High throughput, multiplex analysis of gene expression*

David Goldman,
Laboratory of Neurogenetics, NIH
dgneuro@box-d.nih.gov

*With special thanks to Chiara Mazzanti, Ph.D.

Revolutions in the analysis of gene expression

Protein function

Protein concentration

Post-translational modification

Transcriptional control

Genetic variation

Genomics

Multiplex gene analysis: Why?

Complex diseases involve expression modulation

40,000 genes, perhaps more

We are not so clever

The ultimate goal is to reconstruct the biochemical and functional networks of the cell. One can not understand a network by studying one gene.

Methods for multiplex expression analysis

Two dimensional protein electrophoresis [2DE]

Differential display

Representational difference analysis [RDA]

Serial analysis of gene expression [SAGE]

RNA array methods

Array Systems

cDNA

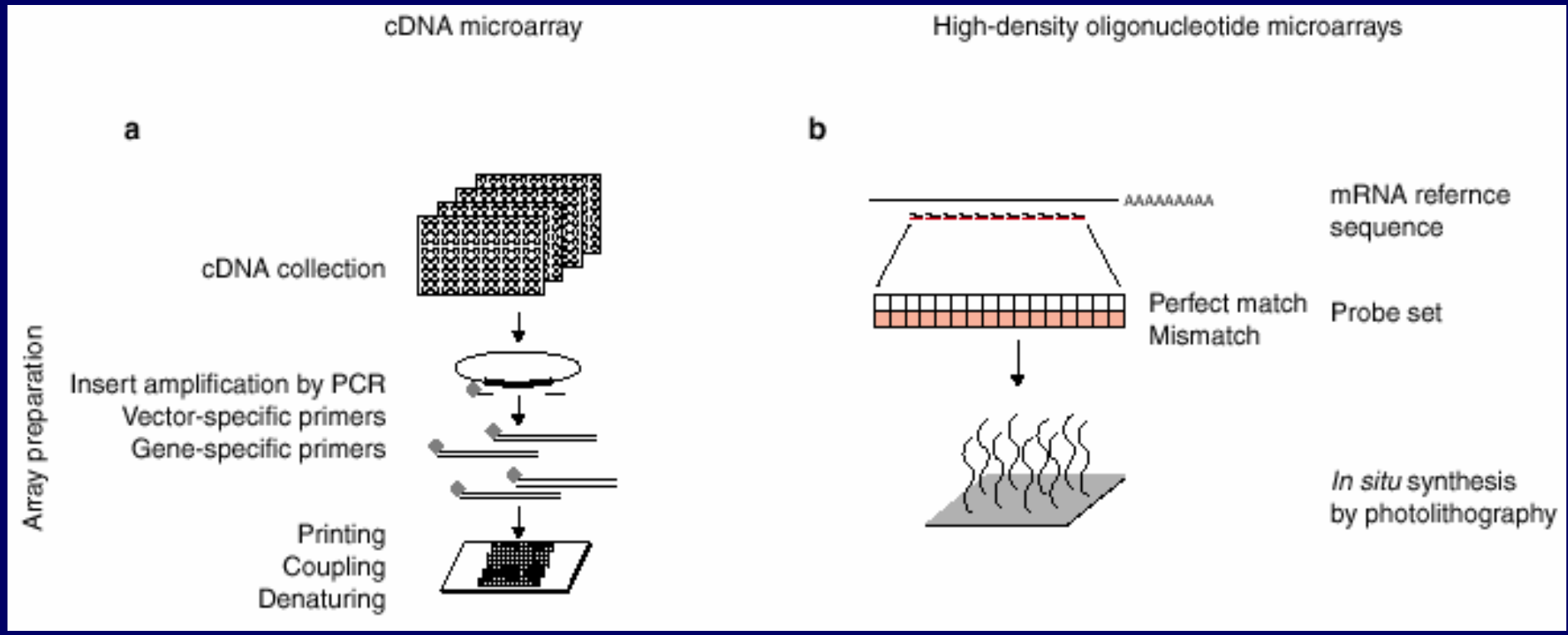
Probes: amplified spotted cDNAs
http://cmgm.standford.edu
http://www.nhgri.nih.gov

Oligonucleotide chips

Probes: Oligos synthesized *in situ* or conventionally & followed by immobilization
Oligos as long as 80bp
http://www.affymetrix.com
http://www.rii.com (Agilent)

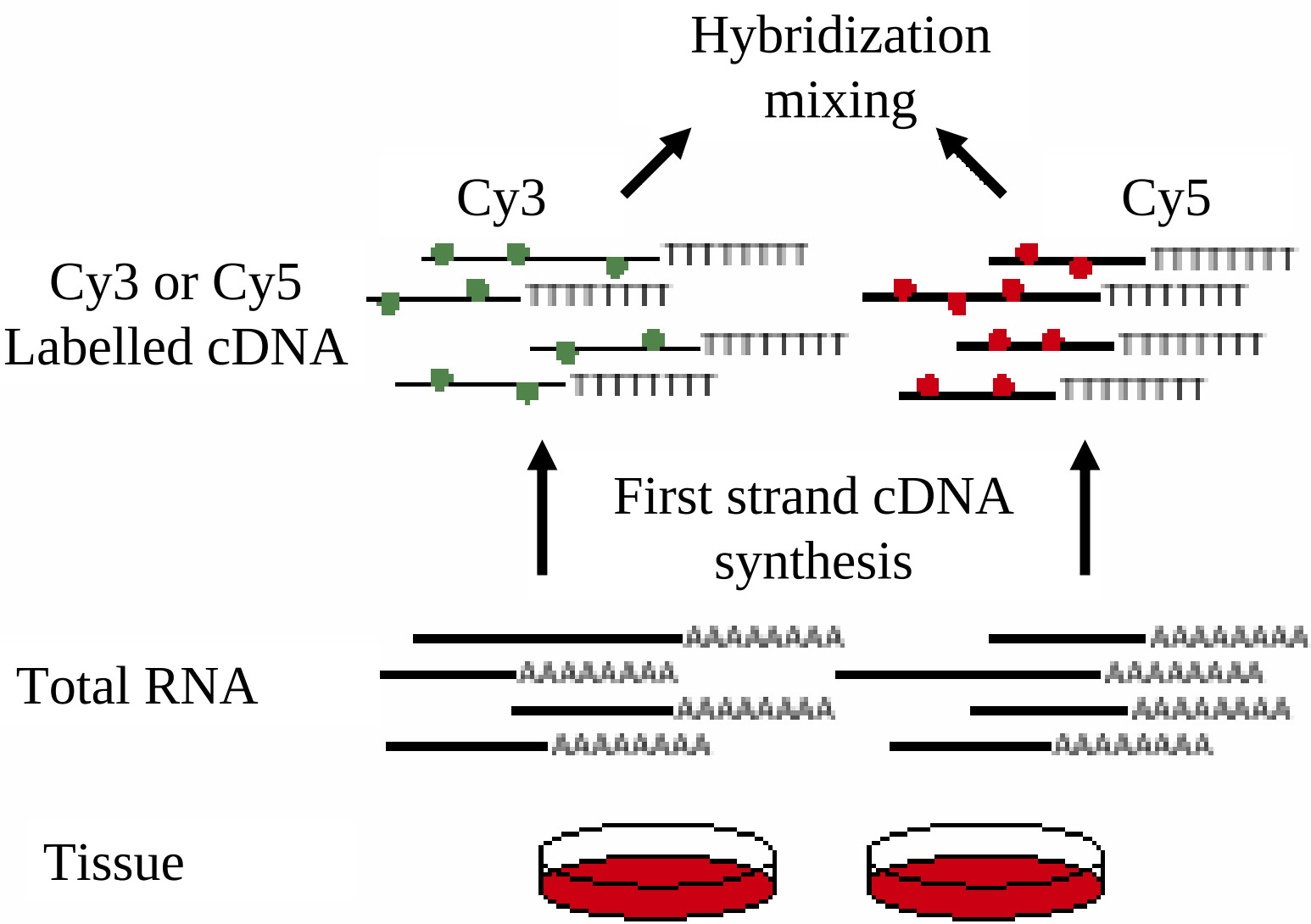
Array preparation

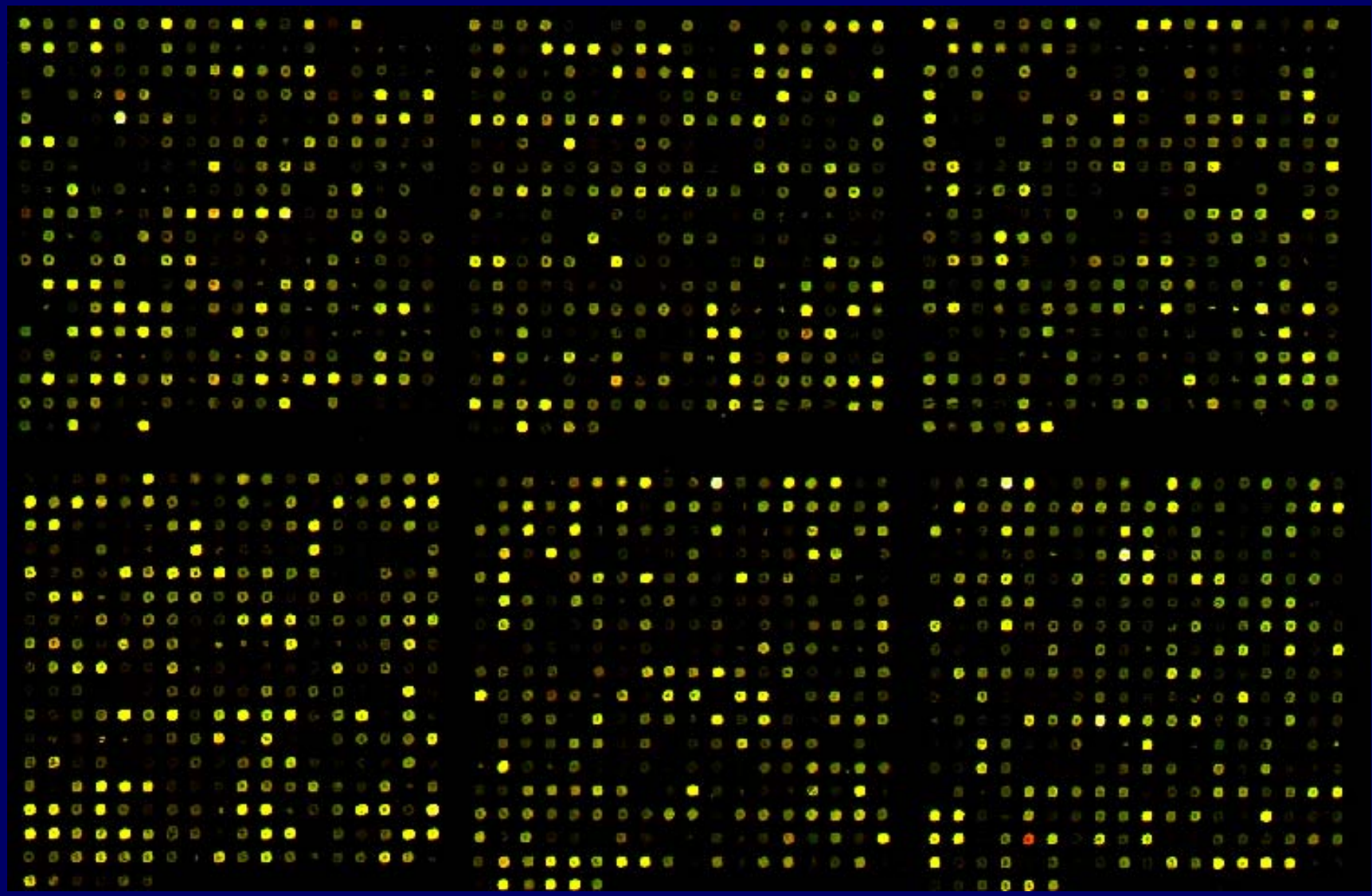
Microarray Genechip

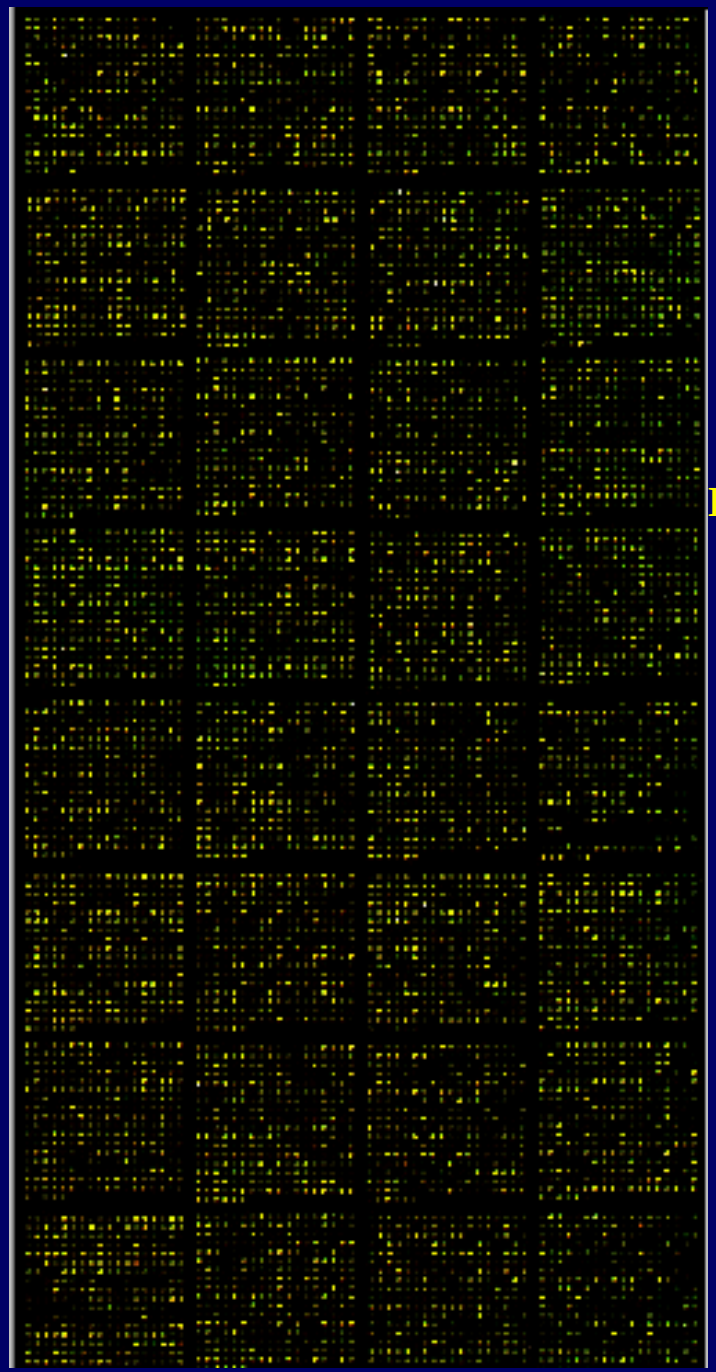


Available genes: 10k – NCI [$50] 20k - Affymetrix

Sample preparation (direct labeling)

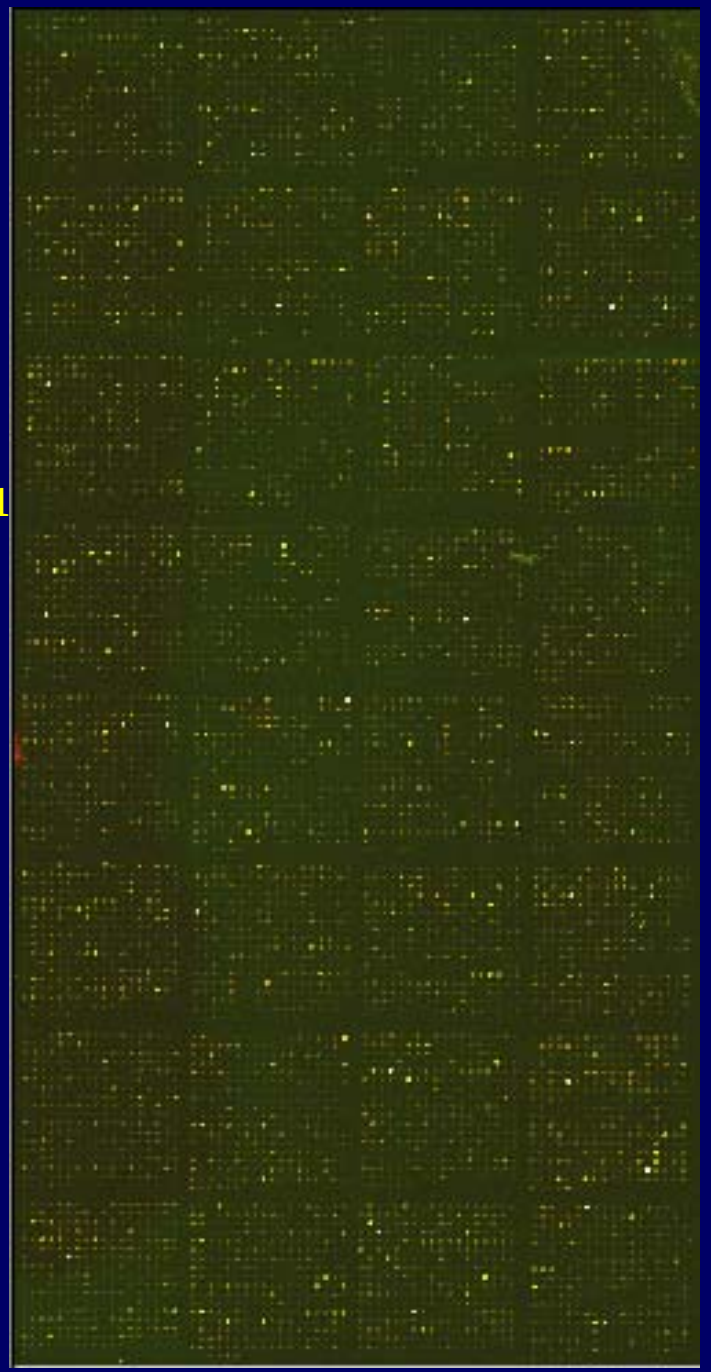






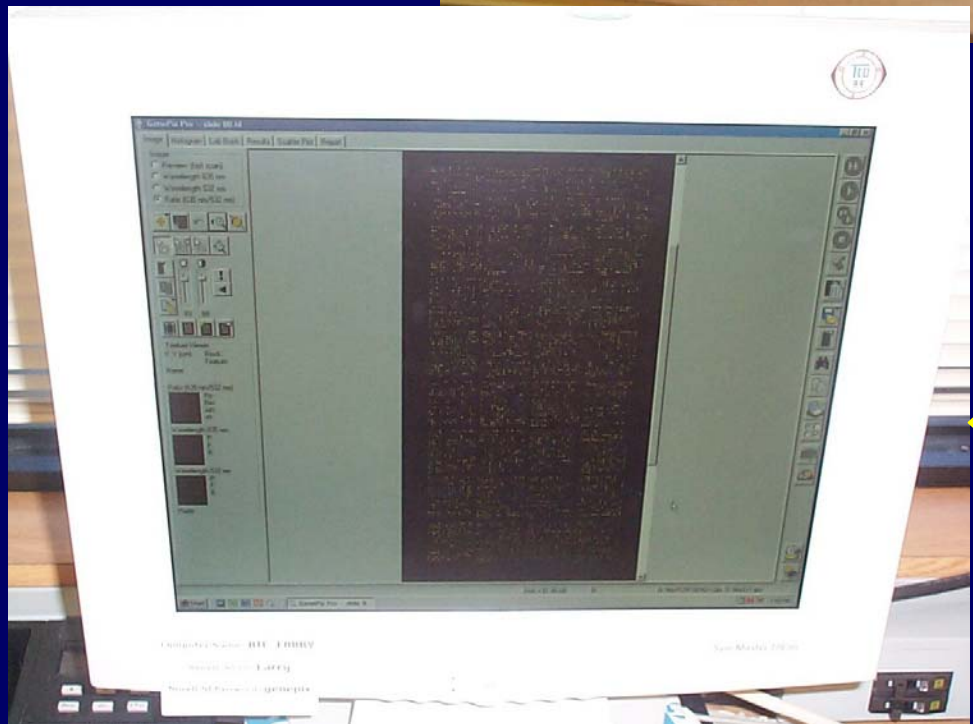
10k Arrays

Hs-UniGem2-v2p9-060701
Blocks: (18x18) x32
Features/Spots: 9984
235um spacing



Scanner
GenePix 4000A

Post-hybridization microarray analysis



GenePix Pro software

Typical software (GenPix Pro)

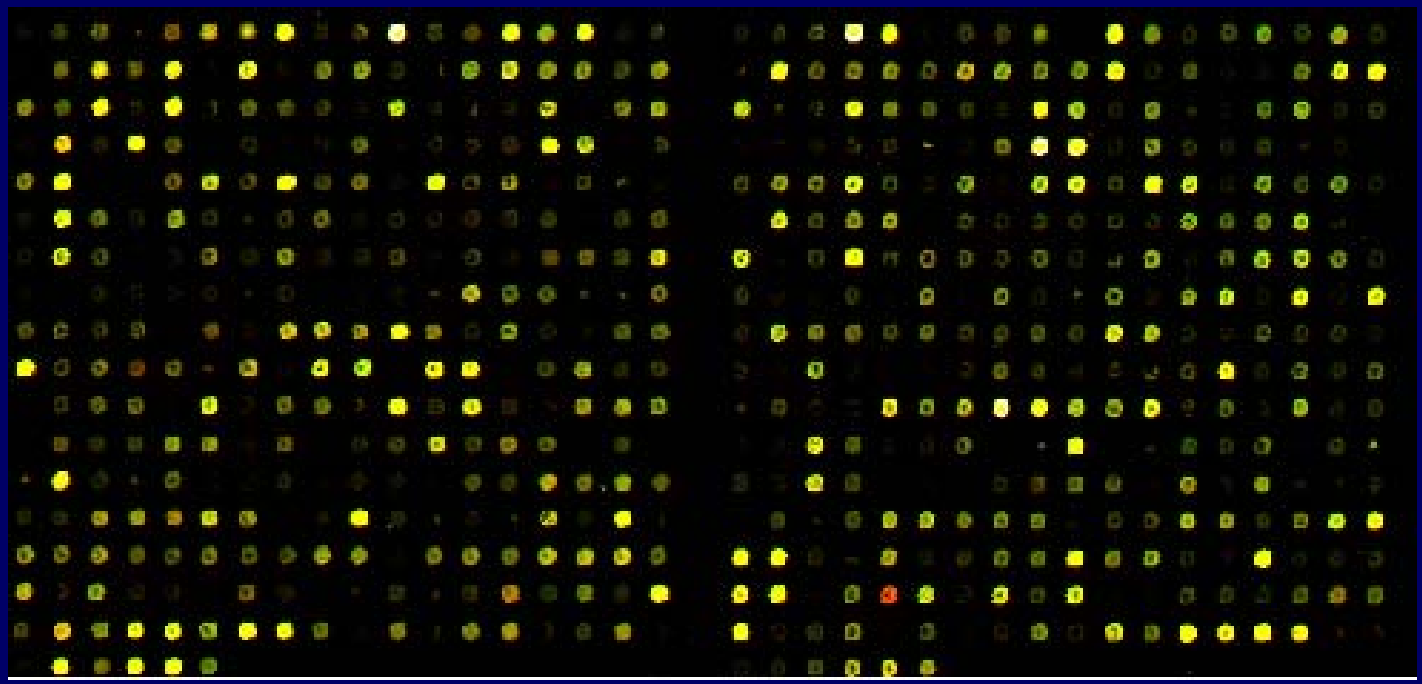
Image scanning

Cy3/Cy5 signal normalization

Image cleaning

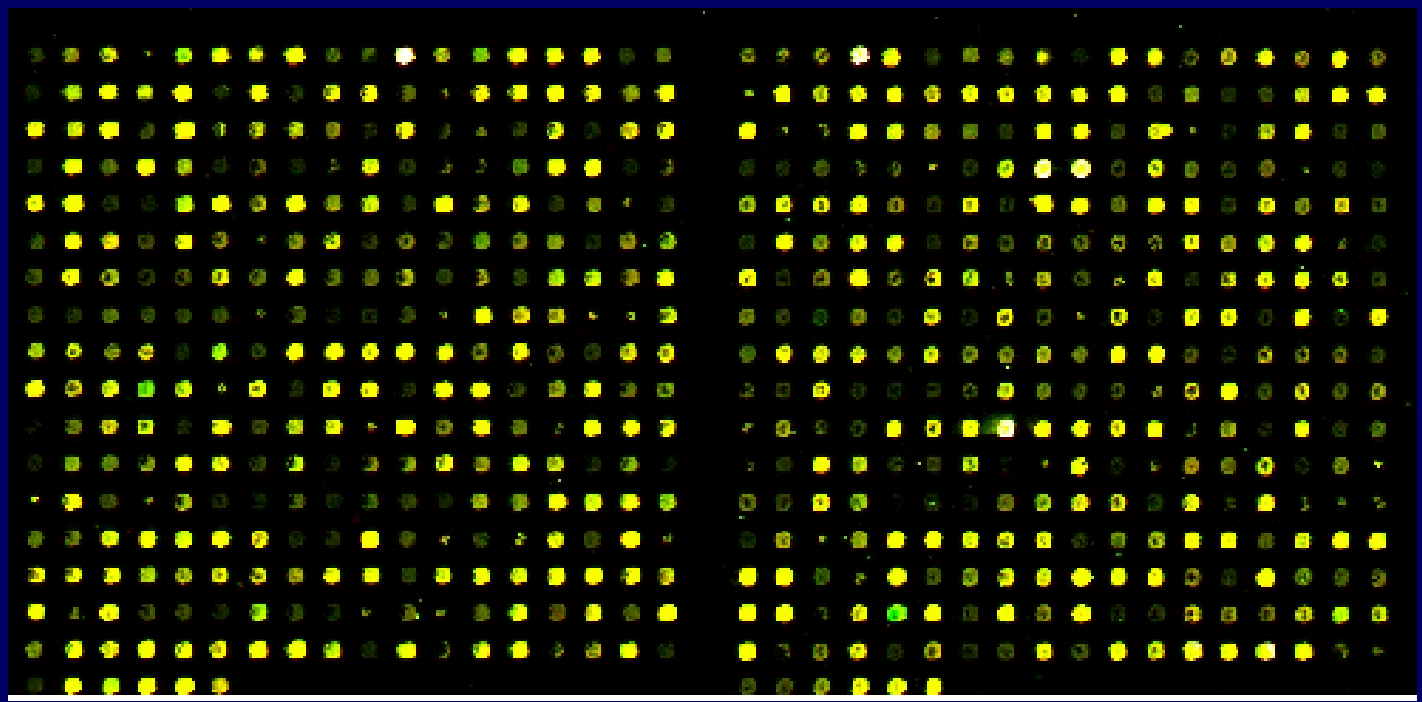
GAL (gene array list) loading

Image analysis



Case Cy5
Control Cy3

Reciprocal labelling



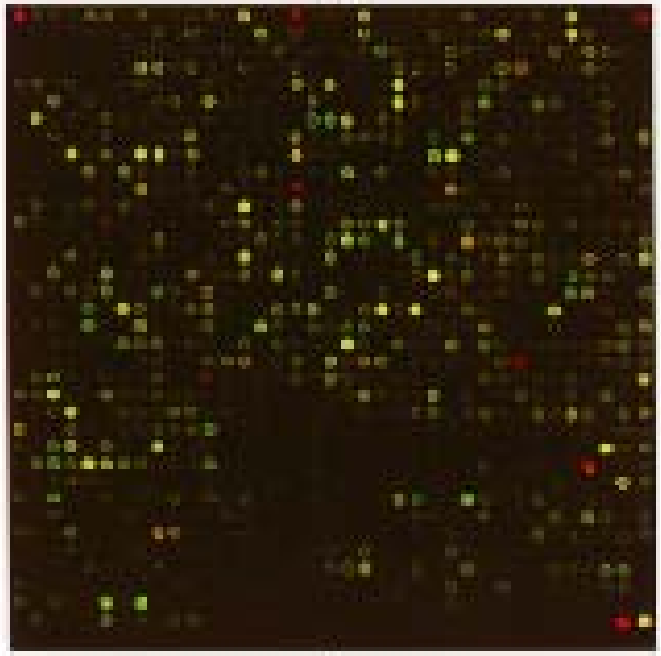
Case Cy3
Control Cy5

Array vs Chip

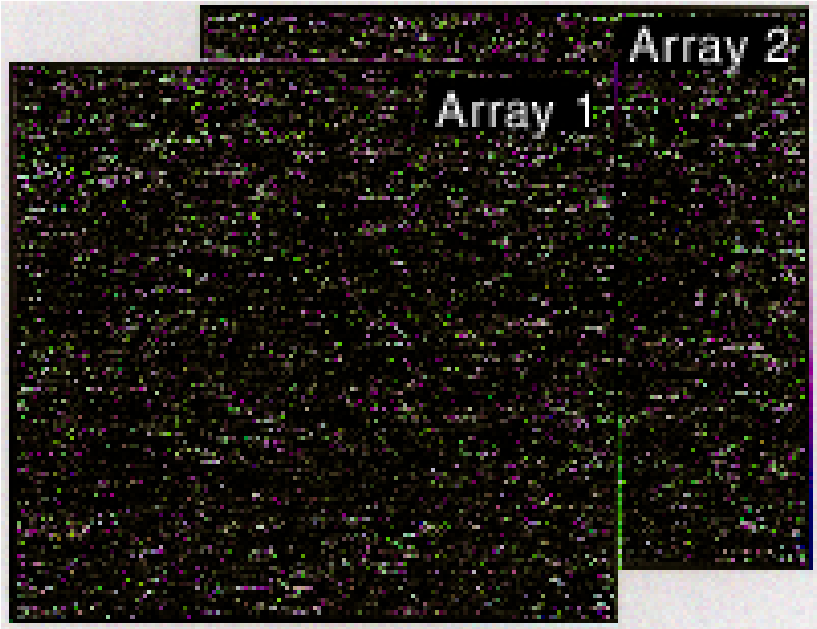
cDNA Microarray

Ratio Cy5/Cy3

Ratio >2, < 0.5

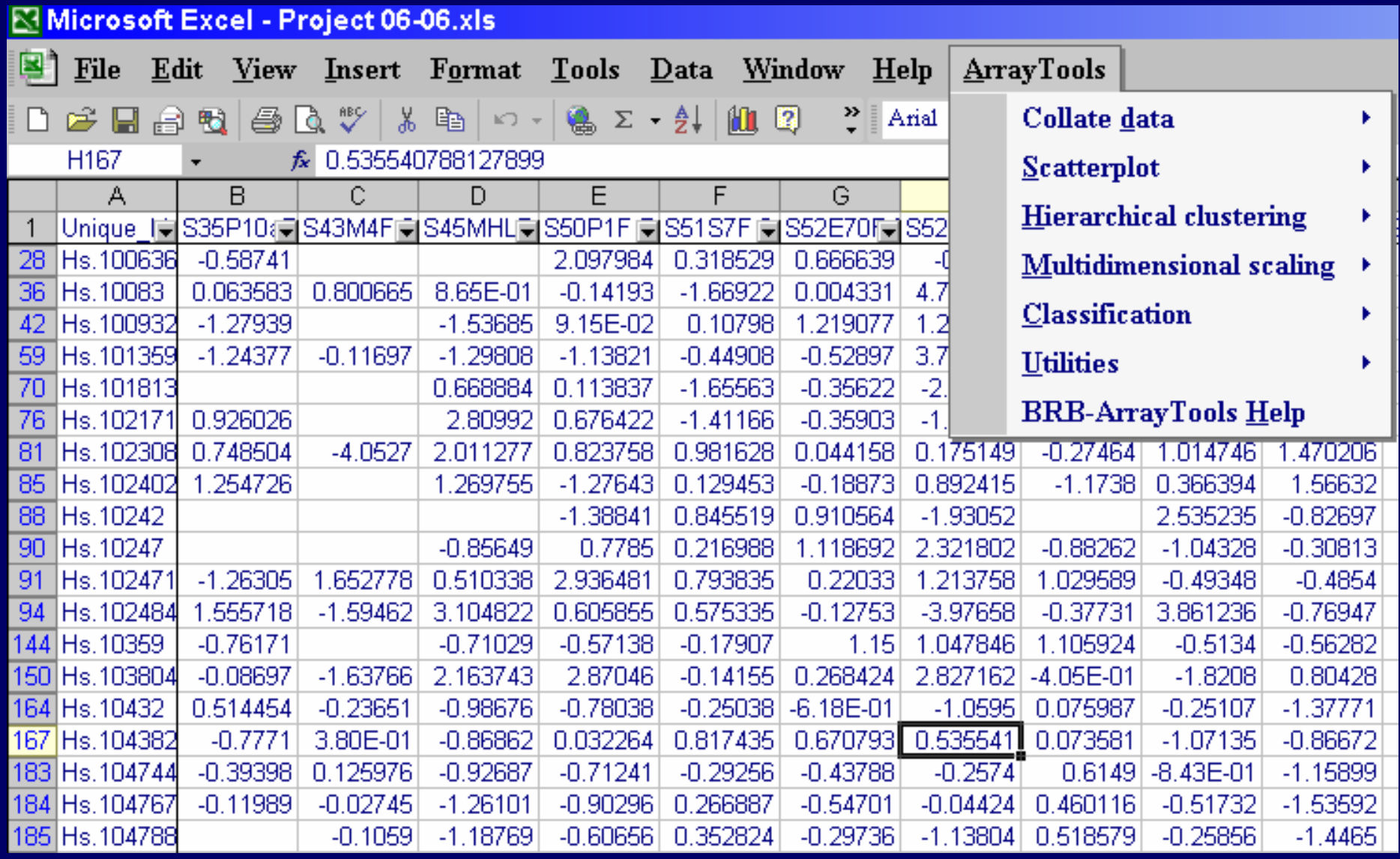


Case Cy5
Control Cy3



Oligo-microchip

Ratio Array1/Array2



	A	B	C	D	E	F	G				
1	Unique_I	S35P10	S43M4F	S45MHL	S50P1F	S51S7F	S52E70F	S52			
28	Hs.100636	-0.58741			2.097984	0.318529	0.666639	-0			
36	Hs.10083	0.063583	0.800665	8.65E-01	-0.14193	-1.66922	0.004331	4.7			
42	Hs.100932	-1.27939		-1.53685	9.15E-02	0.10798	1.219077	1.2			
59	Hs.101359	-1.24377	-0.11697	-1.29808	-1.13821	-0.44908	-0.52897	3.7			
70	Hs.101813			0.668884	0.113837	-1.65563	-0.35622	-2.			
76	Hs.102171	0.926026		2.80992	0.676422	-1.41166	-0.35903	-1.			
81	Hs.102308	0.748504	-4.0527	2.011277	0.823758	0.981628	0.044158	0.175149	-0.27464	1.014746	1.470206
85	Hs.102402	1.254726		1.269755	-1.27643	0.129453	-0.18873	0.892415	-1.1738	0.366394	1.56632
88	Hs.10242				-1.38841	0.845519	0.910564	-1.93052		2.535235	-0.82697
90	Hs.10247			-0.85649	0.7785	0.216988	1.118692	2.321802	-0.88262	-1.04328	-0.30813
91	Hs.102471	-1.26305	1.652778	0.510338	2.936481	0.793835	0.22033	1.213758	1.029589	-0.49348	-0.4854
94	Hs.102484	1.555718	-1.59462	3.104822	0.605855	0.575335	-0.12753	-3.97658	-0.37731	3.861236	-0.76947
144	Hs.10359	-0.76171		-0.71029	-0.57138	-0.17907	1.15	1.047846	1.105924	-0.5134	-0.56282
150	Hs.103804	-0.08697	-1.63766	2.163743	2.87046	-0.14155	0.268424	2.827162	-4.05E-01	-1.8208	0.80428
164	Hs.10432	0.514454	-0.23651	-0.98676	-0.78038	-0.25038	-6.18E-01	-1.0595	0.075987	-0.25107	-1.37771
167	Hs.104382	-0.7771	3.80E-01	-0.86862	0.032264	0.817435	0.670793	0.535541	0.073581	-1.07135	-0.86672
183	Hs.104744	-0.39398	0.125976	-0.92687	-0.71241	-0.29256	-0.43788	-0.2574	0.6149	-8.43E-01	-1.15899
184	Hs.104767	-0.11989	-0.02745	-1.26101	-0.90296	0.266887	-0.54701	-0.04424	0.460116	-0.51732	-1.53592
185	Hs.104788		-0.1059	-1.18769	-0.60656	0.352824	-0.29736	-1.13804	0.518579	-0.25856	-1.4465

http://linus.nci.nih.gov/cgi-bin/brb/download.cgi

Center for Cancer Research NCI

CIT Center for Information Technology

mAdb System Home Page

Page Updated: Friday, 07-Jun-2002 10:42:00 EDT

Welcome to the *mAdb* (aka *Mad Bee*) Home page.
CIT/BIMAS is collaborating with NCI/CCR in the development of the BioInformatics to manage, access and analyze cDNA µArray data generated by the NCI/CCR µArray Center.

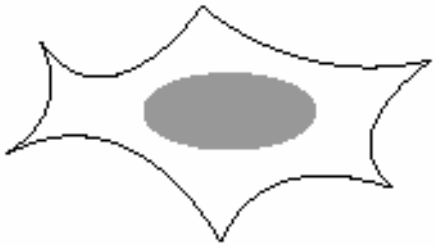
mAdb Amusement

NOTICE mAdb Scheduled system OUTAGE
On Saturday, Jun 15, 2002, the AFS fileservers will be upgraded. The mAdb WEB server depends on these servers and will be unavailable during the upgrade. The outage is scheduled to begin at 9am and last until 3pm (6 hours).
Posted 8:30am 6/11/2002

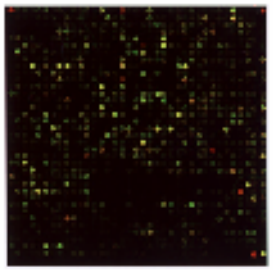
- Gateway - Data Upload, Access and Analysis Tools
  (Note: Must be a registered mAdb user - Login/Password required.)
- Array Ordering/Tracking WEBsite
  (Note: Different WEBsite Administered by the ATC, a separate login needed)
- mAdb Account Request - Request a new user account.
- mAdb Training via CIT
  - mAdb Basic Informatics Course - Description and Sign up for 2 hour introductory class

LOCAL APPROACH

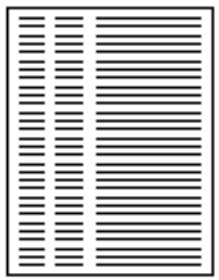
Condition/phenotype



Experiment



Microarray analysis



List of
up- and downregulated genes

Identification of genes
potentially involved in
condition/phenotype

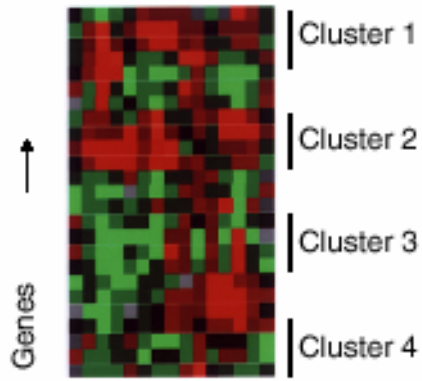
Analysis of selected genes
Overexpression
Gene knockout
RNAi

GLOBAL APPROACH

Different conditions
or phenotypes

Expression profiling

Compendium of data from different experiments



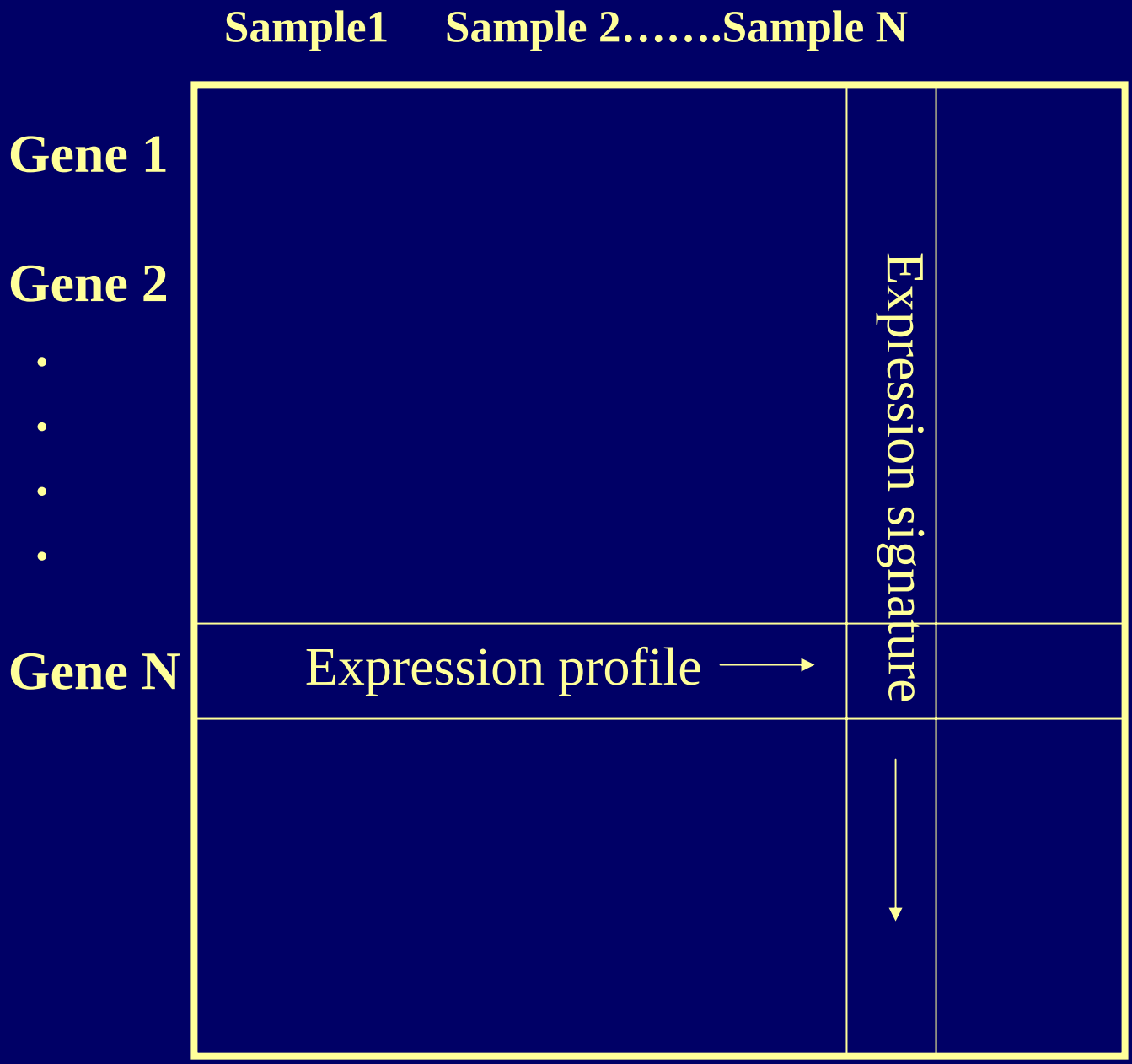
Identification of:
Co-regulated genes
Genes with related functions
Pathway signatures
Tumour classification

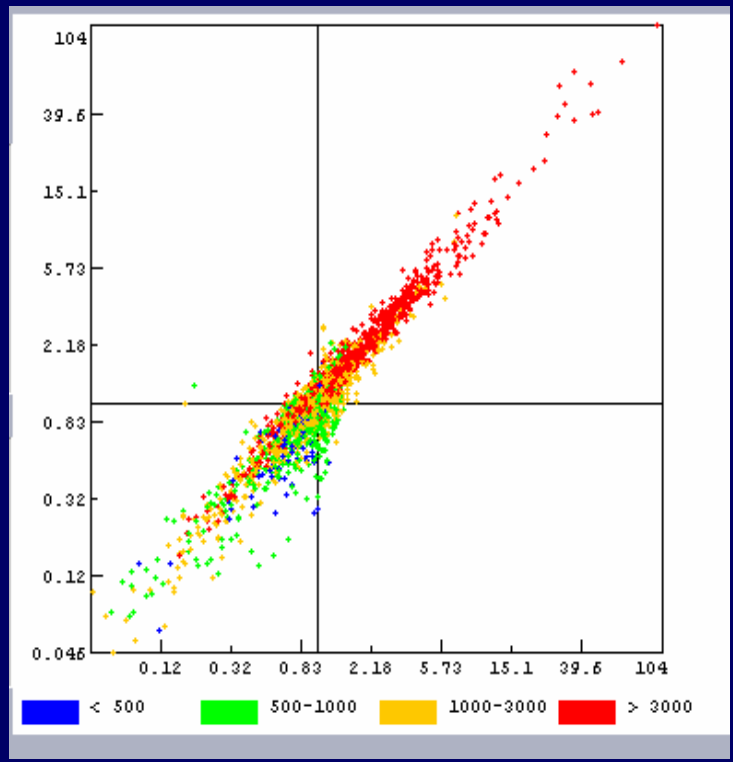
Schulze et al, Nature Cell Biology, August 2001

Array Analysis Methods

- **Gene Discovery**
  - Outlier detection – simple and group logic retrieval tools; single and multiple array viewers
  - Scatter plots
- **Pattern Discovery**
  - Clustering – Hierarchical, K-means
  - Principal Components Analysis; Multidimensional Scaling
- **Pattern Prediction**

Expression signatures and profiles




104
39.6
15.1
5.73
2.18
0.83
0.32
0.12
0.046
0.12
0.32
0.83
2.18
5.73
15.1
39.6
104
< 500
500-1000
1000-3000
> 3000

What is Clustering?

Clustering algorithms are discovery algorithms that find structure in data.

Discovery algorithms develop their own sorting parameters without human intervention.

Hierarchical Clustering

mAdb Hierarchical Tool Form 1.00

Note the --🐝 marks items which lead to additional help when clicked

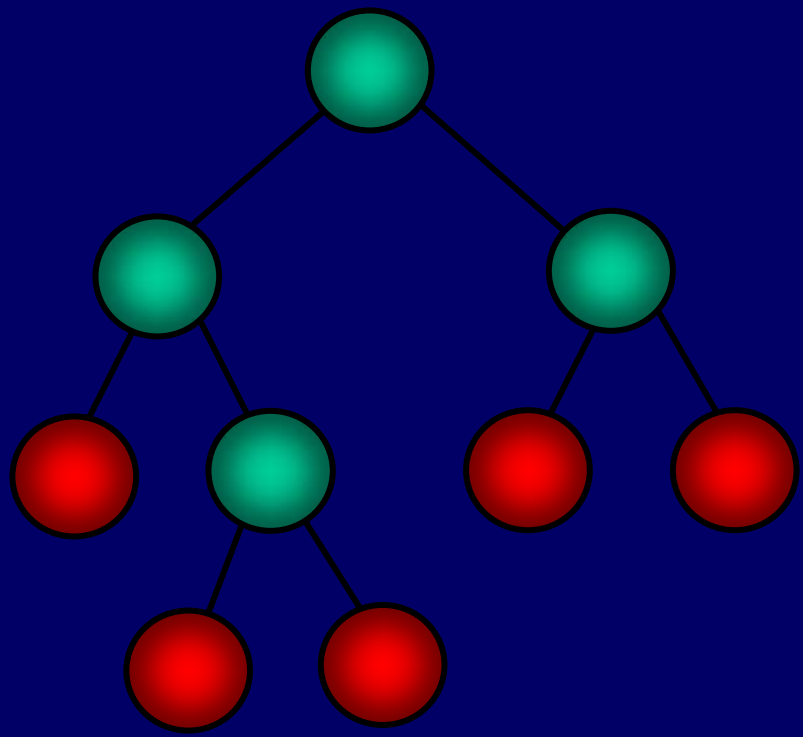
Hierarchical Clustering Options --🐝

Genes	Non-centered Metric
Arrays	Non-centered Metric
Distance Metric	Pearson Correlation
Name (optional)	DEMO

Cluster

Home | Analysis Tools | Forums | Reference Info | Program Downloads | GeneCards

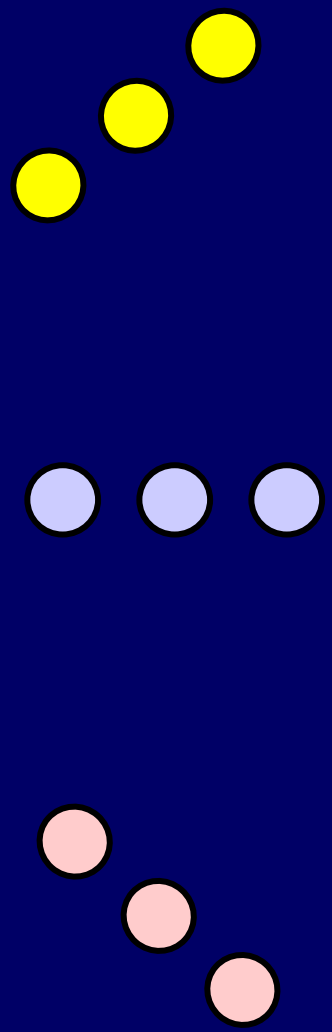
Hierarchical clustering



Mock phylogenetic dendrograms

Dendrogram Construction for Hierarchical Clustering

- Merge nearest neighbors
- Merge more distant datapoints



K-means clustering

Genes are divided into k user-defined, equally-sized groups.

Gene group centroids and individual genes are then reassigned.

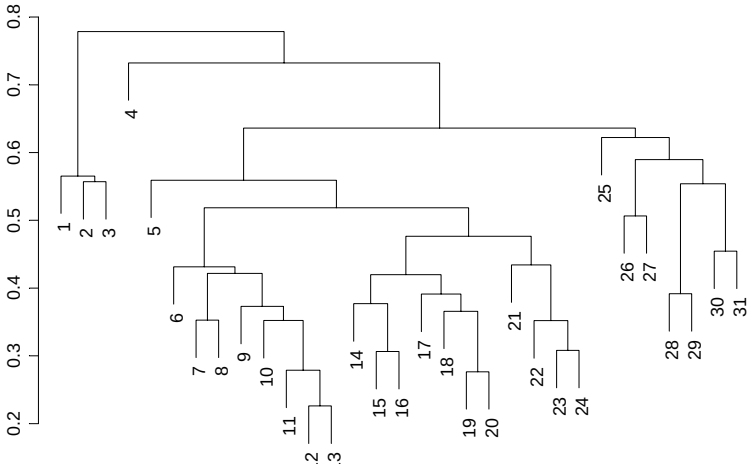
The process is reiterated until group compositions stabilize.

Clustering

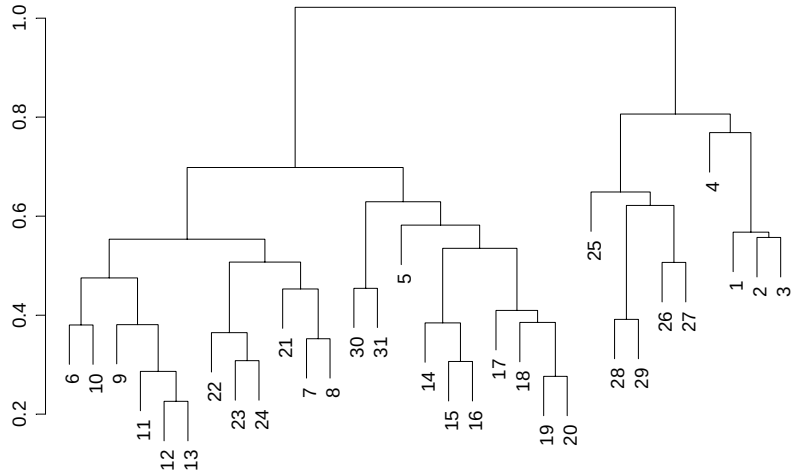
Clusters are also developed from random data:

- Are clustered genes related in function?
- Do clusters group related samples/tissues/diseases/treatments?

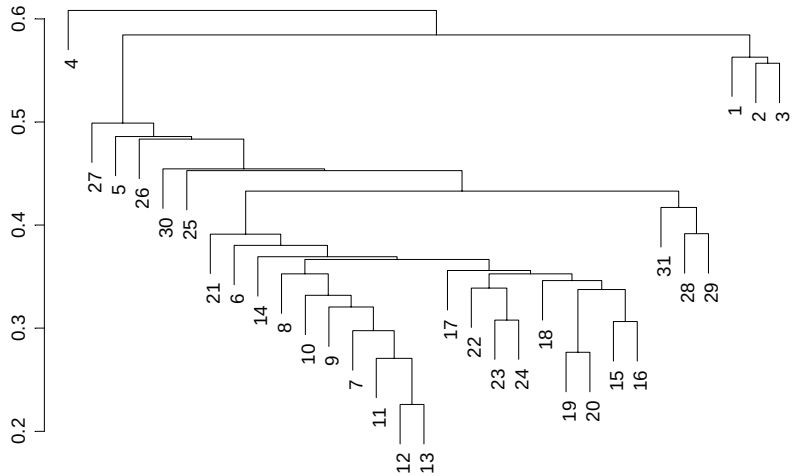
Clustering of Melanoma Tumors Using Average Linkage



Clustering of Melanoma Tumors Using Complete Linkage



Clustering of Melanoma Tumors Using Single Linkage



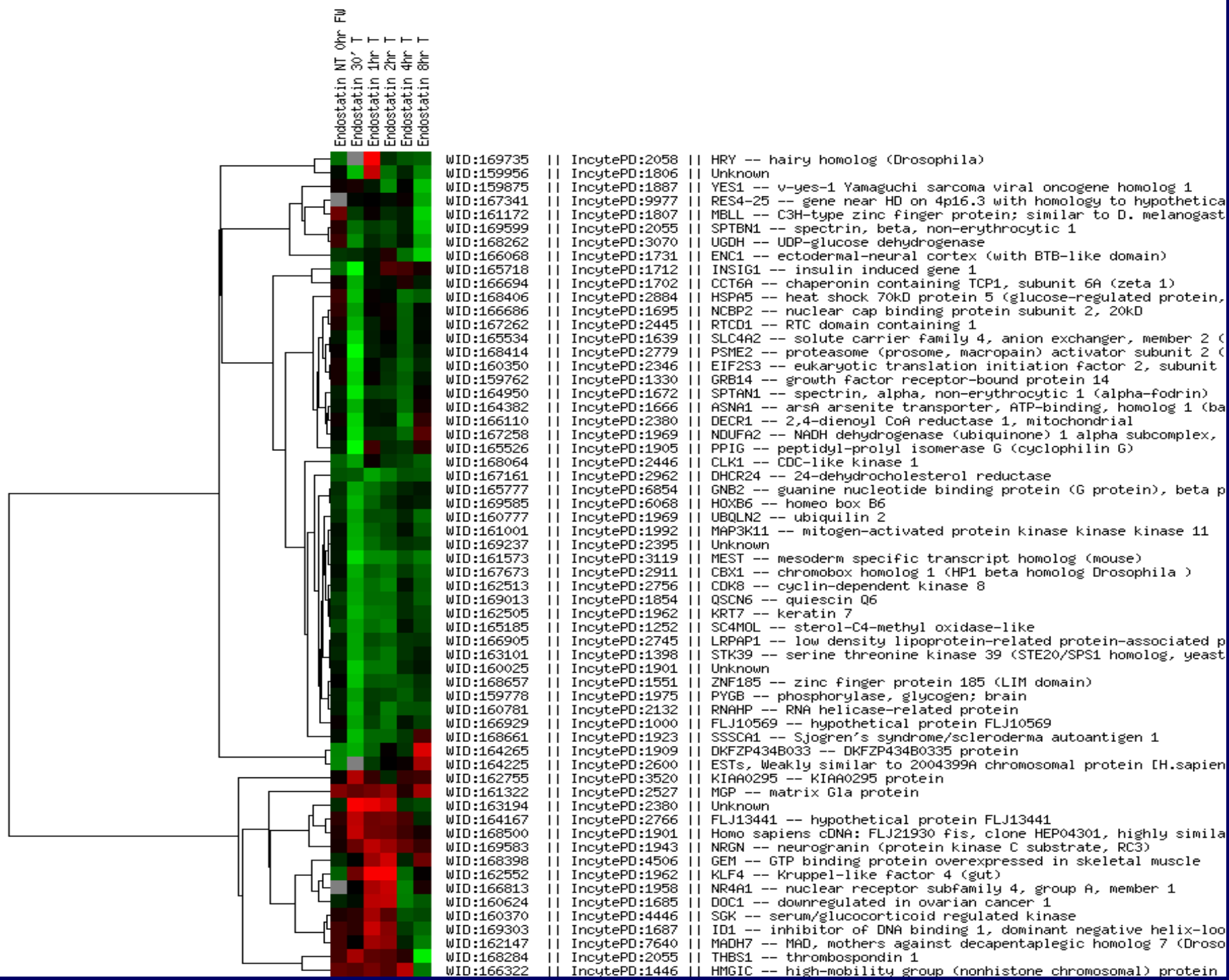
Clustering algorithms: Differing results

(Bittner *et al.*, *Nature*, 2000)

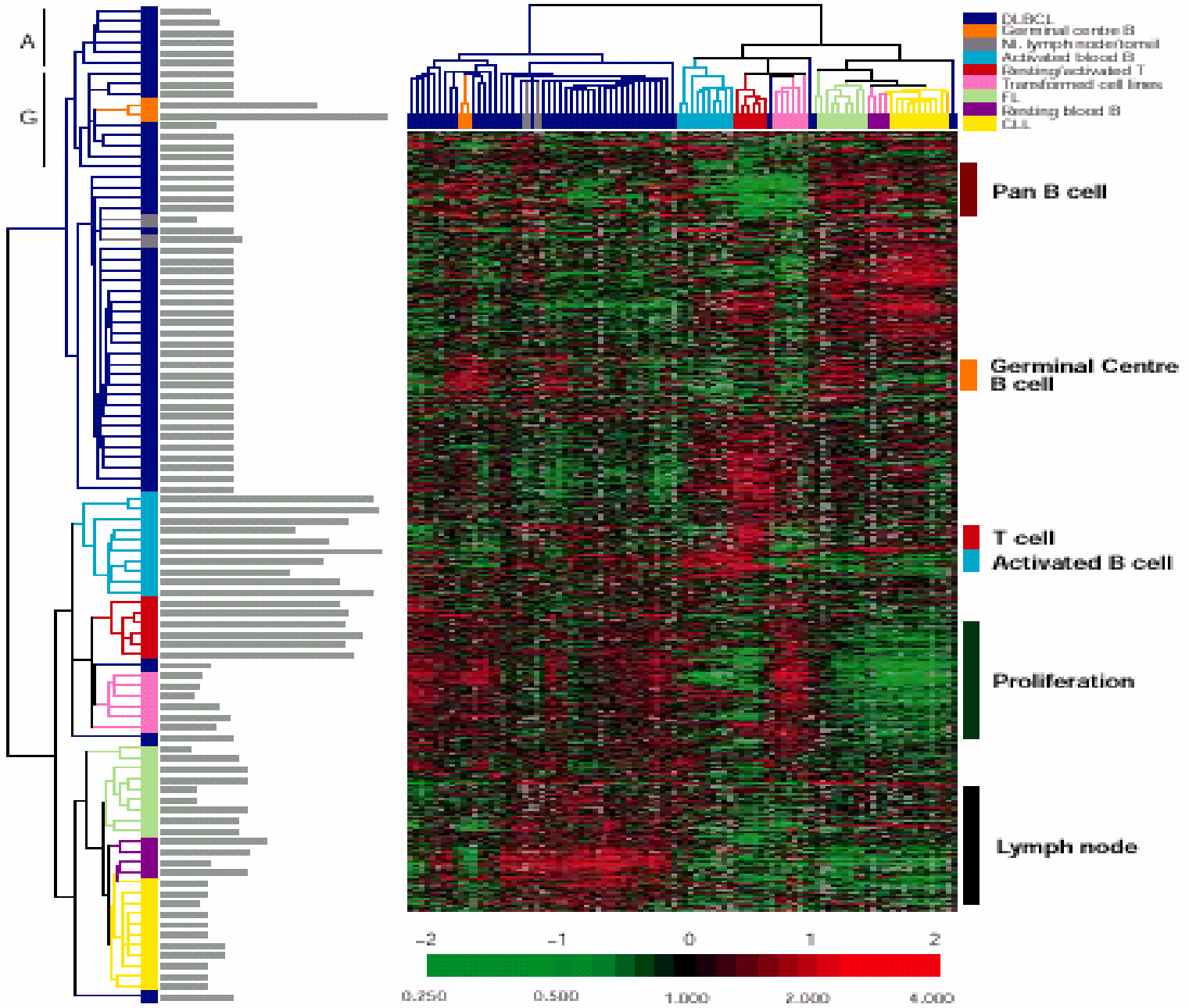
mAdb BioInformatics Project: Current

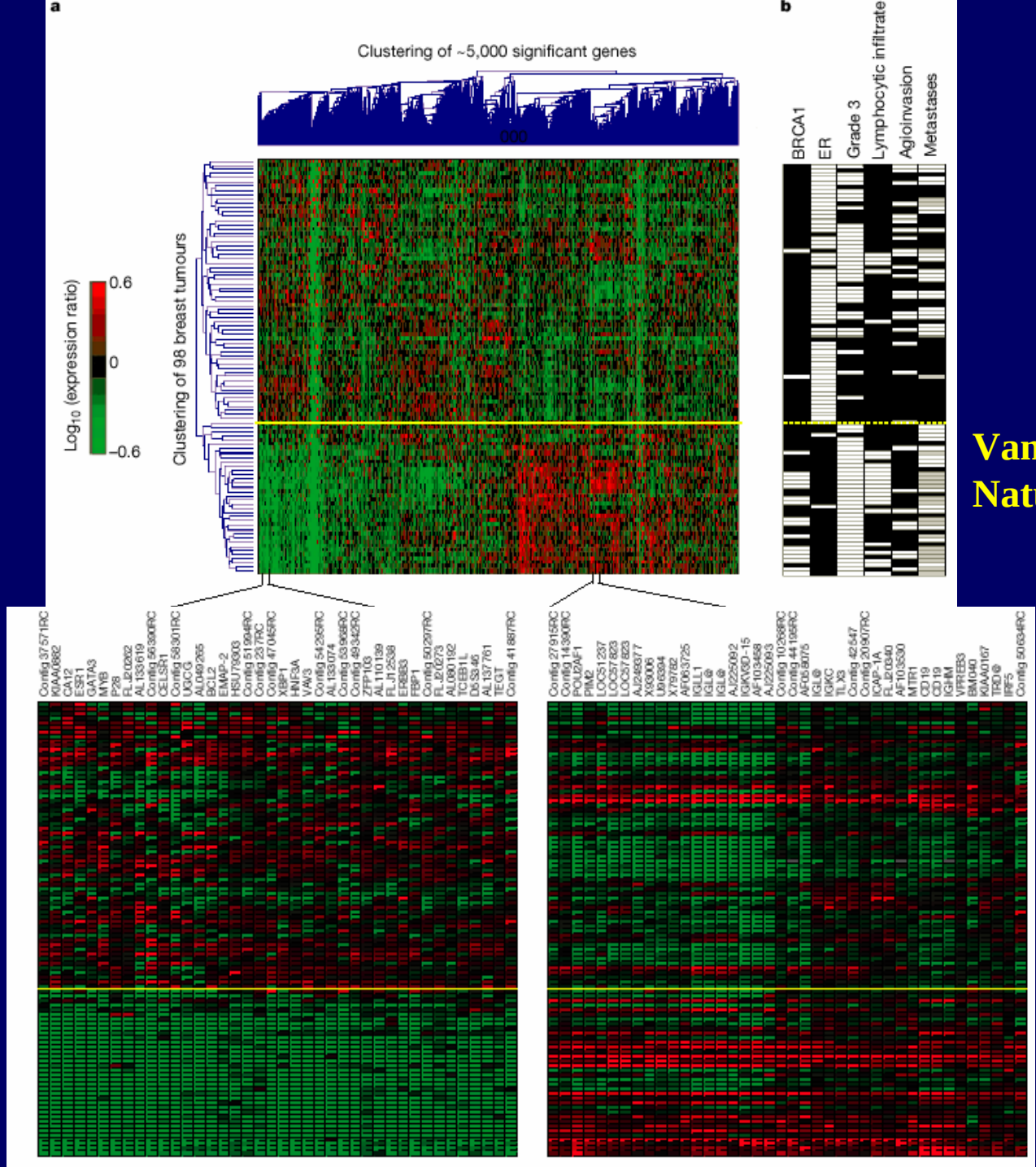
- Links to external data sources
  - dbEST
  - GeneCards
  - LocusLink
  - MGD
  - GenBank
  - Stanford GO (Genome Ontology)
  - PubMed
  - UniGene
- Automatic updates of external data sources

A*	A	A	A*						
#18	#19	#20	#21	Well ID	Feature ID	Map	UniGene	Gene	Descriptions/Gene Ontology Terms
0206	1.0844		2.5936	164548	IncytePD:1804666	2q11.2	Hs.153884	APACD	ATP binding protein associated with cell different
0402	0.0509	0.8579	0.5107	160680	IncytePD:3948420	3q26.2-qter	Hs.75736	APOD	apolipoprotein D lipid metabolism; extracellular space; high-density
6240	0.8441		4.6527	164724	IncytePD:2345343	11p15	Hs.74515	ARNTL	aryl hydrocarbon receptor nuclear translocator-lik circadian rhythm
9690	0.9009	3.5553	2.9125	160419	IncytePD:2056163	1p12	Hs.184270	CAPZA1	capping protein (actin filament) muscle Z-line, alp cell motility; actin binding; actin cytoskeleton; acti
2152	0.2024	0.6577	5.6715	161765	IncytePD:4271973	7q31.1	Hs.74034	CAV1	caveolin 1, caveolae protein, 22kD caveolae; tumor suppressor; structural protein; int
3022	1.3037	2.0487	3.7294	168759	IncytePD:157510	1q31.1	Hs.75184	CHI3L1	chitinase 3-like 1 (cartilage glycoprotein-39) extracellular space; extracellular matrix; extracellu
6701			6.7045	166687	IncytePD:2257362	3q28-q29	Hs.7327	CLDN1	claudin 1 cell adhesion; tight junction; integral membrane pr
3556			9.5932	165783	IncytePD:782235	17q21.3-q22.1	Hs.172928	COL1A1	collagen, type I, alpha 1 collagen; collagen type I; skeletal development; e
2180	0.1037	0.4395	0.3940	169365	IncytePD:585432	15q24	Hs.346950	CRABP1	cellular retinoic acid binding protein 1 cytoplasm; retinoid binding; signal transduction; d
6702	2.1474	3.3447	3.4755	169317	IncytePD:2806166	8p22	Hs.297939	CTSB	cathepsin B lysosome; cathepsin B; intracellular; proteolysis a
3255	0.3610	0.2738	0.1496	165580	IncytePD:1514989	1p31-p22	Hs.8867	CYR61	cysteine-rich, angiogenic inducer, 61 cell proliferation; embryogenesis and morphogene
2378	0.1601	0.2189	1.4210	164649	IncytePD:1969055	1q12-q23	Hs.80552	DPT	dermatopontin cell adhesion; protein binding


Endostatin NT 0hr FW
Endostatin 30' T
Endostatin 1hr T
Endostatin 2hr T
Endostatin 4hr T
Endostatin 8hr T
WID:169735 || IncytePD:2058 || HRY -- hairy homolog (Drosophila)
WID:159956 || IncytePD:1806 || Unknown
WID:159875 || IncytePD:1887 || YES1 -- v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1
WID:167341 || IncytePD:9977 || RES4-25 -- gene near HD on 4p16.3 with homology to hypothetica
WID:161172 || IncytePD:1807 || MBLL -- C3H-type zinc finger protein; similar to D. melanogast
WID:169599 || IncytePD:2055 || SPTBN1 -- spectrin, beta, non-erythrocytic 1
WID:168262 || IncytePD:3070 || UGDH -- UDP-glucose dehydrogenase
WID:166068 || IncytePD:1731 || ENC1 -- ectodermal-neural cortex (with BTB-like domain)
WID:165718 || IncytePD:1712 || INSIG1 -- insulin induced gene 1
WID:166694 || IncytePD:1702 || CCT6A -- chaperonin containing TCP1, subunit 6A (zeta 1)
WID:168406 || IncytePD:2884 || HSPA5 -- heat shock 70kD protein 5 (glucose-regulated protein,
WID:166686 || IncytePD:1695 || NCBP2 -- nuclear cap binding protein subunit 2, 20kD
WID:167262 || IncytePD:2445 || RTCD1 -- RTC domain containing 1
WID:165534 || IncytePD:1639 || SLC4A2 -- solute carrier family 4, anion exchanger, member 2 (
WID:168414 || IncytePD:2779 || PSME2 -- proteasome (prosome, macropain) activator subunit 2 (
WID:160350 || IncytePD:2346 || EIF2S3 -- eukaryotic translation initiation factor 2, subunit
WID:159762 || IncytePD:1330 || GRB14 -- growth factor receptor-bound protein 14
WID:164950 || IncytePD:1672 || SPTAN1 -- spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)
WID:164382 || IncytePD:1666 || ASNA1 -- arsA arsenite transporter, ATP-binding, homolog 1 (ba
WID:166110 || IncytePD:2380 || DECR1 -- 2,4-dienoyl CoA reductase 1, mitochondrial
WID:167258 || IncytePD:1969 || NDUFA2 -- NADH dehydrogenase (ubiquinone) 1 alpha subcomplex,
WID:165526 || IncytePD:1905 || PPIG -- peptidyl-prolyl isomerase G (cyclophilin G)
WID:168064 || IncytePD:2446 || CLK1 -- CDC-like kinase 1
WID:167161 || IncytePD:2962 || DHCR24 -- 24-dehydrocholesterol reductase
WID:165777 || IncytePD:6854 || GNB2 -- guanine nucleotide binding protein (G protein), beta p
WID:169585 || IncytePD:6068 || HOXB6 -- homeo box B6
WID:160777 || IncytePD:1969 || UBQLN2 -- ubiquilin 2
WID:161001 || IncytePD:1992 || MAP3K11 -- mitogen-activated protein kinase kinase kinase 11
WID:169237 || IncytePD:2395 || Unknown
WID:161573 || IncytePD:3119 || MEST -- mesoderm specific transcript homolog (mouse)
WID:167673 || IncytePD:2911 || CBX1 -- chromobox homolog 1 (HP1 beta homolog Drosophila)
WID:162513 || IncytePD:2756 || CDK8 -- cyclin-dependent kinase 8
WID:169013 || IncytePD:1854 || QSCN6 -- quiescin Q6
WID:162505 || IncytePD:1962 || KRT7 -- keratin 7
WID:165185 || IncytePD:1252 || SC4MOL -- sterol-C4-methyl oxidase-like
WID:166905 || IncytePD:2745 || LRPAP1 -- low density lipoprotein-related protein-associated p
WID:163101 || IncytePD:1398 || STK39 -- serine threonine kinase 39 (STE20/SPS1 homolog, yeast
WID:160025 || IncytePD:1901 || Unknown
WID:168657 || IncytePD:1551 || ZNF185 -- zinc finger protein 185 (LIM domain)
WID:159778 || IncytePD:1975 || PYGB -- phosphorylase, glycogen; brain
WID:160781 || IncytePD:2132 || RNAHP -- RNA helicase-related protein
WID:166929 || IncytePD:1000 || FLJ10569 -- hypothetical protein FLJ10569
WID:168661 || IncytePD:1923 || SSSCA1 -- Sjogren's syndrome/scleroderma autoantigen 1
WID:164265 || IncytePD:1909 || DKFZP434B033 -- DKFZP434B0335 protein
WID:164225 || IncytePD:2600 || ESTs, Weakly similar to 2004399A chromosomal protein [H.sapien
WID:162755 || IncytePD:3520 || KIAA0295 -- KIAA0295 protein
WID:161322 || IncytePD:2527 || MGP -- matrix Gla protein
WID:163194 || IncytePD:2380 || Unknown
WID:164167 || IncytePD:2766 || FLJ13441 -- hypothetical protein FLJ13441
WID:168500 || IncytePD:1901 || Homo sapiens cDNA: FLJ21930 fis, clone HEP04301, highly simila
WID:169583 || IncytePD:1943 || NRGN -- neurogranin (protein kinase C substrate, RC3)
WID:168398 || IncytePD:4506 || GEM -- GTP binding protein overexpressed in skeletal muscle
WID:162552 || IncytePD:1962 || KLF4 -- Kruppel-like factor 4 (gut)
WID:166813 || IncytePD:1958 || NR4A1 -- nuclear receptor subfamily 4, group A, member 1
WID:160624 || IncytePD:1685 || DOC1 -- downregulated in ovarian cancer 1
WID:160370 || IncytePD:4446 || SGK -- serum/glucocorticoid regulated kinase
WID:169303 || IncytePD:1687 || ID1 -- inhibitor of DNA binding 1, dominant negative helix-loo
WID:162147 || IncytePD:7640 || MADH7 -- MAD, mothers against decapentaplegic homolog 7 (Droso
WID:168284 || IncytePD:2055 || THBS1 -- thrombospondin 1
WID:166322 || IncytePD:1446 || HMGIC -- high-mobility group (nonhistone chromosomal) protein

Hierarchical Clustering
(Alizadeh *et al.*, Nature, Feb. 2000)





Van't Veer et al, Nature, Jan. 2002

Issues for Study Design

- Appropriateness of model
  - Exposure, genotype, timing
  - Tissue and cells
  - Controls: Sieving the genes
- The issue of completeness
  - # genes expressed in tissue
  - % modulated at chosen level

Pitfalls in multiplex analysis of gene expression

Nonspecific changes in gene expression

Quantitatively small changes can be critical

mRNA modulation is often not the mechanism

- Allosteric regulation of enzymes
- Receptor desensitization

Conclusions

- Multiplex analysis of gene expression has great potential for identifying intermediate processes in neural diseases
- Realization of the potential of these inherently industrial methods requires more incisive experimental designs
- The downstream biology is inherently nonindustrial
- Appropriate controls can narrow the field of candidate genes
- These methods should lead to the identification of new pathways in disease and an understanding of sharing of pathways between diseases