



UNIVERSITY
OF
JOHANNESBURG

Standardisation of Gene Expression Analysis

Jonathan Featherston

University of Johannesburg &
NHLS

Microarray Technology (MA)

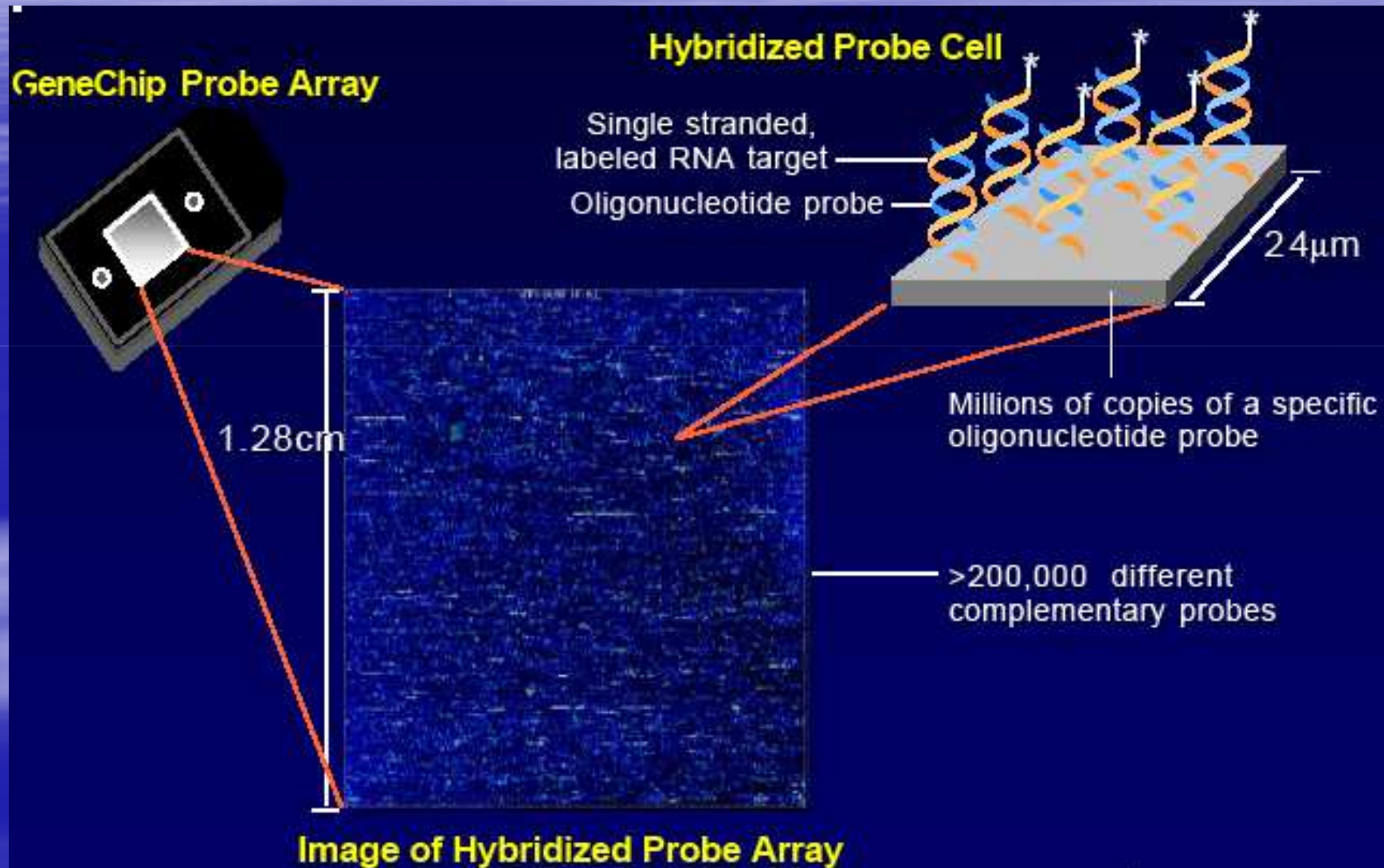
- Microarray or GeneChip technology has revolutionised the way biologists ask questions about the genome.
- Microarrays allow for the simultaneous analysis of thousands of genes.
- Thus, whole-genome gene expression profiles can be generated.
- Numerous applications: hypothesis generation, diagnostics, gene discovery.

Basic procedure

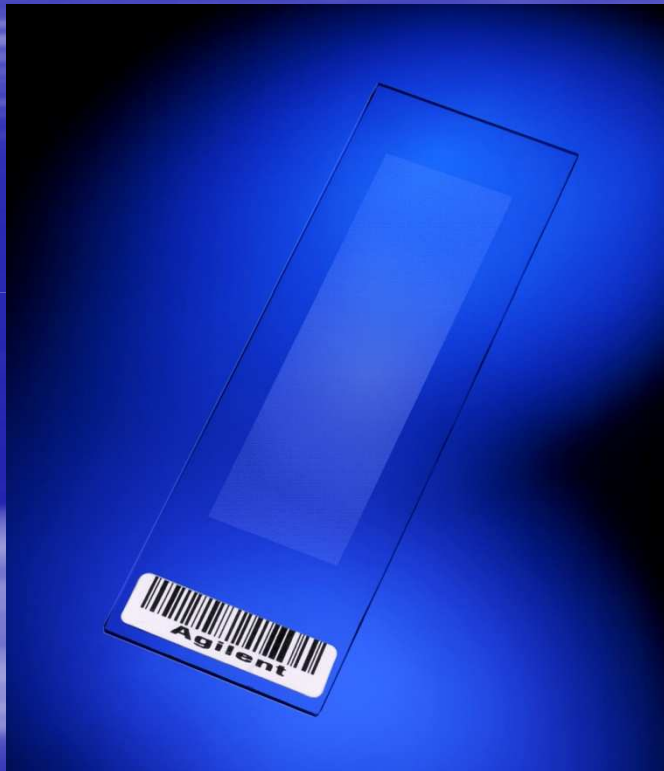


- 1) Preparation of RNA or DNA sample
- 2) QC
- 3) Amplification, hybridisation and labelling
- 4) Laser guided Scanning of Chips to measure signals
- 5) Data analysis

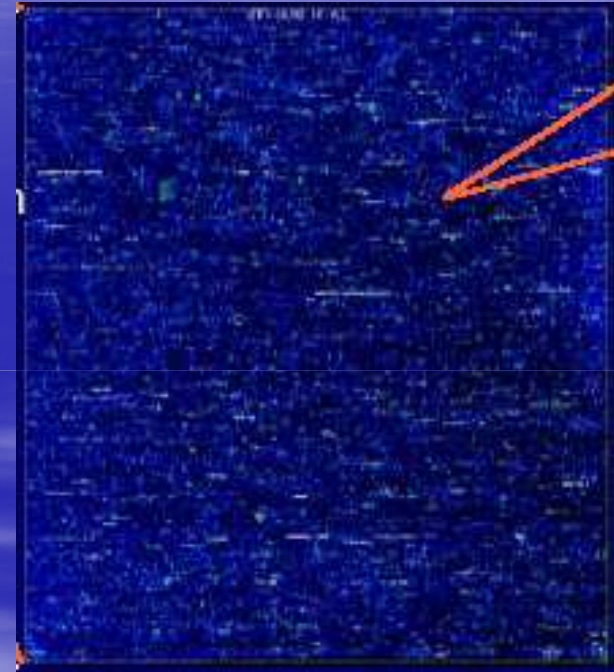
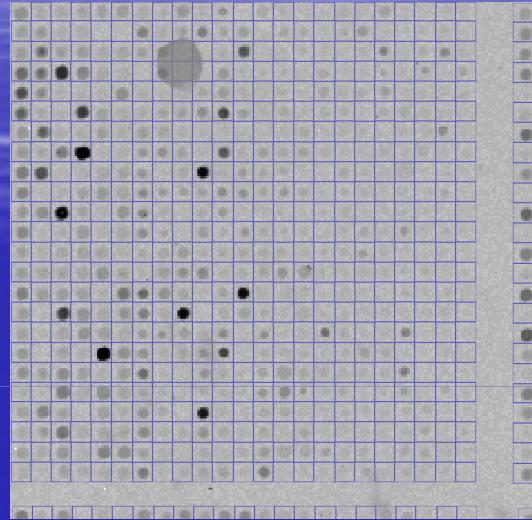
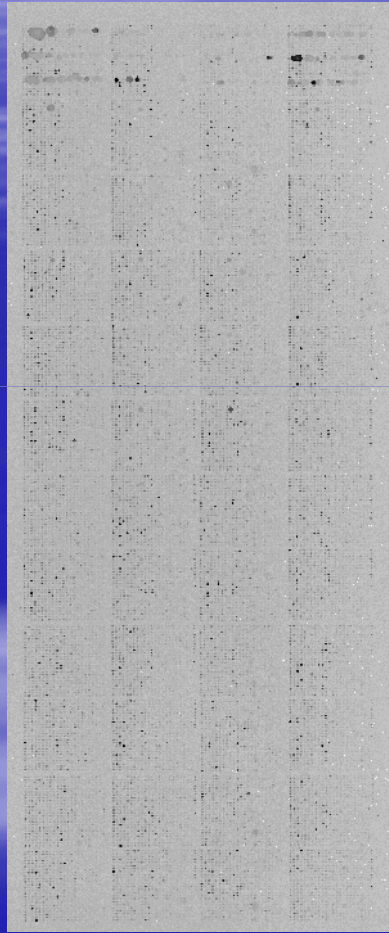
Principle of MA Technology



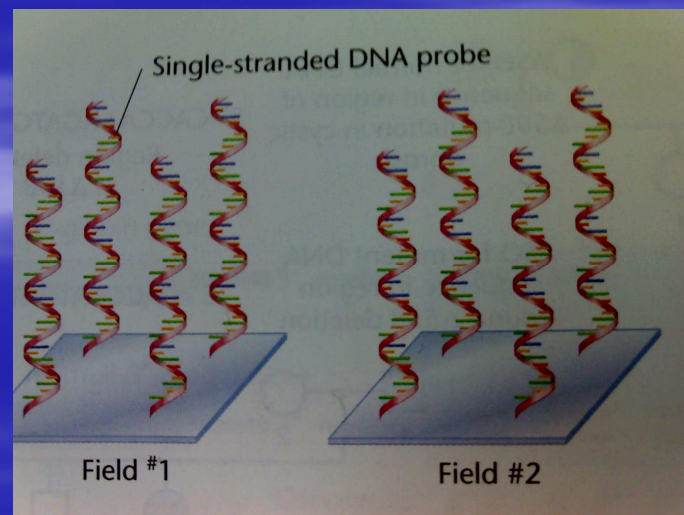
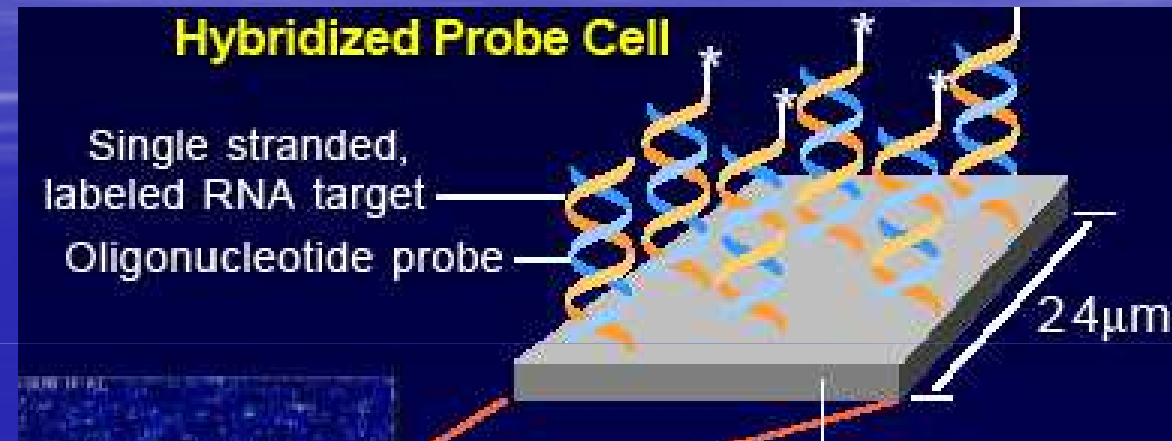
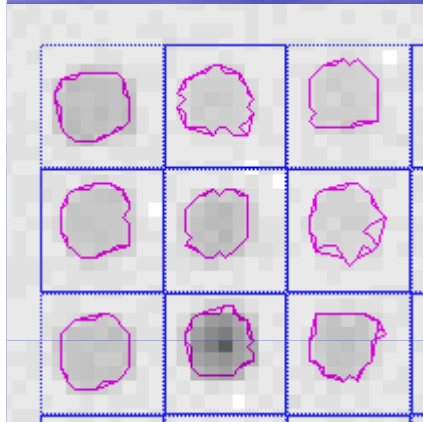
Microarrays



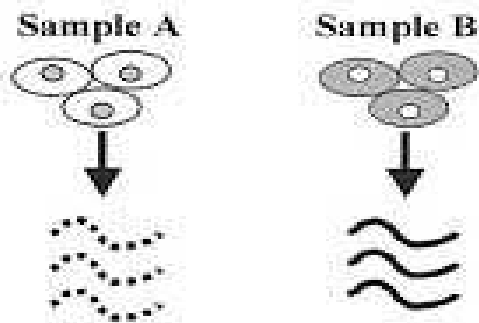
Microarray Substrate



Features/Fields/Spots/Probes



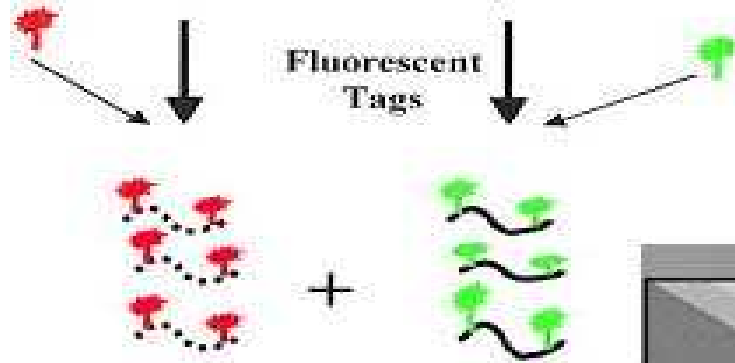
A. RNA Isolation



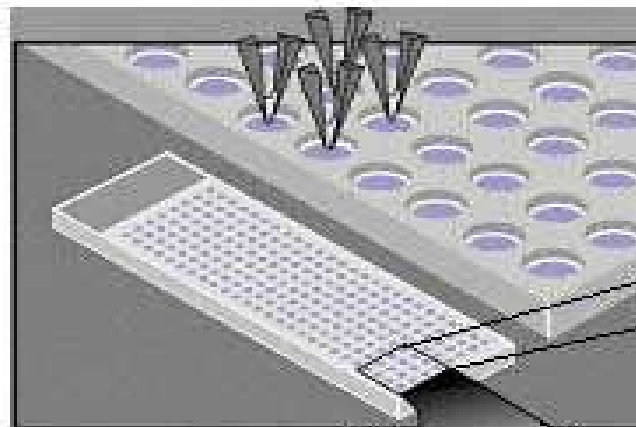
B. cDNA Generation

C. Labeling of Probe

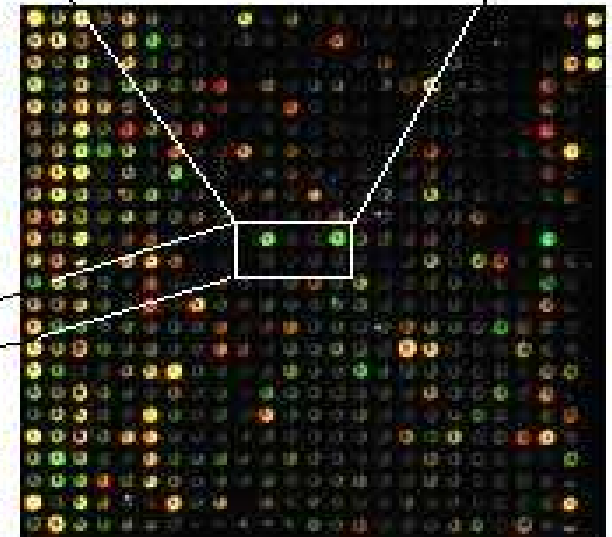
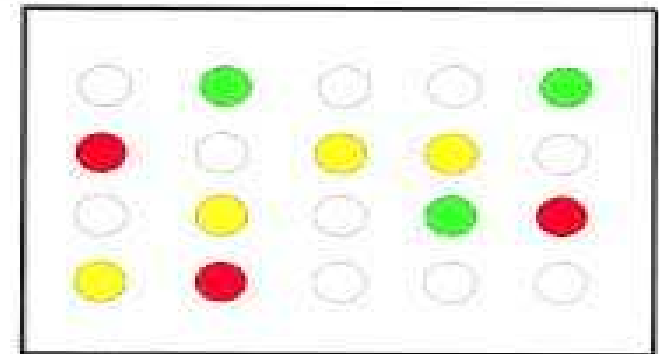
Reverse Transcriptase



D. Hybridization to Array



E. Imaging



Data Analysis

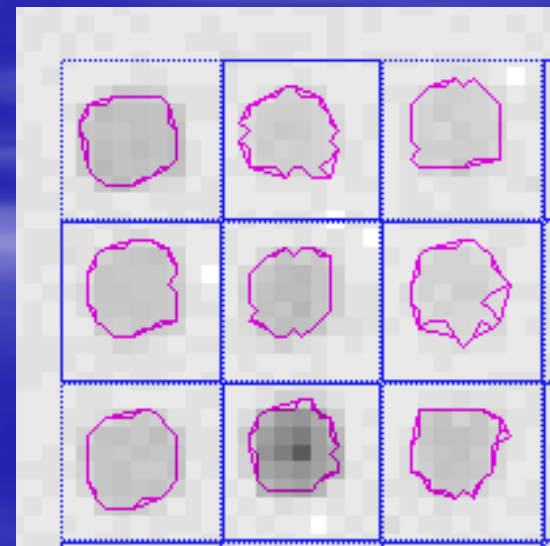
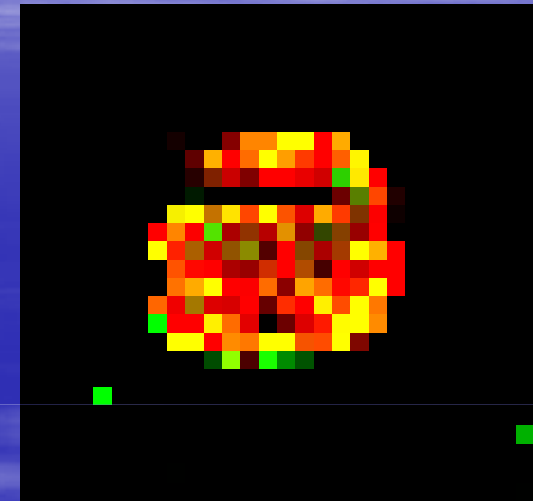
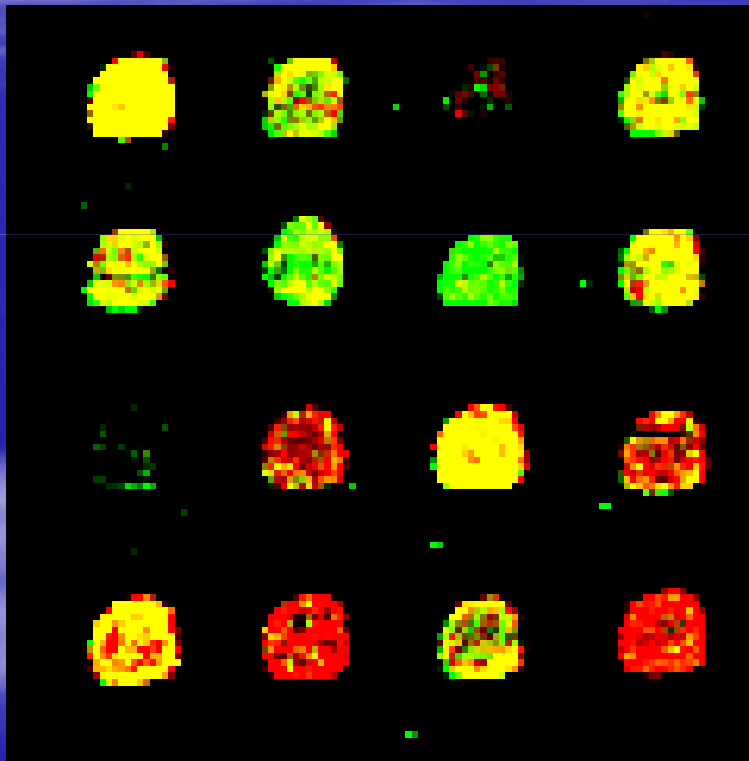
- FDA performed extensive investigation into the validity of microarrays. Results were positive and lead investigators conclude that the largest determinant of reproducibility was data analysis.
- Modern arrays have up to 2.5 million probes.
- Analysis of this dimensionality of data is complex and fraught with pitfalls.
- Earlier studies employed erroneous statistics.
- Recent methods have greatly increased stringency.

Outline of analysis steps

- Scanning
 - Intensity generation
 - Pre-processing
- Gene selection/Expression analysis
 - Biological interpretation
 - Possibly exploratory statistics

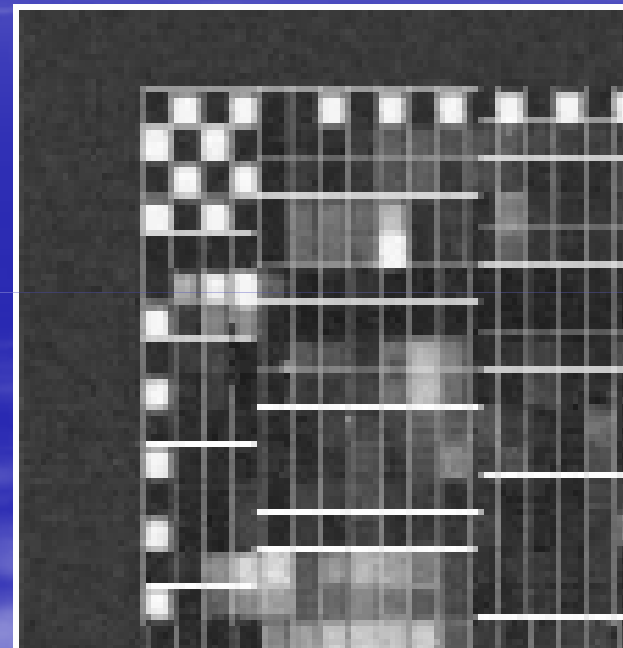
And of course QA throughout analysis!

Spot calling



QA of Microarray Data

	Stat Pairs	Stat Pairs Used	Signal	Detection	Detection p-value
AFFX-BioB-5_at	20	20	338.0	P	0.000972
AFFX-BioB-M_at	20	20	667.6	P	0.000060
AFFX-BioB-3_at	20	20	389.9	P	0.000060
AFFX-BioC-5_at	20	20	893.5	P	0.000110
AFFX-BioC-3_at	20	20	858.9	P	0.000044
AFFX-BioDn-5_at	20	20	1340.8	P	0.000052
AFFX-BioDn-3_at	20	20	4810.4	P	0.000195
AFFX-CreX-5_at	20	20	11307.3	P	0.000052
AFFX-CreX-3_at	20	20	11381.5	P	0.000044
AFFX-DapX-5_at	20	20	21.6	A	0.108979
AFFX-DapX-M_at	20	20	48.7	A	0.131361
AFFX-DapX-3_at	20	20	7.2	A	0.737173
AFFX-LysX-5_at	20	20	14.2	A	0.368438
AFFX-LysX-M_at	20	20	35.8	A	0.544587
AFFX-LysX-3_at	20	20	27.6	A	0.185131
AFFX-PheX-5_at	20	20	4.6	A	0.772364
AFFX-PheX-M_at	20	20	1.9	A	0.910522
AFFX-PheX-3_at	20	20	45.6	A	0.485110
AFFX-ThiX-5_at	20	20	10.9	A	0.529760
AFFX-ThiX-M_at	20	20	37.4	A	0.411380
AFFX-ThiX-3_at	20	20	5.8	A	0.904333
AFFX-TrpX-5_at	20	20	12.1	A	0.440646
AFFX-TrpX-M_at	20	20	2.6	A	0.804734
AFFX-TrpX-3_at	20	20	4.0	A	0.588620
AFFX-r2-Ec-bioB-5_at	11	11	478.1	P	0.000244
AFFX-r2-Ec-bioB-M_at	11	11	707.3	P	0.000244
AFFX-r2-Ec-bioB-3_at	11	11	622.3	P	0.000244
AFFX-r2-Ec-bioC-5_at	11	11	1069.3	P	0.000244
AFFX-r2-Ec-bioC-3_at	11	11	1786.8	P	0.000244
AFFX-r2-Ec-bioD-5_at	11	11	4591.0	P	0.000244
AFFX-r2-Ec-bioD-3_at	11	11	6257.3	P	0.000244
AFFX-r2-Ec-bioE-5_at	11	11	12191.4	P	0.000244



GeneChip HG-U133 Plus 2

Normalisation

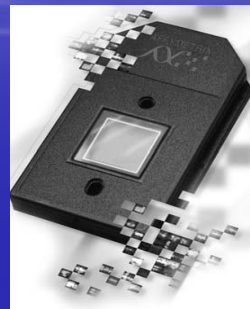
- Adjusting data to compensate for experimental differences (also known as systematic measurement errors) and technical differences so that we can make accurate assumptions.
- We are not interested in these differences (i.e. Not biological).
- Compensating for dye-bias is an example of normalisation.
- The general assumption of microarray data is that most genes do not differ.

Sources of error

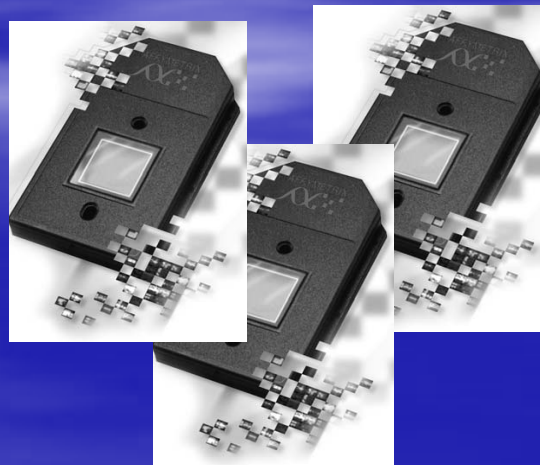
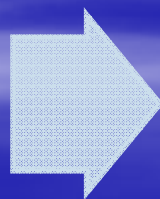
- Position on the slide or chip
- Dye bias
- Labelling methods
- Technician errors or variability
- Total RNA content
- Environmental conditions
- Probe specific affinity

Within and between normalisation

Within

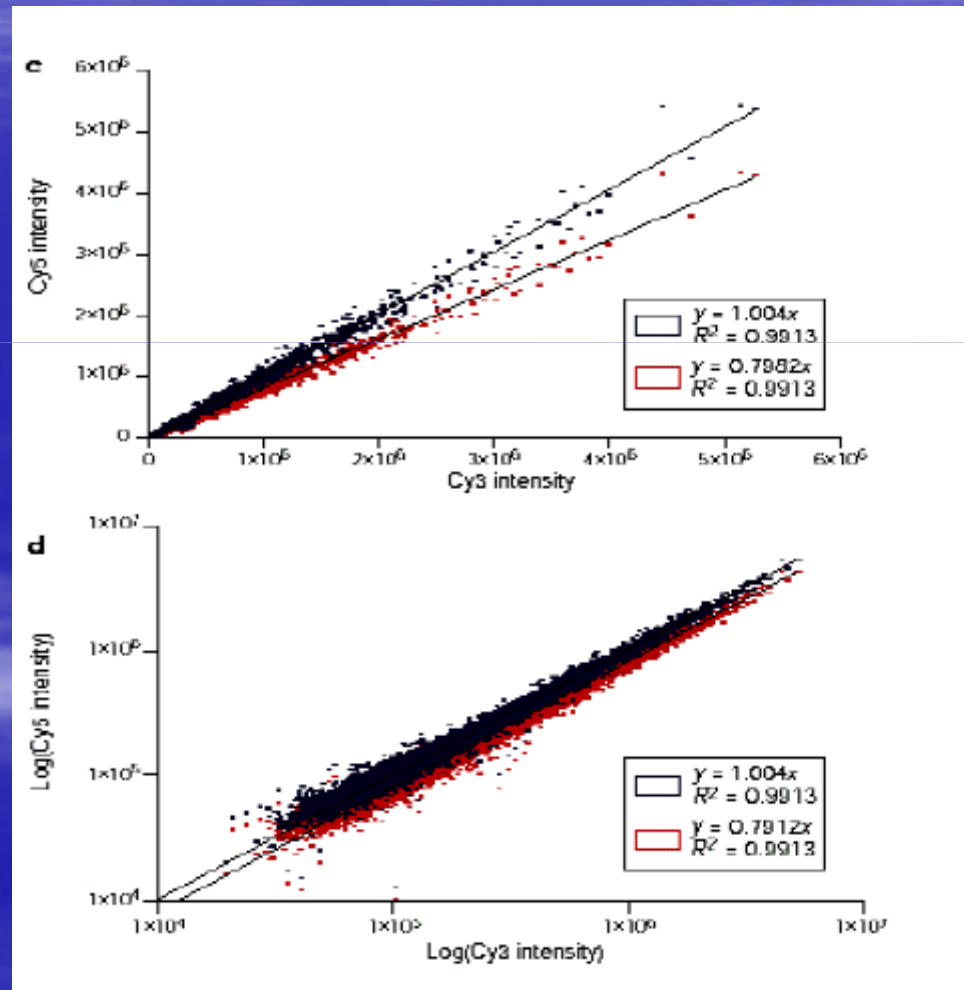


Between



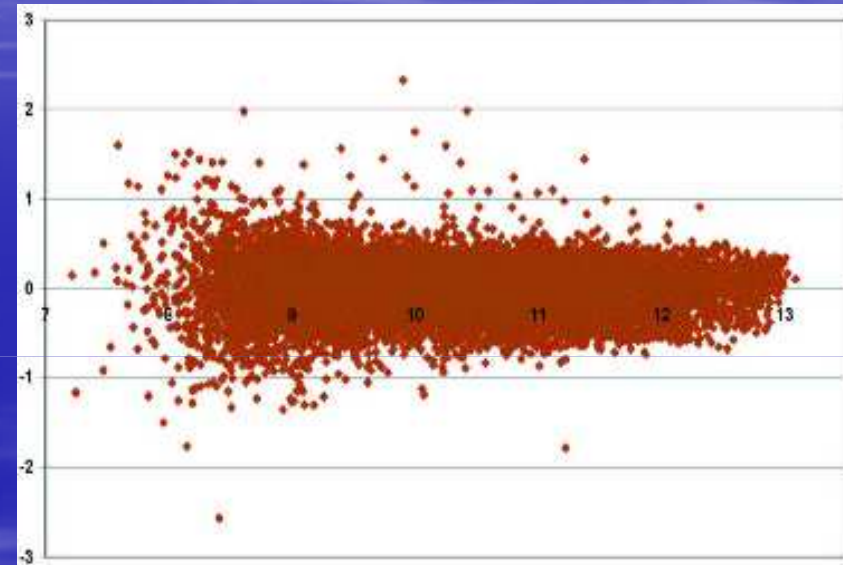
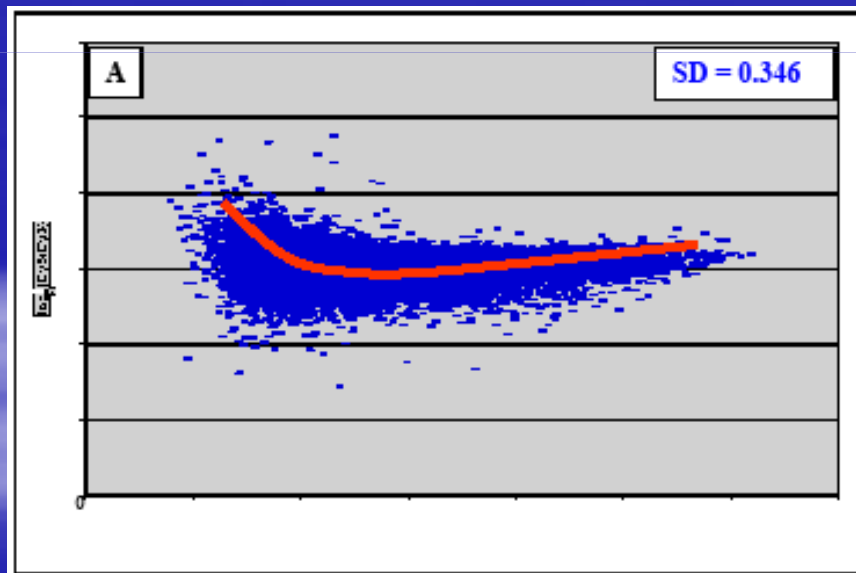
Log Transformation

Linear Data

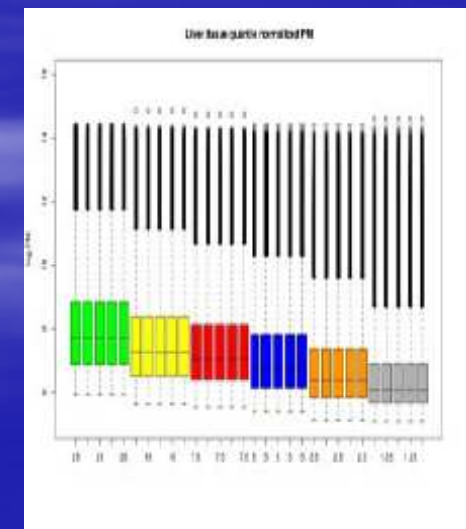
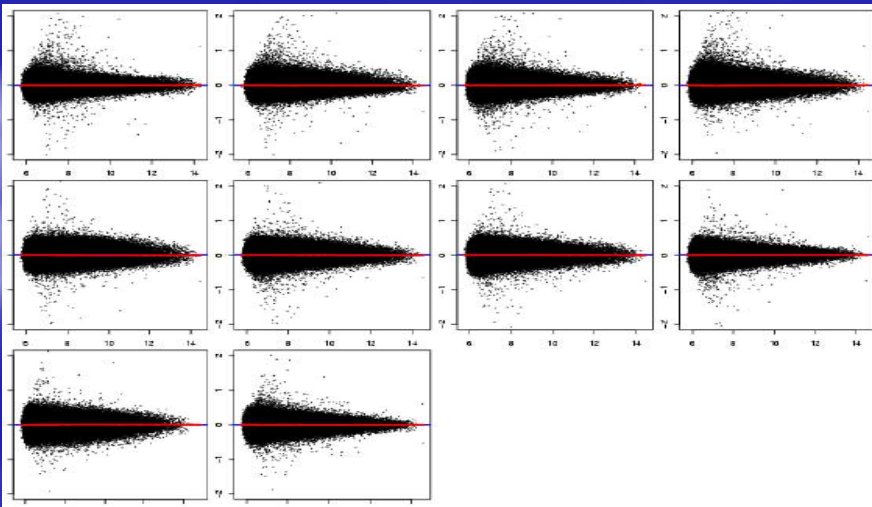
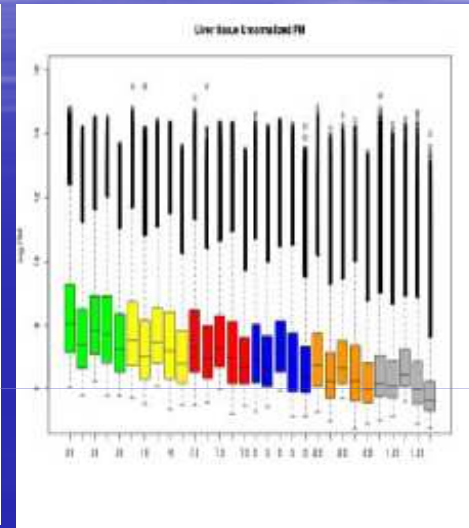
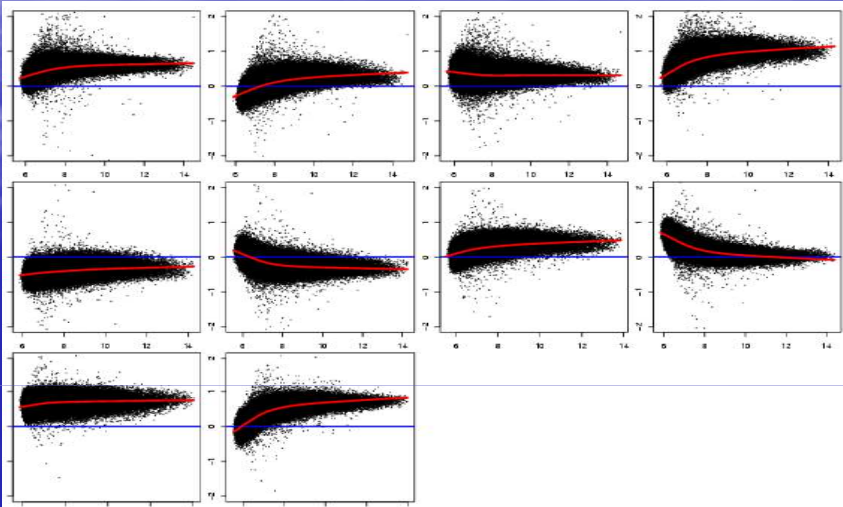


Log Data

Intra-chip Normalisation

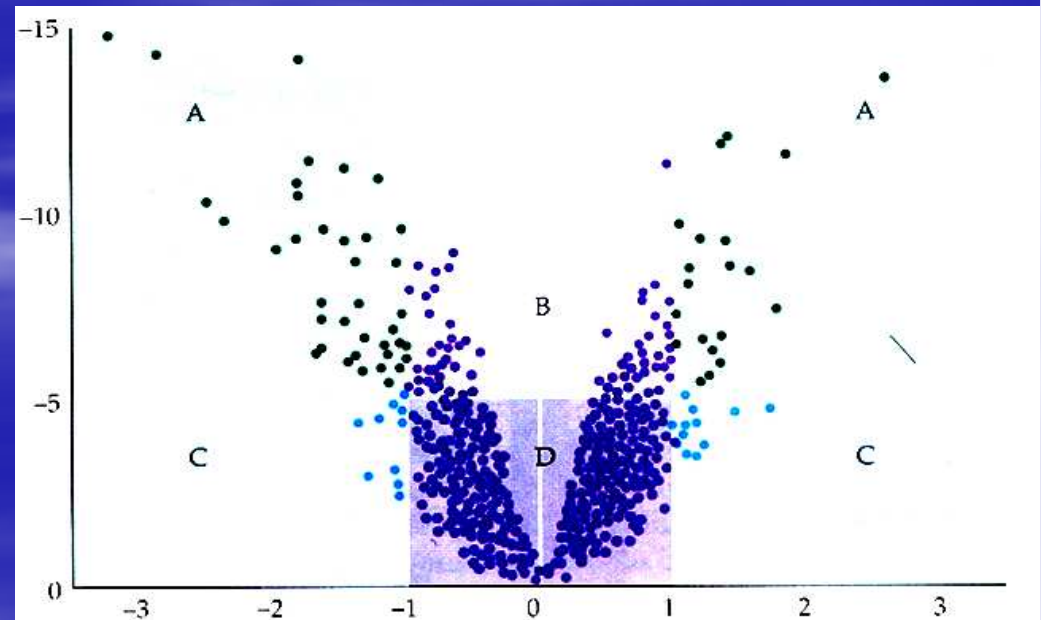


Inter-Chip Normalisation



Gene selection

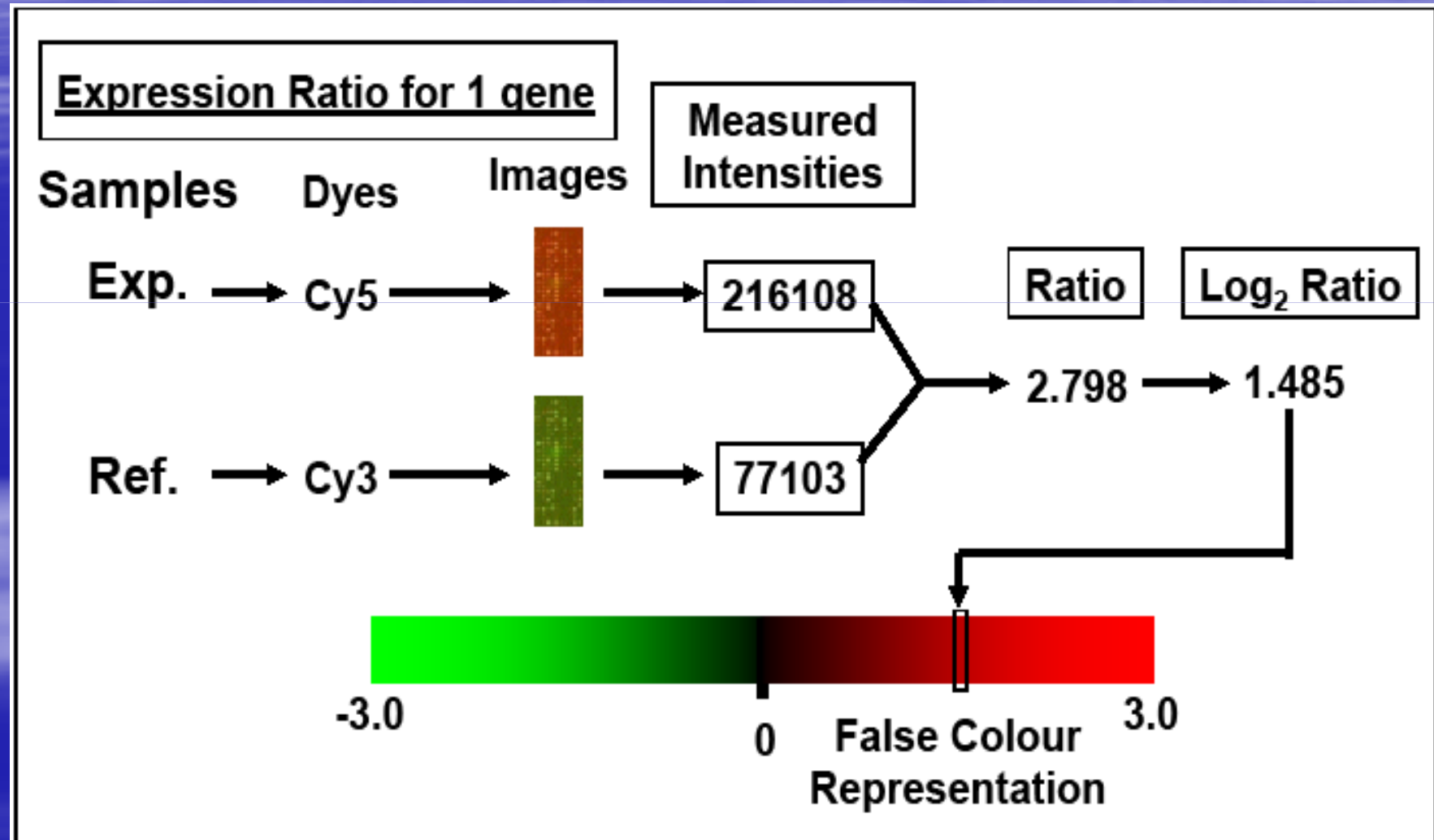
- Differentially expressed genes are generally identified by variance as well as absolute changes in expression i.e. fold changes.
- One colour vs. two.



Multiple-Testing Correction

- Due to the large no. of simultaneous question asked, one needs to adjust p-value cut-offs.
- If $p \leq 0.05$ then 10 000 genes will generate 500 results simply by chance.
- Thus multiple-testing correction is essential.
- Method of choice= False Discovery Rate (Benjamini-Hochberg).

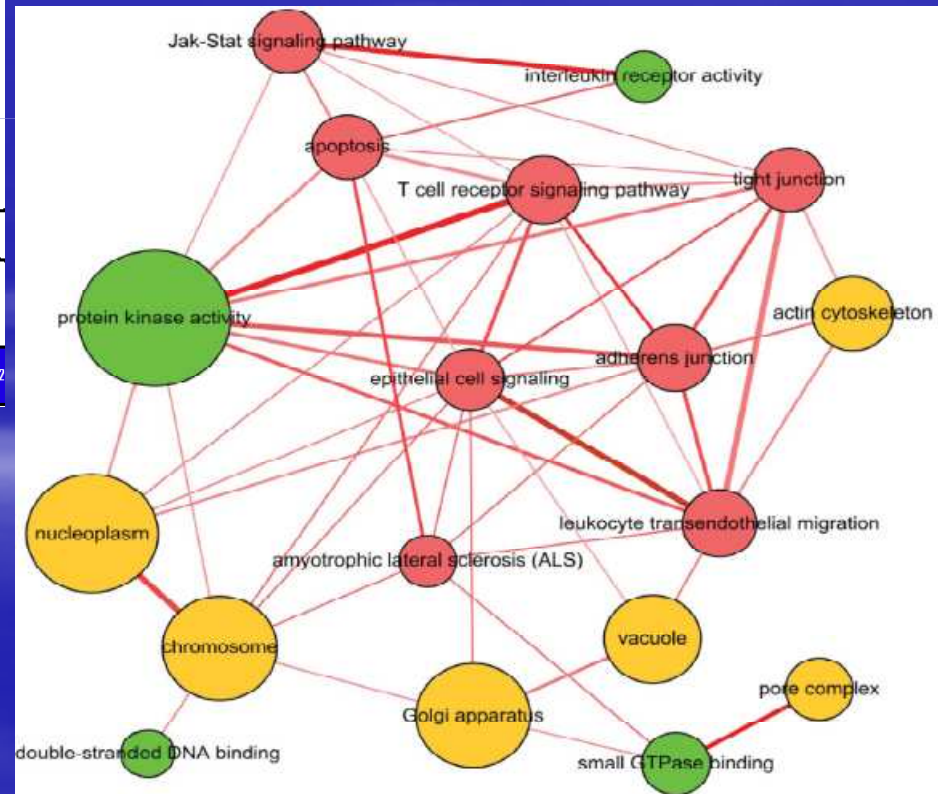
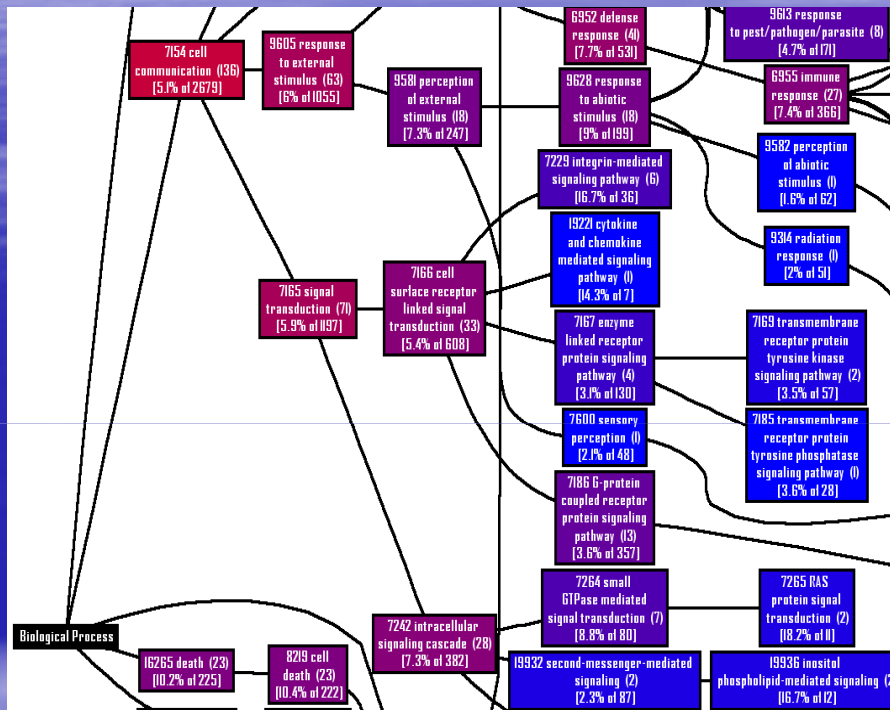
Two Colour Gene Expression Analysis



Functional Enrichment

- Highly necessary for validation of results and must be included.
- Involves comparing the list of differentially expressed genes to gene lists from the genome under investigation and testing for over/under-representation of specific gene ontology (GO) terms.
- This enables researchers to ID GO terms and biological functions that are important in their study.
- Also serves to validate that results are meaningful.
- The same applies for pathway analysis.

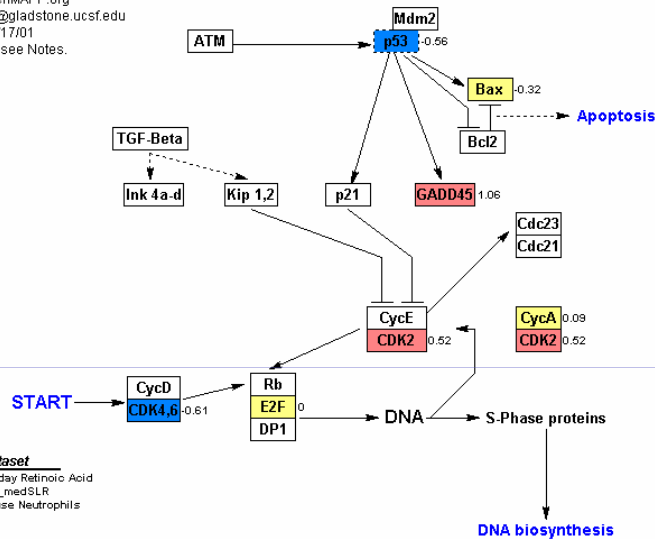
Gene Ontology Results



Pathway Analysis

Author: Adapted from KEGG
 Maintained by: GenMAPP.org
 E-mail: genmapp@gladstone.ucsf.edu
 Last modified: 12/17/01
 Right-click here to see Notes.

Cell Cycle Control in G1/S Phase

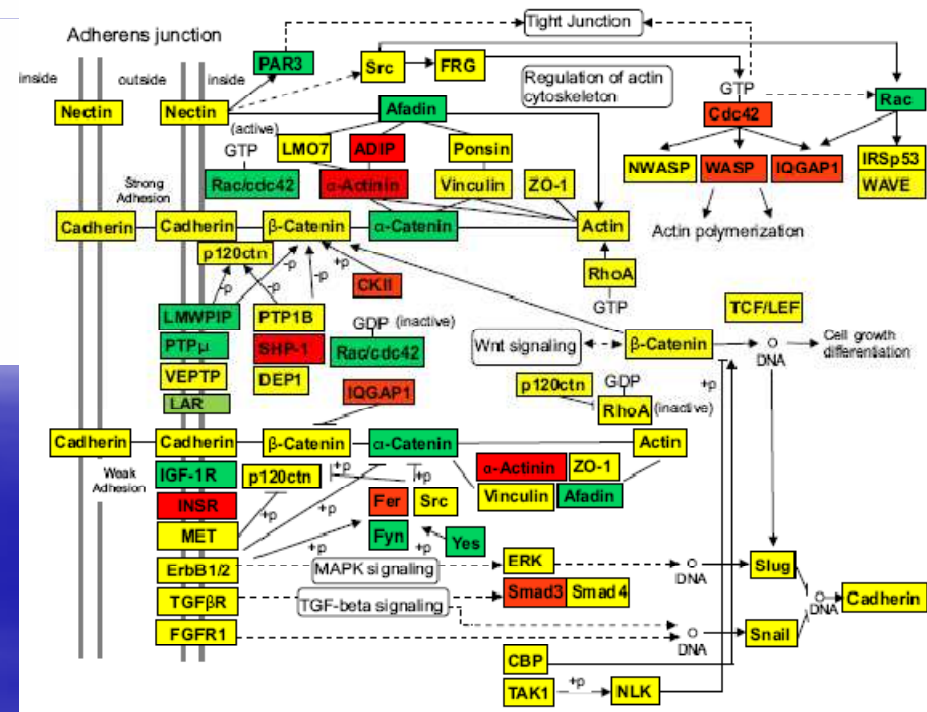


Expression Dataset

Gene Color Set: 4 day Retinoic Acid
 Gene Value: 4dv0_medSLR
 Differentiating Mouse Neutrophils

Legend

- Major Increase
- Moderate Increase
- No Change
- Moderate Decrease
- Major Decrease
- No criteria met
- Not found



Recommendations

- All data quality checks according to manufacturer specifications must be observed.
- Normalisation for intra/inter- sample variability must be applied.
- Multiple testing correction is essential.
- Differentially expressed genes must be assessed for enrichment of biological properties.

Data Analysis Recommendations

QA

Scanning- Manufacturer Specifications

Normalisation Inter/Intra Variability

Statistical Testing $q\text{-value} \leq 0.05$

Fold Change ≥ 2

Functional Enrichment- FDR correction included

Pathway Analysis- FDR correction included