Principy microarrays III

Pavla Gajdušková

Analytická cytometrie, 8. prosince 2009

Obsah přednášky

Technologie přípravy microarrays

Oblasti použití microarrays v biologii

Úvod do statistického hodnocení dat

Příklady konkrétních aplikací z literatury

Úvod do statistického hodnocení dat

Předpříprava dat pro statistické hodnocení

analýza obrazu (měření intenzity bodů a pozadí)

normalizace (nalezení a odstranění systematických chyb, které nejsou způsobeny biologickým objektem)

filtrování dat (odstranění špatných bodů nebo hybridizací ze studie)

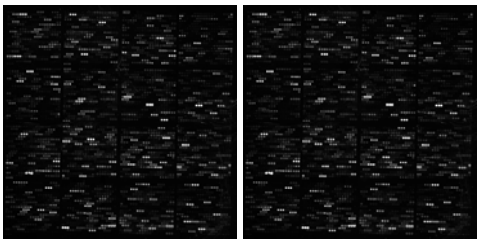
Nalezení rozdílně exprimovaných genů

výpočet zvolené statistiky a následné určení p hodnot

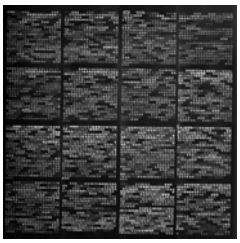
úprava p-hodnot

Analýza obrazu

16-bitový obraz ve stupních šedi
hodnoty intenzity: 0 - 65 536



Red **Green**



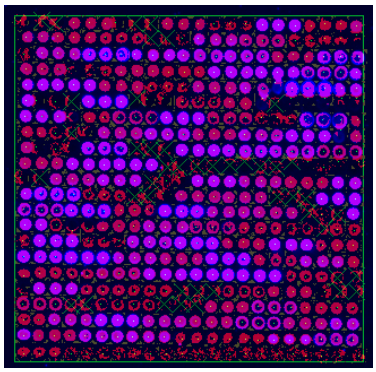
Dapi

Analýza obrazu

rozdělení pixelů v nasnímaném obraze na ty, které nesou informaci o intenzitě bodů na sklíčku nebo pozadí

mnoho programů na analýzu microarray obrazů (GenePix, Spot, ...)

výsledek: txt soubor – každý řádek obsahuje informaci o jednom bodu na sklíčku (průměrná intenzita uvnitř bodu, intenzita okolí, variabilita mezi pixely uvnitř bodu, ...)



Subarray

Analýza obrazu

Nejdůležitější hodnota: poměr mezi intenzitami fluorescence R a G

R/G

Nejčastěji se vyjadřuje pomocí logaritmu o základu 2

$M = \mathrm{Log}_2\ R/G$

Log_2 R/G = 1 — **ve vzorku značeném červeně je dvakrát více kopií specifické mRNA než v zeleně značeném vzorku**

Log_2 R/G =-1 — **ve vzorku značeném červeně je poloviční množství kopií specifické mRNA než v zeleně značeném vzorku**

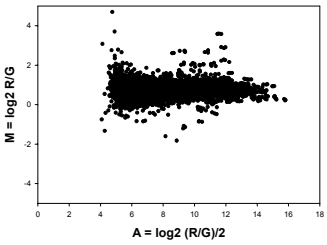
Analýza obrazu

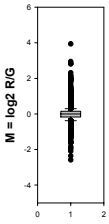
$$M = Log_2 R/G$$

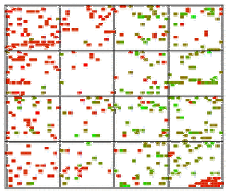
Další důležitá hodnota pro kontrolu kvality hybridizace je

průměrná intenzita bodu v obou snímaných kanálech

$$A = (Log_2R + Log_2G) / 2$$







Důležité předpoklady

Sondy na sklíčku jsou rozmístěny zcela náhodně

do stejné pozice na sklíčku neseskupujeme geny s podobnou funkcí; sekvenčně příbuzné; ležící na stejném chromosomu

Hybridizace byly prováděny v náhodném pořadí

kontroly byly hybridizovány dohromady se zkoumanými vzorky

Předpokládáme, že experiment ovlivní expresi pouze malého počtu genů v daném objektu (většina genů svoji expresi nemění)

průměr (medián) všech poměrů R/G je roven 1
průměr (medián) všech logaritmů poměrů R/G je roven 0

nestačí mít na sklíčku sondy pro geny, které nás zajímají nebo očekáváme, že jejich exprese se bude měnit
pro normalizaci jsou nutné i další geny, jejichž exprese se nemění (těch by měla být většina)

Odstranění „špatných“ bodů

odstranění bodů: body s morfologickými abnormalitami (problematický tisk)

s nízkou intenzitou (není exprese v daném systemu)

s vysokým pozadím (negativní hybridizace)

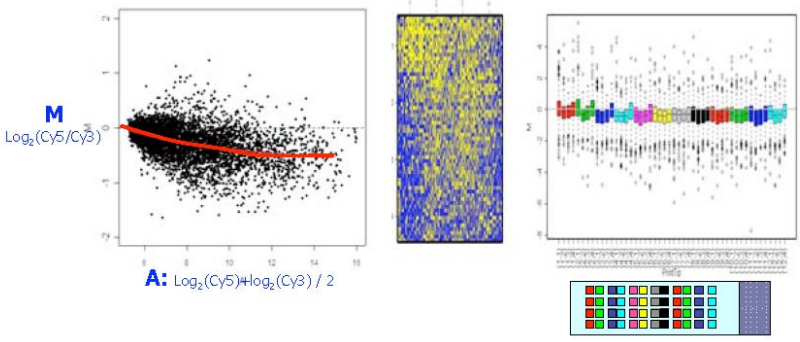
Kontrolní body: prázdné body bez DNA (negativní kontrola)

„spiked“ body (pozitivní kontrola)

stejné sondy na různých místech sklíčka

Normalizace

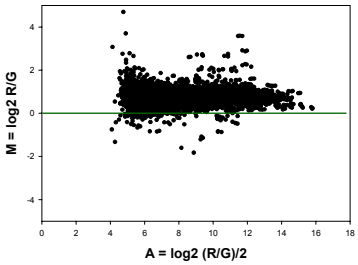
nalezení a odstranění systematických chyb, které nejsou způsobeny biologickým objektem



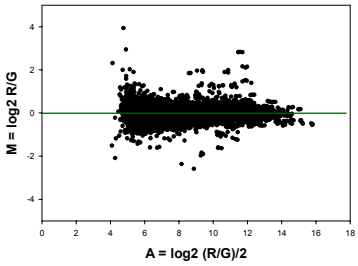
Normalizace

Není splněná podmínka, že průměr (medián) všech logaritmů poměrů R/G je roven 0

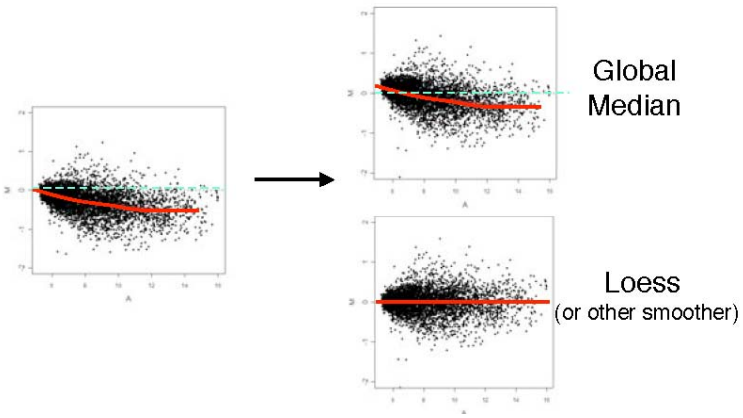
Před normalizací:



Po normalizaci:



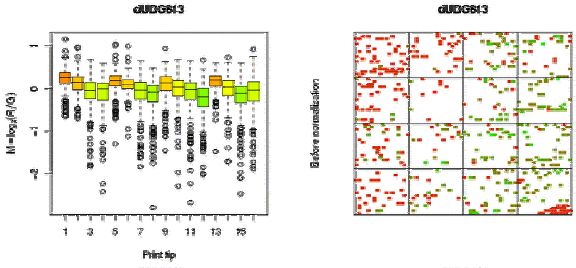
Loess Normalizace



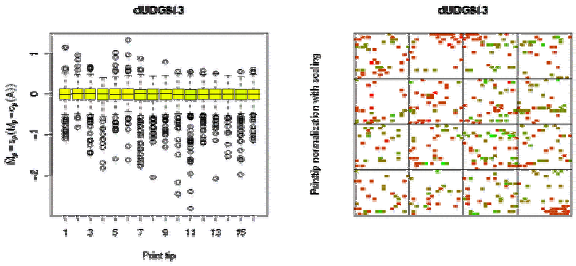
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

“Print Tip” Normalizace

Před normalizací:



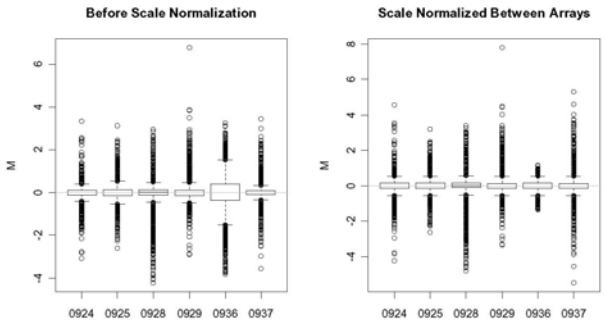
Po normalizaci:



Normalizace mezi arrays

Všechny hybridizace v dané studii by měly mít podobné rozložení hodnot kolem mediánu

“Median Absolute Deviation (MAD) Scaling“



Programy pro předpřípravu dat

Product	Authors/Company/Institute	Interface/Operating System	Reference/Features
ArrayStat 1.0	Imaging Research Inc.	Windows	Software package optimised for statistical analysis of array gene expression data. Quality control, statistical tests of differential expression
Bioconductor	The R Project for Statistical Computing	R-package	An open source and open development software project for the analysis and comprehension of genomic data
BRB ArrayTools 3.2.3	Molecular Statistics and Bioinformatics Section, Biometric Research Branch, NCI	Excel add-in, R-package	Wright GW et al. A random variance model for detection of differential gene expression in samll microarray experiments. Bioinformatics 2003 19:2448-2455.
dCHIP	Wong Lab, Harvard School of Public Health and Dana-Farber Cancer Institute	Windows	Li C and Wong WH (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, Proc. Natl. Acad. Sci. Vol. 98, 31-36
Genetraffic 3.1	Iobion Informatics	Linux server, web client	Analysing and visualizing microarray expression data. Compliant with MIAME & MAML standard.
Lucidea Array Spotfinder 1.0	Amersham Biosciences	Windows	Fully automated image analysis software, taking into account pen effects and calculating various quality metrics
Lucidea Microarray Scorecard 1.0	Amersham Biosciences	Windows	Software package developed to analyse data from two-color experiments, calculate various quality metrics and normalize data using an exponential method
R-package	The R Project for Statistical Computing	R-package	One most famous statistical packages. Most libraries including specific ones for the analysis of microarray data.
SpotFire.net Desktop 5.0	SpotFire	Windows	Asher B. Decision analytics software solutions for proteomics analysis. J Mol Graph Model 2000 18: 79-82.
TIGR Microarray Data Analysis Software (MIDAS)	The Institute for Genomic Research (TIGR)	Java tested on Windows 2000/XP, Linux 7.2, MacOS 10.2	Saeed AI et al. TM4 : a free, open-source system for microarray data management and analysis. Biotechniques 2003 34:274-278
XLstat 3D Plot	Addinsoft	Excel add-in	Xlstat 3D Plot is a complement module for Xlstat Pro that allows to display data in 3 dimension with an intuitive interface.
XLstat Pro 7.1	Addinsoft	Excel add-in	Sotware package for statistical analysis including a wide range of functionalities

http://arraysimage. free.fr/Soft.htm

Nalezení rozdílně exprimovaných genů

odstranění špatných bodů, provedena vhodná normalizace intenzit

Nulová hypotéza: medián exprese daného genu se statisticky neliší od teoretické hodnoty mediánu (v našem případě 0)

Pro každý gen testujeme tuto hypotézu zvlášť

:	Array 1	Array 2	Array 3	Array 4	Medián
Gen 111	0.39		-0.39	0.06	0.06
Gen 112	-0.28	0.33	0.37	0.64	0.35
Gen 113		0.14	0.28	0.44	0.28
Gen 114	-0.19	0.13	-0.13	0.38	0.00
Gen 115	0.88	0.49	0.54	0.45	0.52
:					

Nalezení rozdílně exprimovaných genů

Nulová hypotéza: medián exprese daného genu se statisticky neliší od teoretické hodnoty mediánu (v našem případě 0)

$$T = \frac{\overline{M}}{se(\overline{M})}$$ ..... **p hodnota** pravděpodobnost s jakou lze nulovou hypotézu zamítnout

rozdílně exprimované geny … p hodnota < 0.01 (volitelný práh)

:	Array 1	Array 2	Array 3	Array 4	p hodnota
Gen 111	0.39		-0.39	0.06	0.78
Gen 112	-0.28	0.33	0.37	0.64	0.25
Gen 113		0.14	0.28	0.44	0.38
Gen 114	-0.19	0.13	-0.13	0.38	0.99
Gen 115	0.88	0.49	0.54	0.45	0.02
:					

Statistické problémy při studiu tisíců genů s malým počtem opakování experimentů

rozdílně exprimované geny … p hodnota < 0.01

Příklad:
studujeme 20 000 genů na jednom sklíčku
během normalizace a kontroly kvality vyřadíme 12000 genů
testujeme 8 000 genů (pro každý vypočítáme p hodnotu)

p hodnota < 0.01 připouštíme, že 1% testovaných genů je označeno jako rozdílně exprimované pouze náhodnou variabilitou pokusů

8000 * 0.01 **= 80 genů**

➔ **korekce p hodnot s ohledem k počtu testovaných genů**
➔ **použití alternativních statistik**

Specialized Methods: "Modified" t

- ***Penalized-t*** **(SAM,** Tusher et al 2001, Efron et al 2000**):**

$$t^* = \frac{\overline{M}}{(s+a)/\sqrt{n}}$$

Estimate penalty term a by 90th percentile of s.d. of all genes, or by minimizing the coefficient of variation of the absolute t.

- ***Moderated-t*** **(Limma,** Smyth 2004**):**

$$t^* = \frac{\overline{M}}{\tilde{s}/\sqrt{n}}$$

Use shrinkage s.d. $\tilde{s}^2 = \frac{s^2 d + s_0^2 d_0}{d + d_0}$ estimated by an empirical Bayes method

s_0: pooled s.d., d_0: d.f. of prior

- ***Regularized-t*** **(Cyber-T,** Baldi P & Long AD 2001**):**

$$t^* = \frac{\overline{M}}{\tilde{s}/\sqrt{n}}$$

Use regularized s.d. $\tilde{s}^2 = \frac{\nu_0 \sigma_0^2 + (n-1)s^2}{\nu_0 + n - 2}$

ν_0: prior strength

σ_0^2: background s.d.

From Ru-Fang Yeh presentation: Statistical Methods in Bioinformatics: Case Studies. Center for Bioinformatics & Molecular Biostatistics, UCSF Division of Biostatistics

Alternative Statistics

- ***B-statistic*** (Lonnstedt & Speed 2002): The log posterior odds ratio that a gene is DE vs not DE, estimated by the empirical Bayes method.

$$B = \log \frac{\Pr\{\text{DE}\}}{\Pr\{not\ \text{DE}\}} = \log \frac{p}{1-p} \left(\frac{\nu}{\nu+\nu_0}\right)^{1/2} \left(\frac{t^{*2} + d_0 + d}{t^{*2} \frac{\nu}{\nu+\nu_0} + d_0 + d}\right)^{(1+d+d_0)/2}$$

  - Equivalent to moderated-t in terms of ranking genes.
  - Dependent on p = expected proportion of DE genes

- ***Distance Synthesis*** **(DEDS,** Yang et al 2004**):** Define a *distance* statistic based on measures of choice, and estimate false discovery rates using appropriate null distribution.
- Single channel methods modelling absolute Cy5 & Cy3 expression (Newton et al 2001, Wolfinger et al 2001)

From Ru-Fang Yeh presentation: Statistical Methods in Bioinformatics: Case Studies. Center for Bioinformatics & Molecular Biostatistics, UCSF Division of Biostatistics

Obsah přednášky

Technologie přípravy microarrays

Oblasti použití microarrays v biologii

Úvod do statistického hodnocení dat

Příklady konkrétních aplikací z literatury

Klastrování

Klastrování (shluková analýza) je obecná metoda, kterou je možno použít ke spojování prvků (s podobnými vlastnostmi) do skupin (klastrů)

Microarray analýza:

Klastrování genů (řádků) ➔ identifikace skupin genů, které mohou být společně regulované

Klastrování vzorků (sloupců) ➔ nalezení skupin vzorků, které mají podobné změny v expresi genů (změny na úrovni DNA)

Příklad:
Sorlie et al., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS 98: 10869-10874, 2001.

Design experimentu

78 karcinomů prsu (71 duktálních, 5 lobulárních a 2 in-situ)
3 fibroadenomy
4 vzorky normální tkáně prsu

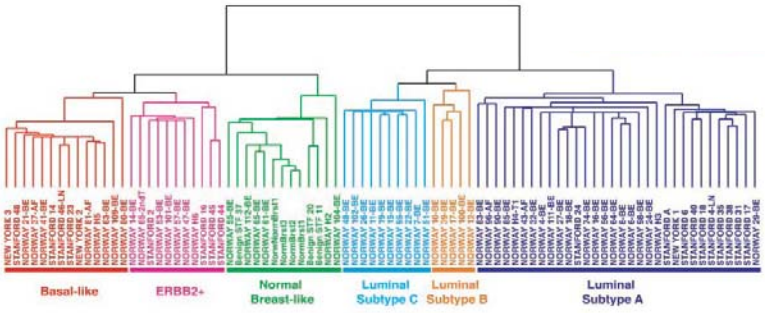
Microarrays: **8 102 cDNA klonů**
každý vzorek (Cy3) hybridizován s referenční RNA (Cy5)

Analýza: **nalezeno 456 cDNA klonů (427 genů) s velkou variabilitou exprese mezi různými vzorky, ale podobnou expresí u příbuzných vzorků**

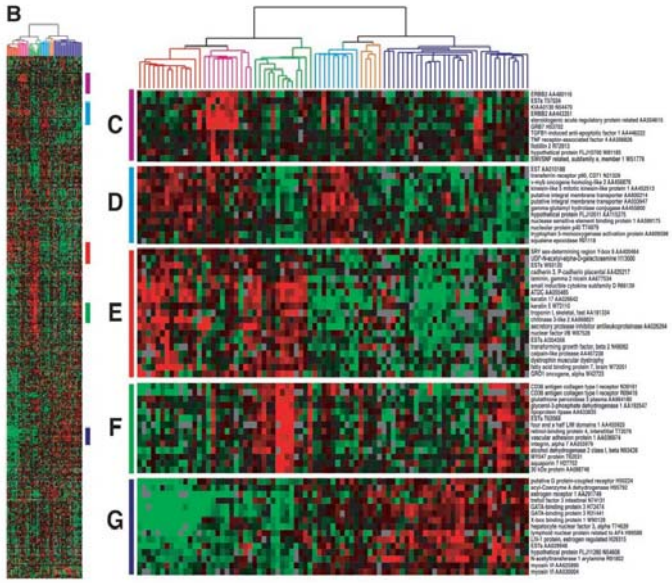
Otázka: **Zda existuje rozdělení karcinomů do podskupin, které mají podobné změny v expresi genů?**

Sorlie et al., PNAS 98: 10869-10874, 2001.

Klastrování

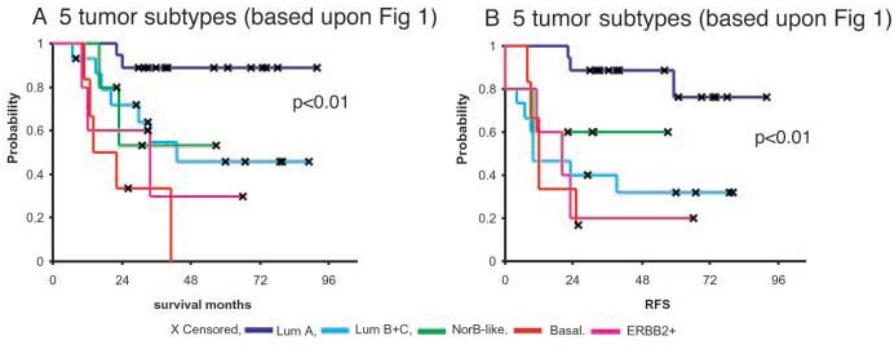


Sorlie et al., PNAS 98: 10869-10874, 2001.



Sorlie et al., PNAS 98: 10869-10874, 2001.

Rozdělení do skupin a prognóza vývoje onemocnění



Sorlie et al., PNAS 98: 10869-10874, 2001.

Programy pro analýzu microarray dat

Product	Authors/Company/Institute	Interface/Operating System	Reference/Features
ArrayStat 1.0	Imaging Research Inc.	Windows NT/2000	Software package that is optimised for statistical analysis of array gene expression data. Quality control, statistical tests of differential expression
BRB ArrayTools 3.2.3	Molecular Statistics and Bioinformatics Section, Biometric Research Branch, NCI	Excel add-in, R-package	Wright GW et al. A random variance model for detection of differential gene expression in samll microarray experiments. Bioinformatics 2003 19:2448-2455.
Cluster	Michael Eisen's lab:Lawrence Berkeley National Lab (LBNL)	Windows 95/98/NT	Eisen MB et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998 95:14863-14868.
Cluster Indentification Tool (CIT)	Van Andel Research Institute	Windows	Rhodes DR et al. CIT: identification of differentially expressed clusters of genes from microarray data. Bioinformatics 2002 18:205-206.
FDR controlling procedure (FDRalgo)		Windows	Adjusts p-values generated in multiple hypothesis testing of gene expression data obtained by cDNA microarray experiment.
Genesis	Bioinformatics Group, Institute of Biomedical Engineering, Graz University of Technology	Java, tested on Windows	Java suite containing various tools such as filters, normalization, visualization tools, clustering, SOM, k-means, PCA, SVM, map onto chromosomal sequences.
Genetraffic 3.1	Iobion Informatics	Linux server, web client	Analysing and visualizing microarray expression data. Compliant with MIAME & MAML standard.
J-express	Bioinformatics research group at the Dept. of Informatics	Java, tested on Windows 2000, LINUX, Thru64 UNIX, Solaris and Irix	Analysing gene expression data giving access to hierarchical clustering, k-means, SOM, PCA, MDS, profile similarity search and visualizing methods.
LACK		Windows	Kim C et al. Significance analysis of lexical bias in microarray data. Bioinformatics 2003, 4:12
Prediction Analysis for Microarray (PAM)	Tibshirani Lab, Departement of Statistics, Stanford University	Excel add-in/ R-package	Narasimhan and Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 2002 99:6567-6572.
R-package	The R Project for Statistical Computing	R-package	One most famous statistical packages. Most libraries including specific ones for the analysis of microarray data.
Significance Analysis of Microarrays (SAM)	Tibshirani Lab, Departement of Statistics, Stanford University	Excel add-in/ R-package	Tibshirani and Chu. Significance analysis of microarrays applied to the ionizing radiation response. PNAS 2001 98: 5116-5121
SpotFire.net Desktop 5.0	SpotFire	Windows	Asher B. Decision analytics software solutions for proteomics analysis. J Mol Graph Model 2000 18: 79-82

http://arraysimage. free.fr/Soft.htm

Veřejné databáze microarray dat

ArrayExpress
ChipDB
ExpressDB
Gene Expression Atlas
Gene Expression Database (GXD)
Gene Expression Omnibus (GEO)
GeneX
GermOnline
Human Gene Expression Index (HuGE Index)
List Of Lists Annotated (LOLA)
M-CHiPS (Multi-Conditional Hybridization Intensity Processing System)
MUSC DNA Microarray Database
NASCArrays
Oncomine
Public Expression Profiling Resource (PEPR)
READ (RIKEN cDNA Expression Array Database)
Rice Expression Database (RED)
RNA Abundance Database (RAD)
Saccharomyces Genome Database (SGD): Expression Connection
SGMD
Standford Microarray Database (SMD)
Yale Microarray Database
yeast Microarray Global Viewer (yMGV)