Principy microarrays

Pavla Gajdušková

Analytická cytometrie, 27. listopadu 2012

Oblasti použití microarrays v biologii

Typ array	Sondy na microarray	Co se fluorescenčně značí a hybridizuje	... analýza čeho
Expresní	DNA (cDNA, oligonucleotidy)	mRNA / cDNA	měření množství mRNA v bunkách, nádorech ...
miRNA	oligonukleotidy	miRNA	měření množství miRNA
CGH	DNA (BAC vektory, oligonukleotidy)	DNA	změny v genomu (zisk, ztráta chromozomů nebo jejich částí
SNP	DNA (oligonukleotidy)	DNA	detekce „Single Nucleotid Polymorphysms"; změny v genomu
Metylace	DNA (CpG islands)	DNA (ovlivněná bisulfidem sodným)	míra metylace promotorových oblastí
Promoter	DNA (promotorové oblasti ~ 1kb)	DNA (ChiP obohacená)	místa vazby transkripčních faktorů, modifikace histonů
Tilling	DNA	všechno dříve zmíněné	všechno dříve zmíněné, sekvenování, anotace genů
Protein	protilátky	protein	exprese proteinů (ELISA)

Použití microarrays ke studiu DNA

Komparativní genomická hybridizace
- BAC arrays
- oligo arrays
- SNP arrays
- tilling arrays (BAC a oligonukleotidy)
- exon-specific arrays
- (dříve i cDNA arrays používané pro expresi)

Genotypování
- SNP arrays

Sekvenování
- Re-Sequencing arrays

ChIP-Chip exprerimenty
- tilling arrays (oligonukleotidy)

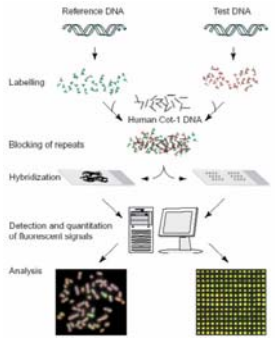
Komparativní genomická hybridizace (CGH)

molekulárně cytogenetická metoda, která slouží k analýze změn obsahu DNA v živých organismech

(delece, zisk, amplifikace různých oblastí genomu)

porovnávání intenzity fluorescence zkoumaného vzorku DNA a normálního diploidního vzorku DNA v různých místech genomu

Komparativní genomická hybridizace (CGH)



Mantripragada et al. Trends in Genetics 2003

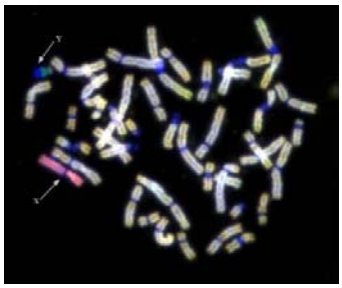
Komparativní genomická hybridizace (CGH)

metafázní chromozomy
- dárce s normálním diploidním karyotypem

DNA:
cy3
zkoumaný vzorek

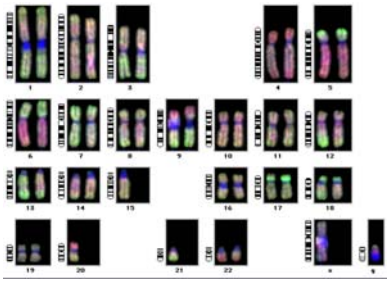
cy5
referenční DNA – 2n

rozlišení ~ 20MB



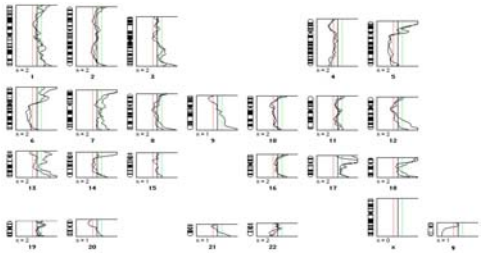
From Szuhai K. presentation: Determination of Genomic Imbalances by Genome-wide Screening Approaches

Komparativní genomická hybridizace (CGH)



From Szuhai K. presentation: Determination of Genomic Imbalances by Genome-wide Screening Approaches

Komparativní genomická hybridizace (CGH)



From Szuhai K. presentation: Determination of Genomic Imbalances by Genome-wide Screening Approaches

„Array“ komparativní genomická hybridizace (Array CGH)

chromozomy nahrazeny body na mikroskopickém sklíčku, které obsahují specifické DNA sekvence

Typy sond natištěných na microarray sklíčku

BAC klony	až 32 000 BAC klonů na jednom sklíčku ~ 160 kb dlouhé úseky DNA
Oligonukleotidy	25 – 80 bazí dlouhé oligonukleotidy mohou pokrývat i celý genom (repetitivní sekvence jsou vynechány)

známe polohu a pořadí všech sond v lidském genomu

Knihovny BAC klonů pro array CGH

BACPAC resources (CHORI)
http://bacpac.chori.org

Research Genetics (Invitrogen)
http://www.resgen.com/resources/index.php3

The Sanger Centre
http://www.geneservice.co.uk/home/

Cheung V. G. et al., Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* 409: 953 – 958, 2001.

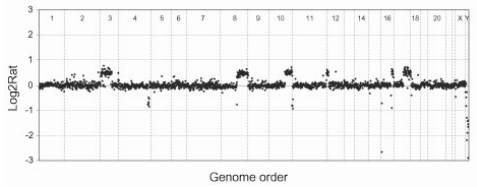
Greshock J. et al., 1-Mb Resolution Array-Based Comparative Genomic Hybridization Using a BAC Clone Set Optimized for Cancer Gene Analysis. Genome Res 14: 179-187, 2004.

Krzywinski M. et al., A set of BAC clones spanning the human genome. Nucleic Acids Res 32: 3651-3660, 2004.

Array CGH s použitím BAC klonů

$Log_2Rat = Log_2 R/G$

$Log_2Rat = 0$	**2 kopie**	**Log2Rat = -1**	**1 kopie (“loss”)**
$Log_2Rat = 0.5$	**3 kopie (“gain”)**	**Log2Rat < -1**	**homozygotní delece**
$Log_2Rat = 1$	**4 kopie (“gain”)**		
Log2Rat = 2	**8 kopií (“amplification”)**		



2464 BAC klonů **UCSF HumArray3.1**

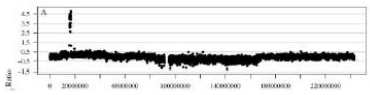
Typy sond natištěných na microarray sklíčku

BAC klony až 32 000 BAC klonů na jednom sklíčku
~ 160 kb dlouhé úseky DNA

Oligonukleotidy 25 – 80 bazí dlouhé oligonukleotidy mohou pokrývat i celý genom (repetitivní sekvence jsou vynechány)

známe polohu a pořadí všech sond v lidském genomu

Array CGH - oligonukleotidy (NimbleGen)



6-kb median probe spacing

Selzer RR et al. Genes Chromosomes Cancer, 2005

SNPs

SNP = single nucleotide polymorphism

jednonukleotidové variace, které jsou náhodně rozmístěny v genomu (bodové mutace rozšířené v populaci)

nukleotidová variace, která se vyskytuje alespoň u 1% jedinců v populaci

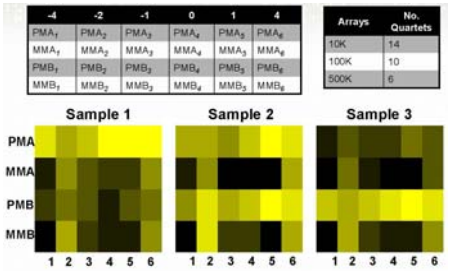
předpokládaný počet SNPs: 10 milionů

výskyt specifických SNP spojen s predispozicí k určitým chorobám

SNP Arrays – probe design (Affymetrix)

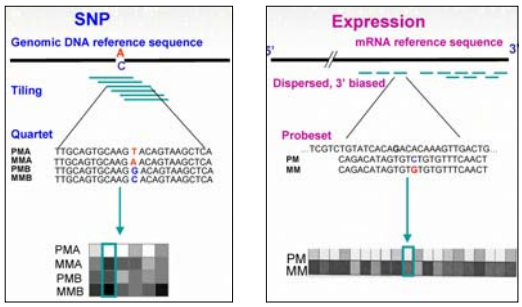
From Xiao Y. presentation: Exploration and Analysis of Affymetrix SNP Arrays. Center for Bioinformatics & Molecular Biostatistics, UCSF Division of Biostatistics

SNP Arrays – probe design



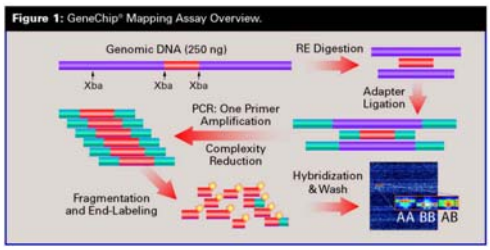
From Xiao Y. presentation: Exploration and Analysis of Affymetrix SNP Arrays. Center for Bioinformatics & Molecular Biostatistics, UCSF Division of Biostatistics

SNP arrays x expression arrays



From Xiao Y. presentation: Exploration and Analysis of Affymetrix SNP Arrays. Center for Bioinformatics & Molecular Biostatistics, UCSF Division of Biostatistics

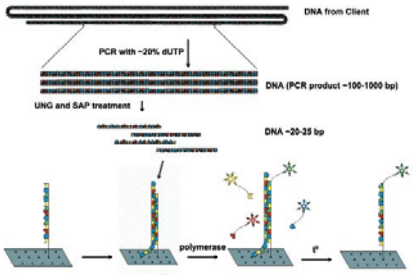
SNP Arrays - labeling



From Xiao Y. presentation: Exploration and Analysis of Affymetrix SNP Arrays. Center for Bioinformatics & Molecular Biostatistics, UCSF Division of Biostatistics

SNP Arrays - APEX technologie

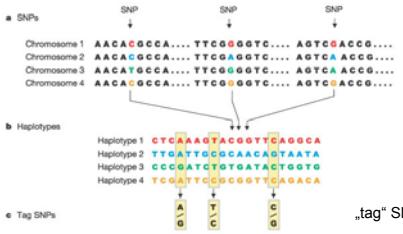
APEX = Arrayed **P**rimer **E**xtension



Kurg A. et al., Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. Genet Test 4:1-7, 2000.

Velké studie SNP

HapMap projekt: mezinárodní projekt, jehož cílem je identifikovat a katalogizovat SNPs v lidské populaci a vybrat z nich **„tag“ SNPs**, kterými se skupiny lidí odlišují



SNPs, které jsou na DNA blízko sebe se také společně dědí a určují **haplotyp** dané skupiny lidí

„tag“ SNPs odlišují dané haplotypy

HapMap projekt

http://www.hapmap.org/index.html.en

HapMap kolekce lidské DNA 270 vzorků DNA

populace:	Nigerie	30 trojic vzorků (matka, otec, dítě)
	Japonsko	45 nepříbuzných vzorků
	Čína	45 nepříbuzných vzorků
	USA	30 trojic vzorků (matka, otec, dítě)

The International HapMap Consortium. **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 449, 851-861. 2007.

The International HapMap Consortium. **A Haplotype Map of the Human Genome.** *Nature* 437, 1299-1320. 2005.

Velké studie SNP

3000 zdravých jedinců

2000 pacientů	bipolar disorder	(1 SNP)
„	coronary artery disease	(1 SNP)
„	Crohn’s disease	(9 SNPs)
„	hypertension	
„	rheumatoid arthritis	(3 SNPs)
„	type 1 diabetes	(1 SNP)
„	type 2 diabetes	(3 SNPs)

P value < 5×10^{-7}

Studovali 500 000 SNPs pomocí Affymetrix microarrays

Wellcome Trust Case control Consortium. **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** Nature. 2007 Jun 7;447(7145):661-78.

Odchylky od referenčního genomu větší než 1kb

ještě v roce 2003 se myslelo, že většina „zdravých“ lidí se od referenčního genomu liší velmi nepatrně (SNPs, mikrosatelity)

array komparativní genomická hybridizace odhalila mnoho větších oblastí DNA, které se u zdravých lidí vyskytují v různém počtu

Copy number variation

DNA segment (většinou větší než 1 kb), který se u daného jedince vyskytuje v jiném počtu kopií než v referenčním lidském genomu

existuje mnoho takových oblastí v genomu (řádově tisíce)

"**Database of Genomic Variants**"
http://projects.tcag.ca/variation/

Copy number polymorphism – výskyt u více než 1% jedinců dané populace

Využití HapMap kolekce ke studiu copy number variant
všichni jedinci v této kolekci byli zdraví, přesto se našlo velké množství oblastí DNA (12% genomu), které se u těchto lidí nacházejí v různém počtu kopií

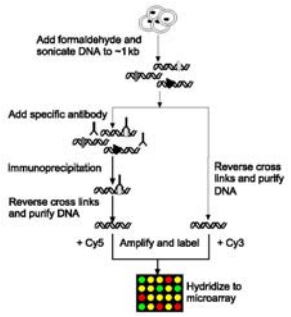
Copy number variation

hledání fenotypových projevů CNV („neškodná" genomová varianta nebo příčina nemoci???)

CNV: pathogenic x benign x unknown clinical significance

vnášejí „zmatek" do experimenů, které např. hledají příčinu vrozených genetických poruch (mentální opožděnost, vývojové odchylky)

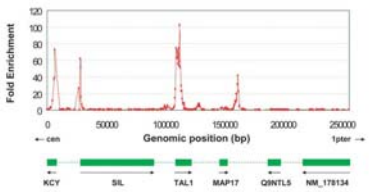
Chromatin ImmunoPrecipitation on chip
ChIP-Chip



Nalezení vazebného místa

256 kb oblast 1p32 pokrytá překrývajícími se PCR produkty (~400 bp)

protilátka: trimethylace histonu H3 Lys4



Carter and Vetrie 2004 Human Mol Genet

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

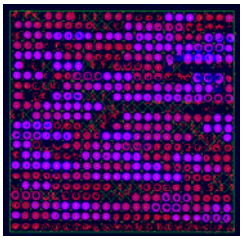
úprava p-hodnot

Analýza obrazu

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

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

$\text{Log}_2\ R/G = 1$ — **ve vzorku značeném červeně je dvakrát více kopií specifické mRNA než v zeleně značeném vzorku**

$\text{Log}_2\ R/G = -1$ — **ve vzorku značeném červeně je poloviční množství kopií specifické mRNA než v zeleně značeném vzorku**

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)

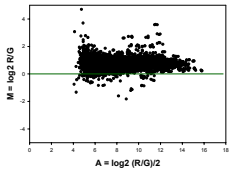
stejné sondy na různých místech sklíčka

Normalizace

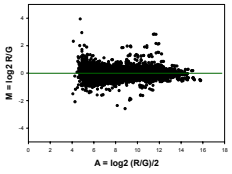
nalezení a odstranění systematických chyb, které nejsou způsobeny biologickým objektem

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

Před normalizací:



Po normalizaci:



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** riziko s jakou lze nulovou hypotézu odmítnout

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

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

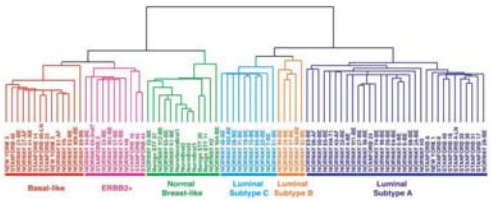
Microarrays: **8 102 cDNA klonů**
každý vzorek (Cy3) hybridizován s referenční RNA (Cy5)

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

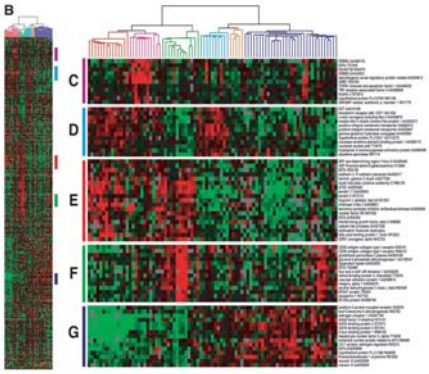
Otázka: **Zda existuje rozdělení karcinomů do podskupin, které mají podobné změny v expresi genů?**

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

Klastrování

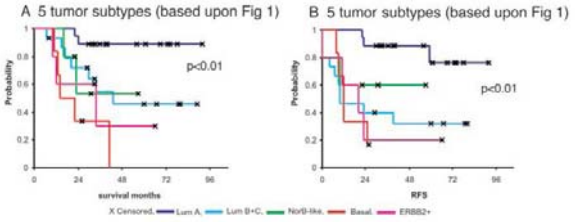


Sorlie et al., PNAS 98: 10869-10874, 2001.



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

Rozdělení do skupin a prognóza vývoje onemocnění



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

Podobné studie

3. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536.
4. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98: 262–272.
5. Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, et al. (2004) Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol 2: E7.
6. Paik S, Shak S, Tang G, Kim C, Baker J, et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351: 2817–2826.
7. Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, et al. (2005) Genes that mediate breast cancer metastasis to lung. Nature 436: 518–524.
8. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671–679.
9. Chi JT, Wang Z, Nuyten DS, Rodriguez EH, Schaner ME, et al. (2006) Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. PLoS Med 3: e47.
10. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, et al. (2007) The prognostic role of a gene signature from tumorigenic breast-cancer cells. N Engl J Med 356: 217–226.
12. Hu Z, Fan C, Oh DS, Marron JS, He X, et al. (2006) The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 7: 96.
21. Finak G, Bertos N, Pepin F, Sadekova S, Souleimanova M, et al. (2008) Stromal gene expression predicts clinical outcome in breast cancer. Nat Med 14: 518–527.
31. van Vliet MH, Reyal F, Horlings HM, van de Vijver MJ, Reinders MJ, et al. (2008) Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. BMC Genomics 9: 375.

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)