

Analýza genomických a proteomických dat

cDNA mikročipy - Kontrola kvality a normalizace

Jaro 2022

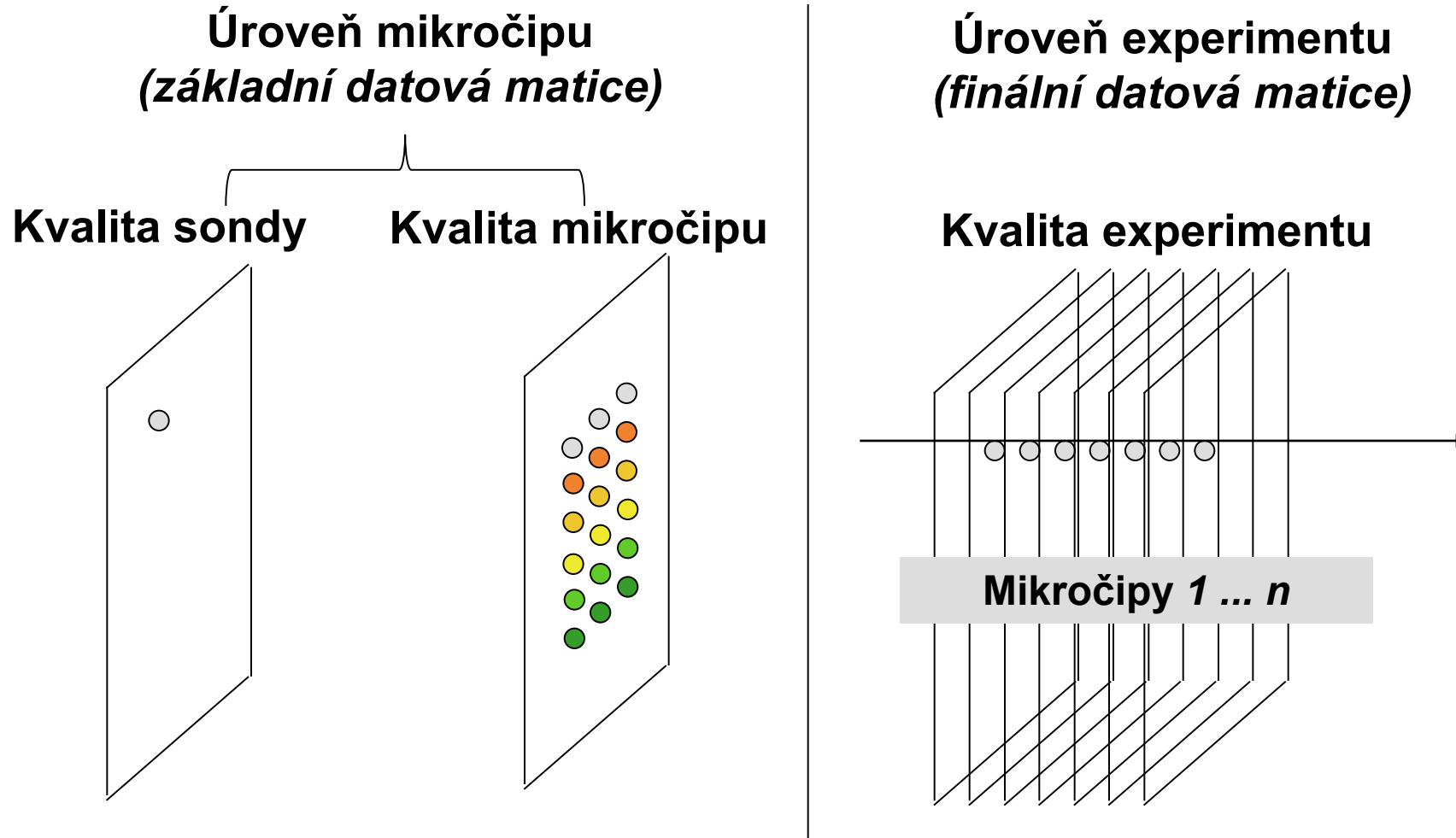
1. a 8. březen 2022

Eva Budinská (budinska@recetox.muni.cz)



cDNA mikročipy – kontrola kvality

Úrovně kontroly kvality

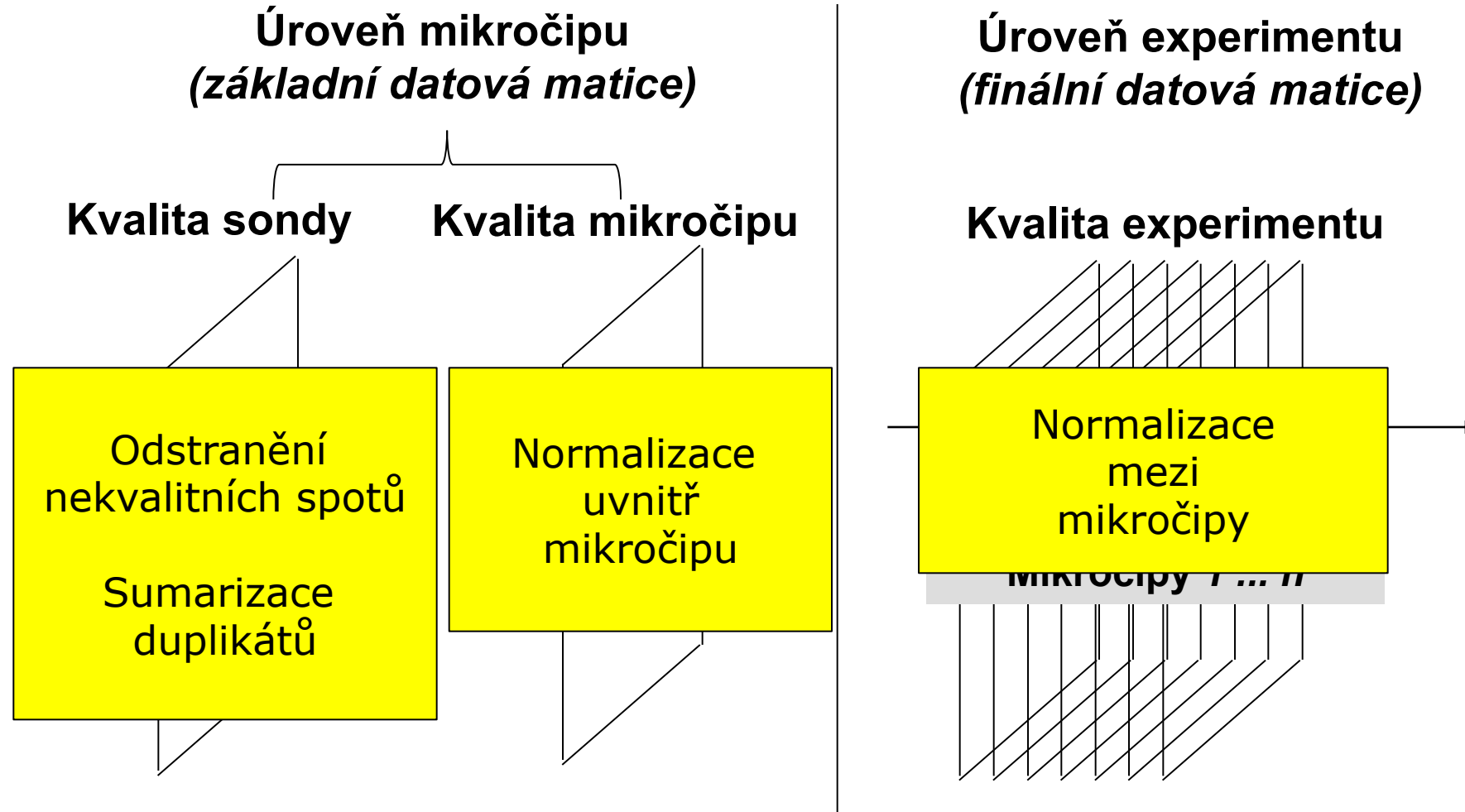


Úroveň sondy: Kvalita jednoho spotu na mikročipu

Úroveň mikročipu: Kvalita celého mikročipu

Úroveň experimentu: Kvalita měření transkriptu všech mikročipů v experimentu

Úrovně úpravy datových souborů

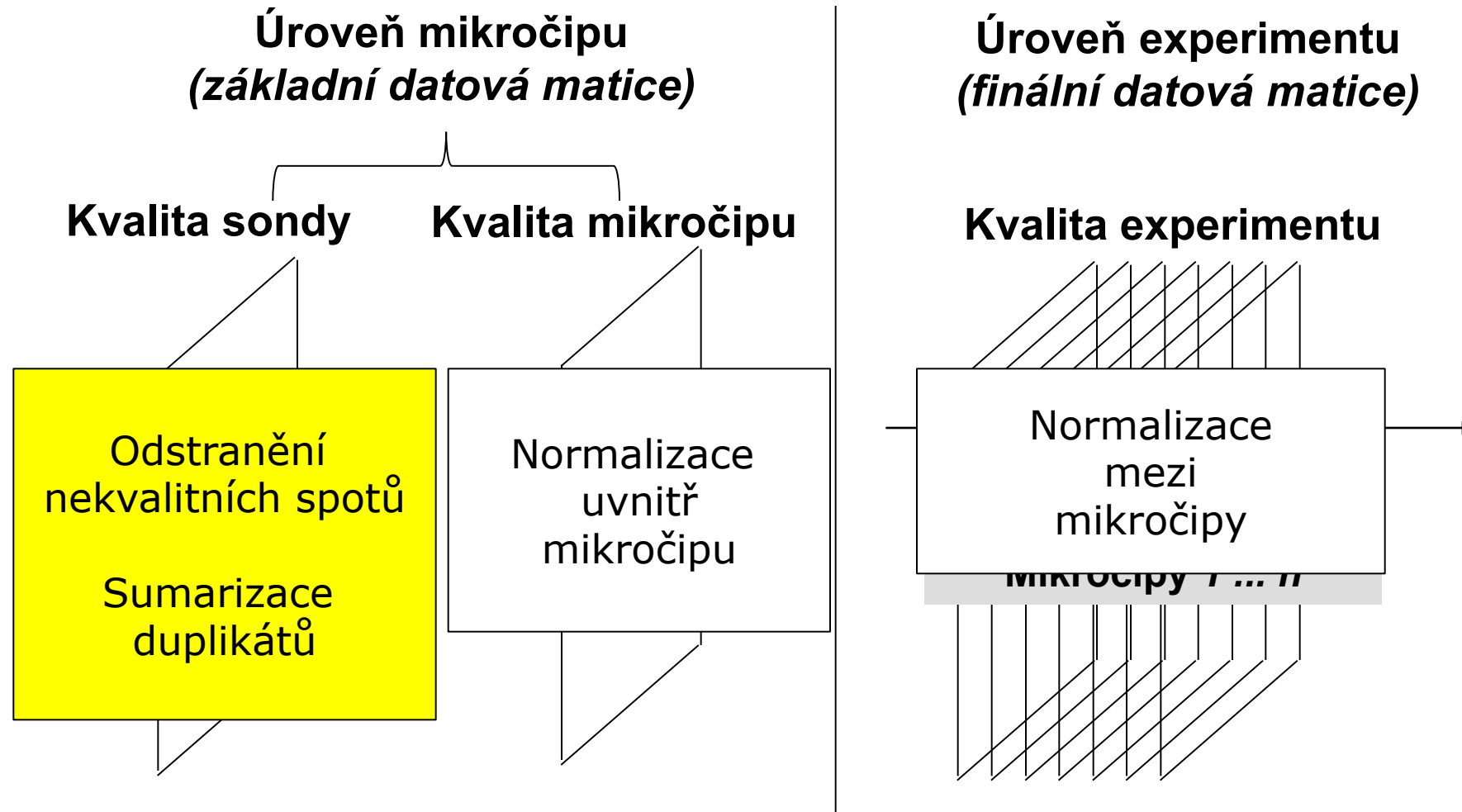


Úroveň sondy: Kvalita jednoho spotu na mikročipu

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)

clone	Replicate			mean	median	SD	No. of non-flagged replicates
	1	2	3				
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

Bud' odstranit sondy s příliš velkou variabilitou mezi replikáty...

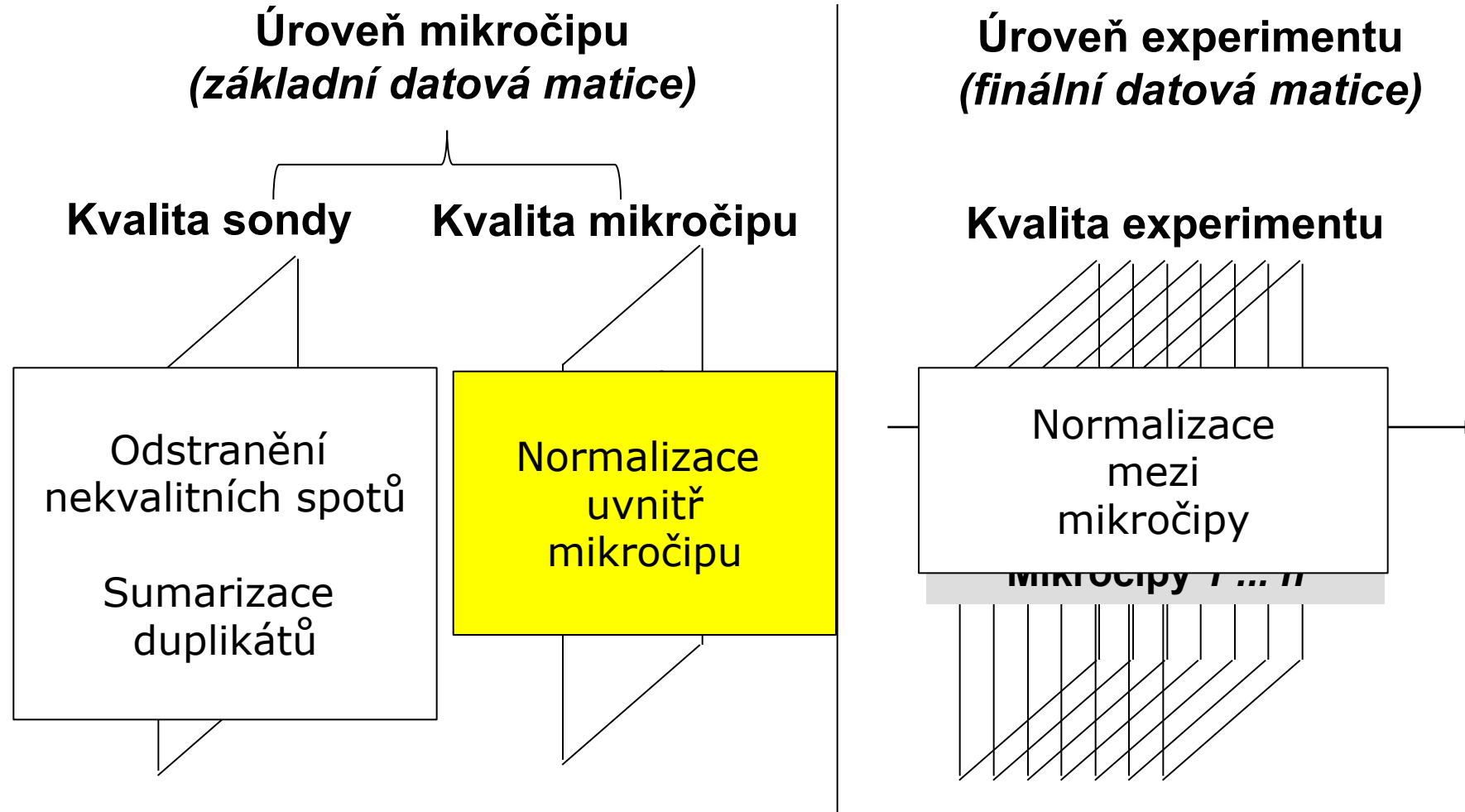
- ...nebo si uschovat informaci o počtu validních replikátů (a vyhodit klony jen s jedním replikátem)

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

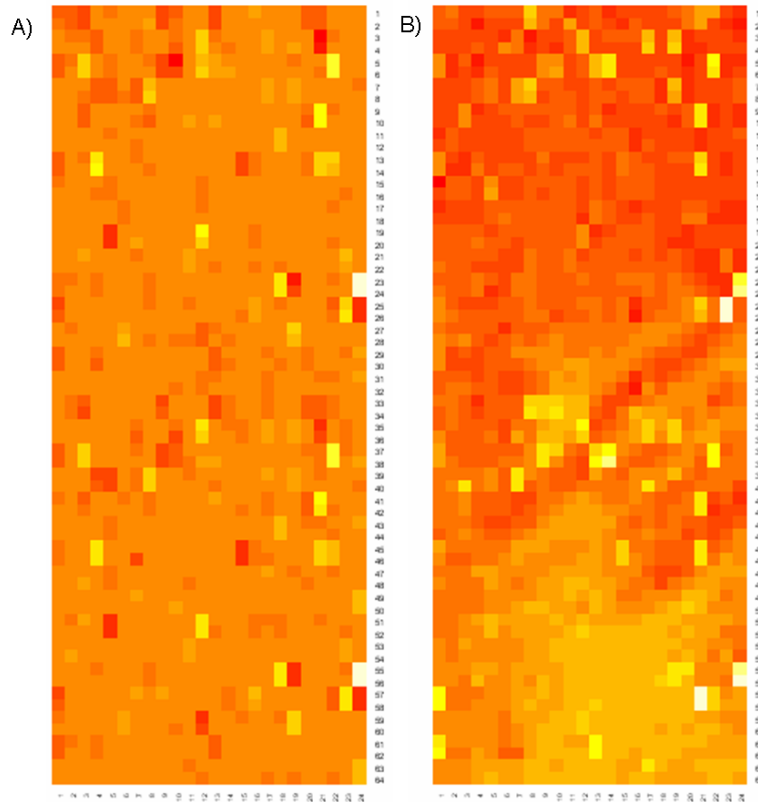
Systematické odchylky uvnitř mikročipu

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

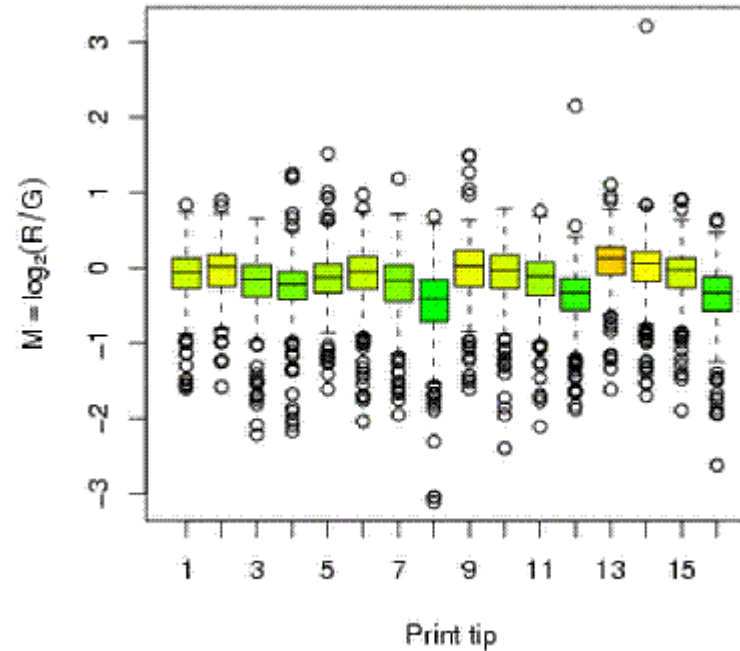
ODHALUJEME GRAFICKOU REPREZENTACÍ DAT

Diagnostika nerovnoměrné hybridizace

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



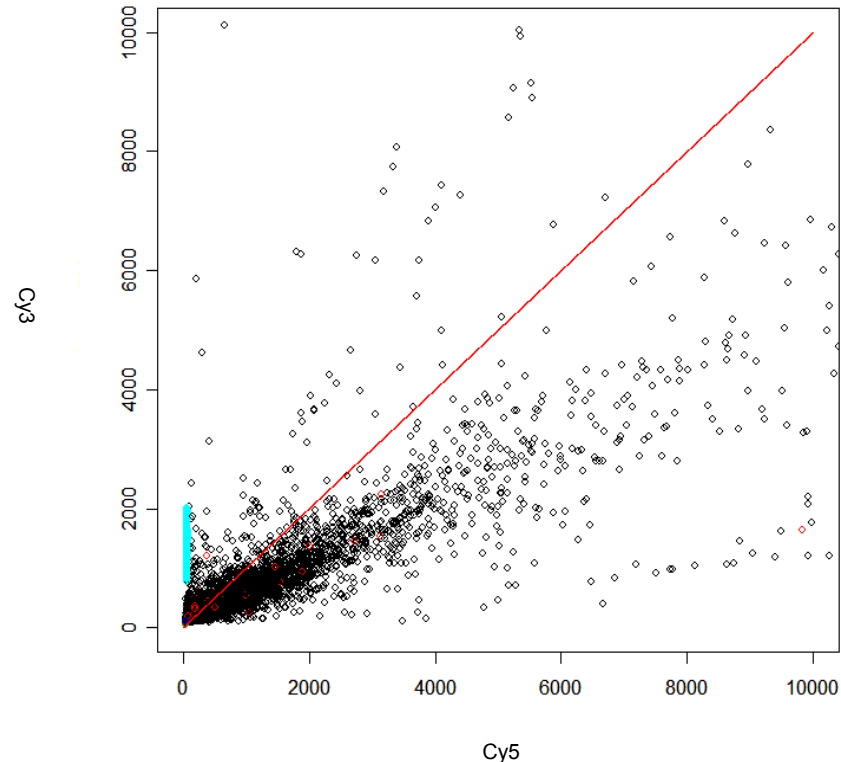
Krabicové grafy jednotlivých
oblastí (nejčastěji print-tip)



Diagnostika efektu barviva

- Často je efekt barviva větší u sond s nízkou expresí

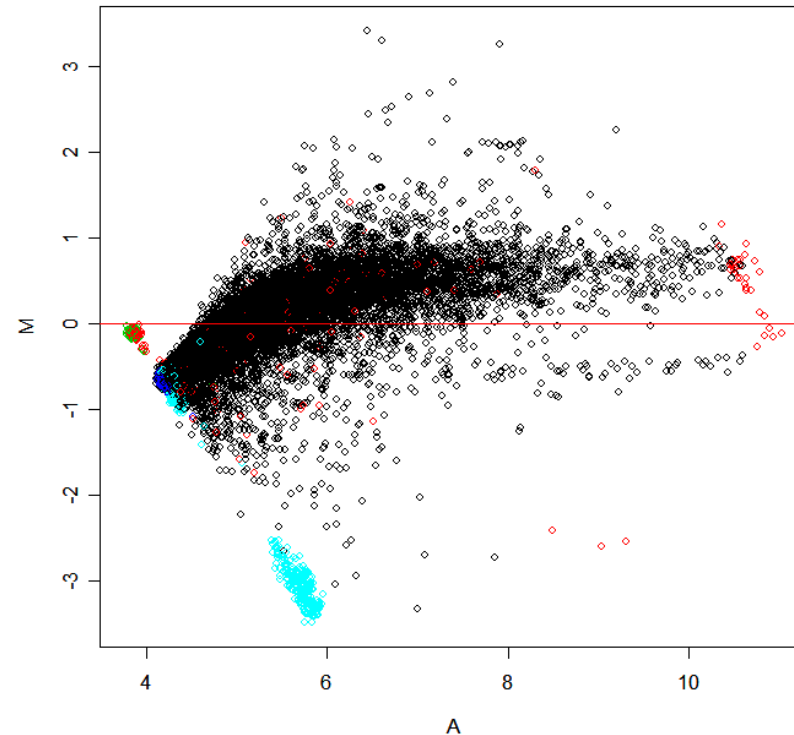
Graf intenzit kanálů



$$\text{Cy3} = B_0 + B_1 \cdot \text{Cy5}$$
$$(\text{Cy3} - B_0) / B_1 = \text{Cy5}'$$

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
- Ukážeme si jak instalovat balíky pro specifické analýzy genomických a proteomických dat
- Na příkladových datech uděláme diagnostiku kvality sklíčka

Bioconductor

- Bioconductor je projekt v R speciálně určený pro analýzu molekulárních dat
- Obsahuje nejenom speciální balíky, ale i typy objektů, smyslem je standardizace a minimalizace chyb!

- Jak instalovat:

- <https://www.bioconductor.org/install/>

- Do R příkazového řádku zadáme:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager") BiocManager::install(version =
"3.16")
```

- Instalace základních balíčků:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager") BiocManager::install()
```

Bioconductor – instalace balíků

Pro instalaci specifického balíku použijeme kód:

```
BiocManager::install(c("nazevbaliku1", "nazevbaliku2"))
```

POZOR NA uvozovky, musí být ", ne ``

Balík marray

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

```
BiocManager::install("marray")
```

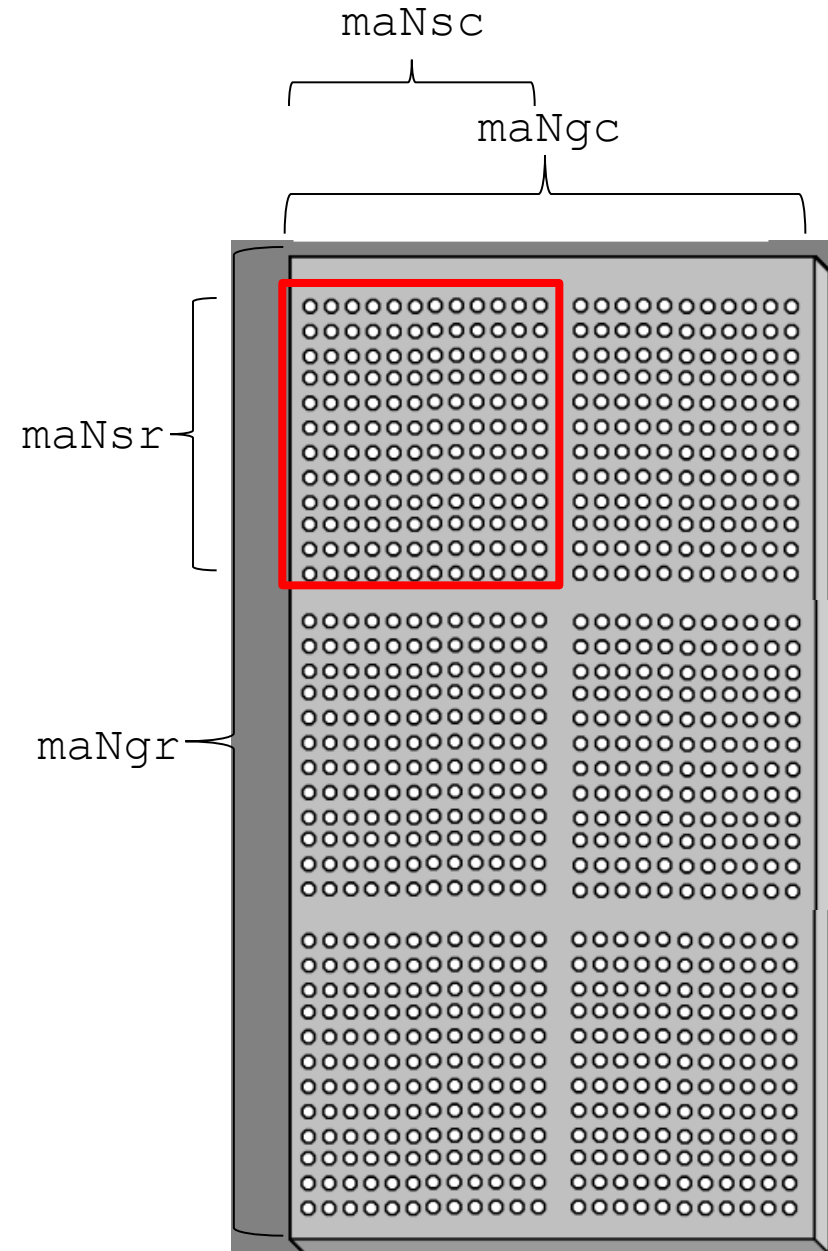
- Základní struktura, se kterou pracuje, a která obsahuje základní data všech matic experimentu je třída `marrayRaw`

```
new('marrayRaw',  
    maRf = ....., # matice intenzit spotů červeného kanálu  
    maGf = ....., # matice intenzit spotů zeleného kanálu  
    maRb = ....., # matice intenzit pozadí červeného kanálu  
    maGb = ....., # matice intenzit 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  
)
```


V Rstudio si otevřeme soubor `cDNA-kontrolaKvality-priklad1.R`

Cvčení

cDNA mikročipy – normalizace

Normalizace uvnitř mikročipu I.

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

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

kde l a s jsou normalizační hodnoty střední hodnoty (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

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

střední hodnota odhadnutého signálu pozadí

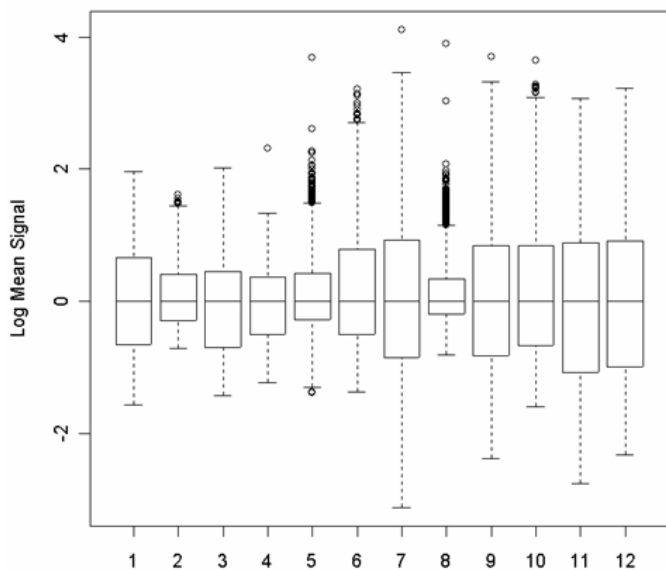
2) bez korekce!

Normalizace uvnitř mikročipu I - metody

3) Normalizace prostorového efektu a rozdílů intenzit mezi kanály

- **Centrování mediánem**

- odečítá medián signálu od intenzit signálu všech spotů
- nejjednodušší, ale není schopný zkorigovat nelinearitu



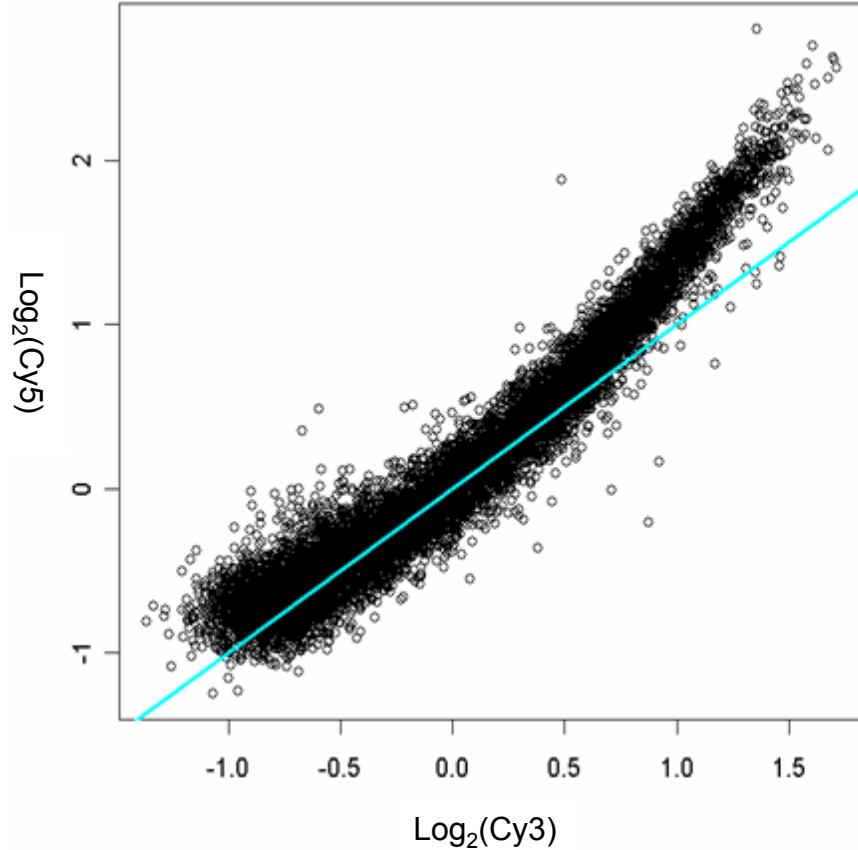
$$M_{norm} = M - l,$$

l je medián intenzit signálu všech spotů

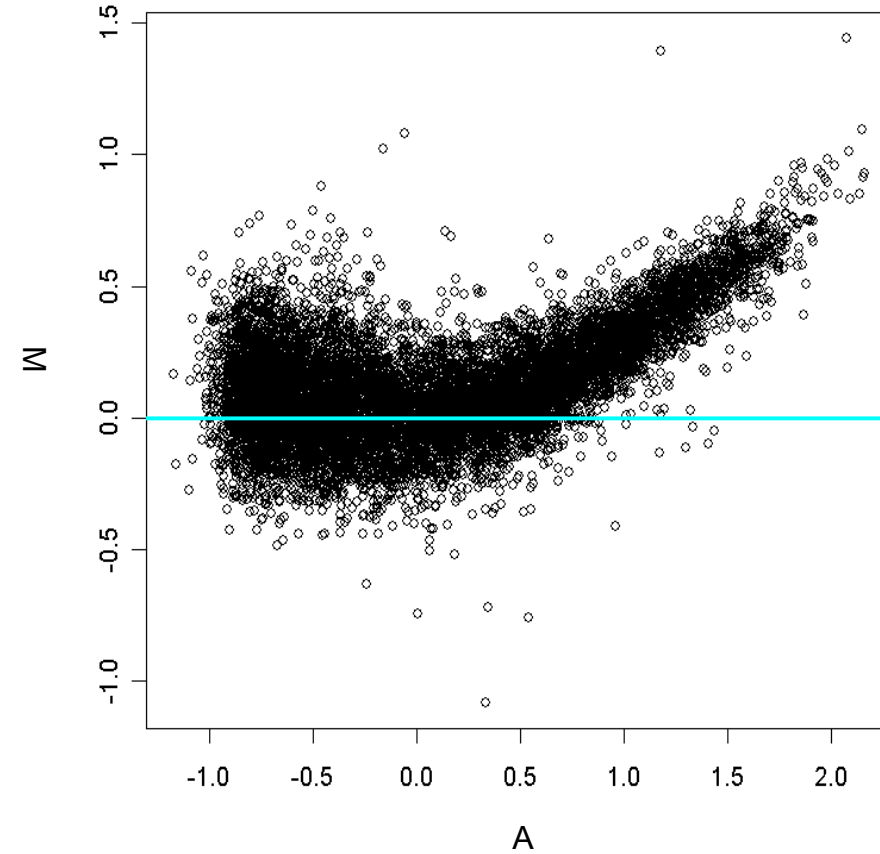
Problémy s mediánovým centrováním

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

Graf intenzit kanálů



MA graf



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

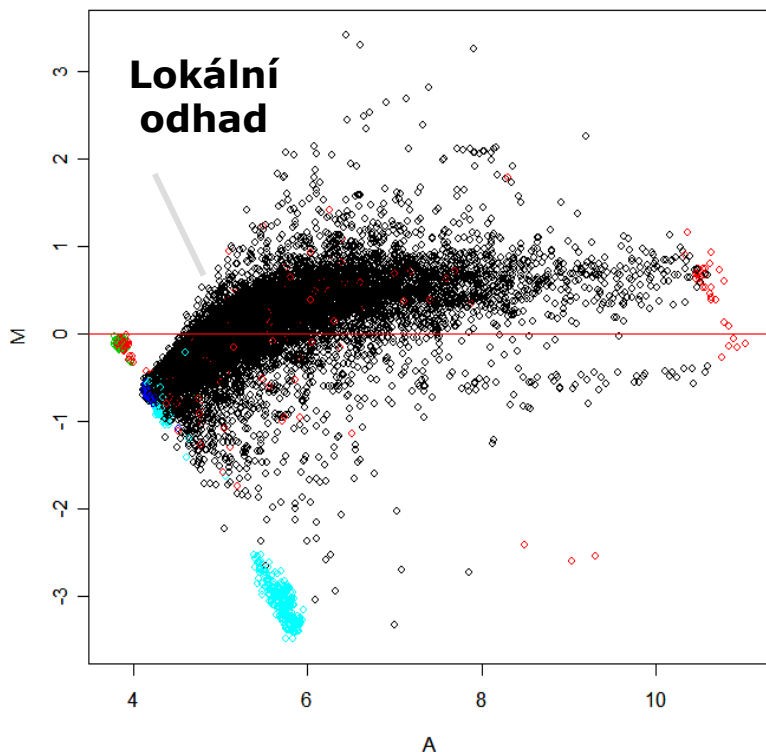
Lowess normalizace I

Princip:

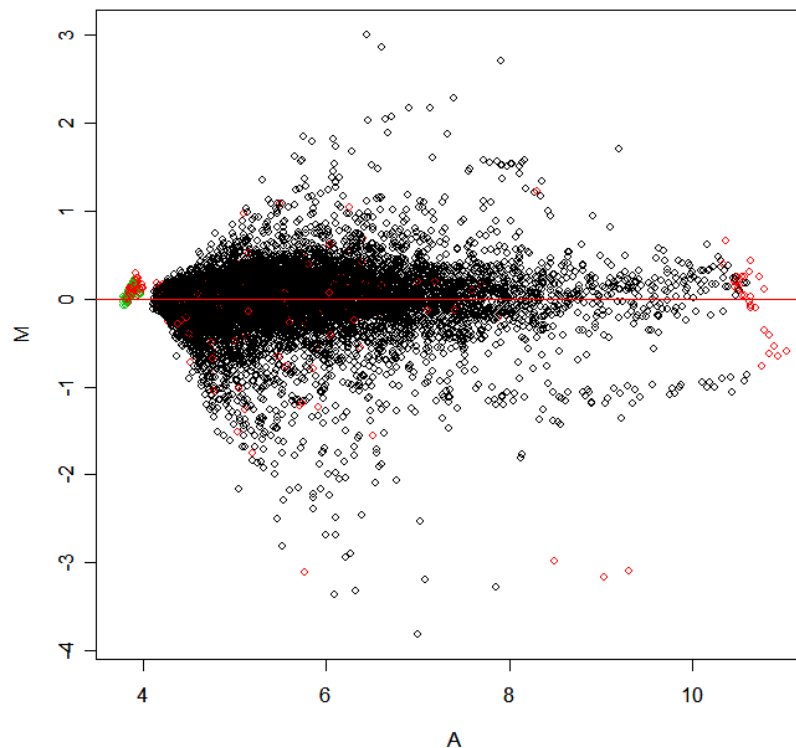
1. Odhad křivky pomocí neparametrické lokálního (váženého) vyhlazování (lo(w)ess - locally (weighted) scatterplot smoothing)
2. Odečtení odhadnuté křivky od naměřených hodnot

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

Před lowess normalizací



Po lowess normalizaci



Lowess normalizace II

Princip lowess

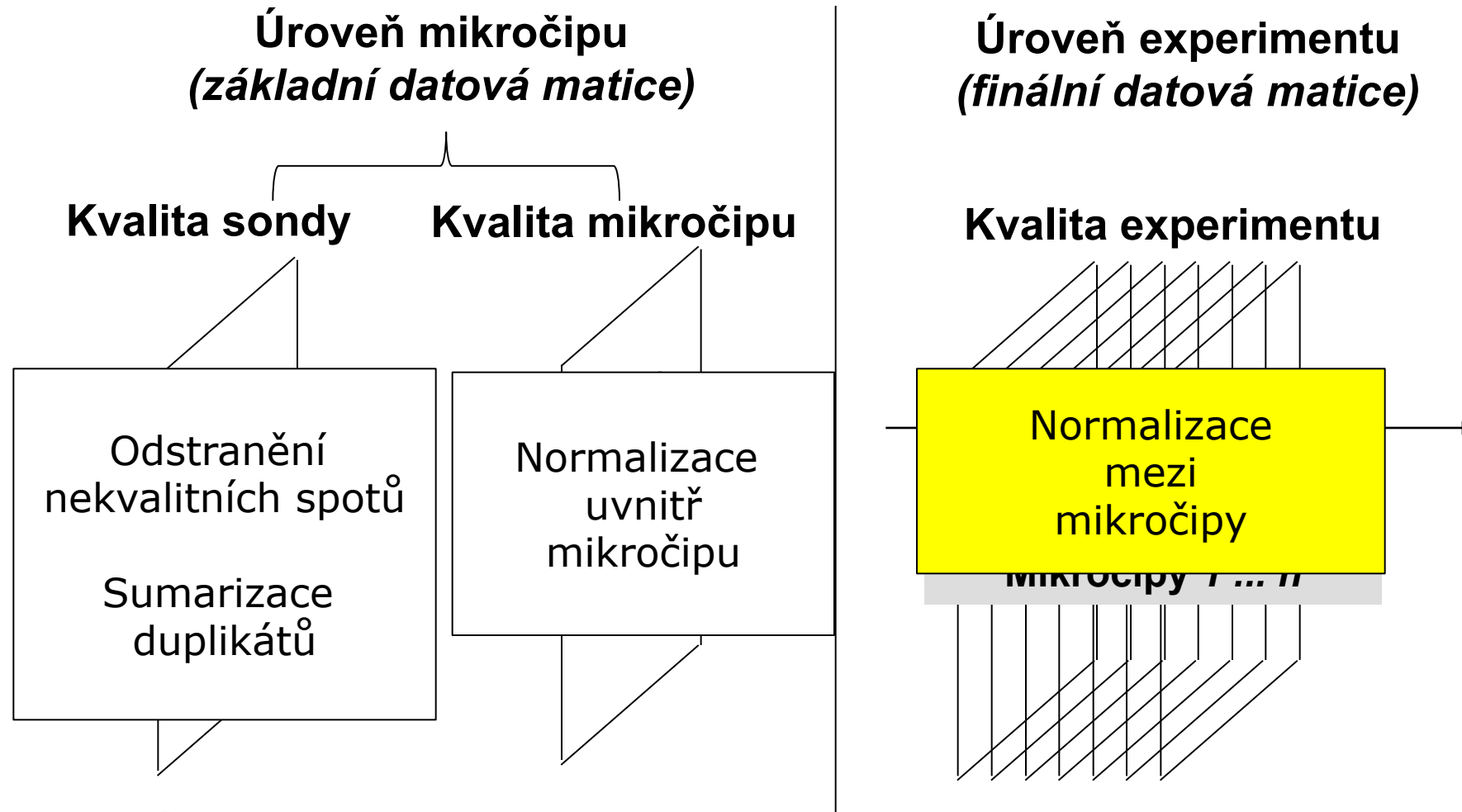
- V každém kroku se určí lokální množina dat, na které se **odhadne křivka s pomocí polynomiálu a metody nejmenších čtverců**
- Parametr λ určuje stupeň polynomiálu ($\lambda=0$ průměr, $\lambda=1$ lineární regrese, $\lambda=2$ kvadratická regrese)
- Množina dat na které se pracuje se určuje pomocí algoritmu nejbližšího souseda
- Vyhlazovací parametr 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 (abychom neodstranili reálné biologické rozdíly)
 - ii) množina by měla být dostatečně velká, aby zachytila variabilitu sklíčka

Např. housekeeping geny

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 sklíčků 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				pořadí				Seřazené hodnoty			
Gen	čip1	čip2	čip3	Gen	čip1	čip2	čip3	čip1	čip2	čip3	
A	5	4	3	A	iv	iii	i	i	2	1	3
B	2	1	4	B	i	i	ii	ii	3	2	4
C	3	4	6	C	ii	iii	ii	iii	4	4	6
D	4	2	8	D	iii	ii	iv	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				pořadí				Seřazené hodnoty			
Gen	čip1	čip2	čip3	Gen	čip1	čip2	čip3		čip1	čip2	čip3
A	5	4	3	A	iv	iii	i	i	2	1	3
B	2	1	4	B	i	i	ii	ii	3	2	4
C	3	4	6	C	ii	iii	ii	iii	4	4	6
D	4	2	8	D	iii	ii	iv	iv	5	4	8

průměr

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

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				pořadí				Seřazené hodnoty			
Gen	čip1	čip2	čip3	Gen	čip1	čip2	čip3	čip1	čip2	čip3	
A	5	4	3	A	iv	iii	i	i	2	1	3
B	2	1	4	B	i	i	ii	ii	3	2	4
C	3	4	6	C	ii	iii	ii	iii	4	4	6
D	4	2	8	D	iii	ii	iv	iv	5	4	8

průměr				normalizované hodnoty			
Gen	čip1	čip2	čip3	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

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

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 abychom odstranili systematické (technické) chyby

Procvičování na doma

- Podíváme se do našeho adresáře s cDNA příkladem a otevřeme cDNA.R v programu Rstudio.
- Postupujeme dle instrukcí, na konci je dobrovolný úkol.