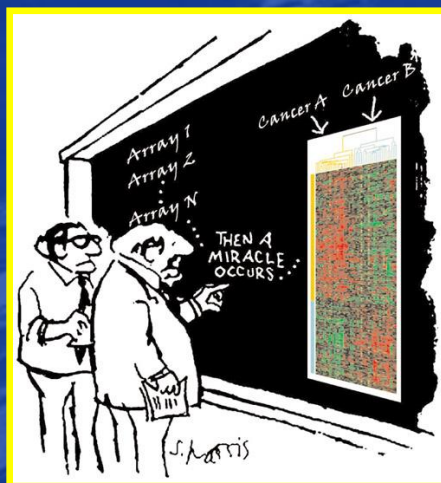


# GENE EXPRESSION MICROARRAY PROFILING IN PRACTICE

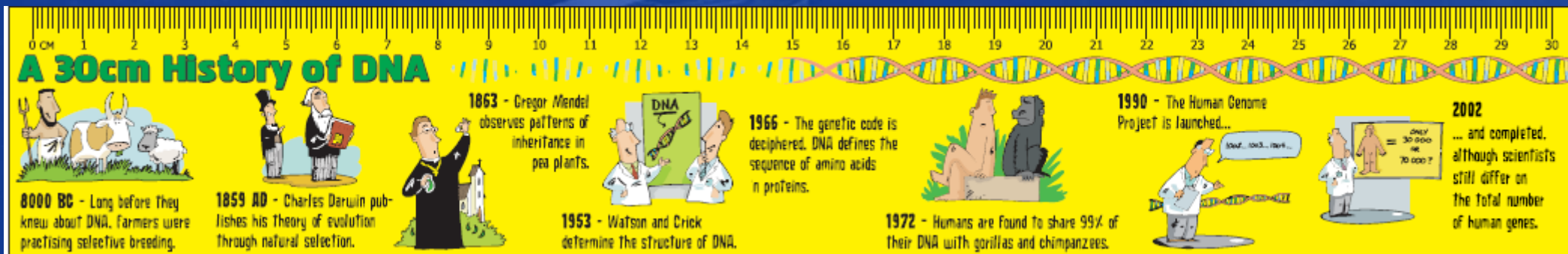


Ingrid Arijs

Translational Research Center for Gastrointestinal Disorders (TARGID), IBD Leuven, KU Leuven,  
Leuven, Belgium  
[ingrid.arijs@med.kuleuven.be](mailto:ingrid.arijs@med.kuleuven.be)

# Introduction:

- Microarray technology is the practical offshoot of the **Human Genome Project (HGP)** which is completed in 2003
- The goals of HGP were to:
  - *identify* all the approximately 20000-25000 genes in human DNA
  - *determine* the sequences of the 3 billion chemical base pairs that make up human DNA
  - *store* this information in databases
  - *improve* tools for data analysis

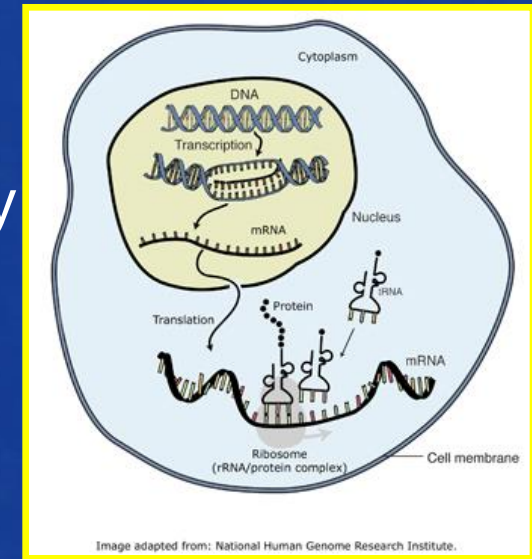


- The completion of HGP was not the end, but merely the start of the unraveling of the information hidden in the DNA

- Buried within the DNA sequences are the genes (i.e. DNA sequences that code for proteins).

**Genes (mRNA)** within the genome only come into play in a cell when they are "expressed."

The differences in gene expression among cells is what makes them unique (i.e. what distinguishes a neuron from a kidney cell).



- To study which are active and which are inactive in different cell types helps scientist to understand both how these cells function normally and how they are affected when various genes do not perform properly.
- Already since the mid-1970s, we were capable of measuring gene expression, with techniques called **Southern blotting** and later with **Northern blotting**.

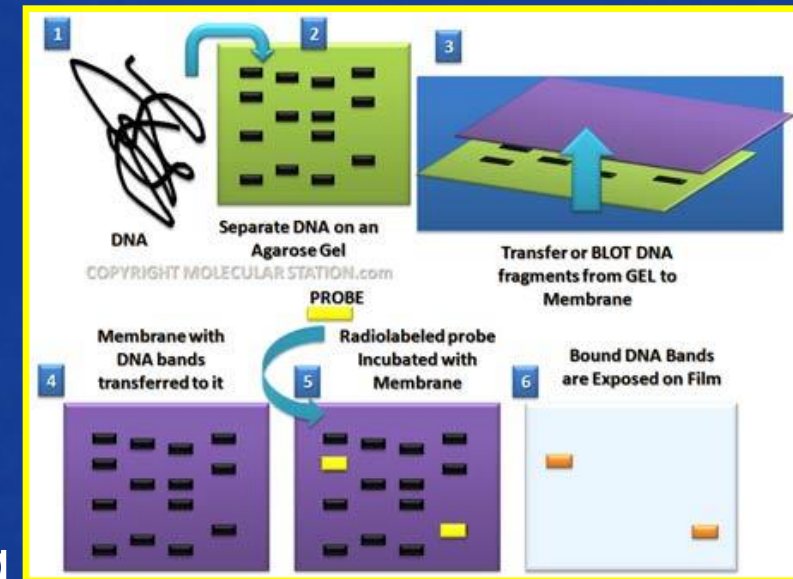


- **Southern blotting** is used to recognize a DNA sequence and uses a piece of DNA as probe of DNA, whereas **Northern blotting** uses a piece of mRNA as probe and is applied to recognize RNA sequences.

For both methods the probe is radioactively labeled and **hybridized** to a complementary fragment of DNA or RNA that has been previously separated according to molecular weight (size) by gel electrophoresis. In Northern blot analysis, the amount of radioactivity is a function of the amount of probe hybridized, which reflect the amount of mRNA in the sample.

→ But these techniques could only focus **one gene at a time**.

- With the development of **gene expression microarray technology**, we can now examine how **active thousands of genes** are at any time.



- A **microarray** can be understood as performing thousands of Southern or Northern blottings in parallel. Instead of distributing one probe over a gel with RNA or DNA, thousands of probes are attached to a solid surface, which will become the microarray, and the RNA sample is spread over these probes.
- The use of microarrays for gene expression profiling was first published in 1995 by Schena et al.

☐ 1: [Science](#). 1995 Oct 20;270(5235):467-70.

Comment in:

[Science](#). 1995 Oct 20;270(5235):368-9, 371.

#### Quantitative monitoring of gene expression patterns with a complementary DNA microarray.

[Schena M](#), [Shalon D](#), [Davis RW](#), [Brown PO](#).

Department of Biochemistry, Beckman Center, Stanford University Medical Center, CA 94305, USA.

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 Arabidopsis genes were made by means of simultaneous, two-color fluorescence hybridization.

PMID: 7569999 [PubMed - indexed for MEDLINE]

- Gene expression microarrays:

- Powerful tool for measuring simultaneously expression of thousands of genes in one single experiment
- Comprehensive picture of gene expression at tissue/cellular level
- Comparing gene expression profiles between clinical conditions
  - To identify diagnostic and prognostic biomarkers

1: [Oncol Rep.](#) 2008 Dec;20(6):1441-7.

**Overexpression of insulin-like growth factor binding protein 3 in oral squamous cell carcinoma.**

[Zhong LP](#), [Yang X](#), [Zhang L](#), [Wei KJ](#), [Pan HY](#), [Zhou XJ](#), [Li J](#), [Chen WT](#), [Zhang ZY](#).

Department of Oral and Maxillofacial Surgery, Ninth People's Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200011, P.R. China.

Previously, we established an in vitro cellular carcinogenesis model of oral squamous cell carcinoma (OSCC), including a human immortalized oral epithelial cell (HIOEC) line and its derived cancerous HB96 cell line. Further cDNA microarray analysis showed a significant up-regulated gene, insulin-like growth factor binding protein 3 (IGFBP3), accompanying with in vitro cancerization from HIOEC to HB96. In order to investigate IGFBP3 up-regulation and its potential usefulness as a molecular marker in OSCC, we detected the IGFBP3 expression with a panel of OSCC lines, and clinical samples of cancerous tissues and paired adjacent non-malignant epithelia from primary OSCC patients. Western blotting and real-time PCR showed increased IGFBP3 mRNA level and protein expression in OSCC cell lines compared with HIOEC in vitro; immunohistochemistry and real-time PCR also showed increased IGFBP3 mRNA level and protein expression in cancerous tissues compared with adjacent non-malignant epithelia from OSCC patients. Positive correlations were found between the IGFBP3 protein-positive grade in cancerous tissue and the tumor size as well as lymph node metastasis, a larger tumor size and positive lymph node metastasis indicating a higher level of IGFBP3 protein-positive grade. Based on these results, IGFBP3 may be used as a positive biomarker for OSCC development and progression.

PMID: 19020726 [PubMed - in process]



- Gene expression microarrays:

- Powerful tool for measuring simultaneously expression of thousands of genes in one single experiment
- Comprehensive picture of gene expression at tissue/cellular level
- Comparing gene expression profiles between clinical conditions
  - To classify diseases (e.g. different types of breast cancer)

1: [N Engl J Med](#). 2001 Feb 22;344(8):539-48.

Comment in:

[N Engl J Med](#). 2001 Feb 22;344(8):601-2.

[N Engl J Med](#). 2001 Jun 28;344(26):2028-9.

[N Engl J Med](#). 2001 Jun 28;344(26):2029.

#### Gene-expression profiles in hereditary breast cancer.

[Hedenfalk I](#), [Duggan D](#), [Chen Y](#), [Radmacher M](#), [Bittner M](#), [Simon R](#), [Meltzer P](#), [Gusterson B](#), [Esteller M](#), [Kallioniemi OP](#), [Wilfond B](#), [Borg A](#), [Trent J](#), [Raffeld M](#), [Yakhini Z](#), [Ben-Dor A](#), [Dougherty E](#), [Kononen J](#), [Bubendorf L](#), [Fehrle W](#), [Pittaluga S](#), [Gruvberger S](#), [Loman N](#), [Johannsson O](#), [Olsson H](#), [Sauter G](#).

Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892-4470, USA.

**BACKGROUND:** Many cases of hereditary breast cancer are due to mutations in either the BRCA1 or the BRCA2 gene. The histopathological changes in these cancers are often characteristic of the mutant gene. We hypothesized that the genes expressed by these two types of tumors are also distinctive, perhaps allowing us to identify cases of hereditary breast cancer on the basis of gene-expression profiles. **METHODS:** RNA from samples of primary tumor from seven carriers of the BRCA1 mutation, seven carriers of the BRCA2 mutation, and seven patients with sporadic cases of breast cancer was compared with a microarray of 6512 complementary DNA clones of 5361 genes. Statistical analyses were used to identify a set of genes that could distinguish the BRCA1 genotype from the BRCA2 genotype. **RESULTS:** Permutation analysis of multivariate classification functions established that the gene-expression profiles of tumors with BRCA1 mutations, tumors with BRCA2 mutations, and sporadic tumors differed significantly from each other. An analysis of variance between the levels of gene expression and the genotype of the samples identified 176 genes that were differentially expressed in tumors with BRCA1 mutations and tumors with BRCA2 mutations. Given the known properties of some of the genes in this panel, our findings indicate that there are functional differences between breast tumors with BRCA1 mutations and those with BRCA2 mutations. **CONCLUSIONS:** Significantly different groups of genes are expressed by breast cancers with BRCA1 mutations and breast cancers with BRCA2 mutations. Our results suggest that a heritable mutation influences the gene-expression profile of the cancer.

PMID: 11207349 [PubMed - indexed for MEDLINE]

- Gene expression microarrays:

- Powerful tool for measuring simultaneously expression of thousands of genes in one single experiment
- Comprehensive picture of gene expression at tissue/cellular level
- Comparing gene expression profiles between clinical conditions

- To monitor response to therapy

Gut. 2009 Dec;58(12):1612-9. Epub 2009 Aug 20.

**Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis.**

Arijs I, Li K, Toedter G, Quintens R, Van Lommel L, Van Steen K, Leemans P, De Hertogh G, Lemaire K, Ferrante M, Schnitzler F, Thorrez L, Ma K, Song XY, Marano C, Van Assche G, Vermeire S, Geboes K, Schuit F, Baribaud F, Rutgeerts P.

Department of Gastroenterology, University of Hospital Gasthuisberg, Herestraat 49, B-3000 Leuven, Belgium.

**Abstract**

**BACKGROUND AND AIMS:** Infliximab is an effective treatment for ulcerative colitis with over 60% of patients responding to treatment and up to 30% reaching remission. The mechanism of resistance to anti-tumour necrosis factor alpha (anti-TNFalpha) is unknown. This study used colonic mucosal gene expression to provide a predictive response signature for infliximab treatment in ulcerative colitis.

**METHODS:** Two cohorts of patients who received their first treatment with infliximab for refractory ulcerative colitis were studied. Response to infliximab was defined as endoscopic and histological healing. Total RNA from pre-treatment colonic mucosal biopsies was analysed with Affymetrix Human Genome U133 Plus 2.0 Arrays. Quantitative RT-PCR was used to confirm microarray data.

**RESULTS:** For predicting response to infliximab treatment, pre-treatment colonic mucosal expression profiles were compared for responders and non-responders. Comparative analysis identified 179 differentially expressed probe sets in cohort A and 361 in cohort B with an overlap of 74 probe sets, representing 53 known genes, between both analyses. Comparative analysis of both cohorts combined, yielded 212 differentially expressed probe sets. The top five differentially expressed genes in a combined analysis of both cohorts were osteoprotegerin, stanniocalcin-1, prostaglandin-endoperoxide synthase 2, interleukin 13 receptor alpha 2 and interleukin 11. All proteins encoded by these genes are involved in the adaptive immune response. These markers separated responders from non-responders with 95% sensitivity and 85% specificity.

**CONCLUSION:** Gene array studies of ulcerative colitis mucosal biopsies identified predictive panels of genes for (non-)response to infliximab. Further study of the pathways involved should allow a better understanding of the mechanisms of resistance to infliximab therapy in ulcerative colitis. ClinicalTrials.gov number, NCT00639821.

**Comment in**

Mucosal gene expression signatures that predict response of ulcerative colitis to infliximab. [Gastroenterology. 2011]



- Gene expression microarrays:

- Powerful tool for measuring simultaneously expression of thousands of genes in one single experiment
- Comprehensive picture of gene expression at tissue/cellular level
- Comparing gene expression profiles between clinical conditions
- To understand mechanisms in the pathogenesis of diseases

PLoS One. 2009 Nov 24;4(11):e7984.

**Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment.**

Arijs I, De Hertogh G, Lemaire K, Quintens R, Van Lommel L, Van Steen K, Leemans P, Cleynen I, Van Assche G, Vermeire S, Geboes K, Schuit F, Rutgeerts P.

Department of Gastroenterology, University Hospital Gasthuisberg, Leuven, Belgium.

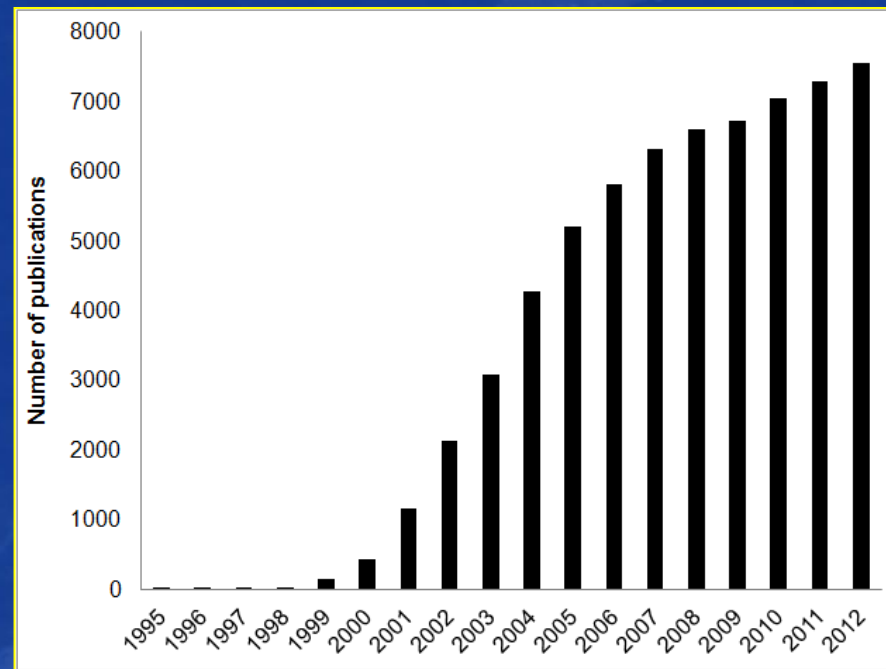
**Abstract**

**BACKGROUND:** Antimicrobial peptides (AMPs) protect the host intestinal mucosa against microorganisms. Abnormal expression of defensins was shown in inflammatory bowel disease (IBD), but it is not clear whether this is a primary defect. We investigated the impact of anti-inflammatory therapy with infliximab on the mucosal gene expression of AMPs in IBD.

**METHODOLOGY/PRINCIPAL FINDINGS:** Mucosal gene expression of 81 AMPs was assessed in 61 IBD patients before and 4-6 weeks after their first infliximab infusion and in 12 control patients, using Affymetrix arrays. Quantitative real-time reverse-transcription PCR and immunohistochemistry were used to confirm microarray data. The dysregulation of many AMPs in colonic IBD in comparison with control colons was widely restored by infliximab therapy, and only DEFB1 expression remained significantly decreased after therapy in the colonic mucosa of IBD responders to infliximab. In ileal Crohn's disease (CD), expression of two neuropeptides with antimicrobial activity, PYY and CHGB, was significantly decreased before therapy compared to control ileums, and ileal PYY expression remained significantly decreased after therapy in CD responders. Expression of the downregulated AMPs before and after treatment (DEFB1 and PYY) correlated with villin 1 expression, a gut epithelial cell marker, indicating that the decrease is a consequence of epithelial damage.

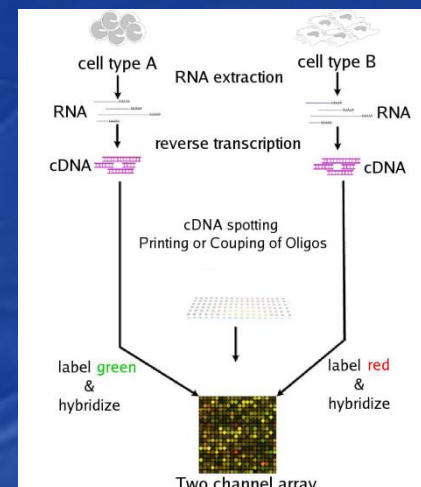
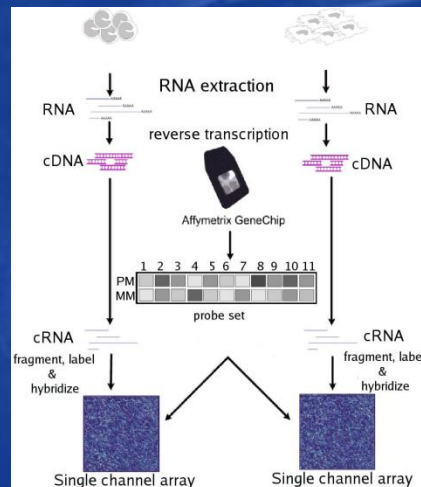
**CONCLUSIONS/SIGNIFICANCE:** Our study shows that the dysregulation of AMPs in IBD mucosa is the consequence of inflammation, but may be responsible for perpetuation of inflammation due to ineffective clearance of microorganisms.

- Microarrays have become an important research tool for **functional genomics** (= deconstruction of the genome to determine the biological function of genes and gene interactions) and will lead to new discoveries in clinical medicine. **The use of microarrays grows exponentially.**



**Figure: Number of publications on microarray related research.** The histogram shows a rapid increase in number of publications involving microarrays in the last 18 years. The numbers correspond to the number of publications containing 'microarray', 'microarrays', 'micro-array', 'micro-arrays' in the titles or abstracts as stored in the pubmed database ([www.pubmed.gov](http://www.pubmed.gov))

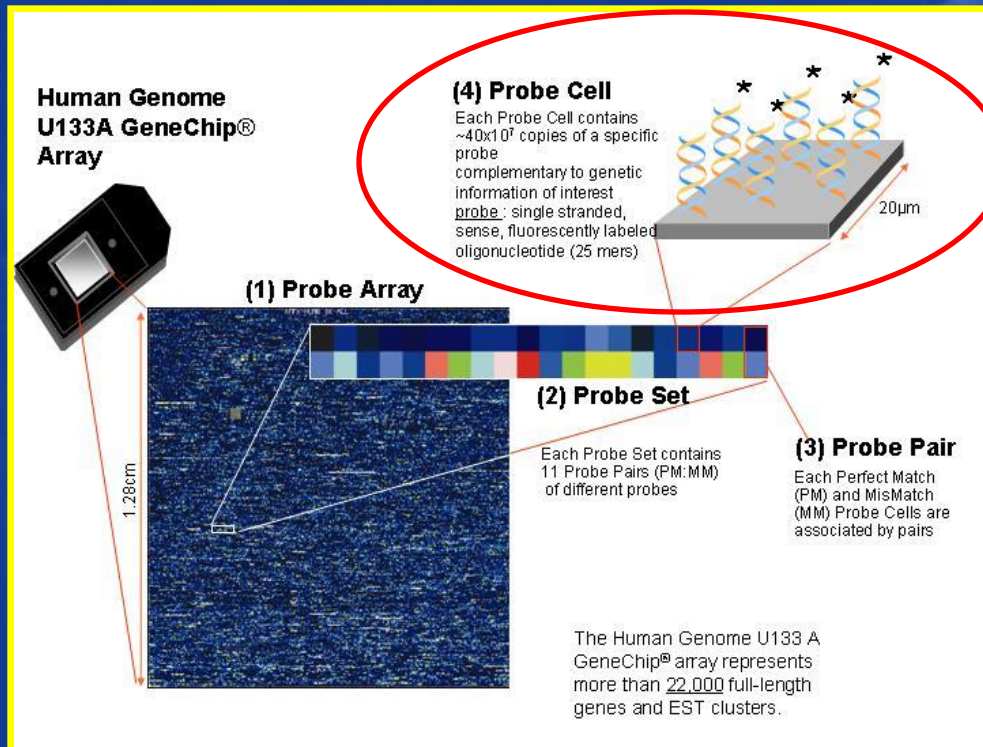
- Types of microarrays: gene expression microarrays can be broadly classified according to at least three criteria
  - Length of the probes**
    - cDNA probe (PCR products)
    - Oligos (oligonucleotides)
  - Manufacturing method**
    - spotting on a glass slide
    - in-situ* synthesis on a glass slide
  - Number of samples that can be simultaneously profiled on one array**
    - single channel arrays: only one sample is hybridized to the arrays labelled with one dye
    - dual channel arrays: two samples are hybridized to the arrays, each labelled with a different dye, this allows the simultaneous measurement of two samples



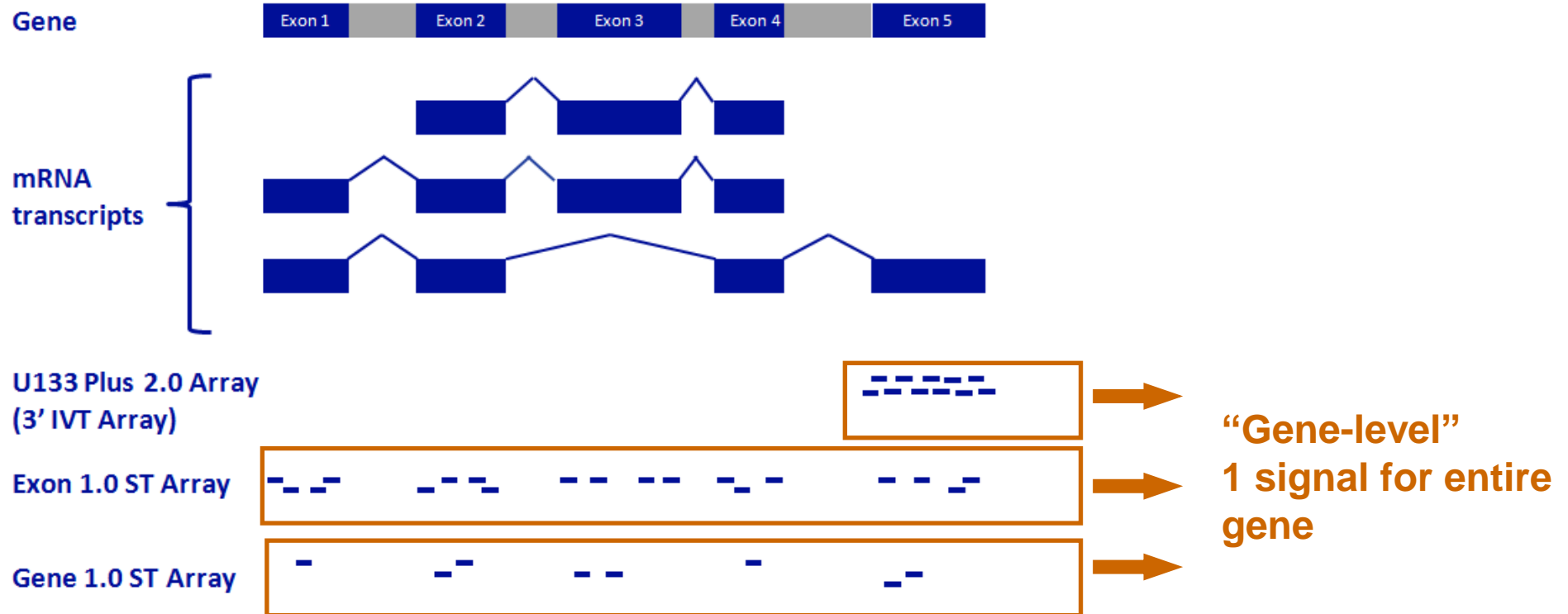


- Types of microarrays: Microarrays can be broadly classified according to at least three criteria
  - Length of the probes
    - cDNA probe (PCR products) (cDNA arrays)
    - Oligos (oligonucleotides) (oligonucleotide arrays)
  - Manufacturing method
    - spotting on a glass slide (cDNA arrays)
    - *in-situ* synthesis on a glass slide (oligonucleotide arrays)
  - Number of samples that can be simultaneously profiled on one array
    - single channel arrays: only one sample is hybridized to the arrays labelled with one dye (oligonucleotide arrays)
    - dual channel arrays: two samples are hybridized to the arrays, each labelled with a different dye, this allows the simultaneous measurement of two samples (cDNA arrays)
- Two basic variations of high-density DNA arrays:
  - cDNA microarrays: robotic spotting and immobilization of purified cDNA inserts to glass slides
  - oligonucleotide arrays (Affymetrix GeneChip array): *in-situ* synthesis of oligonucleotides on silica wafers or “chips”

- Focus today on **Affymetrix gene expression arrays**:
  - These arrays are based on the use of a **single colour label**. The chips use **oligonucleotides** (which are short fragments of a single-stranded DNA) as probes to achieve higher packing density and less manufacturing errors. One feature or probe cell is composed of a large number of identical oligonucleotides of 25 bases that are synthesized onto the chip.

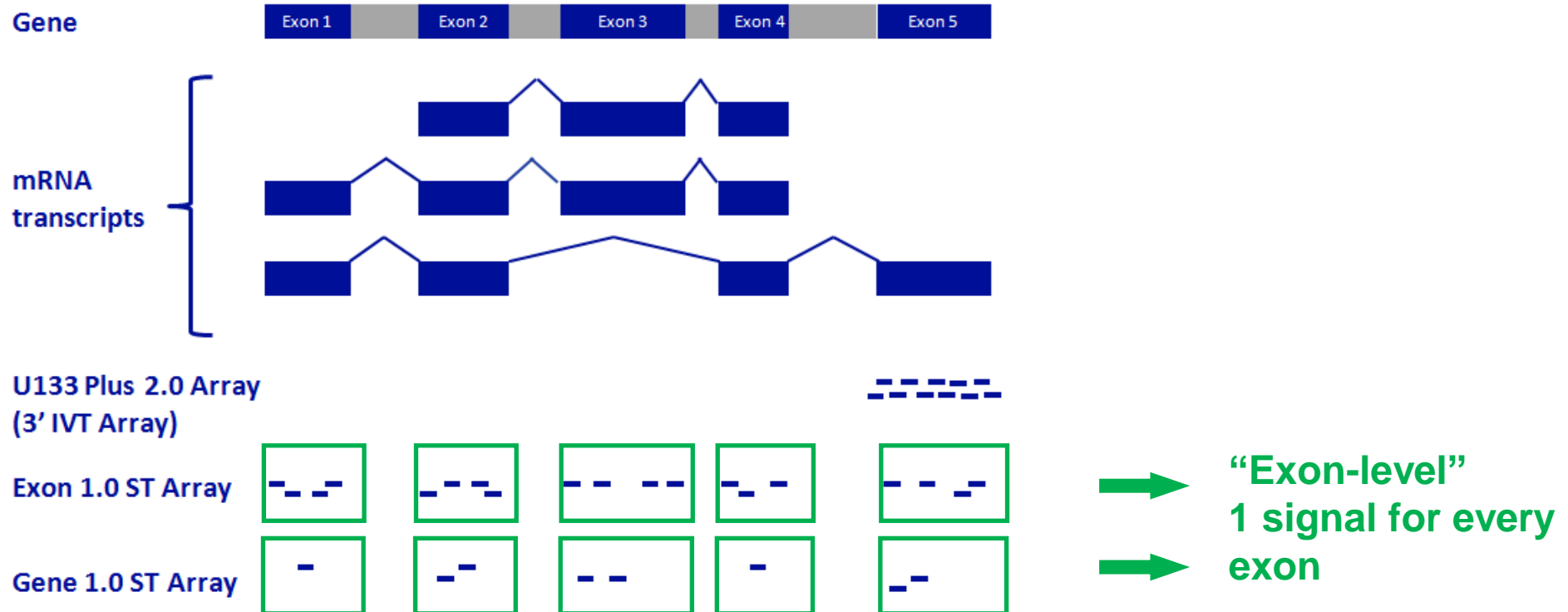


- Different types of Affymetrix gene expression microarrays:

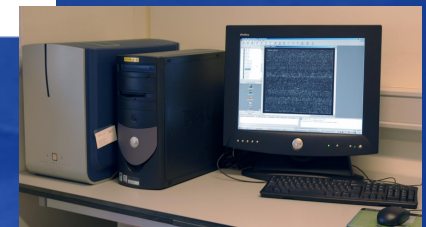
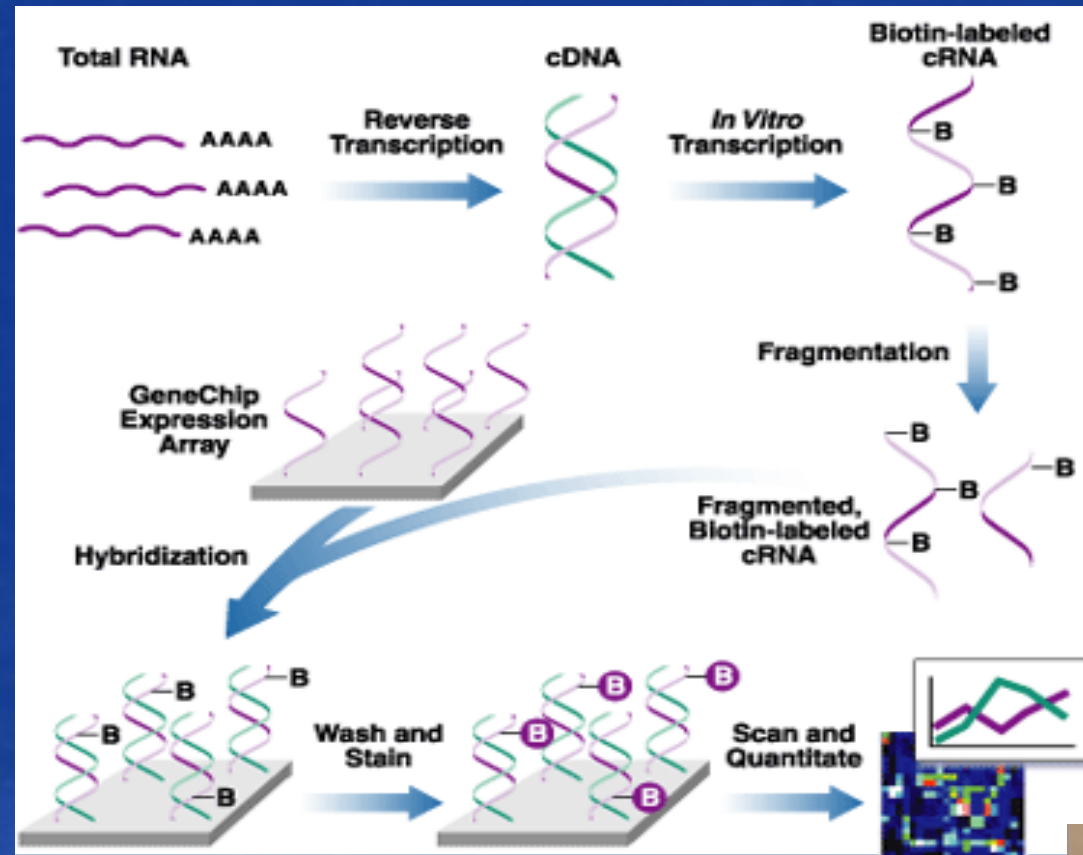




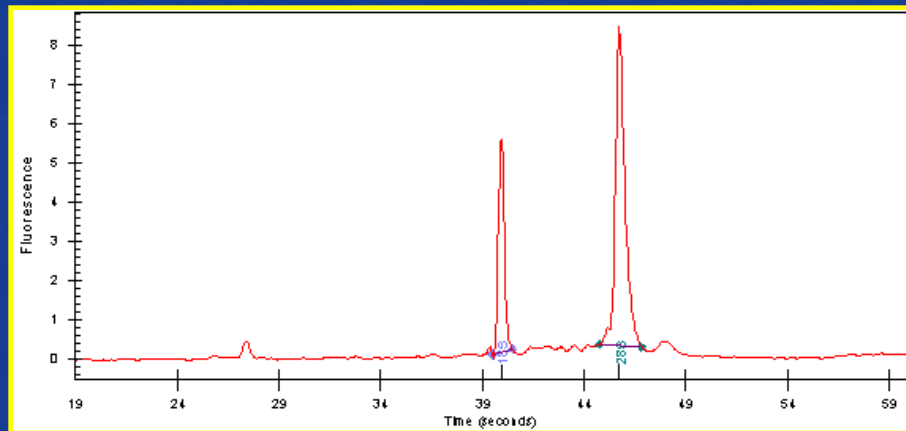
- Different types of Affymetrix gene expression microarrays offering whole-genome wide gene expression analysis:



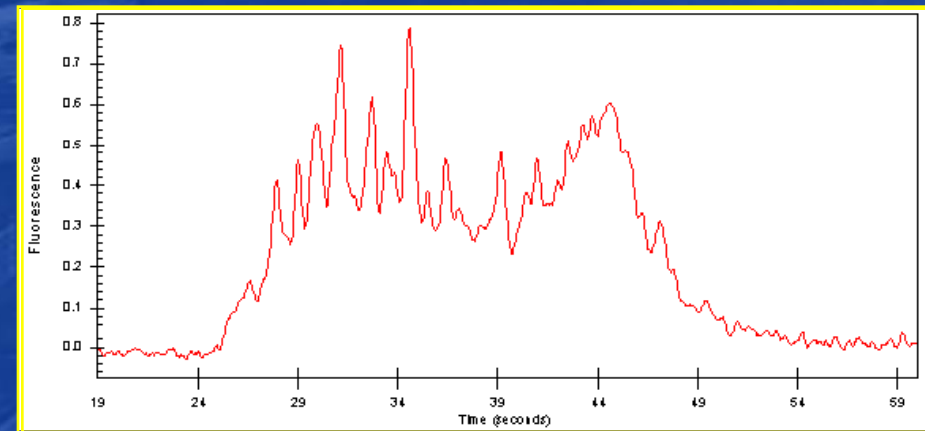
- Schematic overview on the different steps while performing an Affymetrix GeneChip experiment



## RNA quality control: Agilent bioanalyzer 2100



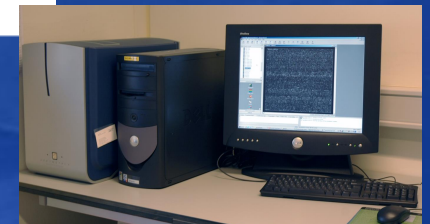
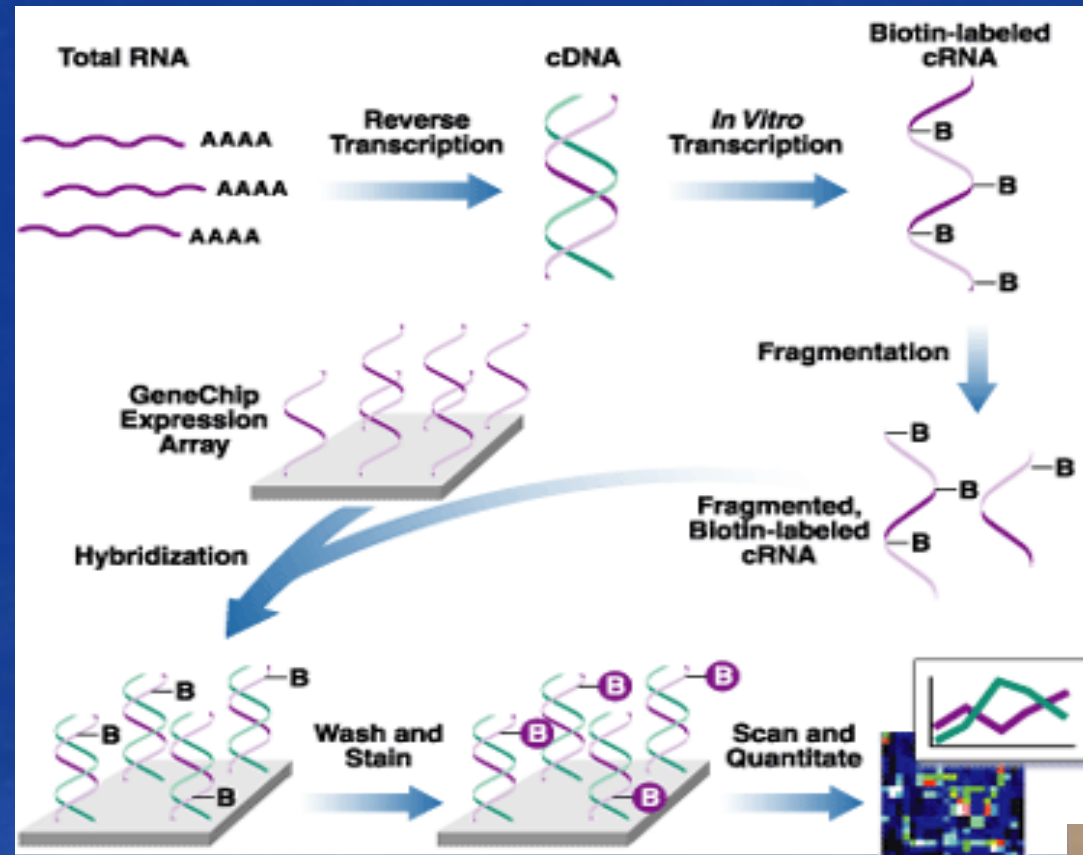
Good RNA sample quality measured with the bioanalyzer



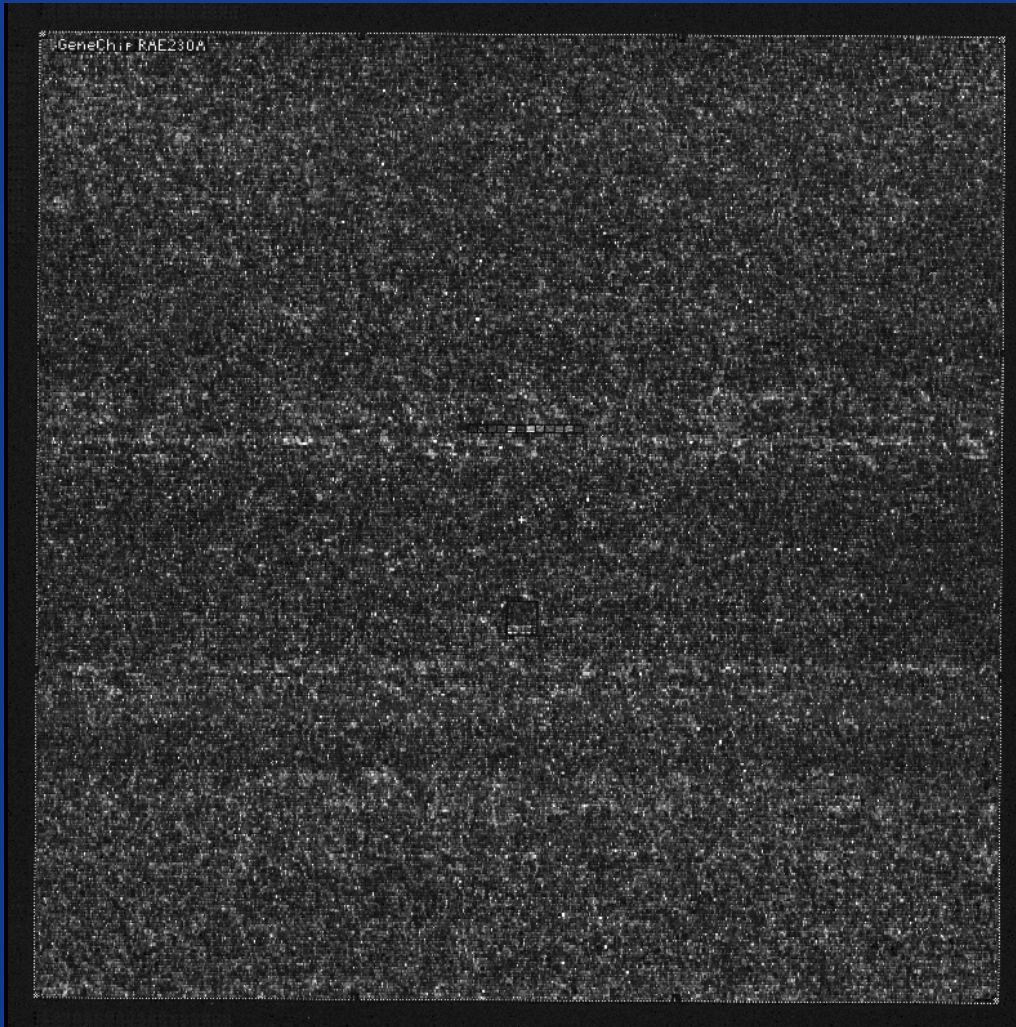
Degraded RNA sample quality measured with the bioanalyzer



- Schematic overview on the different steps while performing an Affymetrix GeneChip experiment



- The result is a **DAT file**: the image of a scanned probe array, analysed with Affymetrix GeneChip® Operating (GCOS) Software

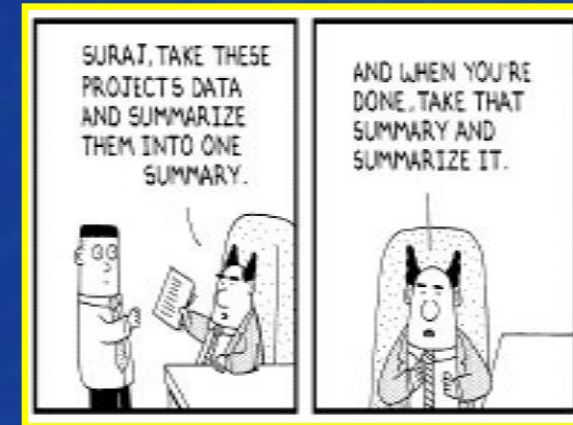
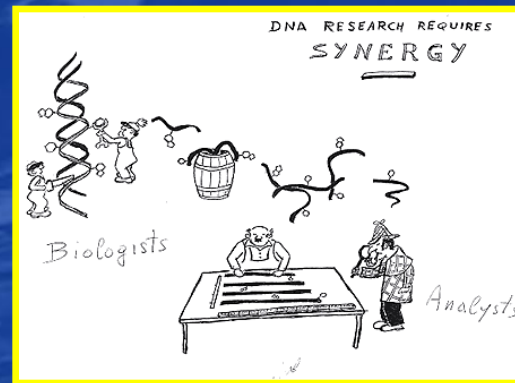




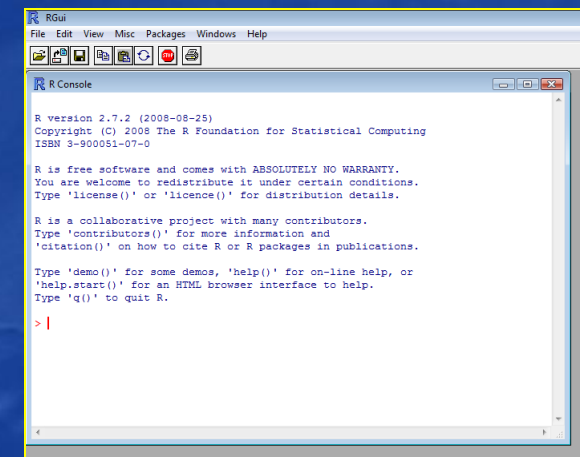
# Microarray data analysis process

## from raw data to biological significance

- Microarray experiments generate **enormous amount of data**. This is a biologist's nightmare, however it is a bioinformatician's dream.



- Focus today on the data analysis of **Affymetrix Human Genome U133 Plus 2.0 arrays** with **Bioconductor software**
  - A free, open source and open development software project for the analysis and comprehension of genomic data. It is based on the statistical R programming language.
  - <http://www.bioconductor.org/>





Biological question



Experimental design



Microarray experiment



Image analysis



Normalization



Data analysis



Biological verification  
and interpretation



Making data public



Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

- 3 major types of applications of microarrays in medicine:
  - **Class comparison** (differential expression) study:
    - It involves the comparison of gene expression profiles of samples from distinct, predefined groups to identify the genes that are differentially expressed (DE) among the groups.
      - e.g. finding DE genes in the colon from normal patients and patients with ulcerative colitis.
  - **Class prediction** (classification) study:
    - It develops a classifier that uses the gene expression level of multiple genes that can be applied to the expression profile of a newly sample to predict its (unknown) class).
      - e.g. a classifier that distinguishes between 2 different disease states
  - **Class discovery** (clustering, unsupervised learning) study:
    - It involves analyzing a given set of gene expression profiles with the goal of discovering subgroups that share common features.
    - It differs from the other studies because the classes are not predefined

Biological question



Experimental design



Microarray experiment



Image analysis



Normalization



Data analysis



Biological verification  
and interpretation



Making data public

- Type of experiment:
  - Two groups
    - Control vs treated
  - Three or more groups, single factor
    - Time series
  - Four or more groups, multiple factors
    - Time series with control and treated cells

***The type of experiment and number of groups and factors will determine the statistical methods needed to detect differential expression***

- Replicates
  - The more the better, but at least 3
  - Biological better than technical
  - Technical replication is in which the same biological sample is assayed several times
  - Biological replication refers to measuring multiple independent biological samples for each category of interest.

- Affymetrix gene expression array experiments:

Biological question

Experimental design

Microarray experiment

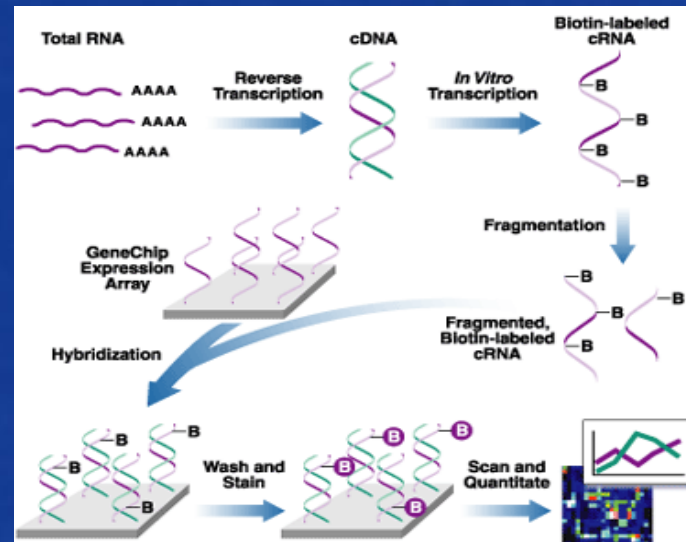
Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public



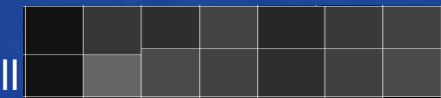
DAT files

The image of a scanned probe array



CEL files = processed .DAT file

It contains a single intensity value for each probe cell delineated by the grid



For normalization and data analysis, the CEL files are loaded in Bioconductor



Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

- Normalisation:

- Goals:

- to compensate for technical differences between chips, to see more clearly the biological differences between samples

- Sources of technical variations:

- Differences in the labelling
    - Differences in the sample preparation
    - Differences in the hybridization
    - Differences in the photo-detection
    - ....

- Allows comparisons across arrays

- Transform intensity values to expression values:

- Algorithms:

- MAS5
      - **RMA** (robust multichip average)
      - GCRMA
      - ....

- I will discuss in more detail the steps in the RMA algorithm

Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

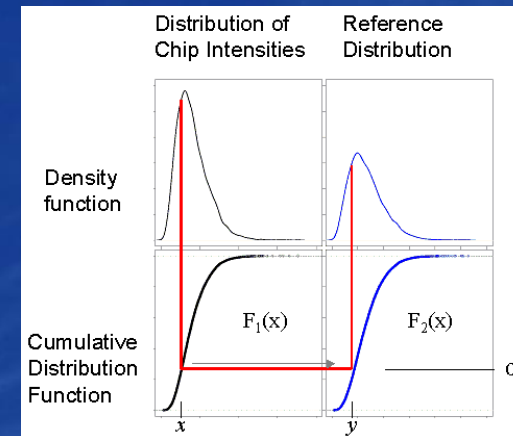
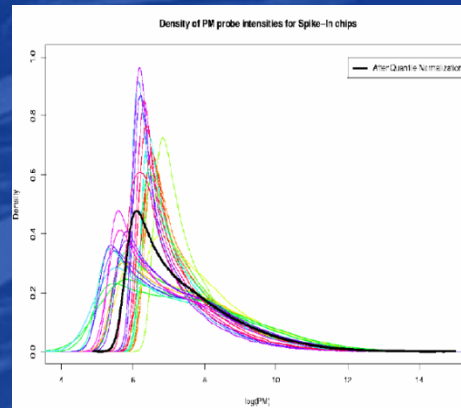
Data analysis

Biological verification  
and interpretation

Making data public

- **RMA** method:

- Robust Multichip Average method\* = a data pre-processing method to compute expression summary values for each probe set
- Implemented in the package Affy in the Bioconductor software
- Consist of 3 steps:
  - Probe-specific **background correction** to compensate for non-specific binding using PM distribution rather than PM-MM values
  - Probe-level multichip **quantile normalization** to unify PM distributions across all chips
  - Fit all the chips to the same distribution



- Robust probe-set **summary** of the log-normalized probe-level data by median polishing

- Class comparison experiment:
  - e.g. to identify the genes that are DE between two groups
    - The “null hypothesis” is that a given gene on the array is not DE between the two conditions under study
    - The “alternative hypothesis” is that the expression level of that gene is different between the two conditions
    - The hypothesis testing is performed by calculating e.g. t-test on the expression values of the gene measured in the two groups:

Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

Calculate t statistic

$$t = \frac{\text{difference between groups}}{\text{variability within groups}} = \frac{\text{Mean grp 1} - \text{Mean grp 2}}{((s_1^2/n_1) + (s_2^2/n_2))^{1/2}}$$

s = variance  
n = size of sample

Determine confidence level for t  
(probability that t could occur by  
chance)

$$df = n_1 + n_2 - 2$$

*The larger the difference  
between the groups and the  
lower the variance the bigger t  
will be and the lower p will be*

Degrees of Freedom	Probability, p			
	0.1	0.05	0.01	0.001
1	6.31	12.71	63.66	638.42
2	2.92	4.30	9.93	31.60
3	2.35	3.18	7.84	12.92
4	2.12	2.78	6.88	9.60
5	2.02	2.57	6.03	8.57
6	1.94	2.45	5.71	7.98
7	1.89	2.37	5.50	7.45
8	1.86	2.31	5.34	7.04
9	1.83	2.26	5.25	6.78
10	1.81	2.23	5.17	6.58
11	1.80	2.20	5.11	6.41
12	1.78	2.18	5.06	6.32
13	1.77	2.16	5.01	6.22
14	1.76	2.14	4.98	6.14
15	1.75	2.13	4.95	6.07

Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

- Class comparison experiment:
  - Two types of errors in hypothesis testing:

Hypothesis Testing		Truth	
		H <sub>0</sub>	H <sub>1</sub>
Decision	Reject H <sub>0</sub>	Type I Error (alpha) (false positive)	Right Decision (true positive)
	Don't Reject H <sub>0</sub>	Right Decision	Type II Error (beta)

- **Multiple testing correction:**
  - Microarray studies typically involve the simultaneous testing of hundreds or thousands of genes for differential expression. This results in a large number of falsely significant results. Therefore we must correct for multiple testing.
  - Multiple testing correction is a method for adjusting the p-value from a comparison test based on the number of test performed. These adjustments help to reduce the number of false positives in an experiment



Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

- Class comparison experiment:
  - **Multiple testing corrections:**
    - **Family Wise Error Rate (FWER)** : the probability of at least one false positive for all comparisons
      - Bonferonni correction ( $\alpha/\text{number comparisons}$ ), Holm's test
    - **False Discovery Rate (FDR)** : the expected proportion of false positives among the results
      - Benjamini and Hochberg (BH), Significance Analysis of Microarrays (SAM)
  - Example: 1000 genes and 50 DE using cutoff of 5%
    - FWER: using 5% FWER means there is a 5% chance that you have at least 1 false positive . For Bonferonni correction the p-value of DE genes is  $< 0.00005$  ( $0.05/1000$ ). This is very good and would be a very conservative requirement, you are confident that all of your results are real.
    - FDR: using 5% FDR you would expect 2.5 false positives (5% of 50). This is probably acceptable, you are confident that most of your results are real.
  - In Bioconductor software: package LIMMA<sup>1</sup> for moderated t-statistics with multiple testing correction and package SIGGENES for SAM<sup>2</sup>

Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

- Class prediction experiment:

- Numerous methods:

- logistic regression
    - linear and quadratic discriminant analysis,
    - nearest neighbor classifiers
    - decision trees
    - shrunken centroids
    - neural networks
    - random forests
    - support vector machines
    - .....

- I used the PAM method in my research

- Predictive Analysis of Microarrays (PAM)<sup>1</sup>

- Bioconductor software: package PAMR
    - A statistical technique for class prediction from gene microarray data using **nearest shrunken centroids**.
    - It identifies a subset of genes that best characterizes the class identity and this subset can be used to predict the class of new samples.

- Class prediction experiment:

- PAM method
- Nearest shrunken centroid method:

- Modification of nearest centroid method, which computes a standardized **centroid** for each class in the training set. This is the **average gene expression for each gene in each class divided by the within-class standard deviation for that gene.**

- Nearest centroid classification takes the gene expression profile of a new sample, and compares it to each of these class centroids. The class, whose centroid it is closest to, in squared distance, is the predicted class for that new sample.

Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

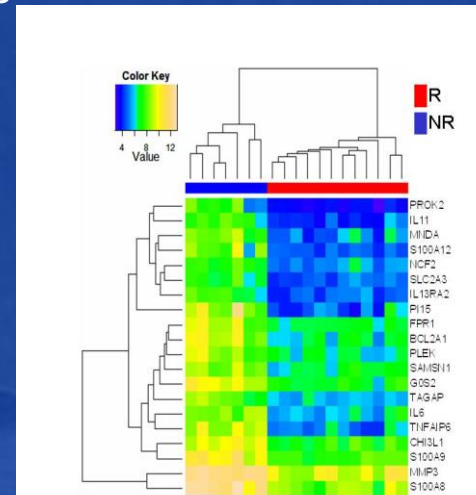
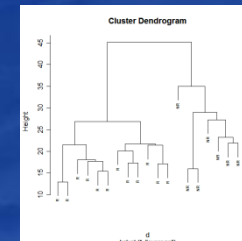
Data analysis

Biological verification  
and interpretation

Making data public

- Class discovery experiment:

- It involves analyzing a given set of gene expression profiles with the goal of discovering subgroups that share common features
- The analyses for class discovery are unsupervised
  - Analysis without prior knowledge of sample identity
  - Clustering analysis
    - It divides data points (genes or samples) into groups (clusters) using measures of similarity, such as correlation or Euclidian distance .
    - Complete, average or single linkage clustering.
- e.g. Hierarchical clustering analysis
  - The end result is a cluster tree or dendrogram
  - The result can also be displayed as a 2-dimensional heatmap with 2 dendrograms, one indicating the similarity between patients and the other indicating the similarity between genes.





- Probe set annotations:

- Affymetrix NetAffx website (<http://www.affymetrix.com/analysis/index.affx>)
  - Example: 206988\_at from Affymetrix Human Genome 133 plus 2.0 array

Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

The screenshot displays the Affymetrix NetAffx Analysis Center website. The top navigation bar includes links for Products & Services, Support, Partners & Programs, About Affymetrix, Careers, and NetAffx. The main content area is titled "NetAffx Query" and features a search bar with the text "206988\_at" entered. Below the search bar, there is a section titled "Select a GeneChip Array:" with a dropdown menu showing several options, including "Human Genome U133 Plus 2.0 Array" (which is selected). To the left of the main content, there is a sidebar with a "NetAffx" section containing links for "Exon/Gene Expression" and "3' IVT Expression", and a "Genotyping" section. At the bottom of the sidebar, there is a "Query History" section with a link for "Expression Queries". The main content area also includes a "Submit" button at the bottom.

- Probe set annotations:

- Affymetrix NetAffx website (<http://www.affymetrix.com/analysis/index.affx>)
  - Example: 206988\_at from Affymetrix Human Genome 133 plus 2.0 array

Biological question

Experimental design

Microarray

Image

Normalization

Data

Bioinformatics

anc

Making data public

**Affymetrix**

Products & Services | Support | Partners & Programs | About Affymetrix | Careers | NetAffx | Shop

日本語

NetAffx™ Analysis Center

Home > Analysis Center > NetAffx > Show Results



## Show Results ?

Refine Query  
Create a Custom View  
Export Results  
Show Orthologs

**Current Query:** All Descriptions (206988\_at)  
**Array(s):** HG-U133\_Plus\_2  
**Probe Sets returned:** 1

Displaying Results: 1-1 of 1.

\* Annotation List \* 50 Remove Checked Save Current List Expanded Mode +

	Probe Set ID	Gene Title	Gene Symbol	go biological process term	go molecular function term	go cellular component term	Pathway
	206988_at	chemokine (C-C motif)	CCL25	chemotaxis	cytokine activity	extracellular region	
	...	...	...	...	...	...	

[Click to Expand/Condense row](#)

NetAffx

- Exon/Gene Expression
- 3' IVT Expression
- NetAffx Query
- Batch Query
- BLAST
- Probe Match
- UCSC Query
- Custom Annotation Views
- Genotyping
- Manage Query Folders
- Query History
- Expression Queries

All Descriptions

- Probe set annotations:

- Affymetrix NetAffx website (<http://www.affymetrix.com/analysis/index.affx>)
  - Example: 206988\_at from Affymetrix Human Genome 133 plus 2.0 array

Biological question

Experimental design

Microarray

Image

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

Assignments			
U86358	<a href="#">NCBI</a>	Human chemokine (TECK) mRNA, complete cds.	11/11 None
+ Cross-hybridizing Transcripts (7)			
Alignment(s) ?			
Position/Genome View	Identity	Coverage	Cytoband
chr19:8023933-8033530 (+) <a href="#">UCSC ENSEMBL IGB</a> *	84.64	98.1	p13.2
* To view alignments in the <a href="#">Integrated Genome Browser (IGB)</a> , you must <a href="#">start IGB</a> first.			



- Probe set annotations:

- Affymetrix NetAffx website (<http://www.affymetrix.com/analysis/index.affx>)
  - Example: 206988\_at from Affymetrix Human Genome 133 plus 2.0 array

Biological question

Experimental design

Microarray experiment

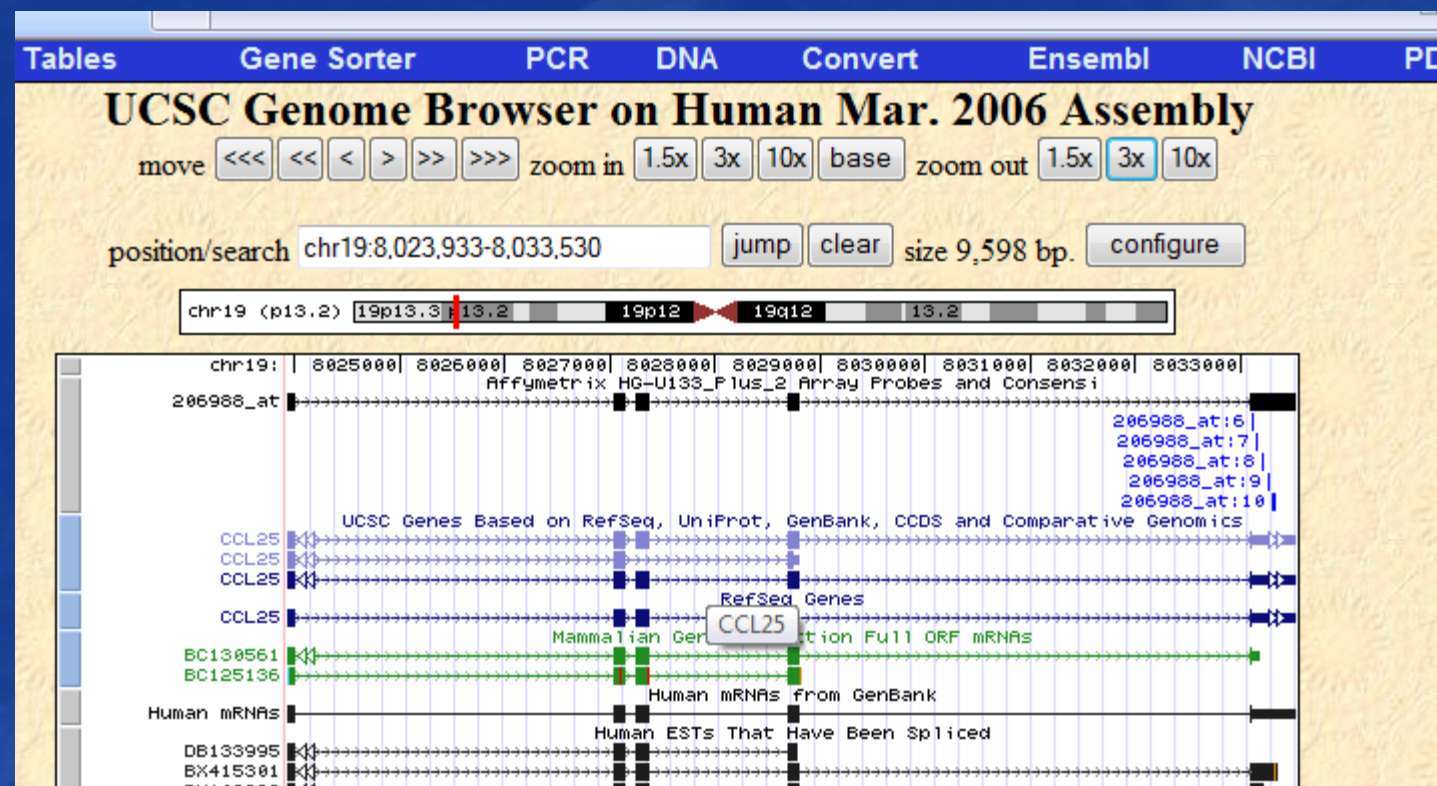
Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public



- Functional analysis:

- to identify which biological processes and/or diseases are overrepresented among sets of significant probe sets
- Programs:
  - Ingenuity Pathway Analysis (Ingenuity Systems®, [www.ingenuity.com](http://www.ingenuity.com))
  - GenMapp (<http://www.genmapp.org/>)
  - DAVID homepage (<http://david.abcc.ncifcrf.gov/home.jsp>)
  - Bioconductor software: package topGO
  - Gene Ontology Enrichment Analysis Software Toolkit (GOEAST)

Biological question



Experimental design



Microarray experiment



Image analysis



Normalization



Data analysis



Biological verification  
and interpretation



Making data public

- **MIAME:**

- For publishing microarray based papers, most of the journals require MIAME compliant data.

Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

1: [Nat Genet.](#) 2001 Dec;29(4):365-71.

Comment in:

[Nat Genet.](#) 2001 Dec;29(4):373.

[Nat Genet.](#) 2006 Oct;38(10):1089.

**Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.**

[Brazma A](#), [Hingamp P](#), [Quackenbush J](#), [Sherlock G](#), [Spellman P](#), [Stoeckert C](#), [Aach J](#), [Ansorge W](#), [Ball CA](#), [Causton HC](#), [Gaasterland T](#), [Glenisson P](#), [Holstege FC](#), [Kim IF](#), [Markowitz V](#), [Matese JC](#), [Parkinson H](#), [Robinson A](#), [Sarkans U](#), [Schulze-Kremer S](#), [Stewart J](#), [Taylor R](#), [Vilo J](#), [Vingron M](#).

European Bioinformatics Institute, EMBL outstation, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.  
brazma@ebi.ac.uk

Microarray analysis has become a widely used tool for the generation of gene expression data on a genomic scale. Although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. Here we present a proposal, the **Minimum Information About a Microarray Experiment (MIAME)**, that describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. The ultimate goal of this work is to establish a standard for recording and reporting microarray-based gene expression data, which will in turn facilitate the establishment of databases and public repositories and enable the development of data analysis tools. With respect to MIAME, we concentrate on defining the content and structure of the necessary information rather than the technical format for capturing it.

PMID: 11726920 [PubMed - indexed for MEDLINE]

- It describes the **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.

- **MIAME:**

- MIAME requirements:

- The **raw data** for each hybridisation (e.g., CEL or GPR files)
- The final processed (**normalised**) **data** for the set of hybridisations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
- The essential **sample annotation** including experimental factors and their values (e.g., compound and dose in a dose response experiment)
- The **experimental design** including sample data relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates)
- Sufficient **annotation of the array** (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
- The essential laboratory and data processing **protocols** (e.g., what normalisation method has been used to obtain the final processed data)

- The public repositories **ArrayExpress** at the EBI (UK), **Gene Expression Omnibus** (GEO) at NCBI (US) and **CIBEX** at DDBJ (Japan) are designed to accept, hold and distribute MIAME compliant microarray data.

Biological question



Experimental design



Microarray experiment



Image analysis



Normalization



Data analysis



Biological verification  
and interpretation



Making data public



# Example:

Gut. 2009 Dec;58(12):1612-9. Epub 2009 Aug 20.

## Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis.

Arijs I, Li K, Toedter G, Quintens R, Van Lommel L, Van Steen K, Leemans P, De Hertogh G, Lemaire K, Ferrante M, Schnitzler F, Thorrez L, Ma K, Song XY, Marano C, Van Assche G, Vermeire S, Geboes K, Schuit F, Baribaud F, Rutgeerts P.

Department of Gastroenterology, University of Hospital Gasthuisberg, Herestraat 49, B-3000 Leuven, Belgium.

### Abstract

**BACKGROUND AND AIMS:** Infliximab is an effective treatment for ulcerative colitis with over 60% of patients responding to treatment and up to 30% reaching remission. The mechanism of resistance to anti-tumour necrosis factor alpha (anti-TNFalpha) is unknown. This study used colonic mucosal gene expression to provide a predictive response signature for infliximab treatment in ulcerative colitis.

**METHODS:** Two cohorts of patients who received their first treatment with infliximab for refractory ulcerative colitis were studied. Response to infliximab was defined as endoscopic and histological healing. Total RNA from pre-treatment colonic mucosal biopsies was analysed with Affymetrix Human Genome U133 Plus 2.0 Arrays. Quantitative RT-PCR was used to confirm microarray data.

**RESULTS:** For predicting response to infliximab treatment, pre-treatment colonic mucosal expression profiles were compared for responders and non-responders. Comparative analysis identified 179 differentially expressed probe sets in cohort A and 361 in cohort B with an overlap of 74 probe sets, representing 53 known genes, between both analyses. Comparative analysis of both cohorts combined, yielded 212 differentially expressed probe sets. The top five differentially expressed genes in a combined analysis of both cohorts were osteoprotegerin, stanniocalcin-1, prostaglandin-endoperoxide synthase 2, interleukin 13 receptor alpha 2 and interleukin 11. All proteins encoded by these genes are involved in the adaptive immune response. These markers separated responders from non-responders with 95% sensitivity and 85% specificity.

**CONCLUSION:** Gene array studies of ulcerative colitis mucosal biopsies identified predictive panels of genes for (non-)response to infliximab. Further study of the pathways involved should allow a better understanding of the mechanisms of resistance to infliximab therapy in ulcerative colitis. ClinicalTrials.gov number, NCT00639821.

### Comment in

Mucosal gene expression signatures that predict response of ulcerative colitis to infliximab. [Gastroenterology. 2011]



# Introduction



- **Ulcerative colitis**

- = a chronic **inflammatory bowel disease** (IBD) involving the colonic mucosa
- Symptoms: abdominal pain, bloody diarrhoea, fatigue, fever and weight loss
- The pathogenesis of UC remains unknown but chronic inflammation probably results from an interaction of genetic factors, the immune response to microbial dysbiosis and environmental factors.

- In the last decades, great progress has been made in the treatment of IBD, especially regarding biological therapeutics

- **Infliximab** (Remicade®):

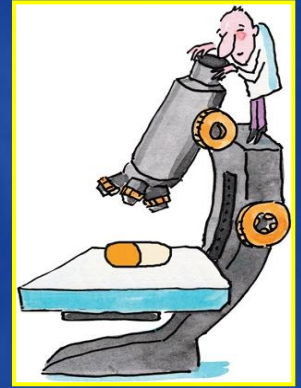
- the first clinically available biological treatment for IBD
- a chimeric monoclonal IgG1 antibody against TNF- $\alpha$
- Up to 30% of the patients do not respond to the treatment



→ It's important to identify markers for response to infliximab (IFX) in order to optimize the use of this costly drug

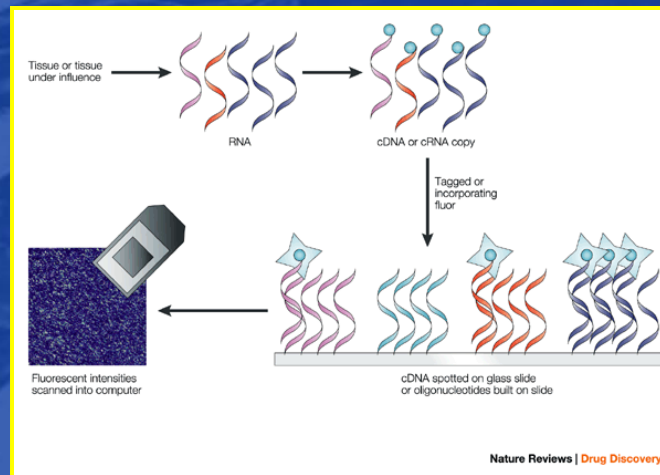
# Aim of the study

- To identify mucosal gene signatures predictive of response to infliximab in UC



using gene expression microarray technology

- A tool that allows a simultaneous wide survey of gene expression

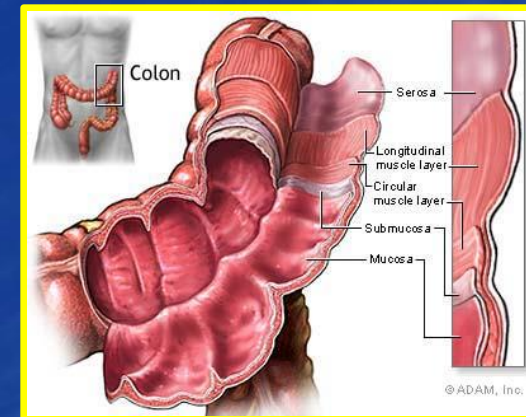
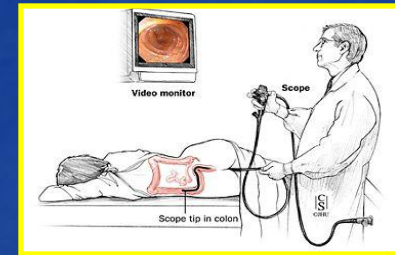




# Methods and results:

## Patients and tissue specimens:

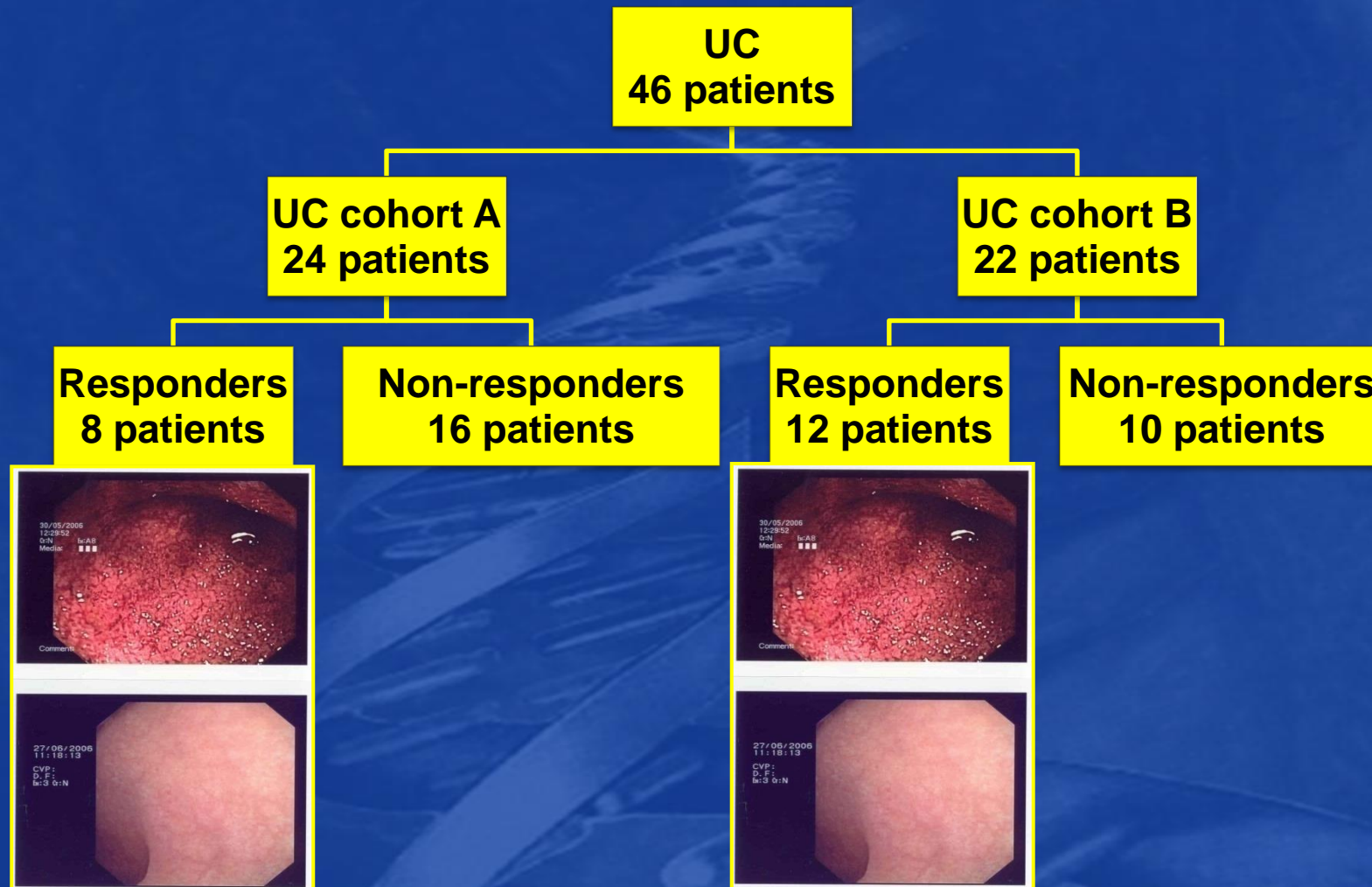
- Two independent cohorts of UC patients who received a first treatment with IFX were studied:
  - Cohort A: 24 UC patients
  - Cohort B: 23 UC patients
- Mucosal biopsies were obtained at routine colonoscopy:
  - from diseased colon before and 4-8 weeks after first IFX (5 or 10 mg per kg body weight) infusion
- All biopsies were blindly scored for inflammation using the histological scoring system from Geboes *et al.*, Gut 2000





## Patients and tissue specimens:

- Response to IFX was assessed at 4-6 weeks after first IFX infusion and defined as complete endoscopic and histologic healing of the lesions:



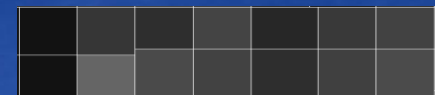
## Microarray analysis:

A. Total RNA was isolated, labelled and hybridized to Affymetrix HGU133 Plus 2.0 array:

- comprised of 54675 probesets
- “whole genome” coverage



Microarray data analyses with the CEL files



It contains a single intensity value for each probe cell delineated by the grid



# Microarray analysis:

## B. Microarray data analysis with R (version R 2.7.2)/Bioconductor software:

- Free download: <http://www.bioconductor.org/docs/install/>

### Installation Instructions

#### Install R

1. Download the most recent version of R from The Comprehensive R Archive Network (CRAN). The R FAQ and the R Installation and Administration Manual contain detailed instructions for installing R on various platforms (Linux, OS X, and Windows being the main ones).
2. Start the R program, on Windows and OS X, this will usually mean double-clicking on the R application, on UNIX-like systems, type "R" at a shell prompt.
3. As a first step with R, start the R help browser by typing "help.start()" in the R command window. For help on any function, e.g. the "mean" function, type "? mean".

#### Install standard Bioconductor packages

Install BioConductor packages using the `biocLite.R` installation script. In an R command window, type the following:

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

This installs the following packages: affy, affydata, affyPLM, annaffy, annotate, Biobase, Biobstrings, DynDoc, gcrma, genefilter, geneplotter, hgu95av2.db, limma, marray, matchprobes, multtest, ROC, vsn, xtable, affyQCRReport. After downloading and installing these packages, the script prints "Installation complete" and TRUE.

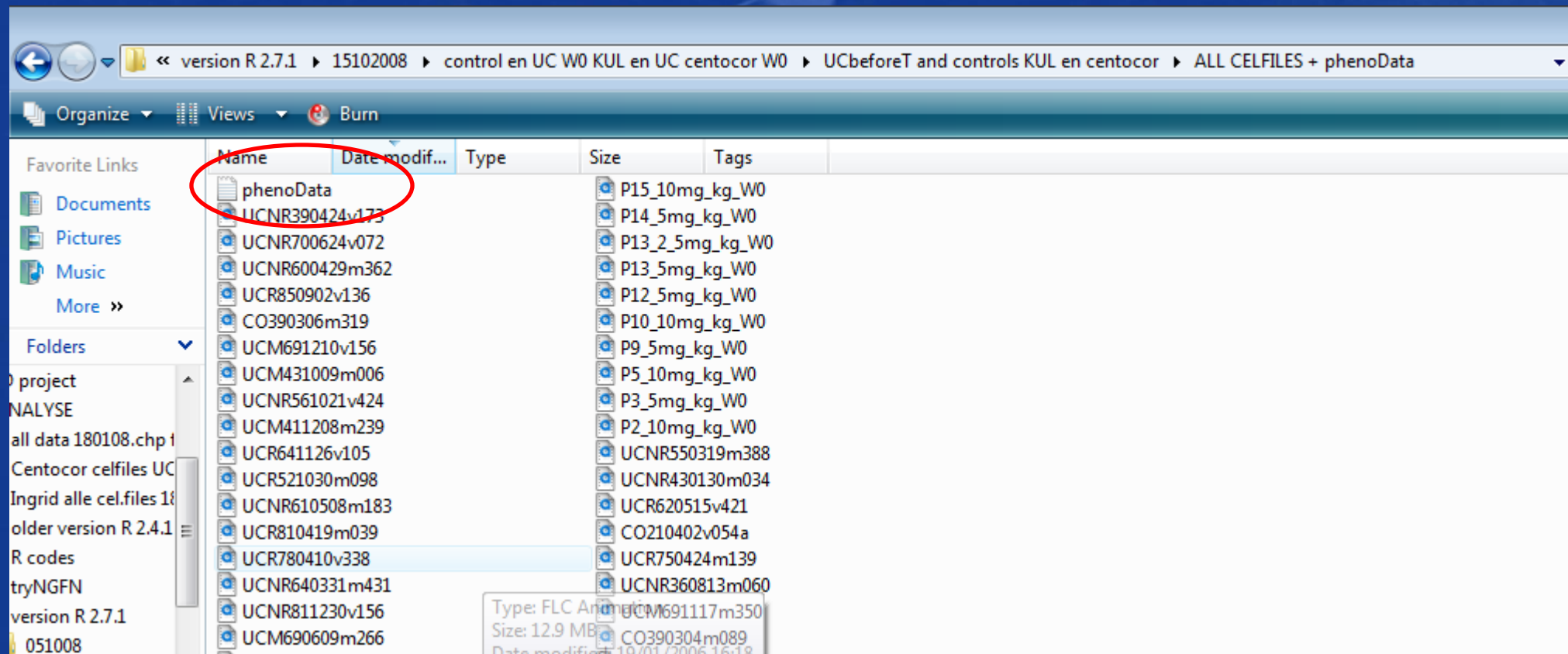
The `biocLite` script has arguments that change the default behavior:



## Microarray analysis:

### B. Microarray data analysis with R (version R 2.7.2)/Bioconductor software:

- Make a file with all the CEL files and phenoData.txt





- Ma

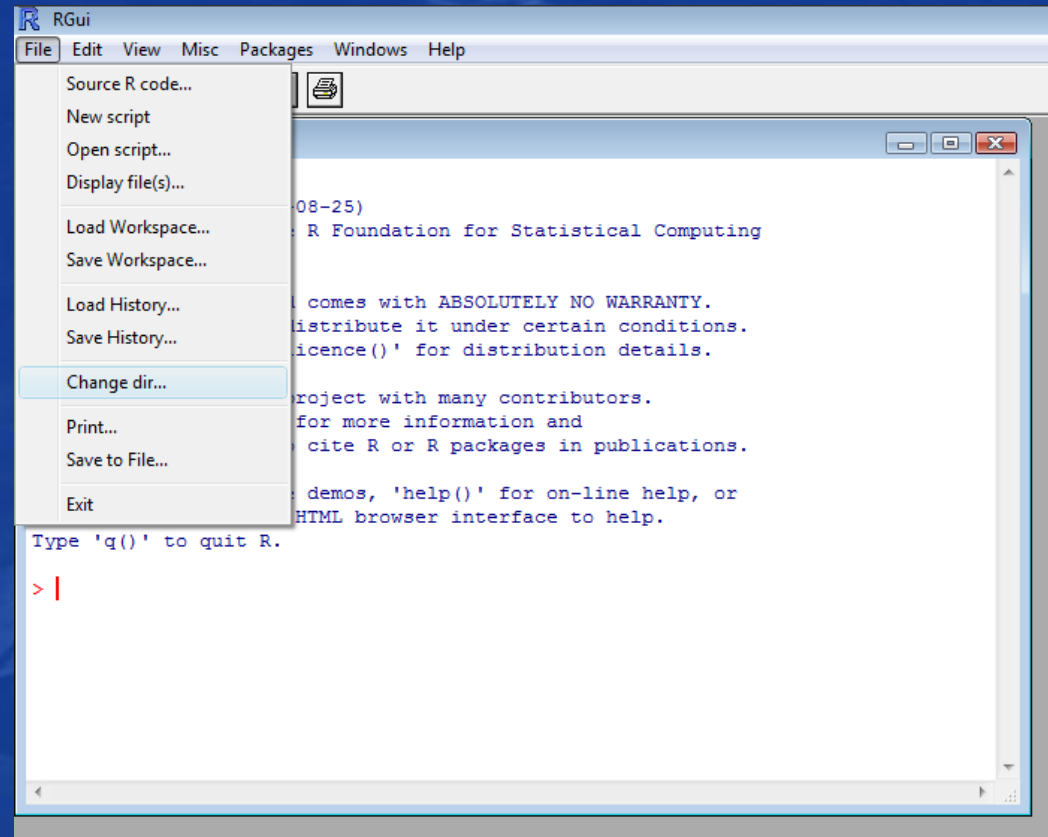
[illegible]



## Microarray analysis:

### B. Microarray data analysis with R (version R 2.7.2)/Bioconductor software:

- **Load the data in R:** Go to file in R → Change dir → menu item (the file with the cel.files and phenoData.txt)



- **Normalisation:** Going from probe level data (.cel files) to expression measures for each probe set with the RMA method in package Affy. The expression measure is the data available for data analysis

### R commands:

- `library(affy)` # Loads affy package
- `library(hgu133plus2.db)` # Loads hgu133plus2.db package
- `pd<-read.AnnotatedDataFrame("phenoData.txt",header=TRUE,row.names=1)`
- `datarma<-justRMA(filenamees=rownames(pData(pd)),phenoData=pd)` # Creates normalized log2 expression values using RMA method
- `write.exprs(datarma, file="datarma.txt")` # Writes expression values to text file in working directory

- **Non-specific filtering: To eliminate non-relevant probe sets.**
  - The probe sets with low overall intensity and variability that are unlikely to carry information about the phenotypes under investigation were removed. A non-specific filtering was applied on the log2 RMA normalized data (54675 probe sets) from the pre-treatment UC samples from both cohorts.
    - Only probe sets with an intensity  $> \log_2(100)$  in at least 10% of the samples and an interquartile range (IQR) of log2 intensities across the samples  $> 0.5$  were included, leaving 9183 probe sets for further data analysis.

## R commands:

```
➤ eset<-  
datarma[,pData(datarma)[,"Disease"]%in%c("UC")&pData(datarma)[,"Treatment"]%in%c("B")&pData(data  
rma)[,"Responsfinal"]%in%c("R","NR")&pData(datarma)[,"celfileszonderP13"]%in%c("centocor","KUL")] #  
select the pre-treatment expression profiles of both cohorts  
➤ eset          } # provides summary information of exprSet object 'eset'  
➤ pData(eset)   }  
➤ library(genefilter)  
➤ f1<-pOverA(0.10,log2(100))  
➤ f2<-function(x)(IQR(x)>0.5)  
➤ ff<-filterfun(f1,f2)  
➤ selected<-genefilter(eset,ff)  
➤ sum(selected)  
➤ esetSub<-eset[selected,]  
➤ table(selected)  
➤ esetSub
```

# non-specific filtering leaving 9183 probe sets for further analysis

- intensity of a gene should be above 100 in at least 25% of the samples
- interquartile range of log2-intensities should be at least 0.5



- Class discovery:

For **comparative analysis**, package **LIMMA (based on moderated t-test)** was used to identify probe sets that are differentially expressed between responders (R) and non-responders (NR) at baseline (before IFX treatment) in cohort A, cohort B and both cohorts combined.

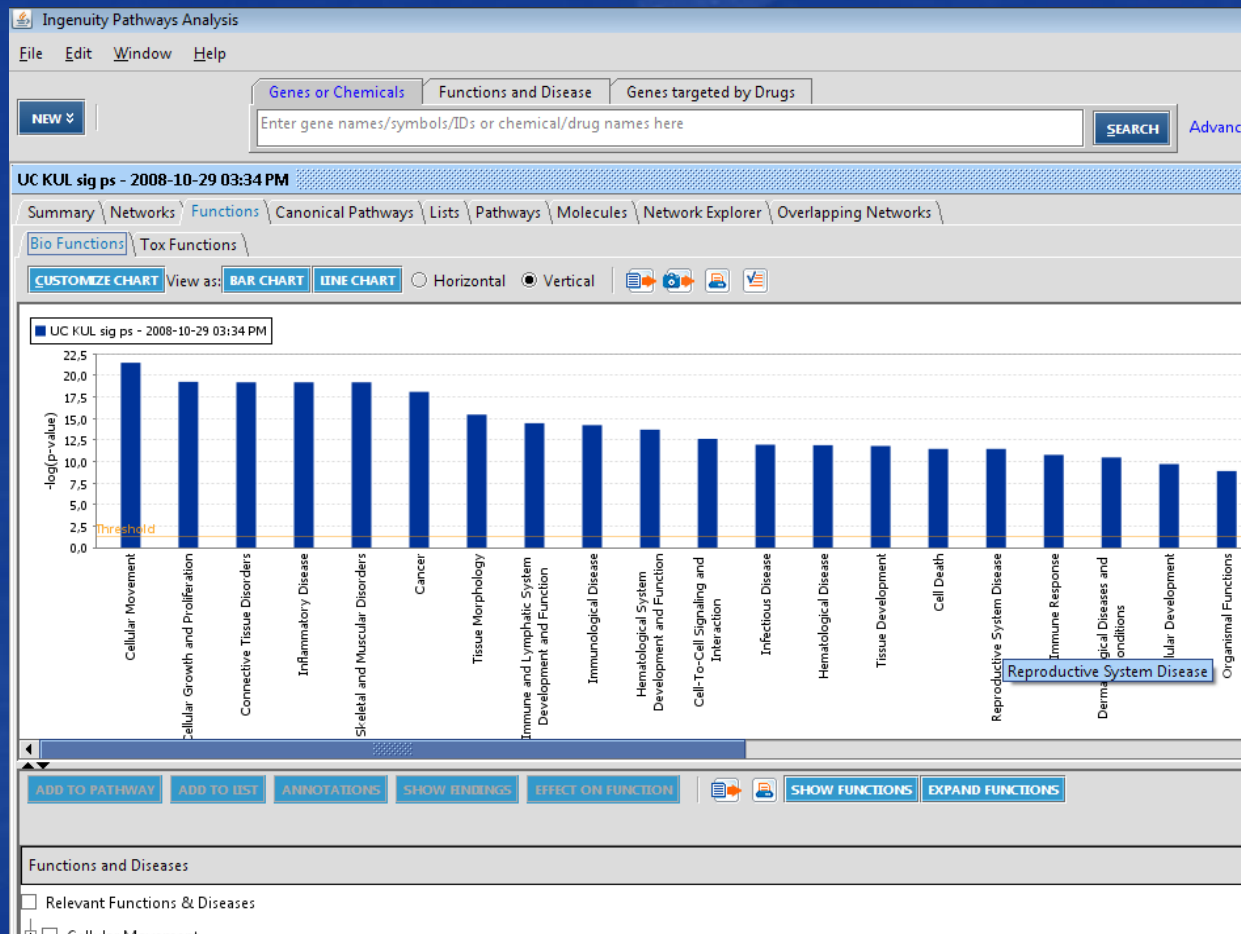
To correct for multiple testing, the **false discovery rate** (FDR) was estimated from p-values derived from the moderated *t*-statistics using the method of **Benjamini and Hochberg** (BH).

- Fold change (average responders/average non-responders)  $>2$  and  $\text{FDR} < 5\%$  were considered statistical significant.

## R commands for comparative analysis between R and NR for example in cohort A:

```
> esetset<-
esetSub[,pData(esetSub)[,"Disease"]%in%c("UC")&pData(esetSub)[,"Treatment"]%in%c("B")&pData(eset
Sub)[,"Responsfinal"]%in%c("R","NR")&pData(esetSub)[,"celfileszonderP13"]%in%c("KUL")] # selects all
pre-treatment UC samples of cohort A
> f<-factor(as.character(esetset$Responsfinal)) # variable that says which sample is R or NR
> design<-model.matrix(~f) # Creates design matrix
> library(limma) # loads limma package
> fit<-lmFit(exprs(esetset),design) # Fits a linear model for each gene based on the given series of arrays.
> fit2<-eBayes(fit) # Computes moderated t-statistics and log-odds of differential expression by empirical Bayes
shrinkage of the standard errors towards a common value.
> options(digits=2)
> topTable(fit2,coef=2,adjust="BH",number=100) # Generates list of top 100 DE probe sets adjusted by BH as
FDR
➤ topTableall<-topTable(fit2,coef=2,adjust="BH",number=9183) # Generates list of all probe sets
➤ write.table(topTableall,file="topTableall.xls",sep="\t",quote=F) # Export limma statistics table to text file
> topTablesig<-topTableall[topTableall$adj.P.Val<0.05&(topTableall$logFC>1|topTableall$logFC<(-1)),]
> dim(topTablesig) # number of significant probe sets wit >2-fold change and FDR<5%
```

- The **Bio Functional Analysis** tool in the Ingenuity Pathway Analysis (IPA) program (Ingenuity Systems®, [www.ingenuity.com](http://www.ingenuity.com)) was used to identify biological functions and/or diseases among sets of significant probe sets.
  - Load the list of significant probe sets in IPA program and run a new core analysis



## Results:

- **Comparative analysis** between R and NR in cohort A, cohort B and both cohorts combined:

Comparative analysis	UC cohort A R(n=8)/NR(n=16)	UC cohort B R(n=12)/NR(n=10)	UC cohort A and B R(n=20)/NR(n=26)
Increased probe sets in R	0	38	5
Decreased probe sets in R	179	323	207
Total	179	361	212

- The significant probe sets of each pairwise comparison showed a predominance of the biological functions: **immune response, cellular movement, cellular growth and proliferation, hematological system development and function, cell-to-cell signalling and interaction, cell death and tissue morphology/development.**



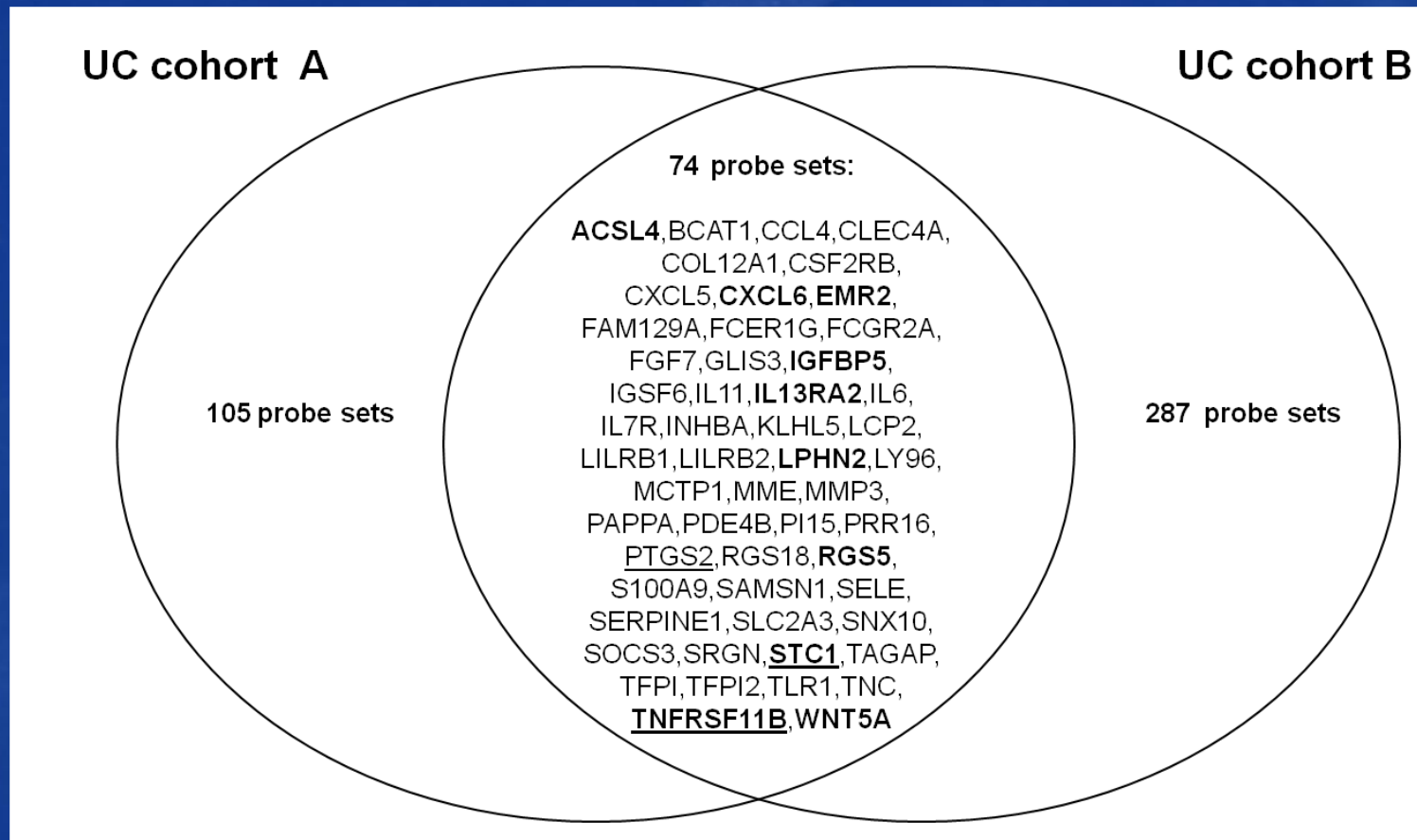
## Results for comparative and IPA analyses:

- **Comparative analysis** between R and NR in cohort A, cohort B and both cohorts combined:

Comparative analysis	UC cohort A R(n=8)/NR(n=16)	UC cohort B R(n=12)/NR(n=10)	UC cohort A and B R(n=20)/NR(n=26)
Increased probe sets in R	0	38	5
Decreased probe sets in R	179	323	207
Total	179	361	212

- The significant probe sets of each pairwise comparison showed a predominance of the biological functions: **immune response, cellular movement, cellular growth and proliferation, hematological system development and function, cell-to-cell signalling and interaction, cell death and tissue morphology/development.**

- There was an **overlap of 74 significant probe sets**, representing 53 different known genes, **between the LIMMA analyses in cohort A and in cohort B**, and these common probe sets were all downregulated in responders as compared to non-responders



- **Class prediction: PAM with leave-one-out cross validation** was carried out on the **top 5 most significantly different known genes**, identified by LIMMA analysis between responders and non-responders in each cohort and both cohorts combined
  - to see if these subsets accurately predict response or non-response to infliximab
  - to identify the lowest misclassification error rate based on these subsets.
- **Unsupervised average-linkage hierarchical clustering, using Euclidian distance as metric**, was applied to the log2 expression values of the top 5 significant genes from the LIMMA analysis in each cohort and both cohorts combined to visualize gene/sample relationship.  
The results of the clustering were visualized as a 2-dimensional heatmap with 2 dendrograms, one indicating the similarity between patients and the other indicating the similarity between genes.

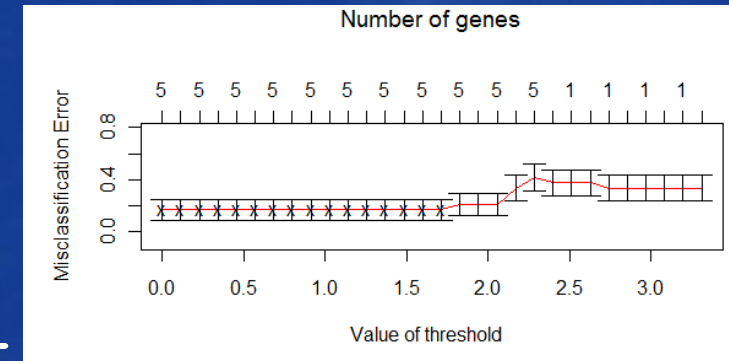
## R commands for PAM with top 5 genes:

```

➤ library(pamr) # loads pamr package
➤ seltop5<-exprs(esetsel)[c("206172_at","210895_s_at","206953_s_at","211671_s_at","213338_at"),] # selects top 5 genes
➤ pData(esetsel)[,"Responsfinal"]<-factor(as.character(esetsel$Responsfinal))
➤ labels<-esetsel$Responsfinal
➤ set.seed(4)
➤ dat<-seltop5
➤ gl<-rownames(dat)
➤ sl<-as.factor(colnames(dat))
➤ train.dat<-list(x=dat,y=labels,geneid=gl,sampleid=sl)
➤ model<-pamr.train(train.dat)
➤ set.seed(4)
➤ model.cv<-pamr.cv(model,train.dat,nfold=10)
➤ model.cv
➤ source("C:\\PhD and postdoc project\\course liege 20112012\\nscv.R")
➤ objects()
➤ ncv <- length(train.dat$y)
➤ ncv
➤ set.seed(4)
➤ model.cv<-MYpamr.cv(model, train.dat, folds=as.list(seq(ncv)))
➤ pamr.plotcv(model.cv)
➤ Delta=1.26
➤ pamr.confusion(model.cv,Delta)
➤ listtop5<-pamr.listgenes(model,train.dat,Delta)
➤ write.table(listtop5, "pamrall_listgenestop5genes.txt", sep='\\t', quote=F)
➤ pamr.plotcen(model,train.dat,Delta)
➤ pamr.geneplot(model,train.dat,Delta)
➤ pamr.predict(model,seltop5,Delta,type="posterior")

```

} # variable for response to IFX



# PAM with leave-one-out cross - validation and threshold 1.26



## R commands for hierarchical cluster analysis with top 5 genes:

```
➤ dat<-seltop5
➤ d<-dist(dat)
➤ hc<-hclust(d,method="average")
➤ d2<-dist(t(dat))
➤ hc2<-hclust(d2,method="average")
➤ library(gplots)
➤ color.map<-function(Responsfinal){if(Responsfinal=="R")"#FF0000"else"#0000FF"}
➤ patientcolors<-unlist(lapply(esetsel$Responsfinal,color.map))
➤ heatmap.2(dat,col=topo.colors(100),cexRow=0.6,cexCol=0.8,scale="none",key=TRUE,symkey=FALSE,density.info="none",trace="none",labCol=
=labels,margins = c(2,8),Rowv=as.dendrogram(hc),Colv=as.dendrogram(hc2),ColSideColors=patientcolors)
```

# average-linkage hierarchical clustering, using  
Euclidian distance as metric with heatmap as  
result

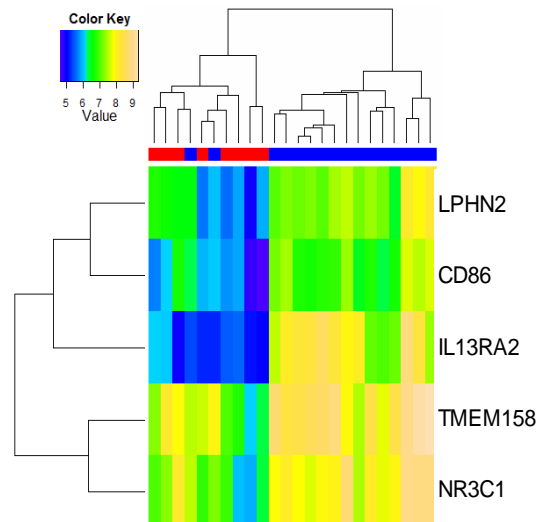
## Results for PAM and hierarchical cluster analyses:

- PAM analysis of the top 5 significantly genes:

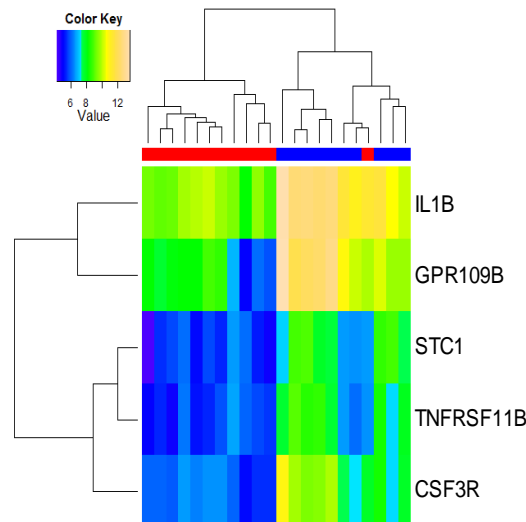
PAM analysis			
UC cohort A (R=8, NR=16)	Sensitivity	Specificity	Overall accuracy
Top 5 genes	75% (6/8)	87.5% (14/16)	83.3% (20/24)
UC cohort B (R=12, NR=10)	Sensitivity	Specificity	Overall accuracy
Top 5 genes	91.7% (11/12)	90% (9/10)	90.9% (20/22)
UC cohort A and B (R=20, NR=26)	Sensitivity	Specificity	Overall accuracy
Top 5 genes	95% (19/20)	84.6% (22/26)	89.1% (41/46)

- Hierarchical cluster analysis of the top 5 significantly genes:

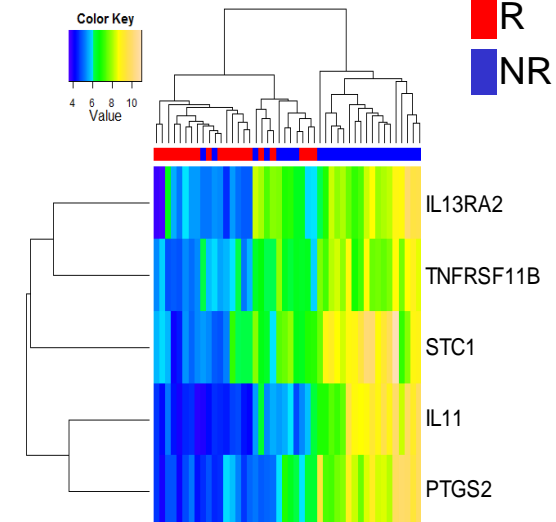
Cohort A:



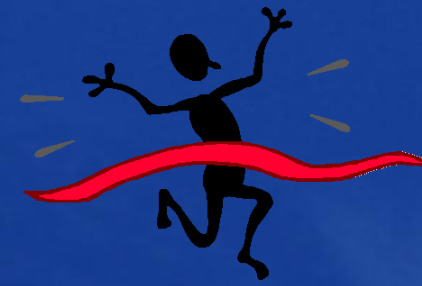
Cohort B:



Cohort A and B:

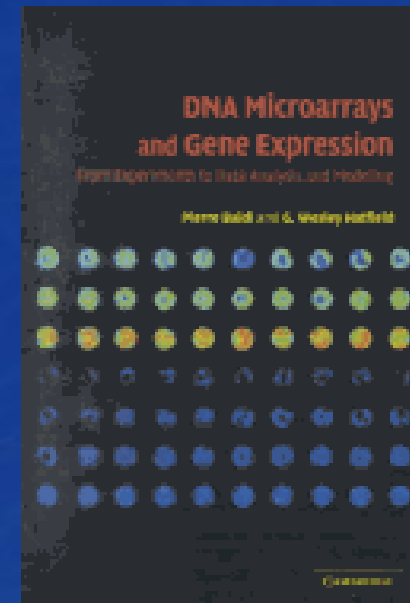
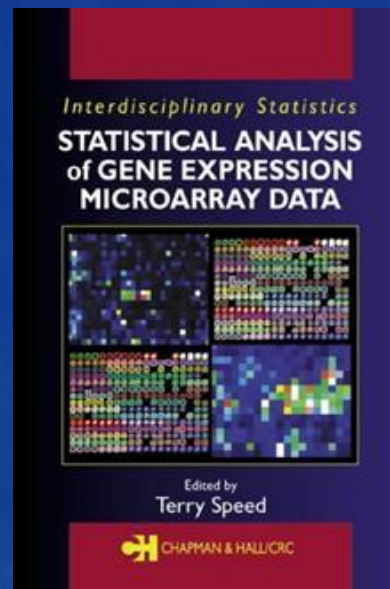
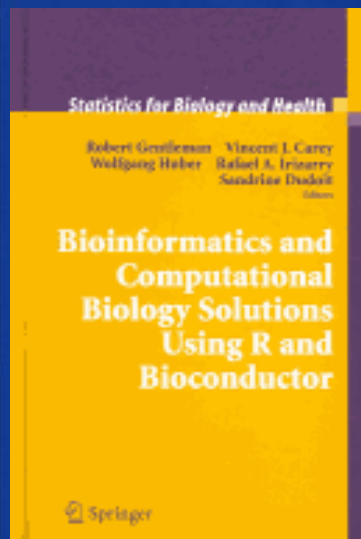


# Conclusion



- Gene array studies of UC mucosal biopsies identified predictive panels of genes for (non-)response to infliximab in pre-treatment mucosal biopsies of patients who received for the first time infliximab therapy in two cohorts of patients.
- Our studies demonstrate that differences in mucosal expression of a limited number of genes involved in the inflammatory cascade account for resistance of UC to respond to infliximab therapy.
- Further study of the pathways involved should allow to better understand mechanisms of the resistance to infliximab therapy in UC.

# Interesting books

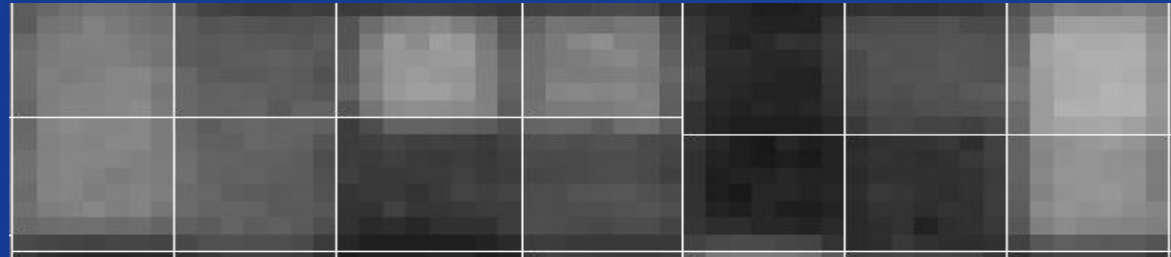




- Affymetrix file types:

### DAT file

The image of a scanned probe array

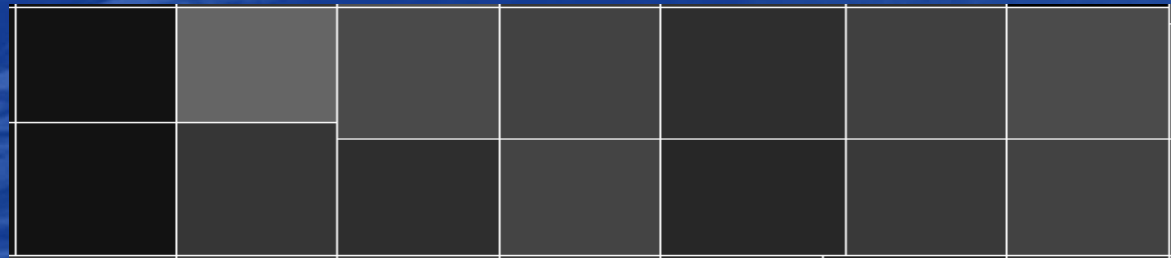


Process image  
(GCOS)



### CEL file = processed .DAT file

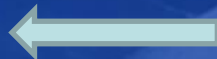
It contains a single intensity value  
for each probe cell delineated by the grid



MAS5  
(GCOS)



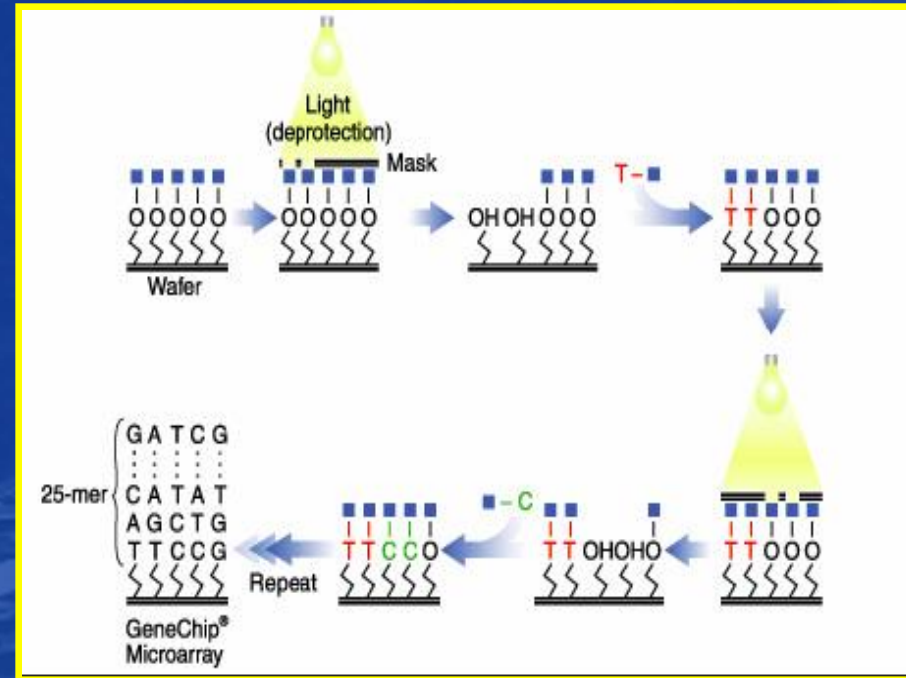
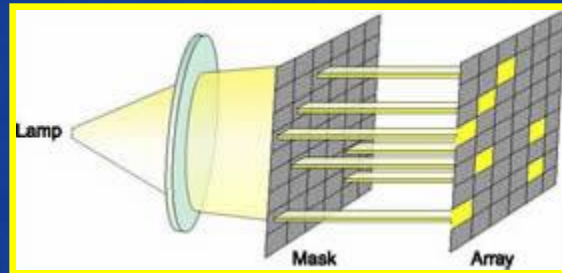
**CDF file** (=chip description file): Provided by  
Affy, describes layout of chip



**CHP file:** Experiment results created from CEL and CDF files

**RPT file:** Generated by GCOS, report of quality control (Quality evaluation included Spike-In controls (BioB, BioC, BioD, and cre), a 3'-5' ratio of GAPDH < 3.0, and average background signal < 100)

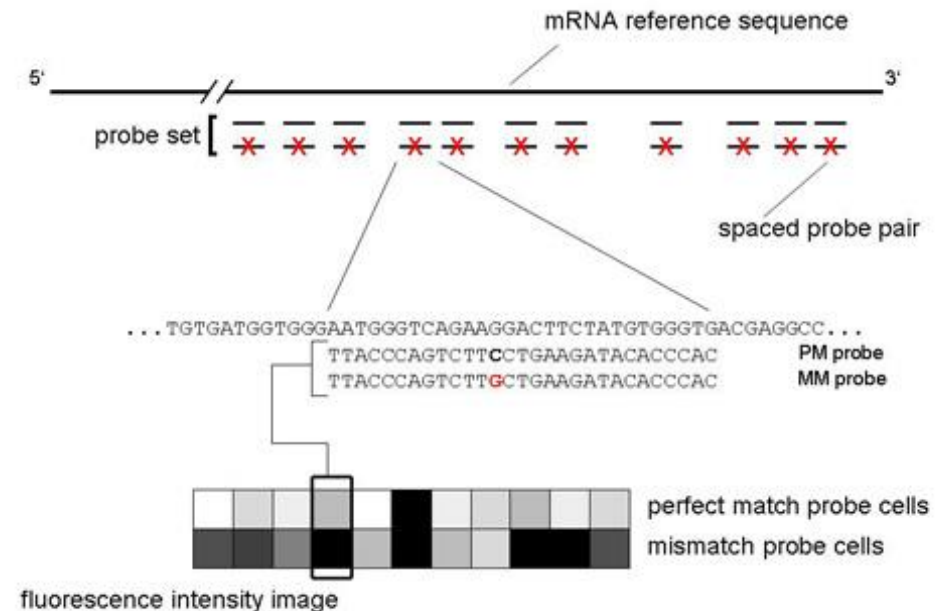
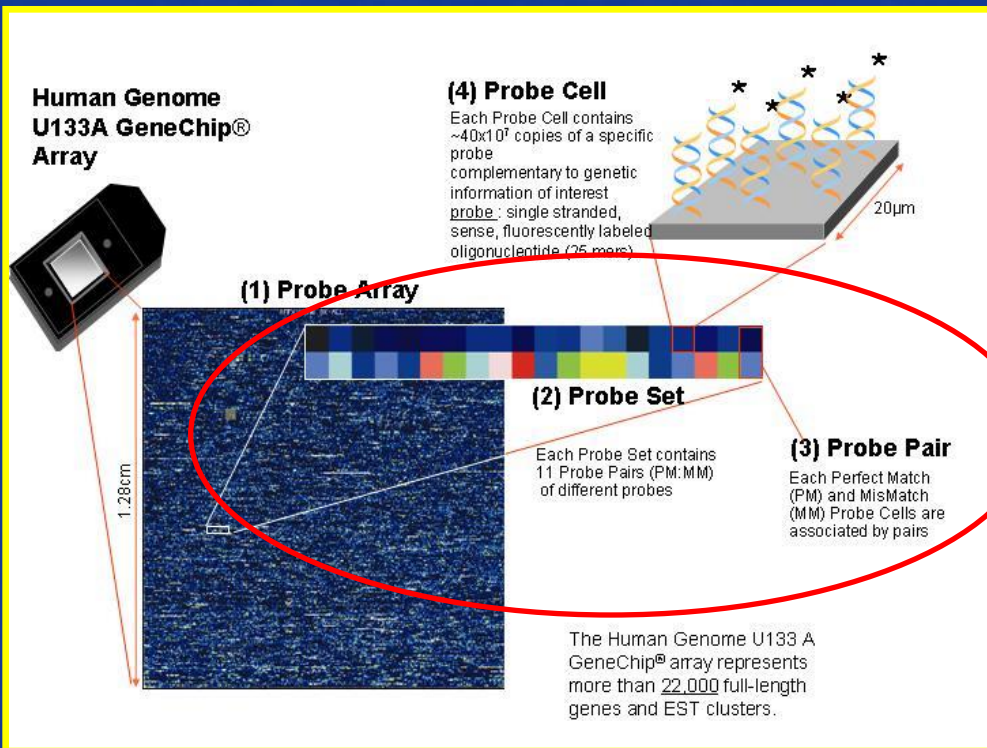
- The synthesis of these oligonucleotides on GeneChip microarrays are based on the concept of **photolithography**



- Light is shined through a mask onto a chip that has initial starting strands where the DNA will be built from
- The mask has specific tiny openings that allow the light to come in contact with the wafer at specific sections (in this diagram there are 5 probes only and each could represent a different feature)
- Any place where light hits, removes a “protective” group from the strands
- Free nucleotides (the red T) are washed over the chip and the nucleotides will combine with any strand that had lost its’ protective group in the previous step
- This is then repeated (shine light through a mask, deprotect the strands, add free nucleotides) numerous times until a each strand built is 25 base pairs long

## Probe set design:

- A group of oligonucleotide-probe pairs designed to detect the expression level of one gene transcript. Two oligonucleotides (25 bases) designed to be complementary to a reference sequence, of which one has a mismatch at the 13th position, define a **probe pair**. Mismatch probes serve as control for non-specific cross-hybridization. A probe set on the Affymetrix Human Genome U133 Plus 2.0 Array contains 11 probe pairs.





- Class prediction experiment:

- PAM method

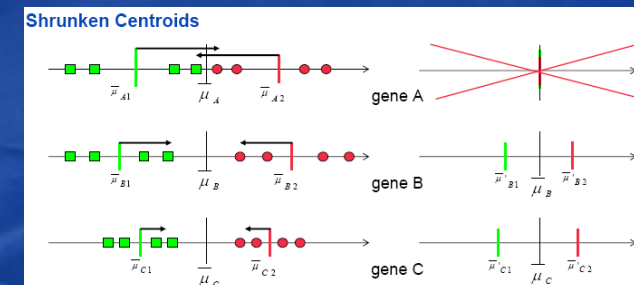
- Nearest shrunken centroid method:

- Modification of nearest centroid method, which computes a standardized **centroid** for each class in the training set. This is the **average gene expression for each gene in each class divided by the within-class standard deviation for that gene.**

- Nearest centroid classification takes the gene expression profile of a new sample, and compares it to each of these class centroids. The class, whose centroid it is closest to, in squared distance, is the predicted class for that new sample.

- Nearest shrunken centroid classification makes one important modification to the standard method. It "shrinks" each of the class centroids toward the overall centroid for all classes by an amount we call the **threshold**. This shrinkage consists of moving the centroid towards zero by subtracting the threshold, setting it equal to zero if it hits zero. For example if threshold was 2.0, a centroid of 3.2 would be shrunk to 1.2, a centroid of  $-3.4$  would be shrunk to  $-1.4$ , and a centroid of 1.2 would be shrunk to zero.

- After shrinking the centroids, the new sample is classified by the usual nearest centroid rule, but using the shrunken class centroids.



Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public



## Introduction

Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

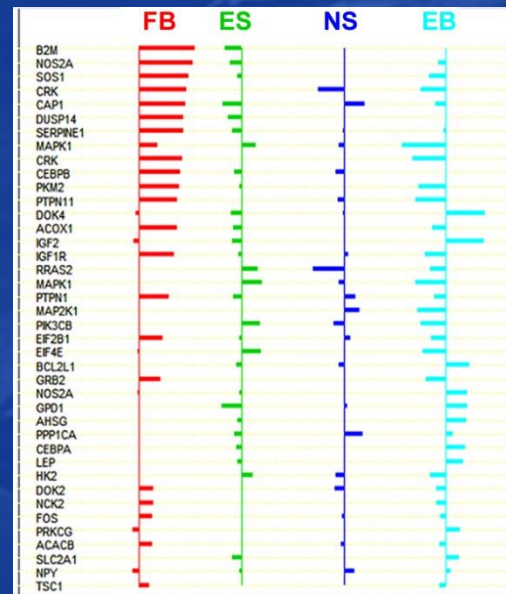
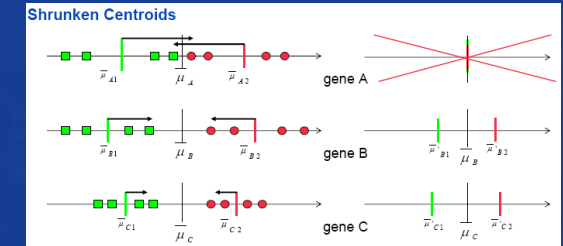
## Data analysis

- Class prediction experiment:

- PAM method
- Nearest shrunken centroid method:
  - The shrinkage has two advantages:

- (1) it can make the classifier more accurate by reducing the effect of noisy genes
- (2) it does automatic gene selection for genes that characterize the classes.
  - In particular, if a gene is shrunk to zero for all classes, then it is eliminated from the prediction rule. Alternatively, it may be set to zero for all classes except one, and we learn that high or low expression for that gene characterizes that class.

## Example

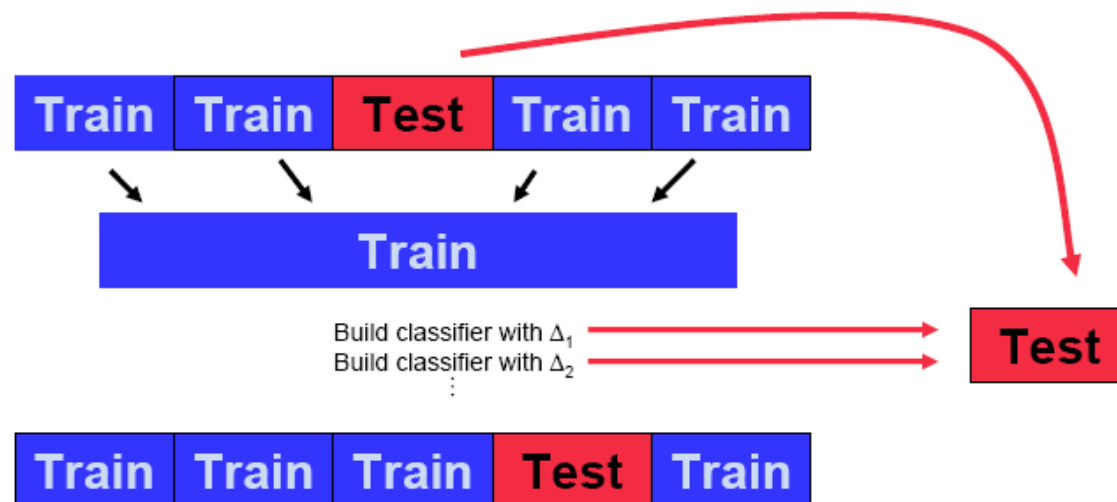


**Figure:** Identification of diagnostic markers by PAM. The shrunken class centroids for genes which have at least one nonzero difference are shown. The genes with nonzero components in each class were almost mutually exclusive and were the candidate molecular markers for the diagnosis of the four groups of cell populations

- Class prediction experiment:

- PAM method
  - Nearest shrunken centroid method:
    - The user decides on the value to use for **threshold  $\Delta$** .
    - Choosing  $\Delta$  with **cross-validation**
      - PAM does **K-fold cross-validation** for a range of threshold values.
        - The samples are divided up at random into K roughly equally sized parts. For each part in turn, the classifier is built on the other K-1 parts then tested on the remaining part. This is done for a range of threshold values, and the cross-validated misclassification error rate is reported for each threshold value.

• Idea: given a set of possible  $\Delta = \{\Delta_1, \dots, \Delta_n\}$  we want to estimate the misclassification rate for each  $\Delta$  and choose the 'best'. Use cross-validation to estimate the misclassification rate.



Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

- Class prediction experiment:

- PAM method
  - Nearest shrunken centroid method:
    - Choosing  $\Delta$  with cross-validation
      - Typically, the user would choose the threshold value giving the minimum cross-validated misclassification error rate.

Biological question

Experimental design

Microarray experiment

Image analysis

Normalization

Data analysis

Biological verification  
and interpretation

Making data public

