Chapter 10 Gene expression profiling

Illumina HiSeq

Ion PGM

Nanopore MinION

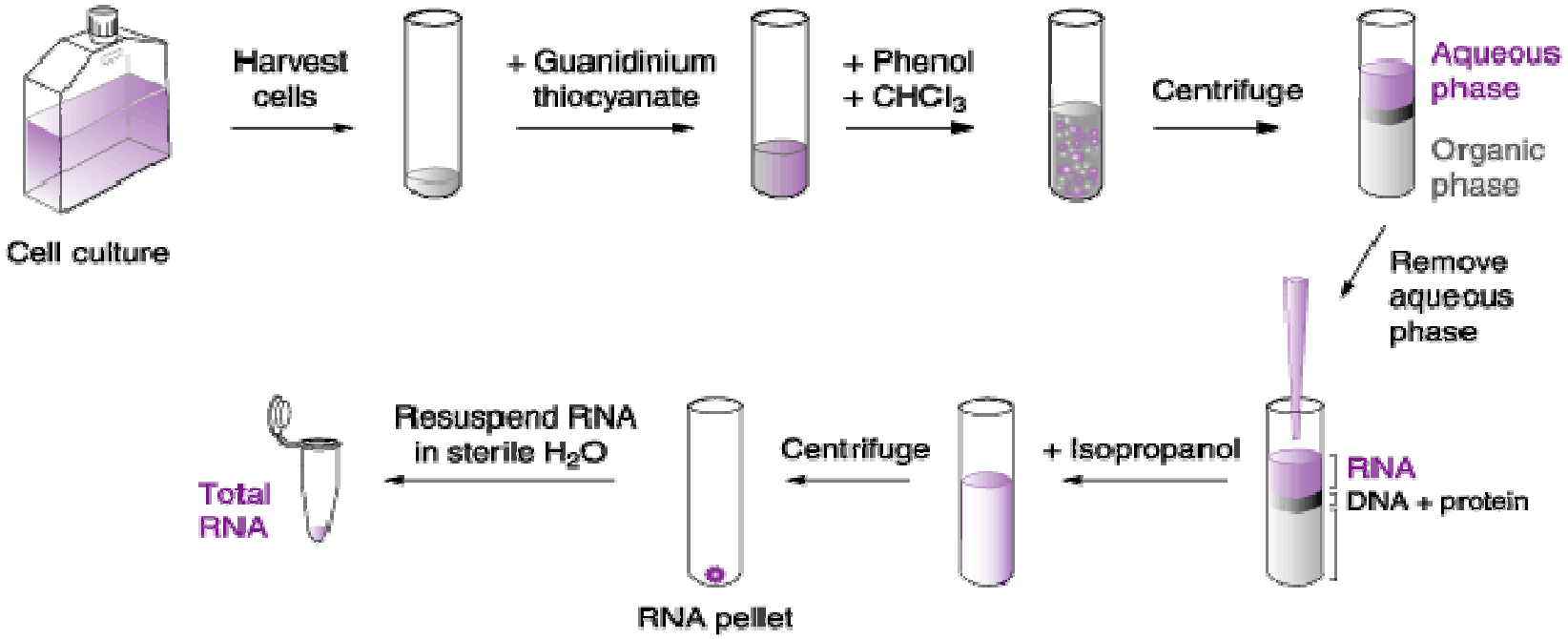
OUTLINE

Measurement of RNA expression levels

- Northern blot: qualitative detection of few targets
- qRT-PCR : quantification of transcript expression
- Genome-wide transcriptome analysis (microarrays)
  - Affimetrix versus glass slide
- Transcriptome analysis using high throughput sequencing technologies (HTS)
  - Amplification-based sequencing methods : Illumina HiSeq, Life Technologies Ion Torrent
  - Single molecule real time sequencing : zero-mode waveguide, Nanopore

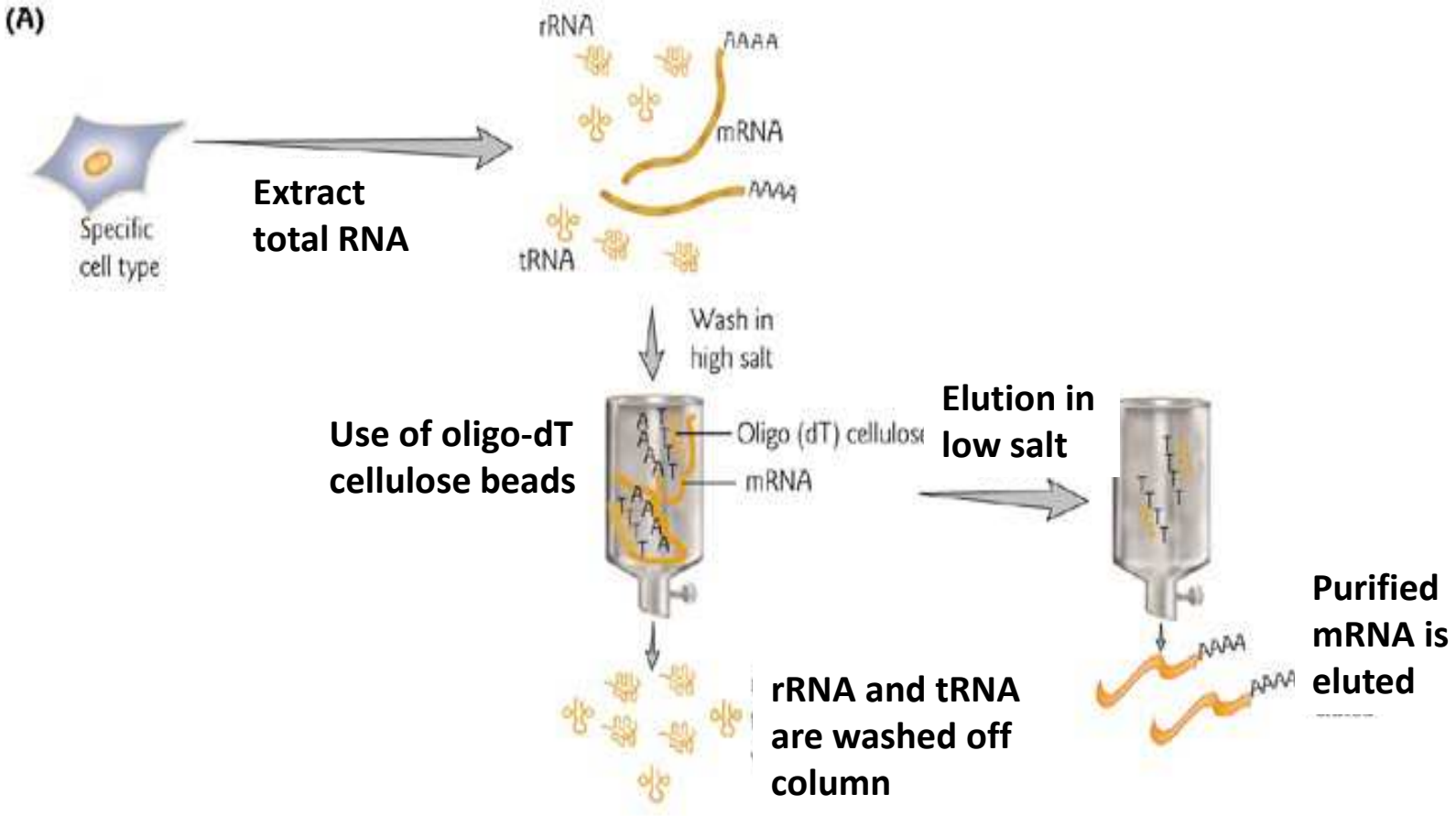
RNA purification and Northern blot analysis

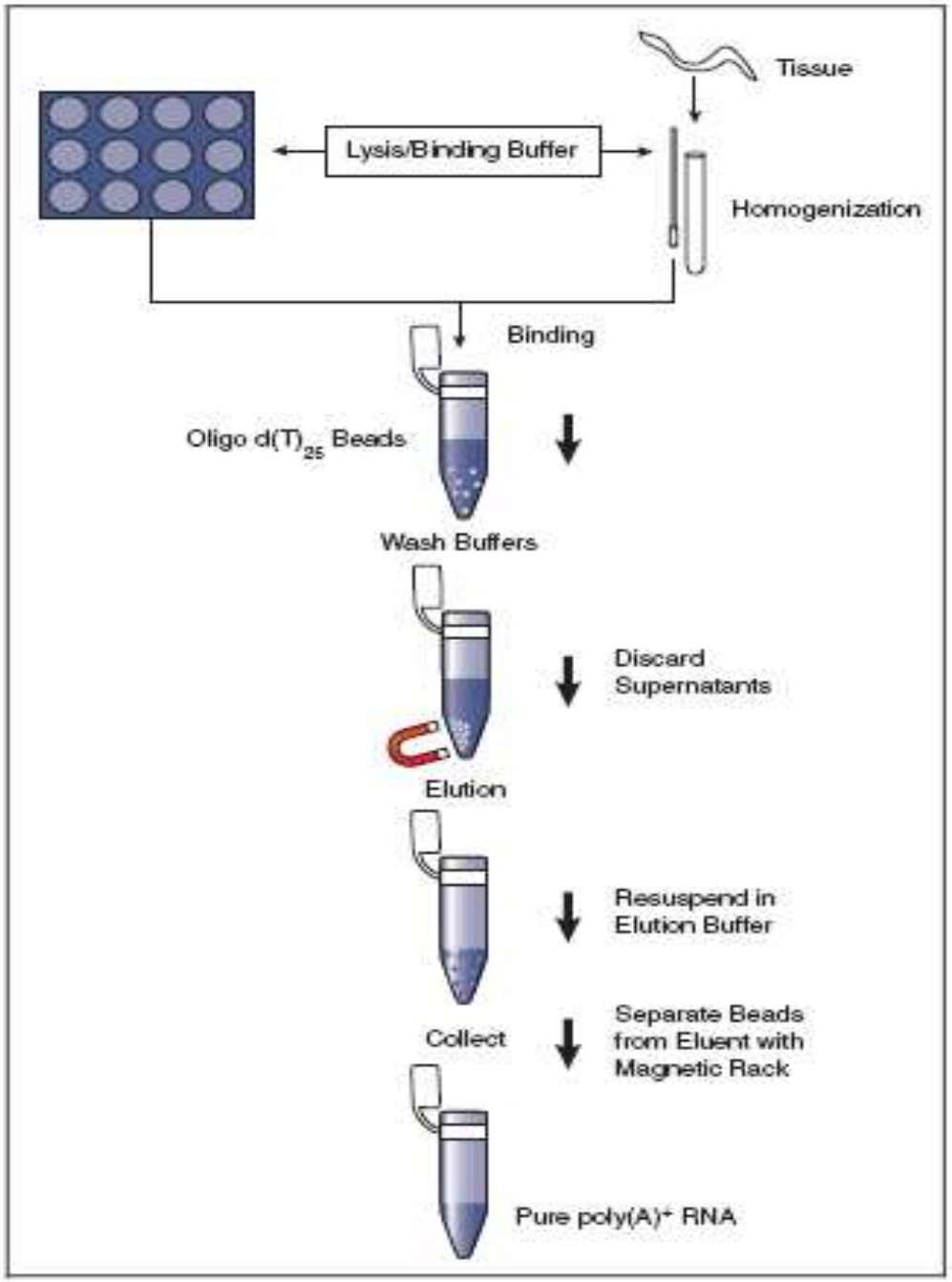
- Tissue homogenization and cell lysis
  - Glass beads
  - Sonication
  - Lytic Enzymes and/or detergents/chaotropic agents
  - Grinding of N_2-frozen sample in a mortar
- Protein denaturation and RNA extraction
- Alcohol precipitation



- Affinity chromatography purification of mRNA

mRNA represents a low fraction of total RNA population (5-10%)
Purification is required for microarray analysis and sequencing

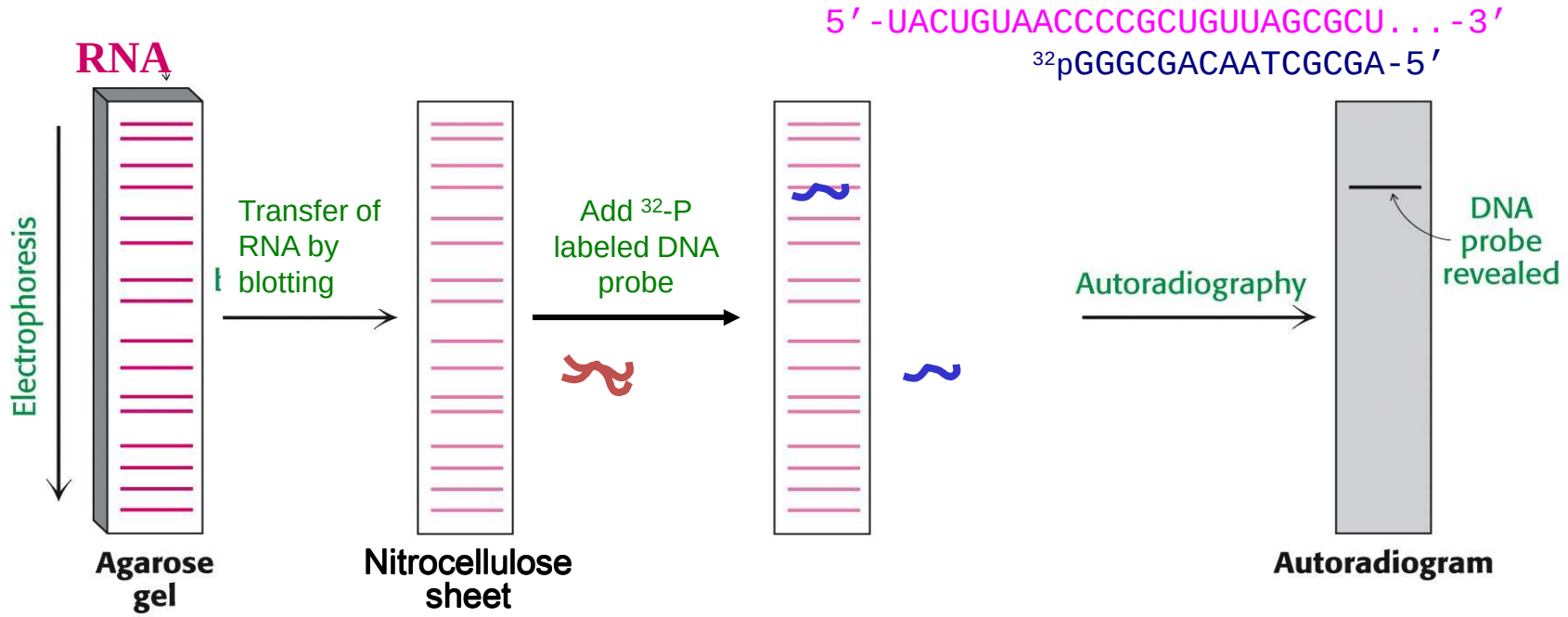




Development of various single-step purification kits (magnet, …)

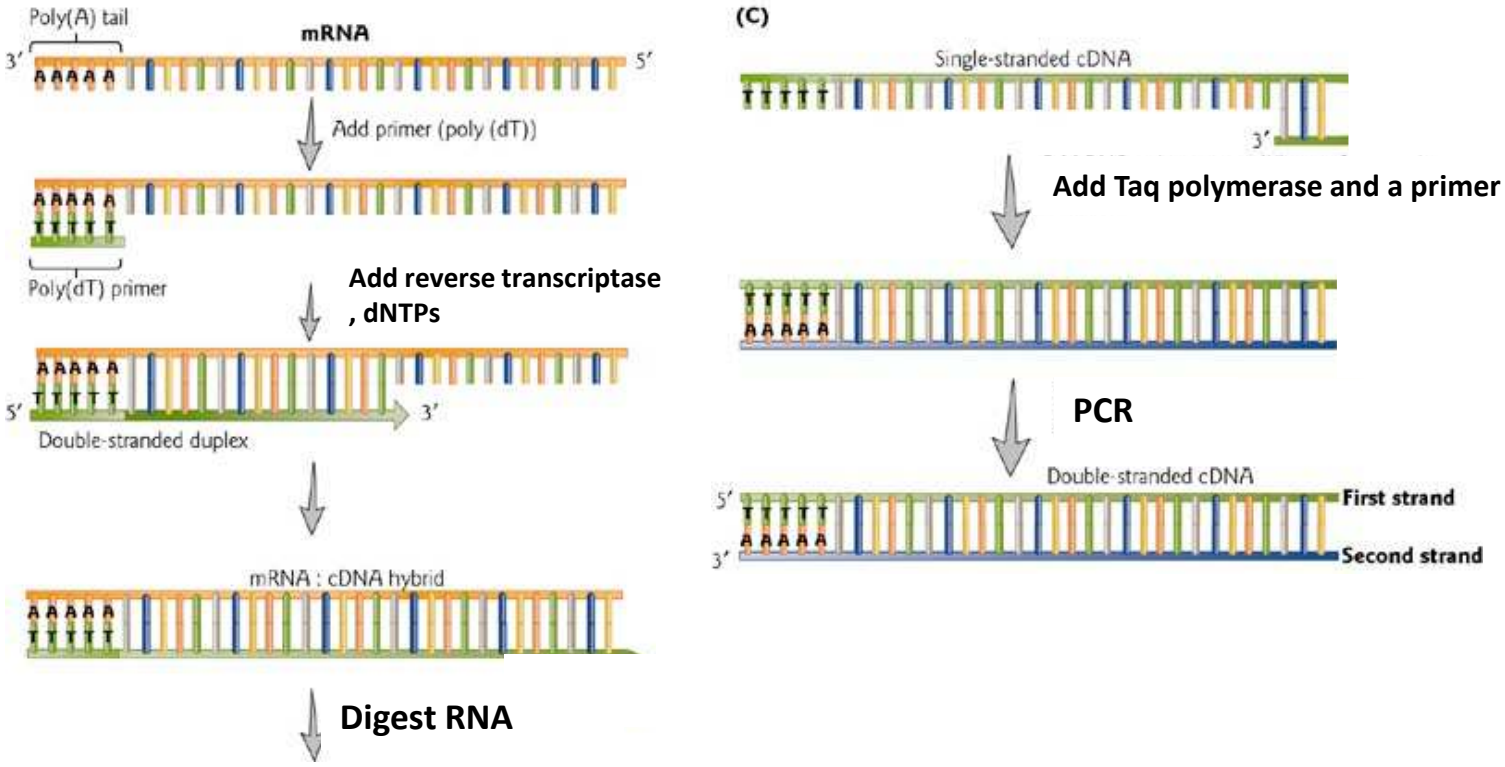
Northern blot analysis

- RNA samples are separated by gel agarose electrophoresis according to size
- RNA is capillary transfered to a nitrocellulose membrane
- Labelled DNA probe that is complementary to the target allows its detection by DNA/RNA hybridization



Reverse transcription polymerase chain reaction

- RT-PCR allows amplification of RNA through synthesis of complementary DNA (cDNA) = reverse transcription
- RT-PCR is used to detect RNA transcripts or virus with an RNA genome



Amplification of RNA lacking a poly(A) (miRNA) requires a preliminar step with a poly A polymerase

The Polymerase Chain Reaction (1984)

Three steps:

- Strand separation (95°C)
- Hybridization of primers (50-60°C)
- DNA synthesis (68-72°C)

Selected DNA sequence (ng) is amplified (µg) in a 2-h reaction (20-30 cycles) using a thermal cycler

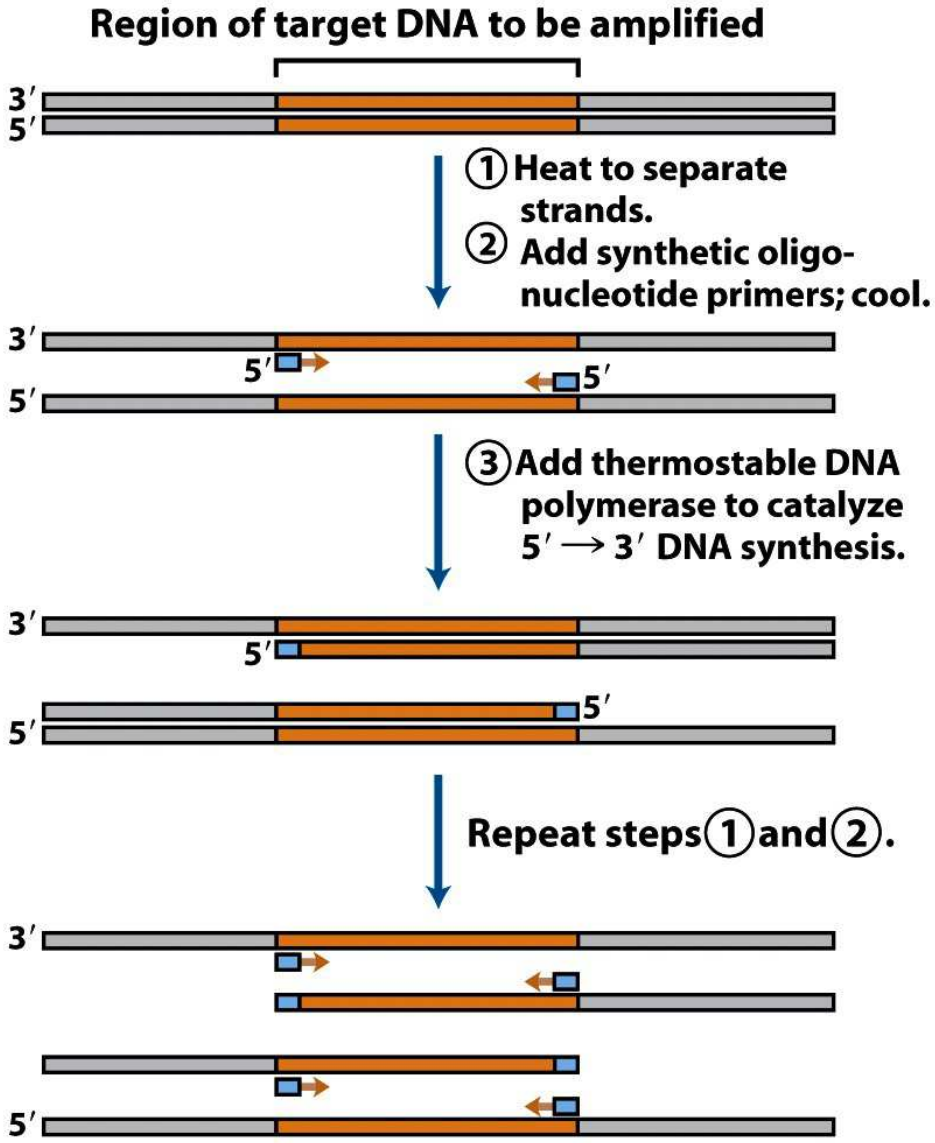


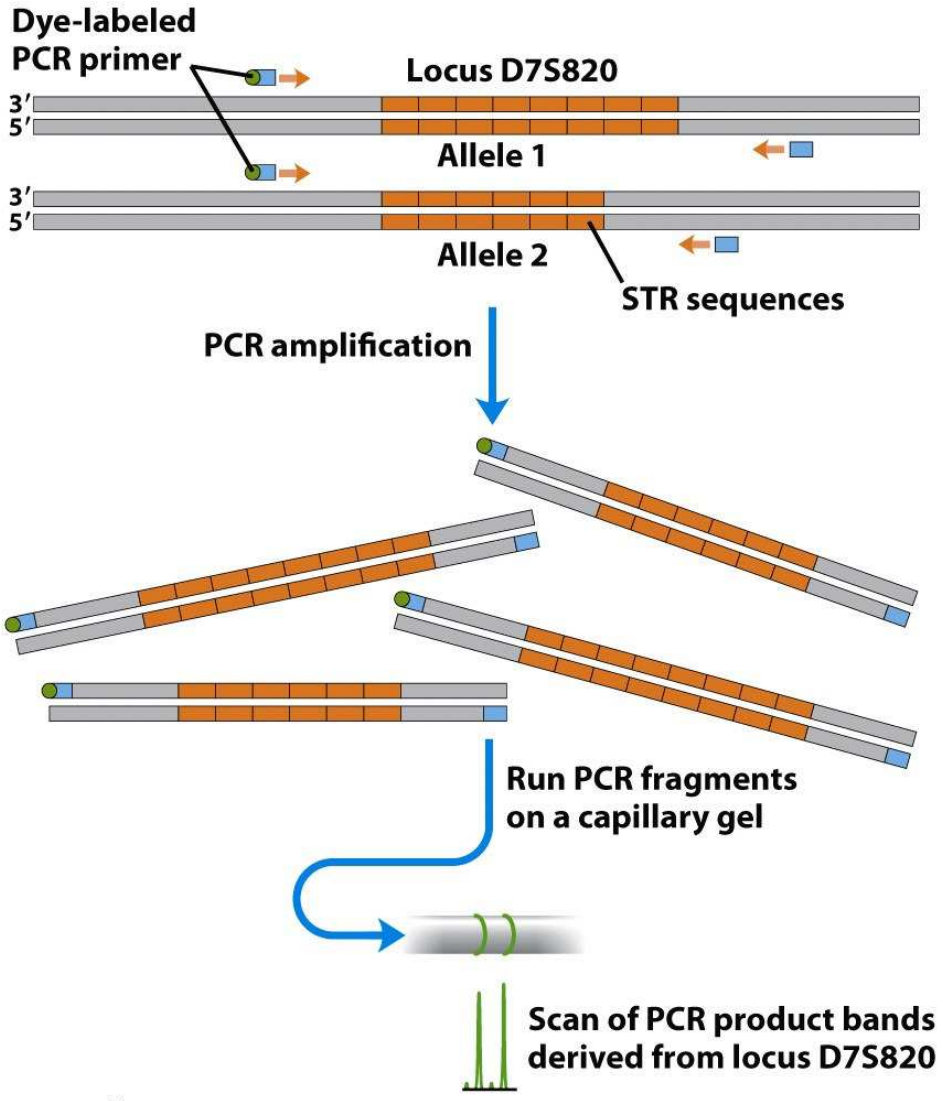
Figure 9-16a part 1
Lehninger Principles of Biochemistry, Fifth Edition

Application: PCR analysis of an Short Tandem Repeat (STR) locus

An STR (*microsatellite* in French) locus is a short DNA sequence (~4 nt), repeated many times in tandem at a particular location in a chromosome

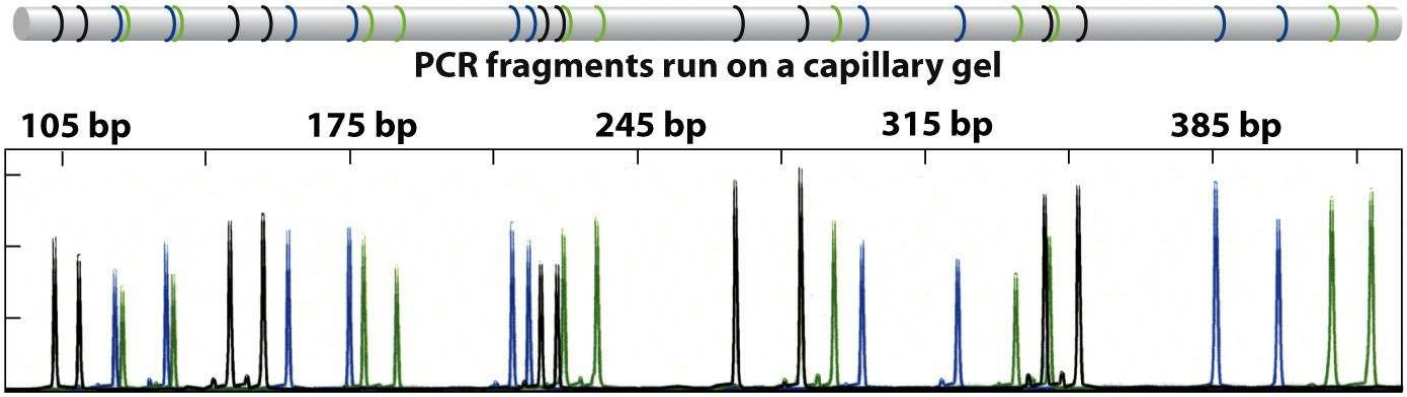
A STR locus from the US CODIS database may have as many as 82 different alleles

The presence of both alleles results in two fluorescence signals separated by electrophoresis on a capillary gel (same machine as the one used for DNA sequencing)



Multiplex-PCR

- Multiplex-PCR consists of multiple primer sets within a single PCR mixture
- The dyes linked to the PCR primers are of several different colors
- The PCR primers are designed to produce products in a size range a distinct as possible from that generated by the primers targeted to other loci
- Annealing temperatures for each of the primer sets must be optimized

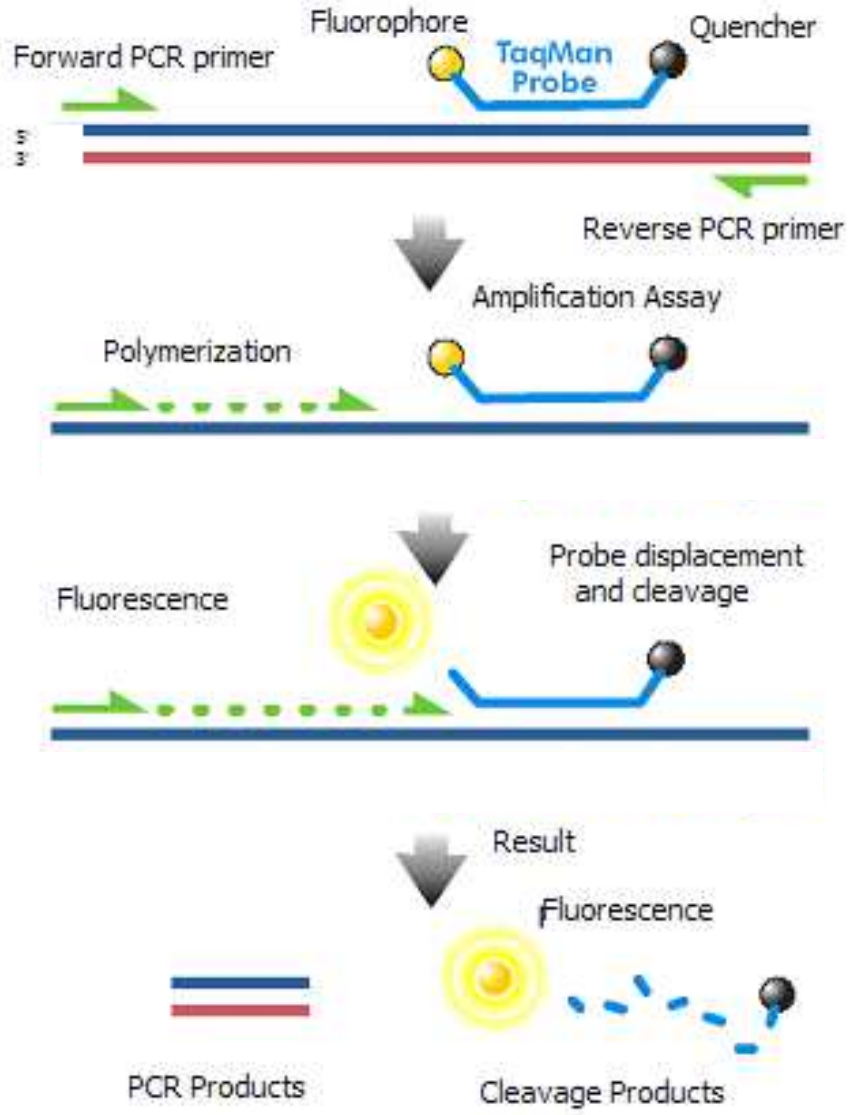


Pattern generated for 16 loci, using a commercial STR analysis kit

- Use of specific fluorescent oligonucleotide probes
  - ✓ Taqman, Molecular beacons and scorpions
- The fluorescent probes detect only the DNA being amplified
- The reporter probe significantly increases specificity, and enables quantification even in the presence of non-specific DNA priming
- Probes with different-colored labels can be used in multiplex (provided that all targeted genes are amplified with similar efficiency)

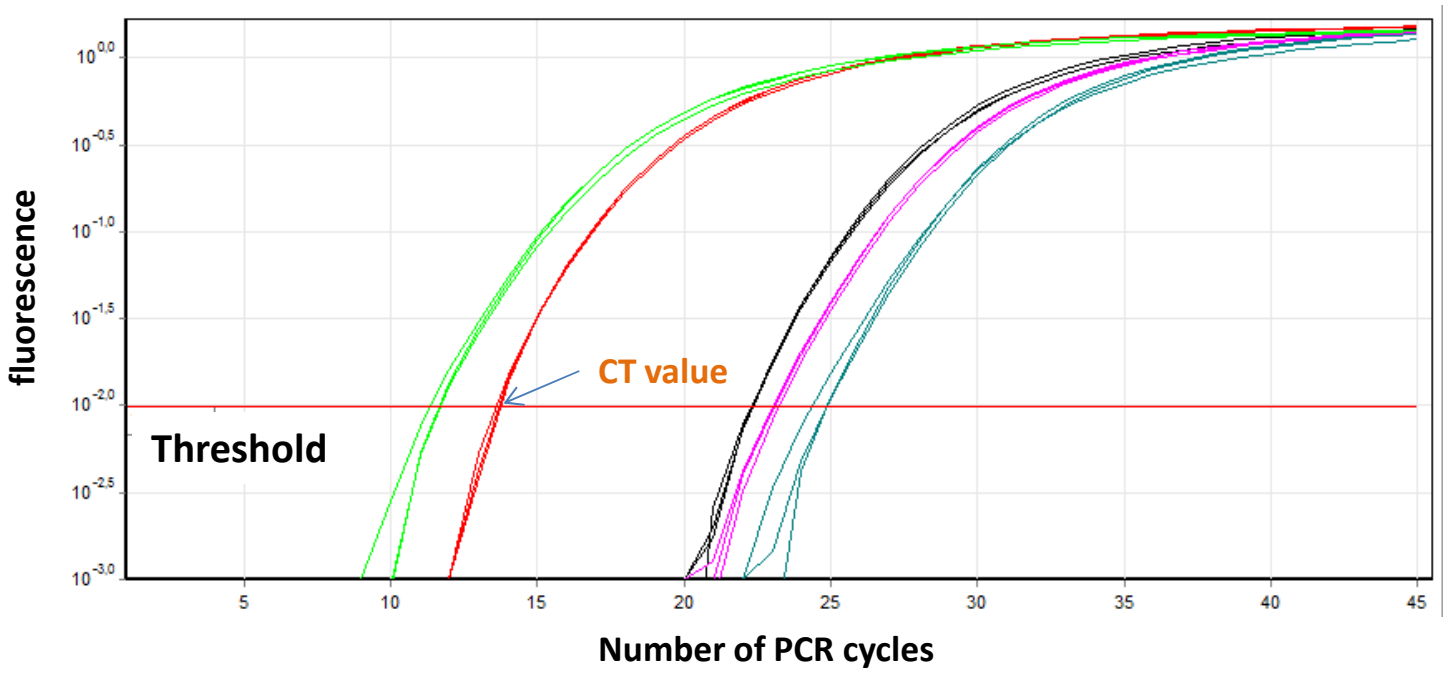
TaqMan probe

- Taqman probes contains a 5′ fluorescent probe and a 3′ quencher
- During the PCR, both primers and the probe hybridize
- The TAQ polymerase cleaves the fluorescent probe
- Fluorescent intensity increases



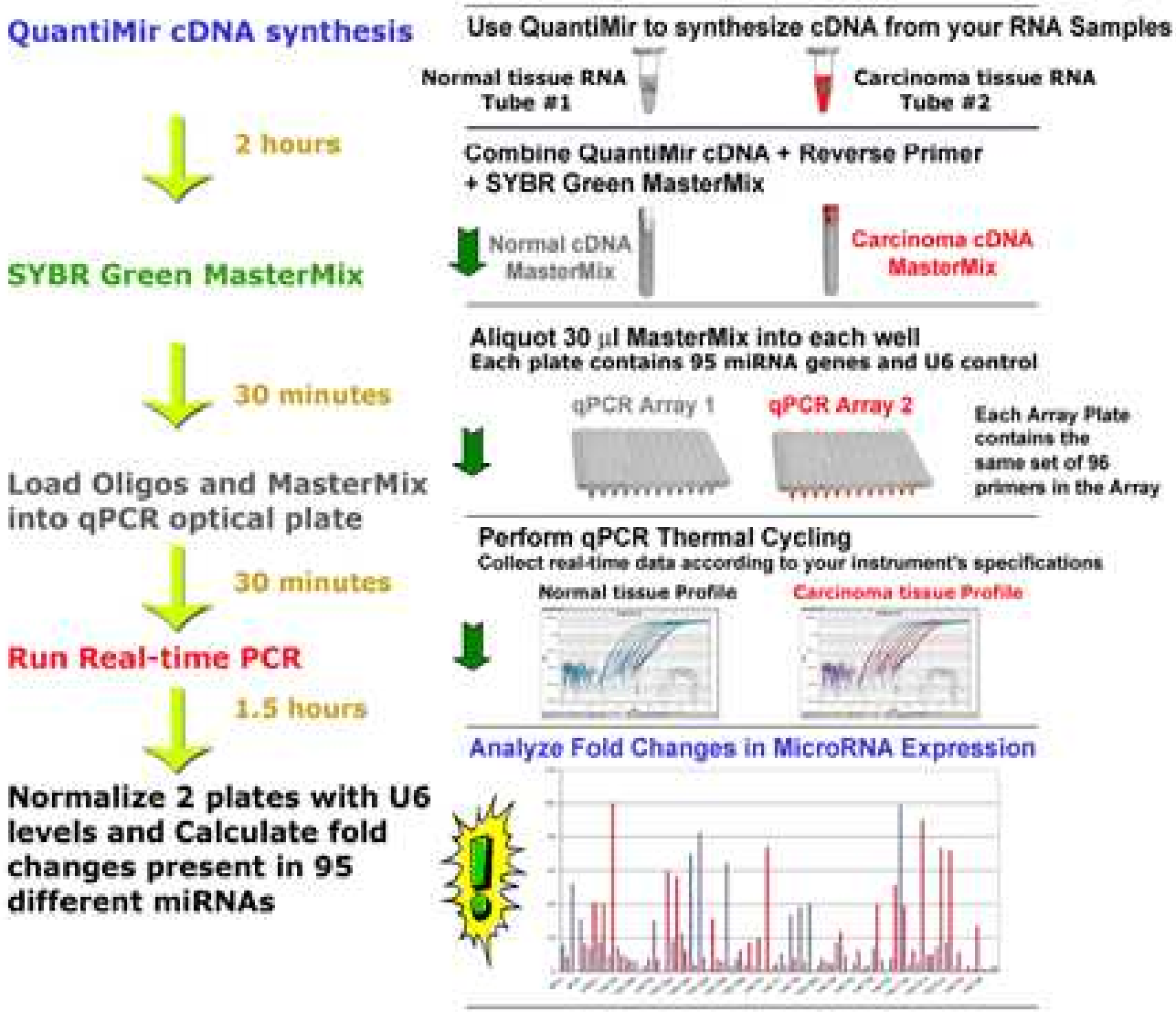
Real-time or quantitative polymerase chain reaction (qPCR)

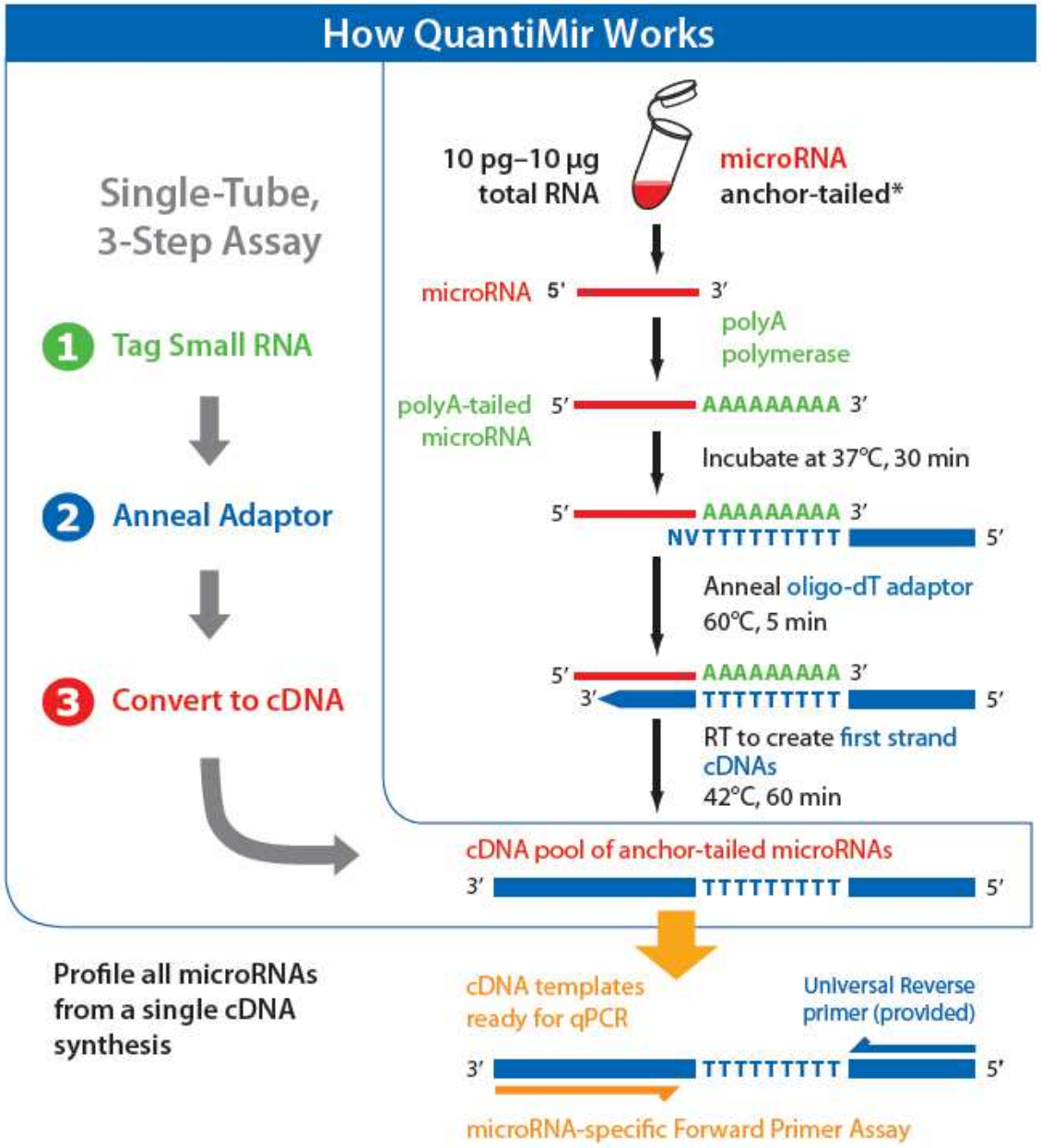
- Amplified DNA is detected as the reaction progresses
- Detection is based on (1) non-specific fluorescent dyes (SYBR Green) that intercalate with dsDNA or (2) specific oligonucleotides (probes)
- The Ct (cycle threshold) is the number of cycles required for the fluorescent signal to cross the threshold
- The quantity can be either an absolute number of copies or a relative amount when normalized to DNA input or additional normalizing genes



5 samples (different colors) with 3 replicates

qRT-PCR analysis of miRNA expression during carcinogenesis (Zhang et al., 2008)



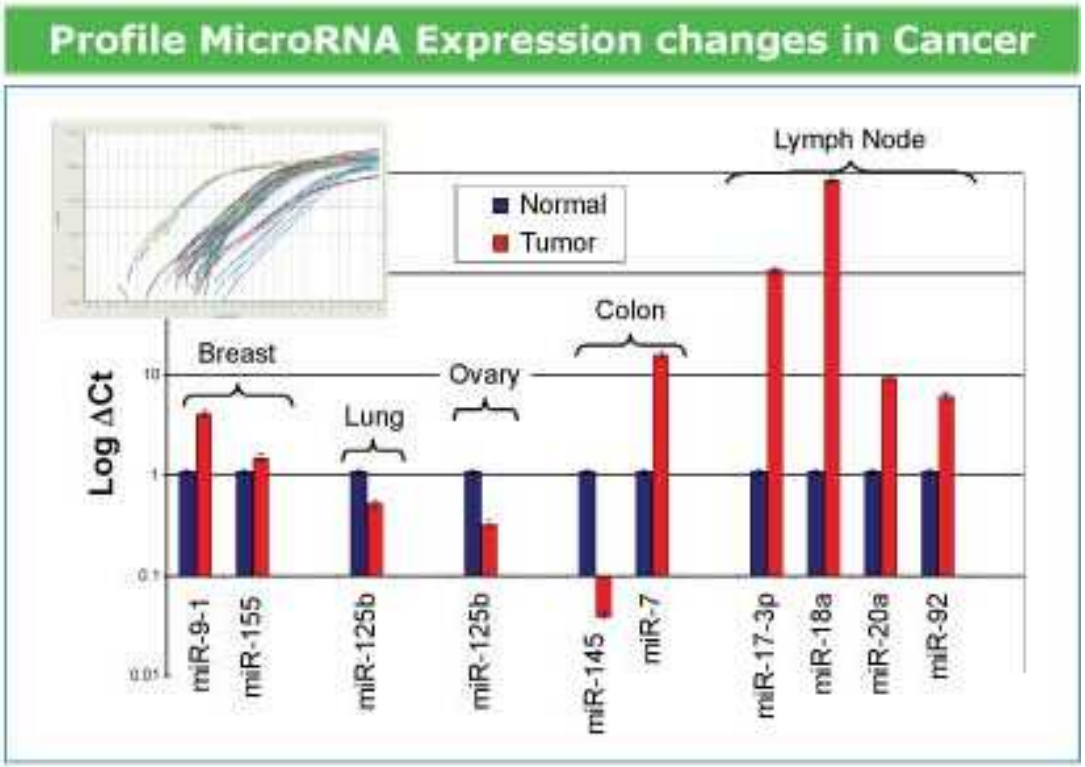


From System Biosciences

Quantification of qRT-PCR analysis

- Ct levels are inversely proportional to the amount of target nucleic acid in the sample
- **Δ Ct** corresponds to the difference between the sample and a control*
- **ΔΔCt** is the difference of the Δ Ct of the sequence of interest (SOI) and Δ Ct of the reference sequence (RS), a house keeping gene sequence usually (U6 snRNA).
- cDNAs are balanced to yield equal Ct values for the reference

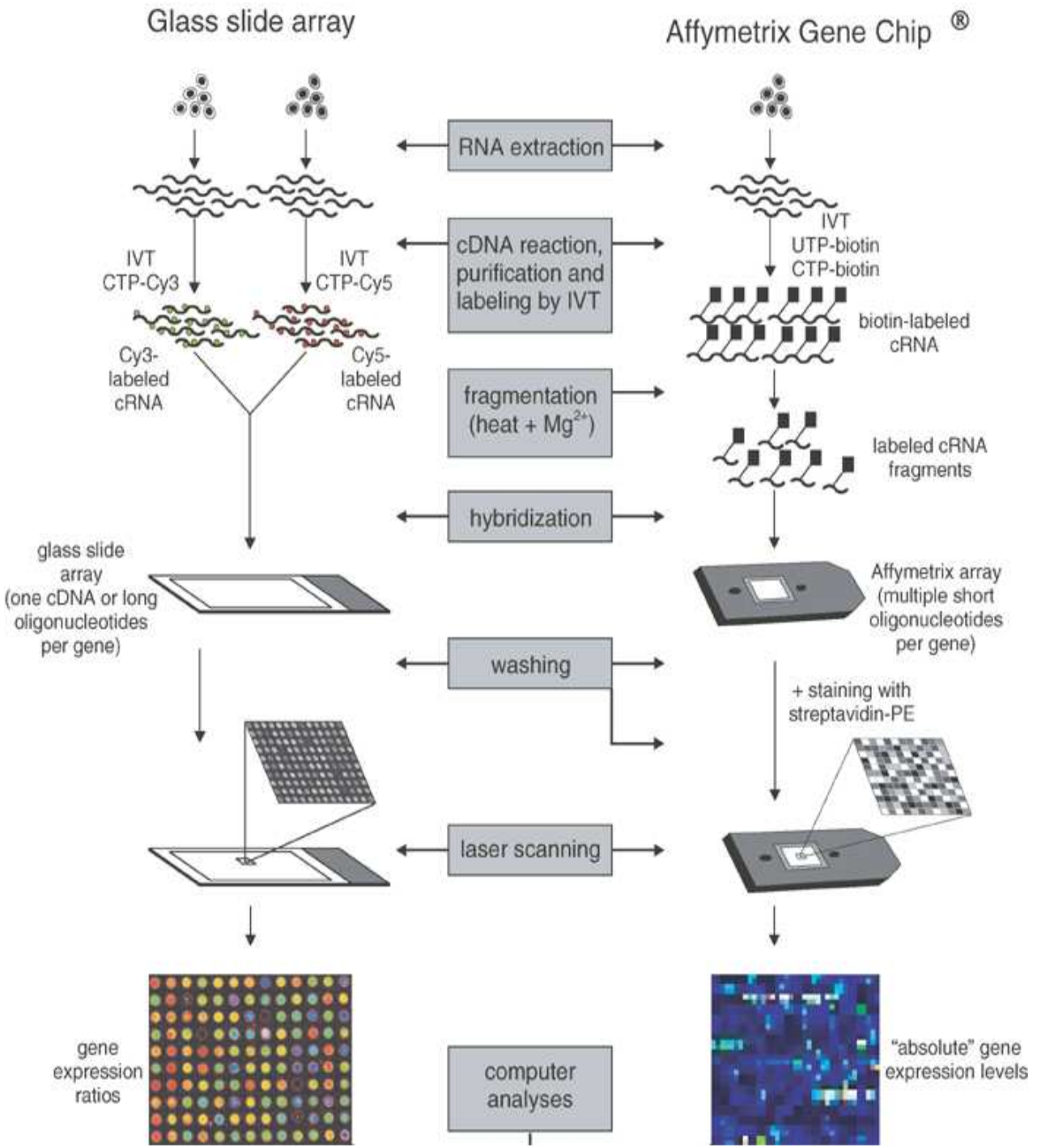
* The reverse relation is also used !!



Example of quantitative miRNA profiling of 9 miRNAs in 5 different Normal and Tumor-derived RNA samples. Several log-fold differences were measured.

Genome-wide analysis of mRNA expression

- RT is followed by *in vitro* transcription (IVT) to allow amplification and labelling of cRNA with (i) Cy3 (green) or Cy5 (red) or (ii) with biotin
- Hybridization to cDNAs or oligonucleotide per gene
- Affymetrix only: staining with streptavidin-pephycoerythrin conjugates

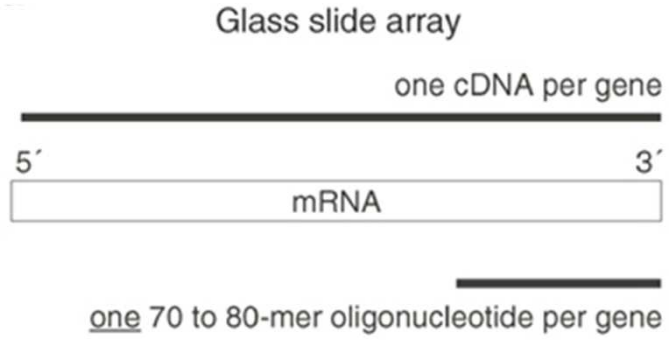


Staal *et al*., 2003

Glass slide experiments

- RNA is extracted from two cell populations, for instance diseased and normal
- cRNA is made with Cy3 (green) or Cy5 (red) labeled nucleotides.
- The two labeled cRNA samples are mixed and hybridized on a glass slide array, which is scanned with a laser, followed by computer analysis of the intensity image.

A glass array uses a single cDNA or long oligonucleotide per gene.

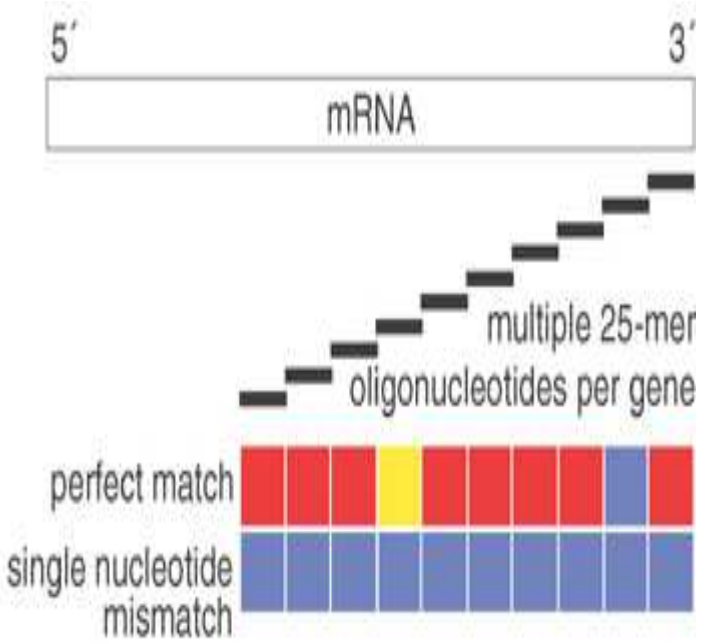


Affymetrix arrays (Genechip)

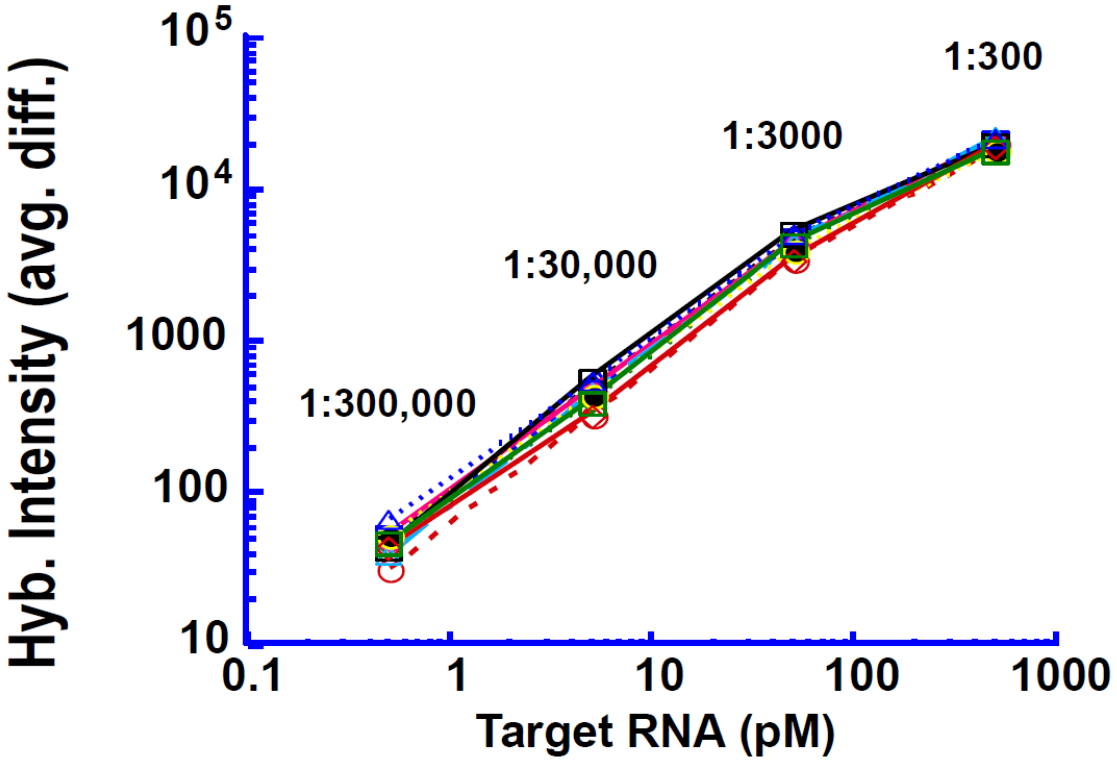
- Total RNA is extracted from one cell population
- Biotinylated cRNA is synthesized via coupled RT/IVT reaction
- After fragmentation, cRNA is hybridized to microarrays, washed and stained with phycoerythrin (PE)-conjugated streptavidin, and subsequently scanned on a laser scanner
- Affymetrix employs 11–20 different and sometimes overlapping 25-mer oligonucleotides per gene

- In addition to perfect-match (PM) oligonucleotides, Affymetrix GeneChips also contain mismatch (MM) oligonucleotides that serve as negative controls. The mismatch oligonucleotides carry a mutation at position 13 of the 25 mers.
- Intensity images of perfect match (upper row of squares) and mismatch (lower row) are indicated. In cases where both perfect match and mismatch yield a strong signal, the contribution of that probe to the overall expression is ignored, because it is not specific.

The absolute expression value is calculated from the combined PM-MM differences of all the pairs in the probe set.
The values can be directly compared to data for any other sample using the same probe sets.



- Detection levels is very low (~1-3 copies per cell)
- The assay is sensitive over 3 log ranges



Log-log plot of the hybridization intensity (average of the PM-MM intensity differences for each gene) versus concentration for cytokine RNA.

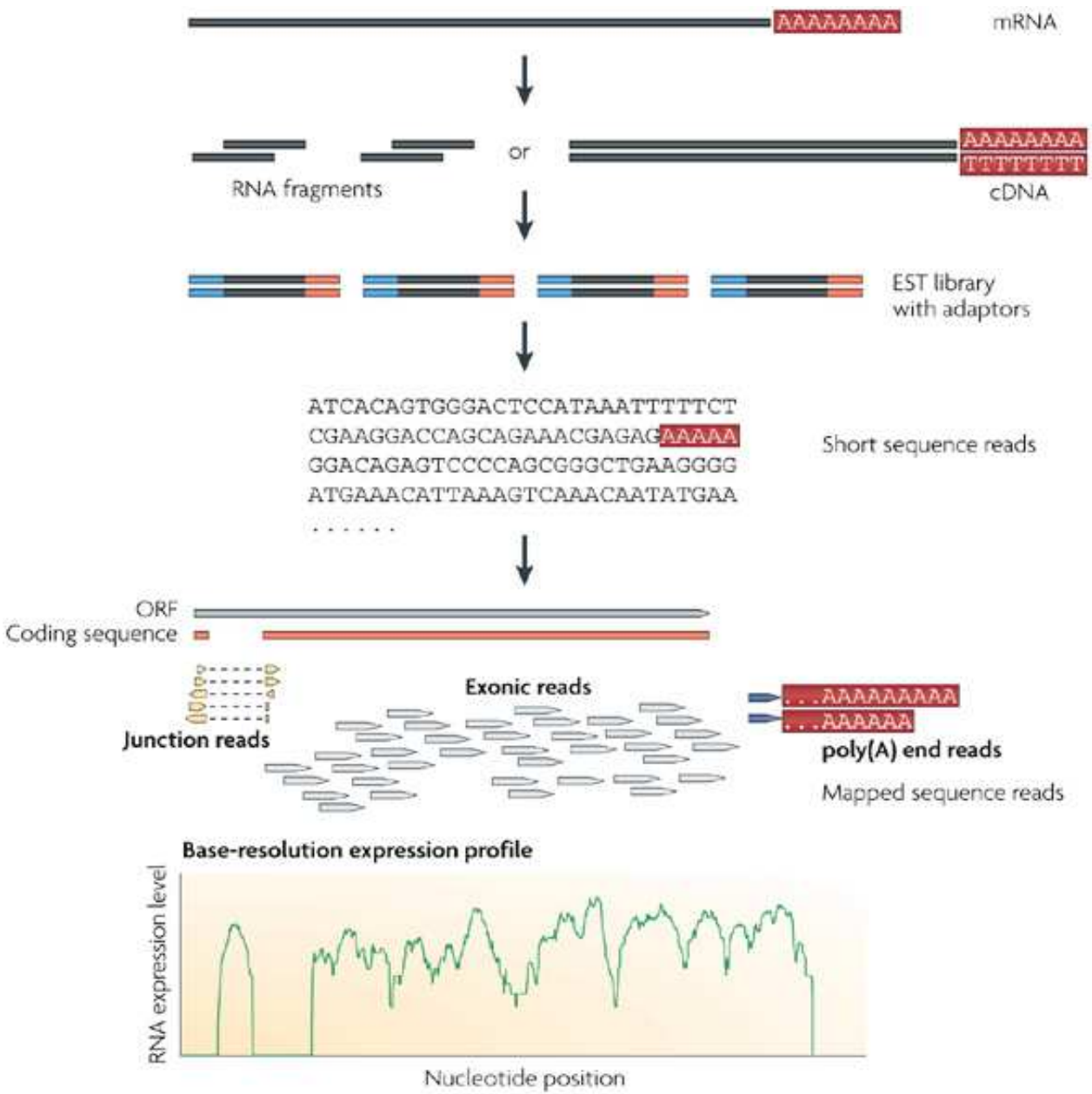
Hybridization-based approaches limitations

- Rely upon existing knowledge about genome sequence (20% of total SNPs known)
- High background level due to cross hybridization
- Limited dynamic range of detection due to signal saturation
- Normalization methods to compare different experiments

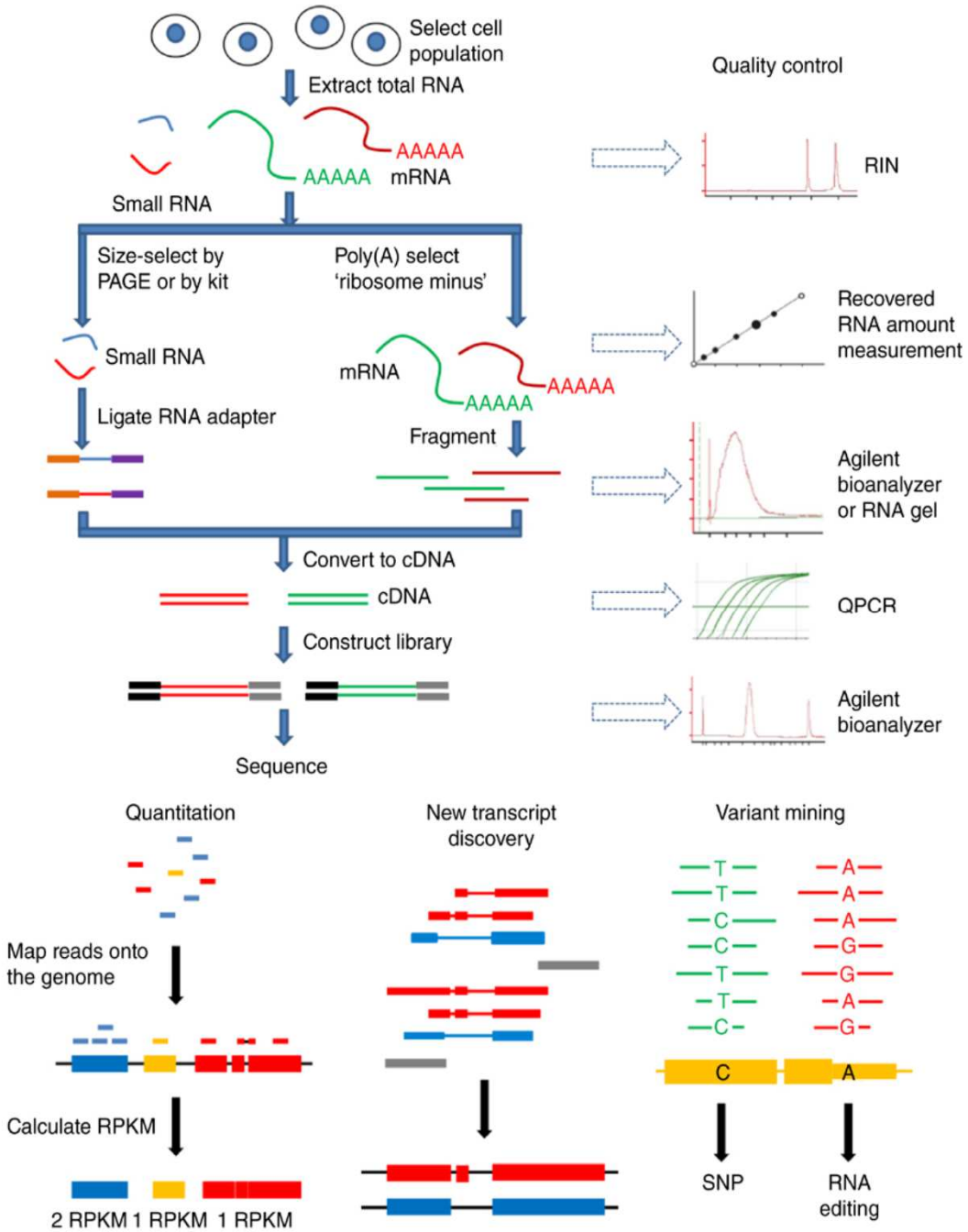
RNA-seq (Whole transcriptome shotgun sequencing WTSS)

- Technology that uses the capabilities of next-generation sequencing (NGS) to reveal RNA presence
- Allows RNA quantification from a genome at a given moment in time
- All transcripts (mRNA and ncRNA) can be analysed
- Detection of fusions and other structural variations (alternative intron splicing)
- High reproducibility and dynamic range (5 log ranges)
- Not limited to genome annotation
- Usually starts with a cDNA library. As RT introduces biases and artifacts single molecule real time (SMRT) sequencing methods have been developped

Overview



Nature Reviews | Genetics



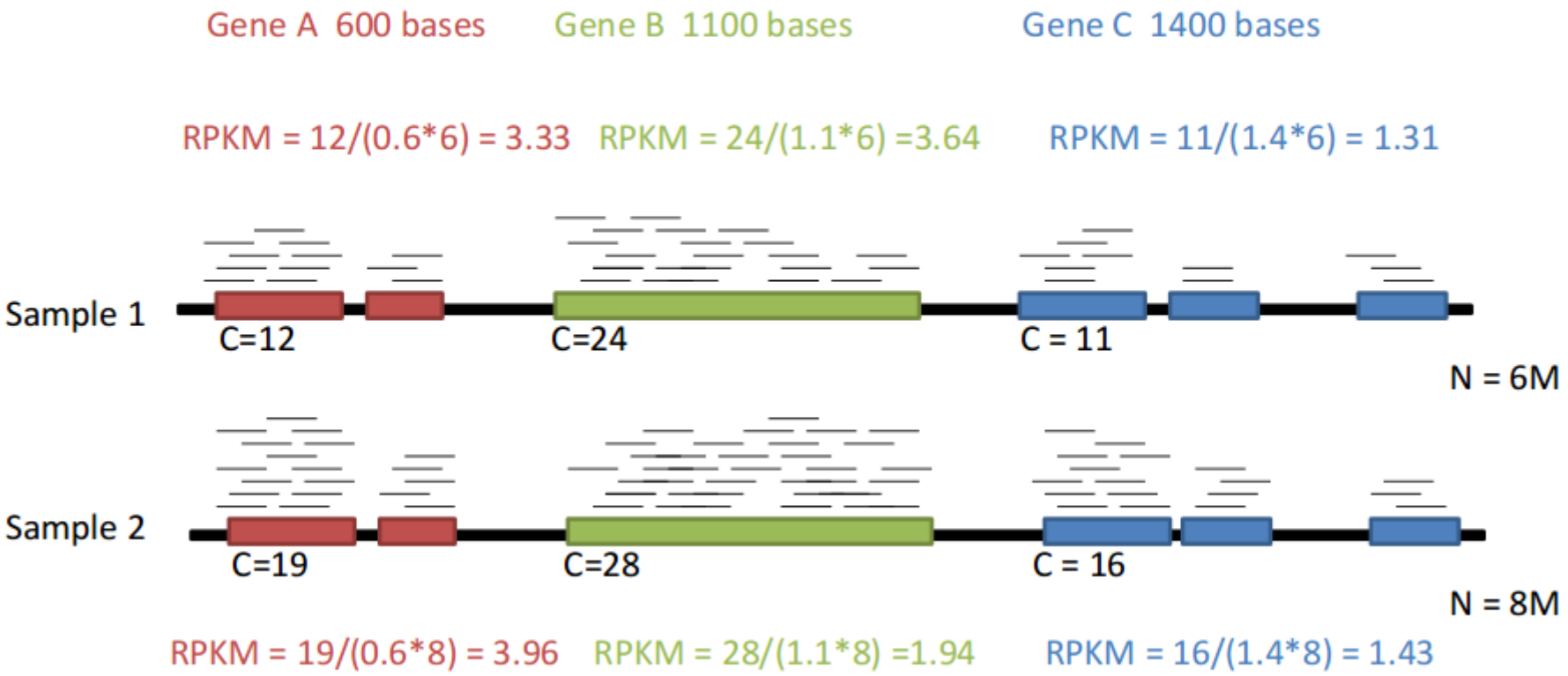
RPKM-values "Reads per kilo base per million mapped reads"

- Amount of reads sequenced per sample = variable
- Statistics needed to compare expression values between different samples / experiments
- Corrects for difference in sequencing depth and gene length

$$RPKM = (10^9 * C)/(N * L)$$

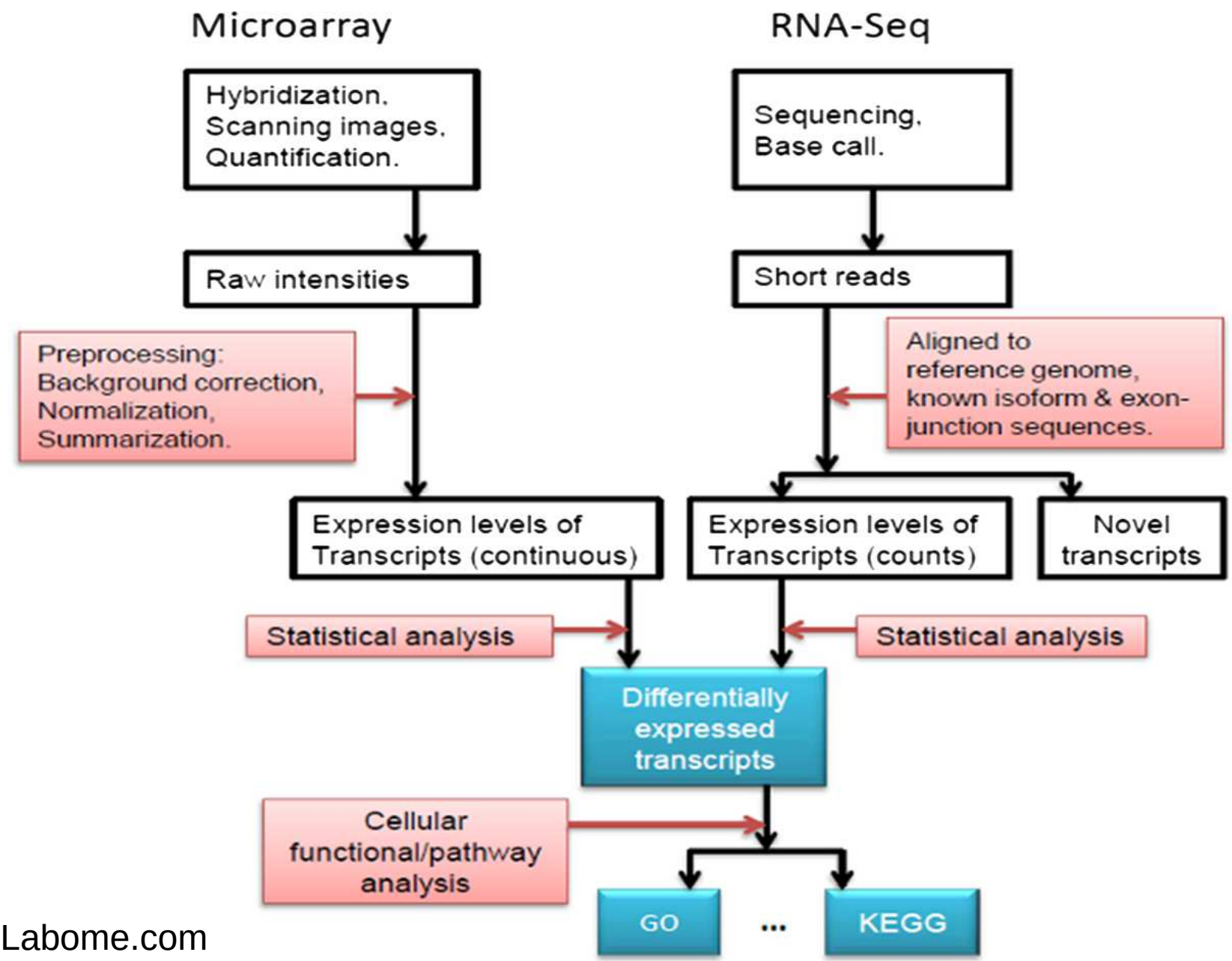
C = Number of reads mapped to a gene
N = Total mapped reads in the experiment
L = exon length in base-pairs for a gene

RPKM Example



IZABELCAVASSIM

Microarray versus RNA-seq



Next-Generation Sequencing technologies

- Automated dideoxy method of Sanger
- Next-generation sequencing technologies
  - Template preparation (PCR amplification of single DNA versus cloned DNA in FGS)
  - Greater sequencing throughput
  - Imaging
- Emphasis on Illumina and IonTorent
  - Need PCR/bridge amplification

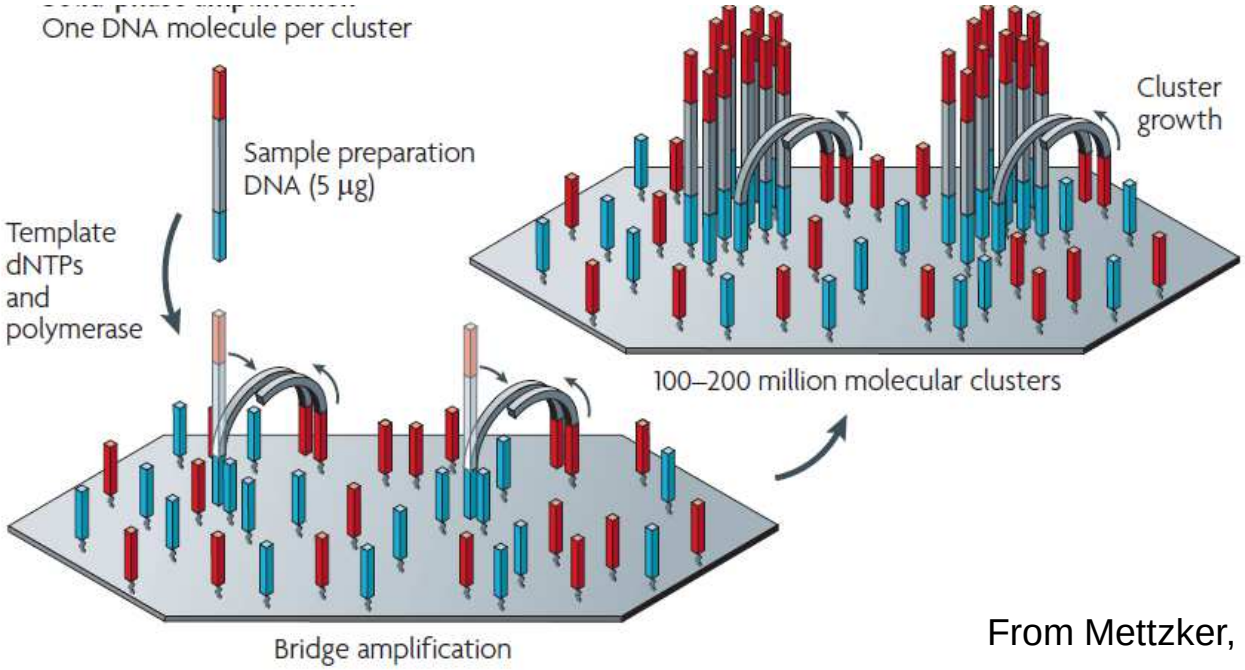
Application of NGS

- Whole-genome assembly (metagenomics)
- Variant detection (SNP, indels, copy number)
- Targeting resequencing (exons)
- ChIP-seq (DNA binding proteins, histone modification)
- Expression profiling (RNA-seq splicing variants)
- Small RNA sequencing

Illumina

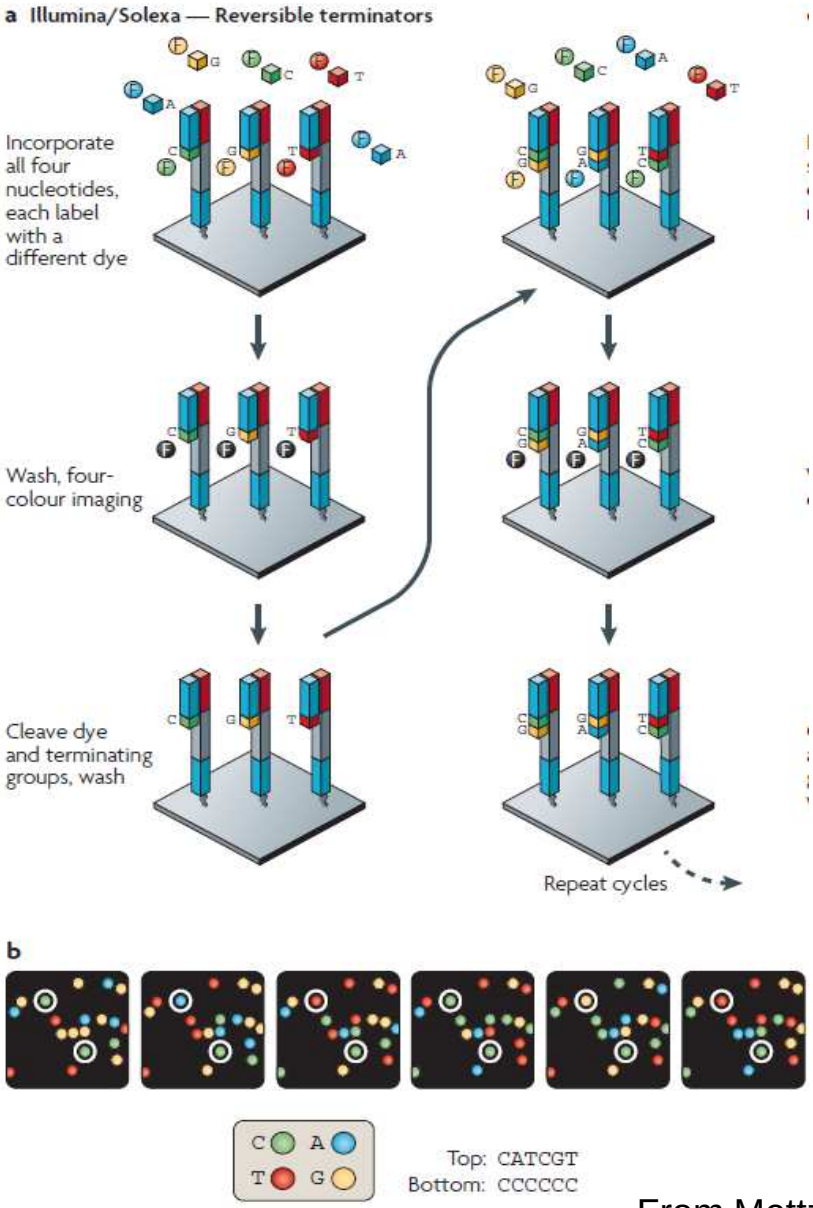
Template preparation

- Genomic DNA is randomly fragmented and adapters are ligated to both ends
- Solid-phase amplification of the template
  a. Bind single-stranded (ss) fragments to the surface of the flow cell channels containing lawns of primers
  b. Initial priming and extending of the single stranded molecule template
  c. Bridge amplification of the immobilized template with immediately adjacent primers to form clusters



From Mettzker, 2010

Cyclic reversible termination



To initiate the first sequencing cycle, add all 4 labeled reversible terminators and DNA polymerase to the flow cell

After laser excitation, capture the image of emitted fluorescence (TIRF) from each cluster on the flow cell

Record the identity of the first base of each cluster

The blocked 3′ terminators and the fluorophore are removed

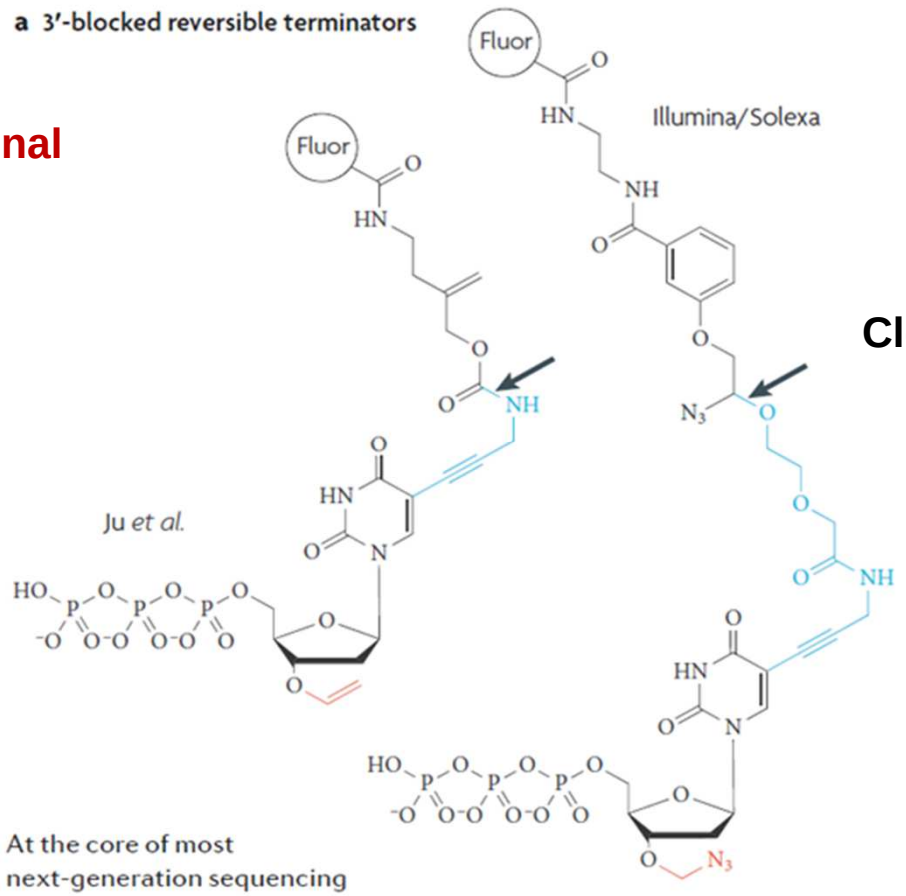
The identity of each base of a cluster is read off from sequential images

From Mettzker, 2010

Modified nucleotides used in next-generation sequencing methods

From Mettzker, 2010

Terminating functional groups



Cleavage site

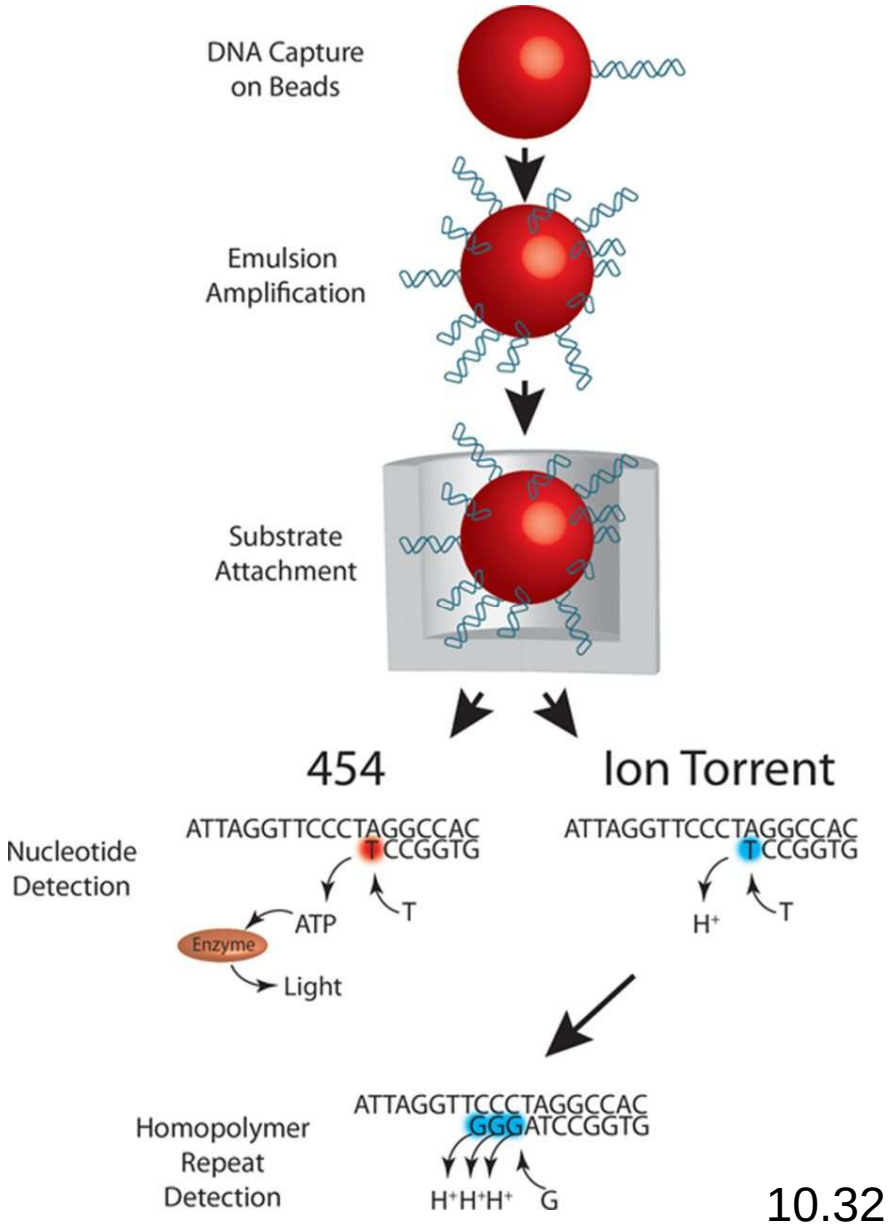
Residual linker structures

Incorporation requires mutated DNA polymerase !

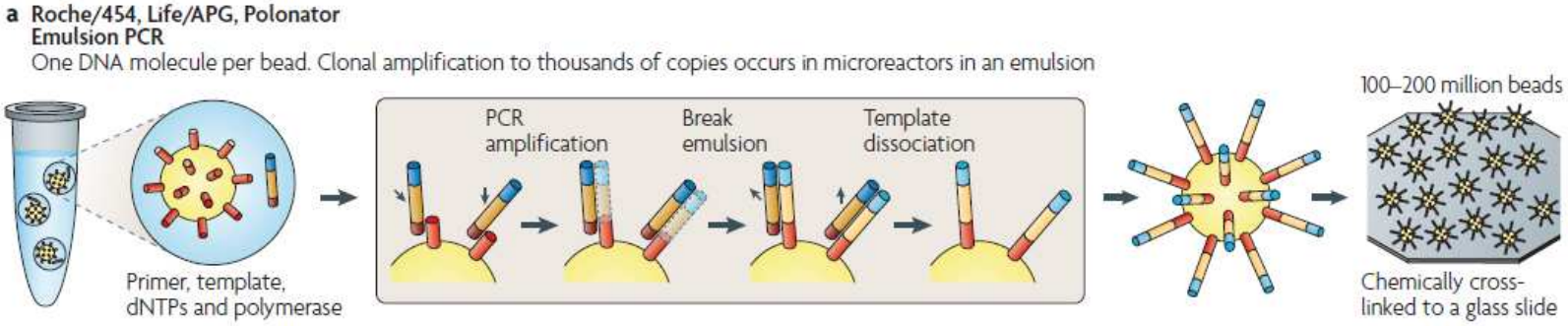
Ion Torrent between 2nd and 3rd generation sequencing

- Clonal amplification occurs on beads in an emulsion by emPCR
- The beads are deposited into picoTiterPlate (PTP) wells
- DNA sequencing using sequential addition of each dNTP in limited amounts
- Extension is measured with a semiconductor apparatus, via the release of H^+

Analogous to pyrosquencing method Roche/454) = the PPi release produces light via a series of enzymatic reaction (luciferase/sulphurylase)
Bioluminescence is imaged with a CCD camera



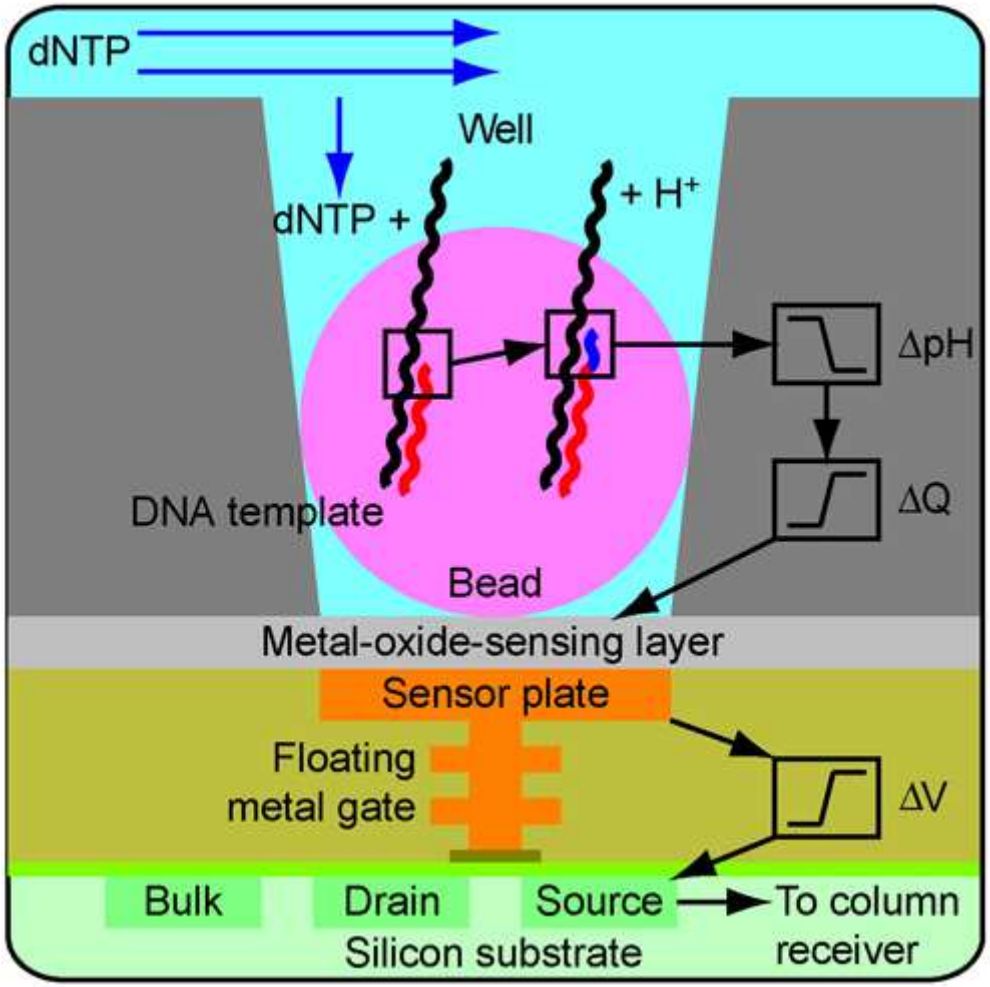
Template preparation



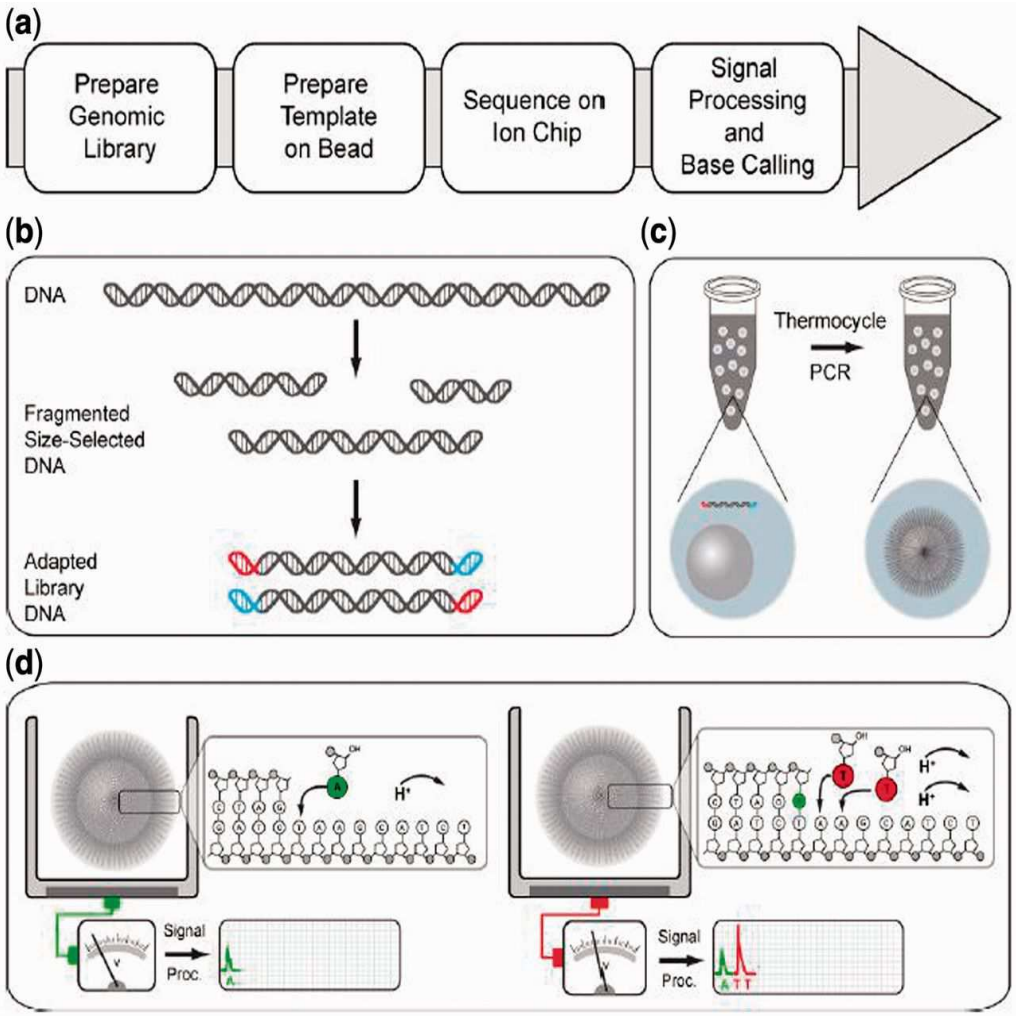
From Mettzker, 2010

- An oil-aqueous emulsion is created to encapsulate bead-DNA complexes into single aqueous droplets
- PCR amplification occurs is performed at the surface of the beads via immobilised primers

Details of the CMOS semiconductor chip

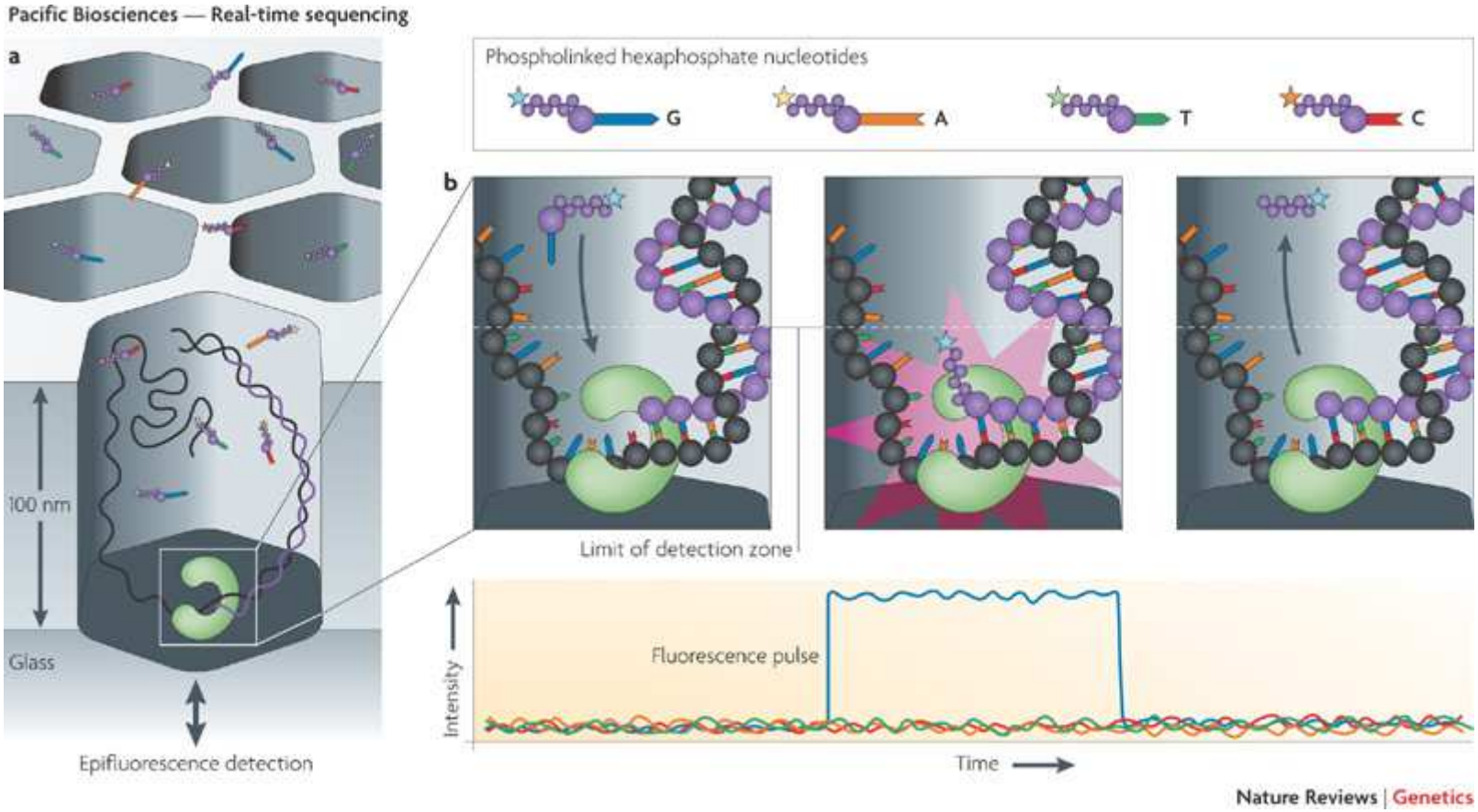


- Single addition of nucleotide in limiting addition
- DNA polymerase extends the primer and pauses
- DNA synthesis is reinitiated following the addition of the next complementary dNTP
- The order and intensity of the intensity peaks are recorded as flowgrams

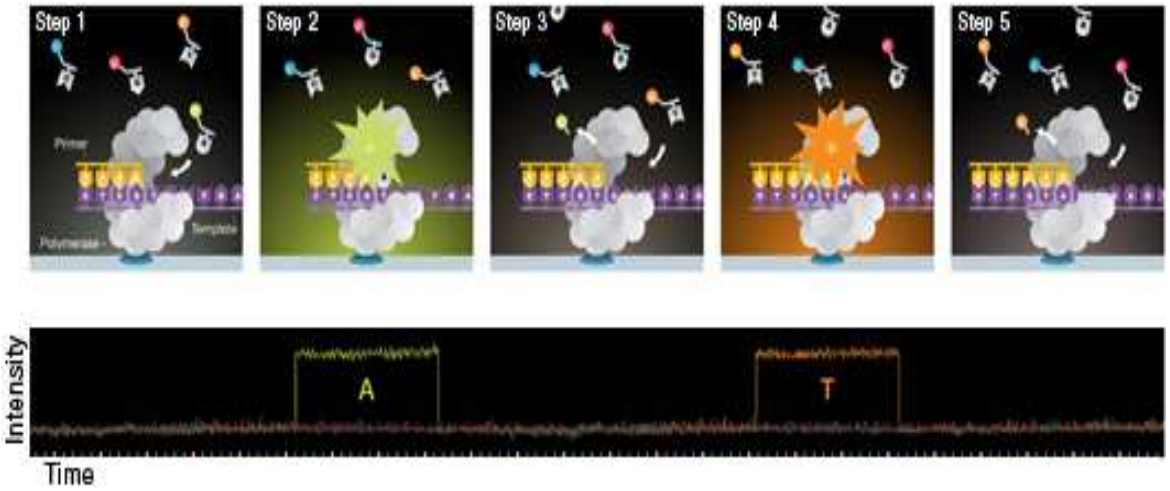


Third generation sequencing : single molecule real time sequencing

- DNA synthesis based sequencing

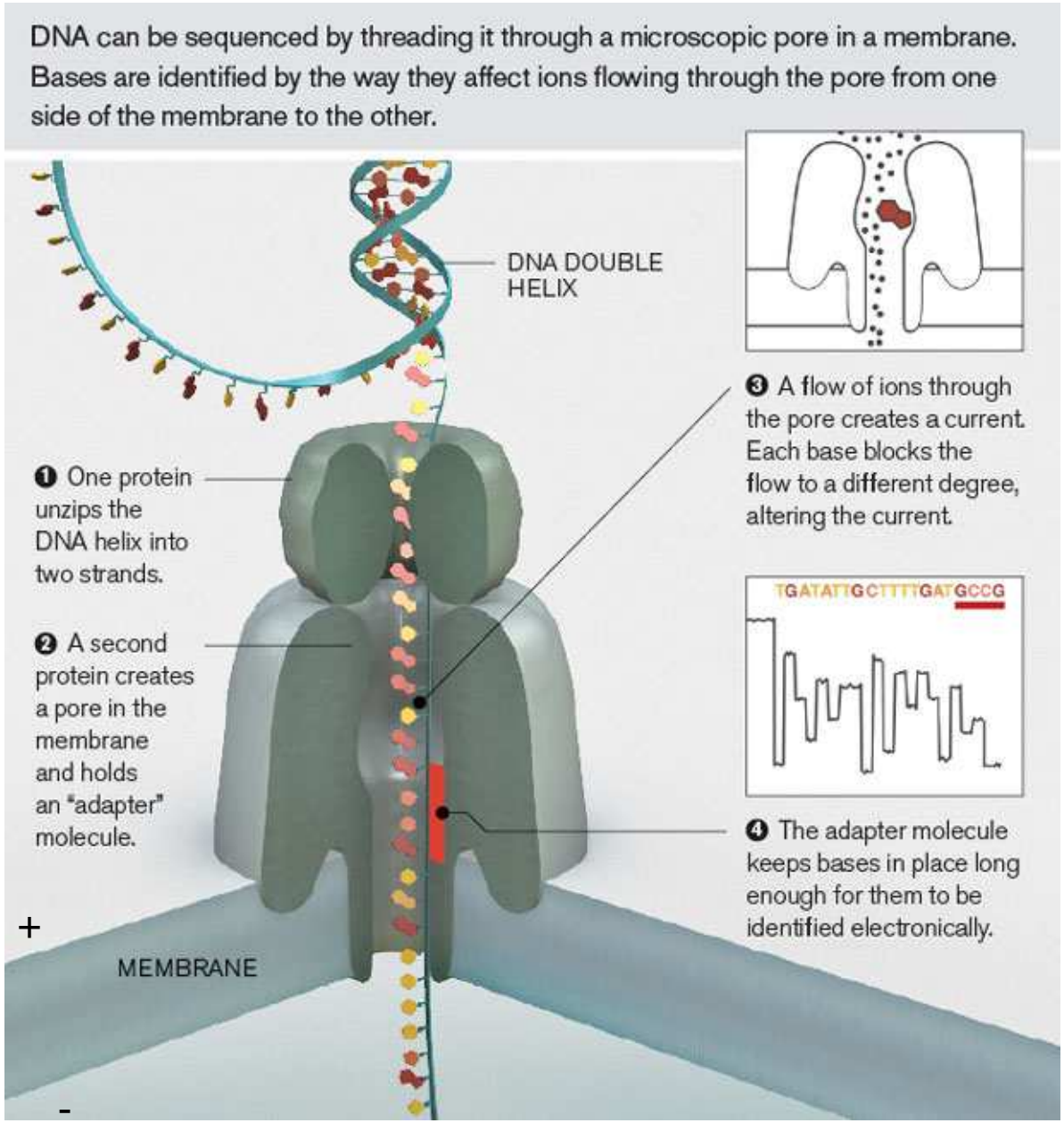


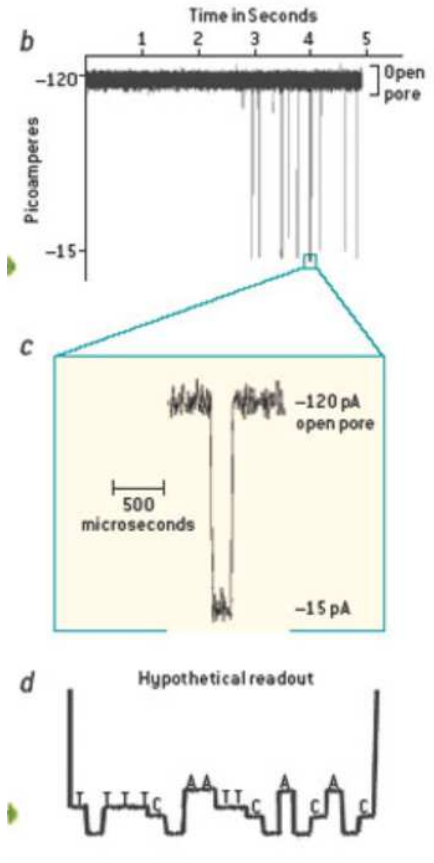
- One SMRT cells contains millions nanowells (70 nm) on alluminium film
- One DNA polymerase per nanowell attached to alluminium film (Zero-Mode Waweguides)
- Fluorescence emission of built-in nucleotides measures



- A single DNA polymerase is affixed at the bottom of a ZMW (zero-mode waveguide on a chip
- ZMW is a structure that creates an illuminated observation volume that is small enough to observe only a single nucleotide
- Each of the four DNA bases is attached to one of 4 different fluorescent dyes
- When a nucleotide is incorporated by the DNA polymerase the fluorescent tag is cleaved off and diffuses out the observation area of the ZMW where its fluorescence is no longer observable
- A detector detects the fluorescence signal

- NanoPore sequencing technologies





www2.technologyreview.com

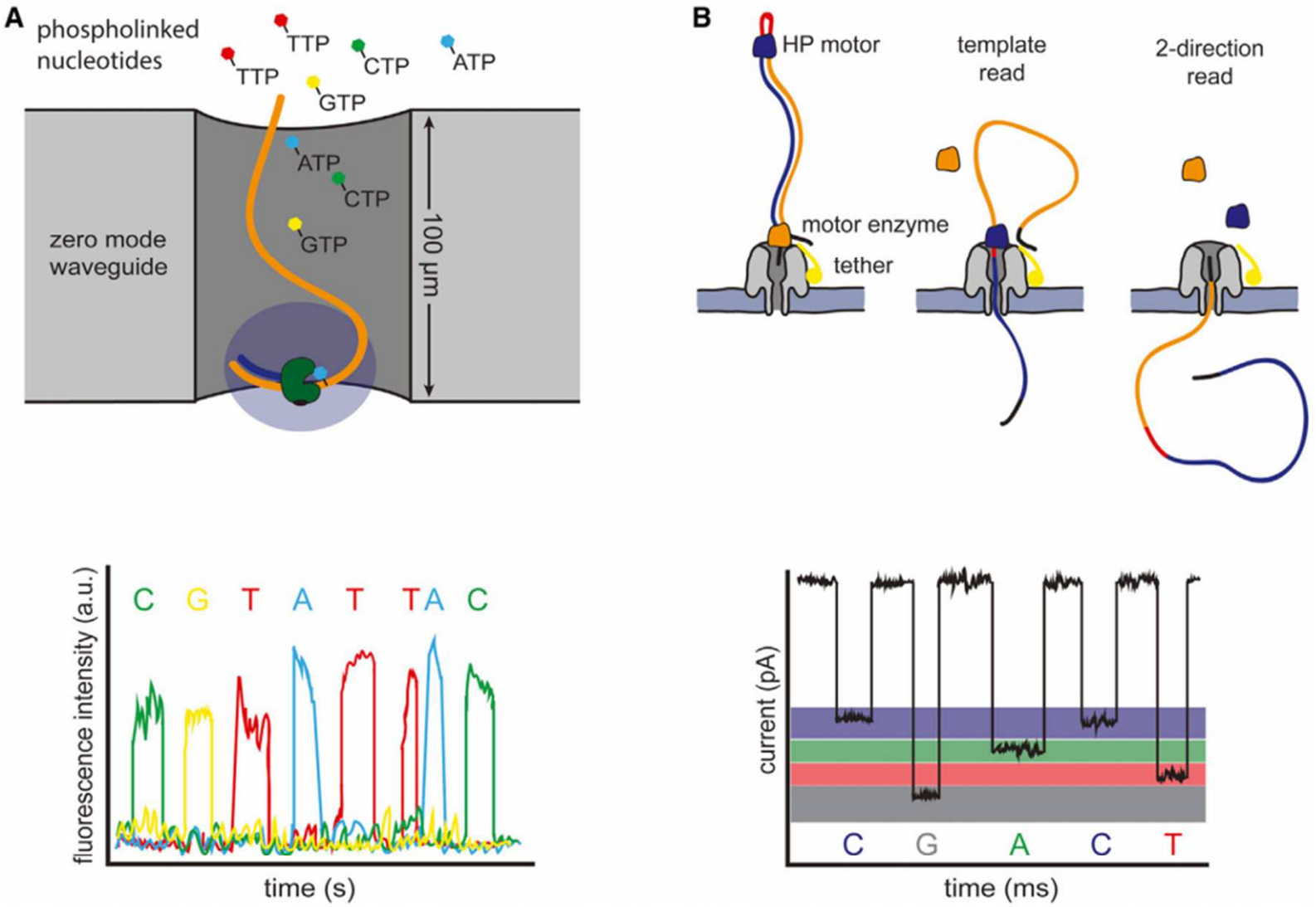


Figure 3. Single Molecule Sequencing Platforms

(A) Pacific Bioscience's SMRT sequencing. A single polymerase is positioned at the bottom of a ZMW. Phosphate-labeled versions of all four nucleotides are present, allowing continuous polymerization of a DNA template. Base incorporation increases the residence time of the nucleotide in the ZMW, resulting in a detectable fluorescent signal that is captured in a video.

(B) Oxford Nanopore's sequencing strategy. DNA templates are ligated with two adapters. The first adaptor is bound with a motor enzyme as well as a tether, whereas the second adaptor is a hairpin oligo that is bound by the HP motor protein. Changes in current that are induced as the nucleotides pass through the pore are used to discriminate bases. The library design allows sequencing of both strands of DNA from a single molecule (two-direction reads).

Comparison of NGS technologies

Character-istic	454 pyro-sequencing	Illumina HiSeq2500	Illumina Miseq	Ion Torrent Proton	PacBio	NanoPore
Read length	1000 bp	2 x 125bp	2 x 300 bp	200 bp	> 10 kb	> 50 kb
Per run / lane	700 Mb	1000 Gb	15 Gb	10 Gb	1 Gb	variable
Run Time	1 day	10 days	2 days	4 hr	2 hr	variable
Number of reads	1 million	2 billion	25 million	80 million	100,000	variable
Error	Homo-polymer	substitutions	substitutions	Homo-polymer	High error rates	High error rates
	Second generation				Third generation	

Computer requirements-data storage

- One RNA-seq sample = ~10 Gb
- Comparing genomic DNA of tumor sample versus normal sample
  - ✓ 30x coverage of both samples: 500 Gb
  - ✓ When testing 20 tumors: 10 Tb raw data
  - ✓ ⇒ storage ! + CPU -speed

Courtesy of SCK-CEN