Babelomics

Microarray Normalization

March 2011 in Valencia

David Montaner
dmontaner@cipf.es
http://bioinfo.cipf.es/dmontaner
Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)

INB

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

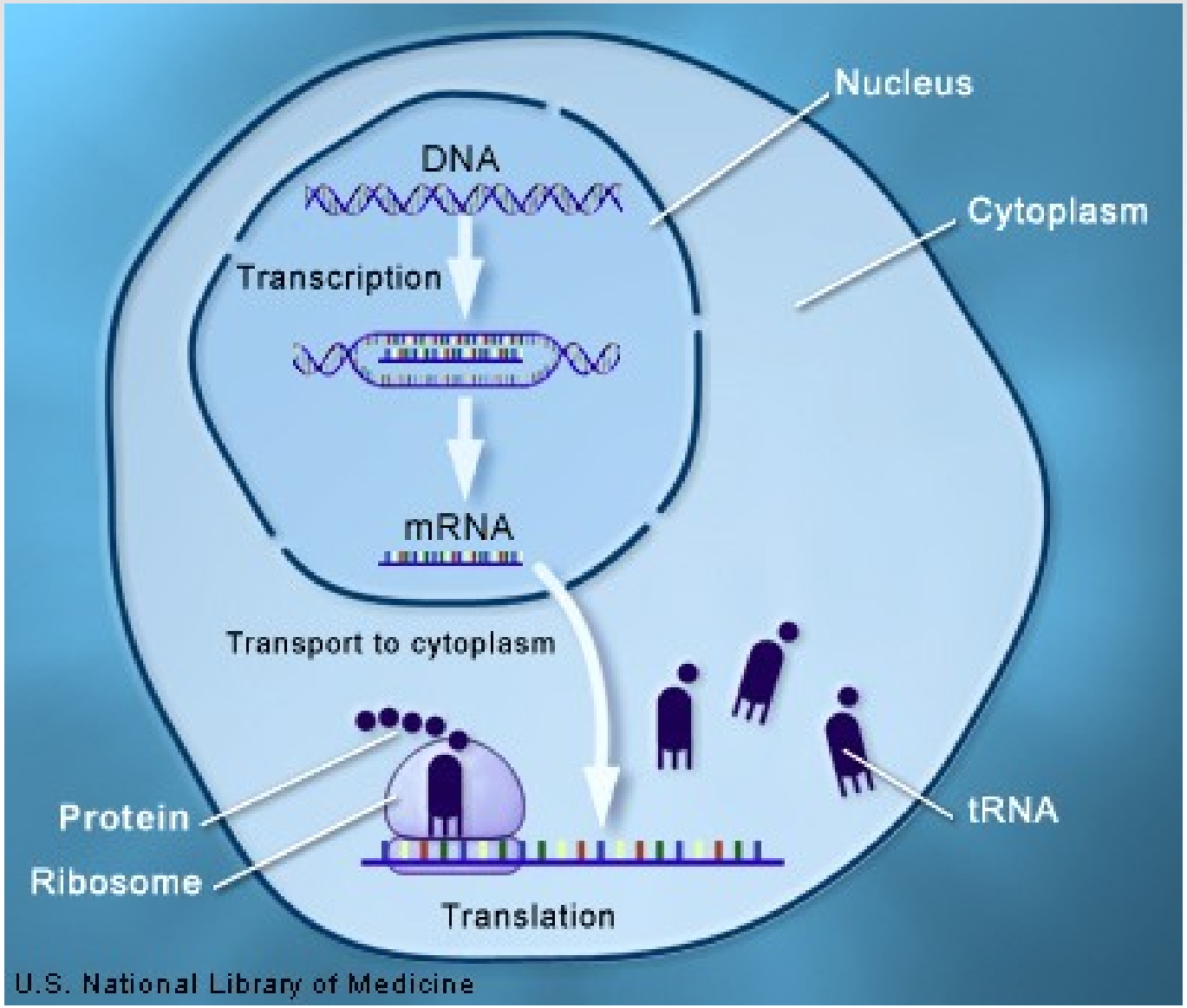
ciberer

DNA Microarrays

- Paradigm of High Throughput Technologies
- Yield concentration measurements for : genes, SNP, exons, mRNA ...
- Measure ells in different biological conditions
- In a genomic scale
- Allow us conducting biological experiments

So... How do they work?

Central Dogma of Molecular Biology

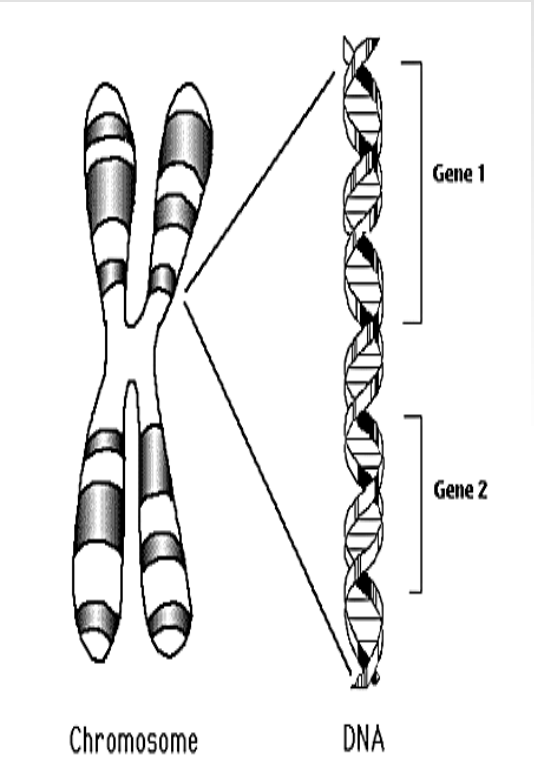


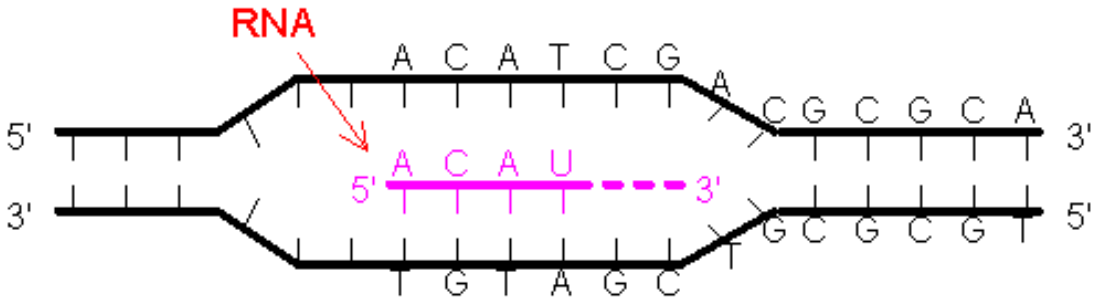
For a cell,
at a particular time,
thousands of mRNA
are created and sent
out of the nucleus to
be translated into
proteins.

Protein concentration
regulates biological
systems.

Human Genome Project

- Determined the sequences that make up human genome.
- Identified the sequence of all genes, transcripts, exons, SNP...
- So the COMPLEMENTARY probes can be build up into the array glass surface.





DNA Microarrays

1.28 cm

1.28 cm

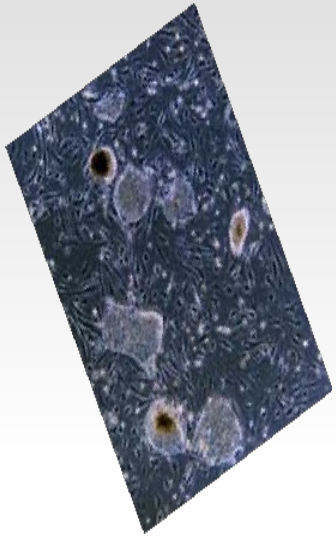
Actual size of GeneChip® array

Millions of DNA strands built up in each location

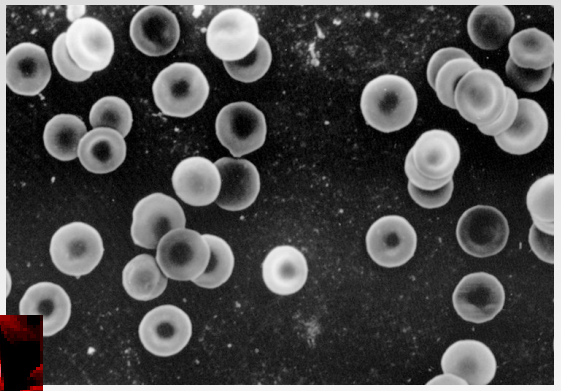
500,000 locations on each GeneChip® array

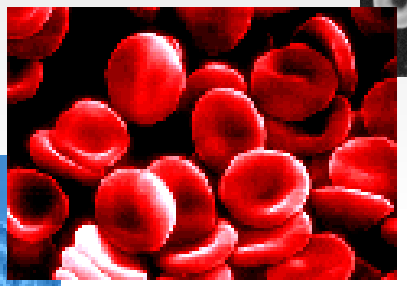
Actual strand = 25 base pairs

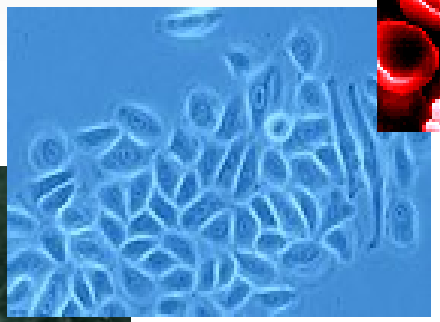
Biological Sample



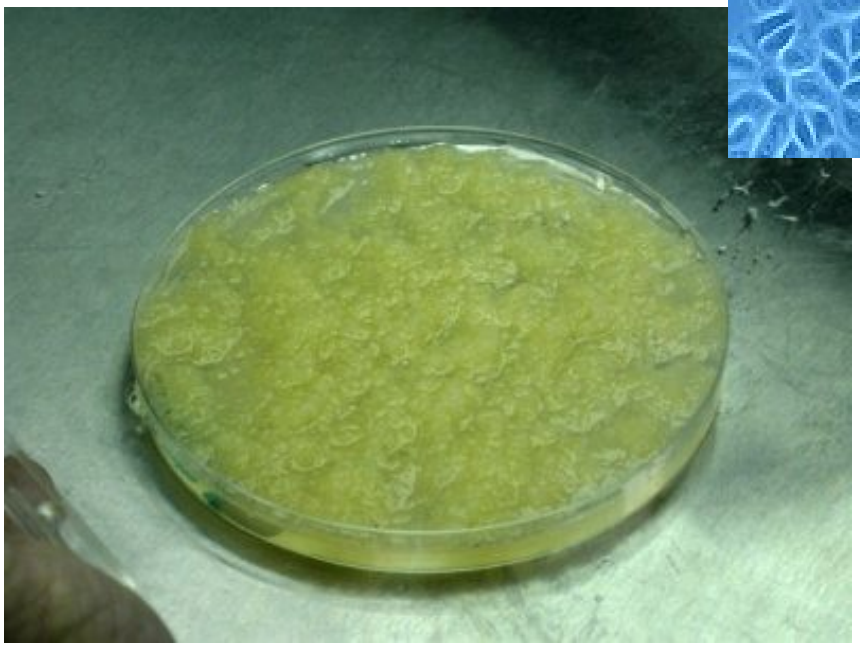
We want to know which genes are expressed under particular biological conditions.







We can extract all mRNA molecules that are being translated within the cells



and provide an expression level indicator of its concentration in the biological sample.

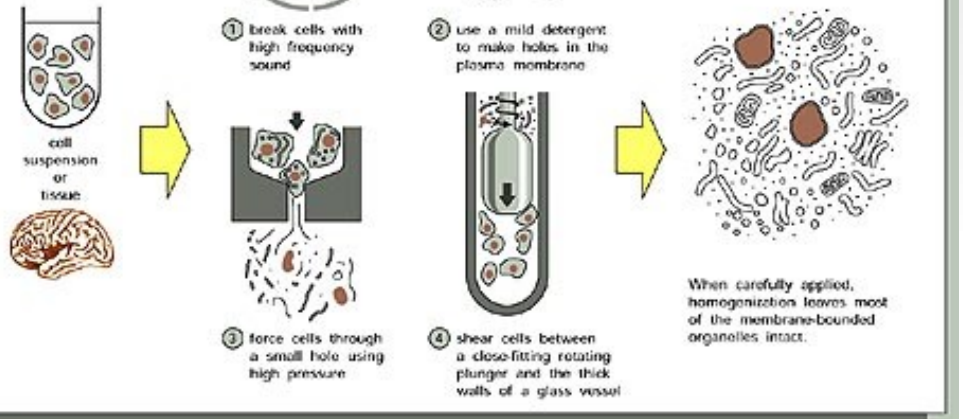
RNA Extraction

BREAKING CELLS AND TISSUES

The first step in the purification of most proteins is to disrupt tissues and cells in a controlled fashion.

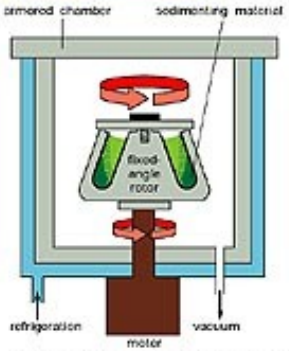
Using gentle mechanical procedures, called homogenization, the plasma membranes of cells can be ruptured so that the cell contents are released. Four commonly used procedures are shown here.

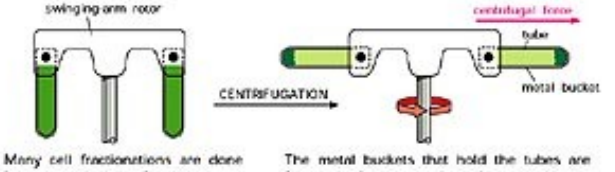
The resulting thick soup (called a homogenate or an extract) contains large and small molecules from the cytosol, such as enzymes, ribosomes, and metabolites, as well as all the membrane-bounded organelles.



When carefully applied, homogenization leaves most of the membrane-bounded organelles intact.

THE CENTRIFUGE





Many cell fractionations are done in a second type of rotor, a swinging-arm rotor.

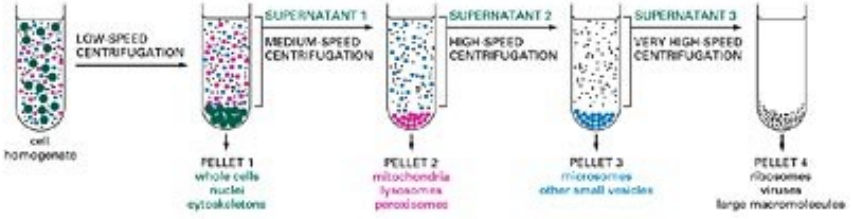
The metal buckets that hold the tubes are free to swing outward as the rotor turns.

Centrifugation is the most widely used procedure to separate the homogenate into different parts, or fractions. The homogenate is placed in test tubes and rotated at high speed in a centrifuge (sometimes called an ultracentrifuge). Present-day ultracentrifuges rotate at speeds up to 100,000 revolutions per minute and produce enormous forces, as high as 600,000 times gravity. At such speeds, centrifuge chambers must be refrigerated and evacuated so that friction does not heat up the homogenate. The centrifuge is surrounded by thick armor plating, since an unbalanced rotor can shatter with an explosive release of energy. A fixed-angle rotor can hold

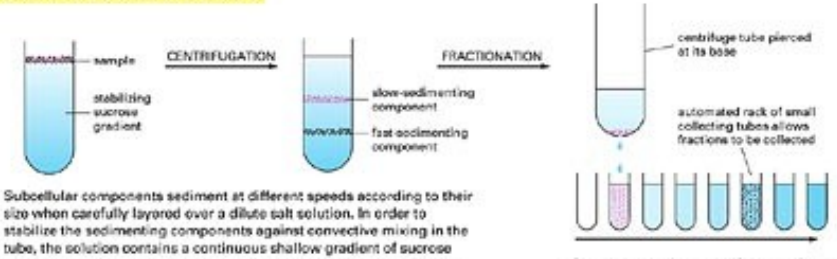
DIFFERENTIAL CENTRIFUGATION

Repeated centrifugation at progressively higher speeds will fractionate cell homogenates into their components.

Centrifugation separates cell components on the basis of size and density. The larger and denser components experience the greatest centrifugal force and move most rapidly. They sediment to form a pellet at the bottom of the tube, while smaller, less dense components remain in suspension above, called the supernatant.



VELOCITY SEDIMENTATION

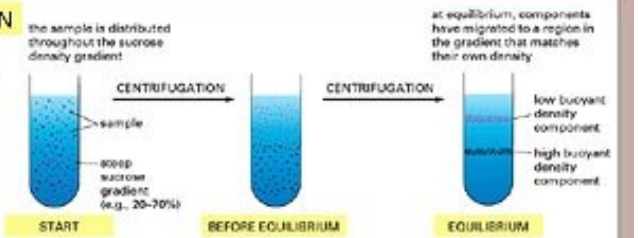


Subcellular components sediment at different speeds according to their size when carefully layered over a dilute salt solution. In order to stabilize the sedimenting components against convective mixing in the tube, the solution contains a continuous shallow gradient of sucrose that increases in concentration toward the bottom of the tube. This is typically 5–20% sucrose. When sedimented through such a dilute sucrose gradient, different cell components separate into distinct bands that can, after an appropriate time, be collected individually.

After an appropriate centrifugation time the bands may be collected, most simply by puncturing the plastic centrifuge tube and collecting drops from the bottom, as shown here.

EQUILIBRIUM SEDIMENTATION

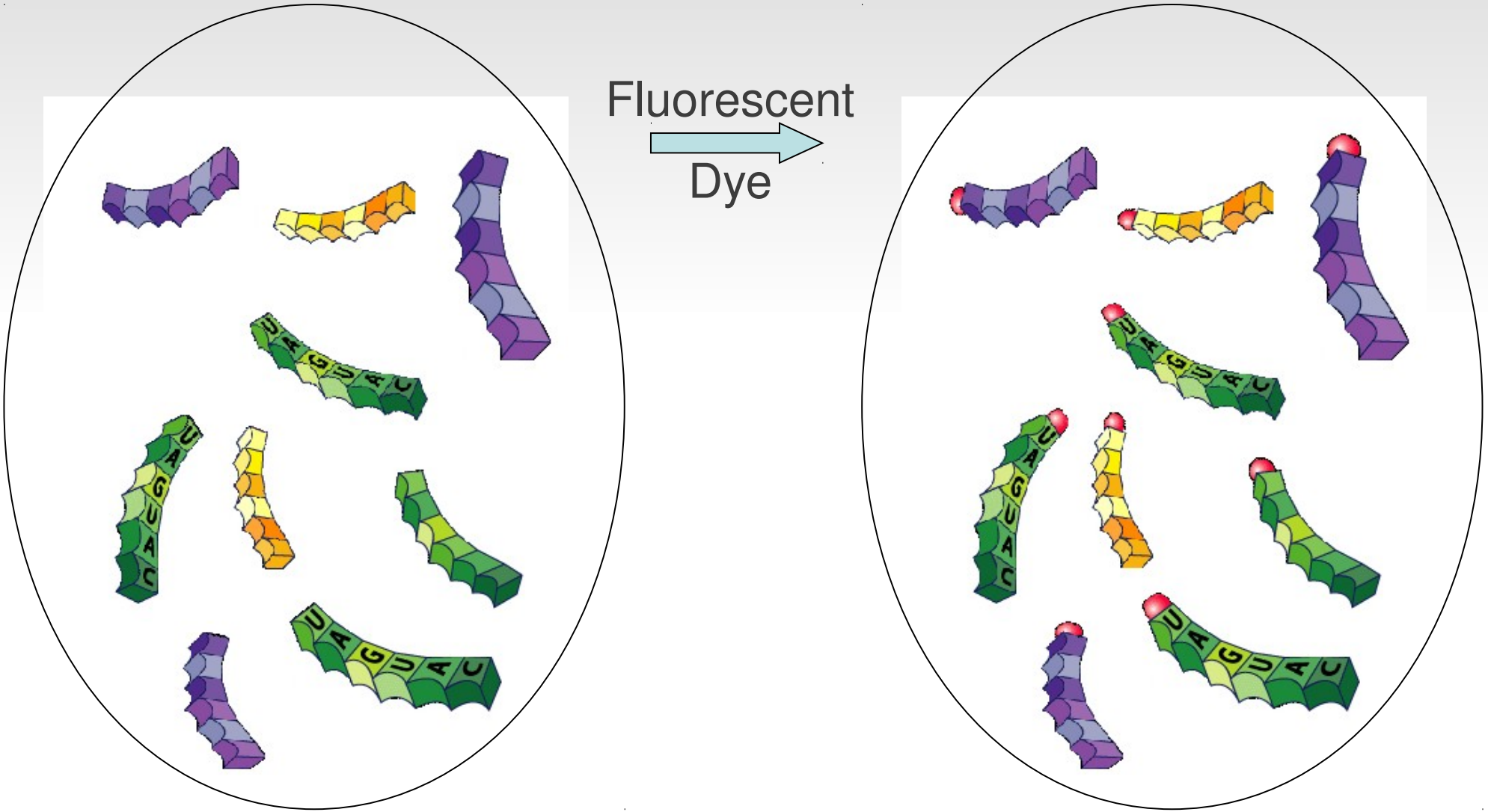
The ultracentrifuge can also be used to separate cellular components on the basis of their buoyant density, independently of their size or shape. The sample is usually either layered on top of, or dispersed within, a steep density gradient that contains a very high concentration of sucrose or cesium chloride. Each subcellular component will move up or down when centrifuged until it reaches a position where its density matches its surroundings and then will move no further. A series of distinct bands will eventually be produced, with those nearest the bottom of the tube containing the components of highest buoyant density. The method is also called density gradient centrifugation.



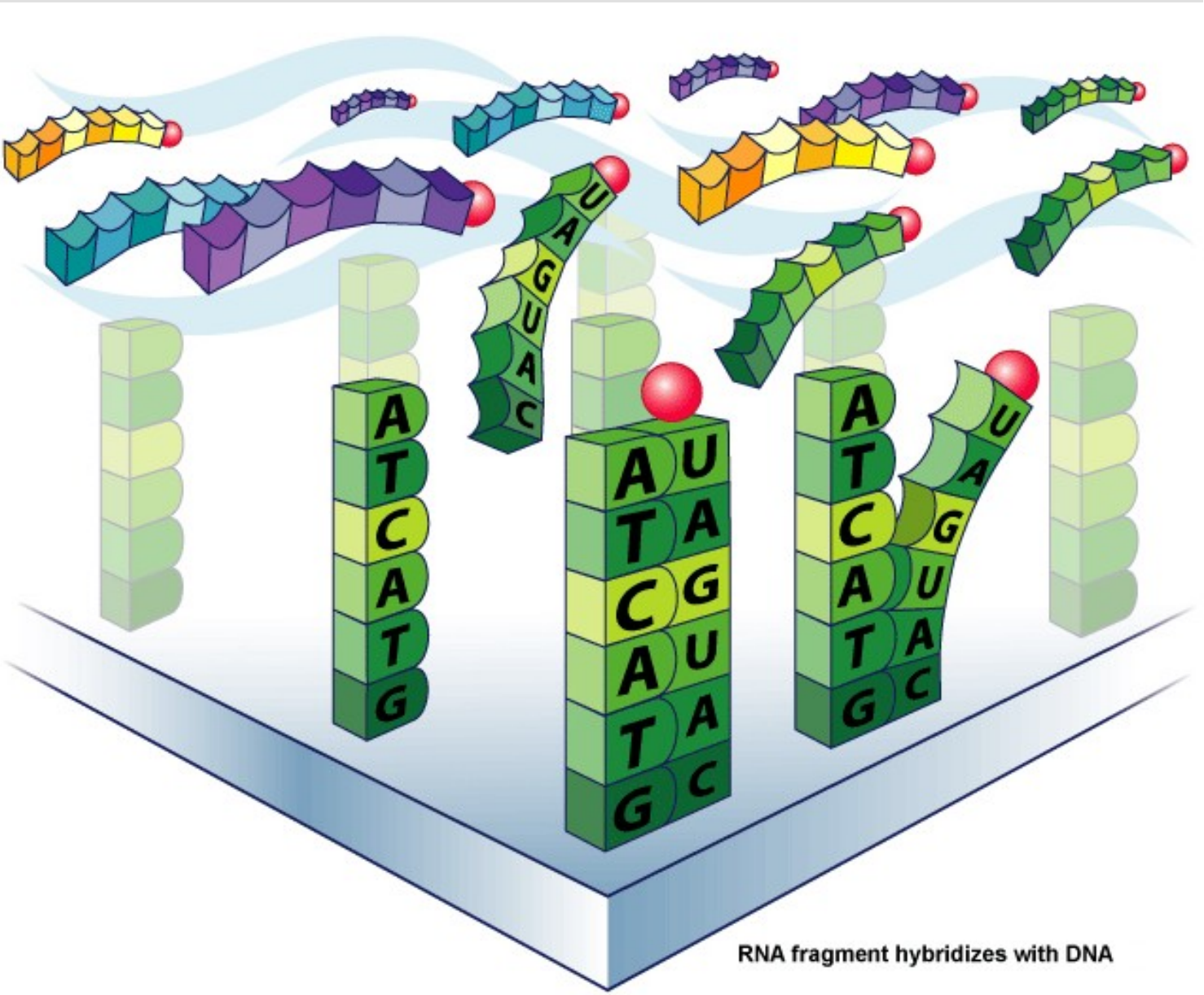
A sucrose gradient is shown here, but denser gradients can be formed with cesium chloride that are particularly useful for separating the nucleic acids (DNA and RNA).

The final bands can be collected from the base of the tube, as shown above.

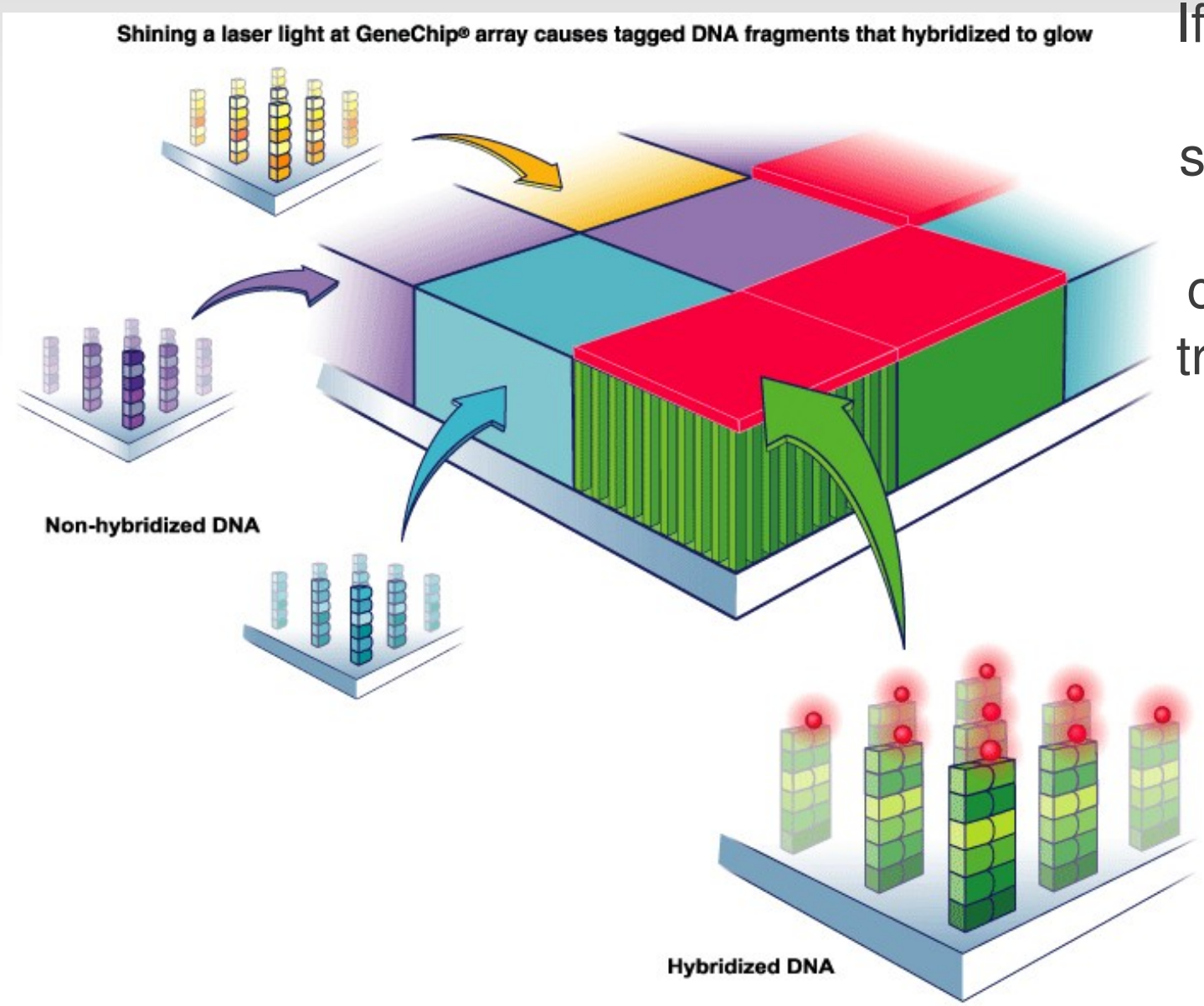
Labelling the Sample



Hybridization



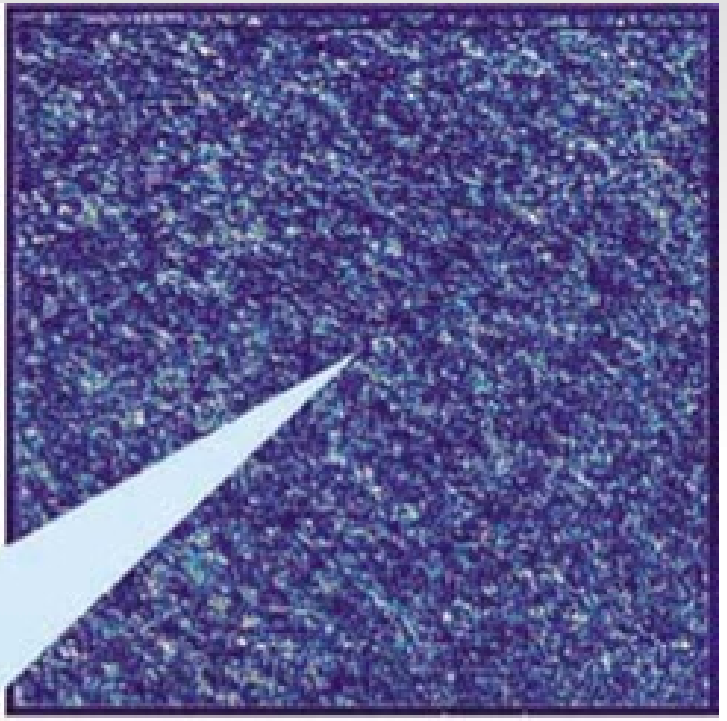
Expression Measurement



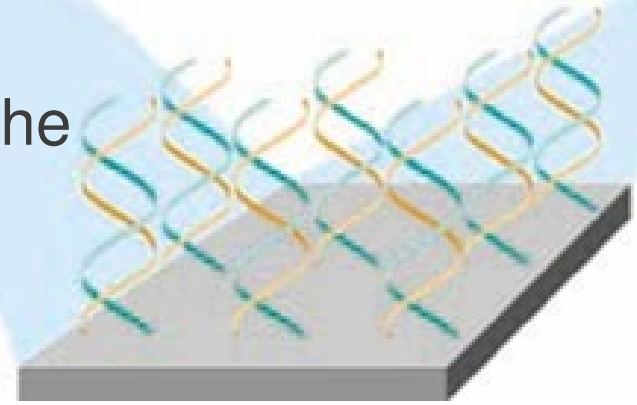
If the fluorescent label is attached to one spot we know that the particular complementary gene transcript was present in our cell sample.

The greater the fluorescence the greater the concentration of the transcript.

Scanning the Microarray



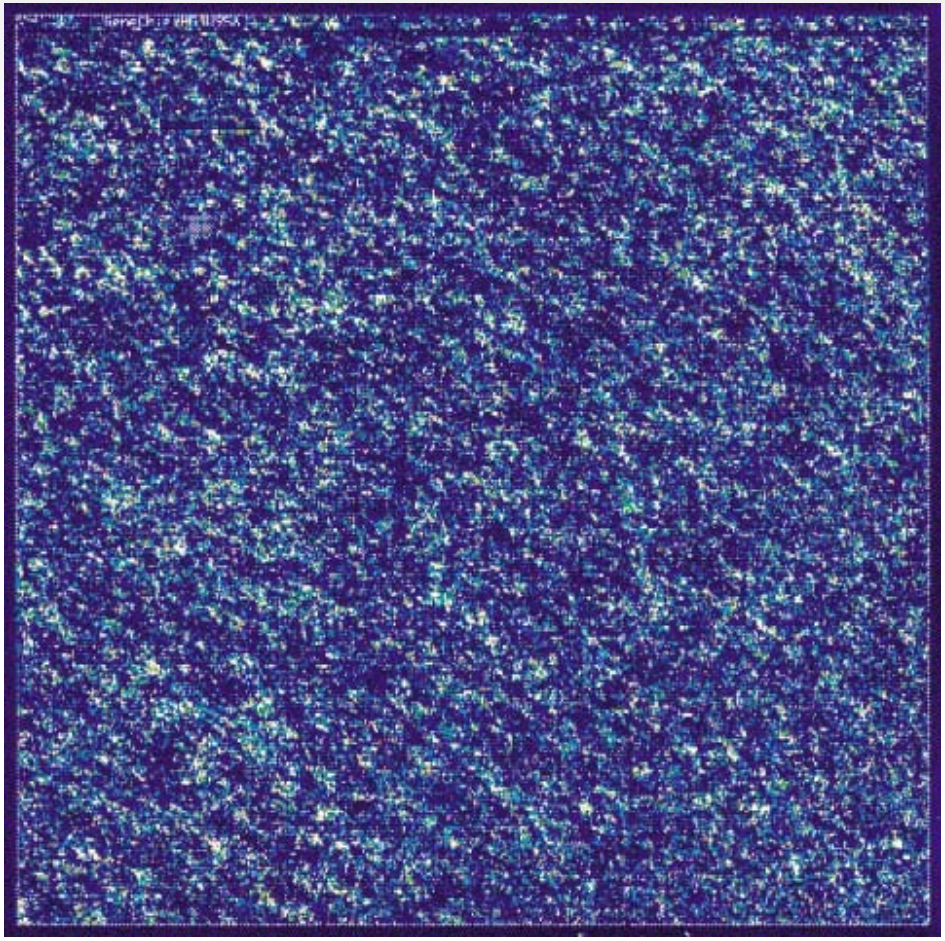
Measuring the intensity of the fluorescence in each spot



We get a measurement of the hybridization in each spot of the microarray

The Data

For each biological sample (individual)

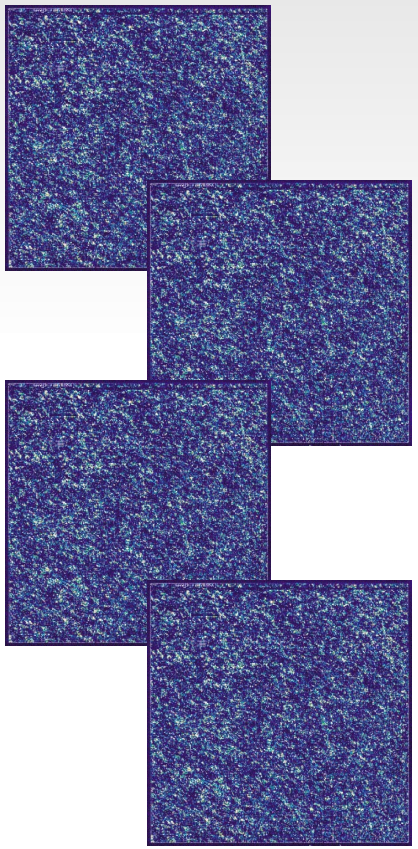


We get intensity measurements for thousands of genetic transcripts.

The measured intensity is used as an indicator of gene expression.

200000_s_at	134.4
200001_at	586.5
200002_at	1868.4
200003_s_at	1232.7
200004_at	1071.6
200005_at	312.8
200006_at	1712.6
200007_at	606.5
200008_s_at	421.9
200009_at	395.6
200010_at	1228.6
200011_s_at	132.5
200012_x_at	2606.3
200013_at	1572.9
200014_s_at	138.7
200015_s_at	124.1
200016_x_at	1058.7
200017_at	889.4
200018_at	3964.2
200019_s_at	1069.9
200020_at	212.1
200021_at	1018.1
200022_at	1254.8
200023_s_at	1202.8
200024_at	2460.6

Several Microarrays

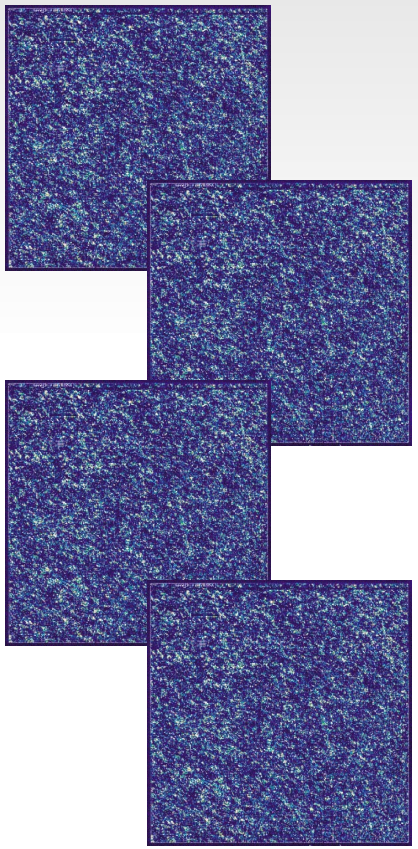


31307_at	5.53
31308_at	7.07
31309_r_at	6.05
31310_at	7.42
31311_at	7.77
31312_at	9.47
31313_at	8.58
31314_at	7.14
31315_at	9.62
31316_at	5.45
31317_r_at	10.27
31318_at	5.7
31319_at	9.25
31320_at	11.5
31321_at	7.79
31322_at	6.98
31323_r_at	11.18
31324_at	7.97
31325_at	9.53
31326_at	9.67
31327_at	6.48
31328_at	8.92
31329_at	6.11
31330_at	14.44
31331_at	6.3

31307_at	5.66
31308_at	7.14
31309_r_at	5.33
31310_at	7.02
31311_at	7.83
31312_at	9.43
31313_at	8.67
31314_at	7.3
31315_at	9.62
31316_at	5.53
31317_r_at	10.75
31318_at	5.53
31319_at	9.19
31320_at	11.51
31321_at	7.91
31322_at	6.93
31323_r_at	10.27
31324_at	8.12
31325_at	9.37
31326_at	10.16
31327_at	6.2
31328_at	9.11
31329_at	5.86
31330_at	14.32
31331_at	6.4

31307_at	5.52
31308_at	7.05
31309_r_at	5.35
31310_at	7.02
31311_at	7.79
31312_at	9.34
31313_at	8.52
31314_at	7.19
31315_at	9.22
31316_at	5.3
31317_r_at	10.41
31318_at	5.59
31319_at	9.24
31320_at	11.35
31321_at	7.76
31322_at	6.91
31323_r_at	10.32
31324_at	8.06
31325_at	9.33
31326_at	9.92
31327_at	6.2
31328_at	8.81
31329_at	5.81
31330_at	14.3
31331_at	6.42

Data Matrix



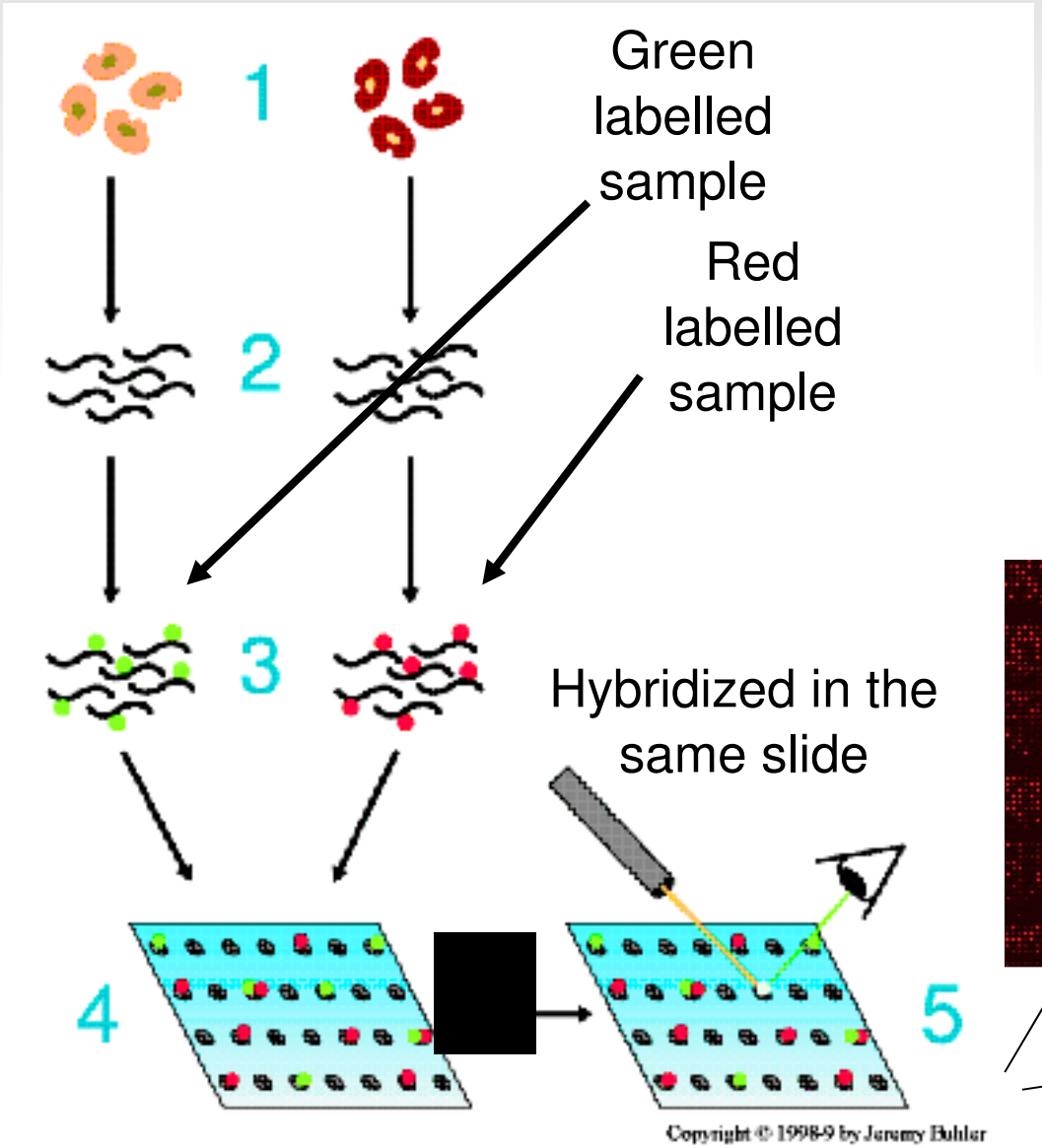
	Array 1	Array 2	Array 3	Array 4	Array 5
Gene 1	5.53	5.66	5.52	5.69	5.62
Gene 2	7.07	7.14	7.05	7.19	7.05
Gene 3	6.05	5.33	5.35	5.07	5.29
Gene 4	7.42	7.02	7.02	7.04	7.22
Gene 5	7.77	7.83	7.79	7.75	7.77
Gene 6	9.47	9.43	9.34	9.37	9.44
Gene 7	8.58	8.67	8.52	8.42	8.52
Gene 8	7.14	7.3	7.19	7.27	7.32
Gene 9	9.62	9.62	9.22	9.44	9.16
Gene 10	5.45	5.53	5.3	5.35	5.44
Gene 11	10.27	10.75	10.41	10.45	10.3
Gene 12	5.7	5.53	5.59	5.58	5.67
Gene 13	9.25	9.19	9.24	8.78	8.86
Gene 14	11.5	11.51	11.35	11.36	11.25
Gene 15	7.79	7.91	7.76	7.82	7.74
Gene 16	6.98	6.93	6.91	7.04	6.8

Single channel hybridization.

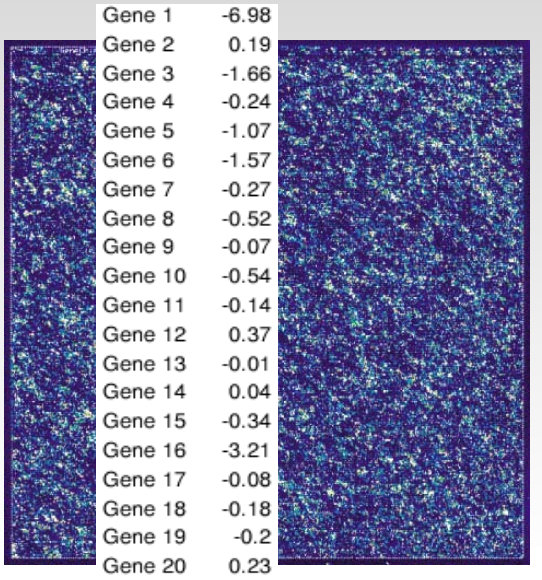
- Each slide is hybridized with a **single biological sample** labelled with a **unique dye**.
- Measured fluorescent intensities ideally represent **molecule abundance** in the sample.
- Most new technologies follow this approach: Affymetrix, Agilent, Codelink.

Competitive hybridization

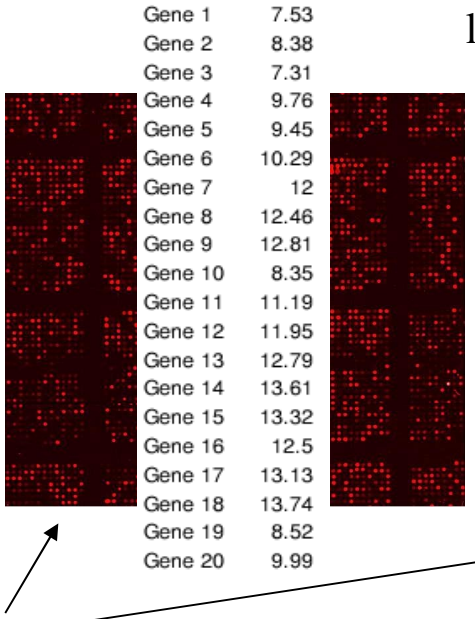
Two different biological samples

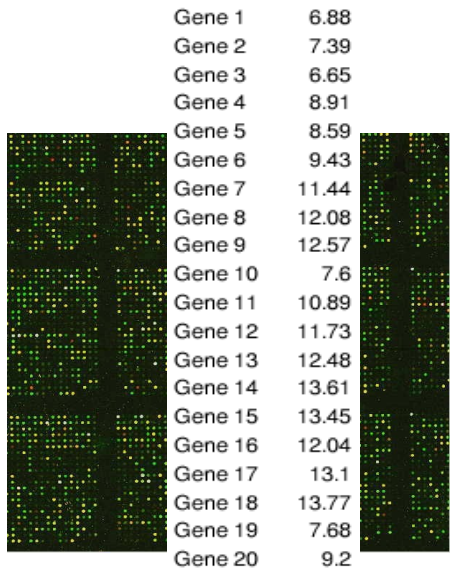


Synthetic image



$$\log_2\left(\frac{\mathrm{Re}\,d}{Green}\right)$$





Competitive hybridization

- Each slide is hybridized with a **two** biological samples each labelled with a different dye.

- Log ratios of the two colour intensities ideally **represent the relative abundance** of the transcripts in one sample compared to the transcripts in the other one.

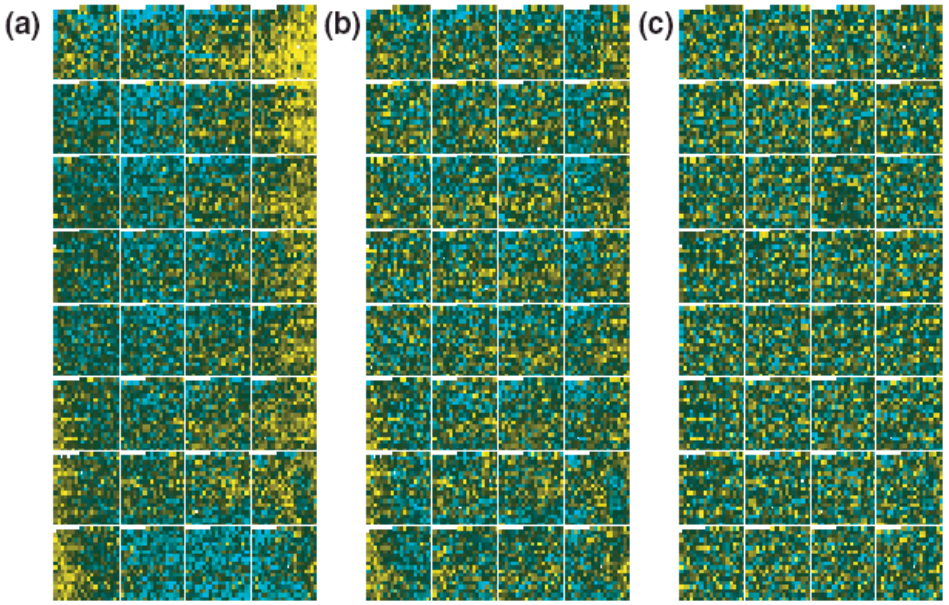
A Noisy Process

+ intensity ⇨ **+ hybridization** ⇨ **+ DNA/RNA concentration in the sample**

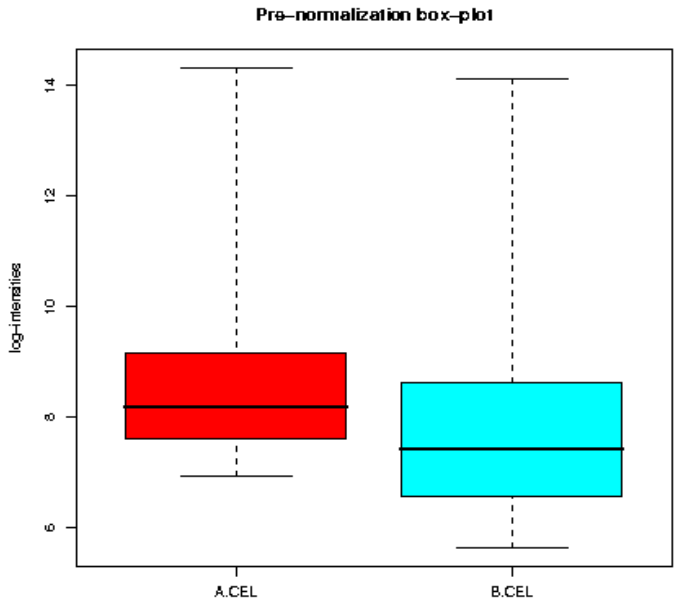
But there is always

- **noise** from technical irregularities
- that produces **signal effects** not due to biological reasons

Background effects

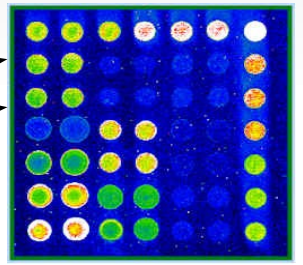


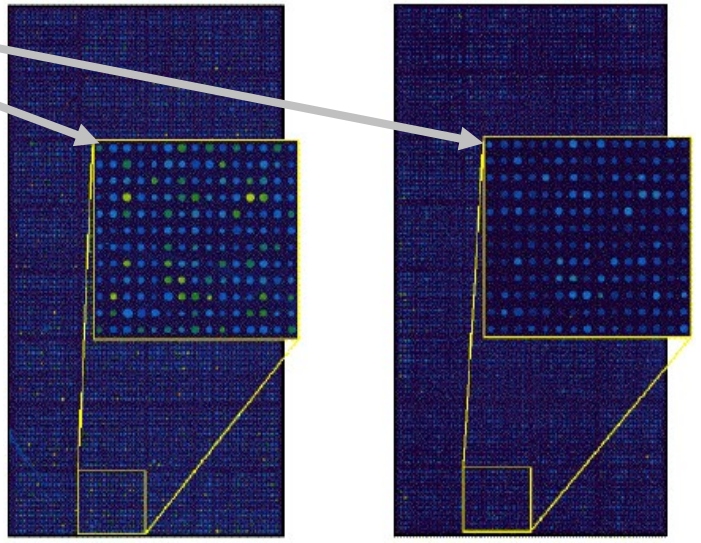
Scale variability



Non calibrated Measurements

- Each spot in the array is measured in its own scale:
  - with different *origin* of the measurements
  - with different *unit* of change
- Problems may arise when comparing:
  - spots within the same array
  - the same spot between arrays





Objective

- Achieve a measurement scale such that:
  - has the same **origin** (zero or other) for all spots
  - uses the same **unit** for all spots and microarrays
  - has a linear relationship with the DNA/RNA biological
  - has good statistical properties (good for future analyses)

- Deal with the particular characteristics of each platform and experiment
  - Colour differences
  - Reference sample
  - Summarize information of each gene
  - Affymetrix PM-MM

Hypotheses (biological)

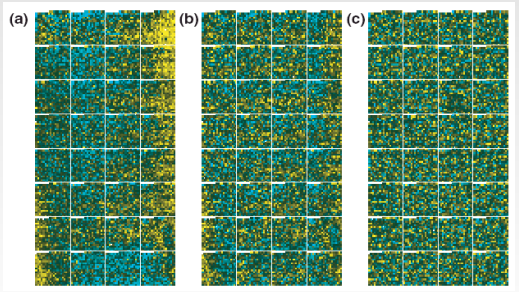
- Most normalization methodologies make two major assumptions about the data.
  - When comparing different samples, only few genes are over-expressed or under-expressed in one relative to the others.
  - The number of genes over-expressed in a condition is similar to the number of genes under-expressed.
- This assumptions should agree with your experimental context.
- They mean that *no overall pattern* should be in your data.

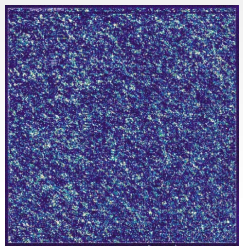
Normalization Steps

- Background Correction
- Within Array Correction - Dye bias correction
- Between Array Scaling - Normalization
- Feature Summarization

Background Correction

- Applied to each intensity channel separately
- Depends on the technology





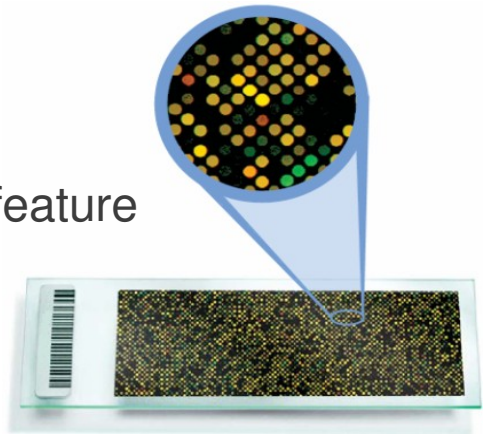
RMA: (robust multiarray average)

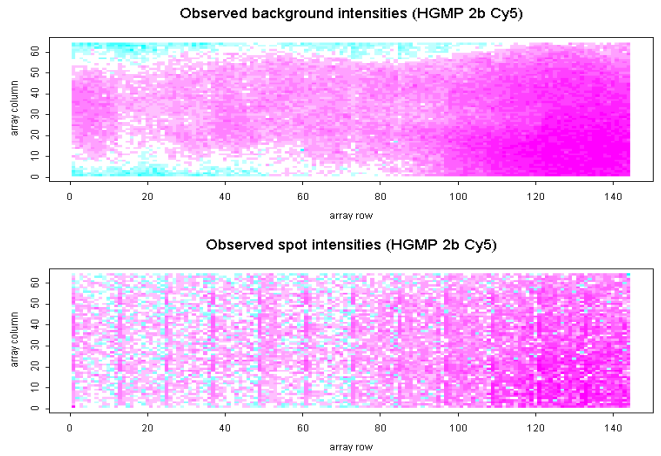
Observed Intensity = Background + Signal

(Normal dist.) (Exp. dist.)

Affymetrix:
- no space between spots
- local background estimation
- MM probes to control crosshybridization

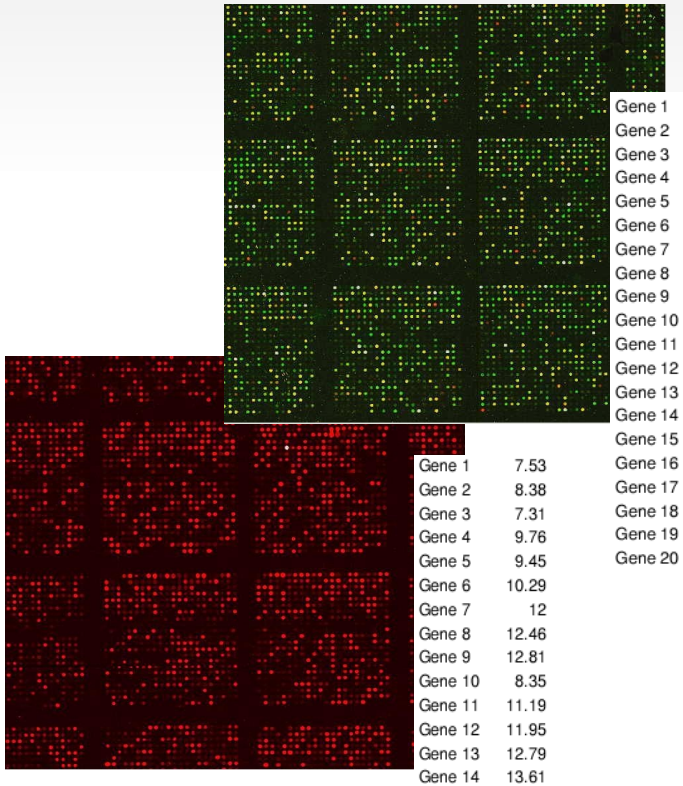
Agilent, GenePix, spotted arrays :
- inter spot gaps
- background estimation for each feature
- Background subtraction (or half)





Within Array Correction

- Just for two color arrays (Agilent, GenePix)
- Provide a unique measurement for each feature **log ratio**

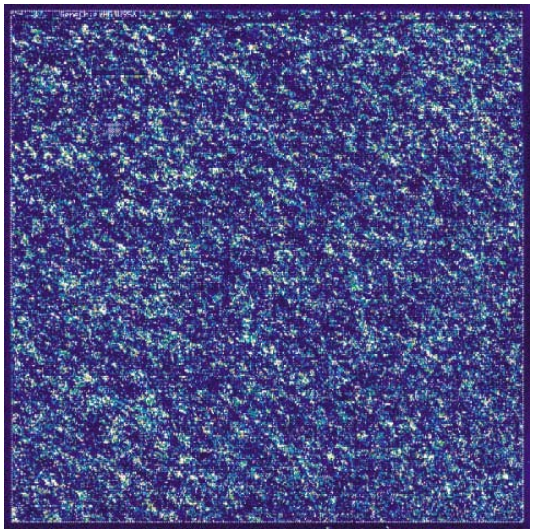


Gene 1	6.88
Gene 2	7.39
Gene 3	6.65
Gene 4	8.91
Gene 5	8.59
Gene 6	9.43
Gene 7	11.44
Gene 8	12.08
Gene 9	12.57
Gene 10	7.6
Gene 11	10.89
Gene 12	11.73
Gene 13	12.48
Gene 14	13.61
Gene 15	13.45
Gene 16	12.04
Gene 17	13.1
Gene 18	13.77
Gene 19	7.68
Gene 20	9.2

Gene 1	7.53
Gene 2	8.38
Gene 3	7.31
Gene 4	9.76
Gene 5	9.45
Gene 6	10.29
Gene 7	12
Gene 8	12.46
Gene 9	12.81
Gene 10	8.35
Gene 11	11.19
Gene 12	11.95
Gene 13	12.79
Gene 14	13.61
Gene 15	13.32
Gene 16	12.5
Gene 17	13.13
Gene 18	13.74
Gene 19	8.52
Gene 20	9.99

$$\log_2\left(\frac{\mathrm{Re}\,d}{Green}\right)$$

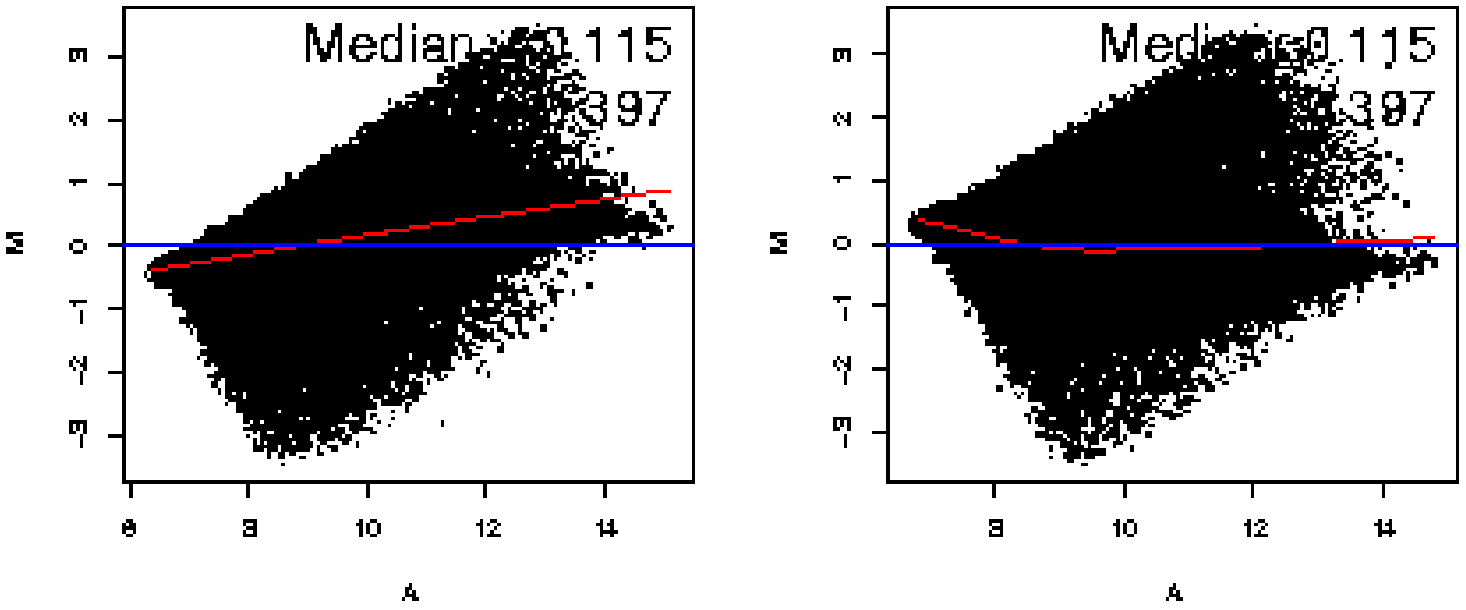
Gene 1	-6.98
Gene 2	0.19
Gene 3	-1.66
Gene 4	-0.24
Gene 5	-1.07
Gene 6	-1.57
Gene 7	-0.27
Gene 8	-0.52
Gene 9	-0.07
Gene 10	-0.54
Gene 11	-0.14
Gene 12	0.37
Gene 13	-0.01
Gene 14	0.04
Gene 15	-0.34
Gene 16	-3.21
Gene 17	-0.08
Gene 18	-0.18
Gene 19	-0.2
Gene 20	0.23



Within Array Correction

- Just for two color arrays (Agilent, GenePix)
- Provide a unique measurement for each feature **log ratio**
- Correct dye-bias. Example: **loess** normalization (MA plots)

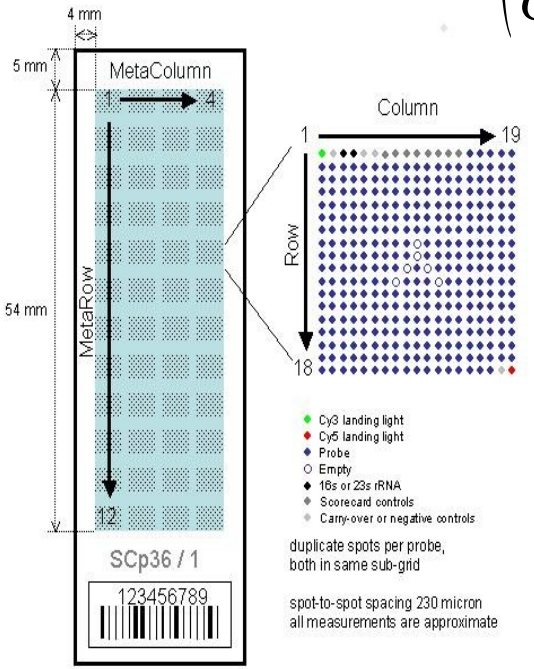
$$M = \log R - \log G = \log\left(\frac{R}{G}\right) \qquad A = \left(\frac{\log R + \log G}{2}\right) = \log\sqrt{R \cdot G}$$



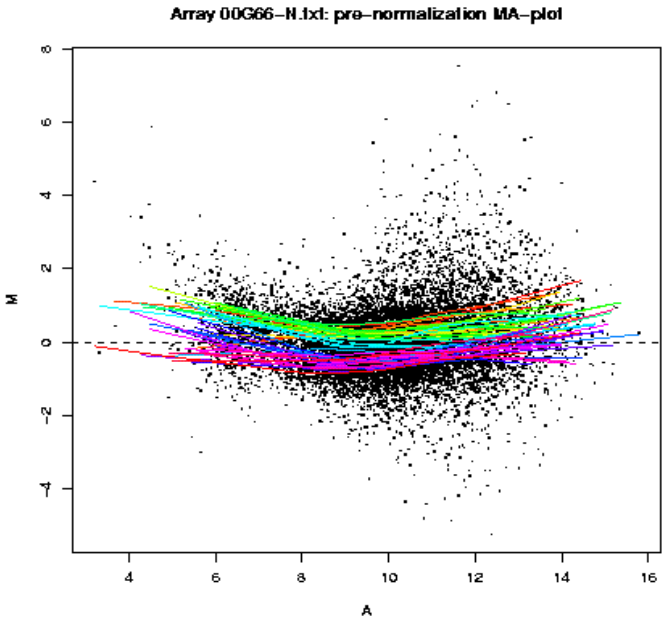
Within Array Correction

- Just for two color arrays (**GenePix**)
- Provide a unique measurement for each feature **log ratio**
- Correct dye-bias. Example: **print tip loess** normalization (MA plots)

$$M = \log R - \log G = \log\left(\frac{R}{G}\right) \qquad A = \left(\frac{\log R + \log G}{2}\right) = \log\sqrt{R \cdot G}$$

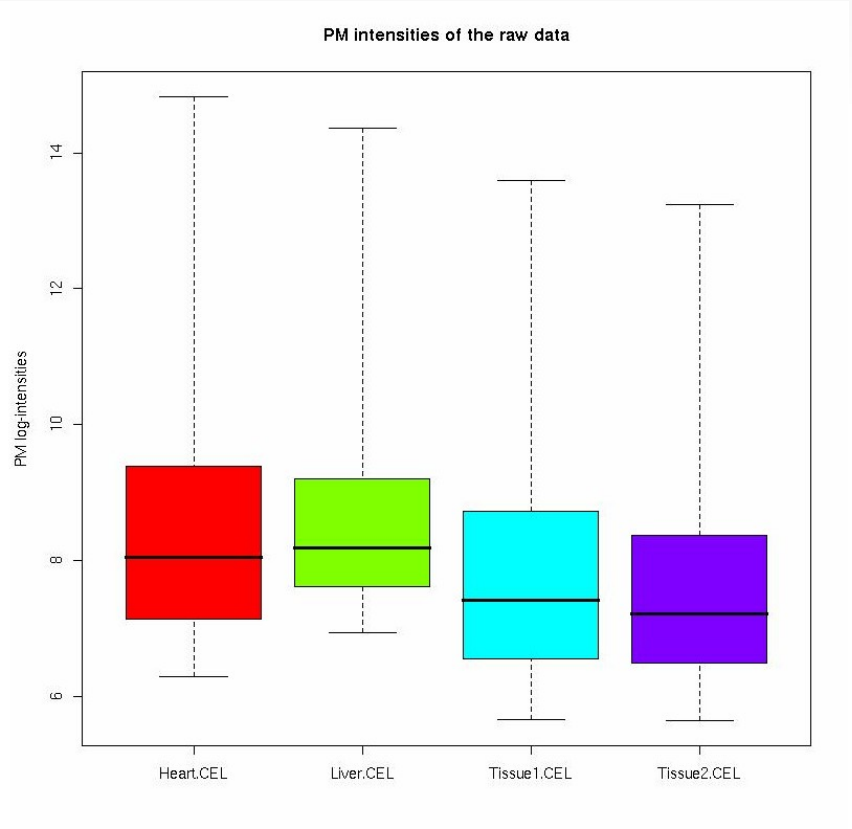


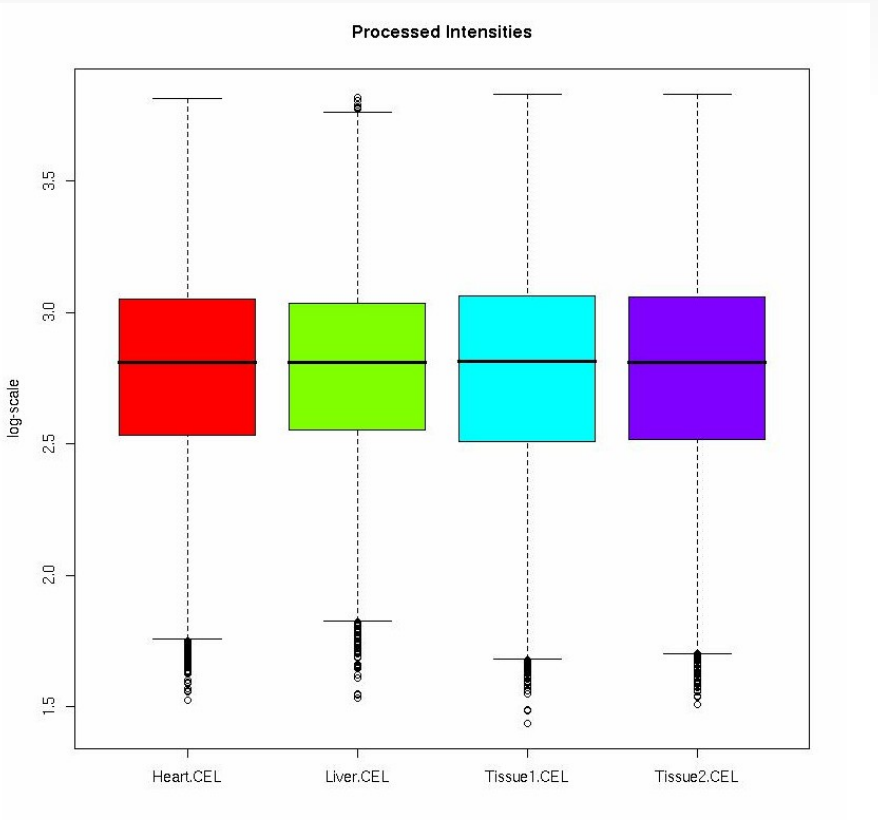
A fit for each print tip block



Between Array Scaling

- Usually just a scale transformation
- Example: **quantile** normalization





Background correction: rma. Normalization: quantiles. PM correction: pmonly. Summary: medianpolish.

Feature Summarization

- Remove control spots and non biological features
- Summarize all probes from the same gene into a unique signal.
- Usually signal averagin.
- There is a choice of the **universe of features** left for the analysis

Steps not always in this order