Λειτουργική γονιδιωματική

Λειτουργική γονιδιωματική: Τι είναι

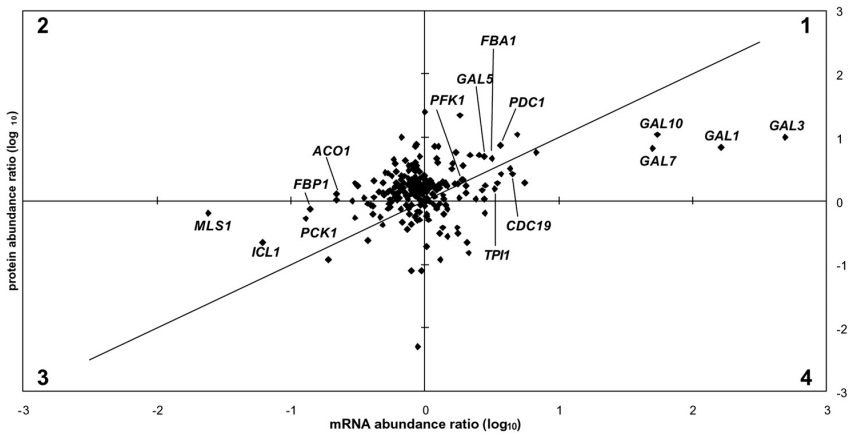
- Προσπαθεί να κατανοήσει τις λειτουργίες των βιολογικών μορίων, σε επίπεδο ολόκληρου του γονιδιώματος.
- Γίνονται μετρήσεις για το σύνολο των γονιδίων, σε μια συγκεκριμένη στιγμή ή κατάσταση.
- Αρχικά, οι μετρήσεις γίνονταν για ένα βιομόριο. Σήμερα μελετάμε την συμπεριφορά ολόκληρου του συστήματος.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

Transcriptomics

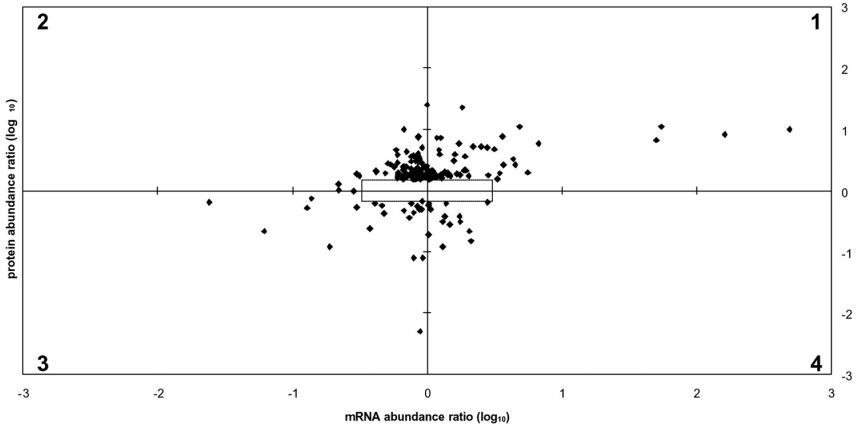
- Expressed sequence tags (ESTs)
- Serial analysis of gene expression (SAGE)
- Μικροσυστοιχίες (microarrays)
- RNA-seq (whole transcriptome shotgun sequencing)

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)

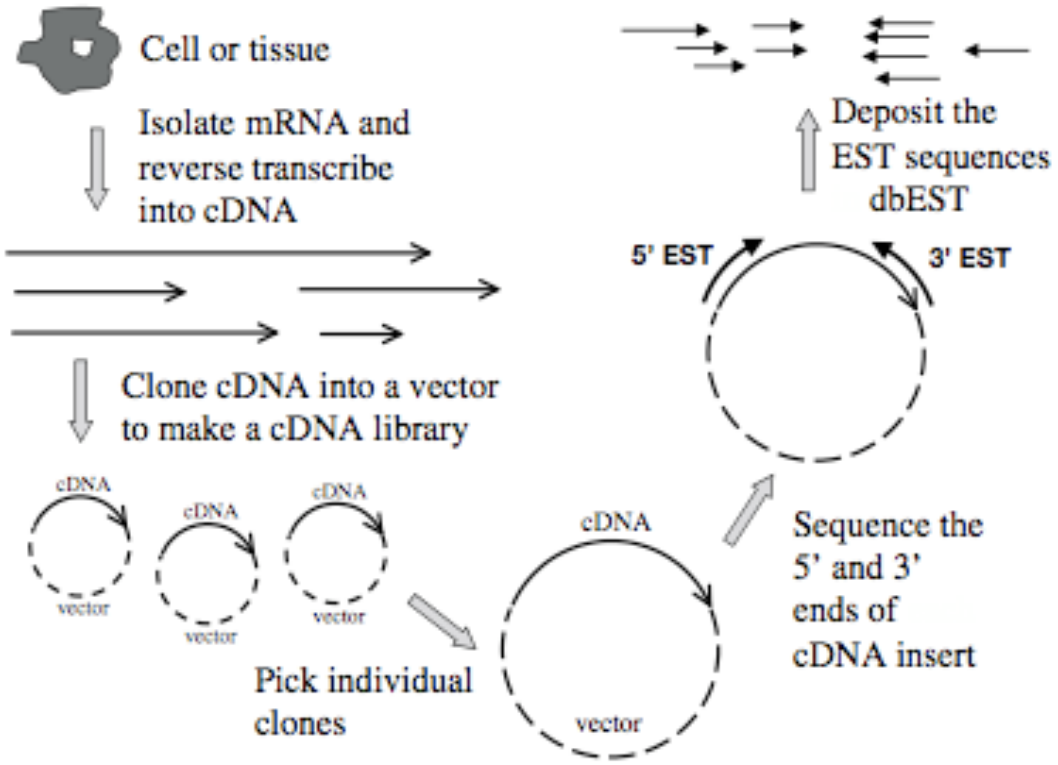


Figure 12.1. Overview of how ESTs are constructed.

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/

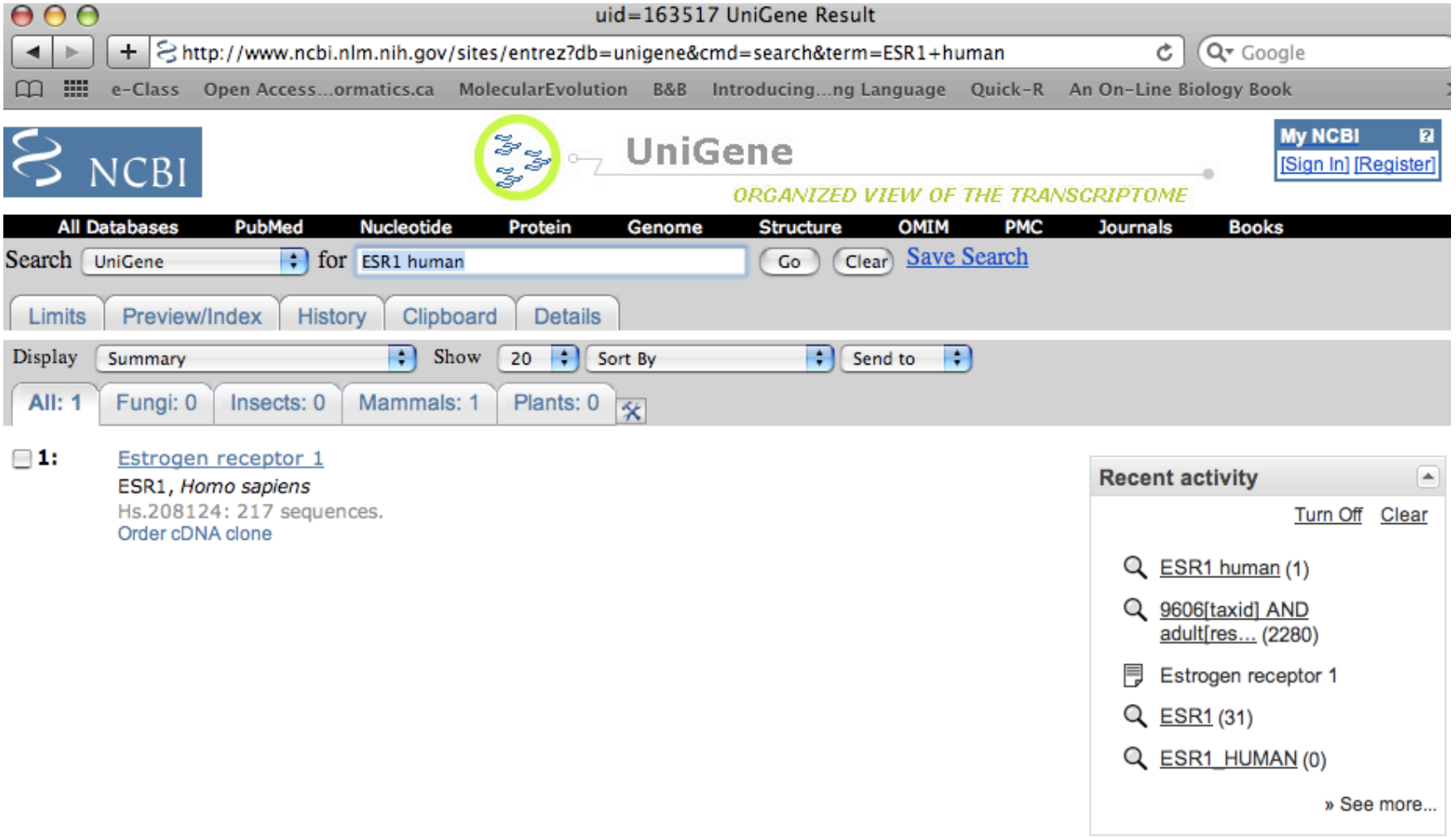
Expressed sequence tags (ESTs)

- 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 συγκεκριμμένο γονίδιο

Unigene


uid=163517 UniGene Result
http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene&cmd=search&term=ESR1+human
Google
e-Class
Open Access...ormatics.ca
MolecularEvolution
B&B
Introducing...ng Language
Quick-R
An On-Line Biology Book
NCBI
UniGene
ORGANIZED VIEW OF THE TRANSCRIPTOME
My NCBI
[Sign In] [Register]
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
OMIM
PMC
Journals
Books
Search
UniGene
for
ESR1 human
Go
Clear
Save Search
Limits
Preview/Index
History
Clipboard
Details
Display
Summary
Show
20
Sort By
Send to
All: 1
Fungi: 0
Insects: 0
Mammals: 1
Plants: 0
1:
Estrogen receptor 1
ESR1, Homo sapiens
Hs.208124: 217 sequences.
Order cDNA clone
Recent activity
Turn Off
Clear
ESR1 human (1)
9606[taxid] AND adult[res... (2280)
Estrogen receptor 1
ESR1 (31)
ESR1_HUMAN (0)
» 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	*H. sapiens*	100.0	594
XP_001097228.1	PREDICTED: estrogen receptor isoform 1	*M. mulatta*	99.7	594
NP_001158059.1	estrogen receptor	*P. anubis*	99.7	594
NP_031982.1	estrogen receptor	*M. musculus*	94.8	598
NP_001083084.2	estrogen receptor 1	*X. laevis*	82.4	584
NP_694491.1	estrogen receptor	*D. rerio*	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

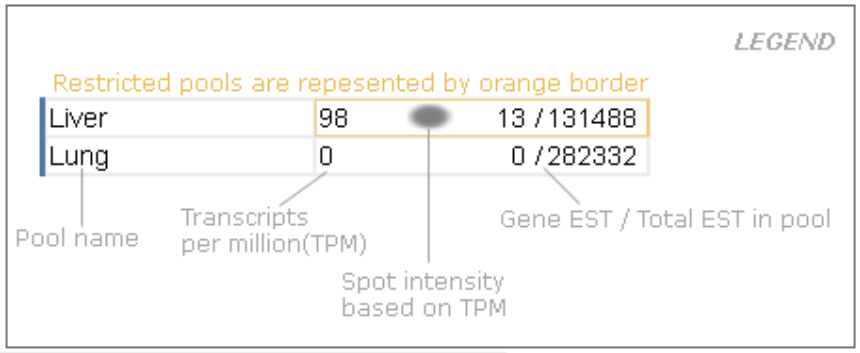
	Hs.208124	
adipose tissue	0	0/13106
adrenal gland	30	1/33197
ascites	0	0/40015
bladder	0	0/29757
blood	0	0/123478
bone	0	0/71655
bone marrow	0	0/48801
brain	0	1/1100989
cervix	0	0/48171
connective tissue	6	1/149255
ear	0	0/16212

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

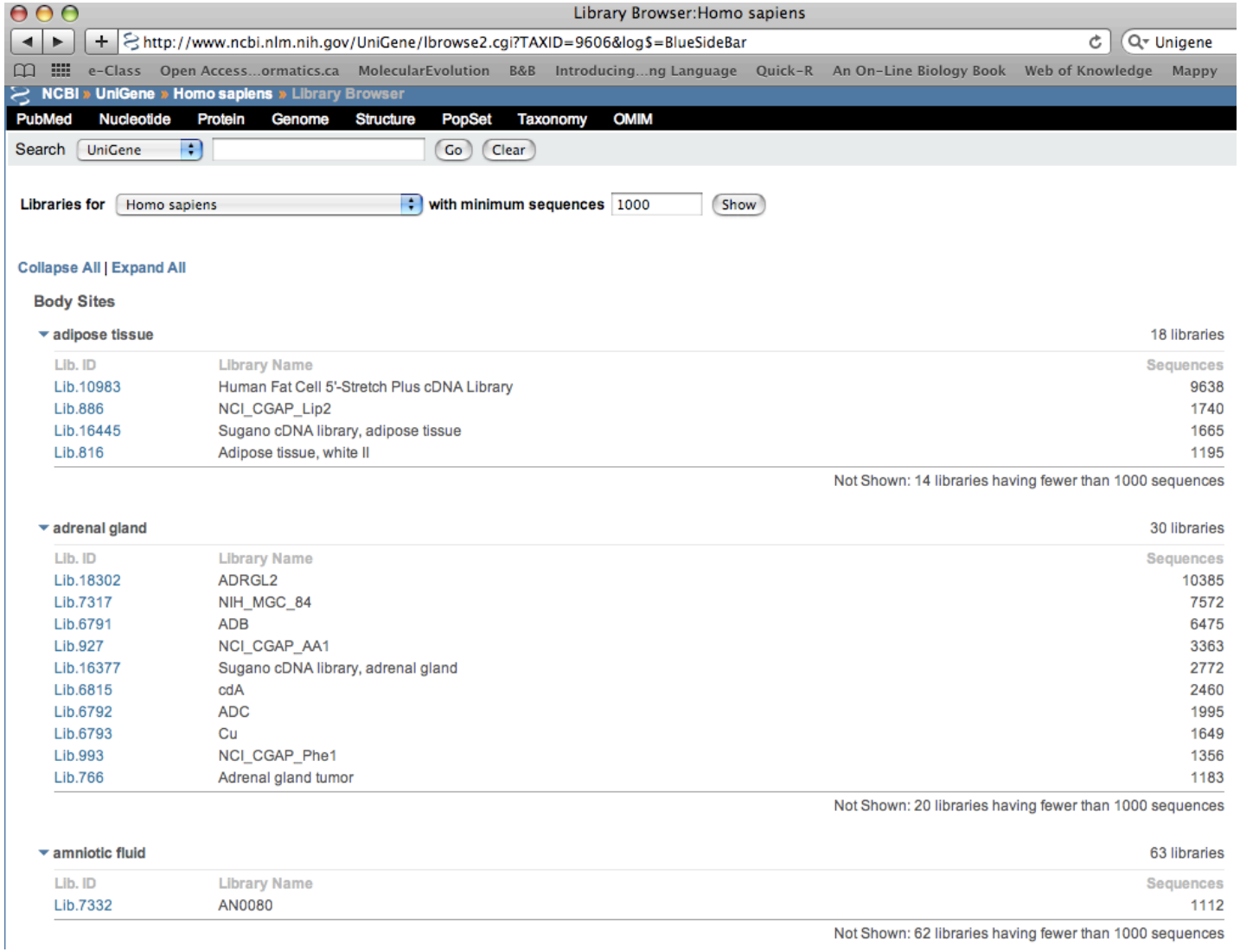
Breakdown by Health State

	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



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

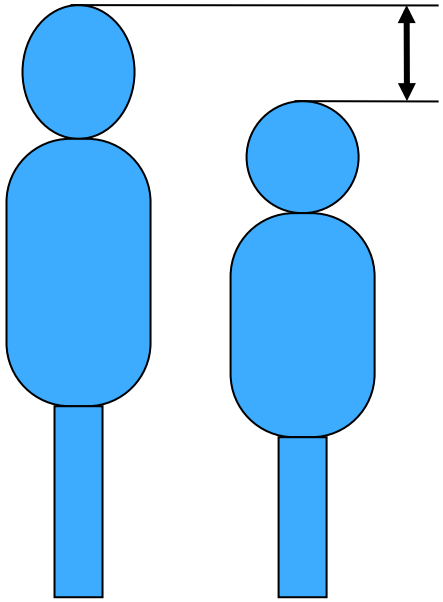
Microarrays

- 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 Α μετά από άσκηση (control: αμέσως πριν την άσκηση).
- Δεν μπορούμε να συγκρίνουμε το γονίδιο Α με το γονίδιο Β

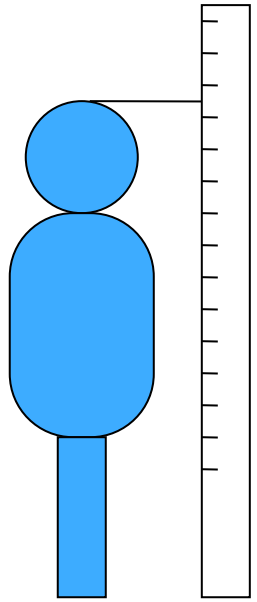
- Στο Chip υπάρχουν για το κάθε μεταγράφημα:
  - cDNAs ολόκληρου του μεταγραφώματος (spotting) (PCR)
  - ολιγομερή (τα λεγόμενα oligo-probes), με μήκος 25-70 βάσεις.
    - Spotting
    - In situ synthesis

Σχετικές διαφορές

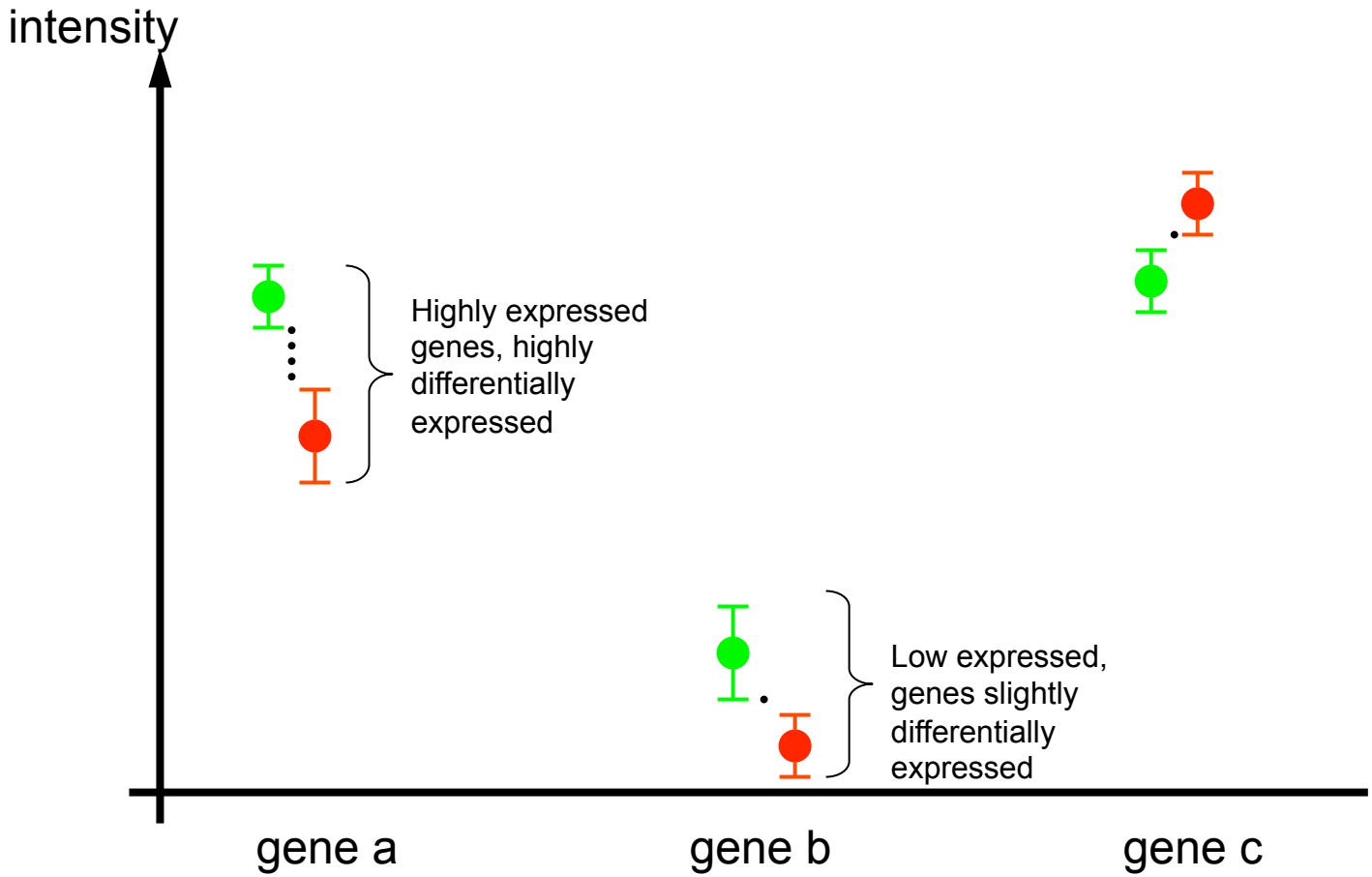
relative measurement 2
COLORS (Cy5 / Cy3)



absolute measurement
1 COLOR (Cy3)

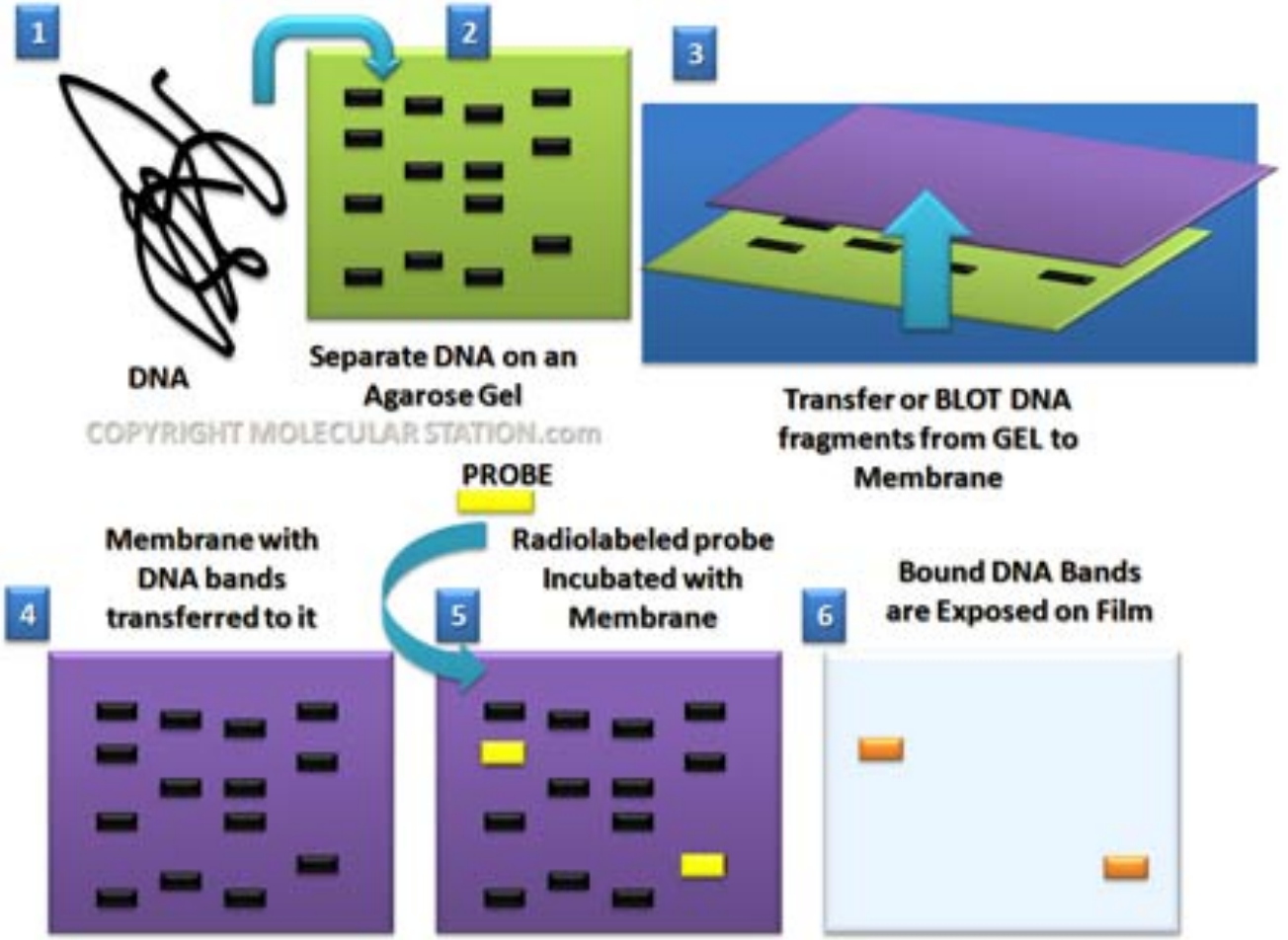


Σχετικές διαφορές



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 συγκεκριμμένο 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.

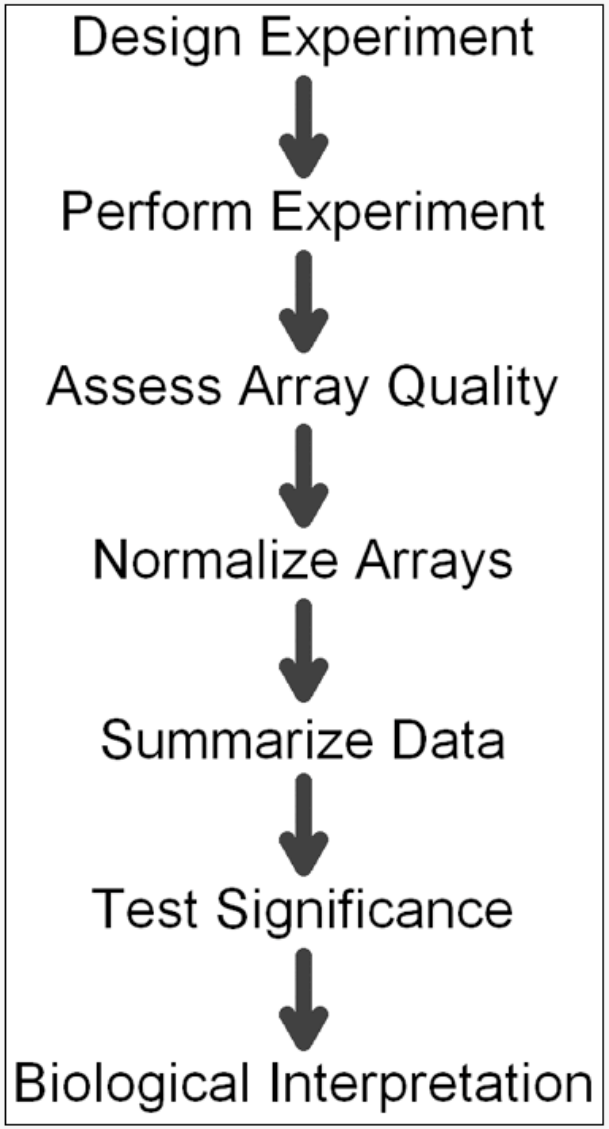


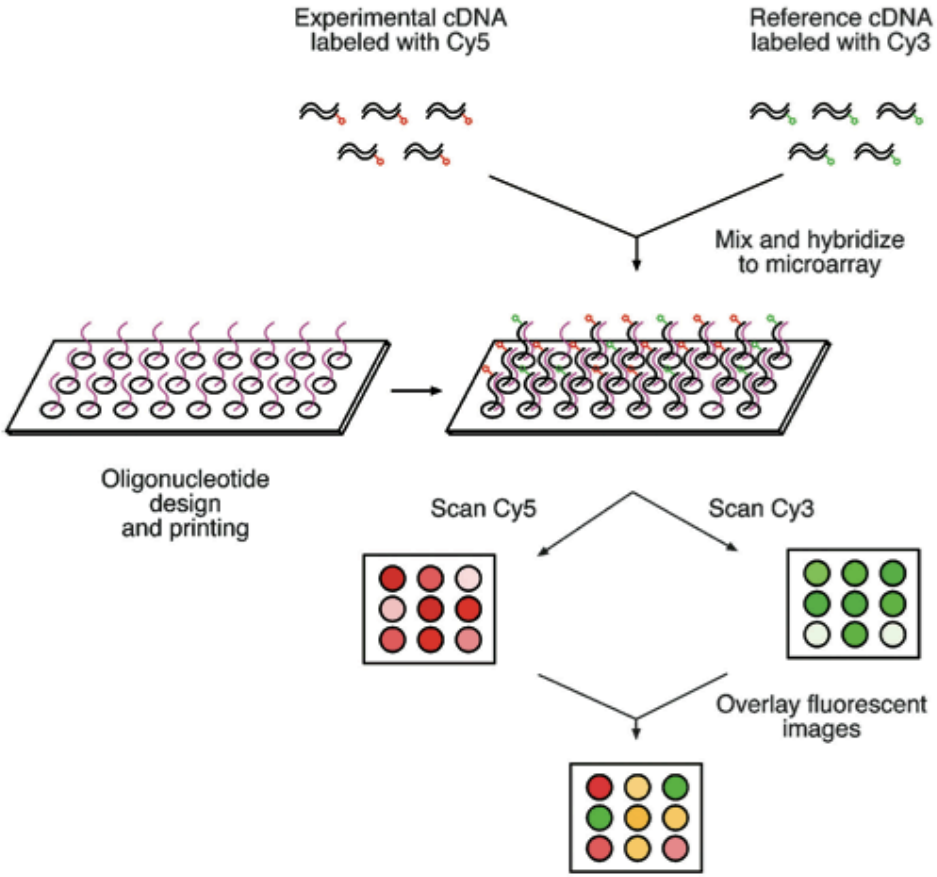
Figure 1

Steps in a typical microarray analysis.

Μικροσυστοιχίες (microarrays)

- Με δύο χρώματα (πράσινο/κόκκινο)
- Το ένα χρώμα είναι για μια συγκεκριμμένη 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 συγκεκριμμένη χρωστική (π.χ. Cy5:red - Cy3:green).
- Τα 2 δείγματα αναμιγνύονται και το μίγμα υπόκειται σε υβριδισμό πάνω στο chip.
- Μετριέται η ένταση της κάθε μιας από τις 2 χρωστικές, για κάθε κηλίδα.
- Υπολογίζεται ο λόγος των εντάσεων των 2 χρωστικών, στην κάθε κηλίδα.

Μικροσυστοιχίες



Μικροσυστοιχίες (microarrays)

- Αν το γονίδιο εκφράζεται περισσότερο στην Α συνθήκη (κόκκινη χρωστική) από ότι στην control (πράσινη χρωστική), τότε ο λόγος συνθήκη_Α/control (κόκκινη/πράσινη) θα είναι $\lambda>1$, αλλιώς σε αντίθετη περίπτωση $0<\lambda<1$.
- Αν το γονίδιο εκφράζεται με διπλάσια ένταση στην συνθήκη Α, σε σχέση με την συνθήκη control, τότε ο λόγος θα είναι $\lambda=2$.
- Αν το γονίδιο εκφράζεται με τη μισή ένταση στην συνθήκη Α, 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 συγκεκριμμένη συνθήκη.
  - $Log_2\lambda > 1$ ή $Log_2\lambda < -1$ (διπλάσια/υποδιπλάσια έκφραση σε σχέση με τη συνθήκη control).
  - Με στατιστικές μεθόδους (t-test, ANOVA).

Ομαδοποίηση γονιδίων με την ίδια συμπεριφορά.

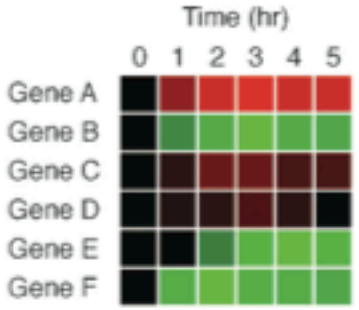
- Χρειαζόμαστε αρκετά σημεία (διαφορετικές συνθήκες ή χρονικές στιγμές)
- Με μεθόδους αποστάσεων, όπου οι μετρήσεις ενός γονιδίου για διαφορετικές συνθήκες αποτελούν ένα διάνυσμα.
- Υπολογίζουμε αποστάσεις μεταξύ διαφορετικών διανυσμάτων (γονιδίων).
  - Ευκλείδια απόσταση
  - Συντελεστής συσχέτισης Pearson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

Μικροσυστοιχίες

	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

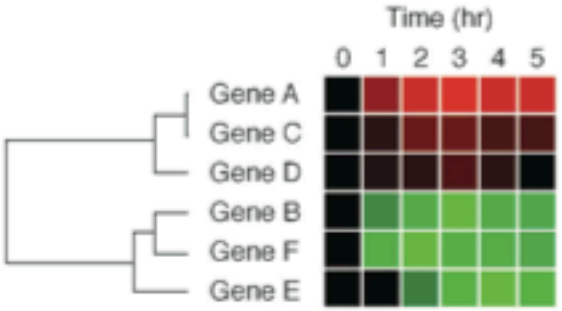
calculating Pearson correlation coefficients between genes

	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

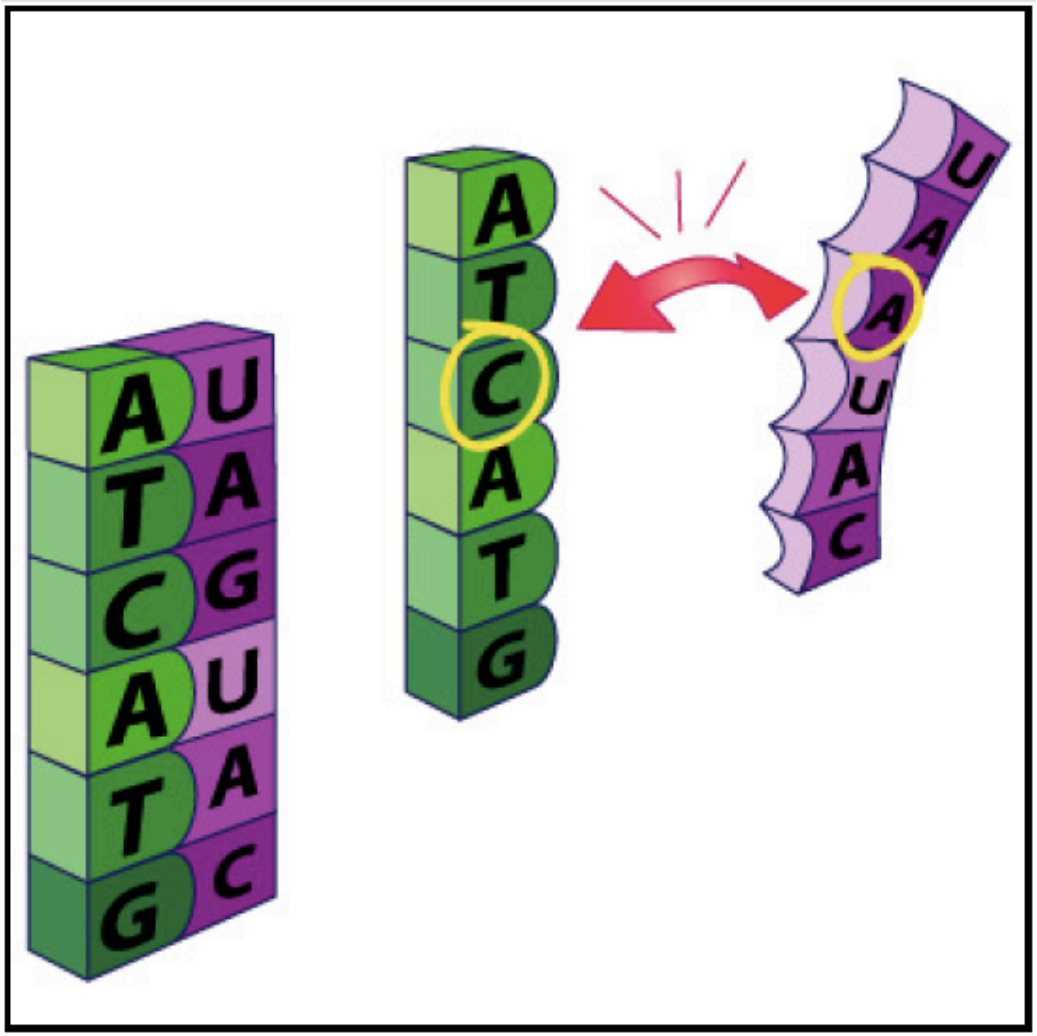
hierarchical clustering



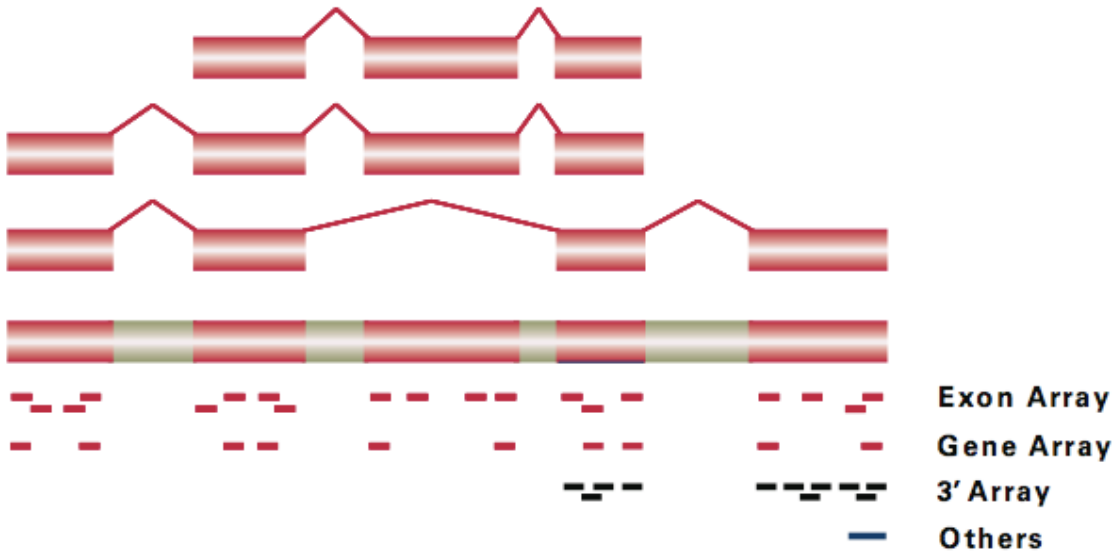
Μήκος των probes

- 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



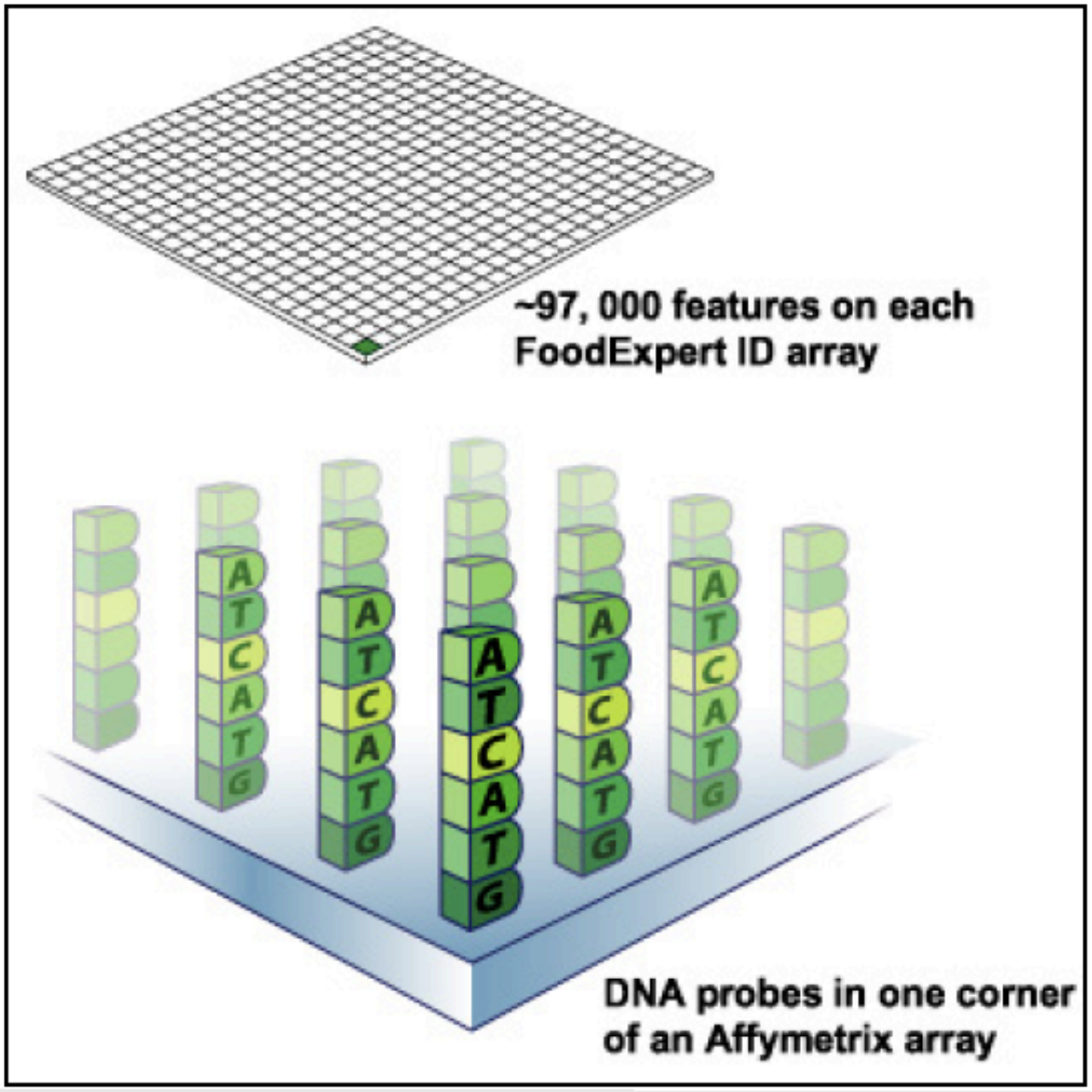
Probes & chips

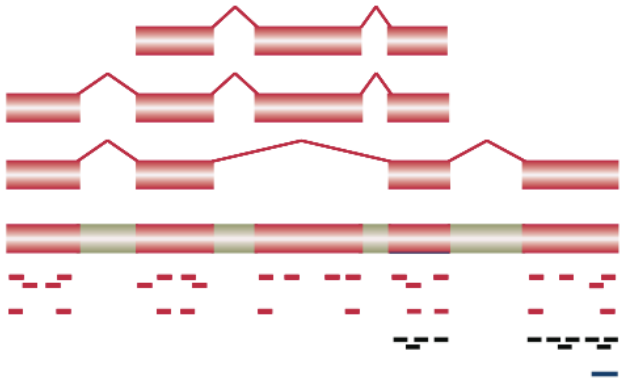


Σχεδιασμός των probes

- 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 Tm (βέλτιστη θερμοκρασία υβριδισμού).
- Τα probes δεν θα πρέπει να σχηματίζουν δευτεροταγείς δομές που τα εμποδίζουν να υβριδιστούν με τον στόχο.

spots


~97, 000 features on each
FoodExpert ID array
DNA probes in one corner
of an Affymetrix array



Ανάλυση δεδομένων

Μετατροπή έντασης σήματος σε $\log_2$

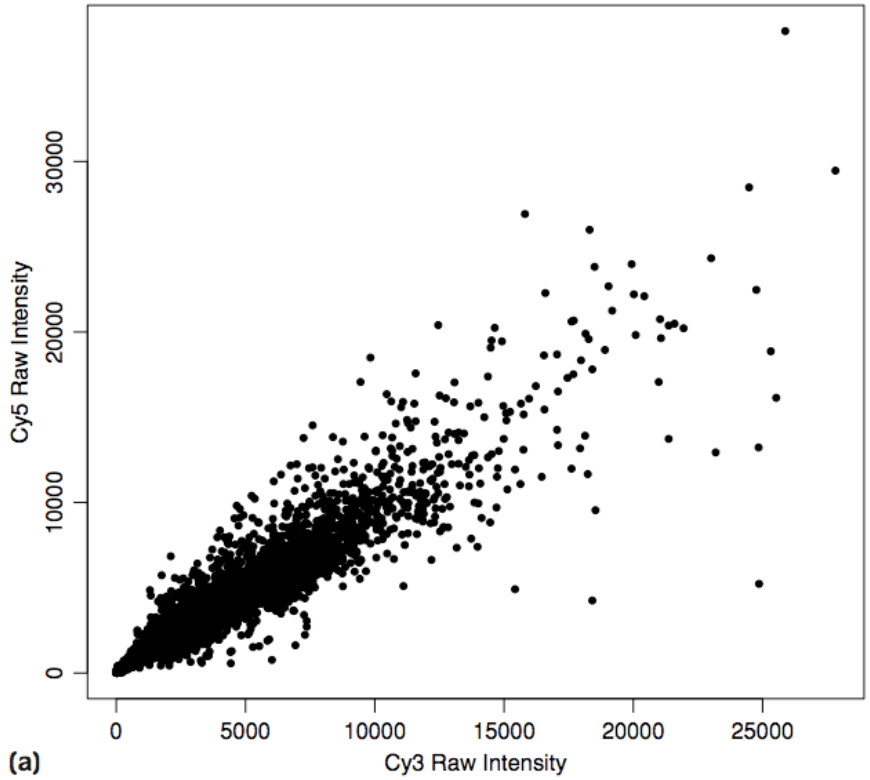
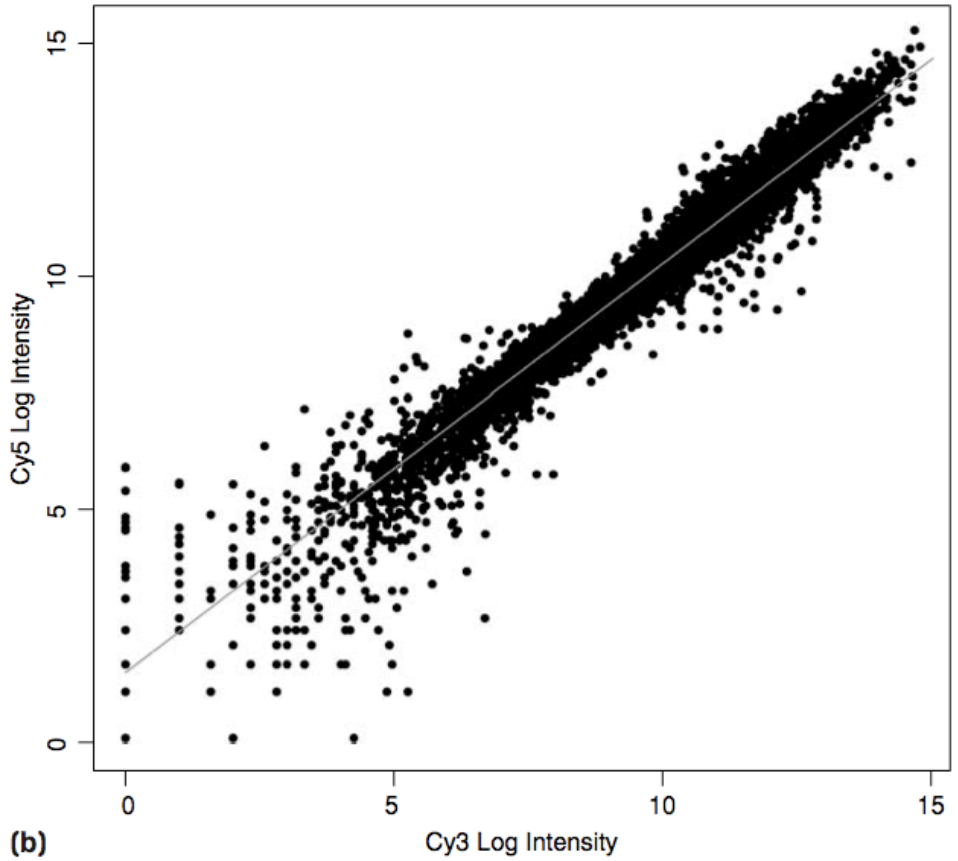


Figure 5.1: Plots of Cy3 vs. Cy5 for data set 5A. Human foreskin fibroblasts have been infected with *Toxoplasma gondii* for a period of 1 hour. A sample has been prepared, labelled with Cy5 (red), and hybridised to a microarray with approximately 23,000 features. The Cy3 (green) channel is a sample 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 $\log_2$

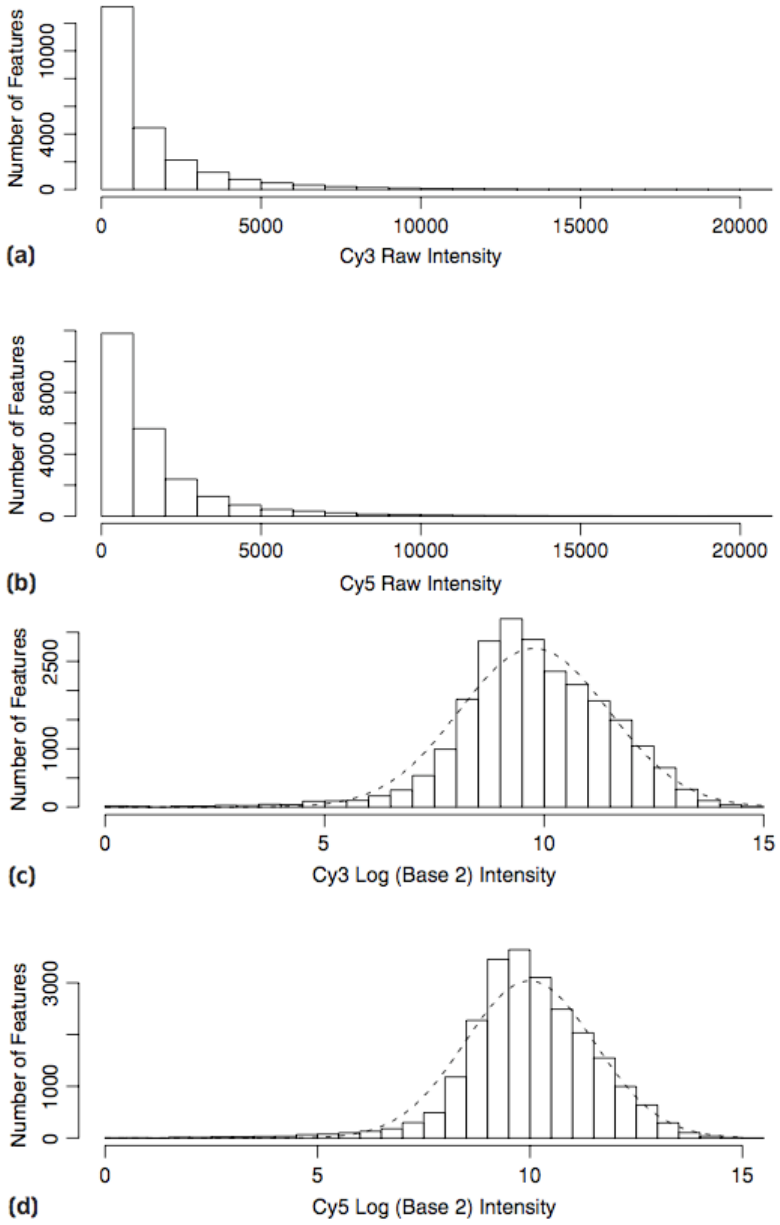


Figure 5.2: Histograms of the raw and log Cy3 and Cy5 intensities. Histograms of the intensities of the features for the human fibroblast data. **(a)** The raw intensities for the Cy3 channel; the data is right-skewed, with the majority of features having low intensity and decreasing numbers of features 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

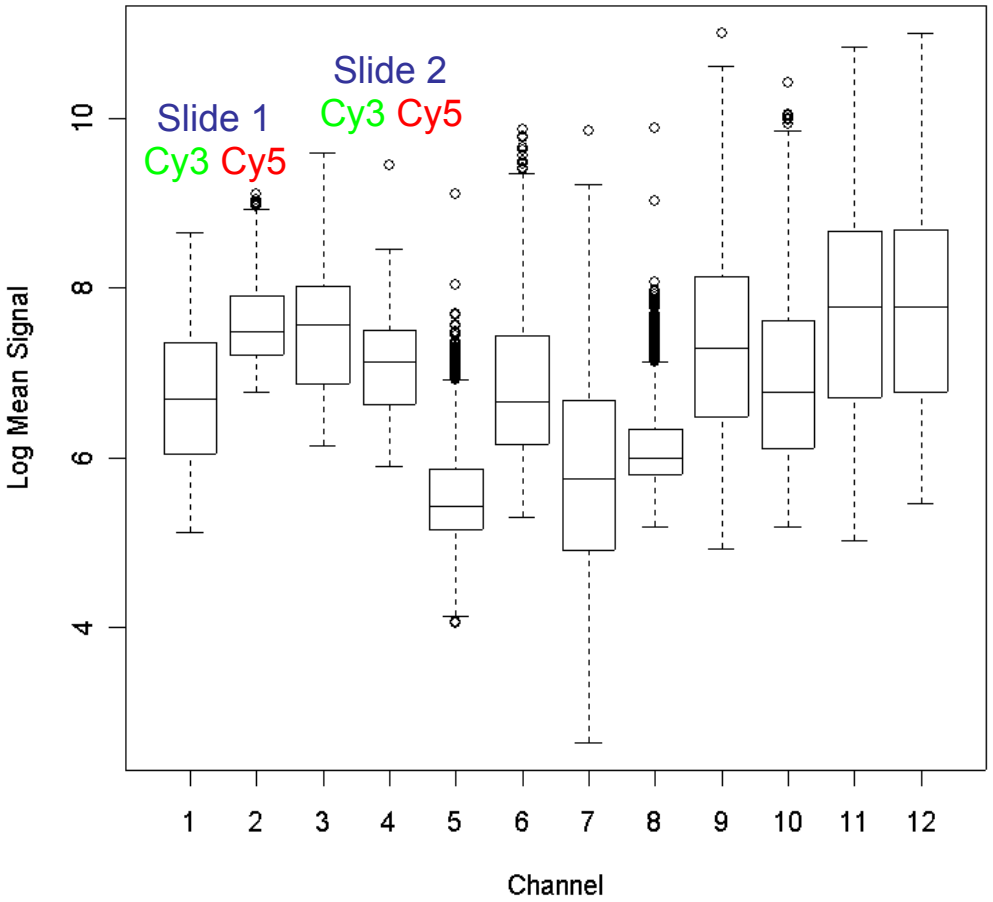
- Η κανονικοποίηση προσπαθεί να ελαχιστοποιήσει/εξαλείψει τον πειραματικό συστηματικό θόρυβο/διακύμανση, ώστε να αποκαλυφθούν οι διαφορές που έχουν βιολογική βάση.

- 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.



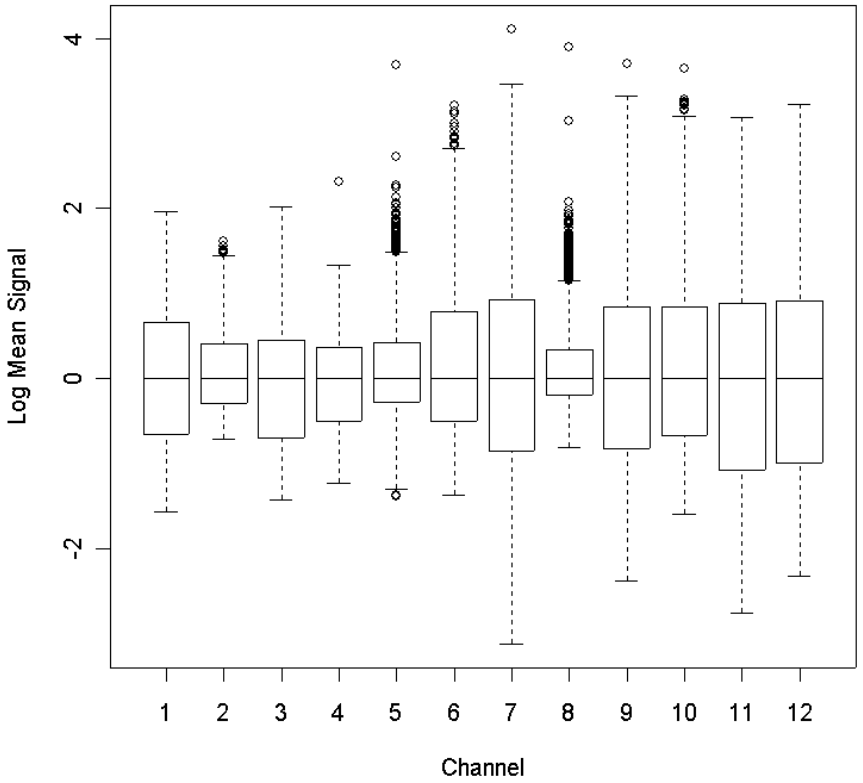
maximum

Q3=75th percentile

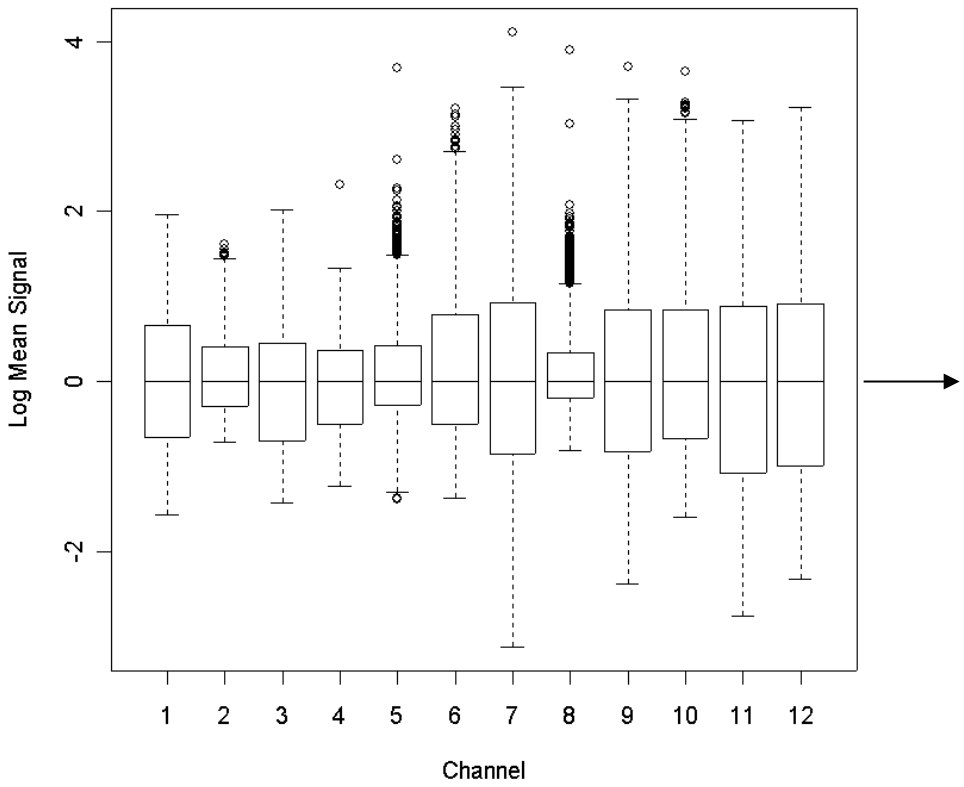
median

Q1=25th percentile

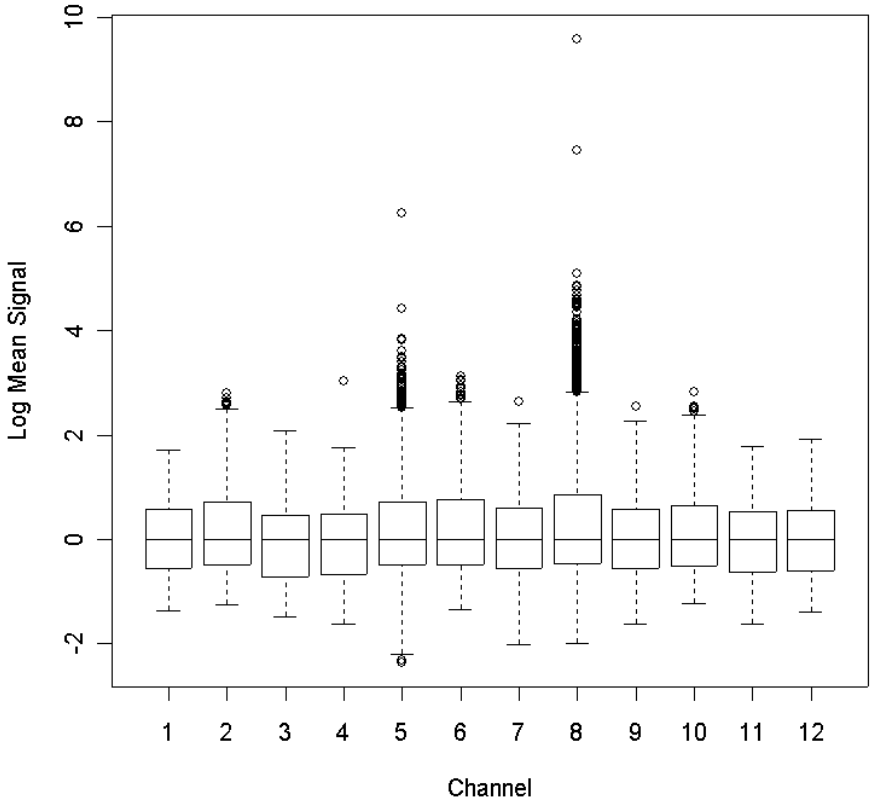
minimum



Κανονικοποίηση κλίμακας
Scale normalization



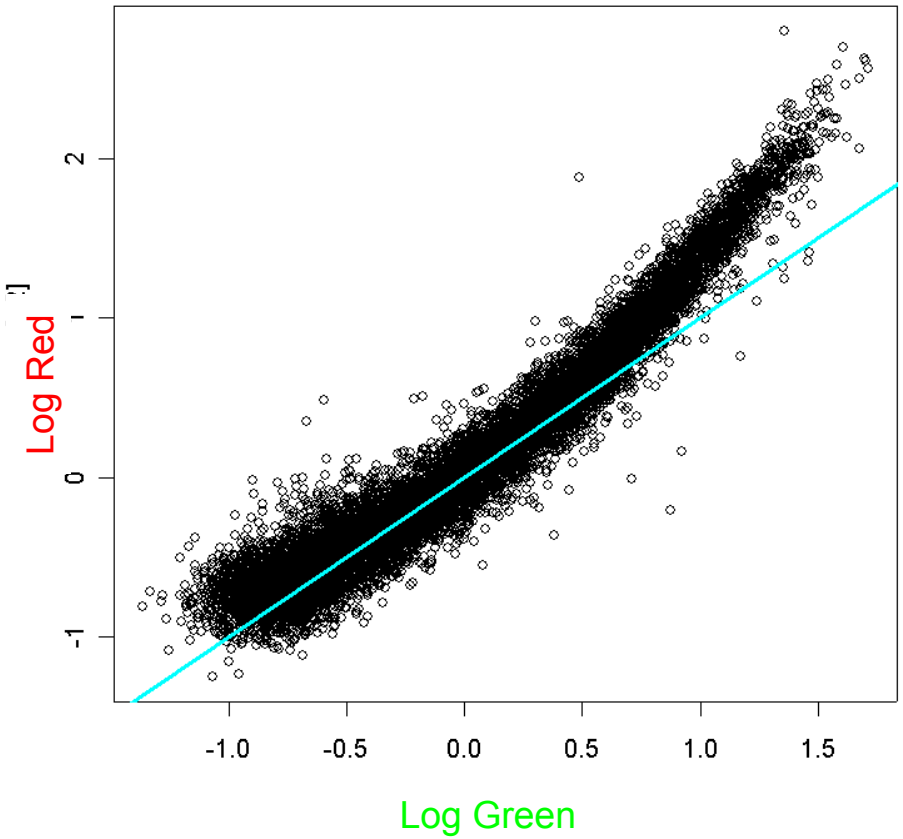
Data after Median Centering and Scale Normalizing



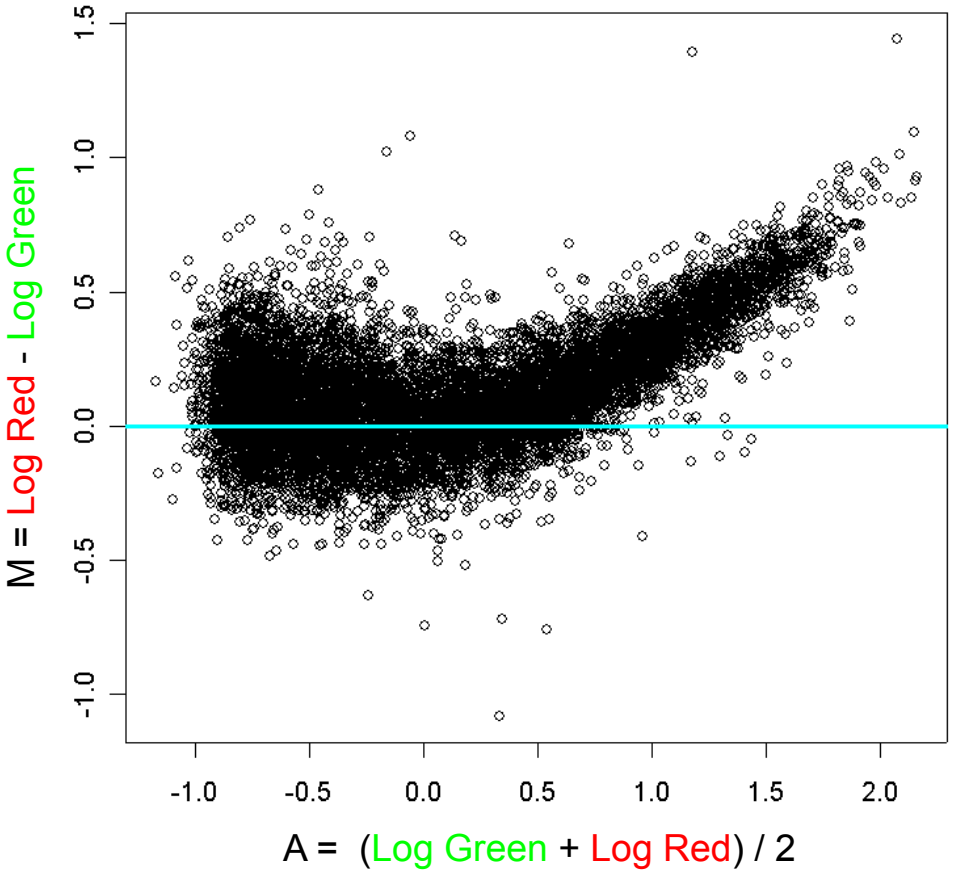
Συστηματικό λάθος στην χρώση

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

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

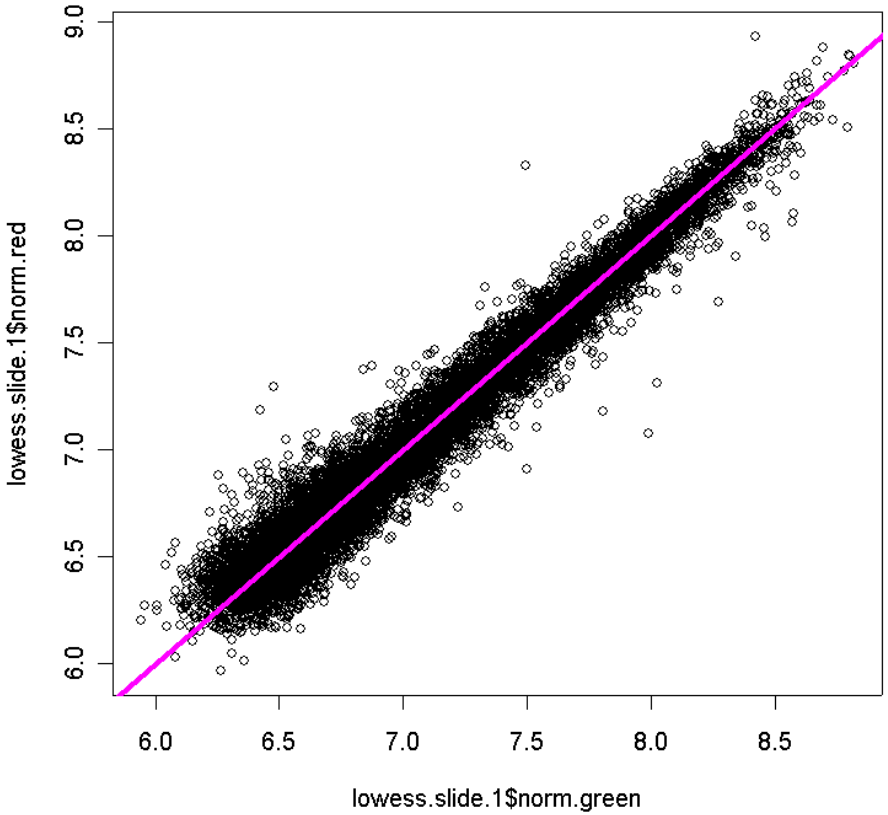


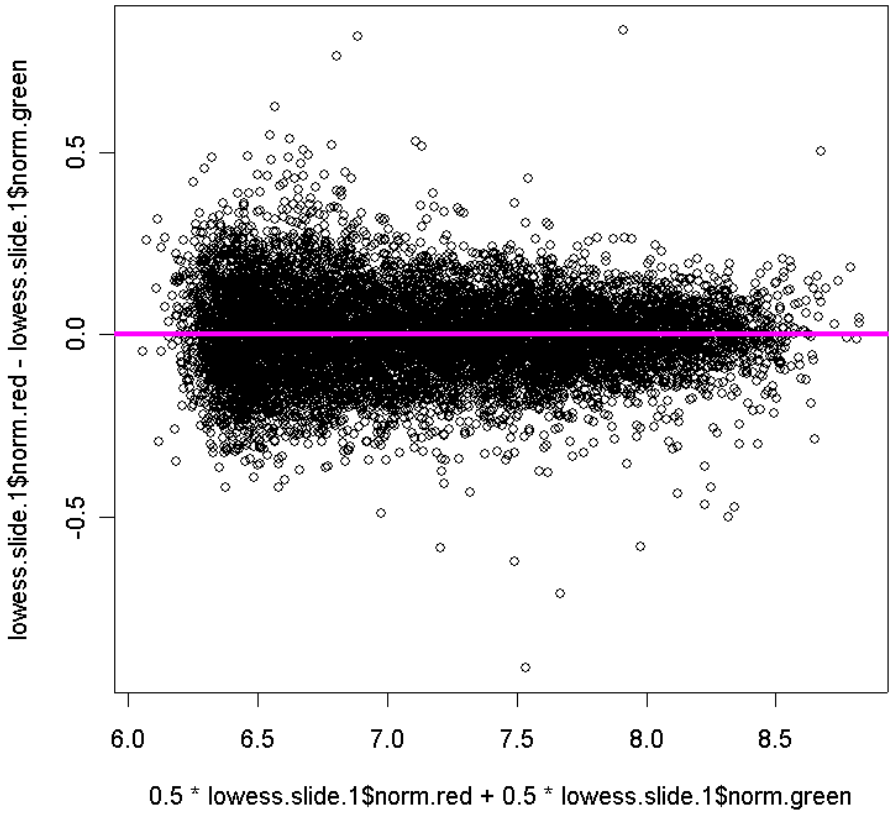
M vs. A Plot of the Logged, Centered, and Scaled Slide 1 Data



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)





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)

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

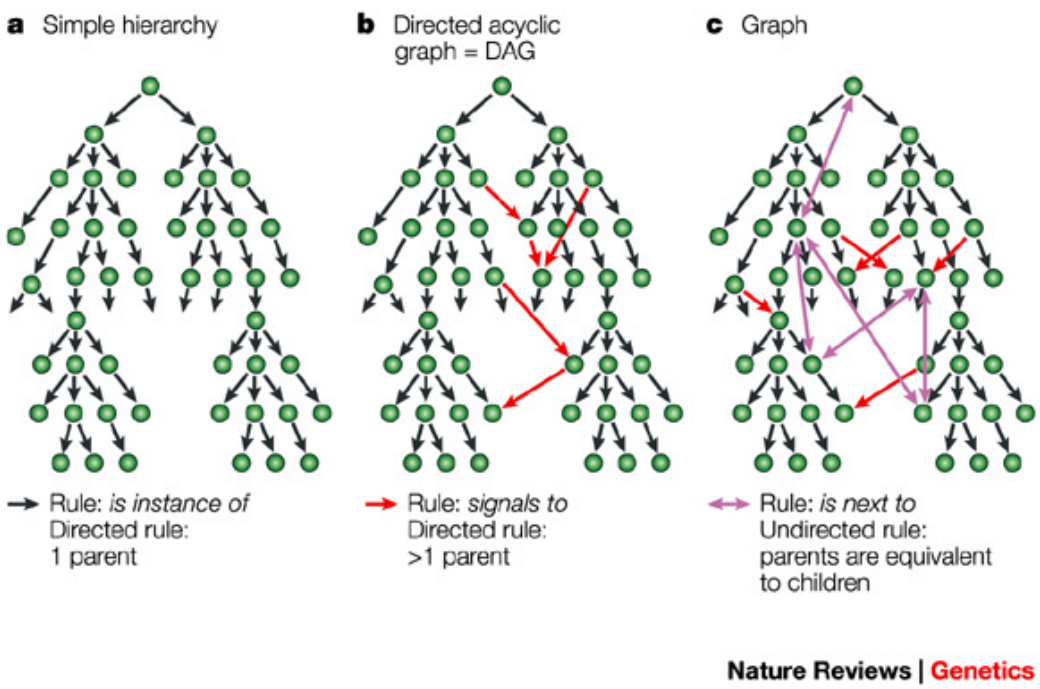
Use and misuse of the gene ontology annotations

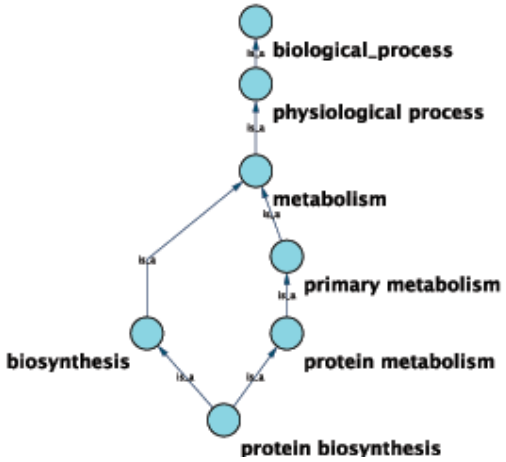
Seung Yon Rhee, Valerie Wood‡, Kara Dolinski§ and Sorin Draghici‖*

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

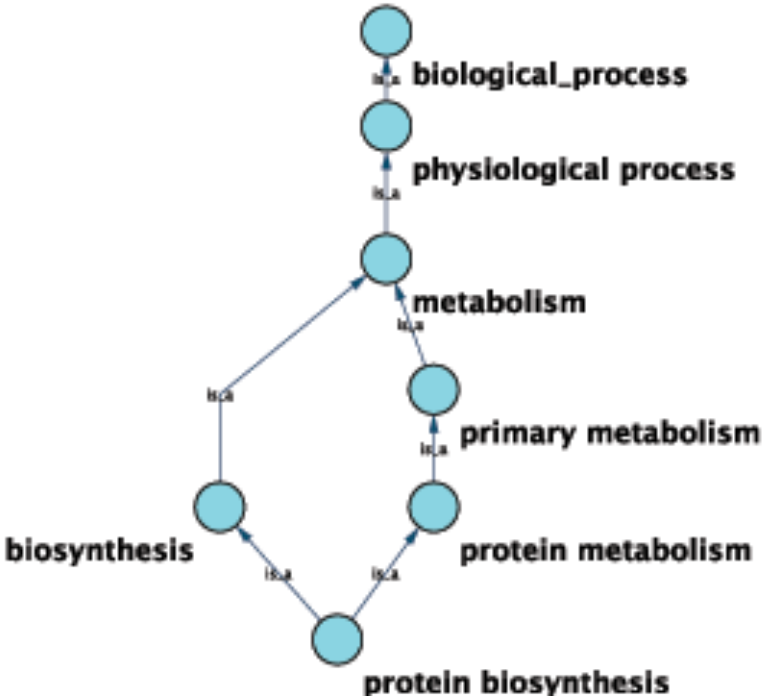
- Δείχνει τις σχέσεις μεταξύ των διαφορετικών όρων.
- Ένας όρος μπορεί να αποτελεί πιο εξειδικευμένη περιγραφή ενός άλλου όρου.
- Είναι κατευθυνόμενα ακυκλικά γραφήματα (DAG).
- Παρόμοια με ιεραρχίες.
- Η διαφορά είναι ότι ένας κόμβος-απόγονος μπορεί να έχει περισσότερους από έναν προγόνους.





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

NATURE REVIEWS | GENETICS VOLUME 9 | JULY 2008 | 509

REVIEWS

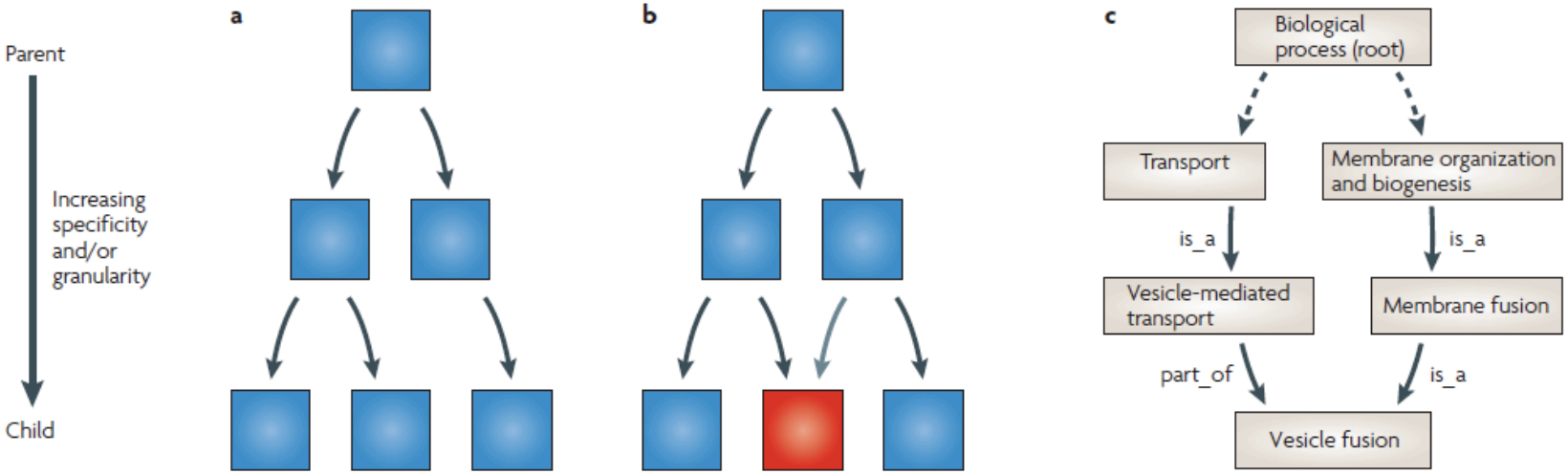


Figure 1 | **Simple trees versus directed acyclic graphs**. Boxes represent nodes and arrows represent edges. **a** | An example of a simple tree, in which each child has only one parent and the edges are directed, that is, there is a source (parent) and a destination (child) for each edge. **b** | A 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 leaf[41]. 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‖
Schizosaccharomyces pombe (4896)	4,482	4,930	90.9%	4,930	100%	90.9%
Saccharomyces cerevisiae (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%
Caenorhabditis elegans (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%
Arabidopsis thaliana# (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%
Candida albicans (5476)	806	3,756	21.4%	6,166	60.9%	13.0%
Pseudomonas aeruginosa 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%
Trypanosoma brucei (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%
Plasmodium falciparum (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

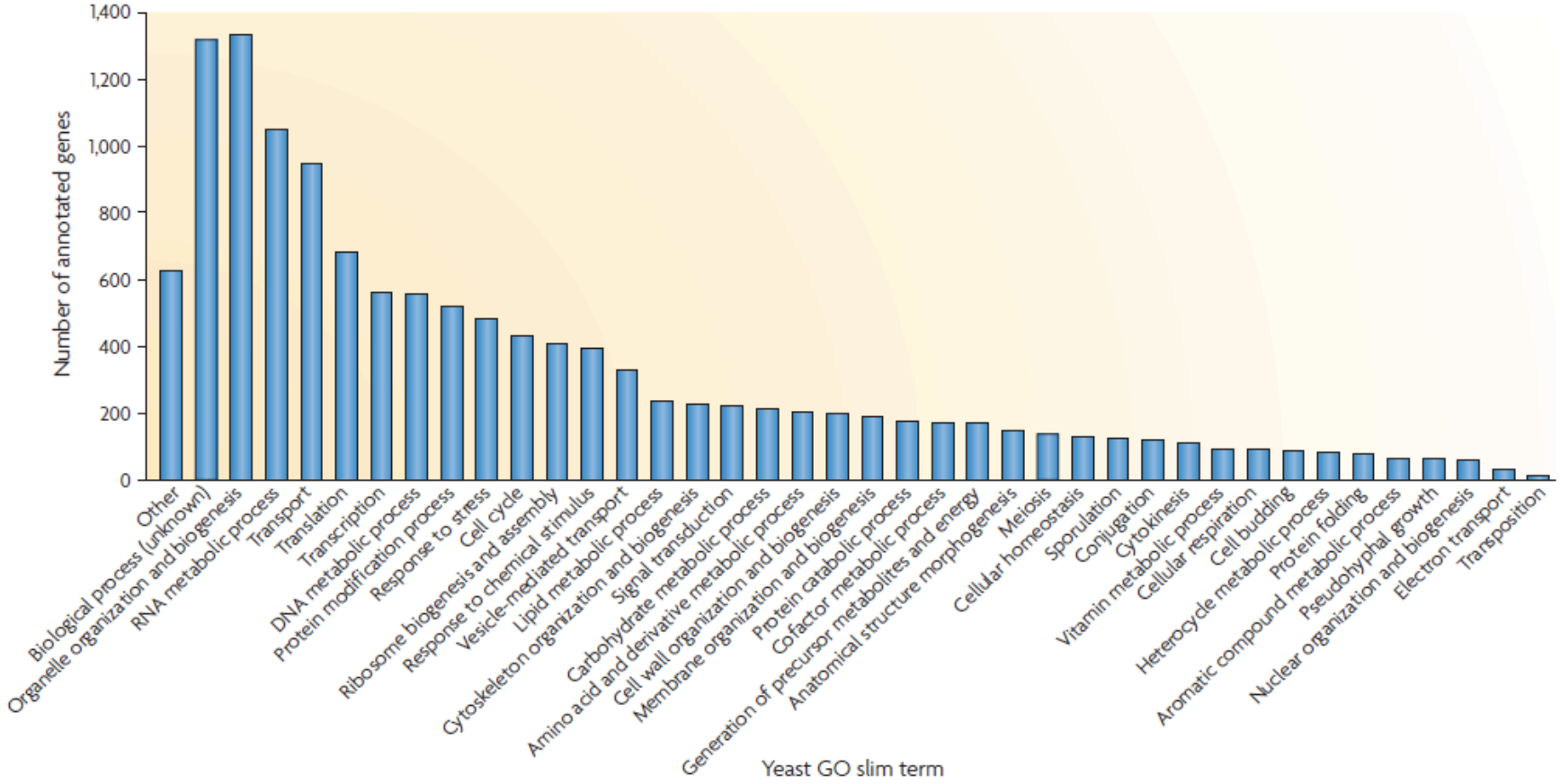


Figure 2 | **Using gene ontology (GO) to bin the yeast genome into broad biological process categories.** This example was generated by downloading the go_slim_mapping.tab file from the *Saccharomyces* genome database ftp site (dated 19 January 2008). This file maps every gene in the yeast genome to 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 Χ 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 Χ αριθμός γονιδίων υπερεκφράζονται.
  - Σε αυτό τον Χ 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-value < 0.05, τότε είναι στατιστικά σημαντικό.

- Αν στις οντολογίες μας είχαμε 100 όρους, θα επαναλαμβάναμε τα παραπάνω τεστς για τον κάθε όρο.
- Όμως, όσο περισσότερα τεστ κάνουμε για το πείραμά μας, τόσο αυξάνει ή πιθανότητα να βρούμε κάτι στατιστικά σημαντικό (p-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

In vitro διαγνωστικά τεστ που βασίζονται σε μικροσυστοιχίες

FDA: In Vitro Diagnostic Multivariate Index Assays (IVDMIAs)

- 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 IVDMIAs 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

- IVDMIAs 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, IVDMIAs frequently have a high risk intended use. FDA is concerned that patients are relying upon IVDMIAs with high risk intended uses to make critical healthcare decisions when FDA has not ensured that the IVDMIA 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 IVDMIA is safe and effective for its intended use.

Mammaprint - Tissue of origin

- http://www.ivdtechnology.com/article/exploring-fda-approved-ivdmias
- **MammaPrint.**
  The first IVDMIA, 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.