MUNI | RECETOX
SCI

Analýza genomických a proteomických dat

Mgr. Eva Budinská, Ph.D.

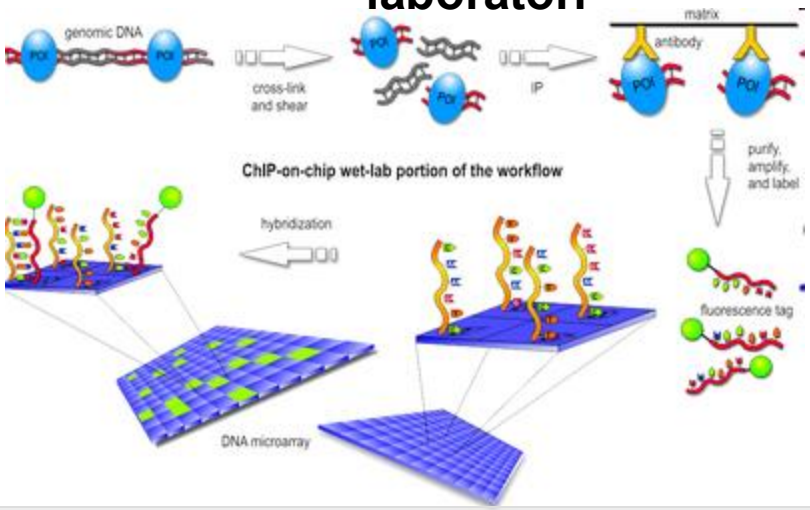
Jaro 2023

Technologie studující genomiku a proteomiku

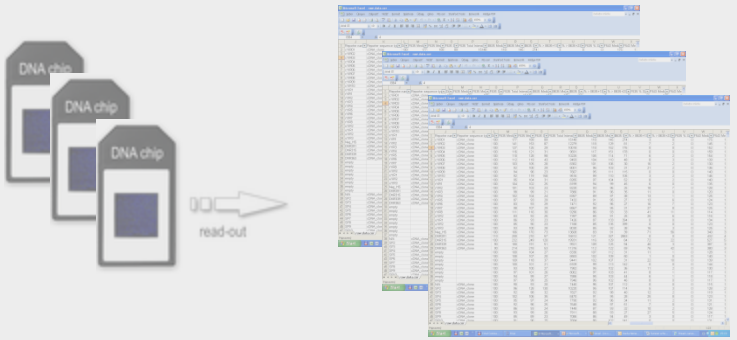
Mikročipy (microarrays)

Průběh genomického experimentu

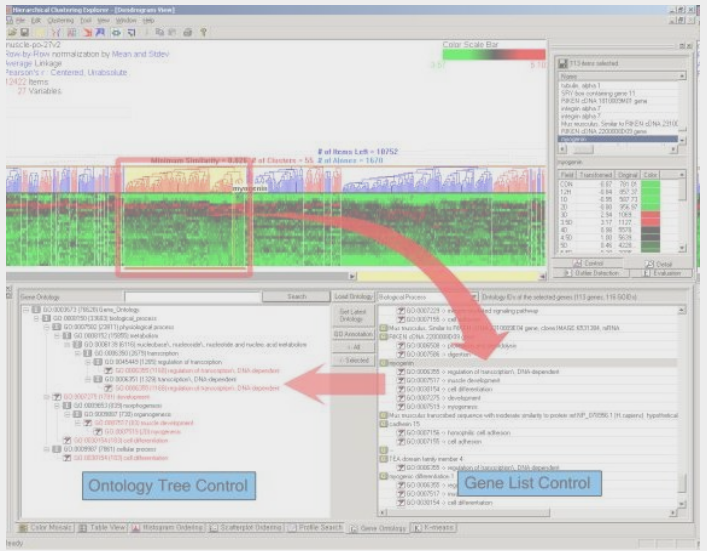
1. Příprava a provedení experimentu v laboratoři



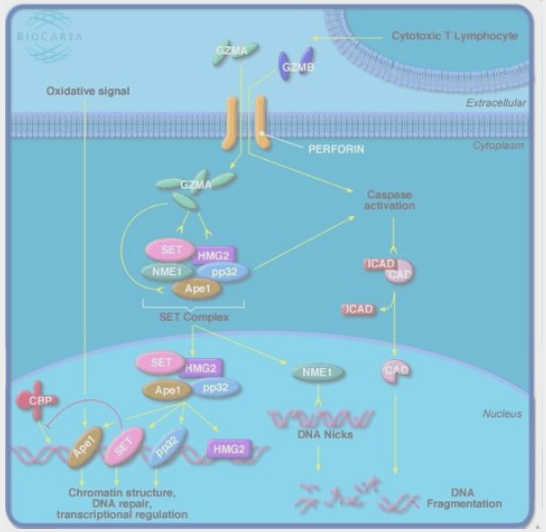
2. Extrakce a úprava dat



3. Statistická analýza dát



4. Biologická a klinická interpretace

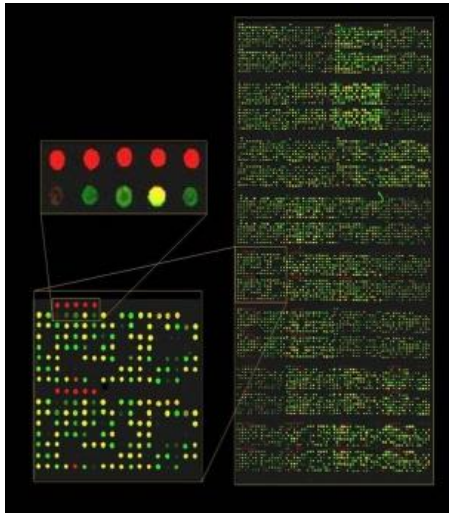


Technologie mikročipů

- **Mikročipy**– biotechnologie simultánně srovnávající biologické objekty (molekuly, tkáně) 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
  - Proteinové mikročipy



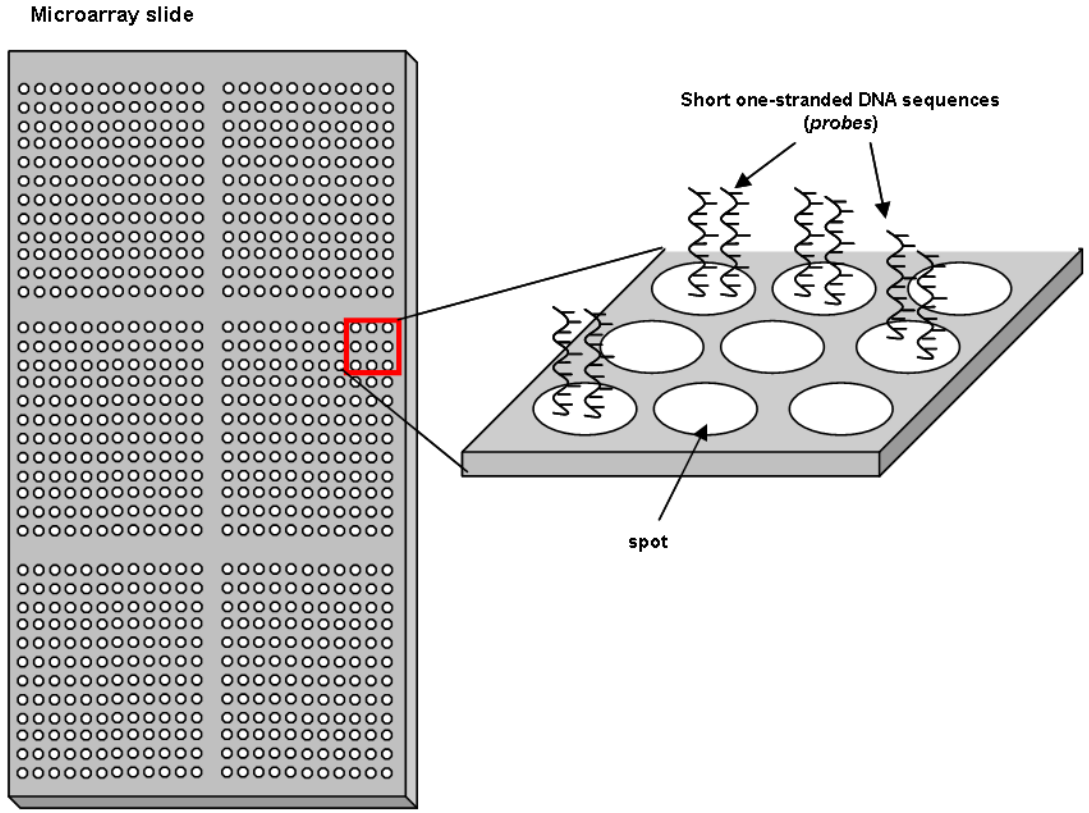
DNA mikročipy

- Serie **krátkých DNA sekvencí** imobilizovaných rovnoměrně na podklad, používaná k detekci DNA nebo RNA (obvykle jako cDNA) ve vzorcích.

- **Využití**
  - Měření změn v hladinách genové exprese (gene expression profiling, detekcia RNA - cDNA) - **expresní mikročipy**
  - detekce strukturních změn v genomu (SNPs- jednonukleotidové polymorfismy nebo změny v počtu kopií genů) – **arrayCGH, SNP arrays**
  - detekci vazebních míst proteinů na genomu (**ChIP-on-chip**)
  - detekci alternativního sestřihu (**exon junction arrays**)
  - přesná detekce neznámých a nepredikovaných transkriptů (**tiling arrays**)

Sonda (probe)

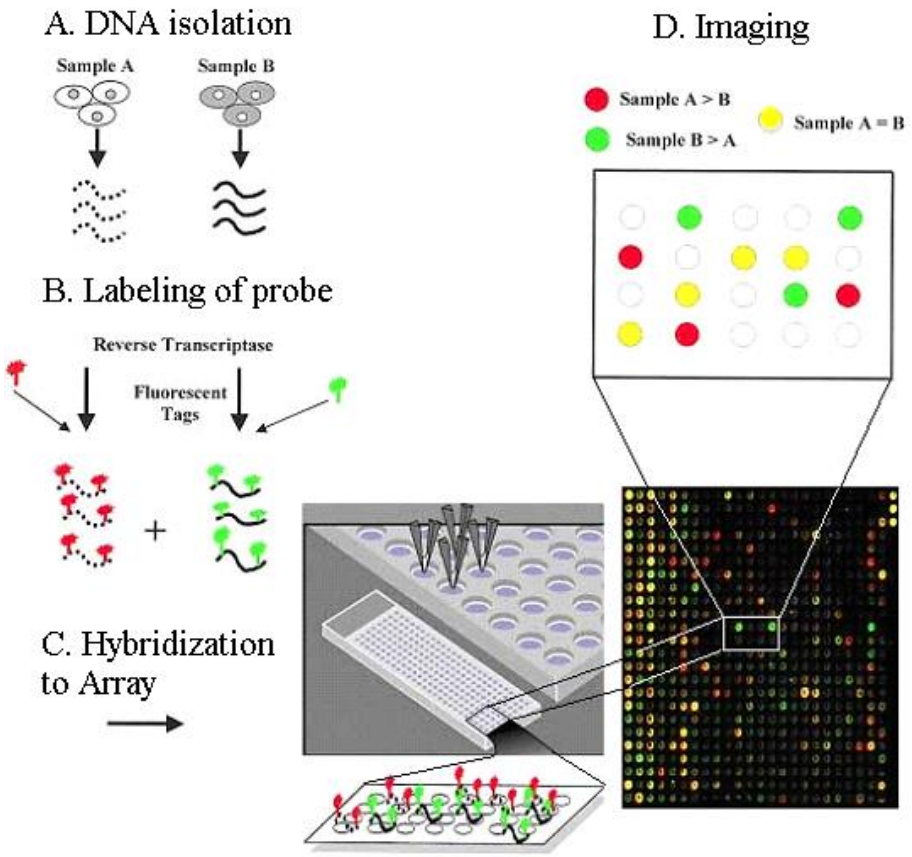
- **Krátké sekvence DNA (oligonukleotidy)** na mikročipu se nazývají ***sondy***, anglicky *probes*
- Každá oblast DNA (obvykle gen), **kterou chceme zkoumat**
- Sondy jsou navrženy tak, aby byly pro daný gen / oblast co nejspecifičtější



Základní princip

1. Fragmenty DNA / cDNA ze vzorku se s**párují s komplementárními sondami** na mikročipu a tím se mobilizují.

2. Imobilizované molekuly DNA, které byly dříve označeny **fluorescenčním barvivem** se pak dají detekovat pomocí **UV skeneru** a **kvantifikovat** tak množství mRNA / DNA s danou sekvencí přítomné ve vzorku.



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)

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)

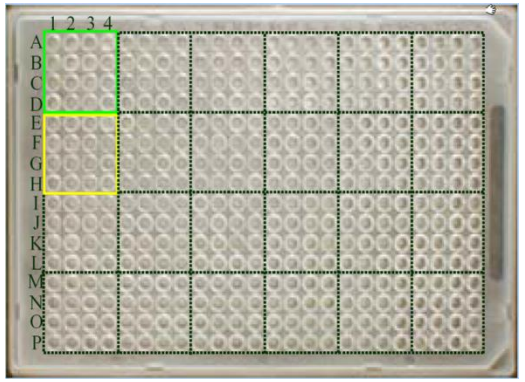
Princip výroby DNA mikročipu

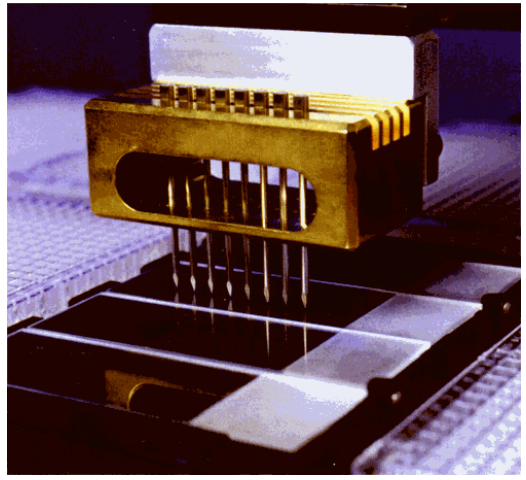
- Výroba sklíčka spočívá v připojení sond na podložné sklíčko do oblastí spotů

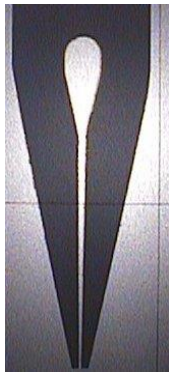
- Dvě hlavní metody:

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

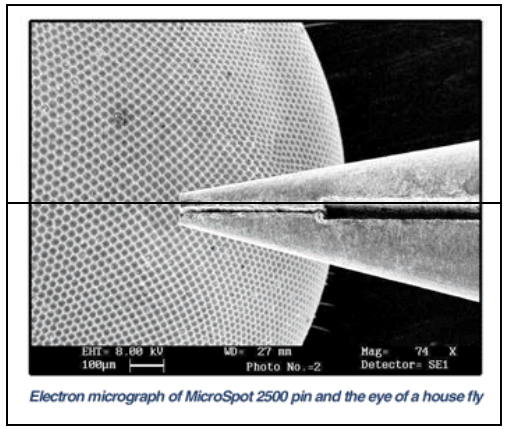
Spotovací robot







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:

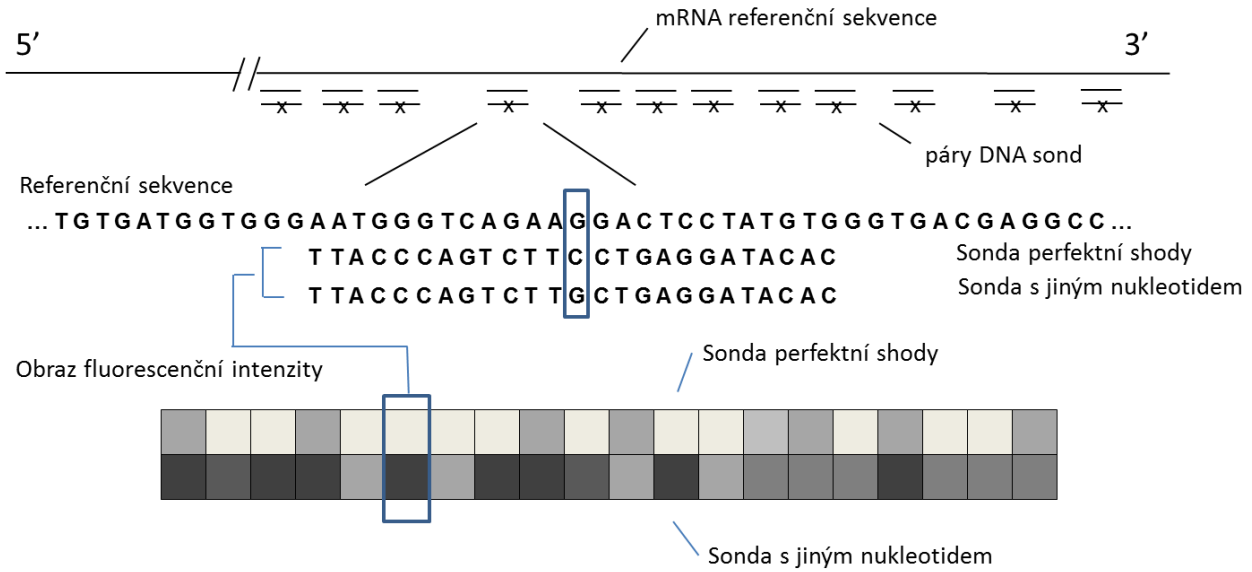
  - *Spotting* – sondy jsou syntetizované PŘED umístěním na microarray sklíčko, potom umístěné na sklíčko pomocí speciálního robota

  - *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 krátké 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

---

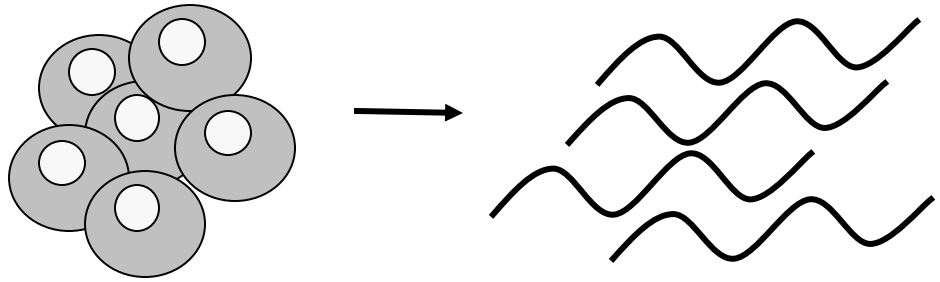
4. Skenování

Vznik dat

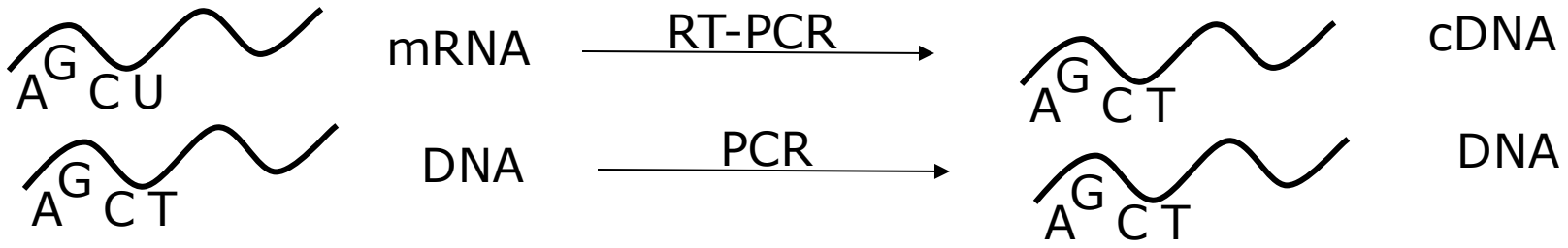
5. Analýza obrazu (kvantifikace signálu, vznik expresních dat)

Příprava vzorků

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

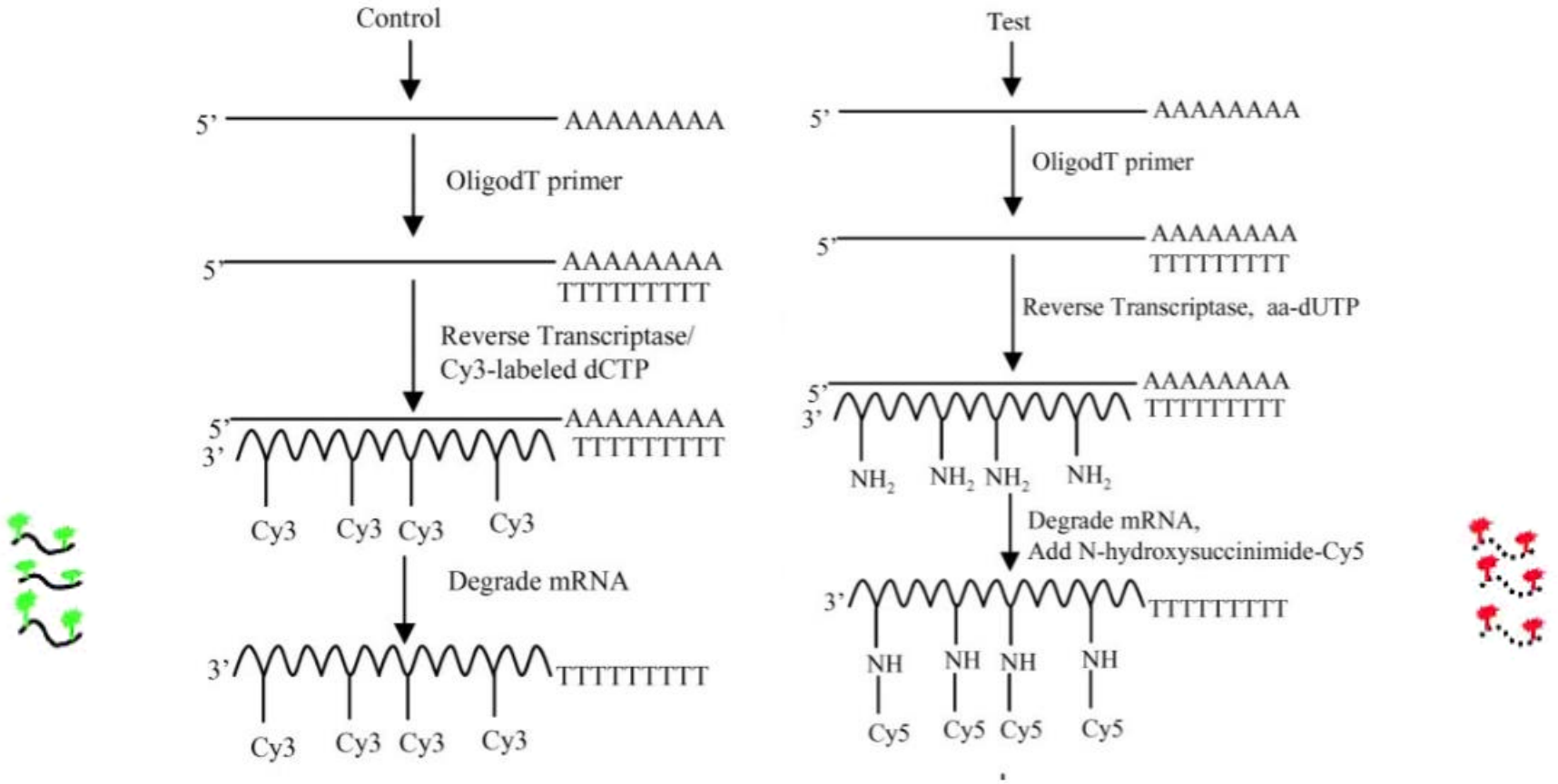


2. Přepis a amplifikace: mRNA se přepisuje do cDNA a amplifikuje se pomocí RT-PCR. DNA zas pomocí PCR.



Příprava vzorků

3. Značení: 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

Příprava čipu a vzorků

3. Hybridizace

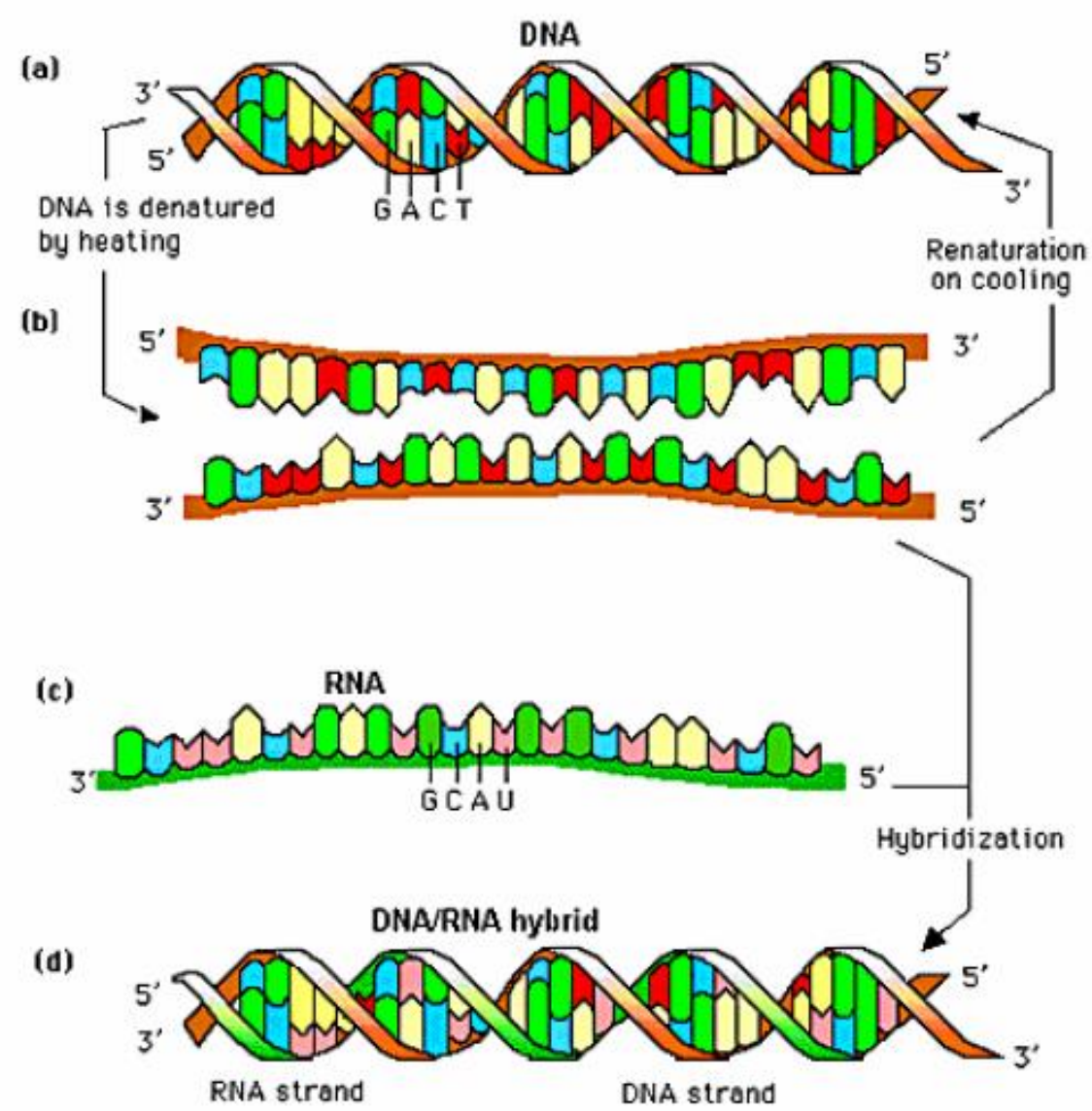
---

4. Skenování

Vznik dat

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

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 nanese na mikročipové sklíčko, kde už jsou předem navázané jednořetězcové sondy.

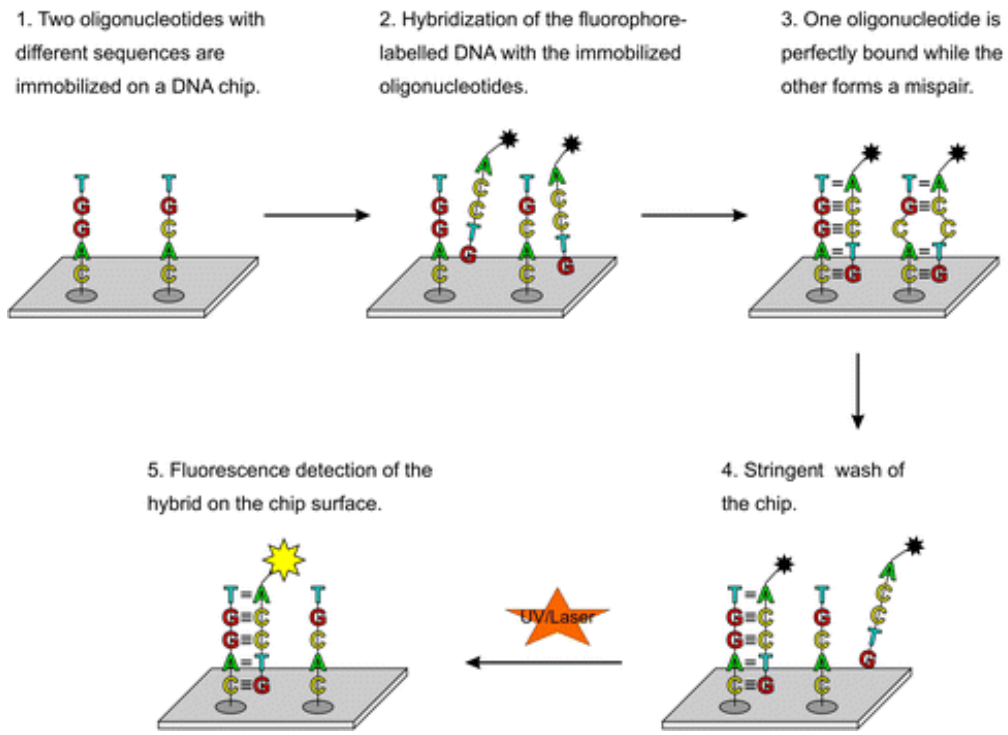
2-. Zahřátím na určitou teplotu se zruší vodíkové vazby mezi řetězci a DNA vzorku se rozplétá na dva samostatné řetězce – tento proces nazýváme **denaturace**.

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

4. Nastává komplementární párování mezi:

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

5. Sklíčko se nakonec omyje a zůstanou



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

Vznik a vlastnosti mikročipových 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

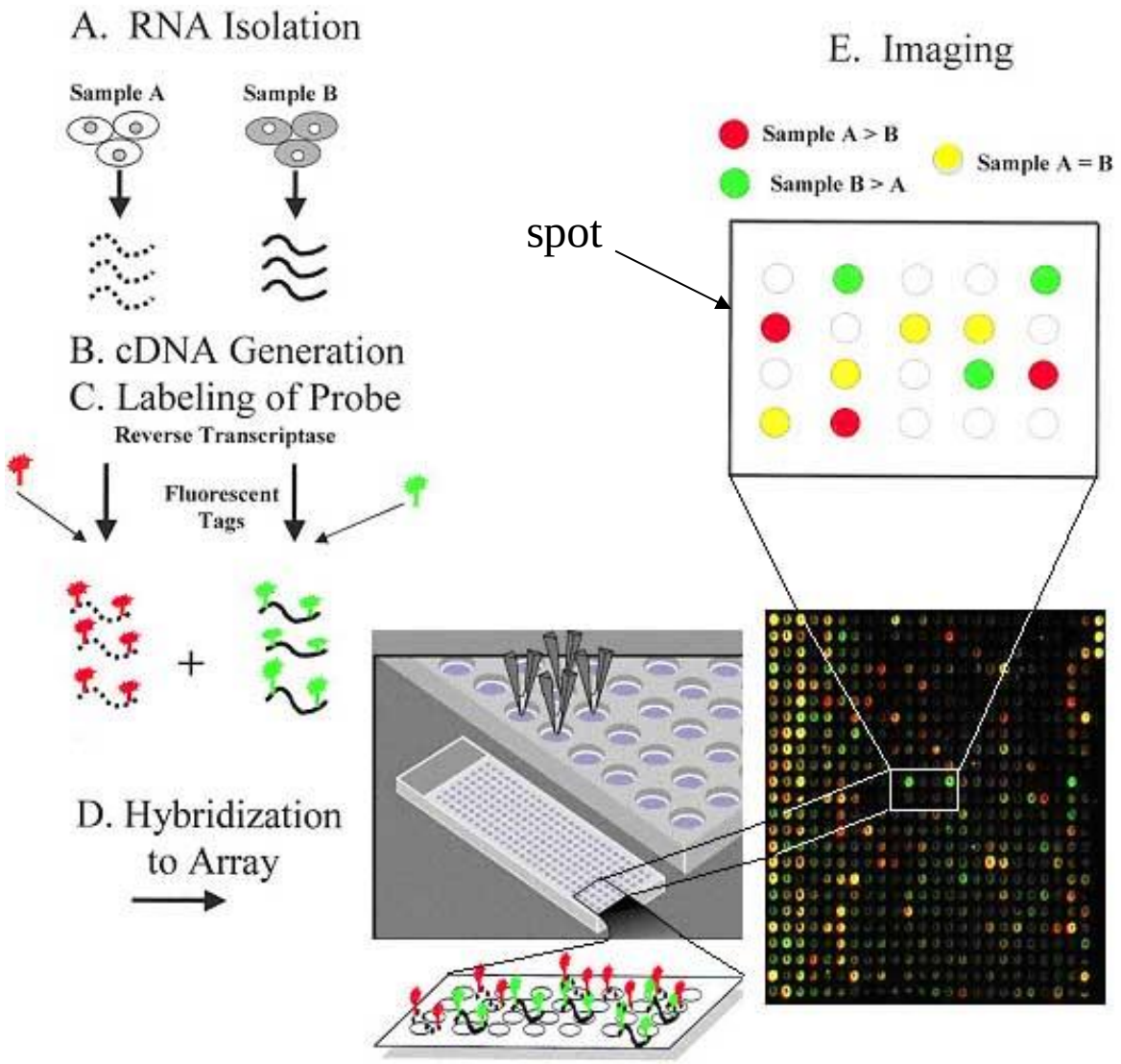
Každá technologie má svůj vlastní způsob **kvantifikace signálu** (teda proměny signálu na čísla – data).

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

Vznik a vlastnosti mikročipových 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
- ...

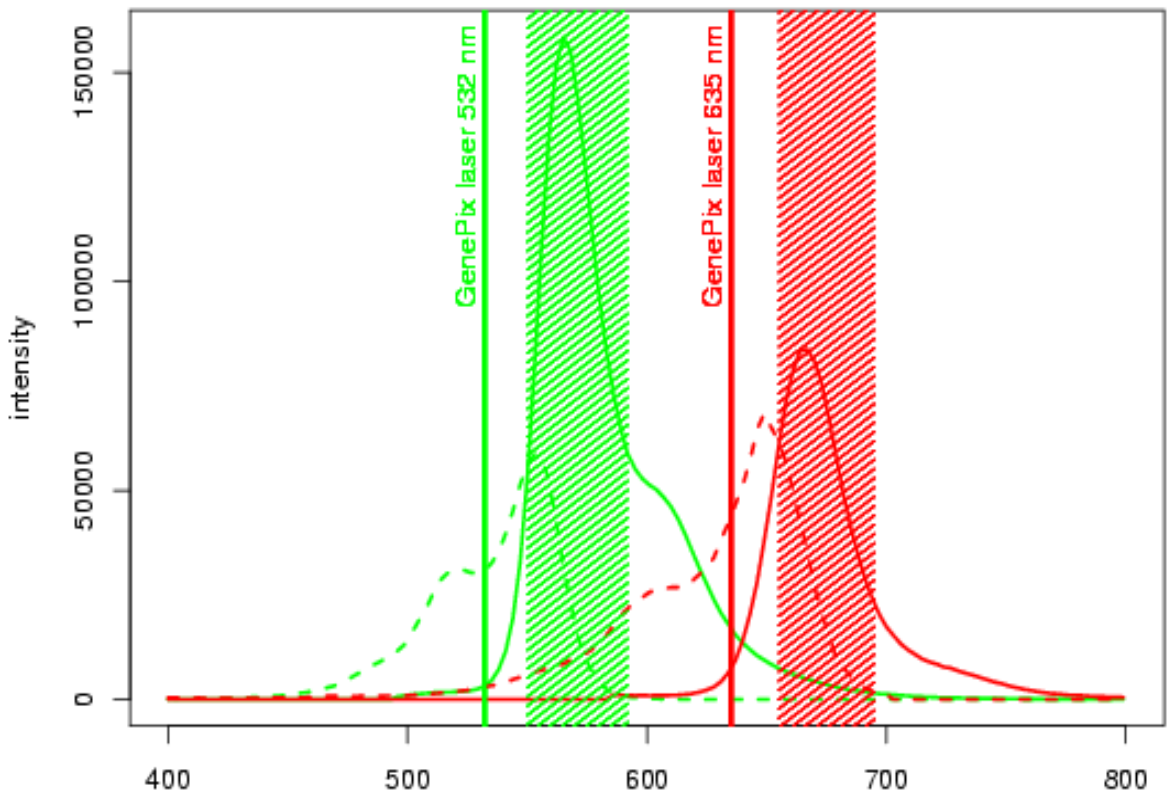
Další analýza

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

Dvoukanálové skenování

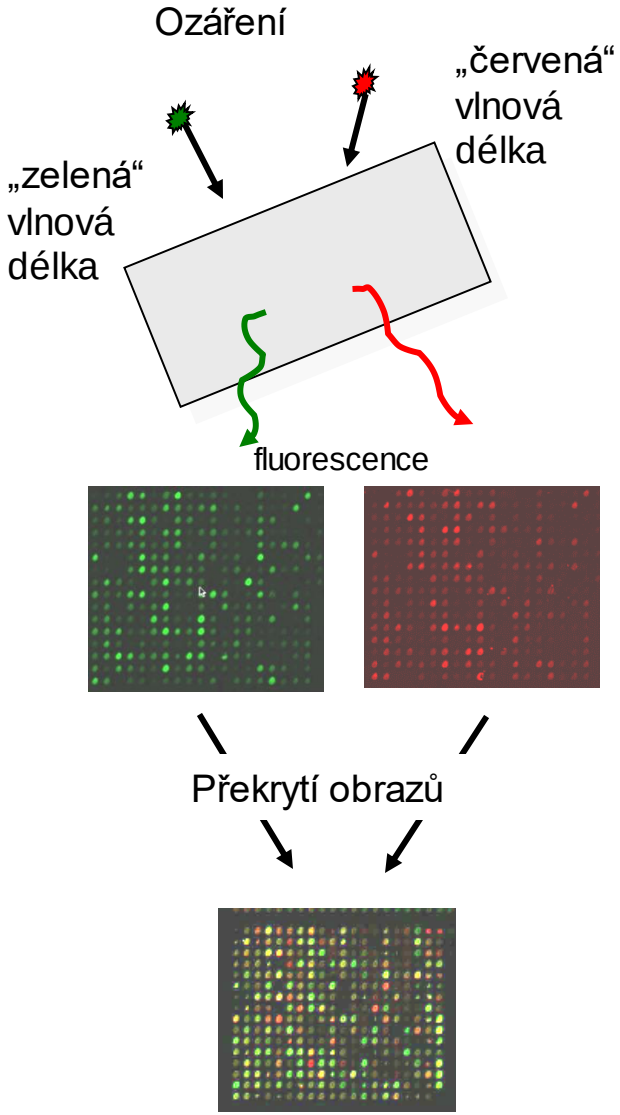
Po hybridizaci vkládáme sklíčko do skeneru abychom vytvořili obrázek mikročipu.

Excitační a emisní spektra Cy3 a Cy5



⇐ Vyšší frekvence, více energie

⇒ Nižší frekvence, méně energie

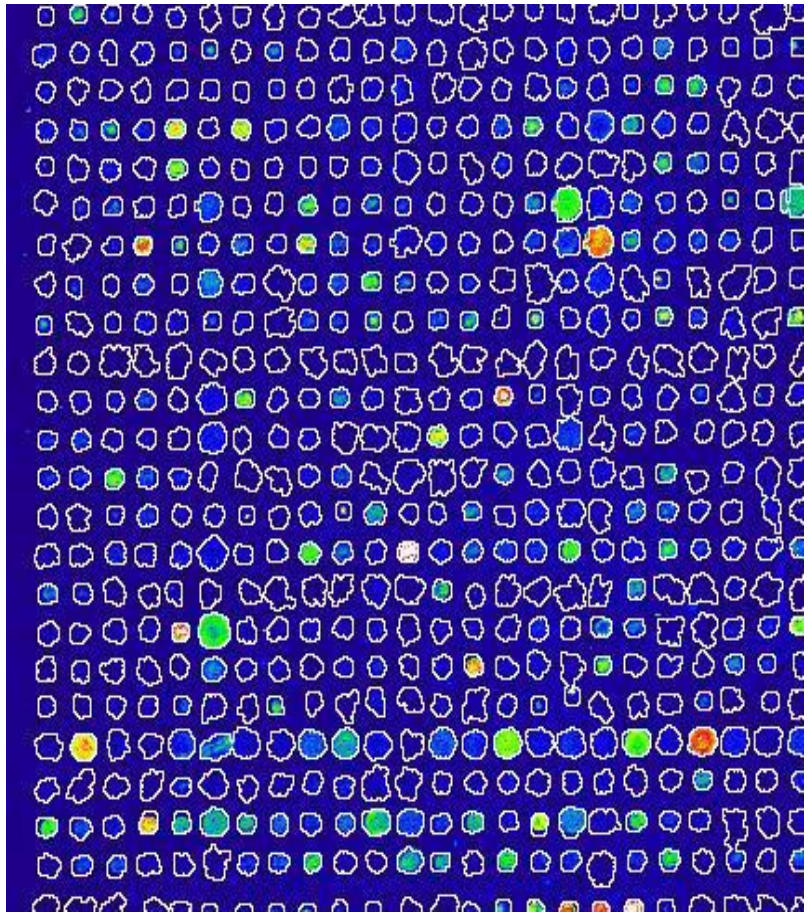


Analýza obrazu

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

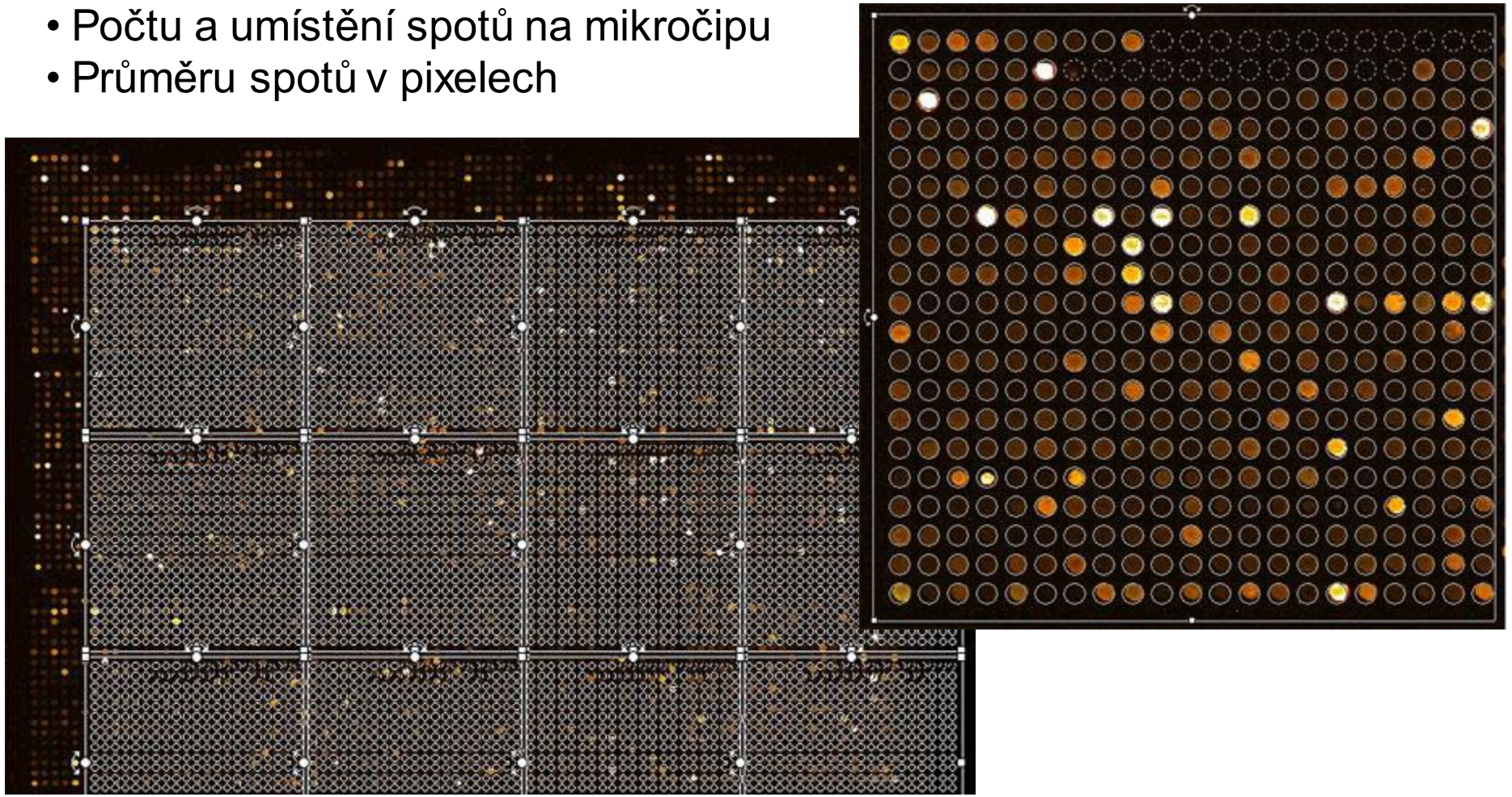


Terry Speed et al.

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



Terry Speed et al.

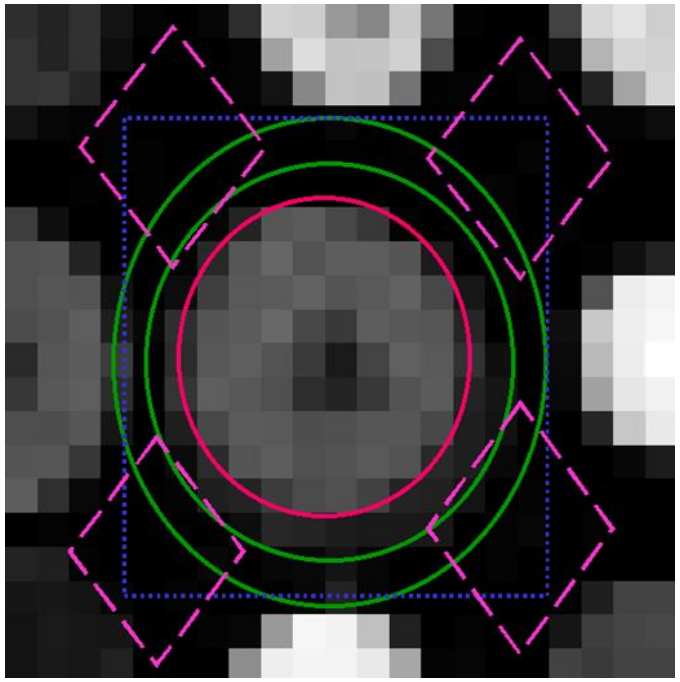
Segmentace

- 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 spotů:
  - Fixed circles
  - Adaptive circles
  - Histogram adaptive
- Různé programy různě definují **pozadí** spotu

GenePix

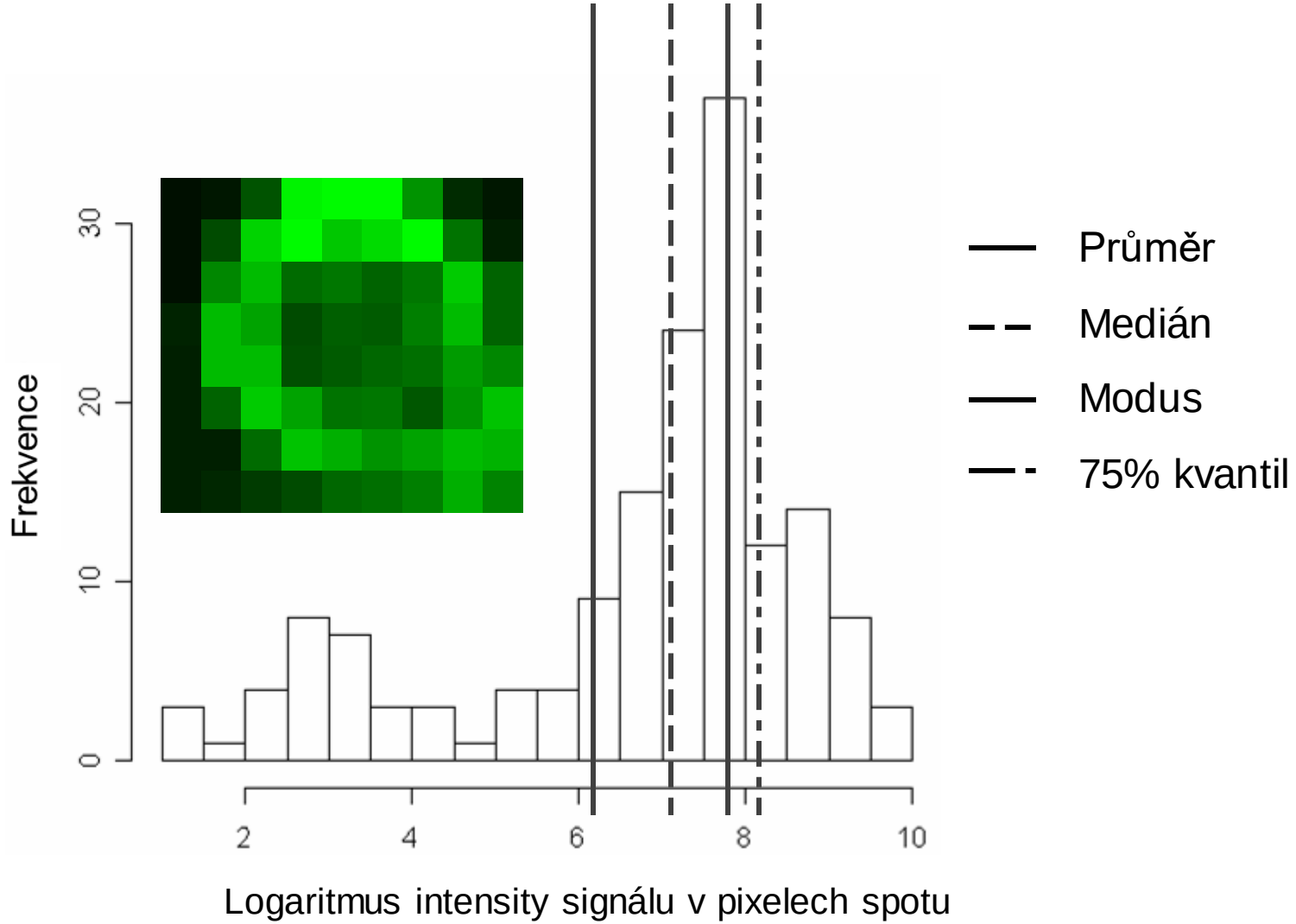
QuantArray

ScanAlyse



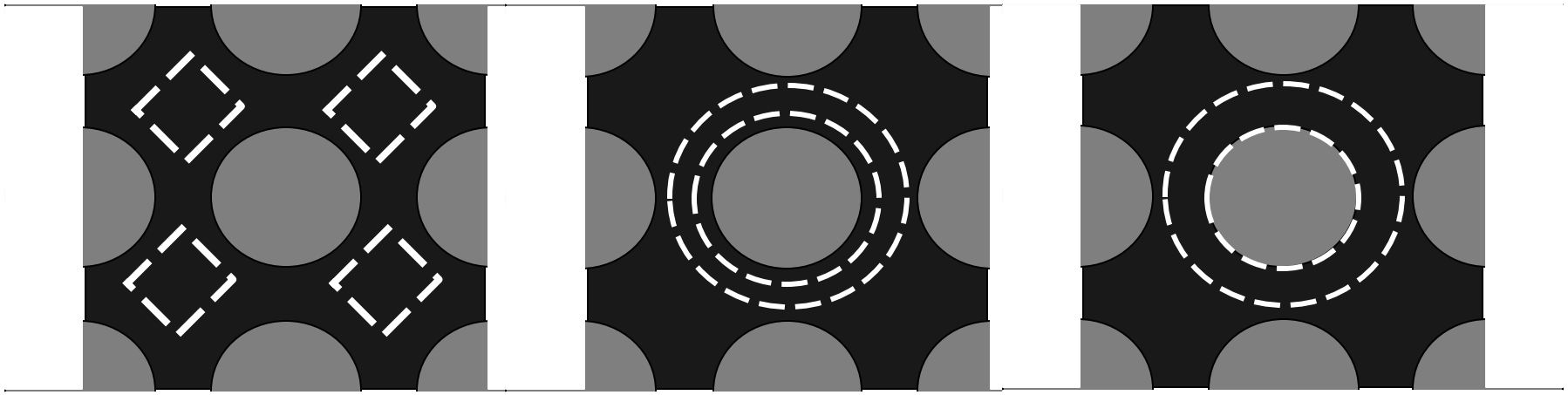
Kvantifikace signálu

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



Kvantifikace signálu pozadí

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

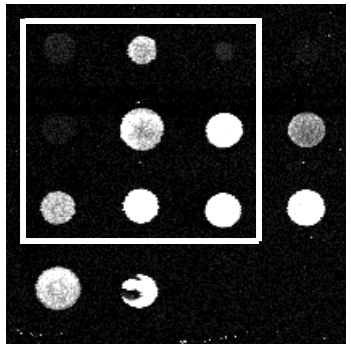


GenePix QuantArray ScanAlyse

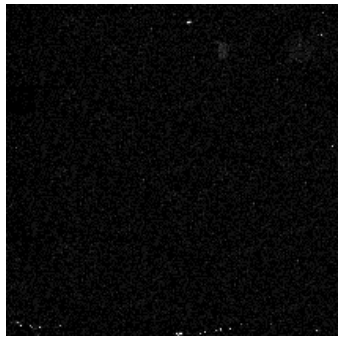
Vizualizace oblastí lokálního odhadu intenzity pozadí u tří různých programů analýzy obrazu cDNA mikročipu

Kvantifikace signálu pozadí 2.

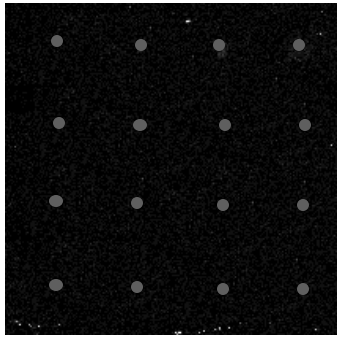
- Tři druhy metod:
  1. Lokální metoda (local background)
  2. **Morfologické otevření (morphological opening)**
  3. Konstantní/globální metoda (constant/global background)



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

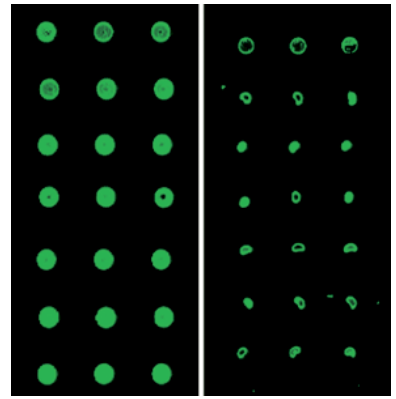
- I po kvantifikaci je možné manuálně označit spoty, které považujeme za nekvalitní

- Spotům, 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:

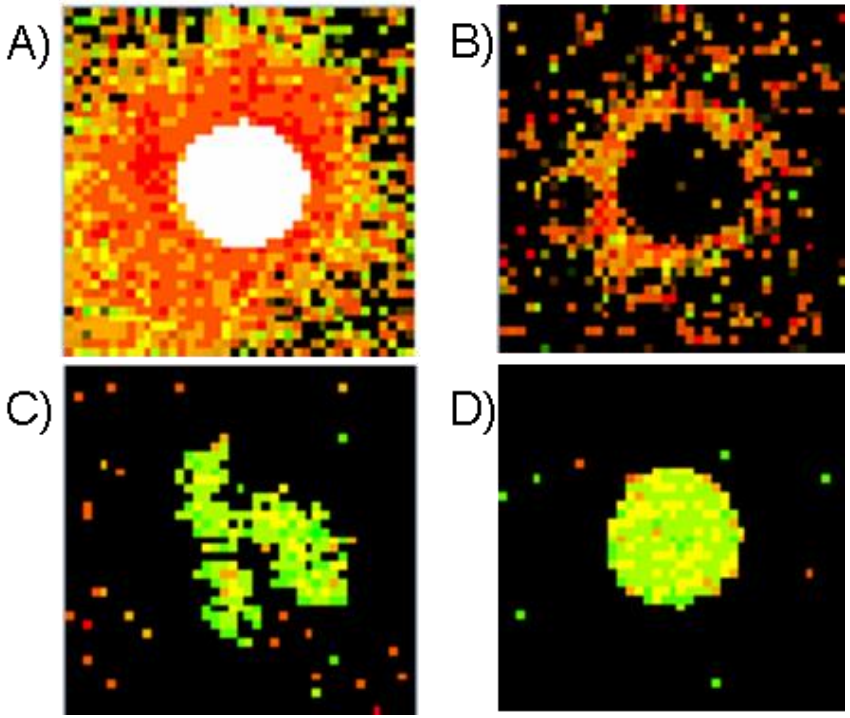
- **Velikost a tvar spotu**
  - Příliš malé spoty neposkytují věrohodné odhady intensity hybridizace (Simon et al., 2003) (spoty menší než < 25 pixelů by měly být odstraněné)
  - Spoty s nepravidelným tvarem, případně "koblihové spoty" by měly být označené jako nekvalitní
- **Intensita signálu**
  - 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 velký
- **Nasycení (saturace) spotu**
  - Spoty by neměly obsahovat nasycené pixely!



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 mikročipového 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

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 stáhněte a rozbalte.

Struktura adresáře:

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

Vyberte jeden ze souborů z adresáře 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
…
```