

Advanced Practical Genetics

Microarray - Hands-on data analysis

203.305

Dr. Pierre-Yves Dupont, Postdoctoral Researcher
Computational Biology Research Group

p.y.dupont@massey.ac.nz

http://bit.ly/massey_203305

Planning

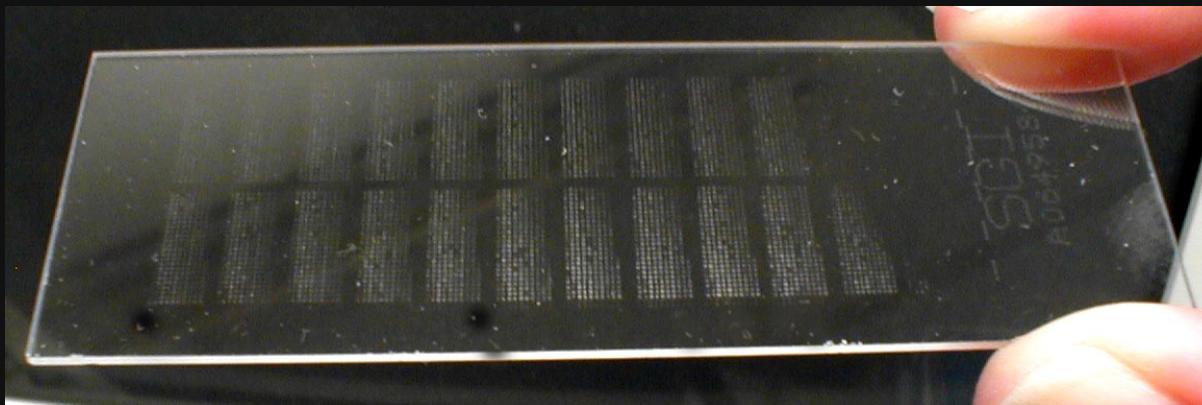
- **25.9.15** – Introduction (lecture), *AHB1.40B*
- **29.9.15** – From raw data to lists of differentially expressed genes (Step by step analysis of a microarray data set using the R language, **3h lab**, *SC5.10*) -
Discussion worksheet due, research proposal due
- **2.10.15** – Paper discussion, *AHB1.40B*
- **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. **Introduction**
2. Microarray technology
3. Analysis
4. MIAME
5. Examples of microarray studies (paper discussion topic and lab topic)

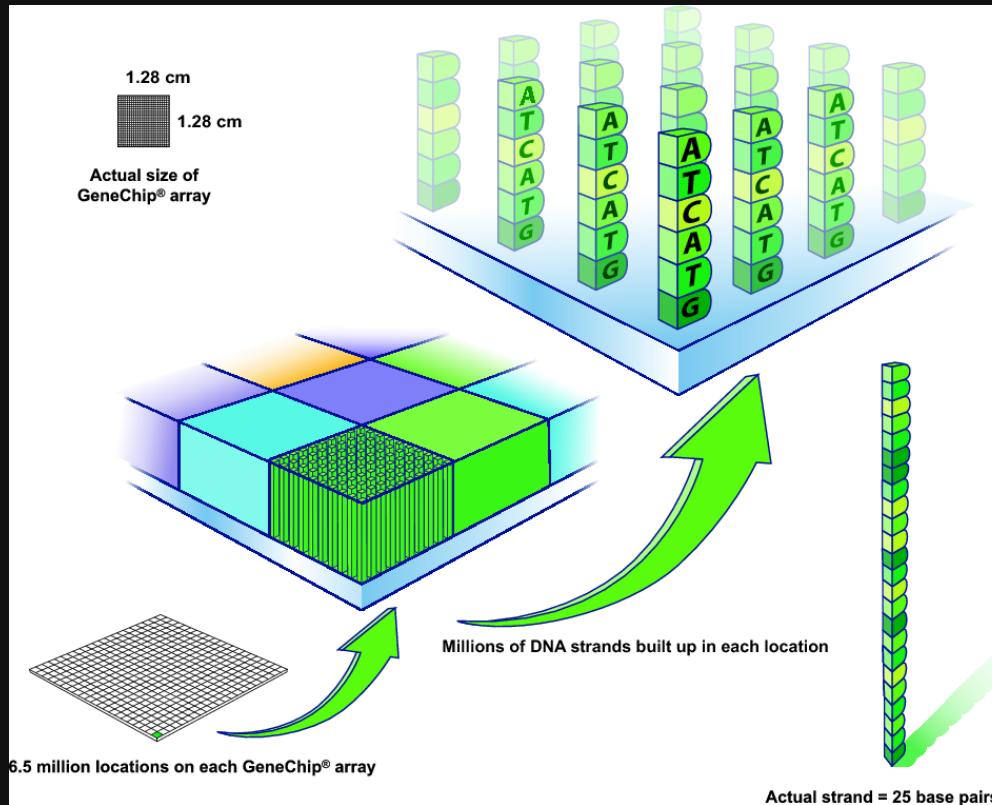
What are microarrays?

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**.



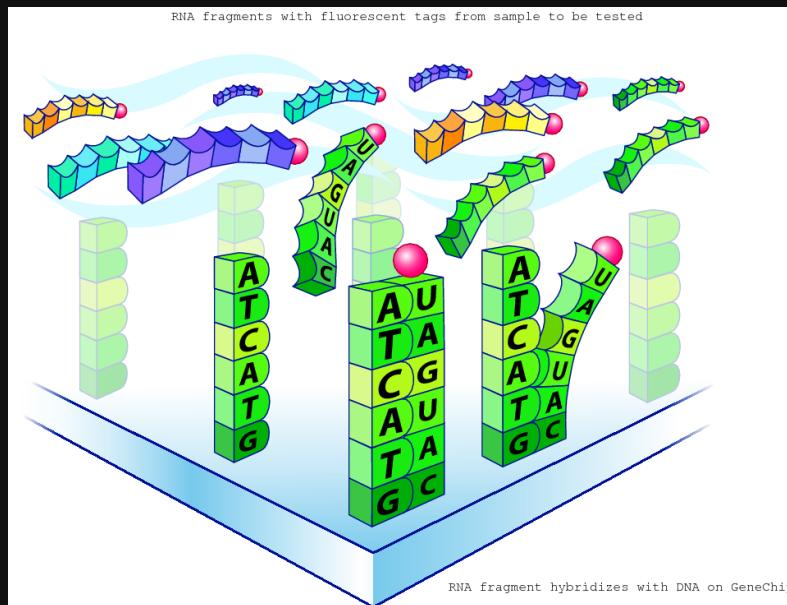
What are microarrays?

DNA microarray



What are microarrays

Hybridisation and transcriptomics?



The amount of RNA hybridised 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

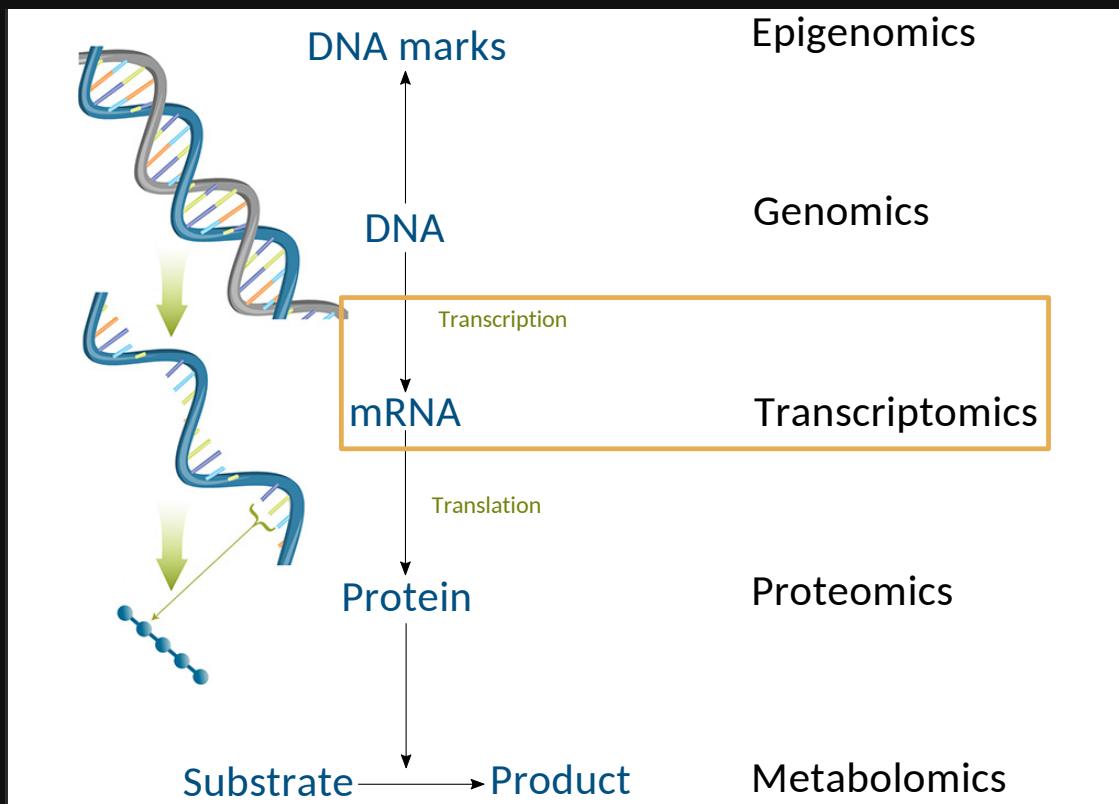


Figure modified from: Katherine Joyce, Woods Hole Oceanographic Institution

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)
- **Hybridisation:** 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 hybridise (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. Introduction
2. **Microarray technology**
3. Analysis
4. MIAME
5. Examples of microarray studies (paper discussion topic and lab topic)

Microarray analysis principle

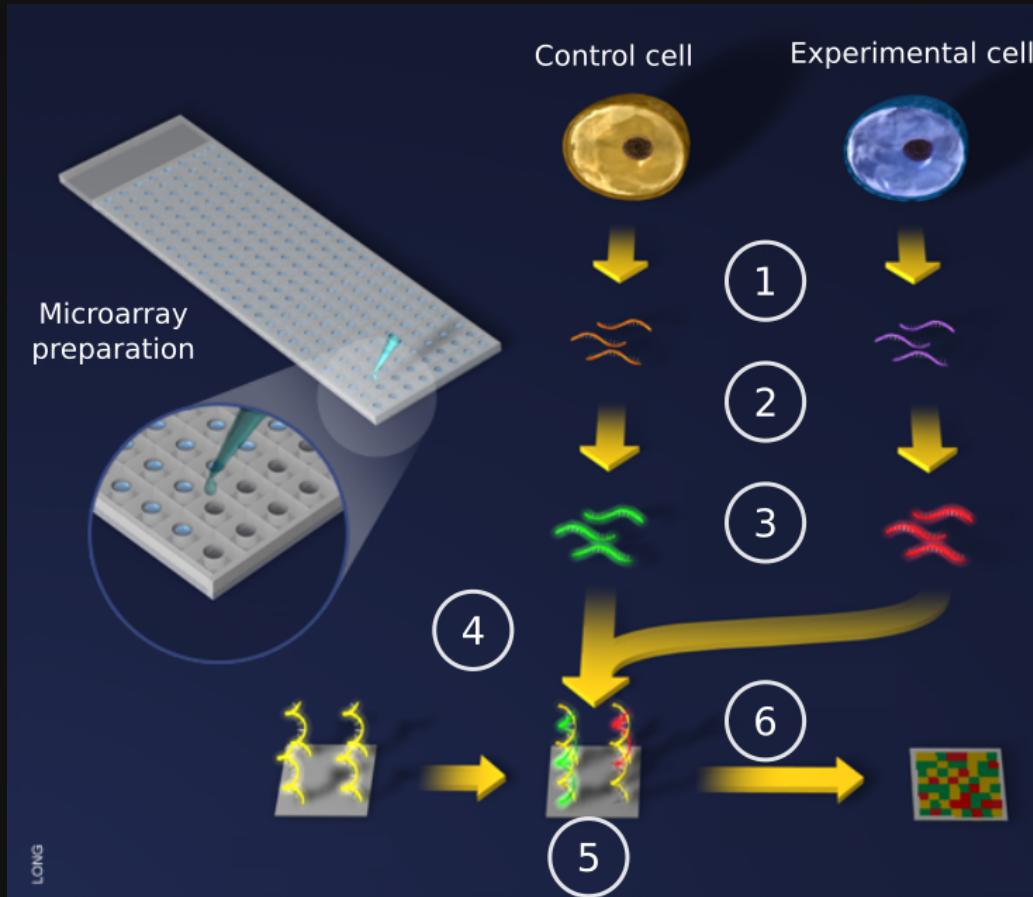


Image from: <http://www.scq.ubc.ca/image-bank/>

Microarray analysis principle

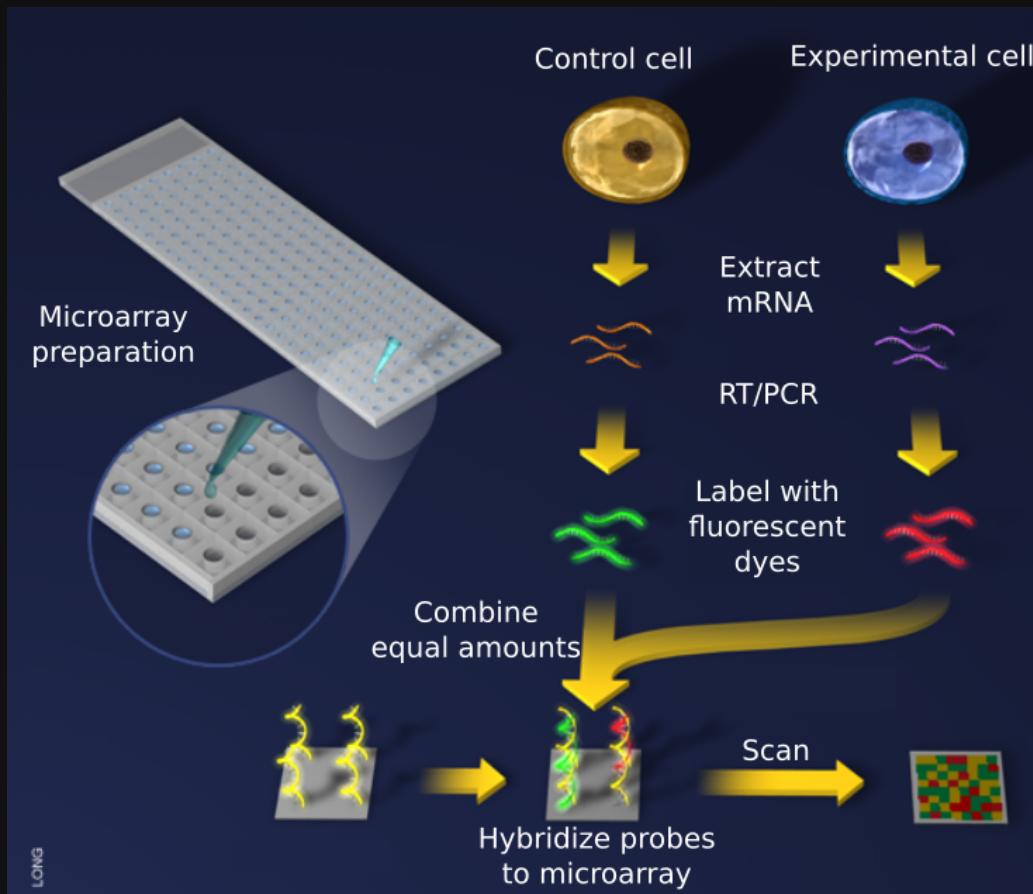


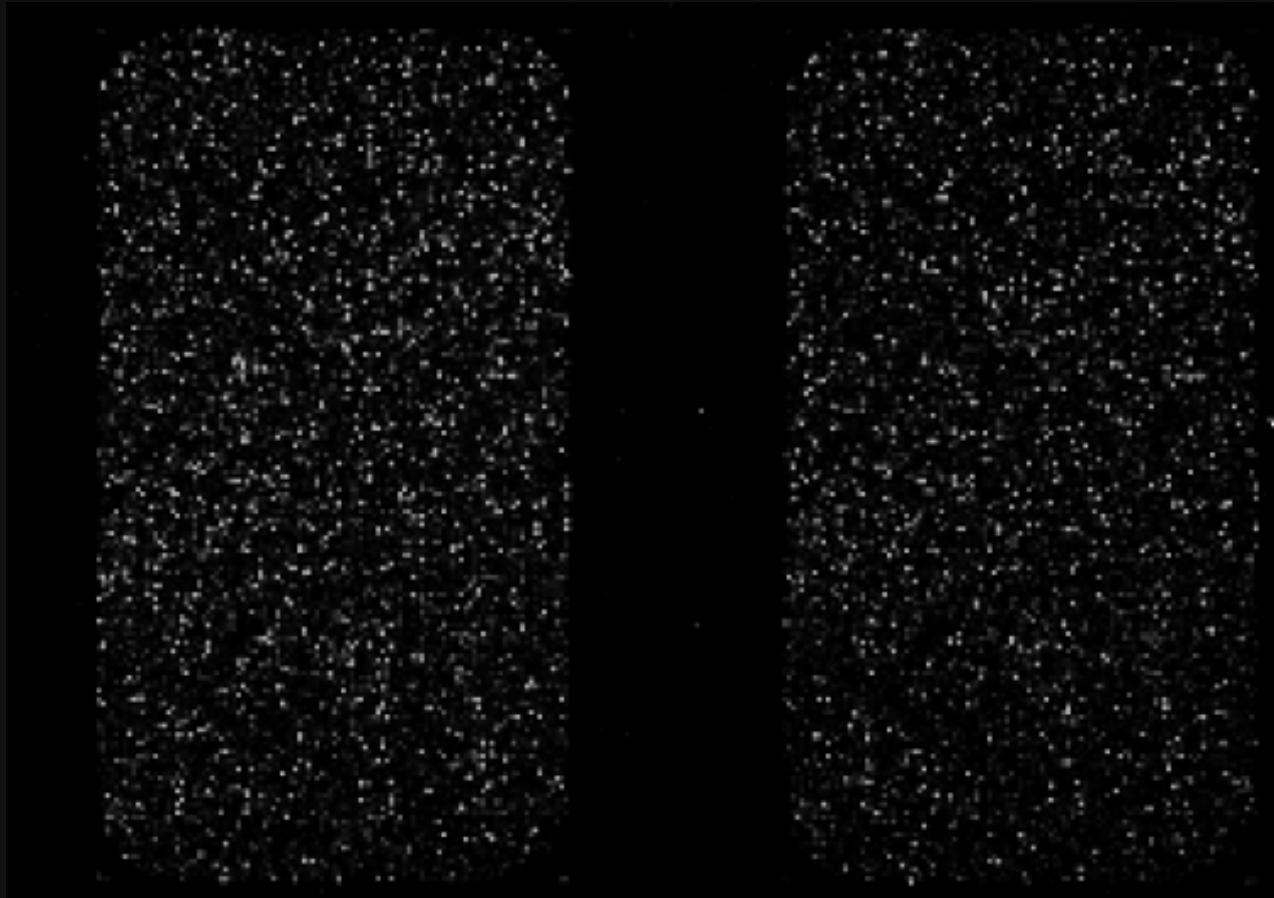
Image from: <http://www.scq.ubc.ca/image-bank/>

Competitive hybridisation

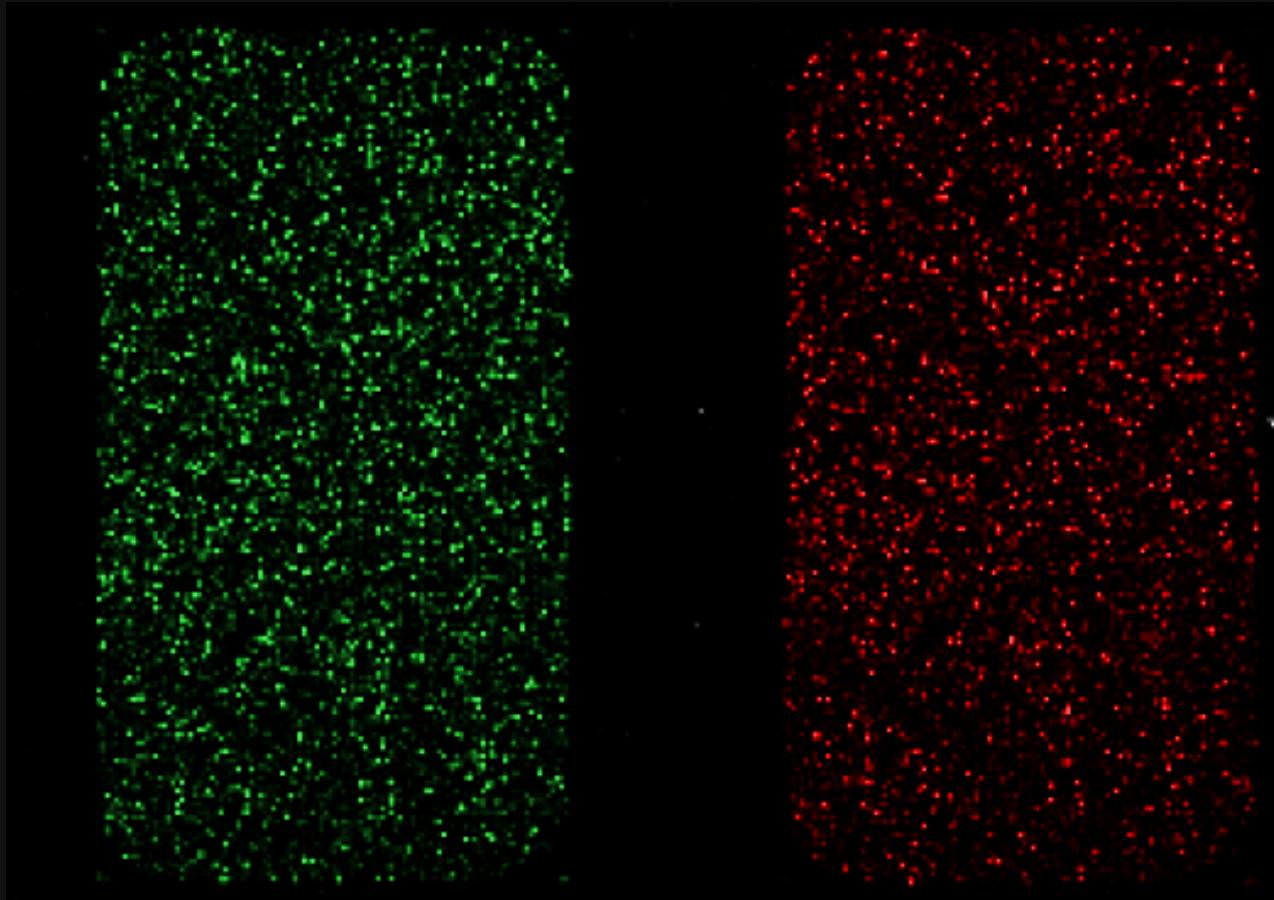
*It is possible to represent **different** samples on **one** microarray using **different fluorescent molecules (fluorophores)***

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

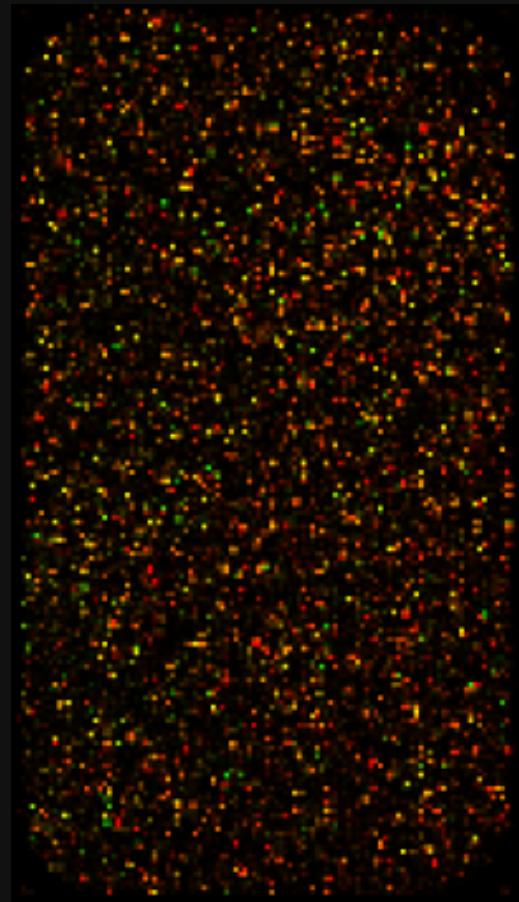
Competitive hybridisation



Competitive hybridisation



Competitive hybridisation



Microarray study pipeline
Question driven
Goals? Hypothesis? Questions?

Microarray study pipeline

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

Microarray study pipeline

- Platform
- **Experimental design**
 - What Analysis?
 - What analysis software?
 - Replication level
 - Hybridisation scheme

Microarray study pipeline

- Platform
- Experimental design
- **Laboratory steps**
 - Sample preparation and labelling
 - Hybridisation
 - 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

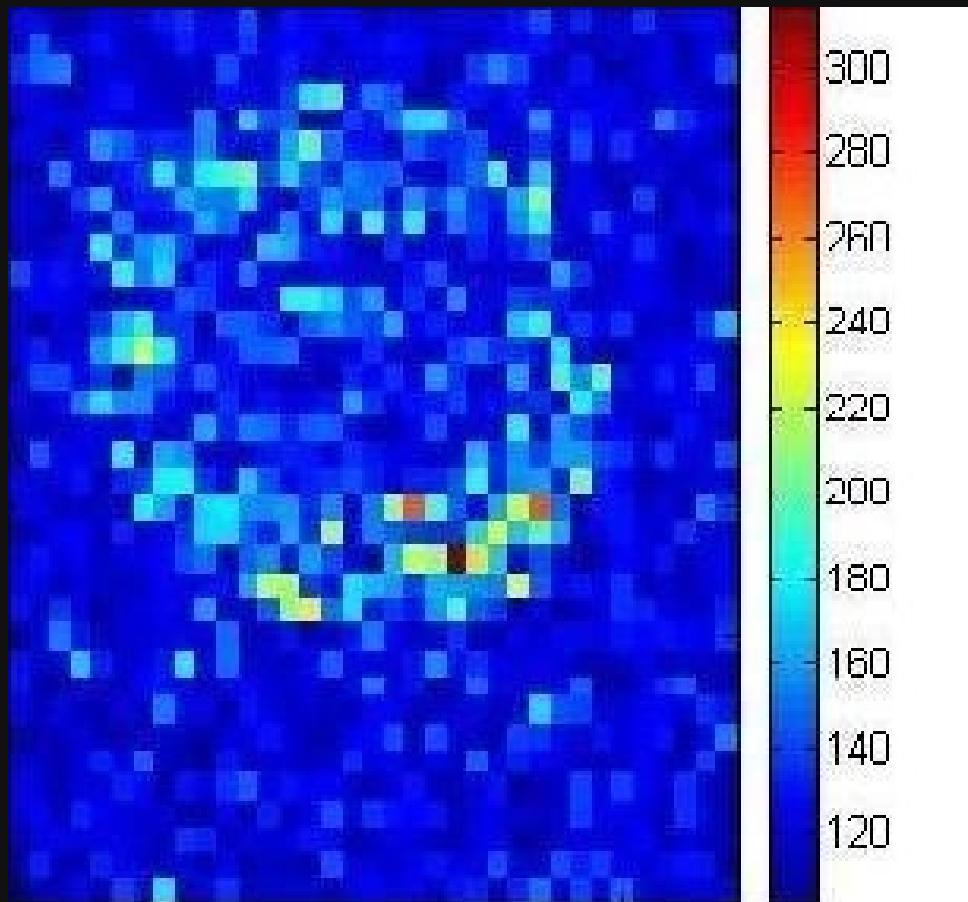
- Platform
- Experimental design
- Laboratory steps
- Bioinformatics steps
- **Data interpretation**
 - Answers?
 - New hypotheses?
 - Follow-up experiments?
 - Validation?

Microarray studies

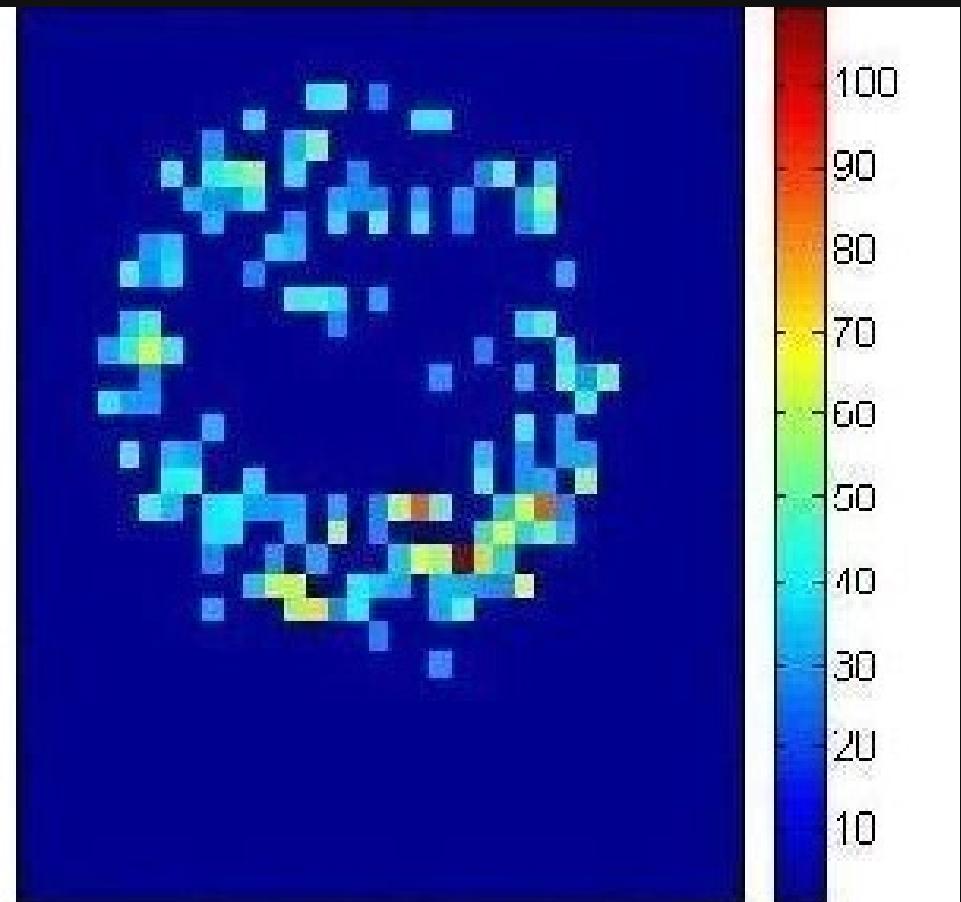
1. Introduction
2. Microarray technology
3. **Analysis**
4. MIAME
5. Examples of microarray studies (paper discussion topic and lab topic)

Noise reduction

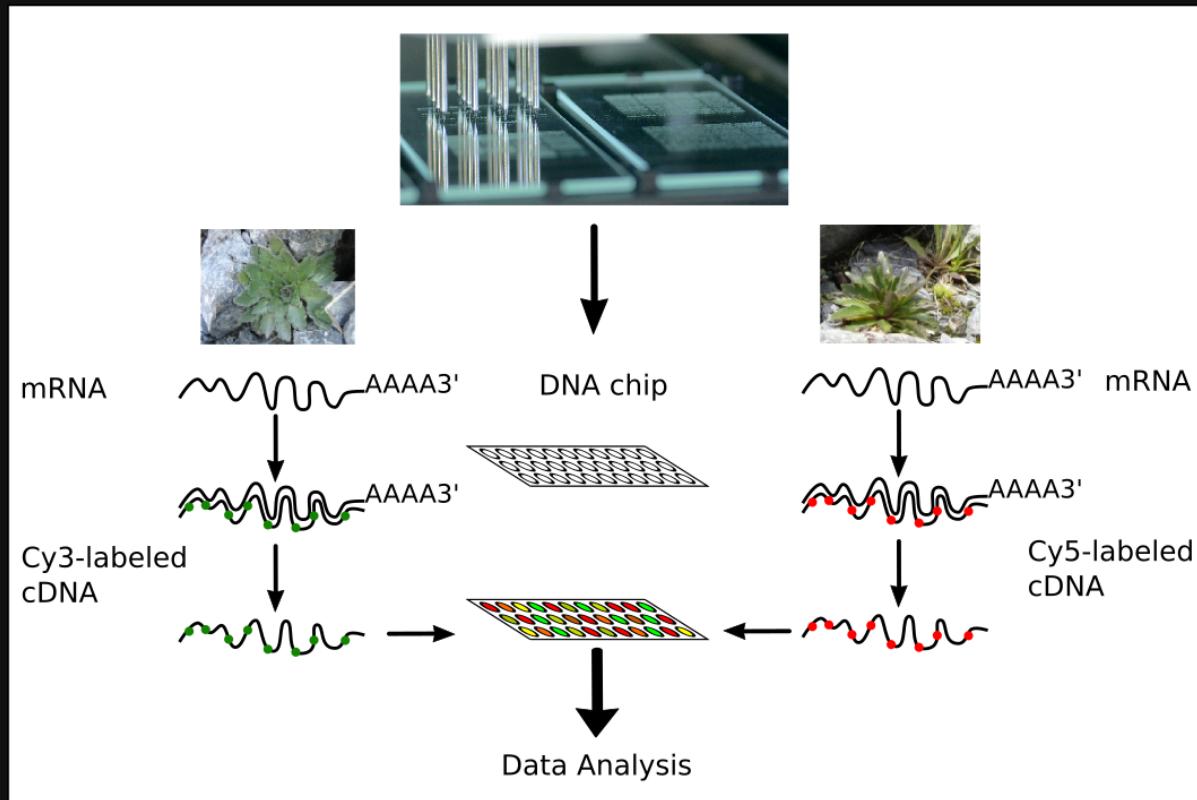
Before



After



Log Fold Ratio



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

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: $M = \log_2\left(\frac{Cy5}{Cy3}\right) = \log_2(Cy5) - \log_2(Cy3)$
- 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 $2^3 = 8$

$$\begin{array}{r} \textcolor{red}{Cy5/Cy3} \quad \log_2(\textcolor{red}{Cy5/Cy3}) \\ \hline 4 & 2 \\ \hline 2 & 1 \\ \hline 1 & 0 \\ \hline 0.5 & -1 \\ \hline 0.25 & -2 \end{array}$$

Hypothesis testing

T-test

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

Mean:

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

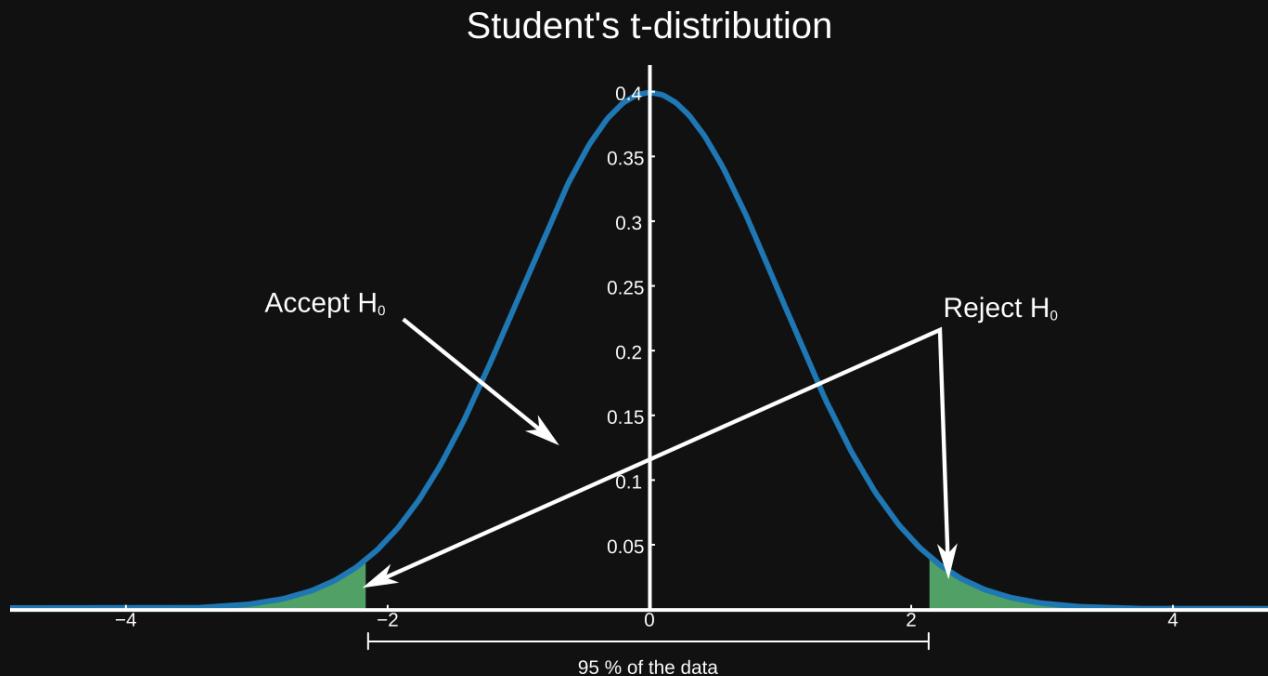
Variance:

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

T-statistic:

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

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 $p < \textbf{0.05}$ 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.

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

*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.

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

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. Introduction
2. Microarray technology
3. Analysis
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 hybridisations 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

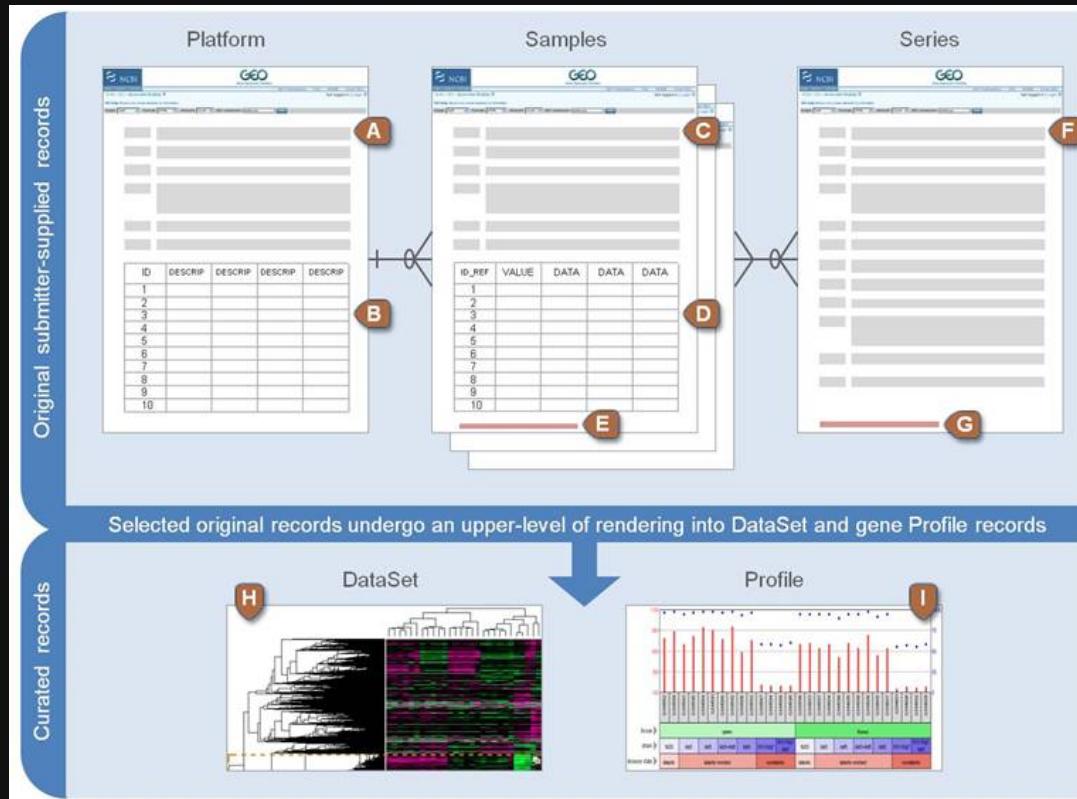
Gene Expression Omnibus (GEO) @ NCBI (<http://www.ncbi.nlm.nih.gov/geo/>)

The screenshot shows the homepage of the Gene Expression Omnibus (GEO) at NCBI. The top navigation bar includes links for NCBI, Resources, How To, GEO Home, Documentation, Query & Browse, Email GEO, and user account options (pydupont, My NCBI, Sign Out). The main title "Gene Expression Omnibus" is displayed above a brief description of the repository's purpose: "GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles." To the right is the GEO logo and a search bar labeled "Keyword or GEO Accession" with a "Search" button. Below the main title, there are three columns of links: "Getting Started" (Overview, FAQ, About GEO DataSets, About GEO Profiles, About GEO2R Analysis, How to Construct a Query, How to Download Data), "Tools" (Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation, Analyze a Study with GEO2R, GEO BLAST, Programmatic Access, FTP Site), and "Browse Content" (Repository Browser, DataSets: 3848, Series: 59282, Platforms: 14769, Samples: 1539231). At the bottom, there is a section for "Information for Submitters" with links to Login to Submit, Submission Guidelines, Update Guidelines, MIAME Standards, Citing and Linking to GEO, Guidelines for Reviewers, and GEO Publications.

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/>)

The screenshot shows the homepage of the ArrayExpress functional genomics data archive. At the top, there's a navigation bar with links for Services, Research, Training, and About us. Below the navigation is a search bar with examples like "E-MEXP-31, cancer, p53, Geuvadis". The main content area features a large title "ArrayExpress – functional genomics data" and a sub-section "ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community." A "Browse ArrayExpress" link is present. To the right, there's a "Data Content" section showing statistics: "Updated today at 04:00", "58541 experiments", "1729714 assays", and "34.89 TB of archived data". Below this, there's a "Latest News" section with a recent entry about a revamped submission tool. The footer contains links for EMBL-EBI services, research, training, industry, and about us, along with copyright information.

EMBL-EBI

ArrayExpress

Services | Research | Training | About us

Search Examples: E-MEXP-31, cancer, p53, Geuvadis Advanced

Feedback | Login

ArrayExpress – functional genomics data

ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.

Browse ArrayExpress

Data Content

Updated today at 04:00

- 58541 experiments
- 1729714 assays
- 34.89 TB of archived data

Latest News

7 July 2015 - Revamped guide for ArrayExpress submission tool "Annotate"

It's been almost a year since we launched Annotate. We have listened to our submitters and rolled out a revamped guide covering many frequently asked questions. Spare a few minutes to pick up some bite-size hints to make your submission experience smoother. For example, did you know we have introduced a few time-saving features in Annotate?

Links

Information about how to search ArrayExpress, understand search results, how to submit data and FAQ can be found in our Help section.

Find out more about the Functional Genomics group.

Tools and Access

ArrayExpress Bioconductor package: an R package to access ArrayExpress and build data structures.

Programmatic access: query and download data using web services or JSON.

FTP access: data can be downloaded directly from our FTP site.

Related Projects

Discover up and down regulated genes in numerous experimental conditions in the Expression Atlas.

Explore the Experimental Factor Ontology used to support queries and annotation of ArrayExpress data.

EMBL-EBI

News
Brochures
Contact us
Intranet

Services

By topic
By name (A-Z)
Help & Support

Research

Overview
Publications
Research groups
Postdocs & PhDs

Training

Overview
Train at EBI
Train outside EBI
Train online
Contact organisers

Industry

Overview
Members Area
Workshops
SME Forum
Contact Industry programme

About us

Overview
Leadership
Funding
Background
Collaboration
Jobs
People & groups
News
Events
Visit us
Contact us

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK +44 (0)1223 49 44 44

Copyright © EMBL-EBI 2015 | EBI is an outstation of the European Molecular Biology Laboratory | Privacy | Cookies | Terms of use

Gene expression databases

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

The screenshot shows the homepage of the Expression Atlas at EBI. At the top, there's a navigation bar with links for Services, Research, Training, and About us. Below the navigation is a search bar with placeholder text "Enter gene query..." and a "Search" button. To the right of the search bar are examples of queries: ASPM, REACT_284558, ENSMUS000000021789, and "zinc finger".

The main content area has a title "Expression Atlas: Differential and Baseline Expression". Below it, a sub-section titled "iRAP: RNA-seq analysis tool" provides a brief description of the tool and its purpose. There are also sections for "Publications" (listing two academic papers), "Browse..." (with categories for Baseline Experiments, Plant Experiments, and All Experiments), and a "Search..." section with fields for Gene query, Organism (set to Homo sapiens), and Sample properties.

At the bottom, there's a footer with links for various EMBL-EBI services like News, Brochures, and Intranet, as well as Research, Training, Industry, and About us sections. The footer also includes copyright information: "EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK +44 (0)1223 49 44 44" and "Copyright © EMBL-EBI 2015 | EBI is an outstation of the European Molecular Biology Laboratory | Privacy | Cookies | Terms of use".

Microarray studies

1. Introduction
2. Microarray technology
3. Analysis
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

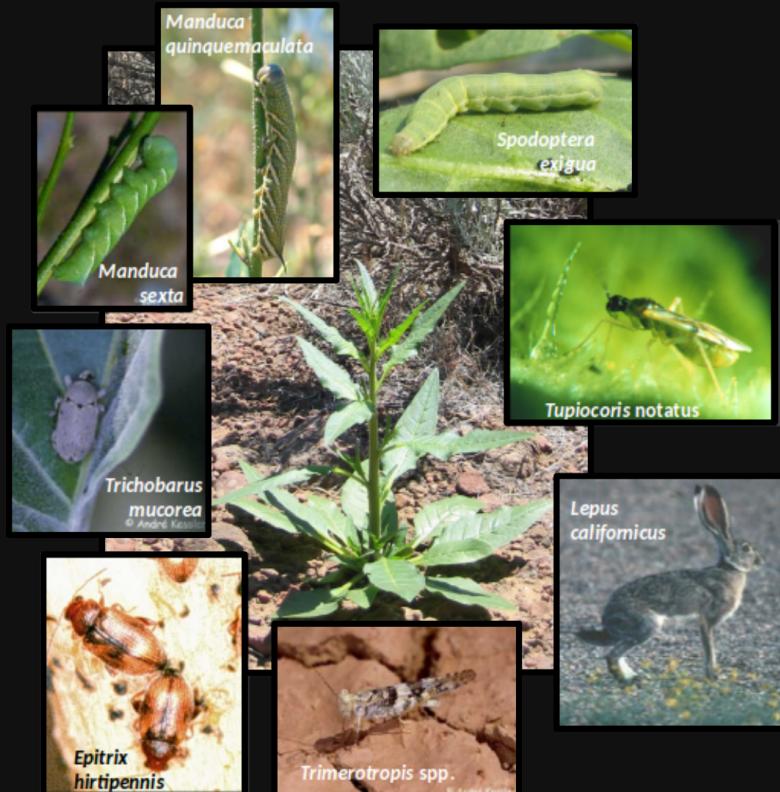
Adaptive differences in gene expression associated with heavy metal tolerance in the soil arthropod *Orchesella cincta*

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

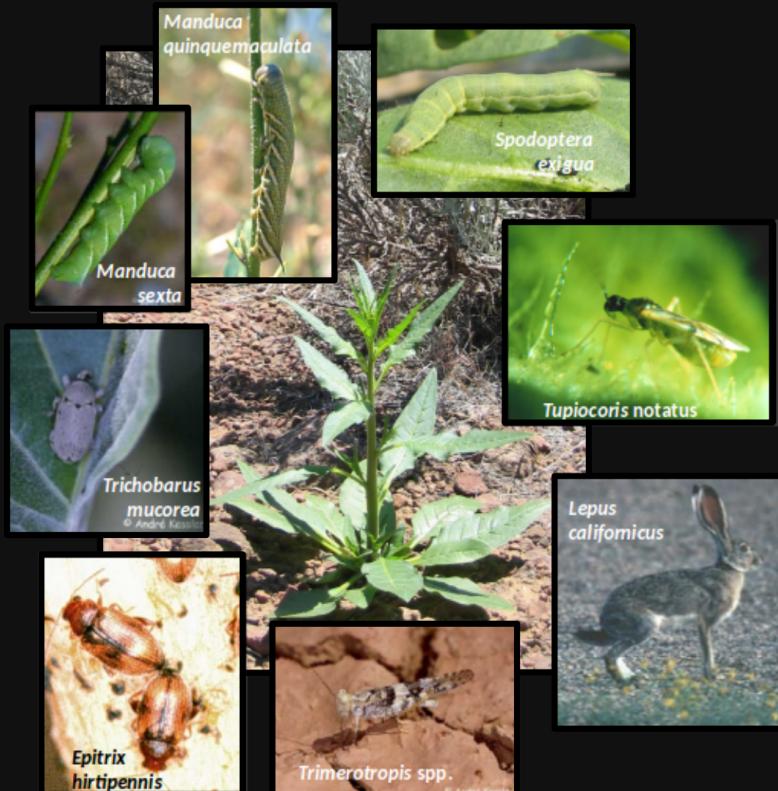
Lab case study: Herbivory in *Nicotiana attenuata* (Solanaceae)



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)
- 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 Voelckel², Markus Hartl, Silvia Schmidt, and Ian T. Baldwin*

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

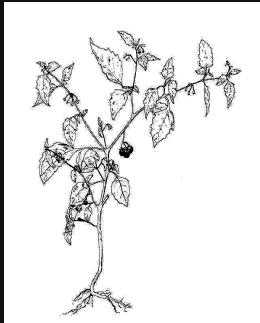
Case study - Chips, veggies & vegetarians

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



The veggies

Solanum nigrum
Black
nightshade

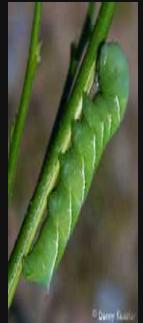


Nicotiana attenuata
Coyote tobacco



The vegetarian

Manduca sexta



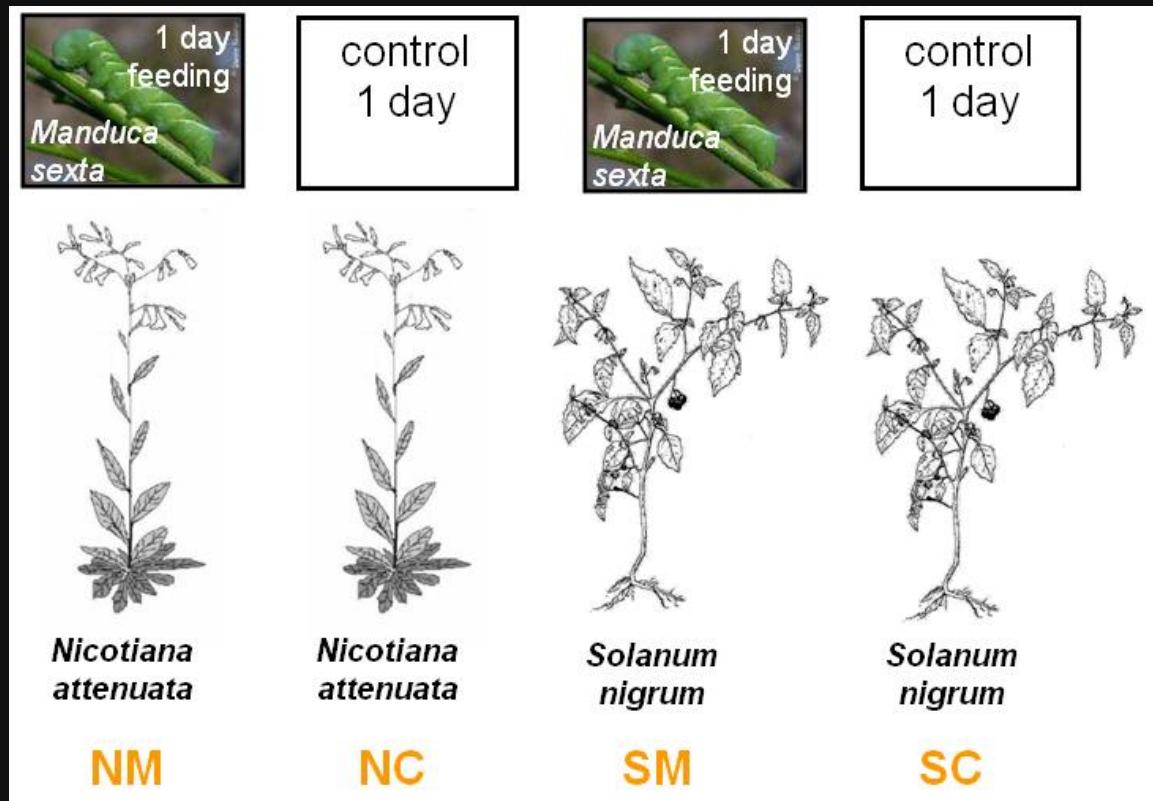
Question:

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

Microarray Case Study

RNA source

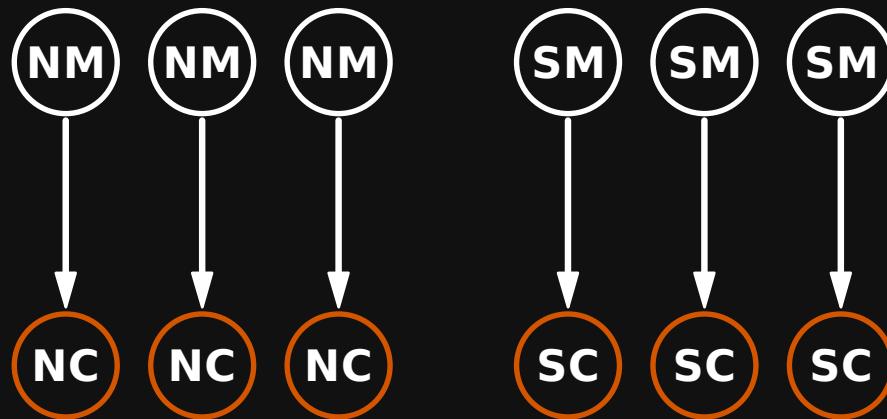
2 herbivore treatments and 2 controls



Microarray Case Study

6 arrays

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



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**

Thanks