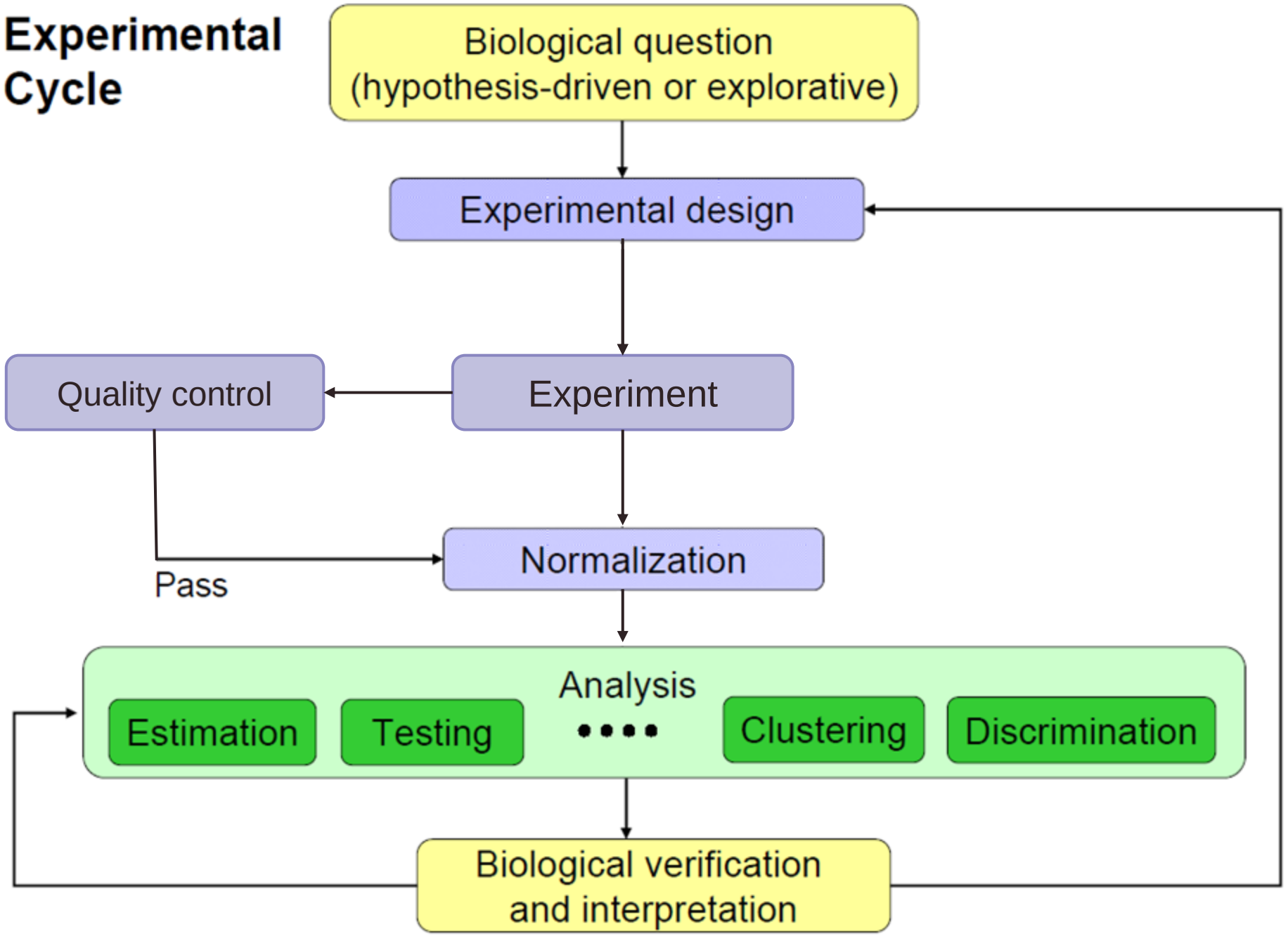
KONTROLA JAKOŚCI DANYCH NORMALIZACJA

Na poprzednim wykładzie … skrót


Experimental Cycle
Biological question (hypothesis-driven or explorative)
Experimental design
Experiment
Quality control
Pass
Normalization
Analysis
Estimation
Testing
Clustering
Discrimination
Biological verification and interpretation

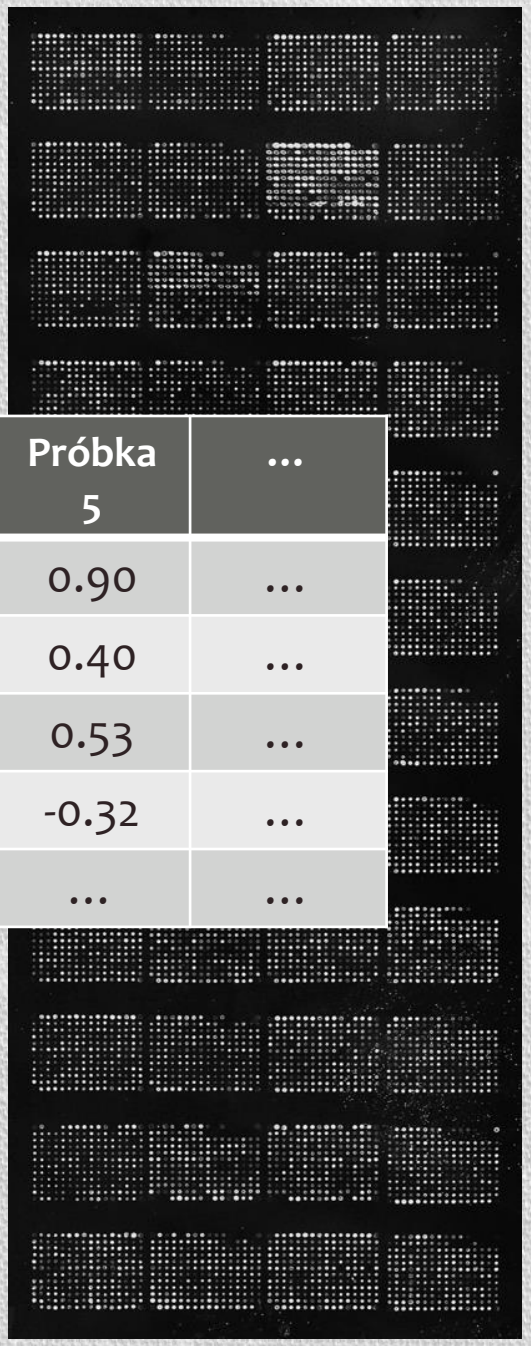
Wstępna analiza mikromacierzowa

- **Przetworzenie danych eksperymentalnych do wartości liczbowych**

  kropka/gen -> liczba

- Wynikiem przetworzenia danych jest **macierz ekspresji genów**, która jest reprezentowana przez macierz utworzoną z *n* wierszy, każdy odpowiadający jednemu genowi, lub punktowi na mikromacierzy, oraz *m* kolumn, każda odpowiadająca warunkom (np. kolejne punkty czasowe), dla których poziom ekspresji genów był mierzony.

Macierz ekspresji genów



Próbki

	Próbka 1	Próbka 2	Próbka 3	Próbka 4	Próbka 5	…
1	0.25	0.30	0.70	1.53	0.90	…
2	-0.12	0.30	0.45	0.12	0.40	…
3	0.13	0.46	0.75	0.32	0.53	…
4	-0.16	-0.43	-0.65	-0.79	-0.32	…
…	…	…	…	…	…	…

Geny/ sondy

Poziom ekspresji genu lub stosunek, dla genu *i*-tego w *j*-tej próbce mRNA

M= { $\log_2$(red intensity/green intensity)
Funkcja (PM,MM) MAS, dchip lub RMA

A= { ½ $\log_2$(red intensity*green intensity)
Funkcja (PM,MM) MAS, dchip lub RMA

Jak przejść od obrazu do liczb?

- Zidentyfikować pozycję punktów na mikromacierzy
- Dla każdego punktu: zidentyfikować piksele, które należą do punktu
- Dla każdego punktu: zidentyfikować piksele sąsiadujące z punktem, które będą używane do obliczenia obrazu tła
- Wyliczenie numerycznych informacji dla intensywności punktów, intensywności tła i informacji kontrolnych o jakości

Kontrola jakości punktów (genów)

- Źródła błędów
  - błędny wydruk, nierówny rozkład, zanieczyszczenie resztkami, znaczenie sygnału w porównaniu do szumu, słaby pomiar punktów
- Inspekcja „naoczna”
  - Włosy, kurz, zadrapania, bąble powietrzne, ciemniejsze regiony na płytce, regiony rozmyte
- Jakość punktów
  - *Jasność*: stosunek punkt/tło (foreground/background)
  - *Jednorodność*: wariacja intensywności pikseli w punkcie
  - *Morfologia*: kształt, obwód, okrąg
  - *Rozmiar punktu*: liczba pikseli punktu (*foreground*)
- Co robić ze złymi punktami?
  - Ustawić pomiar na NA (brakujące wartości)
  - Używanie wag dla pomiarów, które wskażą jakość dla kolejnych etapów

Co otrzymujemy na wyjściu?

- Średnia sygnału punktu
- Średnia sygnału tła
- Mediana sygnału
- Mediana sygnału tła
- Odchylenie standardowe dla punktu (wyznaczone dla wszystkich pikseli z punktu)
- Odchylenie standardowe dla tła (wyznaczone dla wszystkich pikseli tła)
- Średnica – liczba pikseli w poprzek punktu
- Liczba pikseli w punkcie
- Flaga – 0 – jeśli punkt jest dobry, lub inna wartość jeśli punkt oznaczony jako błędny

KONTROLA JAKOŚCI DANYCH

Przyczyny błędów

Dane zaszumione

Lekki szum

„bias” – błąd, odchylenie

bez błędów

Przyczyny błędów

- ilość RNA w biopsji
- wydajność
  - ekstrakcji RNA
  - odwrotnej trankrypcji
  - znakowania
  - fotodetekcji

systematyczne

- podobny efekt dla wielu pomiarów
- poprawki mogą być estymowane z danych

normalizacja

- wynik PCR
- jakość DNA
- wydajność znakowania, rozmiar punktu
- niespecyficzna hybrydyzacja
- błędny (zabłąkany) sygnał

stochastyczne

- efekt dla pojedynczego punktu
- błędy losowe, które nie mogą być estymowane z danych

model błędów

Kontrola jakości danych

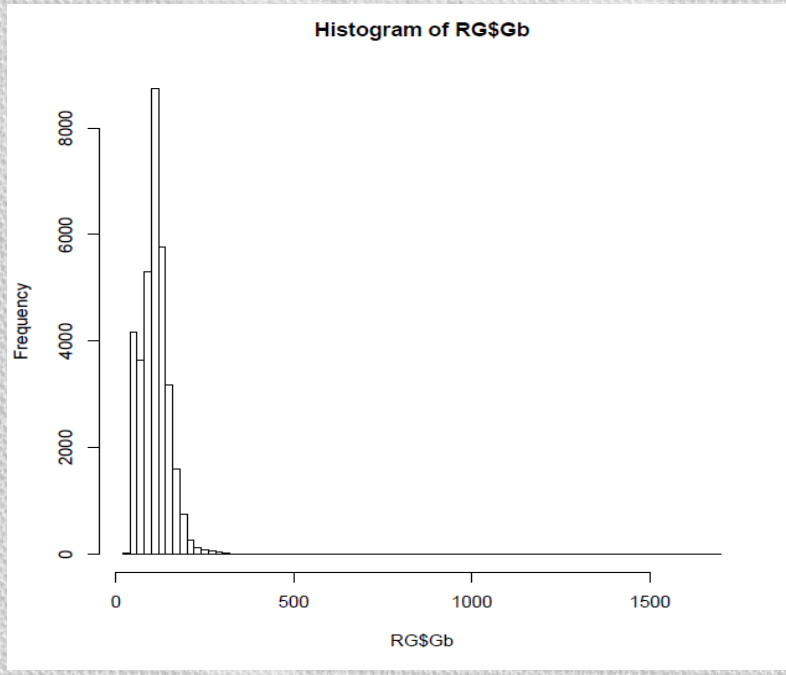
- Histogram
- Przestrzenny rozkład intensywności kolorów
- Boxplot
- Scatterplot
- MA plot

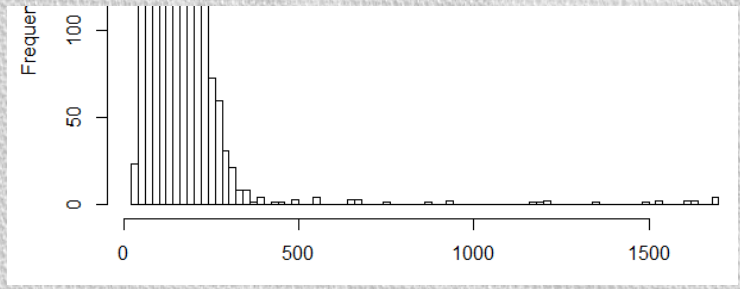
Zestaw danych *swirl zebrafish*

SlideNumber	FileName	Cy3	Cy5	Date
81	swirl.1.spot	swirl	wild type	20/09/2001
82	swirl.2.spot	wild type	swirl	20/09/2001
93	swirl.3.spot	swirl	wild type	8/11/2001
94	swirl.4.spot	wild type	swirl	8/11/2001

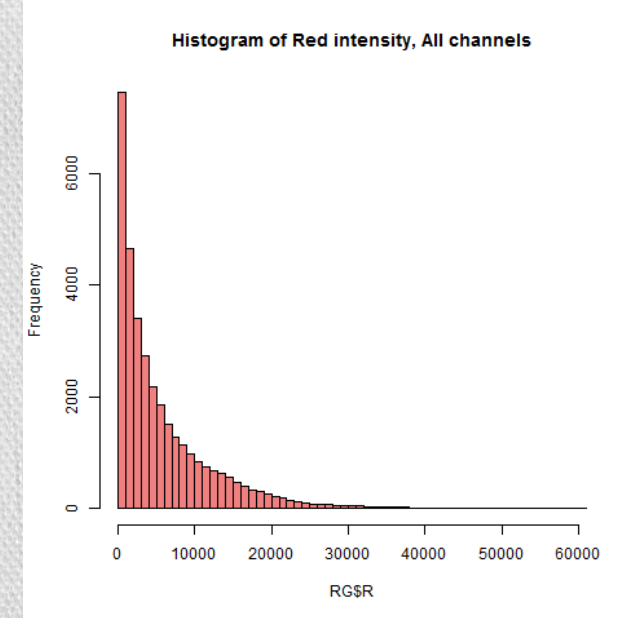
Histogram

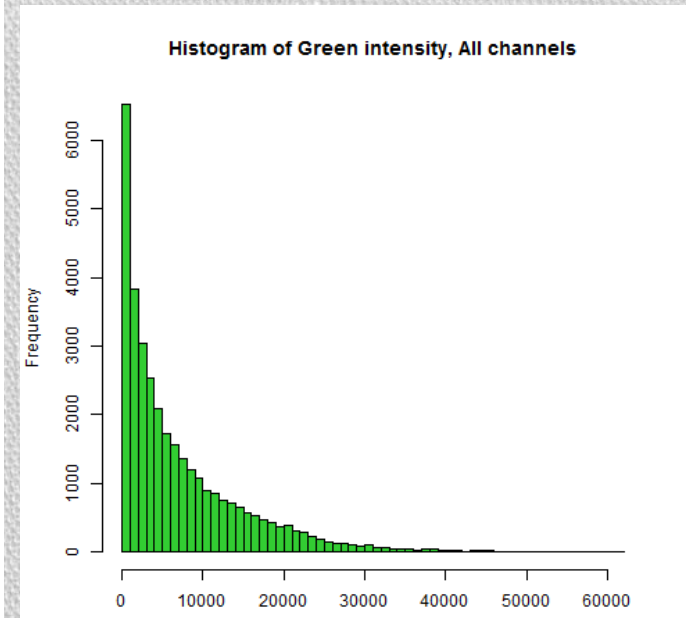
- Histogram – przedstawienie rozkładu intensywności sygnałów genów dla każdej próbki oddzielnie
- Zazwyczaj obserwuje się **unimodalną funkcję rozkładu** (z jednym ekstremum)
- Obecność wielu szczytów na histogramie oznacza zazwyczaj artefakt eksperymentalny
- Większość genów ma słabą intensywność, co oznacza iż geny te nie uległy, bądź też uległy słabej ekspresji (stąd też duży skok z lewej strony wykresu)
- „Długi ogon” z prawej pokazuje geny, które uległy ekspresji na różnych poziomach

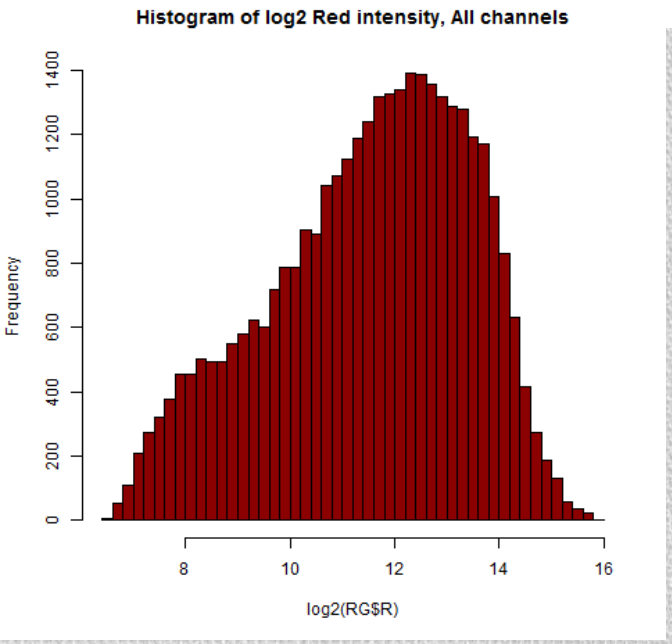


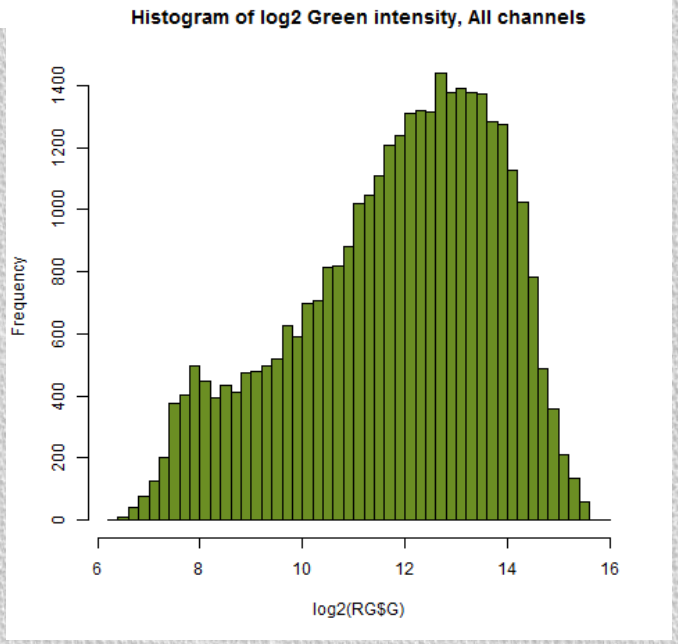


Histogram

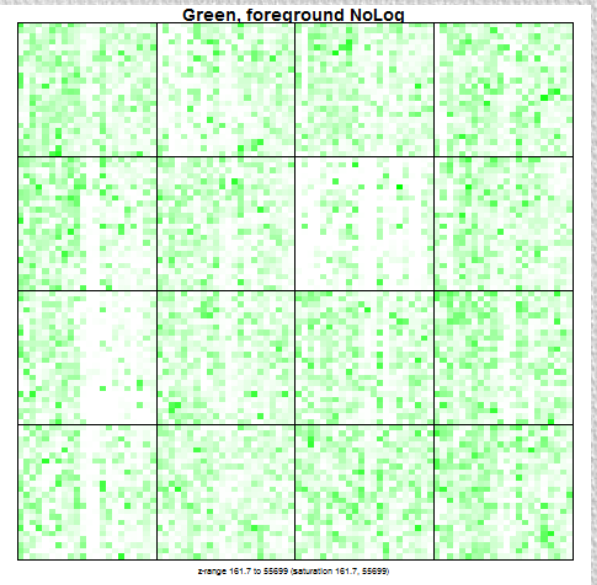

Histogram of Red intensity, All channels
Frequency
0
2000
4000
6000
0
10000
20000
30000
40000
50000
60000
RG$R

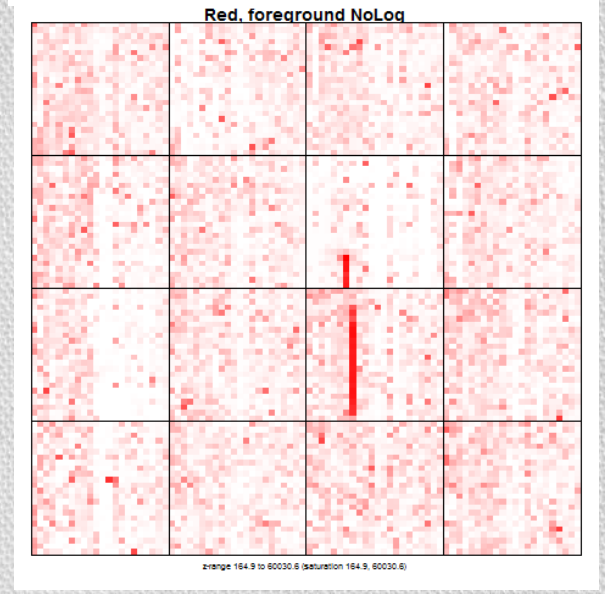

Histogram of Green intensity, All channels
Frequency
0
1000
2000
3000
4000
5000
6000
0
10000
20000
30000
40000
50000
60000


Histogram of log2 Red intensity, All channels
Frequency
0
200
400
600
800
1000
1200
1400
8
10
12
14
16
log2(RG$R)


Histogram of log2 Green intensity, All channels
Frequency
0
200
400
600
800
1000
1200
1400
6
8
10
12
14
16
log2(RG$G)

Przestrzenny rozkład intensywności kolorów punktów i tła





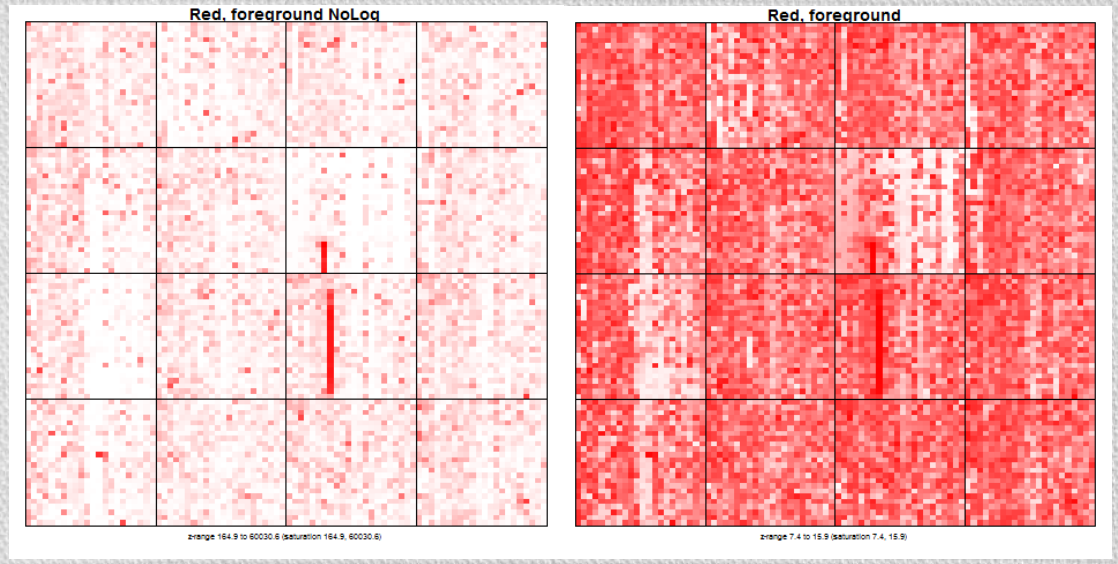
Na obrazkach widać rozkład intensywności kolorów punktów:

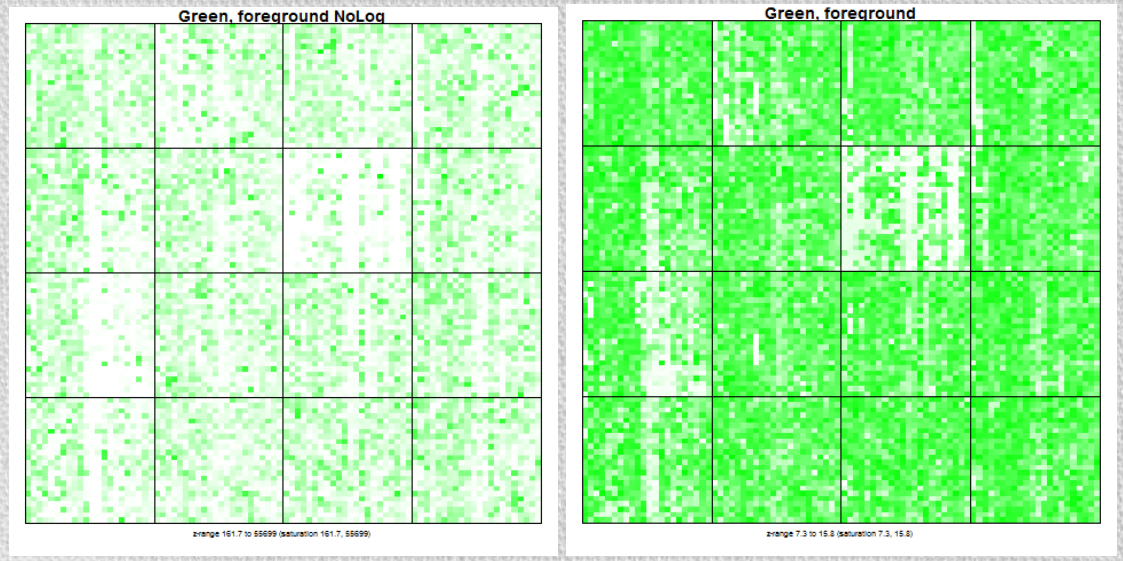
- jest kilka kropek bardzo intensywnych
- jest bardzo dużo kropek o niskiej intensywności, lecz ciężko je odróżnić między sobą

Na drugim obrazku (kolor czerwony) widać wyraźnie kreskę – na mikromacierzy najprawdopodobniej była rysa.

Na obrazku pierwszym (zielonym) rysa nie jest widoczna.

Przestrzenny rozkład intensywności kolorów skala logarytmiczna



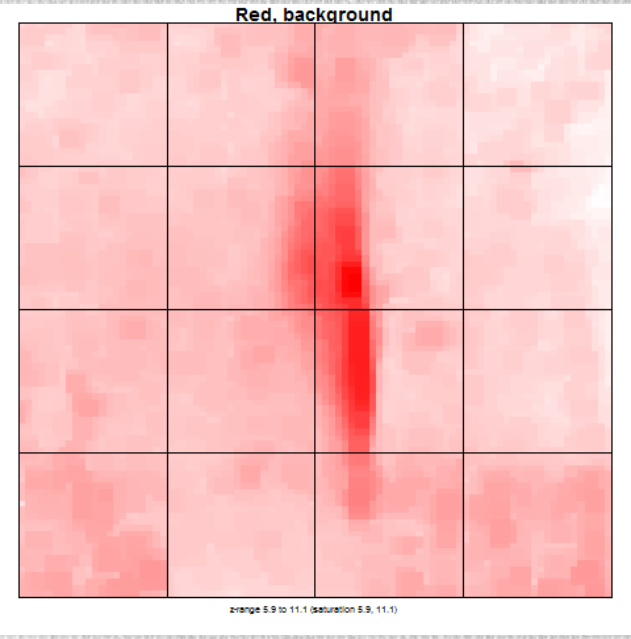


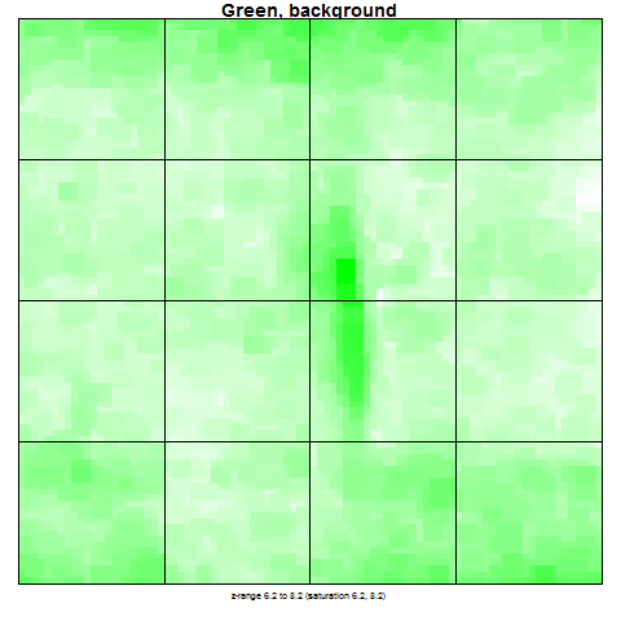
Większość punktów ma dość niską intensywność, która jest rozróżnialna dopiero na skali logarytmicznej

```
>imageplot(RG$G, RG$printer,
low="white",
high="green",main="Green,
foreground")


>imageplot(log2(RG$Gb[,1]),
RG$printer, low="white",
high="green", main="Green,
background")
```

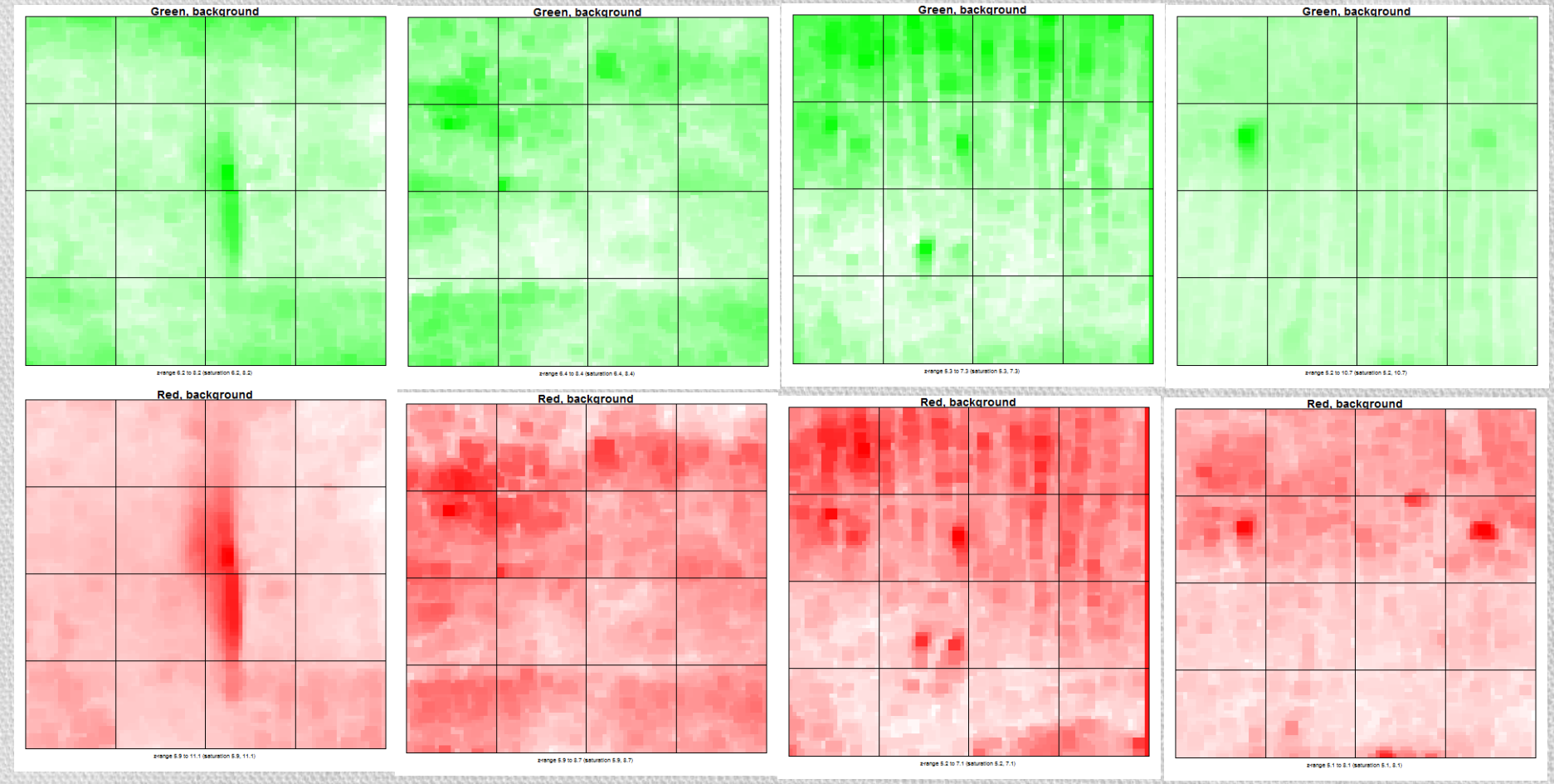
Przestrzenny rozkład intensywności kolorów tła

Rysę widać wyraźnie na obu kanałach mikromacierzy



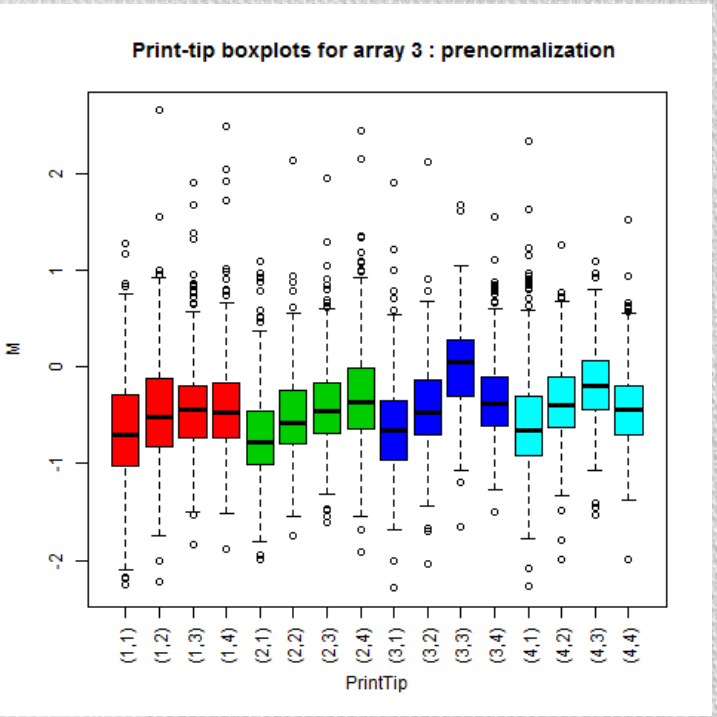


Rozkład intensywności kolorów tła – cały eksperyment



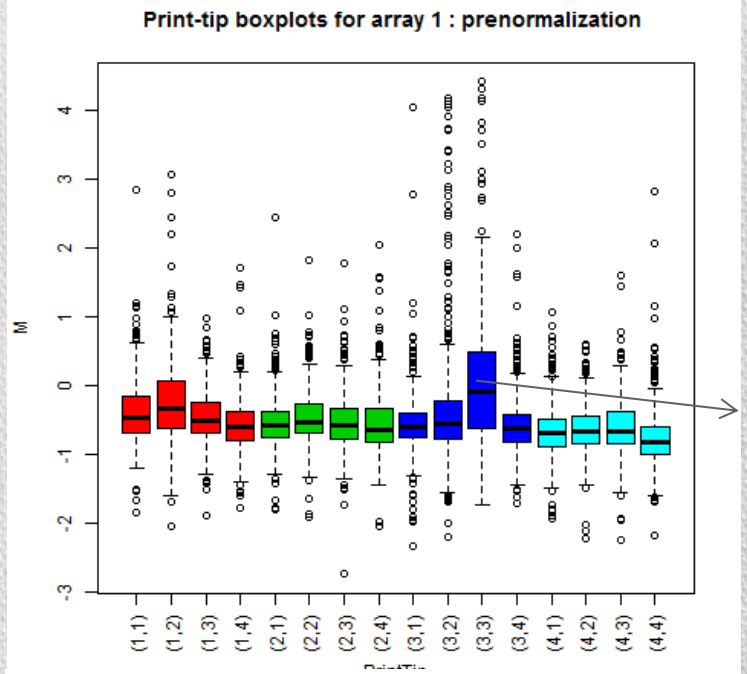
Boxplot

Wykres pudełkowy (boxplot) pozwala zilustrować podstawowe statystyki opisowe w formie charakterystycznych słupków. Pozwala ująć na jednym rysunku wiadomości dotyczące położenia, rozproszenia i kształtu rozkładu badanej zmiennej.

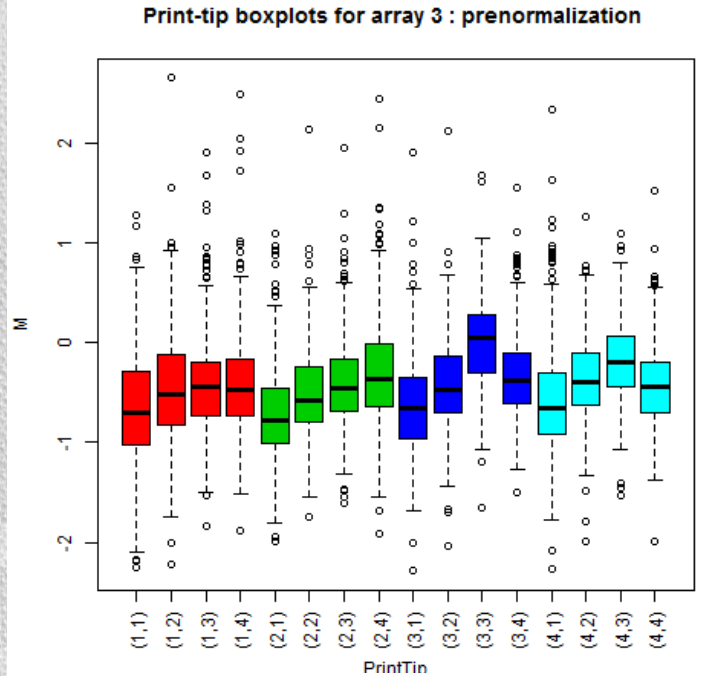


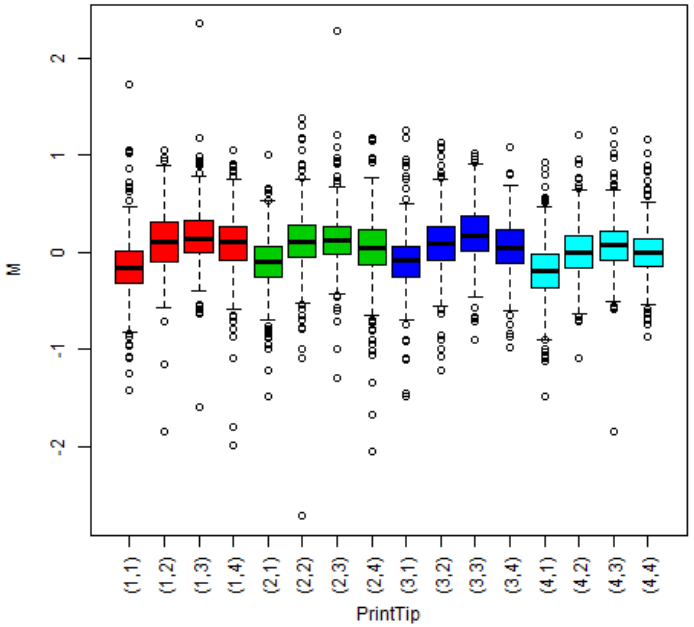

```
>library(convert)

>mraw = as(RGb,"marrayRaw")

>boxplot(mraw[,3],
 xvar="maPrintTip", yvar="maM",
 main="Print-tip boxplots for
 array 3: prenormalization")
```

Konwersja z klasy *RGlist* do *marrayRaw*



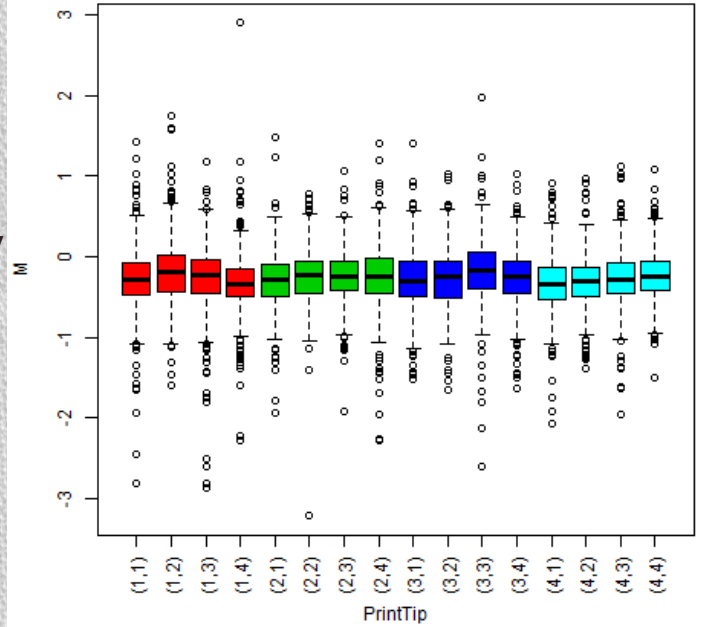
Print-tip jest ewidentnie gorszy przez rysę na płytce



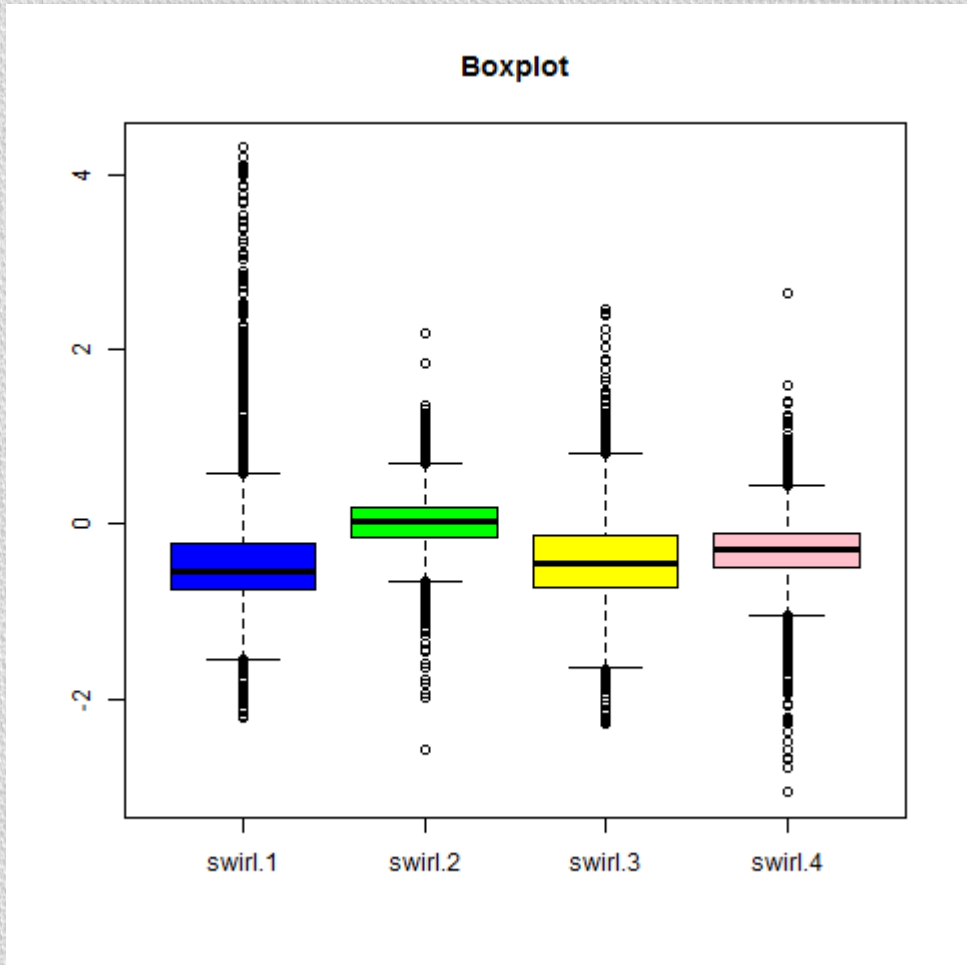


Wykresy 1-3 wyraźnie pokazują potrzebę znormalizowania danych

Na wykresie 4. mediany „pudełek” leżą prawie na jednej linii, lecz wszystkie są poniżej zera



Boxplot dla wszystkich mikromacierzy

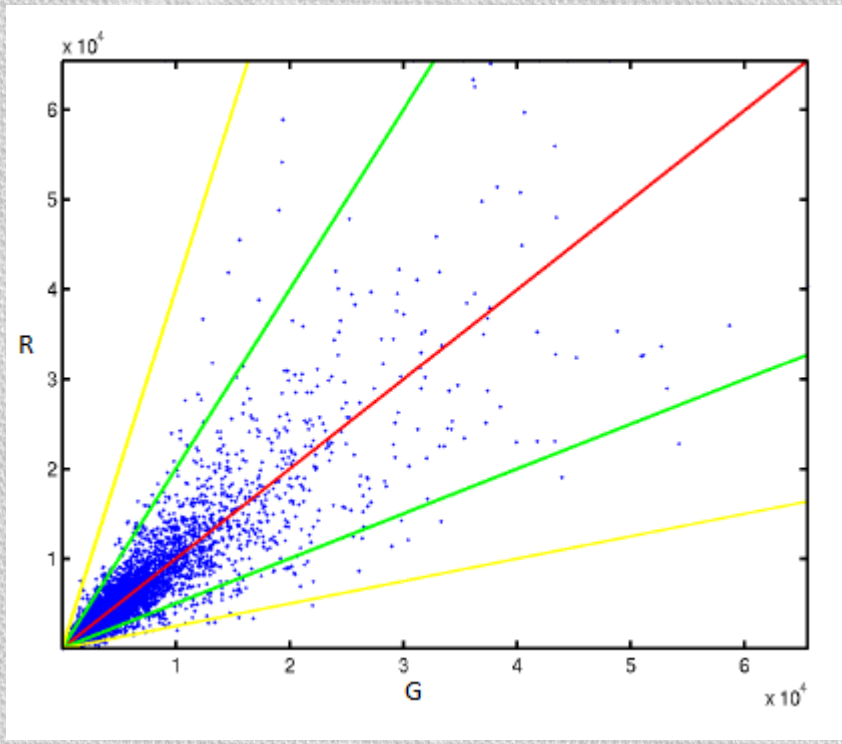



```
>boxplot(log2(RG$R/RG$G),
col=cols, main="Boxplot")
```

Scatterplot

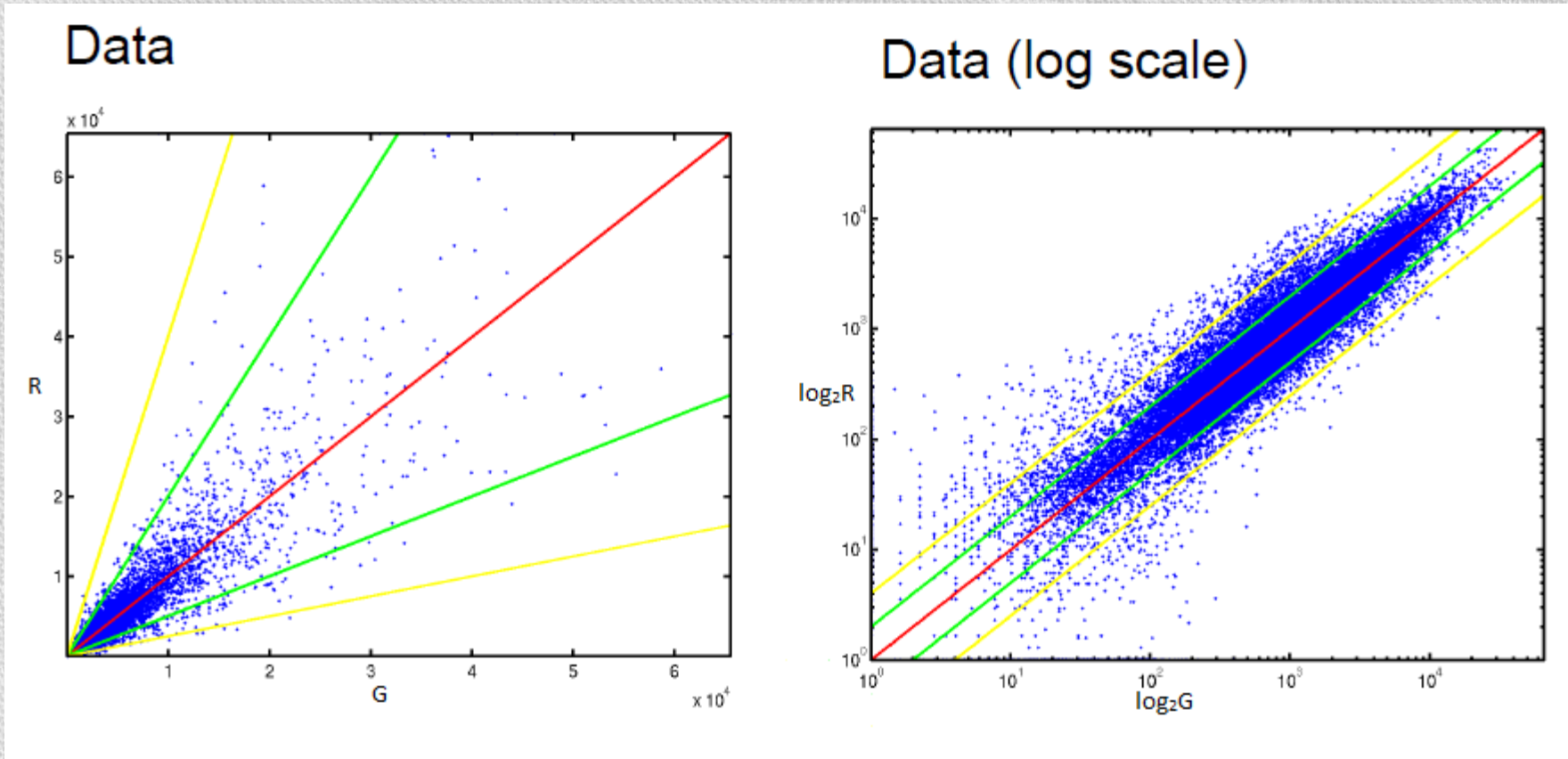
Scatter plot jest to wykres przedstawiający wartości dwóch zmiennych dla zbioru danych. W naszym przypadku będą to zależności intensywności świecenia sond dla pary próbek.

- W idealnym przypadku większość genów, która pozostaje niezmieniona, powinna leżeć na dwusiecznej kąta.
- W rzeczywistości pojawiają się systematyczne lub też przypadkowe odstępstwa. Np. barwnik czerwony świeci mniej intensywnie niż barwnik zielony o współczynnik 0.75, dane nie będą wówczas leżały na dwusiecznej, tylko na linii *y=0.75x*
- Większość punktów jest zgromadzona z lewym dolnym narożniku. Bardziej informatywne są wykresy przeskalowane (log), lub wykresy MAplot



Scatterplot

Jeśli na wykresie damy skalę logarytmiczną, wówczas punkty przesuwają się w górę wykresu wzdłuż dwusiecznej.



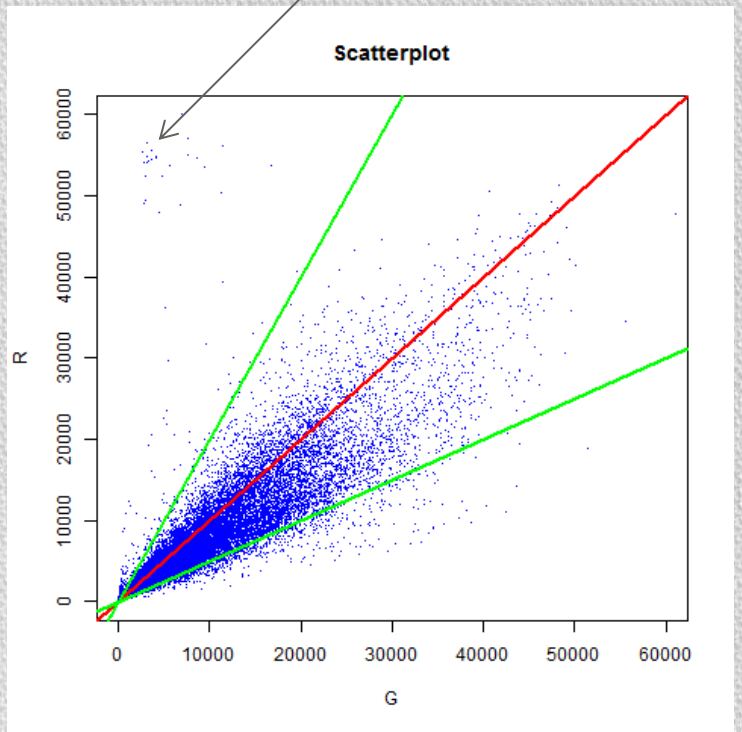
Scatterplot – dla zestawu danych Zebrafish,swirl

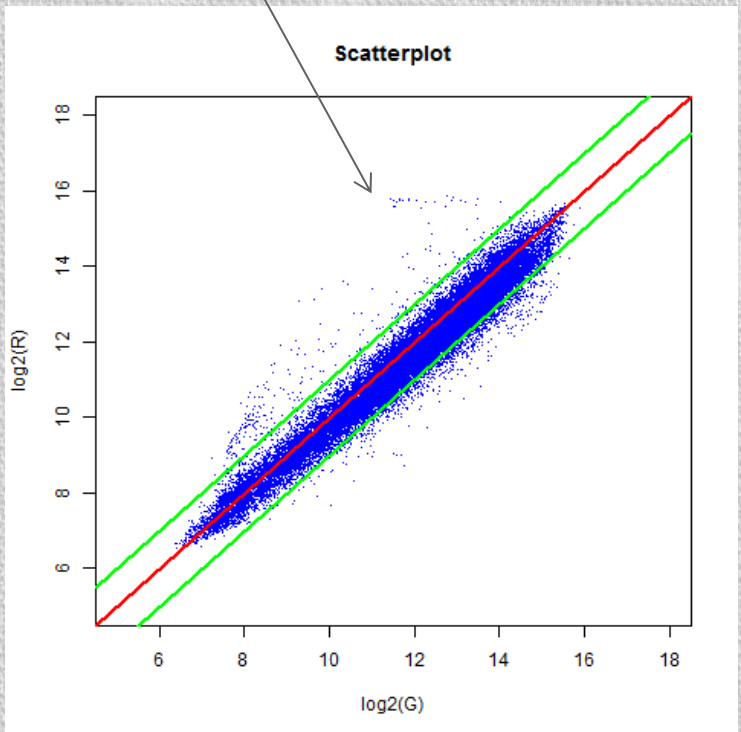
Scatterplot dla naszego zestawu danych

- Punkty są przesunięte nieznacznie por
  zielony jest bardziej intensywny niż cz

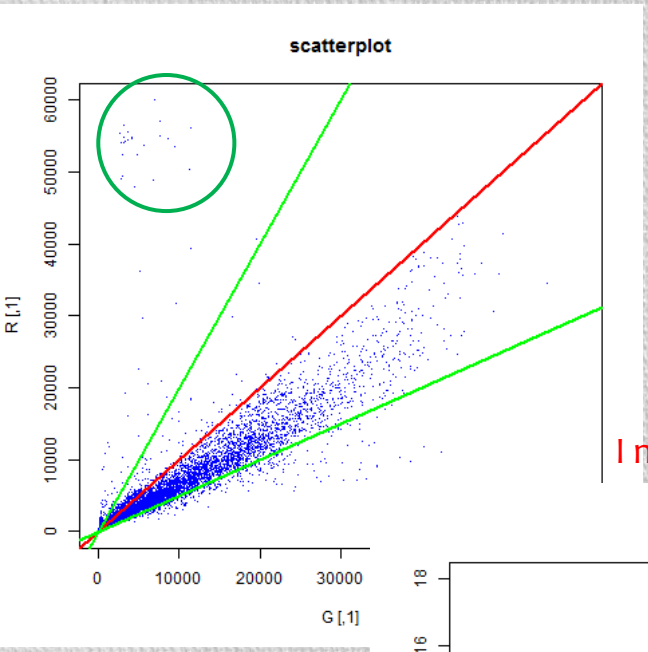
```
>plot(RG$G,RG$R,pch=".",
main="scatterplot", xlab="G", ylab="R",
col="blue", xlim=c(0,60000),
ylim=c(0,60000))
>plot(log2(RG$G),log2(RG$R),pch=".",
main="scatterplot", xlab="log2(G)",
ylab="log2(R)",col="blue", xlim=c(5,18),
ylim=c(5,18))
```

Te punkty są wynikiem rysy

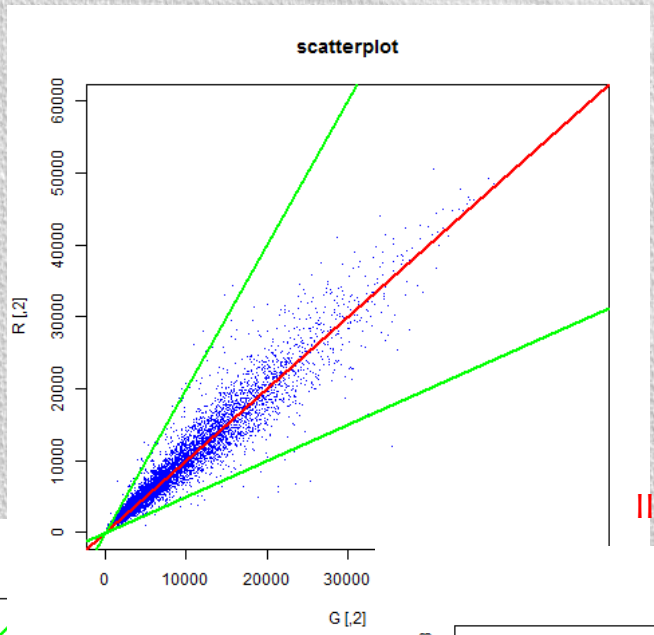




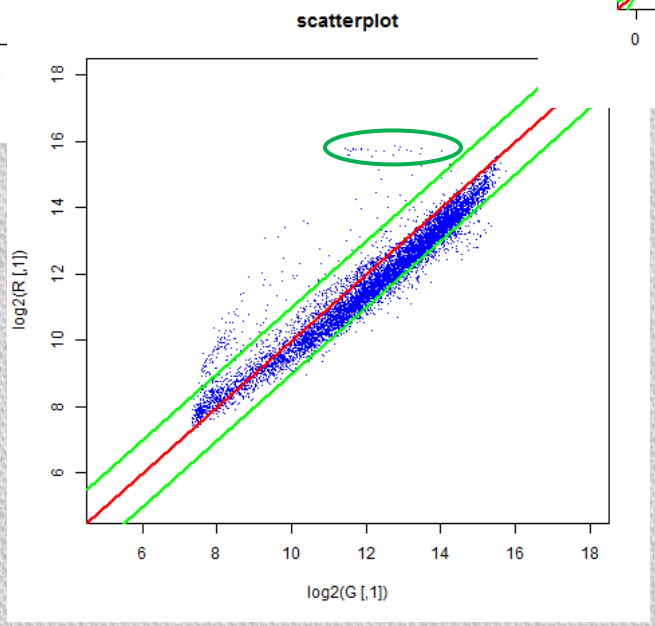
Scatterplot – dla zestawu danych Zebrafish,swirl

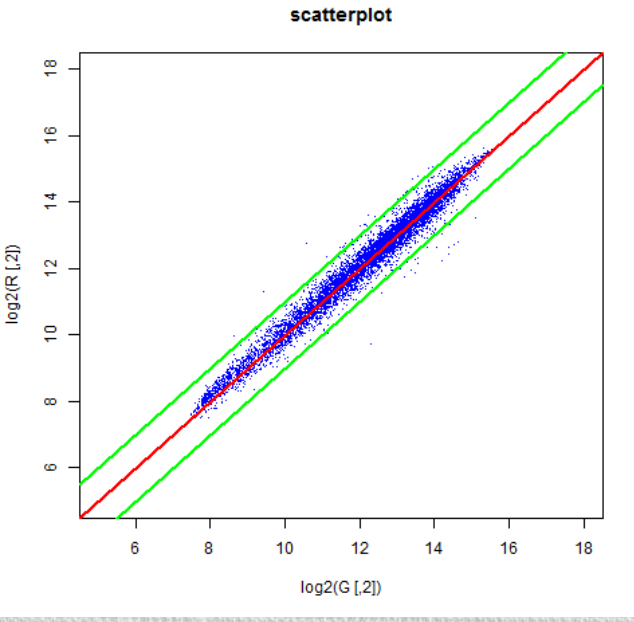


I mikromacierz

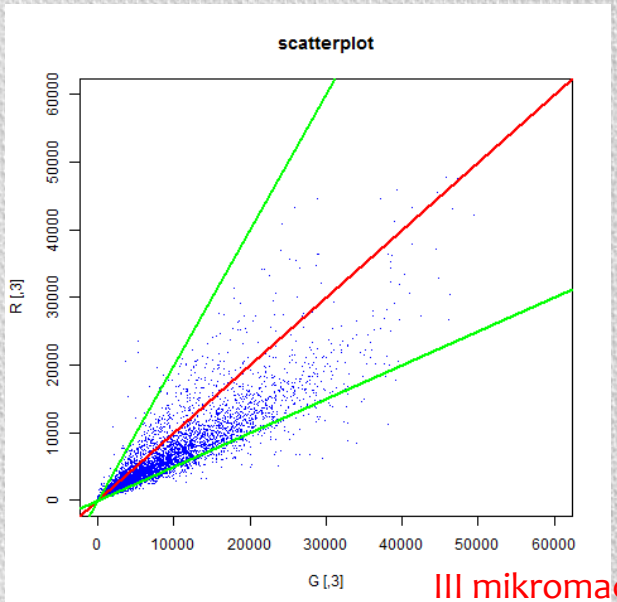


II mikromacierz

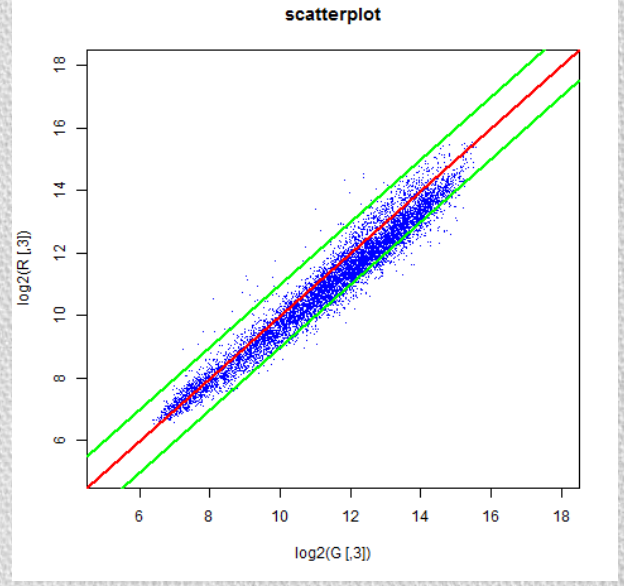


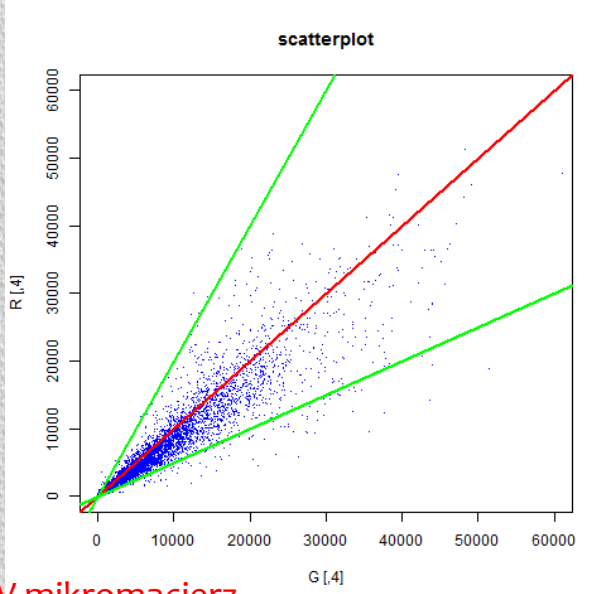


Scatterplot – dla zestawu danych Zebrafish,swirl

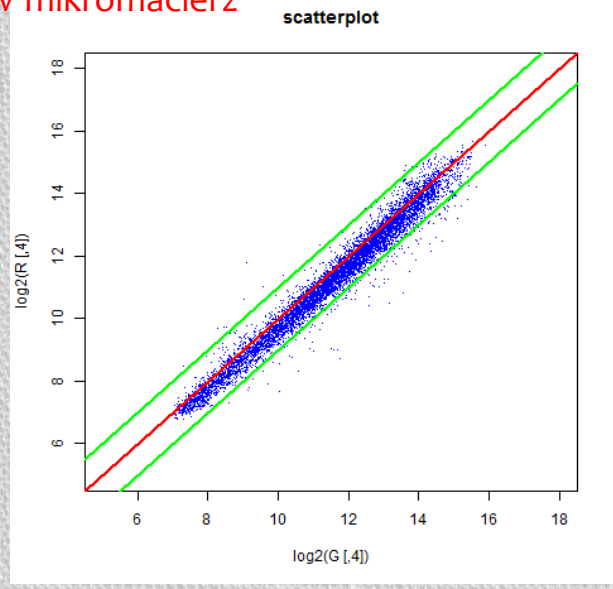


III mikromacierz





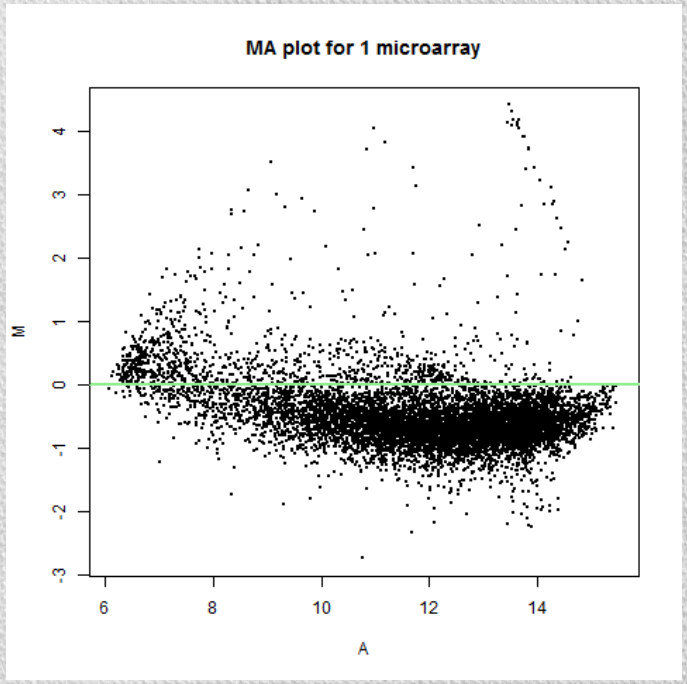
IV mikromacierz

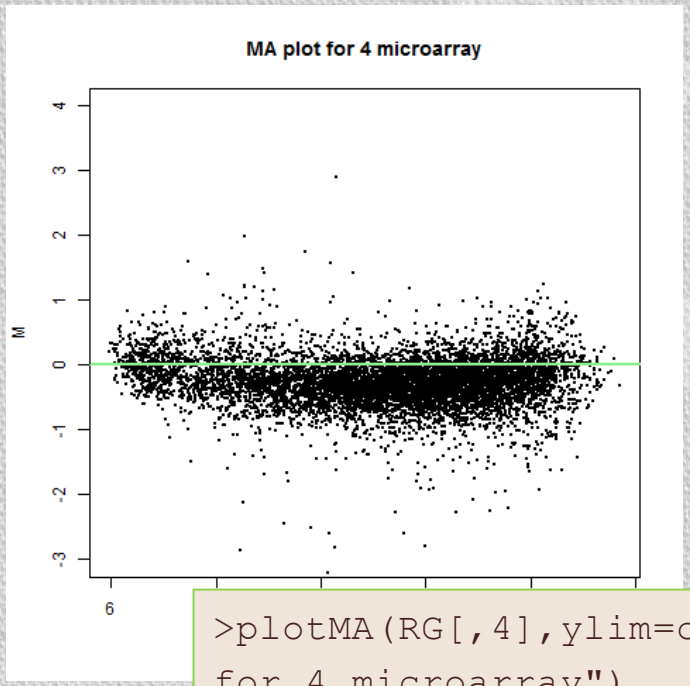


MA plot

$M = \log_2(\text{Red intensity}/\text{Green intensity}) = \log_2(R) - \log_2(G)$ Minus

$A = \log_2\sqrt{(\text{Red intensity} * \text{Green intensity})} = \frac{1}{2}(\log_2(R) + \log_2(G))$ Add





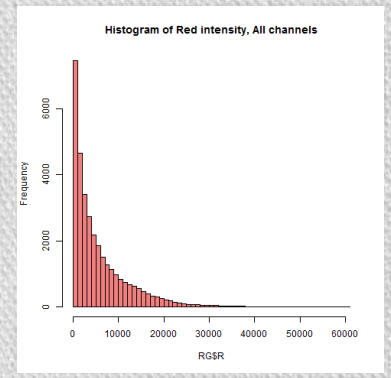
M=0 oznacza brak zmiany w ekspresji genów dla próbek znakowanych kolorami R i G

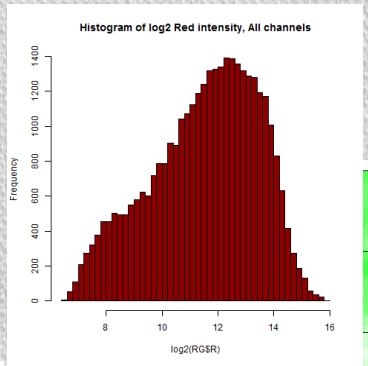
M=1 dla próbki znakowanej kolorem R geny uległy dwukrotnie większej ekspresji niż znakowane kolorem G

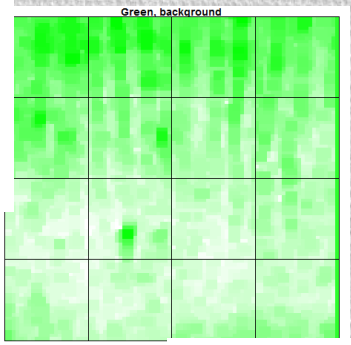
M=-1 geny uległy dwukrotnie niższej ekspresji R i G

```
>plotMA(RG[,4],ylim=c(-3,4),main="MA plot for 4 microarray")
>abline(a=0,b=0,col="lightgreen",lwd=2)
```

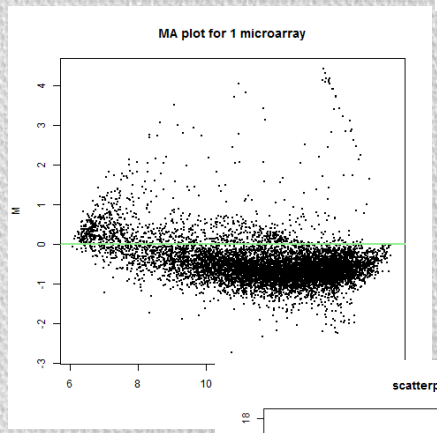
Wyniki eksperymentu obarczone są błędami …

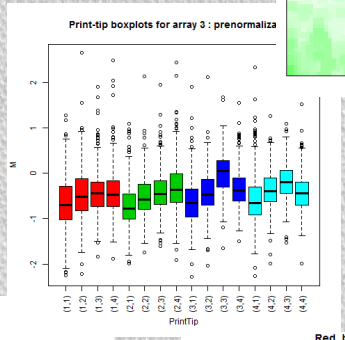


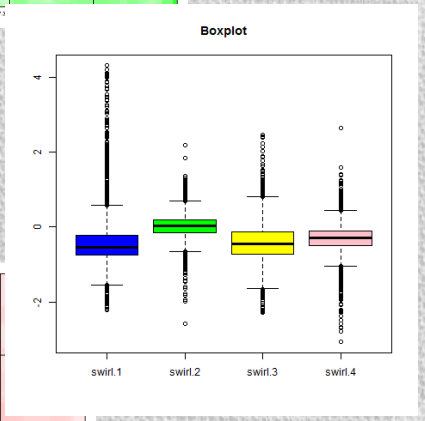


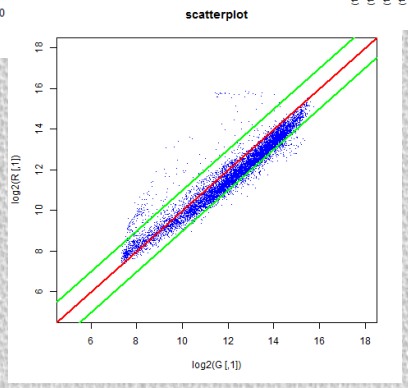


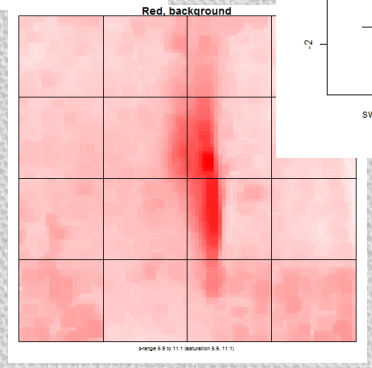
Normalizacja











NORMALIZACJA DANYCH

Normalizacja

Celem normalizacji jest **usunięcie systematycznych błędów** (czyli wynikających z niedoskonałości technologii) **przy zachowaniu informacji biologicznej** i wygenerowanie wartości, które będą mogły być porównane pomiędzy eksperymentami, w szczególności jeśli były wygenerowane w innym czasie, miejscu, na innych mikromacierzach, reagentach.

Rodzaje normalizacji

- globalna – wszystkie geny biorą udział w wyznaczaniu normalizacji w myśl zasady

  większość genów nie uległa zróżnicowanej ekspresji, więc dla większości genów M= 0

- lokalna – w celu wyznaczenia czynnika skalującego (normalizującego) używana jest niewielka pula punktów:
  - *housekeeping genes*: geny o stałej ekspresji, niezależnie od warunków; często nie mogą być brane pod uwagę, gdyż nie reprezentują całej gamy intensywności świecenia
  - *spike controls* – RNA/DNA dodane do wszystkich próbek w równym stopniu, mają swoje odpowiedniki w punktach, do których hybrydyzują; hybrydyzacja powinna być stała dla wszystkich eksperymentów

Rodzaje normalizacji

- within – przeprowadzana jest dla jednego eksperymentu mikromacierzowego, gdy mamy dwie próbki znakowane innymi kolorami
- between – ma na celu znormalizować wyniki ekspresji genów pomiędzy różnymi eksperymentami

Within przeprowadzana jest dla macierzy dwukolorowych (zawsze – trzeba wyrównać różnicę w intensywności kolorów). Jeśli zachodzi taka potrzeba, to przeprowadzana jest również normalizacja *between*.

Between przeprowadzana jest dla macierzy jednokolorowych.

Normalizacja poprzez skalowanie

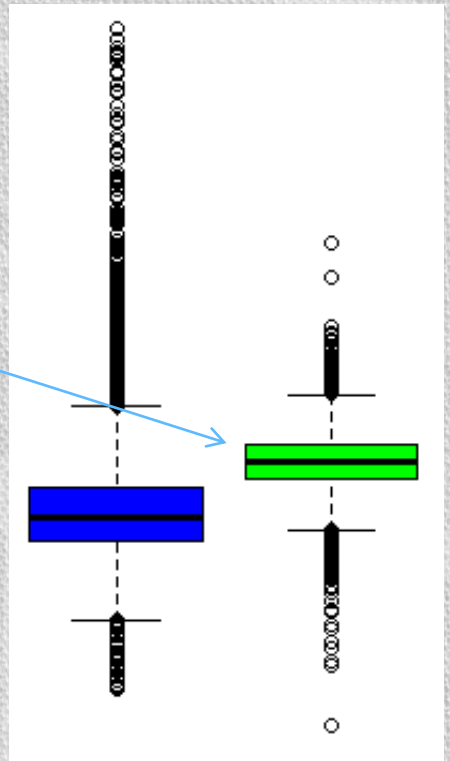
Lokalizacja i skalowanie to podstawowe pojęcia normalizacji

- ***Lokalizacja***

  Poprawia odchylenie przestrzenne lub od barwnika

- ***Skalowanie***

  Ujednolicenie różnorodności pomiędzy macierzami



Znormalizowany stosunek log intensywności

M_{norm}= (M - lokalizacja) / skala

Lokalizacja i skala dla różnych mikromacierzy powinna być (prawie) taka sama

normalizeWithinArrays

```
MAmedian = normalizeWithinArrays (RG, method="median")
```

Klasa „MAlist”

Klasa „RGlist”

Normalizacja *ekspresji stosunku log* dla jednego lub wielu 2-kolorowych eksperymentów mikromacierzowych, w taki sposób, aby średni stosunek log był 0 dla każdej macierzy (lub bloku – print-tip)

- *none* - wyznacza wartości M oraz A, ale nie dokonuje żadnej normalizacji
- *median, loess, printtiploess, composite, control, robustspline*

Globalna normalizacja

Globalna normalizacja zakłada iż intensywności czerwona i zielona są powiązane między sobą przez pewien **stały** czynnik *k*

$$R = kG$$

Środek dystrybucji wartości M (czyli $\log_2(R/G)$) powinien być przesunięty w kierunku 0

$$\log_2 R/G - c = \log_2\{R/(kG)\}$$

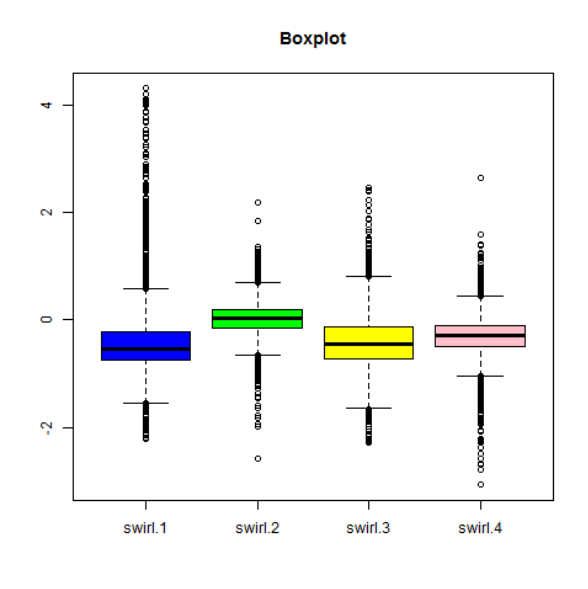
$$c = \log_2 k$$

Często czynnikiem skalującym c jest mediana z wartości M.

Metoda ***median*** (*normalizeWithinArrays*)

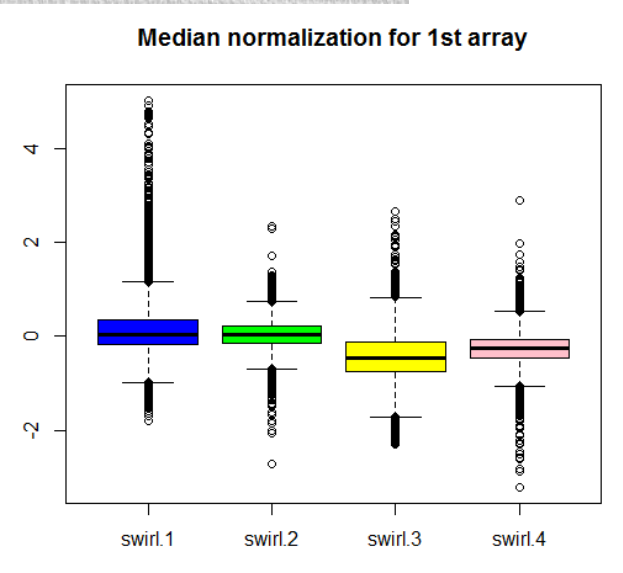
median: normalizacja dla pierwszej mikromacierzy

```
>RGmedian=RG
>for (i in 1:4){
+ m=median(MA$M[,i])
+ s=2^m
+ r=1/s #czynnik skalujacy R
+ RGmedian$R[,i]=RG$R[,i]*r
+ }
```

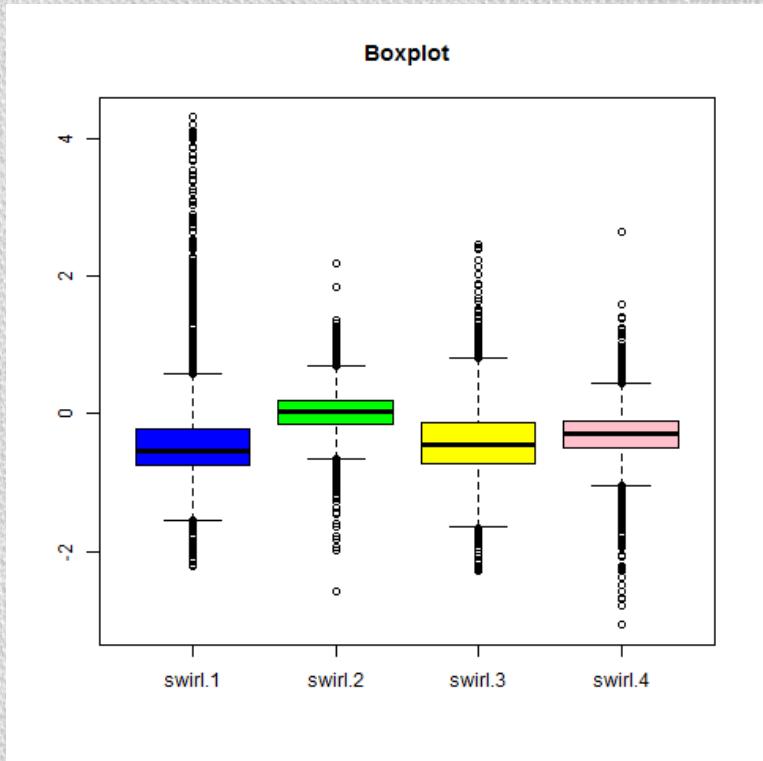


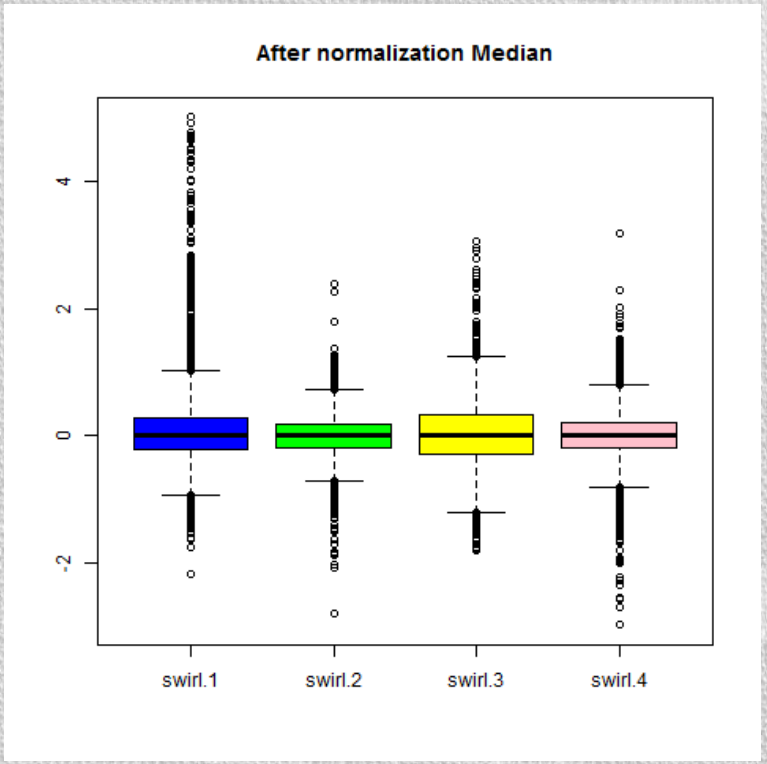
dla i=1

```
> m
[1] -0.5824334
> s
[1] 0.6678364
> r
[1] 1.497373
```



median: boxplot dla wszystkich mikromacierzy

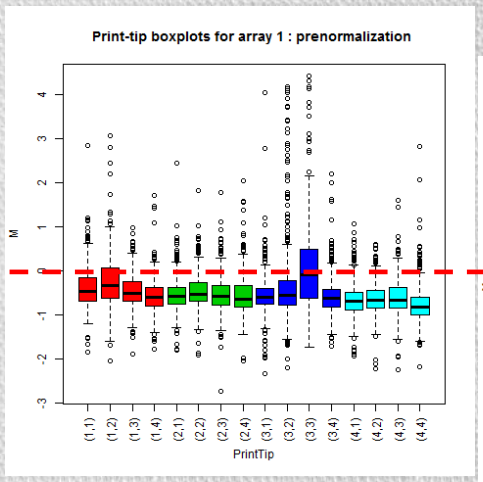


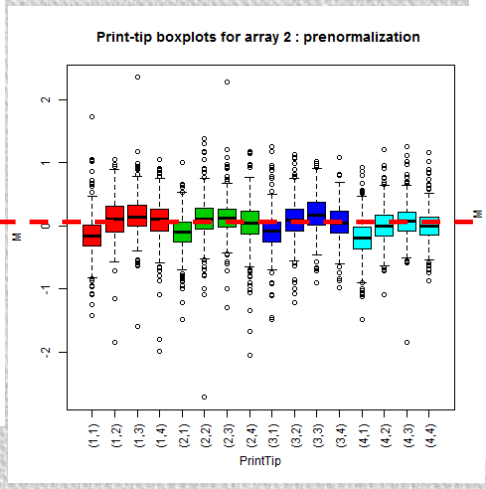


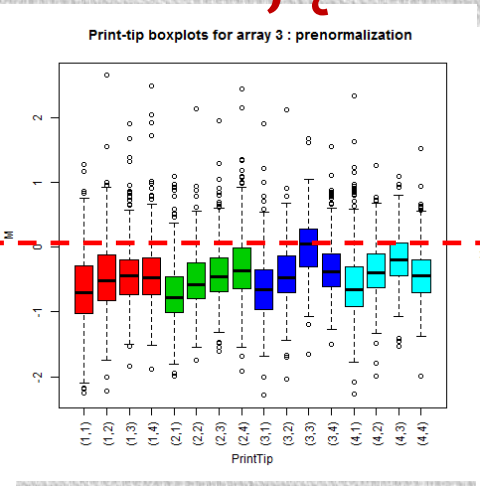
Mediany zostały wycentrowane ☺

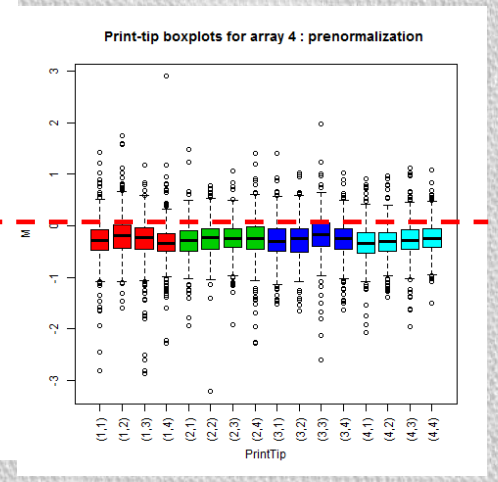
median: boxplot z rozbiciem na bloki (print-tip)

Przed normalizacją

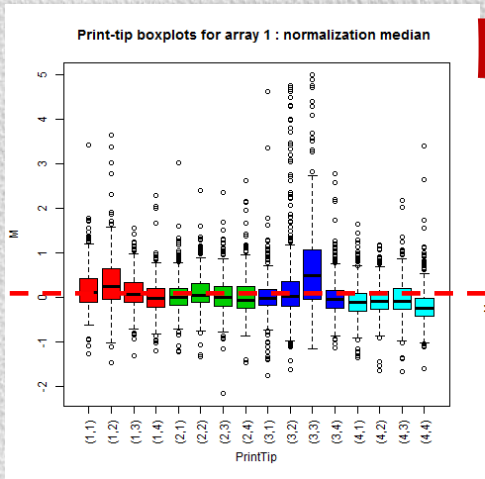


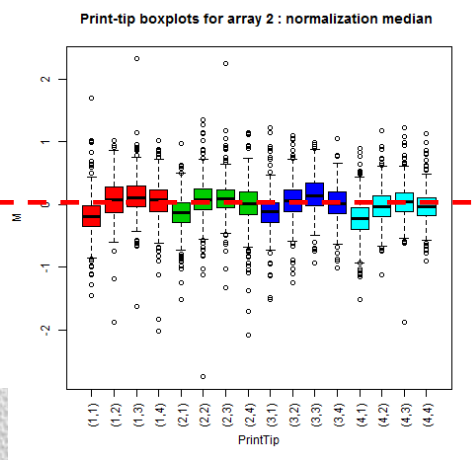


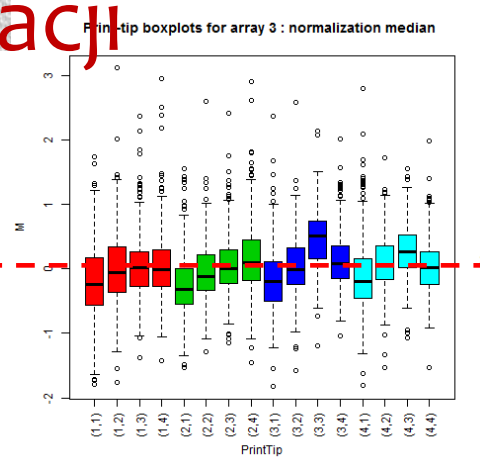


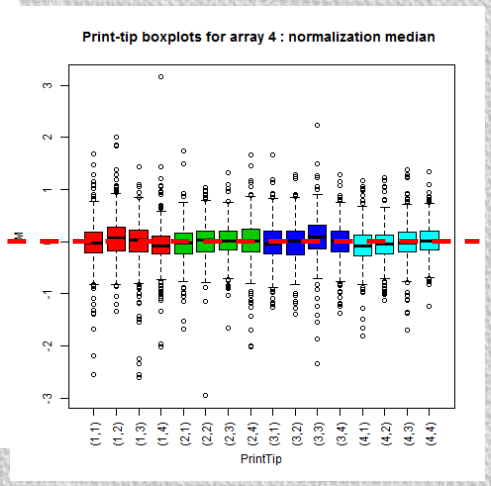


Po normalizacji

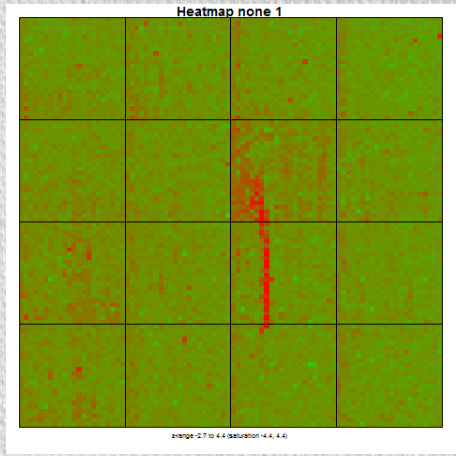


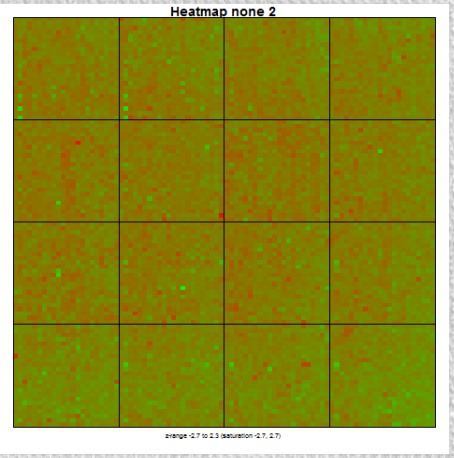


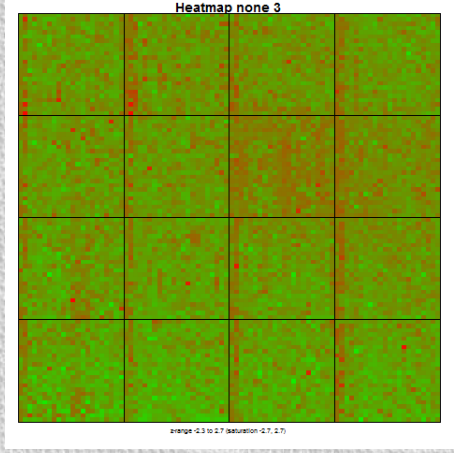


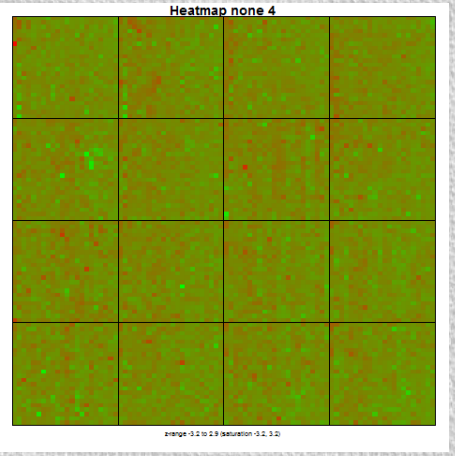


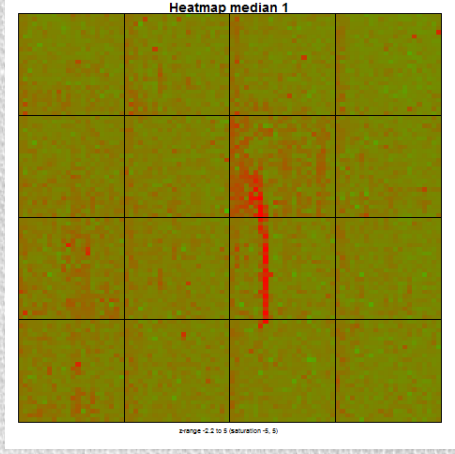
median : wyrównanie sygnału M na mikromacierzy

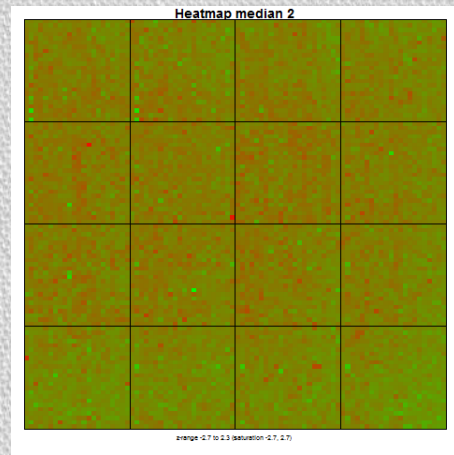


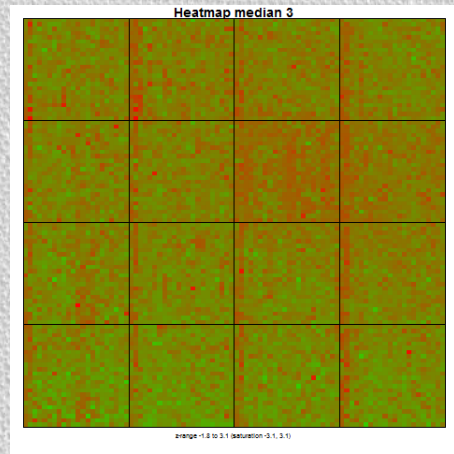


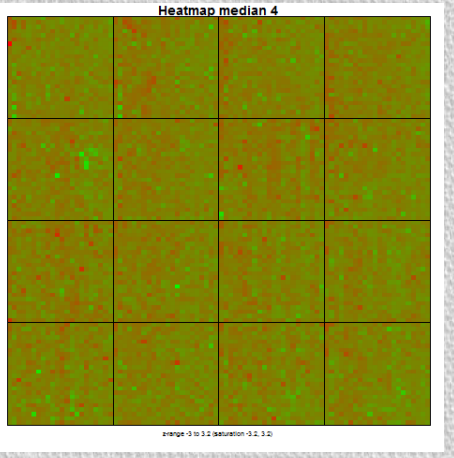




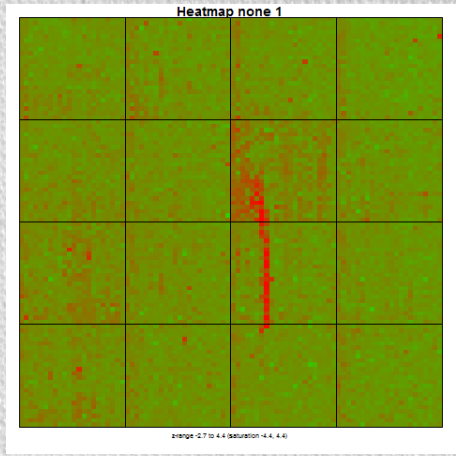


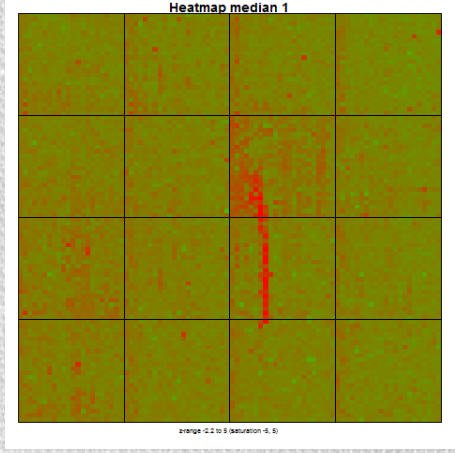






median : wyrównanie sygnału M na mikromacierzy

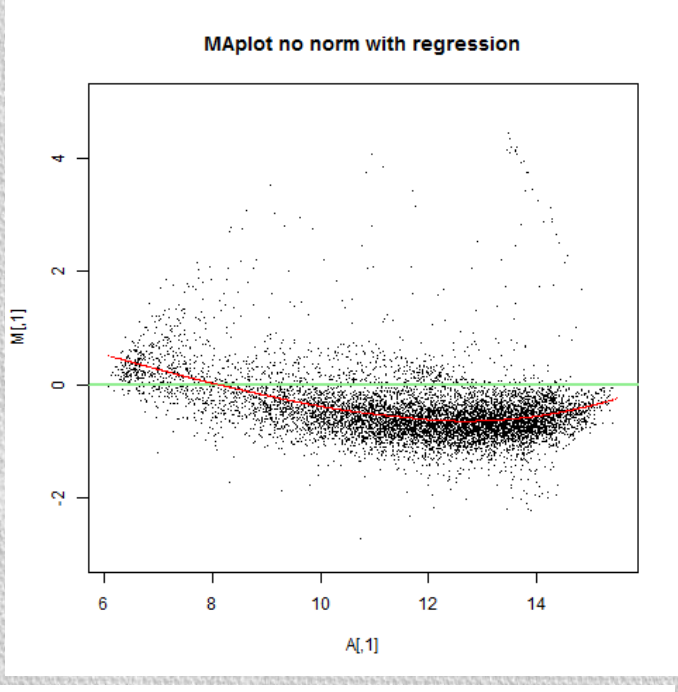




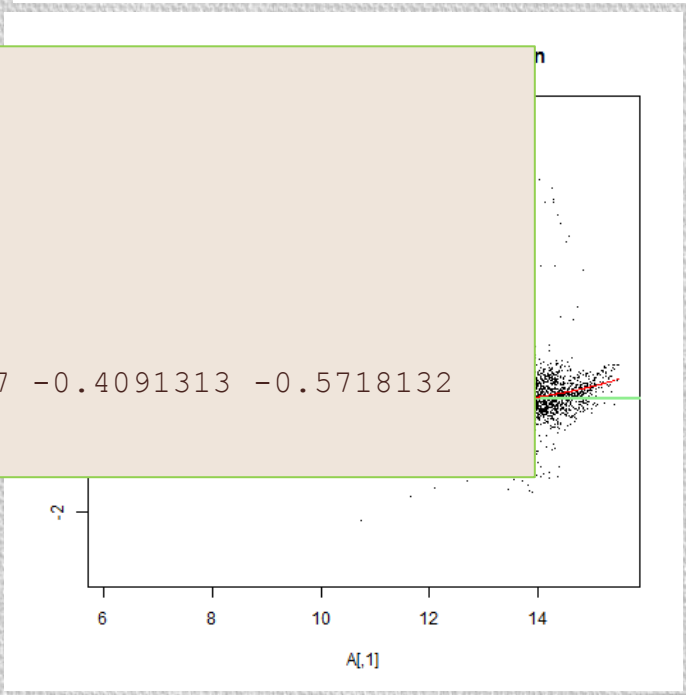
- Normalizacja nie usunęła rysy na mikromacierzy
- Normalizacja nie ma na celu przeniesienie *wszystkich* punktów w okolice 0 (MAplot), lecz *większość*
- Punkty z różną ekspresją genów (jak również artefakty) będą przedstawione na wykresach (MAplot) jako odstające (*outliers*)

MAplot z krzywą regresji

```
>MAmedian=normalizeWithinArrays(RG,method="median")

>z=loess(MAmedian$M[,1]~MAmedian$A[,1])  #loess(y~x)

>tmp=predict(z,MAmedian$A[,1])

>plot(MAmedian$A[,1],MAmedian$M[,1],pch=".",xlab="A[,1]",
ylab="M[,1]", ylim=c(-3,5), main="MAplot median norm with
regression")

>points(MAmedian$A[,1],tmp, col="red",pch=".")
```




```
> z
Call:
loess(formula = MA$M[, 1] ~ MA$A[, 1])

Number of Observations: 8448
Equivalent Number of Parameters: 4.86
Residual Standard Error: 0.4868
> tmp[1:10]
 [1] -0.5261056 -0.4819986 -0.5101984 -0.6517583 -0.6530537 -0.4091313 -0.5718132
 [8] -0.4734599 -0.5146899 -0.6462602
```



Regresja

Regresja dla pewnego zbioru danych to aproksymacja funkcji (pewnego modelu matematycznego), która jak najdokładniej będzie opisywała te dane.

Regresja liniowa, w której funkcja odwzorowania (*y*) jest liniowo zależna od cech (*x*)

$$y=ax+b$$

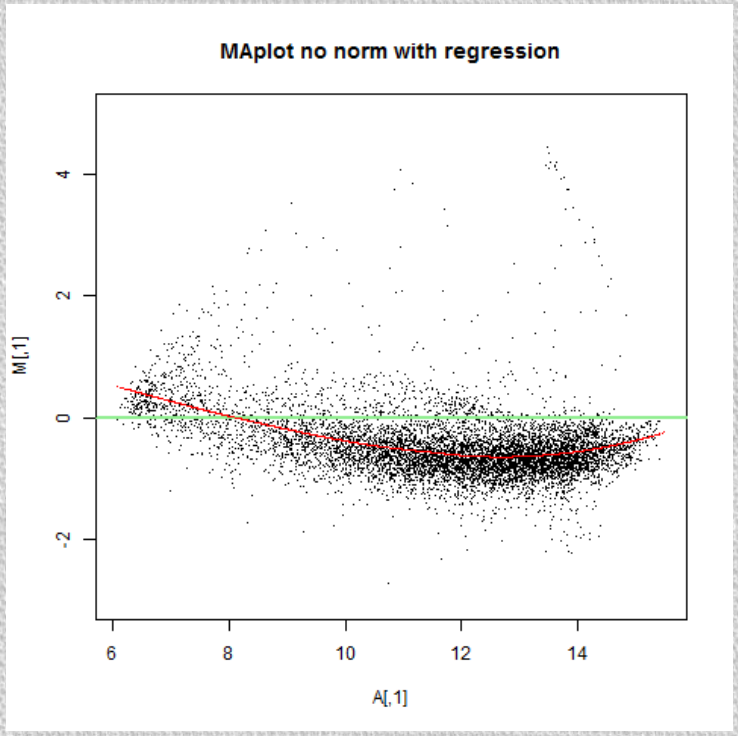
Istnieją również różne formy regresji nieliniowej.

Normalizacja zależna od intensywności

- Normalizacja zależna od A – **dla każdej wartości A średnia wartość M jest równa 0**.

$$M = \log_2 R/G \rightarrow M_{norm} = \log_2 R/G - c(A)$$

c(A) jest to funkcja dopasowania loess na wykresie MAplot.



loess

Loess – ważona regresja lokalnie wielomianowa (local polynomial regression fitting). Przeprowadzana dla każdego punktu, polega na wygładzeniu linii regresji w kierunku zera.

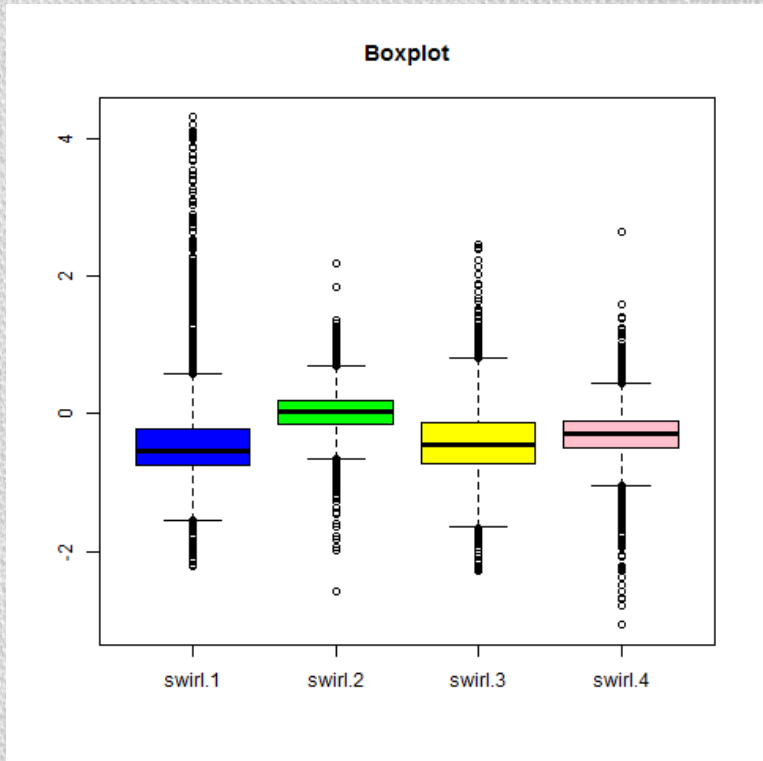
$$\log_2 R/G \rightarrow \log_2 R/G - c(A) = \log_2 R/\{k(A)G\}$$

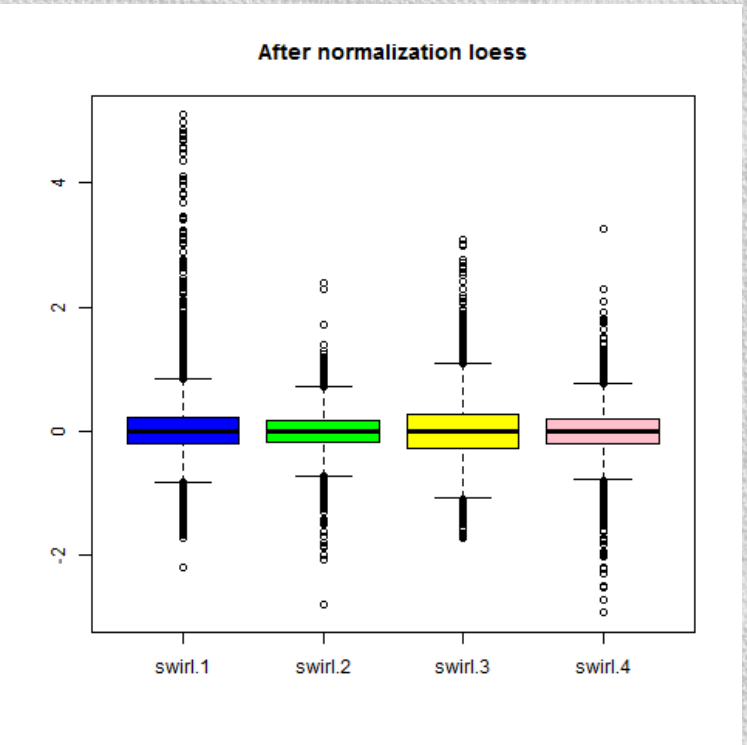
$c(A) = \log_2 k(A)$, gdzie $c(A)$ jest dopasowaniem loess (*loess fit*), uzależnionym od A.

Niski procent genów ulegających zróżnicowanej ekspresji nie będzie miał wpływu na wygładzanie, a punktu te będą przedstawione jako ‚outliers' na wykresie MAplot.

Użytkownik może zdefiniować parametr *f*, który odpowiada za odsetek punktów użytych w procesie wygładzania; im większa wartość *f*, tym gładsza będzie linia. Zazwyczaj przyjmuje się *f*=40%.

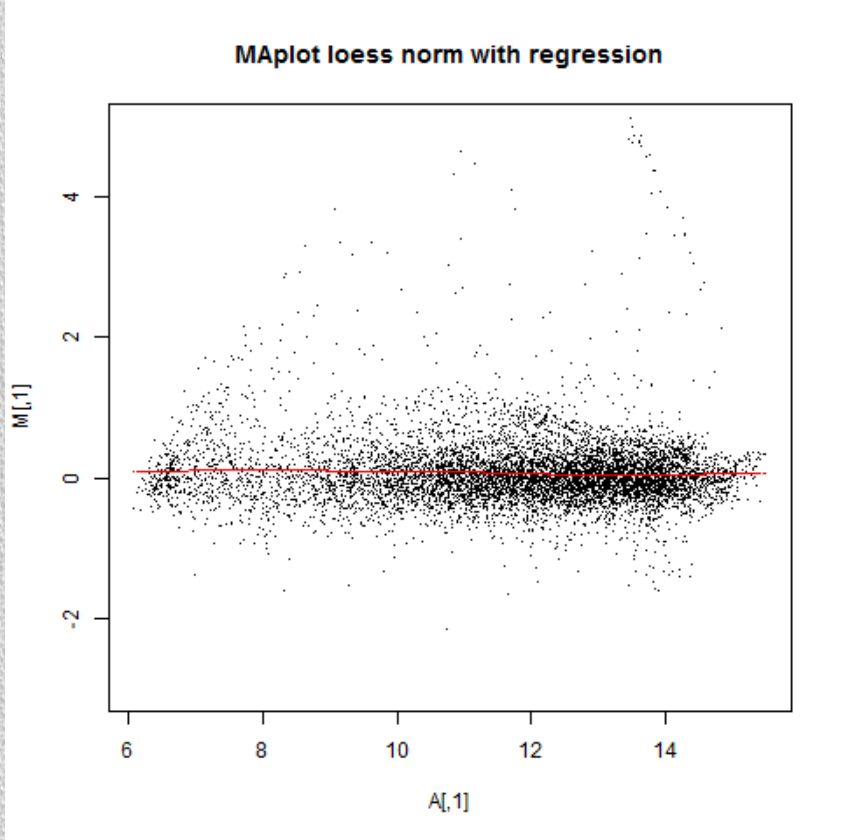
loess: boxplot dla wszystkich mikromacierzy





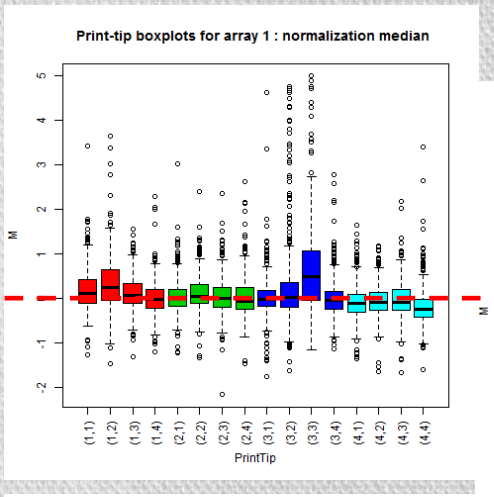
Mediany zostały wycentrowane ☺

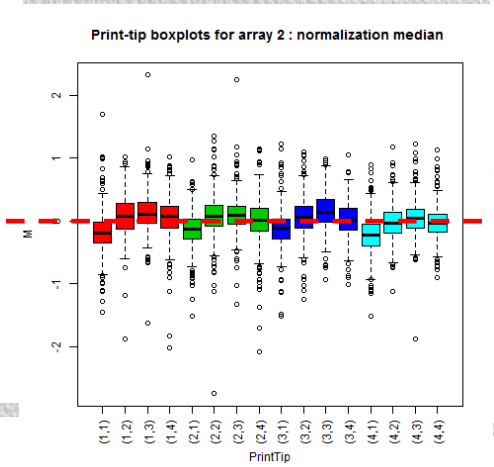
loess: MAplot

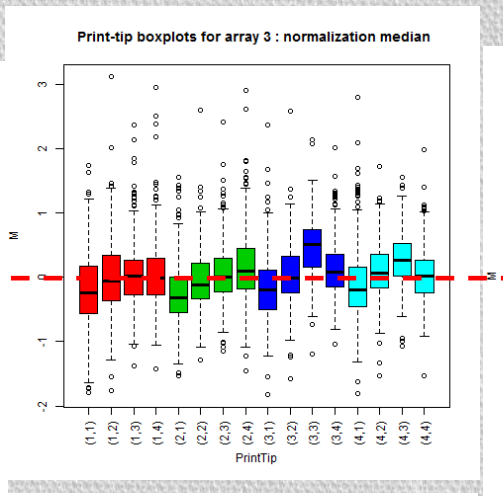


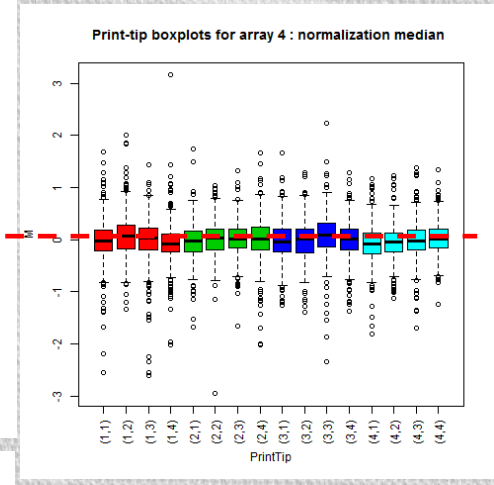
Krzywa loess oscyluje wokół 0 ☺☺

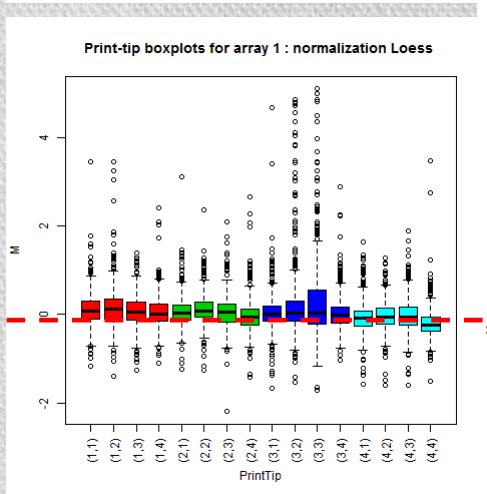
loess: boxplot z rozbiciem na bloki (print-tip)

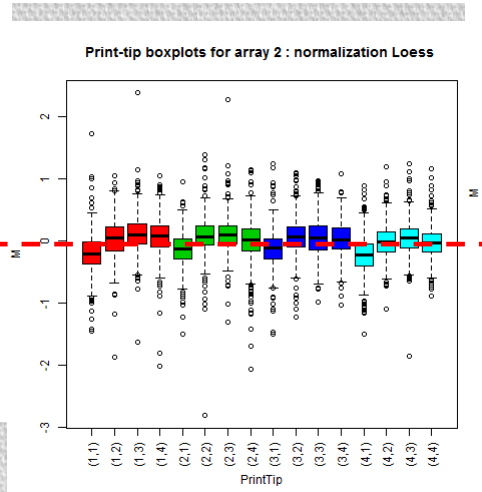


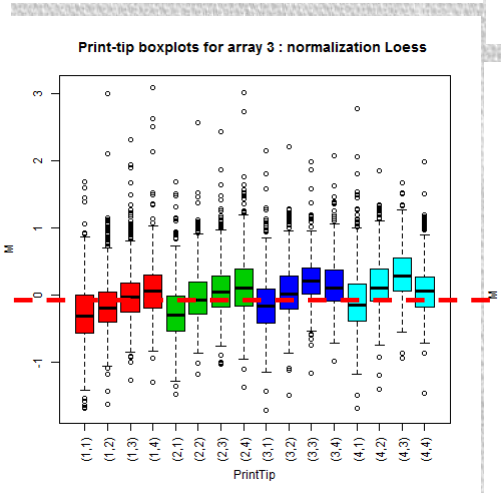


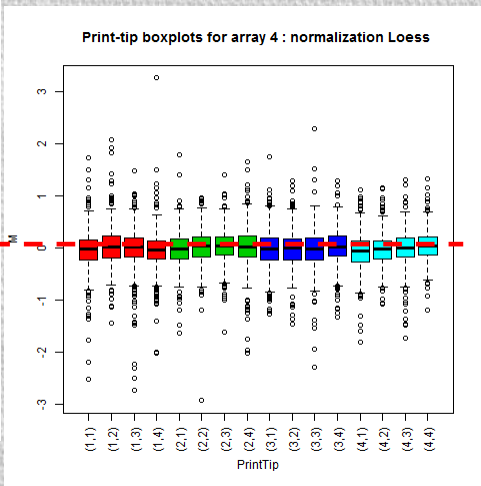










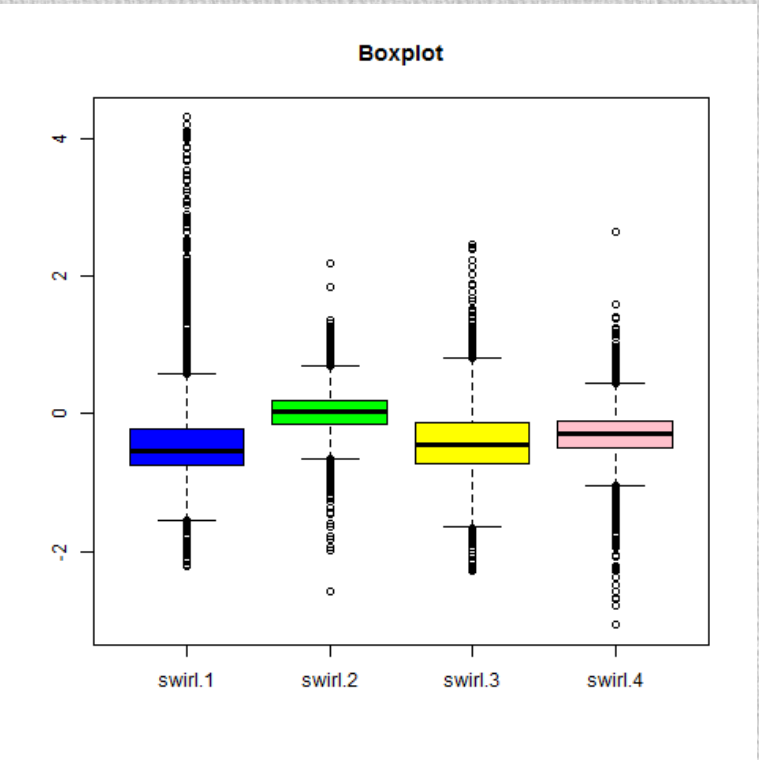


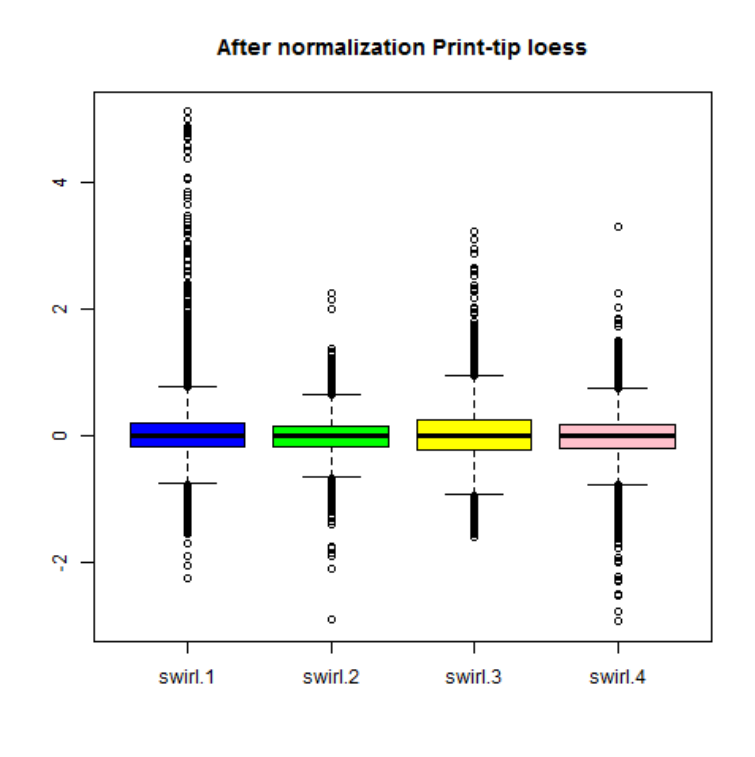
printtiploess: boxplot dla wszystkich mikromacierzy

Normalizacja zależna od A, także od bloku w którym się znajduje (print-tip)

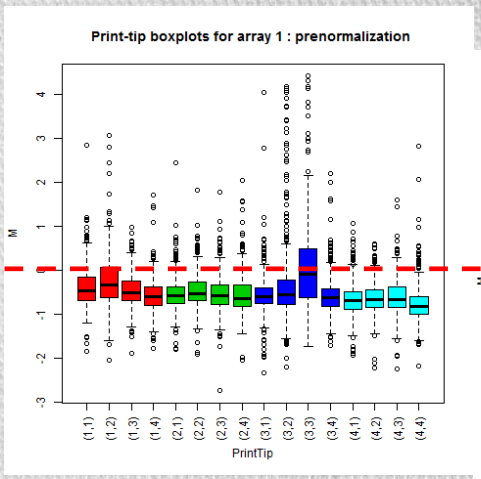
$$\log_2 R/G \rightarrow \log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G)$$

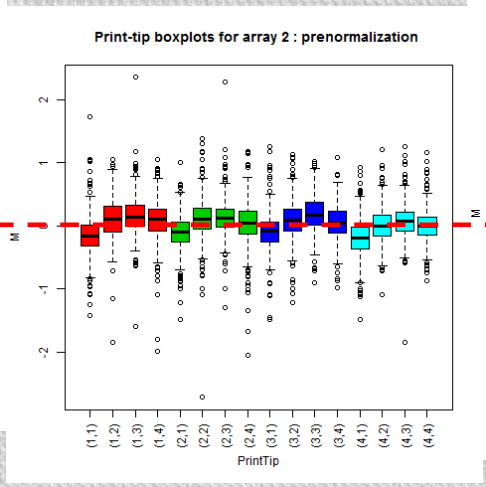
$c_i(A) = \log_2 k_i(A)$, to dopasowanie loess w i-tym bloku

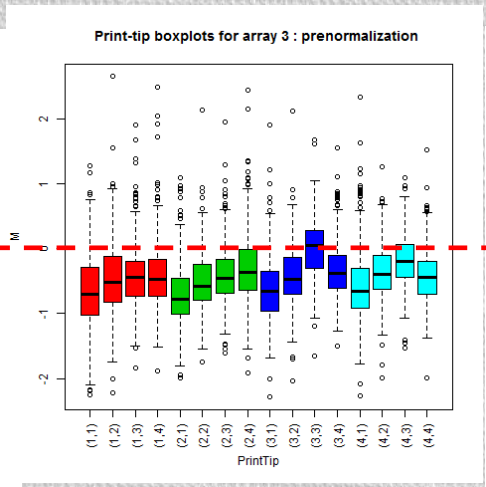


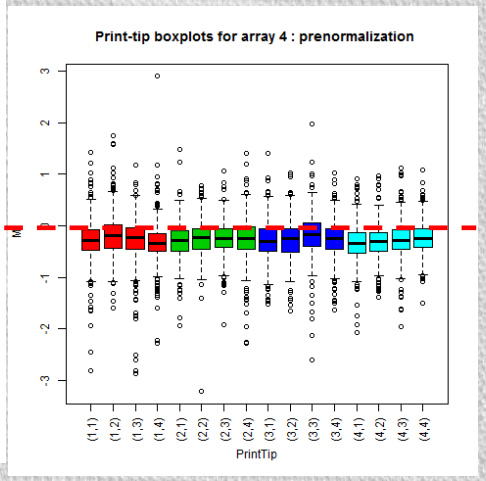


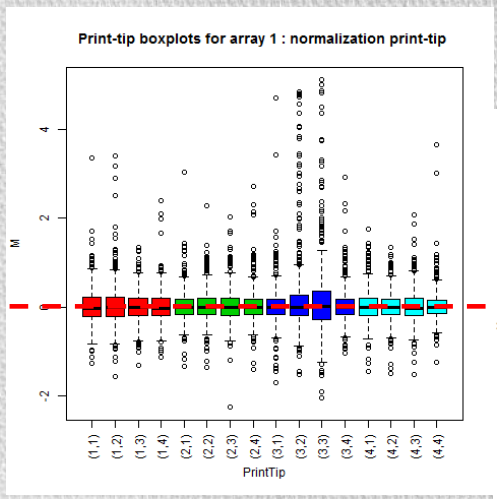
printtiploess: boxplot z rozbiciem na bloki (print-tip)

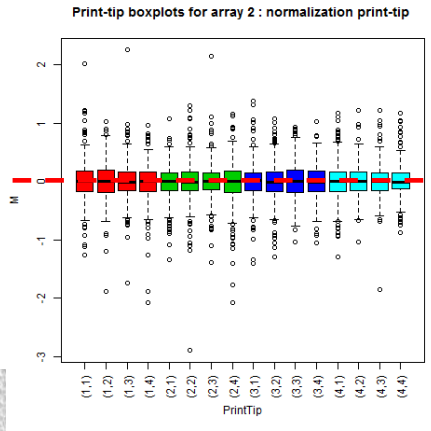


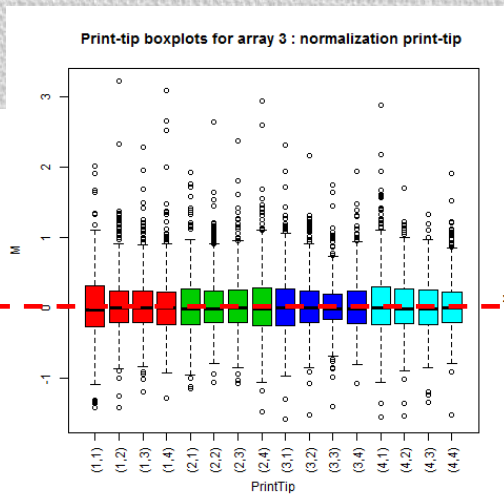


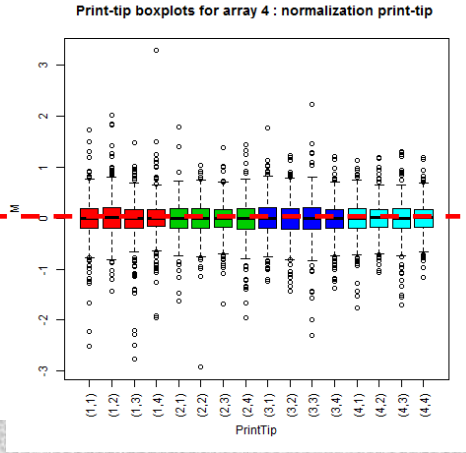






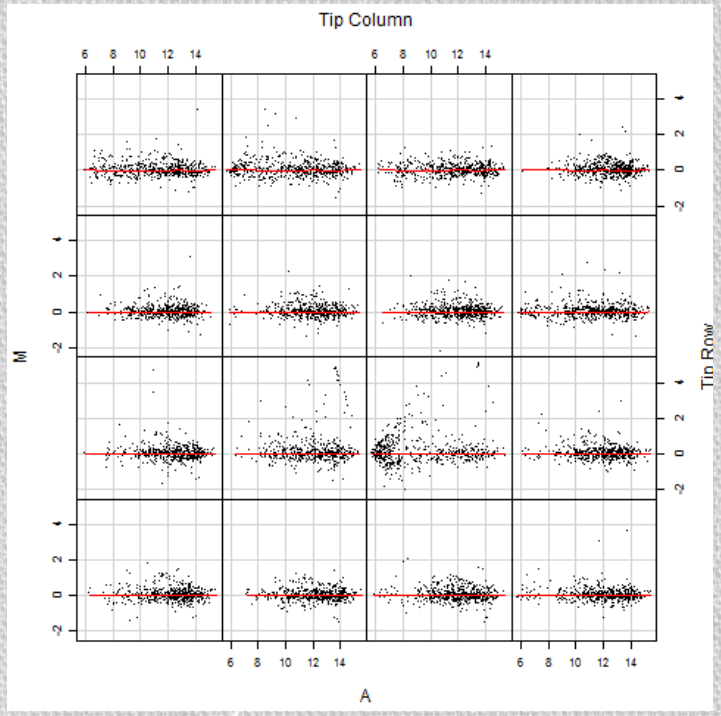


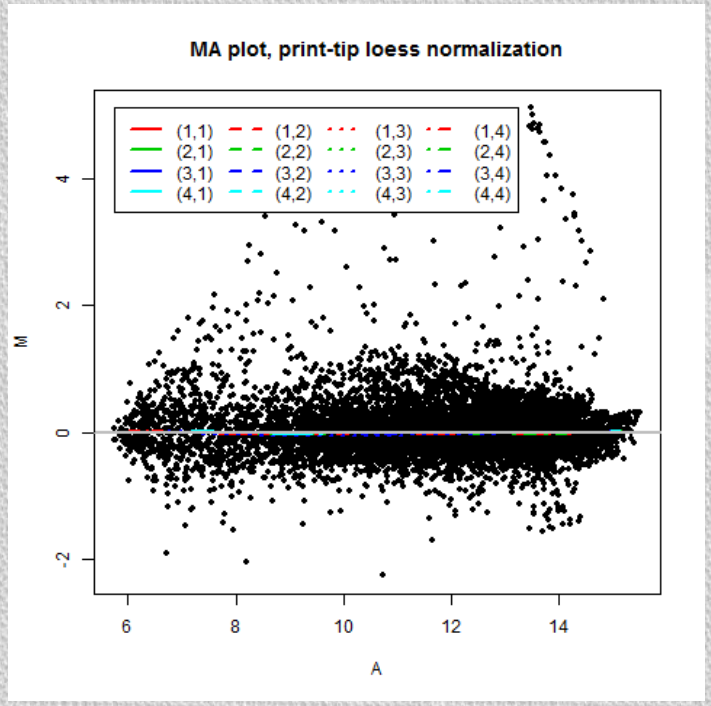




printtiploess

Metoda ta zniwelowała różnice pomiędzy poszczególnymi mikromacierzami (wycentrowane mediany dla całych mikromacierzy) jak i usunęła różnice powstałe w wyniku drukowania przez różne igły (mediany dla poszczególnych print-tipów).





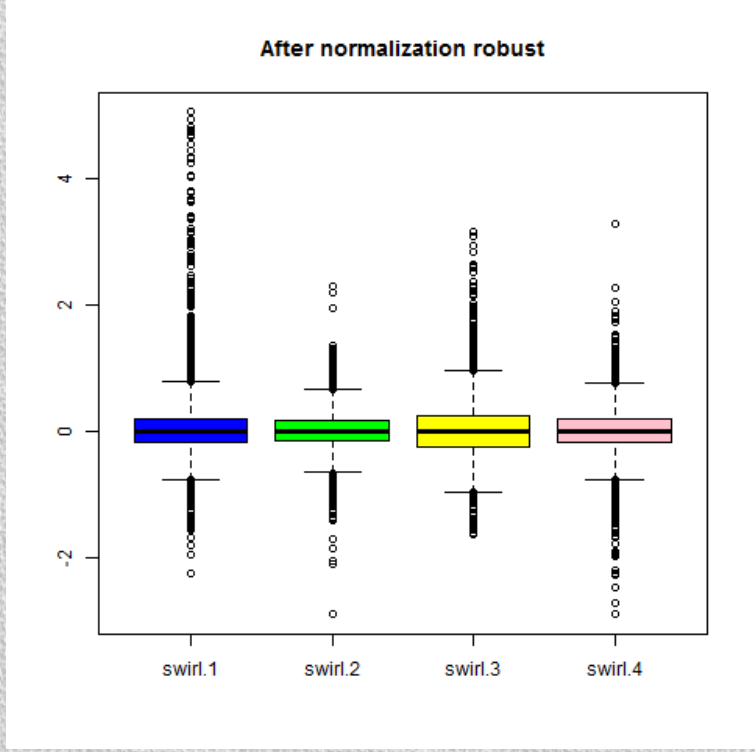
printtiploess

Tę normalizację można przeprowadzić wówczas, gdy jest wystarczająca liczba punktów przypadających na jeden print-tip, czyli ok. 150 lub więcej

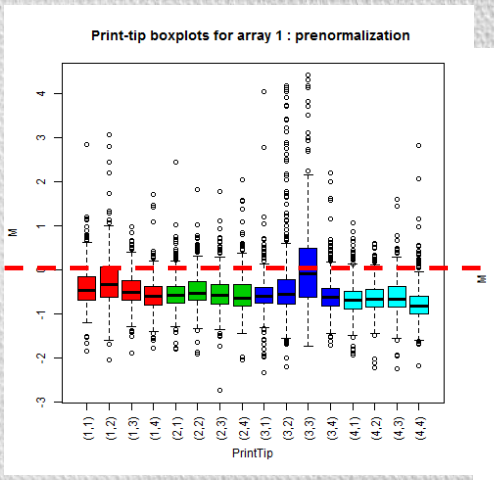
robust: boxplot dla wszystkich mikromacierzy

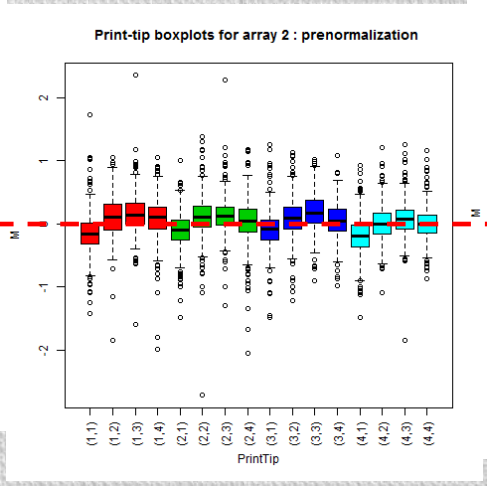
Funkcja ***robust*** jest podobna do normalizacji *print-tip* , z tymże używa pasków regresji zamiast krzywych loess oraz używa funkcji Bayesowskiej, aby indywidualne linie dla bloków (*print-tip*) zbiegły się w kierunku określonej wartości (zera).

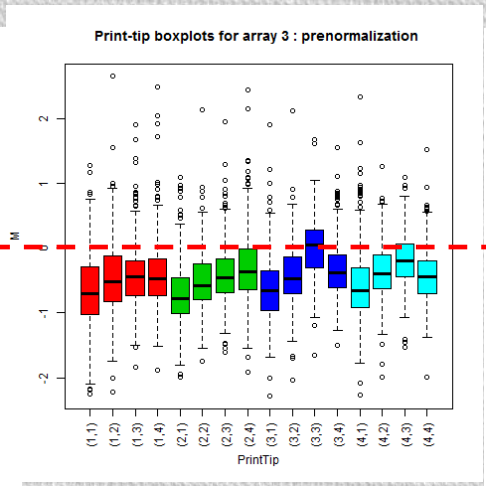
Ta metoda wprowadza mniej błędów do mikromacierzy dobrej jakości, z niewielką wariancją przestrzenną, niż metoda *print-tip*. Również dla punktów z silną wariancją w blokach, funkcja nadal daje dobre rezultaty.

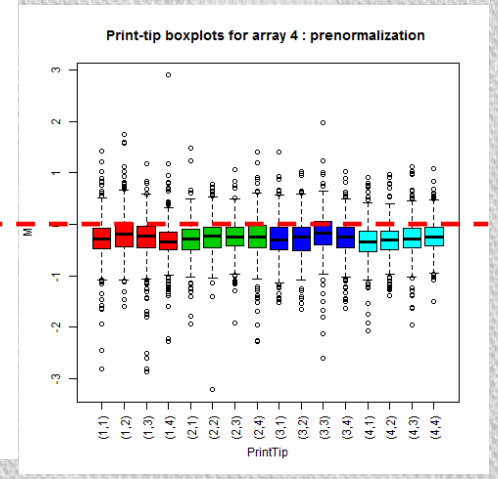


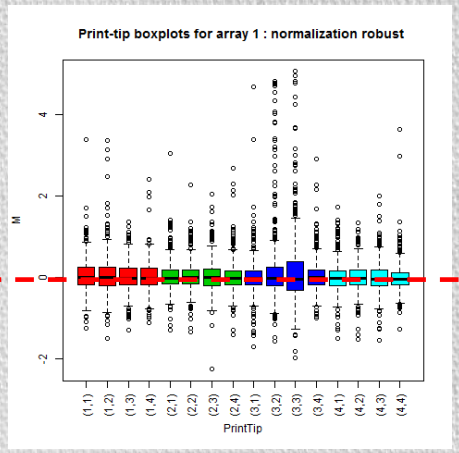
robust: boxplot z rozbiciem na bloki (print-tip)

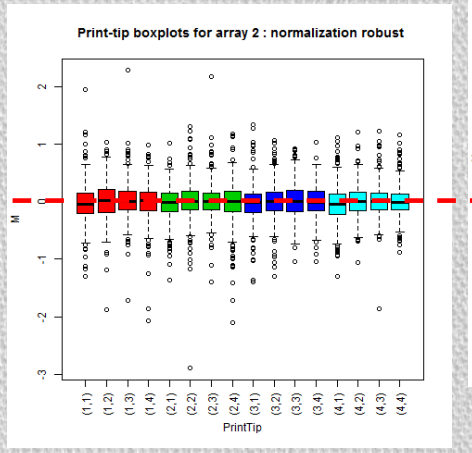


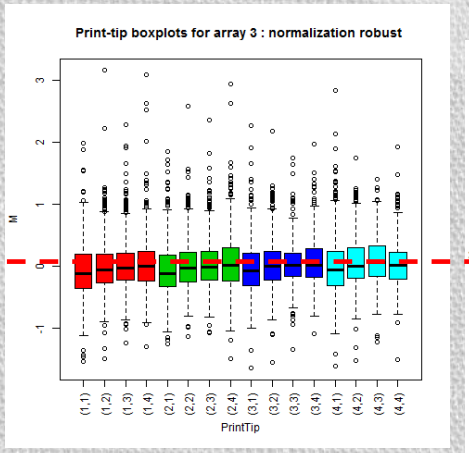


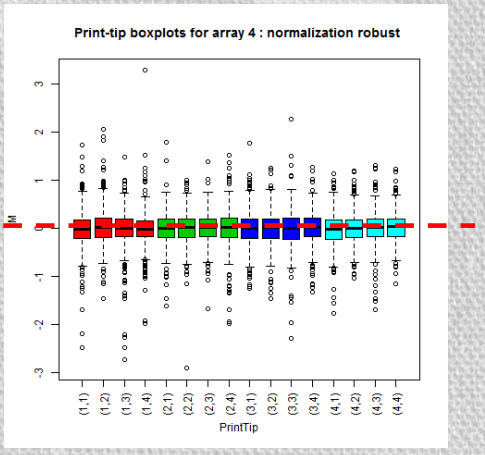












composite

- Z udziałem **spike genes** na mikromacierzy
- Próbki kontrolne (**MSP – microarray sample pools**) z niewielkim odchyleniem zależnym od badanej próbki przy jednoczesnym dużym zakresie intensywności

Algorytm

1) Wyznacz funkcję loess $f_i(A)$ dla i-tego bloku (print-tip) na MAplot
2) Wyznacz funkcje loess dla punktów MSP $g(A)$
3) Oblicz średnią ważoną $c_i(A) = \alpha_A\, g(A) + (1-\alpha_A)\, f_i(A)$

α_A – jest zdefiniowana jako proporcja genów mniejsza od intensywności A

control

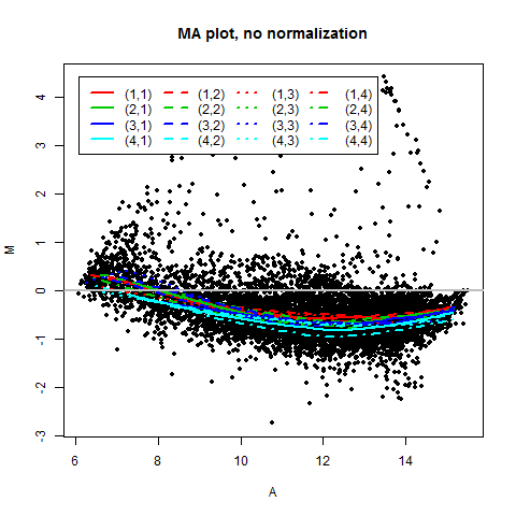
- Dla mikromacierzy, dla której duża część genów uległa zróżnicowanej ekspresji
- Podobne do metody *composite*

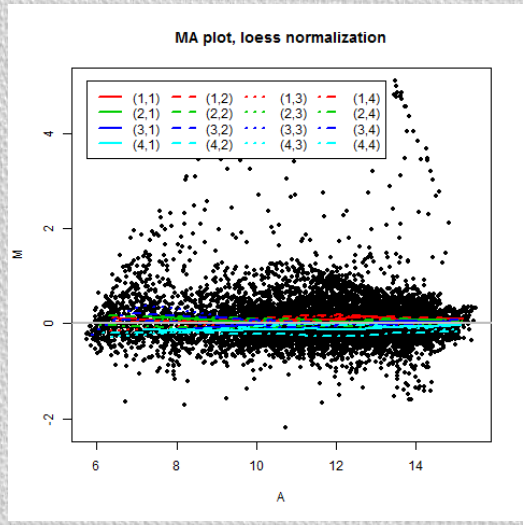
$$c_i(A) = \alpha_A\, g(A) + (1-\alpha_A)\, f_i(A)$$

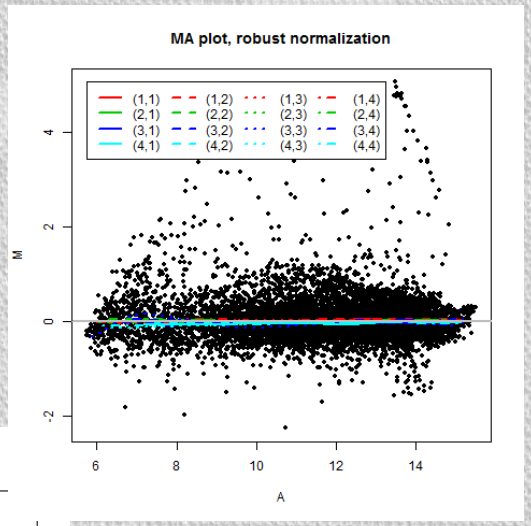
- Wyznaczana jest globalna krzywa loess dla punktów kontrolnych, a następnie korekcja względem tej krzywej jest stosowana do wszystkich punktów

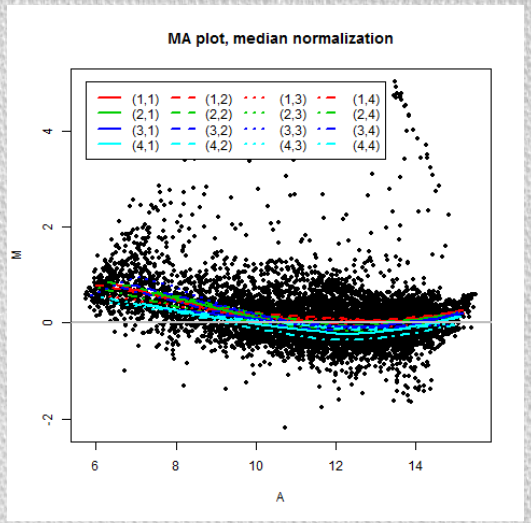
Oshlack et al (2007)

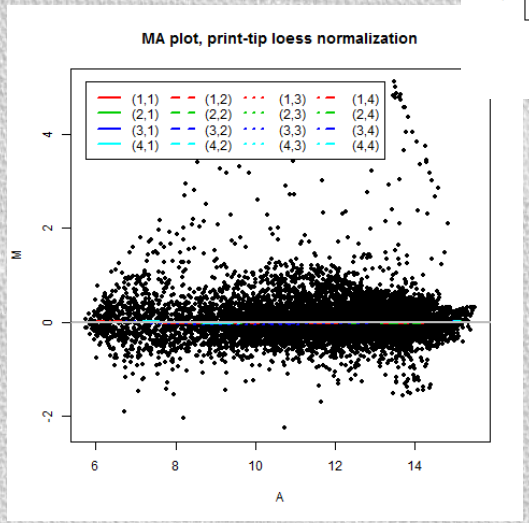
Porównanie MAplot dla różnych metod normalizacji











Tabelka z artykułu Yang et al(2002)

Nucleic Acids Research, 2002, Vol. 30, No. 4 e15

Table 1. The various normalization methods considered in this article

		Within-slide				Multiple slide
		Global, location	Intensity-dependent, location	Print tip-dependent, location	Print tip, location and scale	Scale
		$c(.)$ constant, $a(.) = 1$	$c(.) = c(A)$, $a(.) = 1$	$c(.) = c(A$, print tip), $a(.) = 1$	$c(.) = c(A$, print tip), $a(.) = a$(print tip)	
All genes	Assumes the majority of genes in the two mRNA samples have similar overall expression levels	Yes	Yes	Yes	Yes	Yes
Housekeeping genes	Usually highly expressed and do not capture intensity-dependent structure	Yes	No	No	No	No
MSP titration series	Doesn't require any prior biological assumption, however, estimating $c(A$, print tip) based on a small number of spots may not be very stable	Yes	Yes	No	No	No
Rank-invariant set	May not span the whole intensity range	Yes	Yes	No	No	No

For within-slide normalization, the log ratios are normalized by $\log_2 R/G \rightarrow [\log_2 R/G - c(.)]/a(.)$, where $c(.)$ and $a(.)$ correspond to location and scale adjustment, respectively. The columns refer to different normalization methods and the rows correspond to different sets of control spots. The Yes or No in each cell refers to the feasibility of performing the normalization in practice. For example, it is possible in practice to perform global normalization based only on housekeeping genes, but it is not advisable to perform intensity-dependent normalization on housekeeping genes only.

BetweenArraysNormalization

- Normalizację pomiędzy różnymi mikromacierzami przeprowadzamy jeśli mikromacierze mają różny rozrzut wartości (na boxplocie).
- Jeśli wartości te nie będą przeskalowane, może to doprowadzić do faworyzowania jednego lub kilku eksperymentów ze względu na różnicę log-ratio.
- Czasami rezygnuje się ze skalowania między-mikromacierzowego, jeśli odchylenia nie są zbyt wielkie, aby dodatkowo wprowadzony szum nie był bardziej znaczący niż niewielka początkowa różnica między odchyleniami.

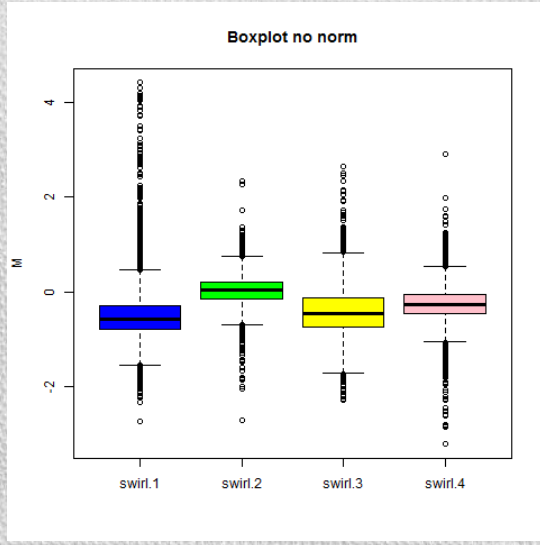
scale, quantile, Aquantile, Gquantile, Rquantile, Tquantile, vsn

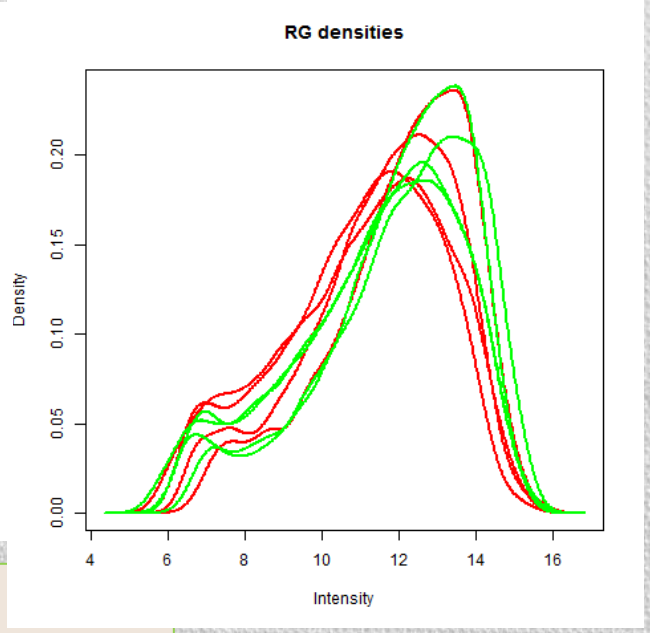
BetweenArraysNormalization

Tabela, w której każda kolumna reprezentuje jedną mikromacierz

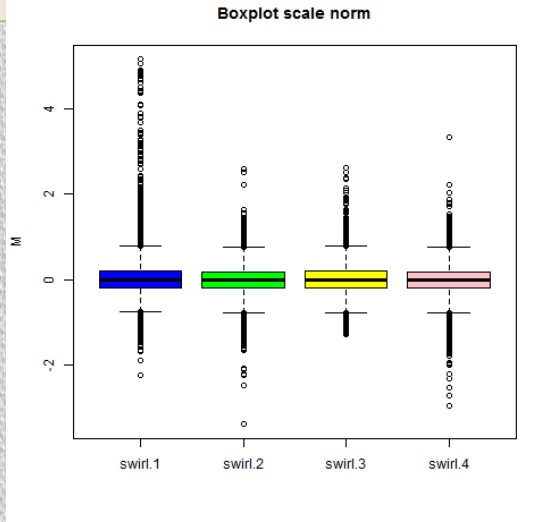
- **scale** – skaluje kolumny, aby miały taką samą medianę
- **quantile** – wymuszenie aby cały rozkład każdej z kolumn był identyczny (a nie tylko 50% punktów)
- **Cyclic loess** – stosuje normalizację loess dla każdej pary mikromacierzy, zazwyczaj cyklicznie dla każdej pary kilkukrotnie wyznaczana jest loess
- **Aquantile** – ustawienie wartości A (średnia intensywność) aby miały taki sam rozkład dla wszystkich mikromacierzy
- **Tquantile** – quantile normalization z podziałem na grupy (mikromacierzy)
- **Gquantile/Rquantile** – metoda zapewnia że kanał zielony/czerwony jest stały dla wszystkich macierzy, podczas gdy wartości M są niezmienione. Metodę tę przeprowadza się, gdy na kanale zielonym/czerwonym jest próbka referencyjna dla wszystkich mikromacierzy

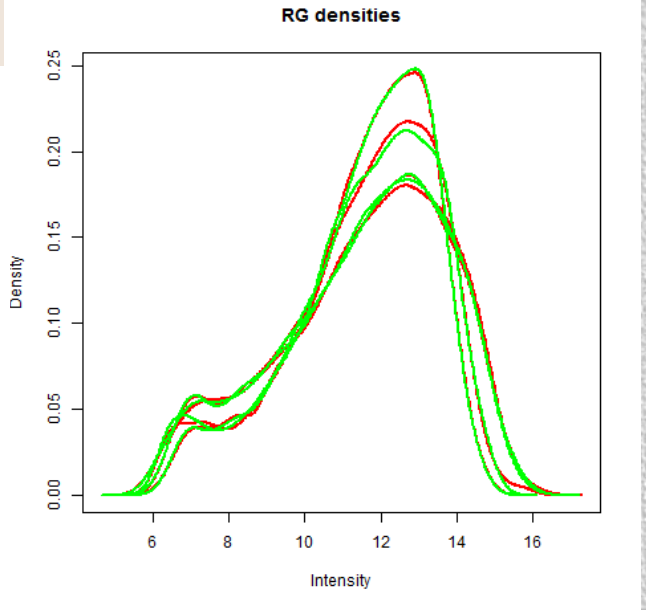
Normalizacja *scale*






```
> MA.ps=normalizeBetweenArrays(MA.p,method="scale")
> plotDensities(MA)
> plotDensities(MA.ps)
> boxplot(MA$M,col=cols, main="Boxplot no norm",ylab="M")
> boxplot(MA.ps$M,col=cols, main="Boxplot scale
norm",ylab="M")
```



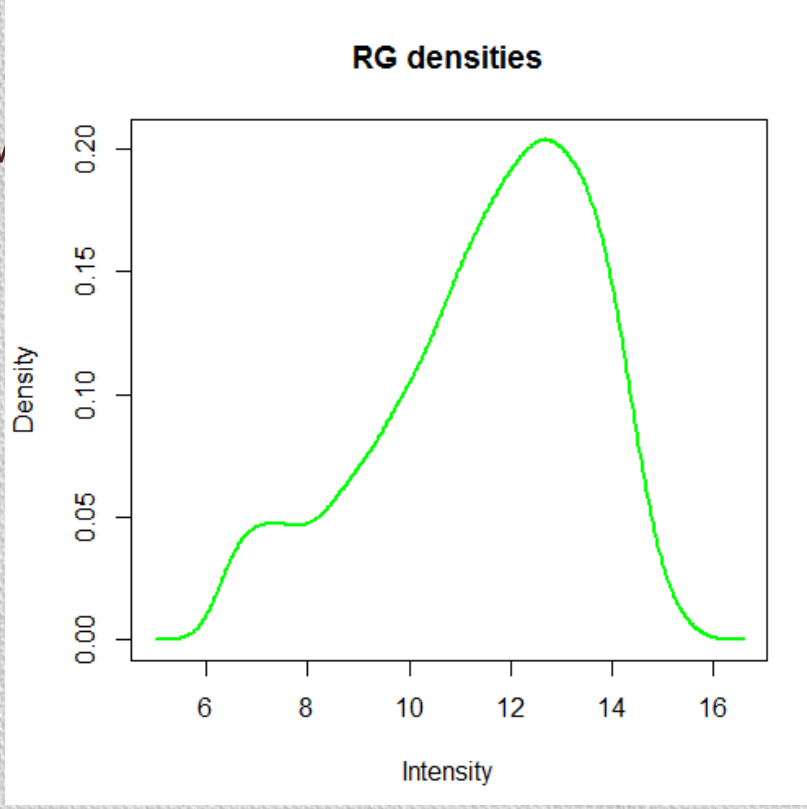


Normalizacja *quantile*

Założenie:

Histogramy dla wszystkich płytek mają być identyczne

Jeszcze bardziej w trowania median



Normalizacja *quantile*

	Próbka A	**Próbka B**	**Próbka C**
Gen 1	100	200	140
Gen 2	10	40	270
Gen 3	100	120	70

Normalizacja *quantile*

	Próbka A	**Próbka B**	**Próbka C**
Gen 1	100	200	140
Gen 2	10	40	270
Gen 3	100	120	70

Normalizacja *quantile*

	Próbka A	**Próbka B**	**Próbka C**
	100 Gen 1	200 Gen 1	140 Gen 1
	10 Gen 2	40 Gen 2	270 Gen 2
	100 Gen 3	120 Gen 3	70 Gen 3

Normalizacja *quantile*

	Próbka A	Próbka B	Próbka C
	10 Gen 2	40 Gen 2	70 Gen 3
	100 Gen 3	120 Gen 3	140 Gen 1
	100 Gen 1	200 Gen 1	270 Gen 2

Średnia
(10+40+70)/3 = 40
(100+120+140)/3= 120
(100+200+270)/3= 190

Normalizacja *quantile*

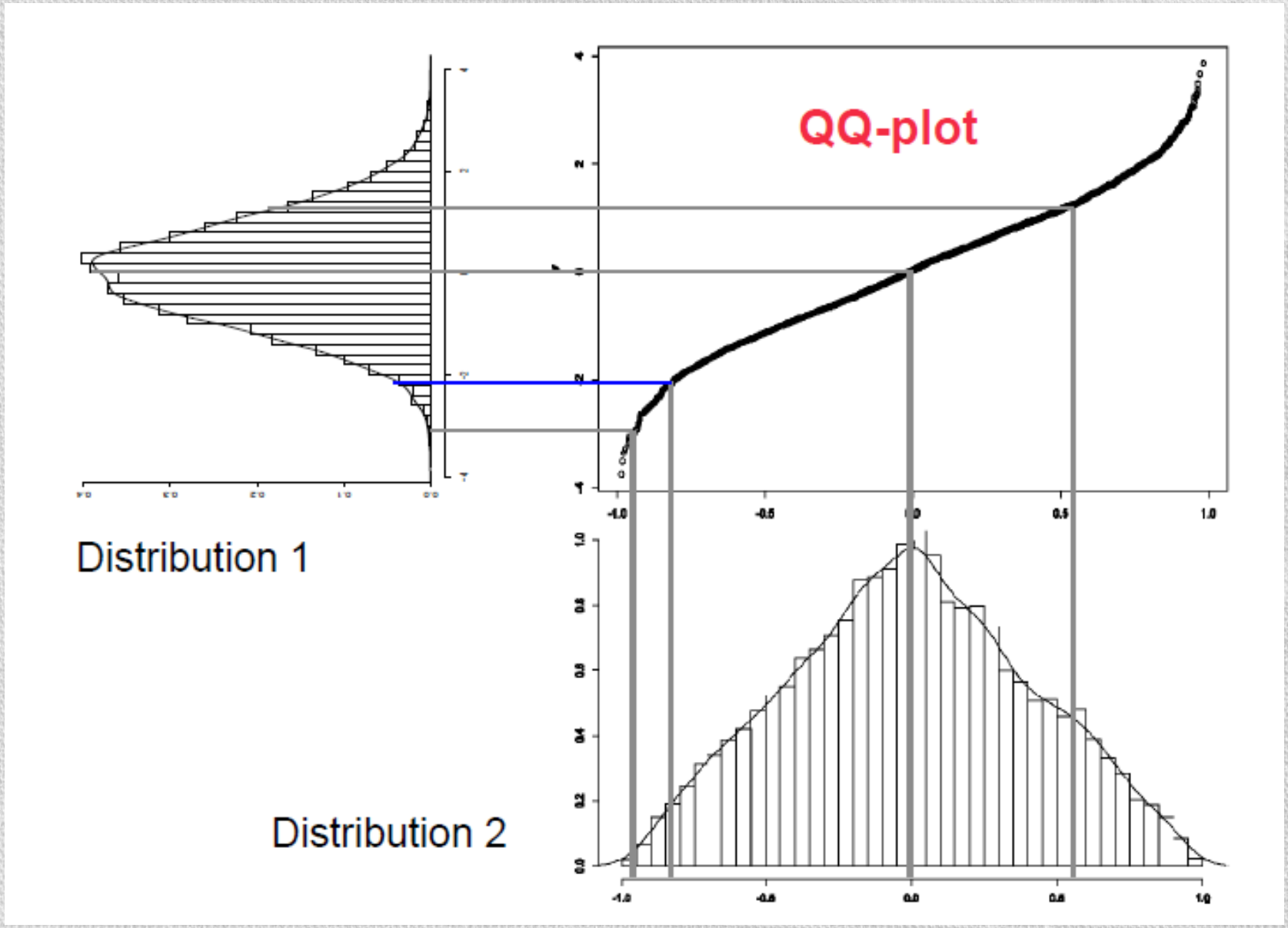
	Próbka A	Próbka B	Próbka C
	40 Gen 2	40 Gen 2	40 Gen 3
	120 Gen 3	120 Gen 3	120 Gen 1
	190 Gen 1	190 Gen 1	190 Gen 2

Średnia
(10+40+70)/3 = 40
(100+120+140)/3= 120
(100+200+270)/3= 190

Normalizacja *quantile*

	Próbka A	**Próbka B**	**Próbka C**
Gen 1	190 (100)	190 (200)	120 (140)
Gen 2	40 (10)	40 (40)	190 (270)
Gen 3	120 (100)	120 (120)	40 (70)

Normalizacja *quantile*



vsn

- Transformacja logarytmiczna zastąpiona jest poprzez transformację arcsinh
- Korekcja tła
- Estymacja parametrów położenia i skalowania poprzez rozwiązanie minimalnej sumy kwadratów regresji
- Dane znormalizowane przez metode vsn mają rozkład zbliżony do normalengo

- Można zastosować jeśli <50% genów uległo różnicowej ekspresji