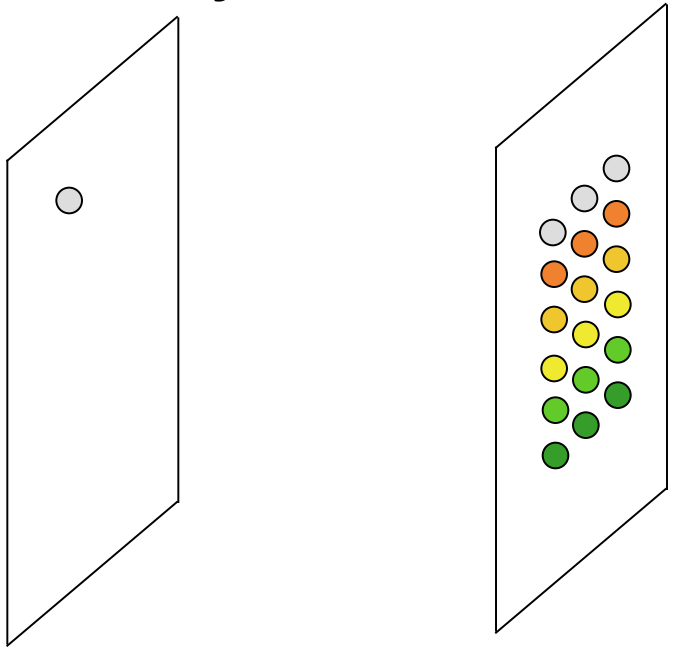
Úrovne kontroly kvality

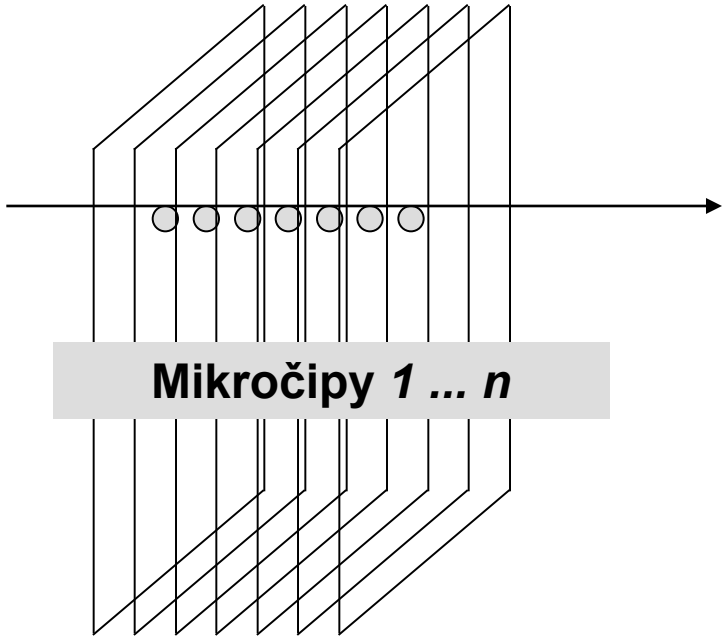
Úroveň mikročipu
(základní datová matice)

Kvalita sondy **Kvalita mikročipu**



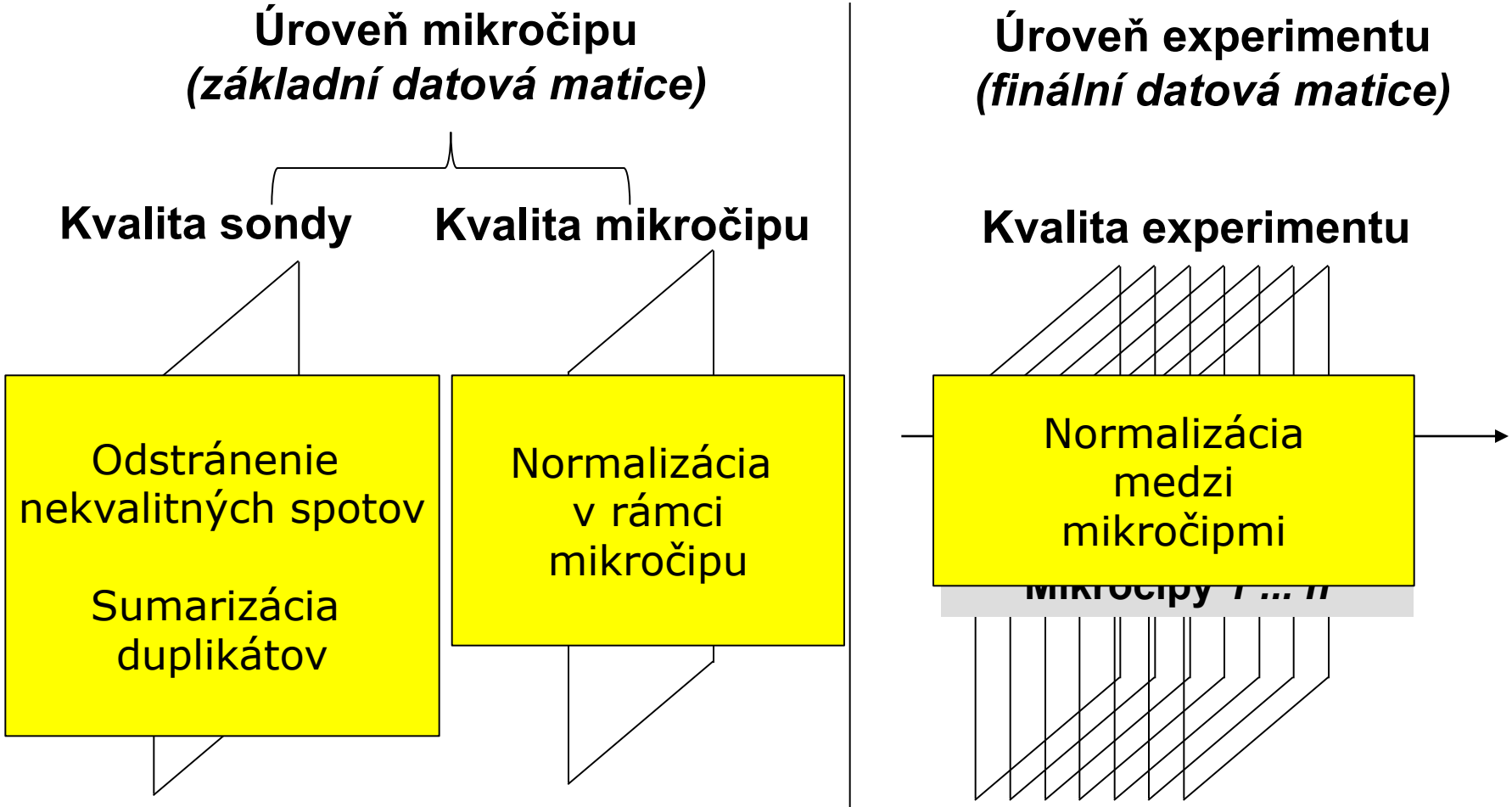
Úroveň experimentu
(finální datová matice)

Kvalita experimentu



Úroveň sondy: Kvalita jedného spotu na mikročipe
Úroveň mikročipu: Kvalita celého mikročipu
Úroveň experimentu: Kvalita měření transkriptú všech mikročipů v experimentu

Úrovne úprav dátových súborov



Úroveň sondy: Kvalita jedného spotu na mikročipe
Úroveň mikročipu: Kvalita celého mikročipu
Úroveň experimentu: Kvalita měření transkriptů všech mikročipů v experimentu

Úrovne úprav dátových súborov

Úroveň mikročipu
(základní datová matice)

Kvalita sondy

Odstránenie nekvalitných spotov

Sumarizácia duplikátov

Kvalita mikročipu

Normalizácia v rámci mikročipu

Úroveň experimentu
(finální datová matice)

Kvalita experimentu

Normalizácia medzi mikročipmi

Úroveň sondy: Kvalita jedného spotu na mikročipe
Úroveň mikročipu: Kvalita celého mikročipu
Úroveň experimentu: Kvalita měření transkriptú všech mikročipů v experimentu

Kontrola dát v rámci microarray sklíčka

- **Replikáty sond**
  - Sumárne štatistiky replikátov spotov (nekvalitné spoty už vylúčené)

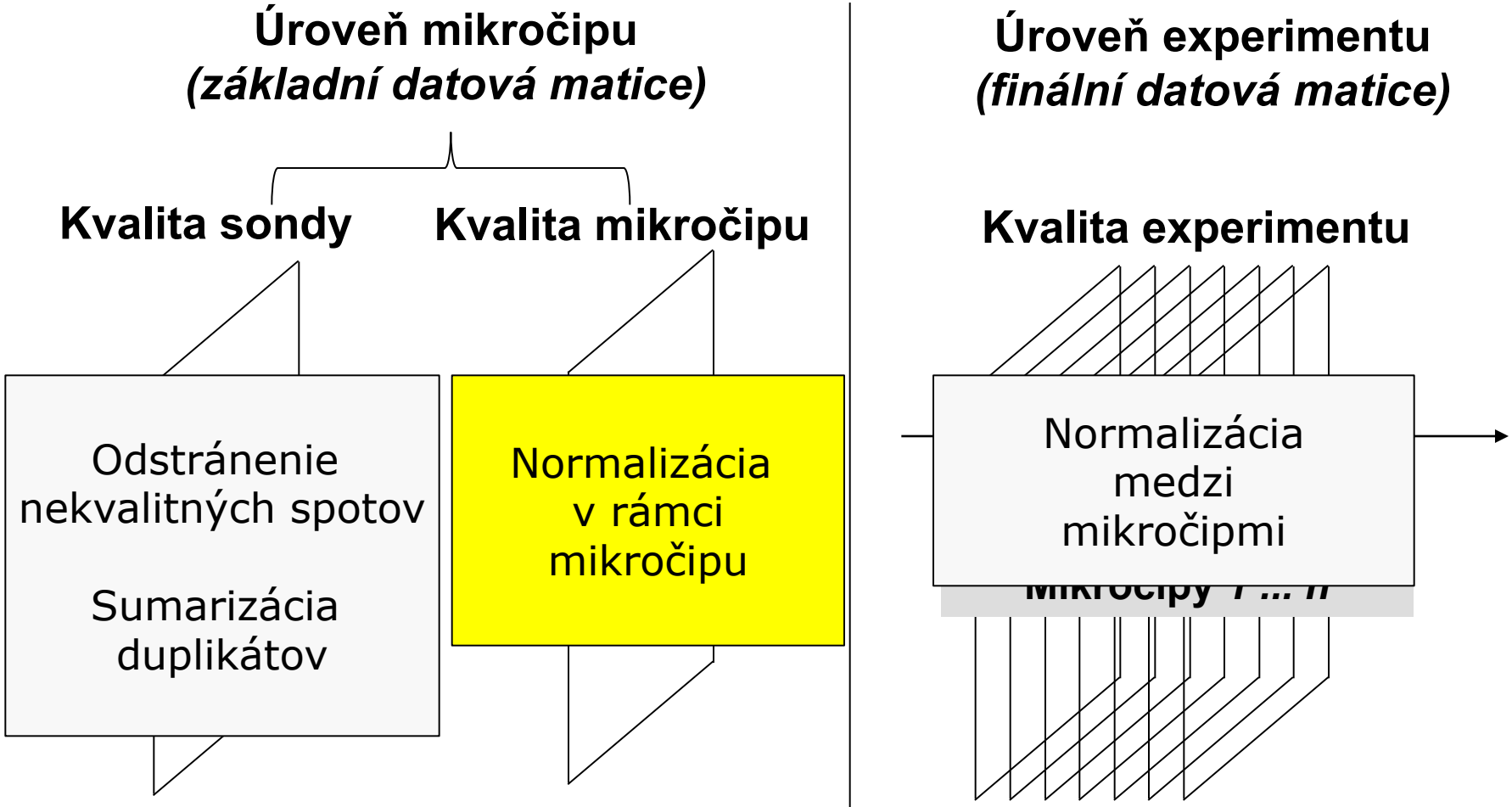
clone	Replicate			mean	median	SD	No. of non-flagged replicates
	1	2	3				
A_23_P347643	-0.186	-0.265	-0.313	-0.254	-0.265	0.052	3
A_23_P60243	0.523	flagged	flagged	0.523	0.523	0	1
A_23_P116057	0.039	-0.978	flagged	-0.495	-0.495	0.5	2
A_23_P203743	-0.614	0.537	1.589	0.504	0.537	0.899	3

  - Buď vyhodiť sondy s príliš veľkou variabilitou medzi replikátmi…
  - …alebo si uschovať informáciu o počte validných replikátov (a vyhodiť klony len s jedným replikátom)

- **Kvalita microarray sklíčka**
  - Percento nekvalitných spotov nesmie byť príliš veľké (<25 %)

- **Systematické odchýlky odstránime procesom NORMALIZÁCIE**

Úrovne úprav dátových súborov



Úroveň sondy: Kvalita jedného spotu na mikročipe
Úroveň mikročipu: Kvalita celého mikročipu
Úroveň experimentu: Kvalita měření transkriptů všech mikročipů v experimentu

Systematické odchýlky v rámci microarray sklíčka

- **Nerovnomerná hybridizácia** (priestorové odchýlky)
  - Príčina: nerovnomerne umytý čip, nerovnomerne distribuovaná vzorka, print-tip efekt (defektná ihla)

- **Signál pozadia**
  - Môže byť veľmi silný, buď zle umytý čip, alebo zlá segmentácia (časť popredia je kvantifikovaná ako pozadie)

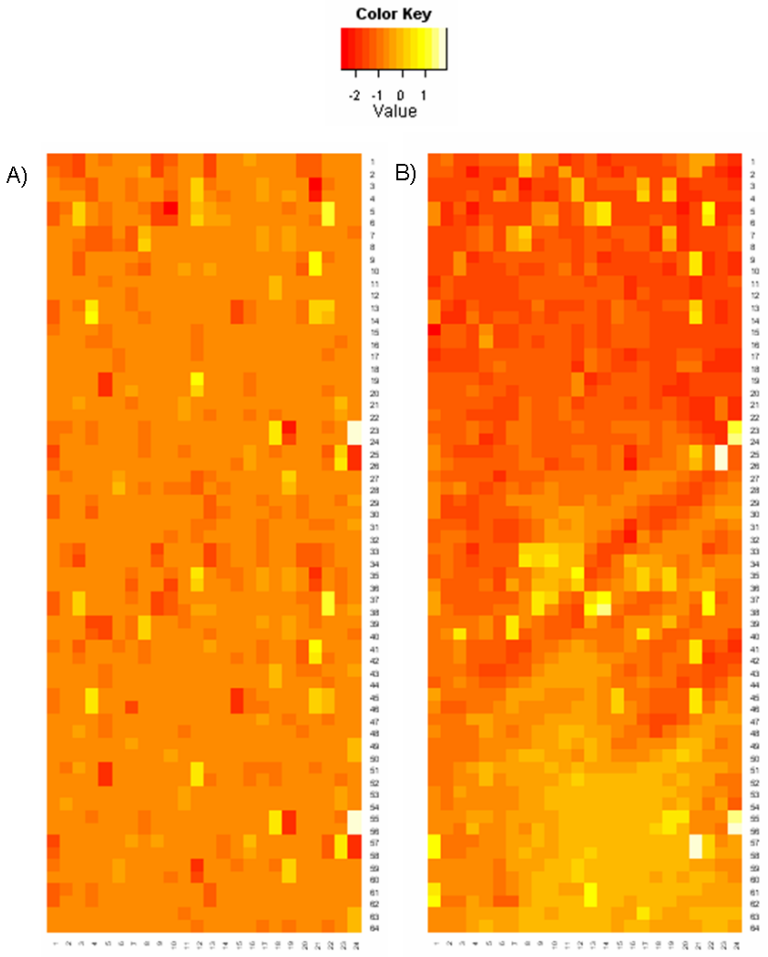
- **Efekt farbiva (rozdiely intenzít medzi kanálmi)**
  - Príčina: odlišná schopnosť inkorporácie molekúl farbiva (Cy3, Cy5)

    odlišná reakcia na excitáciu (slabšia intenzita UV, ...)

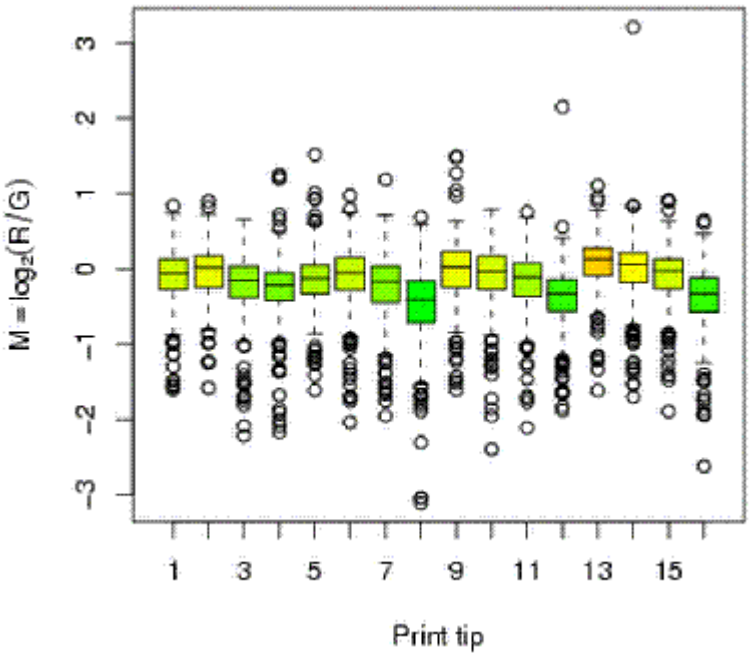
ODHAĽUJEME GRAFICKOU REPREZENTÁCIOU

Diagnostika nerovnomernej hybridizácie

Virtuálna rekonštrukcia microarray sklíčka, vykreslenie **heatmapy log2 pomeru Cy5/Cy3 intenzít** na základe ich pozície na sklíčku



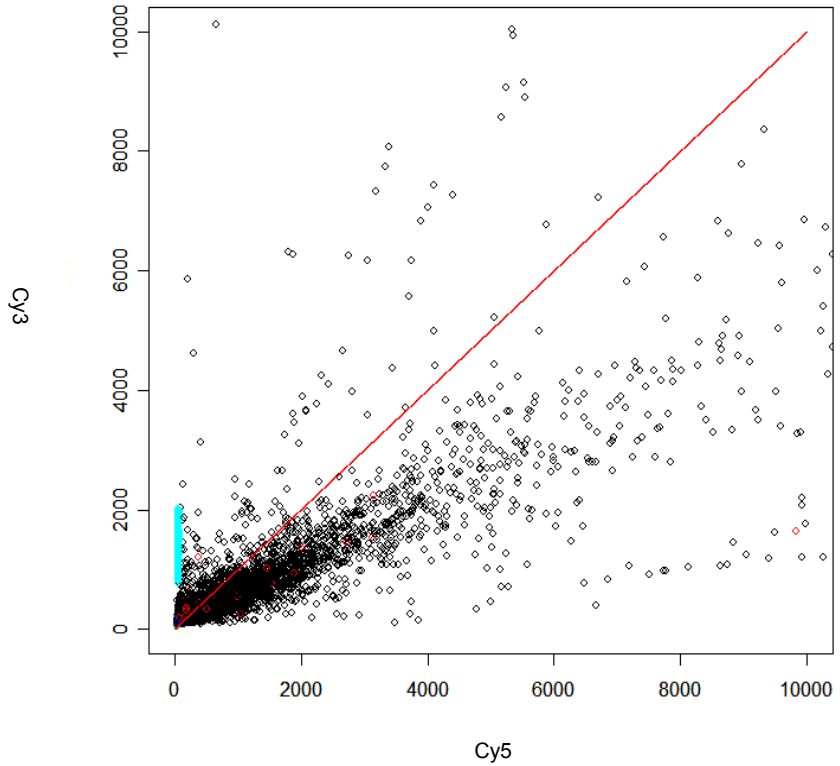
Box-ploty jednotlivých oblastí (najčastejšie print-tip)



Diagnostika efektu farbiva

- **Často je efekt farbiva väčší u sond s nízkou expresiou**

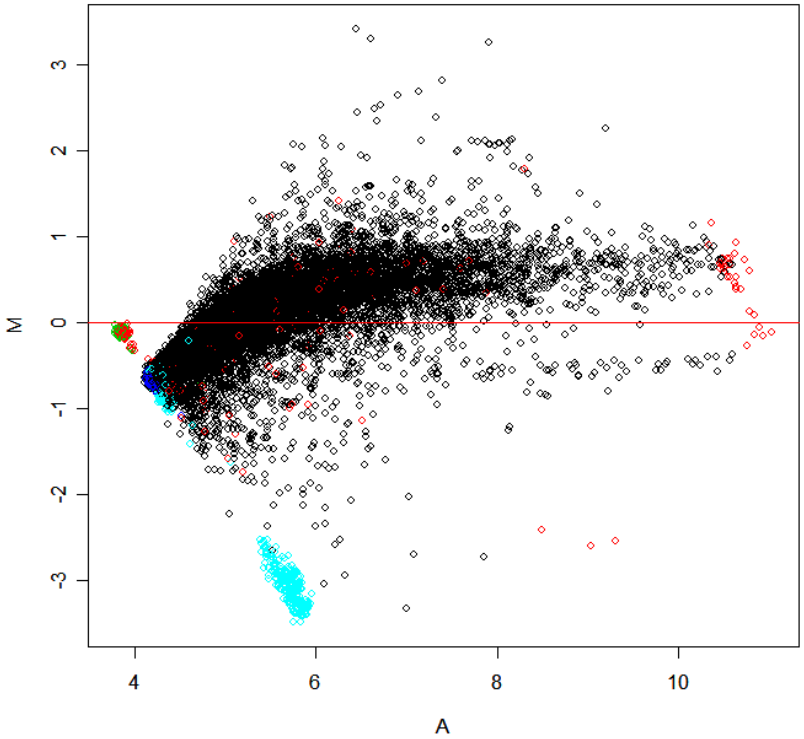
Graf intensit kanálů



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

Neukáže nelineárne trendy

MA graf



M = log (R/G)

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

Ukáže nelineárne trendy!

Normalizácia v rámci microarray sklíčka I.

- Cieľ: Upraviť hodnoty signálu tak, aby sme odstránili sytematické odchýlky v rámci microarray sklíčka
- Princíp: ***Centrovanie* a/alebo *škálovanie*** hodnôt expresie *M*

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

kde *l* a *s* sú normalizačné hodnoty centra a škály

Normalizácia v rámci microarray sklíčka I.

- Typy normalizácie:

1) Logaritmická transformácia - väčšinou používaná z dôvodov transformácie dát na normálne rozdelenie

$$M_{norm} = \log2(M)$$

2) Korekcia na pozadie

- odstraňuje efekt pozadia

- odlišné prístupy:

1) odpočíta sa odhadnutý signál pozadia – založené na predpoklade aditivity signálu

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

TS = OS - BS

- buď pre každý spot osobitne, alebo globálne

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

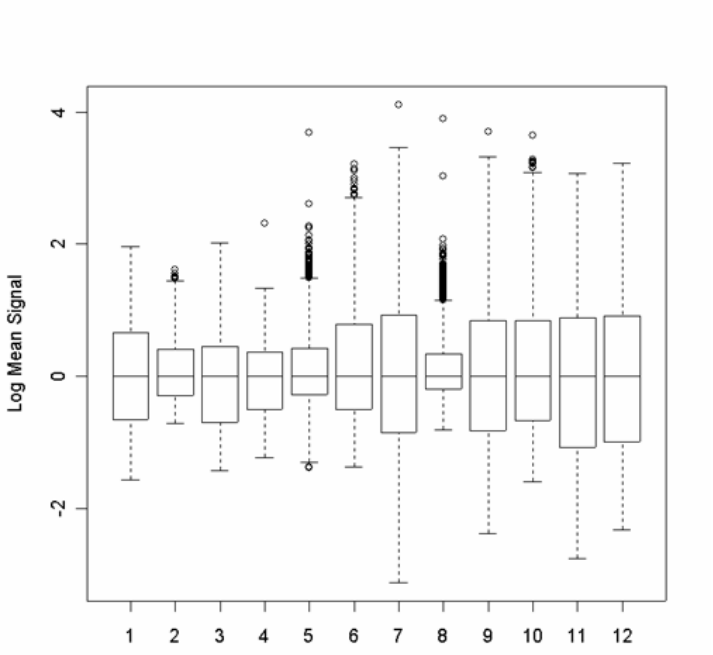
odhadnutý signál pozadia

2) bez korekcie!

Normalizácia v rámci microarray sklíčka II.

3) Normalizácia priestorového efektu a rozdielov intenzít medzi kanálmi

- **Centrovanie mediánom**
  - odčíta medián od intenzít všetkých spotov
  - najjednoduchší, ale nie je schopný skorigovať nelinearitu



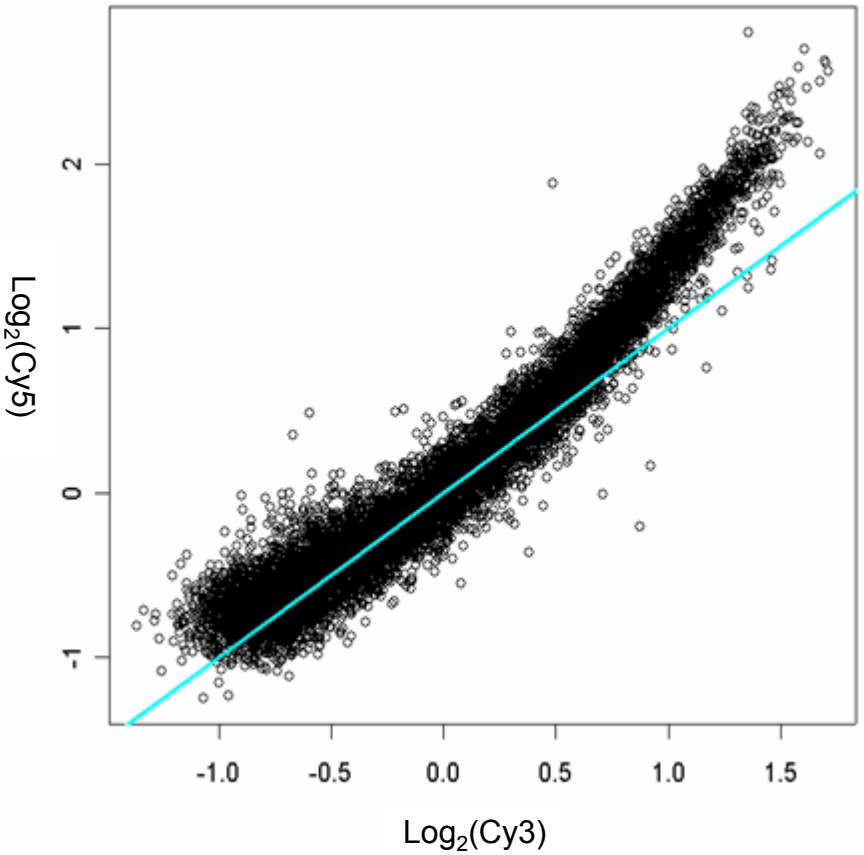
$$M_{norm} = M - l,$$

je medián intenzít všetkých spotov

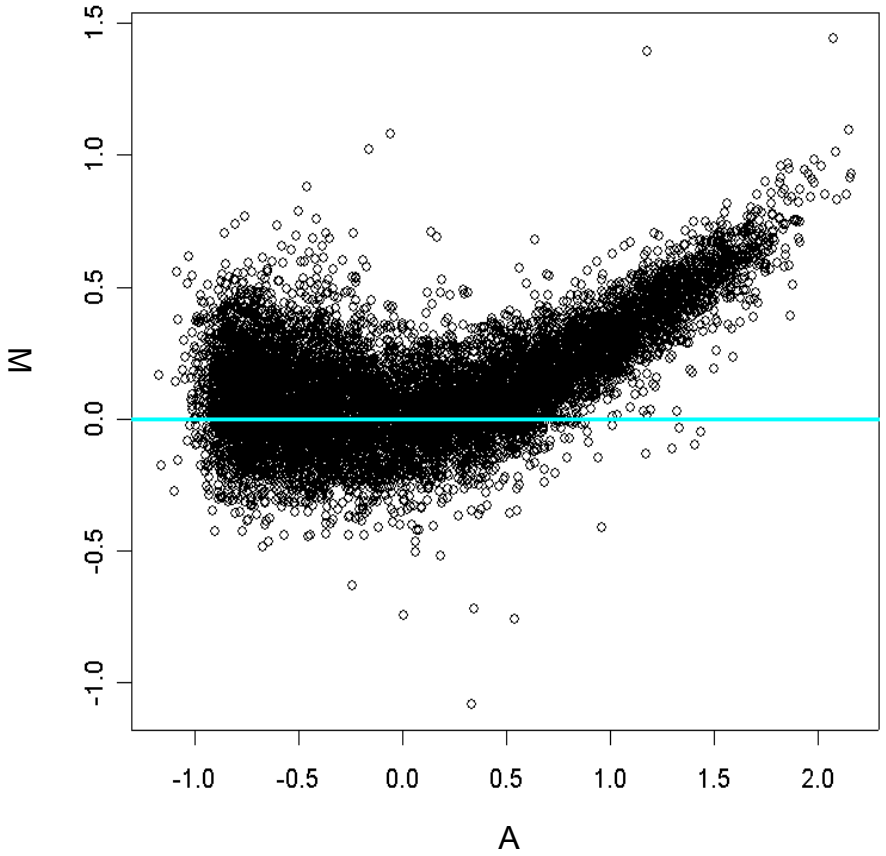
Problémy s mediánovým centrovaním

Jedná sa o globálnu metódu, nie je schopná vyrovnať lokálne efekty, problémy odlišných intenzít, print-tip efekty atd.

Graf intensit kanálů



MA graf



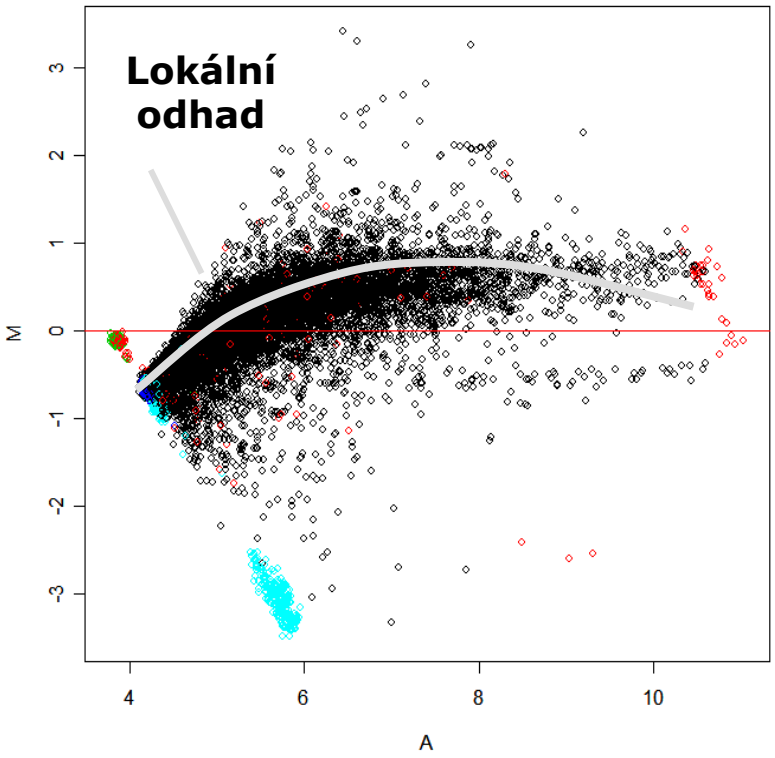
S nelinearitou si vedia poradiť **lokálne regresné metódy (lo(w)ess**)

Lowess normalizácia I

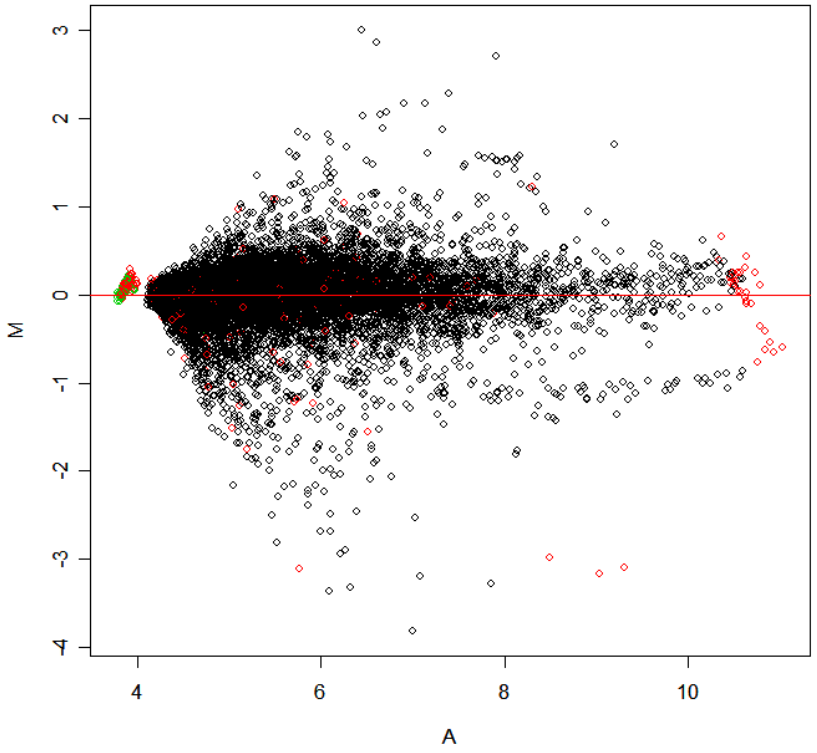
Princíp:
1.Odhad krivky pomocou neparametrickej lokálnej váženej regresie (lowess - locally weighted scatterplot smoothing)
2.Odpočítanie odhadnutej krivky od nameraných hodnôt

Výhoda : nie je nutné poznať funkciu krivky, je odhadovaná z dát!

Před loess normalizací



Po loess normalizaci



Lowess normalizácia II

Princíp lowess

•V každom kroku sa určí lokálna množina dát, na ktorej sa **odhadne krivka s pomocou polynomiálu a metódy najmenších štvorcov**

•Parameter λ určuje stupeň polynomiálu ($\lambda=0$ priemer, $\lambda=1$ lineárna regresia, $\lambda=2$ kvadratická regresia)

•Množina dát na ktorej sa pracuje sa určuje pomocou algoritmu najbližšieho suseda

•Vyhladzovací parameter α určuje veľkosť tejto množiny ($n\alpha$ bodov v okolí odhadovaného bodu)

•α nadobúda hodnoty medzi $(\lambda + 1)/n$ a 1

Normalizácia v rámci microarray sklíčka II.

- Krivky odhadujeme:
  - na základe signálov **všetkých sond na mikročipe**

Predpoklad: expresia väčsiny génov, ktoré sondy predstavujú, nie je zmenená medzi porovnávanými skupinami! (závisí od mikročipu a od testovanej hypotézy)

  - na základe signálu **skupiny sond:**

i) skupina sond by mala mať približne rovnakú expresiu vo všetkých vzorkách (aby sme neodstránili reálne biologické rozdiely)

ii) množina by mala byť dostatočne veľká, aby zachytila variabilitu sklíčka

Napr. housekeeping geny

Příklad IV – normalizace uvnitř čipu

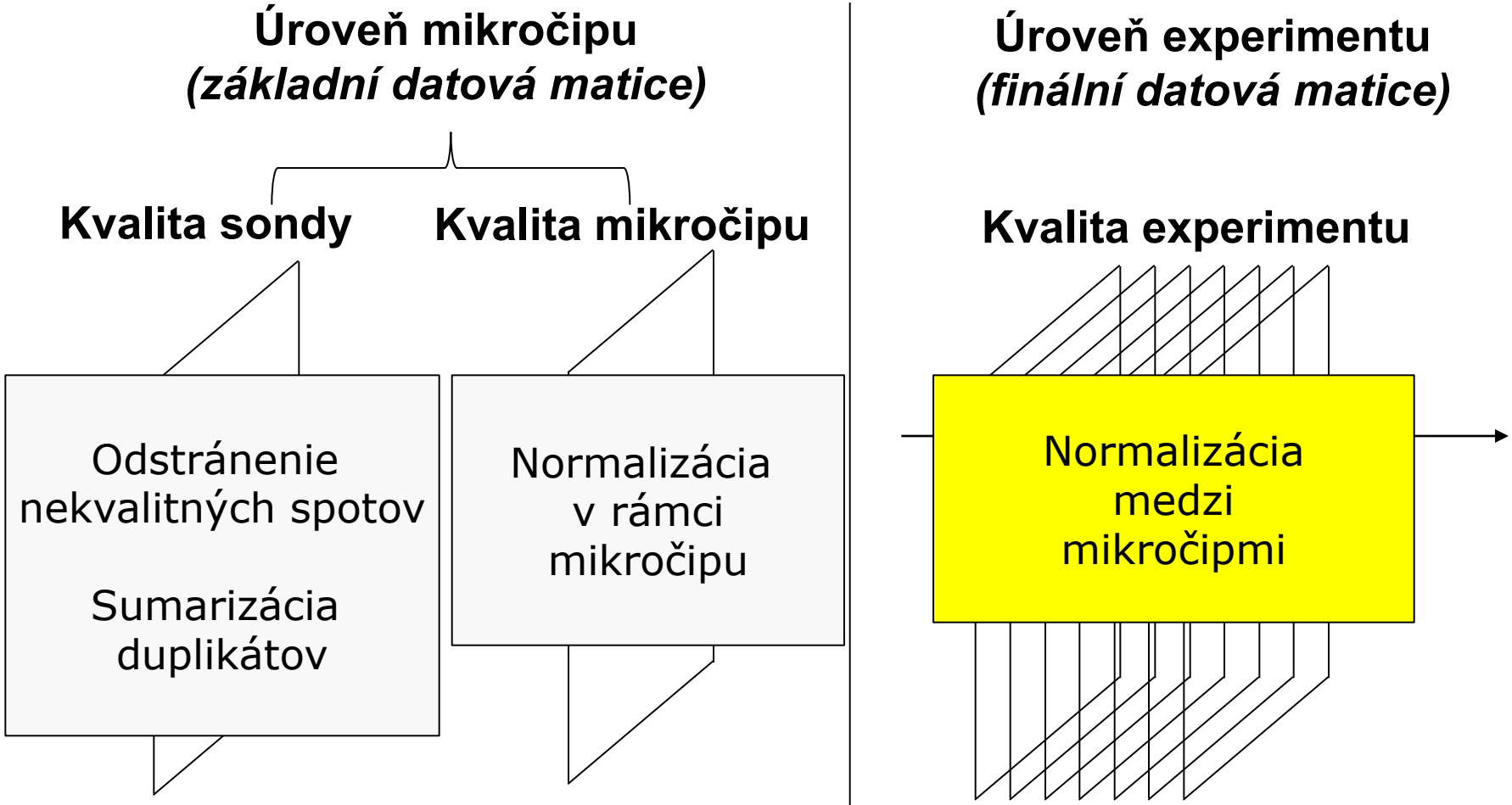
- Aplikujme centrování mediánem na *M* hodnoty prvního mikročipu z příkladu a skontrolujme, jak se normalizace (ne)poprala s nelineárními efekty:

```
> 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)
```

Úrovne úprav dátových súborov



Úroveň sondy: Kvalita jedného spotu na mikročipe
Úroveň mikročipu: Kvalita celého mikročipu
Úroveň experimentu: Kvalita měření transkriptů všech mikročipů v experimentu

Normalizácia medzi sklíčkami I

- Keď sú všetky datové matice mikročipov znormalizované, tak vytvárame **finálnu dátovú maticu**, ktorý použijeme pre následnú analýzu

  riadky ~ vzorky, stĺpce ~ gény

- Jednotlivé súbory musíme normalizovať navzájom, aby sme odstránili efekty medzi sklíčkami, spôsobené rozličnou hybridizáciou, rozličným množstvom vzorky (mRNA), rozličným efektom skenovania, chybami v segmentácii... apod.

- Princíp – zjednotenie rozloženia (priemer, smerodatná odchýlka, prípadne kvantily)

Normalizácia medzi sklíčkami II

- **Globálne centrovanie**

Nastaví priemer a škálu všetkých sklíčok na jednu hodnotu (medián, priemer, orezaný priemer... všetkých čipov alebo hodnoty referenčného čipu)

Nevýhoda: predpokladá, že rozdiely sú len posunové, lineárne

- **Škálovanie**

Táto metoda zjednocuje variabilitu jednotlivých mikročipov, napríklad podelením hodnôt mediánovou absolutnou odchýlkou ich intenzít. Obvykle sa kombinuje s centrovaním.

- **Loess**

Prebieha cyklickým spôsobom – vždy medzi pármi mikročipov až do konvergencie. Takisto je možné vybrať množinu sond na ktorých sa spraví odhad loess krivky

Normalizácia medzi sklíčkami III

- **Kvantilová normalizácia**

Je založená na poradí pozorovaní, a teda neparametrická. Buď na skupine všetkých sond, alebo len na skupine vybraných sond.

Princíp: U každého mikročipu sa zoradia hodnoty expresie a potom sa nahradia priemernou hodnotou kvantilu, ktorý predstavujú v celom sklíčku

hodnoty

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

→

poradie

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

→

zoradené

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

priemer

Gén	
A	(2 1 3)/3 = 2.00 = poradie i
B	(3 2 4)/3 = 3.00 = poradie ii
C	(4 4 6)/3 = 4.67 = poradie iii
D	(5 4 8)/3 = 5.67 = poradie iv

→

normalizované hodnoty

Gén	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)
```

Zhrnutie

- Základné dáta nie sú mRNA koncentrácie
- Musíme skontrolovať kvalitu dát na rôznych úrovniach
  - Úroveň sondy
  - Úroveň sklíčka (všetky sondy na sklíčku)
  - Úroveň génu (gén medzi sklíčkami)
- Vždy transformujte svoje dáta *logaritmom*
- Normalizujte dáta aby ste odstránili systematické (technické) chyby