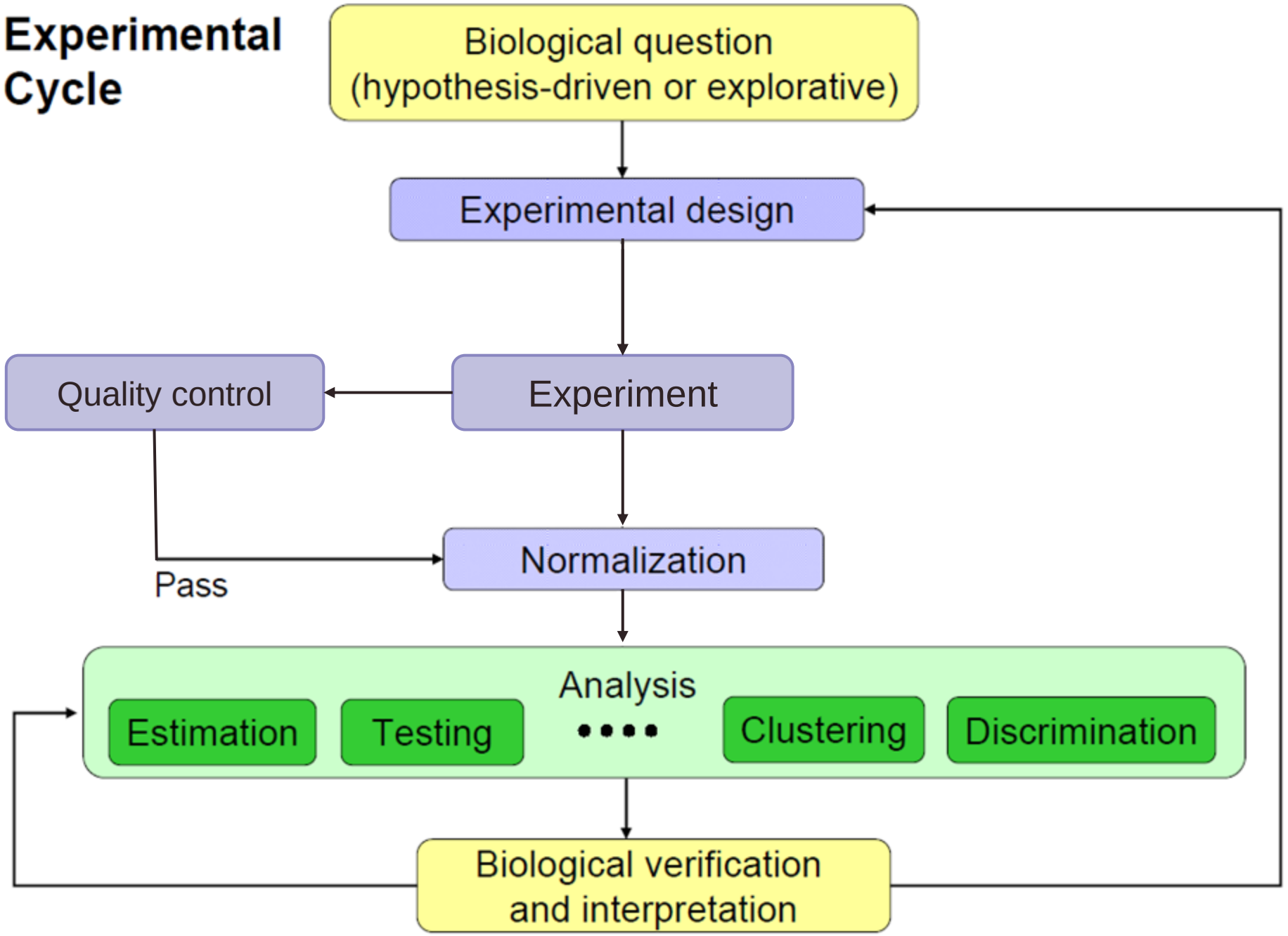
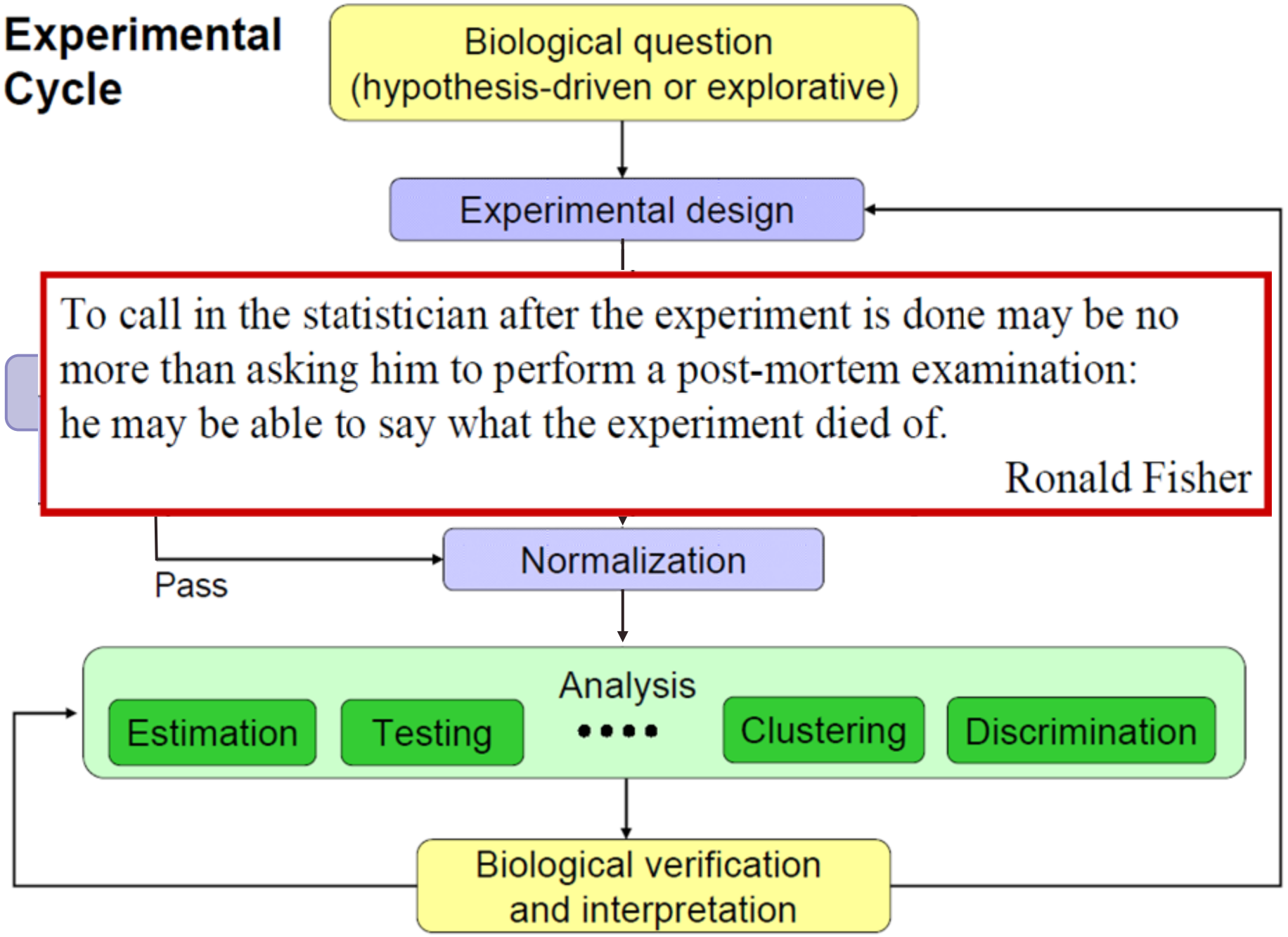
KOREKCJA TŁA
KONTROLA JAKOŚCI DANYCH
NORMALIZACJA


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


Experimental Cycle
Biological question (hypothesis-driven or explorative)
Experimental design
To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.
Ronald Fisher
Pass
Normalization
Analysis
Estimation
Testing
Clustering
Discrimination
Biological verification and interpretation

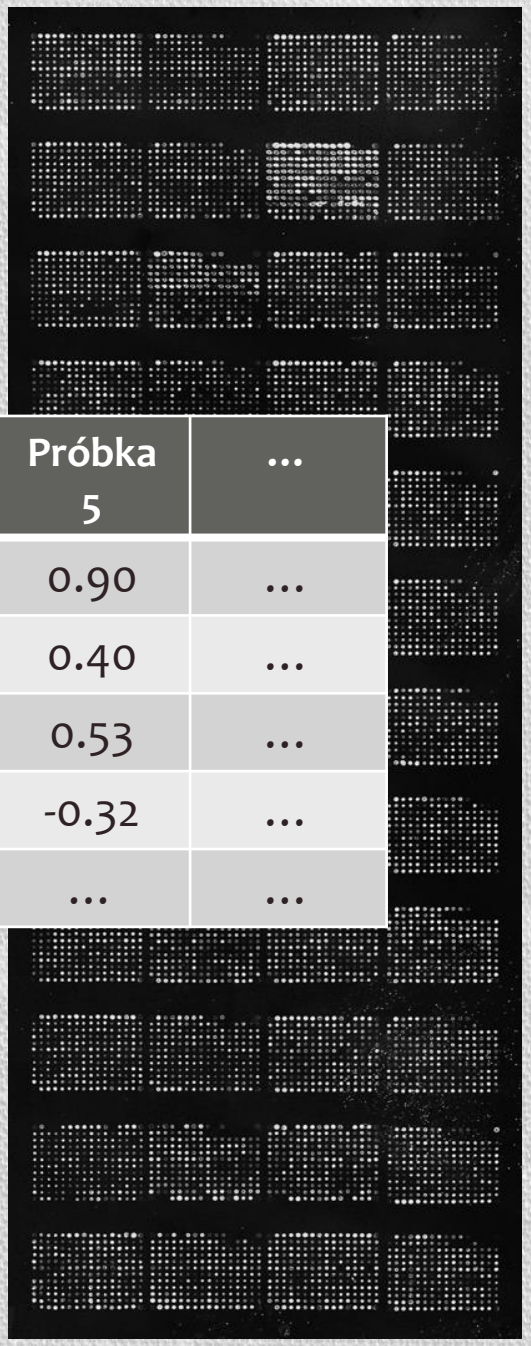
Cel wstępnej analizy mikromacierzowej

- **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.
- Każda wartość w macierzy ekspresji genów jest albo absolutną wartością transkryptu (intensywnością fluorescencyjną), albo stosunkiem dwóch intensywności fluorescencyjnych (dla dwóch różnych warunków eksperymentalnych)
- interpretacja danych intensywności w macierzy ekspresji genów:
  - aby wyciągnąć ciekawe wnioski biologiczne
  - w celu zaproponowania kolejnych eksperymentów.

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	…
…	…	…	…	…	…	…

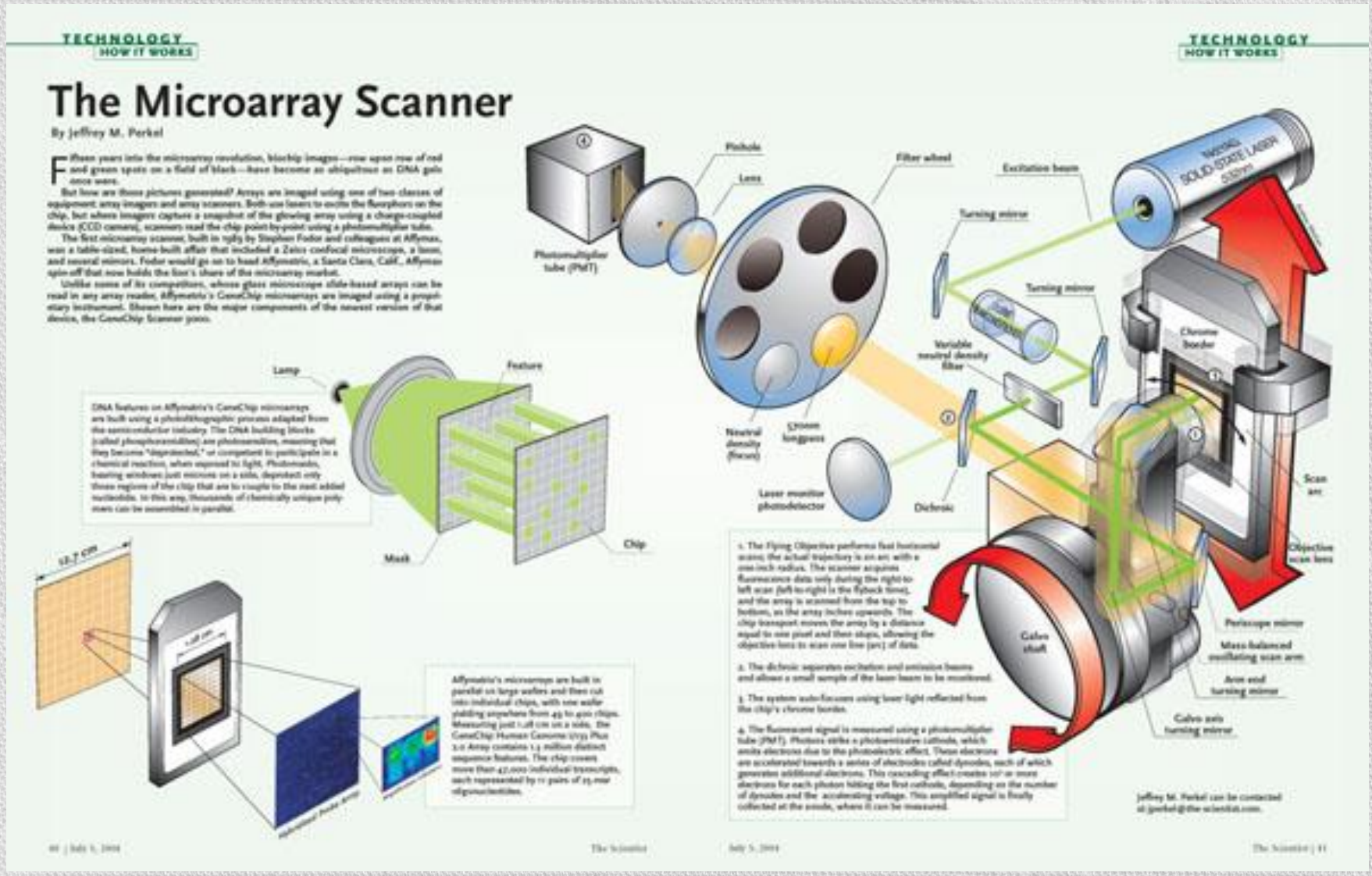
Geny/ sondy

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

Skaner mikromacierzy



TECHNOLOGY
HOW IT WORKS

The Microarray Scanner

By Jeffrey M. Perkel

Fifteen years into the microarray revolution, blockip images—row upon row of red and green spots on a field of black—have become as ubiquitous as DNA gels once were.

But how are those pictures generated? Arrays are imaged using one of two classes of equipment: array imagers and array scanners. Both use lasers to excite the fluorphors on the chip, but where imagers capture a snapshot of the glowing array using a charge-coupled device (CCD camera), scanners read the chip point-by-point using a photomultiplier tube.

The first microarray scanner, built in 1989 by Stephen Fodor and colleagues at Affymax, was a table-sized, home-built affair that included a Zeiss confocal microscope, a laser, and several mirrors. Fodor would go on to head Affymetrix, a Santa Clara, Calif., Affymax spin-off that now holds the lion's share of the microarray market.

Unlike some of its competitors, whose glass microscope slide-based arrays can be read in any array reader, Affymetrix's GeneChip microarrays are imaged using a proprietary instrument. Shown here are the major components of the newest version of that device, the GeneChip Scanner 3000.

Photomultiplier tube (PMT)
Pinhole
Lens
Filter wheel
Excitation beam
Turning mirror
Turning mirror
Variable neutral density filter
Chrome border
Neutral density (focus)
570nm longpass
Laser monitor photodetector
Dichroic
Scan arm
Objective scan lens
Periscope mirror
Mass-balanced oscillating scan arm
Arm-end turning mirror
Galvo axis turning mirror
Galvo shaft
SOLID-STATE LASER

Lamp
Feature
Mask
Chip

DNA features on Affymetrix's GeneChip microarrays are built using a photolithographic process adapted from the semiconductor industry. The DNA building blocks (called phosphoramidites) are photosensitive, meaning that they become "deprotected," or competent to participate in a chemical reaction, when exposed to light. Photomasks, bearing windows just microns on a side, deprotect only those regions of the chip that are to couple to the next added nucleotide. In this way, thousands of chemically unique polymers can be assembled in parallel.

Affymetrix's microarrays are built in parallel on large wafers and then cut into individual chips, with one wafer yielding anywhere from 49 to 400 chips. Measuring just 1.28 cm on a side, the GeneChip Human Genome U133 Plus 2.0 Array contains 1.3 million distinct sequence features. The chip covers more than 47,000 individual transcripts, each represented by 11 pairs of 25-mer oligonucleotides.

1. The Flying Objective performs fast horizontal scans; the actual trajectory is an arc with a one-inch radius. The scanner acquires fluorescence data only during the right-to-left scan (left-to-right is the flyback time), and the array is scanned from the top to bottom, as the array inches upwards. The chip transport moves the array by a distance equal to one pixel and then stops, allowing the objective lens to scan one line (arc) of data.

2. The dichroic separates excitation and emission beams and allows a small sample of the laser beam to be monitored.

3. The system auto-focuses using laser light reflected from the chip's chrome border.

4. The fluorescent signal is measured using a photomultiplier tube (PMT). Photons strike a photosensitive cathode, which emits electrons due to the photoelectric effect. These electrons are accelerated towards a series of electrodes called dynodes, each of which generates additional electrons. This cascading effect creates 10⁷ or more electrons for each photon hitting the first cathode, depending on the number of dynodes and the accelerating voltage. This amplified signal is finally collected at the anode, where it can be measured.

Jeffrey M. Perkel can be contacted at jperkel@the-scientist.com.

Różne skanery

SureScan Microarray Scanner - Agilent

GeneChip Scanner 3000 System - Affymetrix

GenePix 4000B microarray Scanner – Molecular Devices

MS 200 Microarray Scanner – NimbleGen, Roche

HiScanSQ System - Illumina

GenePix Pro

No Need to Manually Scan Your Slides. Use the multiplexed image acquisition with your predefined settings and you do not need to manually scan each individual array anymore. Select multiple scan areas on one microarray and set your scan parameters for independent scans and click on *Data Scan All*. A GenePix® Microarray Scanner controlled by GenePix® Pro 7 Acquisition and Analysis Software will do the image acquisition and save the images to your desired location.

Manual Gridding No Longer Necessary. Rely on the powerful spot finding algorithms. Analyze any microarray TIFF image from any scanner, including hexagonally packed arrays, using a set of proprietary spot-finding algorithms in GenePix® Pro 7 Acquisition and Analysis Software. Global alignment algorithms determine translation, rotation and skew of blocks with features. In addition to finding circular or square features, the software has an edge-detection option for segmenting irregularly shaped features from background.

Define Your Own Flags. Define your own flag with the automated feature flagging. The *User-Defined Flag Features* interface allows you to design and save multiple-parameter Boolean queries for automated feature flagging. Queries can easily be shared between collaborators to help standardize quality control practices within a group of users.

Choose the Best Background Subtraction Method. With each microarray experiment, it is critical to calculate the best-possible numerical representation of biological changes on the array, which is often hampered by substandard array quality. GenePix® Pro 7 Software allows you to choose between several background subtraction methods*: local, global, morphological* or *negative control* methods. Or, where appropriate, choose to have *no background subtraction* at all.

Unattended Batch Analysis. GenePix® Pro 7 Acquisition and Analysis Software simplifies batch analysis. Simply load a group of TIFF images to the Batch Analysis Tab, assign layout files and press Start. GenePix® Pro 7 Acquisition and Analysis Software will automatically run spot-finding algorithms, extract numerical values, flag features and save a results file for each image. With the Browse Tool, users can quickly review results of the completed batch.

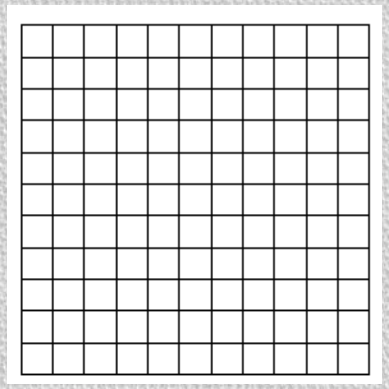
Jak przejść od obrazu do liczb?

- 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

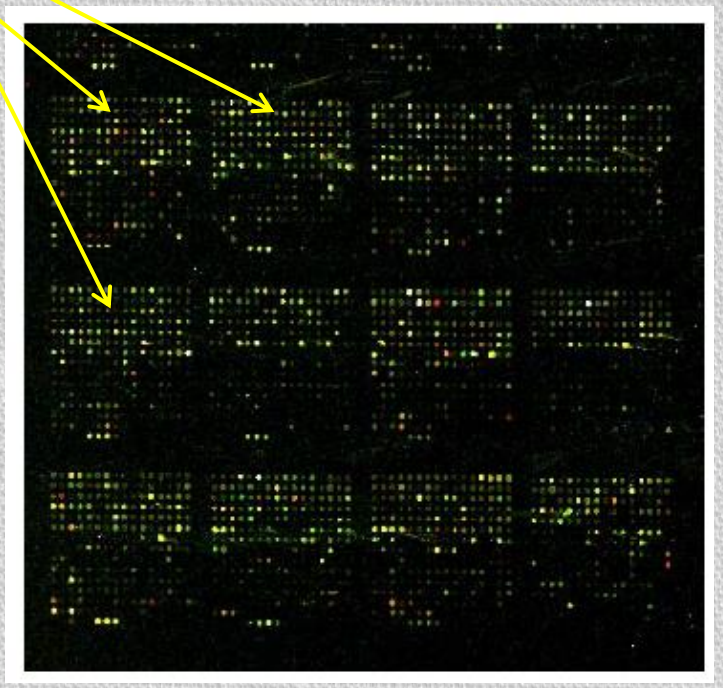
Analiza obrazów – identyfikacja punktów

- Struktura siatki przygotowana jest zazwyczaj przez producenta, lub generowana indywidualnie dla ręcznie robionych mikromacierzy (pliki GAL)
- Siatka ta jest następnie ręcznie bądź też automatycznie nakładana na obraz

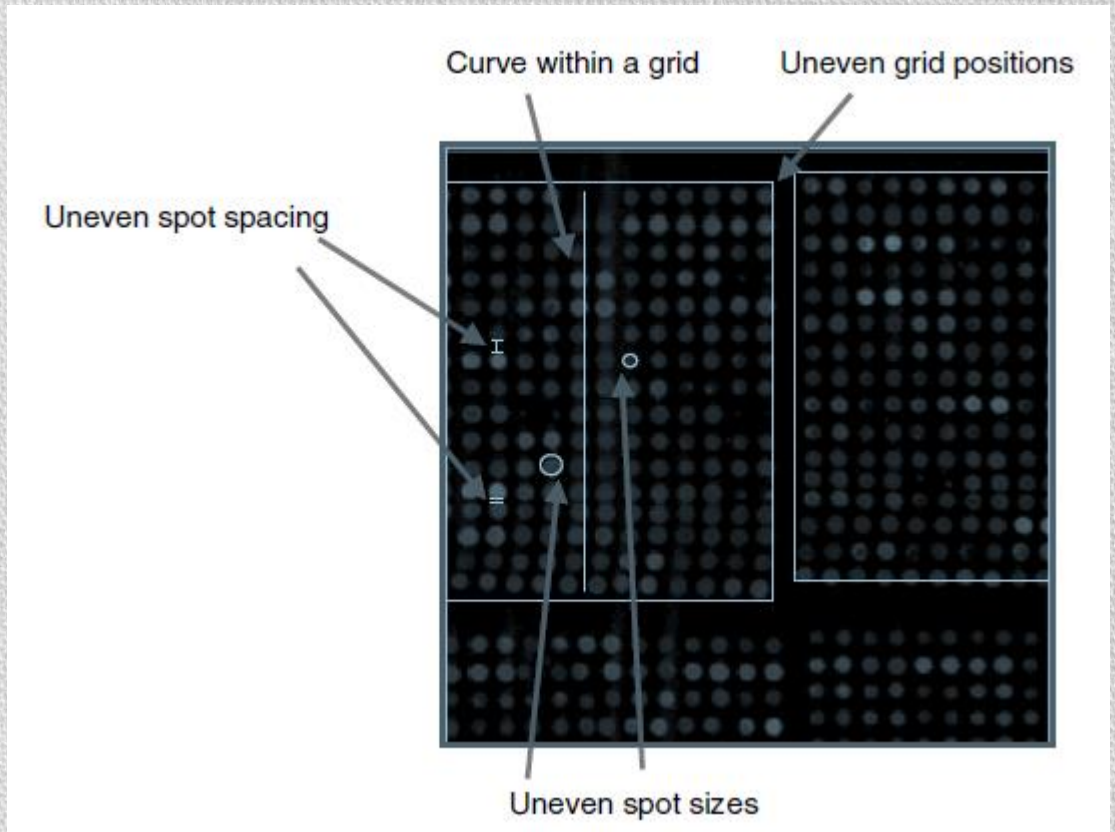
	A	B	C	D	E
1	ATF	1.0			
2	22	5			
3	Type=GenePix ArrayList V1.0				
4	Supplier=Company X				
5	ArrayName=MouseApoptosisProteins 4000				
6	ArrayRevision=2.7				
7	URL=http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Protein&lis				
8	BlockCount=16				
9	Block1=100,100,150,24,180,17,180				
10	Block2=4600,100,150,24,180,17,180				
11	Block3=9100,100,150,24,180,17,180				
12	Block4=13600,100,150,24,180,17,180				
13	Block5=100,4600,150,24,180,17,180				
14	Block6=4600,4600,150,24,180,17,180				
15	Block7=9100,4600,150,24,180,17,180				
16	Block8=13600,4600,150,24,180,17,180				
17	Block9=100,9100,150,24,180,17,180				
18	Block10=4600,9100,150,24,180,17,180				
19	Block11=9100,9100,150,24,180,17,180				
20	Block12=13600,9100,150,24,180,17,180				
21	Block13=100,13600,150,24,180,17,180				
22	Block14=4600,13600,150,24,180,17,180				
23	Block15=9100,13600,150,24,180,17,180				
24	Block16=13600,13600,150,24,180,17,180				
25	Block	Column	Row	Name	ID
26	1	1	1	MAP-1	11139671
27	1	2	1	bad protein	1083224
28	1	3	1	bcl2-like	6753170
29	1	4	1	interleukin-1	2137456
30	1	5	1	caspase 6	6753286



Plik GAL zawiera ID genu oraz jego pozycję na siatce



Macierze – możliwe błędy w nadruku



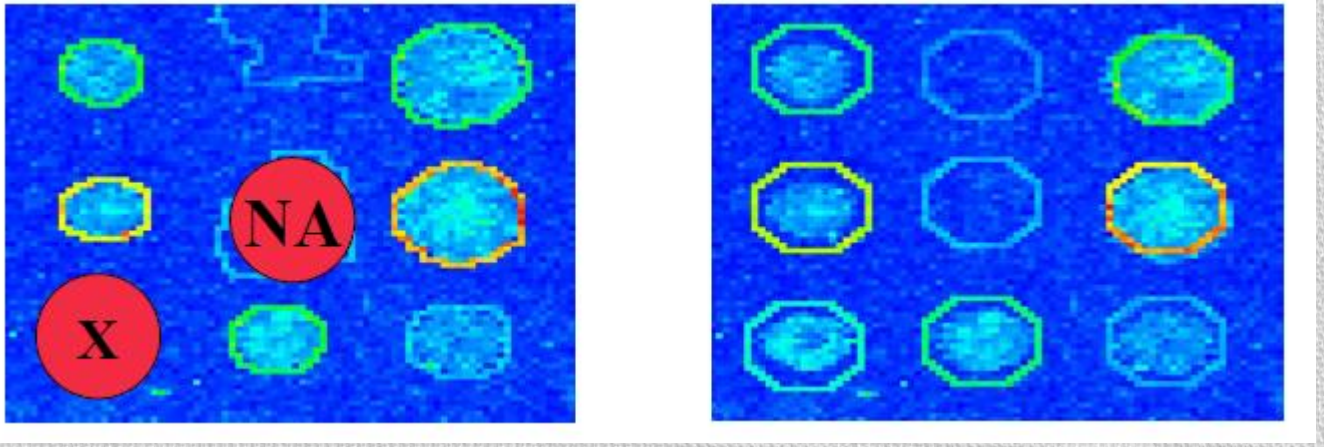
- Nierówne pozycje siatki (igły źle umiejscowione w kasecie)
- Przesunięcie wewnątrz siatki (przesunięcia w trakcie produkcji, płytka nie leży płasko)
- Nierówna przestrzeń między punktami (igła rusza się podczas drukowania, błędne przymocowanie igły)
- Różna wielkość punktów (różna ilość roztworu nadrukowana na płytkę, nierównomierne wysychanie, różna temperatura i wilgotność)

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

Identyfikacja punktów

- Każdy punkt jest rozpoznawany w siatce, a jego wielkość i kształt można dopasować (automatycznie lub ręcznie)
- Punkty mogą być oznaczone jako złe (X) lub nieobecne (NA)



Różne sposoby identyfikacji punktów: okrąg o ustalonej lub zmiennej średnicy, dowolny kształt punktów (można punkty obrysowywać ręcznie)

Fixed circle	ScanAlyze, GenePix, QuantArray
Adaptive circle	GenePix, Dapple
Adaptive shape	Spot, region growing and watershed
Histogram method	ImaGene, QuantArraym DeArray and adaptive thresholding

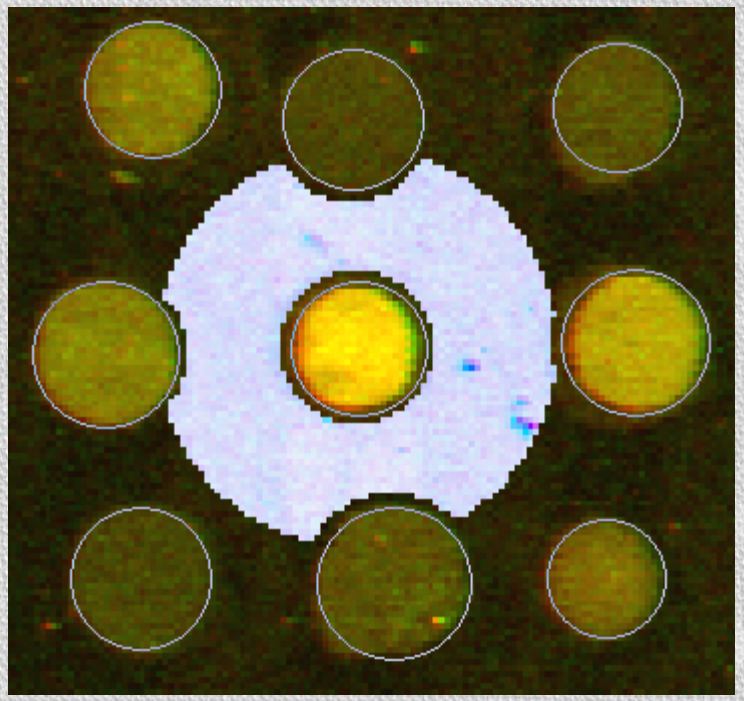
Intensywność punktu

- Całkowita liczba zhybrydyzowanych próbek jest proporcjonalna do intensywności świecenia punktu
- Intensywność punktu = suma intensywności pikseli wewnątrz oznaczonego punktu
- Ponieważ dalsze obliczenia będą się opierały na stosunku między cy5 a cy3, zamiast sumy intensywności pikseli wyznaczana jest średnia lub mediana z intensywności pikseli w punkcie

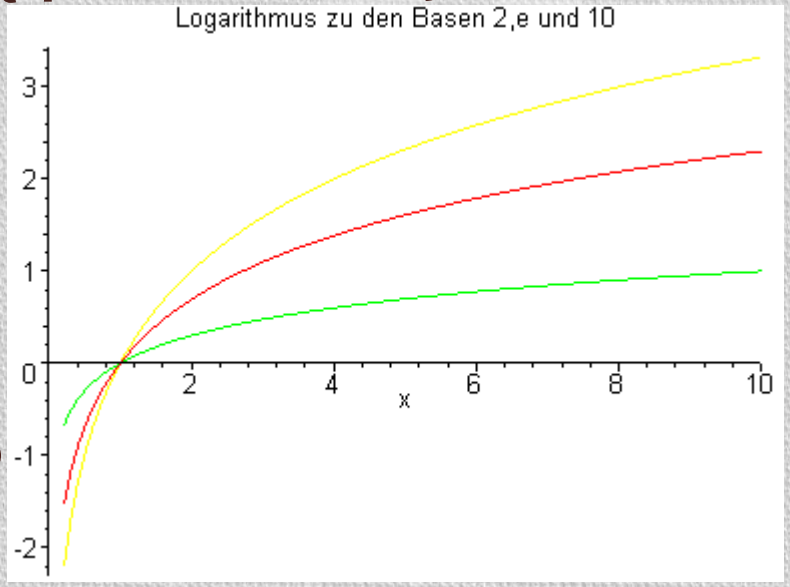
Intensywność świecenia tła

- Na intensywność świecenia punktu składa się również fluorescencja innych chemikaliów takich jak kurz, włókienka osiadające np. podczas przecierania szkiełka
- Sygnał fluorescencyjny w regionie nie zajmowanym przez DNA powinien być inny niż regiony zajmowane przez DNA (w zasadzie najlepiej, gdyby go w ogóle nie było)

Co to jest tło?



Na co trzeba zwrócić uwagę podczas wyboru metod korekcji tła?



Zwykła korekcja tła typu

sygnał kropki – sygnał tła

może doprowadzić do ujemnych warto

Ponieważ dalszą analizę przeprowadza się nie dla intensywności świecenia punktu, ale dla zlogarytmowanego stosunku sygnałów czerwonego i zielonego, nie chcemy dopuszczać ujemnych wartości

$M = \log_2(R/G)$ $A = 1/2\log_2(RG)$

Co robimy w przypadku gdy tło świeci jaśniej niż punkt?

- Usuwamy taki punkt z dalszej analizy (punkt jest uznawany za niewiarygodny)
- Używamy najmniejszą możliwą wartość dla punktu, po zredukowaniu sygnału tła – zazwyczaj jest to ‚1' (zakładamy że gen nie uległ ekspresji, lub tylko w nieznacznym stopniu)
- Używamy bardziej skomplikowanych algorytmów Bayesowskich, w celu estymowania faktycznej intensywności punktu (zakładamy że punkt świeci jaśniej niż tło, a wynik dla intensywności tła jest błędem eksperymentalnym)

Bez korekcji tła (no adjustment)

Udawać że tła w ogóle nie ma i nie trzeba się nim przejmować

$R_B=0$

$G_B=0$

Stałe tło (constant background)

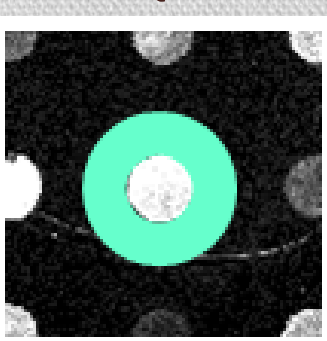
Tło jest stałe dla wszystkich punktów. Od każdego punktu odejmowana jest więc taka sama wartość

- Jeśli na płytce są *negatywne punkty kontrolne* to nie powinny one świecić, a poziom ich intensywności świecenia uznawany jest za szum. Wówczas uśrednia się sygnał dla negatywnych punktów kontrolnych
- Jeśli takich punktów nie ma wówczas aproksymowana jest wartość tła jako trzeci percentyl wartości świecenia punktów

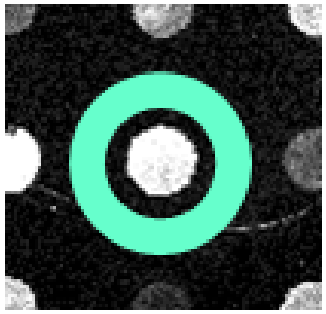
trzeci percentyl to wartość intensywności świecenia poniżej której znajduje się 3% punktów

Tło lokalne (wokół każdego punktu)

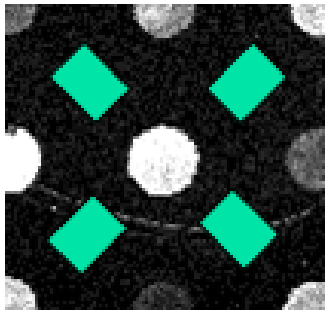
- Intensywność tła jest zależna od regionów znajdującego się wokół punktu
- Wyznaczana jest mediana z wartości pikseli z tego regionu
- Dla większości oprogramowania jest to ustawienie domyślne



ScanAlyze



ImaGene

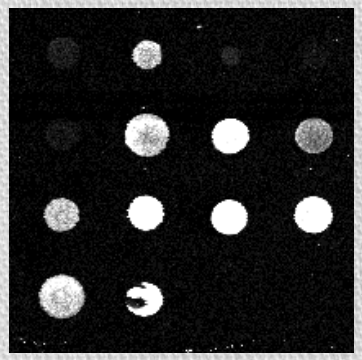


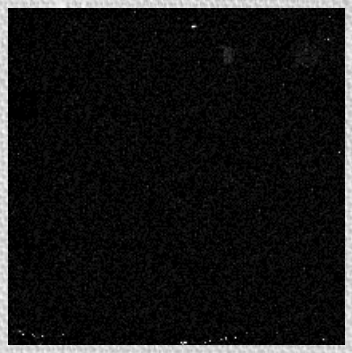
Spot, GenePix

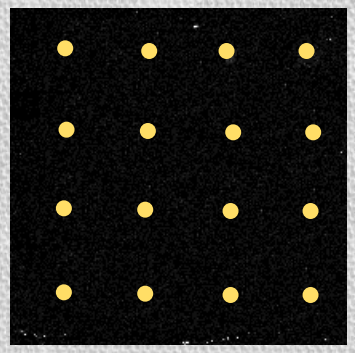
- Jeśli nie są brane pod uwagę piksele znajdujące się zaraz obok punktu, wówczas metoda korekcji tła nie jest tak wrażliwa na sposób zaznaczania punktu
- Wartość tła odejmowana jest od wartości punktu

Morphological opening

- Uzyskane wartości tła są niższe i mniej zróżnicowane niż przy wyznaczaniu tła lokalnego
- Są dwa podstawowe operatory: erozja (*erosion*) i rozszerzanie się (*dilation)* - rankingi
- Erozja: usuwane są wszystkie punkty jak również zbyt jasne piksele
- Rozszerzanie się: używane są kwadratowe okna o wielkości dwóch odległości pomiędzy punktami, z których estymowane jest tło
- W tym celu używana jest transformacja Fouriera







Normexp (+offset)

- Metoda ta stosowana jest do eksperymentów dwu-kolorowych, wzięła początek od metody RMA (robust multi-array average) do przetwarzania danych Affymetrix'owych
- Każdy kolor dla każdego eksperymentu rozpatrywany jest oddzielnie
- Metoda aproksymacji podobieństwa do estymowania parametrów, oparta na aproksymacji siodłowej

Normexp (+offset)

The joint density of B and S is just the product of densities

$$f_{B,S}(b, s; \mu, \sigma, \alpha) = \frac{1}{\alpha} \exp(-s/\alpha)\phi(b; \mu, \sigma^2), \tag{2.2}$$

where $s > 0$ and $\varphi(\cdot)$ is the Gaussian density function. A simple transformation gives the joint density of X and S as

$$f_{X,S}(x, s; \mu, \sigma, \alpha) = \frac{1}{\alpha} \exp\left(\frac{\sigma^2}{2\alpha^2} - \frac{x - \mu}{\alpha}\right)\phi(s; \mu_{S \cdot X}, \sigma^2),$$

where $\mu_{S \cdot X} = x - \mu - \sigma^2/\alpha$. Integrating over s gives the marginal density of X:

$$f_X(x; \mu, \sigma, \alpha) = \frac{1}{\alpha} \exp\left(\frac{\sigma^2}{2\alpha^2} - \frac{x - \mu}{\alpha}\right)[1 - \Phi(0; \mu_{S \cdot X}, \sigma^2)], \tag{2.3}$$

where $\Phi(\cdot)$ is the Gaussian distribution function. Dividing the joint by the marginal gives the conditional density of S given X as

$$f_{S|X}(s|x; \mu, \sigma, \alpha) = \frac{\phi(s; \mu_{S \cdot X}, \sigma^2)}{1 - \Phi(0; \mu_{S \cdot X}, \sigma^2)}$$

for $s > 0$, which is a truncated Gaussian distribution. Our estimate of the signal given the observed intensities is the conditional expectation

$$\mathbb{E}(S|X = x) = \mu_{S \cdot X} + \frac{\sigma^2 \phi(0; \mu_{S \cdot X}, \sigma^2)}{1 - \Phi(0; \mu_{S \cdot X}, \sigma^2)}. \tag{2.4}$$

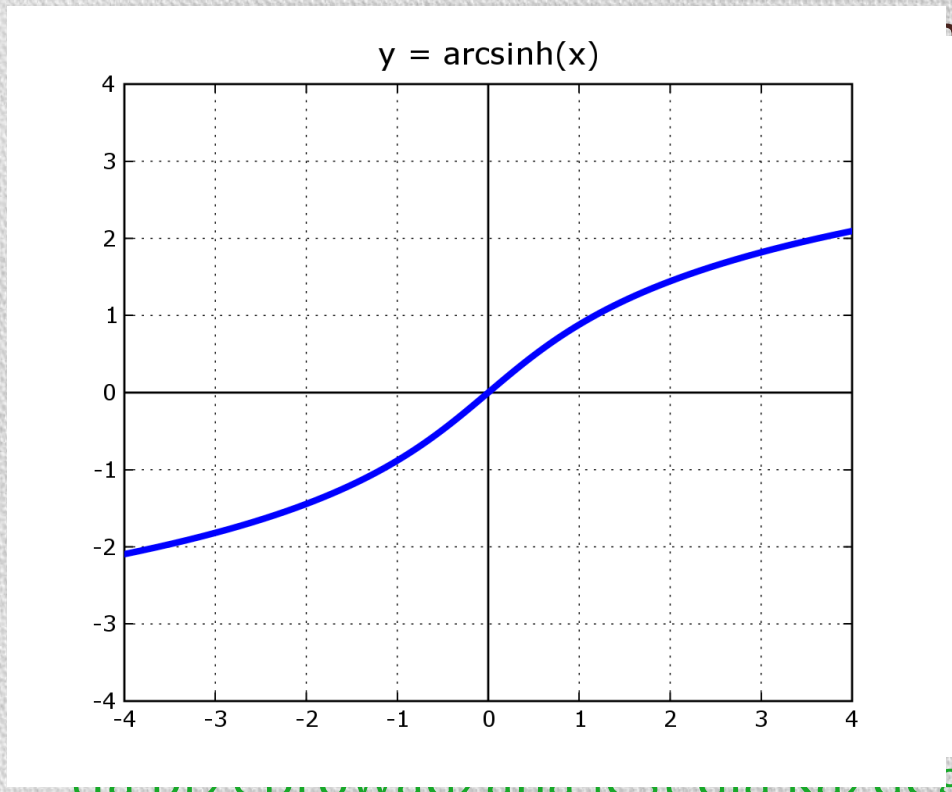
Normexp (+offset)

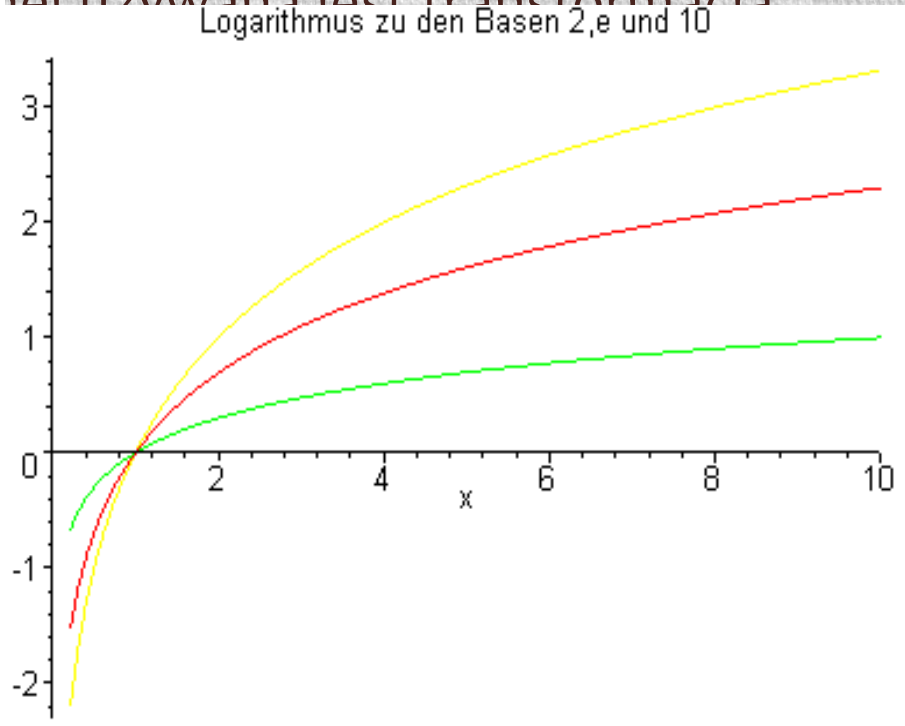
- Metoda ta stosowana jest do eksperymentów dwu-kolorowych, wzięła początek od metody RMA (robust multi-array average) do przetwarzania danych Affymetrix'owych
- Każdy kolor dla każdego eksperymentu rozpatrywany jest oddzielnie
- Metoda aproksymacji podobieństwa do estymowania parametrów, oparta na aproksymacji siodłowej
- Wszystkie wartości intensywności R i G są dodatnie, a następnie zamieniane na stosunki logarytmiczne $M = \log_2(R/G)$ i $A = 1/2\log_2(RG)$
- Offset – przesunięcie wartości intensywności od zera, ma na celu służyć stabilizacji wariancji log-ratio dla niskich intensywności, zazwyczaj $k=50$

$$M = \log_2[(R-k)/(G-k)]$$

Vsn – variance stabilization method

- Dane są kalibrowane dla każdego kanału pomiędzy różnymi mikromacierzami





ej używana jest transformacja

dla przeprowadzana jest dla każdego eksperymentu oddzielnie

Limma – pakiet ‚R' do analizy danych mikromacierzowych

Wszystkie metody korekcji tła dostępne są w pakiecie *limma* w ‚R' jako parametry we funkcji ***backgroundCorrect***

lub też jako parametr przy metodzie normalizacji

Metoda vsn dostępna jest w pakiecie *limma* we funkcji *normalizeBetweenArrays* z parametrem ‚method=vsn'

‚Donuts' - oponki

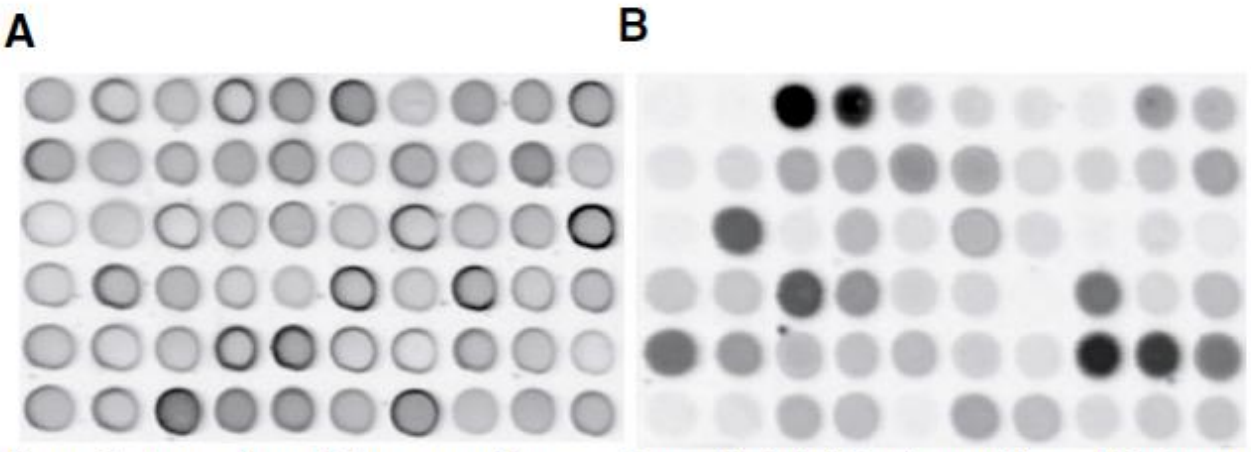


Figure 5. Comparison of the scanned images after a 1-h hybridization on Mouse 5K microarray slides (Clontech, Mountain View, CA, USA) with (A) a 23-base-long fluorescent oligonucleotide and (B) an overnight hybridization with a total RNA sample containing 500-base-long target strands. Both images are identical regions on the microarray slide and contain identical spots (from the same spotting batch), with only the applied target sample differing in length. All hybridizations were performed according to the techniques described in the Materials and Methods section. To check whether this difference is due to the size of the target molecules and not to the hybridization time, overnight hybridizations with the M13 targets gave identical results as the 1-h hybridizations. On the other hand, 1-h hybridizations with the complex target mixtures resulted in very low intensity values that cannot be interpreted.

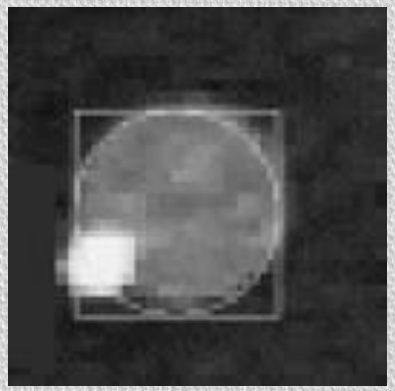
Research Reports

Diffusion limitation: a possible source for the occurrence of doughnut patterns on DNA microarrays

Kris Pappaert[1], Heidi Ottevaere[1], Hugo Thienpont[1], Paul Van Hummelen[2], and Gert Desmet[1]

BioTechniques 41:609-616 (November 2006)
doi 10.2144/000112293

Kurz na punkcie



Flagowanie punktów

- **Błędny kształt** – odchylenie standardowe pikseli jest dużo większe od średniej pikseli
- **Wartość ujemna** – sygnał pochodzący z punktu jest mniejszy od sygnału tła
- **Ciemny punkt** – sygnał pochodzący z punktu jest bardzo ciemny
- **Ręczne flagowanie** – za pomocą programu do przetwarzania obrazów

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

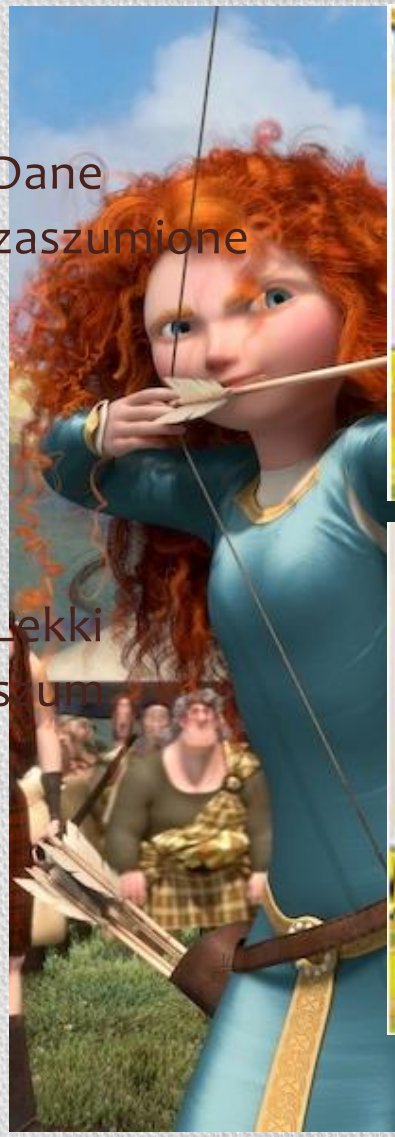
Podsumowanie

- Obraz (skan) mikromacierzy to są dane surowe
- Software do przetwarzania skanów wyznacza numeryczne wartości ekspresji genów z obrazu
- Wybór algorytmu ekstrakcji punktów (sposób wyznaczania kształtu, średnia/mediana/histogram pikseli, sposób wyboru tła, algorytm korekcji tła) ma duży wpływ na wygenerowane dane … oraz dalszą analizę

KONTROLA JAKOŚCI DANYCH
NORMALIZACJA DANYCH

Po co?

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

Oczyszczanie danych i przetwarzanie

- Usuwanie oflagowanych punktów

  Punkty są błędne

- Korekcja tła

  Możemy to zrobić podczas skanowania obrazów lub po wczytaniu danych do ‚R' w czasie normalizacji

- Transformacja liczb (punktów) do logarytmów

Kontrola jakości danych

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

Dane użyte w czasie tego wykładu

- Pakiet Bioconductora – ‚limma'
- Zbiór testowy ‚zebrafish – swirl'

- Dane dla 4 mikromacierzy dwukolorowych

Wczytywanie danych do R

Dane należy ściągnąć dane ze strony:

http://bioinf.wehi.edu.au/limmaGUI/DataSets.html

```
"fish.gal" "swirl.1.spot" "swirl.2.spot" "swirl.3.spot"
"swirl.4.spot" "SwirlSample.txt"
```

```
> library(limma)
> targets <- readTargets("SwirlSample.txt")
> targets
```

	SlideNumber	FileName	Cy3	Cy5	Date
1	81	swirl.1.spot	swirl	wild type	2001/9/20
2	82	swirl.2.spot	wild type	swirl	2001/9/20
3	93	swirl.3.spot	swirl	wild type	2001/11/8
4	94	swirl.4.spot	wild type	swirl	2001/11/8

```
> RG <- read.maimages(targets$FileName, source="spot")
```

Wczytywanie danych do R

```
> RG
An object of class "RGList"
$R
swirl.1 swirl.2 swirl.3 swirl.4
[1,] 19538.470 16138.720 2895.1600 14054.5400
[2,] 23619.820 17247.670 2976.6230 20112.2600
[3,] 21579.950 17317.150 2735.6190 12945.8500
[4,] 8905.143 6794.381 318.9524 524.0476
[5,] 8676.095 6043.542 780.6667 304.6190
8443 more rows ...
$G
swirl.1 swirl.2 swirl.3 swirl.4
[1,] 22028.260 19278.770 2727.5600 19930.6500
[2,] 25613.200 21438.960 2787.0330 25426.5800
[3,] 22652.390 20386.470 2419.8810 16225.9500
[4,] 8929.286 6677.619 383.2381 786.9048
[5,] 8746.476 6576.292 901.0000 468.0476
8443 more rows ...
$Rb
swirl.1 swirl.2 swirl.3 swirl.4
[1,] 174 136 82 48
[2,] 174 133 82 48
[3,] 174 133 76 48
[4,] 163 105 61 48
[5,] 140 105 61 49
8443 more rows ...
$Gb
swirl.1 swirl.2 swirl.3 swirl.4
[1,] 182 175 86 97
[2,] 171 183 86 85
[3,] 153 183 86 85
[4,] 153 142 71 87
[5,] 153 142 71 87
8443 more rows ...
$targets
SlideNumber FileName Cy3 Cy5 Date
1 81 swirl.1.spot swirl wild type 2001/9/20
2 82 swirl.2.spot wild type swirl 2001/9/20
3 93 swirl.3.spot swirl wild type 2001/11/8
4 94 swirl.4.spot wild type swirl 2001/11/8
$source
[1] "spot"
```

Wczytywanie danych do R

```
> RG
An object of class "RGList"
$R
swirl.1 swirl
[1,] 19538.4
[2,] 23619.8
[3,] 21579.950 17317.150 2735.6190 12945.8500
[4,] 8905.143 6794.381 318.9524 524.0476
[5,] 8676.095 6043.542 780.6667 304.6190
8443 more rows ...
$G
swirl.1 swirl
[1,] 22028.2
[2,] 25613.200 21438.960 2787.0330 25426.5800
[3,] 22652.390 20386.470 2419.8810 16225.9500
[4,] 8929.286 6677.619 383.2381 786.9048
[5,] 8746.476 6576.292 901.0000 468.0476
8443 more rows ...
$Rb
swirl.1 swirl.2 swirl.3 swirl.4
[1,] 174 136 82 48
[2,] 174 133 82 48
[3,] 174 133 76 48
[4,] 163 105 61 48
[5,] 140 105 61 49
8443 more rows ...
$Gb
swirl.1 swirl.2 swirl.3 swirl.4
[1,] 182 175 86 9
[2,] 171 183 86 8
[3,] 153 183 86 85
[4,] 153 142 71 87
[5,] 153 142 71 87
8443 more rows ...
$targets
SlideNumber FileName Cy3 Cy5 Date
1 81 swirl.1.spot swirl wild type 2001/9/20
2 82 swirl.2.spot wild type swirl 2001/9/20
3 93 swirl.3.spot swirl wild type 2001/11/8
4 94 swirl.4.spot
$source
[1] "spot"
```

Macierz z intensywnościami sygnału czerwonego cy5

Macierz z intensywnościami sygnału zielonego cy3

Macierz z intensywnościami tła sygnału czerwonego cy5

Macierz z intensywnościami tła sygnału zielonego cy3

Schemat eksperymentu

Jakim programem odczytano intensywności kropek

Wczytywanie danych do R

```
> RG$genes <- readGAL("fish.gal")
> RG$genes[1:5,]
```

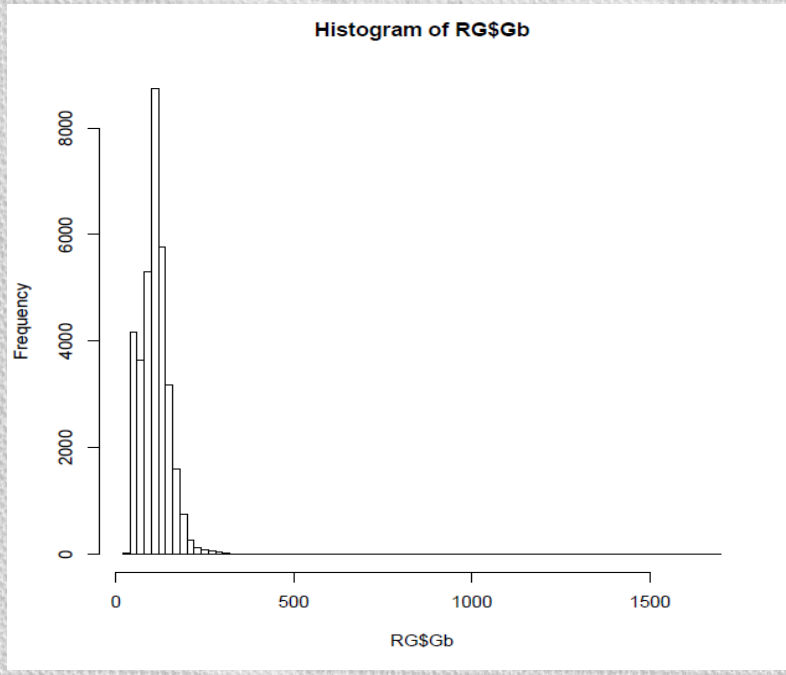
	Block	Row	Column	ID	Name
1	1	1	1	control	geno1
2	1	1	2	control	geno2
3	1	1	3	control	geno3
4	1	1	4	control	3XSSC
5	1	1	5	control	3XSSC

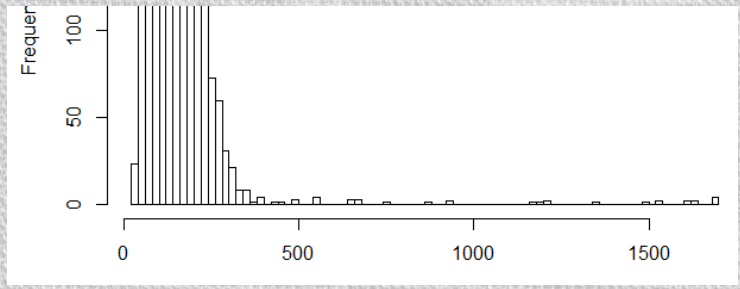
```
Ustalenie siatki 4x4x22x24
> RG$printer <- getLayout(RG$genes)

$ngrid.r
[1] 4
$ngrid.c
[1] 4
$nspot.r
[1] 22
$nspot.c
[1] 24
attr(,"class")
[1] "PrintLayout"
```

Histogram

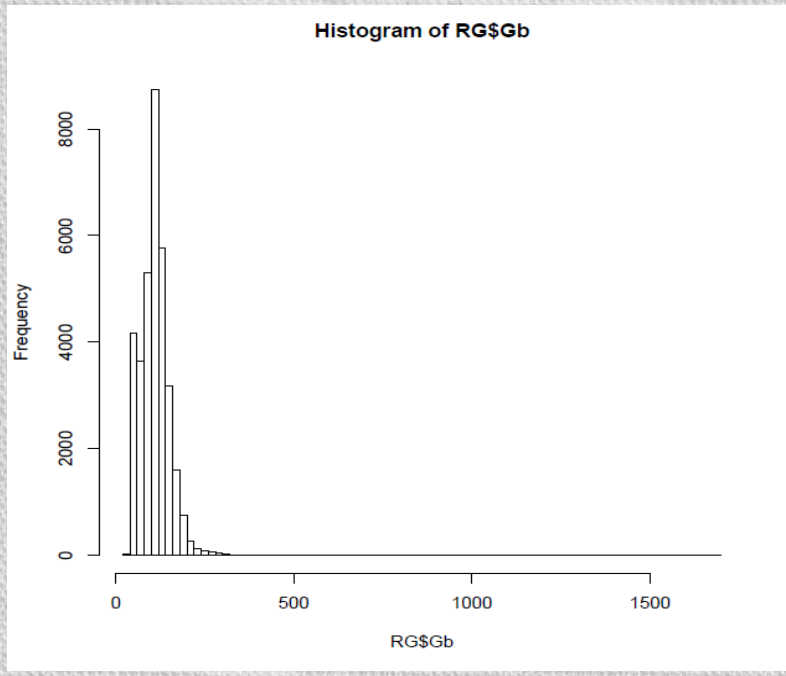
- Histogram – przedstawienie rozkładu intensywności sygnałów genów dla każdej próbki oddzielnie
- Zazwyczaj obserwuje się **unimodalną funkcję rozkładu** (z jednym ekstremum)
- Obecność wielu szczytów na histogramie oznacza zazwyczaj artefakt eksperymentalny
- Większość genów ma słabą intensywność, co 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

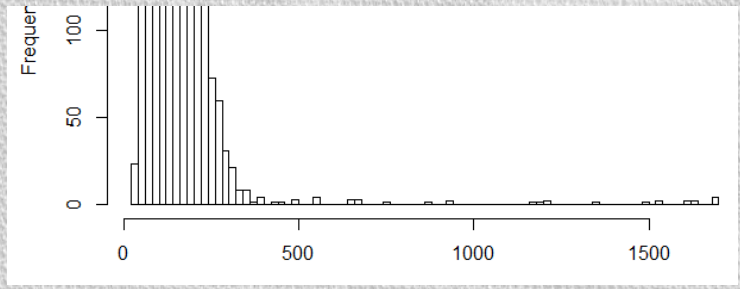




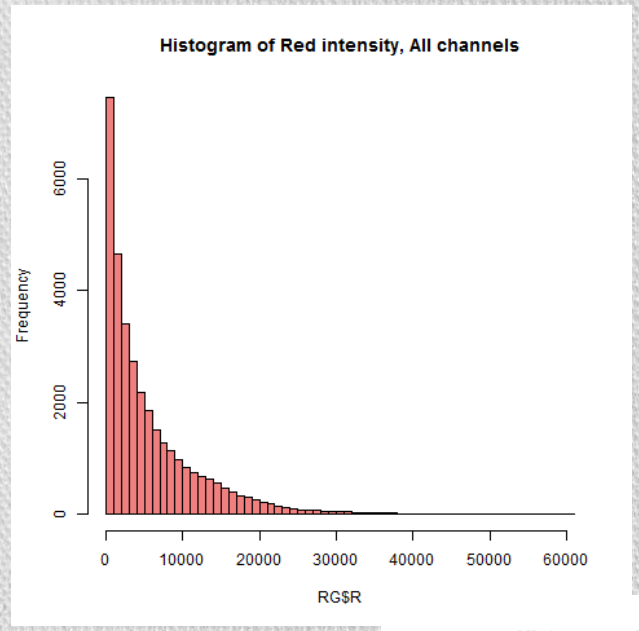
Histogram

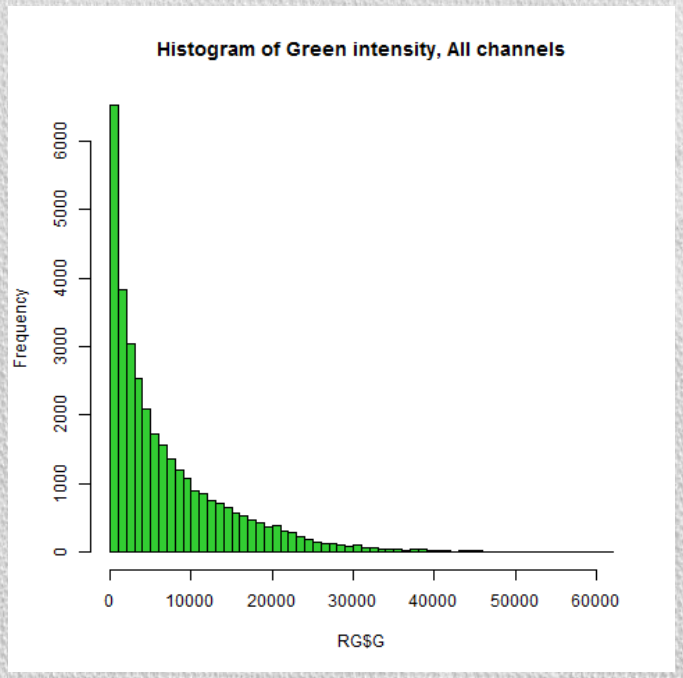
```
> plot(hist(RG$Gb[,1], breaks=50),
main=„Histogram of RG$Gb",
ylim=c(0,200), col=„white")
```

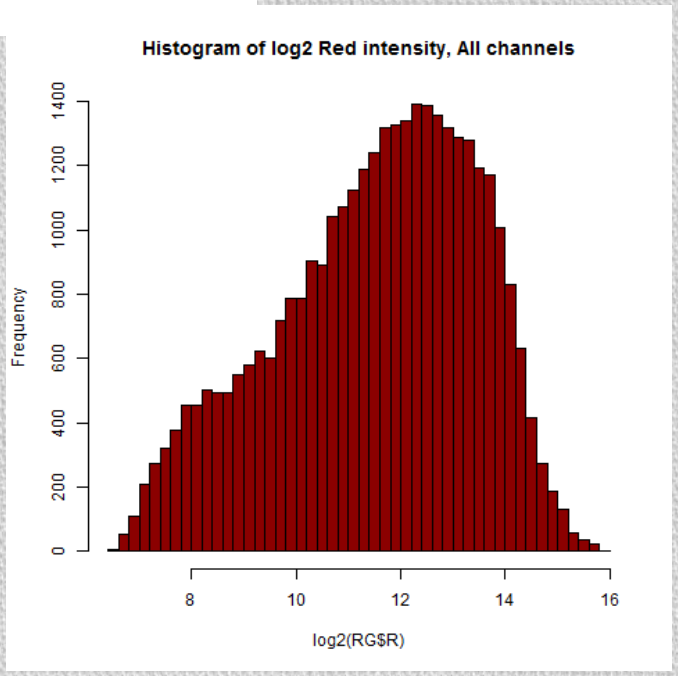


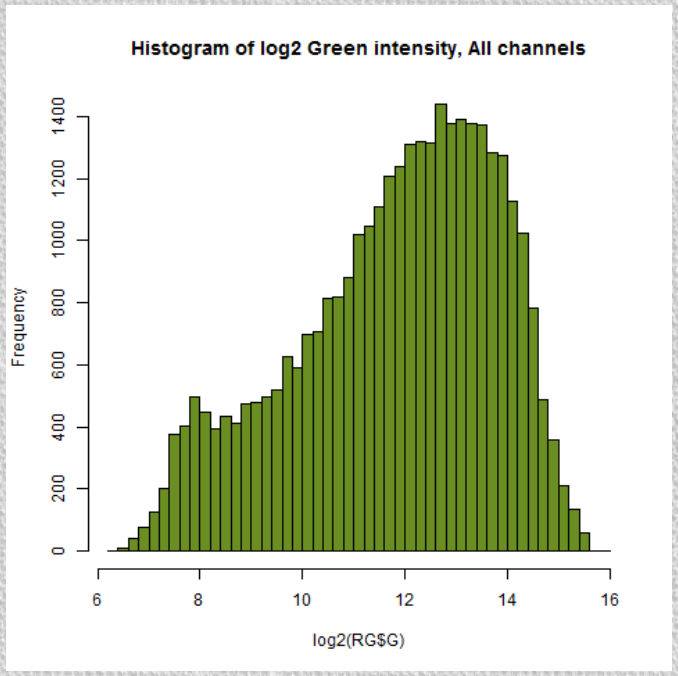


Histogram


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

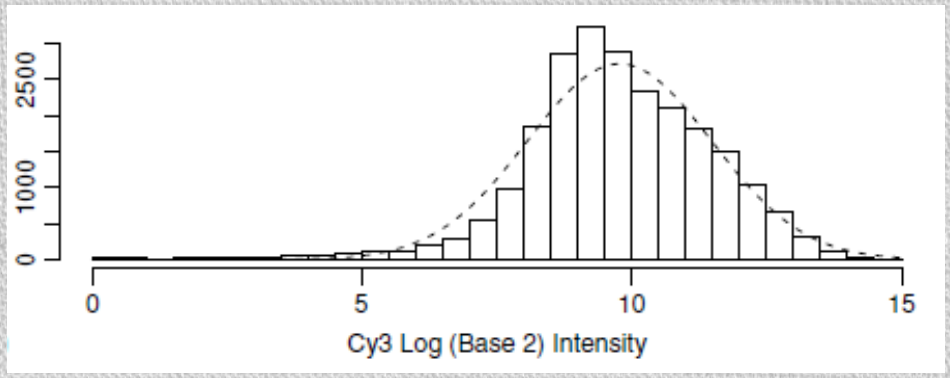

Histogram of Green intensity, All channels
Frequency
0
1000
2000
3000
4000
5000
6000
0
10000
20000
30000
40000
50000
60000
RG$G


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)

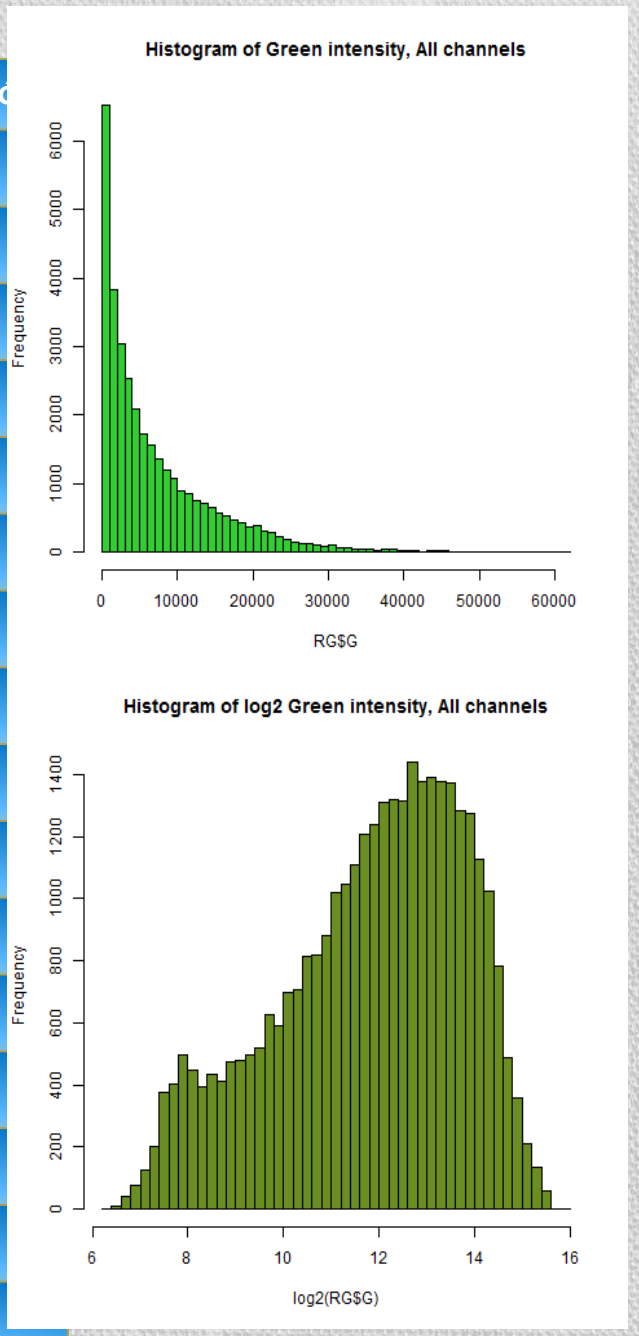
Transformacja do logarytmów

- Wartości intensywności powinny być umiarkowanie rozrzucone na całej skali
- Różnorodność powinna być stała dla każdego poziomu intensywności
- Dystrybucja błędów eksperymentalnych powinna być w przybliżeniu normalna
- Dystrybucja intensywności powinna mieć w przybliżeniu kształt dzwonowaty



Konwersja z logarytmu$_2$ do Intensywności i z powrotem

Log_2 (intensywność)	Intensywność	Intensywność	Log_2(intensywność
0	1	1	0
1	2	2	1
2	4	5	2.32
3	8	10	3.32
4	16	20	4.32
5	32	50	5.64
6	64	100	6.64
7	128	200	7.74
8	256	500	8.97
9	512	1000	9.97
10	1024	2000	10.97
11	2048	5000	12.29
12	4096	10000	13.29
13	8192	20000	14.29
14	16384	50000	15.61
15	32768		



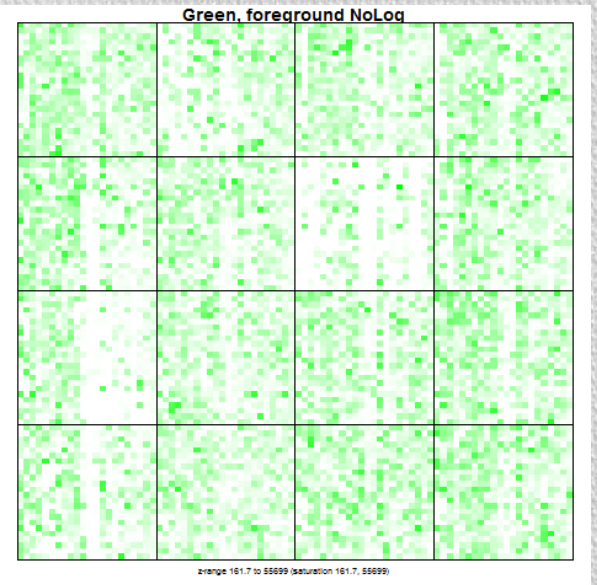
Jak rozumieć stosunek logarytmów w mikromacierzach dwukolorowych?

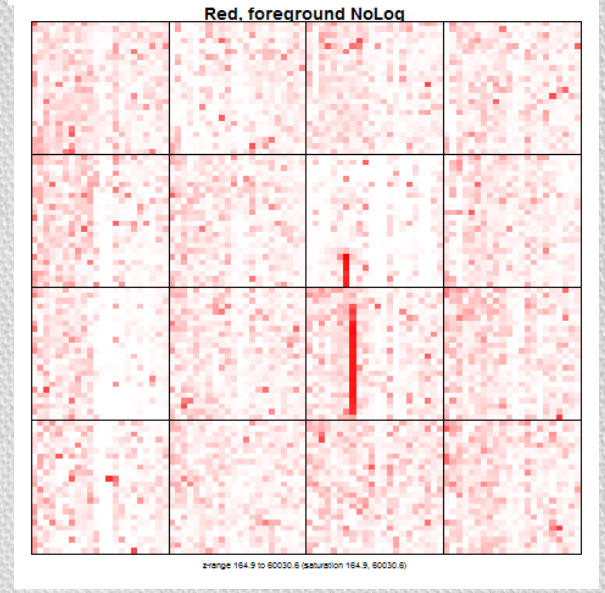
Stosunek surowych intensywności pomiędzy Cy5 i Cy3 jest transformowany do różnicy pomiędzy logarytmami intensywności dla kanałów Cy5 i Cy3

$$\log_2(R/G) = \log_2(R) - \log_2(G)$$

Ekspresja genu	Stosunek $\log_2(R/G)$
4-krotnie **podwyższona** ekspresja	+2
3-krotnie **podwyższona** ekspresja	+1.58
2-krotnie **podwyższona** ekspresja	+1
brak różnicy w ekspresji genów	0
1,5-krotnie **obniżona** ekspresja	-0.58
2-krotnie **obniżona** ekspresja	-1

Przestrzenny rozkład intensywności kolorów





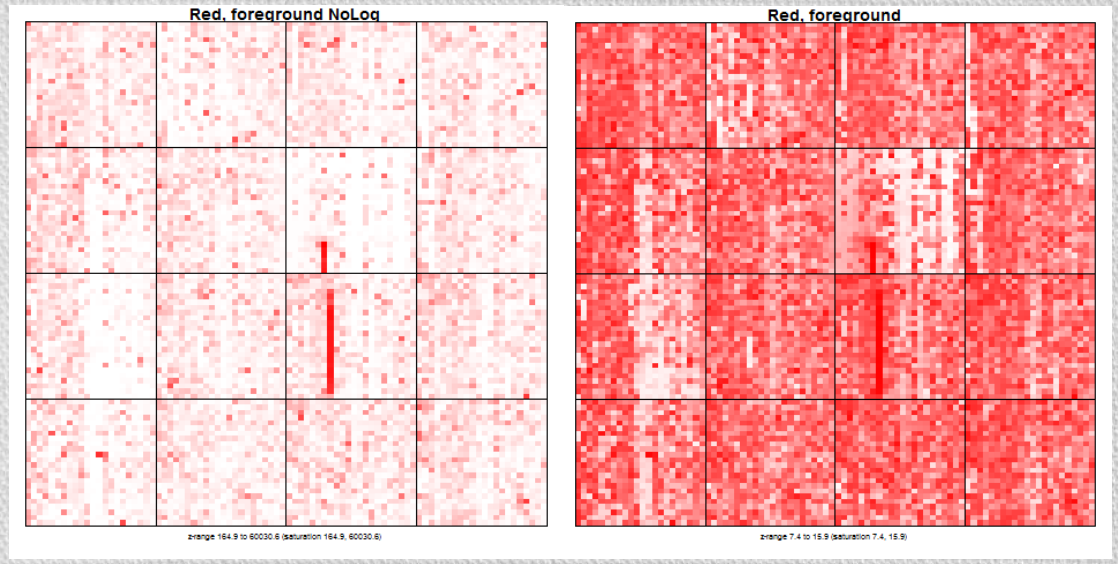
Na obrazkach widać rozkład intensywności kolorów punktów:

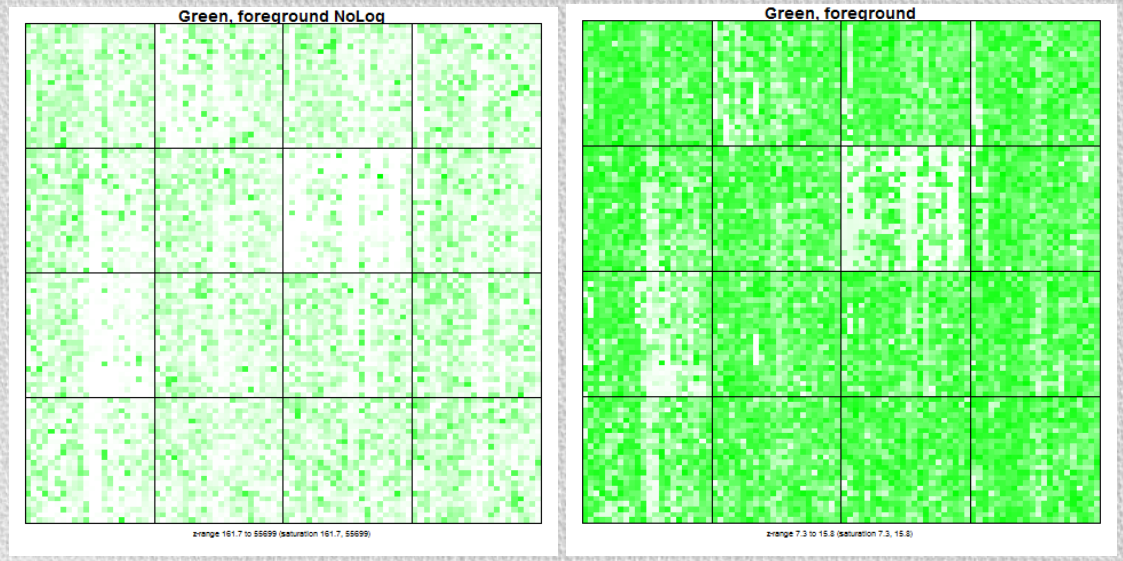
- jest kilka kropek bardzo intensywnych
- jest bardzo dużo kropek o niskiej intensywności, lecz ciężko je odróżnić między sobą

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





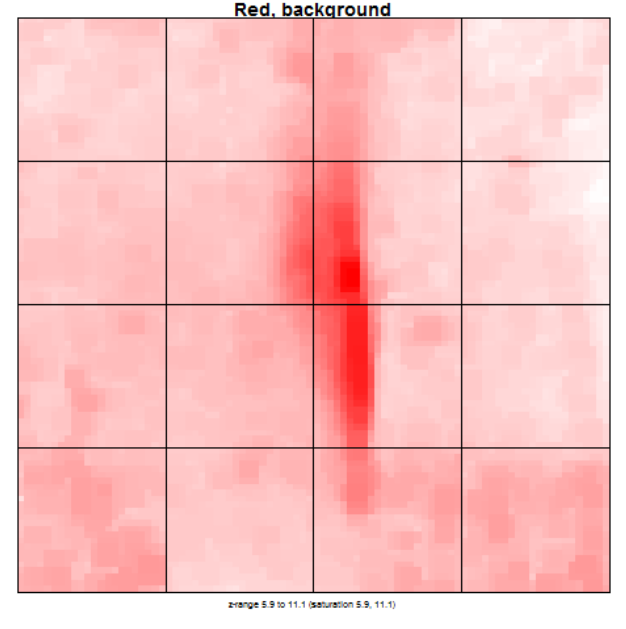
Dlaczego lubimy skalę logarytmiczną?

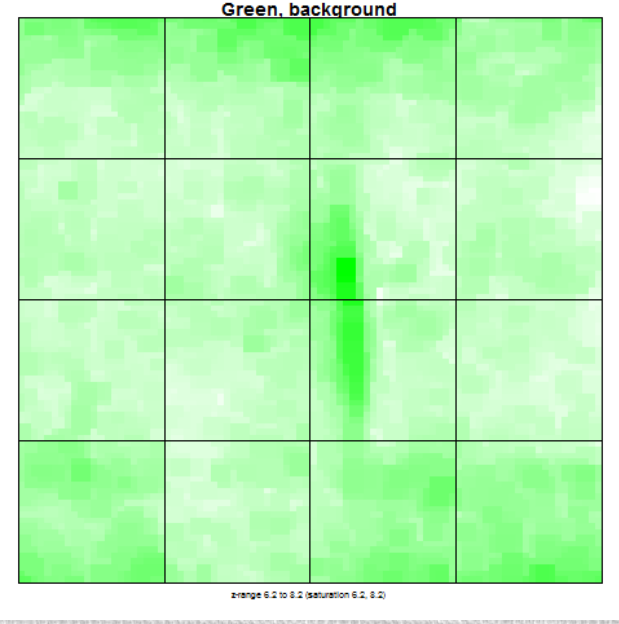
Większość punktów ma dość niską intensywność, która jest rozróżnialna dopiero na skali logarytmicznej ☺

Przestrzenny rozkład intensywności kolorów

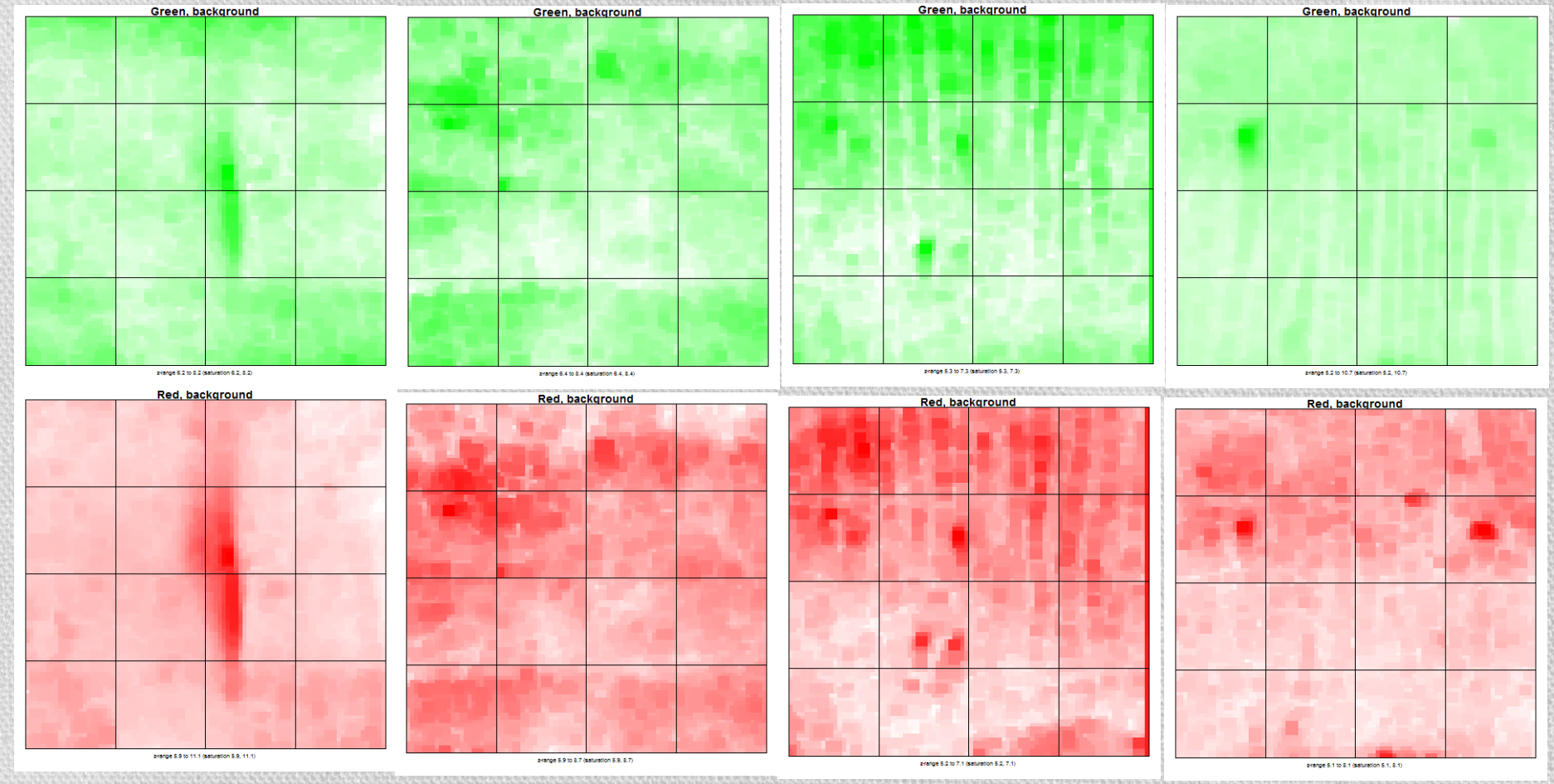
Jak wygląda rozkład intensywności kolorów tła?

Rysę widać wyraźnie na obu mikromacierzach



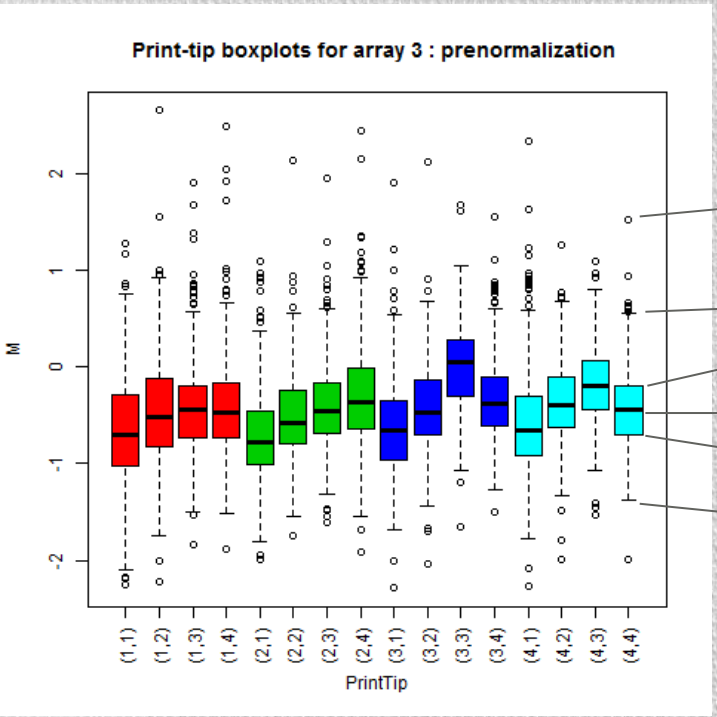


Rozkład intensywności kolorów tła – cały eksperyment



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.



$M = \log_2(R/G)$

outliers, obserwacje odstające

maksymalna wartość

1. kwartyl

mediana

3. kwartyl

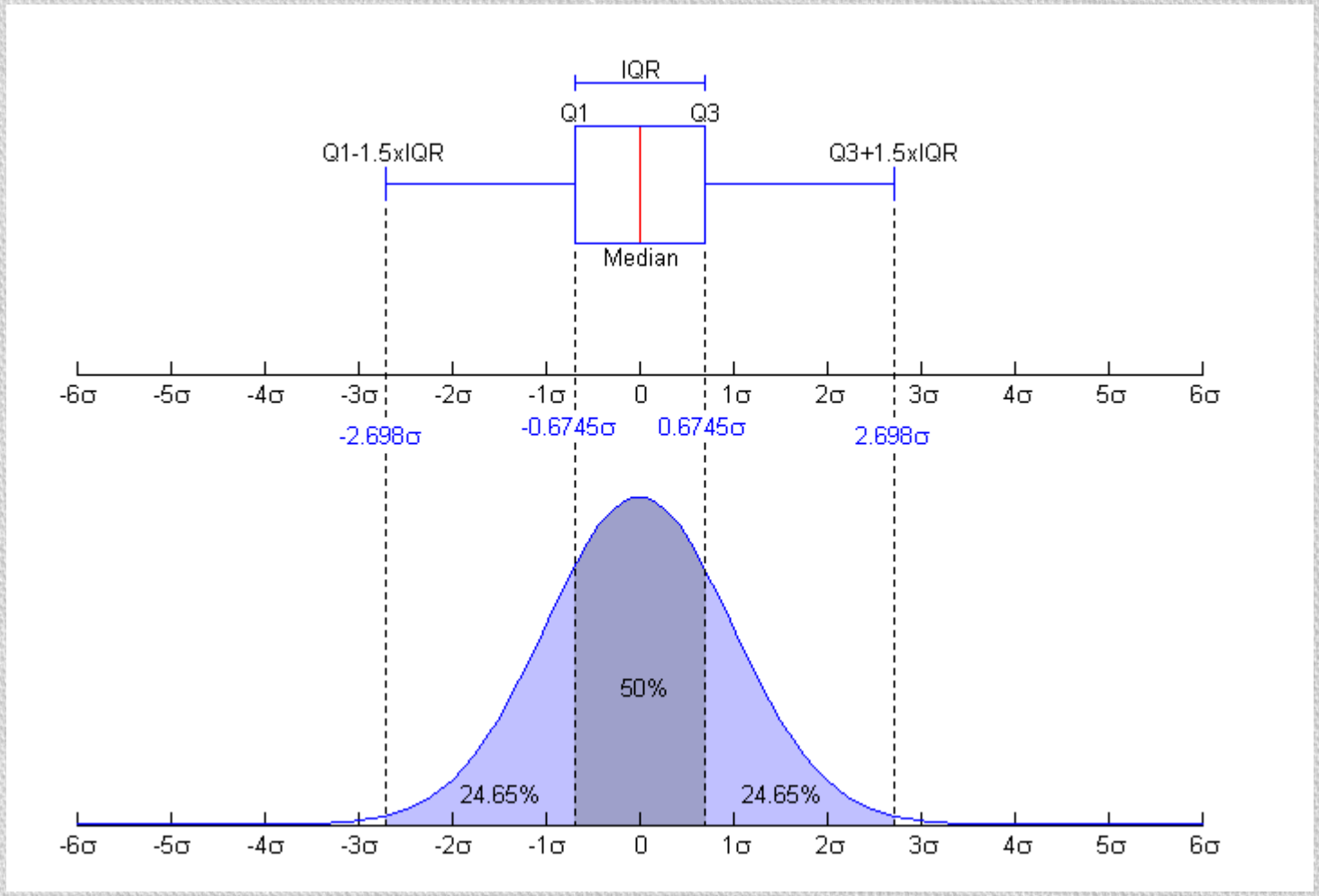
minimalna wartość

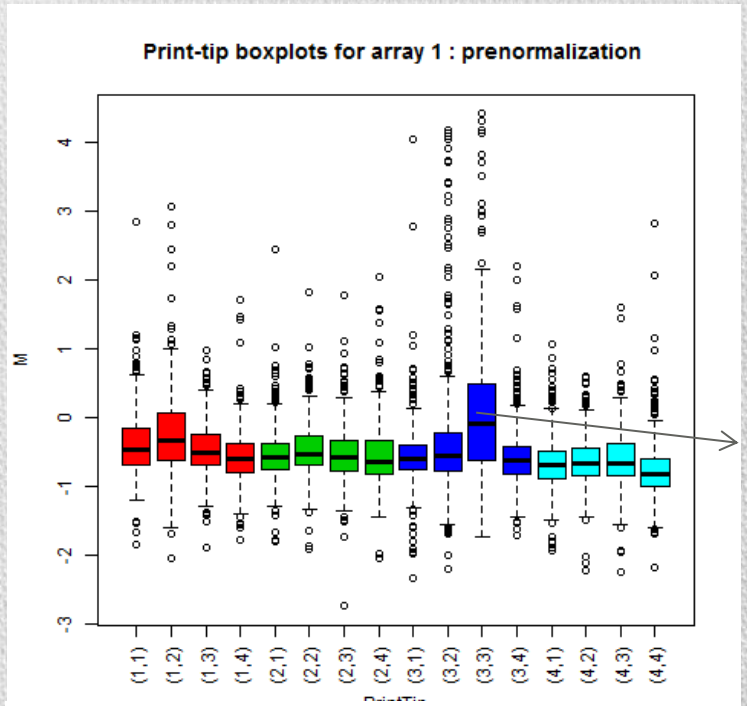
50% obserwacji

Kwantyle:

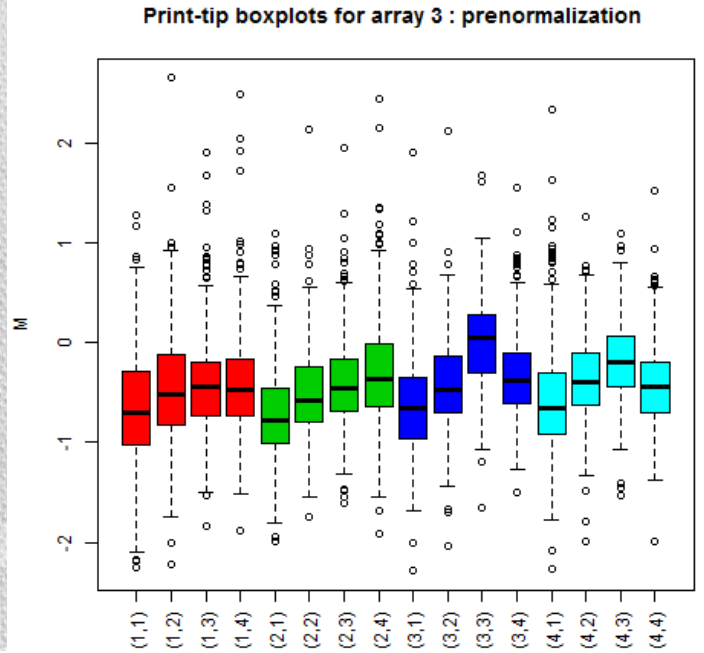
- rzędu 1/2 - mediana
- rzędu 1/4, 2/4, 3/4 - kwartyle
- rzędu 1/10, 2/10, ... 9/10 decyle
- rzędu 1/100, 2/100, ... 99/100 percentyle

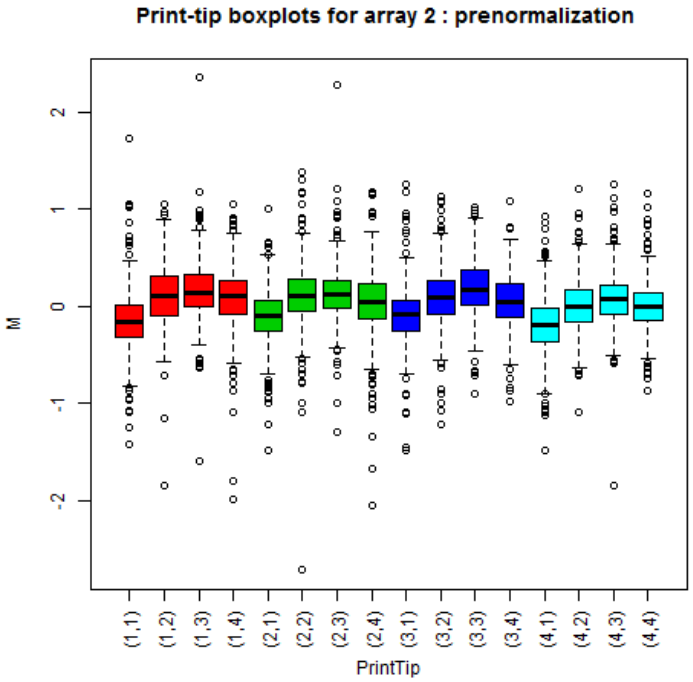
Boxtplot cd.


IQR
Q1
Q3
Q1-1.5xIQR
Q3+1.5xIQR
Median
-6σ
-5σ
-4σ
-3σ
-2σ
-1σ
0
1σ
2σ
3σ
4σ
5σ
6σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
50%
24.65%
24.65%



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