Lecture 1 – concepts

•Comparative genome analysis – size and gene numbers in different organisms

•Comparative gene organization in prokaryotes *versus* eukaryotes

•The Human genome: classification of coding and noncoding sequences

•Origins and different types of repetitive sequences

•Gene families – Functional classification of genes

•Possible mechanisms producing genetic variants and new genes

Accessing to genome sequence

Revision: classical methods to sequence DNA

New alternative methodologies

Large-scale sequencing: Next Generation Sequencing

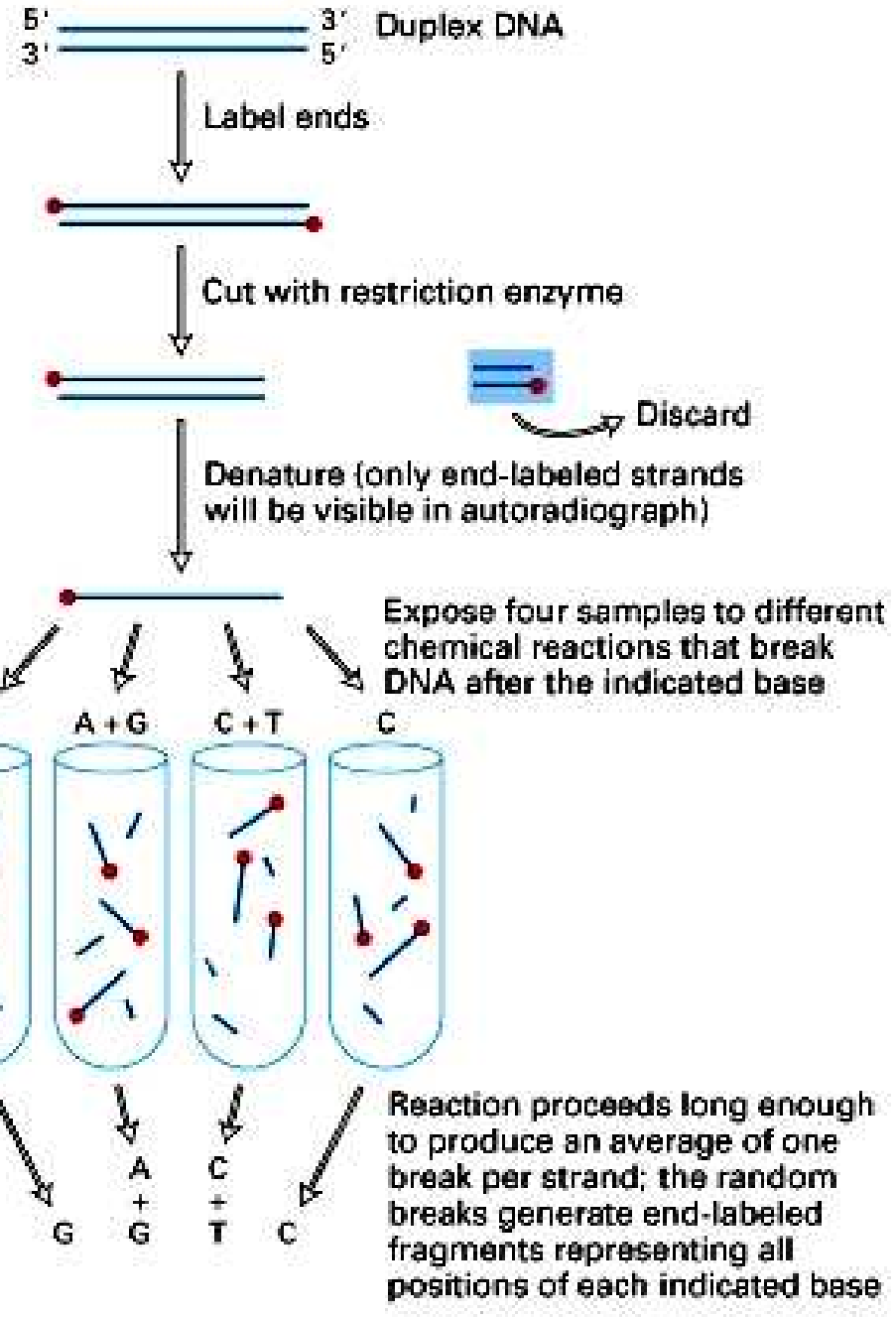
# DNA Sequencing by the Maxam-Gilbert chemical method

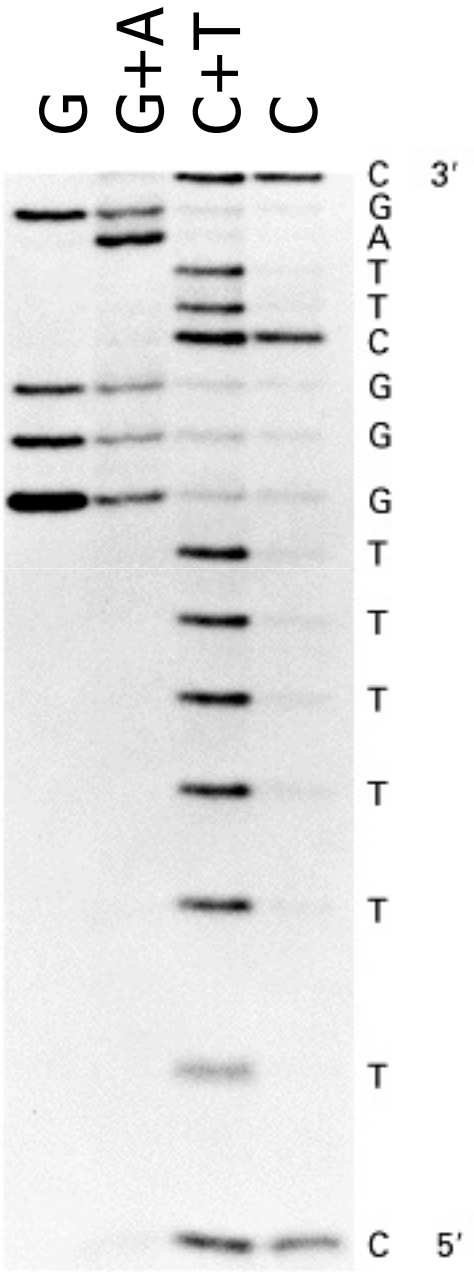
For chemical sequencing, the DNA should be a linear fragment, labeled only at one end

-Restriction fragment of cloned DNA
-PCR product

-Polynucleotide kinase (5')
-Fill-in

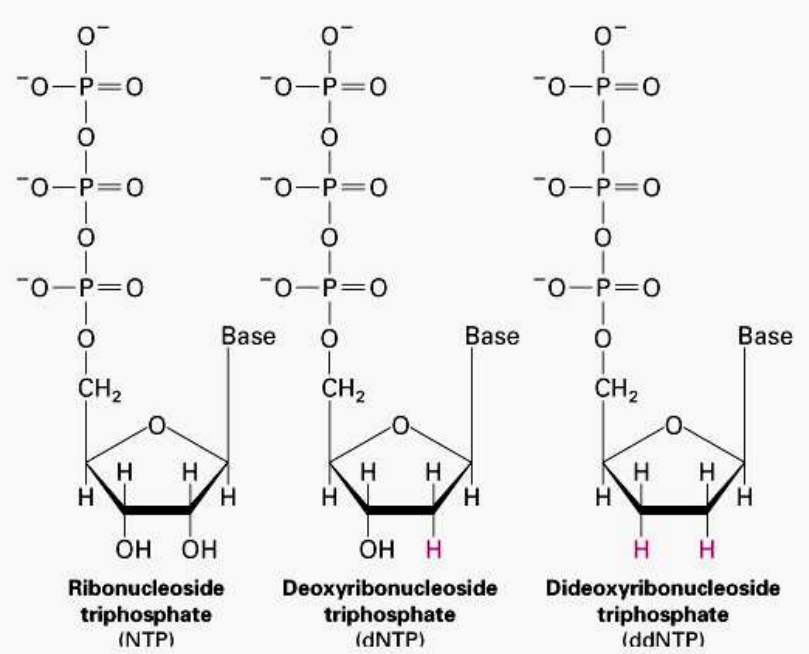
# DNA Sequencing by the Maxam-Gilbert chemical method



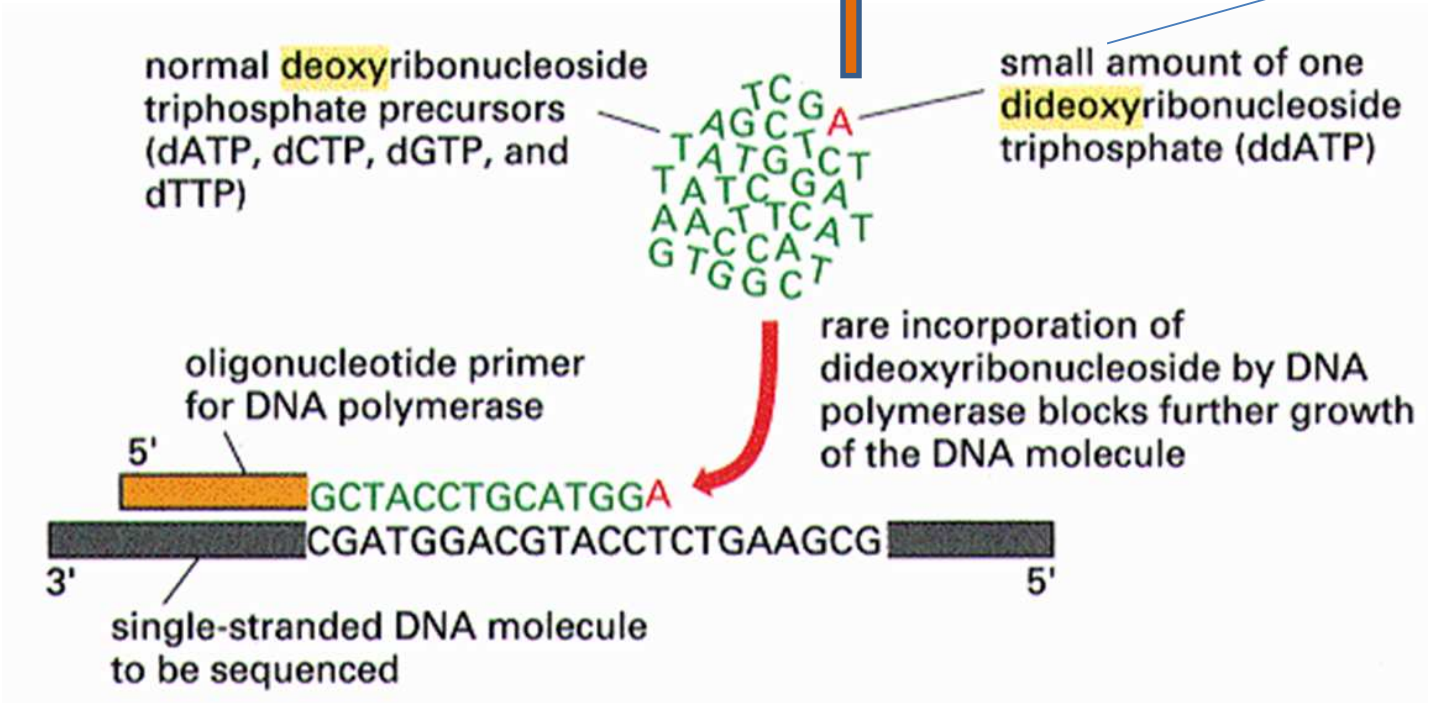



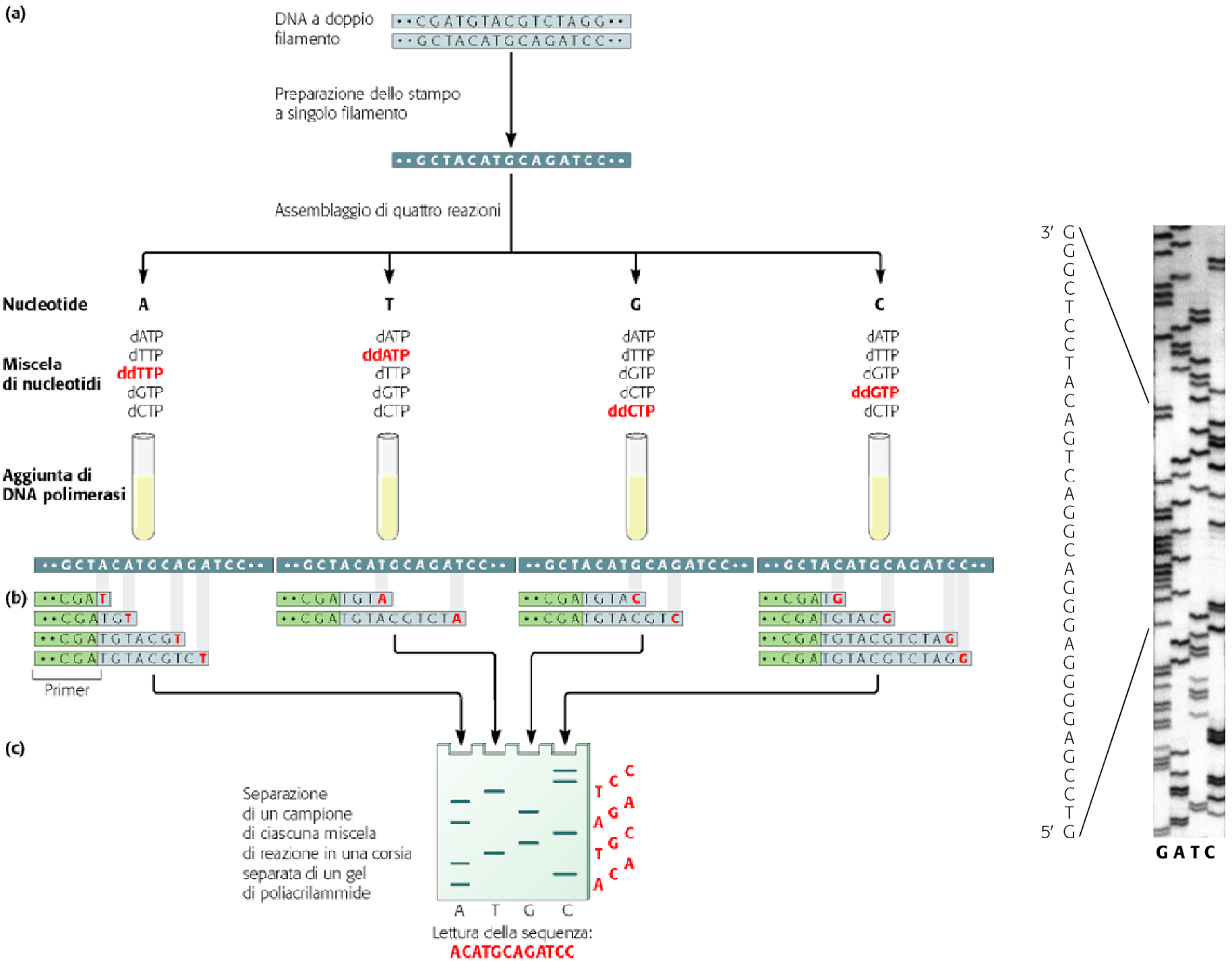

Parallel gel electrophoresis and autoradiography

## Dideoxy-sequencing: the **Sanger** "chain-terminator" method

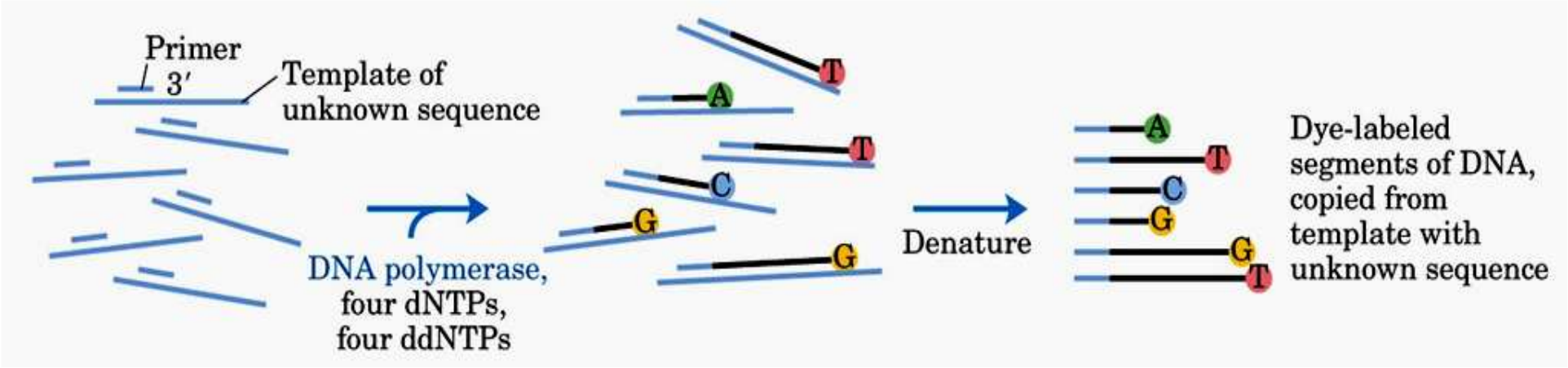



The amount of di-deoxy-nucleotide is low, it is incorporated by chance

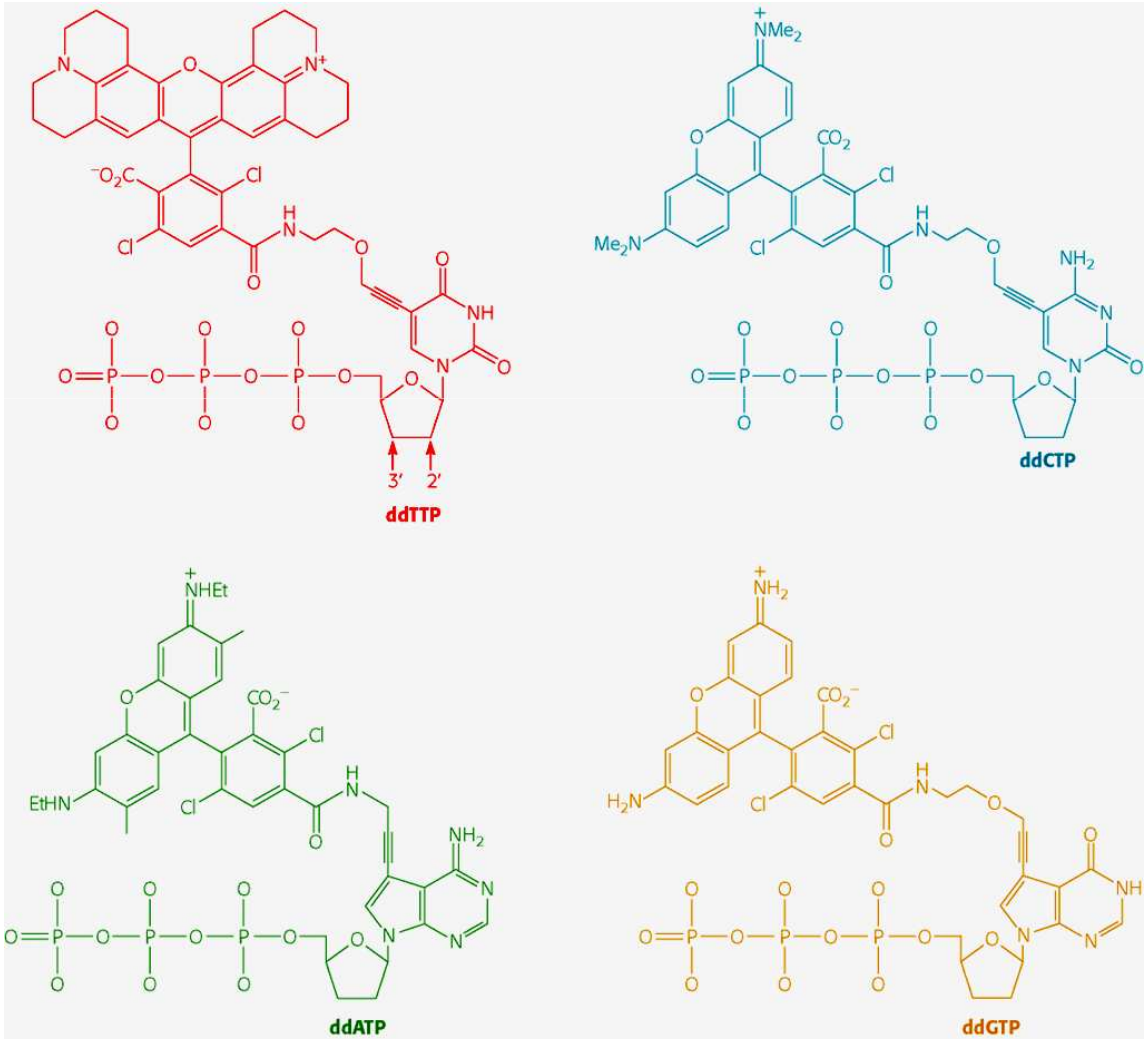


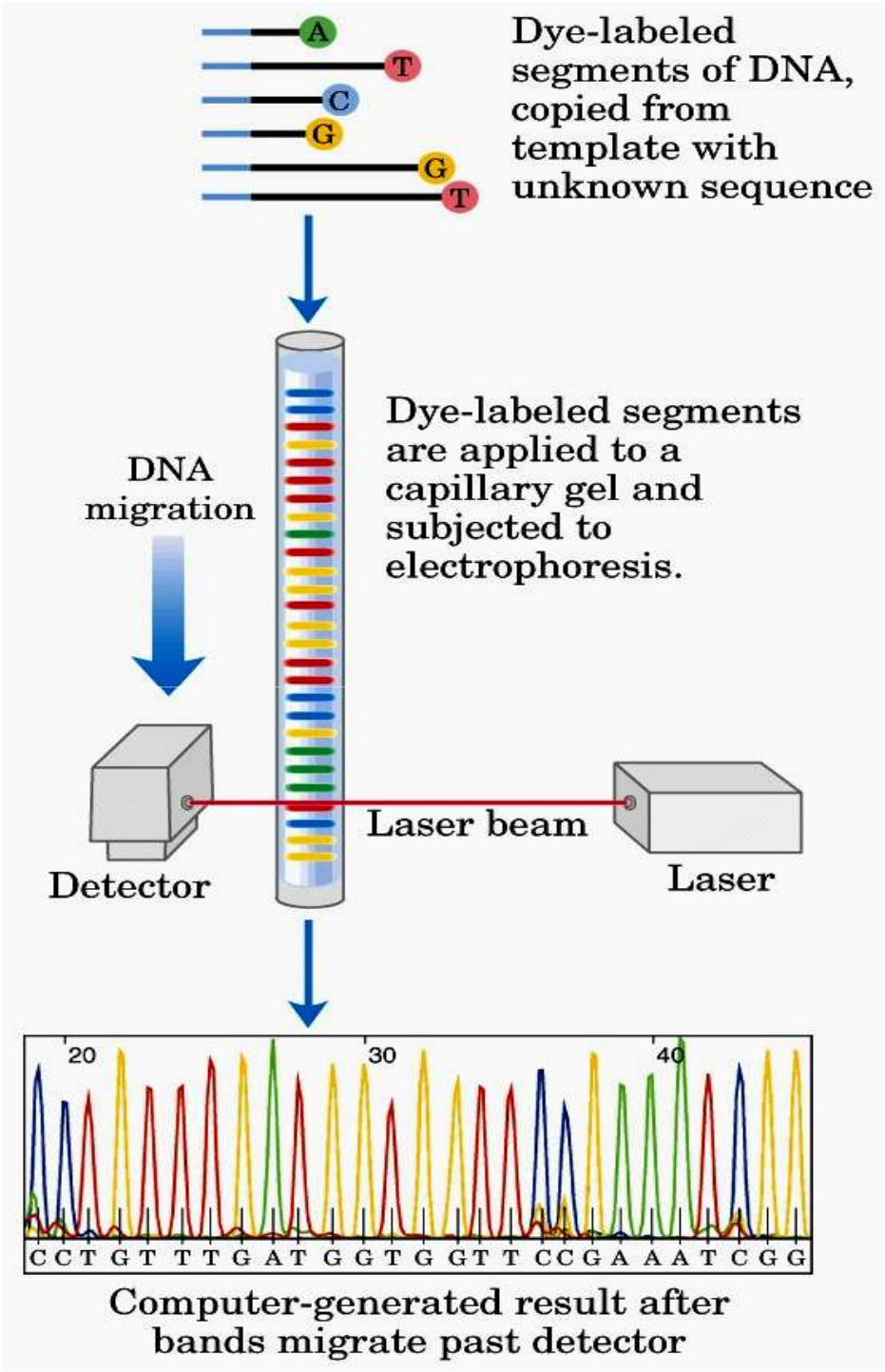

(a)
DNA a doppio filamento
••CGATGTACGTCTAGG••
••GCTACATGCAGATCC••
Preparazione dello stampo a singolo filamento
••GCTACATGCAGATCC••
Assemblaggio di quattro reazioni
Nucleotide
A
T
G
C
Miscela di nucleotidi
dATP dTTP ddTTP dGTP dCTP
dATP ddATP dTTP dGTP dCTP
dATP dTTP dGTP dCTP ddCTP
dATP dTTP dGTP ddGTP dCTP
Aggiunta di DNA polimerasi
(b)
Primer
(c)
Separazione di un campione di ciascuna miscela di reazione in una corsia separata di un gel di poliacrilammide
A T G C
Lettura della sequenza:
ACATGCAGATCC
3′ G G G C T C C T A C A G T C A G G C A G G G G A G G G G A G C C C T G 5′
G A T C

Una successiva evoluzione del metodo ha permesso l'automazione del sequenziamento del DNA. In questo caso, i nucleotidi dideossi-terminatori sono marcati mediante l'addizione di un gruppo chimico fluorogeno, diverso per ogni base.



Sequenziamento di Sanger con dideossiterminatori fluorescenti






*Sequenziamento automatico del DNA con dideossinucleotidi marcati con fluorofori.*

*Le reazioni di terminazione di catena vengono effettuate in un'unica provetta, con ciascun di-dNTP marcato con un fluoroforo diverso.*

*Il frazionamento viene effettuato con elettroforesi capillare continua.*

*La lettura è effettuata in continuo, con laser, man mano che le diverse bande migrando attraversano il raggio.*

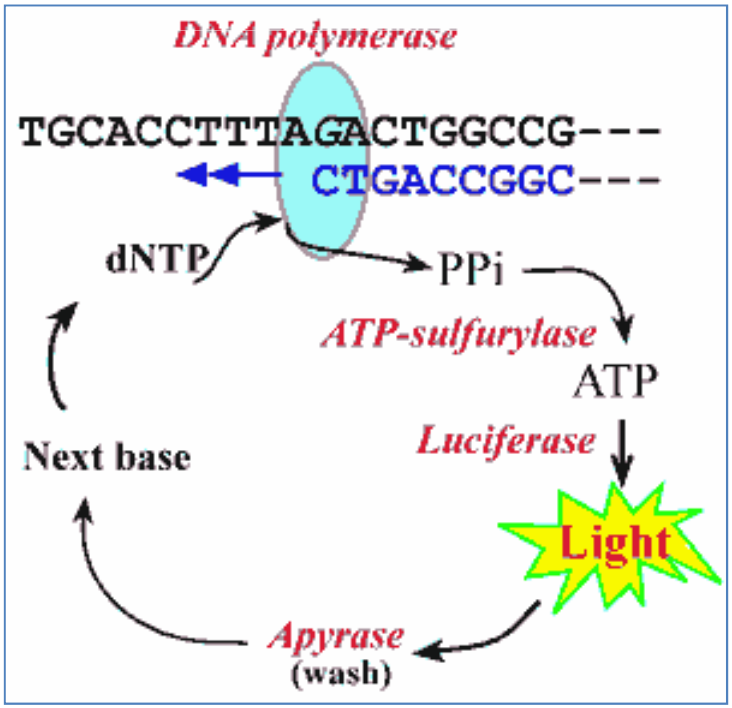
*La portata è di 600-800bp/corsa, a seconda degli apparecchi e della qualità del DNA.*

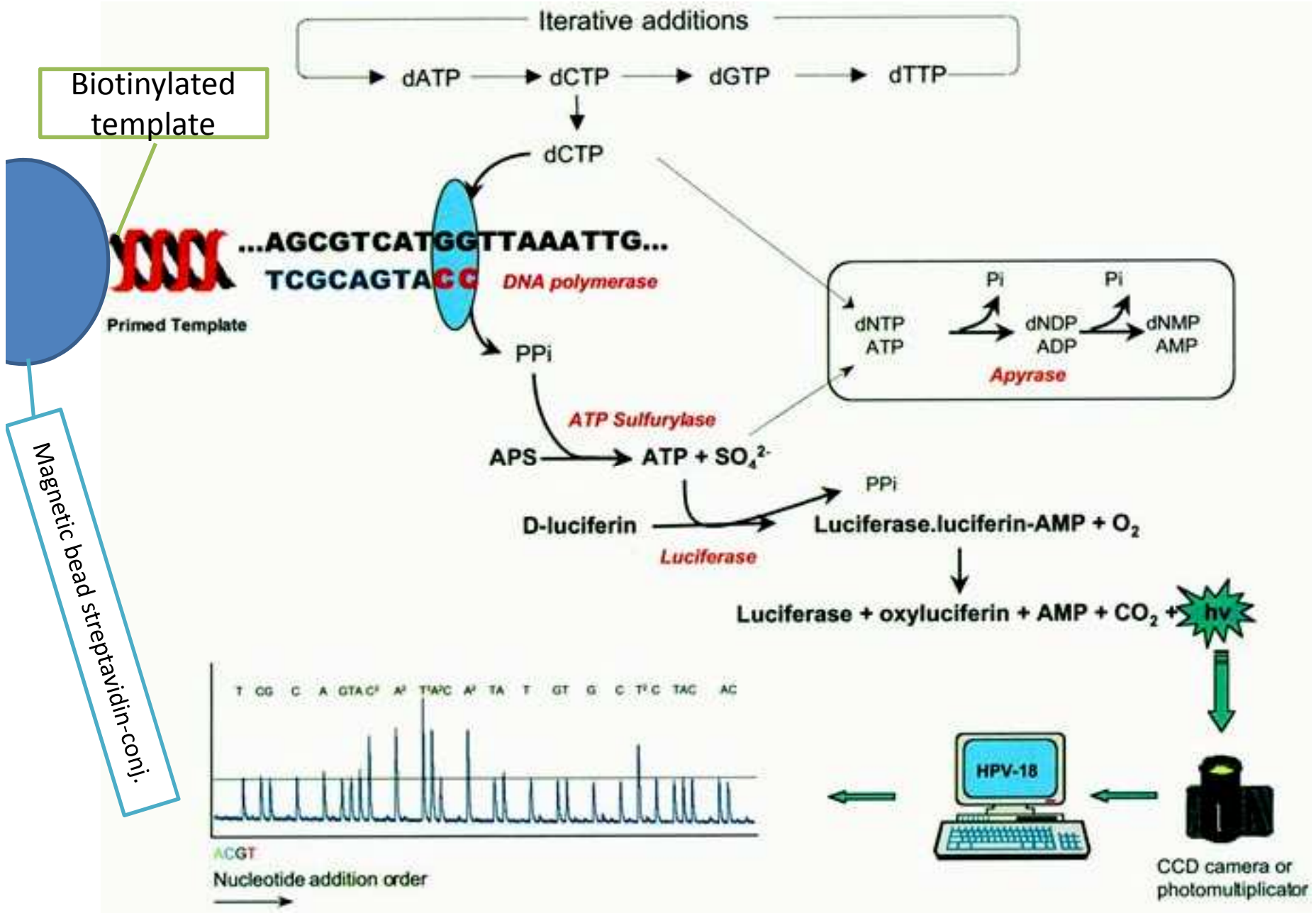
Evolution of sequencing:

. pyrosequencing
. deep-sequencing (solid phase multiple sequencing)

Good for known genomes: re-sequencing

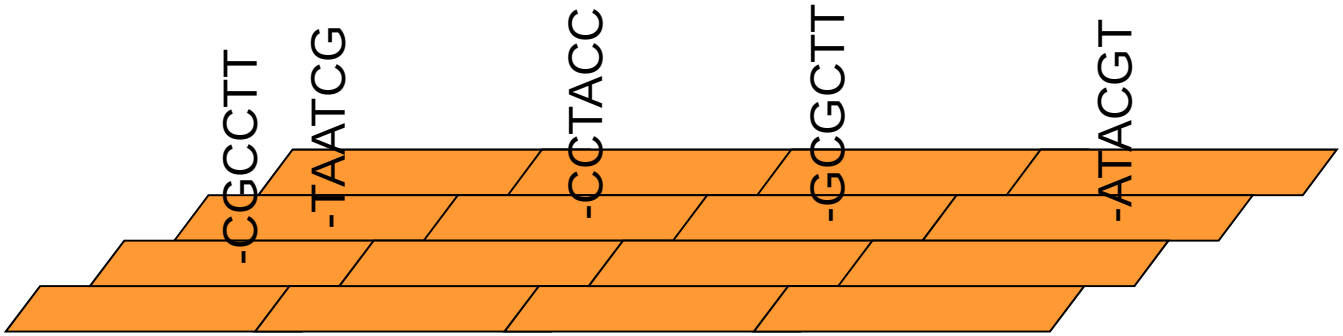
Pyrosequencing is a variant of “sequencing by synthesis”




Iterative additions
dATP
dCTP
dGTP
dTTP
Biotinylated template
dCTP
...AGCGTCATGGTTAAATTG...
TCGCAGTACC
DNA polymerase
Primed Template
PPi
ATP Sulfurylase
APS
ATP + SO4 2-
dNTP
ATP
Pi
Pi
dNDP
ADP
dNMP
AMP
Apyrase
PPi
D-luciferin
Luciferase.luciferin-AMP + O2
Luciferase
Luciferase + oxyluciferin + AMP + CO2 +
hv
HPV-18
CCD camera or photomultiplicator
ACGT
Nucleotide addition order
Magnetic bead streptavidin-conj.

http://www.pyrosequencing.com/DynPage.aspx?id=7454

Solid-phase sequencing with the Sanger method on lab-on-a-chip (microfluidic slides)




Recently developed photolitographic techniques allows the synthesis of oligonucleotides of  desired sequence on solid surfaces, like a glass slide. A “microzone” for synthesis can be as small as 10x10µm or less, so that up to 1 million different oligos can be synthsized on a 1 cmq surface.

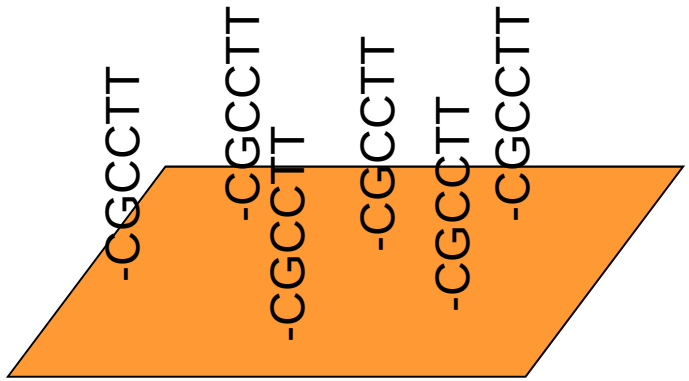
(Obviously, and contrary to what is seen in he figure, there are thousands of copies of the same oligo on each “cell” of the “microarray”

ATACGTCGTACTCGCAAGGCG

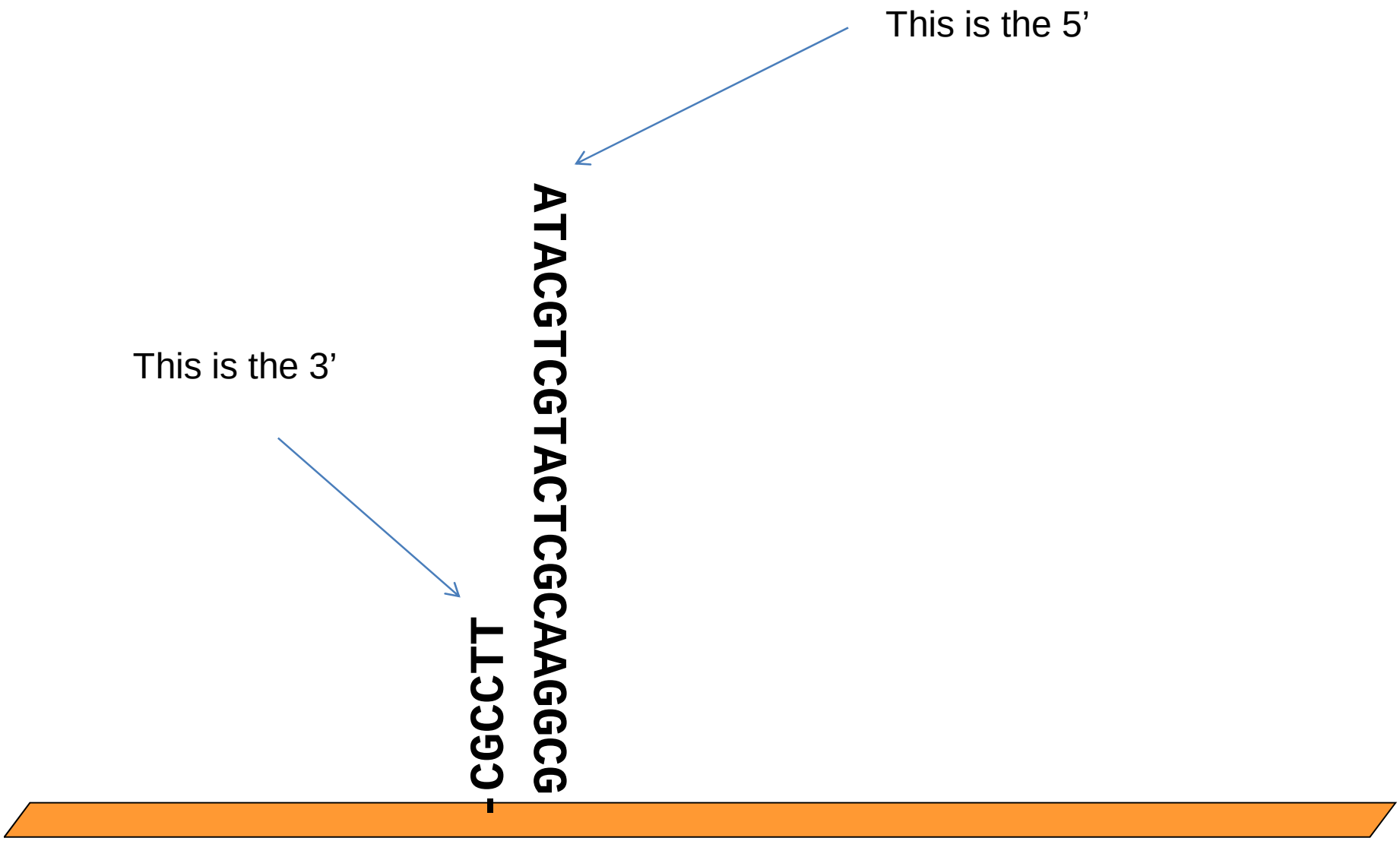
ATACGTCGTACTCGCAAGGCG

ATACGTCGTACTCGCAAGGCG

ATACGTCGTACTCGCAAGGCG




a single cell of the microarray is represented here


This is the 5’
This is the 3’
ATACGTCGTACTCGCAAGGCG
TTCCGC

Nella cella a flusso, introduciamo la Klenow, più il nucleotide marcato A(green)

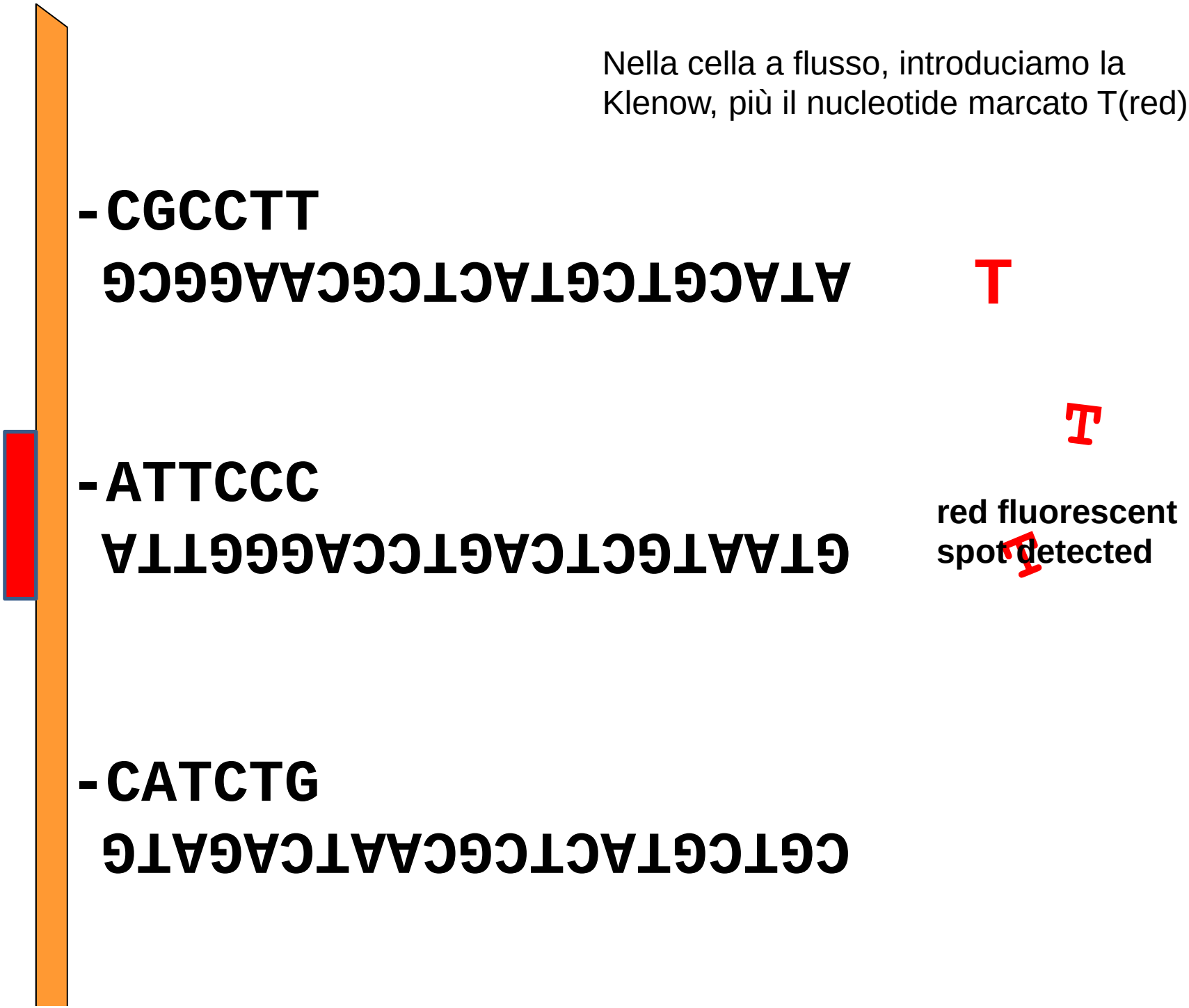
Nella cella a flusso, laviamo via tutto

-CGCCTT
ATACGTCGTACTCGCAAGGCG

-ATTCCC
GTAATGCTCAGTCCAGGGTTA

-CATCTG
CGTCGTACTCGCAATCAGATG

Nella cella a flusso, introduciamo la Klenow, più il nucleotide marcato T(red)



The “C” (yellow) follows, then the “G” (blue).

A new cycle begins.....

This process can work in a highly-parallel fashion is a precise method of detection is available: a laser scanner can do the job!
Tens of thousand sequences can be run in parallel.

Variation of this technology  or slightly different technologies, such as **pyro-sequencing** are today exploited to obtain re-sequencing apparatus that carry out millions of sequencing reaction at a time on solid surfaces.

Today, 3 technologies are available:

Illumina-Solexa Genome Analyzer
Read lengths: 36 bp, 50 bp or 75 bp for fragment or paired sequencing
Throughput (reads):  120 million reads per run, fragment

SoliD – Applied Biosystems
Read lengths: 50 bp fragment, 25 bp and 35 bp paired
Throughput (reads): > 160 million reads per slide, fragment

Roche – 454
Read lengths: Averaging 350 - 400 bp
Throughput (reads): ~1 million reads per run

Next Generation Sequencing
deep-sequencing / mass sequencing

✓generation of “DNA-nanoclones” on distinct solid surfaces by PCR

✓highly parallel fluorescent in situ sequencing and laser detection

✓record read-out i.e. millions or short sequences (“reads”)

✓align reads on genomes

Is sequencing of single molecules possible ?

Helicos site

## Functional genomics

- Functions not associated to transcription
  - centromeric sequences
  - matrix attachment regions
  - other structural elements
  - telomeric sequences
  - .....

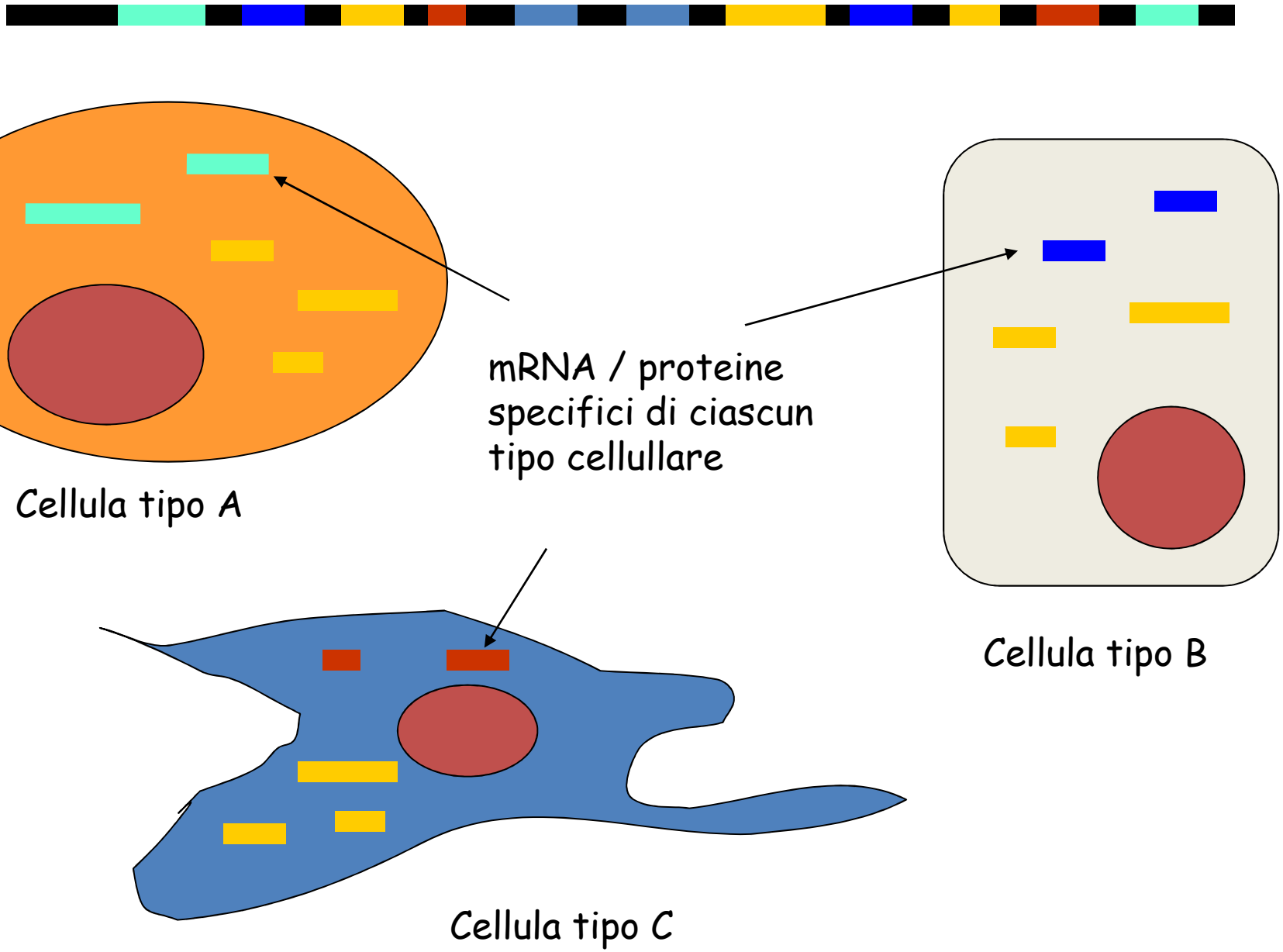
- Functions associated to transcription
  - regulatory
  - coding
  - structural/cofactor RNAs
  - regulatory RNAs

# Genome expression

Development, growth, apoptosis, homeostasis, and all cell activities depend on the qualitative and quantitative control of gene expression.

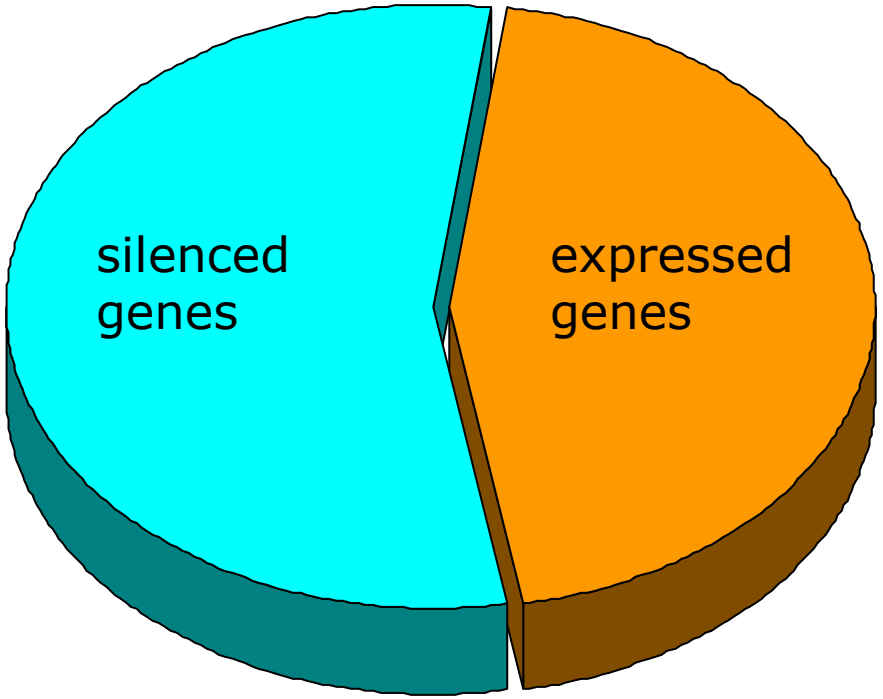
How could we estimate or measure the entire **genome expression** in a cell, tissue or organism ?

**Il DNA di ogni cellula è identico: tutte le cellule contengono l'intera informazione**



In any cell type, a large part of the genes are kept in a silenced form.

…in any differentiated cells:




The primary control is at the genomic level, i.e. transcription

*H. sapiens* genome: 30,000 genes estimated

alternative splicing

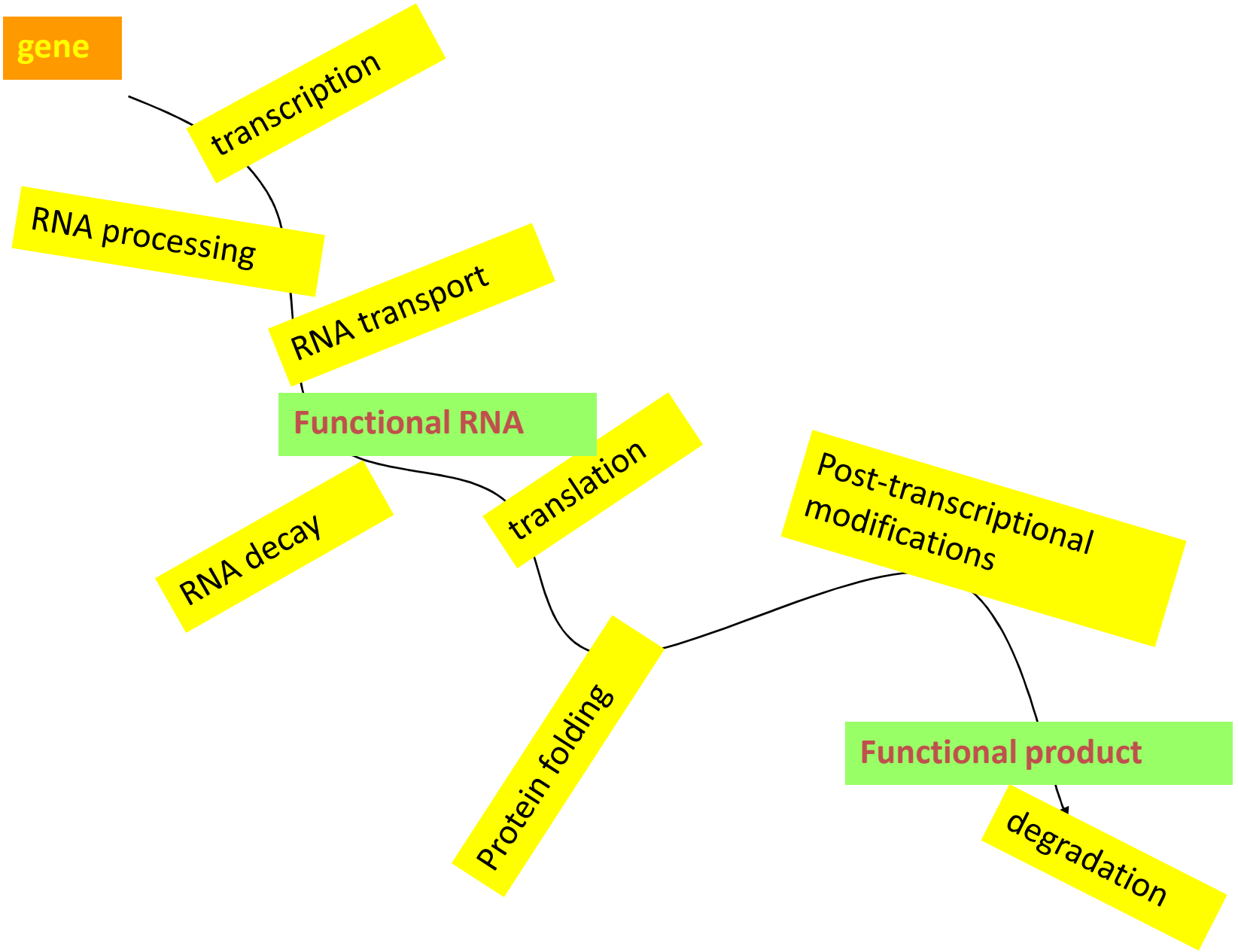
100 - 150,000 proteins estimated

genetic programs for multicellular development

We call **genetic program** the set of genes that a cell uses for expressing a function (e.g. differentiation, cell division, apoptosis, response to a drug, etc…)

We call **gene expression profile** the sum of genes that are expressed in a cell or tissue, weighted by their individual levels of expression.

Transcription is the most important level of **control**



# Large-scale analysis of gene expression

measuring mRNAs

-More mature and technically OK

**Gene-by-gene methods to measure gene expression (mRNA)**

measuring proteins

- Several approaches available, still under development

**Gene-by-gene methods to measure gene expression (proteins)**

**Why mRNA?**

- mRNAs represent a closer mirror of genome activity than proteins
- Homogeneous chemistry makes handling robust
- Complete knowledge of probes and control of hybridization conditions guarantees specificity
- The amount of any mRNA species is not necessarily proportional to the amount of the encoded protein

**Why proteins?**

- Proteins are the real object of gene expression
- Separation and quantitation may be more reliable than for mRNAs
- Not all measured protein may represent “functional” protein.
- Dis-homogeneous chemistry makes it difficult to find procedures equally good for all proteins

Large-scale analysis of gene expression limits

(with microarray methodology → until deep-sequencing methods )

Theoretical limit: measuring all genes

Practical limit: measuring all known genes

Practice: measuring a relevant fraction of genes

How to measure the activity of all genes (genome-wide) in cells/tissues (mRNA).

**Sequencing methods**

(EST) (1980)

SAGE (1995)(LongSAGE, CAGE)

direct re-sequencing (deep sequencing, NGS)(2006)

**Hybridization methods - DNA microarrays, oligonucleotide microarrays.**

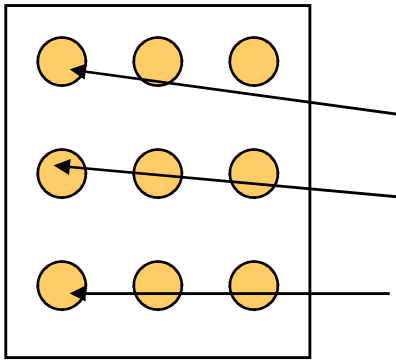
Spotted arrays (1996)

In situ synthesized oligo arrays (1999)

Bead-arrays® (2001)

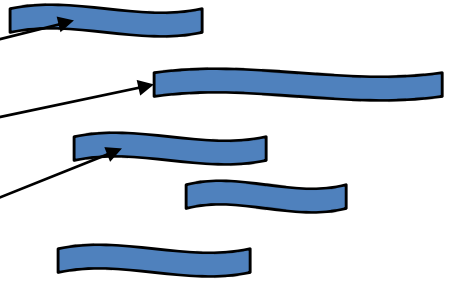
La tecnica di ibridazione su fase solida permette di verificare la presenza di molti geni contemporaneamente

Sonde complementari a diversi geni vengono depositate su un filtro

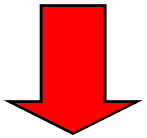


Gene A
Gene B
Gene C
eccetera

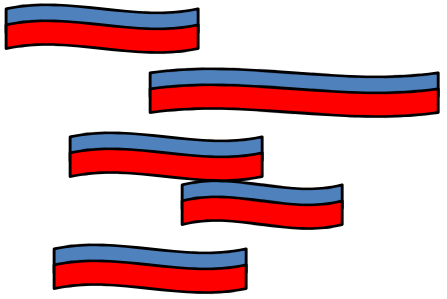
L'RNA o il DNA vengono estratti dalle cellule



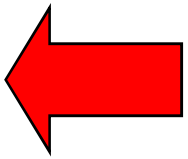
RNA A
RNA B
RNA C
eccetera ...

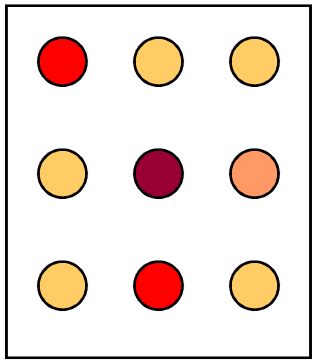


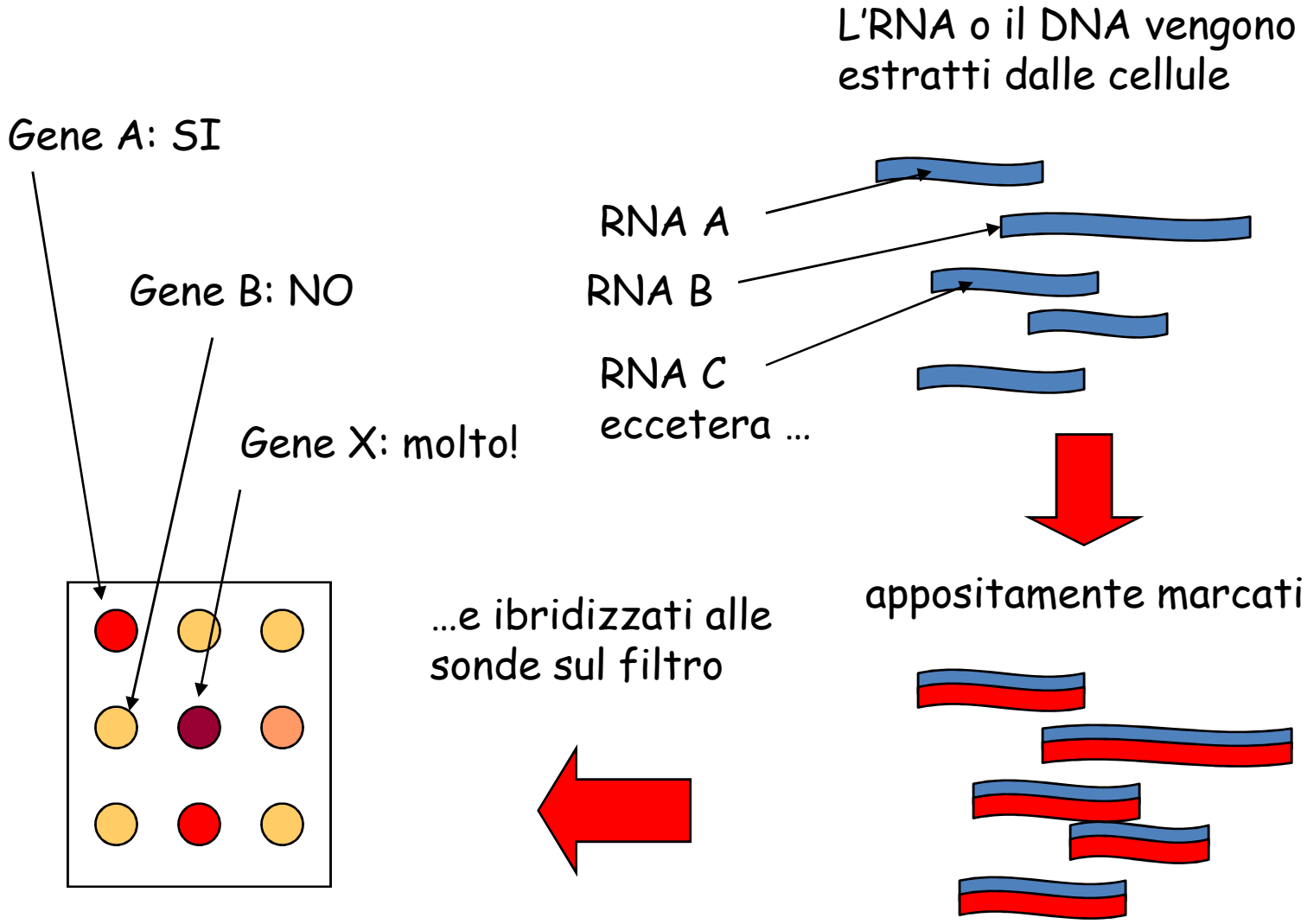
appositamente marcati

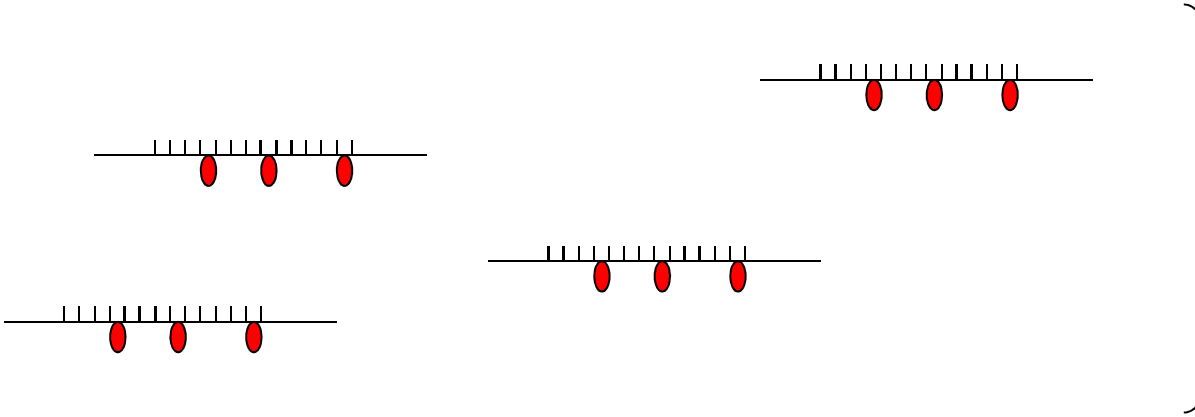


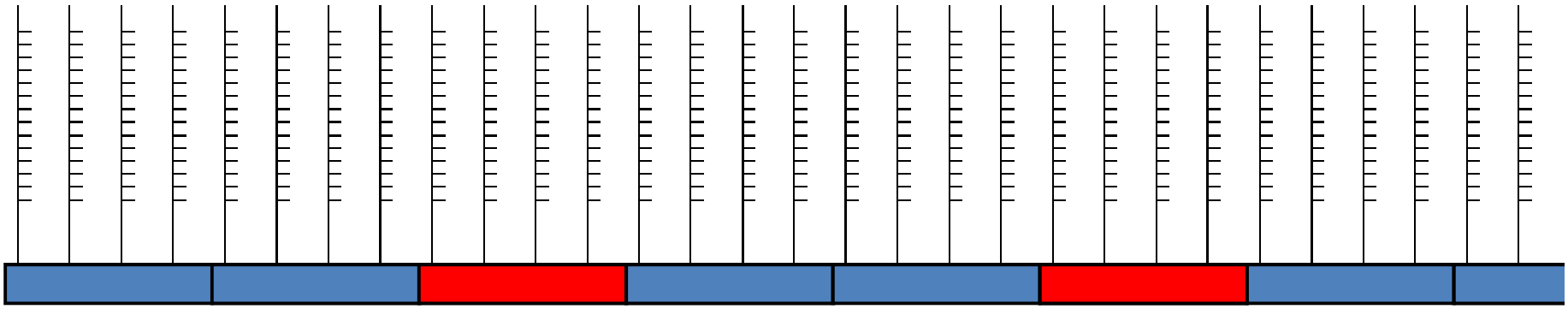
...e ibridizzati alle sonde sul filtro






L'RNA o il DNA vengono estratti dalle cellule
RNA A
RNA B
RNA C
eccetera …
appositamente marcati
…e ibridizzati alle sonde sul filtro
Gene A: SI
Gene B: NO
Gene X: molto!




a section of a oligonucleotide microarray at row "n"

Qualitative and quantitative information on all transcribed sequences ("Transcriptome")

- Changes during developmental processes and cell differentiation
- Cell-type and tissue specific gene expression profiles
- Genes induced and repressed in response to drugs or environmental hits
- Genes regulated in response to cellular communication (signal transduction)
- Genes regulated in response to hormones
- Individual specificity of gene expression
- Changes during neoplastic transformation or other disease

**DNA chips: there are currently several types available:**

## Spotted

cDNA microarrays
DNA microarrays

10,000 - 100,000 probes / $cm^2$ are spotted on glass slides using an automated microarrayer.

Probes are cDNA or PCR products representing known genes or simply EST

Chemically synthesized long oligonucleotides (30 – 70 nt) representing cds of known genes/EST

**DNA chips: there are currently several types available:**

## 2. In situ synthesized

| | |
|---|---|
| Oligonucleotide arrays (Affychip®) Affymetrix | Up to 450,000 20-25nt long oligonucleotides / $cm^2$ are sythesized directly on the chip surface, using a photolitographic technique. Each gene is represented by a "probeset" of 12-13 probes. |
| Long oligonucleotide arrays | Up to 250,000 different 30-60 nt long oligonucleotides / $cm^2$ are sythesized directly on the chip surface, using an ink-jet technique. |
| Bead Arrays® Illumina | Oligonucleotide probes (30-50 nt) are synthesized on beads, with a identification address. Beads are randomly arrayed on surfaces and position of each oligonucleotide determined using addresses. |

## Spotted (pinspotted) DNA microarrays

A

**1 - 100,000 probes / $cm^2$ are spotted on glass slides using an automated microarrayer.**

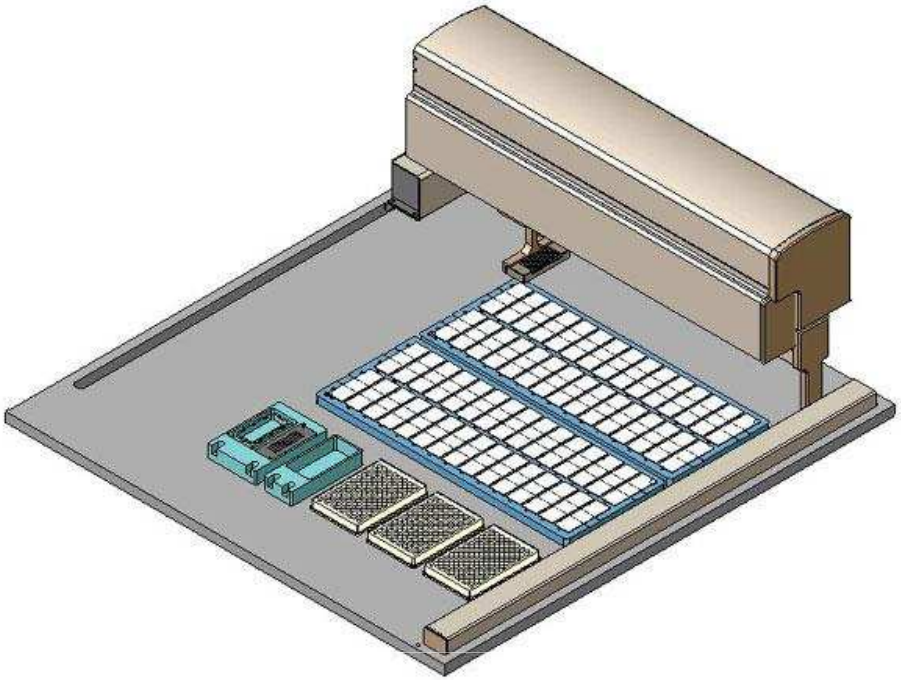
**Probes are cDNA fragments or PCR products representing known genes or simply EST**

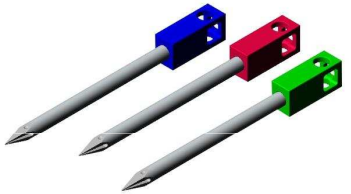
**Probe size: 200-1,000 nt**

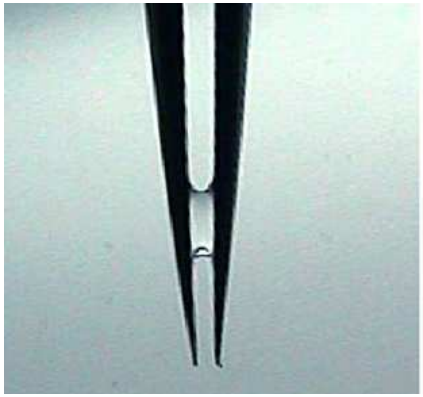
B

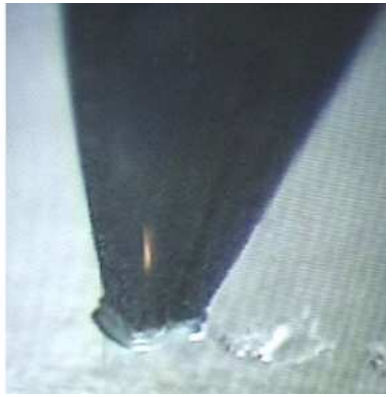
**Same as above, but with:**

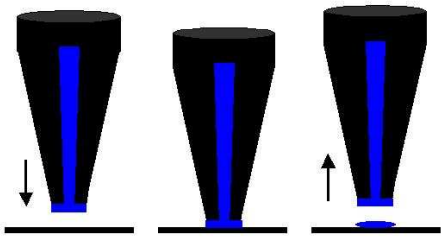
**Oligonucleotide probes (30 – 70 nt in lenght)**



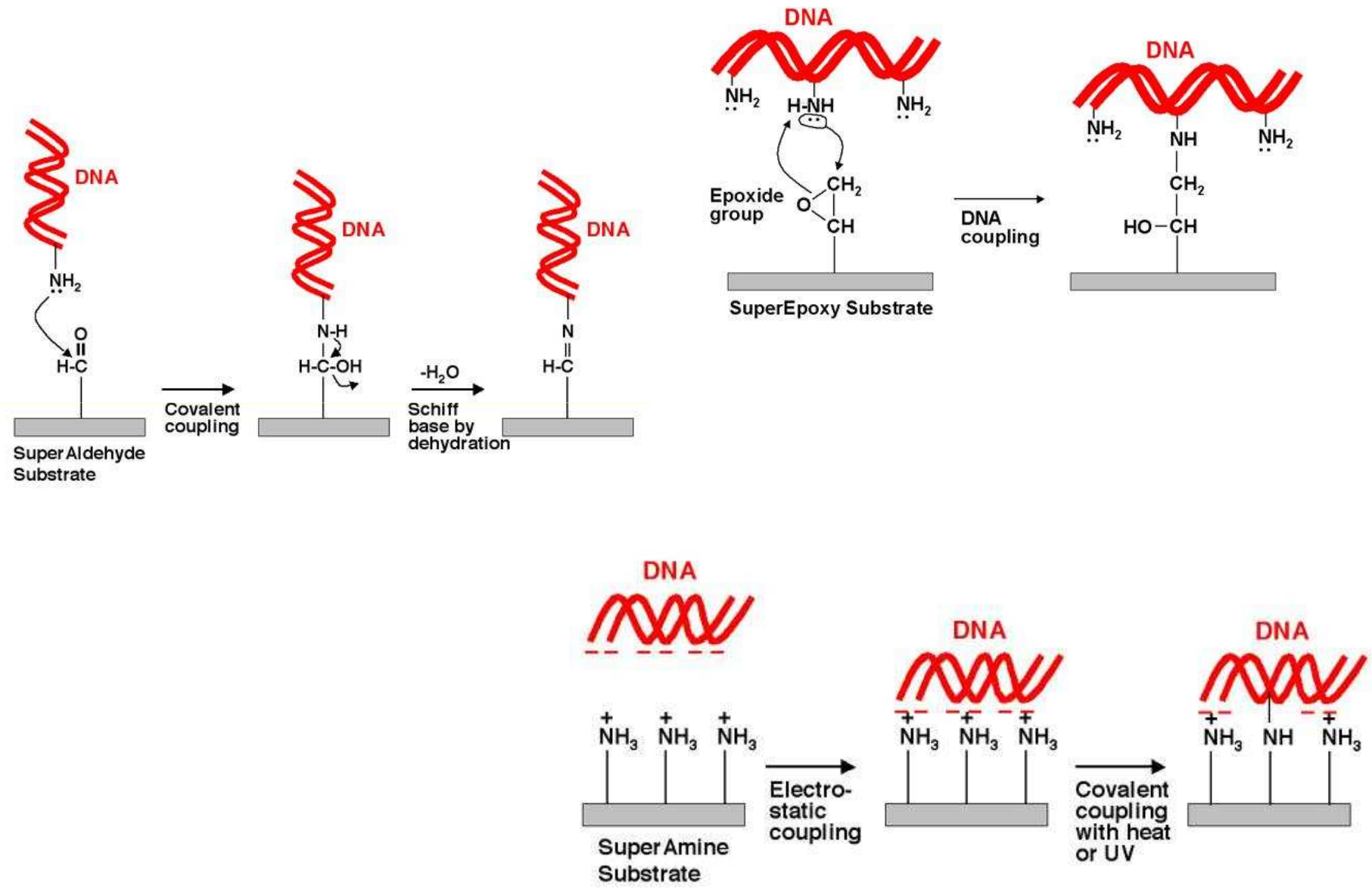








Single-strand



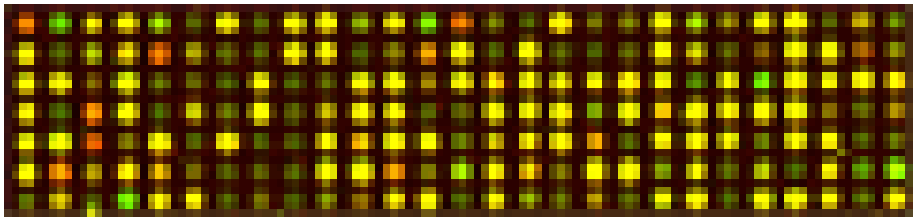
## Long Oligonucleotide arrays

**Up to 250,000 oligonucleotides / $cm^2$ are sythesized directly on the chip surface, using an ink-jet technique.**

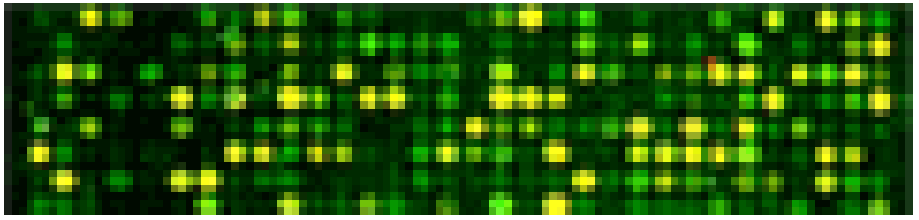
**These 60 nt long oligonucleotides represent sequences of known genes or EST**

**Chemical synthesis of oligonucleotides**

In situ



spotted



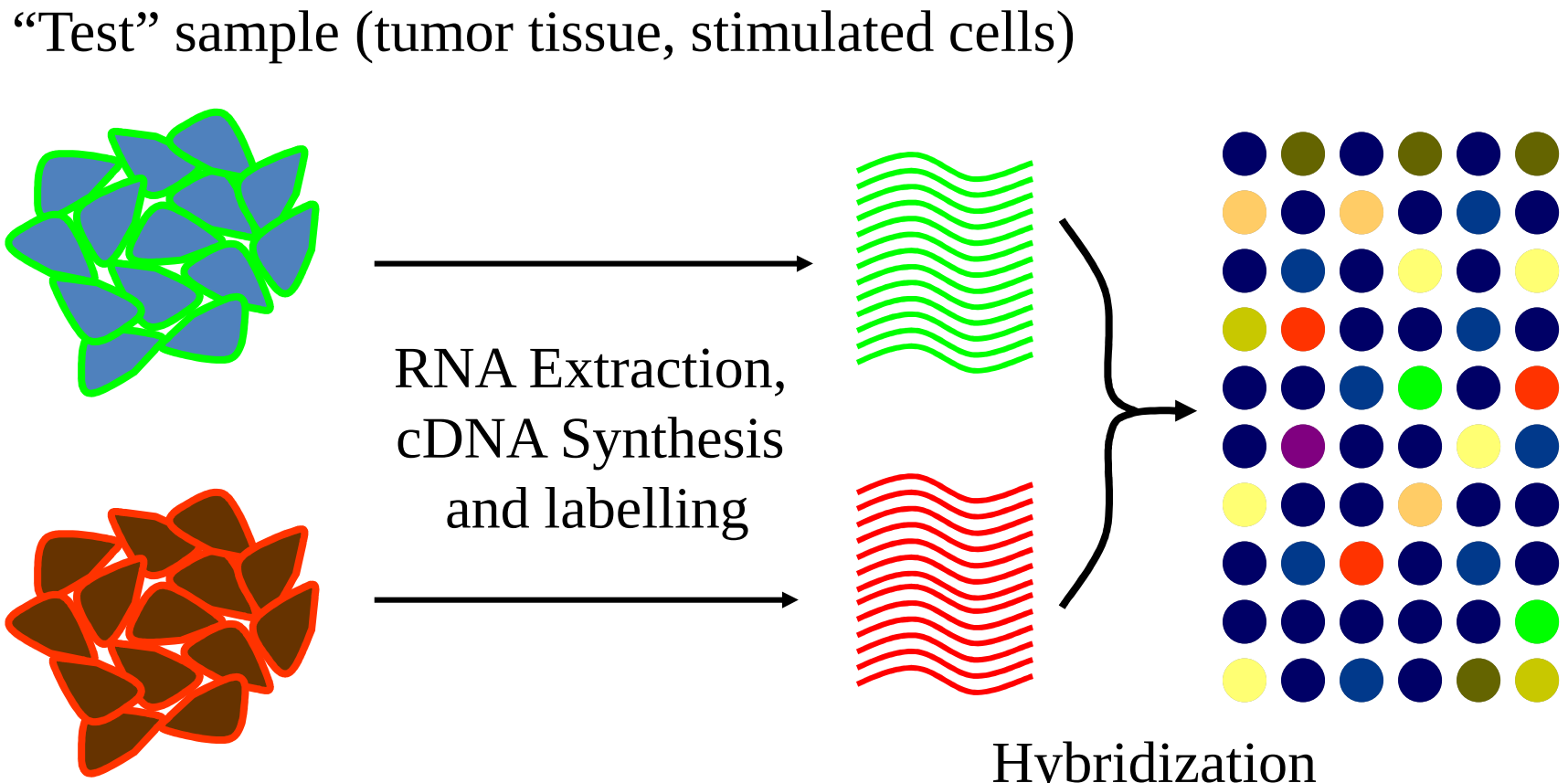
Spotted arrays, or ink-jet oligo arrays are commonly used for **relative** measurements, i.e. to compare gene expression between two biological samples.

RNA from **sample** and from **reference** are labeled by introducing two different fluorochromes.

This allows co-hybridization of the two samples to the same chip, providing direct comparison by two-color analysis

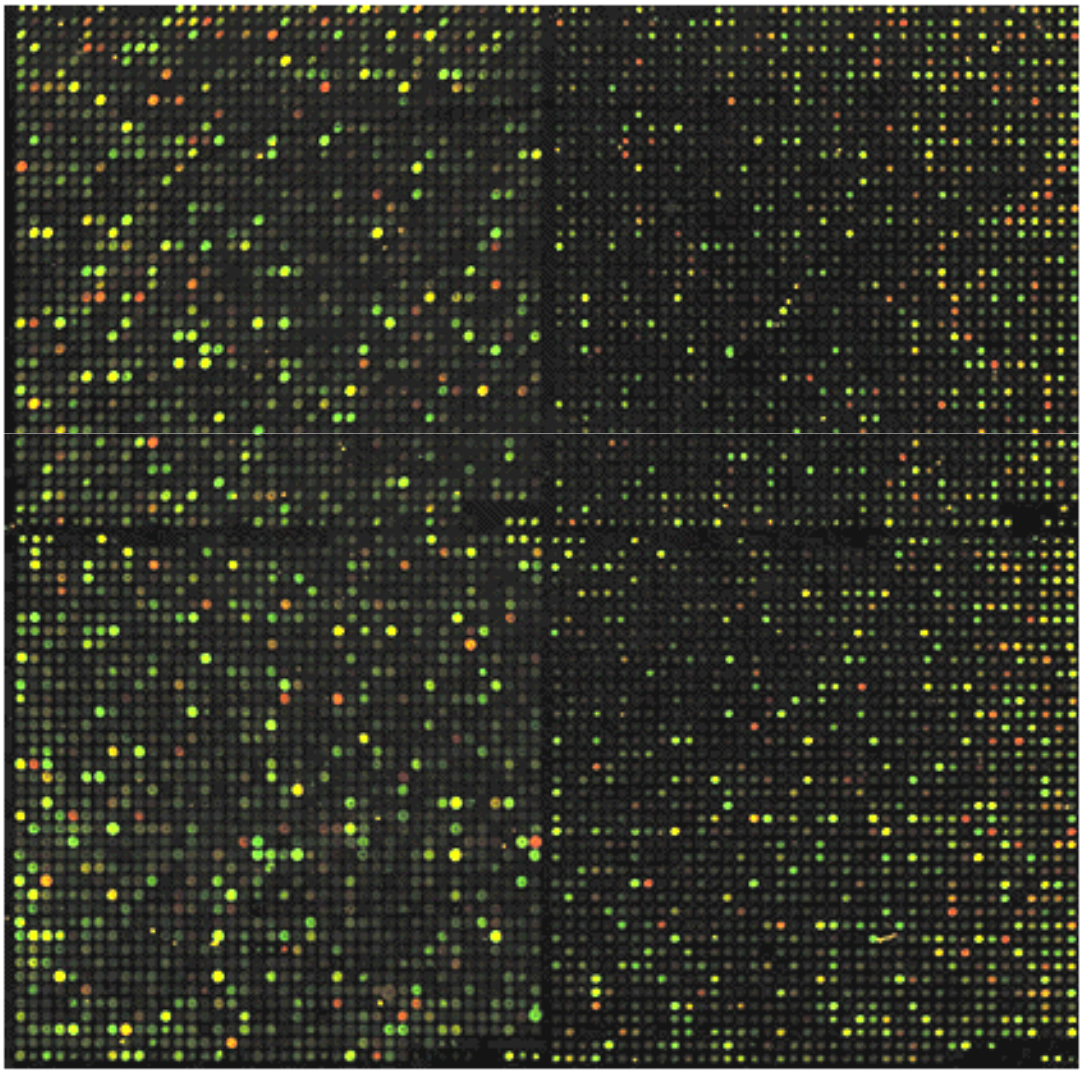
The most common fluorochromes are the cyanines Cy3 (red) and Cy5 (green)

Co-hybridization of double-color labelled test and reference samples



How a spotted microarrays hybridized with two-colors probes looks like

1 cm



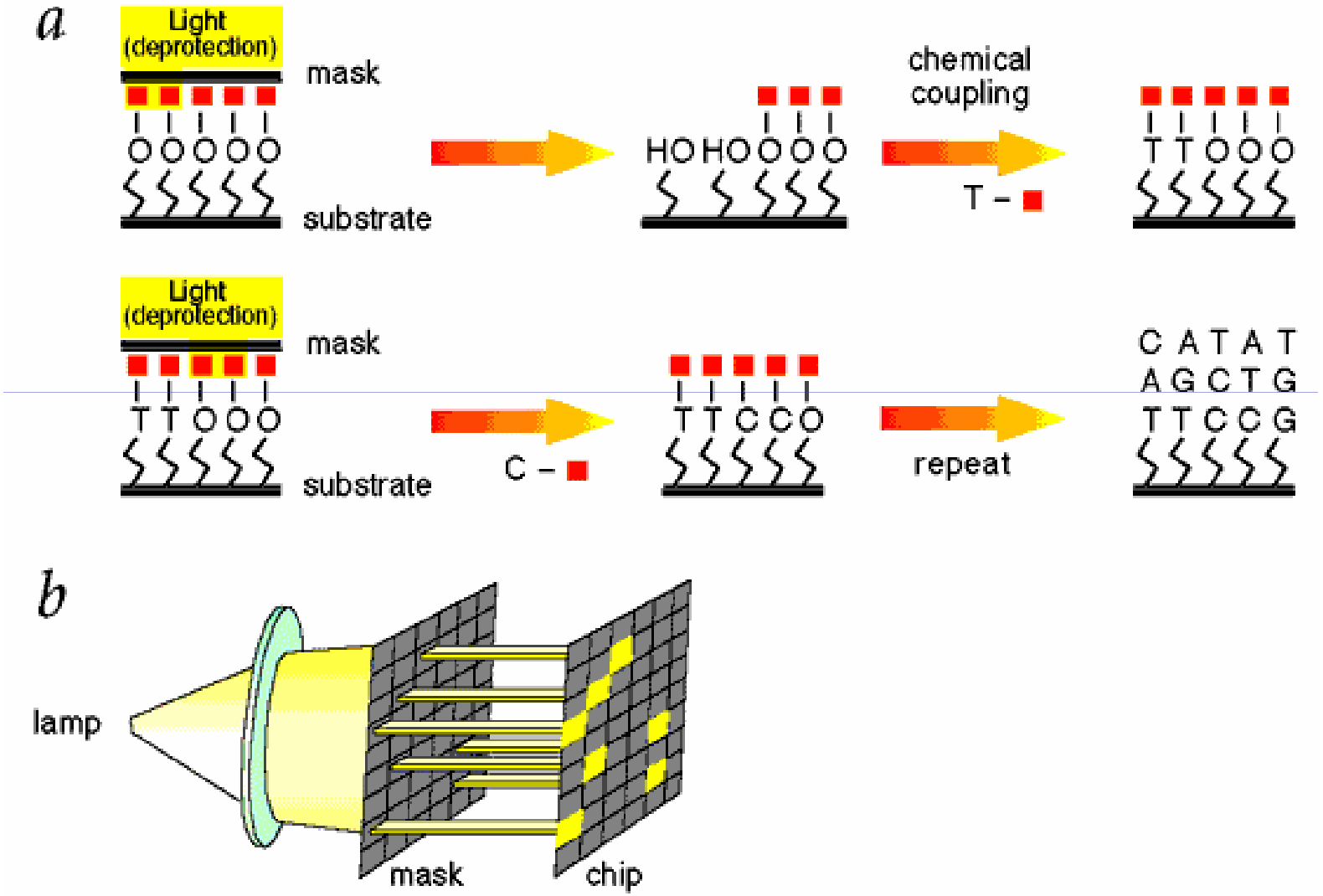
## Short Oligonucleotide arrays (Affymetrix)

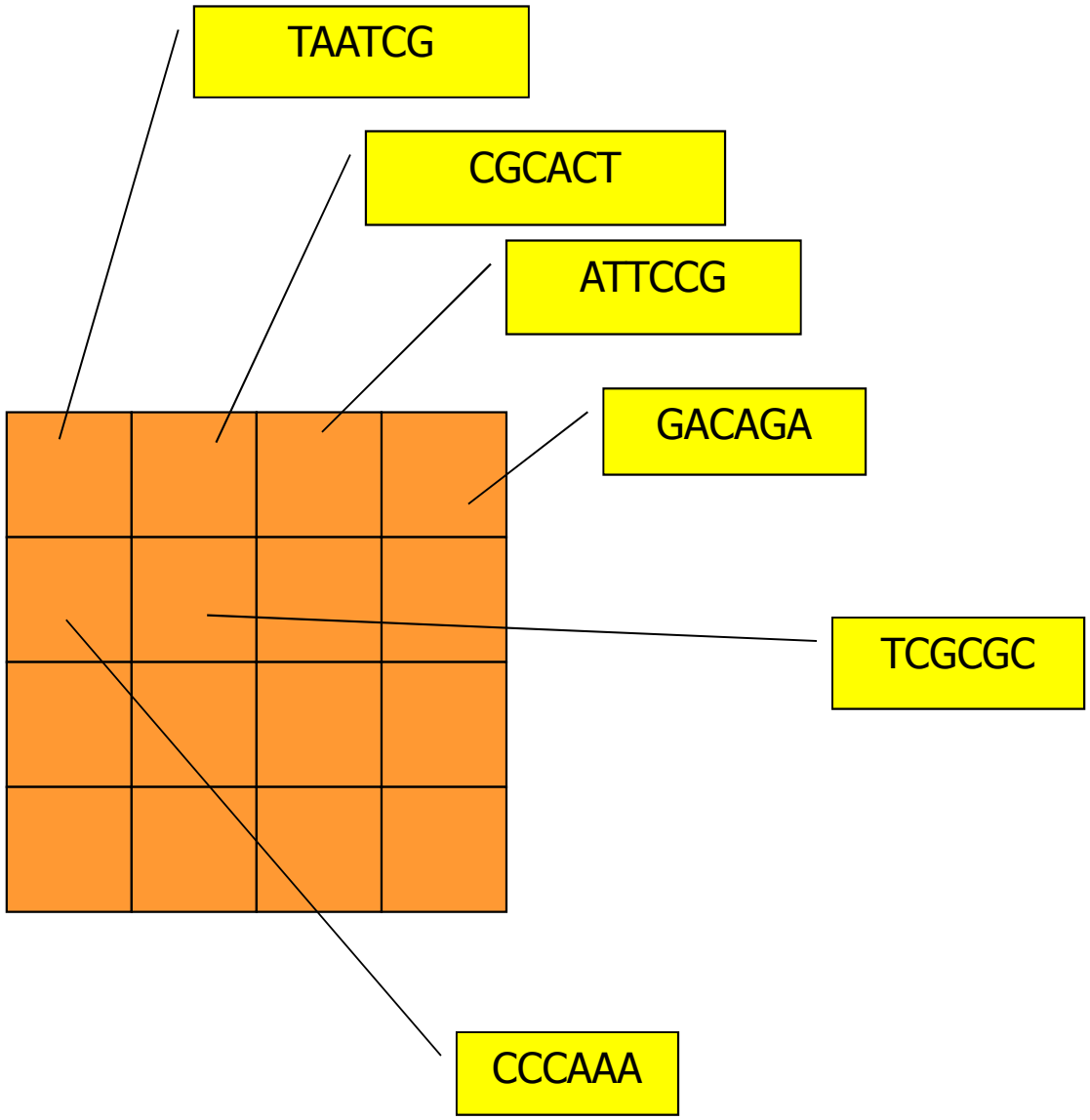
**Up to millions of oligonucleotides / $cm^2$ are synthesized directly on the chip surface, using a photolithographic technique.**

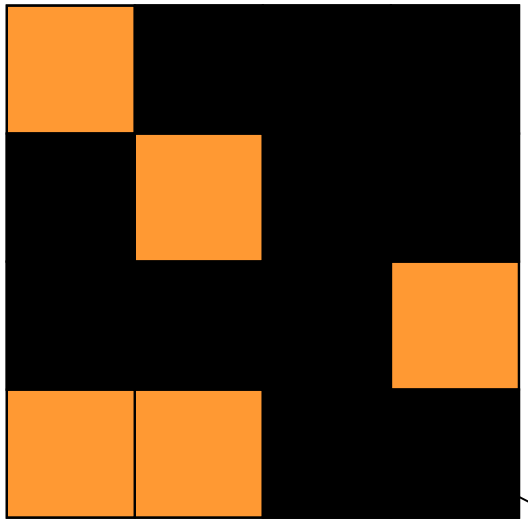
**These 20-25 nt long oligonucleotides represent sequences of known genes or EST**

**Chemical synthesis of oligonucleotides**

Affymetrix technology allows direct synthesis of oligonucleotides in micro-zones on the glass slide by photolithograpy

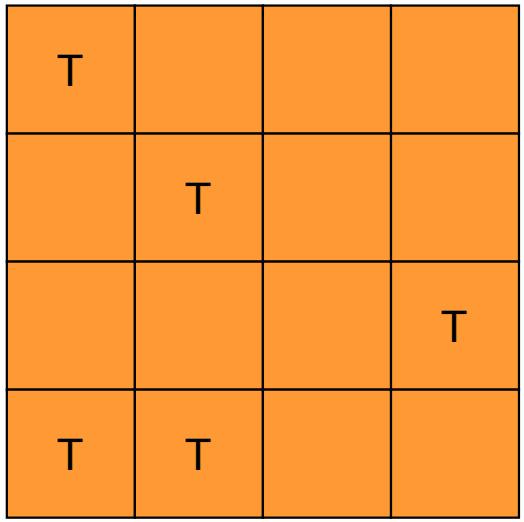


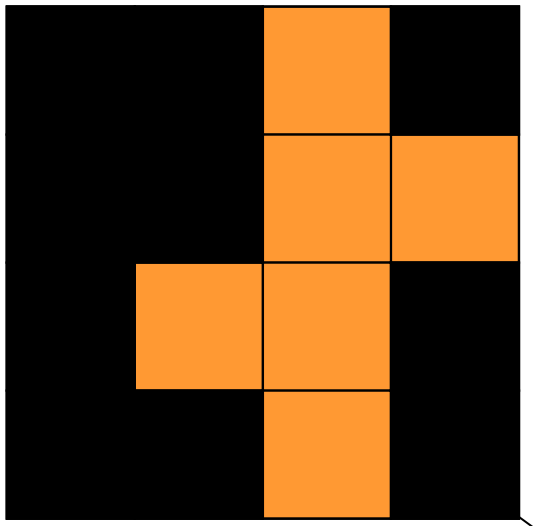

TAATCG
CGCACT
ATTCCG
GACAGA
TCGCGC
CCCAAA



A solution with photoactivatable “T” and reagents is added

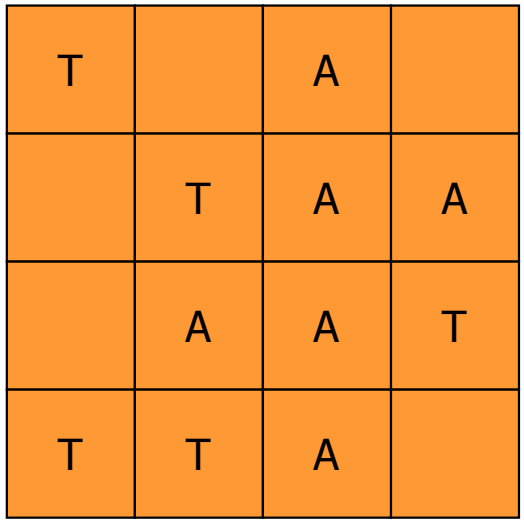
A photolithographic mask is superimposed to avoid light activating the “T” in wrong zones


T
T
T
T
T



A solution with photoactivatable "A" is added with reactants

A second photolithographic mask is used

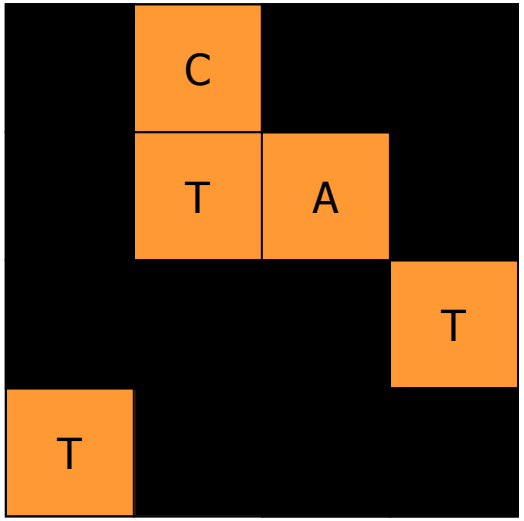

T
A
T
A
A
A
A
T
T
T
A

The same follows for “C” and “G”.

| | | | |
|---|---|---|---|
| T | C | A | G |
| C | T | A | A |
| G | A | A | T |
| T | T | A | G |

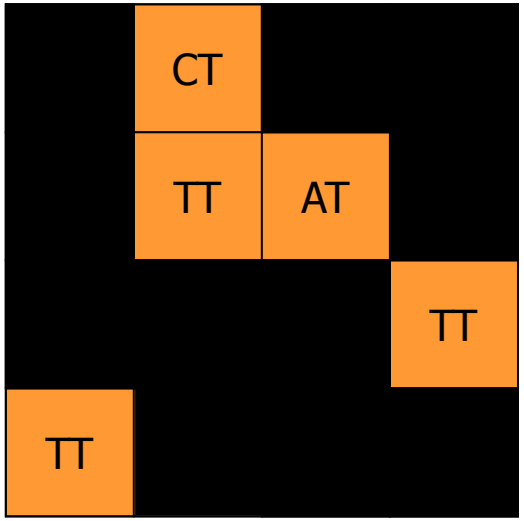
A second cycle is now realized, to add the second nucleotide to each of the zones, an so on for all the lenght required

(practically up to 25 nucleotides)

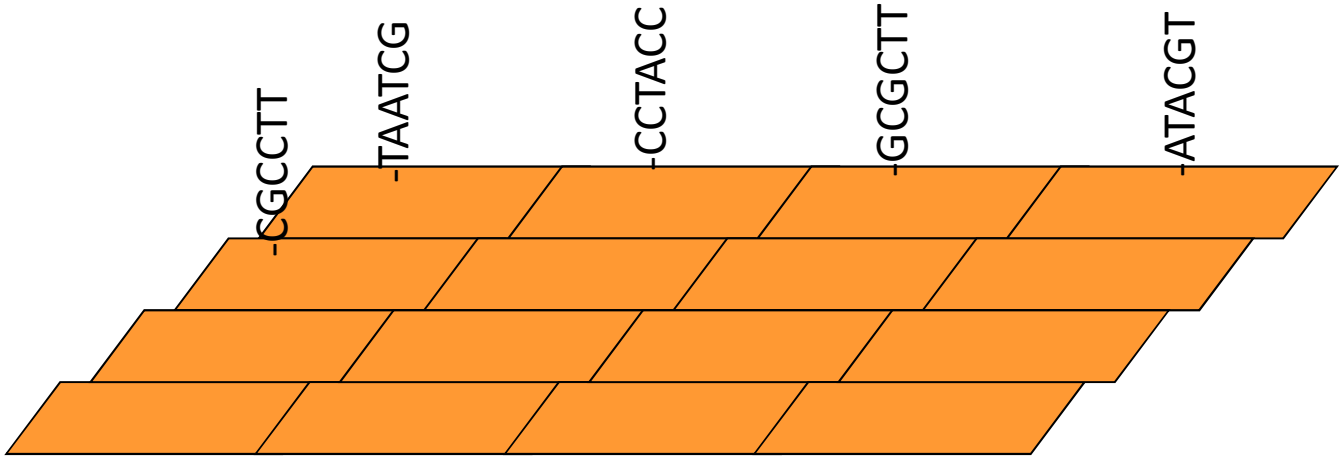



2° cycle:
A solution with photoactivatable "T" and reagents is added

A photolithographic mask is superimposed to avoid light activating the "T" in wrong zones




A “T” is added to programmed positions, then “T” is washed away, a second photolithographic mask is imposed, a “A” is added and over and over.....



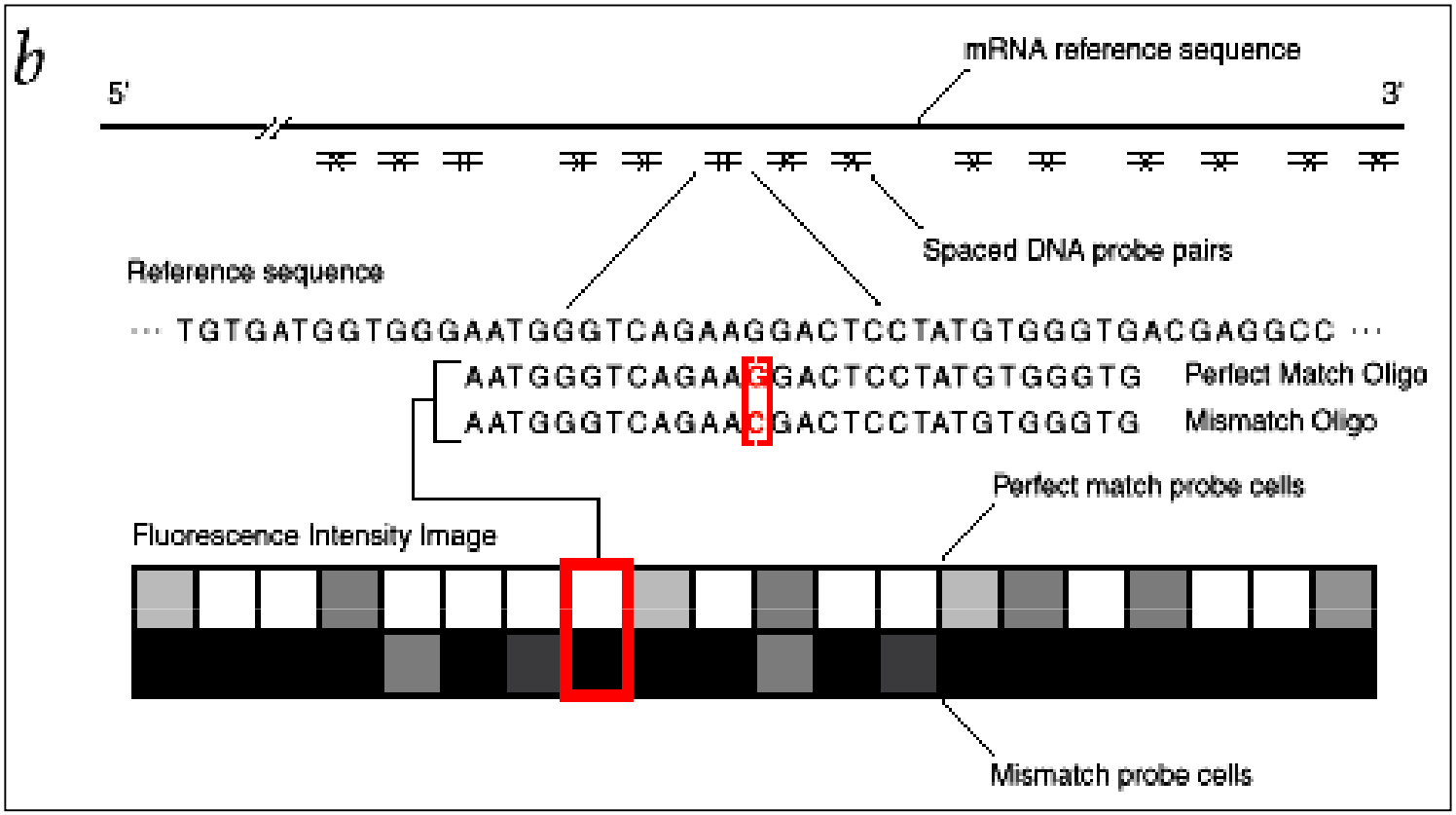

(Ovviamente, in ogni quadratino ci sono migliaia di oligonucleotidi uguali)

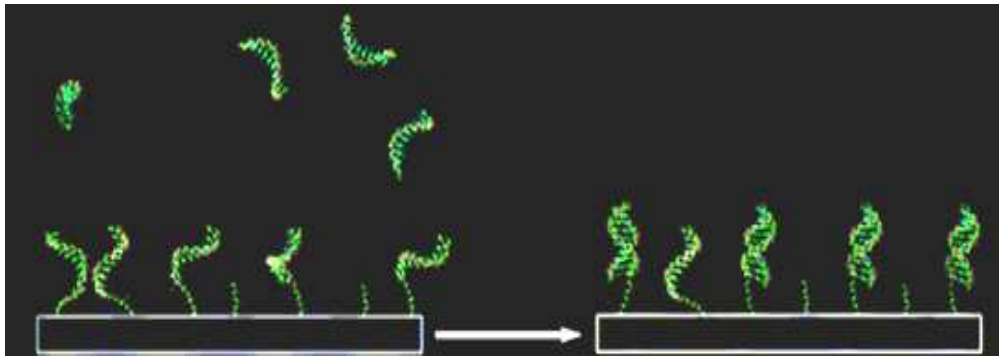
One limitation: the lenght of oligonucleotides (20-25 nt)

This can give problems of aspecific hybridization due to:
-Similar sequences can be present in different mRNAs
-Hybridization is done at the same temperature that is average of the optimal temperature for individual probe-target pairs

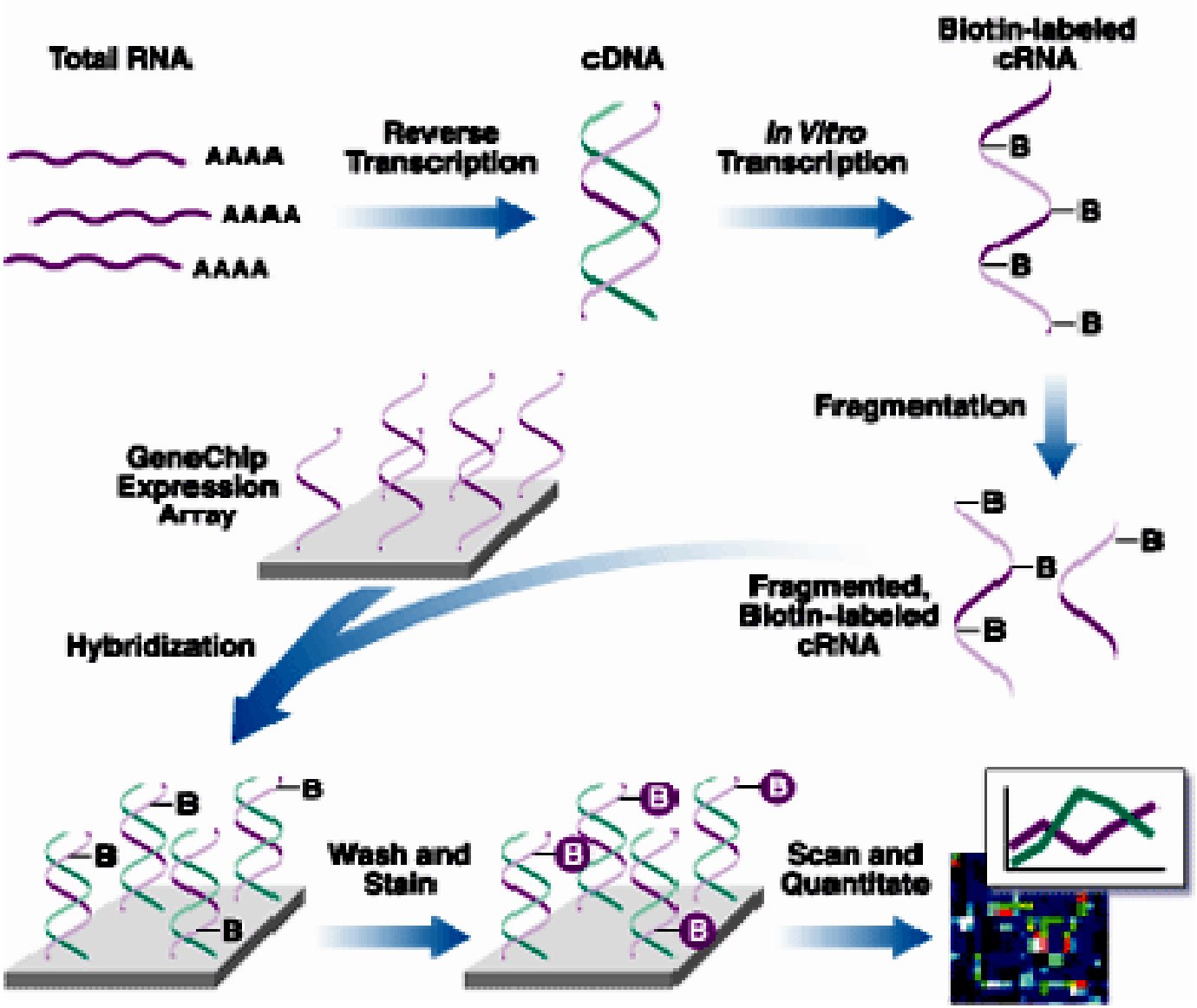
Affymetrix  gives one solution to these problems:

# Affymetrix GeneChip probe set



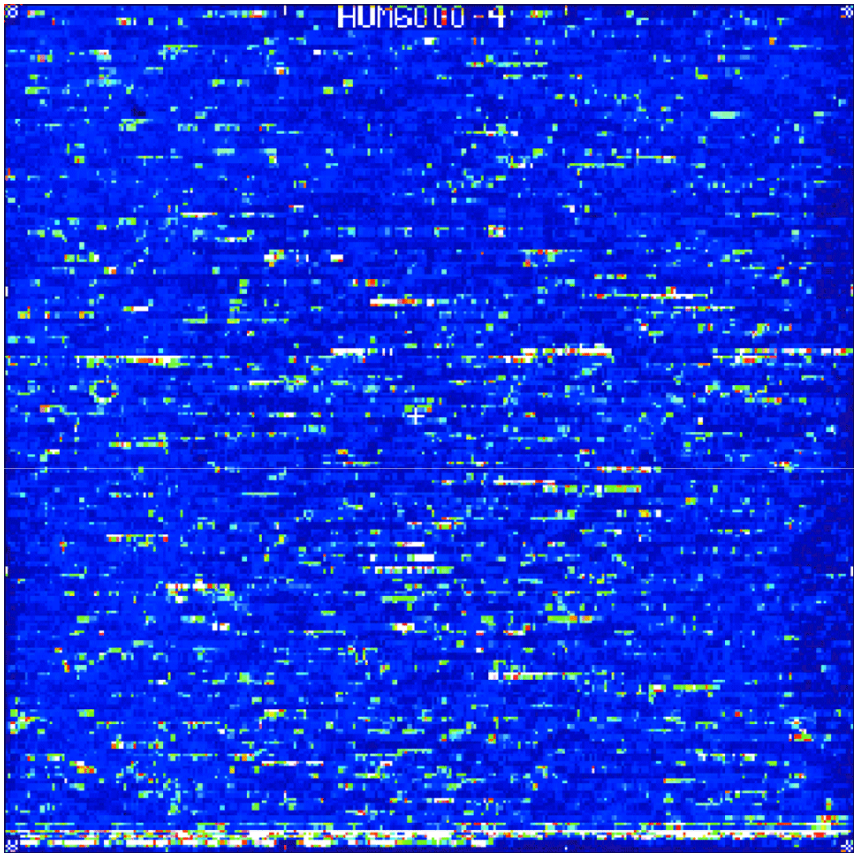



How an Affychip is used



# How an Affychip result looks like

An oligonucleotide array (Affychip®) hybridized to biotin-labelled cRNA and revealed with fluorochrome-conjugated avidin



Bead-arrays

This kind of array was patented by Illumina (San Diego, CA, USA)

Last generation:
Bead-arrays®

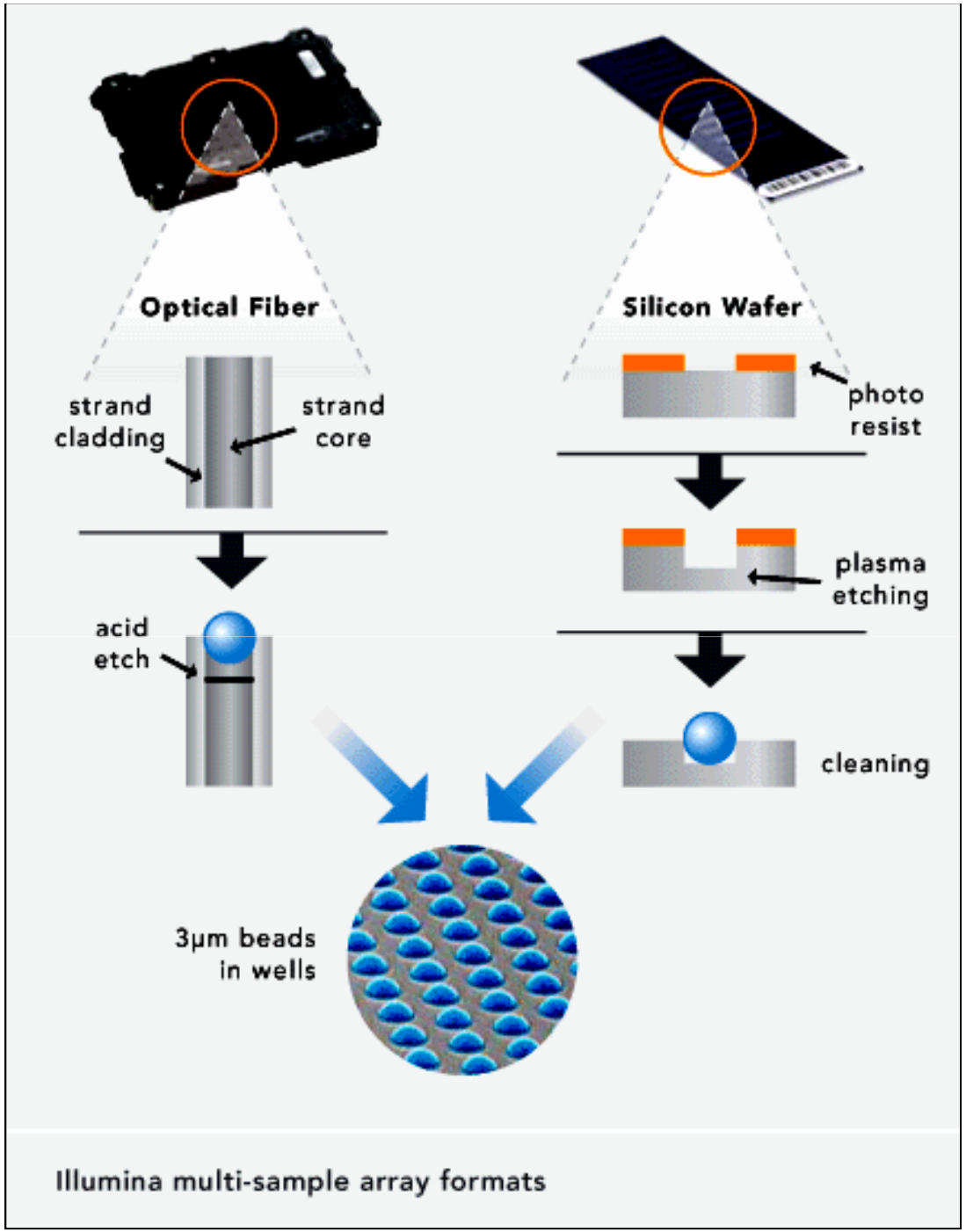
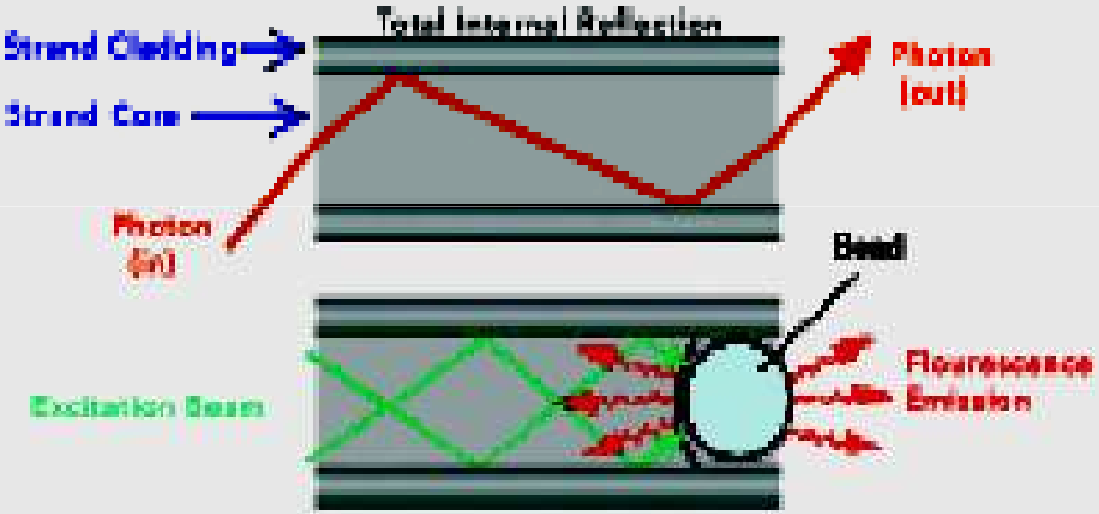
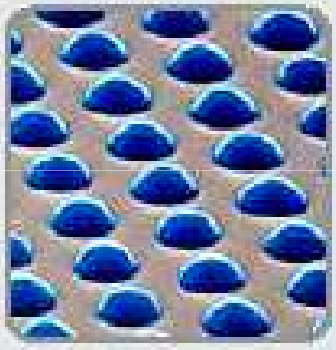


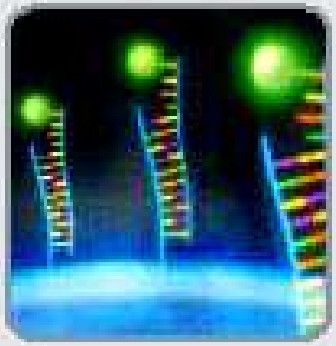
**FIGURE 1: Fiber-optics form the basic substrate for the Sentrix Array Matrix platform.**



1 Each array cluster contains about 50,000 3-micron beads, or features, assembled in dense geometries.

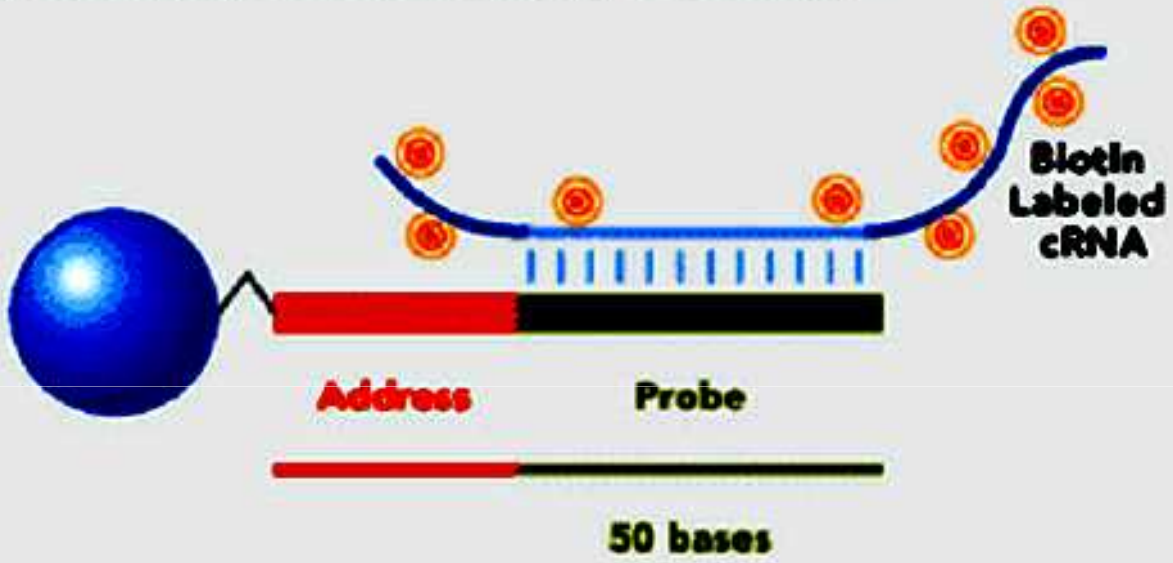


2 Over 1500 probes, or bead types, at >30x average feature redundancy, are represented in each array cluster.

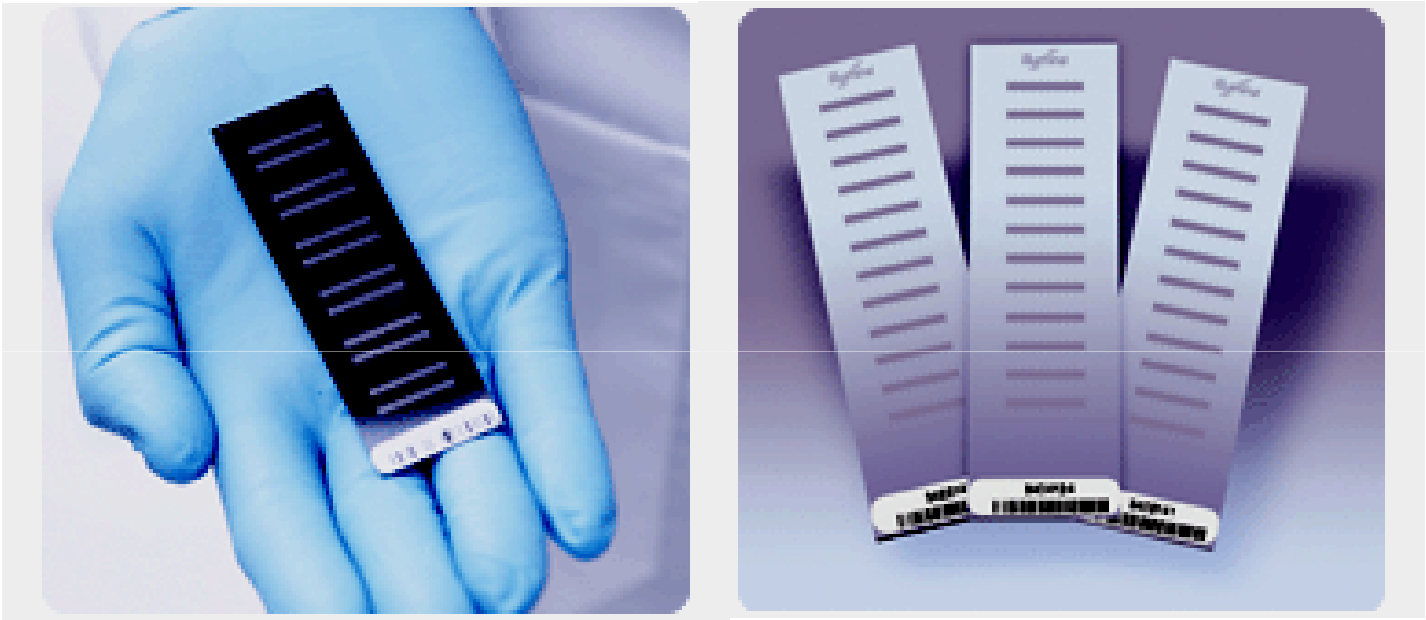


3 Labeled sample targets hybridize to capture probes immobilized on the beads.

## Direct Hybridization Assay Overview




**A 50-base gene-specific probe linked to short address. This probe is hybridized to labeled nucleic acid derived from total RNA.**



Reading of microarrays is performed with laser scanners, which allow a quantitation of fluorescence in different channels

Scanners produce a **table of values** that are intensities at each spot. If using double colors, relative fluorescence intensities in two channels are read.

First important difference to know is therefore:

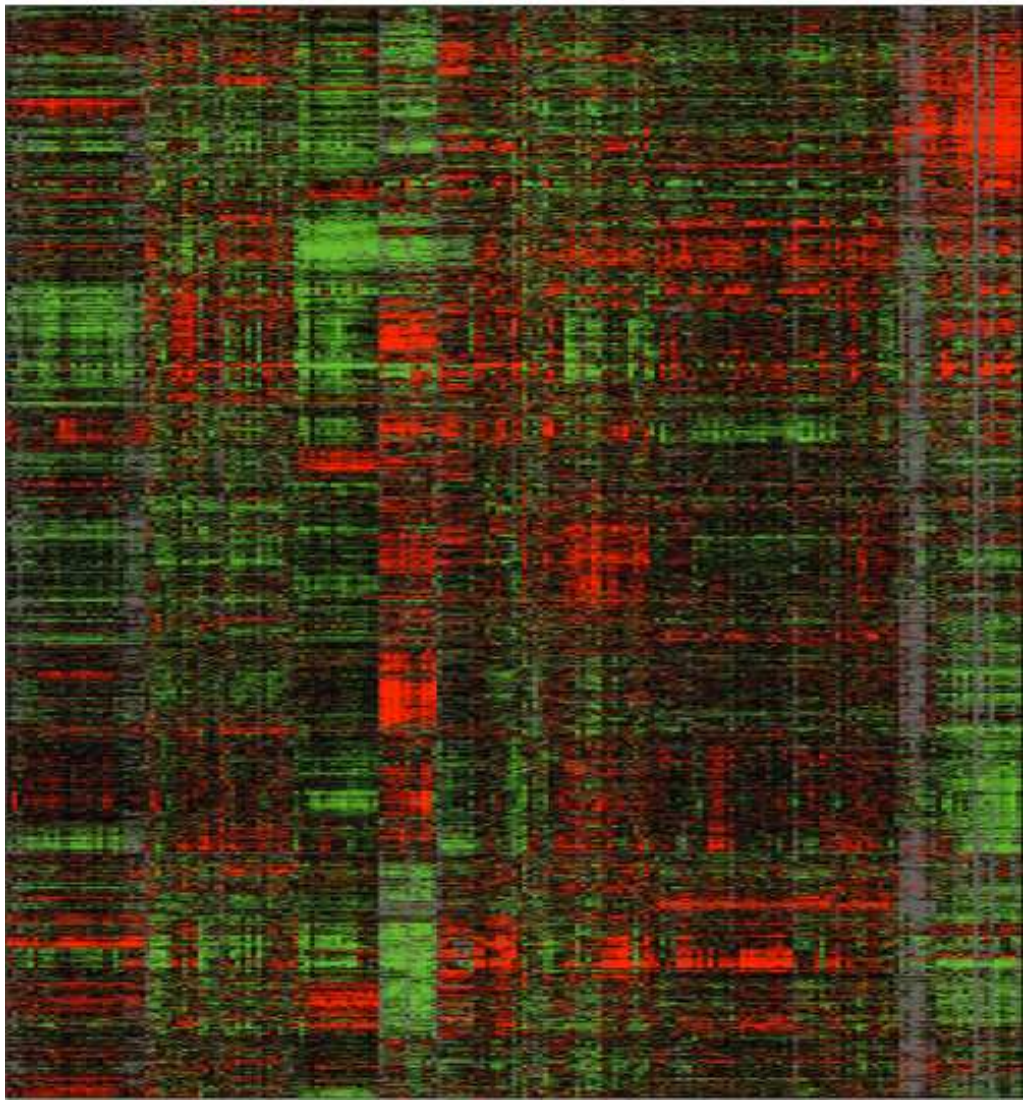
Absolute *versus* relative measurement

Using hybridization reaction as an absolute measurement of RNA requires that the amount of probe on each spot be uniform and reproducible.

This is a requirement fullfilled by Affymetrix arrays and by the latest generation of in situ synthesized long-oligo arrays and bead arrays.

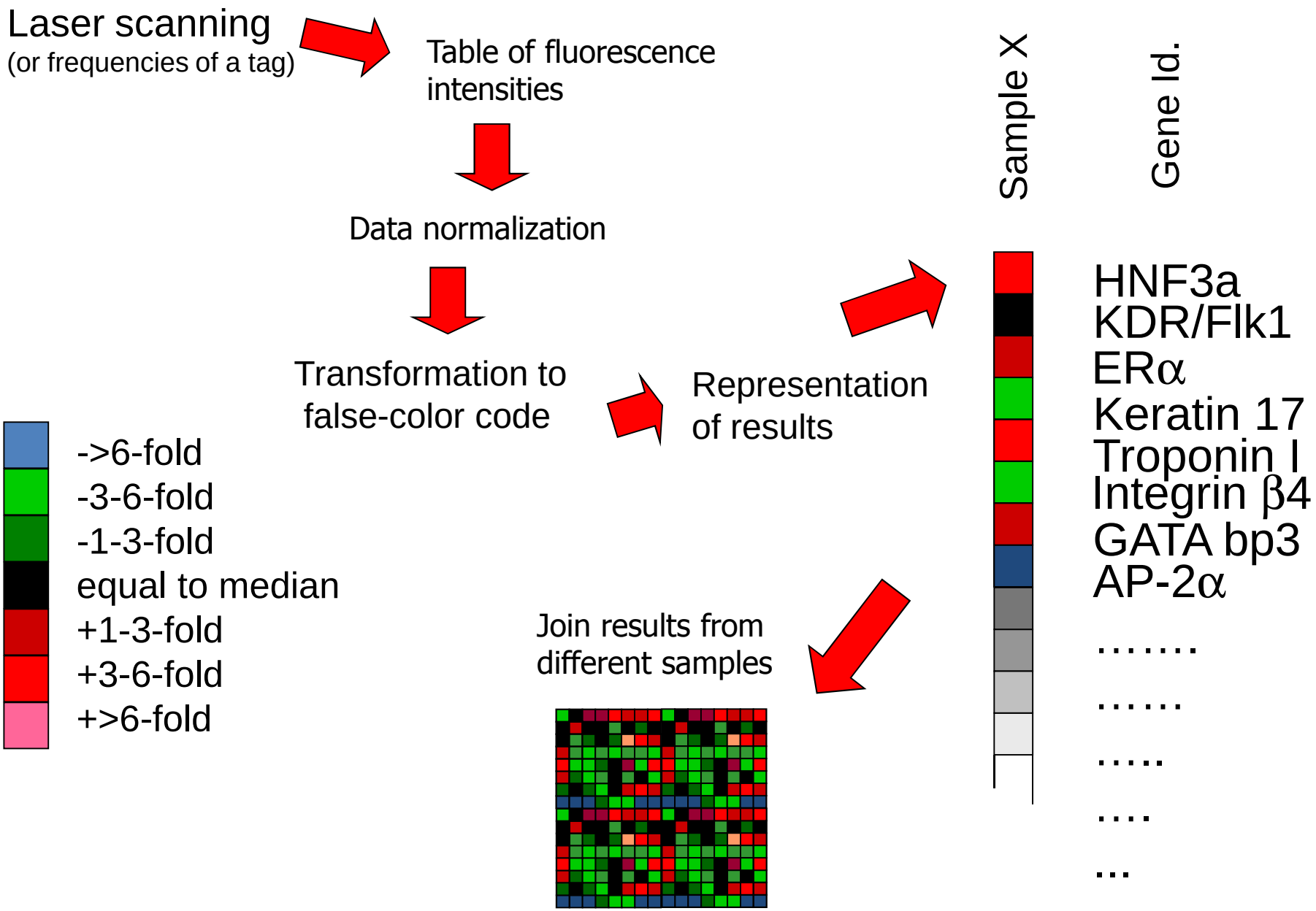
In all the other cases, the amount of probe / spot is variable and unassessable, so that **relative** measurements are necessary.

Each column is a “sample”

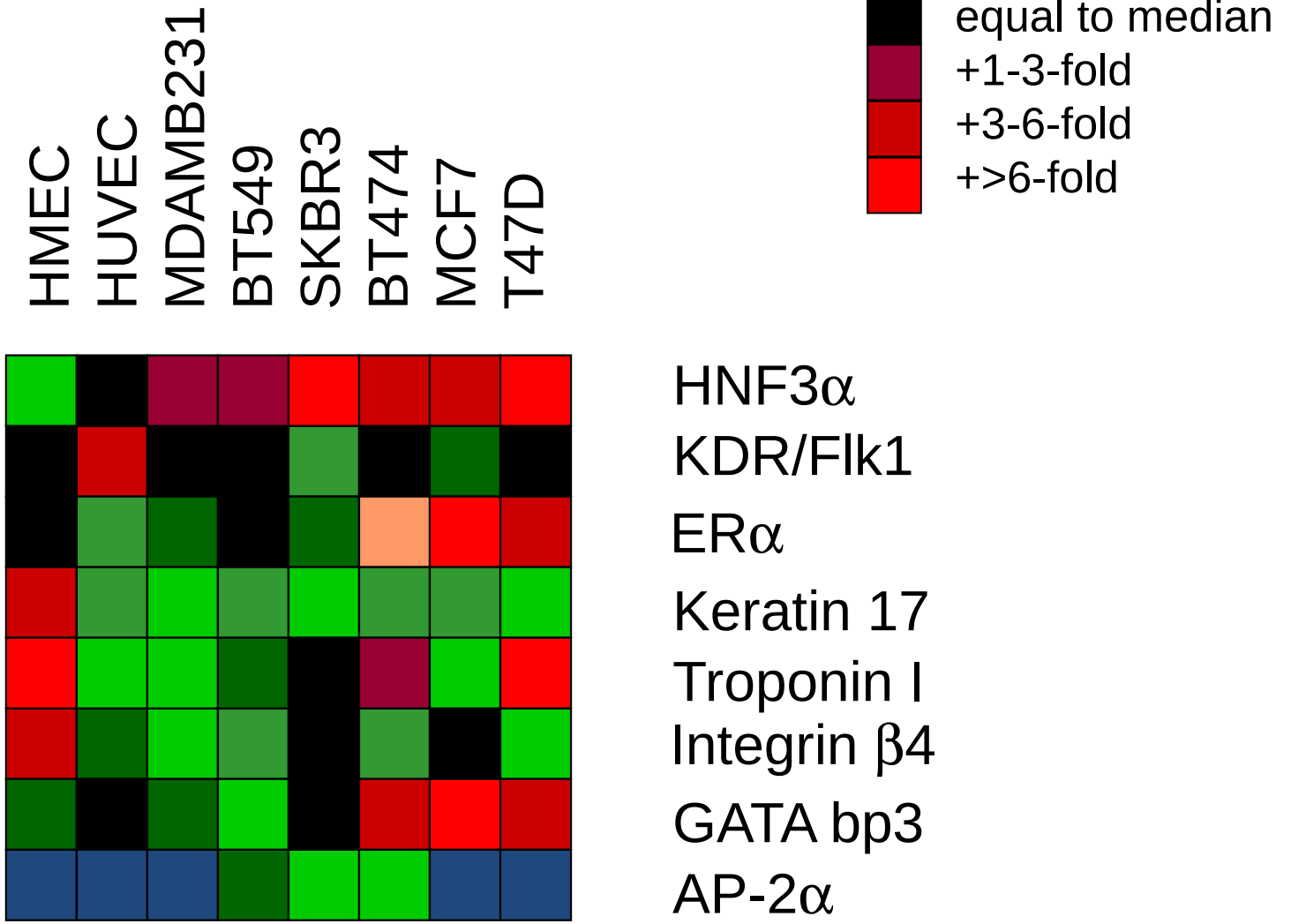


Gene expression are represented as “heat maps”

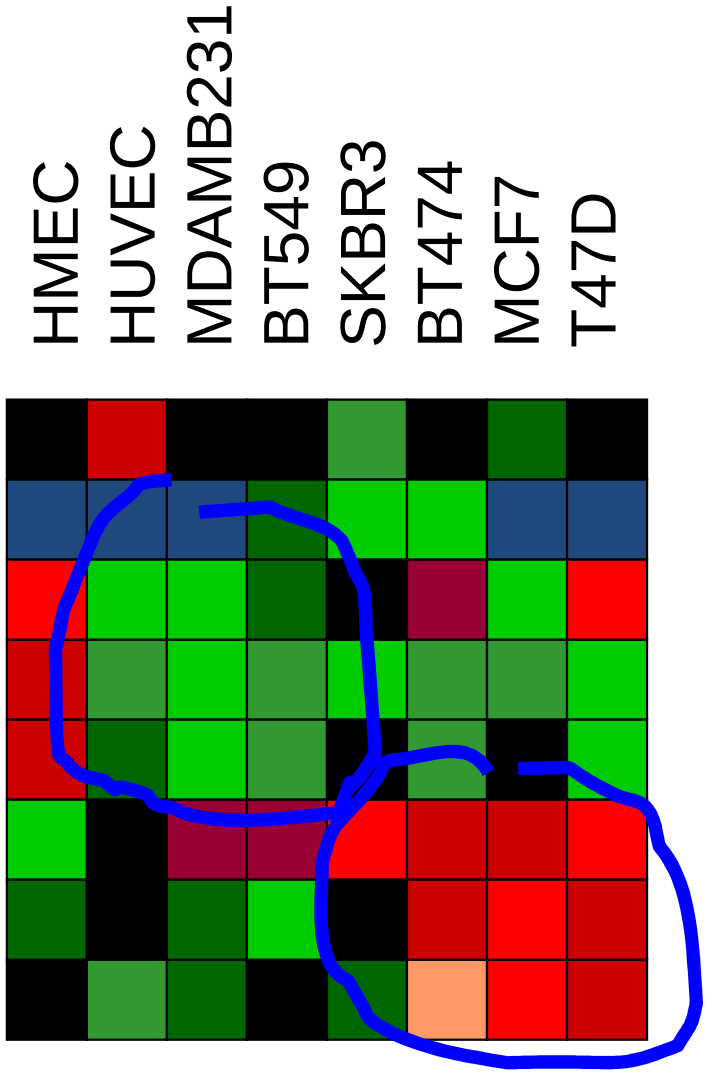
Each row is a “gene”
(better: a “probe”)

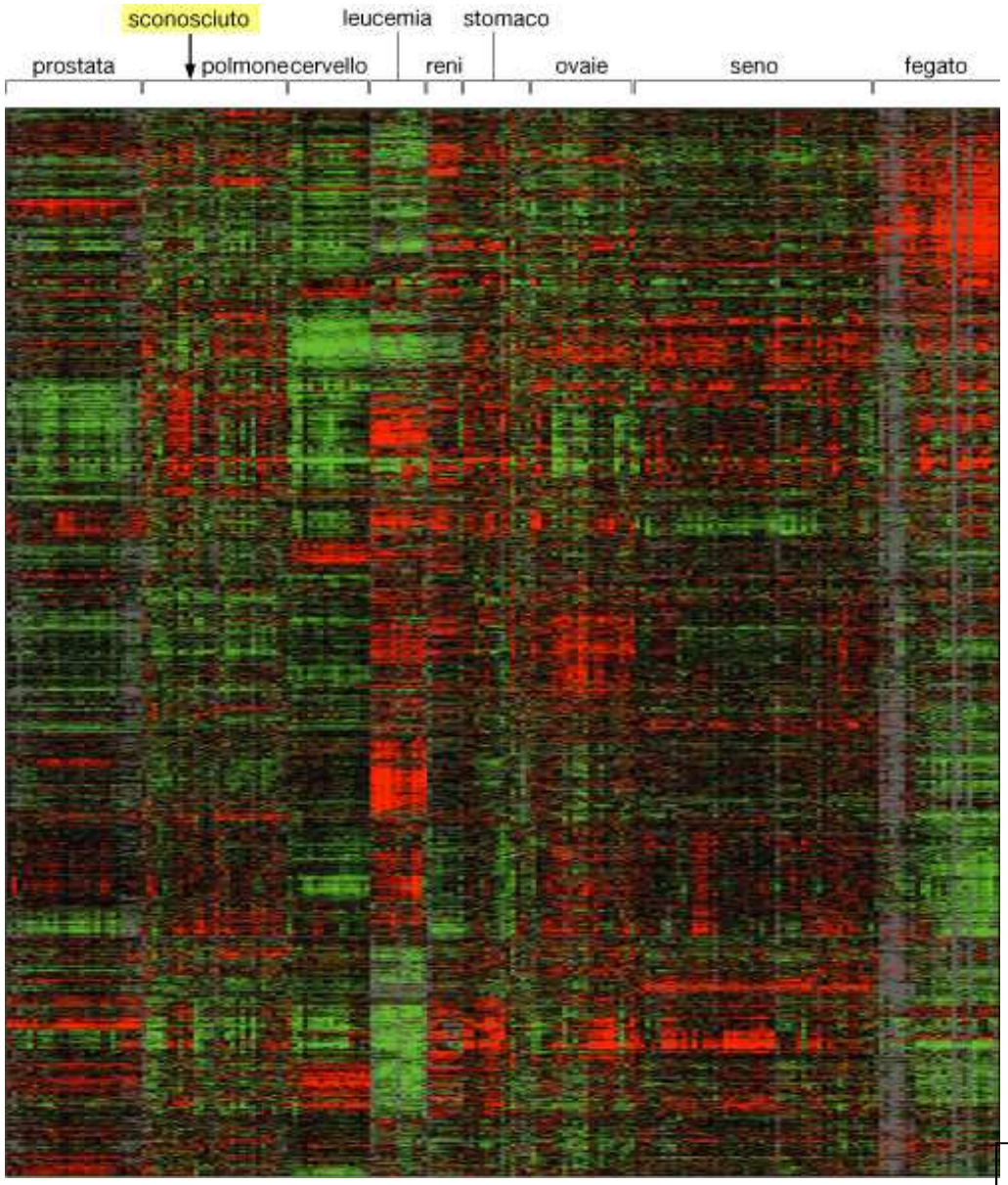

Laser scanning
(or frequencies of a tag)
Table of fluorescence intensities
Data normalization
Transformation to false-color code
Representation of results
Join results from different samples
->6-fold
-3-6-fold
-1-3-fold
equal to median
+1-3-fold
+3-6-fold
+>6-fold
Sample X
Gene Id.
HNF3a
KDR/Flk1
ERα
Keratin 17
Troponin I
Integrin β4
GATA bp3
AP-2α
…….
……
…..
….
…

# Cluster analysis 1



# Cluster analysis 2




Gene expression are represented as “heat maps”

Different expression profiles in human cells of different tissues: 1800 genes probes