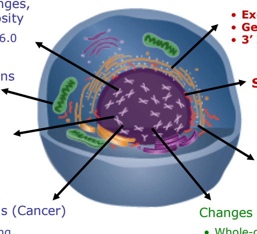
Integrating multiple sources of cellular information

Unraveling the complexities of biology requires the combination of genomic, epigenomic and functional analysis



Copy Number Changes, Loss of Heterozygosity

- Genome-wide SNP 6.0
- Cytov2

Mitochondrial Mutations

- Mitov2 Resequencing

Germline Mutations

- Genome-wide SNP 6.0
- Targeted Genotyping
- Targeted Sequencing

Somatic Mutations (Cancer)

- Targeted Genotyping
- Targeted Sequencing
- Genome-wide SNP 6.0

Differential Gene Expression

- **Exon 1.0**
- **Gene 1.0**
- **3' IVT Expression**

Splice Variant Analysis

- **Exon 1.0**
- ***Gene 1.0***
- **Custom Splice Junctions**

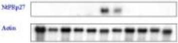
Micro RNA Analysis

- **miRNA**

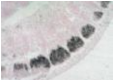
Changes in Methylation Patterns

- Whole-genome Tiling Arrays
- Promoter Tiling Arrays

There is a variety of techniques available to study gene expression

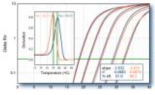


Gel based methods, e.g. Northern blots, RT-PCR

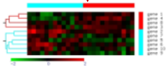


Fluorescent *in situ* hybridisation

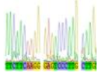
Gene expression



Quantitative RT-PCR



Whole genome microarrays



Next-generation sequencing

History of the microarray

1989:
1st prototype

1996:
WG on 1 array
(~5.6k genes)

2004:
WG on 1 array
(47k transcripts)

2008:
WG 96 array plate

More information on less space

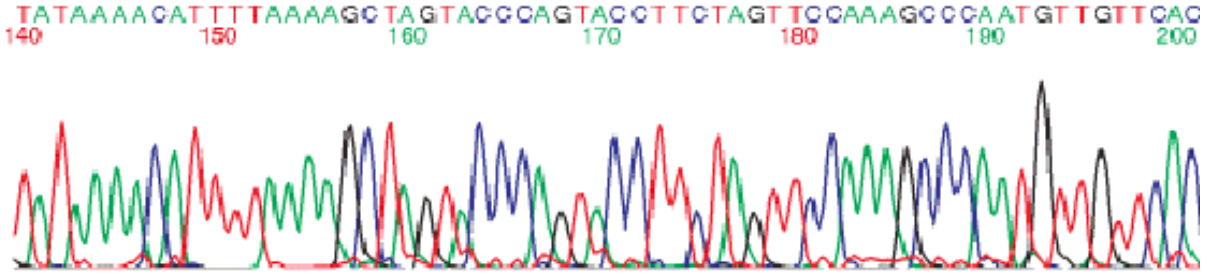
1991:
Landmark
publication

2001:
WG on 2
arrays (39k
transcripts)

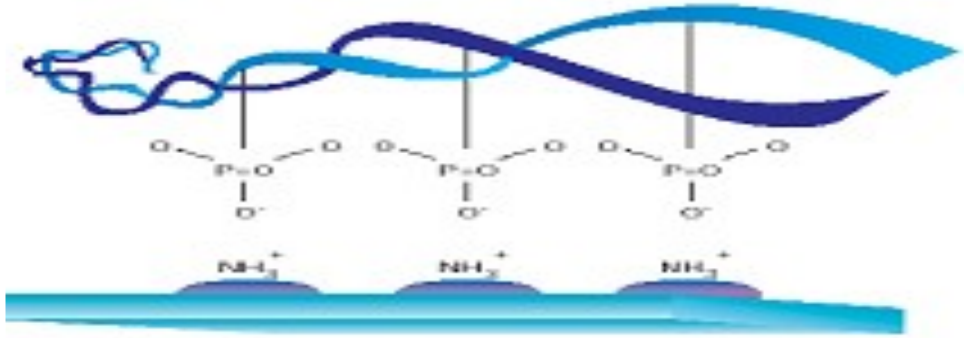
2006:
Exon 1.0 ST
array (1.4m
exons)

Microarray

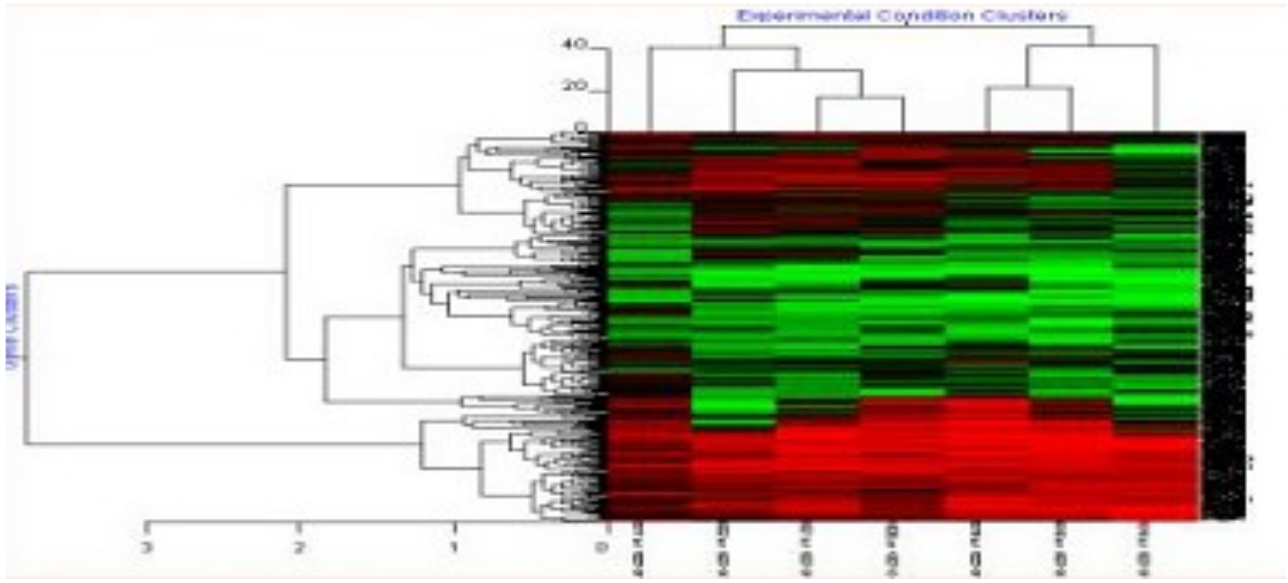
Genomic



Chemistry

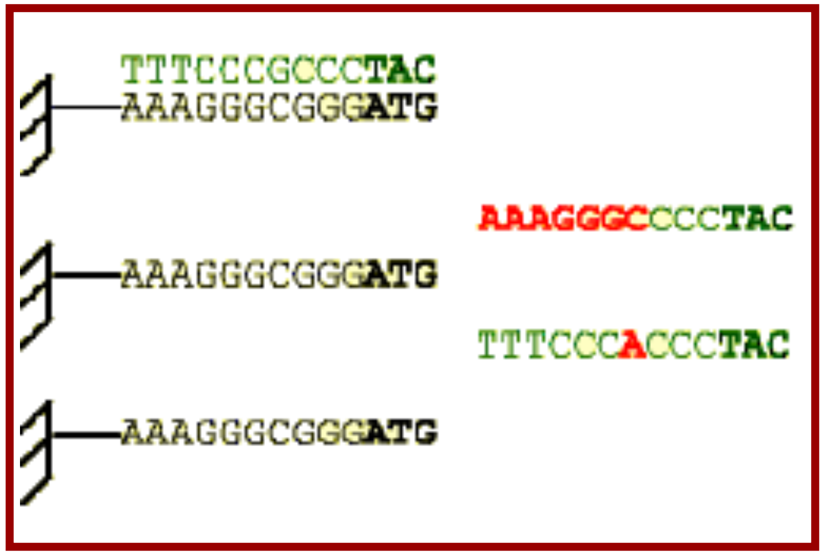


Bioinformatics

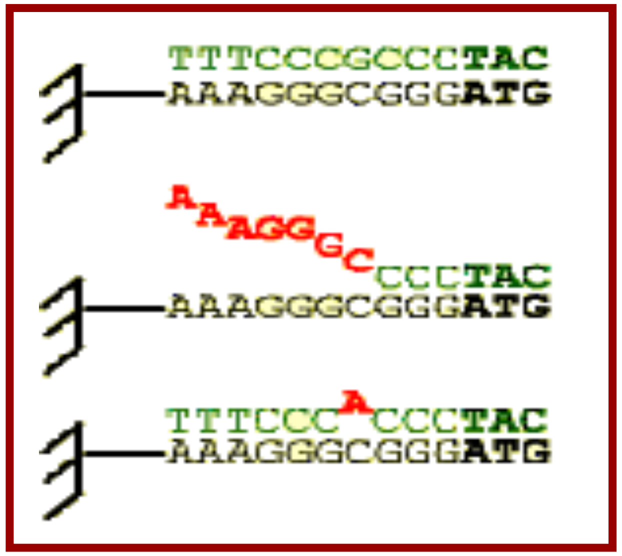


Ibridation

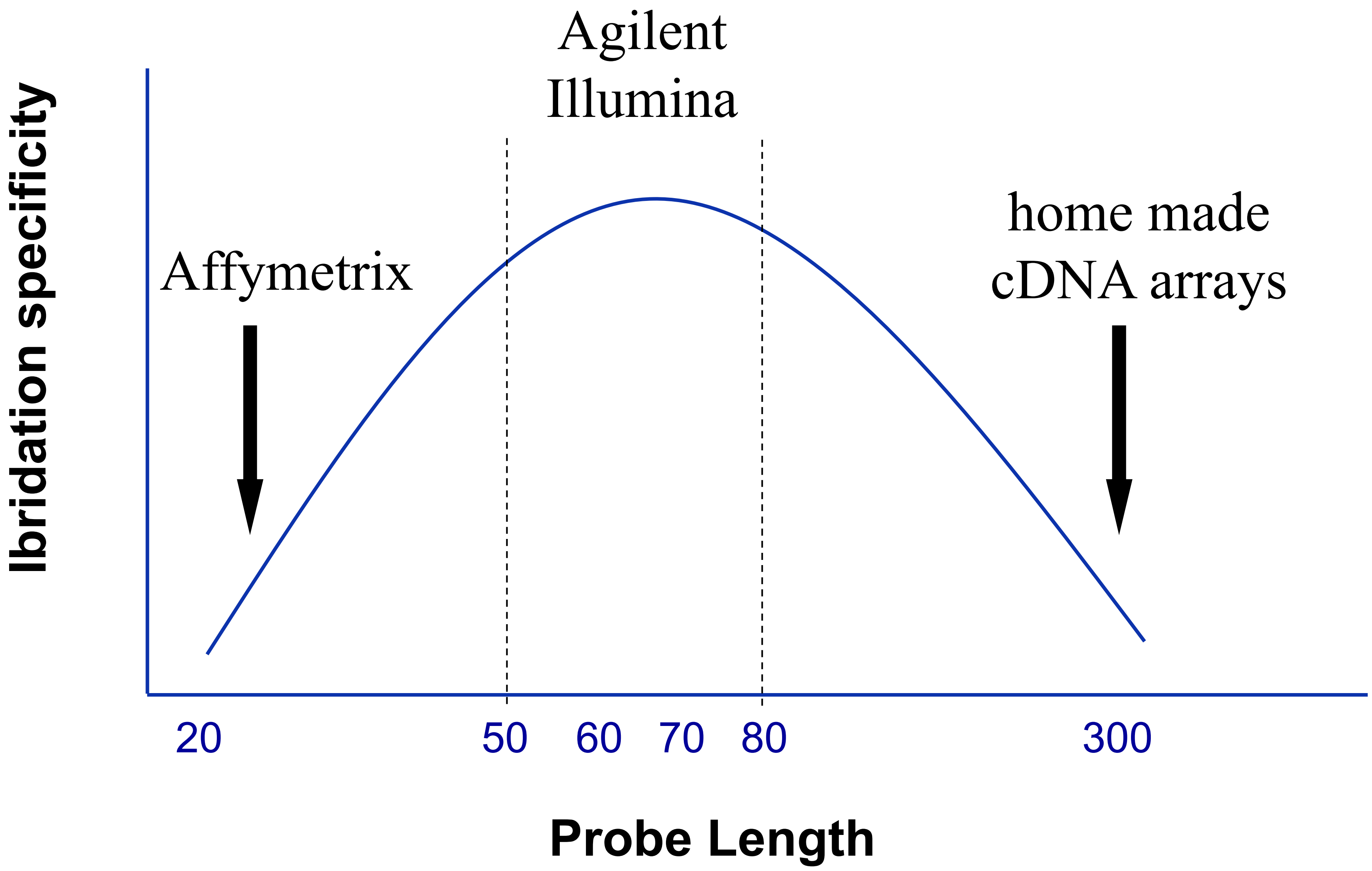
HIGH stringent condition



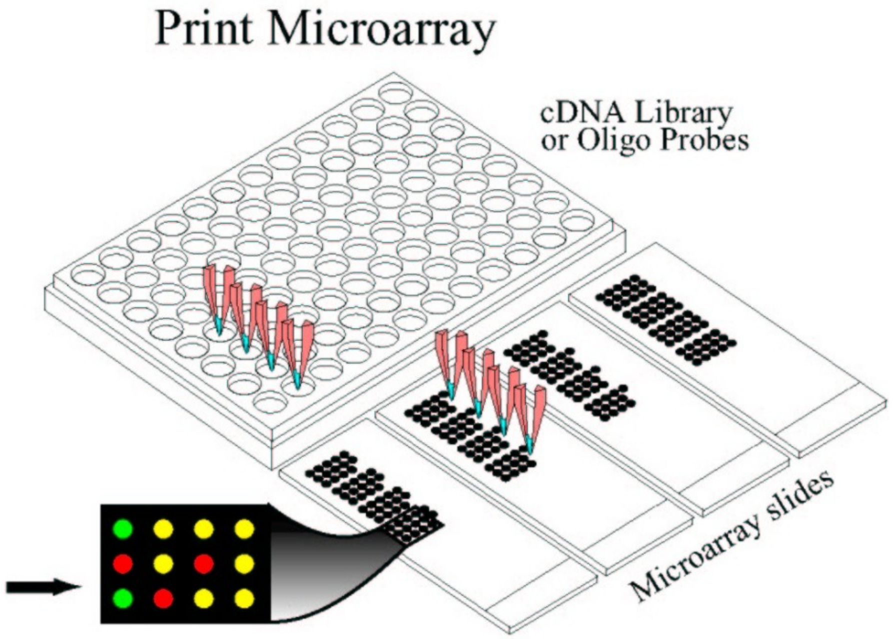
LOW stringent condition

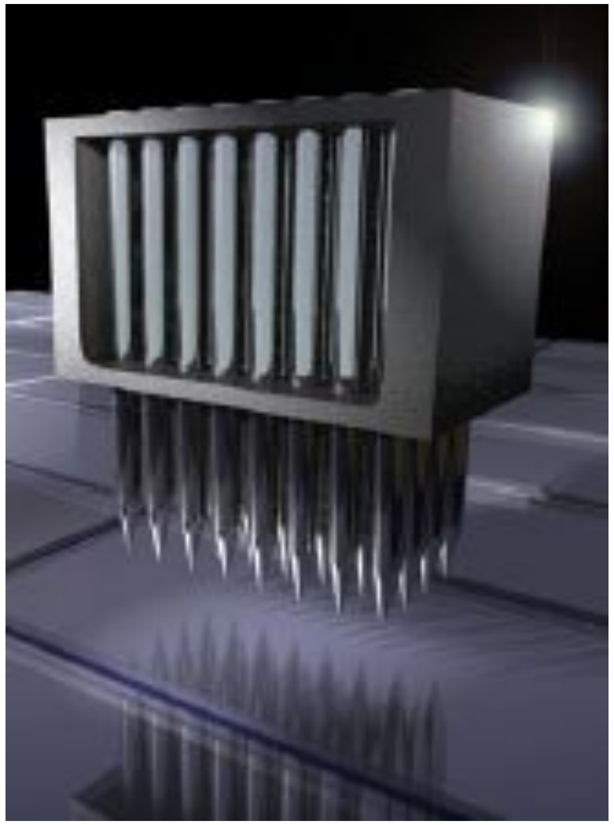


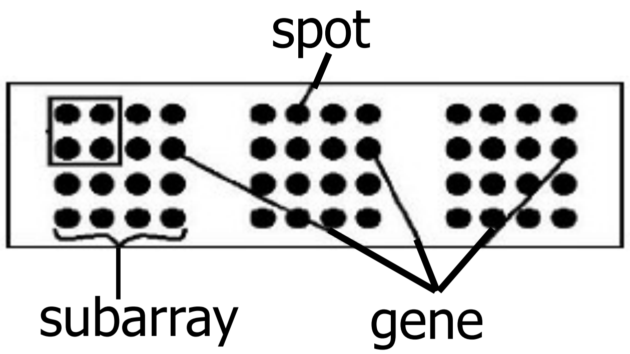
Probes


Agilent
Illumina
Affymetrix
home made
cDNA arrays
Ibridation specificity
20
50
60
70
80
300
Probe Length

Spotted or Printed Array

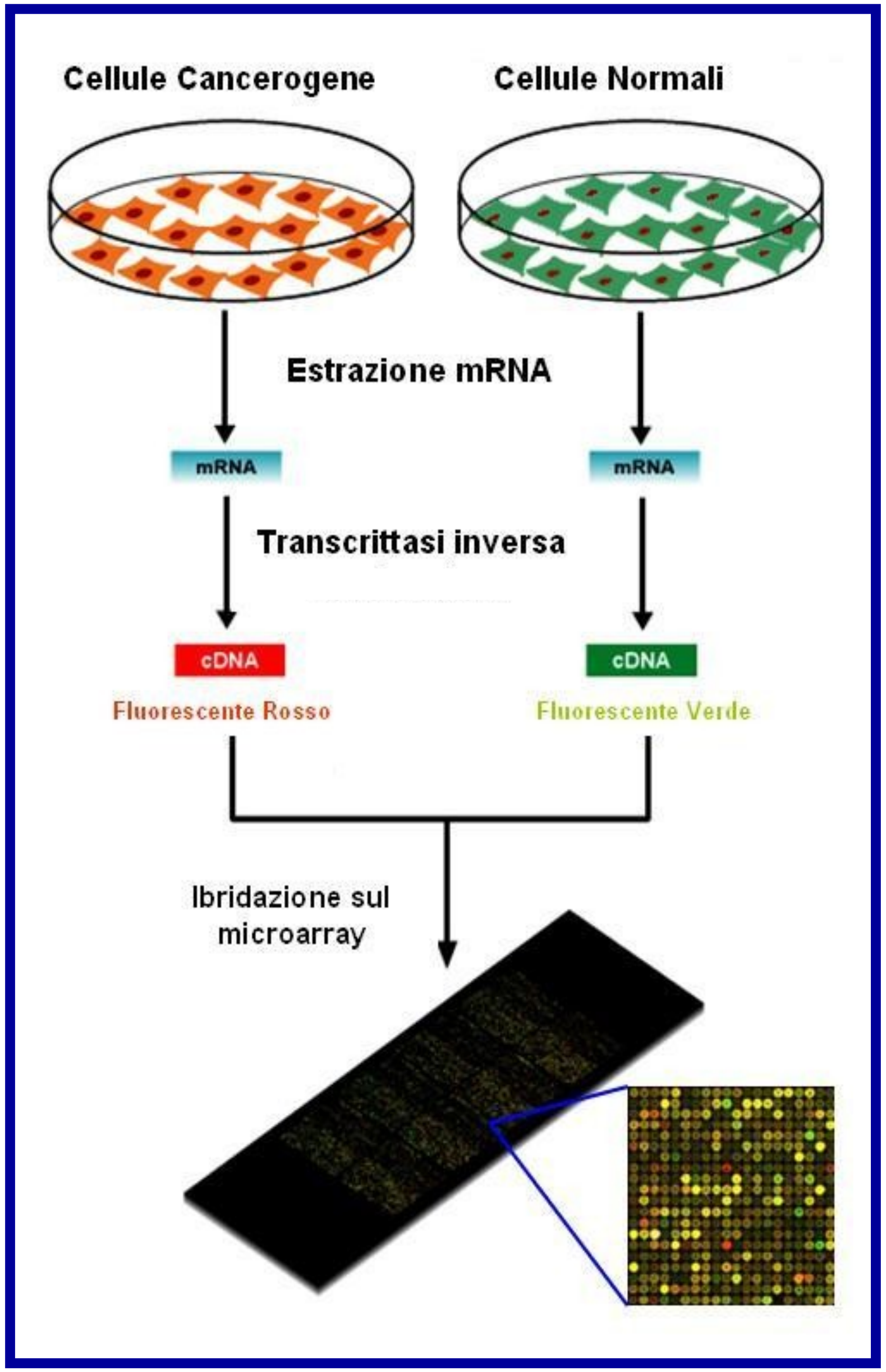






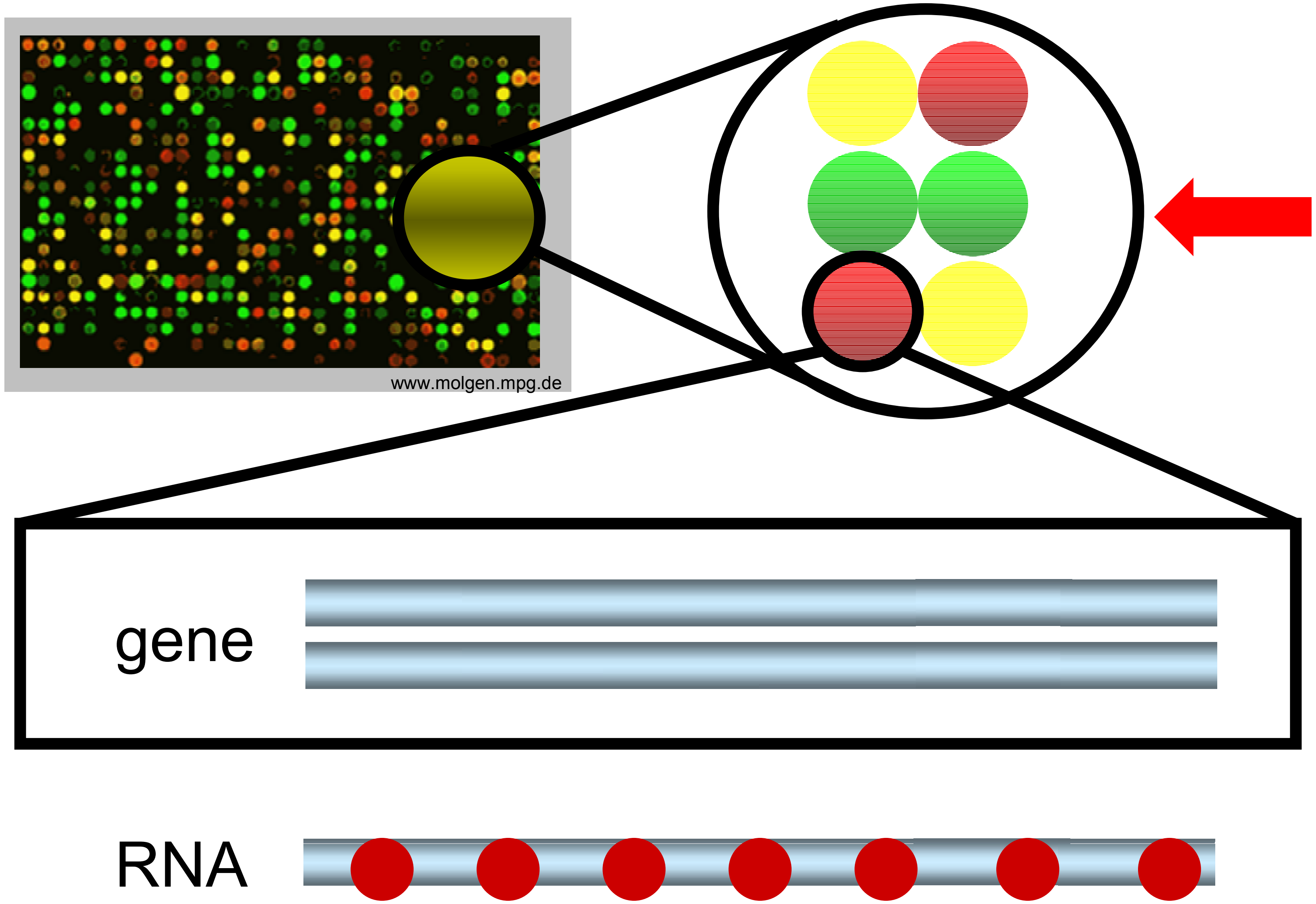
Density
10000-30000

Spotted or Printed Array



Spotted or Printed Array

A closer look at Spotted microarrays
Some nomenclature



www.molgen.mpg.de

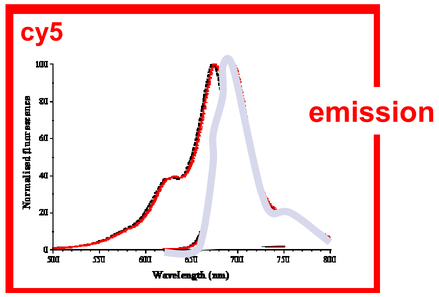
each spot represents a gene or gene fragment

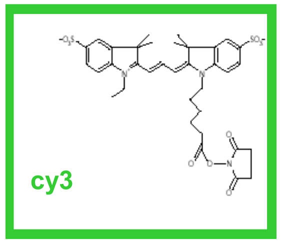
“probe”

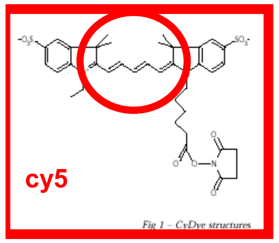
“target”

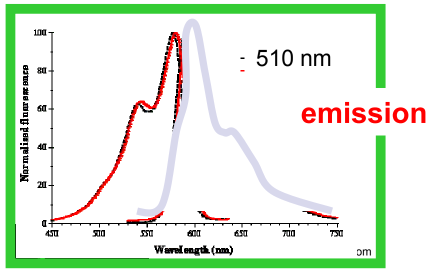
Spotted or Printed Array

cy3 and cy5: Commonly used dyes



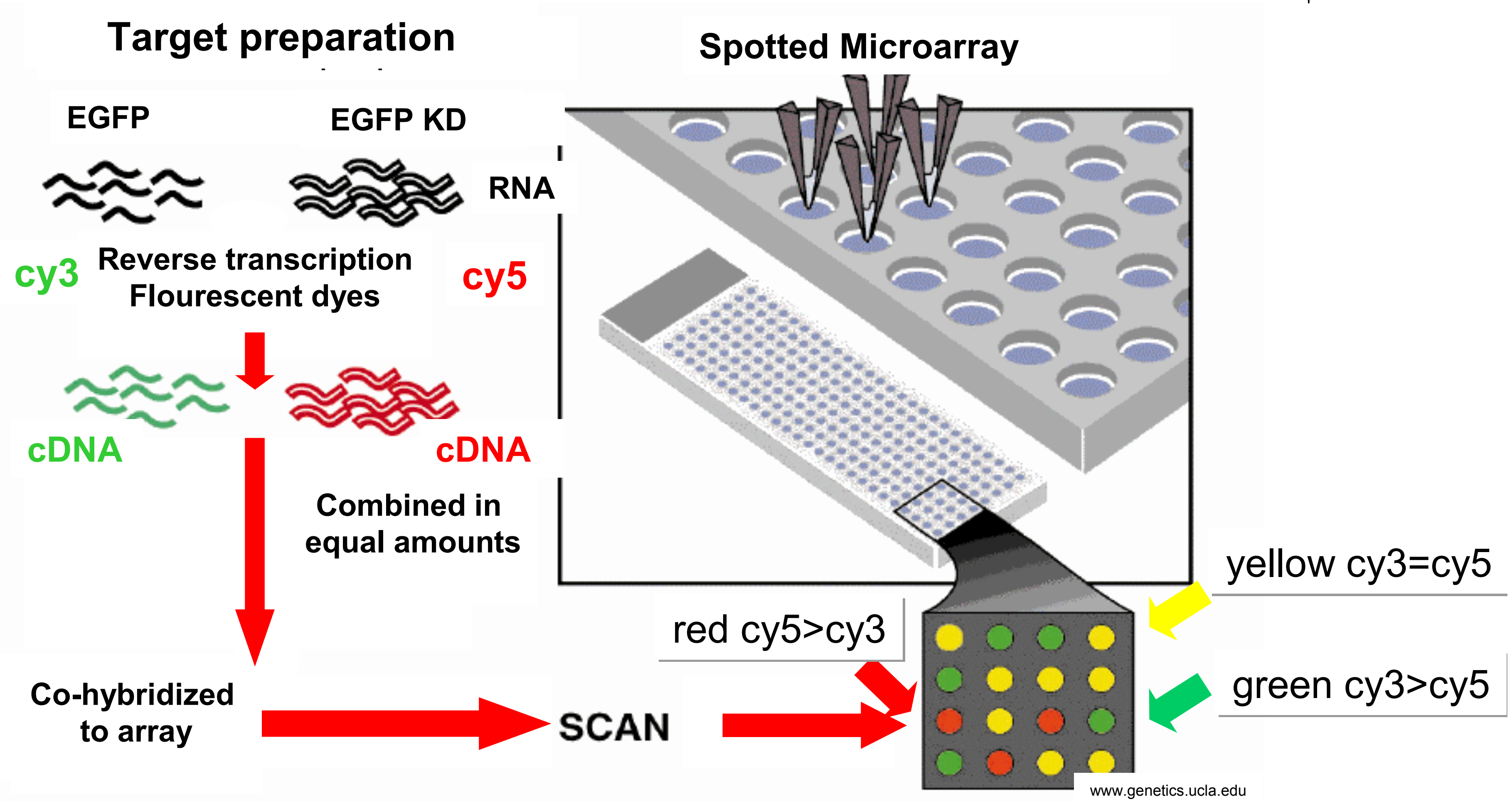






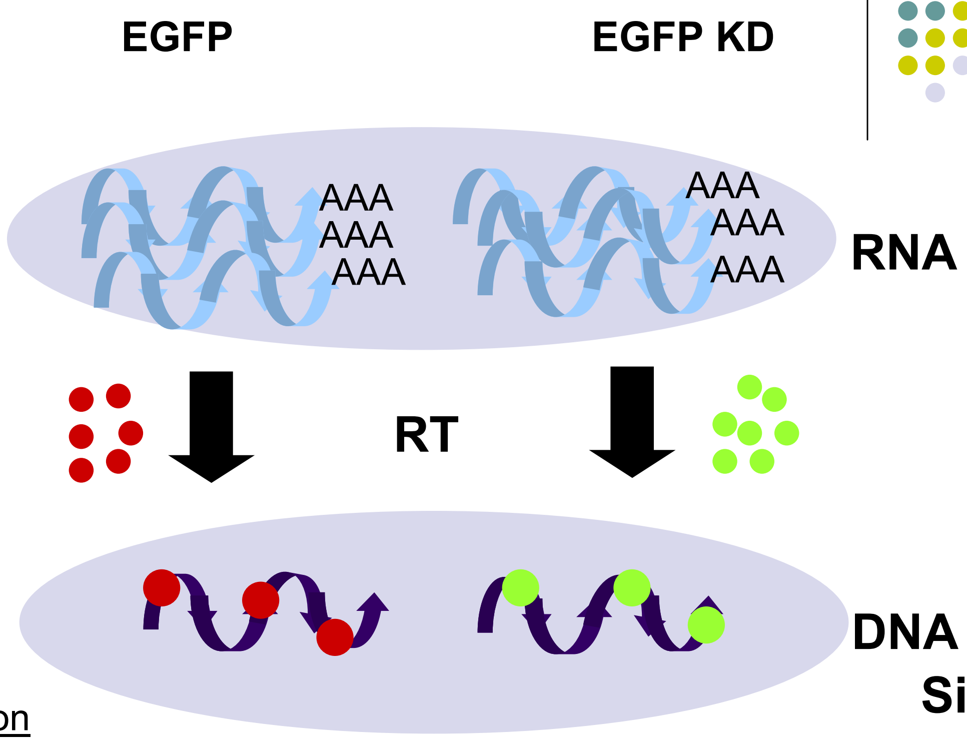
Differential dye incorporation
cy5 less well than cy3
Light sensitivity: cy5 more easily degraded

Spotted microarray target preparation
Direct labeling



www.genetics.ucla.edu

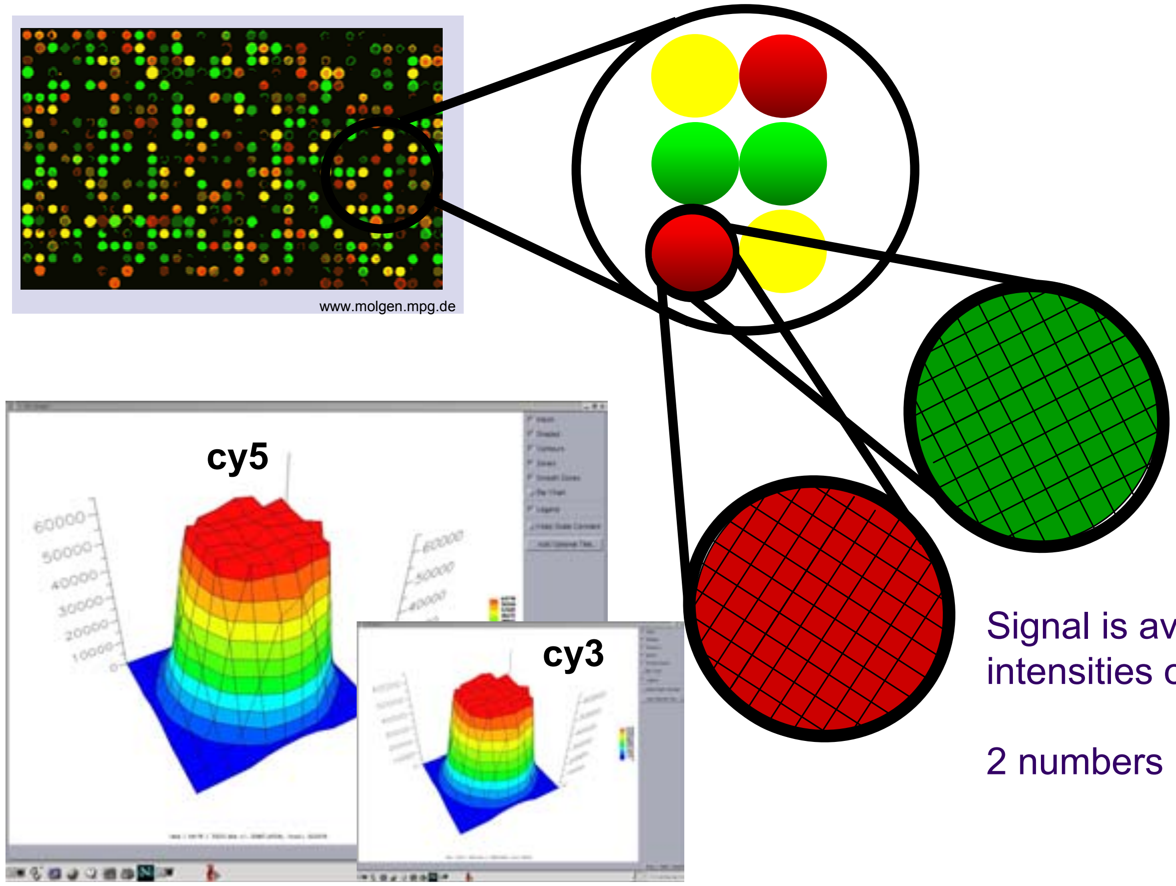
Spotted or Printed Array



Reverse transcription
RNA
Enzyme (Superscript RT)
Dye
Oligo d(T)
nucleotides

Signal: Spotted arrays

Spotted microarrays



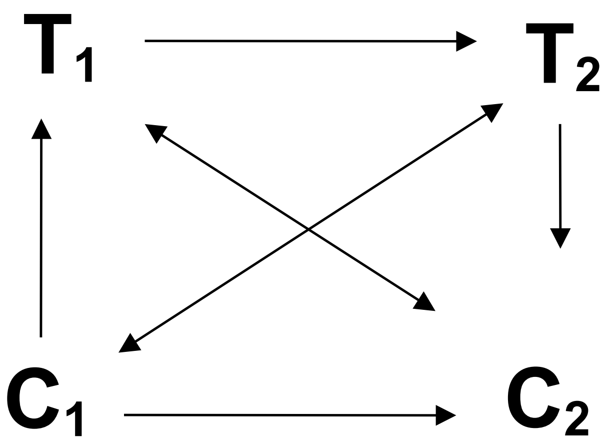
www.molgen.mpg.de

Signal is average of pixel
intensities of spot

2 numbers per spot

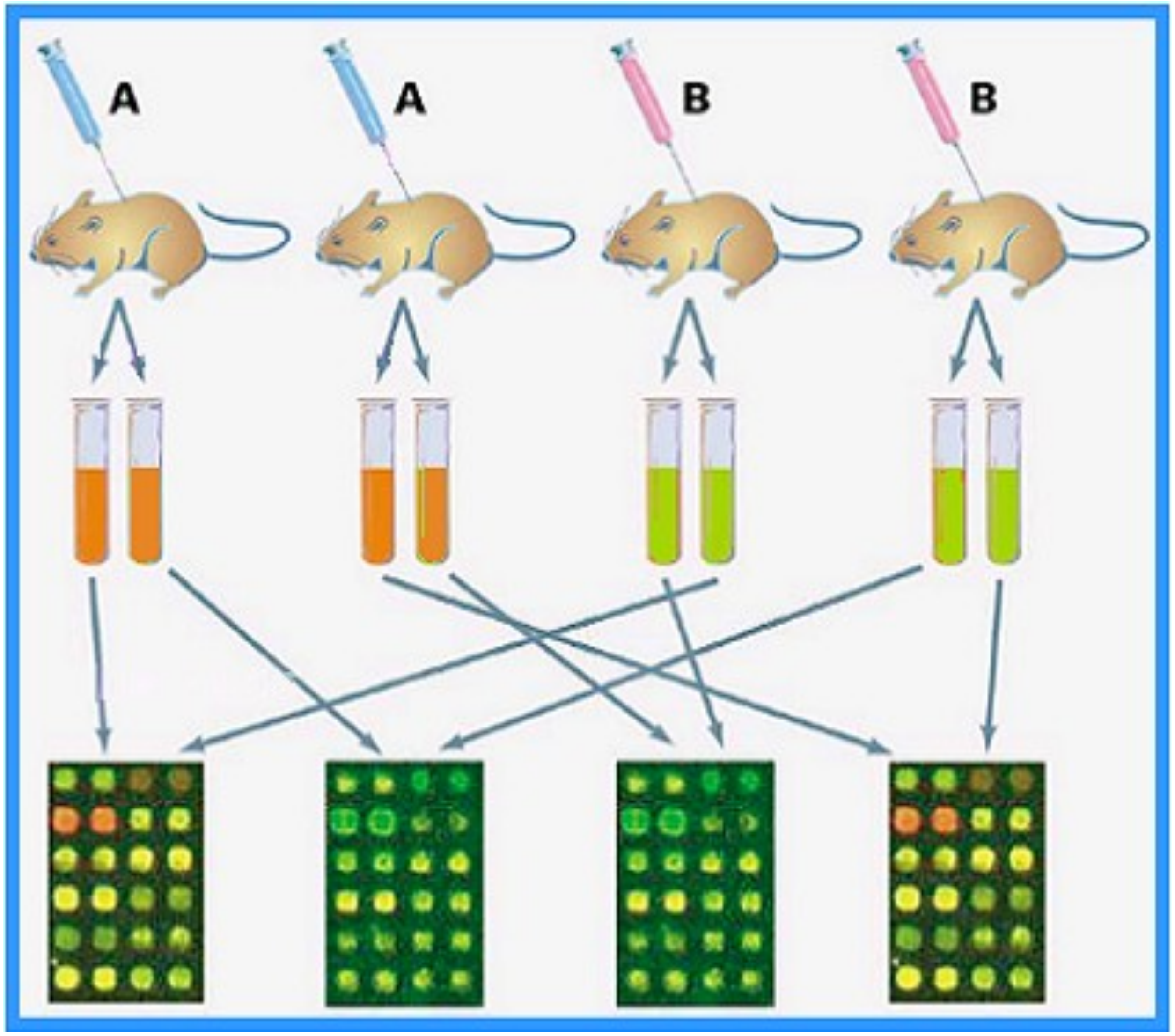
Spotted or Printed Array

Biological and technical replicates are essential



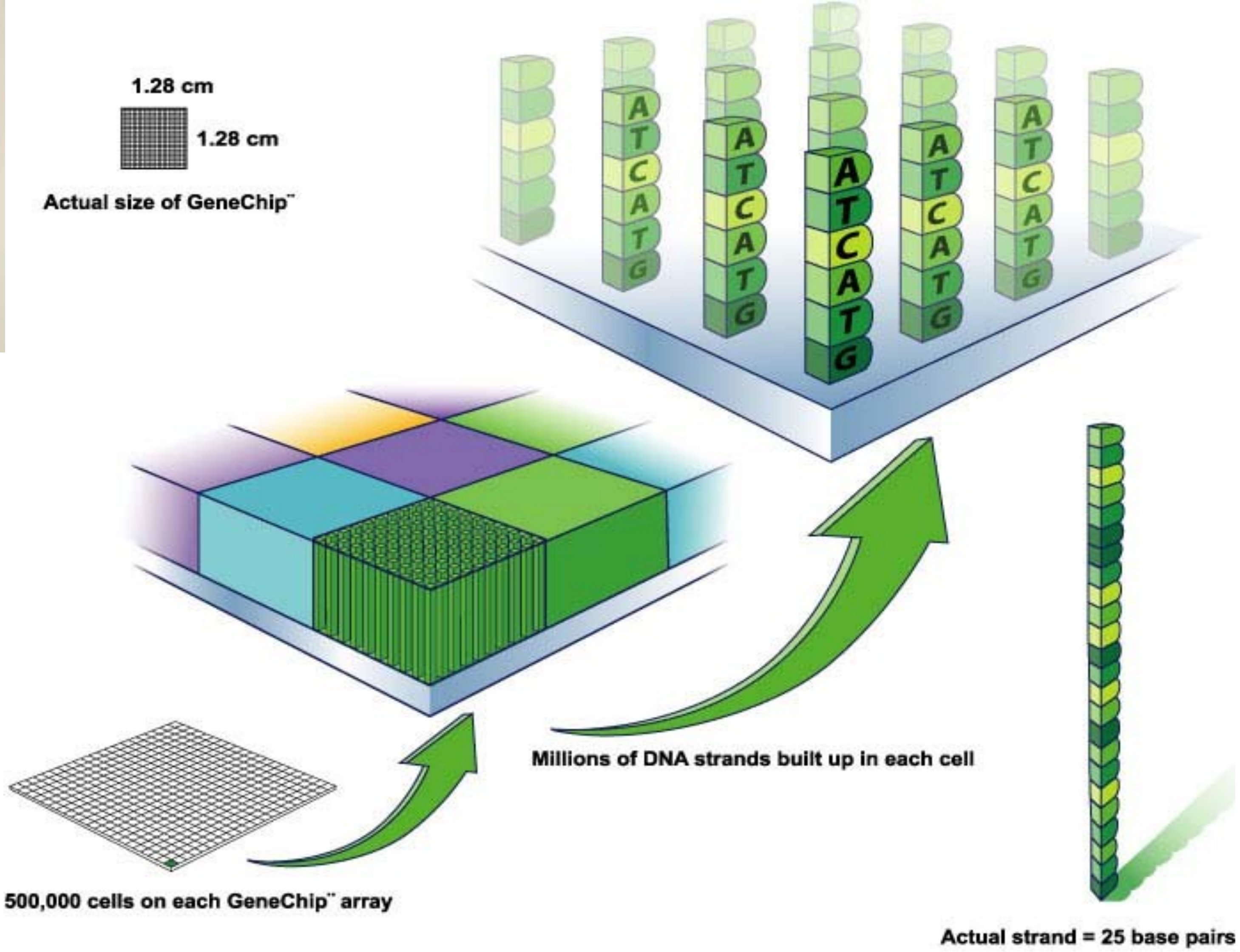
T_1 T_2

C_1 C_2



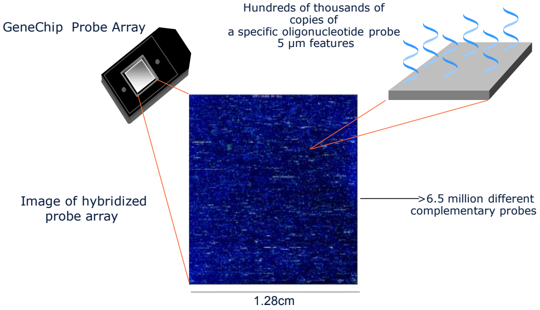
Genechip Array

GENECHIPS ARRAYS


1.28 cm
1.28 cm
Actual size of GeneChip™
A T C A T G
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs

Probe density: 500000 till 10^6

GeneChip® Probe Arrays



GeneChip Probe Array

Hundreds of thousands of copies of a specific oligonucleotide probe 5 µm features

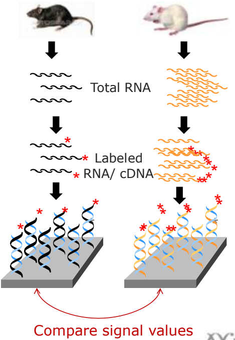
Image of hybridized probe array

>6.5 million different complementary probes

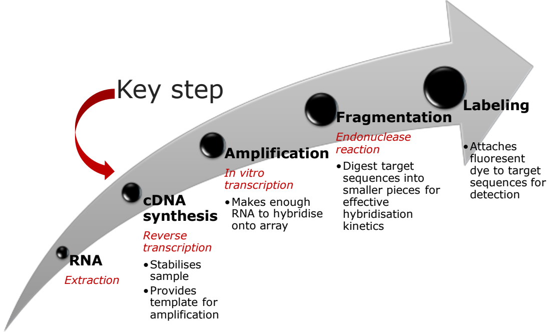
1.28cm

What is a gene expression microarray?

- Powerful tool to simultaneously measure the expression of thousands of genes from a single sample
- Contains thousands of copies of individual oligonucleotide probes
- Each probe is complimentary to a target RNA sequence
- Array applications in research
  - Gene discovery
  - Biomarker/ gene signatures
  - Global expression changes
  - Profiling a large number of genes that are time and cost prohibitive by alternative methods
  - Genotyping



The array assay prepares sample for hybridization



Key step

RNA

Extraction

cDNA synthesis

Reverse transcription

- Stabilises sample
- Provides template for amplification

Amplification

In vitro transcription

- Makes enough RNA to hybridise onto array

Fragmentation

Endonuclease reaction

- Digest target sequences into smaller pieces for effective hybridisation kinetics

Labeling

- Attaches fluoresent dye to target sequences for detection

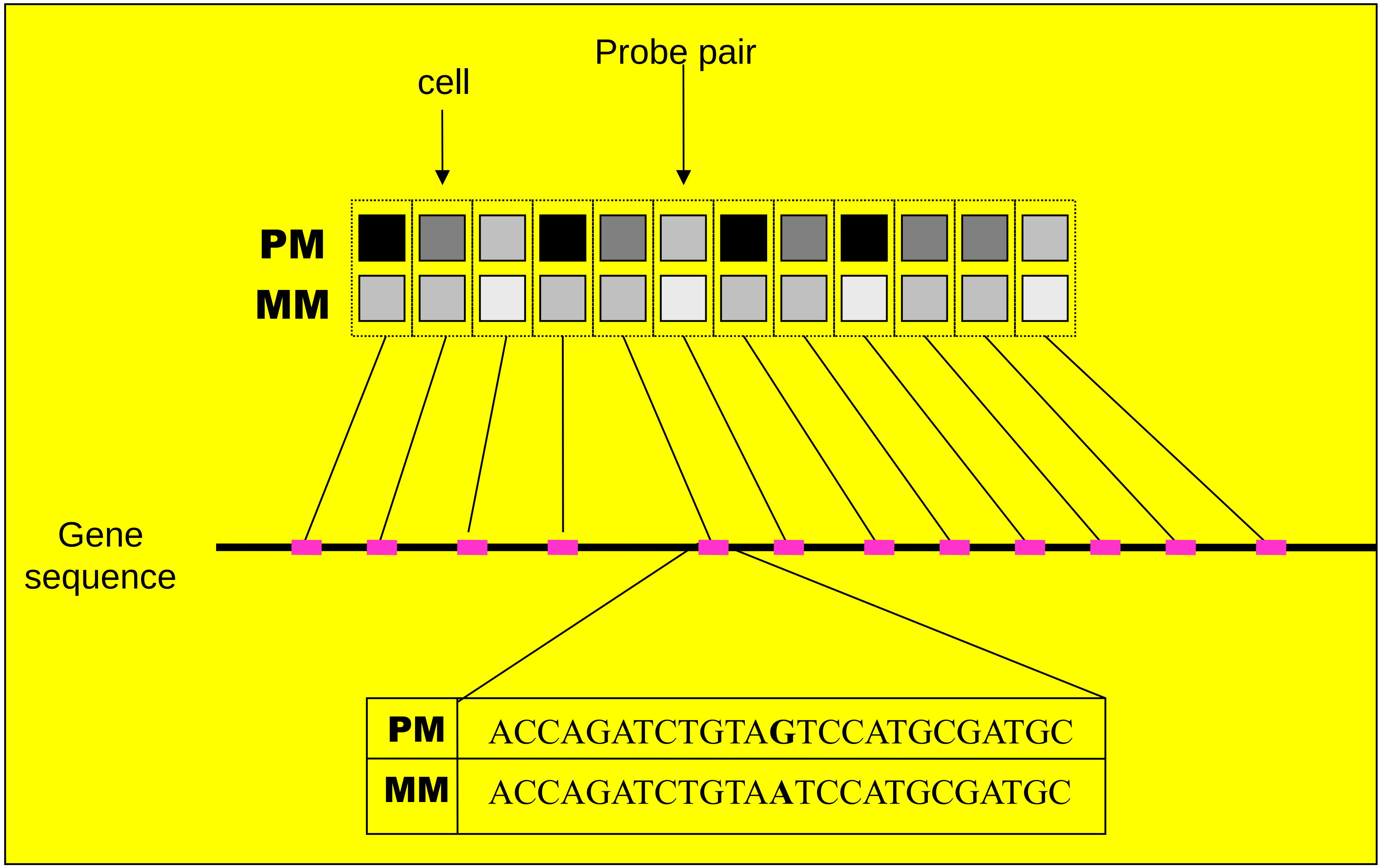
GeneChips detect transcripts using multiple features: **The probe set**

- The power of the probe set
  - Each transcript detected by multiple independent 25mer probes
  - Provides an inherent set of replicate data points
  - Generates high sensitivity without loss of specificity

- Probe set is unique to Affymetrix
  - High densities achievable through photolithographic manufacturing process
  - Features belonging to a probe set are distributed around the array

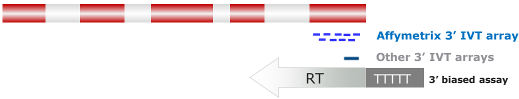
- 25mer oligos are highly specific
  - Differentiate between sequences with 90% identity
  - Highly homogeneous and controlled hybridisation events

Probe set (Affymetrix)

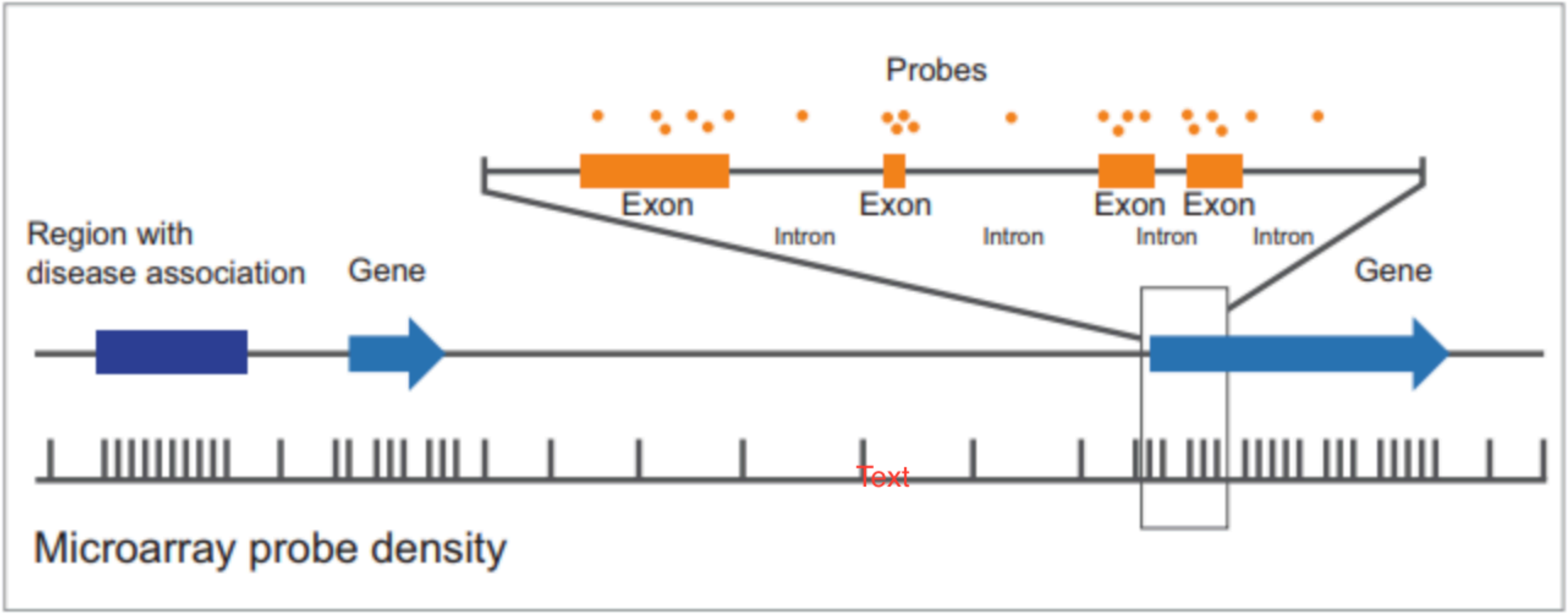


PM	ACCAGATCTGTA**G**TCCATGCGATGC
MM	ACCAGATCTGTA**A**TCCATGCGATGC

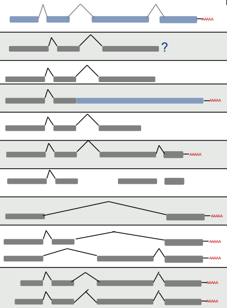
Traditional arrays measure expression at the 3’ end of the gene



- Traditional arrays have probe sets targeted at 3’ end of the gene
- Accompanied by an assay that is 3’biased
- Provide some insight into global gene expression, but assumes:
  - All transcripts have clear, defined 3’ ends
  - All transcripts have a poly-A tail
  - Entire length of a gene is expressed as a single unit

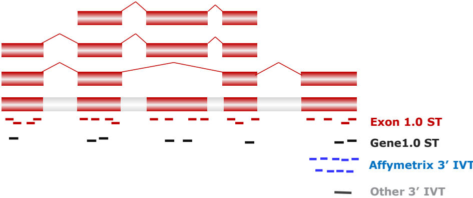

Probes
Exon
Exon
Exon
Exon
Intron
Intron
Intron
Intron
Region with disease association
Gene
Gene
Text
Microarray probe density

Why Limit Your Discoveries to the 3’ End of a Gene?



	Classical 3’ Assay	**WT Assay**
Presumed standard transcript	●	●
Transcripts with undefined 3’ end		●
Non-polyadenylated messages		●
Genomic translocations		●
Truncated transcripts		●
Alternative polyadenylation sites		●
Degraded samples		●
Genomic deletions		●
Alternatively spliced transcripts		●
Alternative 5’ start sites		●

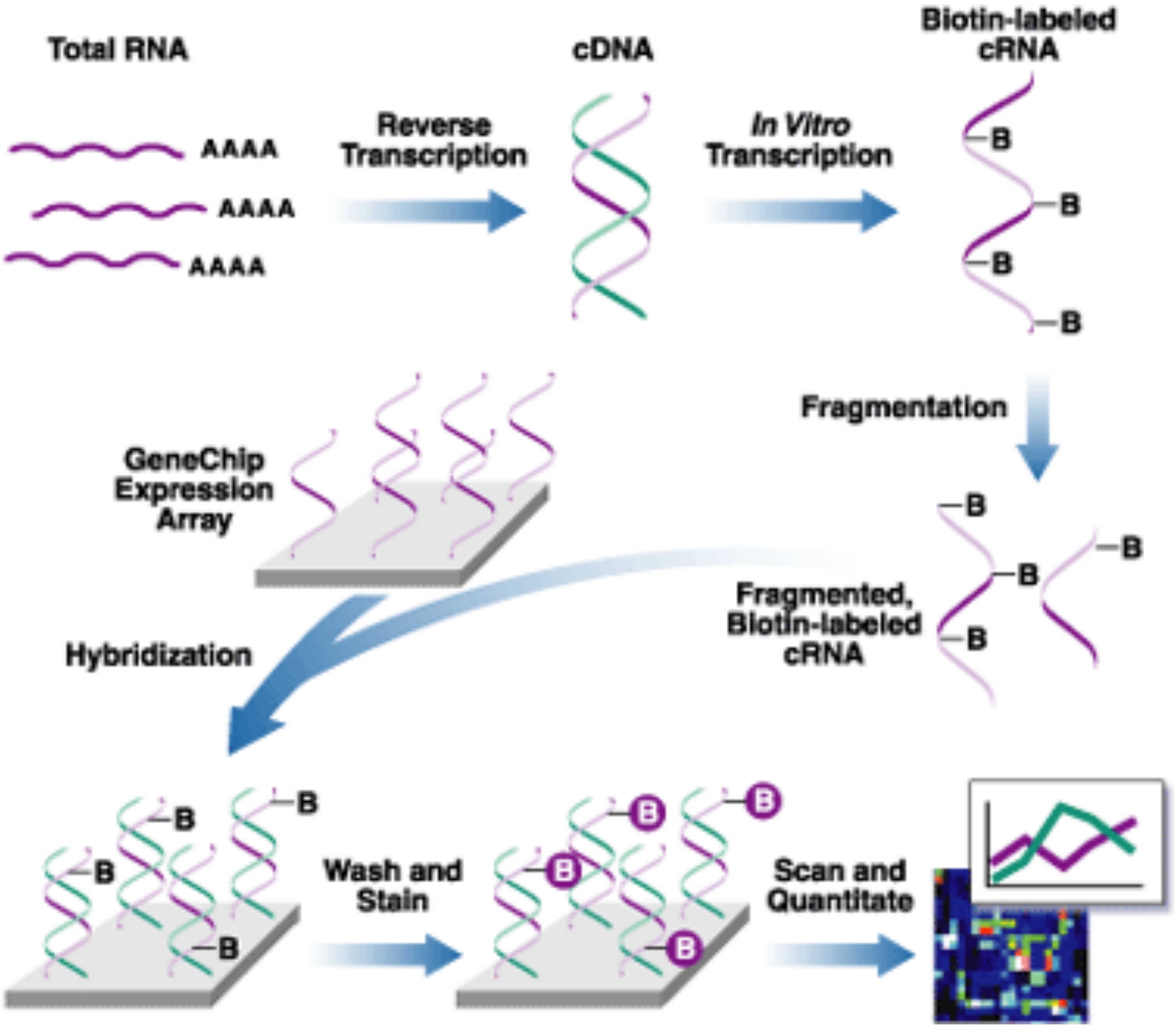
Whole Transcript (WT) arrays have probes throughout the entire transcript



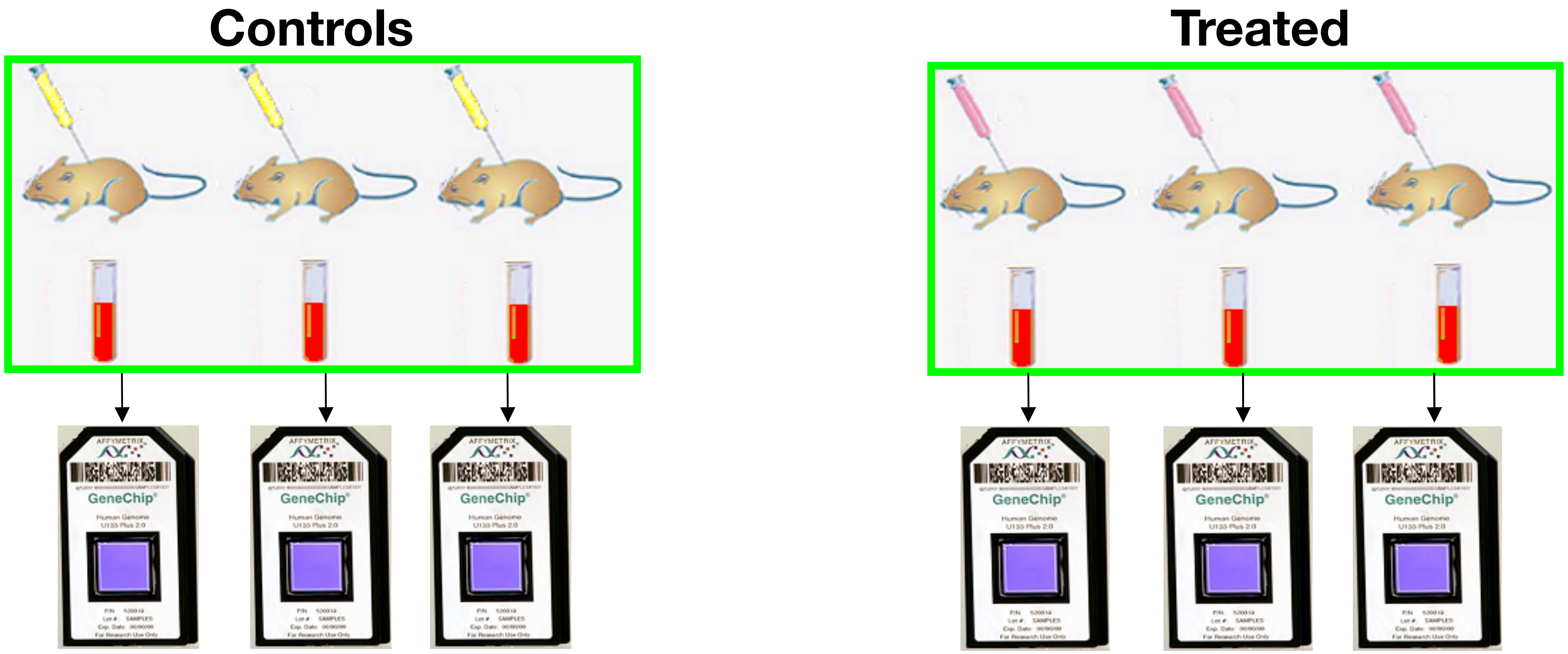
- **Exon 1.0 ST arrays**
  - ~4 probes per **exon**
  - ~40 probes per transcript
  - Predicted & annotated content
- **Affymetrix 3' IVT Arrays**
  - 11 probes per transcript
  - Well annotated content

- **Gene 1.0 ST arrays**
  - ~1-2 probes per **exon**
  - ~26 probes per transcript
  - Well annotated content
- **Other 3' Arrays**
  - ~1-5 probes per transcript
  - Well annotated content

Genechip Array



Genechip Array



No technical replicates
high results reproducibility
Multiple measures for each RNA

Number o replicates:

- cell lines -> 3
- animals -> 3-5
- human sample -> 20/50

Genechip Array

GeneChip® Platform

Design Experiment

Prepare Sample

Probe Array

Hybridize

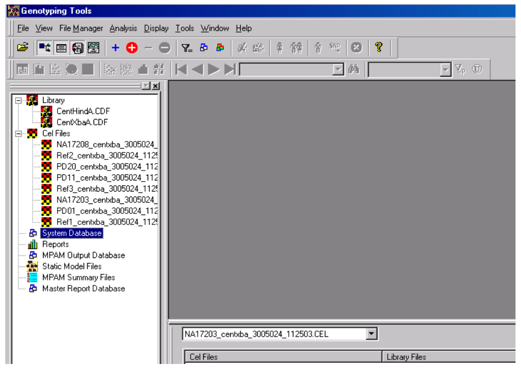
Hybridization Oven

Wash & Stain

Fluidics Station

Scan

Scanner



- Library
  - CentHindA.CDF
  - CentXbaA.CDF
- Cel Files
  - NA17208_centxba_3005024_
  - Ref2_centxba_3005024_112
  - PD20_centxba_3005024_112
  - PD11_centxba_3005024_112
  - Ref3_centxba_3005024_112
  - NA17203_centxba_3005024_
  - PD01_centxba_3005024_112
  - Ref1_centxba_3005024_112
- System Database
- Reports
- MPAM Output Database
- Static Model Files
- MPAM Summary Files
- Master Report Database

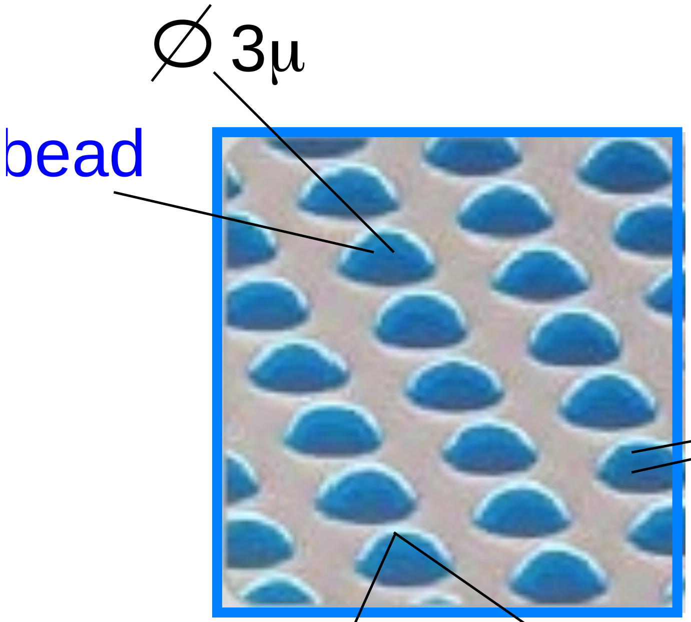
NA17203_centxba_3005024_112503.CEL

Cel Files	Library Files

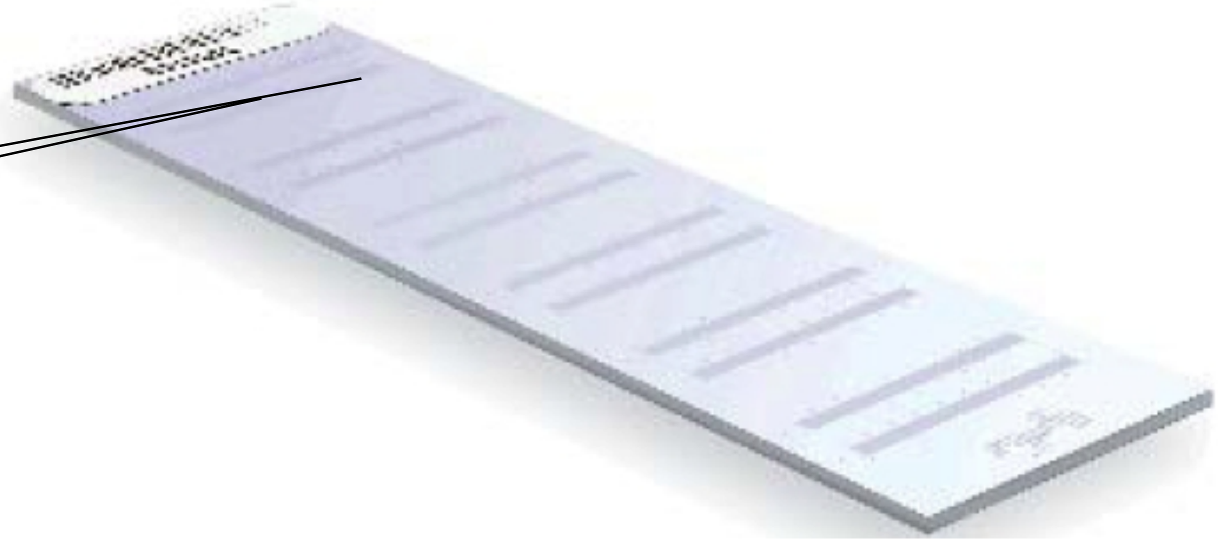
Data Analysis Software

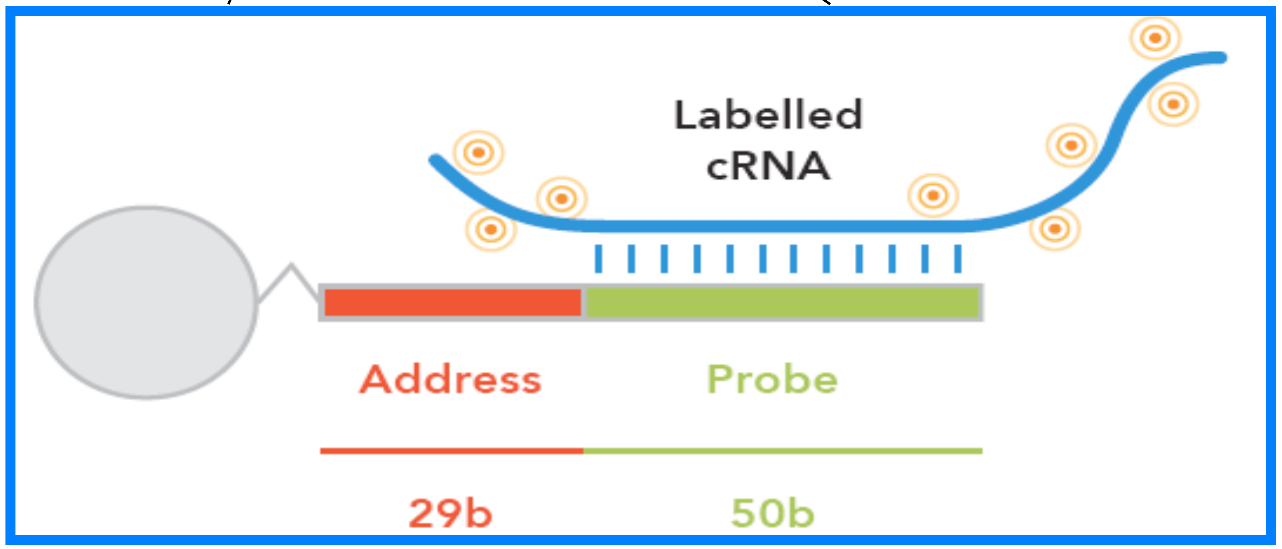
Beads Array

Arrays illumina (beads array)



Illumina Whole-Genome Gene Expression BeadChips consist of oligonucleotides immobilized to beads held in microwells on the surface of an array substrate





Direct Hyb Gene Expression Profiling Bead Design

Beads Array

The basics of the RNA amplification are:

- Hybridization of a oligo-dT oligonucleotide to the polyA component of the total RNA. The oligonucleotide also has the sequence for a viral T7 RNA polymerase promoter.

- Extend the cDNA, then synthesize a second strand to generate double stranded cDNA.

- Add T7 RNA polymerase and nucleotides to linearly amplify the RNA. The nascent aRNA incorporates biotin-modified dUTP.

- Hybridize the biotin-modified aRNA to the BeadChip.

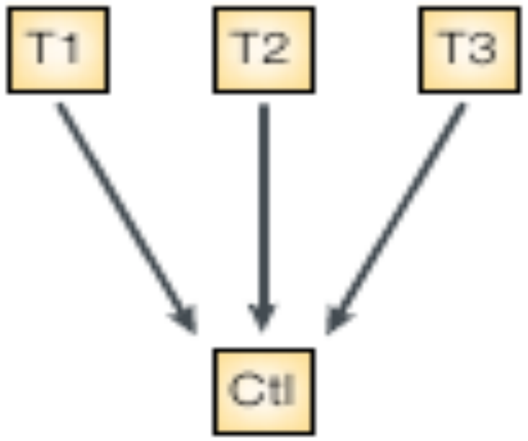
- Stain the BeadChip with Cyanine 3 derivatized to streptavidin.

- Scan on a high resolution Illumina BeadStation scanner.

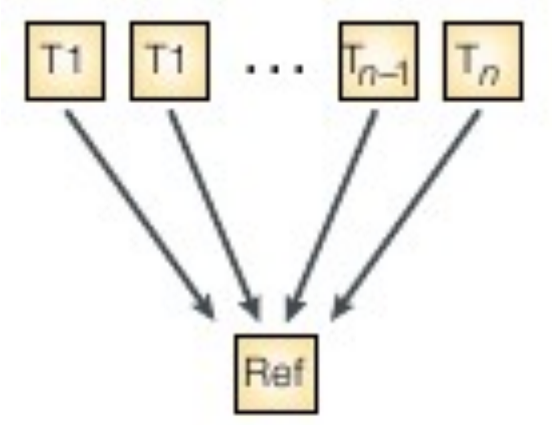
Experimental design

The two designs represented below are best answered by a common reference design.

Case 1:
Use of meaningful biological control (Ctl).
<u>Samples:</u> Liver tissue from mice treated with cholesterol modifying drugs and from untreated (Ctl) mice.
<u>Question 1:</u> The expression of which genes differs between the treated and untreated (Ctl) mice?
<u>Question 2:</u> Which genes respond similarly to two or more treatments, when compared to wild-type?

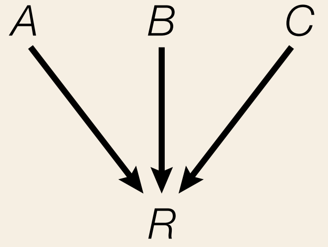
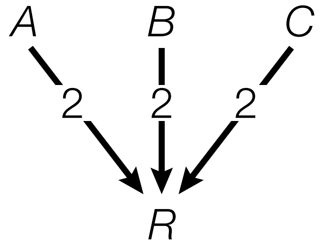
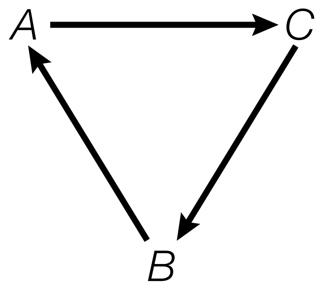


Case 2:
Use of universal reference (Ref).
<u>Samples:</u> Tissue from different tumours.
<u>Question:</u> What are the tumour subtypes?



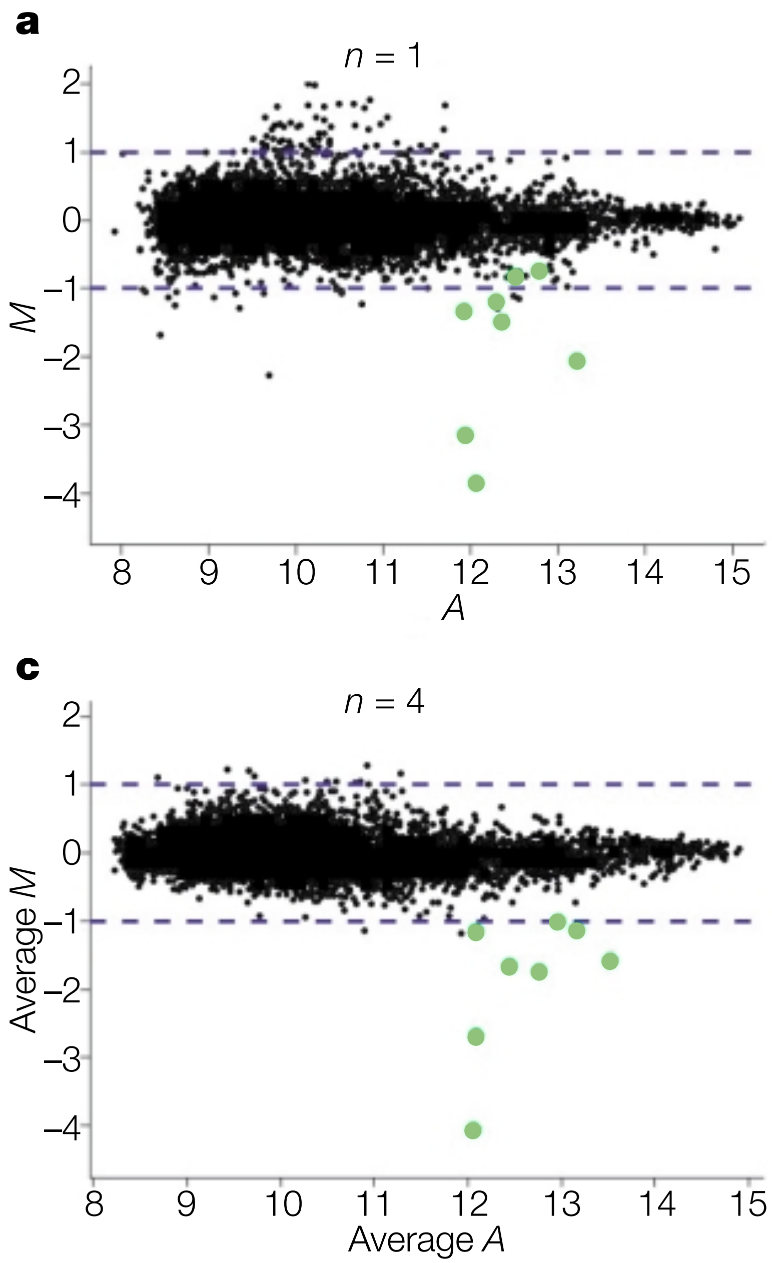
Experimental design

Table 1 | **Single-factor experiments**

Design choices	**Number of slides**	**Units of material (number of samples)**	**Average variance**
Indirect designs			
Design I 	3	$A = B = C = 1$	2.00
Design II 	6	$A = B = C = 2$	1.00
Direct design			
Design III 	3	$A = B = C = 2$	0.67

Variance of estimated effects for three different designs of single-factor experiments. σ^2 was set to 1 throughout.

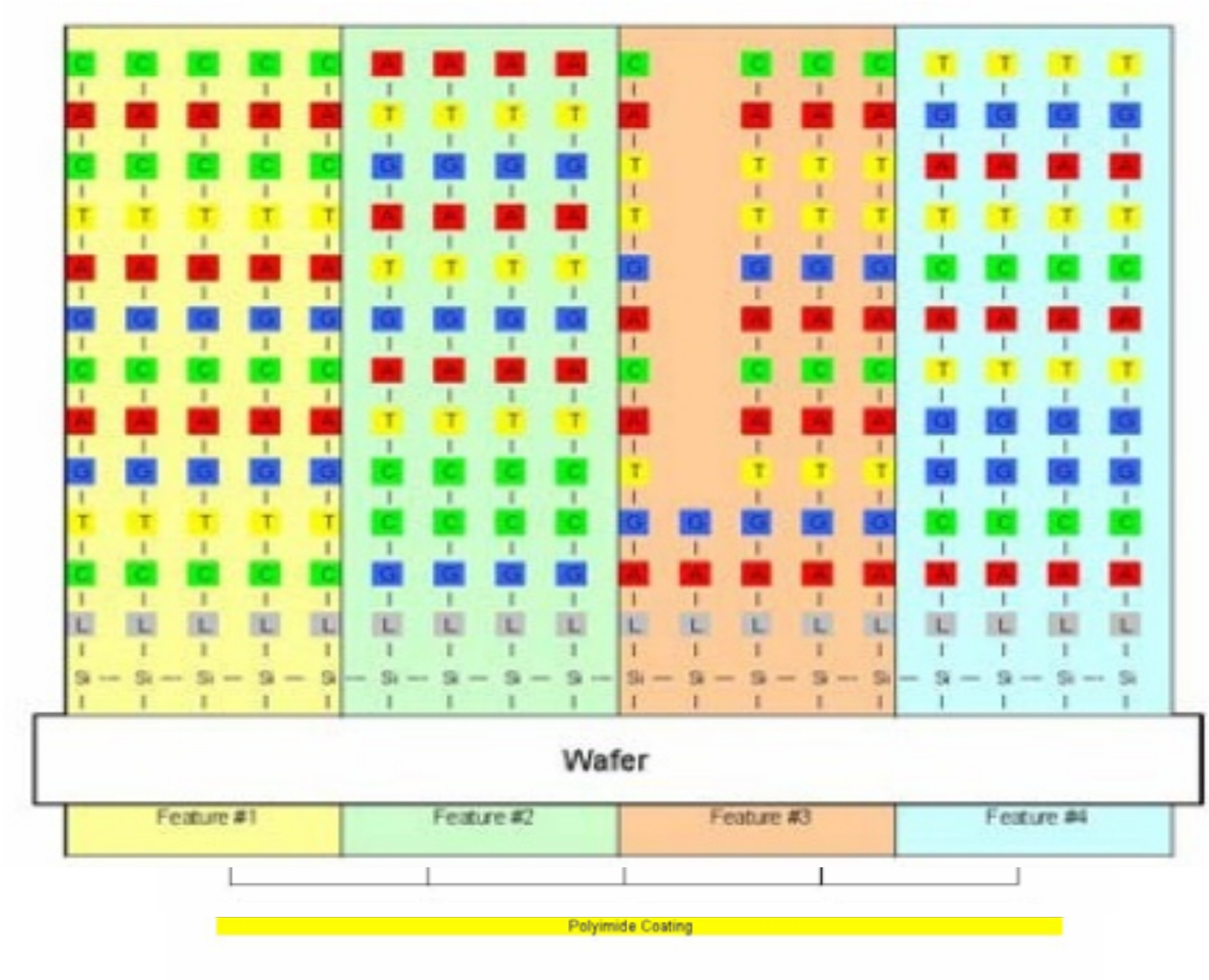
Experimental design

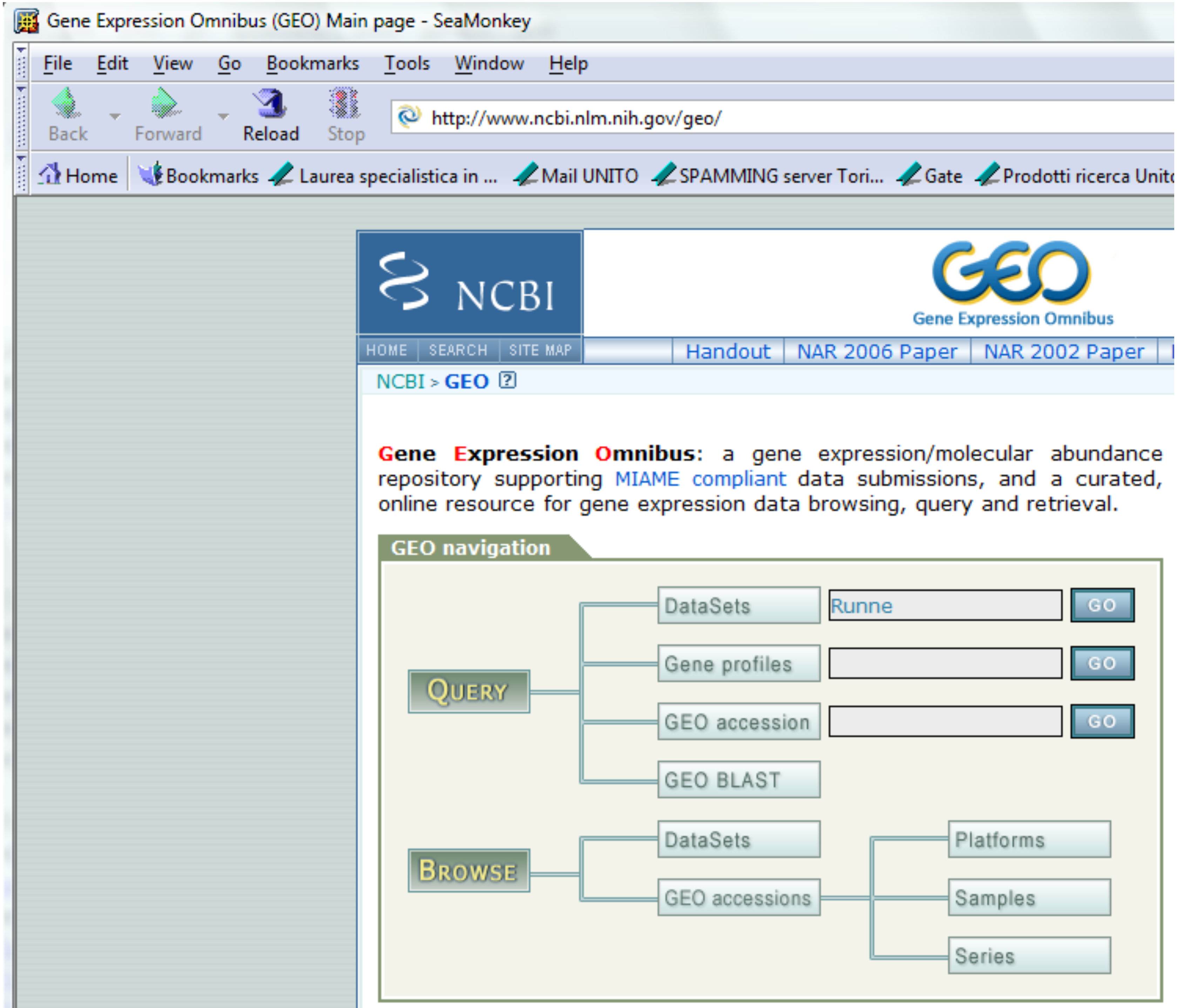


Plots of log ratios M=log2(KO/WT) averaged across replicate slides, against overall intensity A=log2√(KO × WT), similarly averaged.

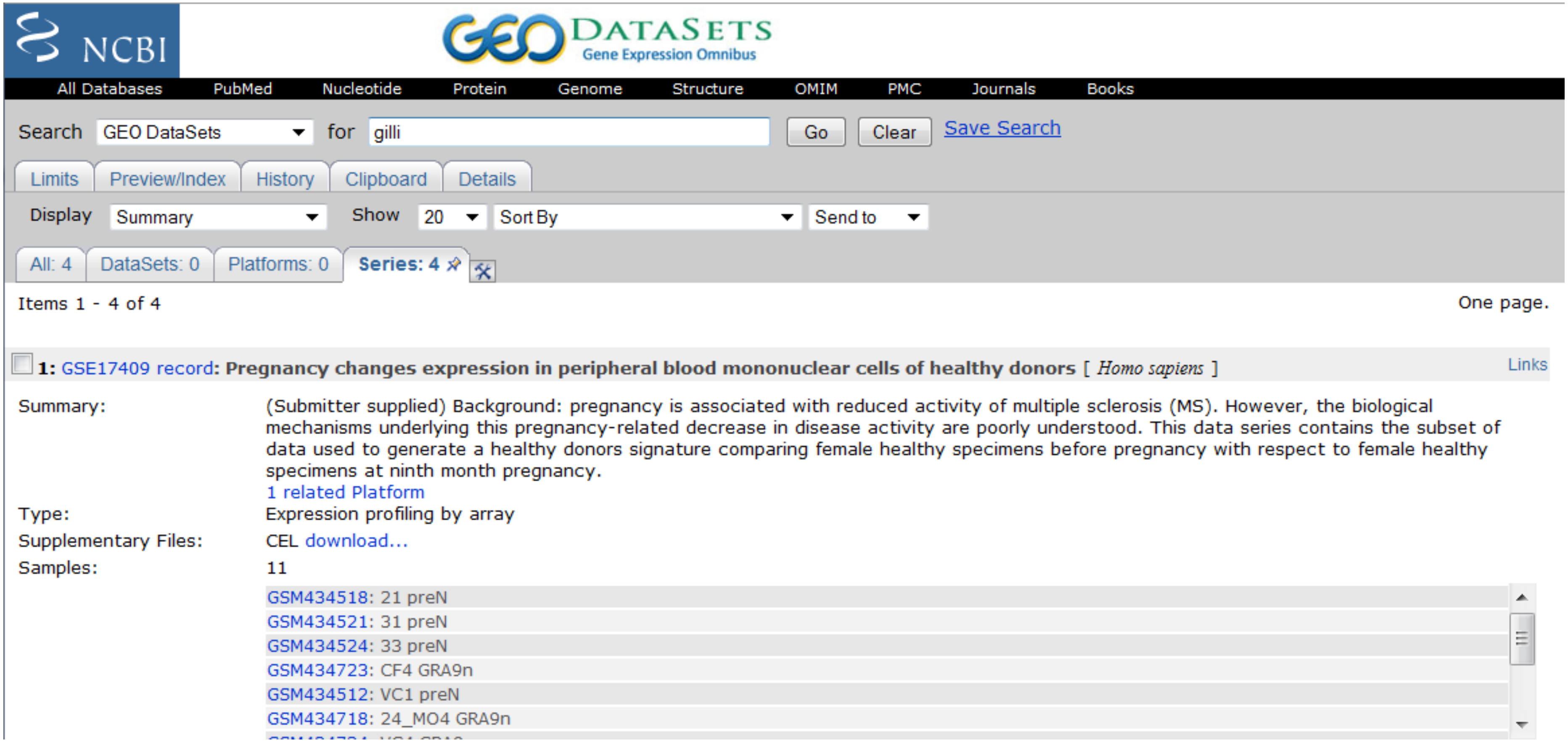
An experimenter will want to use biological replicates to obtain averages of independent data and to validate generalizations of conclusions, and perhaps technical replicates to assist in reducing the variability.

GENECHIPS ARRAYS: PHOTOLITOGRAFIC SYNTHESIS





Gene Expression Omnibus: a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.



Search GEO DataSets for gilli Go Clear Save Search

Display Summary Show 20 Sort By Send to

Items 1 - 4 of 4 One page.

□ **1: GSE17409 record: Pregnancy changes expression in peripheral blood mononuclear cells of healthy donors** [*Homo sapiens*] Links

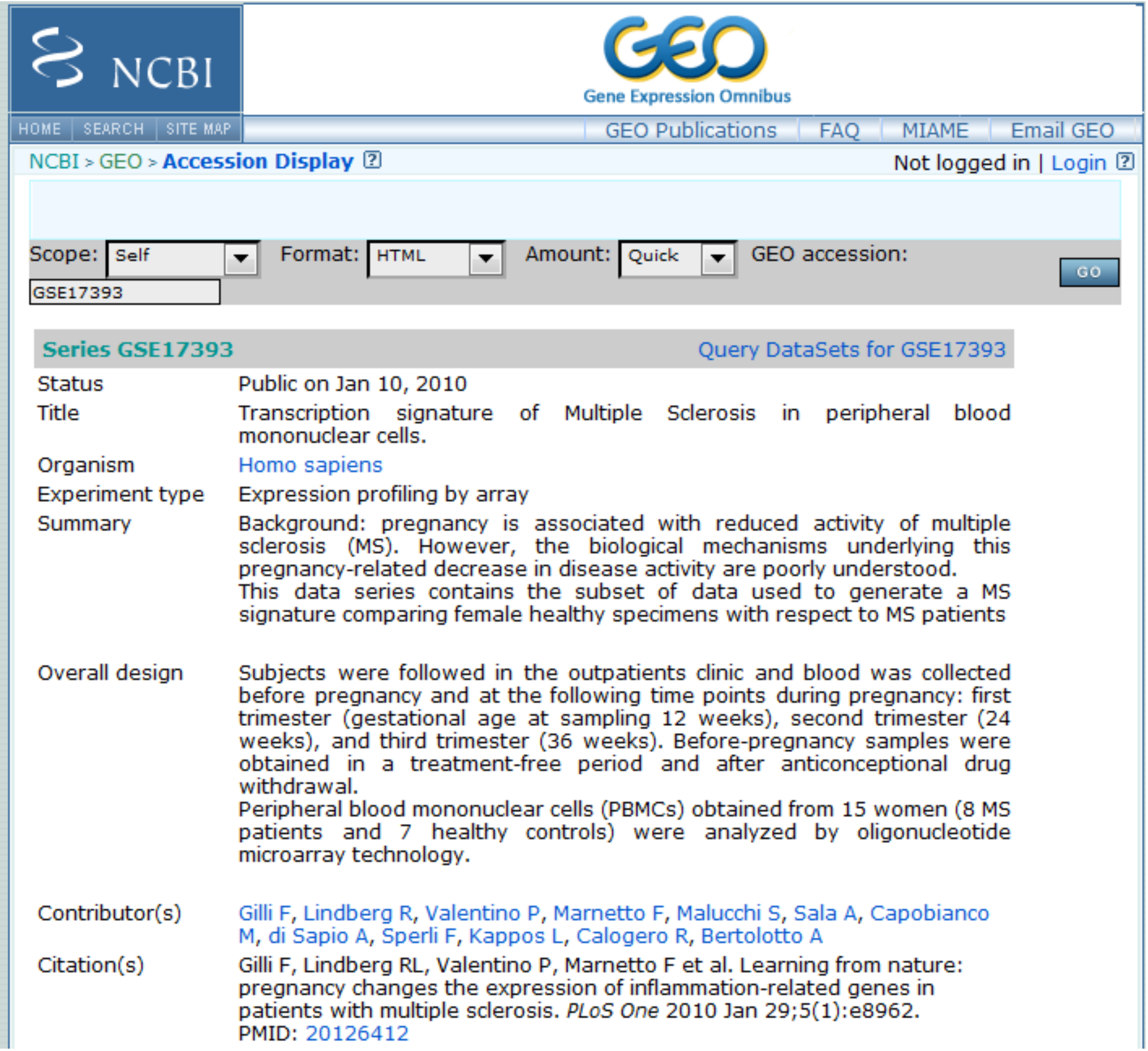
Summary: (Submitter supplied) Background: pregnancy is associated with reduced activity of multiple sclerosis (MS). However, the biological mechanisms underlying this pregnancy-related decrease in disease activity are poorly understood. This data series contains the subset of data used to generate a healthy donors signature comparing female healthy specimens before pregnancy with respect to female healthy specimens at ninth month pregnancy.
1 related Platform

Type: Expression profiling by array

Supplementary Files: CEL download...

Samples: 11

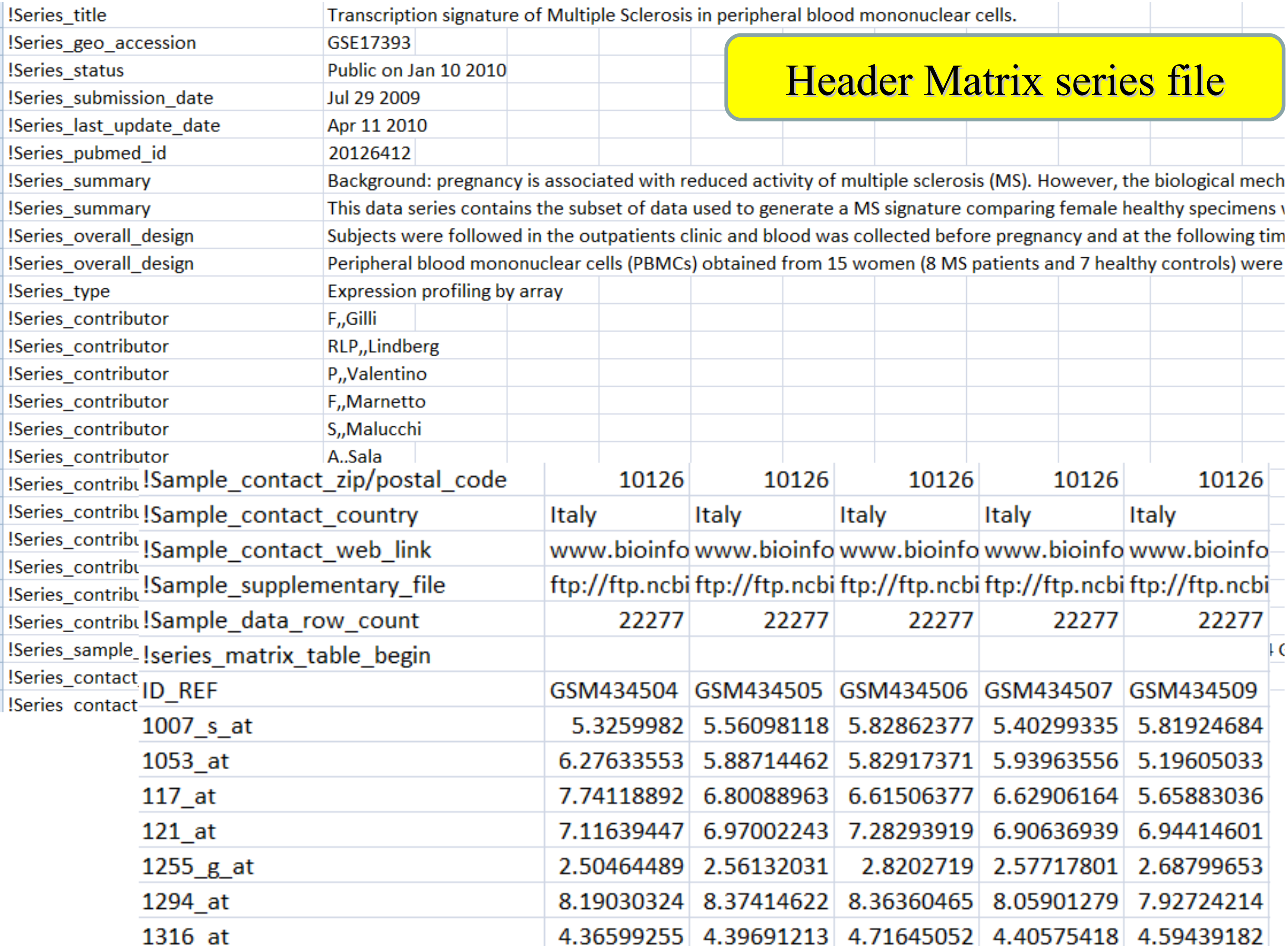
GSM434518: 21 preN
GSM434521: 31 preN
GSM434524: 33 preN
GSM434723: CF4 GRA9n
GSM434512: VC1 preN
GSM434718: 24_MO4 GRA9n



Scope: Self Format: HTML Amount: Quick GEO accession: GSE17393

Series GSE17393 Query DataSets for GSE17393

Status	Public on Jan 10, 2010
Title	Transcription signature of Multiple Sclerosis in peripheral blood mononuclear cells.
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Background: pregnancy is associated with reduced activity of multiple sclerosis (MS). However, the biological mechanisms underlying this pregnancy-related decrease in disease activity are poorly understood. This data series contains the subset of data used to generate a MS signature comparing female healthy specimens with respect to MS patients
Overall design	Subjects were followed in the outpatients clinic and blood was collected before pregnancy and at the following time points during pregnancy: first trimester (gestational age at sampling 12 weeks), second trimester (24 weeks), and third trimester (36 weeks). Before-pregnancy samples were obtained in a treatment-free period and after anticonceptional drug withdrawal. Peripheral blood mononuclear cells (PBMCs) obtained from 15 women (8 MS patients and 7 healthy controls) were analyzed by oligonucleotide microarray technology.
Contributor(s)	Gilli F, Lindberg R, Valentino P, Marnetto F, Malucchi S, Sala A, Capobianco M, di Sapio A, Sperli F, Kappos L, Calogero R, Bertolotto A
Citation(s)	Gilli F, Lindberg RL, Valentino P, Marnetto F et al. Learning from nature: pregnancy changes the expression of inflammation-related genes in patients with multiple sclerosis. *PLoS One* 2010 Jan 29;5(1):e8962. PMID: 20126412

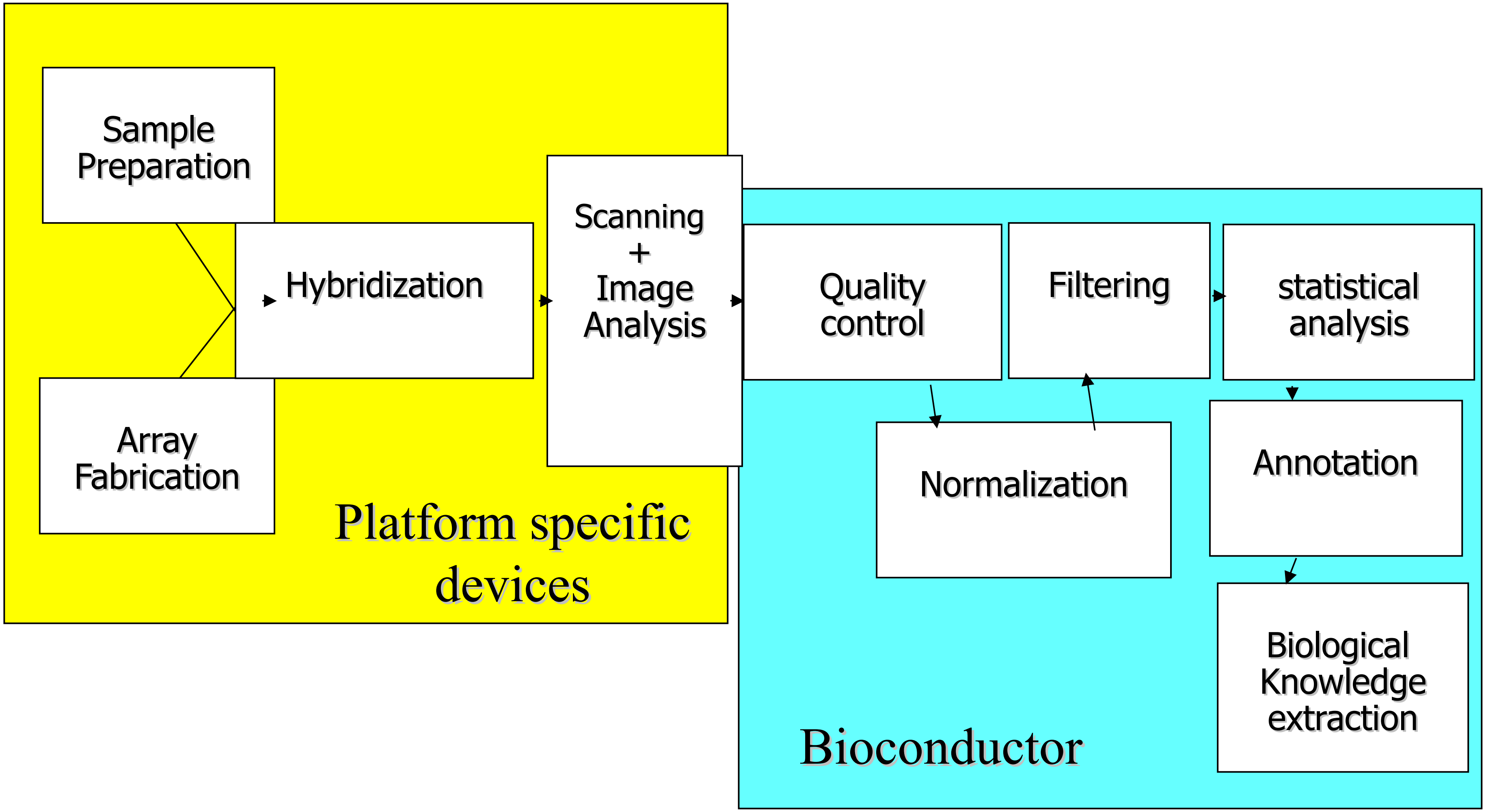


Header Matrix series file

!Series_title	Transcription signature of Multiple Sclerosis in peripheral blood mononuclear cells.
!Series_geo_accession	GSE17393
!Series_status	Public on Jan 10 2010
!Series_submission_date	Jul 29 2009
!Series_last_update_date	Apr 11 2010
!Series_pubmed_id	20126412
!Series_summary	Background: pregnancy is associated with reduced activity of multiple sclerosis (MS). However, the biological mech
!Series_summary	This data series contains the subset of data used to generate a MS signature comparing female healthy specimens
!Series_overall_design	Subjects were followed in the outpatients clinic and blood was collected before pregnancy and at the following tim
!Series_overall_design	Peripheral blood mononuclear cells (PBMCs) obtained from 15 women (8 MS patients and 7 healthy controls) were
!Series_type	Expression profiling by array
!Series_contributor	F,,Gilli
!Series_contributor	RLP,,Lindberg
!Series_contributor	P,,Valentino
!Series_contributor	F,,Marnetto
!Series_contributor	S,,Malucchi
!Series_contributor	A,,Sala
!Series_contribu	
!Series_contribu	
!Series_contribu	
!Series_contribu	
!Series_contribu	
!Series_contribu	
!Series_sample_	
!Series_contact	
!Series_contact	

!Sample_contact_zip/postal_code	10126	10126	10126	10126	10126
!Sample_contact_country	Italy	Italy	Italy	Italy	Italy
!Sample_contact_web_link	www.bioinfo	www.bioinfo	www.bioinfo	www.bioinfo	www.bioinfo
!Sample_supplementary_file	ftp://ftp.ncbi	ftp://ftp.ncbi	ftp://ftp.ncbi	ftp://ftp.ncbi	ftp://ftp.ncbi
!Sample_data_row_count	22277	22277	22277	22277	22277
!series_matrix_table_begin					
ID_REF	GSM434504	GSM434505	GSM434506	GSM434507	GSM434509
1007_s_at	5.3259982	5.56098118	5.82862377	5.40299335	5.81924684
1053_at	6.27633553	5.88714462	5.82917371	5.93963556	5.19605033
117_at	7.74118892	6.80088963	6.61506377	6.62906164	5.65883036
121_at	7.11639447	6.97002243	7.28293919	6.90636939	6.94414601
1255_g_at	2.50464489	2.56132031	2.8202719	2.57717801	2.68799653
1294_at	8.19030324	8.37414622	8.36360465	8.05901279	7.92724214
1316_at	4.36599255	4.39691213	4.71645052	4.40575418	4.59439182

Analysis pipe-line



Data Analysis

Pre-processing microarray data

diagnostic, normalization

Differential Gene Expression

identification of up and down regulated genes

Annotation and metadata

get the DE genes’ id, pathway invovlement, GO

Distances, Prediction, and Cluster Analysis

sample similarity calculation and visulization by heatmap

Class prediction

provide expression profile of type-known samples to computer, train it, and let computer to classify type-unknown samples

Data Analysis

What are the targets genes for my knock-out gene?
Gene discovery, differential expression

Is a specified group of genes (genes from a pathway) all up-regulated in a specified condition?
Gene set enrichment analysis

Can I use the expression profile of cancer patients to predict chemotherapy outcome?
Class prediction, classification

Pathways/network affected?
Kegg, Biocarta
Considering Pathway/network Topology

Data Analysis

Quality Control metrics

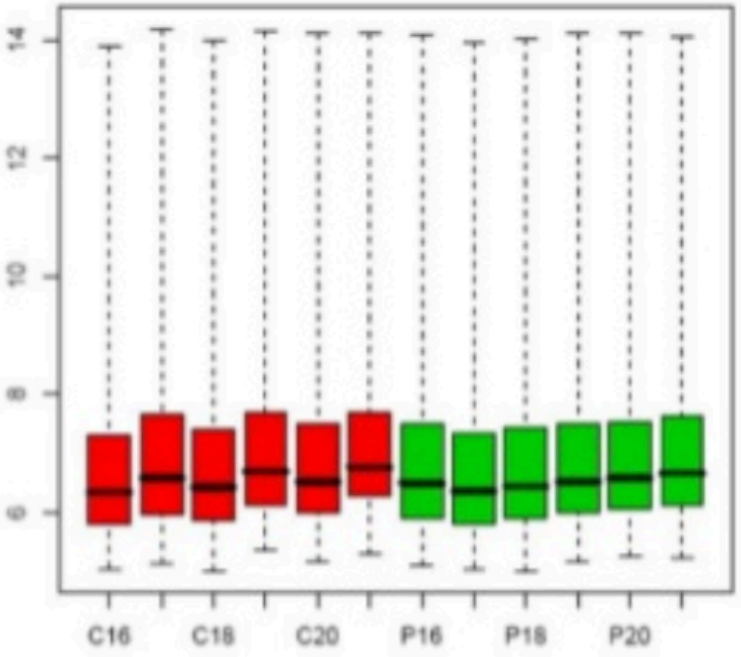
1. Average background

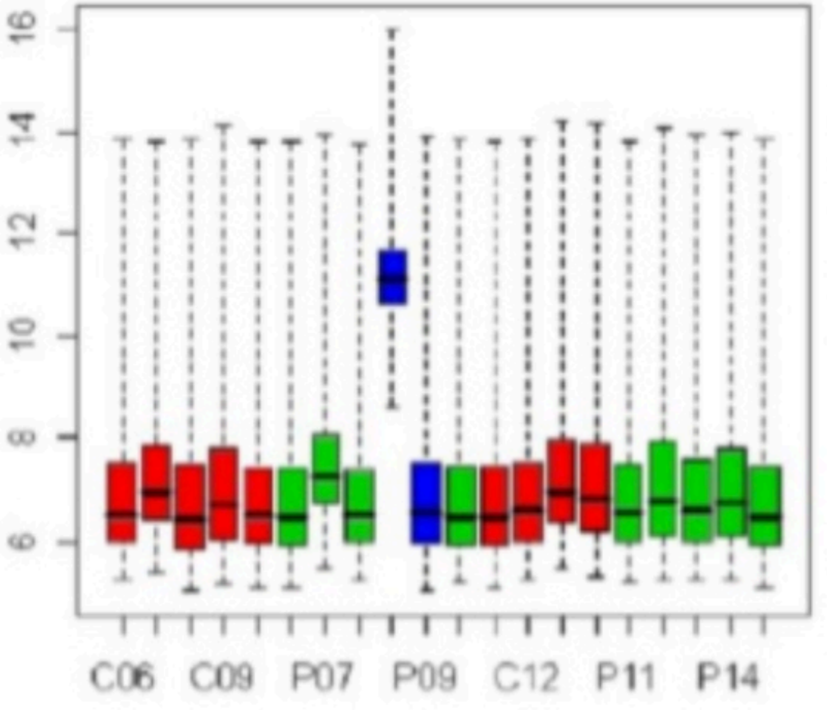
2. Scale factor

3. Number of genes called present

4. 3’ to 5’ ratios of actin and GAPDH

5. Uses ordered probes in all probeset to detect possible RNA degradation.





Normalization

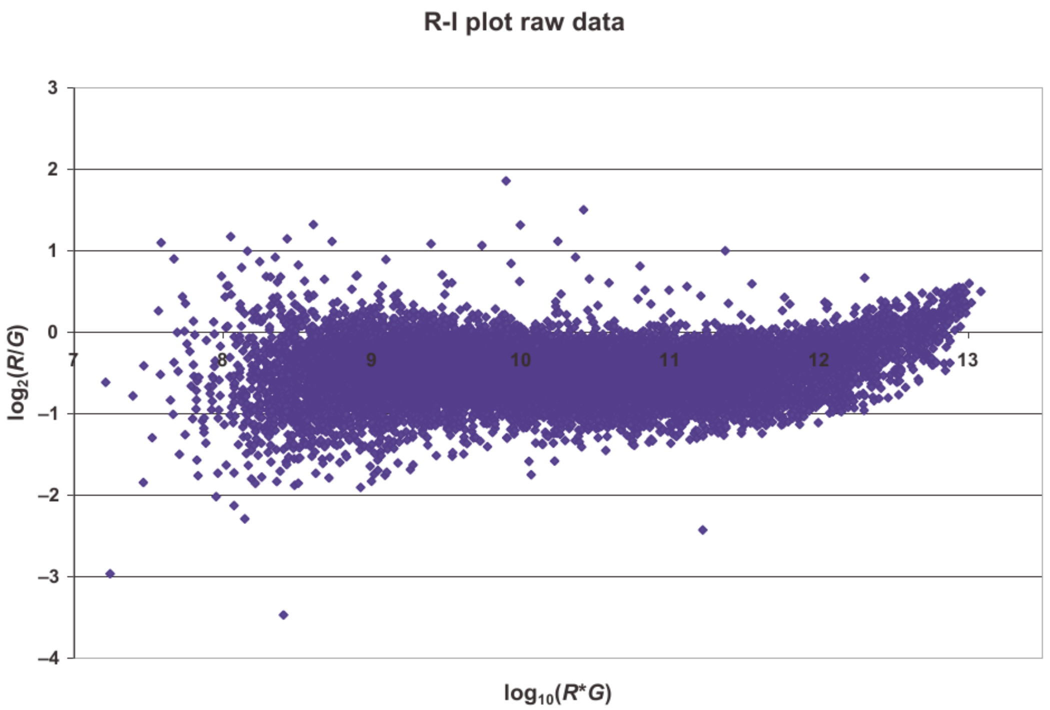
The main goal is to remove the systematic bias in the data as completely as possible, while preserving the variation in gene expression that occurs because of biologically relevant changes in transcription.

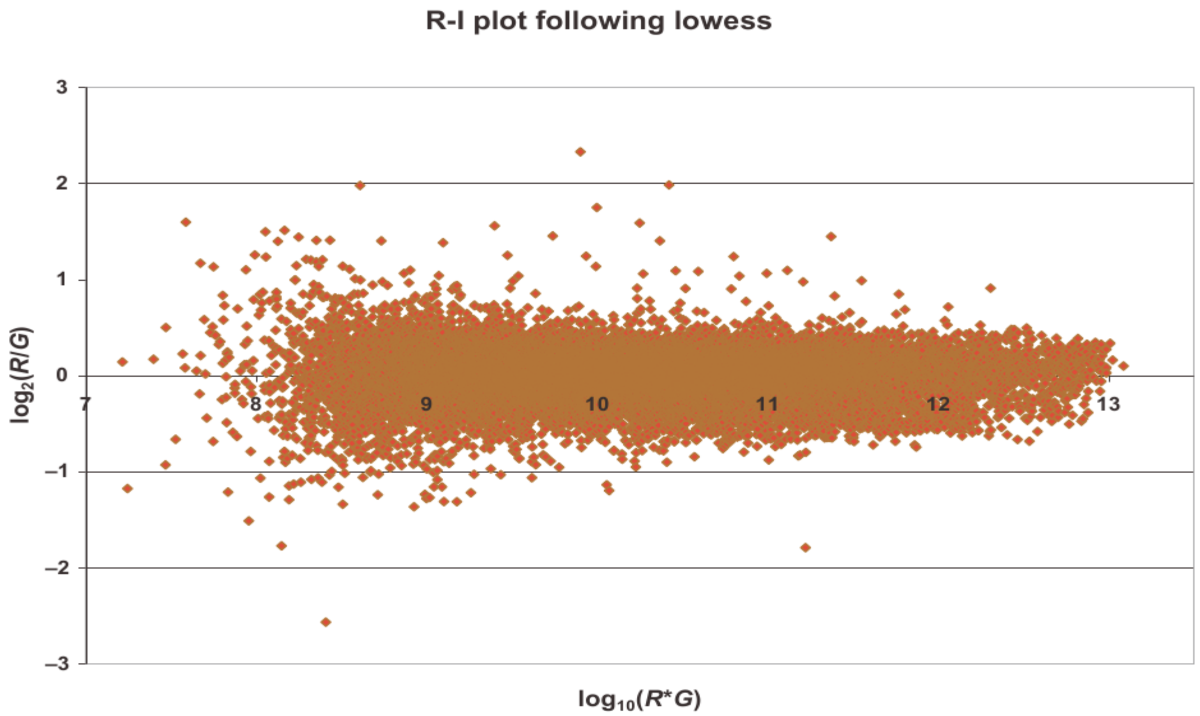
A basic assumption of most normalization procedures is that the average gene expression level **does not change** in an experiment.

Normalization is different in spotted/two-color compared with high-density-oligonucleotides (Affy) technology

$$M = \log_2(R/G) = \log_2(R) - \log_2(G)$$

$$A = \frac{1}{2}\log_2(RG) = \frac{1}{2}(\log_2(R) + \log_2(G))$$





Data Analysis

RMA methodology (Irizarry et al., 2003) performs:
– background correction,
–normalization,
–summarization in a modular way.

RMA does not take in account unspecific probe hybridization in probe set background calculation.

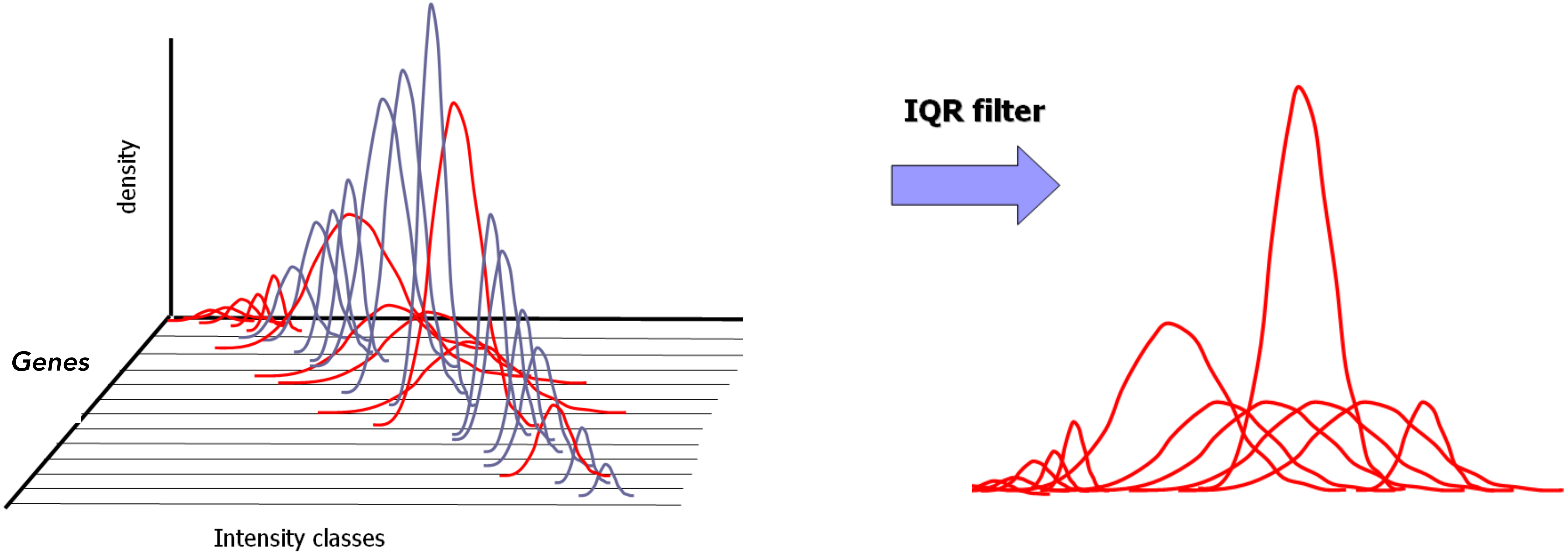
GCRMA is a version of RMA with a background correction component that makes use of probe sequence information (Wu et al., 2004).

Data Analysis

•**Filtering** affects the false discovery rate .

•Researcher is interested in keeping the number of tests/genes as low as possible while keeping the interesting genes in the selected subset.

•If the truly differentially expressed genes are overrepresented among those selected in the filtering step, the FDR associated with a certain threshold of the test statistic will be lowered due to the filtering.



Data Analysis

Statistical Analysis

1. Calculation of a statistic based on replicate array data for ranking genes according to their possibilities of differential expression

2. Selection of a cut-off value for rejecting the null- hypothesis that the gene is not differentially expressed

- The sensitivity of statistical tests is affected by the number of available replicates.
- Replicates can be:
  - Technical
  - Biological

- Biological replicates better summarize the variability of samples belonging to a common group.
- The minimum number of replicates is an important issue!

Data Analysis

Statistical Analysis

- The intensity change between experimental groups (i.e. control versus treated) are known as:

  Fold change.

- Frequently an arbitrary threshold is used to define a significant differential expression

$$\left|\log_2 \frac{\overline{Trtd}}{\overline{Ctrl}}\right| = 1$$

Intensity changes between experimental groups (i.e. control versus treated) are known as:

–Fold change.

–Ranking genes based on fold change alone implicitly assigns equal variance to every gene.

- Fold change alone is not sufficient to indicate the significance of the expression changes, has to be supported by statistical information.

- Statistical validation can be performed using parametric and non-parametric tests.
- Parametric tests:

  –*The populations under analysis are normally distributed.*
- Non parametric tests:

  –*There is no assumption on samples distribution.*
- Non parametric are less sensitive than parametric.

Data Analysis

The limma package allows the construction of linear models and a simple version is implemented in oneChannelGUI.
In case of a C group versus a T group we can build the following model:

$$y_{ij} = \mu_i + \beta_i x_j + \varepsilon_{ij}$$

1) yij is the observed expression level for gene i in sample j (j=1, ...).
2) xj = 1 if T sample and 0 otherwise.
3) µi is the expression level of gene i in C samples
4) βi represents the effects of T on the expression level of gene i
5) ϵij represents random error for gene i and sample j, and is   assumed to be independent for each gene and sample, and normally distributed with mean 0 and variance σi2.

In case of a C group versus a T group we evaluate the following hypotheses:

Ho $$\mu_i + \beta_i 0 = \mu_i + \beta_i 1$$

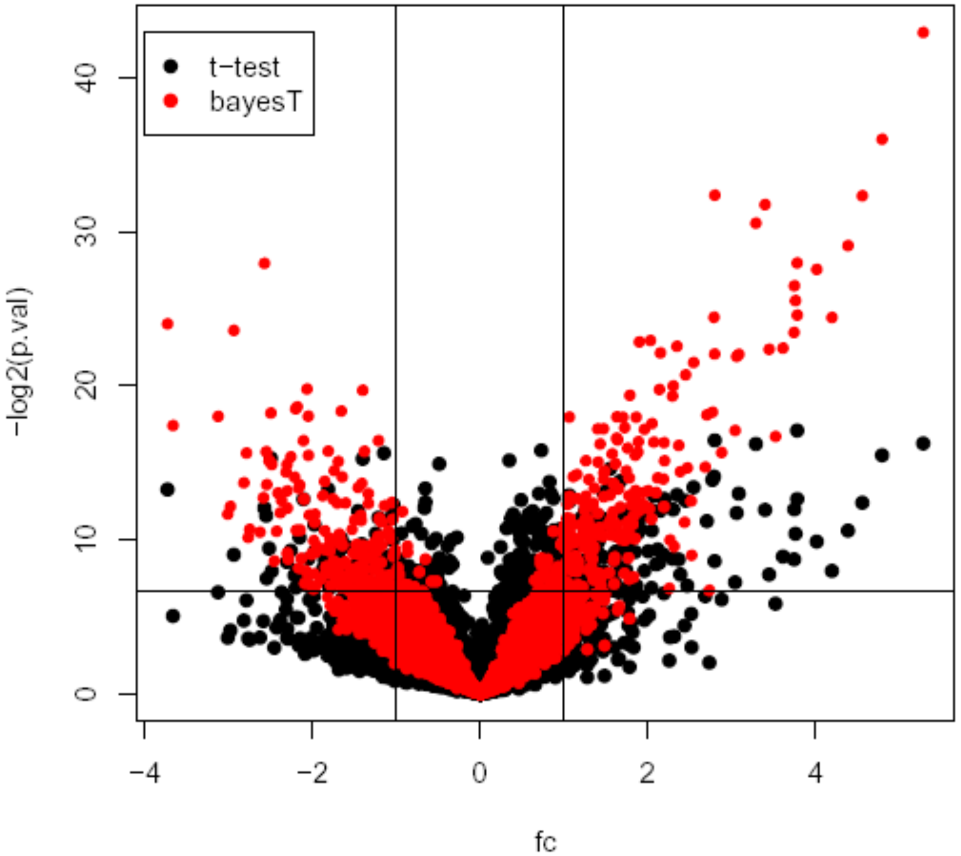
Data Analysis

Formula t-test generale:

$$\boxed{t = \frac{\overline{m_T} - \overline{m_C}}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}}}$$

Formula t-test in linear modelling:

$$t = \frac{(\mu_i + \beta_i) - \mu_i}{se_p}$$



The method tries to decouple the mean–variance dependency by modeling the variance of the expression of a gene as a function of the mean expression of the gene

$$t = \frac{(\mu_i + \beta_i) - \mu_i}{\overset{\approx}{se}_p} \qquad \textbf{where} \qquad \overset{\approx}{s}^2 = \frac{d_0 s_0^2 + d s_p^2}{d_0 + d}$$

d_0: background standard deviation, taking into account a set of genes those expression levels are similar to the gene of interest.

s_0^2: confident factor, it defines the importance of standard deviation w.r.t. the sperimental standard deviation

Data Analysis

Bioconductor aims:

Provide access to powerful statistical and graphical methods for the analysis of genomic data.

o Facilitate the integration of biological metadata (GenBank, GO, LocusLink, PubMed) in the analysis of experimental data.

o Allow the rapid development of extensible, interoperable, and scalable software.

o Promote high-quality documentation and reproducible research.

o Provide training in computational and statistical methods.