

Advanced genetics - 203.305

Microarray - Hands-on data analysis

Dr. Pierre-Yves Dupont, Postdoctoral Researcher

Computer Biology Research Group

p.y.dupont@massey.ac.nz

Planning

- **15.9.14** – Introduction (lecture), *AHB2.38*
- **22.9.14** – Paper discussion, *AHB2.38* - **Discussion worksheet due**
- **23.9.14** – 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*)
- **29.09.14** – **Lab discussion** (feedback!) and new developments in global gene expression analysis, *AHB2.38*
- **30.09.14** – Biological interpretation of microarray data (Gene ontology analysis using the R language + online research of candidate genes, **3h lab**, *SC5.10*)

Microarray studies

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

Microarray applications

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

Expression Studies

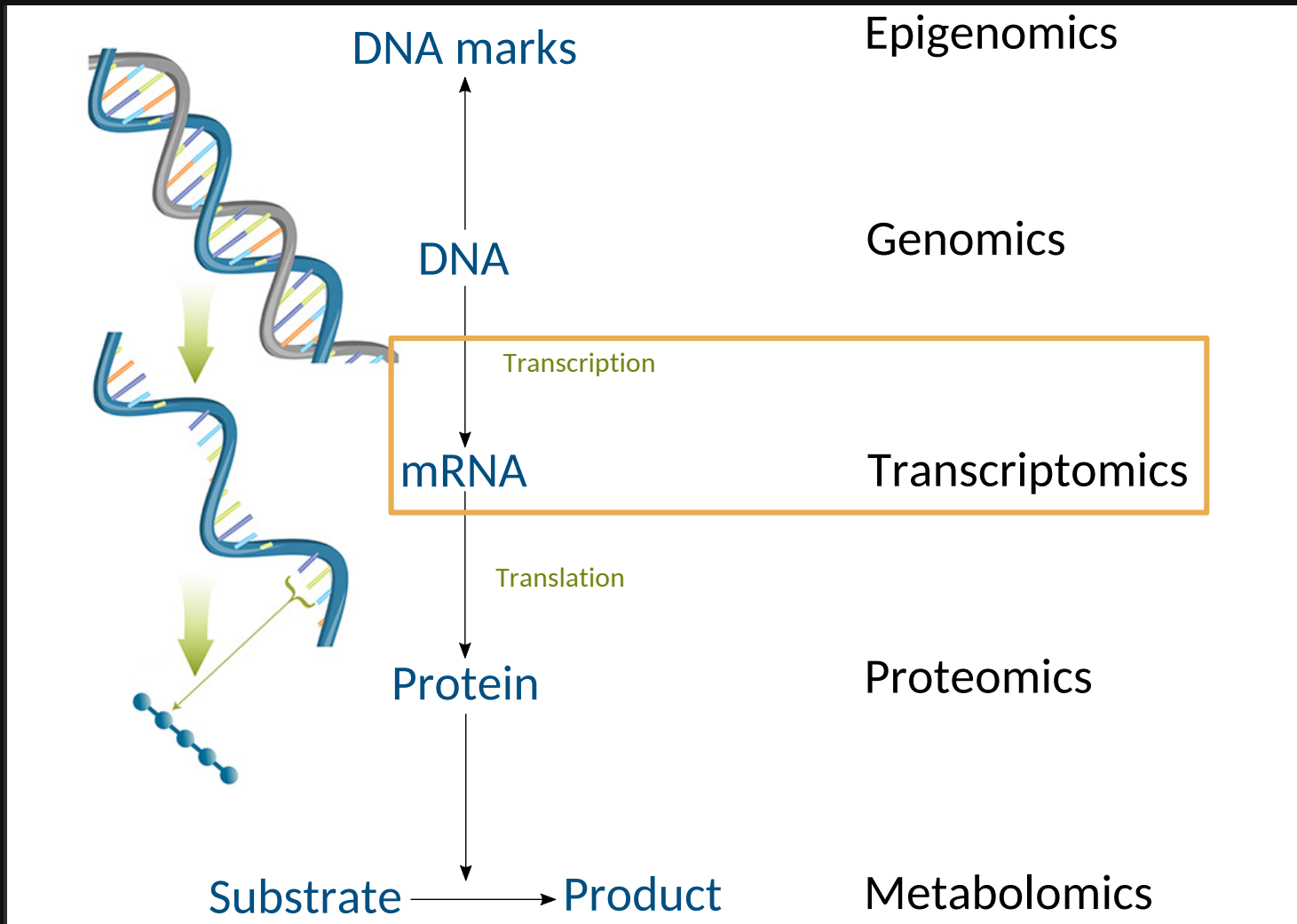


Figure modified from: Katherine Joyce, Woods Hole Oceanographic Institution

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

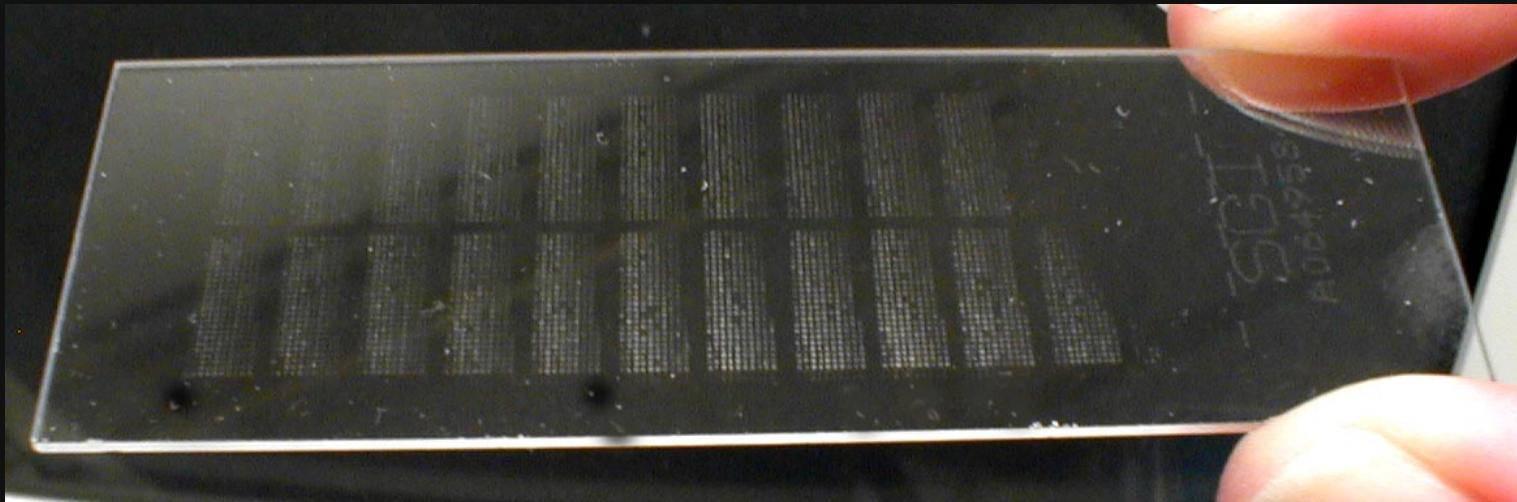
Hypothesis generating tool

Microarray studies

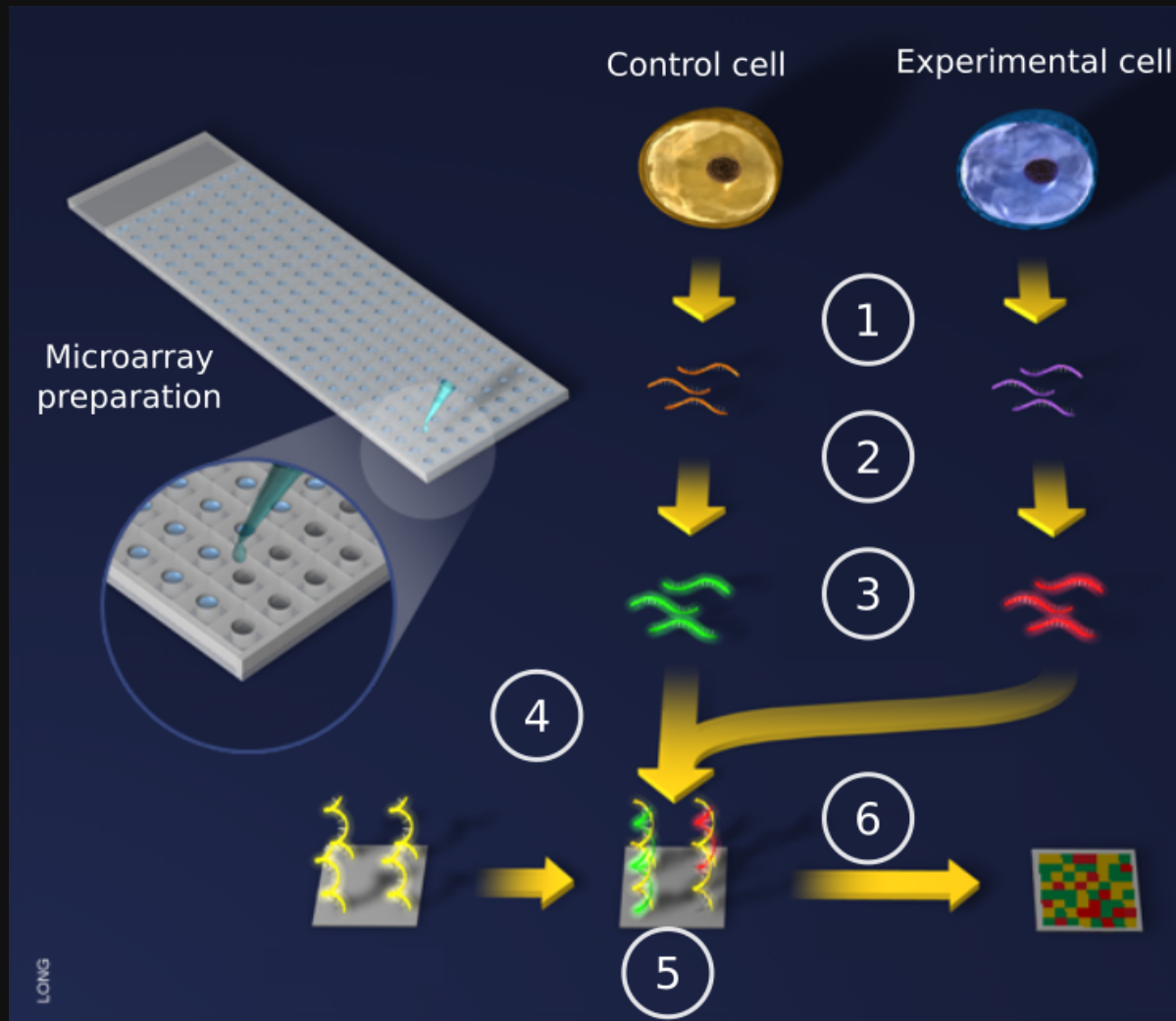
1. Introduction
2. **Microarray technology**
3. Statistics
4. Gene expression databases and MIAME
5. Examples of microarray studies (paper discussion topic and lab topic)

What are microarrays?

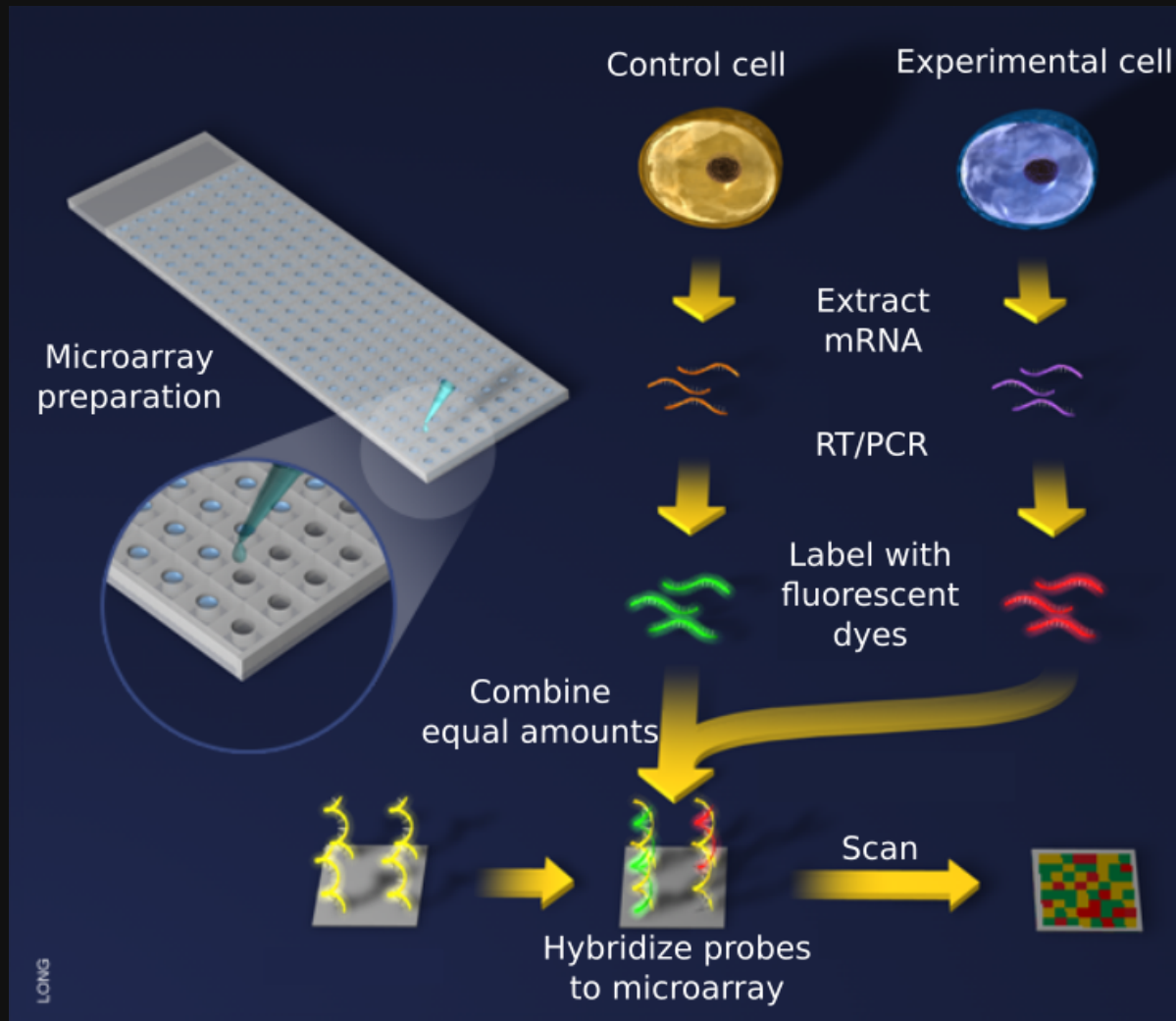
A microarray is a **solid support** (such as a membrane or glass microscope slide) on which **DNA of known sequence** is deposited in a **grid-like array**.



Microarray analysis principle



Microarray analysis principle

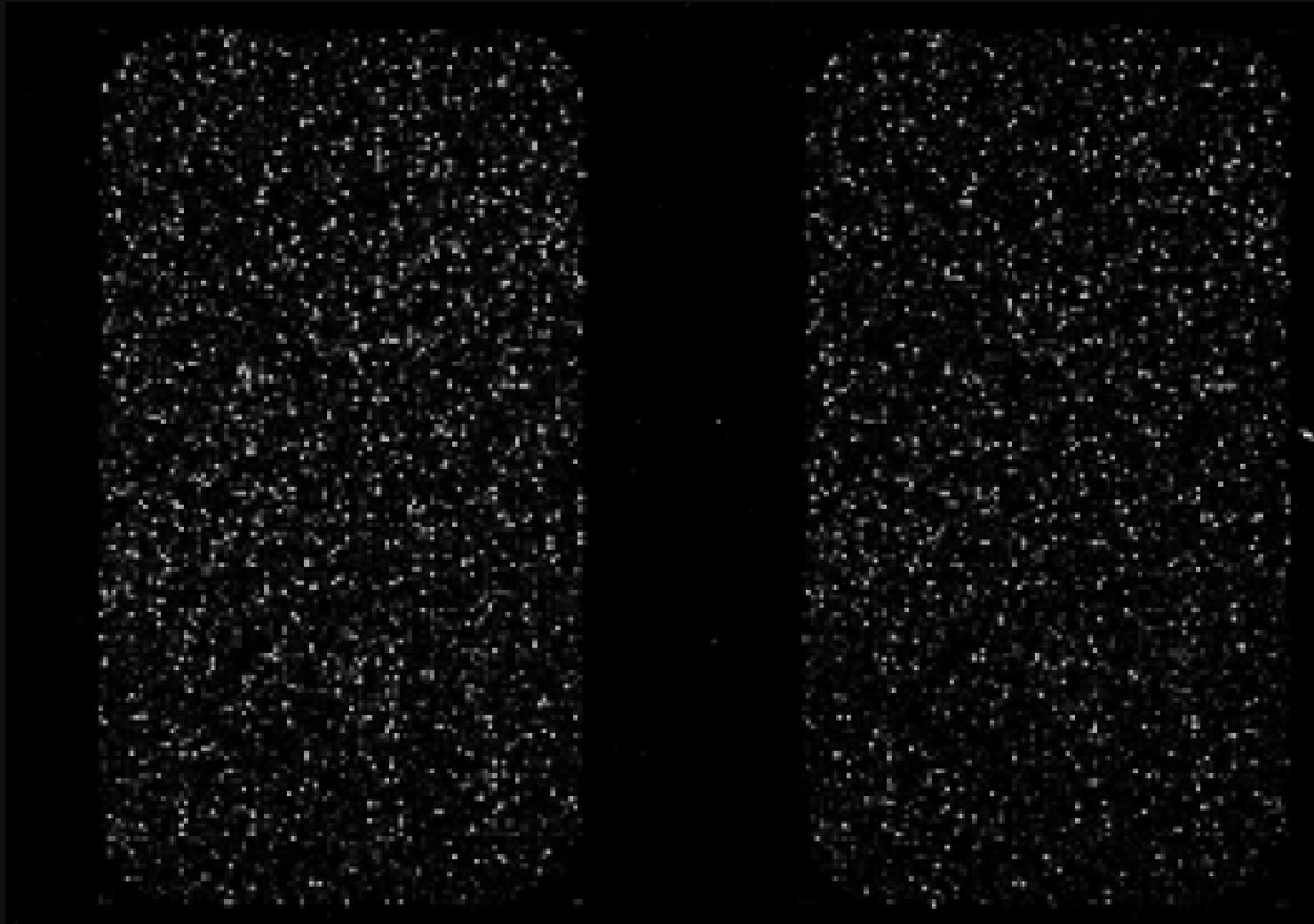


Competitive hybridization

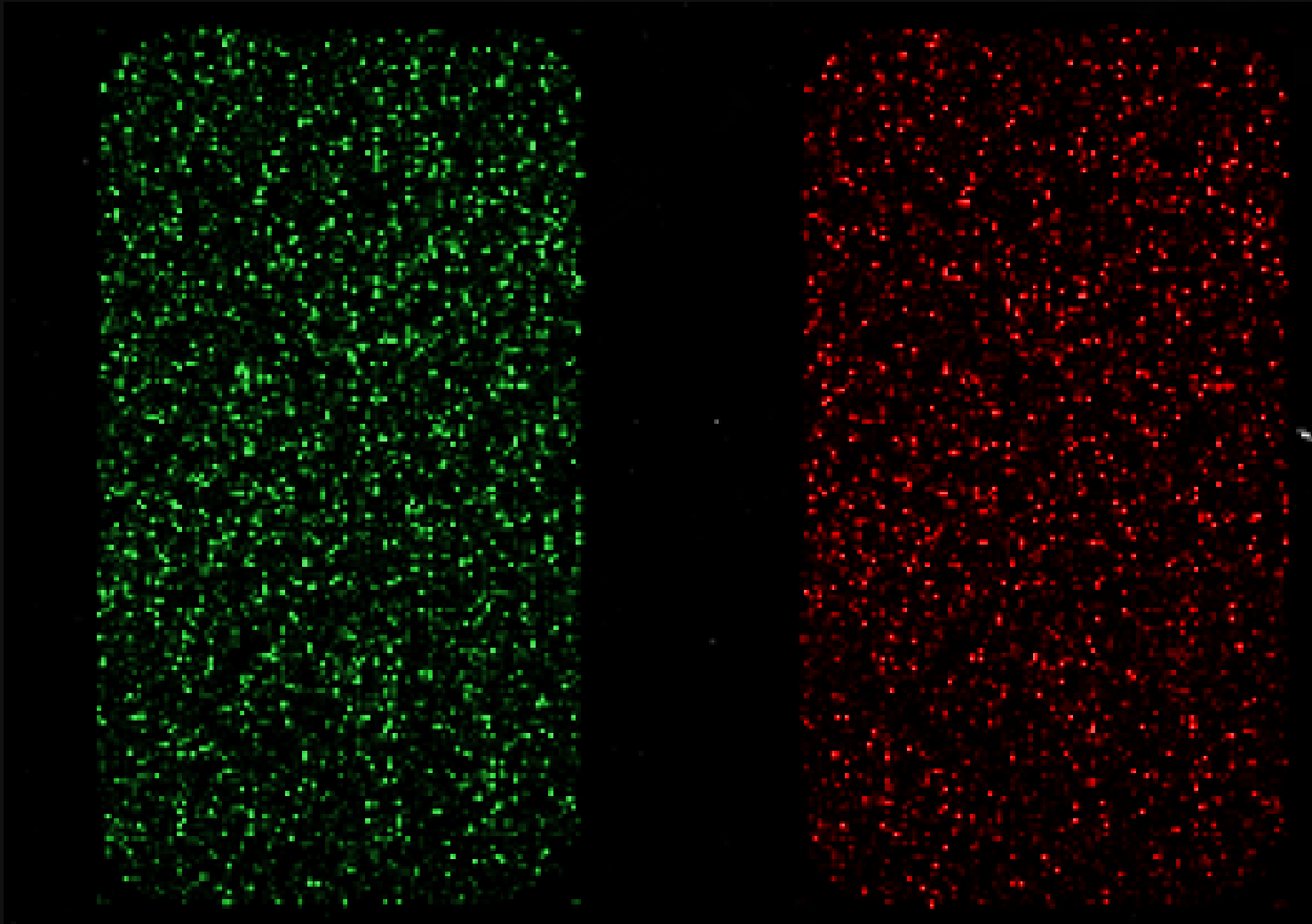
It is possible to represent **different** samples on **one** microarray using **different fluorescent molecules (fluorophores)**

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

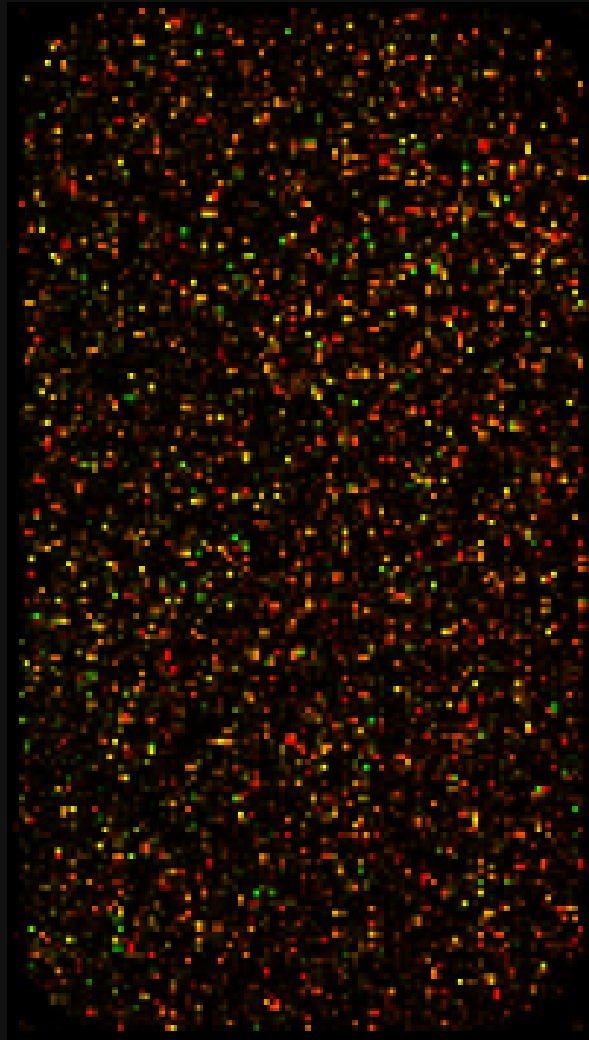
Competitive hybridization



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)

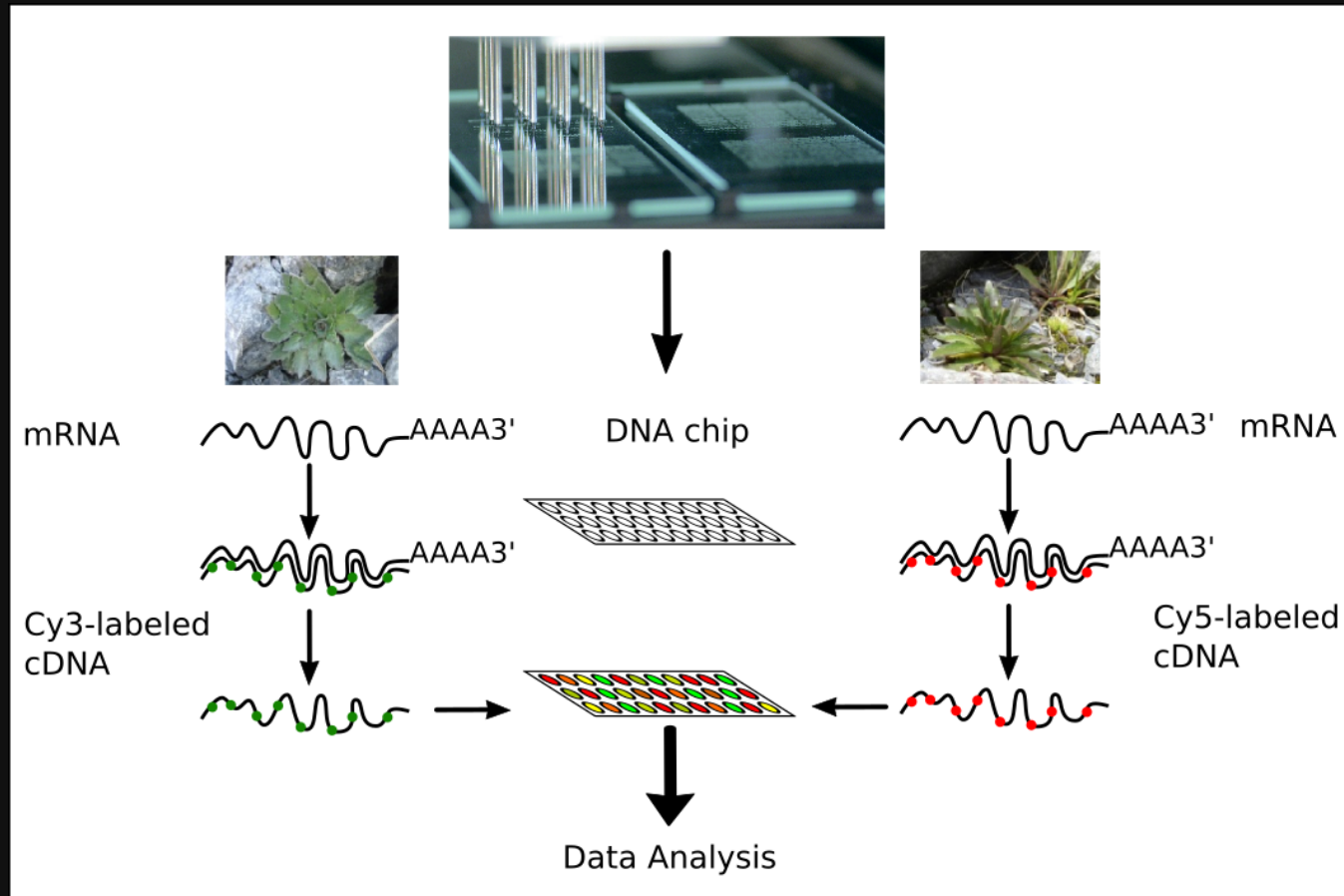
Microarray study pipeline

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

Microarray studies

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

Log Fold Ratio



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

Log Fold Ratio

- $Cy3 = Sample1$ (Green)
- $Cy5 = Sample2$ (Red)
- $Cy5 > Cy3$: higher expression in sample 2
- $Cy3 > Cy5$: higher expression in sample 1
- $M(\log \text{ fold ratio}) = \log_2\left(\frac{Cy5}{Cy3}\right)$
- $M(\log \text{ fold ratio}) = \log_2(Cy5) - \log_2(Cy3)$

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 $2^3 = 8$

$\log_2(\text{Cy5}/\text{Cy3})$	$\text{Cy5}/\text{Cy3}$
2	4
1.58	3
1	2
0.58	1.5
0	1
-0.58	0.66
-1	0.5
-1.58	0.33
-2	0.25

Hypothesis testing

T-test

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

Mean:

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

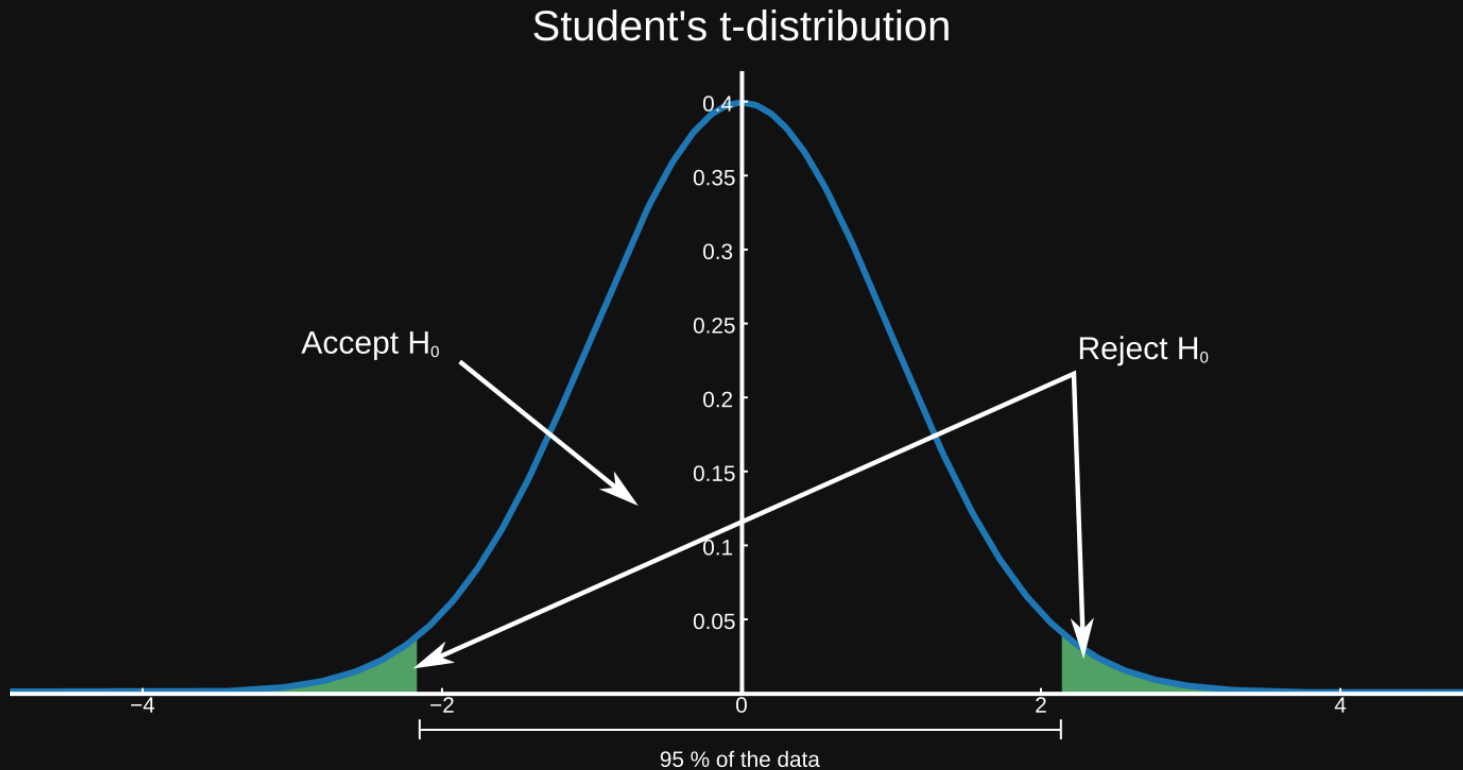
Variance:

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

T-statistic:

$$T_x = \frac{\bar{x}_{S_1} - \bar{x}_{S_2}}{\sqrt{2 \left(\frac{S_x^2 S_1}{M} + \frac{S_x^2 S_2}{N} \right)}}$$

T-test and P-value



*T-test is used only to compare two samples.
To compare more, ANOVA (ANalysis Qf
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 $p < \mathbf{0.05}$ 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.

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

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. Statistics
4. **Gene expression databases and MIAME**
5. Examples of microarray studies (paper discussion topic and lab topic)

Gene expression databases

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

 [Resources](#) [How To](#)

[GEO Home](#) [Documentation](#) [Query & Browse](#) [Email GEO](#)

[pydubont](#) [My NCBI](#) [Sign Out](#)


Gene Expression Omnibus

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](#)

Browse Content

Repository Browser

DataSets:	3848
Series: 	59282
Platforms:	14769
Samples:	1539231

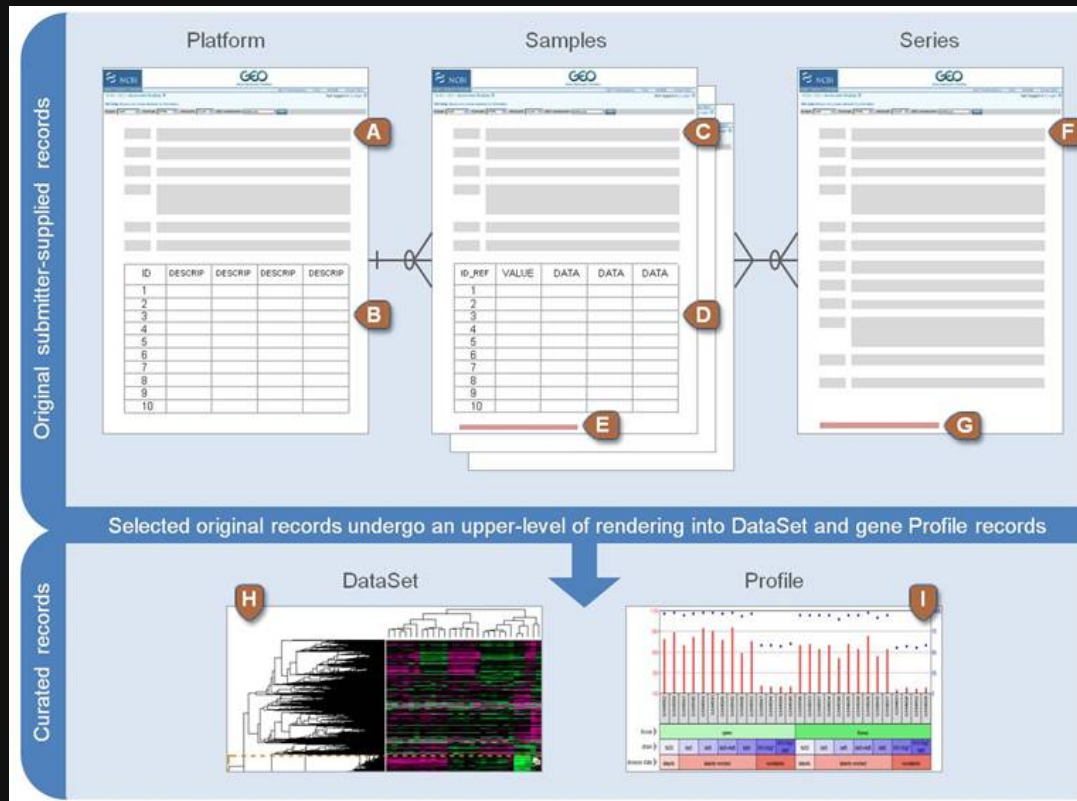
Information for Submitters

Login to Submit	Submission Guidelines	MIAME Standards
	Update Guidelines	Citing and Linking to GEO
		Guidelines for Reviewers
		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 ArrayExpress website homepage. At the top, there's a navigation bar with links for Services, Research, Training, and About us. Below this is the ArrayExpress logo and a search bar with a search button. A search example is provided: "E-MEXP-31, cancer, p53, Geuvade". Below the search bar is a secondary navigation bar with links for Home, Browse, Submit, Help, and About ArrayExpress, along with Feedback and Login links. The main content area features the title "ArrayExpress – functional genomics data" and a brief description of the database. A "Browse ArrayExpress" link is provided. To the right, there's a "Data Content" section with a bar chart icon, stating it was updated today at 04:00 and listing statistics: 58541 experiments, 1729714 assays, and 34.89 TB of archived data. Below this is a "Latest News" section with a globe icon, dated 7 July 2015, announcing a revamped guide for the "Annotare" submission tool. The page is divided into three columns: "Links" (information about searching and submitting data), "Tools and Access" (ArrayExpress Bioconductor package, Programmatic access, and FTP access), and "Related Projects" (Expression Atlas and Experimental Factor Ontology). The footer contains a grid of links for Services, Research, Training, Industry, and About us, each with a list of sub-links. At the very bottom, there's a copyright notice for EMBL-EBI, Wellcome Trust Genome Campus, and a link to the privacy policy.

EMBL-EBI

Services Research Training About us

ArrayExpress

Home Browse Submit Help About ArrayExpress

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 "Annotare"**

It's been almost a year since we launched Annotare. 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 Annotare?

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 outpost 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 Expression Atlas website. At the top, there's a navigation bar with links for Services, Research, Training, and About us. Below this is the main header with the 'Expression Atlas' logo and a search bar. The search bar contains the text 'Enter gene query...' and a 'Search' button. Below the search bar, there are examples of queries: 'ASPM, REACT_284558, ENSMUSG00000021789, "zinc finger"'. The main content area is titled 'Expression Atlas: Differential and Baseline Expression'. It includes a paragraph about the database and a 'Search...' section with a 'Gene query' input field, an 'Organism' dropdown menu (set to 'Homo sapiens'), and a 'Sample properties' input field. There are also 'Search' and 'Reset' buttons. To the right of the search section, there's a 'Browse...' section with links to 'Baseline Experiments', 'Plant Experiments', and 'All Experiments'. The bottom of the page features a footer with the EMBL-EBI logo, contact information, and a list of services, research, training, industry, and about us links.

EMBL-EBI

Services Research Training About us

Expression Atlas

Enter gene query... Search

Examples: ASPM, REACT_284558, ENSMUSG00000021789, "zinc finger"

Home Release notes FAQs Download Help About Feedback

Expression Atlas: Differential and Baseline Expression

The Expression Atlas provides information on gene expression patterns under different biological conditions. Gene expression data is re-analysed in-house to detect genes showing interesting baseline and differential expression patterns. [Read more about Expression Atlas.](#)

Search...

Gene query ? Enter gene query... Organism Homo sapiens Sample properties ? Enter condition query... Search Reset

E.g. SFTPA2, zinc finger ☒ Exact match E.g. lung, leaf, "valproic acid", cancer

IRAP: RNA-seq analysis tool

IRAP is a flexible pipeline for RNA-seq analysis that integrates many existing tools for filtering and mapping reads, quantifying expression and testing for differential expression. IRAP is used to process all RNA-seq data in Expression Atlas.

Publications

- RNA-Seq Gene Profiling - A Systematic Empirical Comparison (PLoS One, 2014).
- Expression Atlas update - a database of gene and transcript expression from microarray and sequencing-based functional genomics experiments (Nucleic Acids Research, 2014).

Browse...

- Baseline Experiments See all baseline expression data sets in Expression Atlas.
- Plant Experiments See all expression data sets in plants in Expression Atlas.
- All Experiments Scroll through the complete list of all data sets in Expression Atlas.

Still need the old Expression Atlas? [Click here.](#)

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](#)

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)

Microarray studies

1. Introduction
2. Microarray technology
3. Statistics
4. Gene expression databases and 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 STRAALLEN*

**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)