Advanced genetics - 203.305

Microarray - Hands-on data analysis

Dr. Pierre-Yves Dupont, Postdoctoral Researcher

Computational Biology Research Group

p.y.dupont@massey.ac.nz

Planning

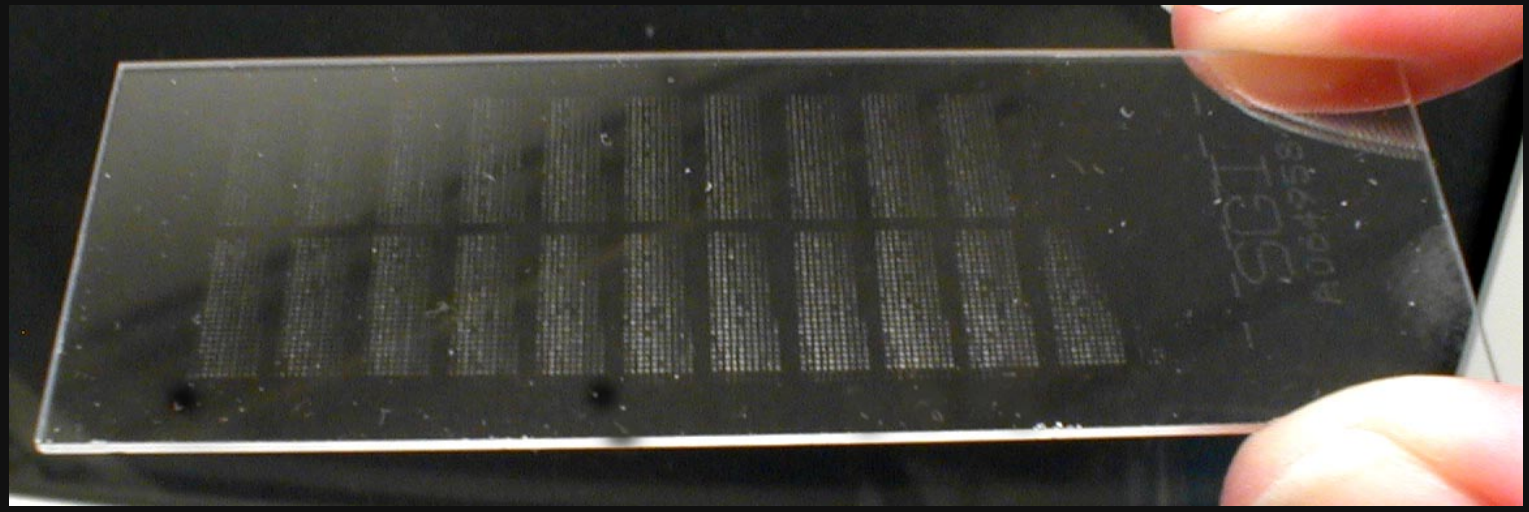
- **25.9.15** – Introduction (lecture), *AHB1.40B*
- **29.9.15** – From raw data to lists of differentially expressed genes (Step by step analysis of a microarray data set using the R language, **3h lab**, *SC5.10*) - **Discussion worksheet due, research proposal due**
- **2.10.15** – Paper discussion, *AHB1.40B*
- **6.10.15** – Biological interpretation of microarray data (Gene ontology analysis using the R language + online research of candidate genes, **3h lab**, *SC5.10*)
- **9.10.15** – **Lab discussion**, *AHB1.40B*

Microarray studies

1. **Indroduction**
2. Microarray technology
3. Statistics
4. MIAME
5. Examples of microarray studies (paper discussion topic and lab topic)

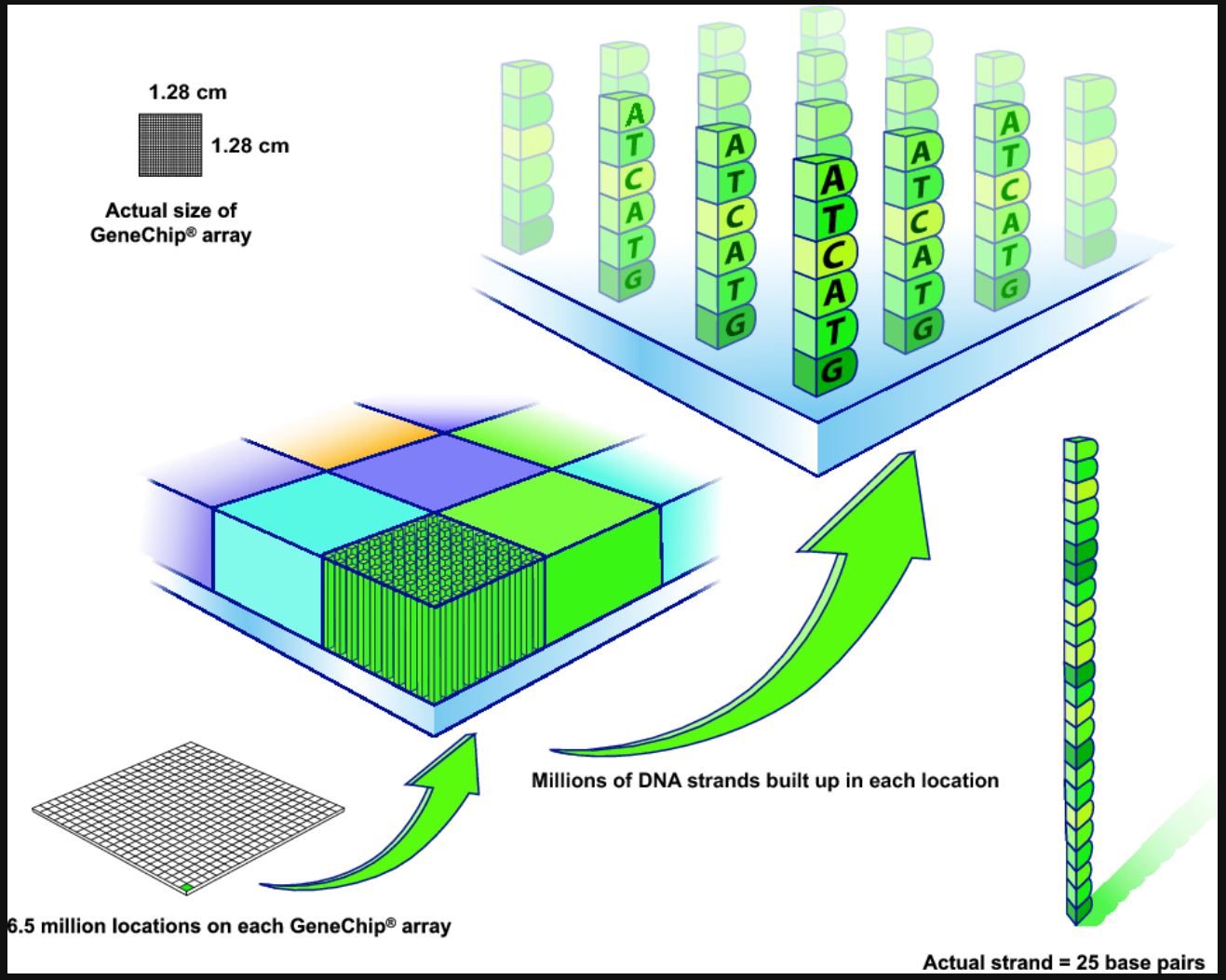
What are microarrays?

*A microarray is a **solid support** (such as a membrane or glass microscope slide) on which **DNA of known sequence** is deposited in a **grid-like array**.*



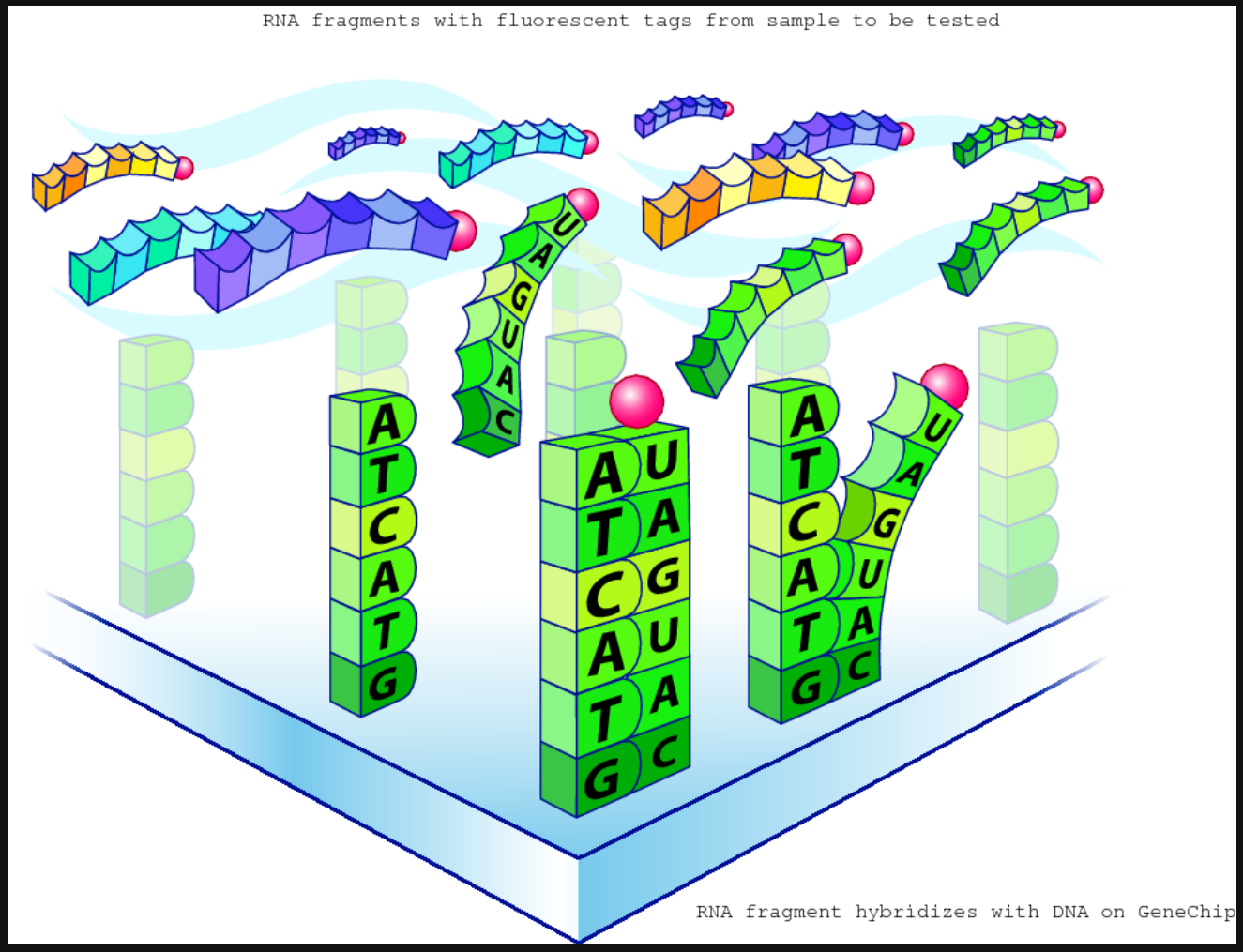
What are microarrays?

DNA microarray



What are microarrays

hybridization and transcriptomics?



The amount of RNA hybridized on each grid location can be measured and is a proxy for the gene expression level

Microarray applications

- **Gene expression analysis**
- Re-sequencing
- SNP-analysis
- DNA-Protein interactions
- Discovery of new transcripts/alternative splice variants

Expression Studies

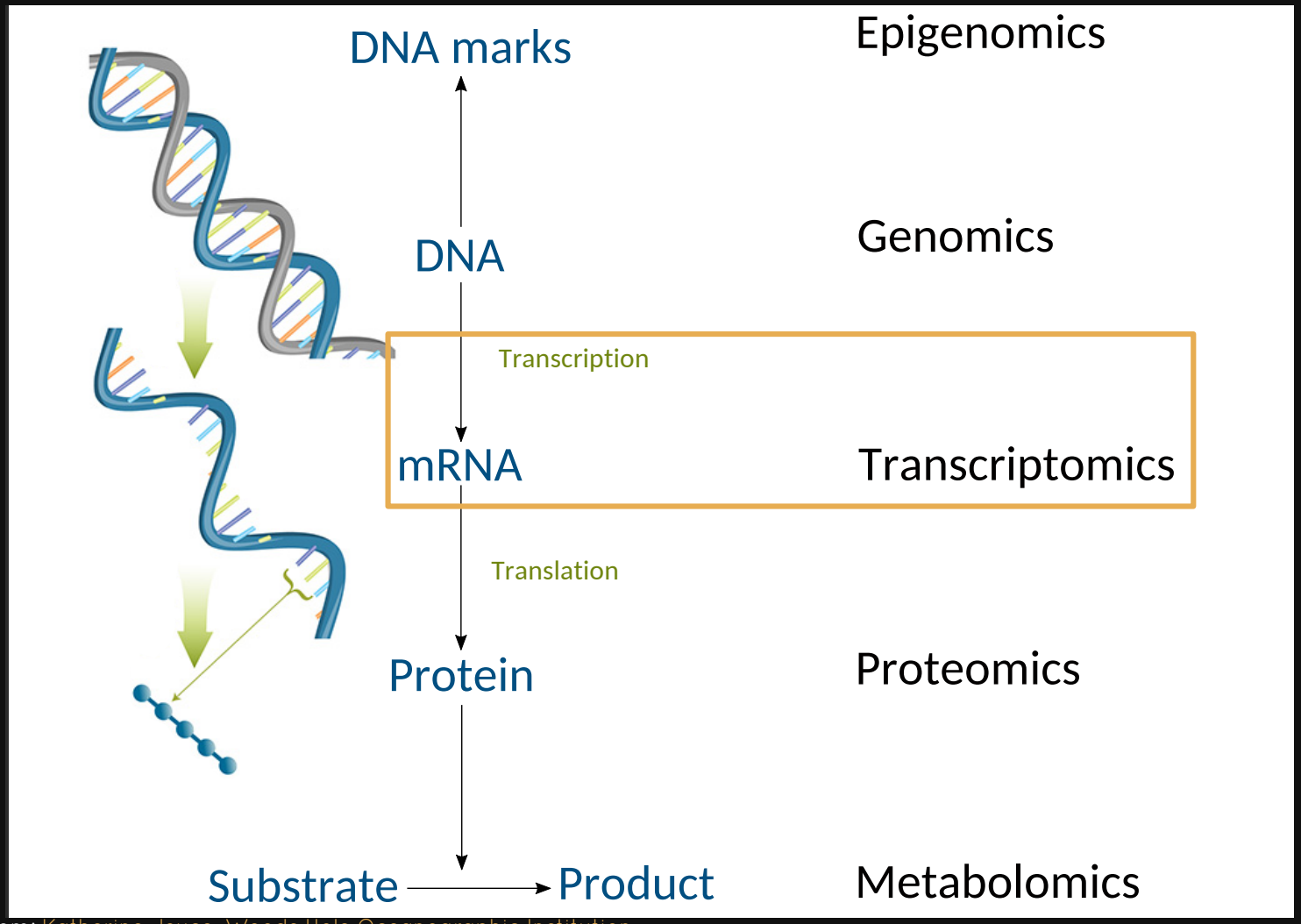


Figure modified from: Katherine Joyce, Woods Hole Oceanographic Institution

Definitions

- **Genome**: entire DNA sequence of an organism
- **Epigenome**: chemical marks of the genome that modify its expression
- **Transcriptome**: all gene transcripts present in a given cell/tissue at a given time ("snapshot")
- **Transcriptomics**: global analysis of gene expression = genome-wide expression profiling

Definitions

- **cDNA**: complementary DNA made from mRNA by the enzyme reverse transcriptase
- **EST**: Expressed Sequence Tag, small pieces of an expressed gene (cDNA)
- **Hybridization**: based on complementary molecules, sequences that are able to base-pair with one another. When two complementary sequences find each other, they will lock together, or hybridize (primer annealing, probe-target binding etc).

Genome-wide expression studies - Medical applications

- **Cancer research**: Cell-cycle monitoring, genetic markers detection
- **Drug development and response**: Treatment-induced expression pattern
- **Diagnosis**: Disease-associated expression patterns

Genome-wide expression studies - Biological applications

- **Development biology**: comparison of different developmental stages
- **Ecology**: interactions between organisms (symbiosis, pathogenicity...) or between organisms and environment (temperature, nutrient...)
- **Evolution**: within and between species variation, hybrids vs. parents, diploids vs. polyploids
- **Functional analyses**: wild type vs. mutant

Microarray studies

1. Indroduction
2. **Microarray technology**
3. Statistics
4. MIAME
5. Examples of microarray studies (paper discussion topic and lab topic)

Microarray analysis principle

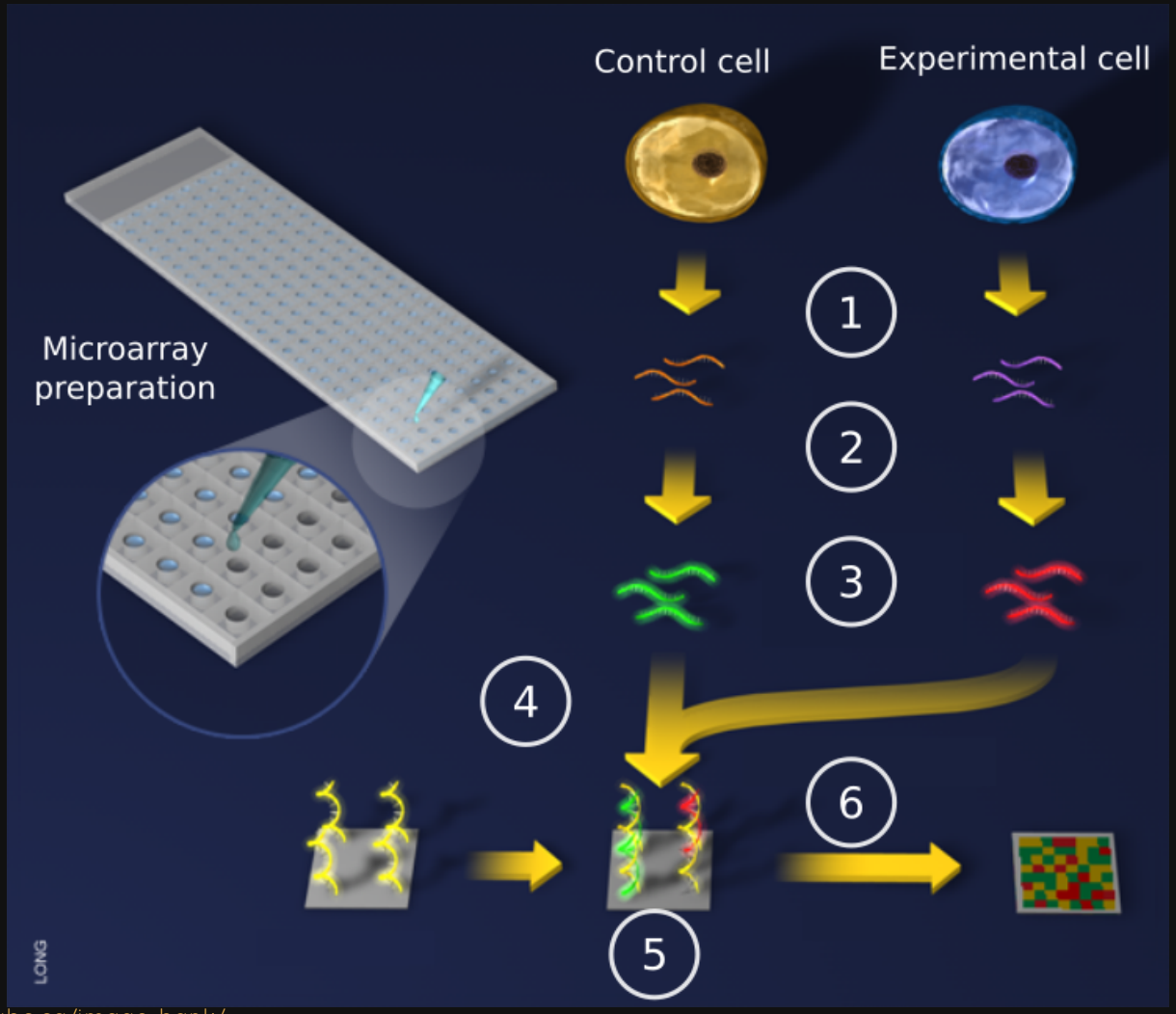


Image from: http://www.scq.ubc.ca/image-bank/

Microarray analysis principle

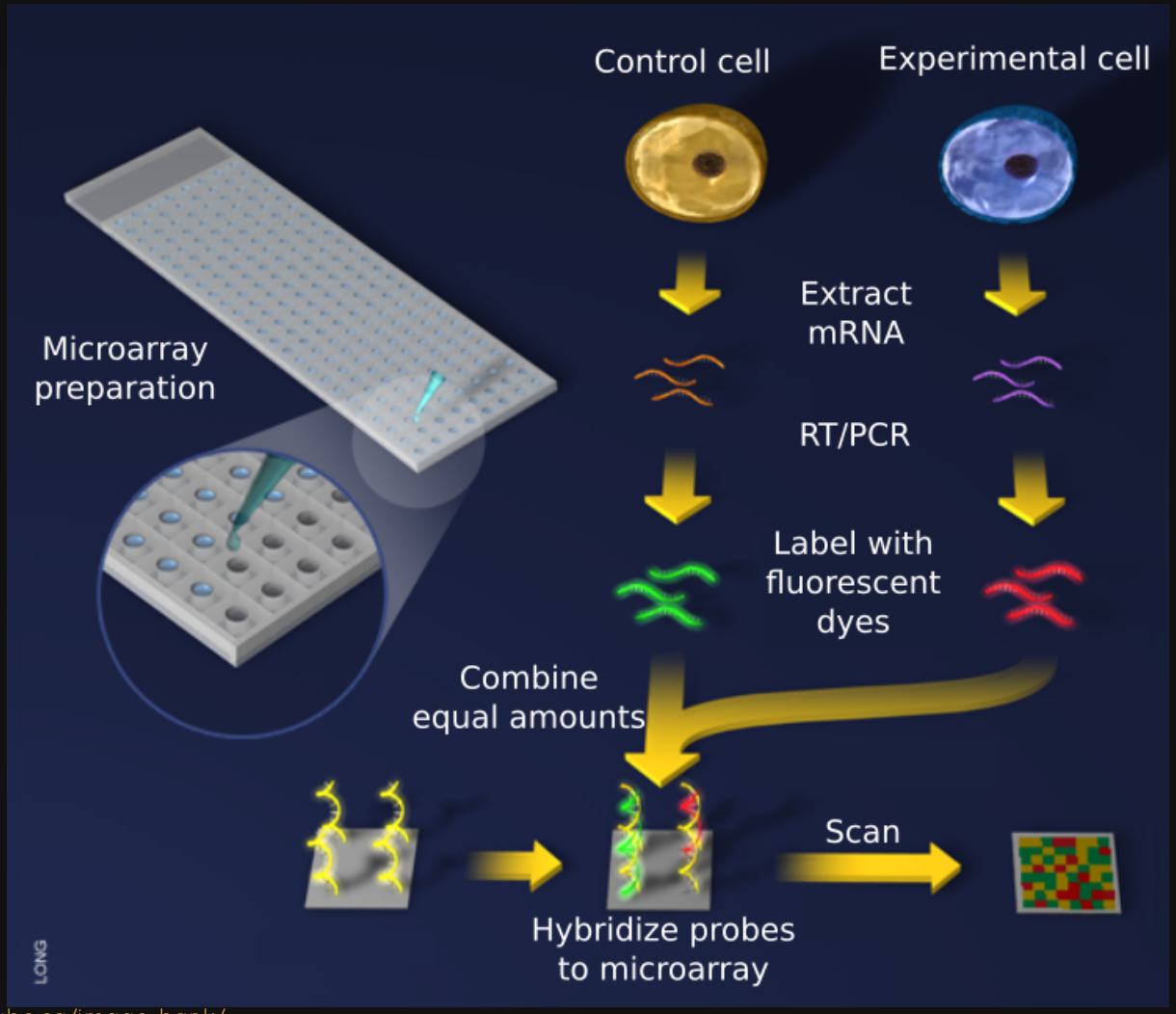


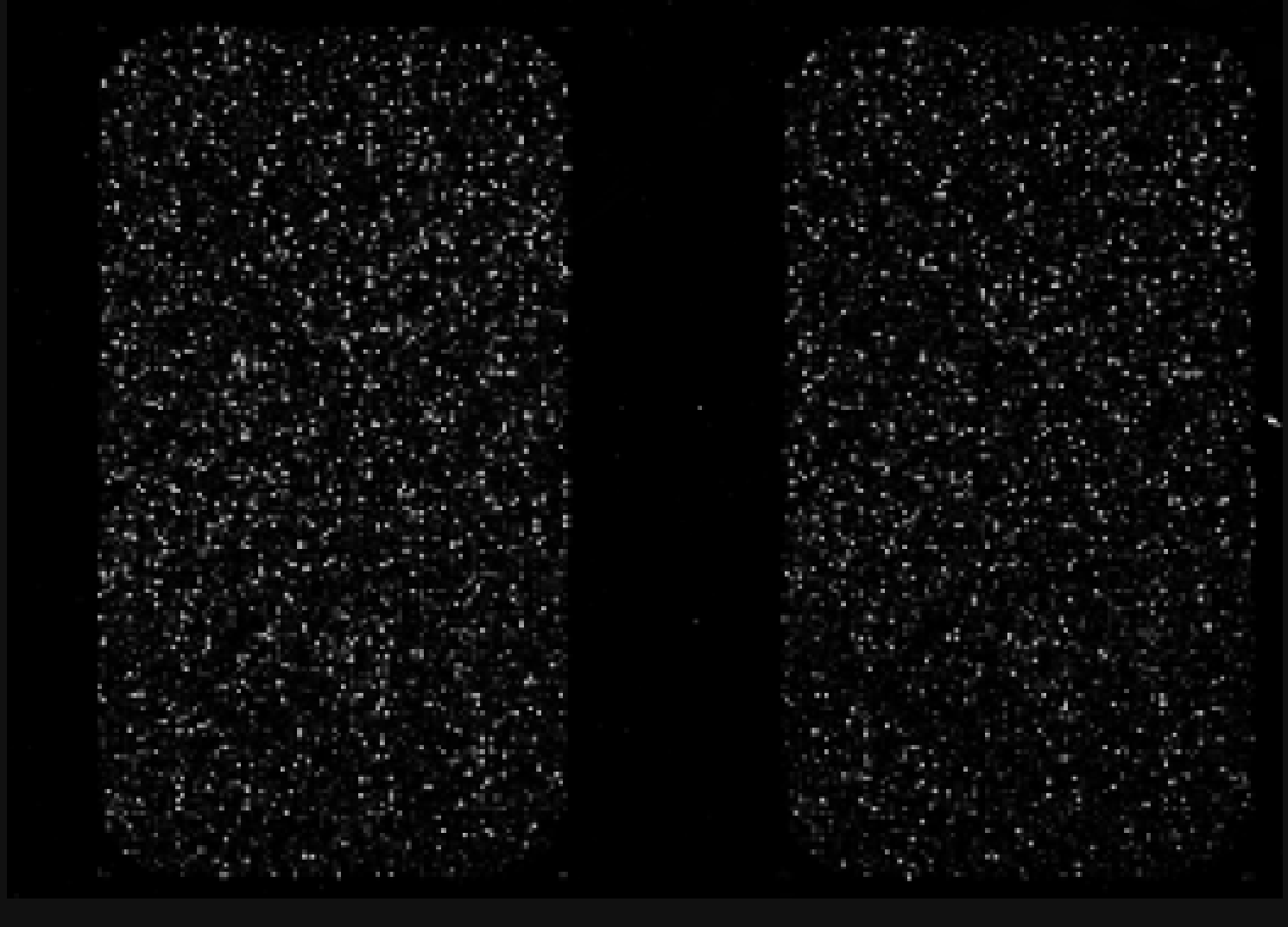
Image from: http://www.scq.ubc.ca/image-bank/

Competitive hybridization

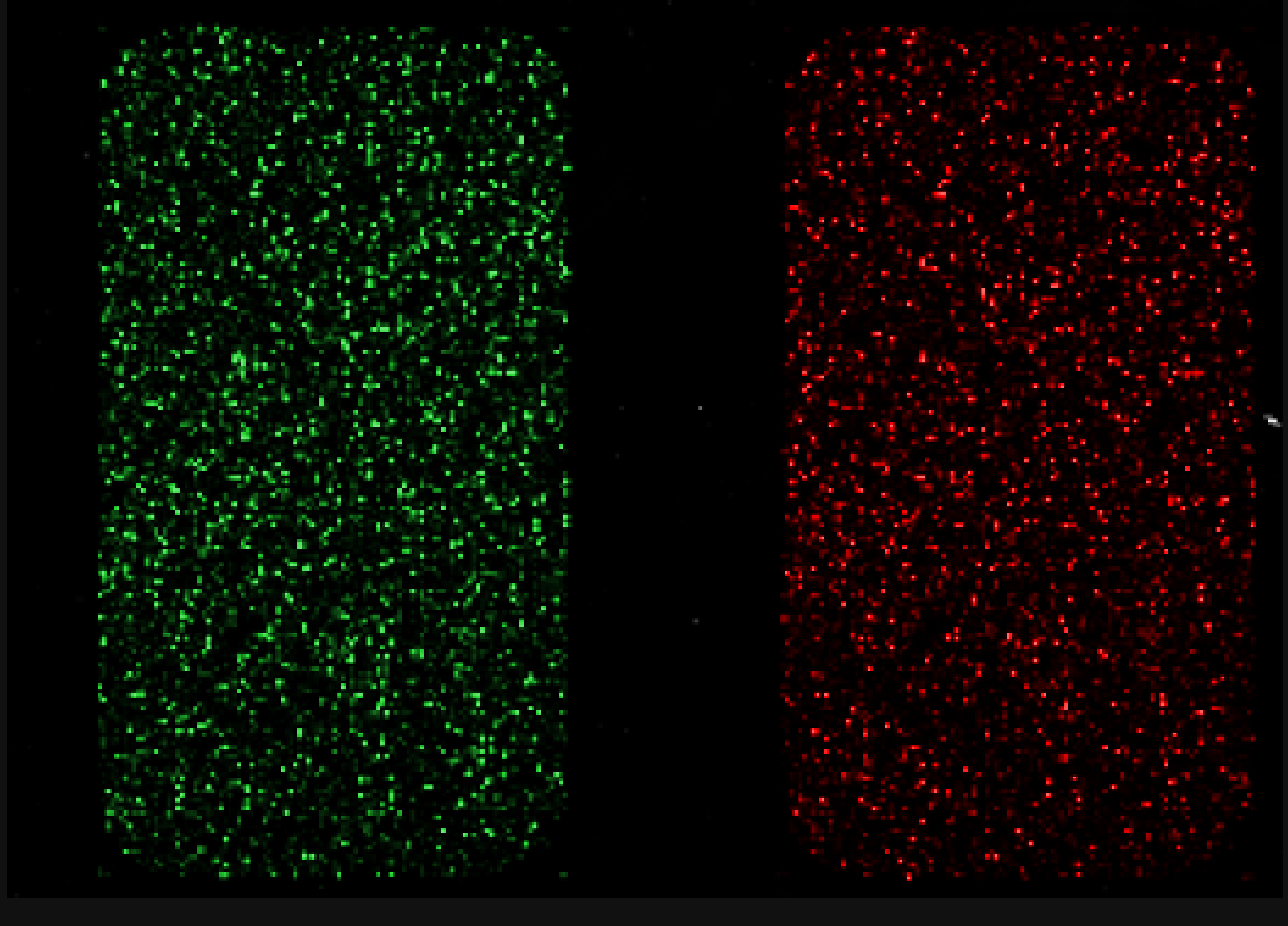
*It is possible to represent **different** samples on **one** microarray using **different** **fluorescent molecules (fluorophores)***

- **Cyanin 3** (Cy3): green fluorescence (excited at 550nm, emission at 570nm)
- **Cyanin 5** (Cy5): red fluorescence (excited at 650nm, emission at 770nm)

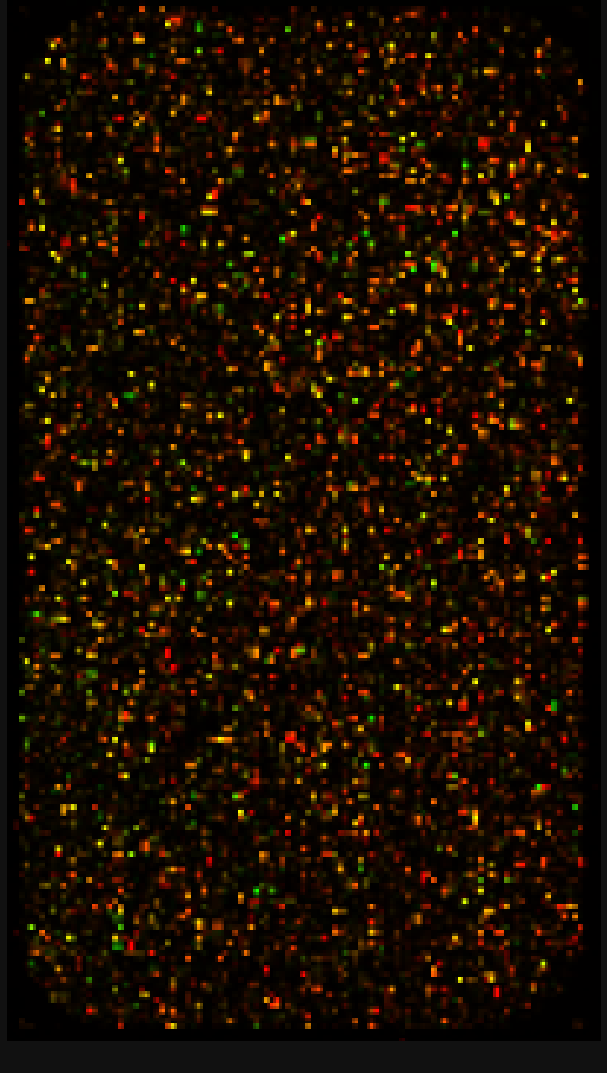
Competitive hybridization



Competitive hybridization



Competitive hybridization



Microarray study pipeline

Question driven

Goals? Hypothesis? Questions?

Microarray study pipeline

- **Platform**
  - What technology?
  - Source of the gene probes?
  - Cross-species hybridization?

Microarray study pipeline

- Platform
- **Experimental design**
  - What statistics?
  - What analysis software?
  - Replication level
  - Hybridization scheme

Microarray study pipeline

- Platform
- Experimental design
- **Laboratory steps**
  - Sample preparation and labelling
  - Hybridization
  - Washing
  - Image acquisition

Microarray study pipeline

- Platform
- Experimental design
- Laboratory steps
- **Bioinformatics steps**
  - Data transformation and normalization
  - Analysis of differentially expressed genes (**statistical tests**, gene ontology, ...)
  - Visualization (graphics)
  - Data storage (databases, MIAME standards)

Microarray study pipeline

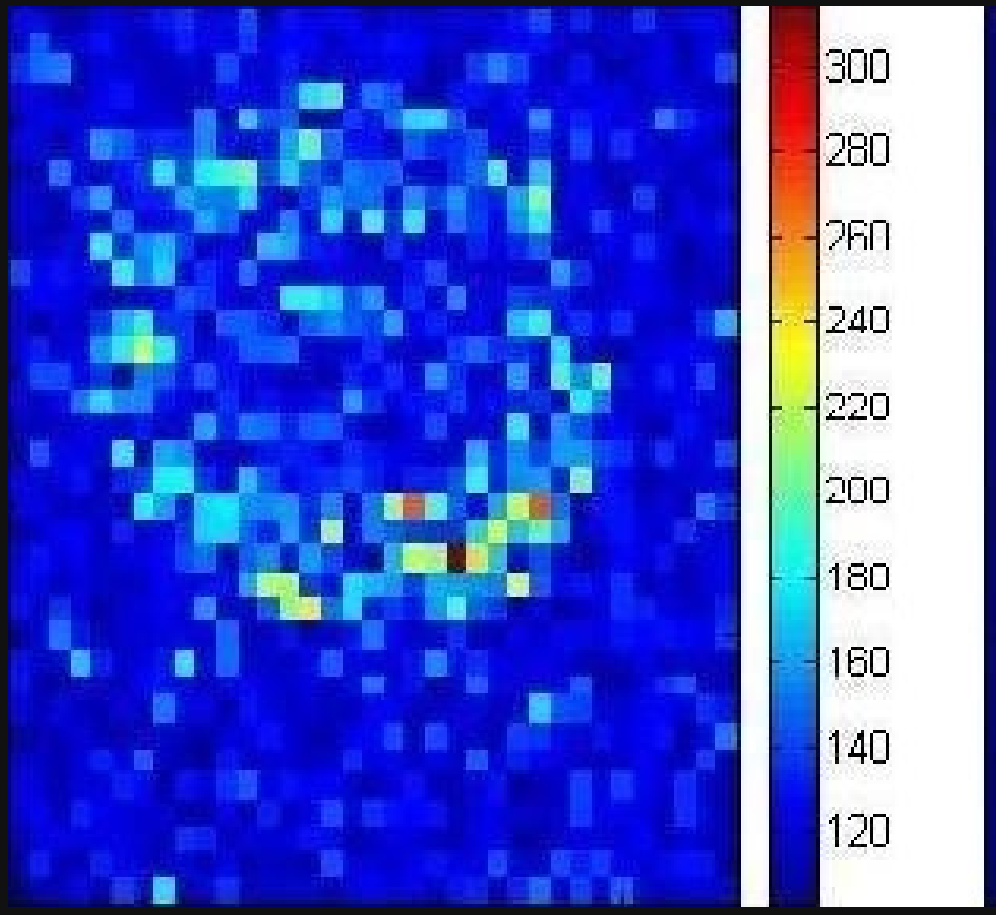
- Platform
- Experimental design
- Laboratory steps
- Bioinformatics steps
- **Data interpretation**
  - Answers?
  - New hypotheses?
  - Follow-up experiments?
  - Validation?

Microarray studies

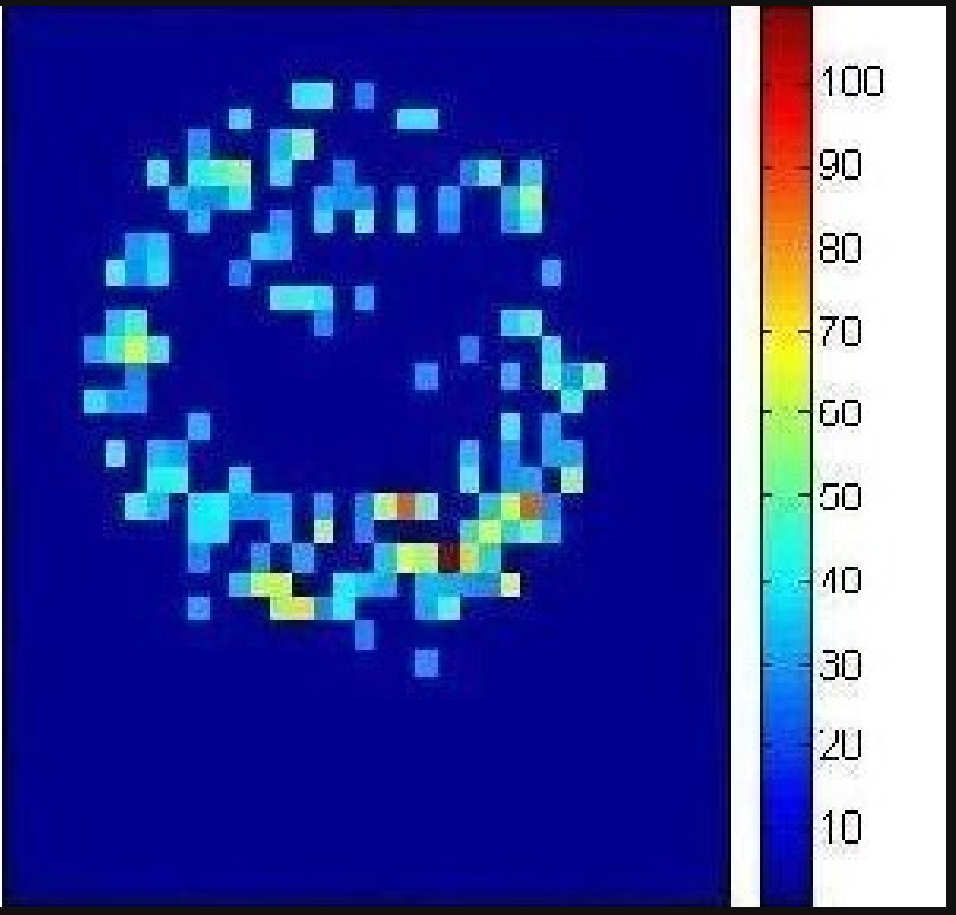
1. Indroduction
2. Microarray technology
3. **Statistics**
4. MIAME
5. Examples of microarray studies (paper discussion topic and lab topic)

Noise reduction

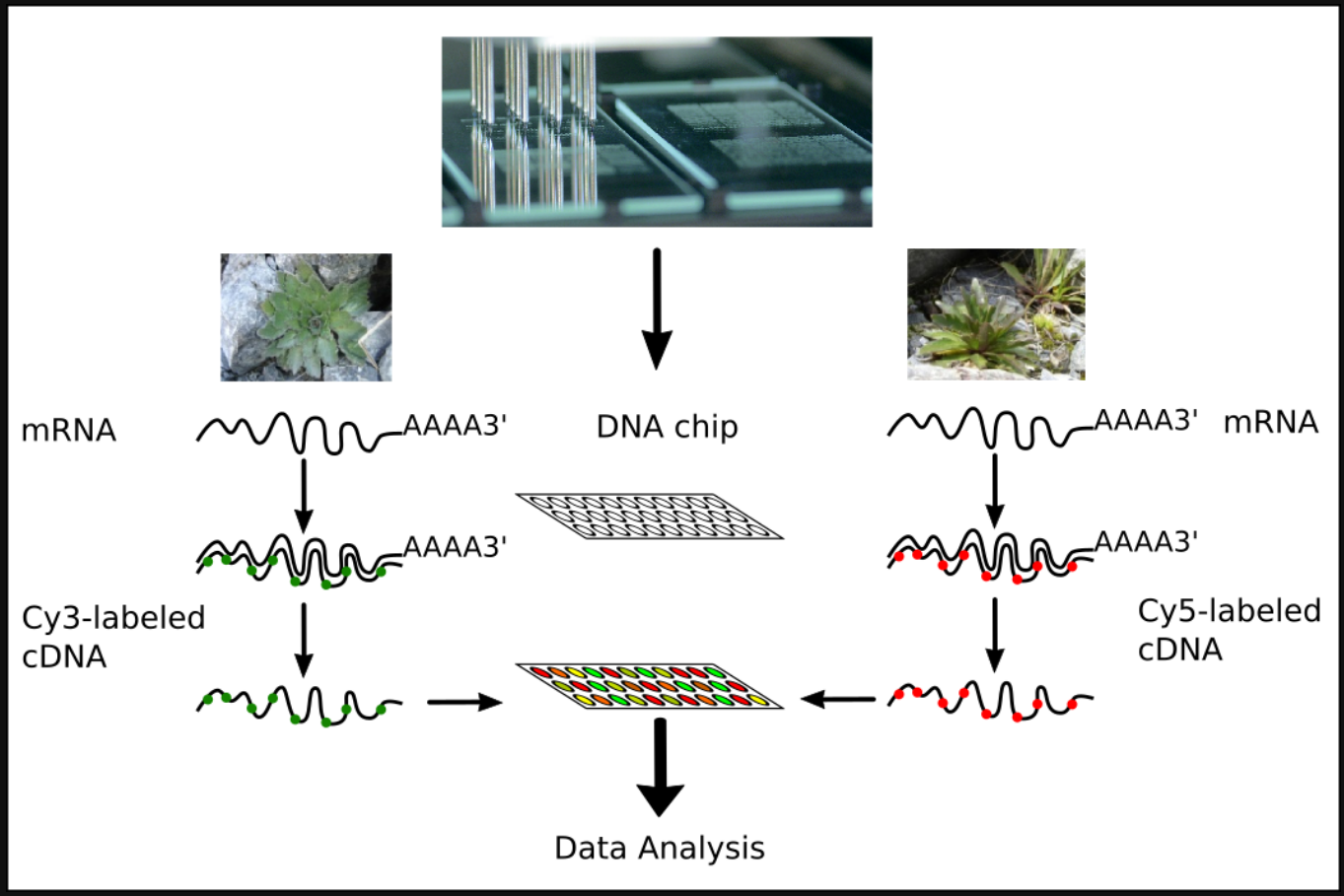
Before



After



Log Fold Ratio



Expression ratio: $log(\frac{Cy5}{Cy3})$

Expression ratios, M & A

- $Cy3 = Sample1$ (Green)
- $Cy5 = Sample2$ (Red)
- $Cy5 > Cy3$: higher expression in sample 2
- $Cy3 > Cy5$: higher expression in sample 1
- Log fold ratio: M = $log_2(\frac{Cy5}{Cy3})$ = $log_2(Cy5) - log_2(Cy3)$
- Expression average: A = $\frac{1}{2}(log_2(Cy5) + log_2(Cy3))$ = $\frac{1}{2}log_2(Cy5Cy3)$

Log Fold Ratio

Reminder: $log_2(x)$ *is the unique real number* y *such that:* $2^y = x$.
For example: $log_2(8) = 3$ *because* $2^3 = 8$

$Cy5/Cy3$	$log_2(Cy5/Cy3)$
4	2
3	1.58
2	1
1.5	0.58
1	0
0.66	-0.58
0.5	-1
0.33	-1.58

0.25 -2

Hypothesis testing

T-test

*Null hypothesis (H_0): gene A is **not** differentially expressed between two treatments*

Mean:

$\overline{x} = \frac{1}{M} \sum_{i=1}^{M} x_i$; *for gene* x *in M* ***replicates***

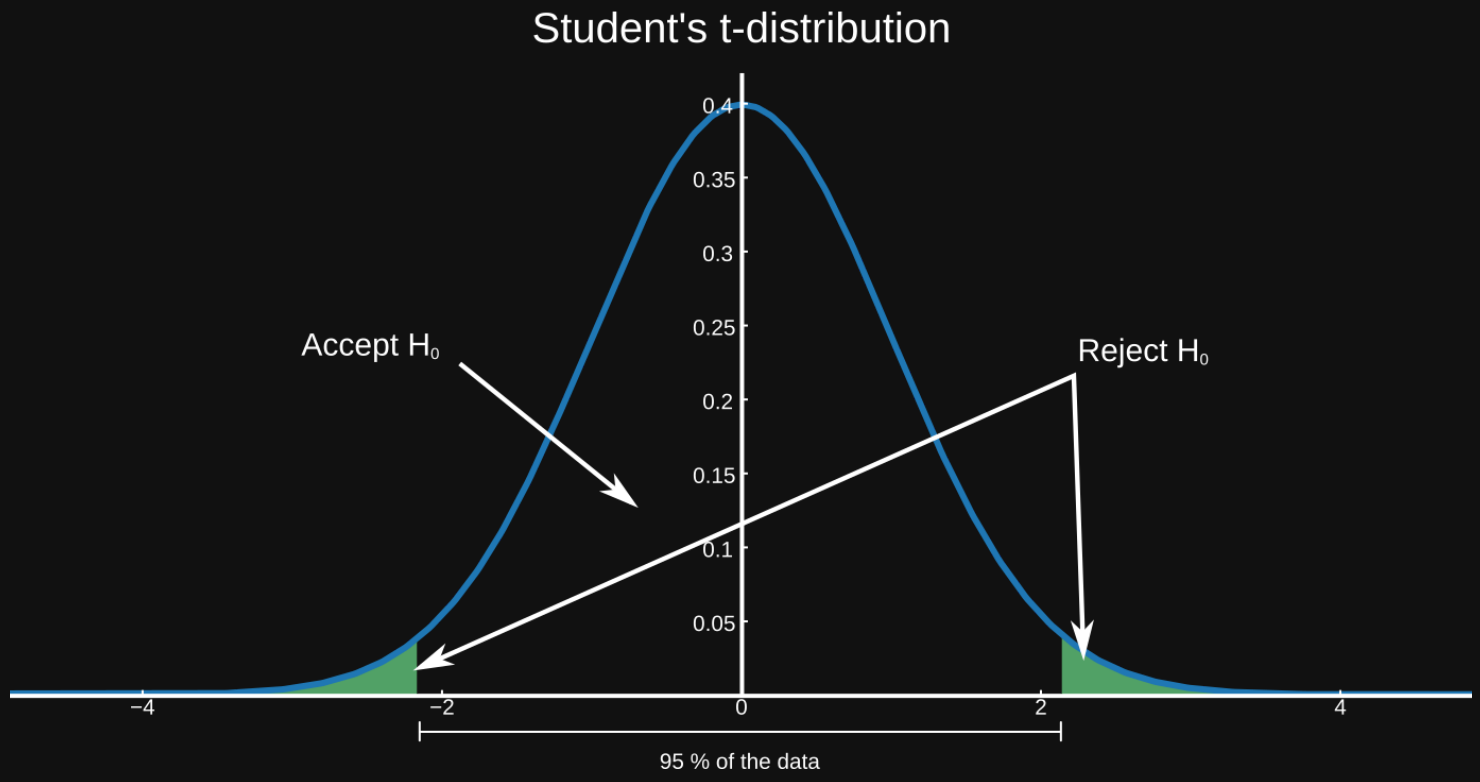
Variance:

$S_x^2 = \frac{1}{M-1} \sum_{i=1}^{M} (x_i - \overline{x}^2)$

T-statistic:

$$T_x = \frac{\overline{x_{S_1}} - \overline{x_{S_2}}}{\sqrt[2]{\frac{S_{x_{S_1}}^2}{M} + \frac{S_{x_{S_2}}^2}{N}}}$$

T-test and P-value



T-test is used only to compare two samples. To compare more, ANOVA (ANalysis Of VAriance) is used.

Hypothesis testing

T-test

*Null hypothesis (H_0): gene A is **not** differentially expressed between two treatments*

1. Compute the signal to noise ratio (difference of the means or medians) for each gene
2. Compute the t-statistic for each gene using the replicates
3. Compare t-statistic with the t-distribution
4. If t-statistic is more extreme than the critical t-statistic at a chosen significance level (e.g. $\alpha = 0.05$) reject the null hypothesis, otherwise accept it. **P-value estimation**

Quiz

Usually, a $p <$ **0.05** is considered small enough to reject the null hypothesis of no biological effect in favour of the alternative hypothesis of a biological effect.
P-values are also known under type **1** error – the probability of rejecting the null hypothesis when it is actually true (= false positive rate).
P-value of 0.01 means a false positive rate of **1** %.
When analyzing multidimensional data sets, p-values need to be adjusted for **multiple testing** .
Two common p-value adjustment methods are **Bonferroni** and **False Discovery Rate** .

Bonferroni Correction

- If you hypothesize that **a specific gene** is up-regulated, $p < 0.05$ is fine.
- If you hypothesize that **any of 10,000 genes** is up-regulated, with $p < 0.05$ you can expect to see 5% (**500 genes**) up-regulated by chance alone.
- To account for the thousands of repeated measurements, some researchers apply a Bonferroni correction.

$$p < (0.05)/10,000$$
$$\text{or } p < 5e^{-6}$$

*The Bonferroni correction is generally considered to be **too** conservative and **False Discovery Rate (FDR)** should be used.*

False Discovery Rate

Benjamini-Hochberg method

Imagine an array with 6400 genes and an experiment where 184 genes are differentially expressed at p=0.01: 64 genes would be expected to appear differentially expressed by chance alone.

FDR = false discovery rate = $\frac{64}{184} * 100 = 35\%$

False Discovery Rate

Benjamini-Hochberg method

P-value	Observed Number of genes	Expected number of False Positives	FDR
10^{-2}	184	64	35
10^{-3}	35	6	18
10^{-4}	15	0.6	4

With decreasing p-value, FDR also decreases, but so does the number of differentially expressed genes – choose a p-value which balances both!

Microarray studies

1. Indroduction
2. Microarray technology
3. Statistics
4. **MIAME**
5. Examples of microarray studies (paper discussion topic and lab topic)

MIAME Standard

Minimum Information About a Microarray Experiment *that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment*

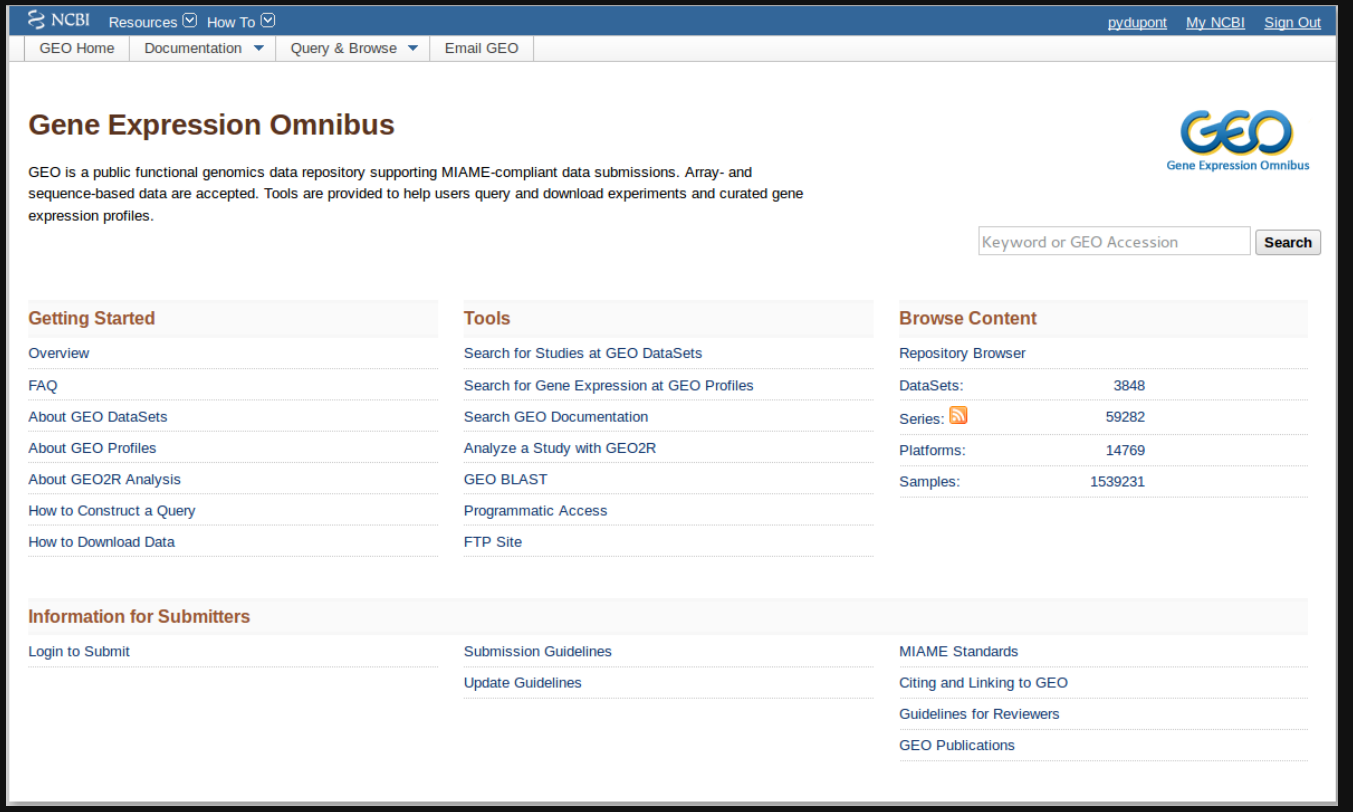
http://fged.org/Workgroups/MIAME/miame.html/

MIAME Standard

1. **Raw data** for each hybridisation (CEL or GPR files)
2. **Processed** (normalised) **data** (used to draw the conclusions from the study)
3. Essential **sample annotation** including experimental factors and their values
4. **Experimental design** including sample data relationships (e.g. which hybridizations are technical and biological replicates)
5. Sufficient **array annotation** (e.g. gene identifiers, probe sequences)
6. Essential **laboratory and data processing protocols** (e.g. normalization method used to obtain the final data)

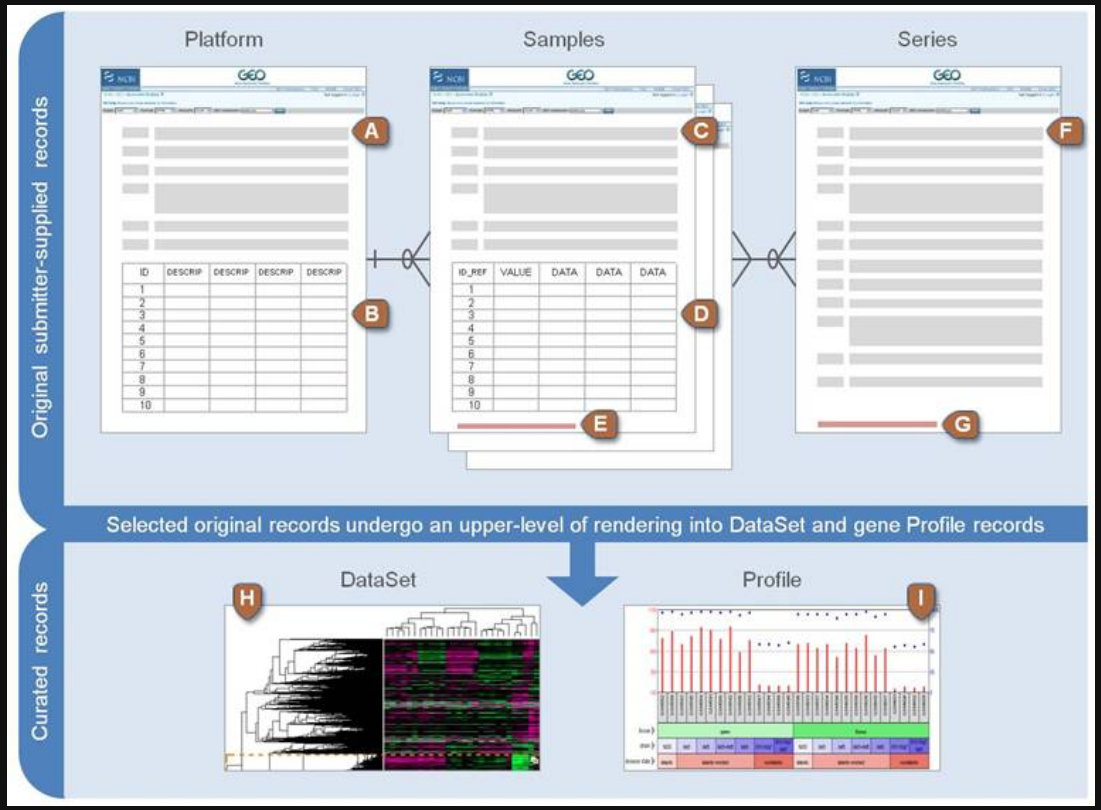
Gene expression databases

Gene Expression Omnibus (GEO) @ NCBI (http://www.ncbi.nlm.nih.gov/geo/)



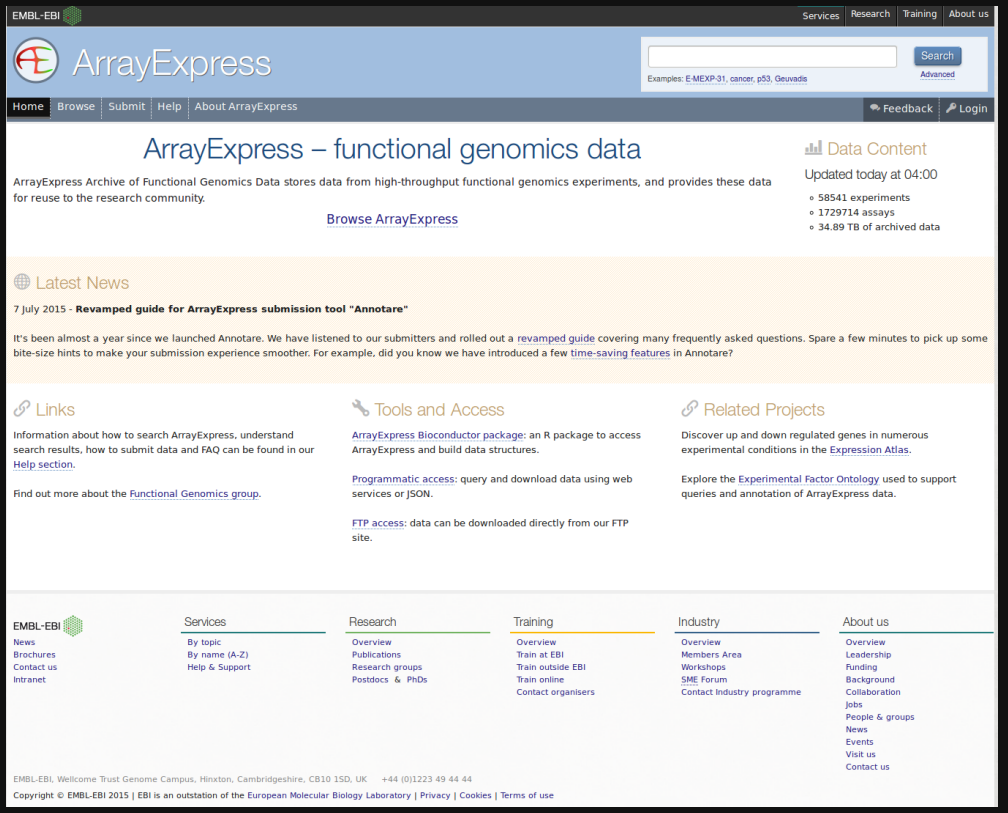
Gene expression databases

Geo Datasets @ NCBI (http://www.ncbi.nlm.nih.gov/gds/)
Geo Profiles @ NCBI (http://www.ncbi.nlm.nih.gov/geoprofiles/)



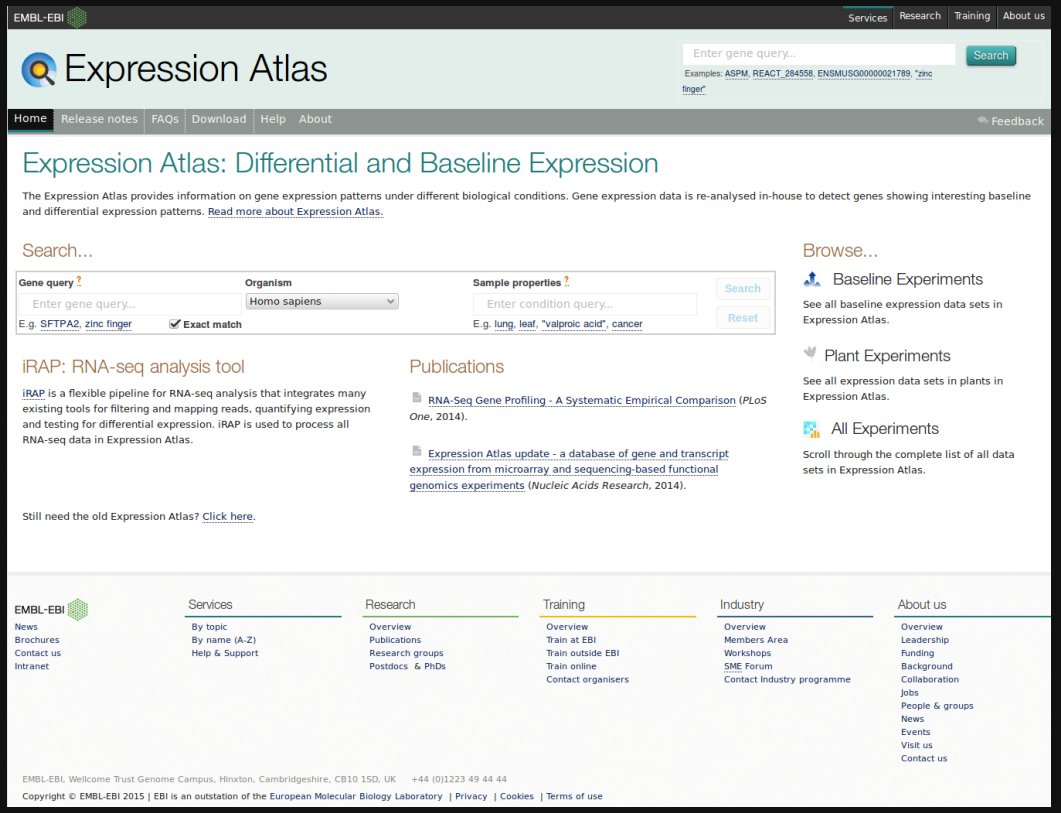
Gene expression databases

ArrayExpress @ EBI (http://www.ebi.ac.uk/arrayexpress/)



Gene expression databases

Expression Atlas @ EBI (http://www.ebi.ac.uk/gxa/)



Microarray studies

1. Indroduction
2. Microarray technology
3. Statistics
4. MIAME
5. **Examples of microarray studies (paper discussion topic and lab topic)**

Microarray paper discussion

MOLECULAR ECOLOGY

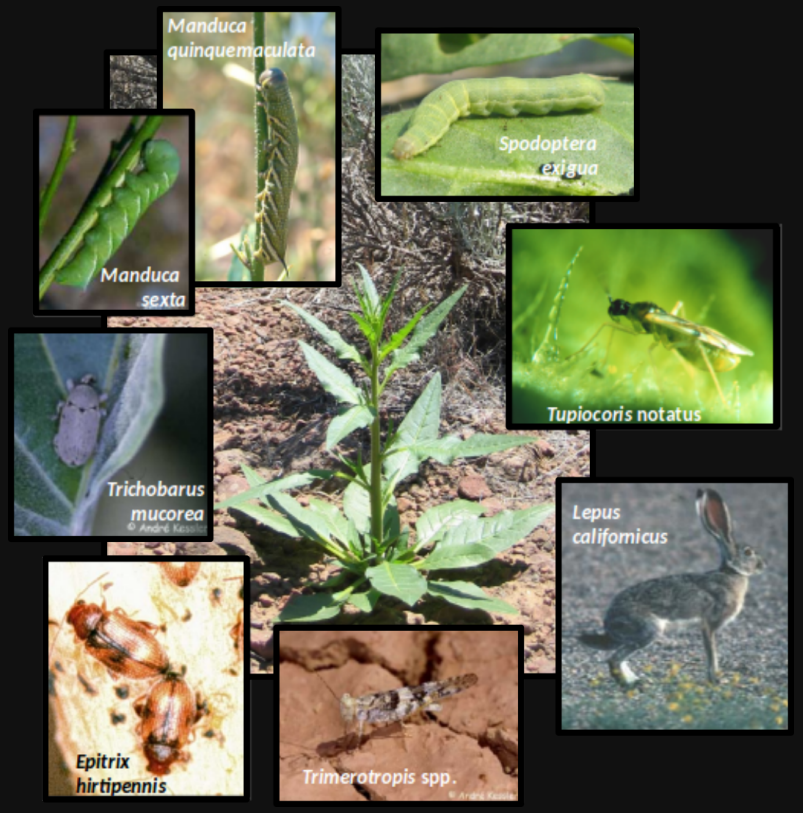
Molecular Ecology (2009) 18, 3227–3239 doi: 10.1111/j.1365-294X.2009.04261.x

Adaptive differences in gene expression associated with heavy metal tolerance in the soil arthropod *Orchesella cincta*

DICK ROELOFS,* THIERRY K. S. JANSSENS,* MARTIJN J. T. N. TIMMERMANS,* BENJAMIN NOTA,* JANINE MARIËN,* ZOLTÁN BOCHDANOVITS,† BAUKE YLSTRA‡ and NICO M. VAN STRAALEN*

**Institute of Ecological Science, VU University Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands, †Department of Clinical Genetics, Section Medical Genomics, VU Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands, ‡Microarray Facility CCA, VU Medical Center, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands*

Lab case study: Herbivory in Nicotiana attenuata (Solanaceae)



What type of research?

- Which genes and metabolites defend plants against insect?
- Costs and benefits of defense
- Genetic engineering of defense traits
- Plant pollination

Lab case study: Herbivory in Nicotiana attenuata (Solanaceae)

Why *N. attenuata*?

- Diverse herbivore community
- High plasticity (direct and indirect defense)
- "Chases" fire
- Easily cultivated annual species

Case study - Chips, veggies & vegetarians

Specificity in Ecological Interactions. Attack from the Same Lepidopteran Herbivore Results in Species-Specific Transcriptional Responses in Two Solanaceous Host Plants[1[w]]

Dominik D. Schmidt[2,3], Claudia Voelckel[2], Markus Hartl, Silvia Schmidt, and Ian T. Baldwin*

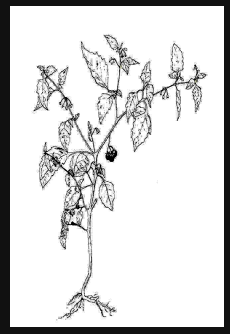
Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Beutenberg Campus, 07745 Jena, Germany

Case study - Chips, veggies & vegetarians

The chip: cDNA array with 15,264 potato genes from TIGR (The Institute for Genomic Research)

The veggies

Solanum nigrum
Black nightshade



Nicotiana attenuata
Coyote tobacco

The vegetarian

Manduca sexta

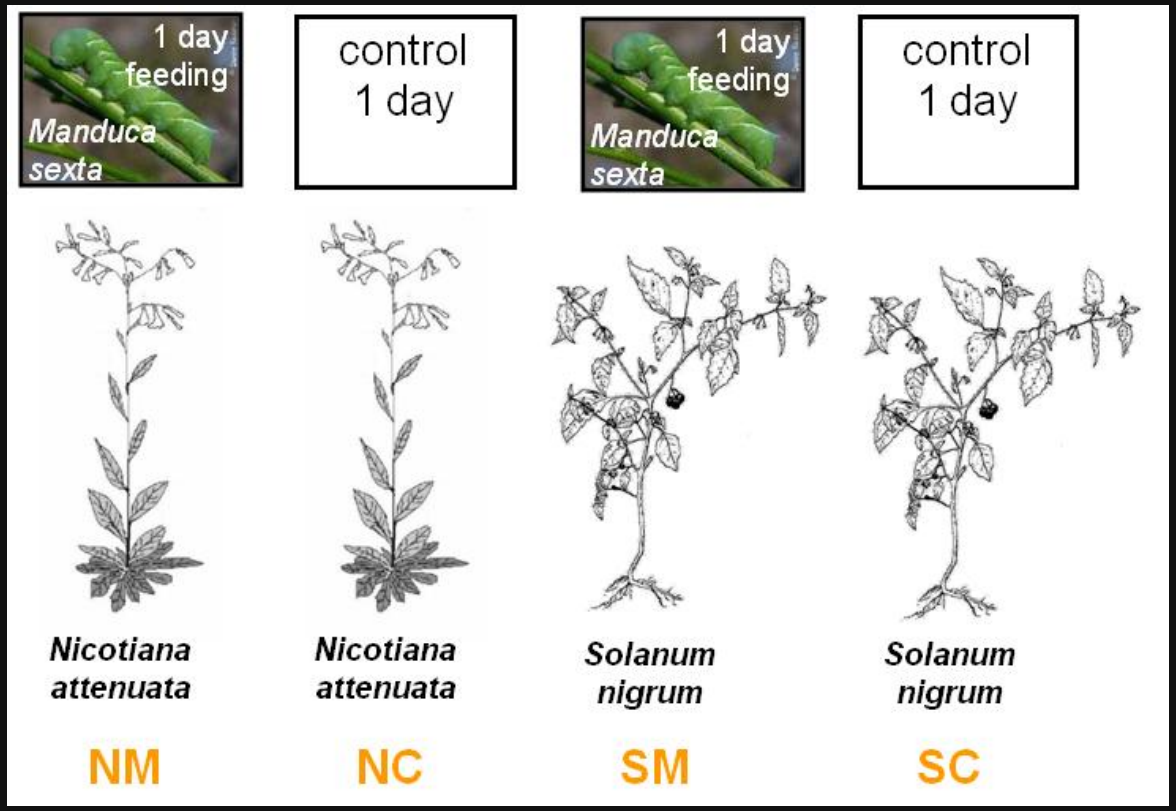
Question:

Do tobacco and black nightshade plants respond differently to caterpillar attack?

Microarray Case Study

RNA source

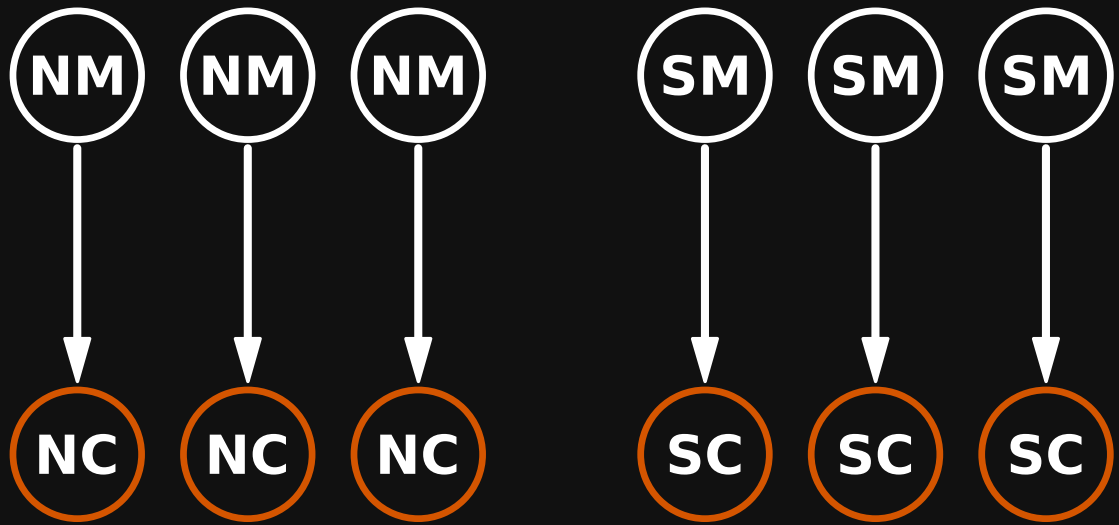
2 herbivore treatments and 2 controls



Microarray Case Study

6 arrays

Each arrow represents one array. Herbivore-induced tissue (cy3) was co-hybridized control tissue (cy5). Each comparison was replicated three times.


NM
NM
NM
SM
SM
SM
NC
NC
NC
SC
SC
SC

What will you do in the lab?

Lab 1

R warm-up exercise. Identification of **differentially expressed genes**

Lab 2

Identification of **differentially expressed biological processes**