

Λειτουργική γονιδιωματική

ΜΠΣ2

Γρ. Αμούτζιας

Λειτουργική γονιδιωματική: Τι είναι

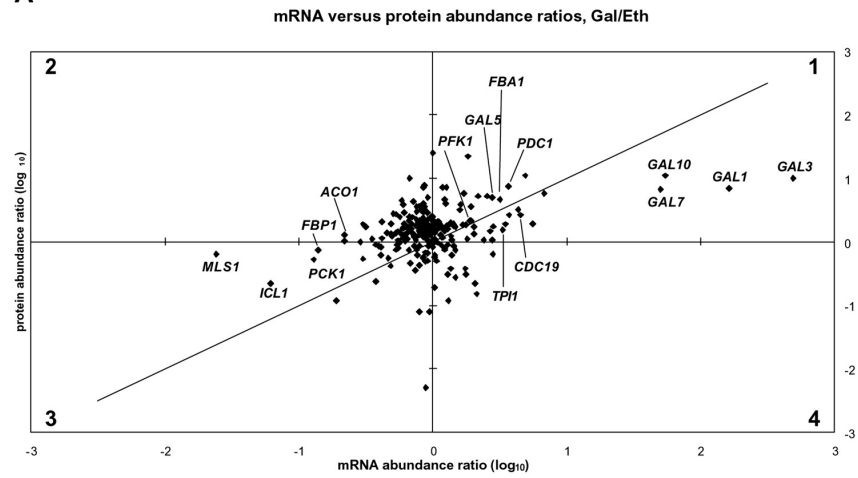
- Προσπαθεί να κατανοήσει τις λειτουργίες των βιολογικών μορίων, σε επίπεδο ολόκληρου του γονιδιώματος.
- Γίνονται μετρήσεις για το σύνολο των γονιδίων, σε μια συγκεκριμένη στιγμή ή κατάσταση.
- Αρχικά, οι μετρήσεις γίνονταν για ένα βιομόριο. Σήμερα μελετάμε την συμπεριφορά ολόκληρου του συστήματος.
- Η μελέτη της μεταγραφής του συνόλου των γονιδίων ονομάζεται μεταγραφωματική ή transcriptomics.

Transcriptomics

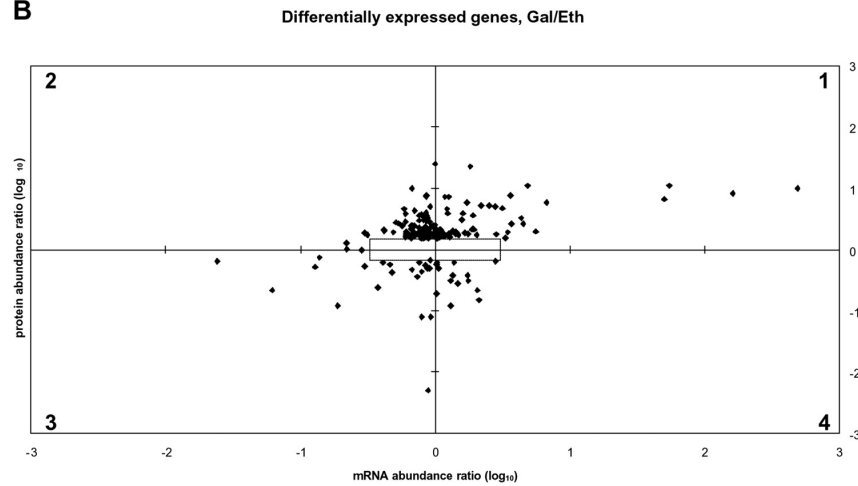
- Expressed sequence tags (ESTs)
- Serial analysis of gene expression (SAGE)
- Μικροσυστοιχίες (microarrays)
- RNA-seq (whole transcriptome shotgun sequencing)

mRNA abundance ratios versus protein-abundance ratios.

A



B



Griffin T J et al. Mol Cell Proteomics 2002;1:323-333



Expressed sequence tags (ESTs)

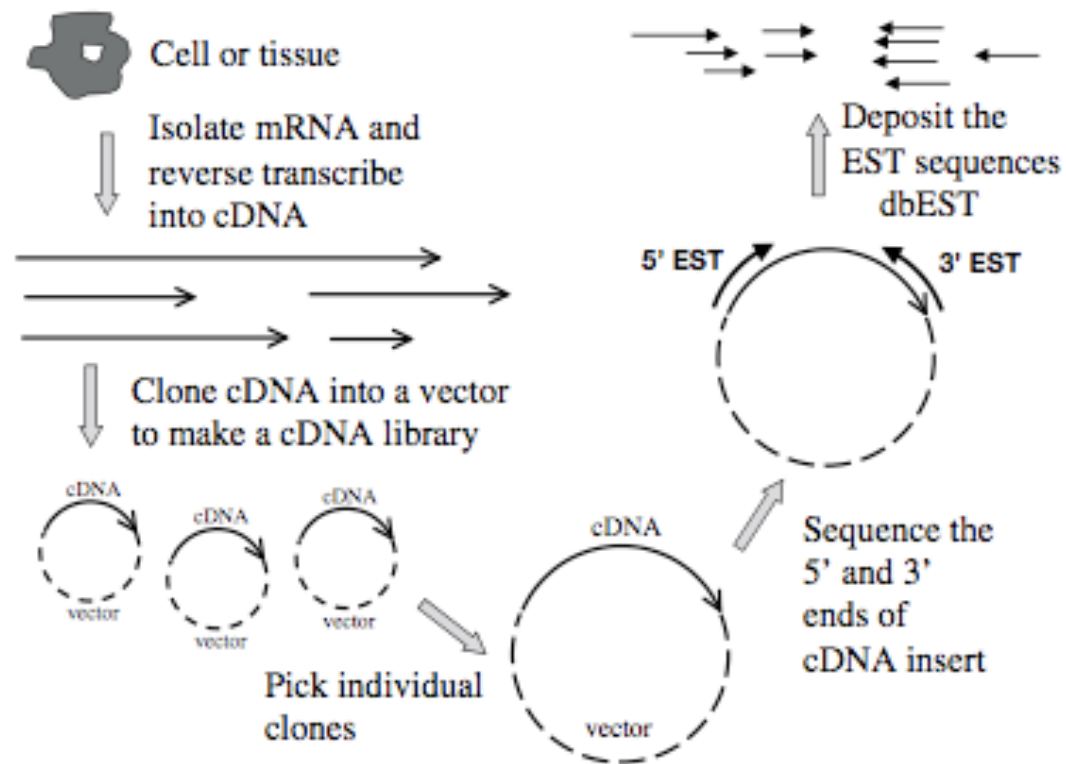


Figure 12.1. Overview of how ESTs are constructed.

Expressed sequence tags (ESTs)

- Το cDNA κλωνοποιείται σε φορείς (vectors) και δημιουργείται μια βιβλιοθήκη.
- Από την βιβλιοθήκη, επιλέγονται τυχαία κλώνοι για αλληλούχιση.
- Η αλληλούχιση ξεκινάει είτε από το 5' ή το 3' άκρο του cDNA.
- Τα παραγώμενα ESTs έχουν μήκος 400-600 νουκλεοτίδια.

Expressed sequence tags (ESTs)

- Η συχνότητα των διαφόρων ESTs που αντιστοιχούν σε ένα γονίδιο αντικατοπτρίζει σε κάποιο βαθμό το επίπεδο της μεταγραφής του γονιδίου.
- Χαμηλά εκφρασμένα γονίδια συνήθως δεν ανιχνεύονται με την μέθοδο των ESTs.
- Κυρίως ανιχνεύονται υψηλά εκφρασμένα γονίδια.
- Οι ακολουθίες των ESTs είναι χαμηλής ποιότητας και μπορεί να εμπεριέχουν λάθη (λάθος νουκλεοτίδια, μετακίνηση αναγνωστικού πλαισίου, λάθος κωδικόνια τερματισμού, χιμαιρικοί κλώνοι).
- Αν και τα επιμέρους ESTs μπορεί να εμπεριέχουν λάθη, μια συλλογή από αλληλο-επικαλυπτόμενα ESTs επιτρέπει την διόρθωση λαθών και ίσως και την ανακατασκευή της ακολουθίας ολόκληρου του cDNA.
- Παρ' όλα τα προβλήματα, τα ESTs χρησιμοποιούνται συχνά, λόγω της ευκολίας που δημιουργούνται οι βιβλιοθήκες σε διάφορες συνθήκες.
- Επιτρέπουν την ανίχνευση νέων γονιδίων.

Expressed sequence tags (ESTs)

- Κατάθεση των δεδομένων στην ΒΔ:
 - NCBI: dbEST (1992)
 - www.ncbi.nlm.nih.gov/dbEST/

Expressed sequence tags (ESTs)

- Οι ακολουθίες:
 - φιλτράρονται.
 - Ομαδοποιούνται και τυχόν λάθη διορθώνονται -> EST contigs.
 - Εντοπίζονται οι περιοχές που κωδικοποιούν την πρωτεΐνη με αλγόριθμους πρόβλεψης γονιδίων.
 - Η προβλεπόμενη πρωτεΐνη επιτρέπει την αναζήτηση ομολόγων σε ΒΔ και την πρόβλεψη της λειτουργίας.
 - Αν είναι διαθέσιμο το γονιδίωμα, τότε τα ESTs βοηθούν στην ανίχνευση γονιδίων και στον καθορισμό των ορίων μεταξύ ιντρονίων-εξονίων.

Expressed sequence tags (ESTs)

- Unigene
- www.ncbi.nlm.nih.gov/UniGene/
- Κάθε ομάδα ESTs είναι αλληλοεπικαλυπτόμενα ESTs που αντιπροσωπεύουν ένα συγκεκριμένο γονίδιο

Unigene

The screenshot shows a web browser window with the address bar containing the URL: `http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene&cmd=search&term=ESR1+human`. The browser title is "uid=163517 UniGene Result". The page header includes the NCBI logo and the UniGene logo with the tagline "ORGANIZED VIEW OF THE TRANSCRIPTOME". Navigation links for "My NCBI" (Sign In, Register) and a menu of databases (All Databases, PubMed, Nucleotide, Protein, Genome, Structure, OMIM, PMC, Journals, Books) are visible.

The search interface shows the search term "ESR1 human" entered in the search box. Below the search box are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The display settings are set to "Summary", "Show 20", and "Sort By". Taxonomic filters are shown as: "All: 1", "Fungi: 0", "Insects: 0", "Mammals: 1", "Plants: 0".

The search results list one entry:

- 1:** [Estrogen receptor 1](#)
 ESR1, *Homo sapiens*
 Hs.208124: 217 sequences.
 Order cDNA clone

A "Recent activity" sidebar on the right lists previous searches:

- ESR1 human (1)
- 9606[taxid] AND adult[res...] (2280)
- Estrogen receptor 1
- ESR1 (31)
- ESR1_HUMAN (0)

At the bottom of the sidebar, there is a link: » See more...

Unigene

Estrogen receptor 1 (ESR1)

SELECTED PROTEIN SIMILARITIES

Comparison of sequences in UniGene with selected protein reference sequences. The alignments can suggest function of a gene.

	Reference Protein	Species	Id(%)	Len(aa)
NP_001116212.1	estrogen receptor isoform 2	<i>H. sapiens</i>	100.0	594
XP_001097228.1	PREDICTED: estrogen receptor isoform 1	<i>M. mulatta</i>	99.7	594
NP_001158059.1	estrogen receptor	<i>P. anubis</i>	99.7	594
NP_031982.1	estrogen receptor	<i>M. musculus</i>	94.8	598
NP_001083084.2	estrogen receptor 1	<i>X. laevis</i>	82.4	584
NP_694491.1	estrogen receptor	<i>D. rerio</i>	69.9	531

GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

Restricted Expression: adult [\[show more like this\]](#)

[EST Profile:](#) Approximate expression patterns inferred from EST sources.
[\[Show more entries with profiles like this\]](#)

[GEO profiles:](#) Experimental gene expression data (Gene Expression Omnibus).

[cDNA Sources:](#) uterus; mammary gland; ovary; uncharacterized tissue; prostate; pancreas; muscle; kidney; testis; lung; brain; spleen; trachea; thymus; heart; pituitary gland; eye; connective tissue; adrenal gland; embryonic tissue; lymph node

MAPPING POSITION

Genomic location specified by transcript mapping, radiation hybrid mapping, genetic mapping or cytogenetic mapping.

Chromosome: 6
 Map position: 6q25.1
 UniSTS entry: Chr 6 [PMC108984P2](#)
 UniSTS entry: Chr 6 [RH103609](#)
 UniSTS entry: [BARC0078](#)
 UniSTS entry: [Esr1](#)
 UniSTS entry: [Esr1](#)

Unigene

SEQUENCES

Sequences representing this gene; mRNAs, ESTs, and gene predictions supported by transcribed sequences.

mRNA sequences (78)

BX640939.1	Homo sapiens mRNA; cDNA DKFZp686N23123 (from clone DKFZp686N23123)	PA
NM_000125.3	Homo sapiens estrogen receptor 1 (ESR1), transcript variant 1, mRNA	PA
M12674.1	Human estrogen receptor mRNA, complete cds	P
AF120105.1	Homo sapiens alternatively-spliced estrogen receptor alpha mRNA, partial cds	P
AY750962.1	Homo sapiens estrogen receptor alpha mamillary body 1 isoform mRNA, partial cds, alternatively spliced	P
AF258449.1	Homo sapiens estrogen receptor alpha mRNA, complete cds, alternatively spliced	P
AF258450.1	Homo sapiens estrogen receptor alpha mRNA, complete cds, alternatively spliced	P
AF258451.1	Homo sapiens estrogen receptor alpha mRNA, complete cds, alternatively spliced	P
DQ163909.1	Homo sapiens hippocampal estrogen receptor alpha isoform TADDI (ESR1) mRNA, partial cds, alternatively spliced	
BC128574.1	Homo sapiens estrogen receptor 1, mRNA (cDNA clone MGC:157709 IMAGE:40128595), complete cds	P

EST sequences (10 of 140) [[Show all sequences](#)]

AI025006.1	Clone IMAGE:1631712	mammary gland	3' read	P
AI073549.1	Clone IMAGE:1640294	testis	3' read	PA
AI127412.1	Clone IMAGE:1705898	heart	3' read	P
BX108369.1	Clone IMAGp998C181779_._; IMAGE:725321	ovary		A
AI202659.1	Clone IMAGE:1943657	prostate	3' read	A
AI274727.1	Clone IMAGE:1986505	uterus	3' read	A
AI273871.1	Clone IMAGE:1964251	ovary	3' read	A
AI370308.1	Clone IMAGE:1987501	uterus	3' read	A
CB215772.1	Clone IMAGE:5937582	uterus	5' read	
AI524356.1	Clone IMAGE:2118591	prostate	3' read	A

[Download Sequences](#)

Key to Symbols

- P** Has similarity to known **P**roteins (after translation)
- A** Contains a poly-**A**denylation signal
- S** Sequence is a **S**uboptimal member of this cluster
- M** Clone is putatively CDS-complete by **M**GC criteria

Unigene

EST Profile

Hs.208124 - ESR1: Estrogen receptor 1

Breakdown by Body Sites

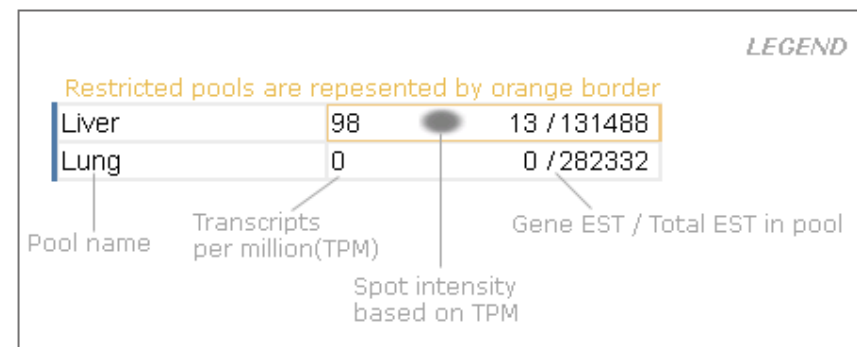
Hs.208124		
adipose tissue	0	0 / 13106
adrenal gland	30	1 / 33197
ascites	0	0 / 40015
bladder	0	0 / 29757
blood	0	0 / 123478
bone	0	0 / 71655
bone marrow	0	0 / 48801
brain	0	1 / 1100989
cervix	0	0 / 48171
connective tissue	6	1 / 149255
ear	0	0 / 16212

Breakdown by Developmental Stage

Hs.208124		
embryoid body	0	0 / 70761
blastocyst	0	0 / 62319
fetus	5	3 / 564012
neonate	0	0 / 31097
infant	0	0 / 23620
juvenile	0	0 / 55556
adult	21	42 / 1939121

Breakdown by Health State

Hs.208124		
adrenal tumor	78	1 / 12794
bladder carcinoma	0	0 / 17475
breast (mammary gland) tumor	63	6 / 94178
cervical tumor	0	0 / 34366
chondrosarcoma	12	1 / 82823
colorectal tumor	0	0 / 114246
esophageal tumor	0	0 / 17290
gastrointestinal tumor	0	0 / 119369
germ cell tumor	0	0 / 263845
glioma	0	0 / 106883



Unigene

- Οι βιβλιοθήκες είναι οργανωμένες στο library browser

[http://www.ncbi.nlm.nih.gov/UniGene/lbrowse2.cgi?TAXID=9606&log\\$=BlueSideBar](http://www.ncbi.nlm.nih.gov/UniGene/lbrowse2.cgi?TAXID=9606&log$=BlueSideBar)

Library Browser: Homo sapiens

http://www.ncbi.nlm.nih.gov/UniGene/lbrowse2.cgi?TAXID=9606&log\$=BlueSideBar

NCBI » UniGene » Homo sapiens » Library Browser

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search UniGene

Libraries for with minimum sequences

[Collapse All](#) | [Expand All](#)

Body Sites

▼ **adipose tissue** 18 libraries

Lib. ID	Library Name	Sequences
Lib.10983	Human Fat Cell 5'-Stretch Plus cDNA Library	9638
Lib.886	NCI_CGAP_Lip2	1740
Lib.16445	Sugano cDNA library, adipose tissue	1665
Lib.816	Adipose tissue, white II	1195

Not Shown: 14 libraries having fewer than 1000 sequences

▼ **adrenal gland** 30 libraries

Lib. ID	Library Name	Sequences
Lib.18302	ADRGL2	10385
Lib.7317	NIH_MGC_84	7572
Lib.6791	ADB	6475
Lib.927	NCI_CGAP_AA1	3363
Lib.16377	Sugano cDNA library, adrenal gland	2772
Lib.6815	cdA	2460
Lib.6792	ADC	1995
Lib.6793	Cu	1649
Lib.993	NCI_CGAP_Phe1	1356
Lib.766	Adrenal gland tumor	1183

Not Shown: 20 libraries having fewer than 1000 sequences

▼ **amniotic fluid** 63 libraries

Lib. ID	Library Name	Sequences
Lib.7332	AN0080	1112

Not Shown: 62 libraries having fewer than 1000 sequences

Digital differential display

- <http://www.ncbi.nlm.nih.gov/UniGene/help.cgi?item=DDD>
- Συγκρίνει ομάδες βιβλιοθηκών ESTs μεταξύ τους και βρίσκει ποιά γονίδια είναι περισσότερο εκφρασμένα στην κάθε μια από τις βιβλιοθήκες.
- Η σύγκριση γίνεται με Fisher's exact test.
- Για να γίνουν οι συγκρίσεις, θα πρέπει ο αριθμός των ESTs που αντιστοιχούν σε ένα γονίδιο, για την κάθε βιβλιοθήκη, να ξεπερνάει ένα κατώφλι, αλλιώς δεν μπορούν να φανούν οι διαφορές.
- Οι συγκρίσεις μπορούν να αποκαλύψουν γονίδια που παίζουν σημαντικό ρόλο σε κάποιο ιστό ή περιβαλλοντική κατάσταση ή ασθένεια.

Digital differential display

Digital Differential Display (DDD)

DDD is a tool for comparing EST profiles in order to identify genes with significantly different expression levels ([More about DDD](#)).

Species: *Homo sapiens* (human)

[Start Over](#)

Pool A: Muscle

2 libraries, 5084 ESTs

[Edit Pool](#)

Pool B: Skin

4 libraries, 35274 ESTs

[Edit Pool](#)

[New Pool](#)

Differential Display Results

The following genes (UniGene entries) display statistically significant differences in EST counts by the Fisher Exact Test.

A Muscle	B Skin	UniGene Entry
0.0519 ●	0.0000 ●	Hs.728212 Transcribed locus, strongly similar to NP_001091.1 actin, alpha skeletal muscle [Homo sapiens]
0.0228 ●	0.0001 ●	Hs.134602 Titin (TTN)
0.0142 ●	0.0000 ●	Hs.726317 Transcribed locus, strongly similar to NP_036662.1 creatine kinase M-type [Rattus norvegicus]
0.0134 ●	0.0000 ●	Hs.631558 Troponin T type 1 (skeletal, slow) (TNNT1)
0.0124 ●	0.0000 ●	Hs.320890 Troponin I type 1 (skeletal, slow) (TNNI1)
0.0104 ●	0.0000 ●	Hs.719946 Myosin, heavy chain 7, cardiac muscle, beta (MYH7)
0.0094 ●	0.0000 ●	Hs.517586 Myoglobin (MB)
0.0092 ●	0.0001 ●	Hs.598320 Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)

Microarrays

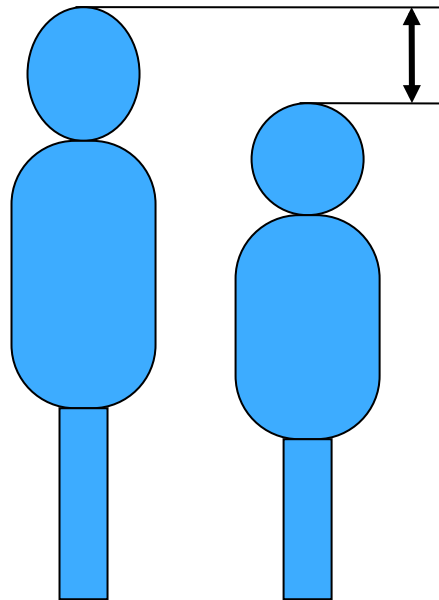
Microarrays

- Chips ή αντικειμενοφόροι πλάκες που μετράνε την γονιδιακή έκφραση (και όχι μόνο).
- Δεν μετράνε απόλυτες τιμές συγκεντρώσεων μεταγραφημάτων.
- Μετράνε σχετικές αλλαγές.
 - Για μια συνθήκη, πόσο πιο αυξημένη ή μειωμένη είναι η έκφραση ενός γονιδίου, σε σχέση με μια άλλη συνθήκη.
 - Π.χ. Πόσο πιο αυξημένη/μειωμένη είναι η έκφραση ενός γονιδίου A μετά από άσκηση (control: αμέσως πριν την άσκηση).
- Δεν μπορούμε να συγκρίνουμε το γονίδιο A με το γονίδιο B

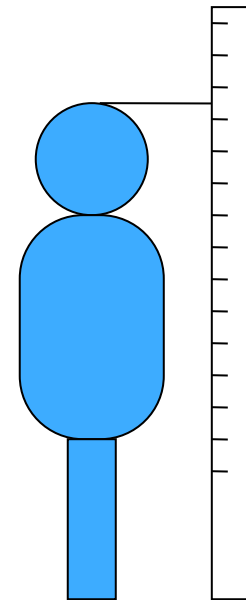
- Στο Chip υπάρχουν για το κάθε μεταγράφημα:
 - cDNAs ολόκληρου του μεταγραφώματος (spotting) (PCR)
 - ολιγομερή (τα λεγόμενα oligo-probes), με μήκος 25-70 βάσεις.
 - Spotting
 - In situ synthesis

Σχετικές διαφορές

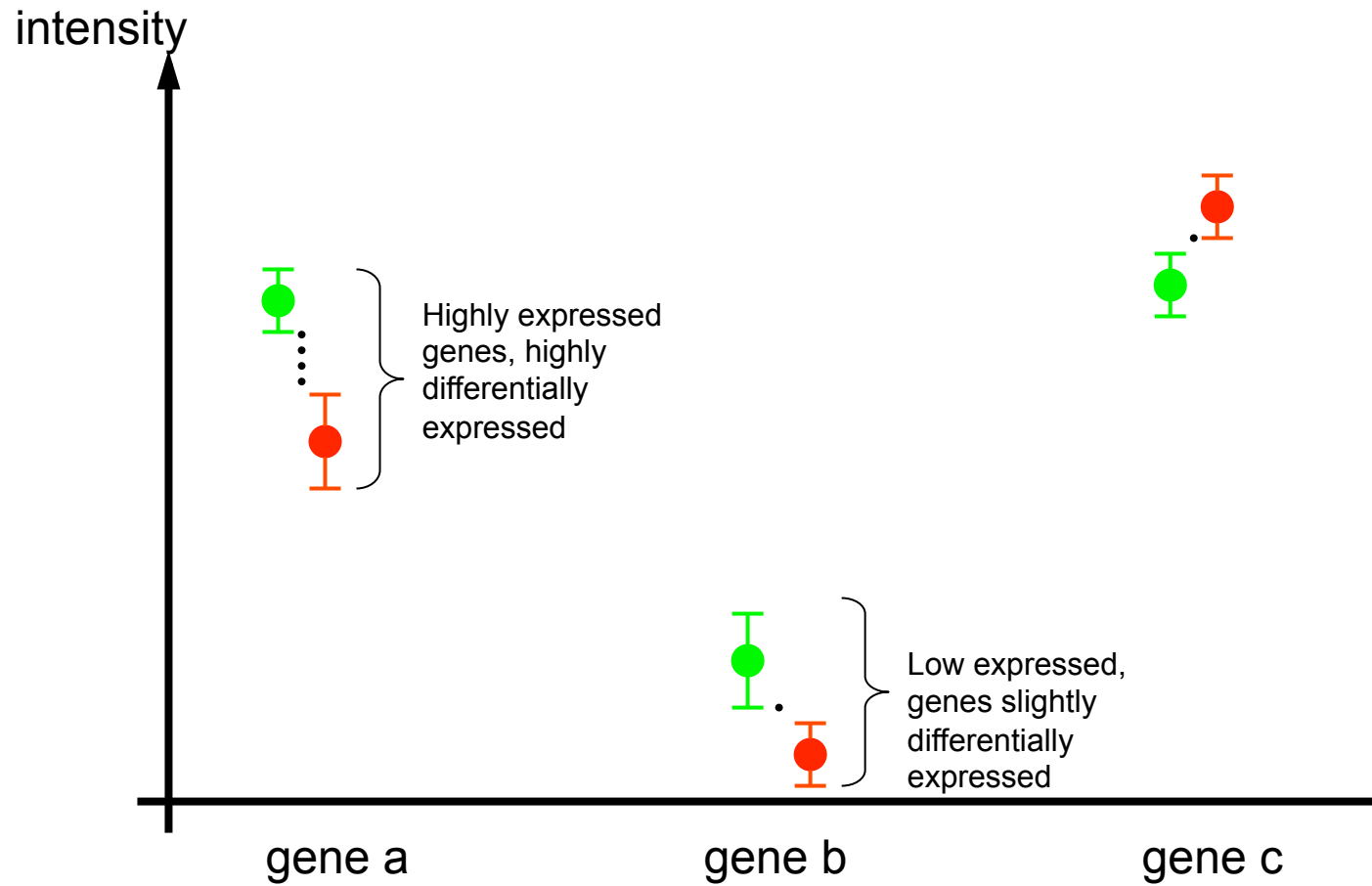
relative measurement 2
COLORS (Cy5 / Cy3)



absolute measurement
1 COLOR (Cy3)

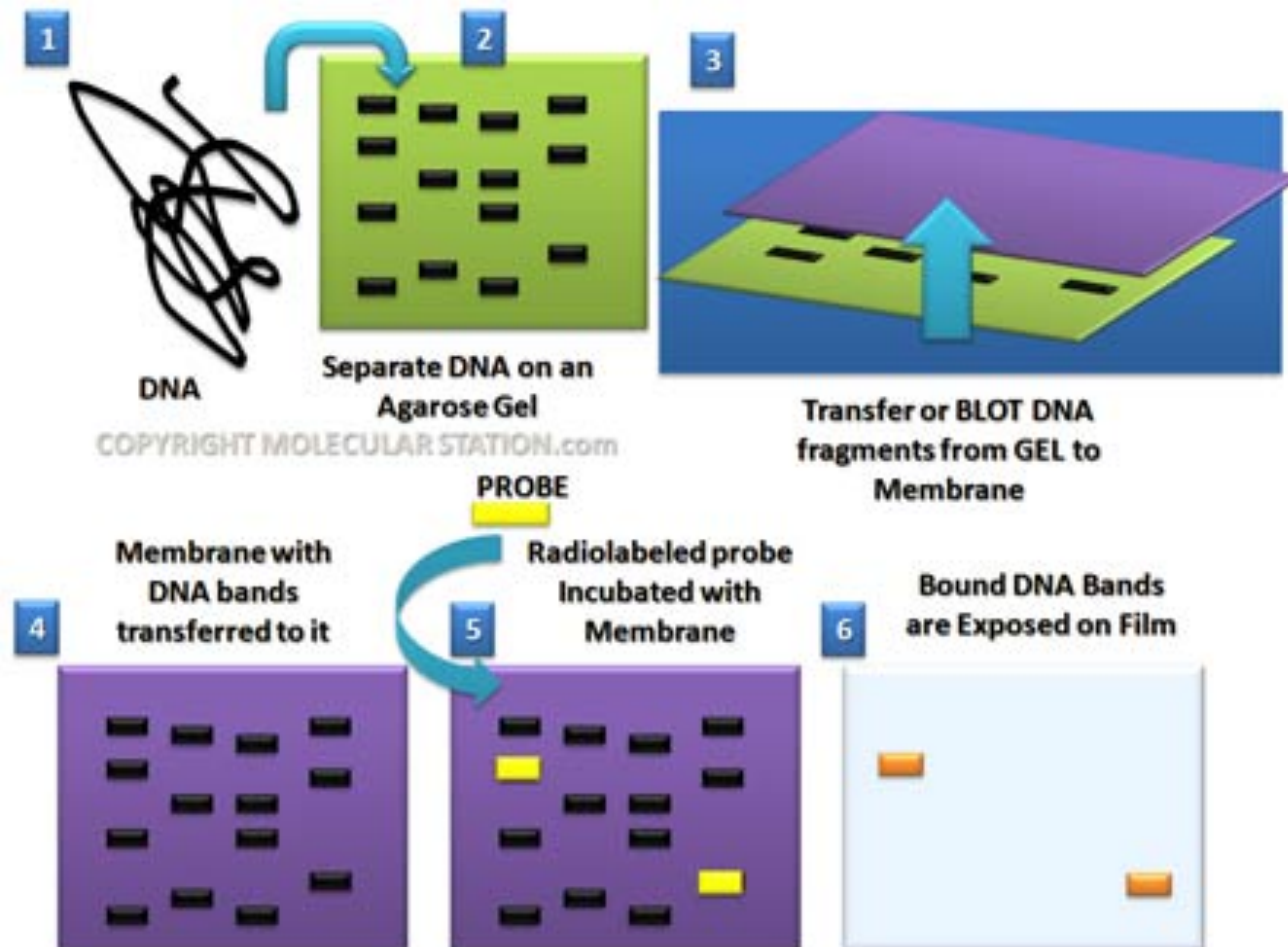


Σχετικές διαφορές



Πρόγονοι των microarrays

- Southern blot (DNA - DNA) - 1975 (Ed Southern)
- Ελέγχει για την παρουσία μιας συγκεκριμένης DNA ακολουθίας σε ένα δείγμα



Όλιγο-Μικροσυστοιχίες (oligo-microarrays)

- Chip στο οποίο υπάρχουν DNA ολιγομερή (τα λεγόμενα probes), με μήκος 25-70 βάσεις που αντιπροσωπεύουν το γονιδίωμα ενός οργανισμού.
- Περισσότερα από ένα διαφορετικά probes μπορεί να υπάρχουν για ένα γονίδιο.
- Πάνω σε αυτά τα probes υβριδίζονται τα συμπληρωματικά τους cDNAs.
- Το κάθε cDNA είναι συνδεδεμένο με φθορίζουσα χρωστική.
- Κάθε κηλίδα (spot) στο chip αντιστοιχεί σε ένα συγκεκριμένο είδος probes.
- Η ένταση της φθορίζουσας χρωστικής υποδηλώνει την ένταση με την οποία εκφράζεται το mRNA για τον αντίστοιχο probe.

Τι μπορούν να ελέγξουν

- Έκφραση mRNA.
- Έκφραση micro-RNA.
- CGH (comparative genomic hybridization).
- SNPs.
- DNA μεθυλίωση.
- ChIP-on-chip.

Τα βήματα

From: [PLoS Comput Biol. 2010 May; 6\(5\): e1000786.](#)
Published online 2010 May 27. doi: 10.1371/journal.pcbi.1000786.
[Copyright/License](#)

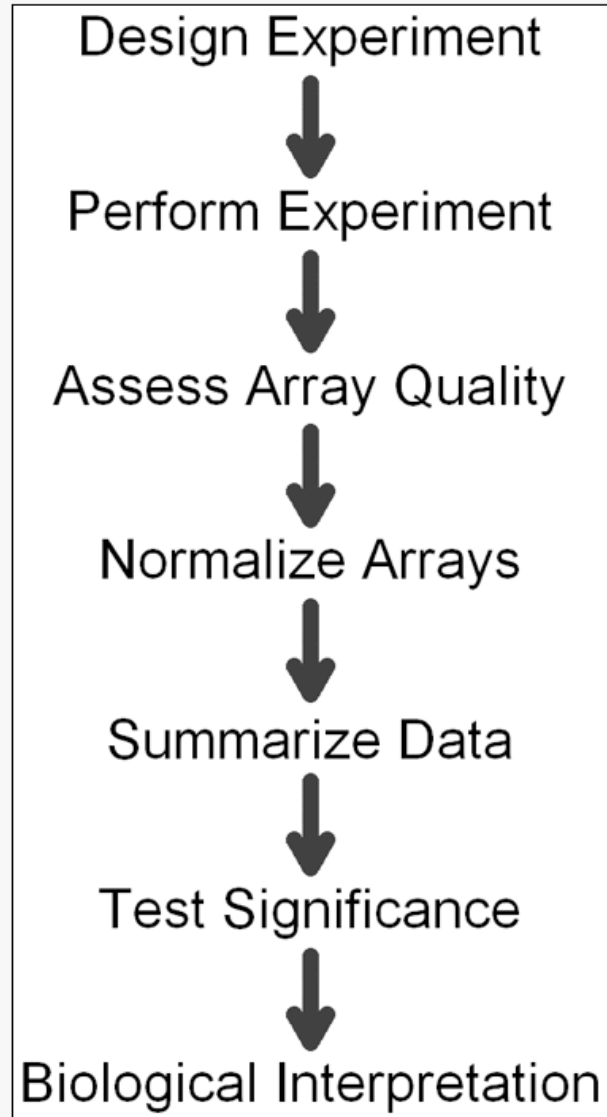
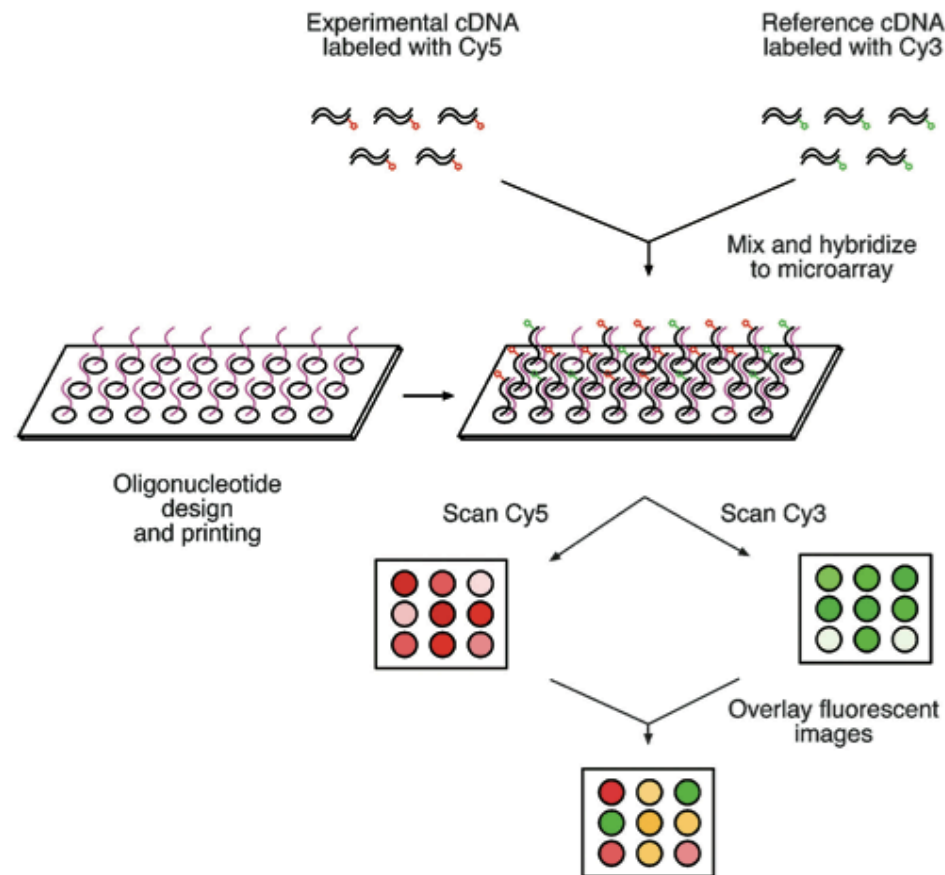


Figure 1
Steps in a typical microarray analysis.

Μικροσυστοιχίες (microarrays)

- Με δύο χρώματα (πράσινο/κόκκινο)
- Το ένα χρώμα είναι για μια συγκεκριμένη συνθήκη και το άλλο χρώμα για μια άλλη (συνήθως control).
- Εξάγεται το ολικό RNA για την κάθε συνθήκη.
- Ελέγχεται η ποσότητα και ποιότητα του RNA.
- Γίνεται σήμανση του κάθε δείγματος RNA με συγκεκριμένη χρωστική (π.χ. Cy5:red - Cy3:green).
- Τα 2 δείγματα αναμιγνύονται και το μίγμα υπόκειται σε υβριδισμό πάνω στο chip.
- Μετρίεται η ένταση της κάθε μιας από τις 2 χρωστικές, για κάθε κηλίδα.
- Υπολογίζεται ο λόγος των εντάσεων των 2 χρωστικών, στην κάθε κηλίδα.

Μικροσυτοιχίες



Μικροσυστοιχίες (microarrays)

- Αν το γονίδιο εκφράζεται περισσότερο στην A συνθήκη (κόκκινη χρωστική) από ότι στην control (πράσινη χρωστική), τότε ο λόγος συνθήκη_A/control (κόκκινη/πράσινη) θα είναι $\lambda > 1$, αλλιώς σε αντίθετη περίπτωση $0 < \lambda < 1$.
- Αν το γονίδιο εκφράζεται με διπλάσια ένταση στην συνθήκη A, σε σχέση με την συνθήκη control, τότε ο λόγος θα είναι $\lambda = 2$.
- Αν το γονίδιο εκφράζεται με τη μισή ένταση στην συνθήκη A, σε σχέση με την συνθήκη control, τότε ο λόγος θα είναι $\lambda = 0.5$.
- Μετατρέποντας τους λόγους σε \log_2 , έχουμε:
 - $\lambda = 2 \rightarrow \log_2 \lambda = 1$
 - $\lambda = 0.5 \rightarrow \log_2 \lambda = -1$
 - Με την κανονικοποίηση σε \log_2 τα δεδομένα γίνονται συμμετρικά.

Μικροσυστοιχίες (microarrays)

- Πότε θεωρούμε ότι ένα γονίδιο υπερ/υπό-εκφράζεται σε μια συγκεκριμένη συνθήκη.
 - $\text{Log}_2\lambda > 1$ ή $\text{Log}_2\lambda < -1$ (διπλάσια/υποδιπλάσια έκφραση σε σχέση με τη συνθήκη control).
 - Με στατιστικές μεθόδους (t-test, ANOVA).

microarrays

- Κατανομή των σημάτων.
- Μετατροπή - κανονικοποίηση των σημάτων.

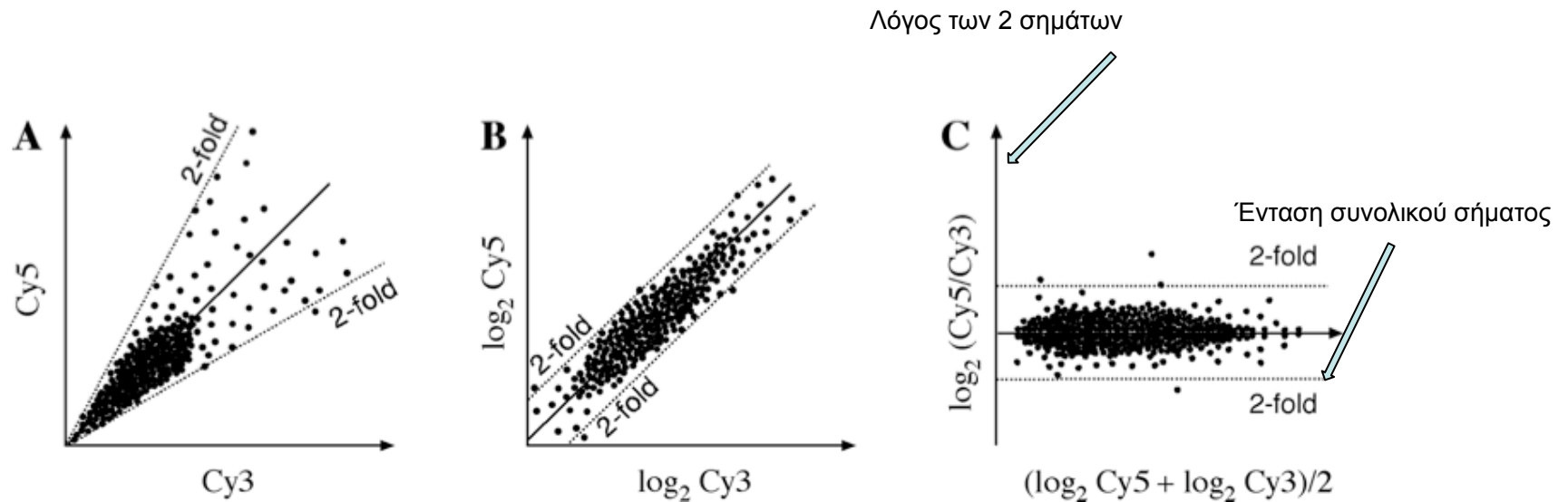


Figure 18.5: Scatter plot of gene expression analysis showing the process of data normalization. The solid line indicates linear regression of the data points; dashed lines show the cutoff for a twofold change in expression. **(A)** Plot of raw fluorescence signal intensities of Cy5 versus Cy3. **(B)** Plot of the same data after log transformation to the base of 2. **(C)** Plot of mean log intensity versus log ratio of the two fluorescence intensities, which shifts the data points to around the horizontal axis, making them easier to visualize.

Ομαδοποίηση γονιδίων με την ίδια συμπεριφορά.

- Χρειαζόμαστε αρκετά σημεία (διαφορετικές συνθήκες ή χρονικές στιγμές)
- Με μεθόδους αποστάσεων, όπου οι μετρήσεις ενός γονιδίου για διαφορετικές συνθήκες αποτελούν ένα διάνυσμα.
- Υπολογίζουμε αποστάσεις μεταξύ διαφορετικών διανυσμάτων (γονιδίων).
 - Ευκλείδια απόσταση
 - Συντελεστής συσχέτισης Pearson (Pearson correlation coefficient).
 - Δημιουργείται πίνακας αποστάσεων μεταξύ των γονιδίων.
 - Το αντίστοιχο μπορεί να γίνει και για να ομαδοποιήσουμε κοινές συνθήκες.

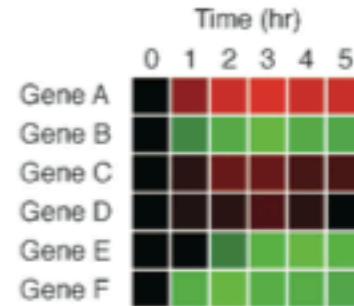
Μικροστοιχίες

	0 hr	1 hr	2 hr	3 hr	4 hr	5 hr
Gene A	1	4	6	8	6	6
Gene B	1	0.6	0.3	0.1	0.3	0.4
Gene C	1	2	4	4	3	3
Gene D	1	1.5	2	3	2	1
Gene E	1	1	0.5	0.2	0.1	0.2
Gene F	1	0.3	0.1	0.2	0.3	0.4

convert to false colors



\log_2 conversion



	Gene B	Gene C	Gene D	Gene E	Gene F
Gene A	-0.82	0.96	0.65	-0.68	-0.79
Gene B		-0.85	-0.86	0.66	0.67
Gene C			0.70	-0.65	-0.87
Gene D				-0.41	-0.72
Gene E					0.26

calculating Pearson correlation coefficients between genes



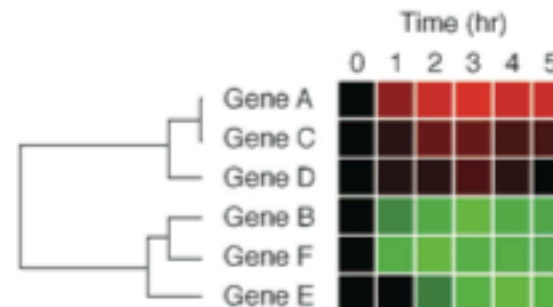
	0 hr	1 hr	2 hr	3 hr	4 hr	5 hr
Gene A	0	2	2.6	3	2.6	2.6
Gene B	0	-0.7	-1.7	-3.3	-1.7	-1.3
Gene C	0	1	2	2	1.6	1.6
Gene D	0	0.6	1	1.6	1	0
Gene E	0	0	-1	-2.3	-3.3	-2.3
Gene F	0	-1.7	-3.3	-2.3	-1.7	-1.3



conversion of coefficients to positive distance values

	Gene B	Gene C	Gene D	Gene E	Gene F
Gene A	1.82	0.04	0.35	1.68	1.79
Gene B		1.85	1.86	0.34	0.33
Gene C			0.30	1.65	1.87
Gene D				1.41	1.72
Gene E					0.74

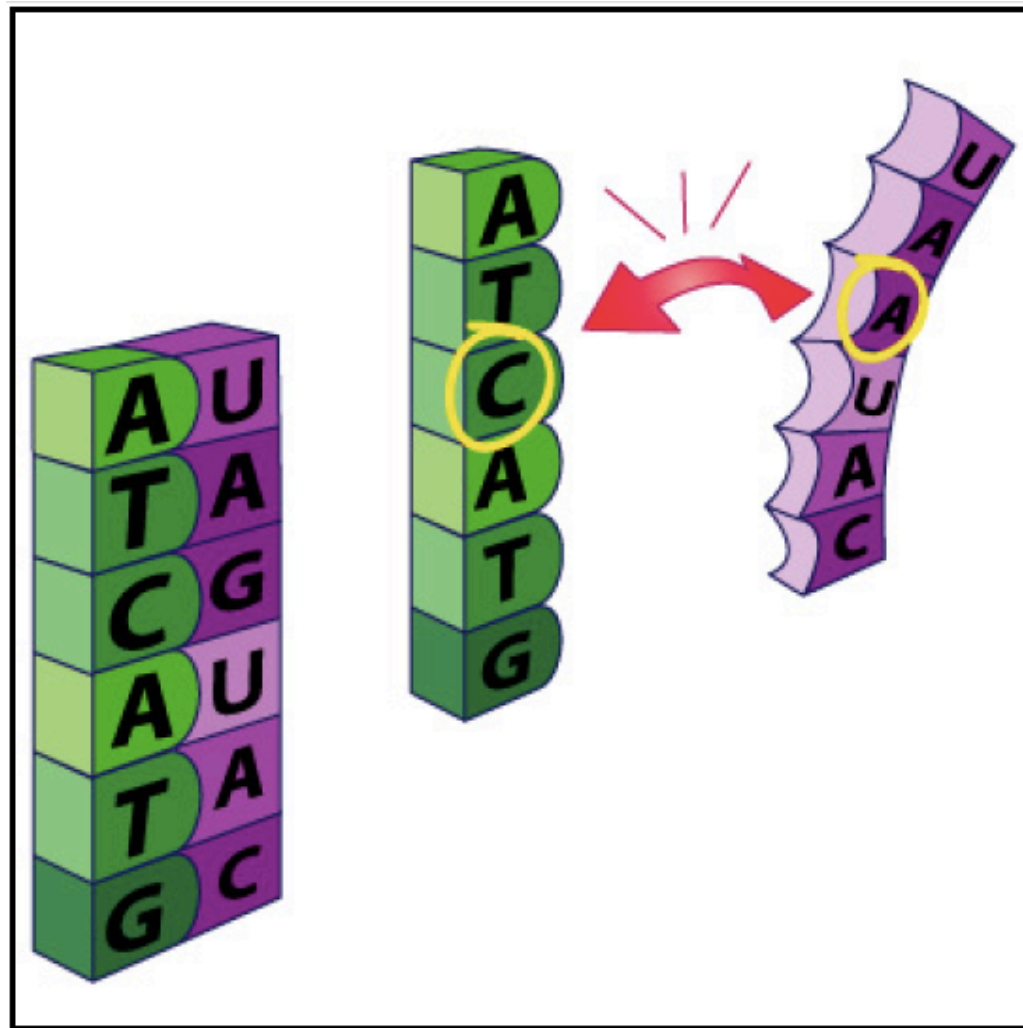
hierarchical clustering



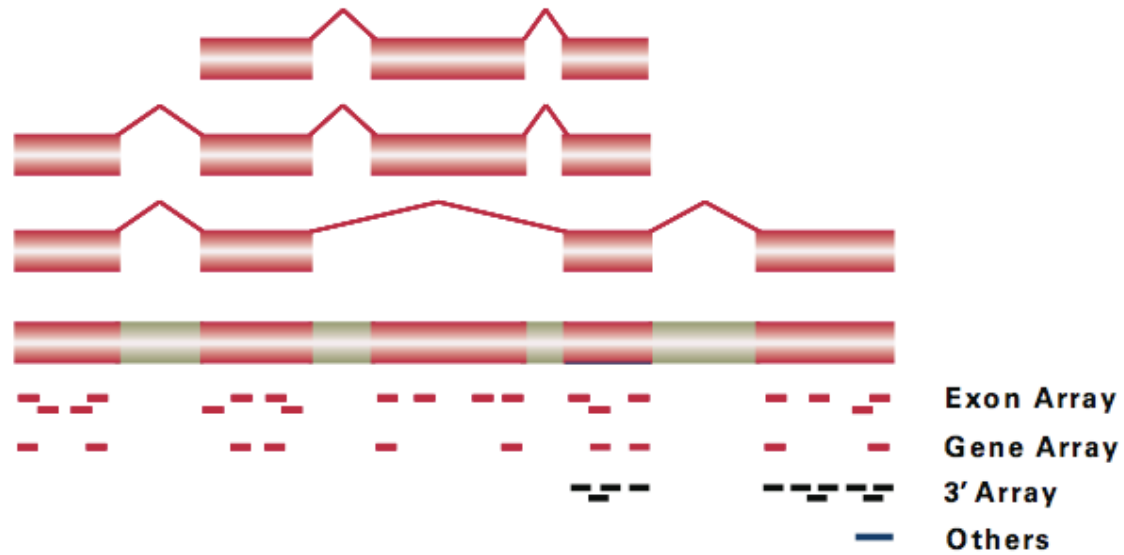
Μήκος των probes

- Affymetrix: 25mer
- Agilent: 60mer
- Illumina: 50mer
- Όσο μεγαλύτερο το probe, τόσο πιο ευαίσθητο (μπορεί να ανιχνεύσει targets σε μικρότερες συγκεντρώσεις).
- Όμως, μεγαλύτερο probe μπορεί και να ανεχτεί περισσότερους πολυμορφισμούς (ή mismatches).

Υβριδισμός των probes



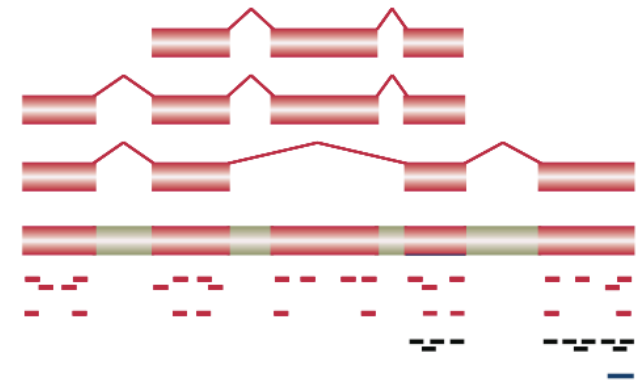
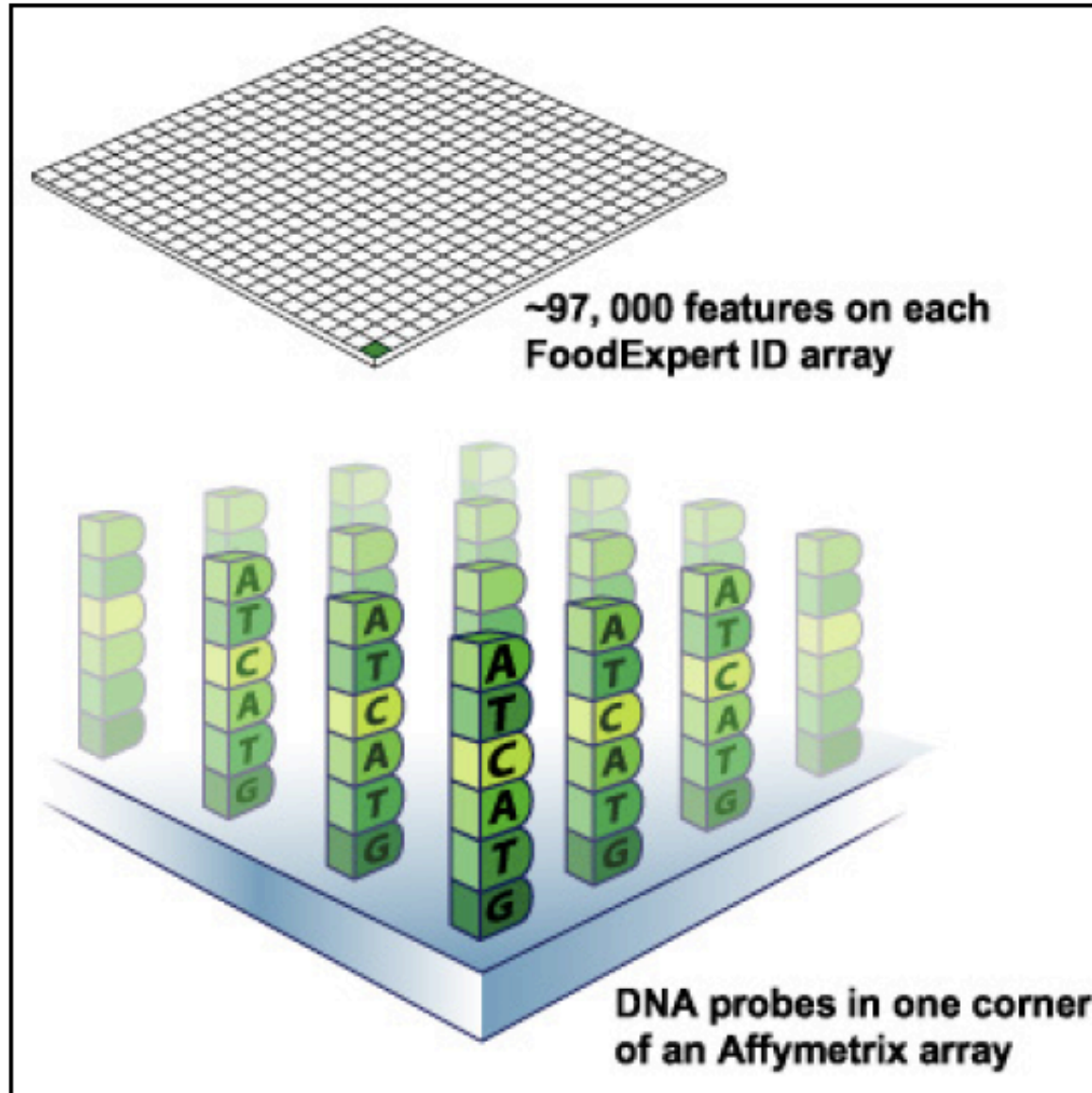
Probes & chips



Σχεδιασμός των probes

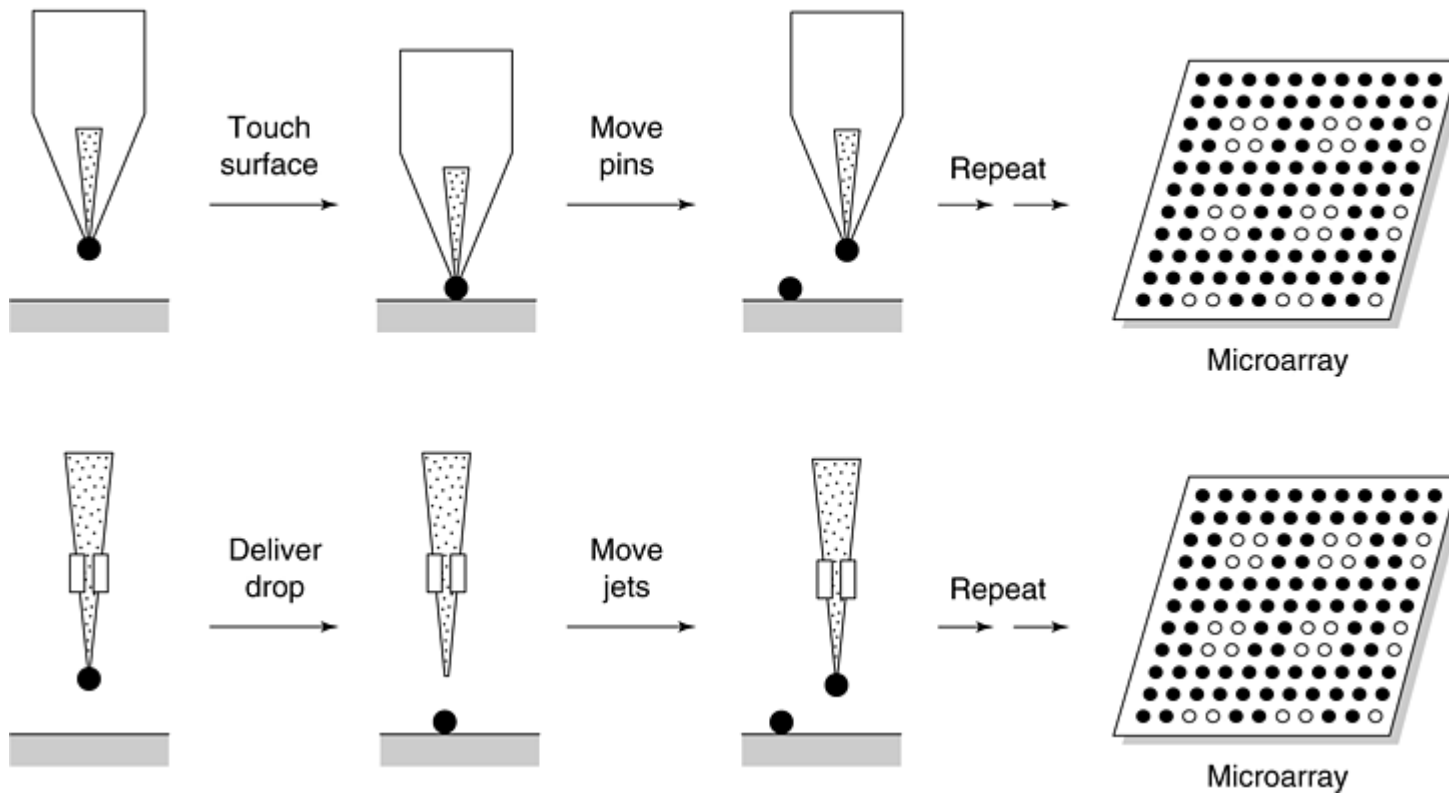
- Θα πρέπει η ακολουθία τους να είναι μοναδική ώστε να αντικατοπτρίζει την έκφραση για το γονίδιο που σχεδιάστηκαν.
- Όλοι οι probes θα πρέπει να έχουν παρόμοια T_m (βέλτιστη θερμοκρασία υβριδισμού).
- Τα probes δεν θα πρέπει να σχηματίζουν δευτεροταγείς δομές που τα εμποδίζουν να υβριδιστούν με τον στόχο.

spots



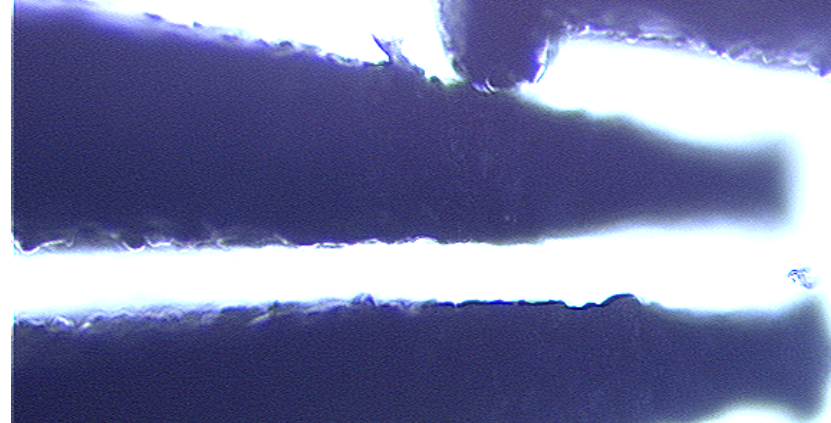
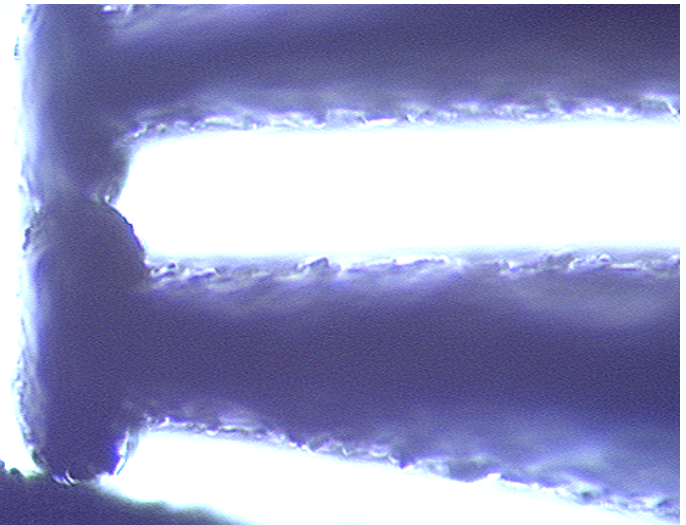
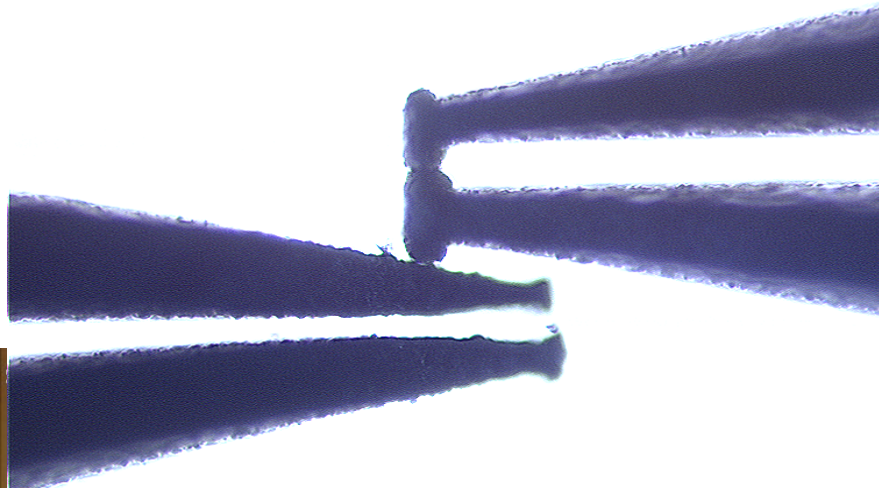
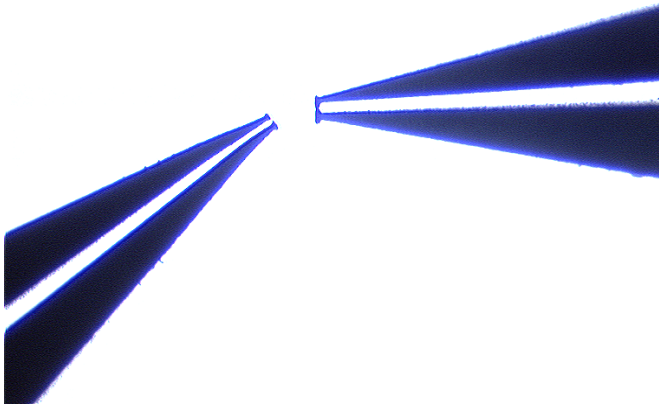
Array Spotting

- Με βελόνες (pins) (απώθηση έτοιμων cDNAs ή oligos).
- Με inkjet (Agilent: spin-off της Hewlett-Packard) (in situ synthesized oligos)



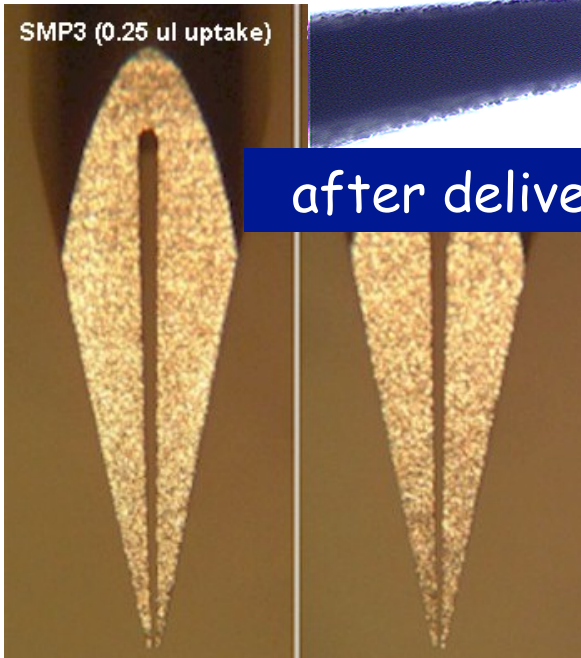
spotting pin quality decline

after delivery of 5×10^5 spots



SMP3 (0.25 ul uptake)

after delivery of 3×10^5 spots



Inkjet printing - Agilent

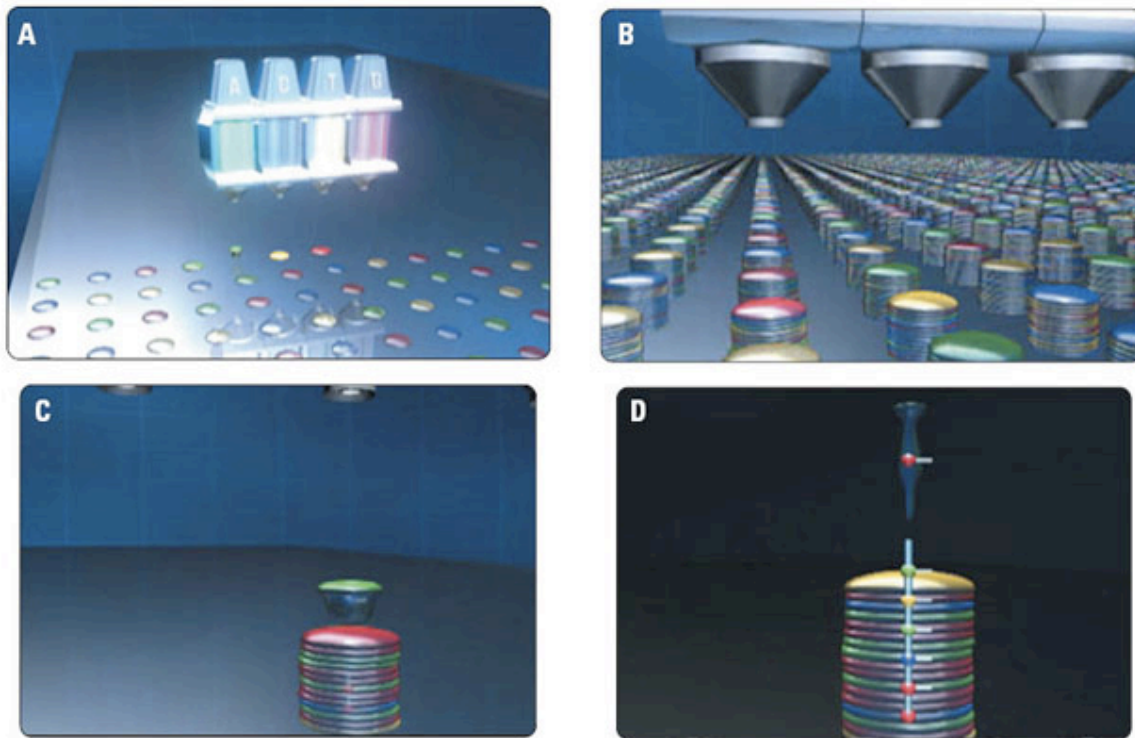


Figure 1. These four images communicate the general mechanism for oligo synthesis via inkjet printing. A: the first layer of nucleotides is deposited on the activated microarray surface. B: growth of the oligos is shown after multiple layers of nucleotides have been precisely printed. C: close-up of one oligo as a new base is being added to the chain, which is shown in figure D.

Φωτολιθογραφία - Affymetrix

Youtube:

<http://www.youtube.com/watch?v=ui4BOtwJEXs&feature=related>

Affymetrix:

http://media.affymetrix.com:80/about_affymetrix/outreach/lesson_plan/downloads/student_manual_activities/activity3/activity3_manufacturing_background.pdf

Φωτολιθογραφία

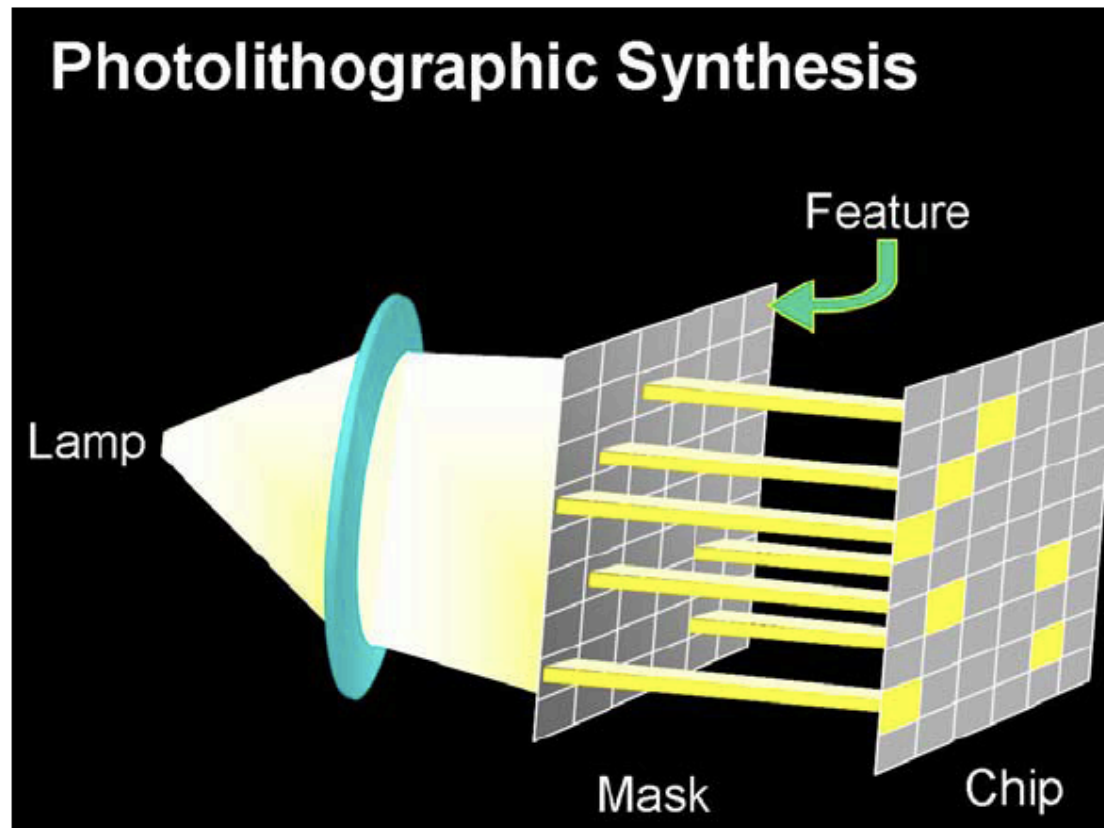


Fig. 1 - The mask only allows light to pass to specific features on the chip

Φωτολιθογραφία

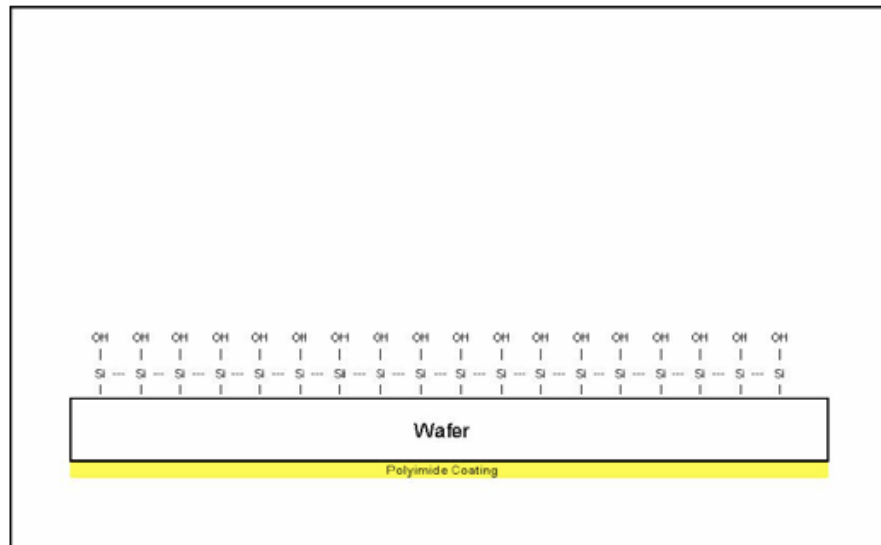


Fig. 3 - Silanation - each "Si" is a starting point

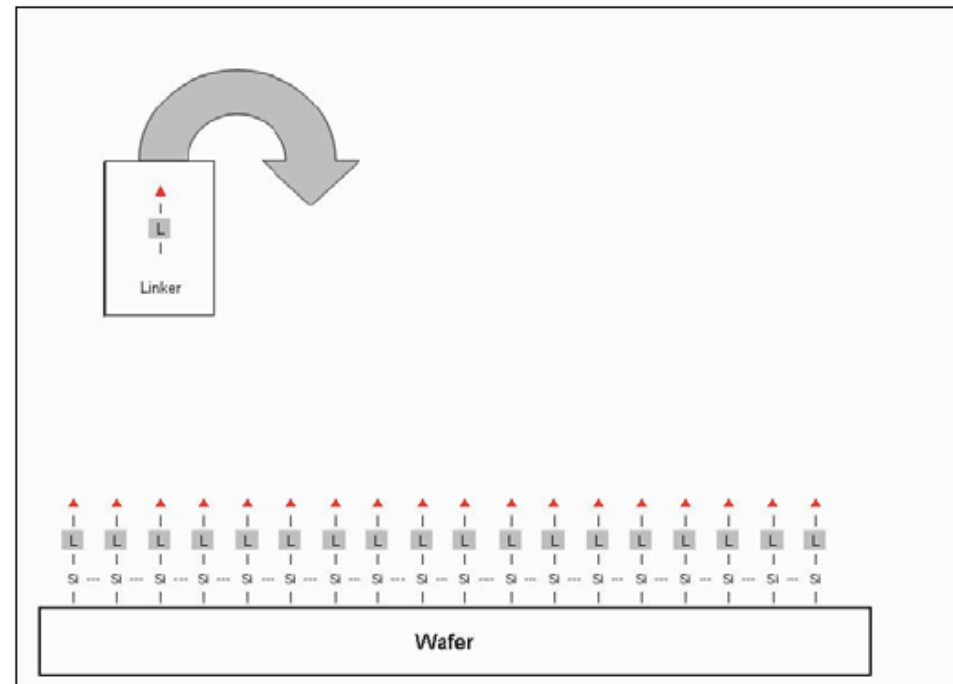


Fig. 4 - The addition of the linker molecule

Φωτολιθογραφία

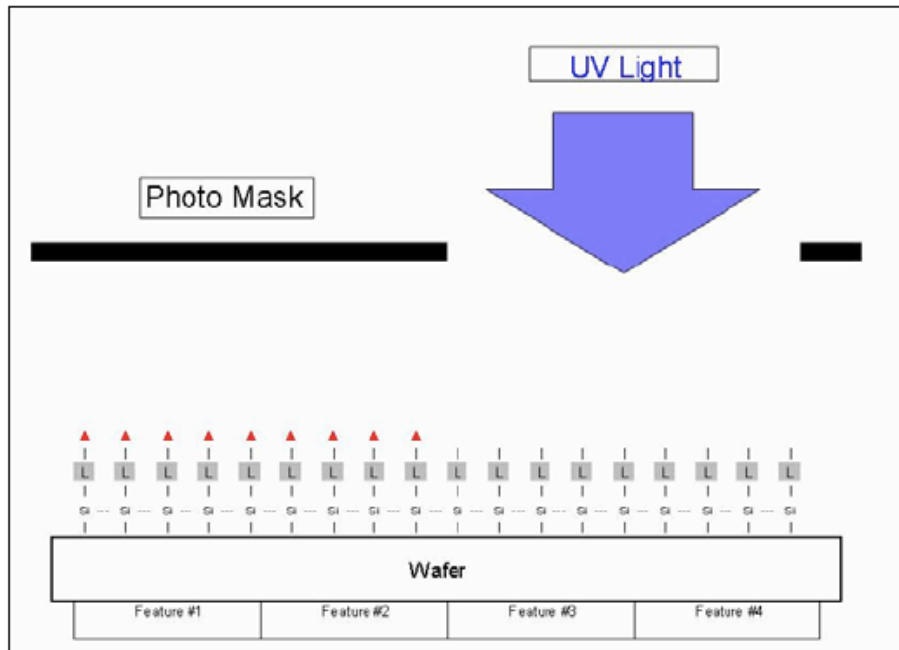


Fig. 5 - Deprotection of Feature #3 and 4

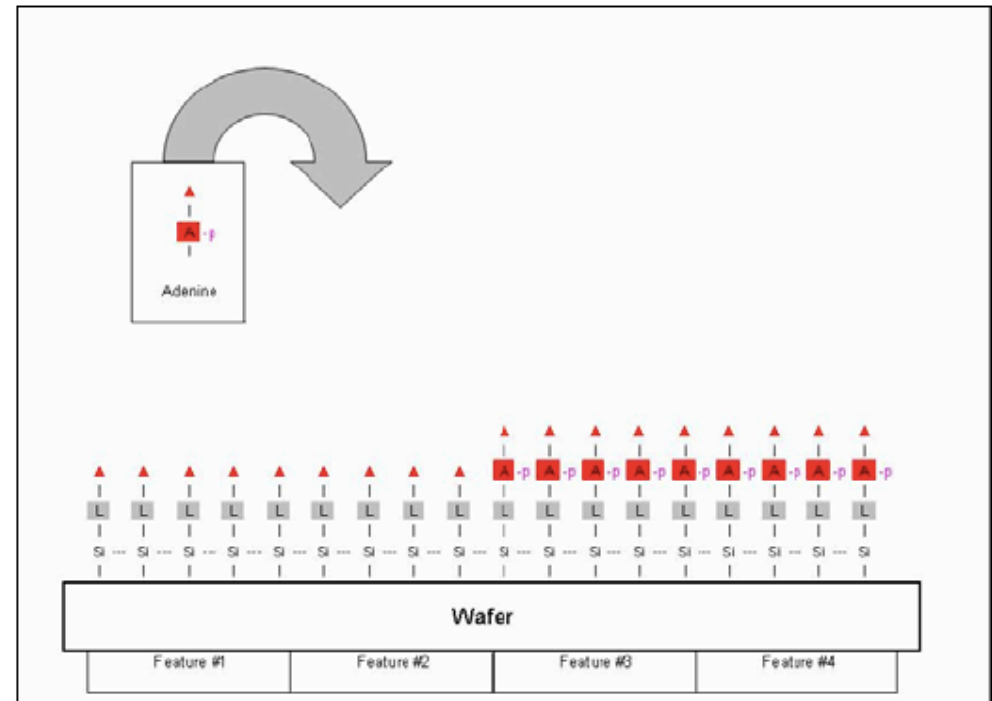


Fig.6 - Addition of Adenine (A) Nucleotides

Φωτολιθογραφία

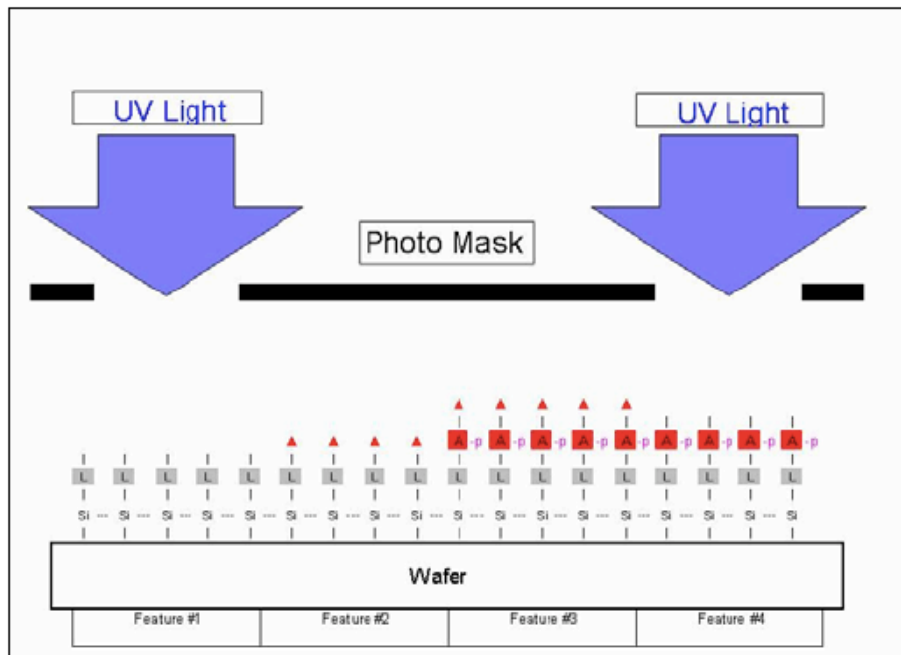


Fig.7 - 2nd mask (deprotecting Feature 1 and 4)

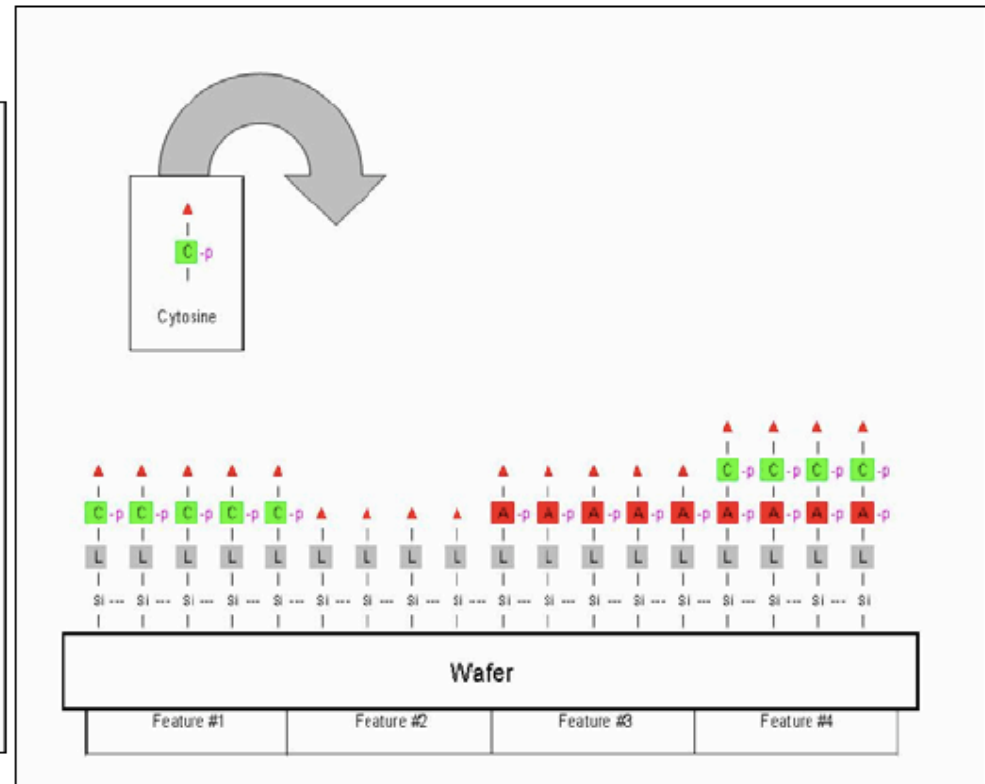


Fig. 8 - Addition of the 2nd nucleotide (C)

Feature

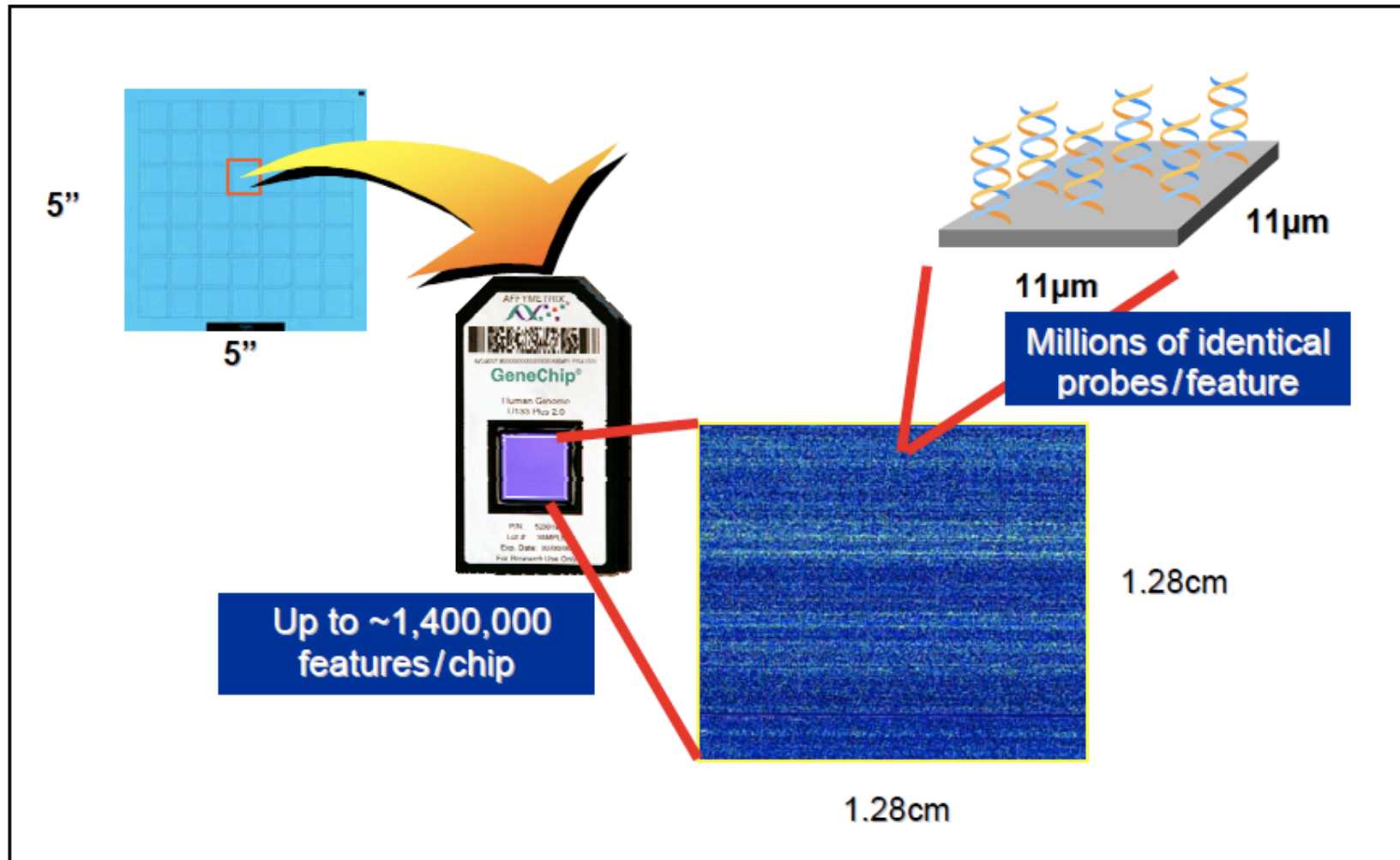
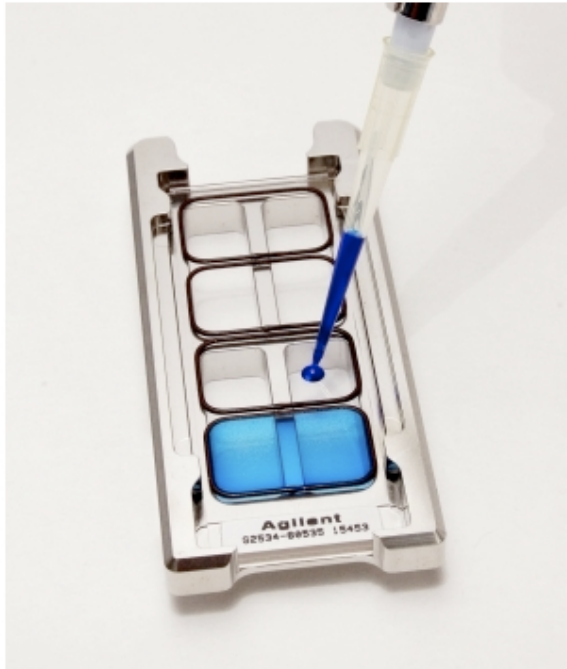
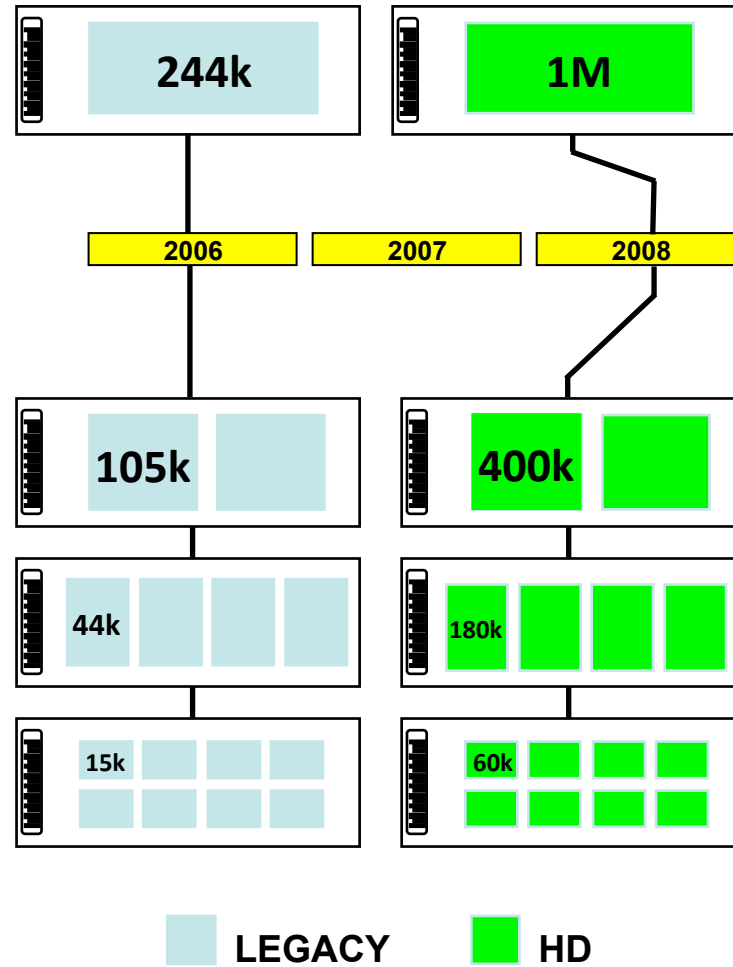


Fig. 17 - Dimension of a GeneChip: Wafer to Chip to a Single Feature

Agilent chips



Agilent chips



Agilent platform

- Φούρνος υβριδισμού



- Laser Scanner



Affymetrix Arrays/strips/plates

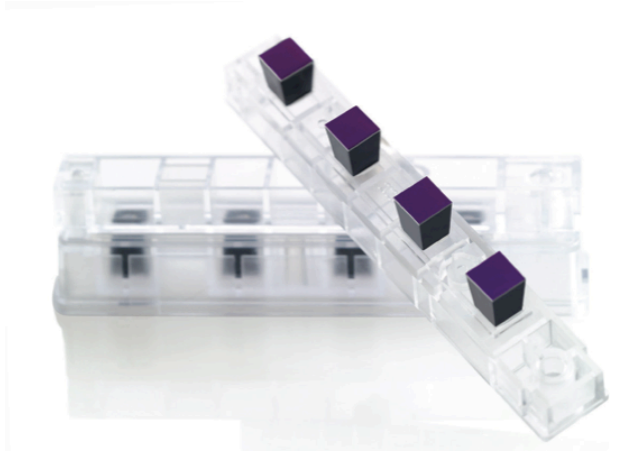


Figure 1: GeneChip® Human Genome U133 Arrays shown in cartridge and plate formats.



Affymetrix platform



Affymetrix platforms



GeneAtlas™ System & array strips



GeneTitan® System & array plates

Έλεγχος Ποσότητας/Ποιότητας ολικού RNA

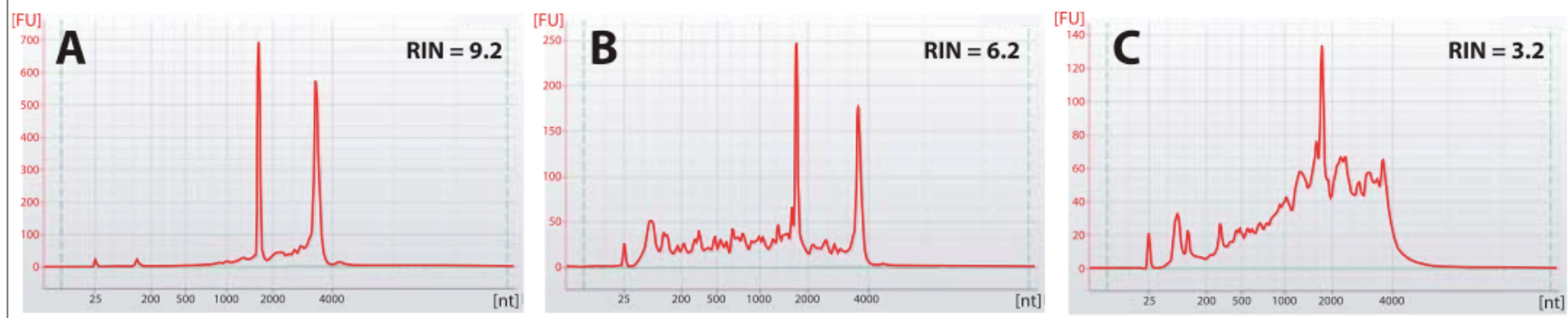
- Nanodrop (φωτόμετρο που χρησιμοποιεί ελάχιστο όγκο δείγματος - 1λ) για ακριβής ποσοτικοποίηση.
 - A260 nm -> νουκλεοτίδια, RNA, ssDNA, dsDNA
 - A280 nm -> πρωτεΐνες, φαινόλες κ.α.
 - A230 nm -> EDTA, φαινόλες, υδατάνθρακες.
 - A260/A280 nm ~ 1.8 - 2.1
 - 260/230 nm ~ 2 - 2.2



Έλεγχος Ποσότητας/Ποιότητας ολικού RNA

- Agilent 2100 Bioanalyzer για έλεγχο ποιότητας του RNA (τυχόν αποδόμιση). (~ €20.000)
- Ηλεκτροφόρηση σε τριχοειδή
 - Λόγος 28S/18S rRNA 2:1
 - RNA integrity number (RIN)

Figure 1.2 Example Agilent Bioanalyzer Electropherograms from three different total RNAs of varying integrity. Panel [A] represents a highly intact total RNA (RIN = 9.2), panel [B] represents a moderately intact total RNA (RIN = 6.2), and panel [C] represents a degraded total RNA sample (RIN = 3.2).



NOTE: Total RNAs with lower RIN values may require increased input amounts to generate enough aRNA for hybridization to an array.

Έλεγχος Ποσότητας/Ποιότητας ολικού RNA

1 Fast and easy operation Add sample



- Ready-to-use reagent kits
- Quick-start instructions
- Chip preparation in less than 5 minutes
- Minimal use of hazardous chemicals and waste disposal
- Sample volumes in the μL -range

2 Automation Start chip run



- Start analysis at the press of a button
- Predefined protocols
- System uses internal standards to calculate results

3 Digital data in 30 minutes Watch real-time data display



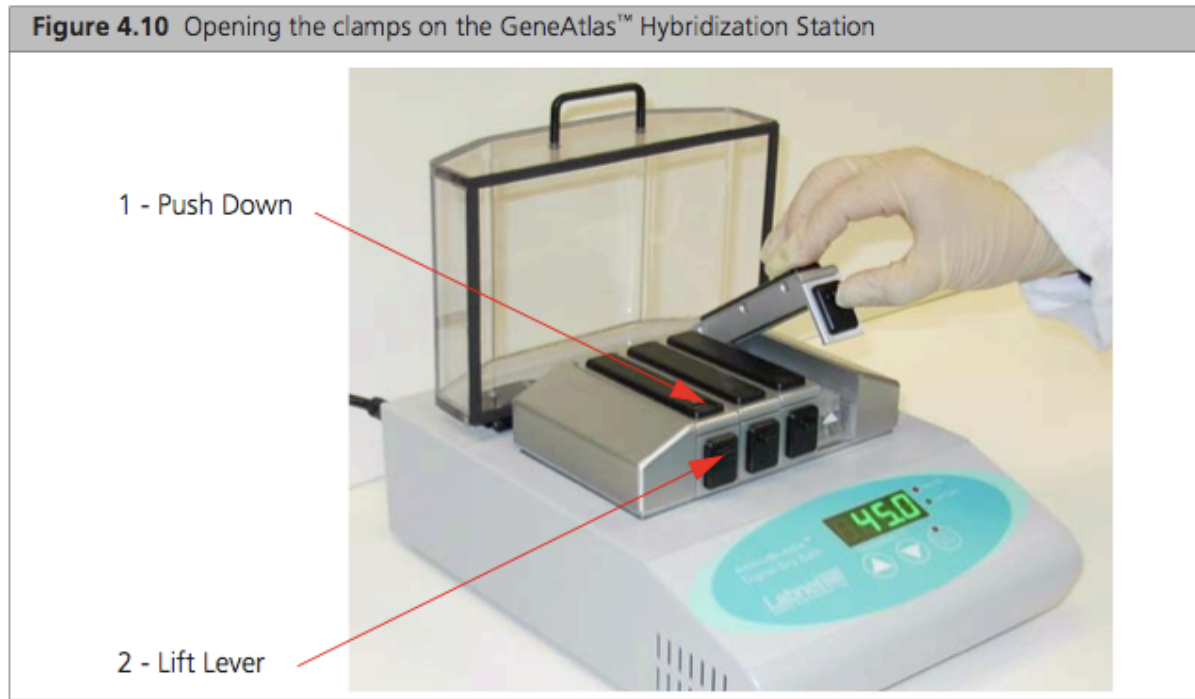
- Automated data analysis
- Digital data can be filed in a database or shared
- No user-dependent data interpretation

Υβριδισμός

Όλα τα probes υβριδίζονται στην ίδια θερμοκρασία.

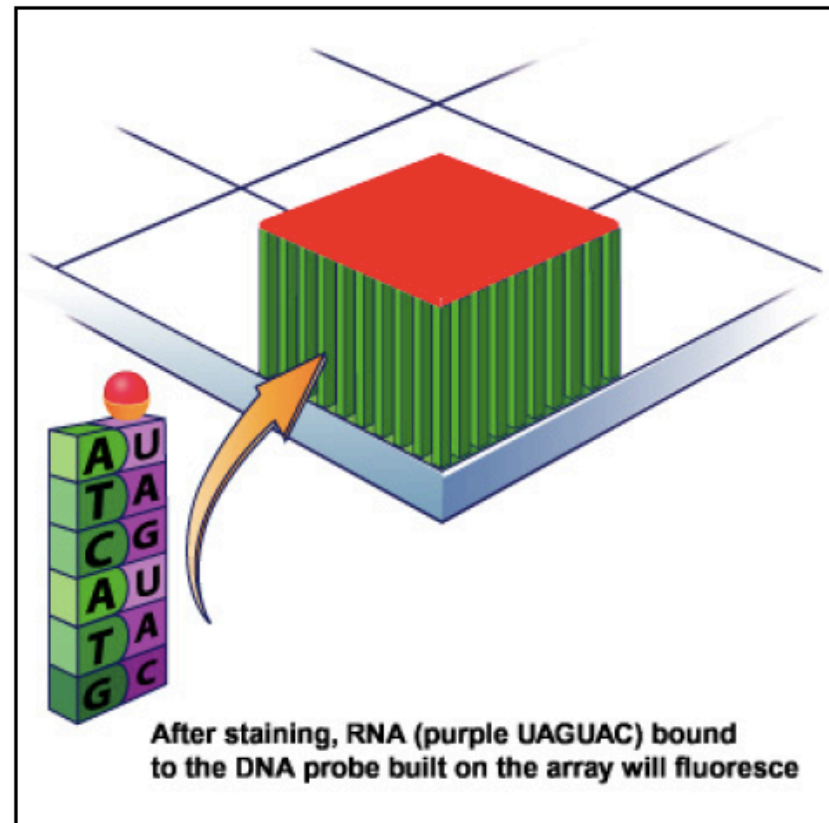
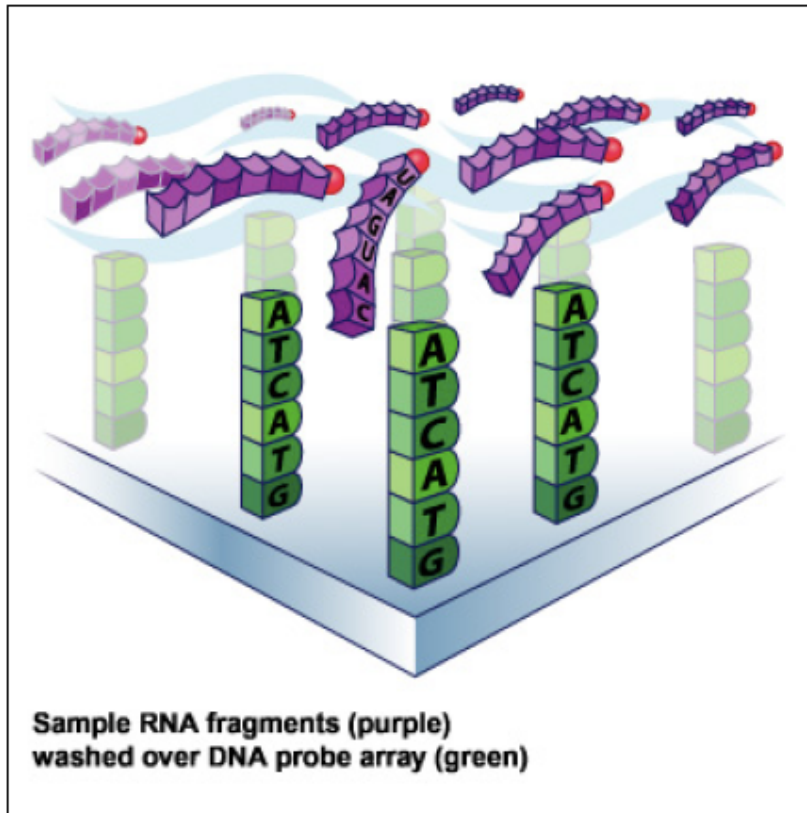
Όμως, όλα τα probes δεν έχουν την ίδια βέλτιστη θερμοκρασία υβριδισμού.

Άρα, υπάρχουν προβλήματα για επιμέρους probes



Φούρνος υβριδισμού του Affymetrix GeneAtlas

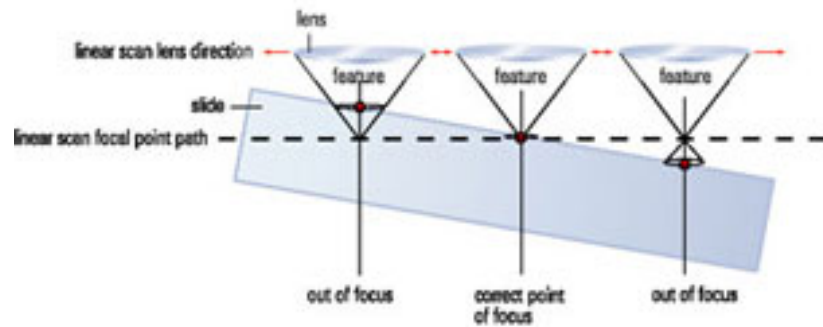
Υβριδισμός



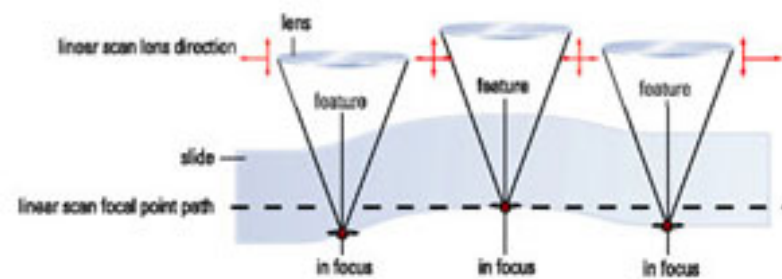
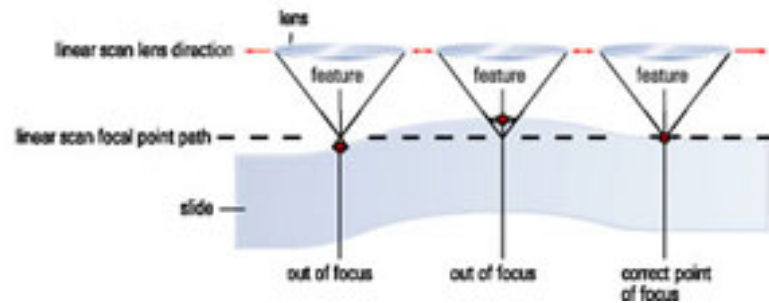
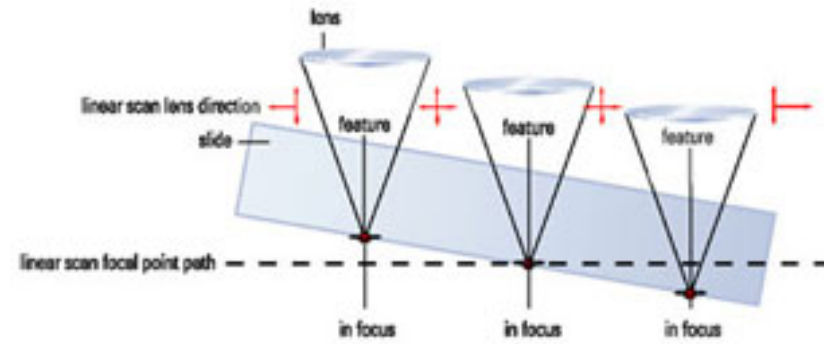
Laser scanning

Agilent autofocus

Single Point Focus
Out of focus, resulting in questionable data.



Agilent Dynamic Autofocus
Always in focus, for results you can trust.



Μέτρηση σήματος στα spots

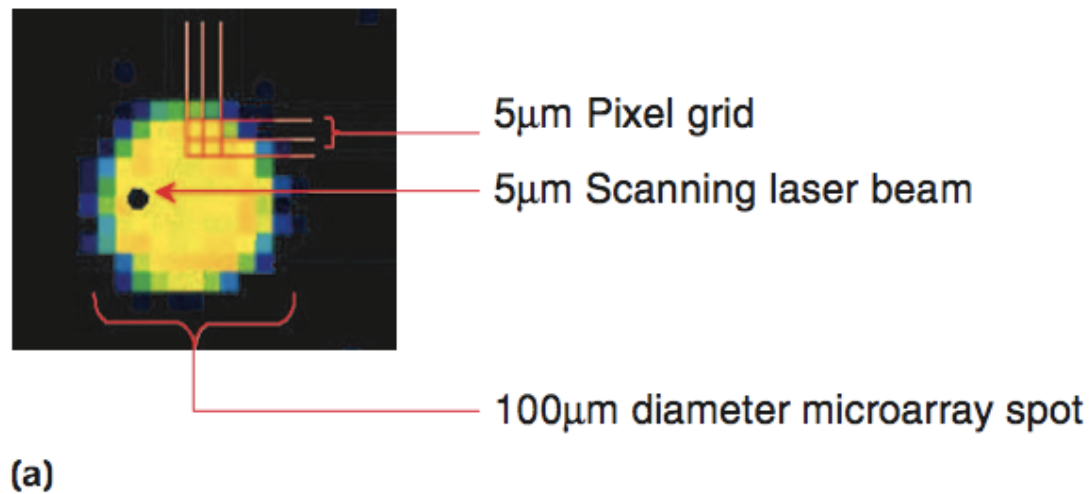


Figure 1.12: The pixels comprising a feature. (a) A false-colour image of the pixels from a single scan of a 100- μm microarray feature. The size of the laser spot is 5 μm . The pixel size has been set to 5 μm so that each pixel represents the area from the size of the laser spot. **(b)** and **(c)** See pp. 9 and 10.

Συνένωση των σημάτων

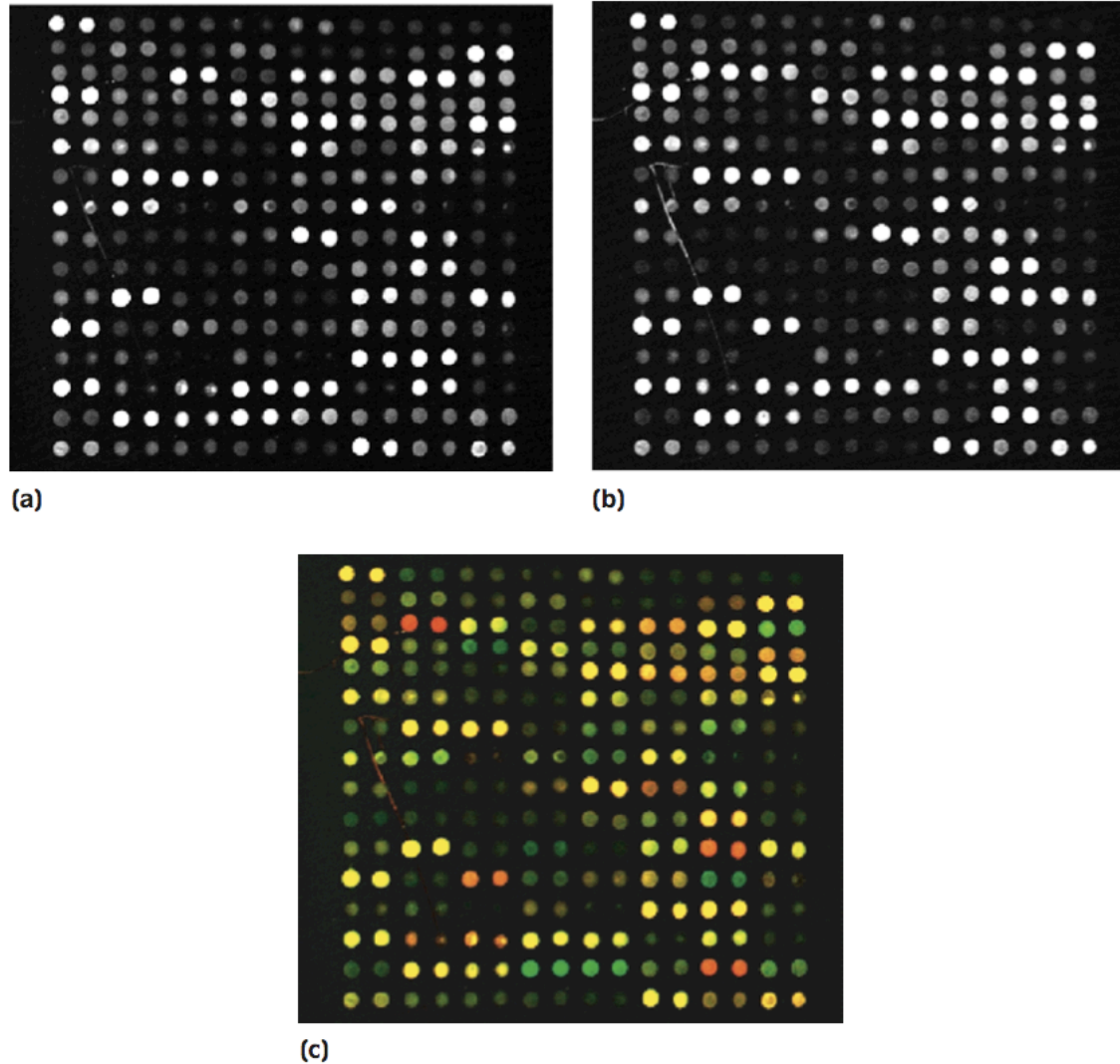


Figure 1.13: Output of scanners. (a) This is the scanner output for a part of a microarray – in this case one of twelve 16×16 blocks of features. This is the monochrome image of the Cy3 (green) channel. (b) The scanner output for the same part of the array but using the Cy5 (red) channel. (c) It is usual to combine the two monochrome images into a composite false-colour image of the array. Green features correspond to features that are expressed more in the sample labelled with Cy3 than the sample labelled with Cy5, and so will be bright in (a) and dark in (b). Similarly, red spots will be bright in (b) and dark in (a). Yellow features have a similar level of expression in both samples. Dark features are low expressed in both samples.

Ποιότητα των spots

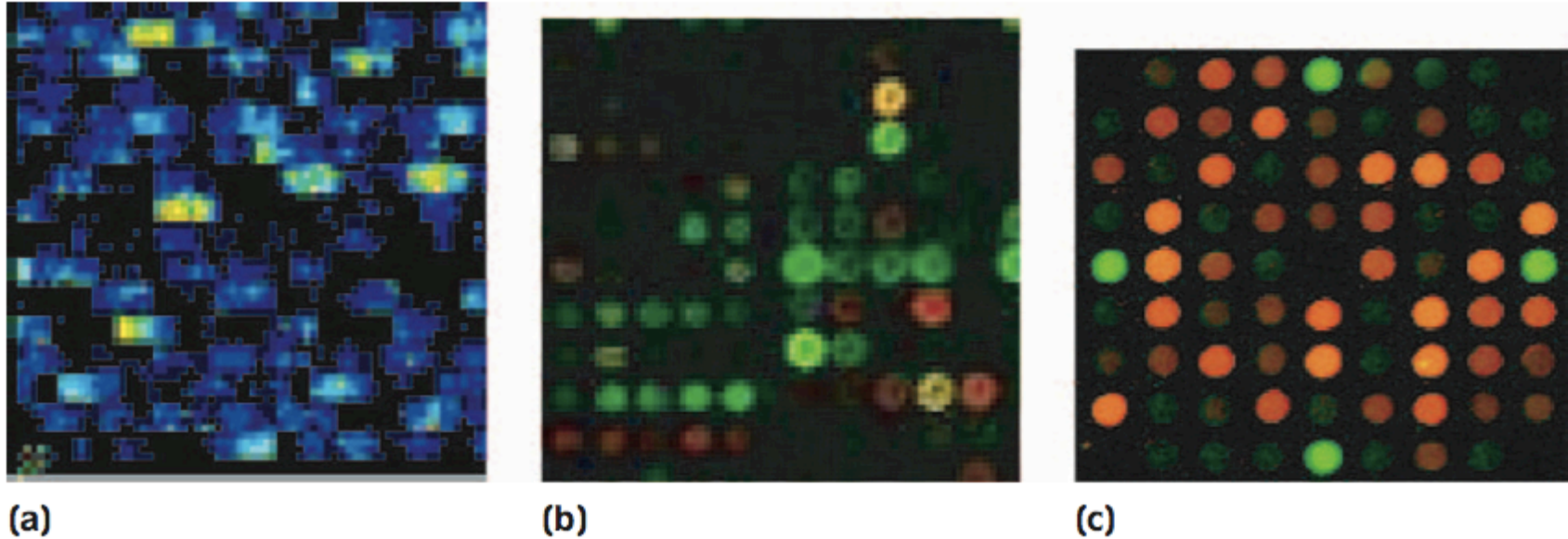
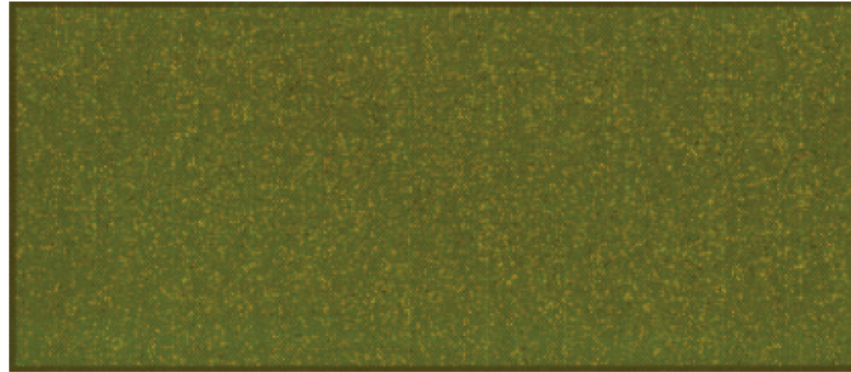
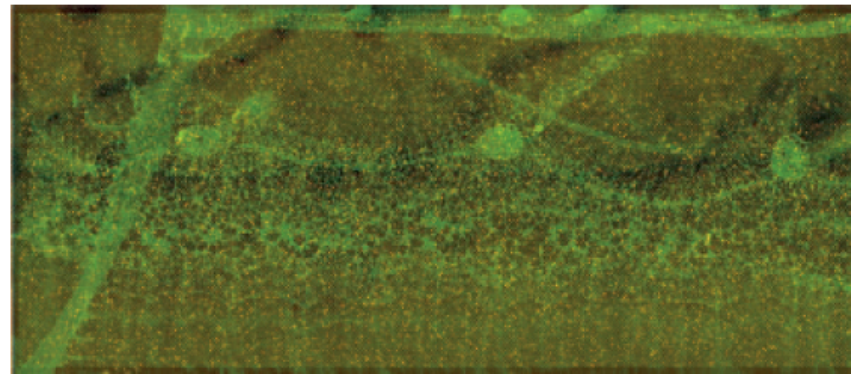


Figure 1.7: Array quality. (a) On Affymetrix arrays the features are rectangular regions. The masks refract light, so there is leakage of signal from one feature to the next. The Affymetrix image-processing software compensates for this by using only the interior portions of the features. (b) Spotted arrays produce spots of variable size and quality. This image shows some of this variation; we cover image processing of spotted arrays in detail in Chapter 4. (c) Inkjet arrays tend to be of the highest quality, with regular, even spots.

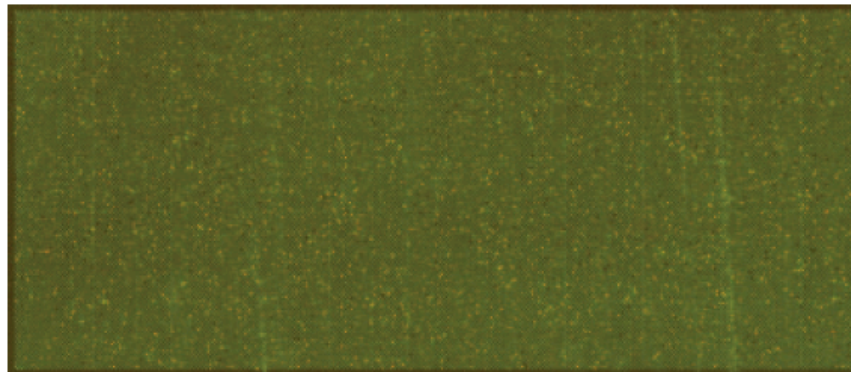
Microarray Quality control



Control, normal microarray



Microarray with wash artifacts



Microarray with degraded total RNA

Ανάλυση προβληματικών Spots

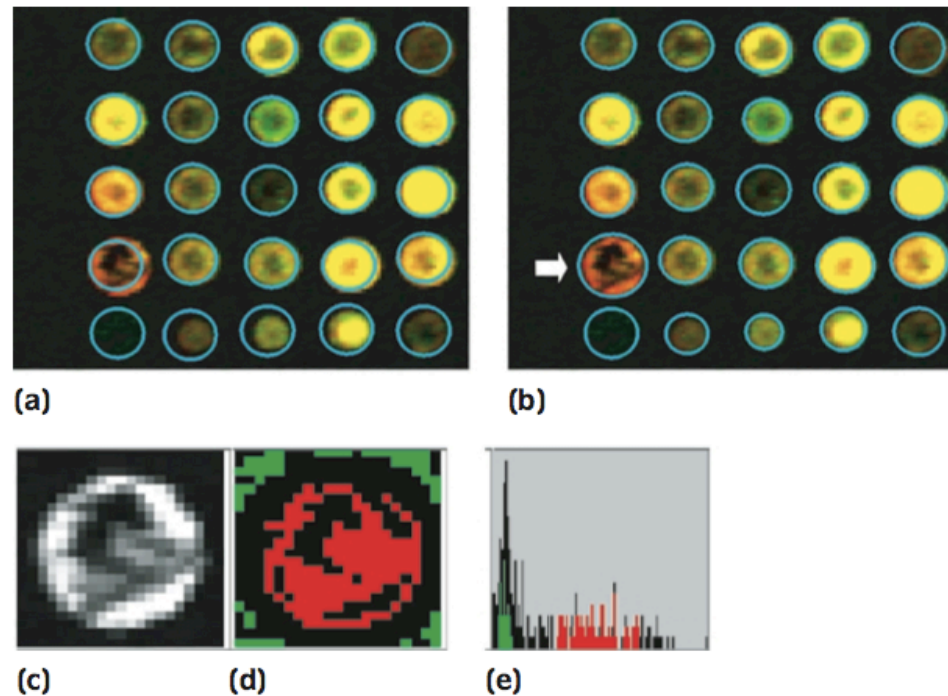


Figure 4.3: (a) Fixed circle segmentation. A circle of the same size is placed on every feature on the array and the pixels inside the circle are used to determine the intensity of the feature. This is not a good method because the circle will be too large for some features and too small for others. (b) Variable circle segmentation. A circle of different size is applied to each feature and the pixels inside the circle are used to determine the intensity of the feature. This performs better on different size features but does not perform so well on features with irregular shapes, for example, the irregular red feature that is marked with an arrow. (c) Zoom in on the red channel of the irregularly shaped feature marked with the arrow in (b). Note the black region where there is no hybridisation, probably because there is no probe attached to the glass in that area. (d) Histogram method applied to that feature. The red pixels are the ones that have been used to calculate the feature signal; the green pixels have been used to calculate the feature background. The black pixels are unused. The area corresponding to the black region in (c) is not used for calculating the feature intensity. The brightest features have also been excluded. The red-to-green ratio of this feature calculated by fixed circle segmentation is 1.8, variable circle segmentation is 1.9, and histogram segmentation is 2.6; so the measured differential gene expression between the samples is different with the different algorithms. Because of the irregular shape of the feature, the histogram method probably gives the most realistic measurement. (e) Histogram of the intensities of the pixels in the irregularly shaped feature. The red bars represent pixels used for the signal intensity; the green bars represent pixels used for the background intensity; the black bars are unused pixels. The brightest and darkest pixels are not used, thus giving a better measurement of hybridisation intensity.

Υπολογισμός & αφαίρεση θορύβου υποβάθρου

IMAGE PROCESSING

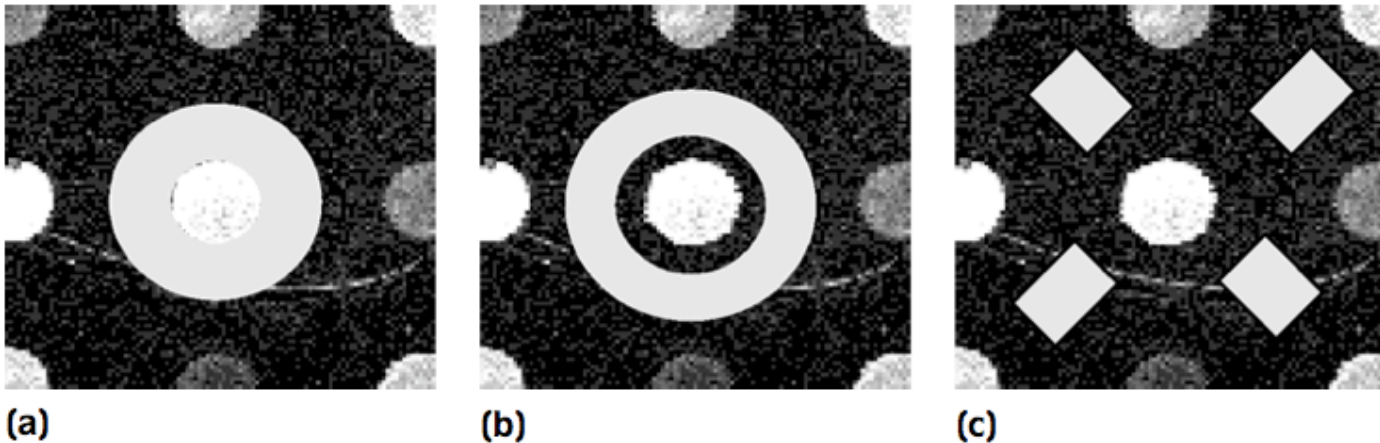


Figure 4.4: Background regions used by different software. Different software packages use different pixel regions surrounding the feature to determine the background intensity. **(a)** ScanAlyze: the region is adjacent to the feature. This will be inaccurate if the feature is larger than the fixed size of the circle used for segmentation. **(b)** ImaGene: there is a space between the feature and the background. This is a better method than (a). **(c)** Spot and GenePix: the background region is in between the features. This is also a good method.

Παράγοντες που επηρεάζουν τα microarrays

- Περιβαλλοντικοί
 - Όζον
 - Υγρασία
 - Υπεριώδης ακτινοβολία
 - Στατικός ηλεκτρισμός
 - Μόλυνση της ατμόσφαιρας (οξειδωτικοί παράγοντες).
 - Σκόνη (μπορεί να επηρεάσει τις μετρήσεις του Laser)
 - Πούδρα από τα γάντια.
- Ενδογενείς
 - Πρόβλημα σε κάποιο βήμα των χημικών αντιδράσεων.
 - Ο τύπος του ιστού.
- Dye swap για να αποκλείσουμε παράγοντες που πιθανών επηρέασαν το πείραμα

Microarrays & Ozone

- Η χρωστική Cy5(red) είναι συνήθως πιο ασταθής σε σχέση με τη χρωστική Cy3(green).
- Το όζον μπορεί να προκαλέσει περισσότερη αστάθεια της Cy5.
- Σε ένα πείραμα που το control χρωματίστηκε με Cy5 και το δείγμα με Cy3, αν υποθέσουμε ότι το όζον προκάλεσε την αποδόμιση του Cy5, τότε περιμένουμε να δούμε (από λάθος) τα γονίδια του δείγματος να:
 - A) υπερεκφράζονται σε σχέση με το control?
 - B) υποεκφράζονται σε σχέση με το control?

Microarrays & ozone

Anal Chem. 2003 Sep 1;75(17):4672-5.

Effects of atmospheric ozone on microarray data quality.

Fare TL, Coffey EM, Dai H, He YD, Kessler DA, Killian KA, Koch JE, LeProust E, Marton MJ, Meyer MR, Stoughton RB, Tokiwa GY, Wang Y.

Rosetta Inpharmatics LLC, 12040 115th Avenue NE, Kirkland, Washington 98034, USA.

Abstract

A data anomaly was observed that affected the uniformity and reproducibility of fluorescent signal across DNA microarrays. Results from experimental sets designed to identify potential causes (from microarray production to array scanning) indicated that the anomaly was linked to a batch process; further work allowed us to localize the effect to the posthybridization array stringency washes. Ozone levels were monitored and highly correlated with the batch effect. Controlled exposures of microarrays to ozone confirmed this factor as the root cause, and we present data that show susceptibility of a class of cyanine dyes (e.g., Cy5, Alexa 647) to ozone levels as low as 5-10 ppb for periods as short as 10-30 s. Other cyanine dyes (e.g., Cy3, Alexa 555) were not significantly affected until higher ozone levels (> 100 ppb). To address this environmental effect, laboratory ozone levels should be kept below 2 ppb (e.g., with filters in HVAC) to achieve high quality microarray data.

PMID: 14632079 [PubMed - indexed for MEDLINE]

Microarrays & ozone

From: [BMC Biotechnol. 2007; 7: 8.](#)

Published online 2007 February 12. doi: 10.1186/1472-6750-7-8.

[Copyright/License](#)

[Permissions](#)

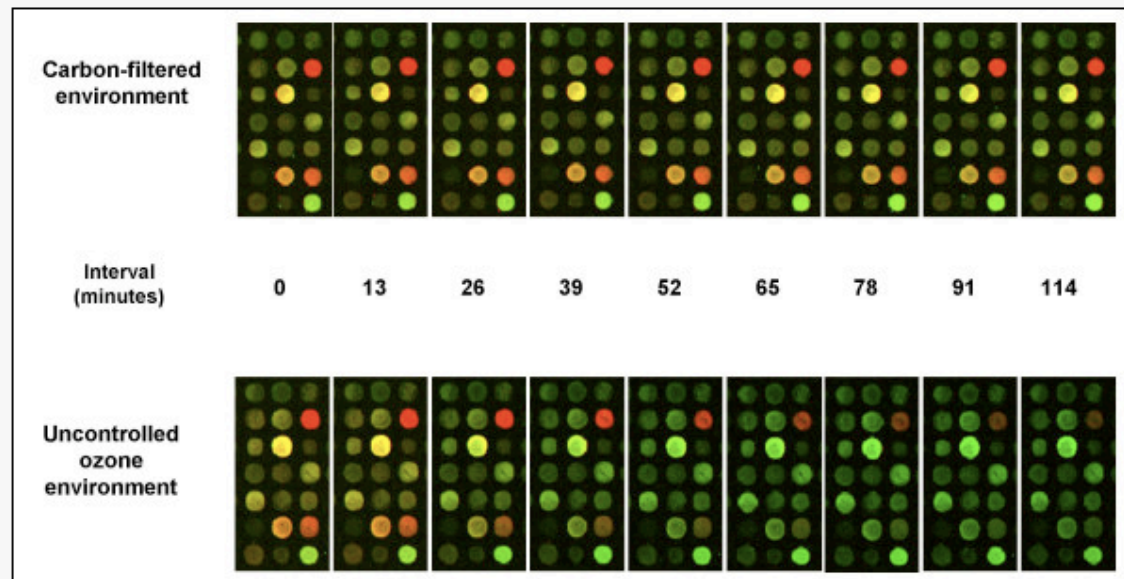


Figure 2

Reduction of Cy5 signal for in-house fabricated microarrays kept in a none-ozone controlled environment. These figures show a region of two 20 K mouse microarrays, selected because of their wide range of both Cy5 and Cy3 signals. The microarray images were from scans made during a 114 minute experiment interval. While the microarrays maintained in the carbon-filtered laboratory environment (ozone ~2–4 ppb) remained relatively unchanged with time (top images), reduction in Cy5 in the uncontrolled ozone environment (ozone ~25 ppb) causes the microarray image to have a predominantly green cast (bottom images).

Ανάλυση δεδομένων

Μετατροπή έντασης σήματος σε \log_2

SECTION 5.2 DATA CLEANING AND TRANSFORMATION

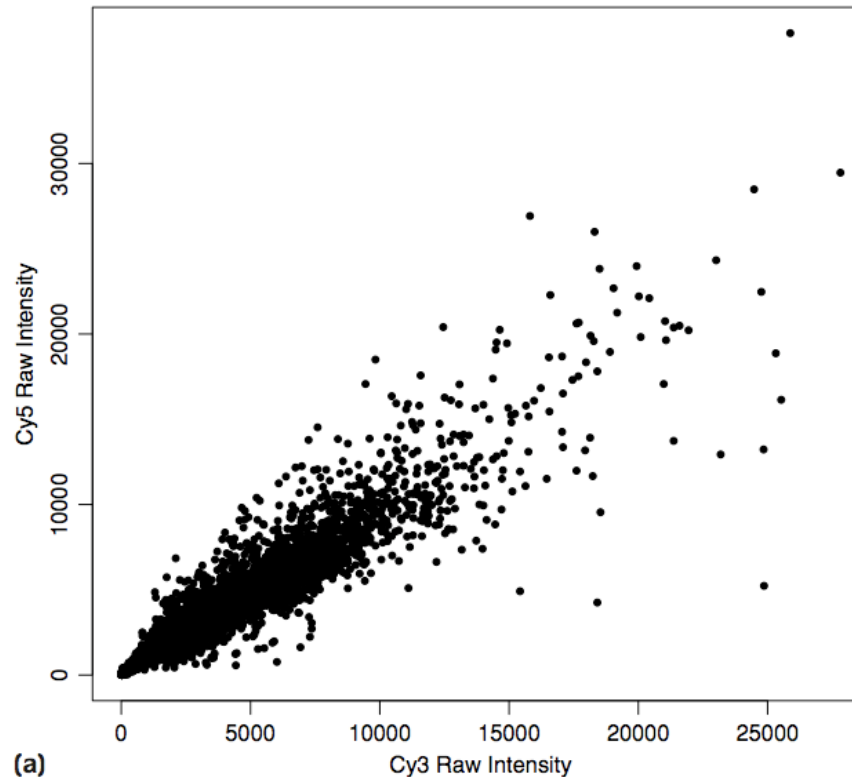
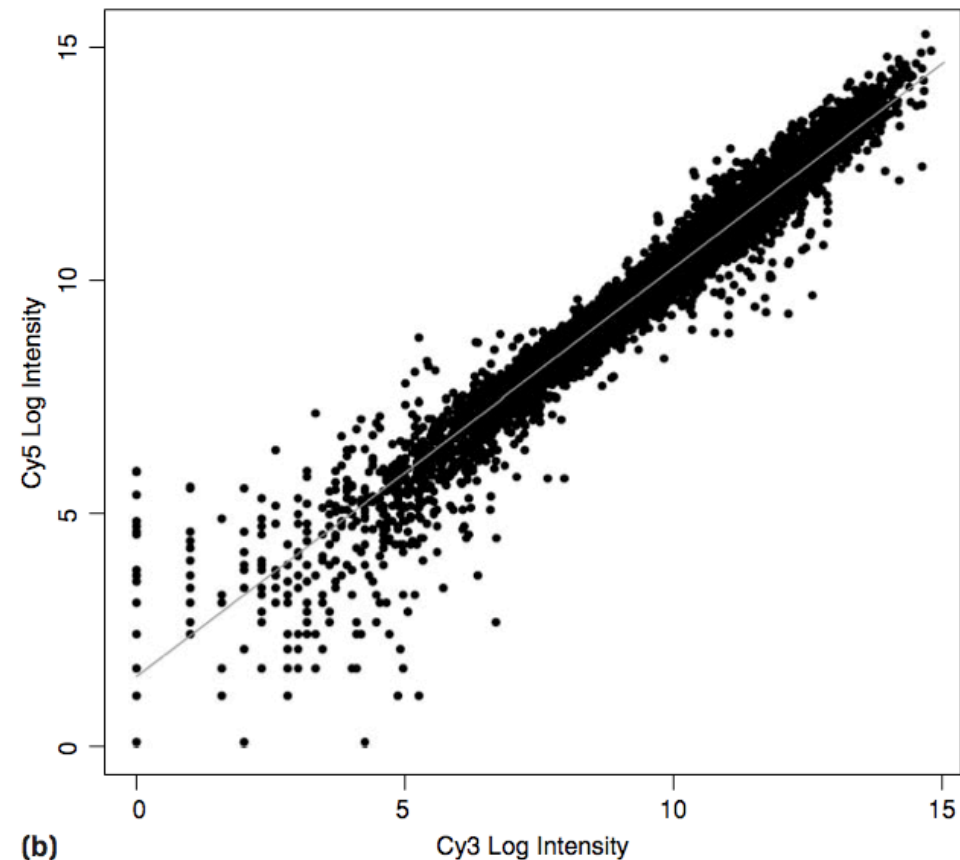


Figure 5.1: Plots of Cy3 vs. Cy5 for data set 5A. Human foreskin fibroblasts have been infected with *Toxoplasma gondii* for a period of 1 hour. A sample has been prepared, labelled with Cy5 (red), and hybridised to a microarray with approximately 23,000 features. The Cy3 (green) channel is a sample prepared from uninfected fibroblasts. Because the infectious period is short, most genes in this experiment are not differentially expressed. **(a)** Scatterplot of the (background-subtracted) raw intensities; each point on the graph represents a feature on the array, with the x coordinate representing the Cy3 intensity, and the y coordinate representing the Cy5 intensity. The graph shows two weaknesses of the raw data that would have a negative impact on further data analysis:

1. Most of the data is bunched in the bottom-left-hand corner, with very little data in the majority of the plot.
2. The variability of the data increases with intensity, so that it is very small when the intensity is small and very large when the intensity is large.

Μετατροπή έντασης σήματος σε \log_2

◇	A	B
1	Cy5 raw intensity	Cy5 log2 transformed intensity
2	1	0
3	2	1
4	4	2
5	8	3
6	16	4
7	32	5
8	64	6
9	128	7
10	256	8
11	512	9
12	1024	10
13	2048	11
14	4096	12
15	8192	13
16	16384	14
17	32768	15



(b) Figure 5.1: (continued)

(b) Scatterplot of the log (to base 2) intensities. This plot is better than (a). The data is spread evenly across the intensity range, and the variability of the data is the same at most intensities. The genes with log intensity less than 5 have slightly higher variability, but these genes are very low expressed and are below the detection level of microarray technology.

The straight line is a linear regression through the data. The linear regression is not perfect (the data appears to bend upwards away from the line at high intensities), but is approximately right. The intercept is 1.4, and the gradient is 0.88. If the two channels were behaving identically, the intercept would be 0 and the gradient would be 1. We conclude that the two Cy dyes behave differently at different intensities; this could result from differential dye incorporation or different responses of the dyes to the lasers.

Μετατροπή έντασης σήματος σε \log_2

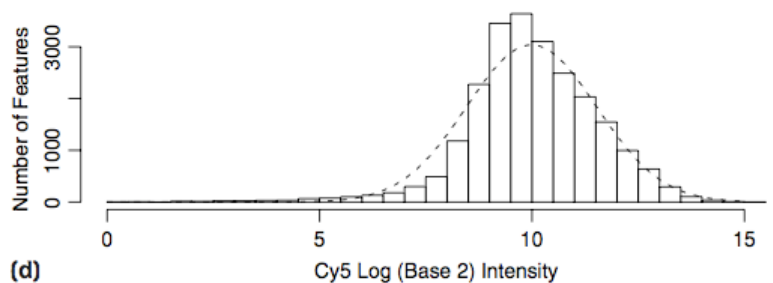
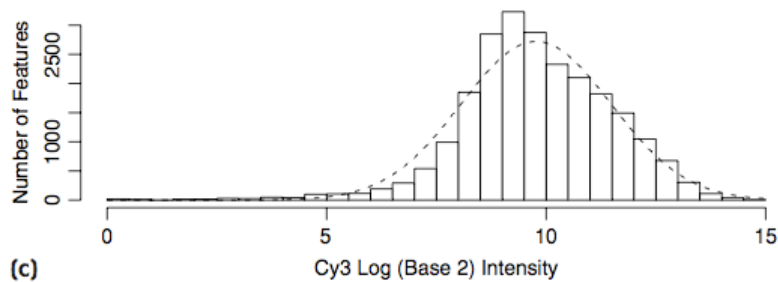
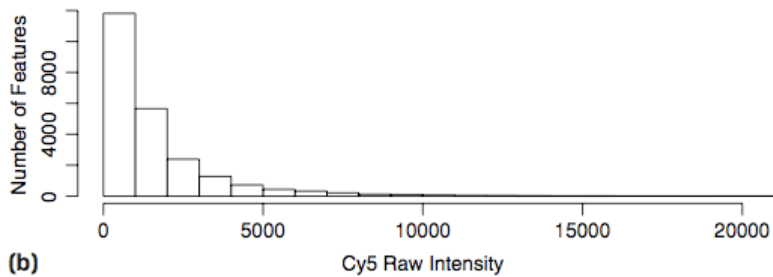
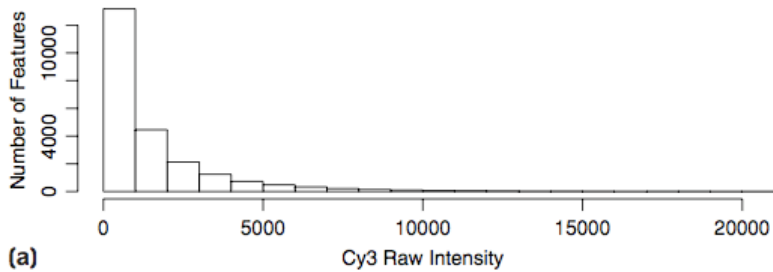


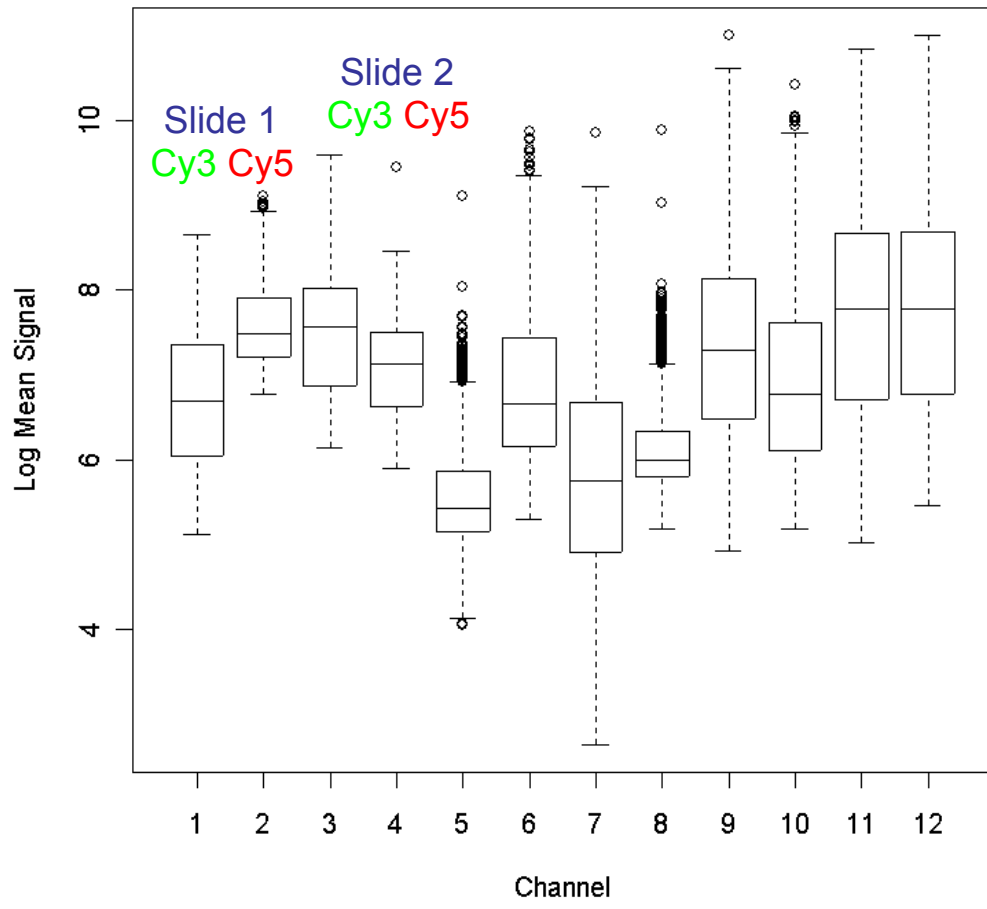
Figure 5.2: Histograms of the raw and log Cy3 and Cy5 intensities. Histograms of the intensities of the features for the human fibroblast data. **(a)** The raw intensities for the Cy3 channel; the data is right-skewed, with the majority of features having low intensity and decreasing numbers of features having higher intensity. **(b)** The raw intensities for the Cy5 channel; the pattern is the same as (a). **(c)** The log intensities for the Cy3 channel; the intensities are closer to a bell-shaped normal curve (shown as a dashed line). There is still a slight right skew, but the logged data is better for data analysis than the raw data. **(d)** The log intensities for the Cy5 channel, along with a normal curve (dashed line). As with (c), the intensities are approximately normal, with a slight right skew.

Κανονικοποίηση δεδομένων

- Η κανονικοποίηση προσπαθεί να ελαχιστοποιήσει/εξαλείψει τον πειραματικό συστηματικό θόρυβο/διακύμανση, ώστε να αποκαλυφθούν οι διαφορές που έχουν βιολογική βάση.
- Η κανονικοποίηση αυτού του είδους δεν πρέπει να συγχέεται με την κανονικοποίηση στις κανονικές κατανομές.

Πηγές Πειραματικών συστηματικών λαθών

- Διακυμάνσεις μεταξύ πειραματικών επαναλήψεων
 - Slides
 - Συνθήκες υβριδισμού
 - Scanning
 - Διαφορετικοί άνθρωποι μπορεί να δουλεύουν με διαφορετικά slides
- Χρωστικές
 - Το φως και η ζέστη μπορεί να επηρεάσουν με διαφορετικό τρόπο την αποτελεσματικότητα ενσωμάτωσης της κάθε χρωστικής.
 - Διαφορές στην συνολική ποσότητα επισημασμένου cDNA μεταξύ των δύο καναλιών/χρωστικών.



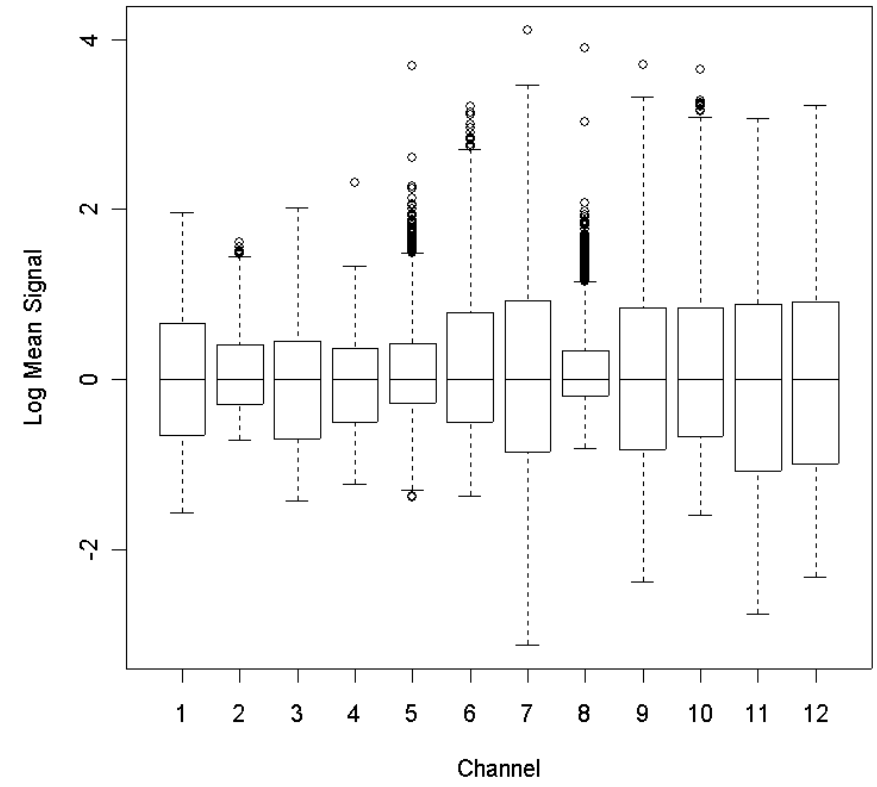
← maximum

← Q3=75th percentile

← median

← Q1=25th percentile

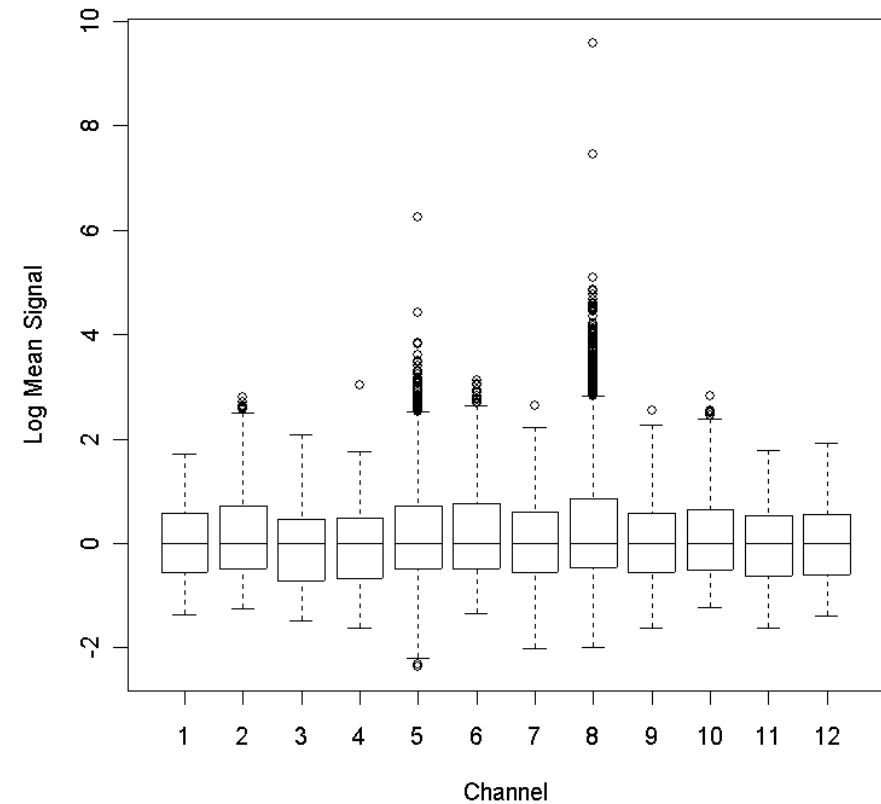
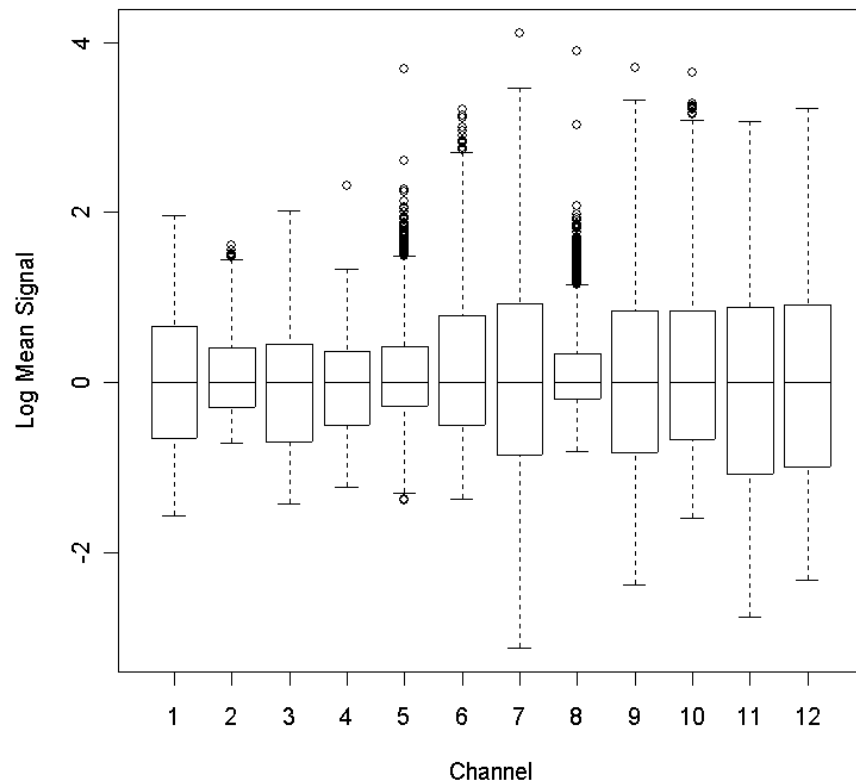
← minimum



Κανονικοποίηση κλίμακας

Scale normalization

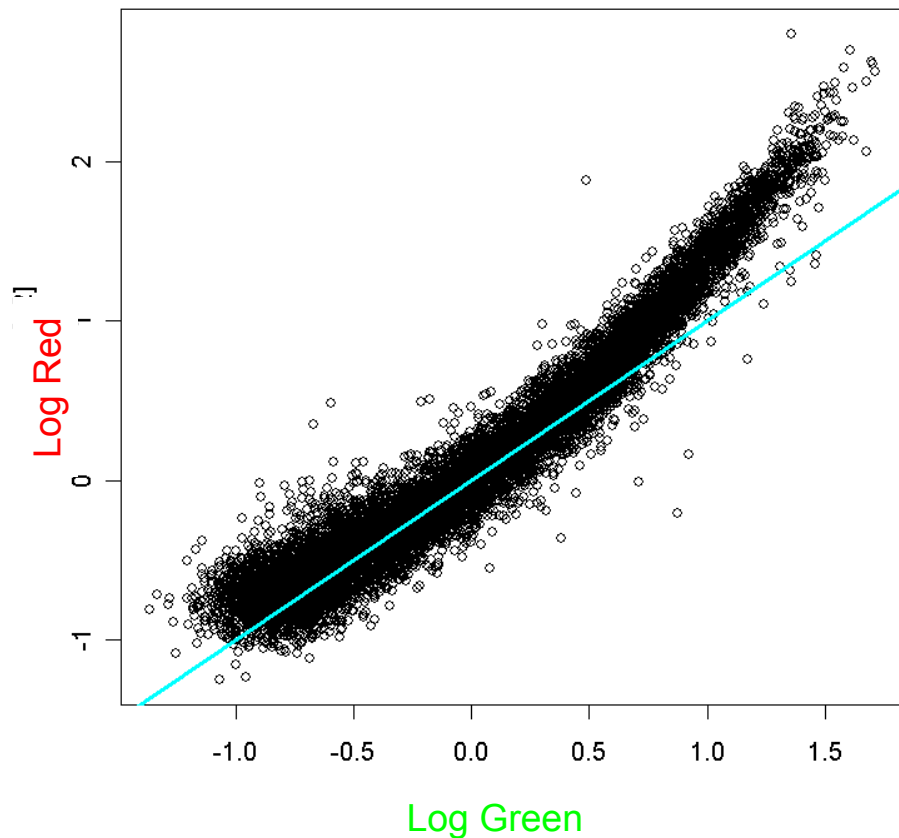
Data after Median Centering and Scale Normalizing



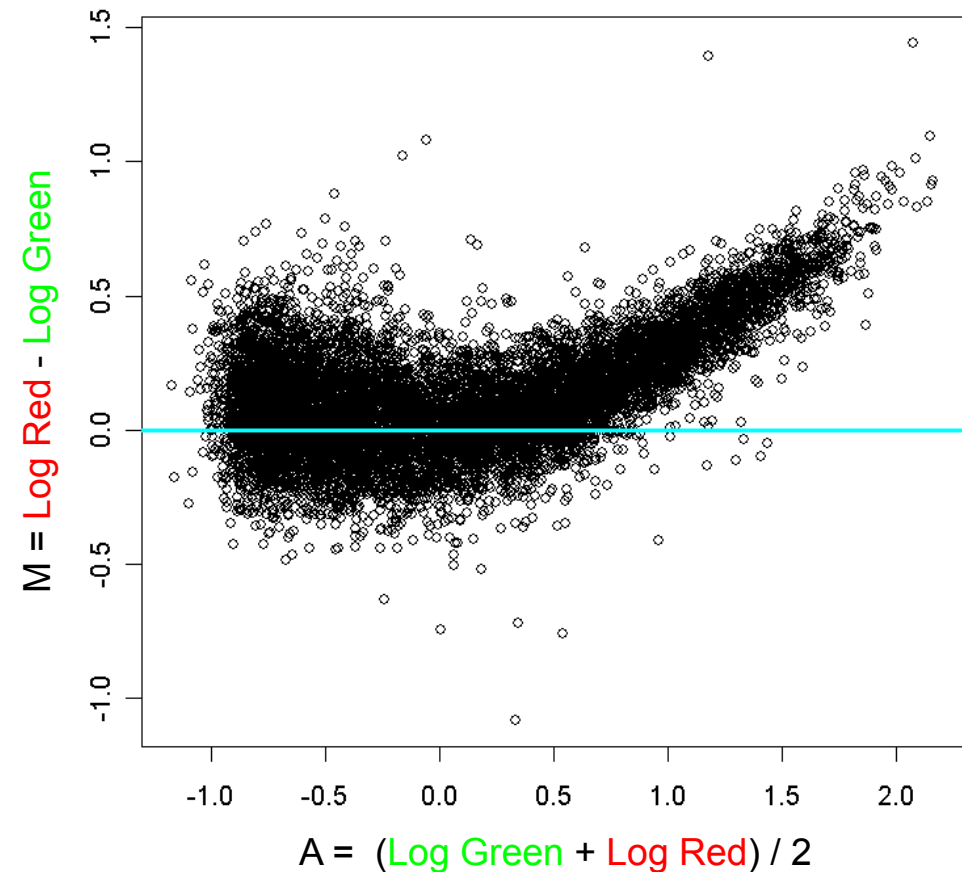
Συστηματικό λάθος στην χρώση

Στο παράδειγμα, η πράσινη χρωστική δεν δουλεύει σωστά για μεγάλες τιμές έντασης

Slide 1 Log Signal Means after Median Centering and Scaling All Channels

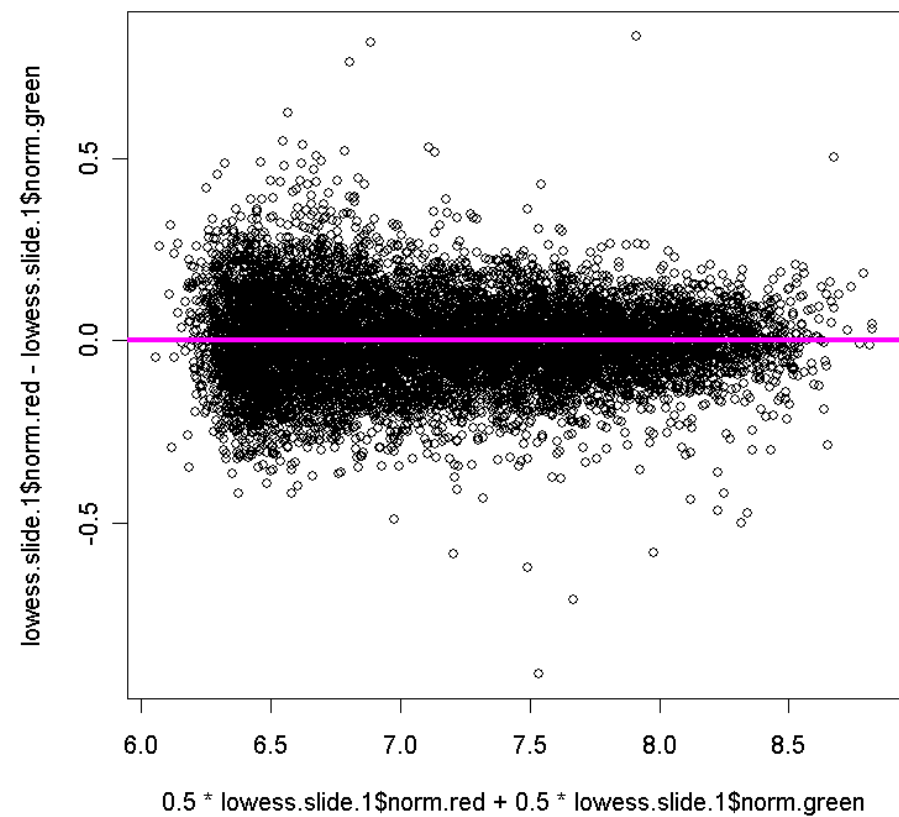
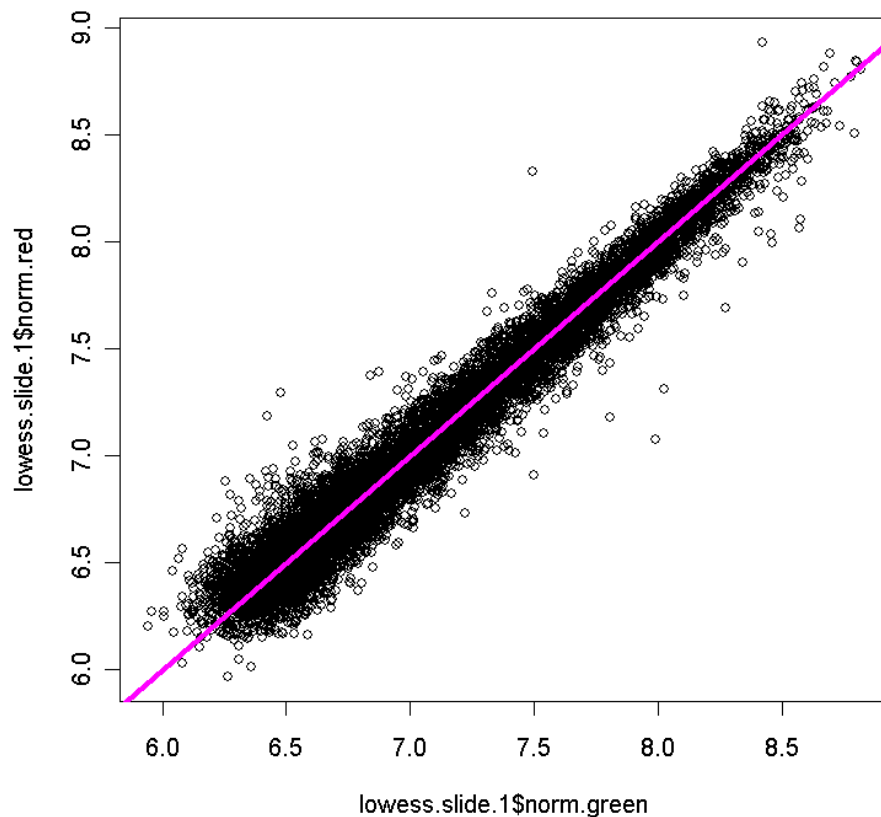


M vs. A Plot of the Logged, Centered, and Scaled Slide 1 Data

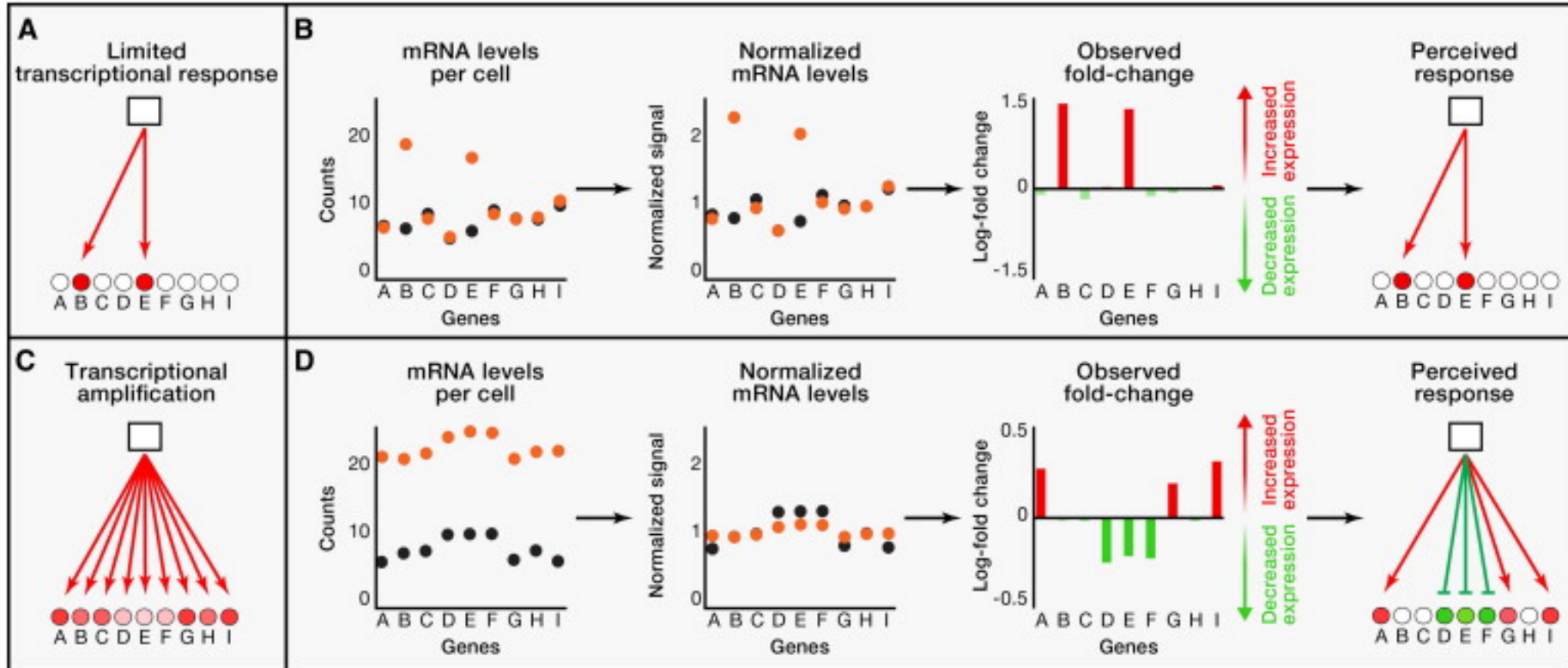


Κανονικοποίηση με Lowess

Lowess για κανονικοποίηση συστηματικών λαθών που σχετίζονται με την ένταση της χρωστικής
LOcally WEighted polynomial regreSSion.
(2002. *Nucleic Acids Research*, **30**, 4 e15)



Gene expression normalization



Transcriptional amplification
Normalization and Interpretation of Expression Data.

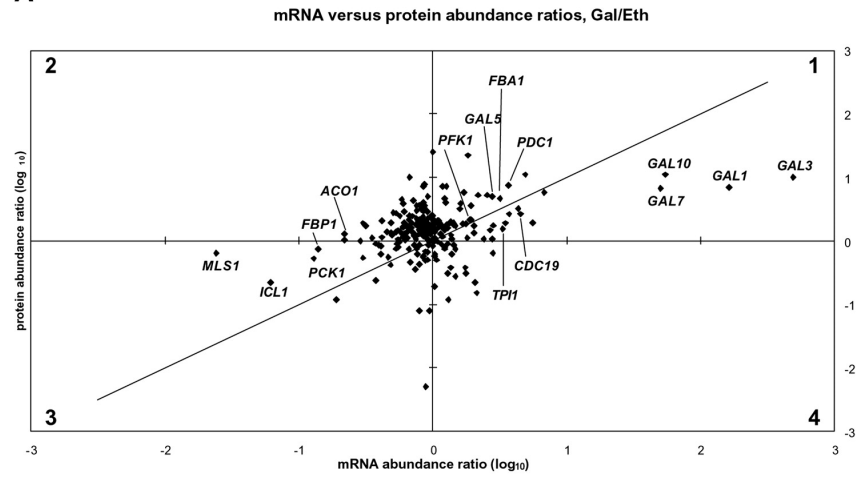
[Cell , Volume 151, Issue 3, Pages 476-482](#)
Revisiting Global Gene Expression Analysis

Κατάθεση των δεδομένων

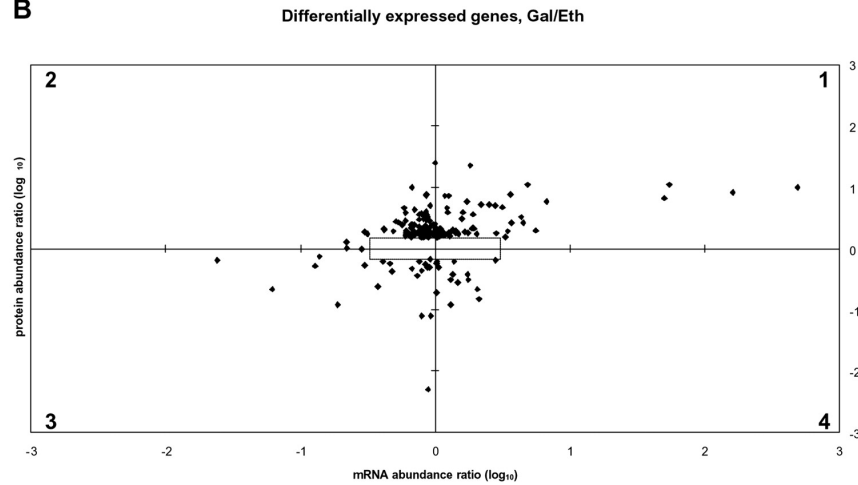
- MIAME (minimum information about a microarray experiment) (αναπτύχθηκε στο EBI).
 - Πληροφορίες σχετικά με το πως έγινε το πείραμα, τι πλατφόρμα/ συνθήκες χρησιμοποιήθηκαν.
- Πολλά περιοδικά απαιτούν πριν την δημοσίευση να έχουν κατατεθεί τα πειραματικά δεδομένα σε μια Β.Δ. με ελεύθερη πρόσβαση.
- Βάσεις δεδομένων
 - ArrayExpress (EBI)
 - Gene expression omnibus (GEO) (USA)
 - Center for information biology gene expression database (CIBEX) (Ιαπωνία)

mRNA abundance ratios versus protein-abundance ratios.

A



B



Griffin T J et al. Mol Cell Proteomics 2002;1:323-333



Οντολογίες

- www.geneontology.org
- Ελεγχόμενο λεξιλόγιο για την περιγραφή των ιδιοτήτων των γονιδίων και των πρωτεϊνών.
- Περιγράφουν:
 - Μοριακές λειτουργίες του βιομορίου (1 ή περισσότερες).
 - Βιολογικές διαδικασίες στις οποίες εμπλέκεται το βιομόριο (1 ή περισσότερες).
 - Κυτταρικό διαμέρισμα στο οποίο συναντάται το βιομόριο (1 ή περισσότερα).

Gene ontology

REVIEWS

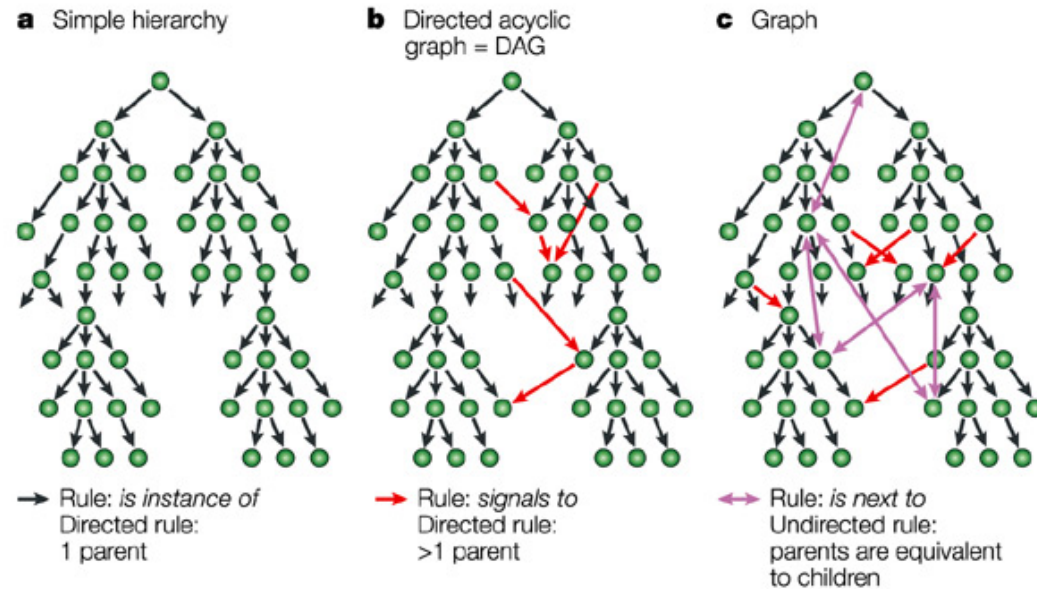
Use and misuse of the gene ontology annotations

*Seung Yon Rhee**, *Valerie Wood*[‡], *Kara Dolinski*[§] and *Sorin Draghici*^{||}

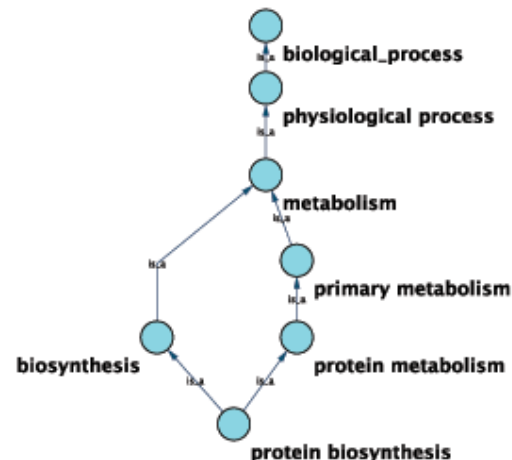
Abstract | The Gene Ontology (GO) project is a collaboration among model organism databases to describe gene products from all organisms using a consistent and computable language. GO produces sets of explicitly defined, structured vocabularies that describe biological processes, molecular functions and cellular components of gene products in both a computer- and human-readable manner. Here we describe key aspects of GO, which, when overlooked, can cause erroneous results, and address how these pitfalls can be avoided.

Οντολογίες: Η δομή τους

- Δείχνει τις σχέσεις μεταξύ των διαφορετικών όρων.
- Ένας όρος μπορεί να αποτελεί πιο εξειδικευμένη περιγραφή ενός άλλου όρου.
- Είναι κατευθυνόμενα ακυκλικά γραφήματα (DAG).
- Παρόμοια με ιεραρχίες.
- Η διαφορά είναι ότι ένας κόμβος-απόγονος μπορεί να έχει περισσότερους από έναν προγόνους.

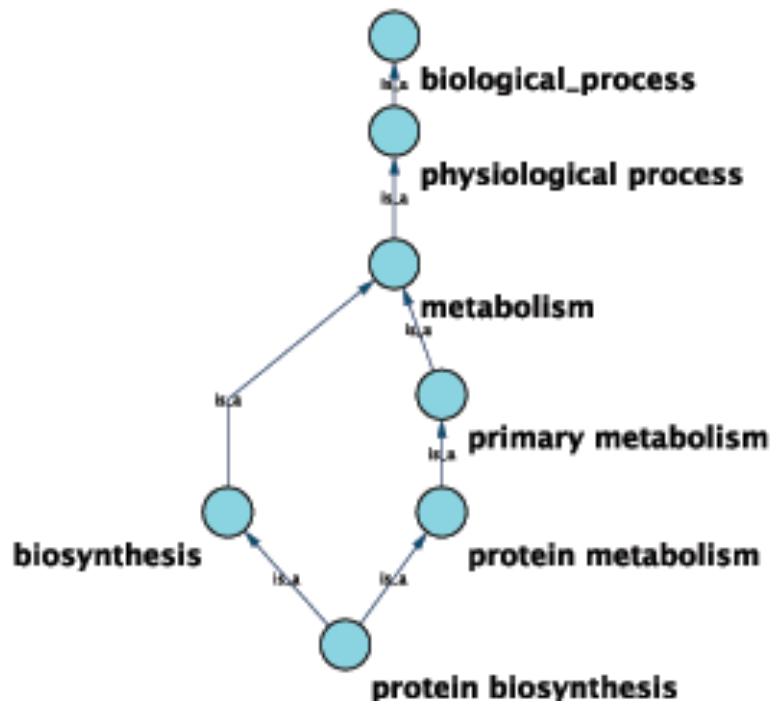


Nature Reviews | **Genetics**



Οντολογίες: Η δομή τους

- Θεωρούμε ότι αν σε ένα βιομόριο αντιστοιχεί ένα όρος-οντολογία, τότε σε αυτό το βιομόριο ανήκουν και όλοι οι πρόγονοι του όρου-οντολογίας.



Gene ontology

REVIEWS

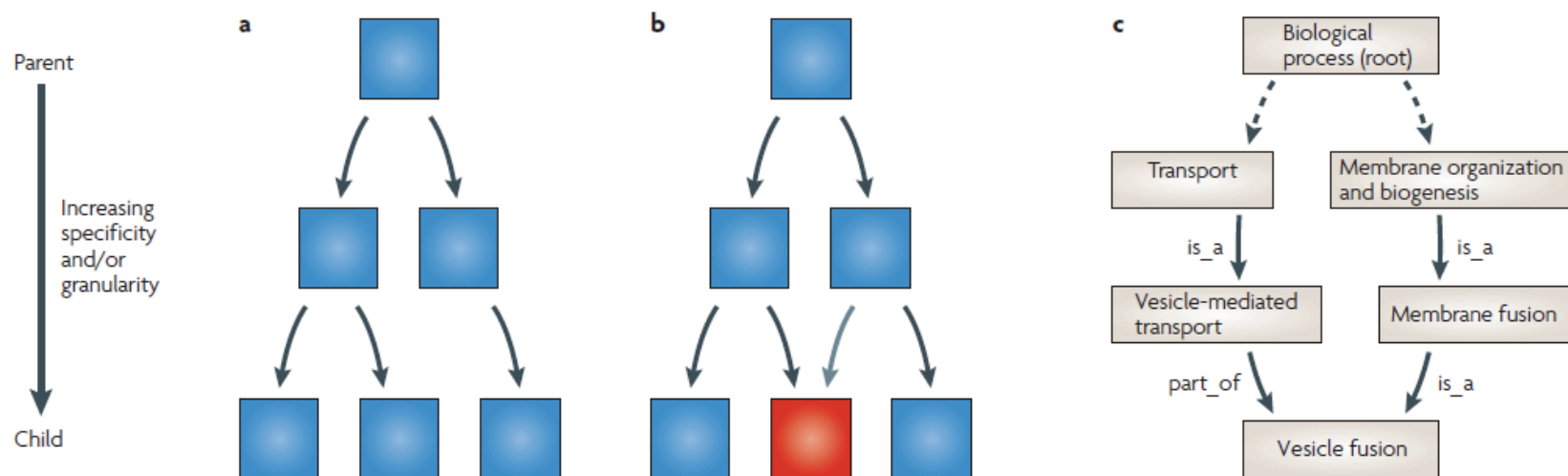


Figure 1 | Simple trees versus directed acyclic graphs. Boxes represent nodes and arrows represent edges. **a** | An example of a simple tree, in which each child has only one parent and the edges are directed, that is, there is a source (parent) and a destination (child) for each edge. **b** | A directed acyclic graph (DAG), in which each child can have one or more parents. The node with multiple parents is coloured red and the additional edge is coloured grey. **c** | An example of a node, vesicle fusion, in the biological process ontology with multiple parentage. The dashed edges indicate that there are other nodes not shown between the nodes and the root node (biological process). A root is a node with no incoming edges, and at least one leaf (also called a sink). A leaf node is a node with no outgoing edges, that is, a terminal node with no children (vesicle fusion). Similar to a simple tree, A DAG has directed edges and does not have cycles, that is, no path starts and ends at the same node, and will always have at least one root node. The depth of a node is the length of the longest path from the root to that node, whereas the height is the length of the longest path from that node to a leaf⁴¹. *is_a* and *part_of* are types of relationships that link the terms in the GO ontology. More information about the relationships between GO terms are found online ([An Introduction to the Gene Ontology](#)).

Gene ontology

Table 1 | Evidence codes used by GO

Evidence code	Evidence code description	Source of evidence	Manually checked	Current number of annotations*
IDA	Inferred from direct assay	Experimental	Yes	71,050
IEP	Inferred from expression pattern	Experimental	Yes	4,598
IGI	Inferred from genetic interaction	Experimental	Yes	8,311
IMP	Inferred from mutant phenotype	Experimental	Yes	61,549
IPI	Inferred from physical interaction	Experimental	Yes	17,043
ISS	Inferred from sequence or structural similarity	Computational	Yes	196,643
RCA	Inferred from reviewed computational analysis	Computational	Yes	103,792
IGC	Inferred from genomic context	Computational	Yes	4
IEA	Inferred from electronic annotation	Computational	No	15,687,382
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes	5,167
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes	44,564
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes	25,656
ND	No biological data available	No information available	Yes	132,192
NR	Not recorded	Unknown	Yes	1,185

*October 2007 release

Gene ontology

Table 2 | **Distribution of gene ontology (GO) annotations for species with more than 5,000 annotations**

Species (NCBI taxon ID)	Genes* with experimental annotations [‡]	Total annotated genes*	Percentage of genes* with at least one experimental annotation	Total genes*	Percentage annotated [§]	Percentage known in genome
<i>Schizosaccharomyces pombe</i> (4896)	4,482	4,930	90.9%	4,930	100%	90.9%
<i>Saccharomyces cerevisiae</i> (4932)	4,947	5,794	85.4%	5,794	100%	85.4%
Mouse (10090)	10,621	18,386	57.8%	27,289	67.4%	38.9%
<i>Caenorhabditis elegans</i> (6239)	4,614	14,154	32.6%	20,163	70.2%	22.9%
Human [¶] (9606)	4,780	17,021	28.1%	20,887	81.5%	22.9%
<i>Arabidopsis thaliana</i> [#] (3702)	5,530	26,637	20.8%	27,029	98.5%	20.5%
Rat (10116)	3,566	17,243	20.7%	17,993	95.8%	19.8%
Fruitfly (7227)**	2,790	9,563	29.2%	14,141	67.6%	19.7%
<i>Candida albicans</i> (5476)	806	3,756	21.4%	6,166	60.9%	13.0%
<i>Pseudomonas aeruginosa</i> PAO1 (208964)	491	2,506	19.6%	5,568	45.0%	8.82%
Slime mold (44689)	797	6,892	11.6%	13,625	50.6%	5.9%
<i>Trypanosoma brucei</i> (5691)	449	3,914	11.5%	9,154	42.8%	4.92%
Zebrafish (7955)	1,235	13,574	5.8%	21,322	63.7%	3.7%
<i>Plasmodium falciparum</i> (5833)	188	3,243	5.8%	5,420	59.8%	3.47%
Rice (39947)	654	29,877	2.2%	41,908	71.3%	1.57%
Chicken [¶] (9031)	75	6,063	1.2%	16,737	36.2%	0.4%
Cow [¶] (9913)	96	8,536	1.1%	21,756	39.2%	0.4%

*Total genes in genomes include only those that encode proteins. These numbers were obtained from the databases that contribute annotations to GO and are listed on the GO annotations download page (<http://www.geneontology.org/GO.current.annotations.shtml>). [‡]Experimental annotations include those only with the following evidence codes: IDA (inferred from direct assay), IEP (inferred from expression pattern), IGI (inferred from genetic interaction), IMP (inferred from mutant phenotype) and IPI (inferred from physical interaction). [§]Percentage annotated is determined by dividing the number of genes annotated by total genes. ^{||}Percentage known in genome is determined by multiplying the percentage of experimentally derived annotations by the percentage of the genome annotated. This is an approximation of the extent of knowledge about the portion of the genome that encodes proteins in an organism with a complete genome sequence that is captured by annotation. [¶]Numbers are from the GO annotation project at the European Bioinformatics Institute, human data last updated 14 September 2007, cow data last updated 17 January 2007, chicken data last updated 10 July 2007. [#]Numbers are from The *Arabidopsis* Information Resource (TAIR), last updated 14 December 2007. ^{**}Numbers are based on release 5.4 of the *Drosophila melanogaster* genome and GO annotations from FlyBase release FB2007_03 (dated 11 January 2007). NCBI, National Center for Biotechnology Information.

Gene ontology

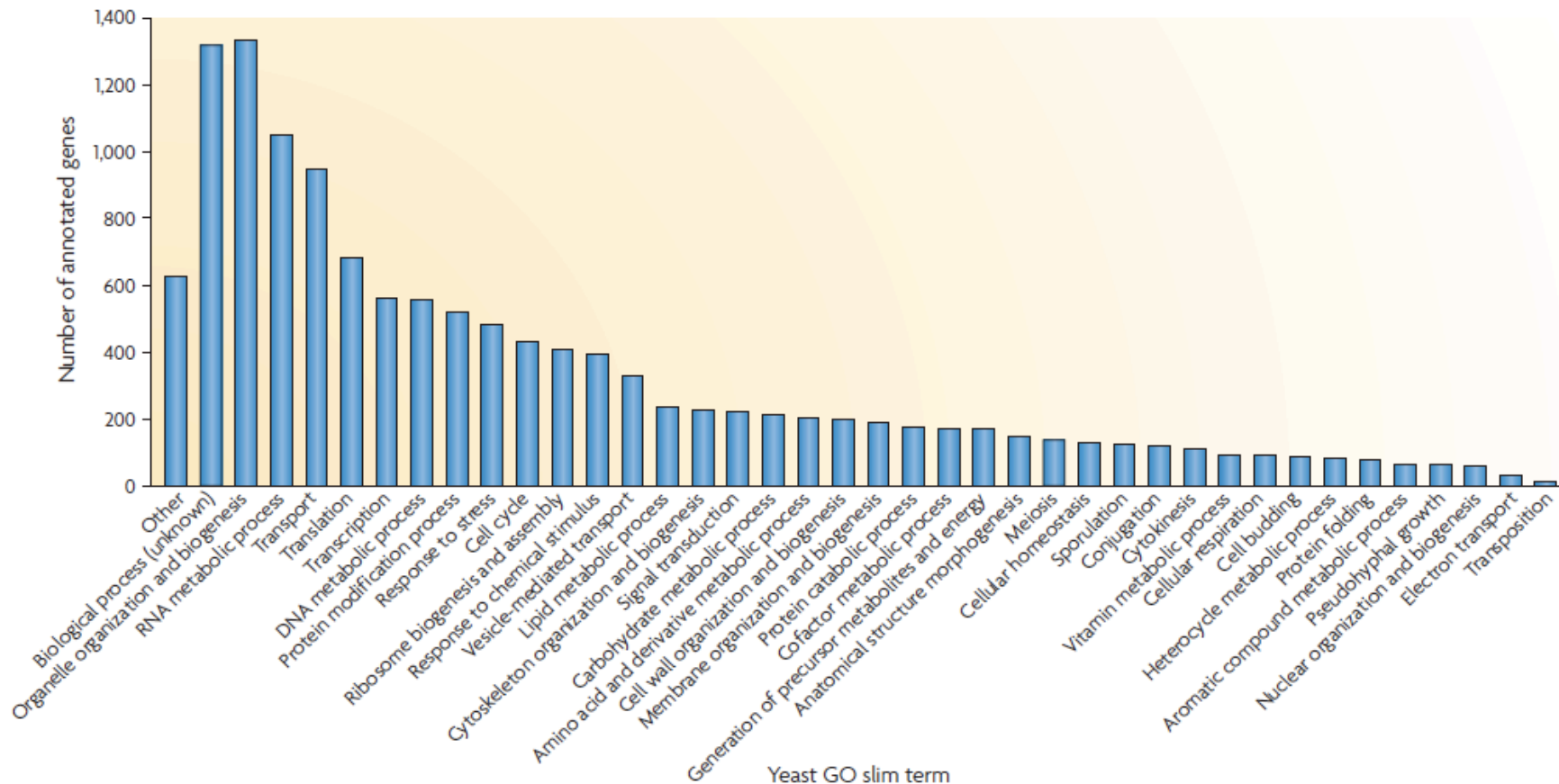


Figure 2 | Using gene ontology (GO) to bin the yeast genome into broad biological process categories. This example was generated by downloading the go_slim_mapping.tab file from the *Saccharomyces* genome database ftp site (dated 19 January 2008). This file maps every gene in the yeast genome to the yeast GO slim ontology available from the GO website. The number of genes (6,200 in total, including RNAs but excluding 'dubious' genes) annotated to a particular term in the yeast GO slim ontology is indicated on the graph. Dubious genes are those that were originally predicted to exist, on the basis of ORF length, but that are now thought to be unlikely to encode an expressed protein, on the basis of functional and comparative genomics data. The 'other' term is used when genes are annotated to terms other than those included in the GO slim ontology, and the 'biological process' term, the root node in the biological process ontology, indicates that genes annotated to it are not yet characterized. Note that because genes can be binned to more than one category, there are more annotations (13,074) than total genes (6,200) with annotations.

Οντολογίες: στατιστική ανάλυση

- Παράδειγμα:
 - 1 γονιδίωμα με 10.000 γονίδια.
 - 1.000 γονίδια εμπλέκονται στον κυτταρικό κύκλο (GO_term: cell-cycle). (10% του γονιδιώματος).
 - Αν επιλέξουμε τυχαία έναν αριθμό X γονιδίων, θα περιμέναμε (από τύχη) περίπου το 10% (με κάποιες διακυμάνσεις) να έχουν τον όρο “κυτταρικός κύκλος”.
 - Η τυχαία διακύμανση εξαρτάται από τον αριθμό των γονιδίων.
 - Έστω ότι με τα microarrays σε ένα πείραμα βρήκαμε ότι X αριθμός γονιδίων υπερεκφράζονται.
 - Σε αυτό τον X αριθμό, βρήκαμε ότι 20% των γονιδίων ανήκουν στον κυτταρικό κύκλο.
 - Αυτή η απόκλιση (20% παρατηρούμενο - 10% αναμενόμενο) είναι στα όρια των τυχαίων διακυμάνσεων, ή είναι στατιστικά σημαντική?
 - Στατιστικά σημαντική, σημαίνει ότι τα υπερεκφρασμένα γονίδια είναι εμπλουτισμένα για την κατηγορία “κυτταρικός κύκλος”. Δηλαδή, ο κυτταρικός κύκλος εμπλέκεται στην διαδικασία που μελετάμε.

Οντολογίες: στατιστική ανάλυση

- Η στατιστική ανάλυση γίνεται με το υπεργεωμετρικό τεστ.
- Παίρνουμε ένα p-value.
- Αν $p\text{-value} < 0.05$, τότε είναι στατιστικά σημαντικό.

- Αν στις οντολογίες μας είχαμε 100 όρους, θα επαναλαμβάναμε τα παραπάνω τεστ για τον κάθε όρο.
- Όμως, όσο περισσότερα τεστ κάνουμε για το πείραμά μας, τόσο αυξάνει ή πιθανότητα να βρούμε κάτι στατιστικά σημαντικό ($p\text{-value} < 0.05$) καθαρά από λάθος.
- Άρα, πρέπει να λάβουμε υπόψην μας πόσα τεστ διενεργούμε και να διορθώσουμε τα p-values (multiple testing correction).
 - False discovery rate (Benjamini-Hochberger)
 - Bonferroni correction

Μοριακά μονοπάτια και πρωτεϊνικές αλληλεπιδράσεις

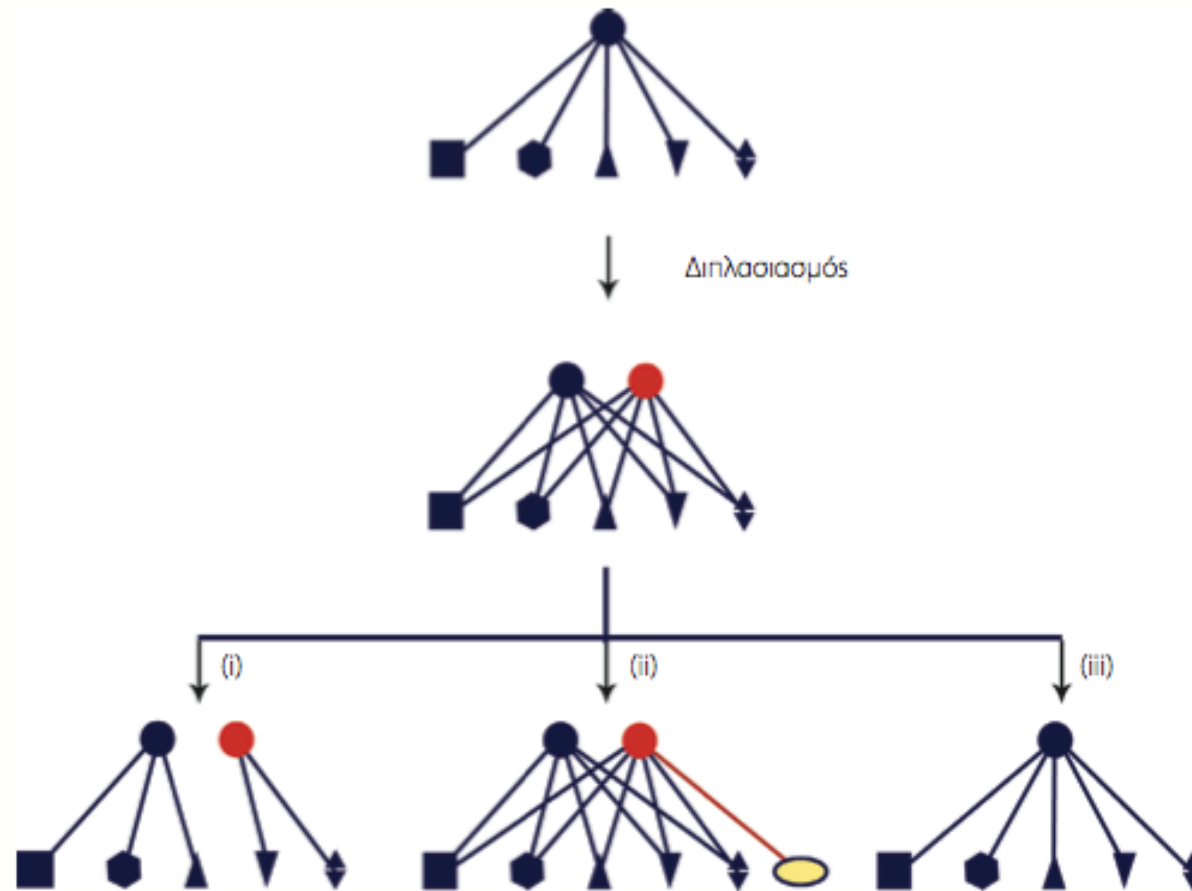
- Οι πρωτεΐνες αλληλεπιδρούν μεταξύ τους και δημιουργούν σύμπλοκα
 - Πρόσκαιρα (π.χ. Μεταγωγή σήματος/κινάσες)
 - Σταθερά

Τα σύμπλοκα μπορεί να είναι διμερή ή πολυμερή.

Οι πρωτεϊνικές αλληλεπιδράσεις εξελίσσονται, μέσω των γονιδιακών διπλασιασμών και των μεταλλάξεων στις επιφάνειες αλληλεπίδρασης.

Οι πρωτεΐνες δημιουργούν δίκτυα πρωτεϊνικών αλληλεπιδράσεων, όπου οι κόμβοι συμβολίζουν τις πρωτεΐνες και οι συνδέσεις τις αλληλεπιδράσεις.

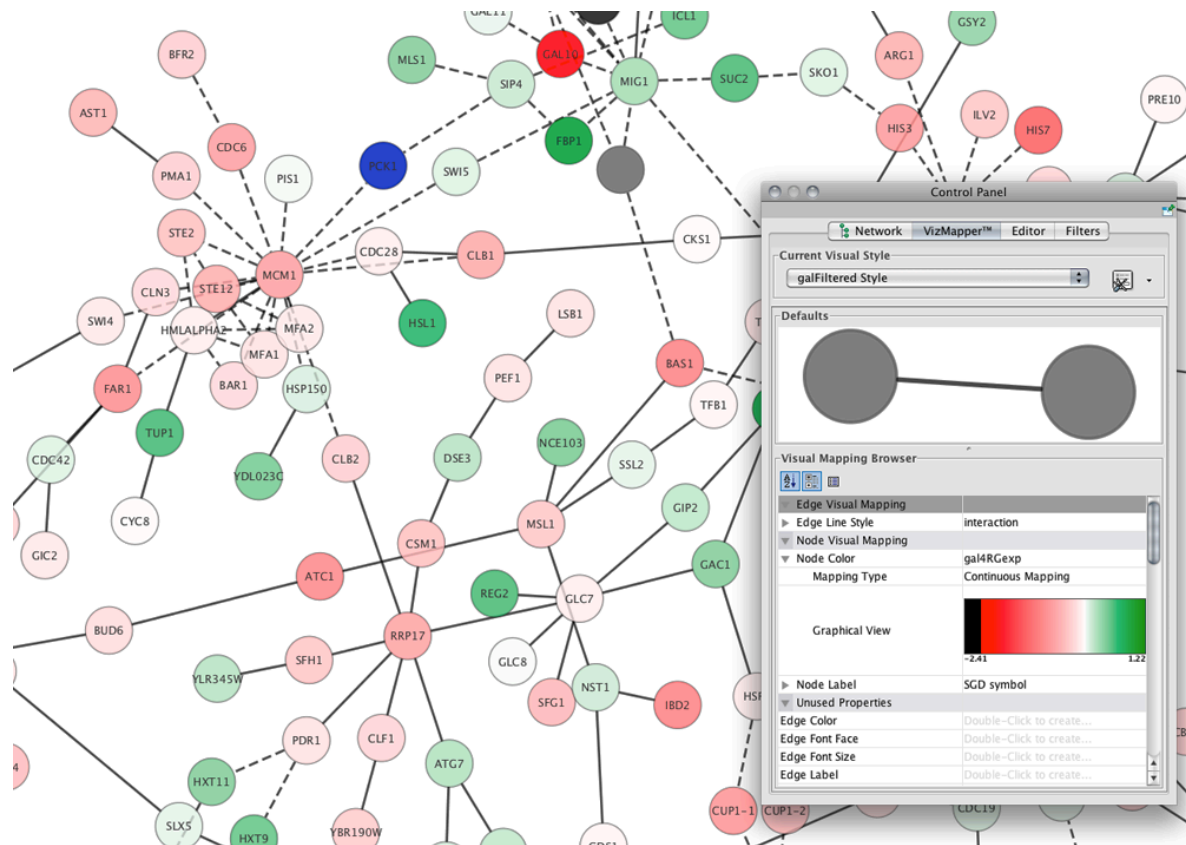
Εξέλιξη των πρωτεϊνικών αλληλεπιδράσεων

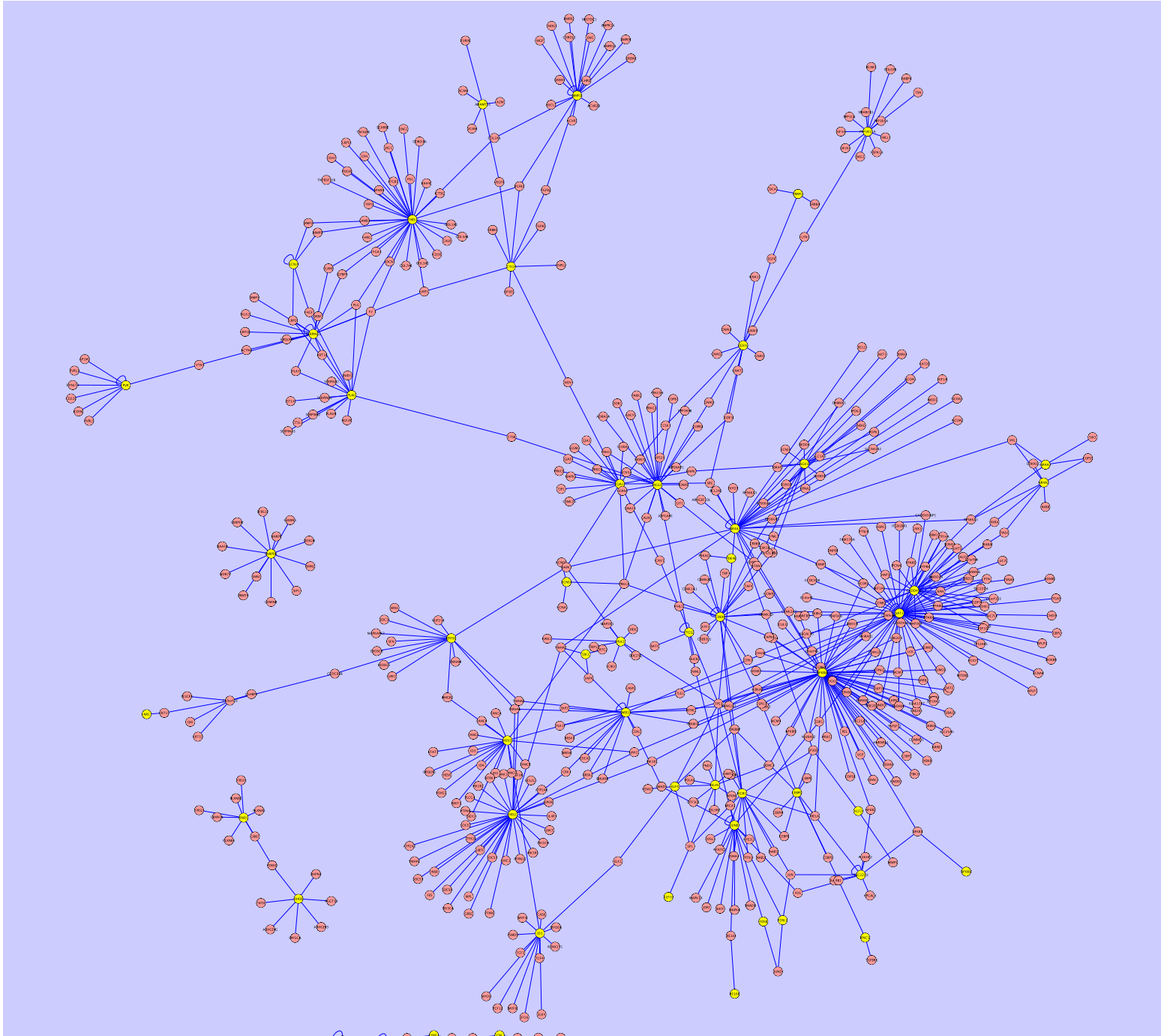


Εικόνα 1: Οι τρεις πιο πιθανές εξελικτικές πορείες ενός ζεύγους διπλασιασμένων γονιδίων: i) οι λειτουργίες του προγονικού γονιδίου μοιράζονται μεταξύ του ζεύγους, ii) το ένα από τα δύο γονίδια διατηρεί τις λειτουργίες του προγονικού γονιδίου, ενώ το δεύτερο αφήνεται ελεύθερο να αναπτύξει κάποιες νέες λειτουργίες, iii) ένα από τα γονίδια διατηρεί τις λειτουργίες του προγονικού γονιδίου, ενώ το άλλο συσσωρεύει επιβλαβείς μεταλλάξεις και τελικά χάνεται.

Cytoscape

- Εργαλείο για την απεικόνιση μοριακών μονοπατιών και πρωτεϊνικών αλληλεπιδράσεων.

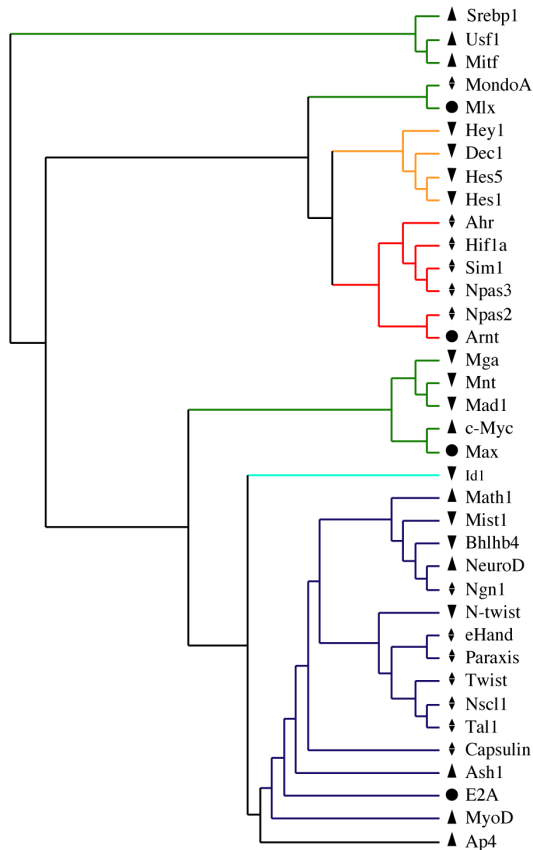




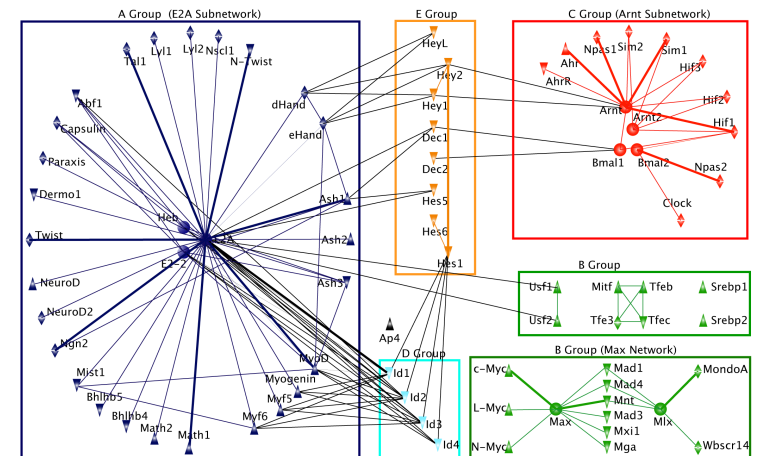
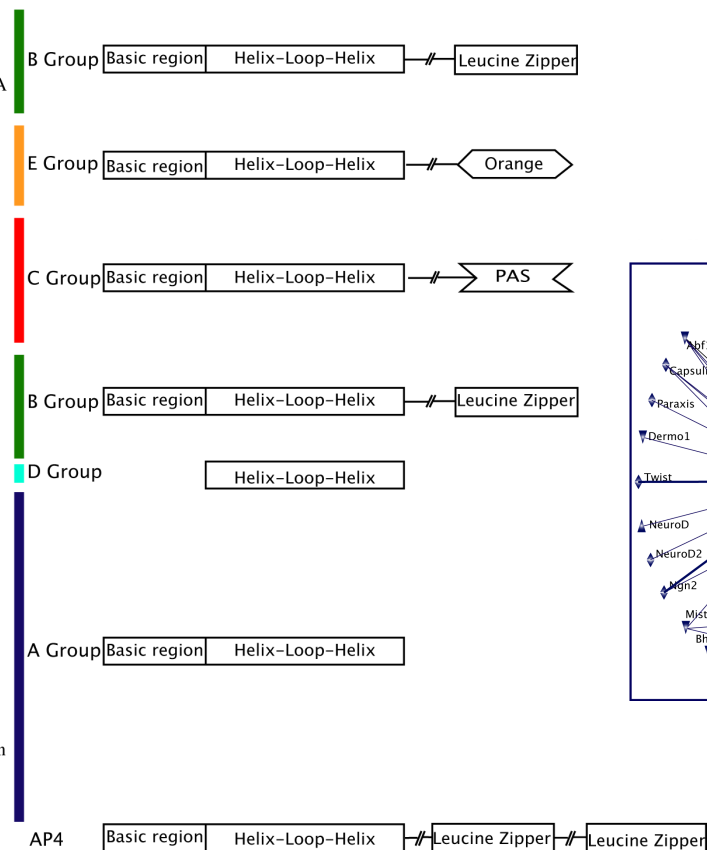
Σύγκλιση στο μοριακό επίπεδο Δίκτυα πρωτεϊνικών αλληλεπιδράσεων

The bHLH heterodimerization network

a) Cladogram of human bHLH domain

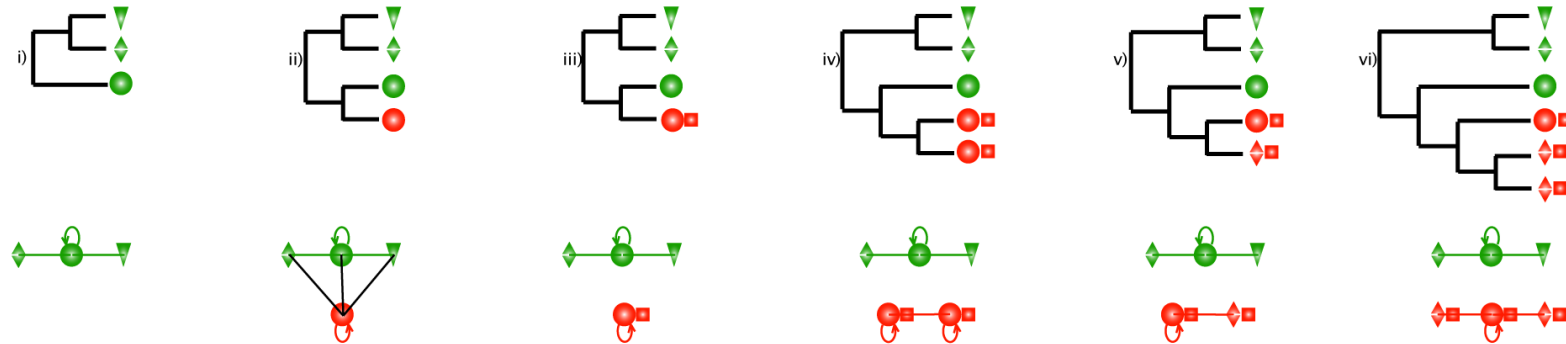


b) Architecture of bHLH dimerization domains

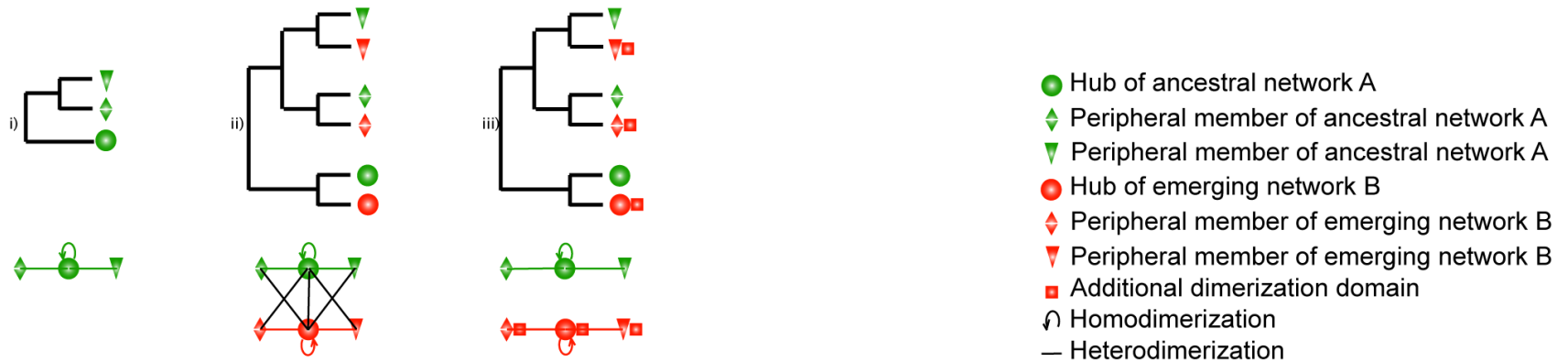


Σύγκλιση στο μοριακό επίπεδο Δίκτυα πρωτεϊνικών αλληλεπιδράσεων

a) Evolution of a heterodimerization network by single gene duplication



b) Evolution of a heterodimerization network by large-scale gene duplication



In vitro

ΔΙΑΓΝΩΣΤΙΚΑ ΤΕΣΤ
ΠΟΥ ΒΑΣΙΖΟΝΤΑΙ ΣΕ
ΜΙΚΡΟΣΥΣΤΟΙΧΙΕΣ

FDA: In Vitro Diagnostic Multivariate Index Assays (IVDMIAAs)

- FDA's In Vitro Diagnostic Product Database
- <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfivd/index.cfm>
- <http://www.ivdtechnology.com/article/exploring-fda-approved-ivdmias>
- Some IVDMIAAs are laboratory-developed tests (LDTs). LDTs are tests that are developed by a single clinical laboratory for use only in that laboratory.
- <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm079148.htm>
- IVDMIAAs raise significant issues of safety and effectiveness. These types of tests are developed based on observed correlations between multivariate data and clinical outcome, such that the clinical validity of the claims is not transparent to patients, laboratorians, and clinicians who order these tests. Additionally, IVDMIAAs frequently have a high risk intended use. FDA is concerned that patients are relying upon IVDMIAAs with high risk intended uses to make critical healthcare decisions when FDA has not ensured that the IVDMIAA has been clinically validated and the healthcare practitioners are unable to clinically validate the test themselves. Therefore, there is a need for FDA to regulate these devices to ensure that the IVDMIAA is safe and effective for its intended use.

Mammaprint - Tissue of origin

- <http://www.ivdtechnology.com/article/exploring-fda-approved-ivdmias>
- **MammaPrint.**

The first IVDMA, the MammaPrint system, made by Agendia Inc., is a qualitative IVD test service performed in a single lab outside the United States using a 70-gene expression profile of fresh frozen breast cancer tissue samples to assess a breast cancer patient's risk for distant metastasis. FDA approved MammaPrint in February 2007 under de novo classification procedures.
- **Tissue of Origin Test**

In July 2008, the Tissue of Origin Test, made by Pathwork Diagnostics, was cleared. This microarray RNA profiling test is to be used on clinical, formalin-fixed, paraffin-embedded (FFPE) biopsy tissue to aid in the classification of the origin of the tumor tissue. In June 2010 a second clearance introduced a different specimen and specimen-preparation method, and the algorithm for analysis of the expression data to create a diagnostics report and interpretation. The test uses microarray technology by Affymetrix Inc. and advanced analytics to measure the gene-expression patterns of challenging tumors, including metastatic, poorly differentiated, and undifferentiated cancer. It is intended to measure the degree of similarity between the RNA expression patterns in a patient's tumor tissue with the RNA expression patterns in a database of fifteen known tumor types.

Καρκίνοι αγνώστου προελεύσεως

- Σε κάποιες περιπτώσεις εμφάνισης/επανεμφάνισης καρκίνου είναι άγνωστη η πρωταρχική πηγή (ιστός), ακόμα και μετά από μια σειρά διαγνωστικών τεστ/βιοψία.
- Αυτό δεν επιτρέπει να χρησιμοποιηθεί ένα κατάλληλο θεραπευτικό σχήμα.
- Οι μικροσυστοιχίες επιτρέπουν να δημιουργηθεί το προφίλ γονιδιακής έκφρασης του συγκεκριμένου καρκίνου και να συγκριθεί με το προφίλ καρκίνων γνωστής προέλευσης.

Καρκίνοι αγνώστου προελεύσεως

- Δημιουργείται μια βάση από δεδομένα μεταγραφωμικής (από άλλες βάσεις δεδομένων και βιβλιογραφία).
- Τα δεδομένα είναι από γνωστούς καρκίνους, κανονικούς ιστούς, και από άλλες ασθένειες.
- Τα δεδομένα φιλτράρονται, κανονικοποιούνται.
- Στη συνέχεια γίνεται σύγκριση.

Καρκίνοι αγνώστου προελεύσεως

- <http://genomemedicine.com/content/3/9/63/abstract>
- **Classification of unknown primary tumors with a data-driven method based on a large microarray reference database**
- **Kalle A Ojala, Sami K Kilpinen and Olli P Kallioniemi**

IVDMIA - FDA

- <http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2007/ucm108836.htm>
- The MammaPrint is the first cleared in vitro diagnostic multivariate index assay (IVDMIA) device.
- <http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2008/ucm116931.htm>
- **FDA Clears Test that Helps Identify Type of Cancer in Tumor Sample**
- The Pathwork Tissue of Origin test compares the genetic material of a patient's tumor with genetic information on malignant tumor types stored in a database. It uses a microarray technology to analyze thousands of pieces of genetic material at one time. The test considers 15 common malignant tumor types, including bladder, breast, and colorectal tumors.

Εφαρμογές στην τοξικολογία/ τοξικογενωμική

- Μέτρηση της γονιδιακής έκφρασης μετά από έκθεση σε τοξικό παράγοντα μπορεί να δείξει τον μοριακό μηχανισμό δράσης του παράγοντα.
- Μπορεί να αποτελέσει μοναδική μοριακή υπογραφή του συγκεκριμένου τοξικού παράγοντα, για μελλοντική ανίχνευσή του.
 - Ομαδοποίηση τοξικών παραγόντων με κοινή δράση, με βάση την ομοιότητα των μοριακών προφίλ τους

Μοριακό προφίλ τοξικότητας

Toxicology Letters

Volume 120, Issues 1-3, 31 March 2001, Pages 359-368

doi:10.1016/S0378-4274(01)00267-3 | How to Cite or Link Using DOI

Permissions & Reprints

Microarray analysis of hepatotoxins in vitro reveals a correlation between gene expression profiles and mechanisms of toxicity

Jeffrey F. Waring, Rita Ciurlionis, Robert A. Jolly, Matthew Heindel and Roger G. Ulrich  

Department of Cellular and Molecular Toxicology, Abbott Laboratories, D468 AP13A, 100 Abbott Park Road, Abbott Park, IL 60064-6104, USA

- Ηπατοκύτταρα αρουραίων εκτέθηκαν σε 15 γνωστές ηπατο-τοξίνες.
- Για την κάθε μία, δημιουργήθηκε το μοριακό προφίλ γονιδιακής έκφρασης (ποιά γονίδια υπερ/υπο-εκφράστηκαν).
- Τοξίνες με παρόμοιο μηχανισμό δράσης είχαν παρόμοια (όχι όμως ακριβώς ίδια) προφίλ γονιδιακής έκφρασης και ομαδοποιούνταν.

Μοριακό προφίλ τοξικότητας

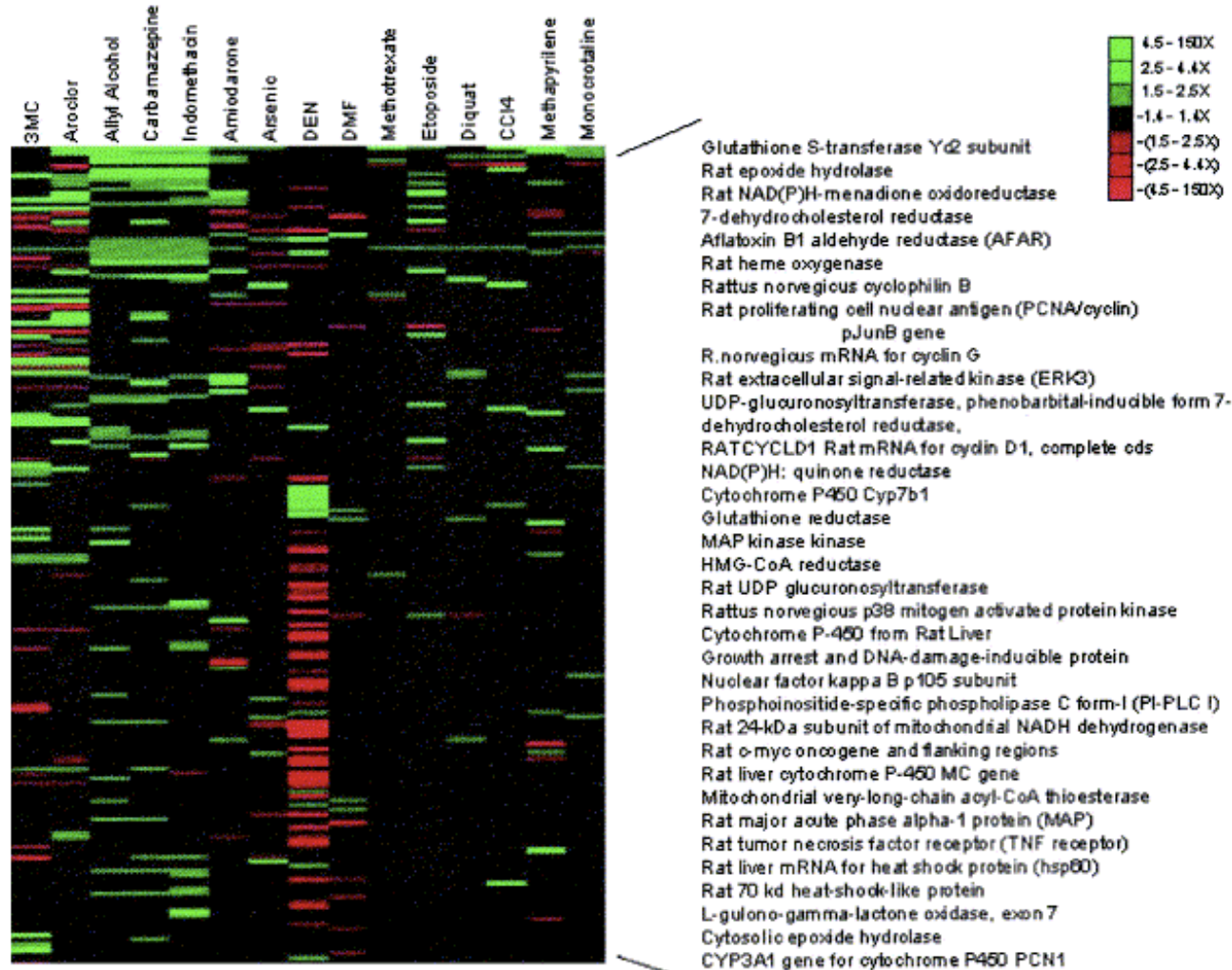


Fig. 2. Graph showing the gene changes occurring in livers from rats treated with the 15 known hepatotoxins. A total of 179 genes were shown to be regulated at least two-fold by at least one compound. Some of these genes are shown to the right of the figure.

Μοριακό προφίλ τοξικότητας

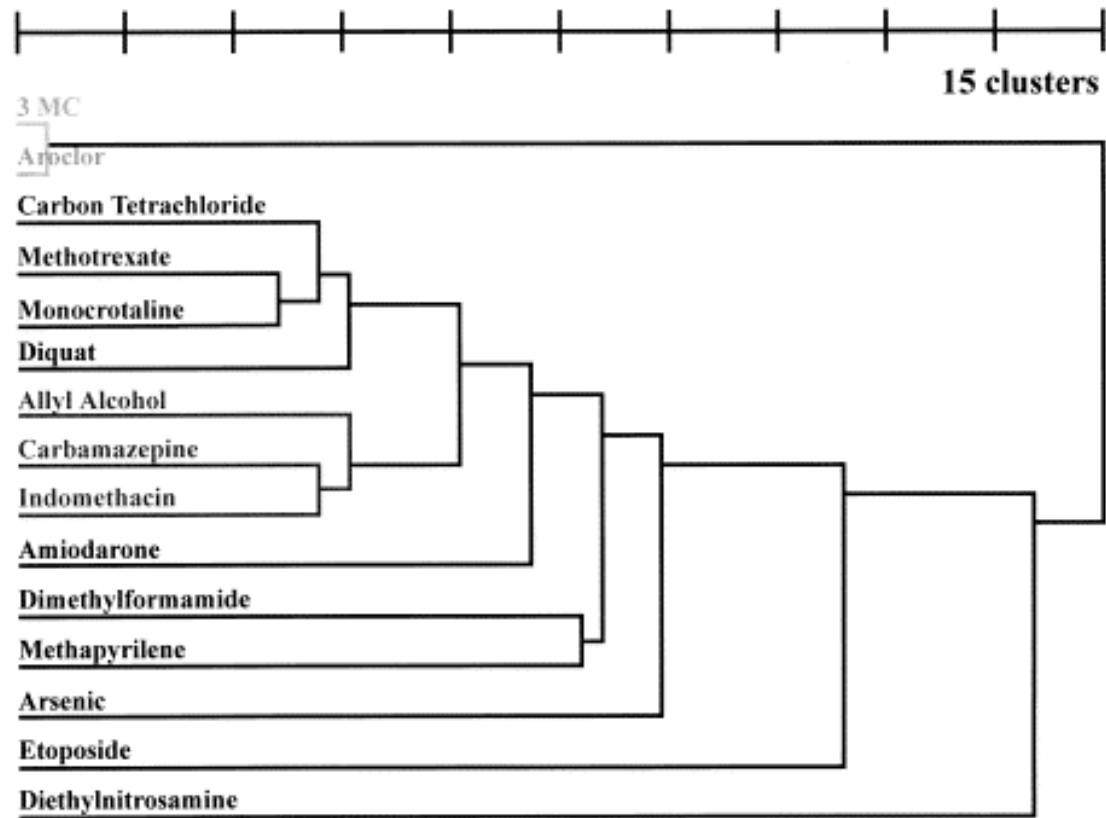
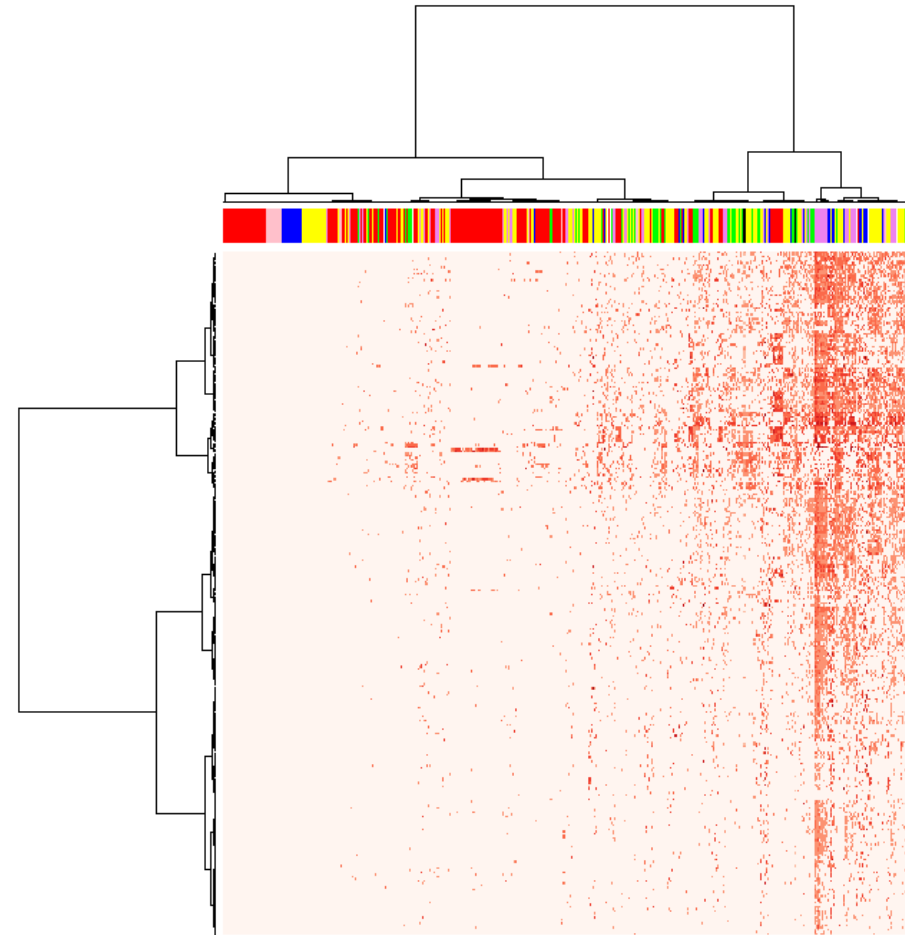


Fig. 3. Dendrogram showing the clustering of the hepatotoxins based on gene regulation. The clustering was hierarchical using correlation as the distance (see [Section 2](#)).

Hierarchical cluster analysis showed a close association in gene expressional responses between aroclor 1254 and 3-methylcholanthrene.

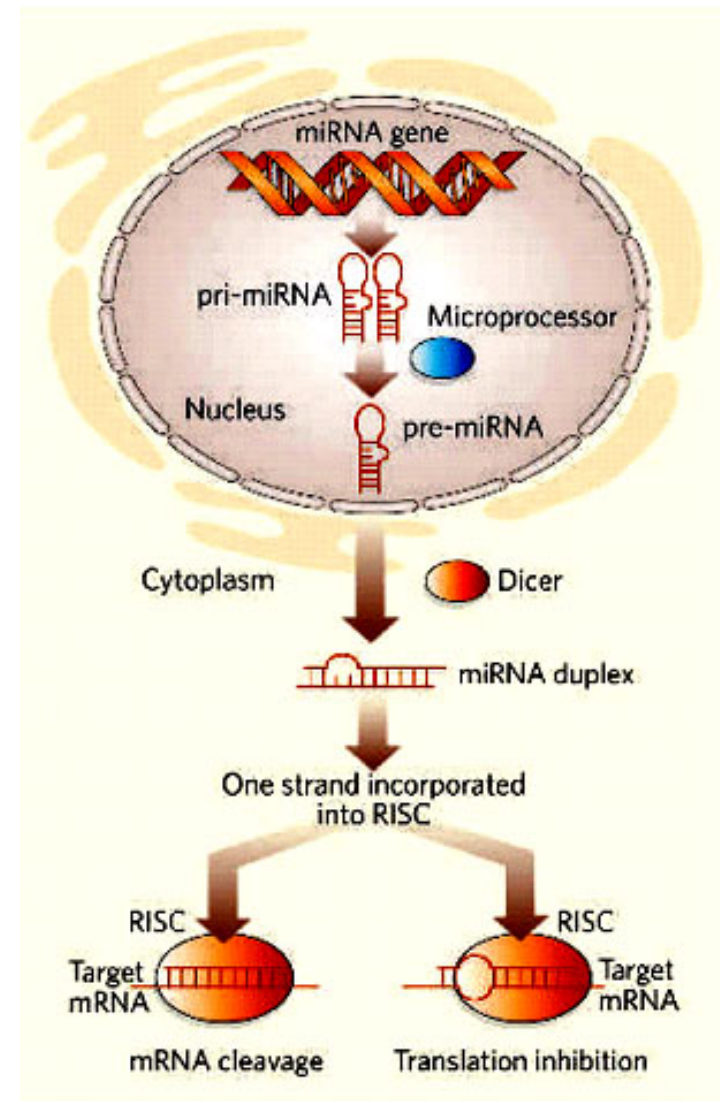
ToxCastDB

- Β.Δ. Που έχει αποτελέσματα 624 τοξικολογικών εξετάσεων (οριζόντιος άξονας) για διάφορες χημικές ουσίες (κάθετος άξονας).



Micro-RNAs


- Μικρά RNAs που ρυθμίζουν (κυρίως καταστέλουν) την μεταγραφή των mRNAs
 - Αποδόμιση του mRNA.
 - Αναστολή της μεταγραφής του mRNA.
- Μέγεθος ~ 22 nt.
- ~1000 στον άνθρωπο.
- Ρυθμίζουν πολλά mRNAs (30%-100%).
- Ένα mi-RNA -> στοχεύει πολλά mRNAs.
- Πολλά mi-RNAs -> στοχεύουν το ίδιο mRNA (hunting pack).
- Έχουν εξειδικευμένο προφίλ έκφρασης για τον κάθε ιστό (και καρκίνο) (συνήθως σχετίζονται με συγκεκριμένες αναπτυξιακές διαδικασίες).
- Μπορούν να λειτουργήσουν ως διαγνωστικοί ή προγνωστικοί μοριακοί δείκτες.



Toxicology Applications

Circulating microRNAs, potential biomarkers for drug-induced liver injury

Kai Wang,¹ Shile Zhang, Bruz Marzolf, Pamela Troisch, Amy Brightman, Zhiyuan Hu, Leroy E. Hood,¹ and David J. Galas

 Author Affiliations

Contributed by Leroy E. Hood, January 16, 2009 (received for review December 9, 2008)

Abstract

Drug-induced liver injury is a frequent side effect of many drugs, constitutes a significant threat to patient health and has an enormous economic impact on health care expenditures. Numerous efforts have been made to identify reliable and predictive markers to detect the early signs of drug-induced injury to the liver, one of the most vulnerable organs in the body. These studies have, however, not delivered any more informative candidates than the serum aminotransferase markers that have been available for ≈ 30 years. Using acetaminophen overdose-induced liver injury in the mouse as a model system, we have observed highly significant differences in the spectrum and levels of microRNAs in both liver tissues and in plasma between control and overdosed animals. Based on our survey of microRNA expression among normal tissues, some of the microRNAs, like messenger RNAs, display restricted tissue distributions. A number of elevated circulating microRNAs in plasma collected from acetaminophen-overdosed animals are highly expressed in the liver. We have demonstrated that specific microRNA species, such as mir-122 and mir-192, both are enriched in the liver tissue and exhibit dose- and exposure duration-dependent changes in the plasma that parallel serum aminotransferase levels and the histopathology of liver degeneration, but their changes can be detected significantly earlier. These findings suggest the potential of using specific circulating microRNAs as sensitive and informative biomarkers for drug-induced liver injury.

Τοξική δράση acetaminophen

- Acetaminophen: Είναι αναλγητικό/αντιπυρετικό
- Έλεγχος με microarrays αν τα mi-RNA μπορούν να καταδείξουν την ηπατική καταστροφή μετά από χορήγηση φαρμάκου (acetaminophen overdose).
- Σε κατάσταση υπερχορήγησης, τα μεταβολικά μονοπάτια που καταβολίζουν το φάρμακο είναι σε κορεσμό.
- Το πλεονάζον φάρμακο υπόκειται σε καταβολισμό από ένζυμα των P450 -> NAPQI (N-acetyl-p-benzoquinone imine) (τοξικό).
- NAPQI εξουδετερώνεται από ενδοκυτταρική γλουταθιόνη (GSH).
- Ότι NAPQI περισσεύει, προκαλεί ηπατική καταστροφή.
- Σε περίπτωση υπερβολικής δόσης acetaminophen (-> υπερβολικό NAPQI), χορηγούμε N-acetylcysteine (NAC) -> ενδοκυτταρικό GSH.

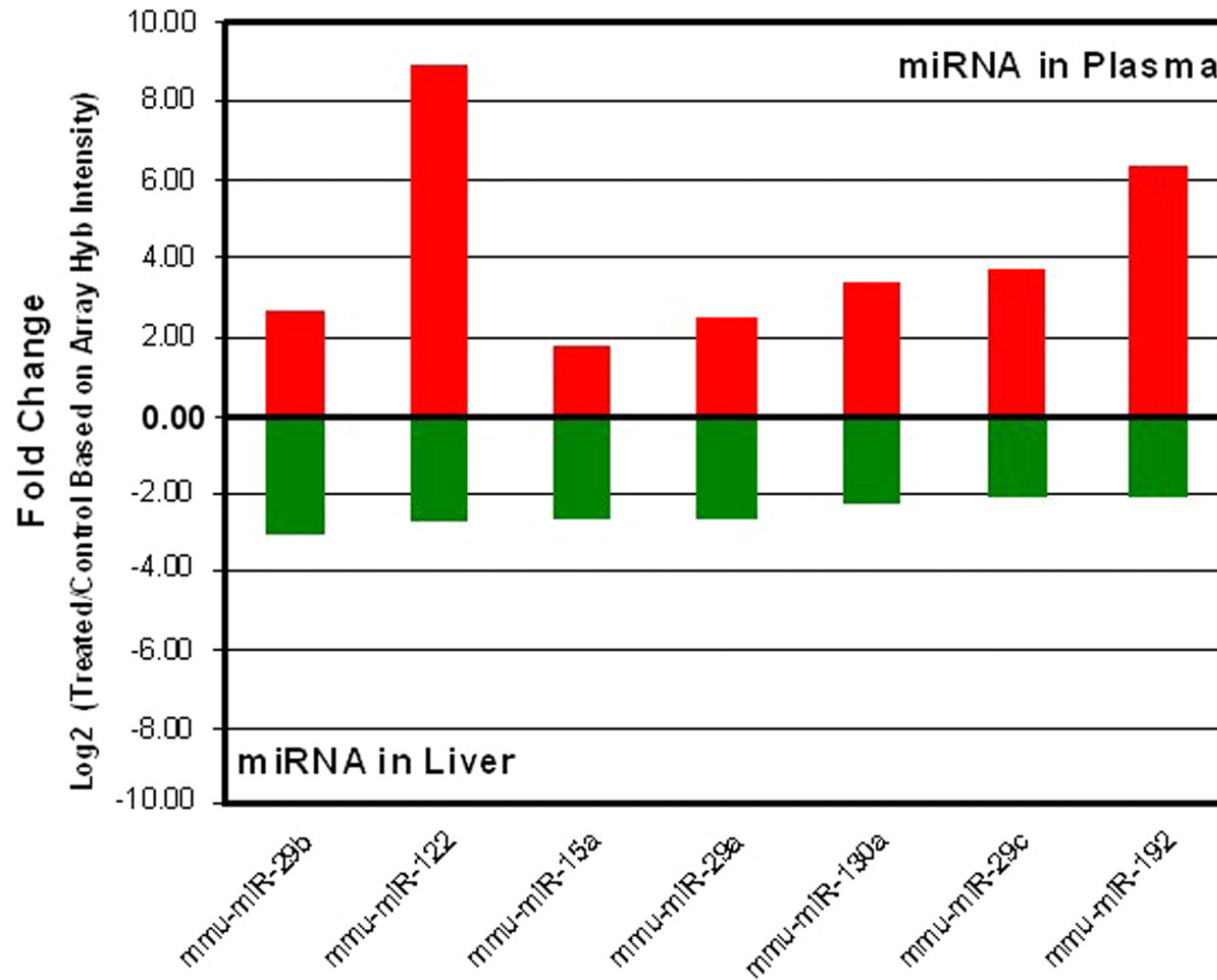
Μέτρηση της τοξικής δράσης του acetaminophen

- Η τοξικότητα δεν εμφανίζεται αμέσως.
- Χρειάζεται γρήγορη και αξιόπιστη μέτρηση της τοξικότητας με αιματολογικό έλεγχο (ορό).
- Μέτρηση της αμινοτρανσφεράσης του ορού (κλασσική μέθοδος).

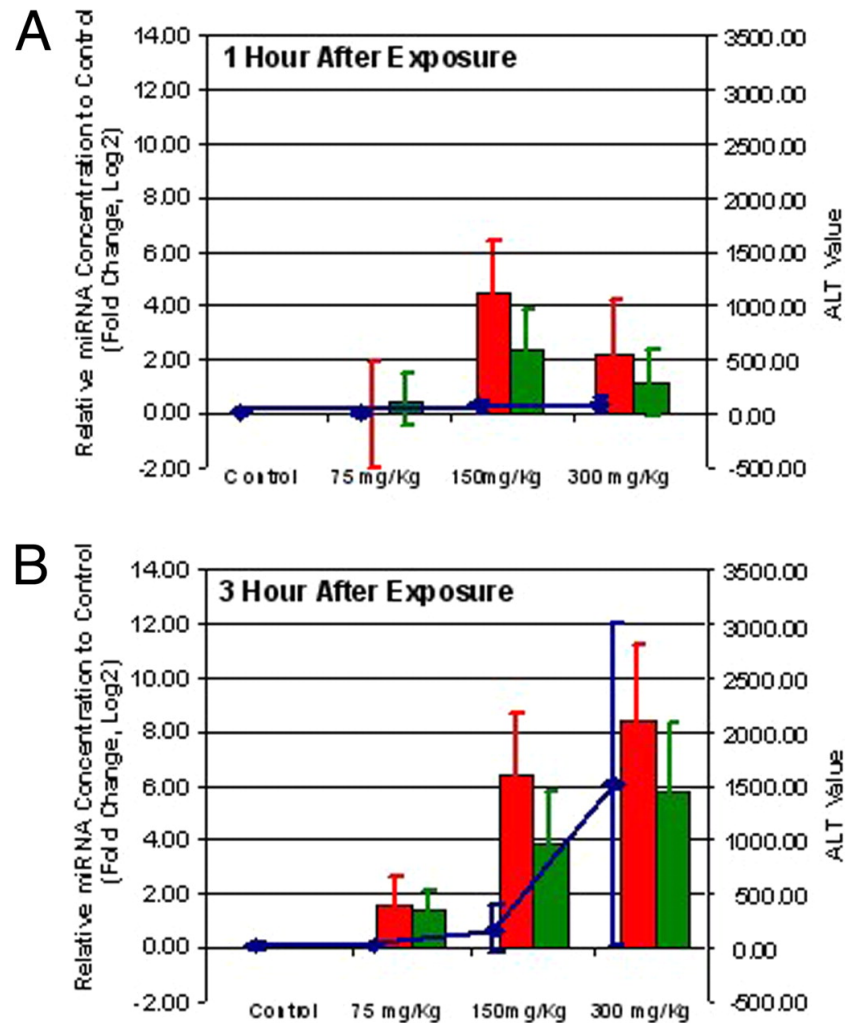
- Βρήκαν mi-RNAs που εκφράζονται σημαντικά στο ήπαρ.
- Παρακολούθησαν την έκφρασή τους στο ήπαρ μετά τη χορήγηση acetaminophen και επίσης παρακολούθησαν την ανίχνευσή τους στον ορό (λόγω πιθανής ηπατικής κυτταρικής καταστροφής).

Toxicology Applications

Selected miRNA shows opposite changes between liver and plasma samples based on microarray results.



Τα mi-RNAs αποδείχθηκαν πιο αξιόπιστοι και πιο γρήγοροι δείκτες ηπατικής καταστροφής



MicroRNAs are more sensitive markers than ALT for liver injury. The comparison among the levels of mir-122 (red bars), mir-192 (green bars), and ALT (blue line) in plasma samples collected from mice at 1 (A) and 3 (B) h after exposure to different doses of acetaminophen (indicated on x axis). The relative change of miRNA expression levels (ratio in log 2 compared with control) is indicated on the left side of the figure and the scale of ALT level is on the right. The relative change of miRNA levels is expressed in log 2 ratio of each treatment condition compared with the corresponding control. The values of miRNA fold change and ALT levels are the average of 4 independent samples from each time point, and the standard derivations are shown as error bars.