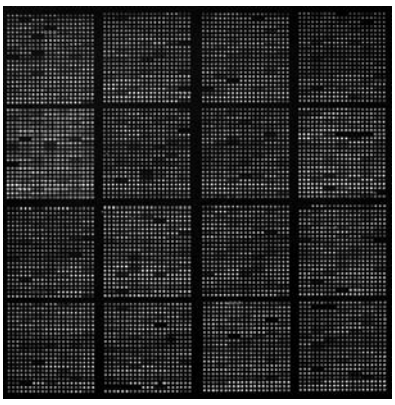
Principy microarrays

Pavla Gajdušková

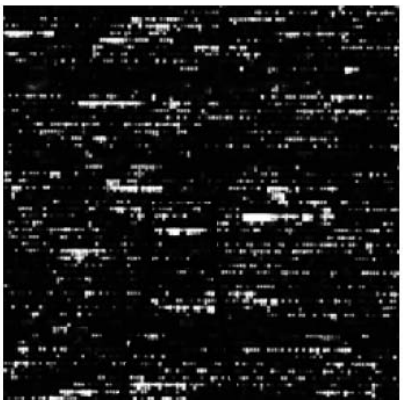
Analytická cytometrie, 15., 22. a 29. listopadu 2011

Microarrays

Kolekce DNA sond přichycených k pevnému podkladu



„Tištěná“ microarrays



Fotolitografie

Microarray technologie

I. Výběr sond (probes): cDNA vektory, BAC vektory, krátké nebo dlouhé oligonukleotidy, proteiny, tkáně

II. Příprava microarray: nanesení sond na sklo nebo membránu

III. Design experimentu: zvolení správné metody, použití referečního vzorku, záměna fluorescenčních barev

IV. Fluorescenční značení vzorků

V. Analýza microarray obrazů: nalezení sond v obraze, korekce pozadí, výpočet intenzity v jednotlivých bodech

IV. Analýza dat: filtrování, normalizace, porovnání výsledků získaných z více microarray experimentů – klastrovací analýza

Obsah přednášky

Technologie přípravy microarrays

Oblasti použití microarrays v biologii

Úvod do statistického hodnocení dat

Příklady konkrétních aplikací z literatury

Technologie přípravy microarrays

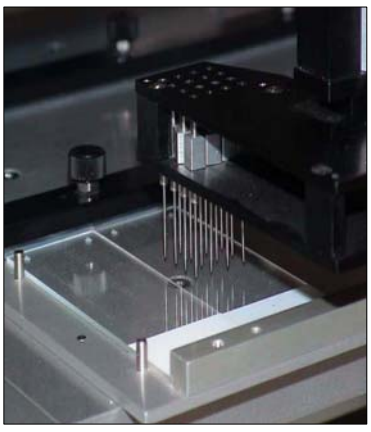
I. tisk pomocí skleněných kapilár (na podložní skla) (výzkumné laboratoře)

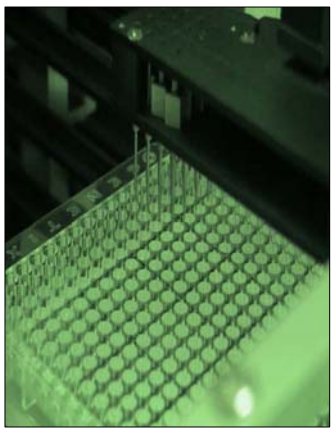
II. ink-jet tisk (Agilent)

III. fotolitografie (Affymetrix, NymbleGen)

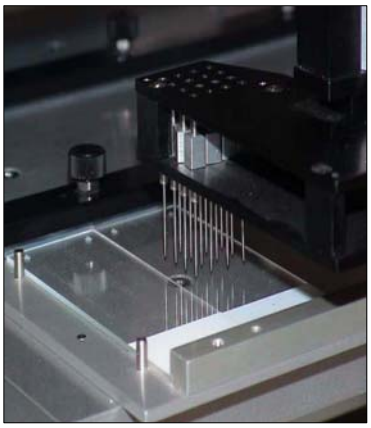
IV. samosestavování silikonových kuliček (Illumina)

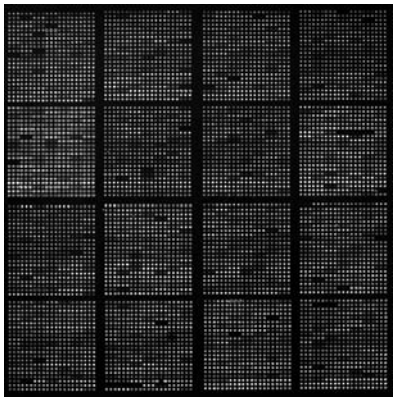
Microarrays tištěná pomocí skleněných kapilár





Microarrays tištěná pomocí skleněných kapilár



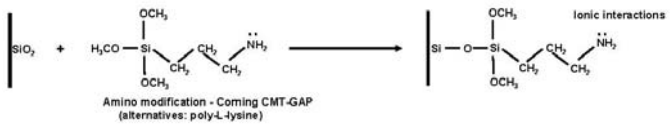


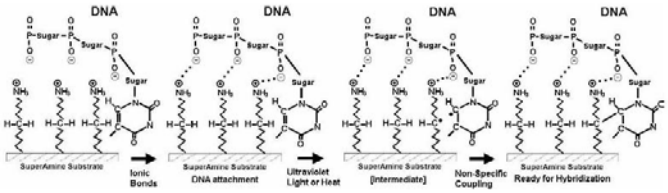
Úprava povrchu sklíček pro tisk arrays I

povrchová úprava skla: amino modifikace, poly-L-lysine

modifikace povrchu → natisknutí DNA sond → UV ozáření

Sklo





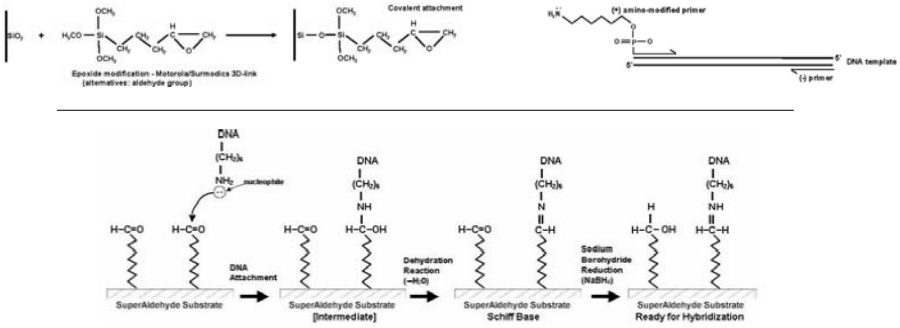
From Lee N. H. presentation: Introduction to High Density Microarrays

Úprava povrchu sklíček pro tisk arrays II

povrchová úprava skla: epoxidová modifikace

úprava DNA: amino-modifikace DNA

Sklo **DNA**



From Lee N. H. presentation: Introduction to High Density Microarrays

Zdroj DNA pro tisk pomocí kapilár

I. **dlouhé oligonukleotidy:**
~ 60 - 70mers
komerčně dostupné (Operon, Agilent)

II. **cDNA**:
knihovny cDNA vektorů (IMAGE, MGC)
dostatečné množství DNA se vyprodukuje pomocí PCR (univerzální primery pro daný typ vektorů)

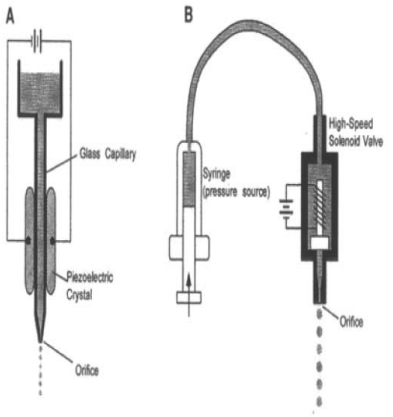
III. **BAC** (**B**acterial **A**rtificial **C**lones): malý výtěžek při izolaci, vysokomolekulární (lepivá) DNA, nutná následná amplifikace DNA spojená s rozdělením na menší úseky (DOP-PCR, ligation-mediated PCR)

“ink-jet” tisk

oligonukleotidy 60 bazí

pravidelnější tvar
a rozmístění bodů

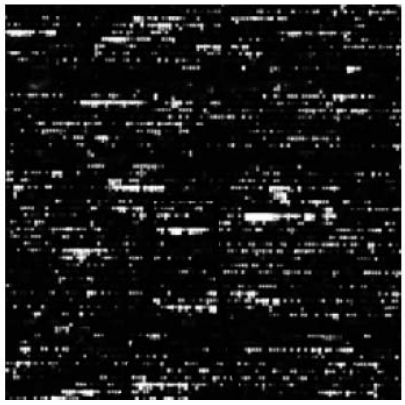
firma: Agilent



(From *Microarray Biochip Technology*, ed. M. Schena)

Fotolitografický způsob přípravy

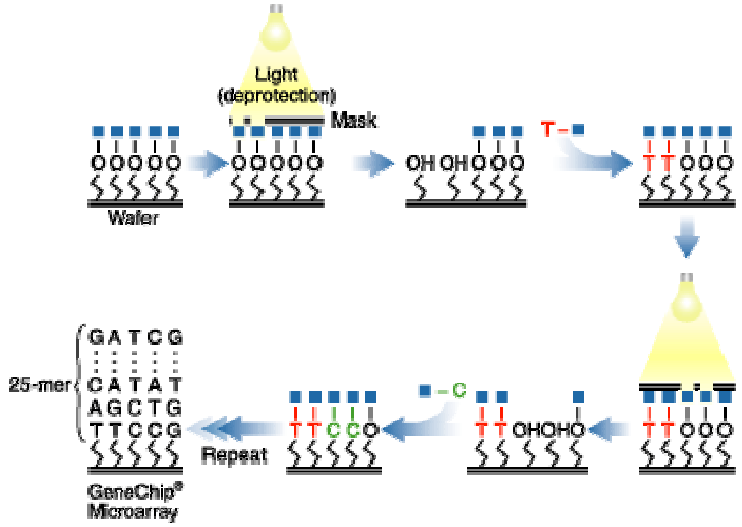
syntéza oligonukleotidů přímo na membráně



„In-situ“ syntéza

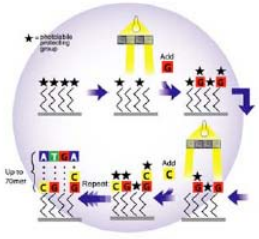
Fotolitografický způsob přípravy (Affymetrix)

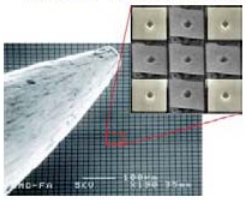
sondy = oligonukleotidy délky 25 bazí

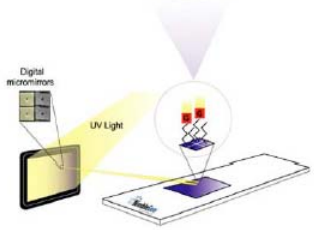


Fotolitografický způsob přípravy (NimbleGen)

sondy = oligonucleotidy 45-85 bazí
podobná teplota tání (T_m)



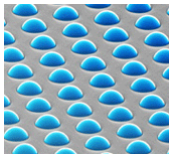




Samosestavování silikonových kuliček

základní stavební jednotka: silikonová kulička (3uM), která je pokryta mnoha kopiemi stejných specifických oligonukleotidů

kulička nemá přesně dané místo na sklíčku, po fixaci na sklíčku je její typ identifikován díky sekvenci části oligonukleotidu



Obsah přednášky

Technologie přípravy microarrays

Oblasti použití microarrays v biologii

Úvod do statistického hodnocení dat

Příklady konkrétních aplikací z literatury

Oblasti použití microarrays v biologii

Typ array	Sondy na microarray	Co se fluorescenčně značí a hybridizuje	... analýza čeho
Expresní	DNA (cDNA, oligonucleotidy)	cDNA / mRNA	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)

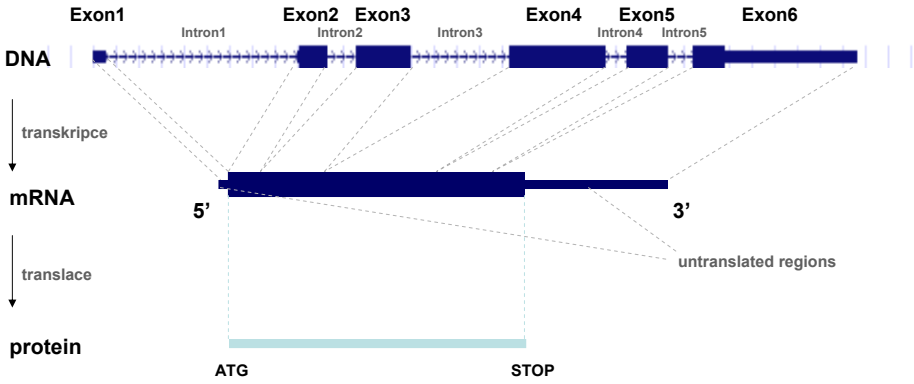
Oblasti použití microarrays v biologii

DNA —transkripce→ **RNA** —translace→ **protein**

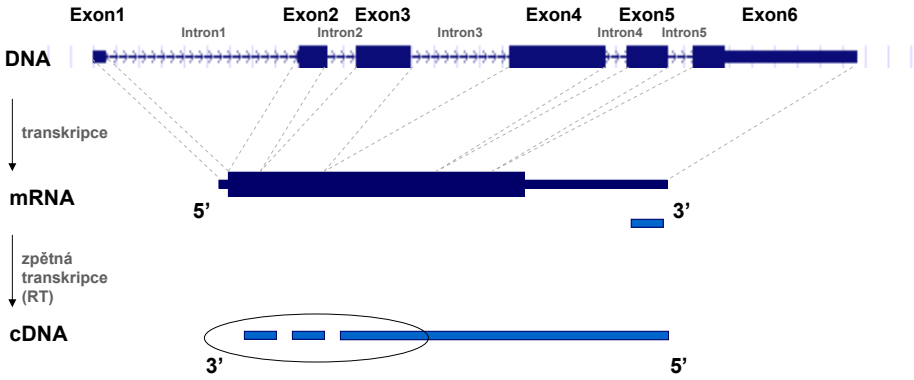
Oblasti použití microarrays v biologii

Schena M., Shalon D., Davis R. W., Brown P. O.
Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-70, 1995.

Genová exprese



Genová exprese



cDNA: jednořetězcová DNA (v dalším kroku je možné syntetizovat druhý řetězec)
u genů s dlouhou mRNA nemusí vznikat vždy celá cDNA

Metody měření množství mRNA

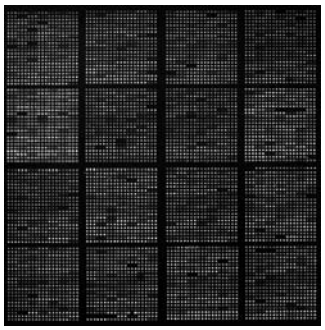
Method	Typical Throughput	Comments
•Northern blot	1 gene	•Standard procedure; "Gold standard", Low throughput
•Subtractive cloning		•Mid-1980's, Not comprehensive, High FP
•Differential display		•1992, Follow up cloning required; Potential to identify rare mRNAs, High FP (Liang & Pardee, Science 257: 967-71, 1992)
•RT-PCR and Real-time RT-PCR		
•2D protein gel/Mass Spec		•Sequence I.D. & semi-quantification, FP?
•ICAT/Tandem Mass Spec		•2001, Sequence I.D. & quantification, FP? (Han et al., Nature Biotech 19: 946-951, 2001)
•EST/SAGE		•1995, Prior sequence knowledge not mandatory, Moderate FP - depends on level of survey (Lee et al., PNAS 92: 8303-7, 1995; Velculescu et al. Science 270: 484-7, 1995)
•High density arrays	20,000-40,000 genes	•1995, Identification of differentially expressed genes dependent on arrayed elements, Low FP (Schena et al. Science 270: 467-70, 1995)

Increasing throughput

From Lee N. H. presentation: Introduction to High Density Microarrays

Měření množství mRNA

(microarrays tištěné pomocí skleněných kapilár)

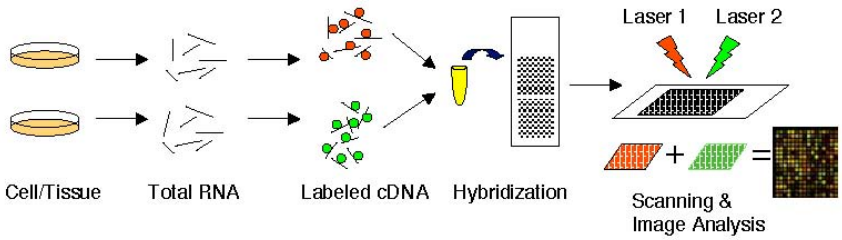


Typ sond

I. **dlouhé oligonukleotidy:**
~ 60 - 70mers
komerčně dostupné (Operon, Agilent)

II. **cDNA**:
knihovny cDNA vektorů (IMAGE, MGC)
dostatečné množství DNA se vyprodukuje pomocí PCR
(univerzální primery pro daný typ vektorů)

Experimentální design



Příklady použití v molekulární biologii (na úrovni mRNA):

- aplikace chemické látky na buněčnou kulturu a její vliv na expresi různých genů (najít geny, které sníží nebo naopak zvýší expresi mRNA)
- zvýšení exprese mRNA zvoleného genu vnesením plasmidu → nalezení dalších genů se změněnou expresí
- snížení exprese mRNA zvoleného genu po vnesení specifické siRNA → nalezení dalších genů se změněnou expresí

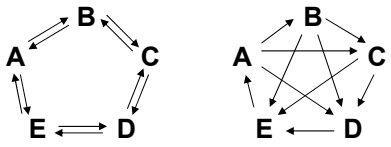
Experimentální design

Porovnání exprese mezi vzorky:

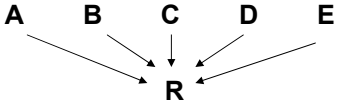
A ⇄ B

B D ??? E A C

1. **Loop design:** každé dva vzorky jsou hybridizovány na jedno sklo (plus vzájemná záměna fluorochromů)



2. **Reference design:** každý vzorek je hybridizován s referenčním vzorkem, který pak slouží jako převodník mezi různými vzorky



Experimentální design

Loop design

poskytuje přímé srovnání mezi vzorky
o každém vzorku získáme více informací - kontrola
vyžaduje větší množství RNA z každého vzorku
špatný vzorek více ovlivní celý experiment

Reference design

lze jednoduše rozšířit o nový vzorek
jednodušší interpretace výsledků
vyžaduje méně RNA ze vzorků
špatný vzorek méně ovlivní celý experiment

Fluorescenční značení

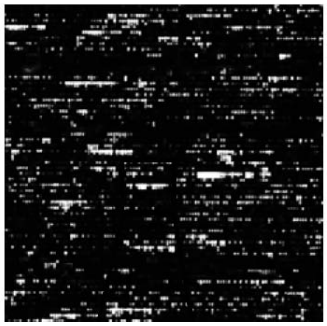
Značení:

Přímé: jeden z nukleotidů je značen fluorescenční značkou
nukleotid s fluorescenční barvou zaujímá více místa→
značen každý 30-35 nukleotid → nižší intenzita fluorescence než nepřímé značení

Nepřímé: jeden z nukleotidů modifikován reaktivní amino skupinou, na kterou se potom váže fluorochrom (NHS ester forma)
pracnější v laboratoři než přímé značení

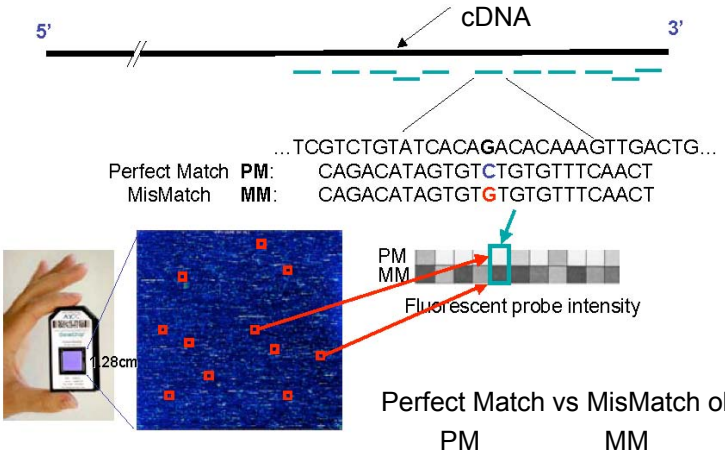
Měření množství mRNA

(fotolitograficky připravené microarrays)



Typ sond

oligonukleotidy 25 bazí



Perfect Match vs MisMatch oligonukleotidy

PM MM

Typ sond

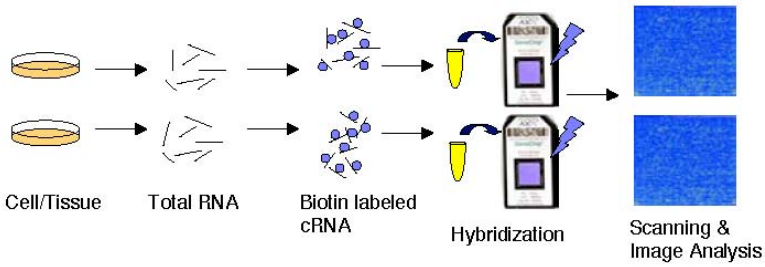
oligonukleotidy 25 bazí

Dříve:
sondy blíže k 3' konci mRNA
11-16 na jeden gen
PM, MM sondy

Nyní:
sondy v různých exonech genu (ideálně 4 sondy v každém exonu)
jenom PM sondy

umožňuje studovat alternativní sestřih

Experimentální design



Příklady použití v molekulární biologii (na úrovni mRNA):

- aplikace chemické látky na buněčnou kulturu a její vliv na expresi různých genů (najít geny, které sníží nebo naopak zvýší expresi mRNA)
- zvýšení exprese mRNA zvoleného genu vnesením plasmidu → nalezení dalších genů se změněnou expresí
- snížení exprese mRNA zvoleného genu po vnesení specifické siRNA → nalezení dalších genů se změněnou expresí

Fluorescenční značení

Nepřímé: jeden z nukleotidů modifikován biotinem, který se detekuje pomocí fluorescenčně značené protilátky až po hybridizaci

biotinem se značí cRNA (in vitro transcription)

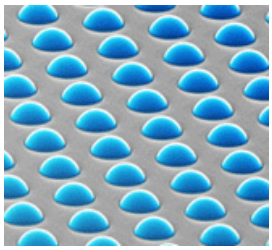
mRNA → first strand cDNA → double strand cDNA → cRNA

Porovnání tištěných a fotolitograficky připravených microarrays

	tisk kapilárami	fotolitografie
počet sond	až 33 000	až 6 500 000
příprava	náročná práce s knihovnami (neplatí pro dlouhé oligo)	jednodušší
tisk	větší variabilita mezi sklíčky	menší variabilita mezi sklíčky
design experimentu	umožňuje přímé srovnání	nepřímé srovnání
alternativní sestřih	nelze studovat (neplatí pro dlouhé oligo)	možné studovat
úprava podle požadavků	jednoduchá	dříve nemožná, dnes možná

Měření množství mRNA

(Allumina samosestavovací arrays)



Samosestavování silikonových kuliček

základní stavební jednotka: silikonová kulička (3uM)

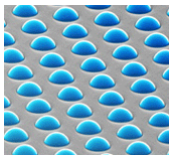
kulička nemá přesně dané místo na sklíčku, po fixaci na sklíčku je její typ identifikován díky sekvenci části oligonukleotidu

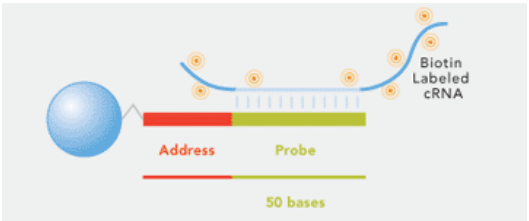
oligonukleotid:

I. **adresa** (definuje typ kuličky)

II. **vlastní sonda** - oligonucleotid (50 bp), který je specifický pro jednotlivé transkripty

míra exprese mRNA = intenzita fluorescence navázané cRNA





Objevování nových transkriptů

objevování nových transkriptů, které nejsou ještě ve veřejných databázích (např. SeqRef, Emsembl)

nebylo to možné pomocí výše zmíněných technologií, protože ty jsou založené na znalostech obsažených v databázích

Řešení:

tilling arrays (Affymetrix)

mRNA sequencing (Illumina)

„Tilling“ arrays

sondy na sklíčku pokrývají kompletně určitou oblast genomu popř. celý genom

repetitivní sekvence nejsou pokryty (před návrhem sond jsou odstraněny pomocí programu „RepeatMasker“)

sondy: oligonukleotidy
  např: 14 arrays, každé obsahuje 2x 3 250 000 sond
  25 bazí sonda, PM a MM, mezera mezi sondami 10 bazí

po hybridizaci s fluorescenčně značenou cRNA „svítí“ sondy, které představují transkribovaná místa ve studované oblasti (genomu)

sondy v místech „bez transkripce“ mají intenzitu fluorescence na úrovni pozadí

lze detekovat nové exony, jejich alternativní sestřih

mRNA sequencing

objevování nových transkriptů pomocí Illumina sekvenační technologie

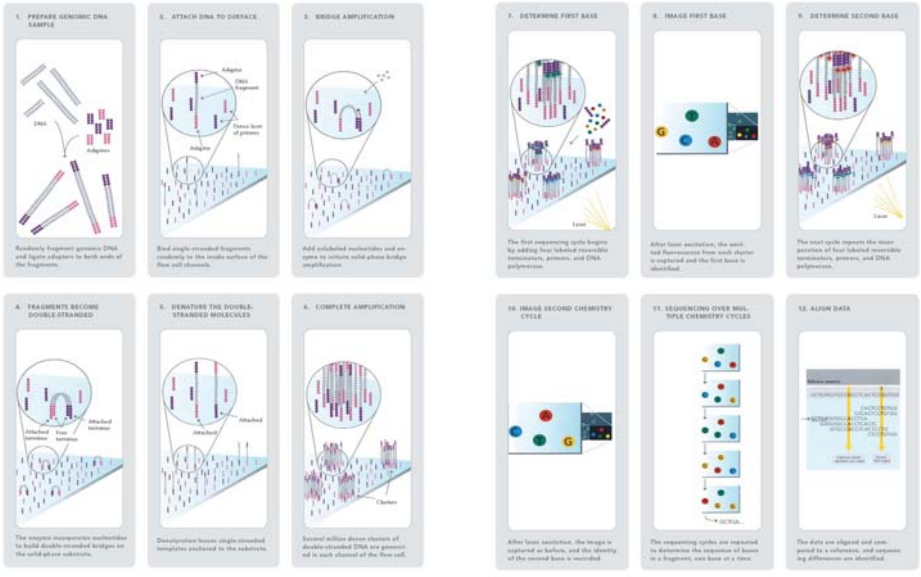
není potřeba navrhovat, tisknout nebo syntetizovat sondy

mRNA → first strand cDNA → double-stranded cDNA → fragmentace

↓

sekvenace (Illumina technologie) ← ligace adapterů

mRNA sequencing



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)

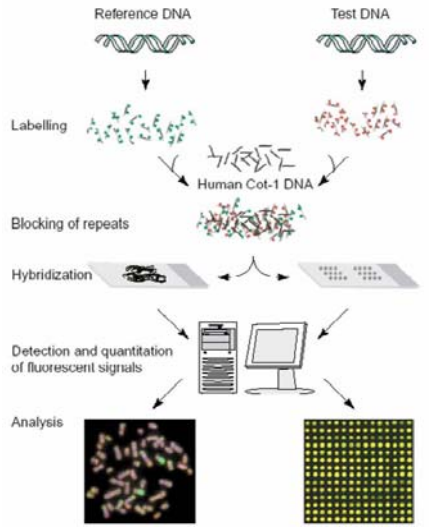
Komparativní genomická hybridizace (CGH)

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

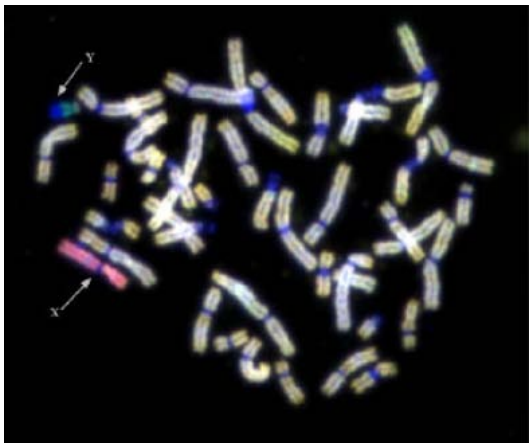
Komparativní genomická hybridizace (CGH)

metafázní chromozomy
- dárce s normálním diploidním karyotypem

DNA:
cy3
zkoumaný vzorek

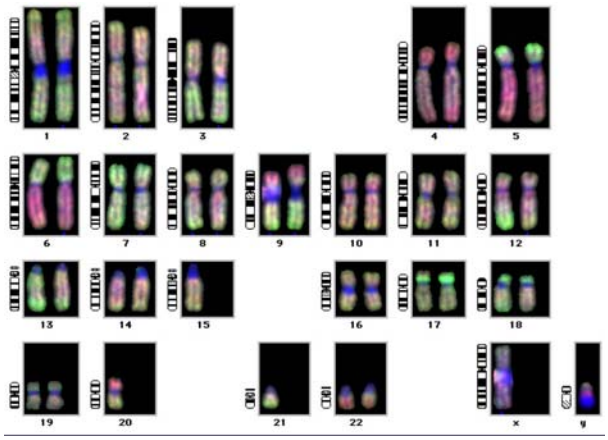
cy5
referenční DNA – 2n

rozlišení ~ 20MB



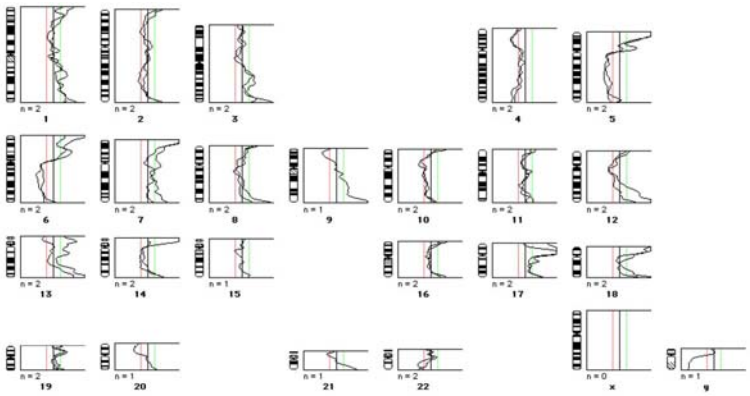
From Szuhai K. presentation: Determination of Genomic Imbalances by Genome-wide Screening Approaches

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.

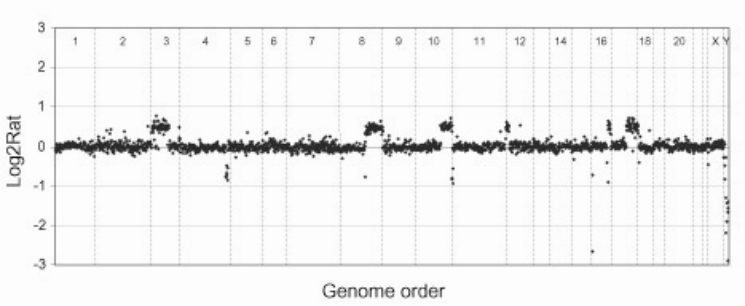
Greshock J. et al., 1-Mb Resolution Array-Based Comparative Genomic Hybridization Using a BAC Clone Set Optimized for Cancer Gene Analysis. Genome Res 14: 179-187, 2004.

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**

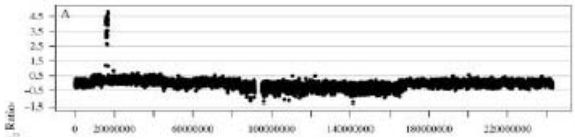
Typy sond natištěných na microarray sklíčku

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

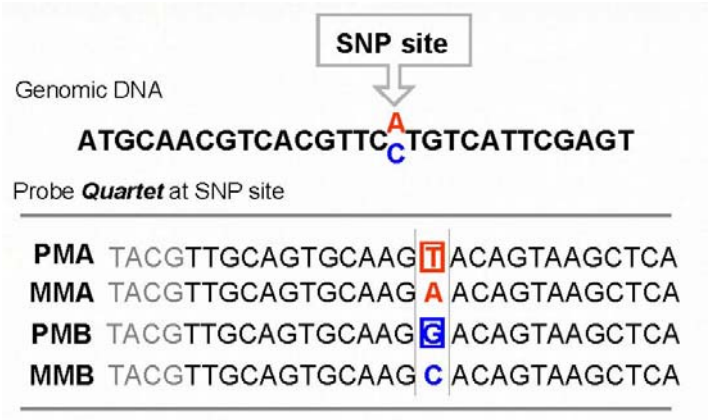
jednonukleotidové variace, které jsou náhodně rozmístěny v genomu (bodové mutace rozšířené v populaci)

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)

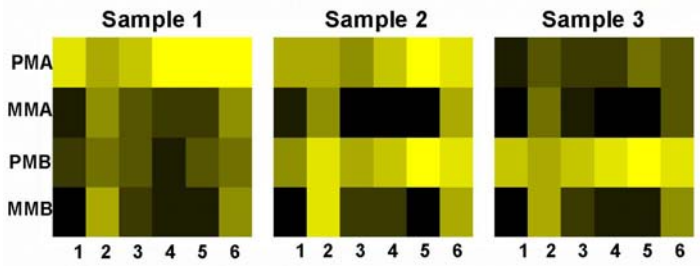


From Xiao Y. presentation: Exploration and Analysis of Affymetrix SNP Arrays. Center for Bioinformatics & Molecular Biostatistics, UCSF Division of Biostatistics

SNP Arrays – probe design

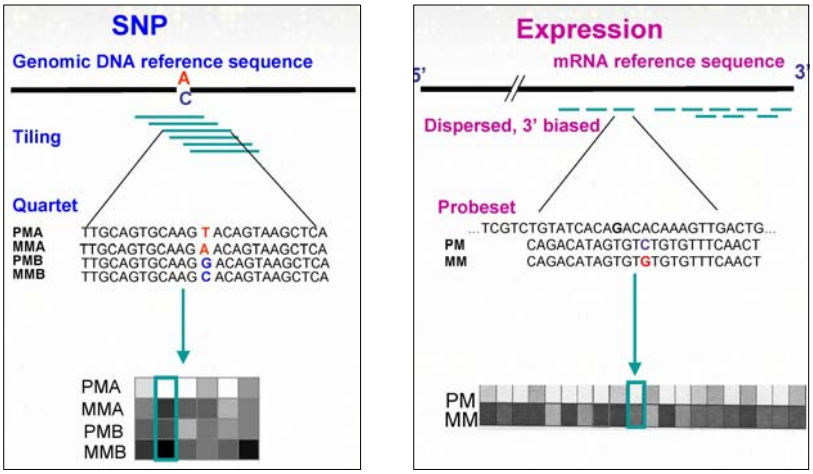
-4	-2	-1	0	1	4
PMA_1	PMA_2	PMA_3	PMA_4	PMA_5	PMA_6
MMA_1	MMA_2	MMA_3	MMA_4	MMA_5	MMA_6
PMB_1	PMB_2	PMB_3	PMB_4	PMB_5	PMB_6
MMB_1	MMB_2	MMB_3	MMB_4	MMB_5	MMB_6

Arrays	No. Quartets
10K	14
100K	10
500K	6



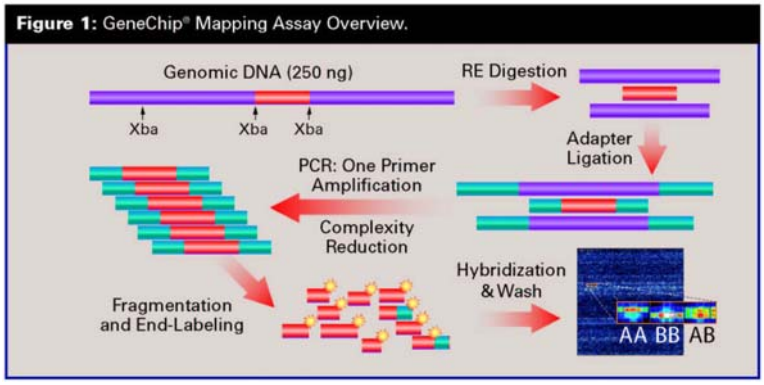
From Xiao Y. presentation: Exploration and Analysis of Affymetrix SNP Arrays. Center for Bioinformatics & Molecular Biostatistics, UCSF Division of Biostatistics

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

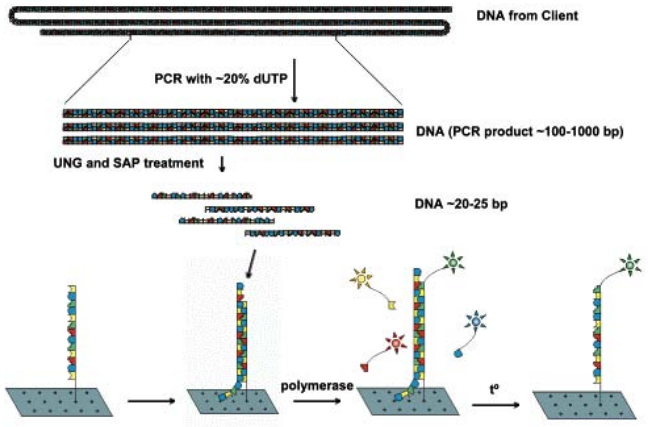
SNP Arrays - labeling



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

SNP Arrays - APEX technologie

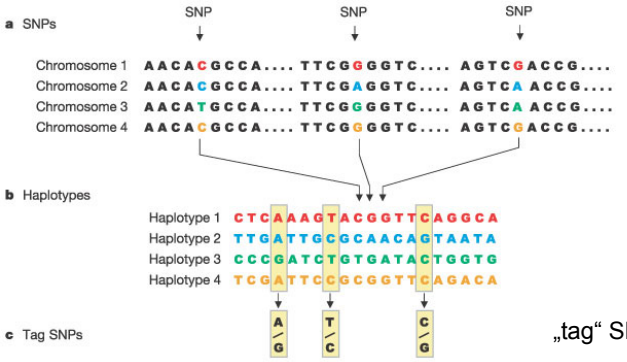
APEX = Arrayed **P**rimer **E**xtension



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 < $5x10^{-7}$

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

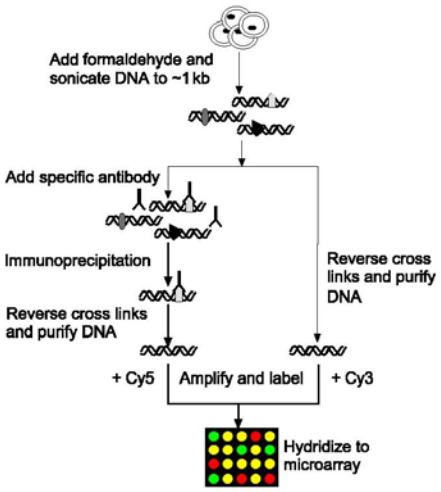
Copy number variation

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)

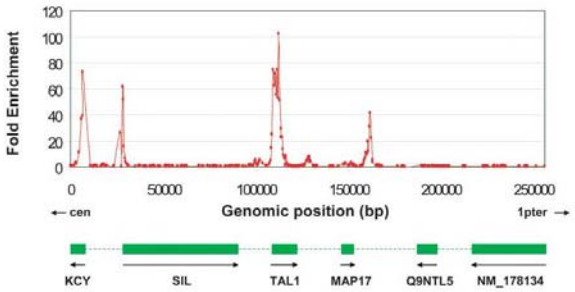
Chromatin ImmunoPrecipitation on chip
ChIP-Chip



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)

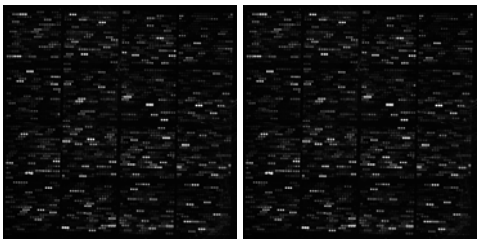
Nalezení rozdílně exprimovaných genů

výpočet zvolené statistiky a následné určení p hodnot

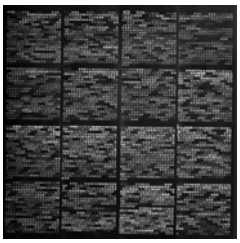
úprava p-hodnot

Analýza obrazu

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



Red **Green**



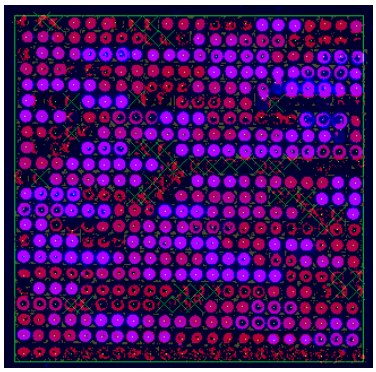
Dapi

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**

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

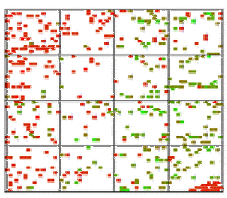
Předpokládáme, že experiment ovlivní expresi pouze malého počtu genů v daném objektu (většina genů svoji expresi nemění)

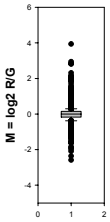
průměr (medián) všech poměrů R/G je roven 1
průměr (medián) všech logaritmů poměrů R/G je roven 0

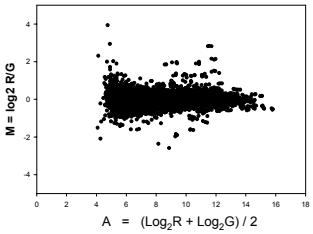
nestačí mít na sklíčku sondy pro geny, které nás zajímají nebo očekáváme, že jejich exprese se bude měnit
pro normalizaci jsou nutné i další geny, jejichž exprese se nemění (těch by měla být většina)

Analýza obrazu

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







Další důležitá hodnota pro kontrolu kvality hybridizace je

průměrná intenzita bodu v obou snímaných kanálech

$$A = (\text{Log}_2 R + \text{Log}_2 G) / 2$$

Odstranění „špatných“ bodů

odstranění bodů: body s morfologickými abnormalitami (problematický tisk)

s nízkou intenzitou (není exprese v daném systemu)

s vysokým pozadím (negativní hybridizace)

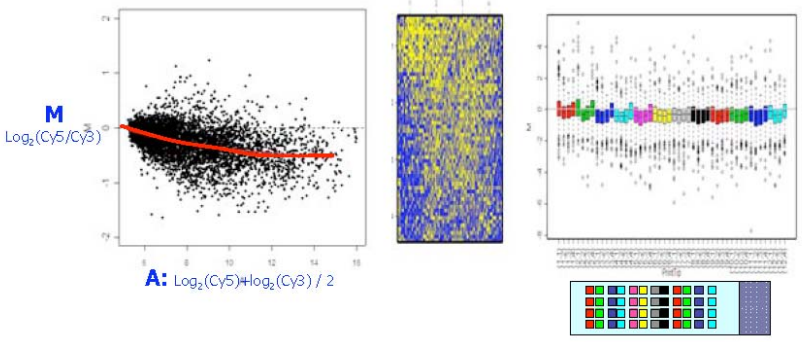
Kontrolní body: prázdné body bez DNA (negativní kontrola)

„spiked“ body (pozitivní kontrola)

stejné sondy na různých místech sklíčka

Normalizace

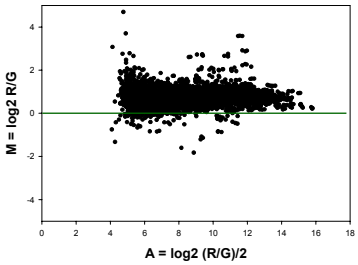
nalezení a odstranění systematických chyb, které nejsou způsobeny biologickým objektem



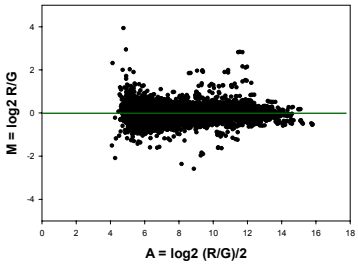
Normalizace

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

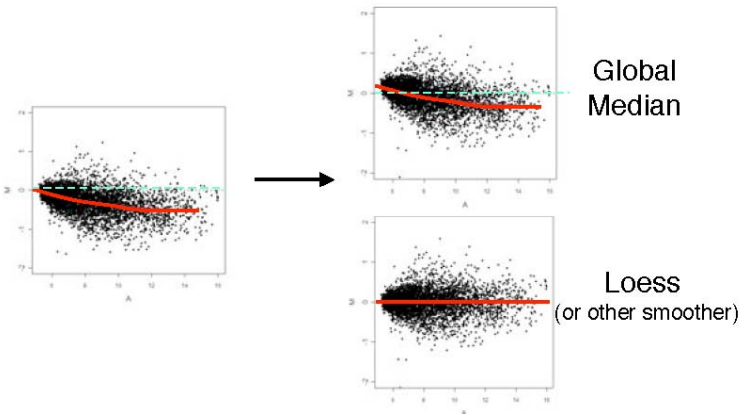
Před normalizací:



Po normalizaci:



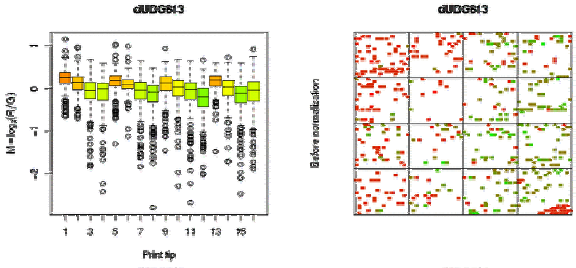
Loess Normalizace



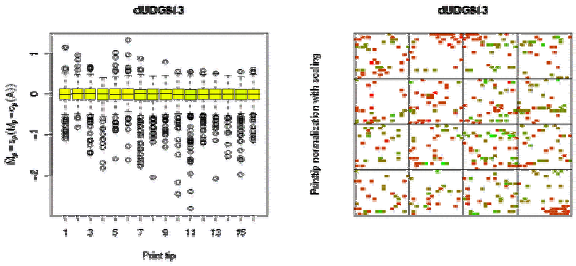
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

“Print Tip” Normalizace

Před normalizací:



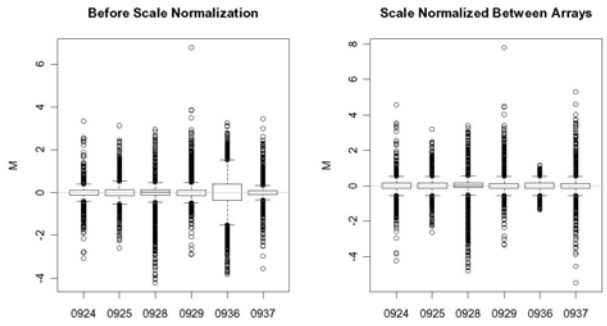
Po normalizaci:



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** 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**

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

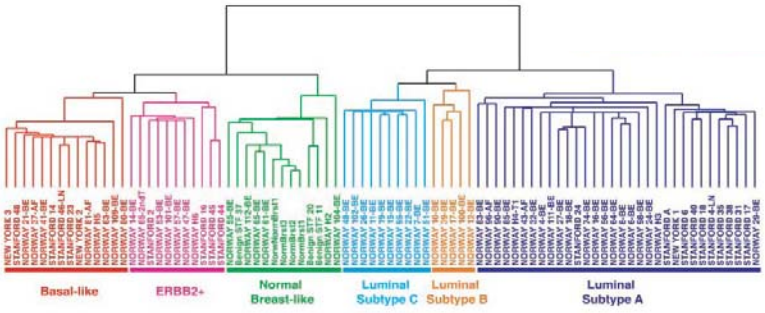
Microarrays: **8 102 cDNA klonů**
každý vzorek (Cy3) hybridizován s referenční RNA (Cy5)

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

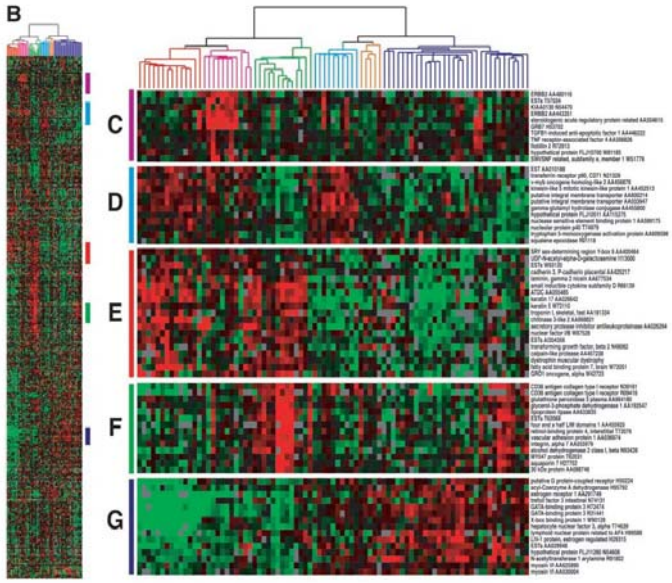
Otázka: **Zda existuje rozdělení karcinomů do podskupin, které mají podobné změny v expresi genů?**

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

Klastrování

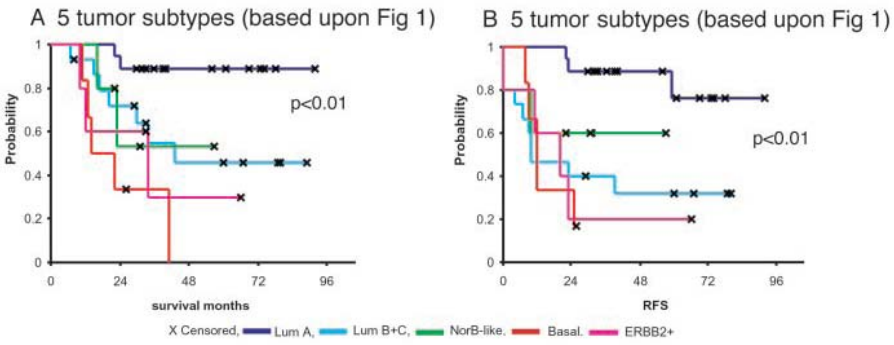


Sorlie et al., PNAS 98: 10869-10874, 2001.



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)