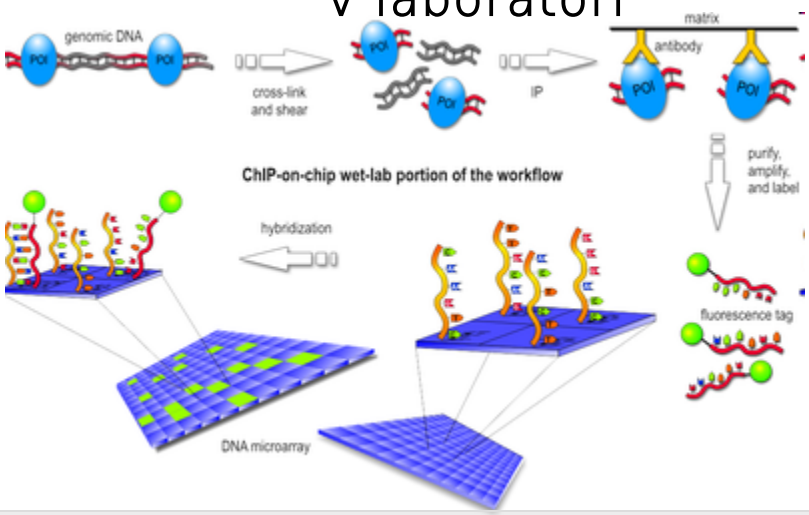
Kapitola II

Technologie studující genomiku a proteomiku

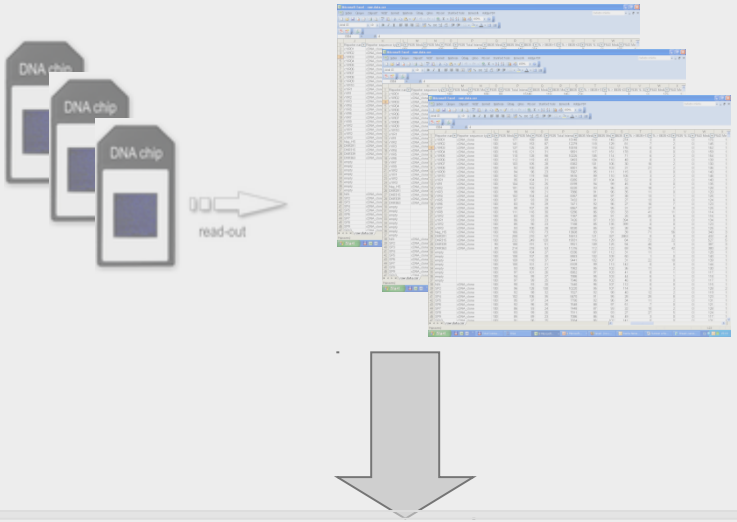
Mikročipy (microarrays)

Průběh genomického experimentu

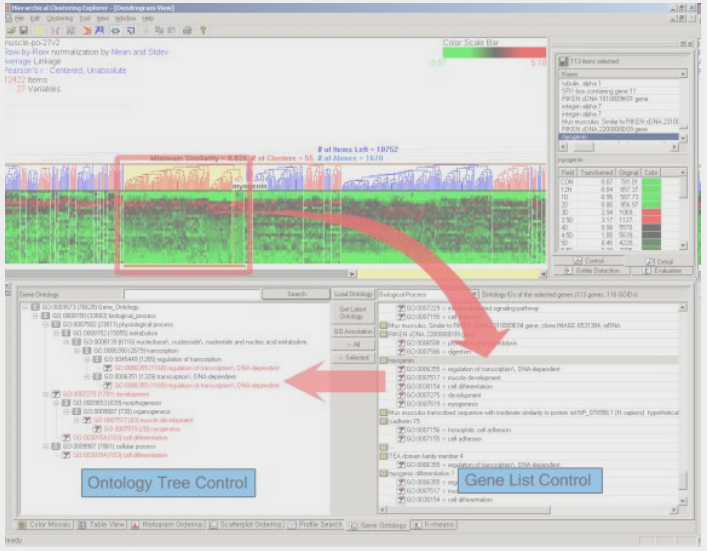
1. Příprava a provedení experimentu v laboratoři



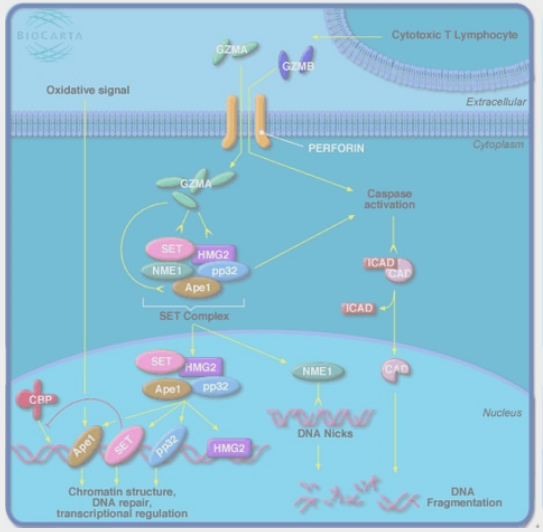
2. Extrakce a úprava dat



3. Statistická analýza dát



4. Biologická a klinická interpretace

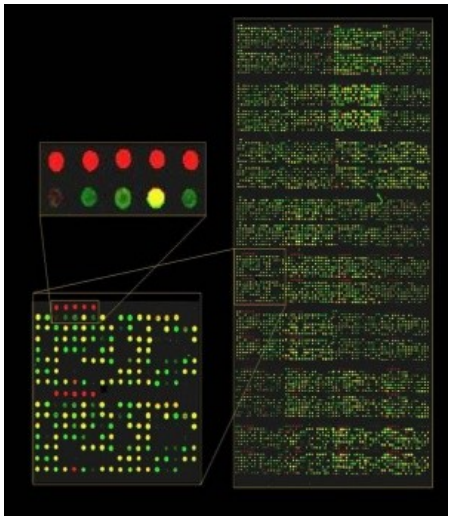


Technika mikročipů

- Mikročipy– biotechnologie simultánně srovnávající biologické objekty (molekuly, tkanivá) na základě jejich imobilizace na jediný **podklad** do oblastí (spotů) které jsou pravidelně uspořádány do řádků a sloupců

- **Podklad**: sklo, gel, parafin, ...

- Mikročipy v genomice a proteomice:
  - DNA mikročipy
  - Proteínové mikročipy



Kapitola II.1

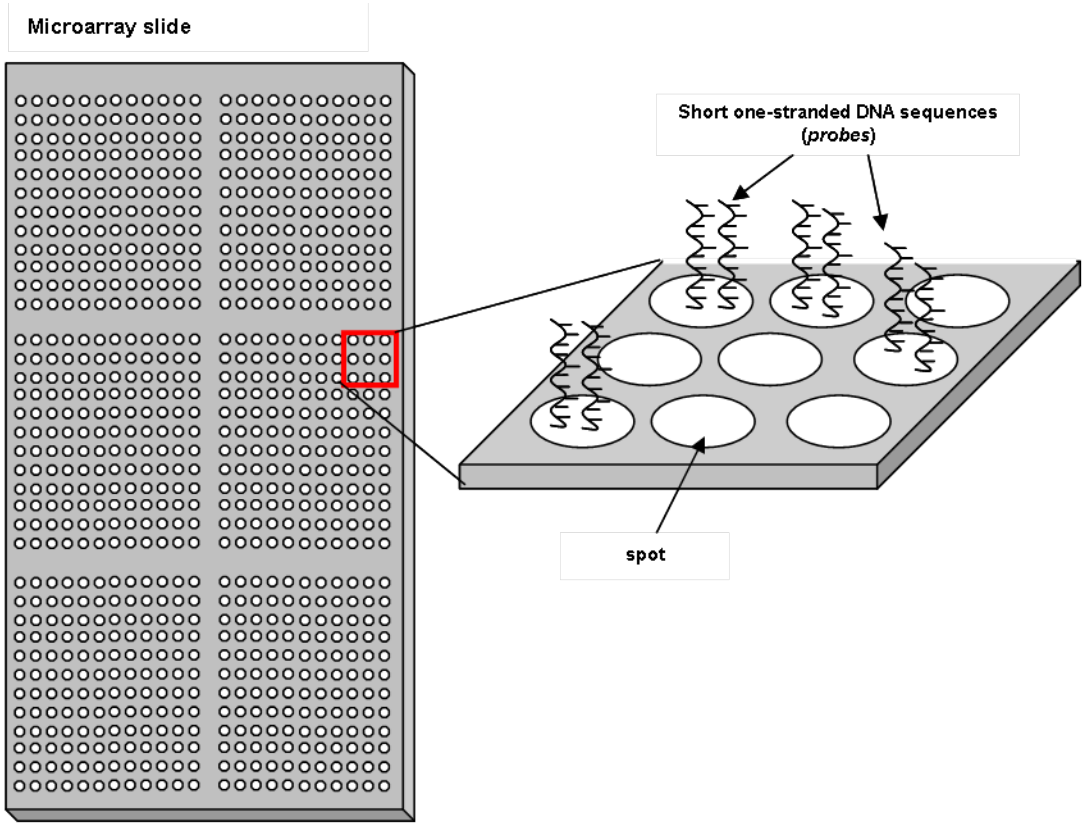
Princip a rozdělení DNA mikročipů

DNA mikročipy

- Séria krátkych DNA sekvencií imobilizovaných rovnomerne na podklad, používaná na detekciu DNA alebo RNA (obvykle vo forme cDNA) vo vzorkách. Najčastejšie aplikovaná na:
  - meranie zmien v hladinách génovej expresie (gene expression profiling, detekcia RNA - cDNA) - **expresné arraye**
  - detekciu štruktúrnych zmien genómu (SNPs- jednonukleotidové polymorfizmy alebo zmeny v počte kópií génov) – **arrayCGH, SNP arrays**
- Taktiež sa úspešne používa na detekciu väzbových miest proteínov na genóme (**ChIP-on-chip**), detekciu alternatívneho zostrihu (**exon junction arrays**) a takisto na presnú detekciu neznámych a nepredikovaných transkriptov alebo alternatívnych foriem zostrihu (**tiling arrays**)

Sonda (probe)

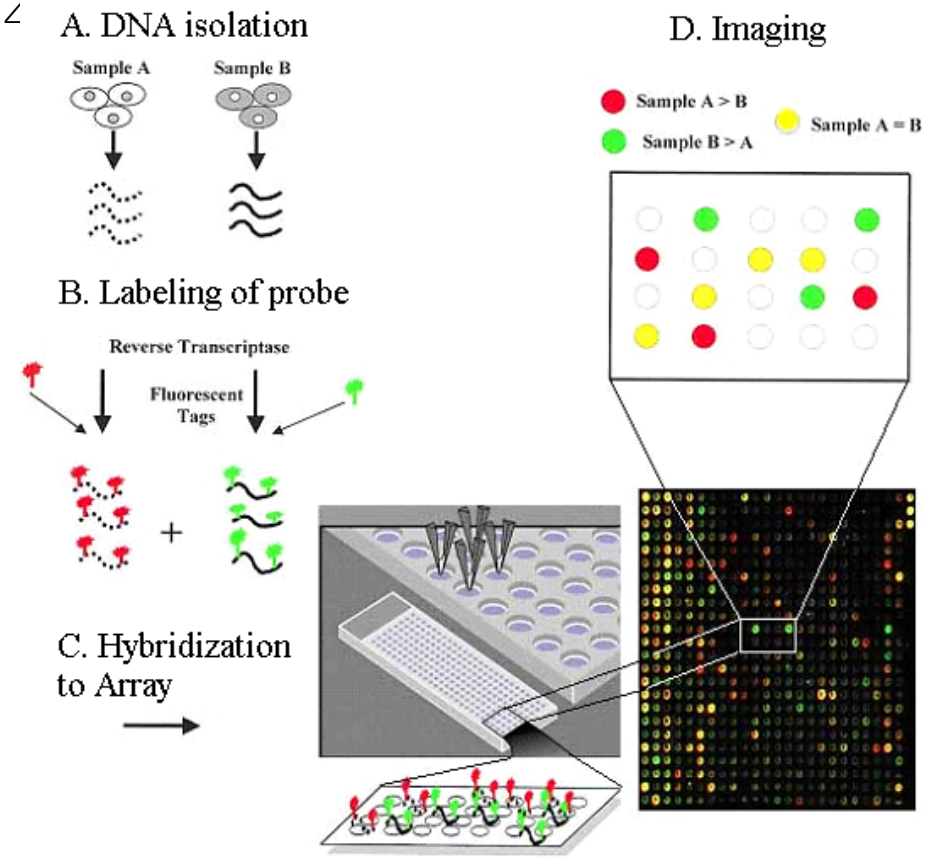
- Krátke DNA sekvencie (oligonukleotidy) na microarray sklíčku sa nazývajú ***sondy***, anglicky *probes*
- Každá oblasť DNA (obvykle gén), ktorú chceme skúmať
- Sondy sú navrhnuté tak, aby boli pre daný gén/oblasť čo najšpecifickejšie



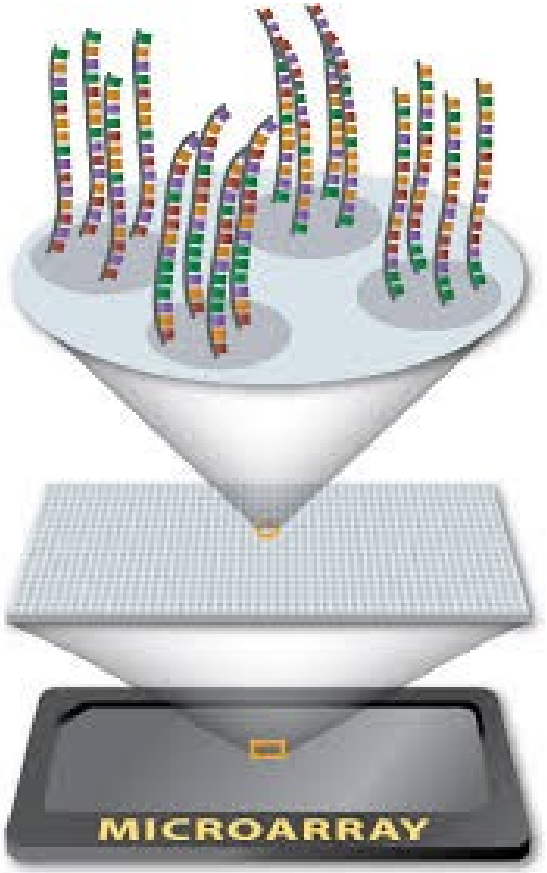
Základný princíp

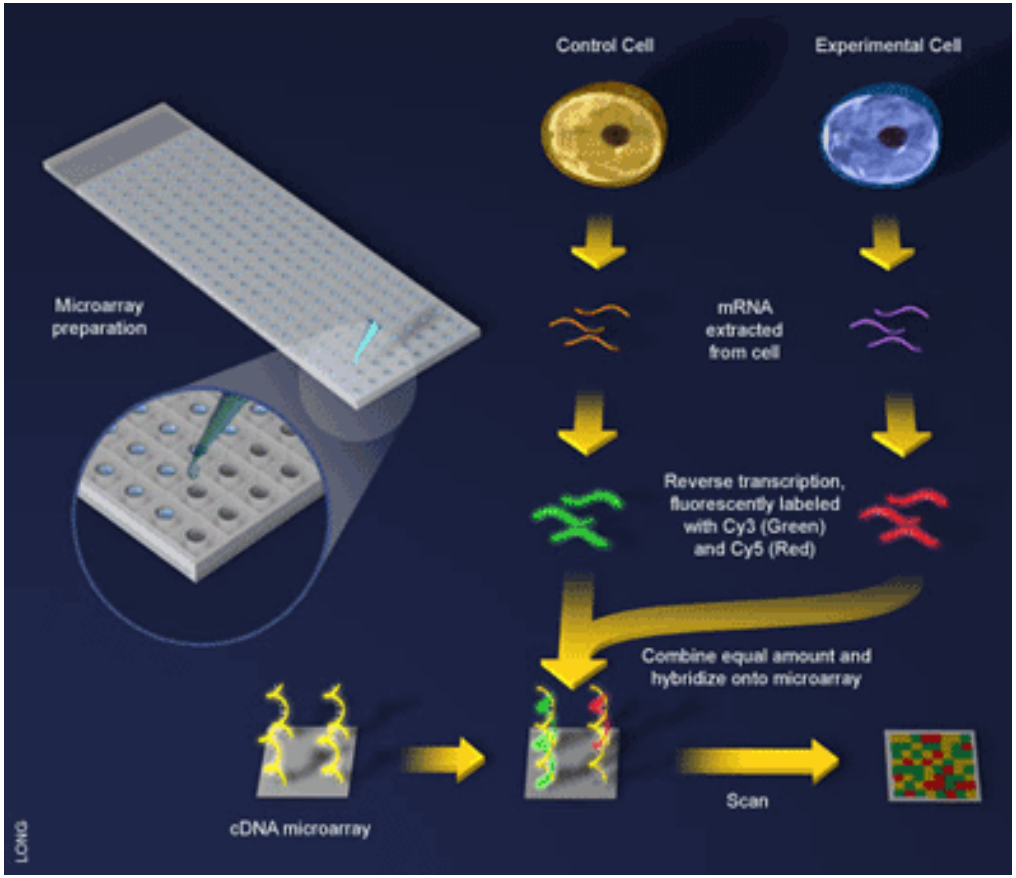
- Fragmenty DNA/cDNA zo vzorky sa **spárujú** s **komplementárnymi** sondami na microarray sklíčku a tým sa **imobilizujú**.

- Imobilizované molekuly DNA, ktoré boli predtým označené **fluorescenčným farbivom** sa potom dajú detekovať pomocou **UV skenera** a kvantifikovať tak množstvo mRNA/DNA s danou sekvenciou prítomnej vo vz



Mikročipy


MICROARRAY


Control Cell
Experimental Cell
Microarray preparation
mRNA extracted from cell
Reverse transcription, fluorescently labeled with Cy3 (Green) and Cy5 (Red)
Combine equal amount and hybridize onto microarray
cDNA microarray
Scan
LOHG

Postup mikročipového experimentu

1. Výroba mikročipového sklíčka
2. Příprava vzorků
3. Hybridizace

Příprava čipu a vzorků

4. Skenování
5. Analýza obrazu (kvantifikace signálu, vznik expresních dat)

Vznik dat

Postup mikročipového experimentu

1. Výroba mikročipového sklíčka
2. Příprava vzorků

Příprava čipu a vzorků

3. Hybridizace

---

4. Skenování

Vznik dat

5. Analýza obrazu (kvantifikace signálu, vznik expresních dat)

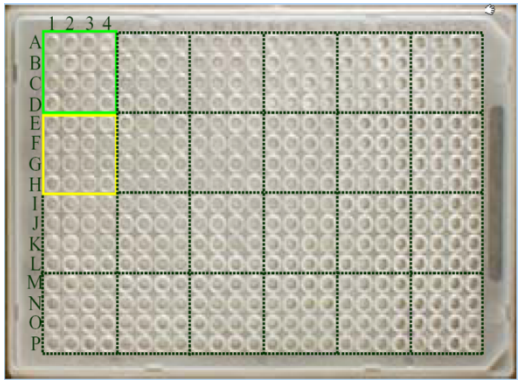
Princip výroby DNA mikročipu

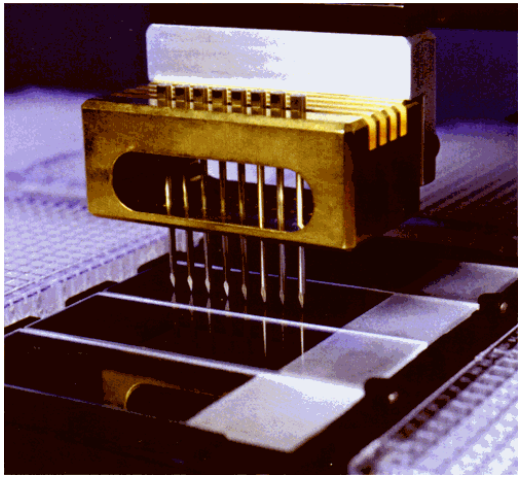
- Výroba sklíčka spočívá v připojení sond na podložné sklíčko do oblastí spotů

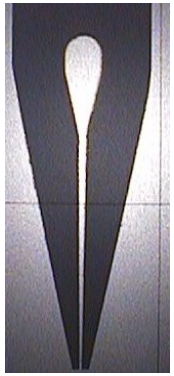
- Dvě hlavní metody:

  - *Spotting* – sondy jsou syntetizované PŘED umístěním na microarray sklíčko, potom umístěné na sklíčko pomocí speciálního robota

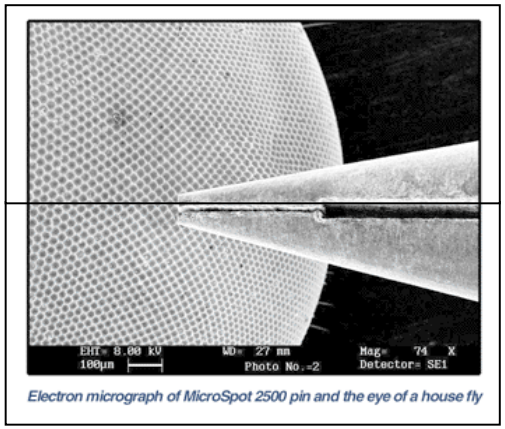
Spotovací robot



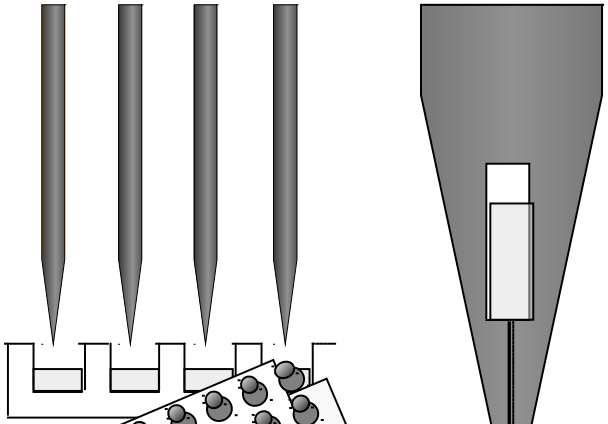


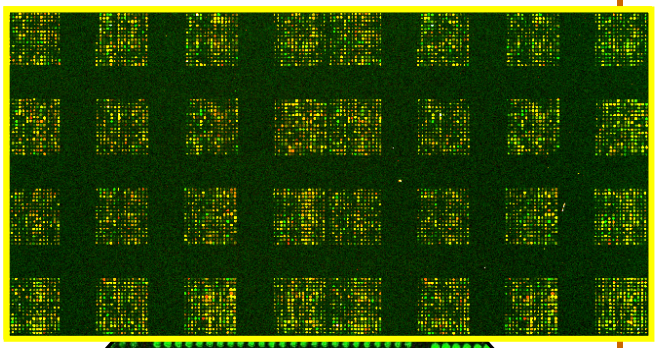


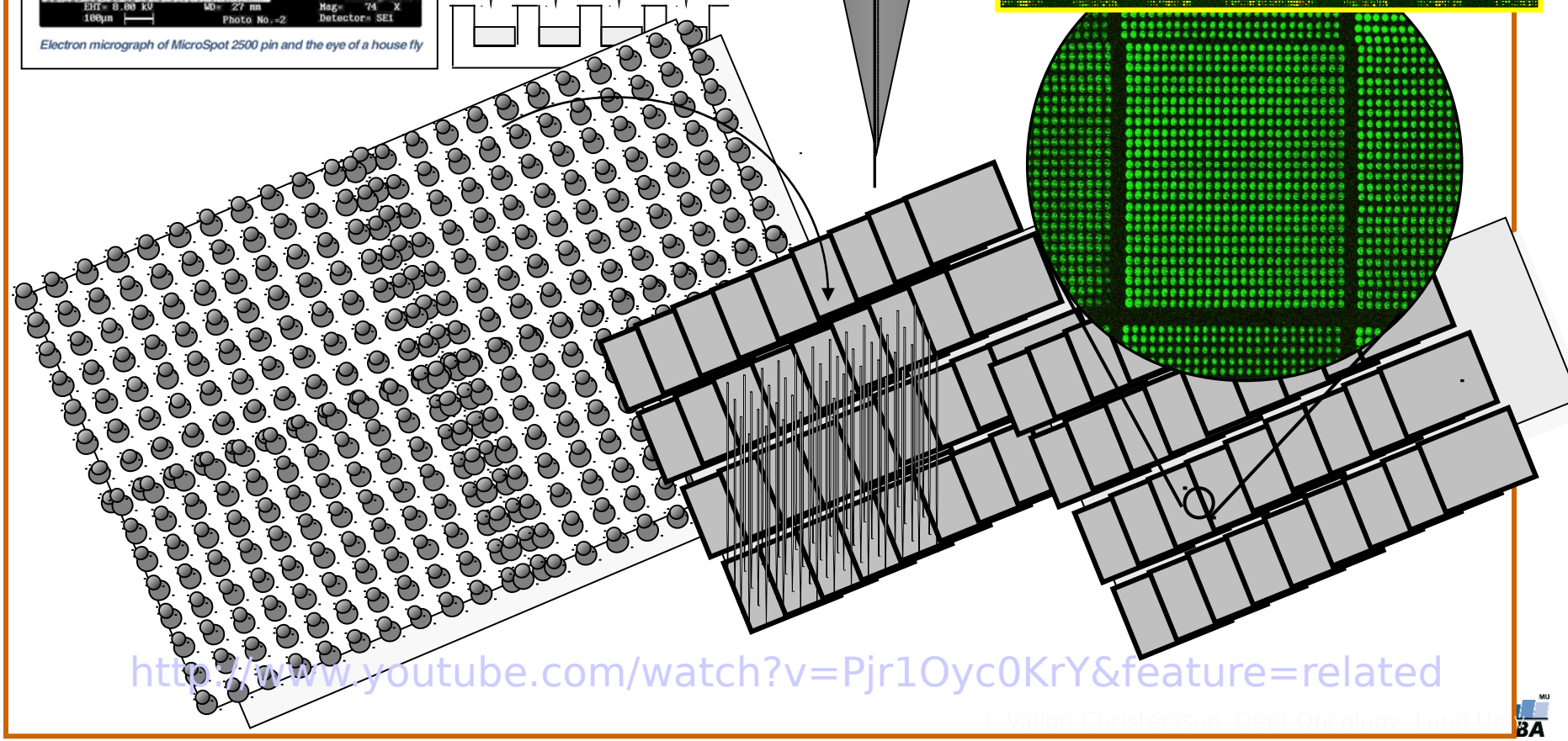
Princip spotování



Electron micrograph of MicroSpot 2500 pin and the eye of a house fly







http://www.youtube.com/watch?v=Pjr1Oyc0KrY&feature=related

Princip výroby DNA mikročipu

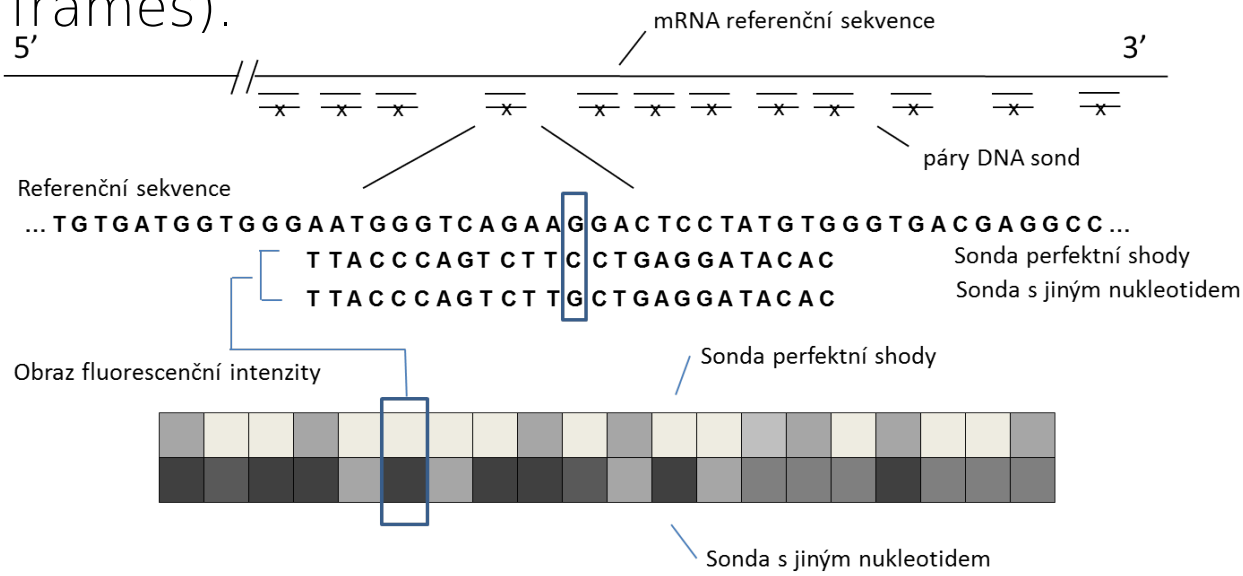
- Výroba sklíčka spočívá v připojení sond na podložné sklíčko do oblastí spotů
- Dvě hlavní metody:
  - *Spotting* – sondy jsou syntetizované PŘED umístěním na microarray sklíčko, potom umístěné na sklíčko pomocí speciálního robota
  - *In-situ syntéza* – sondy jsou syntetizované přímo na podklad, *fotolitografickou syntézou*

    http://www.youtube.com/watch?v=ui4BOtwJEXs&feature=related

- Spotting – u delších cDNA sekvencí
- In-situ syntéza – pro krátke oligonukleotidy

Typy sond

- cDNA sondy - 500-5000 párů bazí dlouhé cDNA klony cílového genu nebo známé sekvence. Obvykle syntetizované před umístěním na microarray sklíčko pomocí spotovacího robota
  - Výhoda: jsou více specifické, a v případě úspěšné hybridizace s cílovou DNA můžeme téměř s jistotou říct, že se spojily právě s daným genem
- Oligonukleotidové sondy – maximálně 25 párů bazí dlouhé sekvence, které jsou designované tak, aby odpovídaly jen částem sekvence známých kódujících genových ORF (open reading frames).



Typ mikročipů dle typu sondy

- Podle typu sondy rozlišujeme:
  - **cDNA mikročipy** – používají cDNA sondu

    - hybridizace závislá na délce sond

  - neznáme přesný počet klonů v každém spotu

Hybridizaci nutno stanovit relativně (k referenci). Tato relativní informace je robustnější než absolutní informace o intenzitě každého spotu. Proto jsou tyto experimenty obvykle **dvoukanálové** (jeden kanál pro DNA, kterou zkoumáme, druhý kanál pro referenční DNA).

- **Oligonukleotidové mikročipy** – oligonukleotidové sondy, obvykle syntetizované in-situ

  - známe přesný počet klonů

    - stejná délka sondy

Není nutná reference, proto jsou **jednokanálové** (jeden vzorek na čip bez reference).

Postup mikročipového experimentu

1. Výroba mikročipového sklíčka
2. Příprava vzorků

Příprava čipu a vzorků

3. Hybridizace

---

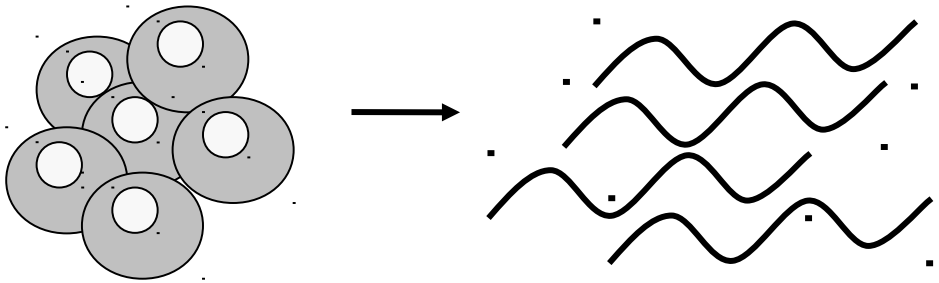
4. Skenování

Vznik dat

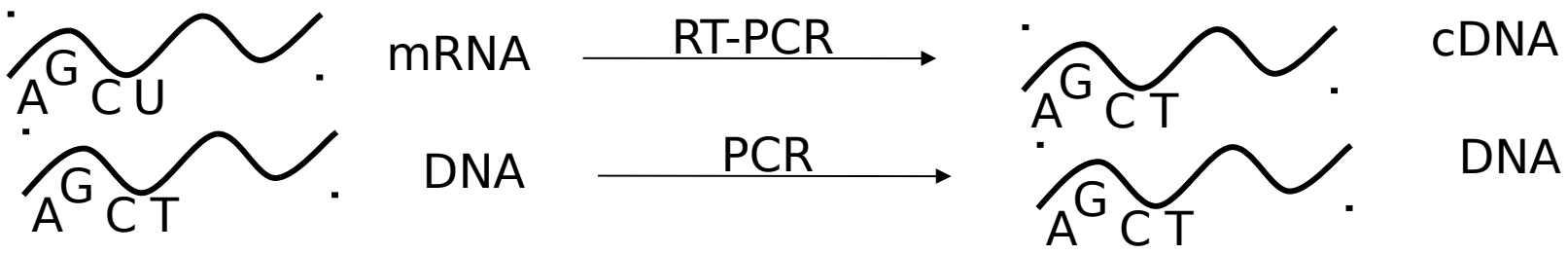
5. Analýza obrazu (kvantifikace signálu, vznik expresních dat)

Příprava vzorků

1. Izolace DNA/RNA: molekuly které chceme zkoumat (DNA či mRNA) jsou extrahované ze vzorku.

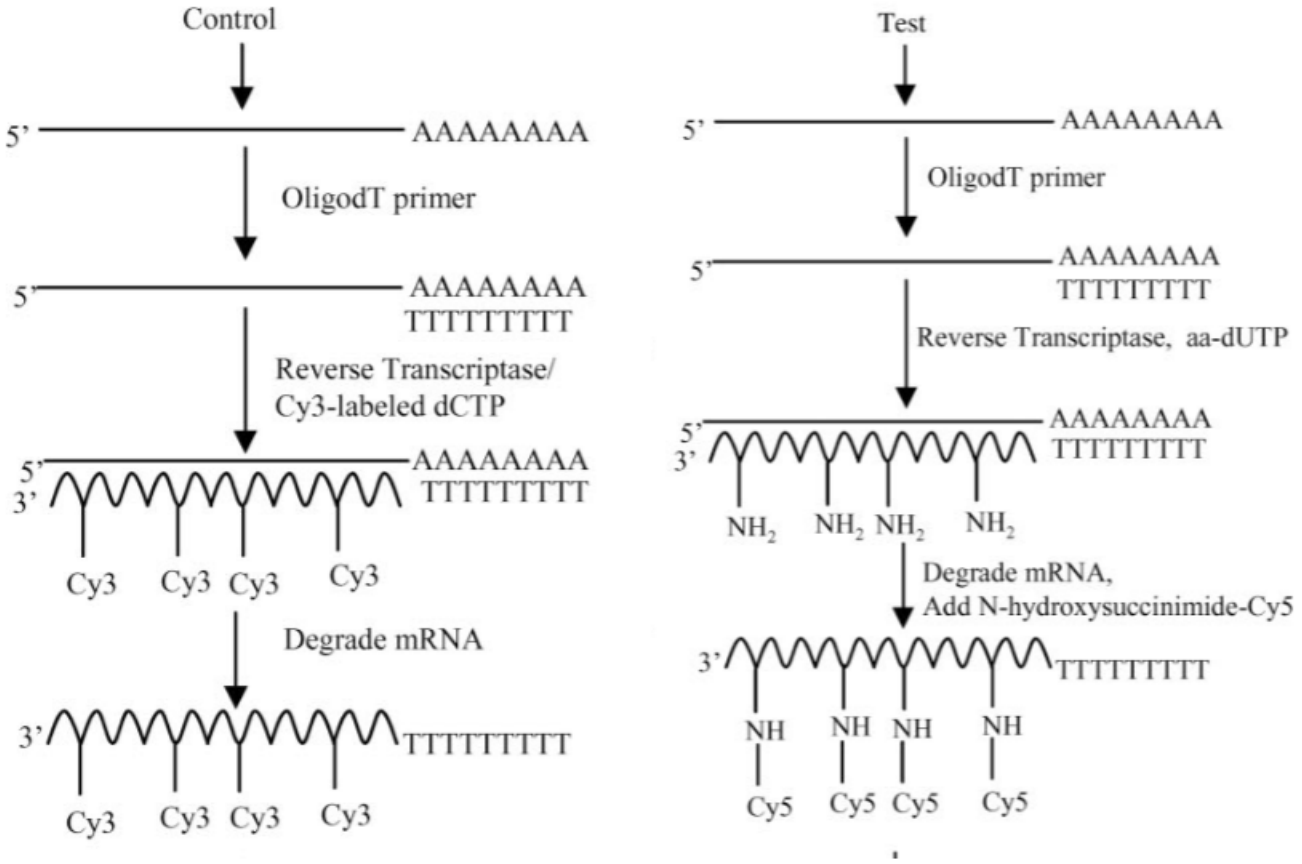


2. Přepis a amplifikace: mRNA se přepisuje do cDNA a amplifikuje se pomocí RT-PCR. DNA zas pomocí PCR.



Příprava vzorků: 3. značení

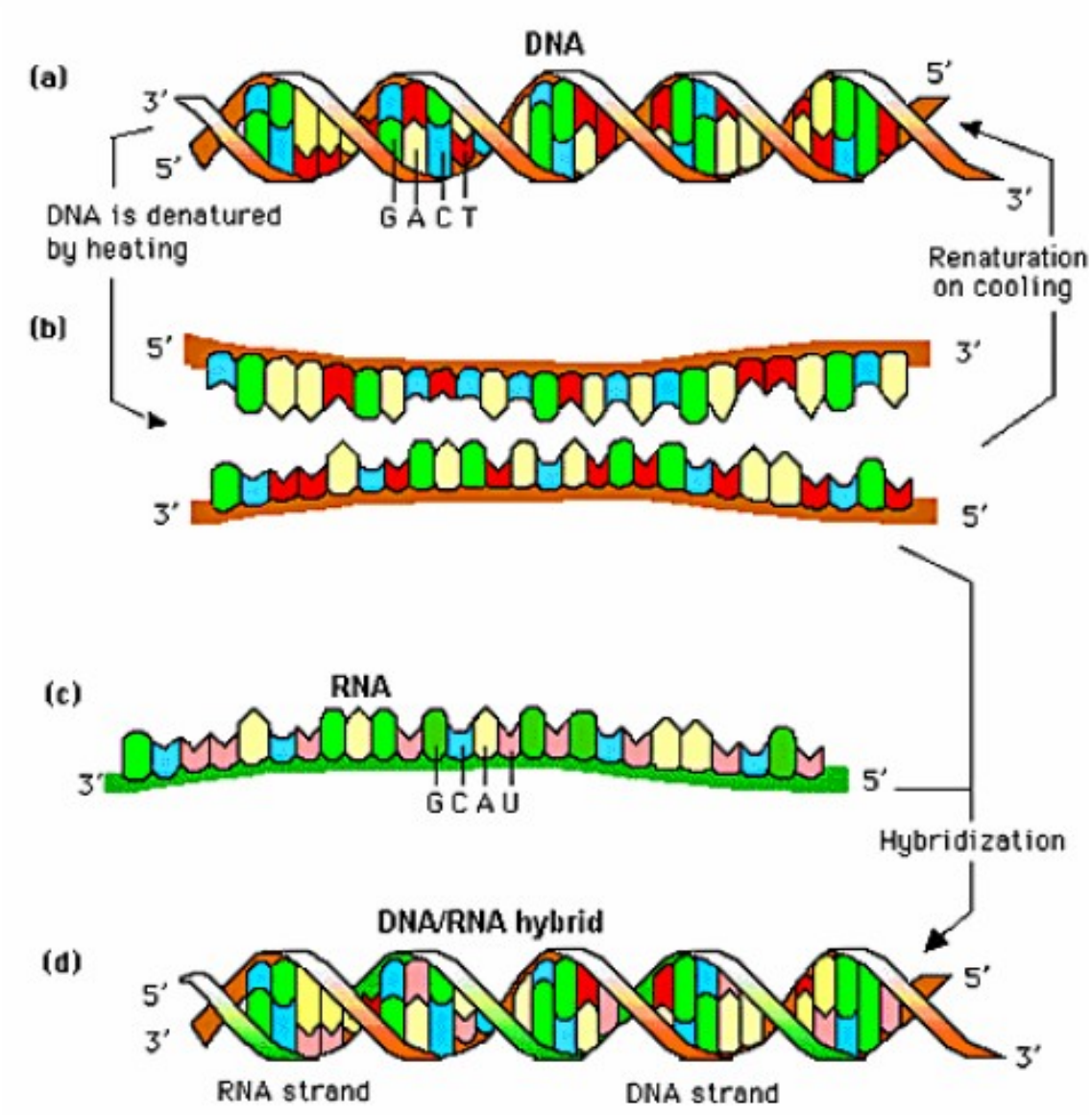
3. Značení: Amplifikovaná DNA (cDNA) je obarvená fluorescenčním barvivem (nejčastěji Cy3 nebo Cy5). Toto s nazývá přímé označení. U nepřímého značení nejdříve skupina, většinou primární amin je inkorporovaná do cDNA a Cy3/Cy5 jsou potom inkorporované do cDNA při následné reakci.



Postup mikročipového experimentu

1. Výroba mikročipového sklíčka
2. Příprava vzorků
3. Hybridizace
4. Skenování
5. Analýza obrazu
6. Kvantifikace obrázku na hodnoty exprese

Hybridizace DNA



- DNA mikročipová technologie je založená na hybridizaci
- **Hybridizace** je proces komplementárního párování dvou jednořetězcových nukleových kyselin do dvouřetězcové molekuly (duplexu) na základě párování bazí.

Hybridizace na mikročipu

1. Fragmentovaná a namnožená cDNA(DNA) vzorku se vylije na microarray sklíčko, kde už jsou navázané jednořetězcové sondy.

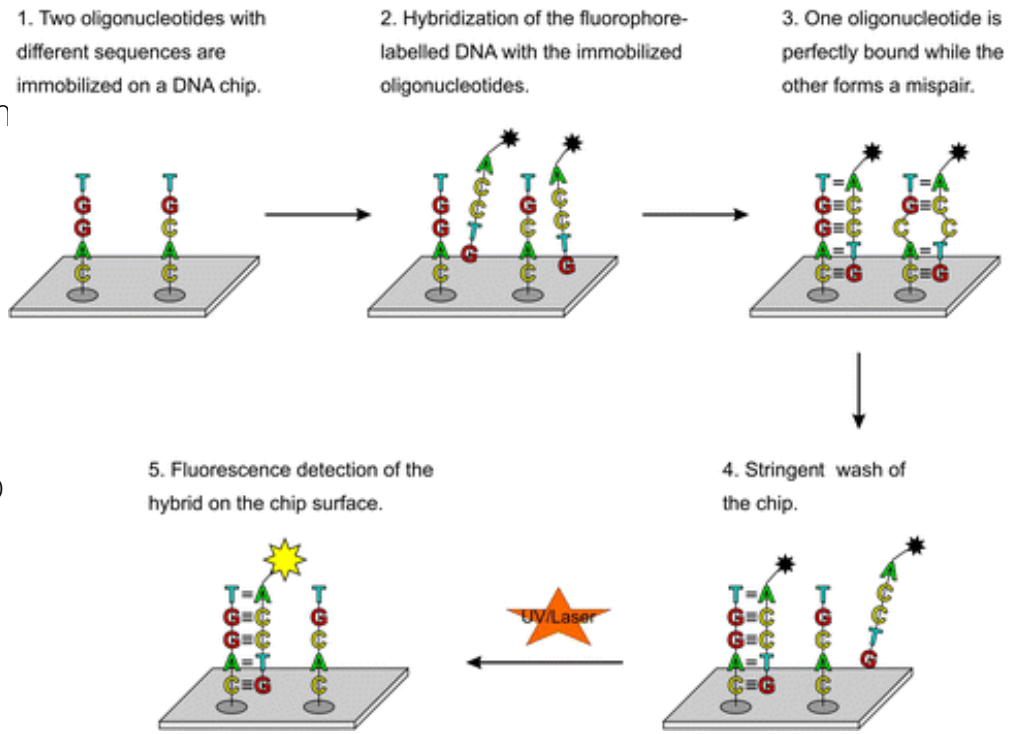
2-. Zahřátím na určitou teplotu se zruší vodíkové vazby mezi řetězci a DNA vzorku se rozplétá na dva samostatné řetězce – tento proces nazýváme **denaturace**.

3. Teplota se zase sníží a jednořetězcové molekuly se snaží znovu spárovat se svými komplementárními řetězci

4. Nastává komplementární párování n

- původním párem DNA řetězců
- DNA a sondou – vzniká **hybrid**

5. Sklíčko se nakonec omyje a zůstano pouze hybridizované řetězce.



Zdroj: http://pubs.rsc.org/en/content/articlehtml/2008/mb/b713259j

Kapitola II.1

Vznik a charakter dat

Postup mikročipového experimentu

1. Výroba mikročipového sklíčka
2. Příprava vzorků

Příprava čipu a vzorků

3. Hybridizace

---

4. Skenování

Vznik dat

5. Analýza obrazu (kvantifikace signálu, vznik expresních dat)

Vznik a charakter dat

Každá technologie má svůj vlastní způsob kvantifikace signálu (teda proměny signálu na čísla – data).

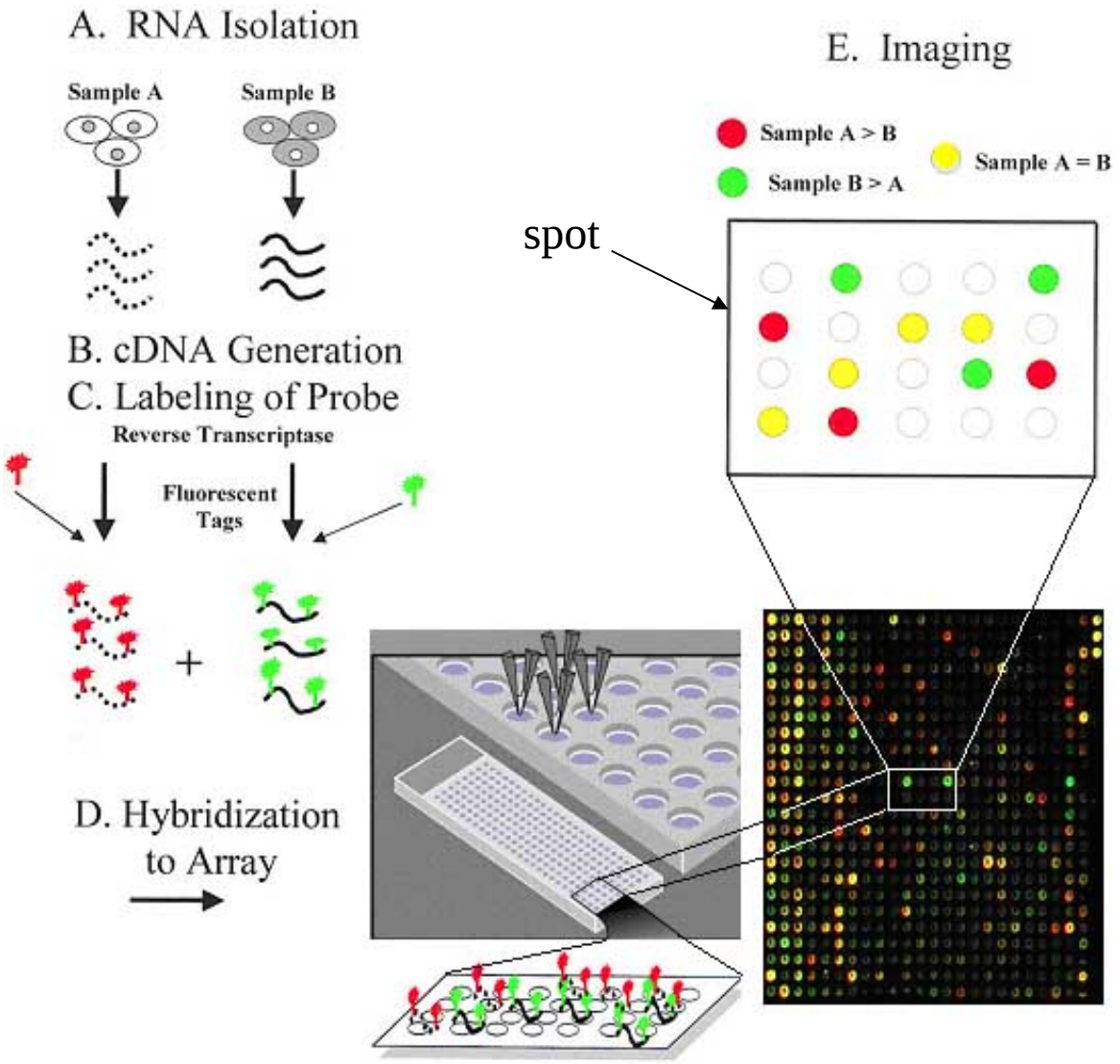
Mnohé principy jsou společné.

1. Fluorescenční signál je excitován s pomocí laseru
2. Elektrony jsou zachycené mikroskopem přes filtry do obrazu
3. Tyto obrazová data se kvantifikují

Kapitola II.1.1

Vznik a charakter dat -> cDNA mikročipy

Jak získáváme základní data z cDNA



F. Analýza obrazu (snímání intenzit jednotlivých kanálů)

Datový soubor:

tisíce řádků (genů)
X desítky sloupců

- číselné hodnoty intenzit testované a referenční RNA (+ hodnoty pozadí…)
- kontrola kvality spotů
- …

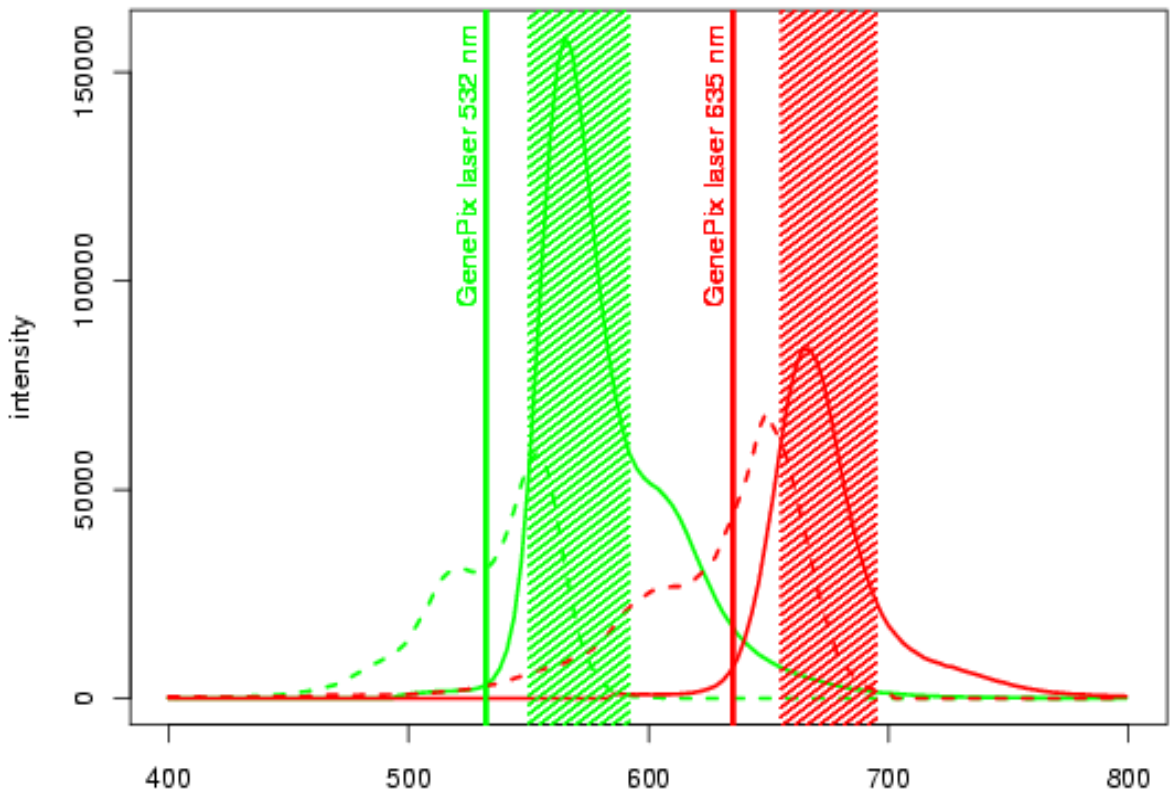
Další analýza

1. úpravy datového souboru
2. určení odlišných genů
3. klasifikace, predikce….

Dvoukanálové skenování

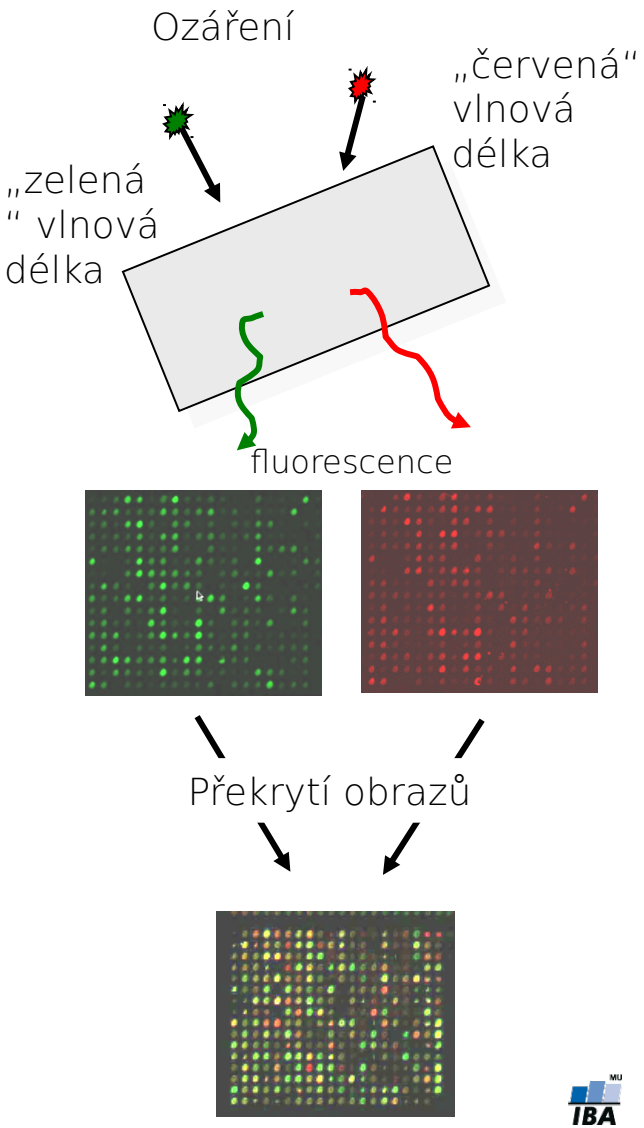
Po hybridizaci vkládáme sklíčko do skeneru abychom vytvořili obrázek mikročipu.

Excitační a emisní spektra Cy3 a Cy5



⇐ Vyšší frekvence, více energie

⇒ Nižší frekvence, méně energie



Analýza obrazu

Po skenování se uloží obrázek mikročipového sklíčka ve formátu .tiff, který se vloží do programu pro analýzu obrazu. Následuje kvantifikace signálu.

Kroky kvantifikace:

1. Lokalizace center spotů

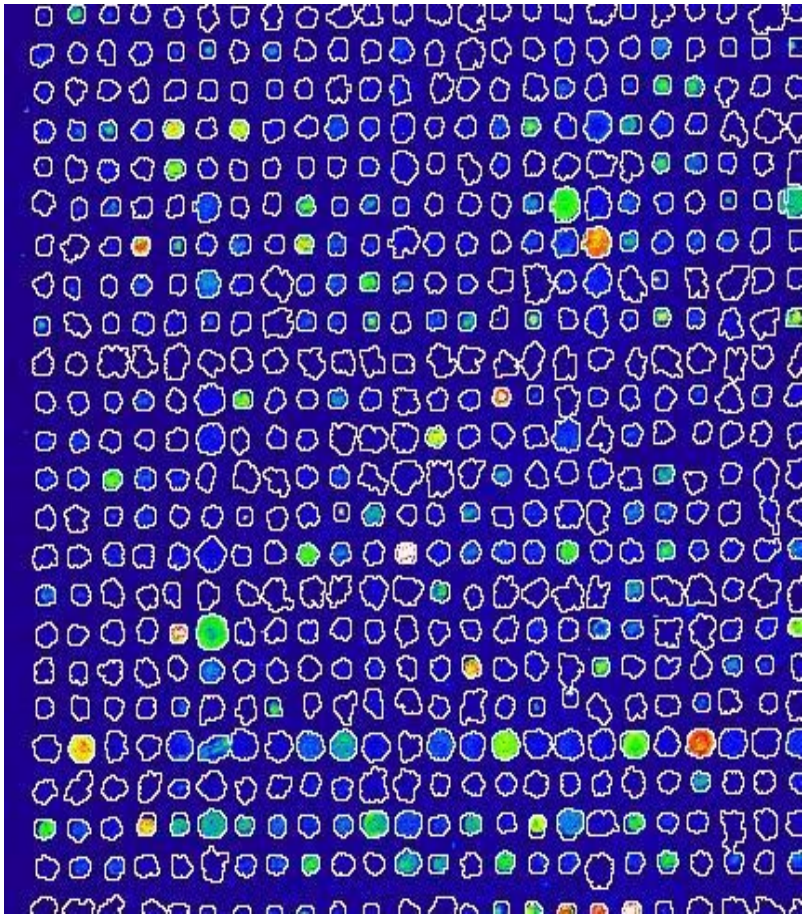
Automaticky pomocí *grid* (síťky), a manuální úpravou

2. Segmentace

Klasifikace spotů, odlišené intenzity pozadí od popředí (pomocí kruhů, etc...).

3. Kvantifikace signálu

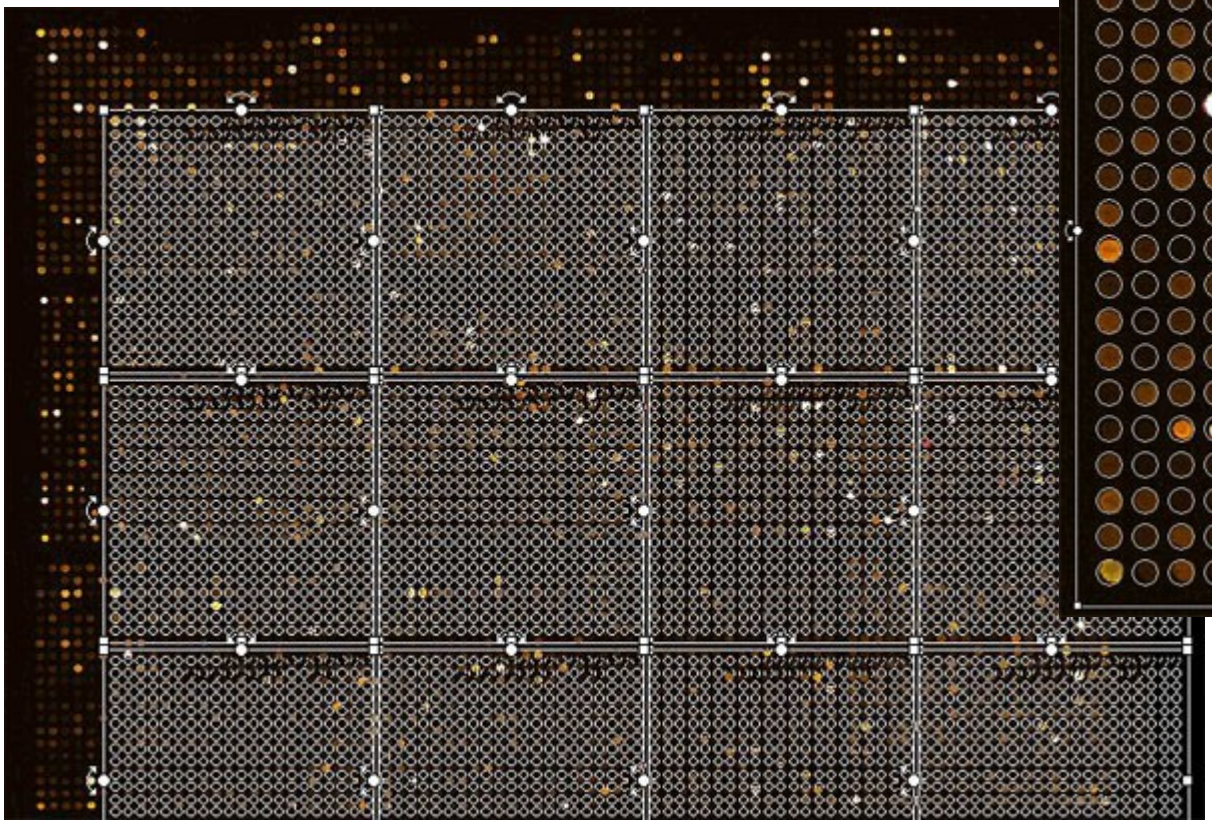
V popředí i v pozadí spotu

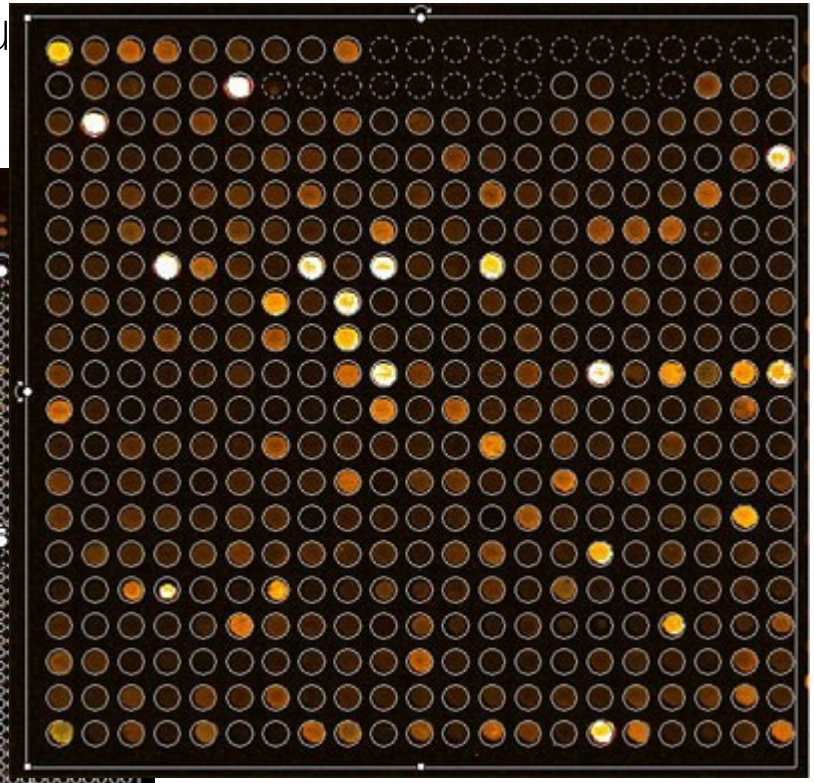


Lokalizace center spotů

Automaticky pomocí speciálního souboru *grid* (od výrobců mikročipu), který obsahuje informaci o:

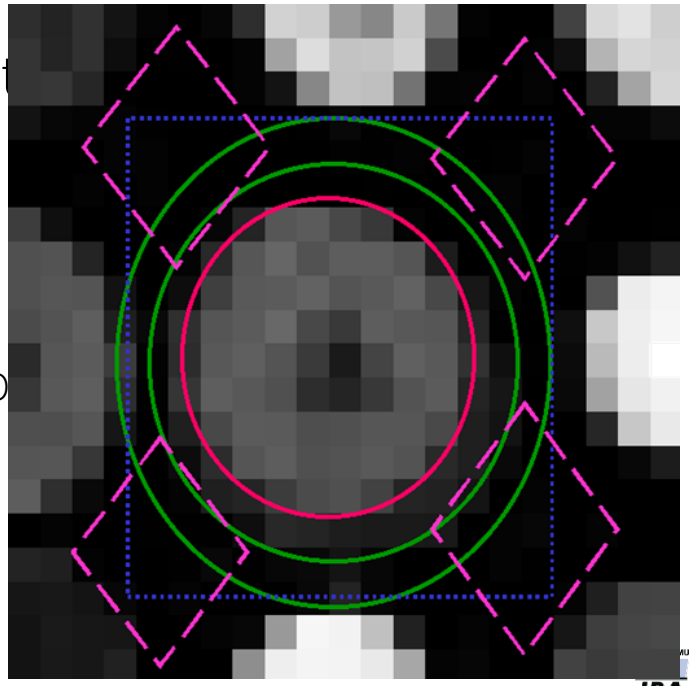
- Počtu a umístění spotů na mikročipu
- Průměru spotů v pixelech





Segmentace

- V tomto kroku jsou programem pro analýzu obrazu rozpoznávané oblasti **spotů** a **pozadí**
- Nastavení velikosti a pozice spotů – probíhá nejprve automaticky
- Obvykle nutná vizuální inspekce a další přizpůsobení ručně
- Navíc – nutné manuální označování špatných, případně prázdných spotů
- Nejčastější algoritmy vyhledávání spot
  - Fixed circles
  - Adaptive circles
  - Histogram adaptive
- Různé programy různě definují **pozadí** spo



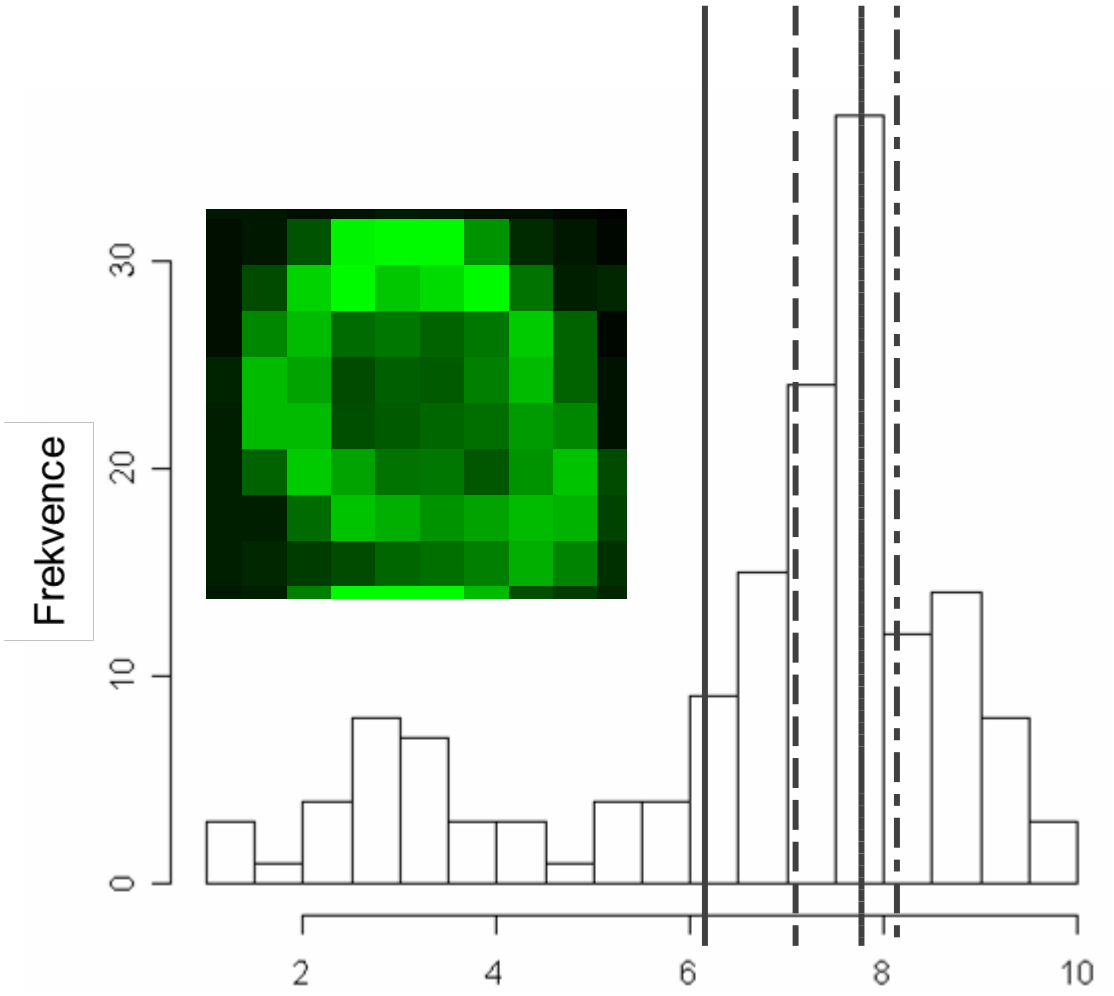
GenePix

QuantArray

ScanAlyse

Kvantifikace signálu

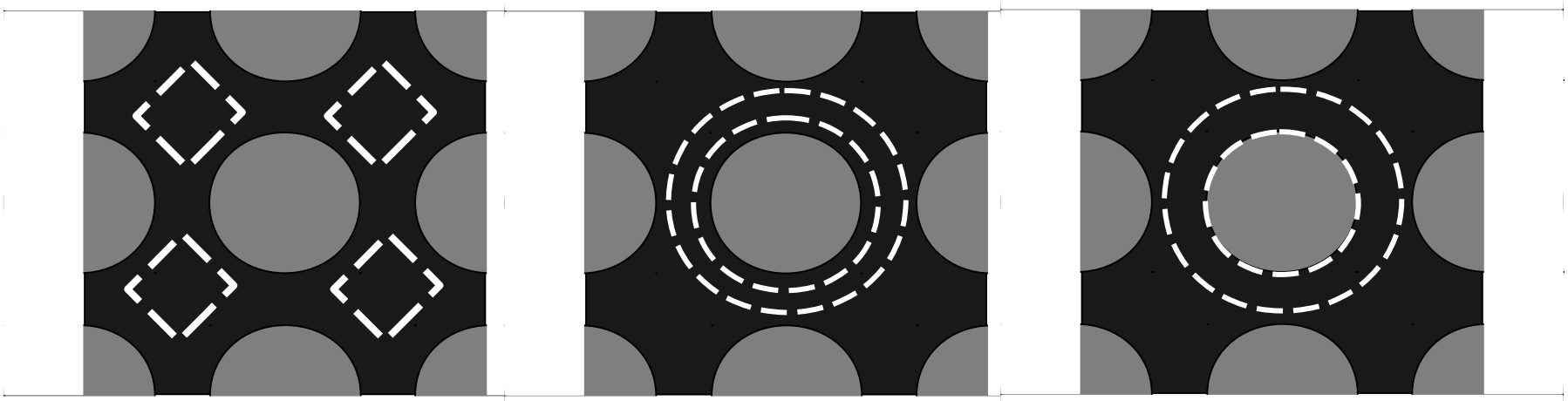
- V této fázi se kvantifikuje signál spotu, používají se různé charakteristiky (průměr, medián, modus, kvantily)



Logaritmus intensity signálu v pixelech spotu

Kvantifikace signálu pozadí

- Tři druhy metod:
  1. **Lokální metoda (local background)**
  2. Morfologické otevření (morphological opening)
  3. Konstantní/globální metoda (constant/global background)

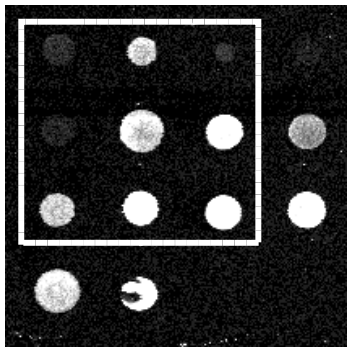


GenePix QuantArray ScanAlyse

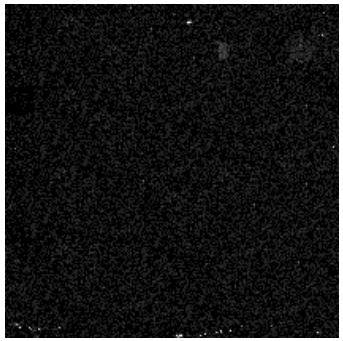
Vizualizace oblastí lokálního odhadu intenzity pozadí u tří různých programů analýzy obrazu cDNA mikročipu

Kvantifikace signálu pozadí 2.

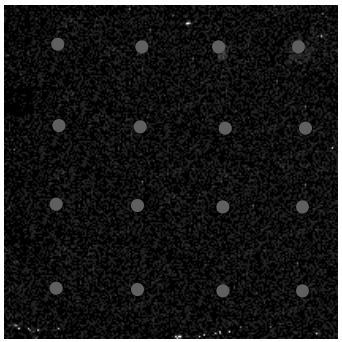
- Tři druhy metod:
  1. Lokální metoda (local background)
  2. **Morfologické otevření (morphological opening)**
  3. Konstantní/globální metoda (constant/global background)



Čtvercový element



Nový obraz
s odhadnutým
signálem pozadí



Schematické znázornění
Center spotů, ze kterých
je odhadnutý
signál pozadí pro spot

Kvantifikace signálu pozadí 3.

- Tři druhy metod:
  1. Lokální metoda (local background)
  2. Morfologické otevření (morphological opening)
  3. **Konstantní/globální metoda (constant/global background)**

Signál je odhadnutý jako jediná hodnota pro všechny spoty:

- Jako průměr intenzit signálů negativních kontrol (sondy jiného organismu, které **by neměly** hybridizovat se vzorkem)
- Nebo jako 3% kvantil rozdělení signálu všech spotů

Kontrola kvality spotů I.

- Po dobu kvantifikace intenzit probíhá ještě **inspekce kvality** spotů na základě **parametrů** zadaných do algoritmu

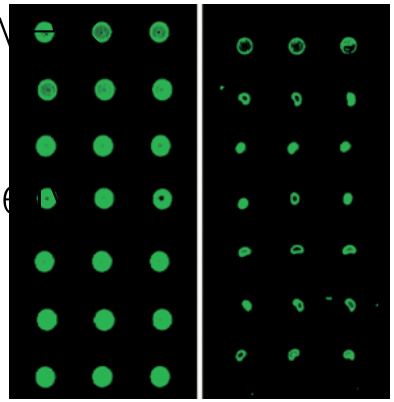
- I po kvantifikaci je možné manuálně označit spoty, které považujeme za nekvalitní

- Spotem, které neprojdou kontrolou kvality je přiřazená příslušná hodnota v proměnné Flags:
- Např.
  - 100 ~ good ;
  - -100 ~ bad ;
  - -75 ~ absent;
  - -50 ~ not found;
  - 0 ~ unflagged;

Kontrola kvality spotů II.

Charakteristiky kontroly kvality:

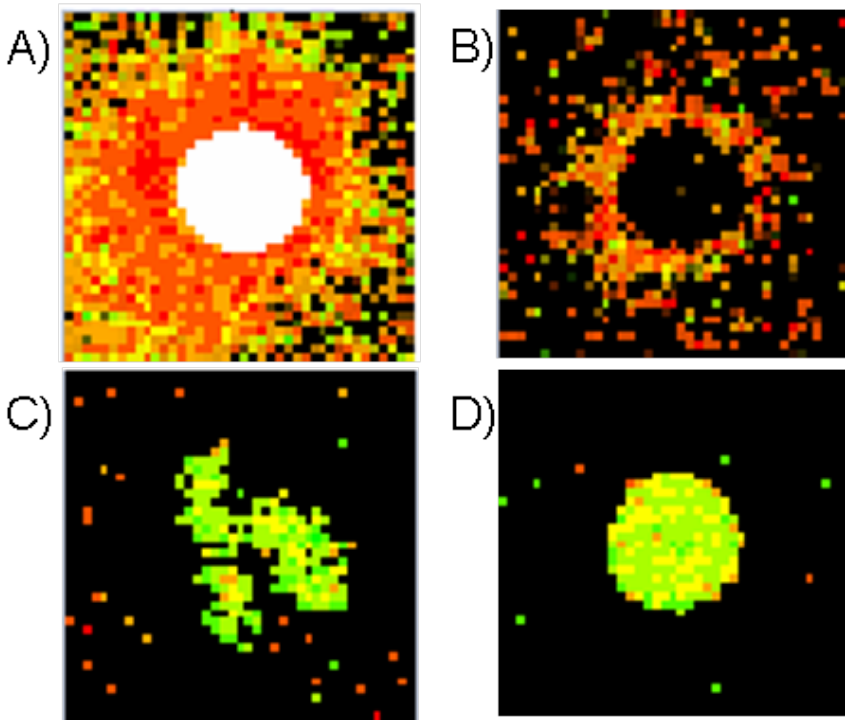
- Velikost a tvar spotu
  - Příliš malé spoty neposkytují věrohodné odhady intensity hybridizace (Simon et al., 2003) (spoty menší než < 25 pixelů by měly být odstraněné)
  - Spoty s nepravidelným tvarem, případně "koblihové spoty" by měly být označené jako nekvalitní
- Intenzita signálu
  - Spoty s příliš malou intenzitou signálu v obou kanálech
    - log2(610/590) = 0.048, ale log2(30/10) = 1.58
  - Poměr signál/šum by měl být dostatečně v
- Nasycení (saturace) spotu
  - Spoty by neměly obsahovat nasycené pixe



Source: http://probes.invitrogen.com/lit/catalog/2/images/g002230.gif

Kontrola kvality spotů III.

Příklady nekvalitních spotů (A-C) v porovnání s ideálním spotem (D)



- A) nasycený (saturovaný) spot, B) koblihový spot, C) spot s nepravidelnou strukturou, D) dobrý spot

Ukázka základních cDNA mikročipových dat

Po kvantifikaci a kontrole získáváme základní datový soubor.

- Data z jednoho cDNA mikročipovéhp sklíčka

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Unique position	ID	Chromosome	Mb positio	SES end	Plate info	Block	Column	Row	Name	X	Y	Dia.
2	44	RP11-195a8	1	37581779	37726637	NKI2C1	26	11	19	44	8600	35890	140
3	44	RP11-195a8	1	37581779	37726637	NKI2C1	26	10	19	44	8370	35890	140
4	44	RP11-195a8	1	37581779	37726637	NKI2C1	26	12	19	44	8820	35890	140
5	102	RP11-124d4	1	87374825	87558032	NKI2B12	4	7	19	102	16600	8970	120
6	102	RP11-124d4	1	87374825	87558032	NKI2B12	4	9	19	102	17060	8970	130
7	102	RP11-124d4	1	87374825	87558032	NKI2B12	4	8	19	102	16830	8970	120
8	154	RP11-145H4	1	1.52E+08	1.52E+08	NKI2G5	26	11	20	154	8600	36110	150
9	154	RP11-145H4	1	1.52E+08	1.52E+08	NKI2G5	26	13	20	154	9040	36110	140
10	154	RP11-145H4	1	1.52E+08	1.52E+08	NKI2G5	26	12	20	154	8820	36110	150
11	187	RP11-1122M4	1	1.83E+08	1.83E+08	NKI2F10	20	7	20	187	16690	27120	130
12	187	RP11-1122M4	1	1.83E+08	1.83E+08	NKI2F10	20	6	20	187	16460	27120	130
13	187	RP11-1122M4	1	1.83E+08	1.83E+08	NKI2F10	20	5	20	187	16240	27120	130
14	196	RP11-66f8	1	1.89E+08	1.9E+08	NKI2C2	18	10	19	196	8330	26880	130
15	196	RP11-66f8	1	1.89E+08	1.9E+08	NKI2C2	18	11	19	196	8560	26890	130
16	196	RP11-66f8	1	1.89E+08	1.9E+08	NKI2C2	18	12	19	196	8780	26880	130
17	236	RP11-845b6	1	2.27E+08	2.27E+08	NKI2C3	10	10	19	236	8330	17960	140
18	236	RP11-845b6	1	2.27E+08	2.27E+08	NKI2C3	10	10	19	236	8330	17960	140
19	236	RP11-845b6	1	2.27E+08	2.27E+08	NKI2C3	10	12	19	236	8780	17960	150
20	236	RP11-845b6	1	2.27E+08	2.27E+08	NKI2C3	10	12	19	236	8780	17960	150
21	236	RP11-845b6	1	2.27E+08	2.27E+08	NKI2C3	10	11	19	236	8550	17960	140
22	236	RP11-845b6	1	2.27E+08	2.27E+08	NKI2C3	10	11	19	236	8550	17960	140
23	320	RP11-1084a2	2	47485695	47697380	NKI1F10	24	7	20	320	16660	31610	130
24	320	RP11-1084a2	2	47485695	47697380	NKI1F10	24	6	20	320	16440	31610	130
25	320	RP11-1084a2	2	47485695	47697380	NKI1F10	24	5	20	320	16220	31610	130
26	323	RP11-460n15	2	47854784	48034160	NKI2H8	4	12	20	323	17720	9190	130
27	323	RP11-460n15	2	47854784	48034160	NKI2H8	4	11	20	323	17500	9190	130
28	323	RP11-460n15	2	47854784	48034160	NKI2H8	4	13	20	323	17940	9190	130
29	324	RP11-3g11	2	47946940	48102089	NKI2H7	12	11	20	324	17540	18150	130
30	324	RP11-3g11	2	47946940	48102089	NKI2H7	12	12	20	324	17760	18160	140
31	324	RP11-3g11	2	47946940	48102089	NKI2H7	12	13	20	324	17990	18160	140
32	361	RP11-232j18	2	71372264	71537932	NKI1F4	8	20	19	361	19530	13430	130
33	361	RP11-232j18	2	71372264	71537932	NKI1F4	8	1	20	361	15250	13660	130
34	361	RP11-232j18	2	71372264	71537932	NKI1F4	8	19	19	361	19290	13430	130

Nossek_results_vyradene_komplet

Podívejme se na reálná data!

V učebních materiálech k předmětu naleznete soubor **cDNApriklad.zi**

Soubor stáhneme a rozbalíme.

Struktura adresáře:

```
raw/
cDNA.R
E-GEOD-45596.idf.txt
E-GEOD-45596.sdrf.txt
SampleInfo.txt
```

Vyberte jeden ze souborů nachádzejících se v adresáři raw/ a otevřete v EXCELu

```
GSM1110303_Texas_Tech_251485034901_S01_GE2-v5_91_0806_1_1.txt
GSM1110304_Texas_Tech_251485036824_S01_GE2-v5_91_0806_1_1.txt
GSM1110305_Texas_Tech_251485034901_S01_GE2-v5_91_0806_1_2.txt
GSM1110306_Texas_Tech_251485036824_S01_GE2-v5_91_0806_1_2.txt
...
```

Základní datový soubor

Obsahuje (příklad GenePix 6.0)

- Pozice spotu
- Jméno a další identifikátory sondy na spotu
- Další charakteristiky spotu: (průměr, tvar, cirkularita, saturace, ...)
- Informace o intenzitě signálu pozadí, popředí (medián, průměr, suma, SD)
- Počet saturovaných pixelů
- Odvozené charakteristiky

  i) % pixelů signálu s intenzitami většími než 1SD (2SD) intenzity pozadí

  ii) intenzita signálu mínus intenzita pozadí

  iii) poměr mediánů/průměrů obou kanálů

  iv) logaritmus báze 2 tohoto poměru
- Informace o kvalitě spotu
- Proměnnou Flags

Základní data

- Data v základním souboru **NEJSOU** koncentrace mRNA!
- Hodnoty získané z microarray experimentu jsou pozitivně korelované s množstvím přítomné mRNA, ale navíc v sobě nesou **ŠUM**, související s:

- Kontaminací tkaniva
- RNA degradací
- Efektivitou
  - amplifikace DNA
  - reverzní transkripce
  - hybridizace a specificitou sond
- Výběrem a identifikací sond
- PCR výsledkem

- Efektivitou spotování
- Dalšími technickými vlivy při zpracování
- Segmentací obrazu
- Kvantifikací signálu
- Korekcí na pozadí

NUTNÁ KONTROLA KVALITY A ÚPRAVA DAT

Podívejme se na reálná data!

V učebních materiálech k předmětu naleznete soubor **cDNApriklad.zip**

Soubor stáhneme a rozbalíme.

Struktura adresáře:

```
raw/
cDNA.R
E-GEOD-45596.idf.txt
E-GEOD-45596.sdrf.txt
SampleInfo.txt
```

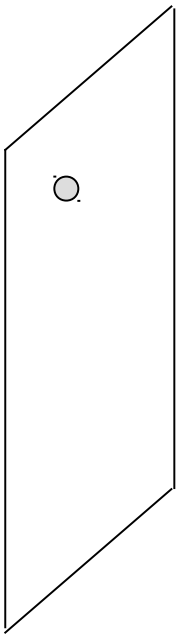
Vyberte jeden ze souborů nachádzejících se v adresáři raw/ a otevřete ho v EXCELu

```
GSM1110303_Texas_Tech_251485034901_S01_GE2-v5_91_0806_1_1.txt
GSM1110304_Texas_Tech_251485036824_S01_GE2-v5_91_0806_1_1.txt
GSM1110305_Texas_Tech_251485034901_S01_GE2-v5_91_0806_1_2.txt
GSM1110306_Texas_Tech_251485036824_S01_GE2-v5_91_0806_1_2.txt
...
```

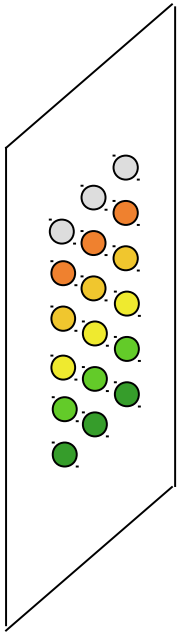
Úrovně kontroly kvality

Úroveň mikročipu
(základní datová matice)

Kvalita sondy

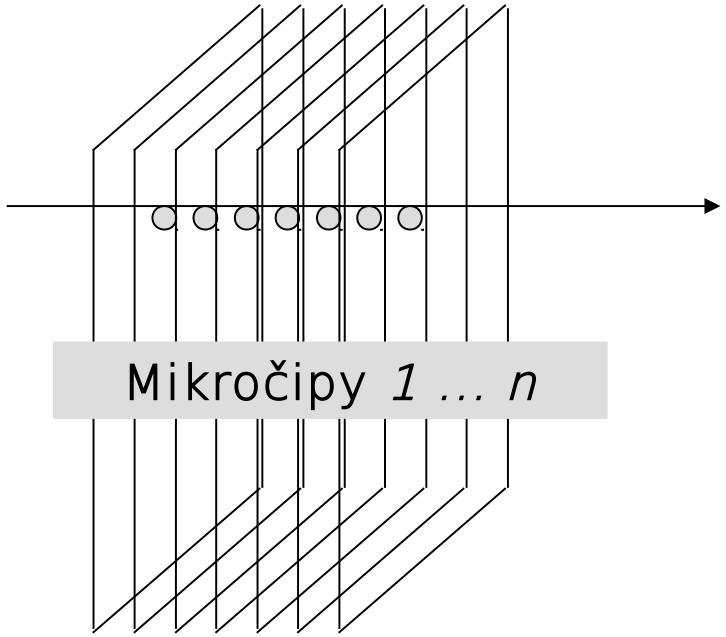


Kvalita mikročipu



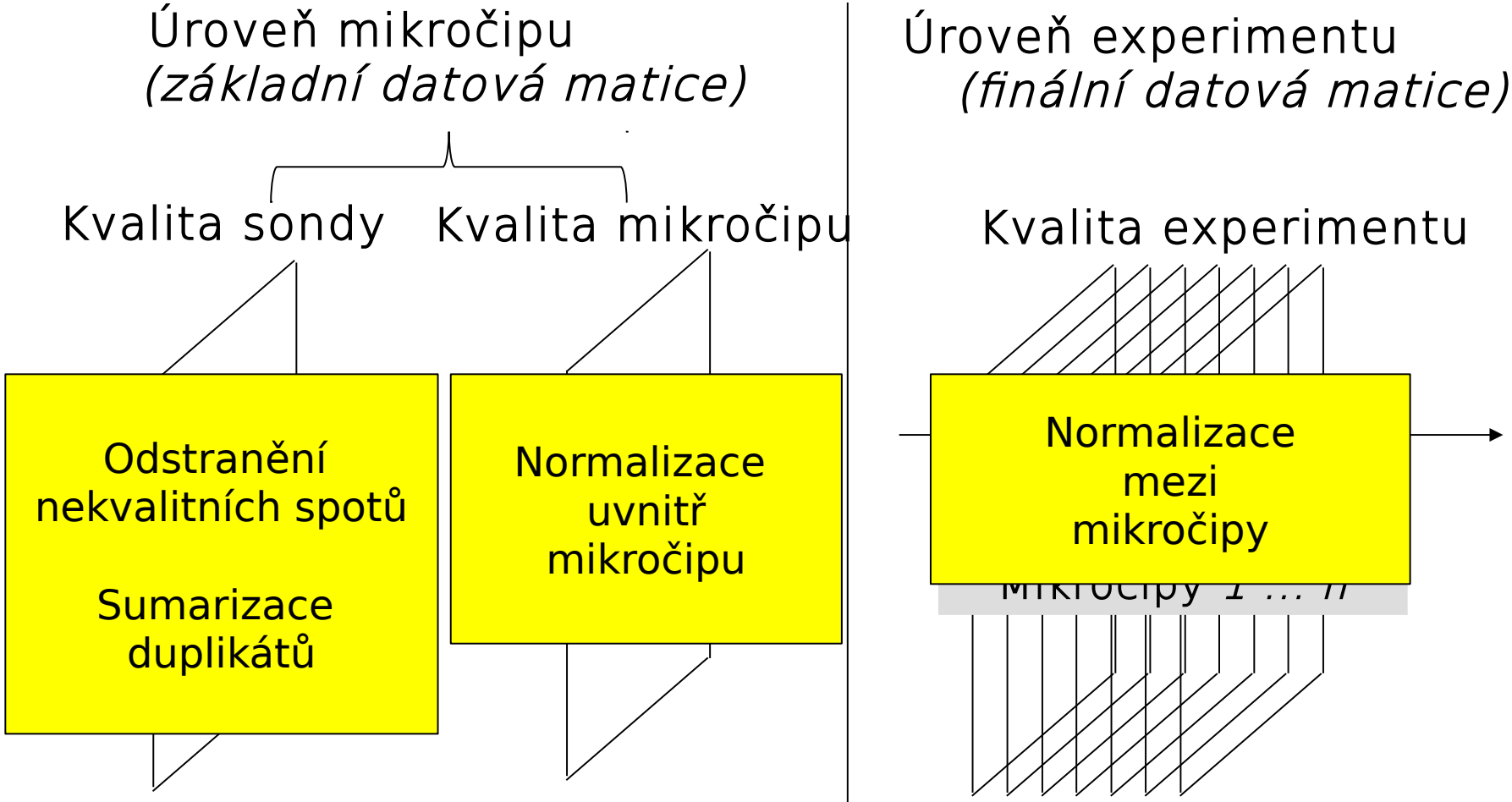
Úroveň experimentu
(finální datová matice)

Kvalita experimentu



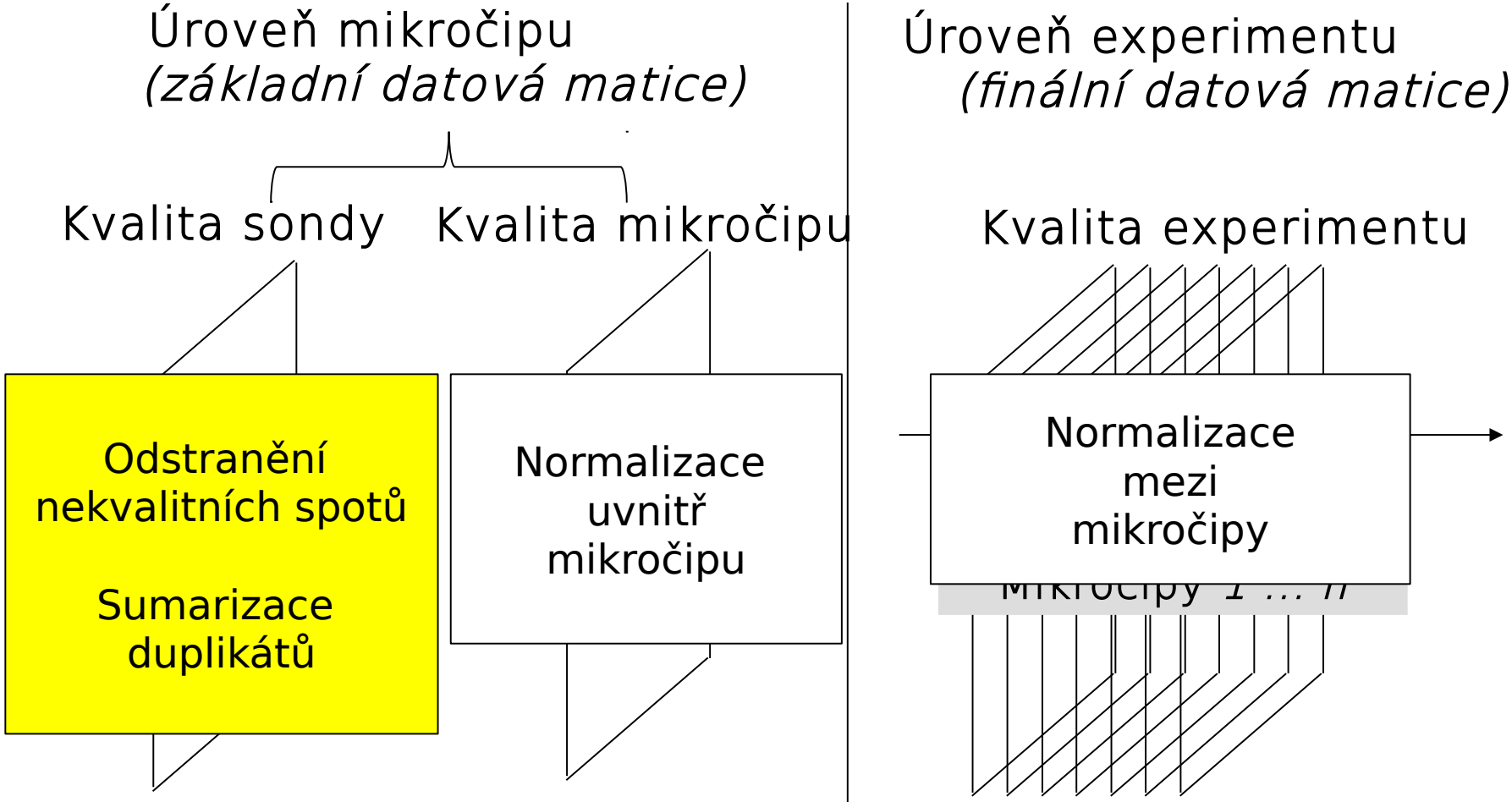
Úroveň sondy: Kvalita jednoho spotu na mikročipu
Úroveň mikročipu: Kvalita celého mikročipu
Úroveň experimentu: Kvalita měření transkriptu všech mikročipů v experimentu

Úrovně úpravy datových souborů



Úroveň sondy: Kvalita jednoho spotu na mikročipu
Úroveň mikročipu: Kvalita celého mikročipu
Úroveň experimentu: Kvalita měření transkriptu všech mikročipů v experimentu

Úrovně úpravy datových souborů



Úroveň sondy: Kvalita jednoho spotu na mikročipu
Úroveň mikročipu: Kvalita celého mikročipu
Úroveň experimentu: Kvalita měření transkriptu všech mikročipů v experimentu

Kontrola dat v rámci mikročipového sklíčka

- Replikáty sond
  - Sumární statistiky replikátů spotů (nekvalitní spoty už vyloučené)

	Replicate						
clone	1	2	3	mean	median	SD	No. of non-flagged replicates
A_23_P347643	-0.186	-0.265	-0.313	-0.254	-0.265	0.052	3
A_23_P60243	0.523	flagged	flagged	0.523	0.523	0	1
A_23_P116057	0.039	-0.978	flagged	-0.495	-0.495	0.5	2
A_23_P203743	-0.614	0.537	1.589	0.504	0.537	0.899	3

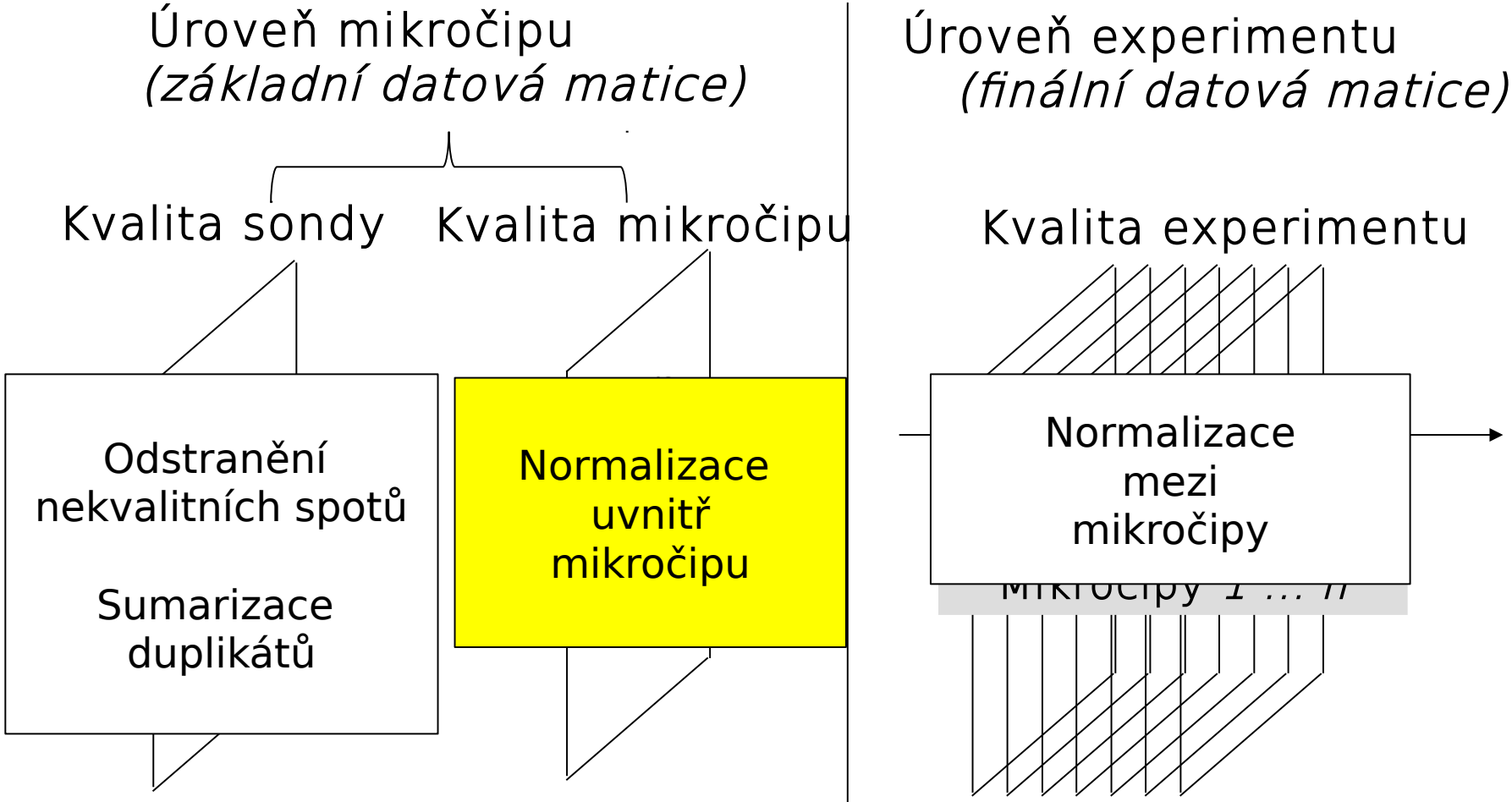
  - Buď odstranit sondy s příliš velkou variabilitou mezi replikáty…
  - …nebo si uschovat informaci o počtu validních replikátů (a vyhodit klony jen s jedním replikátem)

Kvalita mikročipového sklíčka

- Procento nekvalitních spotů nesmí být příliš velké (<25 %)

- Systematické odchylky odstraníme procesem NORMALIZACE

Úrovně úpravy datových souborů



Úroveň sondy: Kvalita jednoho spotu na mikročipu
Úroveň mikročipu: Kvalita celého mikročipu
Úroveň experimentu: Kvalita měření transkriptu všech mikročipů v experimentu

Systematické odchylky uvnitř mikročipu

- **Nerovnoměrná hybridizace** (prostorové odchylky)
  - Příčina: nerovnoměrně umytý čip, nerovnoměrně distribuovaný vzorek, print-tip efekt (defektní jehla)

- **Signál pozadí**
  - Může být velmi silný, buď špatně umytý čip, nebo špatná segmentace (část popředí je kvantifikovaná jako pozadí)

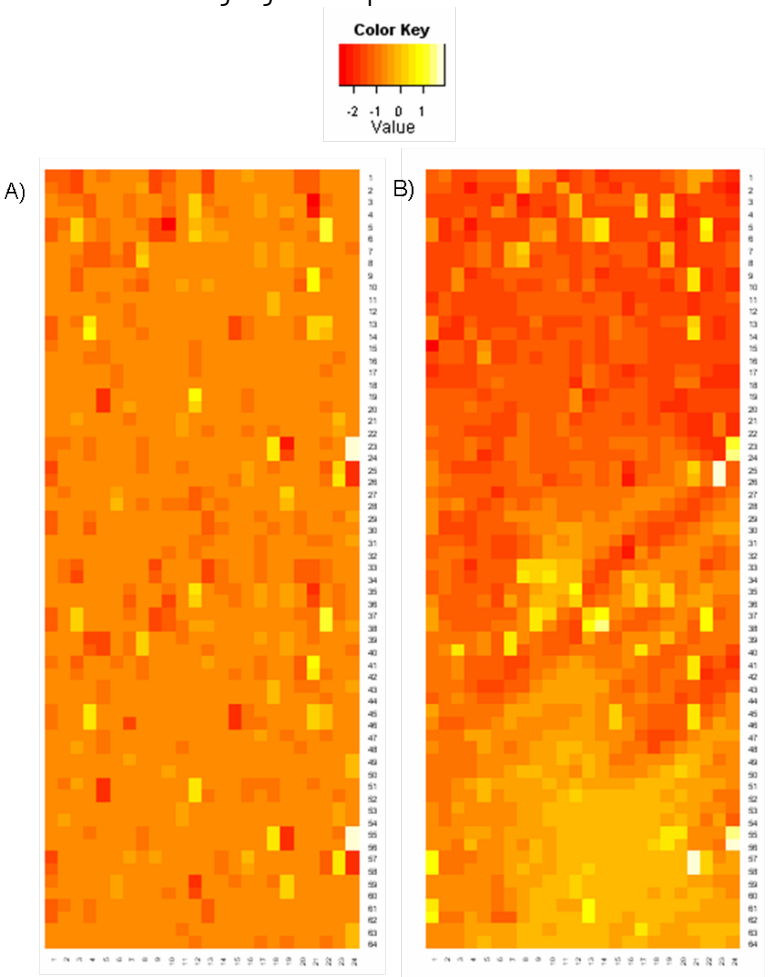
- **Efekt barviva (rozdíly intenzit mezi kanály)**
  - Příčina: odlišná schopnost inkorporace molekul barviva (Cy3, Cy5)

    odlišná reakce na excitaci (slabší intenzita UV, …)

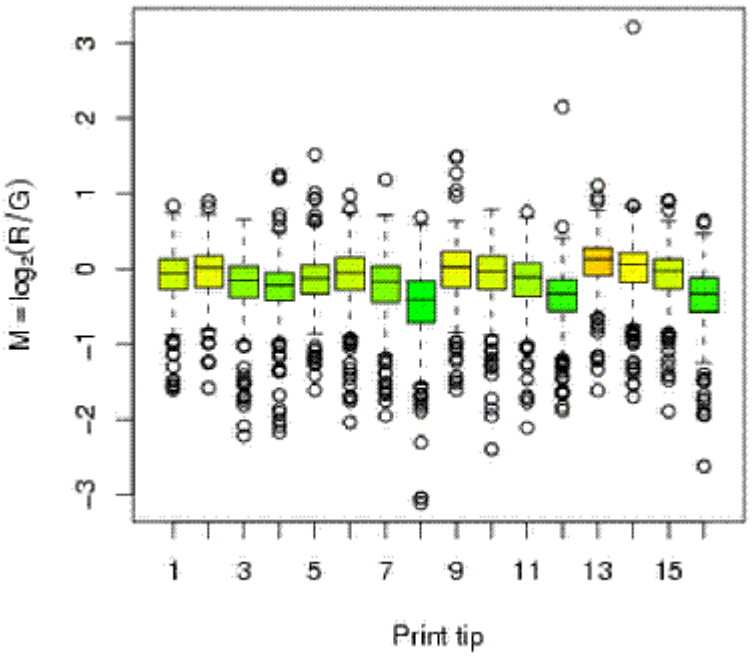
ODHALUJEME GRAFICKOU REPREZENTACÍ DAT

Diagnostika nerovnoměrné hybridizace

Virtuální rekonstrukce mikročipu, vykreslení **heatmapy log2 poměru Cy5/Cy3 intenzit** na základě jejich pozice na sklíčku



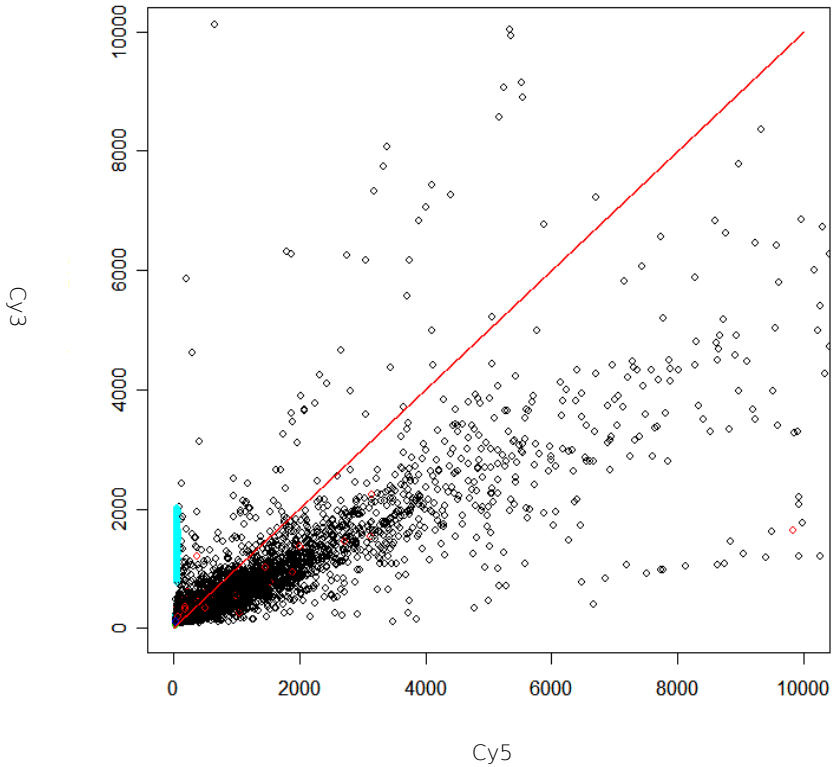
Box-ploty jednotlivých oblastí (najčastejší print-tip)



Diagnostika efektu barviva

- Často je efekt barviva větší u sond s nízkou expresí

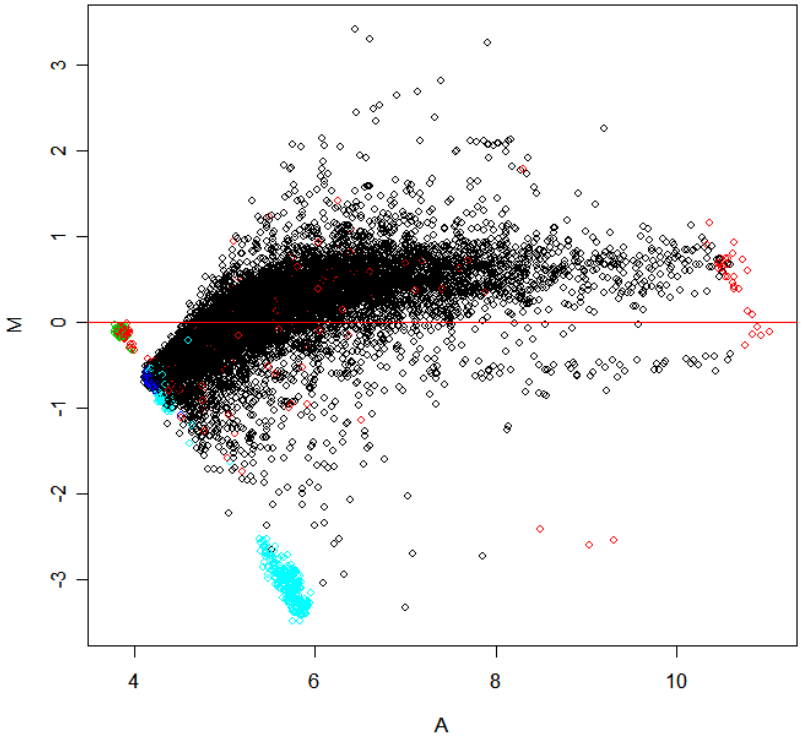
Graf intensit kanálů



Cy3 = B0 + B1*Cy5
(Cy3-B0)/B1=Cy5'

Neukáže nelineární trendy

MA graf



M = log (R/G)

A = 1/2 (log(R)+log(G))

Ukáže nelineární trendy!

Cvičení!

- Budeme pracovat v programu R-Studio
- 10 minutový krátký úvod do R – SW pro analýzu dat
- Ukážeme si jak instalovat balíky pro specifické analýzy genomických a proteomických dat
- Na příkladových datech uděláme diagnostiku kvality sklíčka

Balík **marray**

- Balík **marray** poskytuje sadu funkcí pro analýzu cDNA čipů

Instalace‘:

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("marray")
```

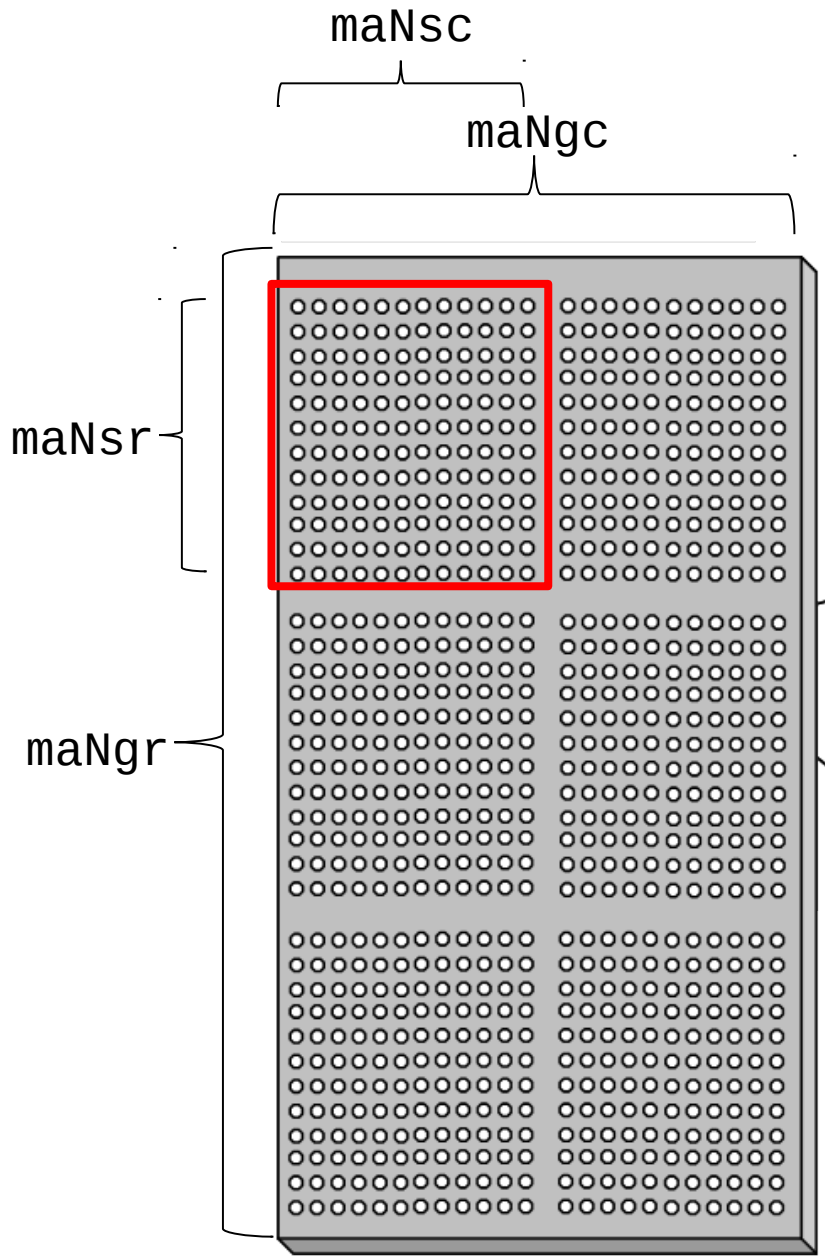
- Základní strukturou, s kterou pracuje a která obsahuje základní data všech matic experimentu je třída **marrayRaw**

```
new('marrayRaw',
   maRf = ...., # matice intensit spotů červeného kanálu
   maGf = ...., # matice intensit spotů zeleného kanálu
   maRb = ...., # matice intensit pozadí červeného kanálu
   maGb = ...., # matice intensit pozadí zeleného kanálu
   maLayout = ...., # objekt třídy marrayLayout, popis mikročipu
   maGnames = ...., # objekt třídy marrayInfo, popis sond
   maTargets = ...., # objekt třídy marrayInfo, popis vzorků
   maNotes = ...., # text - poznámky )
```

Další objekty balíku **marray**

- marrayLayout - popisuje mikročip, umístění spotů a jejich sondy

```
new('marrayLayout',
    maNgr = ... , #počet řádků matic
    maNgc = ..., #počet sloupců matic
    maNsr = ..., #počet řádků v matici
    maNsc = ..., #počet sloupců v matici
    maNspots = ..., # maNgr x maNgc x maNsr x
    maNsc
    maSub = ..., # vektor TRUE/FALSE, které
    spoty se používají
    maPlate = ..., # faktor – print tip
    maControls = ..., # faktor – status sondy
    (kontrolná nebo ne?)
    maNotes = ..., # Object of class
    character)
```



Další objekty balíku **marray**

- marrayInfo - popisuje vzorky nebo sondy

```
new('marrayInfo',
   maLabels = ...., # vektor jmen/názvů
maInfo = ...., # datová tabulka s dalšími
   charakteristikami
   maNotes = ...., # text s poznámkami
   )
```

Příklad I

- Načtěme si data **`swirl`**, které představují mikročipový experiment, porovnávající genovou expresi divokého druhu rybky Dánio pruhované a jejího mutanta v genu *BMP2*. Experiment byl proveden v *dye swap* designu, dohromady jsou k dispozici 4 mikročipy:

```
library(marray)
data(swirl)
str(swirl)
```

- Vytvořme si paletu barev a provedeme kontrolu kvality čipů

```
Gcol <- maPalette(low = "white", high = "green", k = 50)
Rcol <- maPalette(low = "white", high = "red", k = 50)
RGcol <- maPalette(low = "green", high = "red", k = 50)
```

Příklad II – kontrola prostorových efektů

- Vykreslíme si heatmapu třetího mikročipu s pomocí funkce **maImage**

```
maImage(swirl[, 3], x = "maRb") # vykreslíme pozadí červeného kanálu
maImage(swirl[, 3], x = "maGb") # vykreslíme pozadí zeleného kanálu
maImage(swirl[, 3], x = "maM") #  vykreslíme poměr intensit spotů obou kanálů (M hodnoty)
```

- Funkce **maImage** dokáže vykreslit i efekt print-tipu:

```
maImage(swirl[, 1],x="maPrintTip")
```

- Funkce **maBoxplot** vykreslí krabicové grafy

```
maBoxplot(swirl[,1])
```

Příklad III – efekt barviva

- Vykreslíme jednoduše pomocí základní funkce **plot**, a dvou funkcí, kterými z **marrayRaw** objektu extrahujeme intensity spotů červeného a zeleného kanálu:

```
R = maRf(swirl[,1])
G = maGf(swirl[,1])
plot(R,G)
abline(a=0, b=1) # vykreslíme diagonálu
```

- Funkce plot aplikována přímo na objekt třídy **marrayRaw** vykreslí MA graf, s odhadem křivek podle jednotlivých print-tipů

```
plot(swirl[,1])
```

- Jiným způsobem je prvně vypočítat hodnoty *A* a *M*, a pak je vykreslit

```
A = maA(swirl[,3])
M = maM(swirl[,3])
plot(A,M)
```

Normalizace uvnitř mikročipu I.

- Cíl: Upravit hodnoty signálu tak, abychom odstranili sytematické odchylky uvnitř mikročipu
- Princip: *Centrování* a/nebo *škálování* hodnot exprese *M*

$$M_{norm} = \frac{M - l}{s},$$

kde *l* a *s* jsou normalizační hodnoty centra (l) a škály (s)

Normalizace uvnitř mikročipu I - metody

- Typy normalizace:

  1) Logaritmická transformace – většinou používaná z důvodu transformace dat na normální rozdělení

$$M_{norm} = \log 2(M)$$

Normalizace uvnitř mikročipu I - metody

- Typy normalizace:

  1) Logaritmická transformace – většinou používaná z důvodu transformace dat na normální rozdělení

$$M_{norm} = \log 2(M)$$

  2) Korekce na pozadí

  - odstraňuje efekt pozadí
  - odlišné přístupy:

    1) odpočítá se odhadnutý signál pozadí – založené na předpokladu aditivity signálu

    Pozorovaný signál (OS) = Signál pozadí (BS) + Signál sondy (TS)

    TS = OS - BS

    - buď pro každý spot zvlášť, nebo globálně

$$M_{norm} = M - l$$

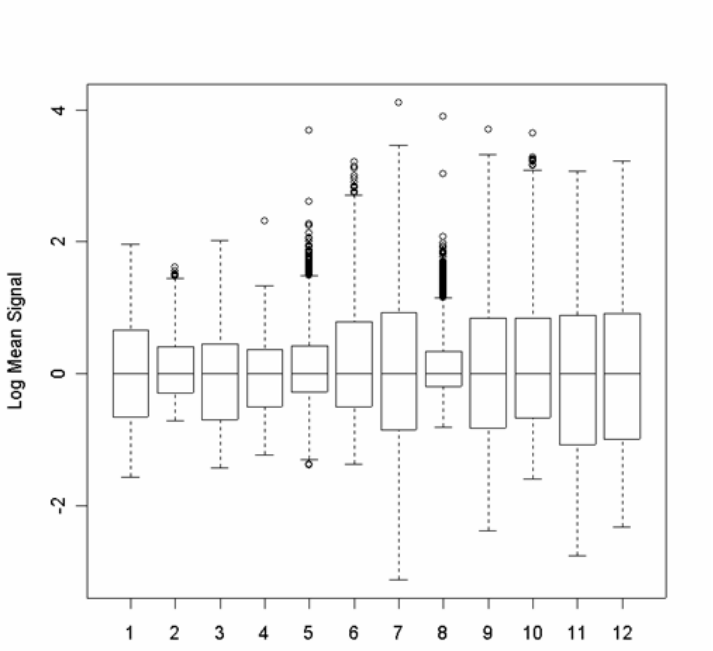
odhadnutý signál pozadí

    2) bez korekce!

Normalizace uvnitř mikročipu I - metody

3) Normalizace prostorového efektu a rozdílů intenzit mezi kanály

- Centrování mediánem
  - odčítá medián od intenzit všech spotů
  - nejjednodušší, ale není schopný zkorigovat nelinearitu



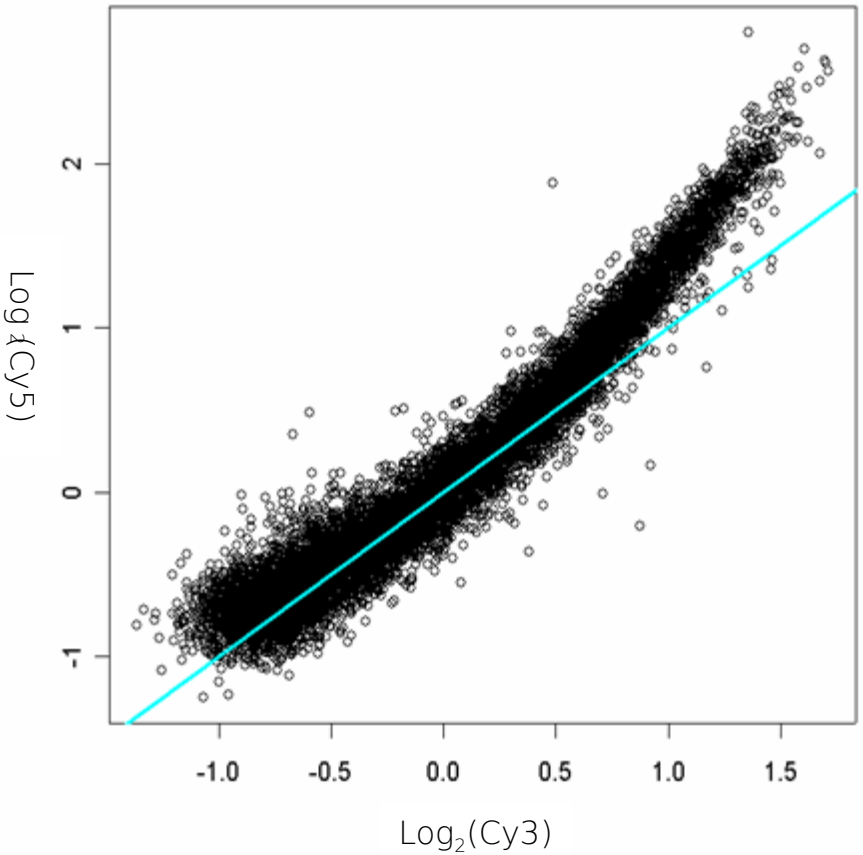
$$M_{norm} = M - l,$$

l je medián intenzit všech spotů

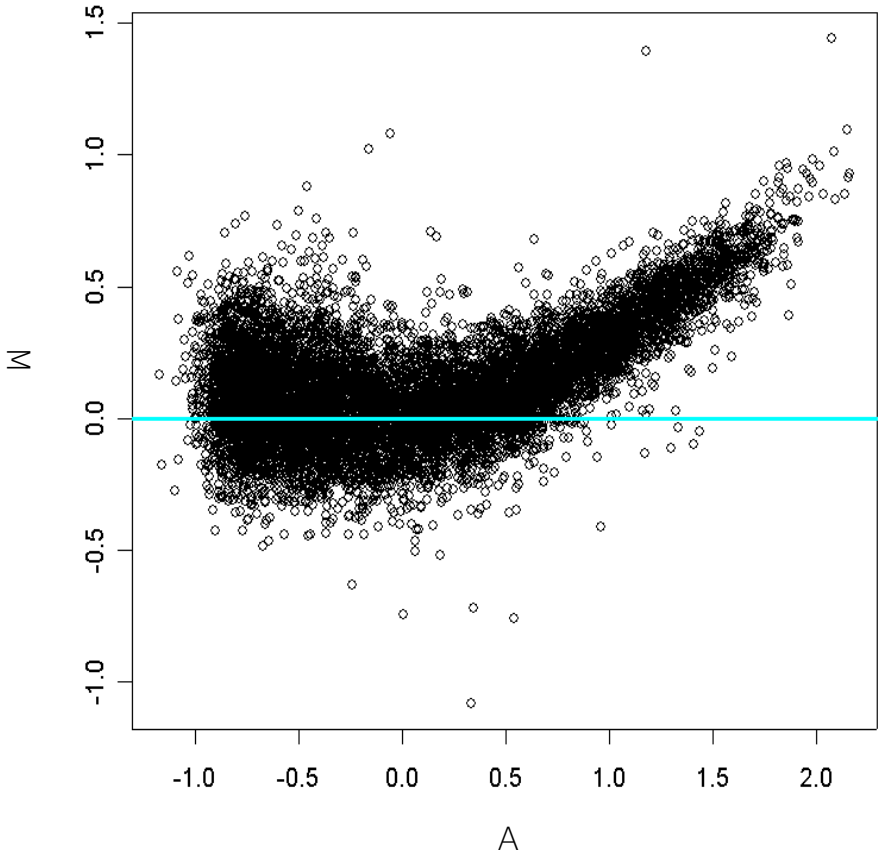
Problémy s mediánovým centrováním

Jedná sa o globální metodu, není schopná vyrovnat lokální efekty, problémy odlišných intenzit, print-tip efekty atd.

Graf intensit kanálů



MA graf



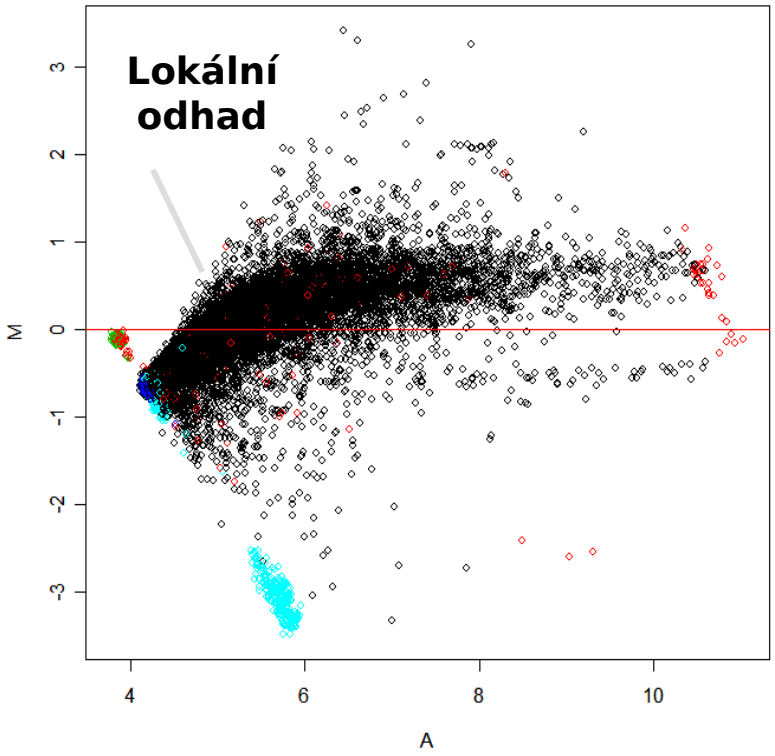
S nelinearitou si umí poradit **lokálně regresní metody (lo(w)ess**)

Lowess normalizace I

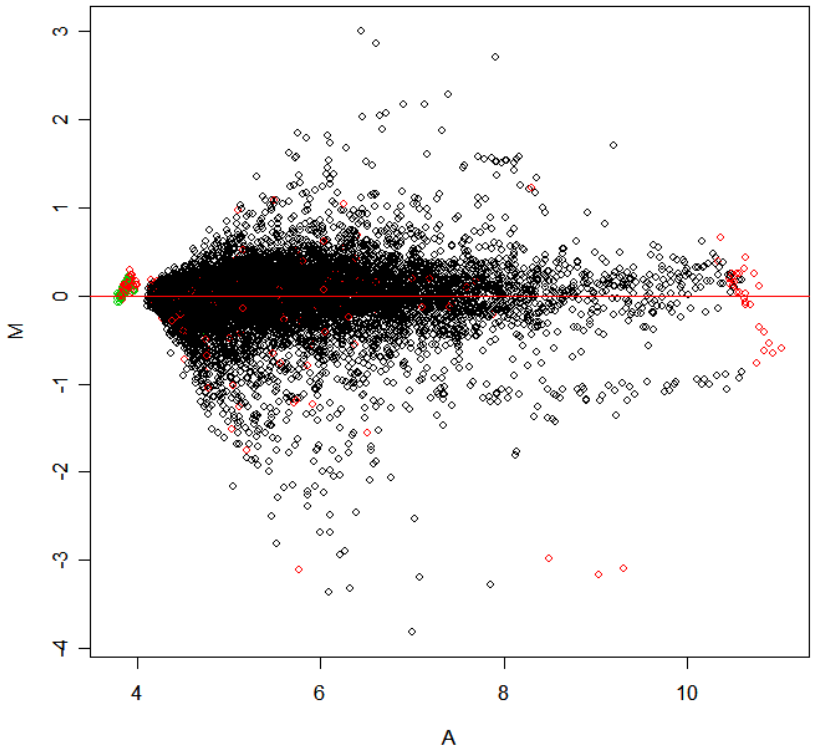
Princip:
1.Odhad křivky pomocí neparametrické lokální vážené regrese (lowess - locally weighted scatterplot smoothing)
2.Odečtení odhadnuté křivky od naměřených hodnot

Výhoda : není nutné znát funkci křivky, je odhadnuta z dat!

Před lowess normalizací



Po lowess normalizaci



Lowess normalizace II

Princip lowess

•V každém kroku se určí lokální množina dat, na které se odhadne křivka s pomocí polynomiálu a metody nejmenších čtverců

•Parameter λ určuje stupeň polynomiálu ($\lambda=0$ průměr, $\lambda=1$ lineární regrese, $\lambda=2$ kvadratická regrese)

•Množina dat na které se pracuje se určuje pomocí algoritmu nejbližšího souseda

•Vyhlazovací parameter α určuje velikost této množiny ($n\alpha$ bodů v okolí odhadovaného bodu)

•α nabývá hodnot mezi $(\lambda + 1)/n$ a 1

Normalizace uvnitř mikročipu II.

- Křivky odhadujeme:
  - na základě signálů **všech sond na mikročipu**

  Předpoklad: exprese většiny genů, které sondy představují, není změněná mezi porovnávanými skupinami! (závisí od mikročipu a od testované hypotézy)

  - na základě signálu **skupiny sond**:

i) skupina sond by měla mít přibližně stejnou expresi ve všech vzorcích (aby jsme neodstranili reálné biologické rozdíly)

ii) množina by měla být dostatečně velká, aby zachytila variabilitu sklíčka

Napr. housekeeping geny

Příklad IV – normalizace uvnitř mikročipu

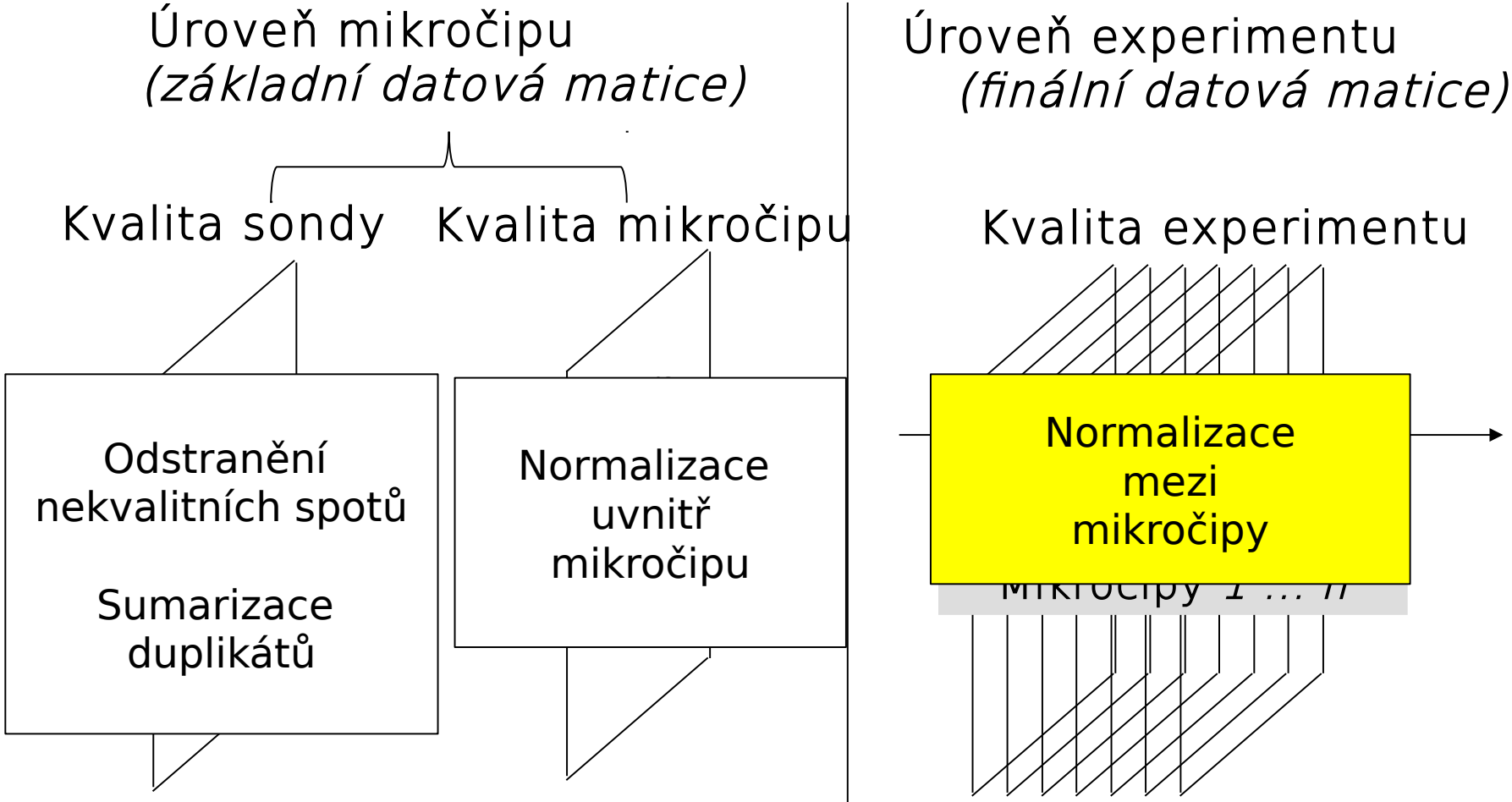
- Aplikujme centrování mediánem na *M* hodnoty prvního mikročipu z příkladu a zkontrolujme, jak se normalizace (ne)poprala s nelineárními efekty:

```
plot(swirl[,1])
swirl.norm <- maNormMain(swirl[,1], f.loc =
   list(maNormMed(x=NULL,y="maM")))
plot(swirl.norm)
```

- A teď aplikujme normalizaci pomocí loess:

```
swirl.norm.loess <- maNormMain(swirl[,1], f.loc =
   list(maNormLoess()))
plot(swirl.norm.loess)
```

Úrovně úpravy datových souborů



Úroveň sondy: Kvalita jednoho spotu na mikročipu
Úroveň mikročipu: Kvalita celého mikročipu
Úroveň experimentu: Kvalita měření transkriptu všech mikročipů v experimentu

Normalizace mezi mikročipy

- Když jsou všechny datové matice mikročipů znormalizované, tak vytváříme **finální datovou matici**, kterou použijeme pro následnou analýzu

  řádky ~ vzorky, sloupce ~ geny

- Jednotlivé soubory musíme normalizovat navzájem, abychom odstranili efekty mezi sklíčky, způsobené rozdílnou hybridizací, rozdílným množstvím vzorku (mRNA), rozdílným efektem skenování, chybami v segmentaci... apod.

- Princip – sjednocení rozložení (průměr, směrodatná odchylka, případně kvantily)

Metody normalizace mezi mikročipy

- Globální centrování

Nastaví průměr a škálu všech sklíček na jednu hodnotu (medián, průměr, ořezaný průměr… všech čipů nebo hodnoty referenčního čipu)

Nevýhoda: předpokládá, že rozdíly jsou jen posunové, lineární

- Škálování

Tato metoda sjednocuje variabilitu jednotlivých mikročipů, například podělením hodnot mediánovou absolutní odchylkou jejich intenzit. Obvykle se kombinuje s centrováním.

- Loess

Probíhá cyklickým způsobem – vždy mezi páry mikročipů až do konvergence. Také je možné vybrat množinu sond na kterých se udělá odhad loess křivky

- Kvantilová normalizace

Kvantilová normalizace

Je založena na **pořadí** pozorování, je tedy **neparametrická**. Buď na skupině všech sond, nebo jen na skupině vybraných sond.

Princip: U každého mikročipu se geny seřadí dle hodnoty exprese a tyto hodnoty se potom nahradí průměrnou hodnotou kvantilu, který představuje v celém čipu

hodnoty

Gen	čip1	čip2	čip3
A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

→

pořadí

Gen	čip1	čip2	čip3
A	iv	iii	i
B	i	ii	ii
C	ii	iii	iii
D	iii	i	iv

→

Seřazené hodnoty

	čip1	čip2	čip3
i	2	1	3
ii	3	2	4
iii	4	4	6
iv	5	4	8

Kvantilová normalizace

Je založena na **pořadí** pozorování, je tedy **neparametrická**. Buď na skupině všech sond, nebo jen na skupině vybraných sond.

Princip: U každého mikročipu se geny seřadí dle hodnoty exprese a tyto hodnoty se potom nahradí průměrnou hodnotou kvantilu, který představuje v celém čipu

hodnoty

Gen	čip1	čip2	čip3
A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

→

pořadí

Gen	čip1	čip2	čip3
A	iv	iii	i
B	i	i	ii
C	ii	iii	iii
D	iii	ii	iv

→

Seřazené hodnoty

	čip1	čip2	čip3
i	2	1	3
ii	3	2	4
iii	4	4	6
iv	5	4	8

průměr

(2+1+3)/3 = 2.00 = pořadí i
(3+2+4)/3 = 3.00 = pořadí ii
(4+4+6)/3 = 4.67 = pořadí iii
(5+4+8)/3 = 5.67 = pořadí iv

Kvantilová normalizace

Je založena na **pořadí** pozorování, je tedy **neparametrická**. Buď na skupině všech sond, nebo jen na skupině vybraných sond.

Princip: U každého mikročipu se geny seřadí dle hodnoty exprese a tyto hodnoty se potom nahradí průměrnou hodnotou kvantilu, který představuje v celém čipu

hodnoty

Gen	čip1	čip2	čip3
A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

pořadí

Gen	čip1	čip2	čip3
A	iv	iii	i
B	i	i	ii
C	ii	iii	iii
D	iii	ii	iv

Seřazené hodnoty

	čip1	čip2	čip3
i	2	1	3
ii	3	2	4
iii	4	4	6
iv	5	4	8

průměr

(2+1+3)/3 = 2.00 = pořadí i
(3+2+4)/3 = 3.00 = pořadí ii
(4+4+6)/3 = 4.67 = pořadí iii
(5+4+8)/3 = 5.67 = pořadí iv

normalizované hodnoty

Gen	čip1	čip2	čip3
A	5.67	4.67	2.00
B	2.00	2.00	3.00
C	3.00	4.67	4.67
D	4.67	3.00	5.67

Příklad V – normalizace mezi čipy

- Provedeme normalizaci pomocí loess a následně škálovou normalizaci mezi čipy a znovu vykreslíme krabicové grafy.

```
swirl.norm <- maNormMain(swirl)
swirl.norm.scale = maNormScale(swirl.norm)
maBoxplot(swirl.norm.scale)
```

Shrnutí

- Základní data nejsou mRNA koncentrace
- Musíme zkontrolovat kvalitu dat na různých úrovních
  - Úroveň sondy
  - Úroveň sklíčka (všechny sondy na sklíčku)
  - Úroveň genu (gen mezi sklíčky)
- Data vždy transformujeme *logaritmem*, abychom zabezpečili normální rozložení hodnot
- Data normalizujeme aby jsme odstranili systematické (technické) chyby

Příklad

- Podívame se do našeho adresáře s cDNA příkladem a otevřeme cDNA.R v programu Rstudio.

- Postupujeme dle instrukcí, na konci je dobrovolný úkol.

Do konce hodiny máte čas na práci na projektu