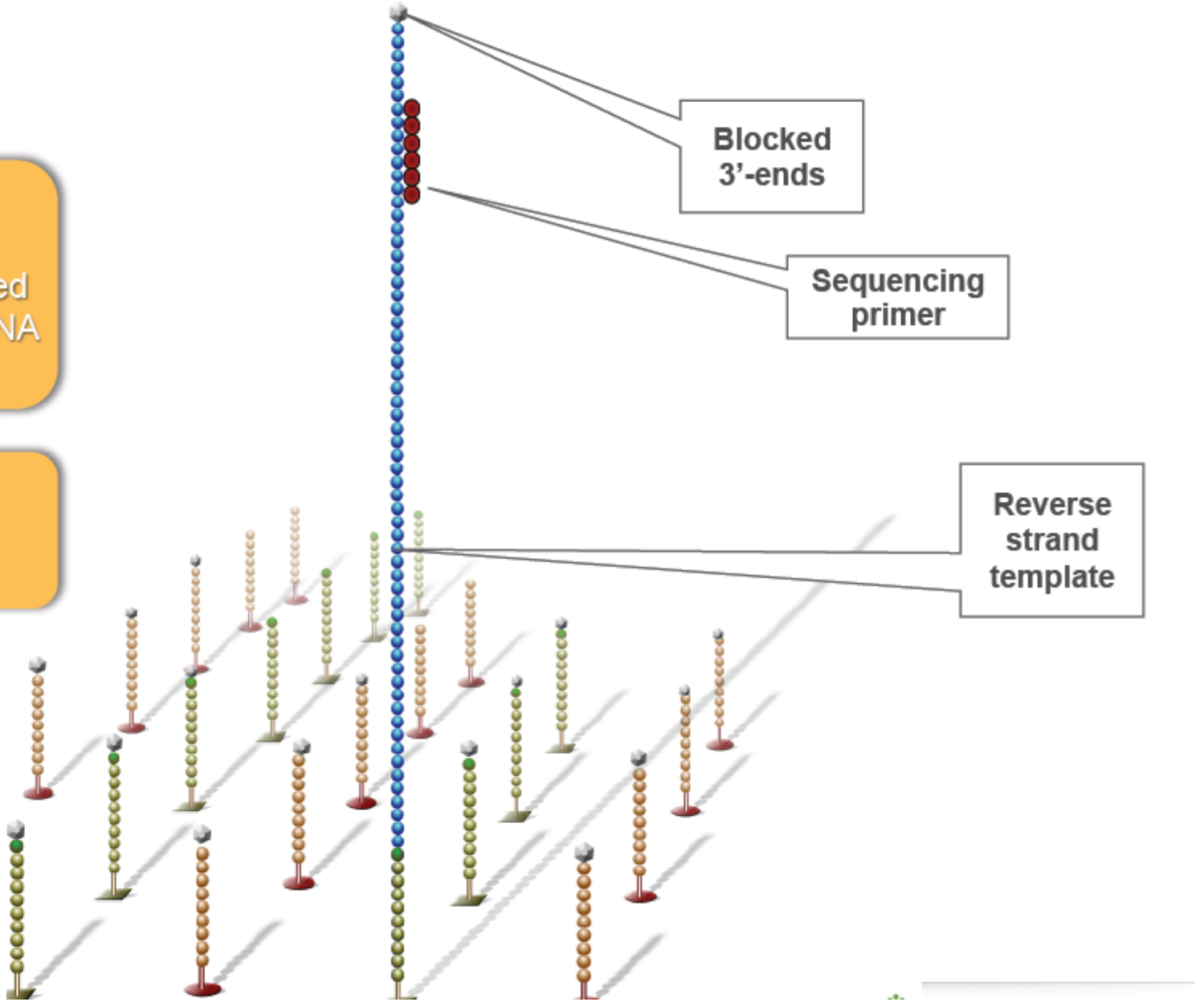
Bridges are linearized and the original forward template is cleaved

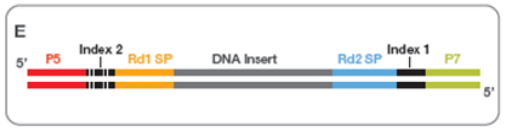


Paired End Sequencing

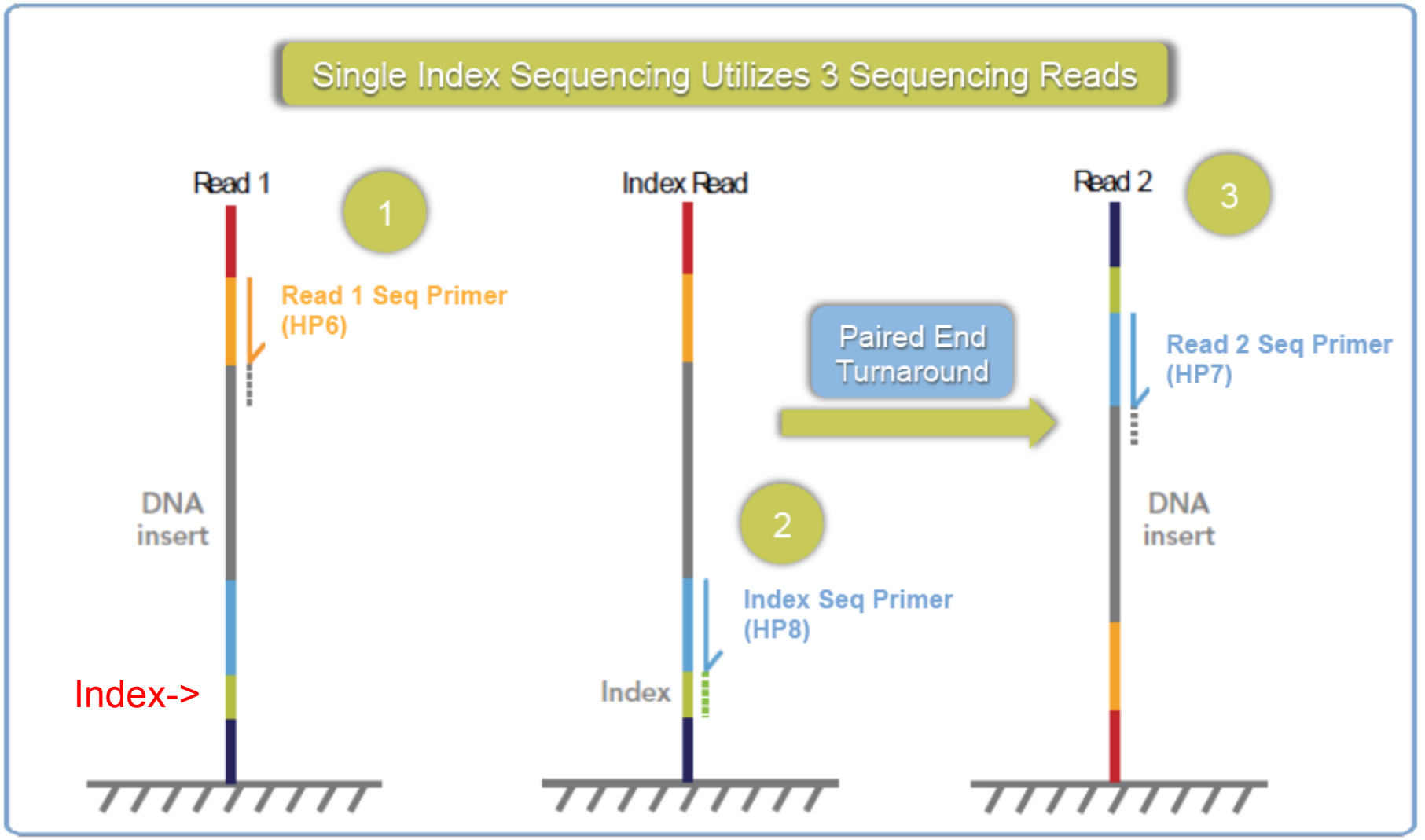
Free 3’ ends of the reverse template and lawn primers are blocked to prevent unwanted DNA priming

Sequencing primer is hybridized to adapter sequence

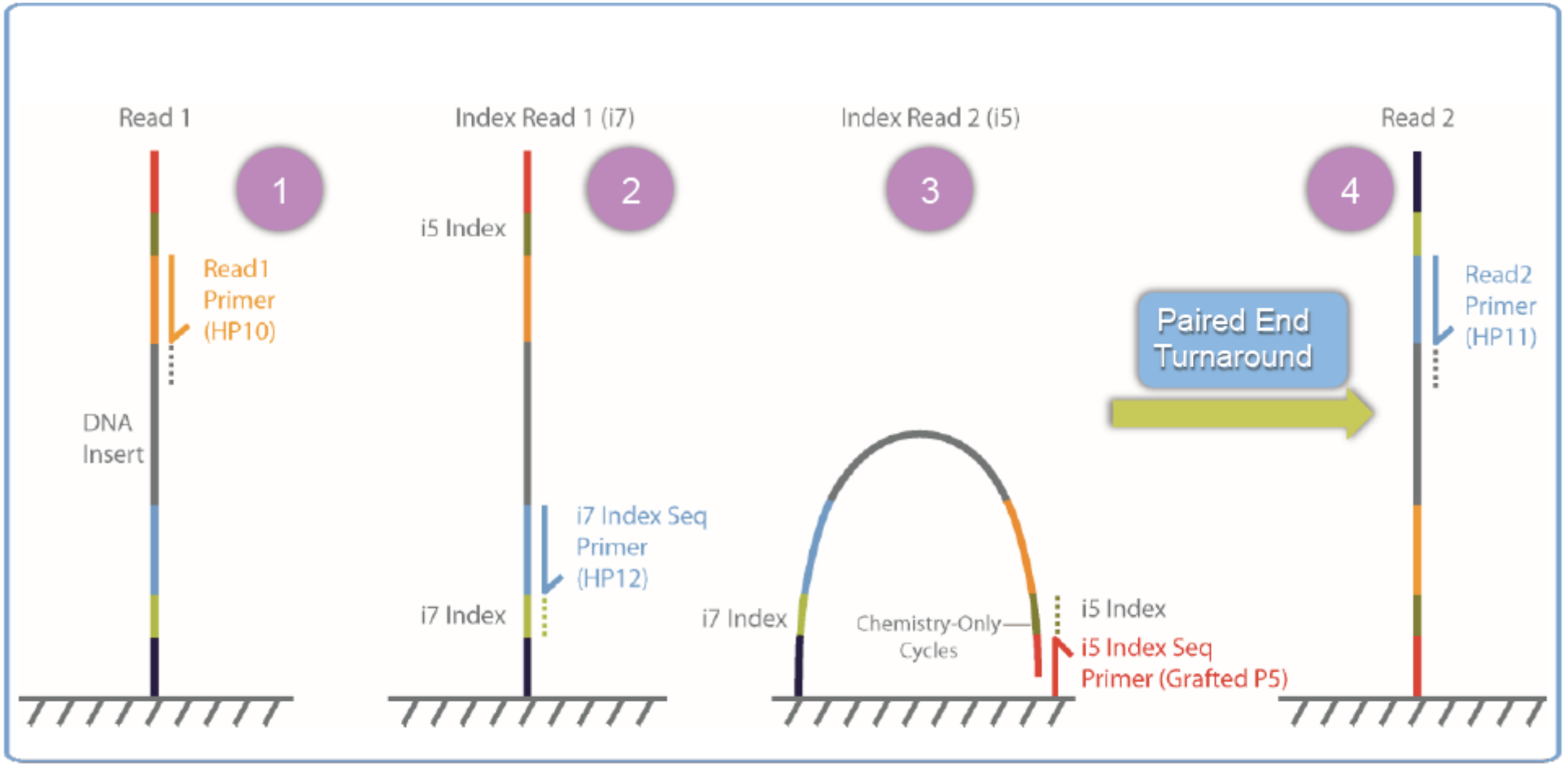




Sequencing with Paired Ends



Sequencing Paired End Libraries with Dual Index Read



Dual Index Sequencing Utilizes 4 Sequencing Reads

video

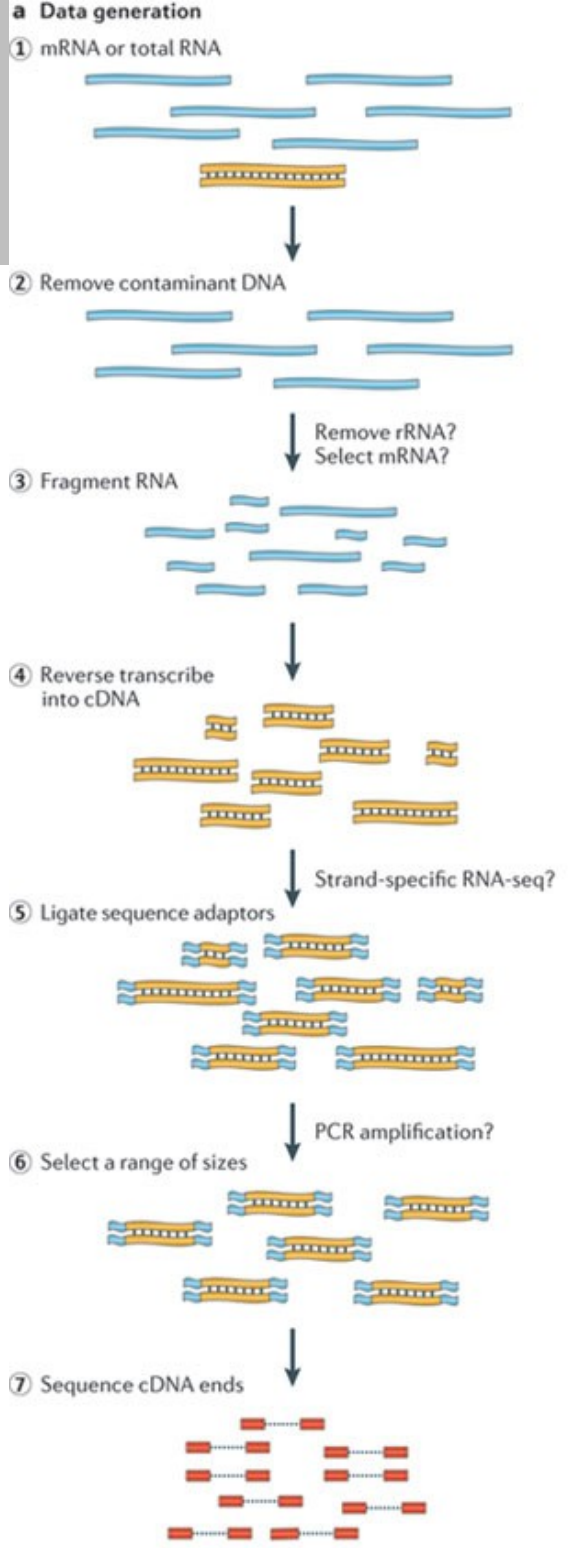
- https://www.youtube.com/watch?v=womKfikWlxM

RNA Seq

RNAseq

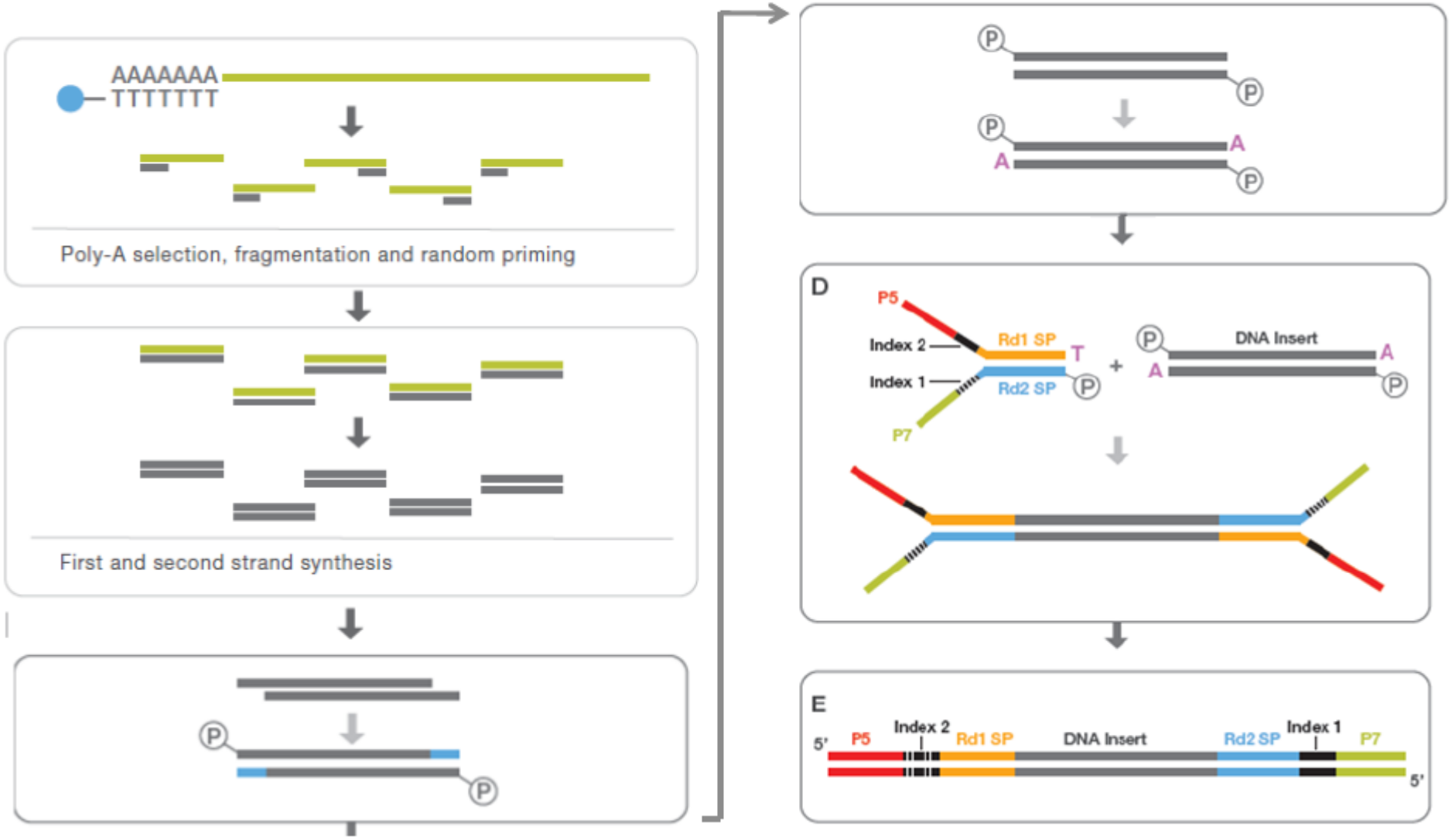
The general experimental procedure for RNA

Transcriptom = sum of all RNA (mRNA, rRNA, tRNA and noncoding RNA)

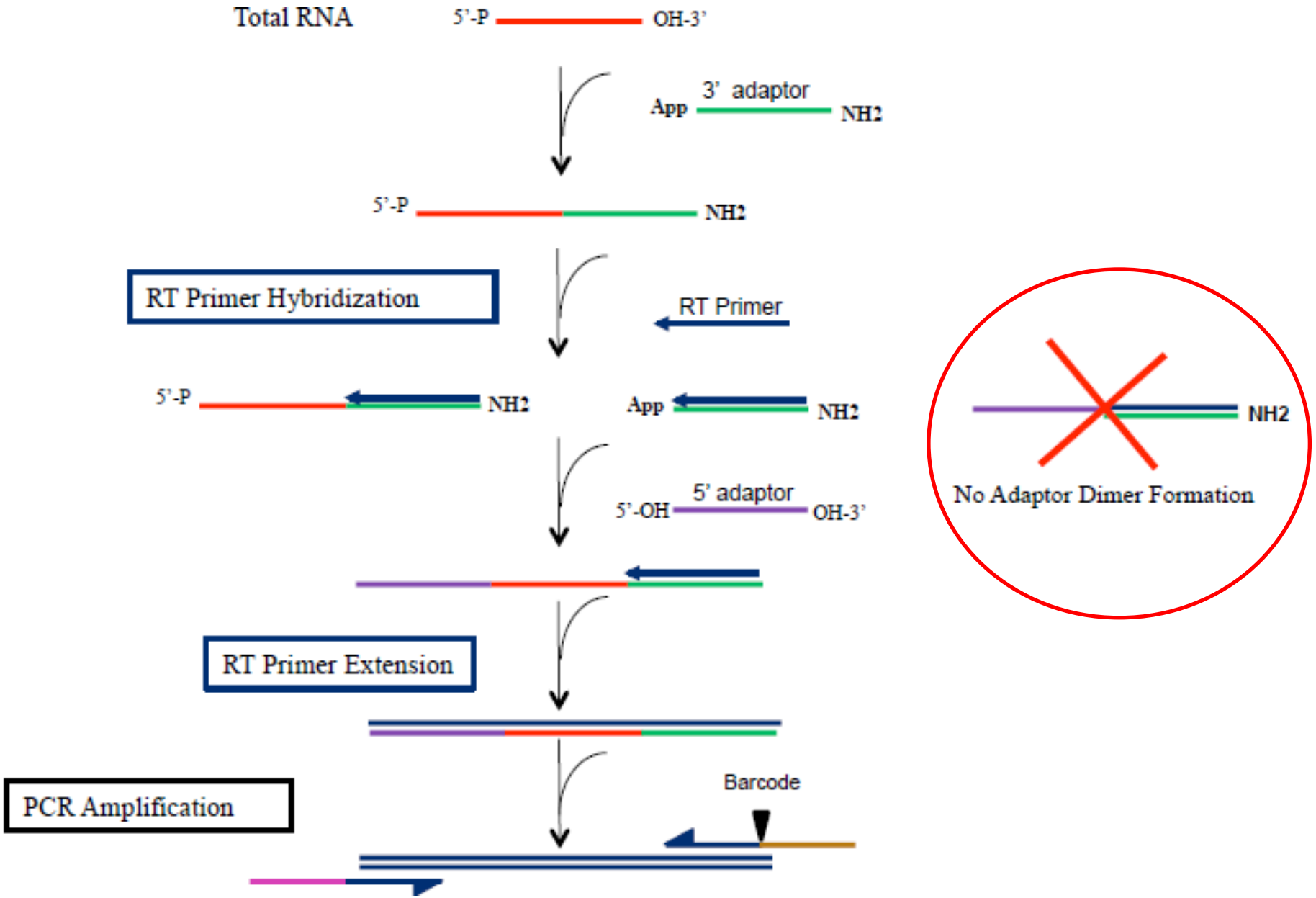


TruSeq RNA v2 Sample Prep Workflow

Blunt end, adenylation



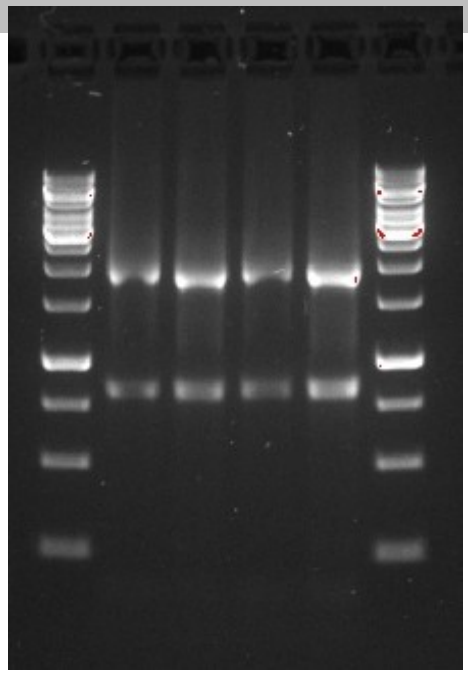
The general experimental procedure for miRNA

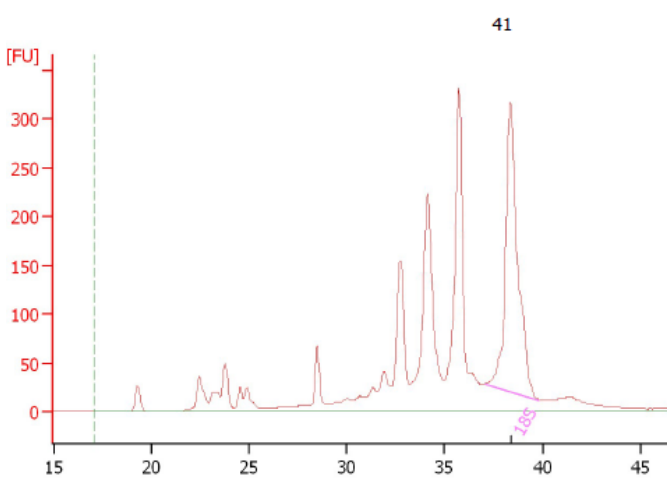


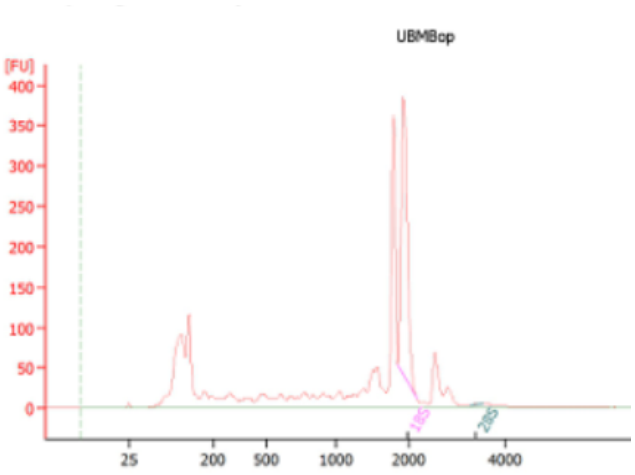
Library preparation

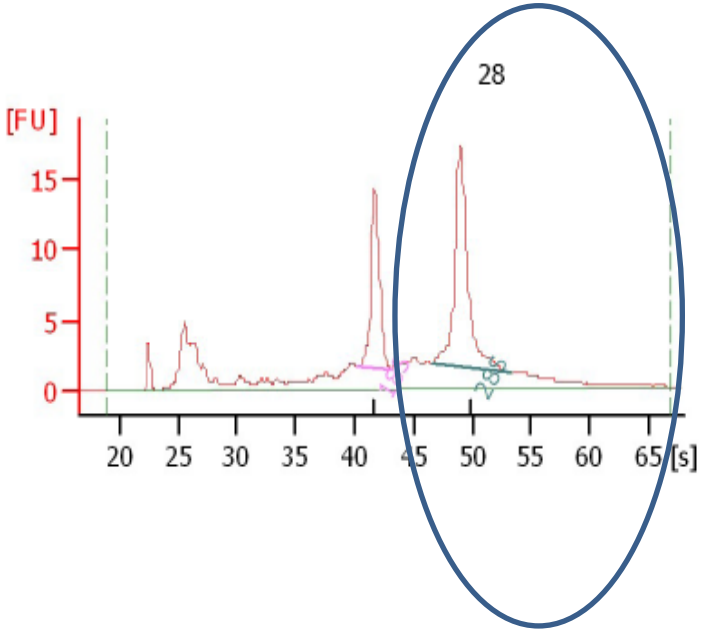
Strict QC of starting material

- appropriate quantification
- gel images, bioanalyzer traces
- which carrier was used – salmon sperm DNA, yeast RNA ☹, linear acrylamide ☺
- How to get rid of rRNA…



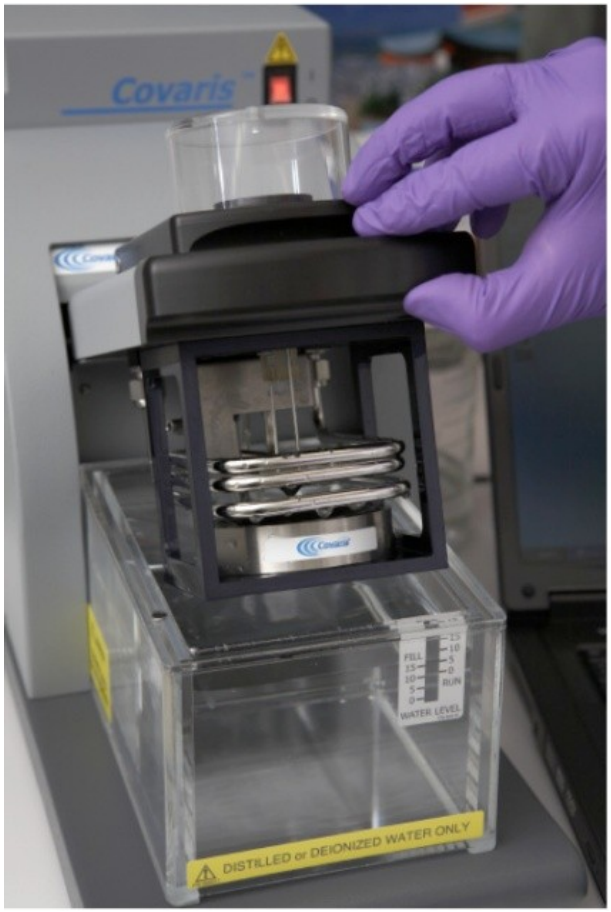






Library preparation

- Fragmentation: Covaris, enzymes, for RNA ions+heat
- Size selection: gel vs beads

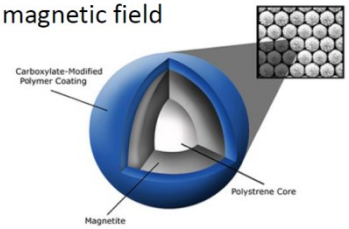


Covaris

Library preparation

How do SPRI beads work?

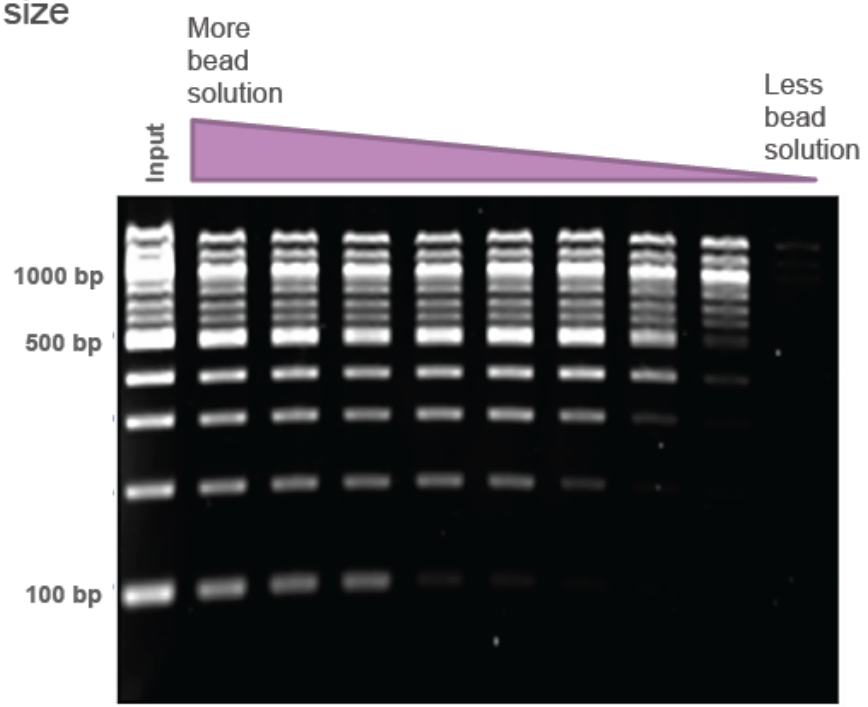
- **S**olid **P**hase **R**eversible **I**mmobilisation beads
- Polysterene core covered with magnetite
- Outer polymere coating
- Only magnetic in a magnetic field
  → Paramagnetic



- Depending on how much SPB are added, the DNA of interest might be bound to the beads or found in the cleared supernatant

Volumetric ratio of SPB to DNA solution is critical. Adding too much or too little SPB could affect final library size

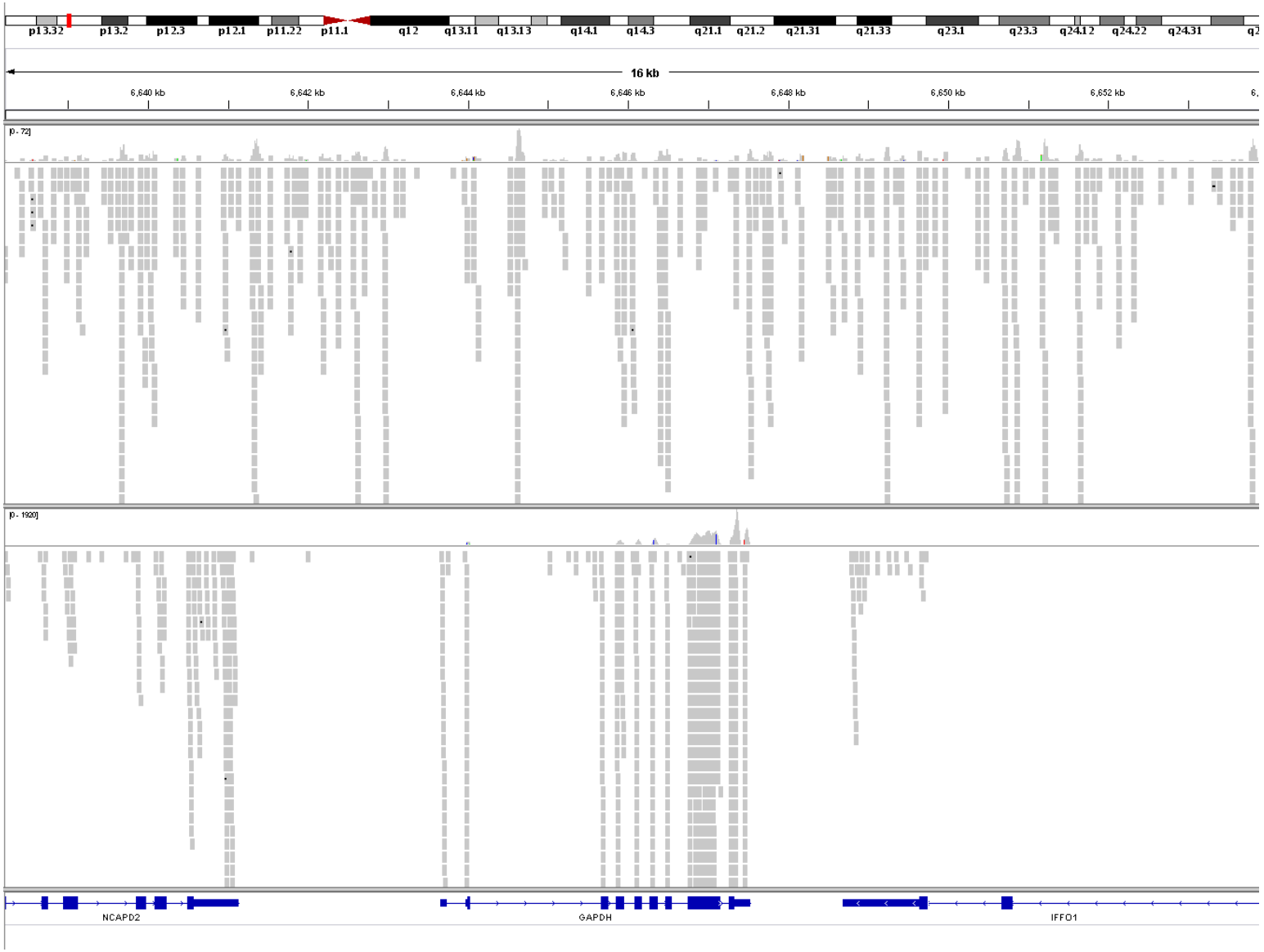
E-gel



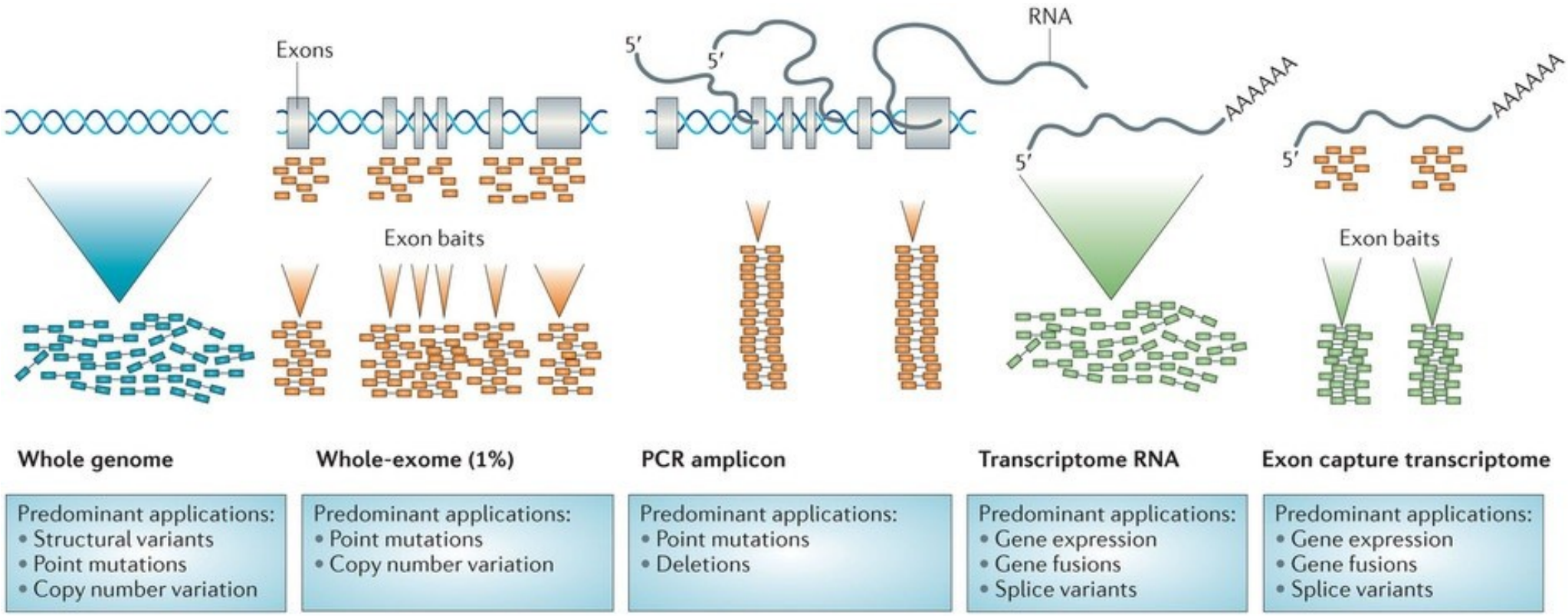
GAPDH

No DNase

After DNase



APPLICATIONS: NGS is good for many things



Genom Exom Amplicon Transcriptom

Simon 2013

Applications

- **De novo genome assembly**
- **Genome re-sequencing :**
  - SNV = single nucleotide variants (mutation/SNP)
  - CNV = copy number variation (insertion/deletion)
  - structural aberation (translocation/inversion)
- **RNA-Seq** (gene expression, exon-intron structure, small RNA profiling, and mutation)
- **CHIP-Seq** (protein-DNA interaction)
- **Epigenetic profiling**

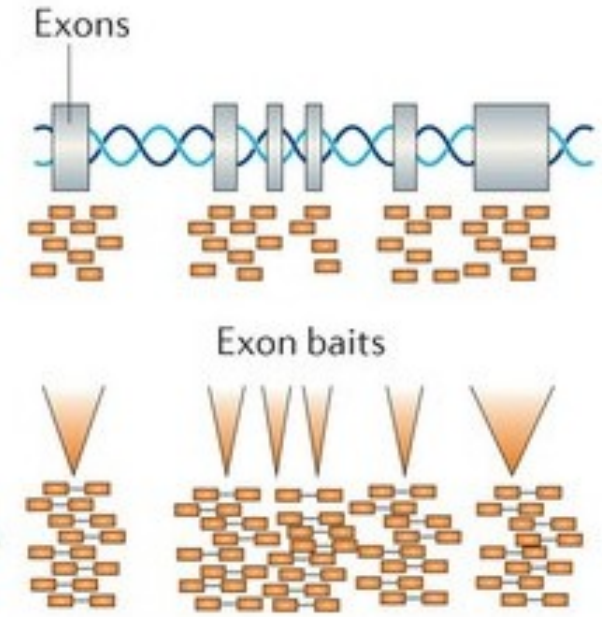
Whole Genome Sequencing

- You sequence all of that – including the „junk“

1. ***De novo* asembly** – using the overlap of the reads to assemble a genome – needs a good coverage
2. **Re-sequencing** – mapping to your reference genome …you need to have one

WES = whole exome sequencing

- You sequence only the coding regions of genes…exons (approx. 2 % of the genome)
- **Effective and cheap**
- **Probably the most widely used**



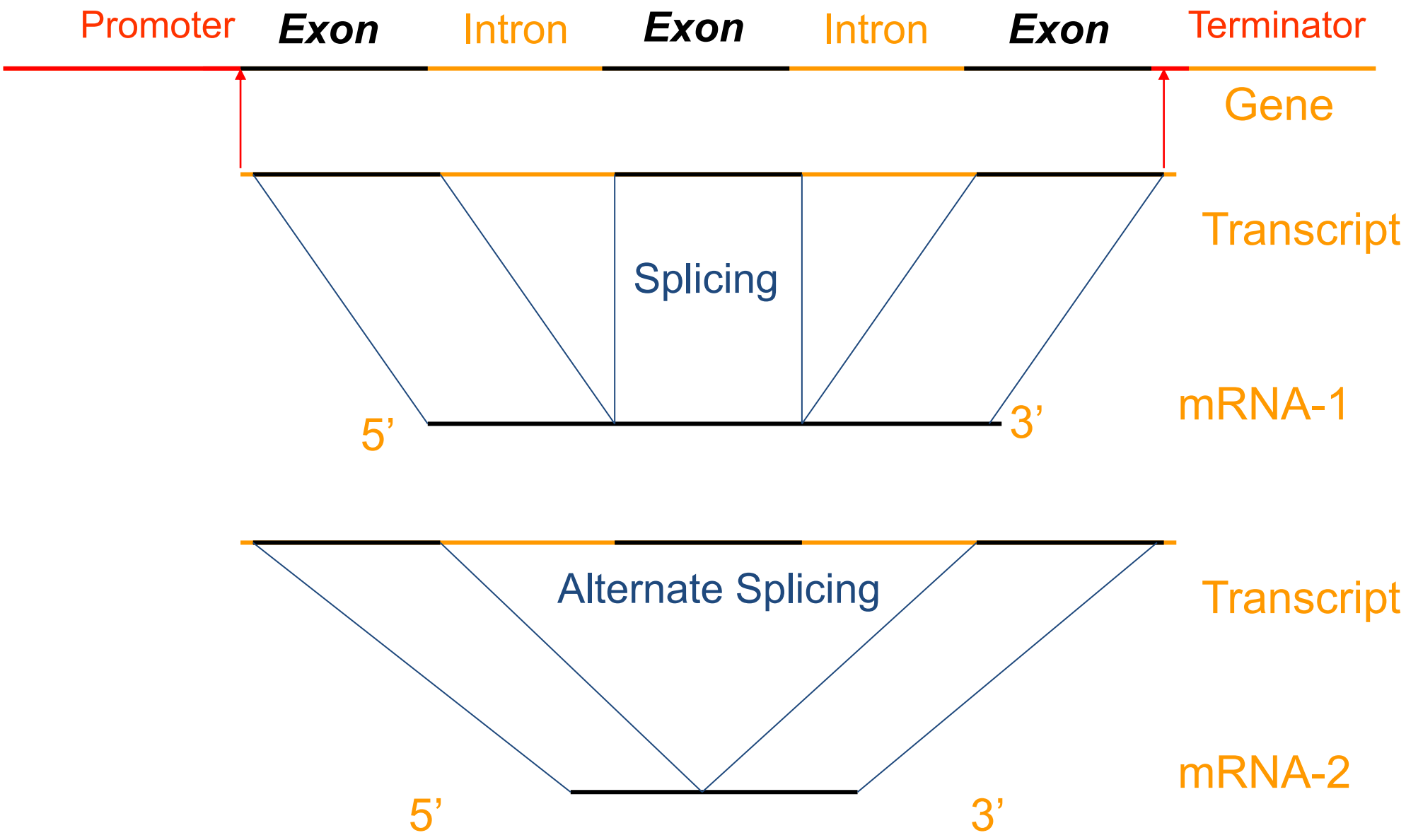
Targeted sequencing

- You already know the exact gene
- And you want to screen
- You are typically looking for a causative mutation that you know in advance can be there
- **Cheap and fast…. Good for detection of small clones using high coverage (but polymerase makes mistakes)**

RNA sequencing

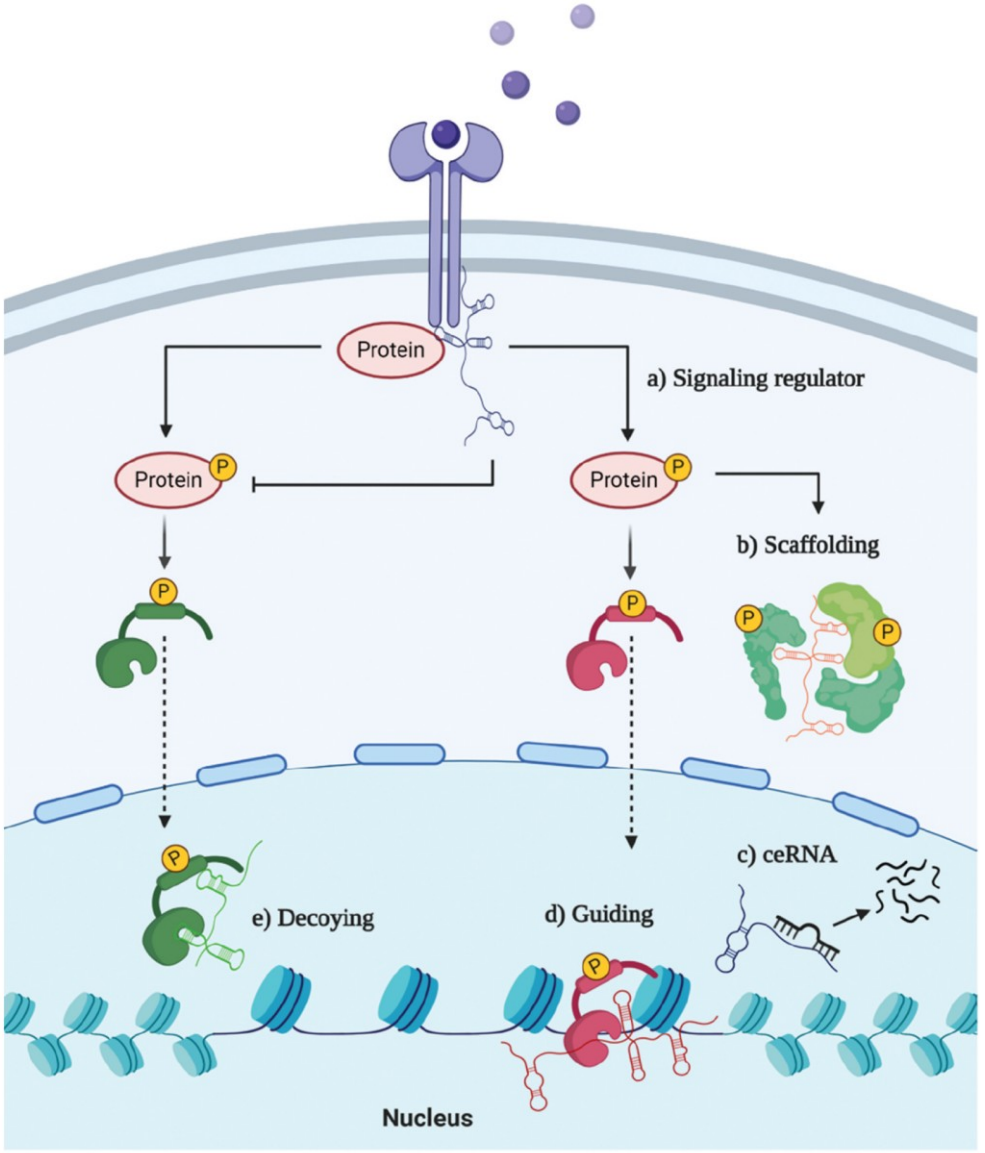
- Detection of expression levels…counting reads that map
- Somatic mutations (of expressed genes)
- Gene fusions
- **Alternative splicing**
- **ncRNA…a whole new universe**

Alternative Splicing Generates Distinct Proteins in Different Tissues



Discovering noncoding RNAs

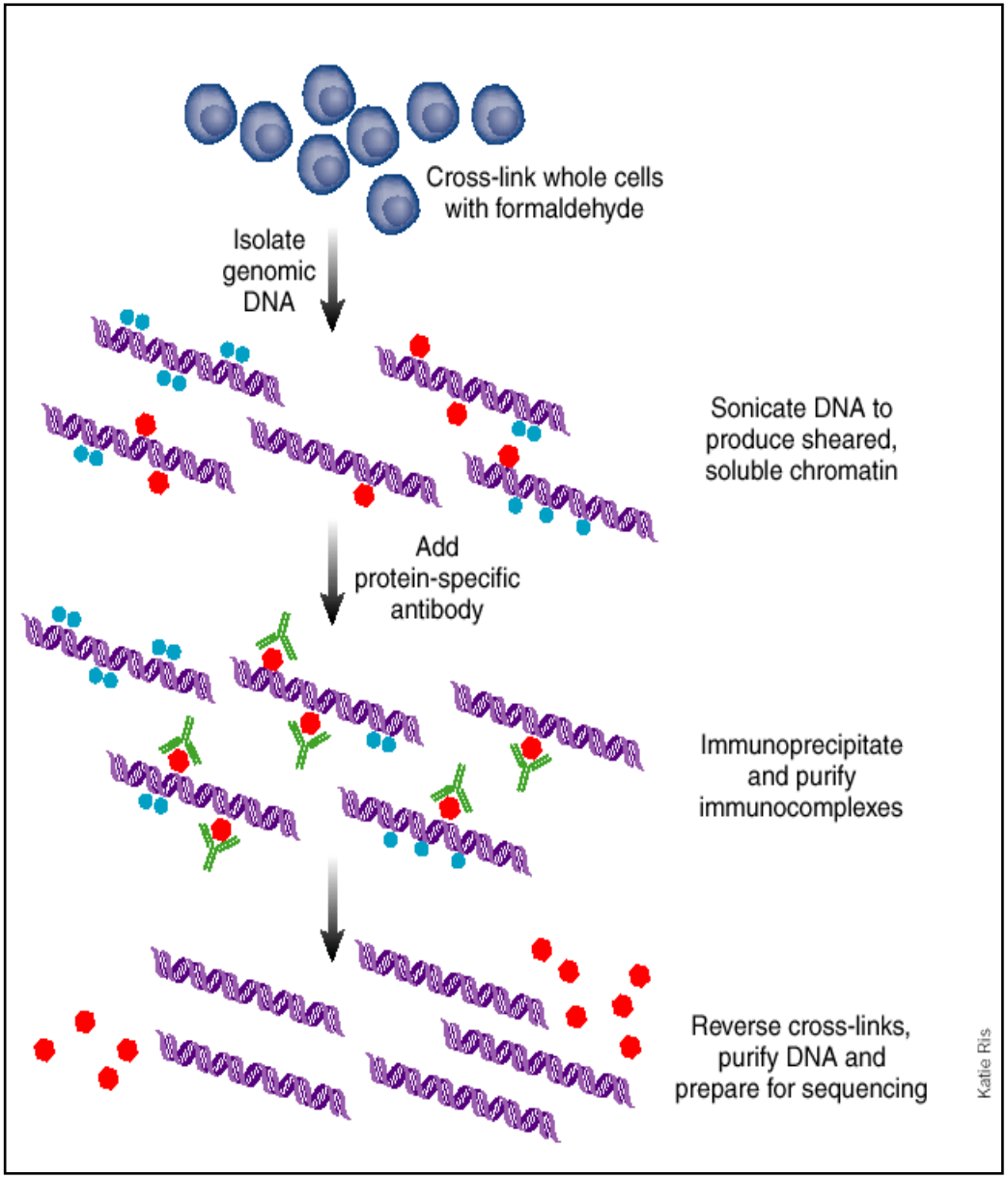
- ncRNA presence in genome difficult to predict by computational methods with high certainty because the evolutionary diversity
- Most have unknown function



Zeni and Mraz,. 2020

Elucidating DNA-protein interactions through chromatin immunoprecipitation sequencing

- Key part in regulating gene expression
- Chip: technique to study DNA-protein interaccions
- Readout of ChIP-derived DNA sequences onto NGS platforms
- Insights into transcription factor/histone binding sites in the human genome

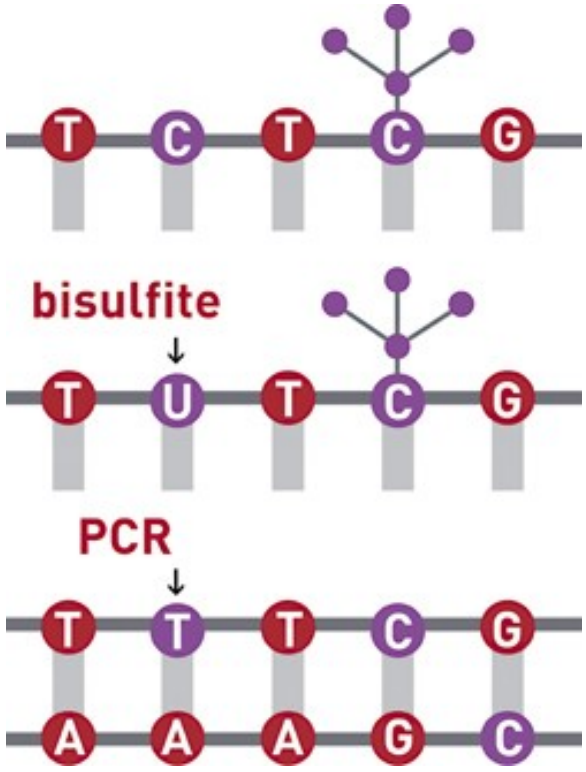


Epigenomic variation

- Enable of genome-wide patterns of methylation and how this patterns change through the course of an organism's development/cancer etc.

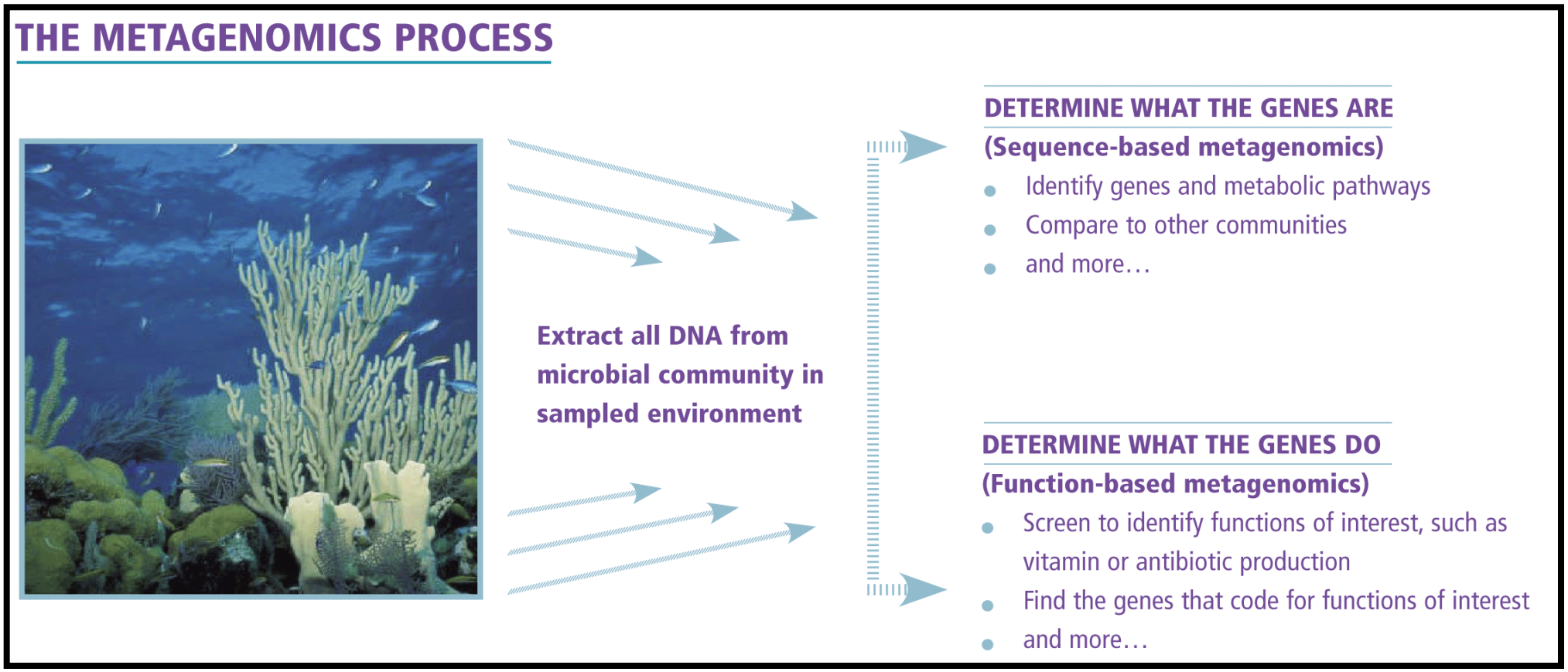
Bisulfit conversion + NGS:

- conversion C $\rightarrow$ U, Met-C not changes
- Identification of methylated bases

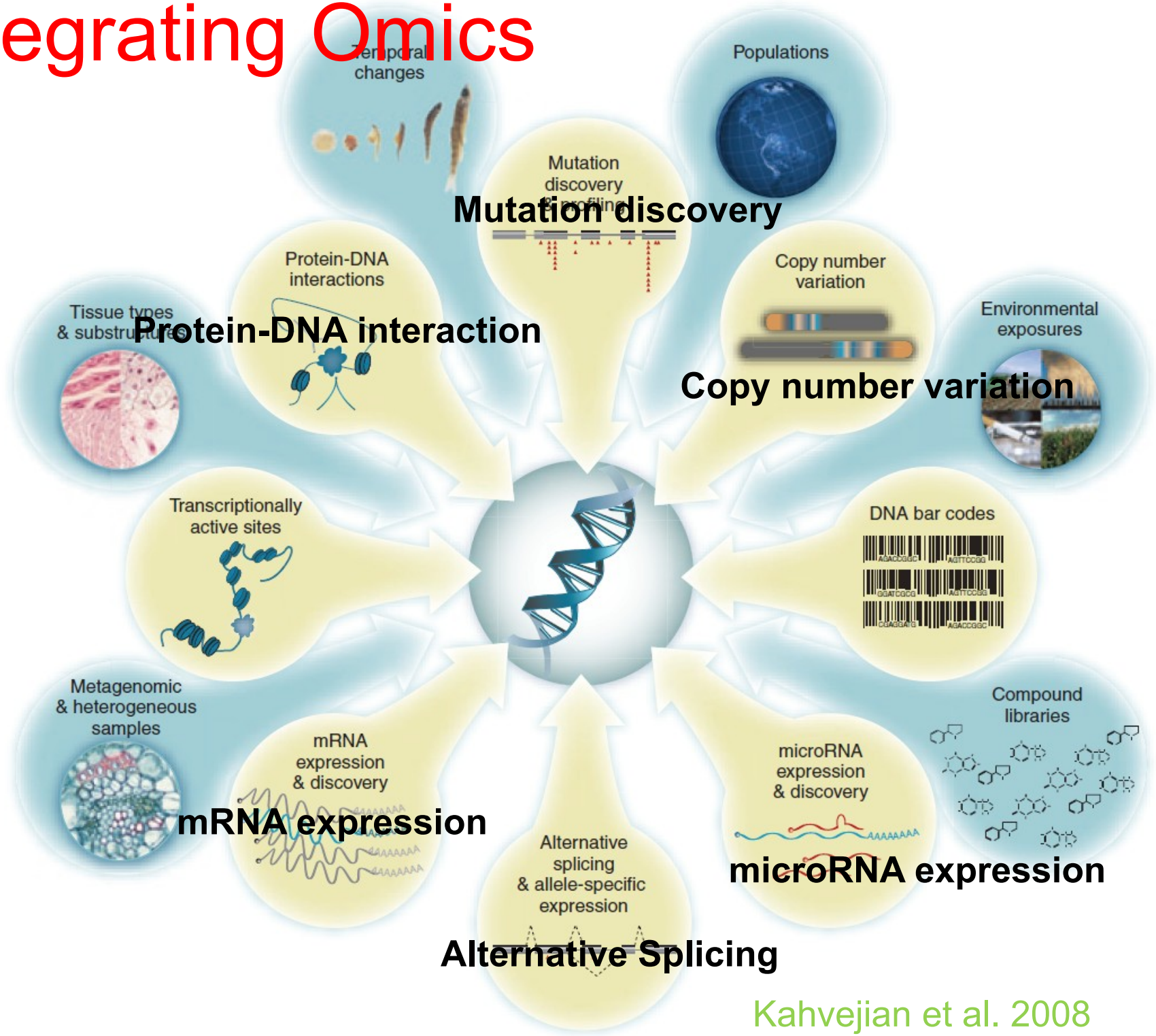


Metagenomics

- Examples: ocean, acid mine site, soil, coral reefs, human microbiome which may vary according to the health status of the individual

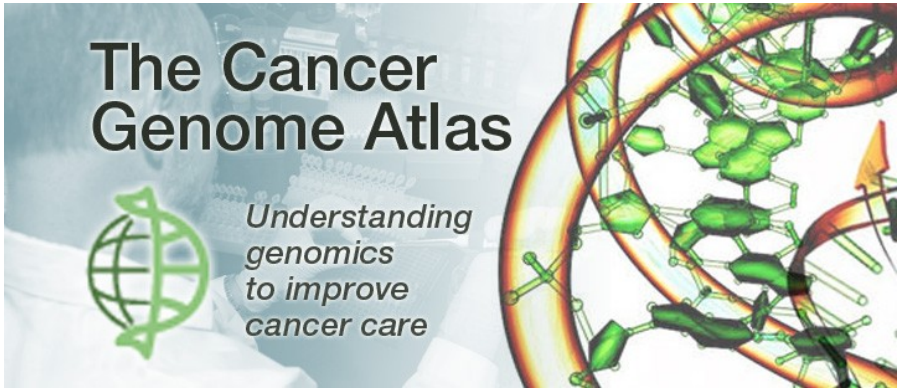


Integrating Omics



Kahvejian et al. 2008

- National Cancer Institute (NCI):– The Cancer Genom Atlas (TCGA)



- International Cancer Genome Consortium (ICGC): Cancer GenomeProject – genome, transcriptom and epigenom in 50 most common tumors

International
Cancer Genome
Consortium

Examples of NGS applications in Oncology

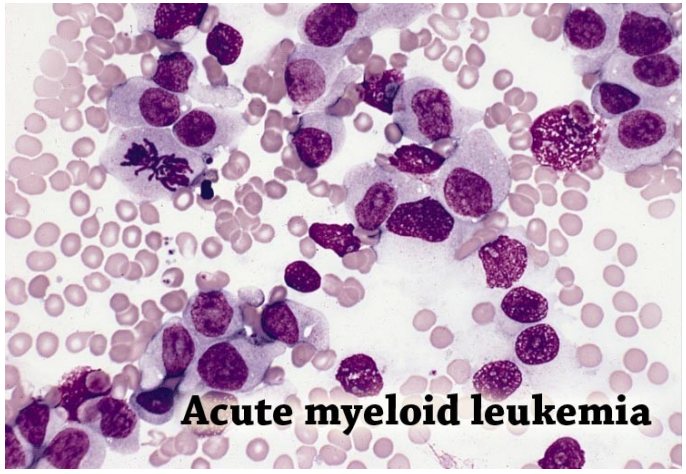
- **Molecular diagnostics**…mutations: known, novel and subclonal
- **RNA seq:** new fusion genes
  - *Fusion EML4- ALK* in lung cancer
  - translocation *TMPRSS2- ERG* in prostate cancer(Dong 2012)
  - microRNA expression, gene patterns

- **Identification of germinal mutations (WES):**
  - Familiar pancreas cancer(PALB2)
  - Feochromocytoma inherited (MAX)
  - Familiar melanom (MITF)
  - ……..screening of large cohorts/families

- **Targeted sequencing**
  - BRCA1 mutations associated with breast and ovarian cancer (difficult to detect by sanger) (Walsh 2010)
  - ……good for huge genes

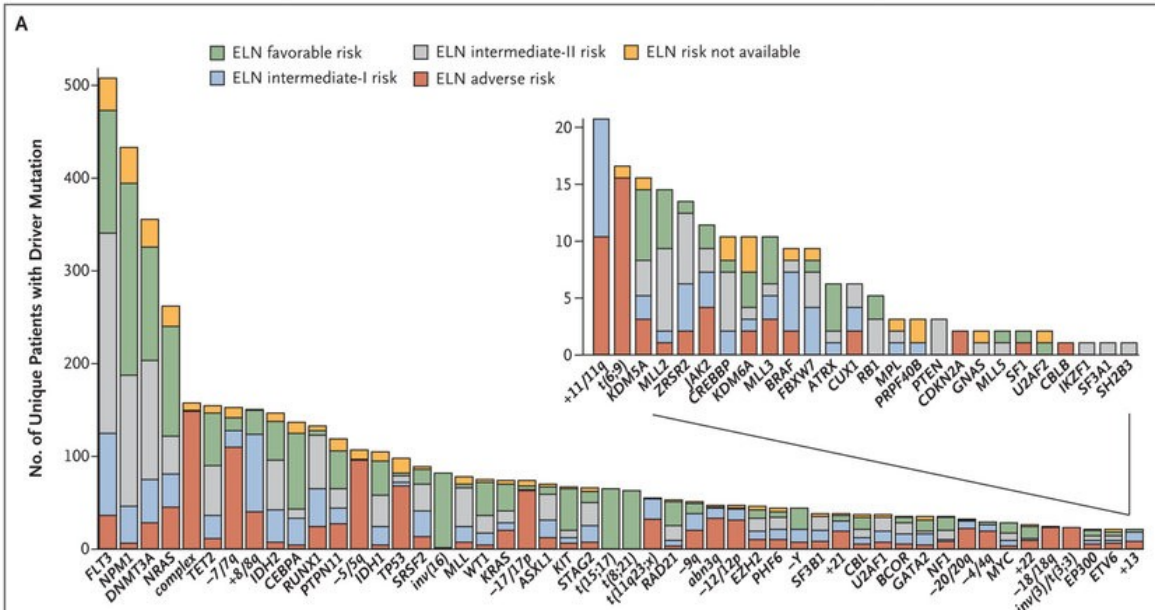
Hematooncology

- First genome of a cancer patient (WGS, 2008): normal cells vs AML cell $\rightarrow$ 8 new somatic mutations (Ley, Nature 2008)

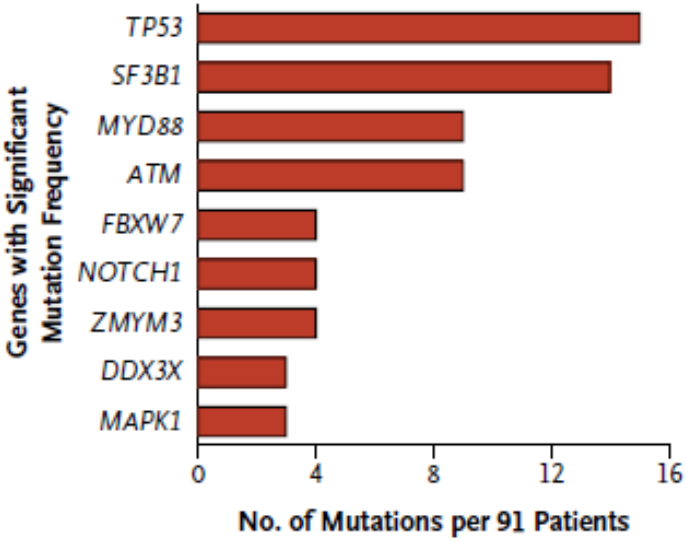


Identification of novel recurrently mutated genes by WES….

AML

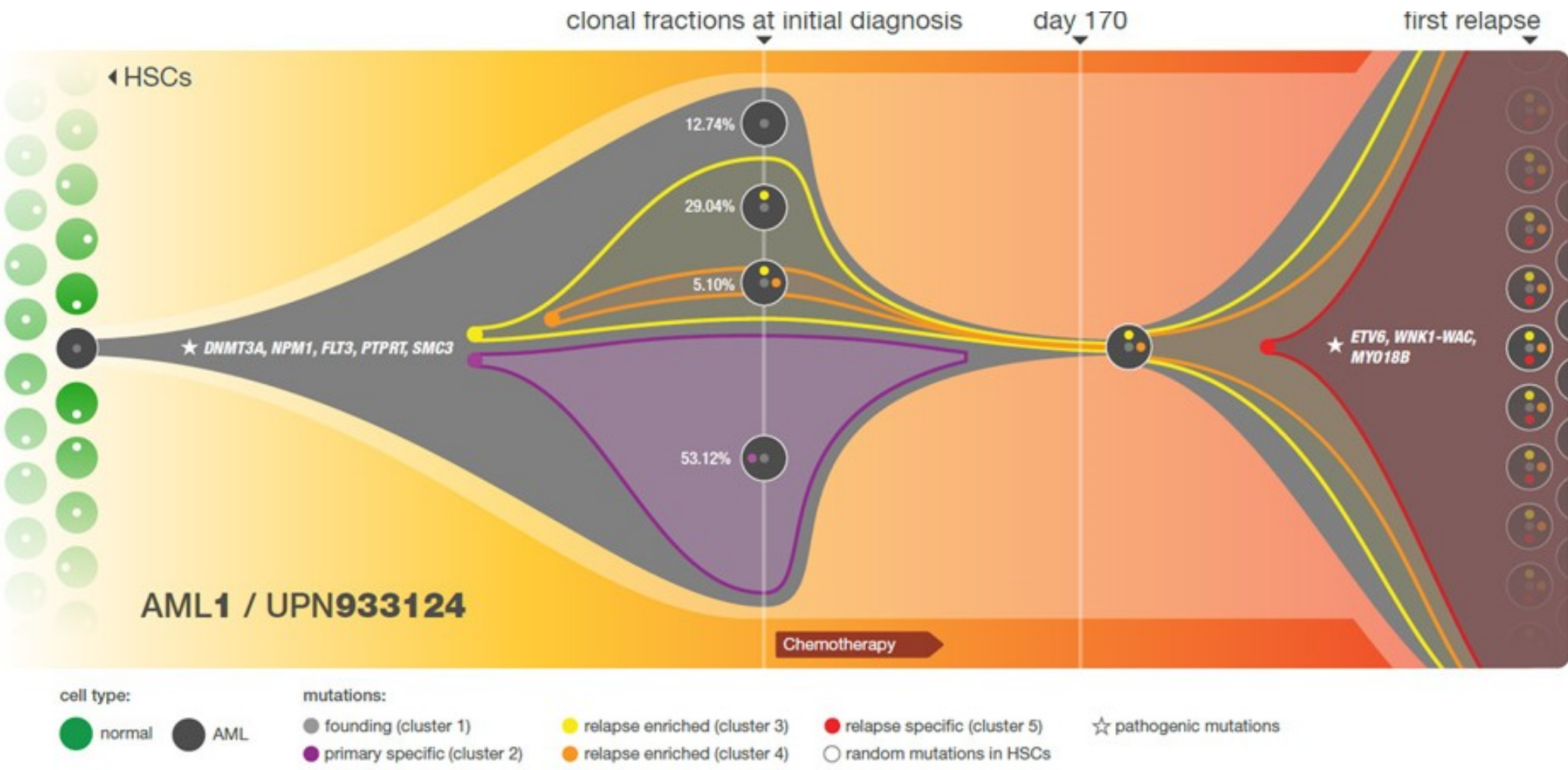


CLL



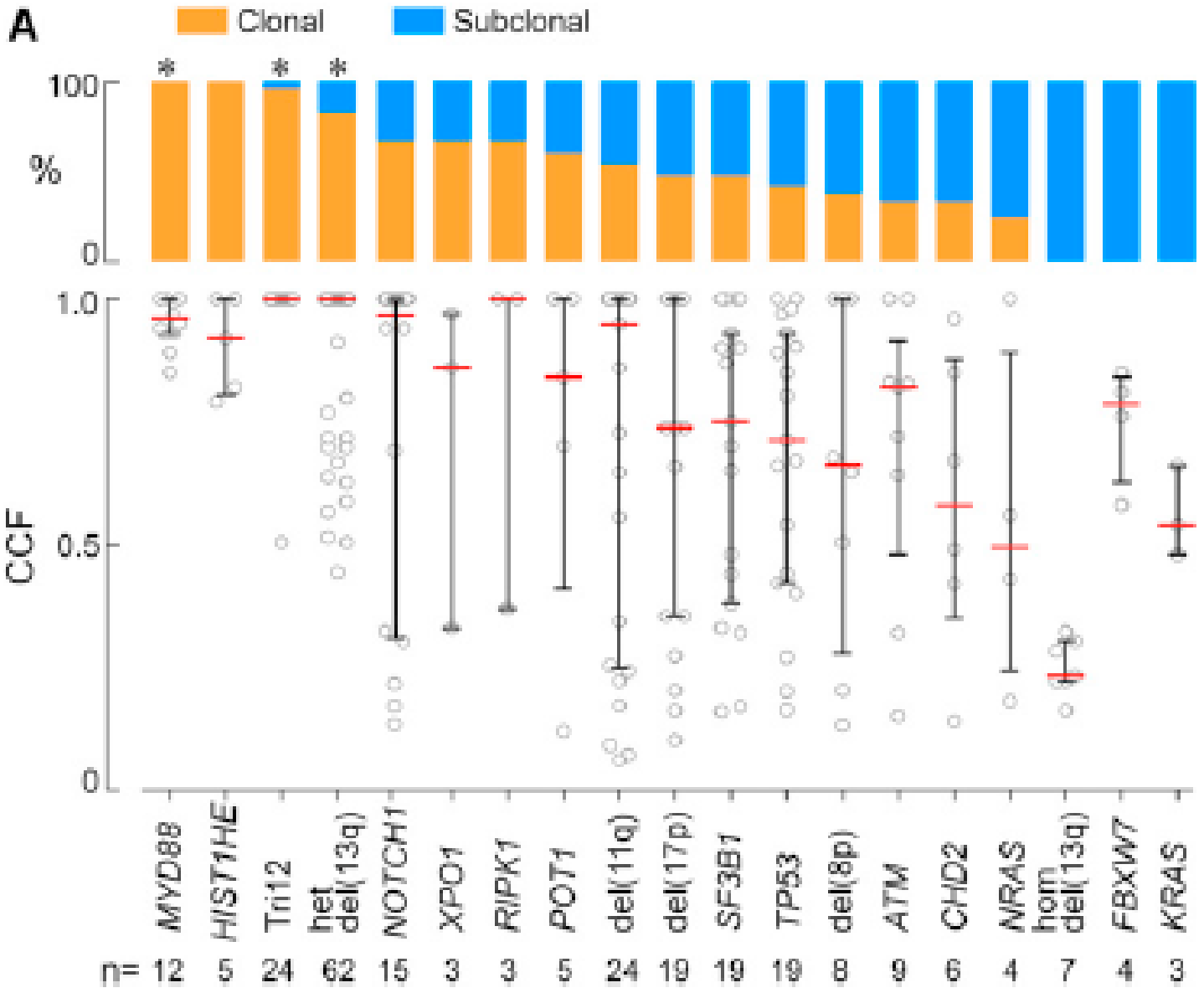
Wang 2011

clonal evolution in cancer : AML (WGS)



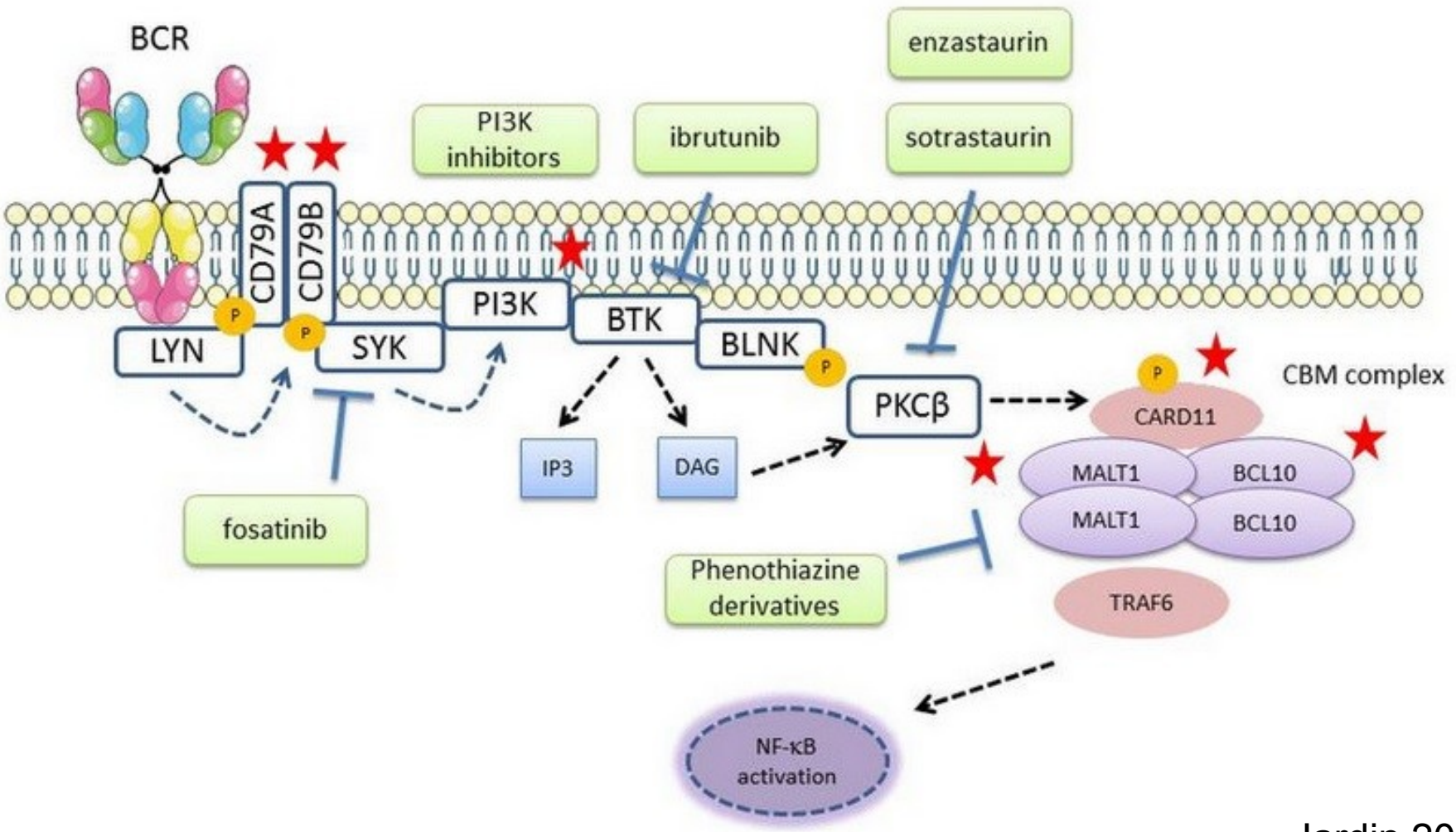
Ding 2012

Subclonal architecture of your tumor



Landau 2013

…. Including new therapeutic targets



Jardin 2014

Thank you for your attention

In summary: there is a whole new universe in front of you…. A one that nobody has ever seen

New technologies: https://nanoporetech.com/how-it-works

Marek Mraz

CEITEC and University Hospital Brno

marek.mraz@email.cz