

# Bioinformatyka dla studiów podyplomowych

Adrian Kania<sup>1</sup>

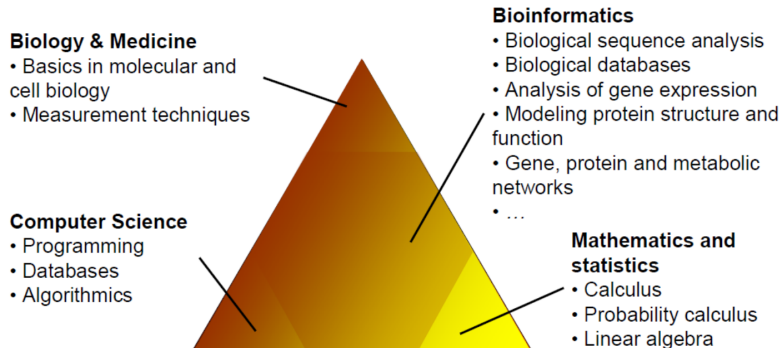
<sup>1</sup>Zakład Biofizyki Obliczeniowej i Bioinformatyki

2022/2023

# Czym jest bioinformatyka?

- The science of information and information flow in biological systems, especially of the use of computational methods in genetics and genomics. (Oxford English Dictionary)
- The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information. (Fred J. Tekaia)
- "I do not think all biological computing is bioinformatics, e.g. mathematical modelling is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, particular genetic information." (Richard Durbin)

Terminy pokrewne: Biologia obliczeniowa, Biometria, Biologia matematyczna, Biologia systemów.



# Umiejętności

Dlaczego bioinformatyka jest ważna?

- Najnowsze techniki eksperymentalne wytwarzają ogromne ilości danych.
- Zaawansowane analizy są konieczne do zrozumienia danych.
- Typowe dane są często wybrakowane (missing values).

Co powinien umieć bioinformatyk?

- statystyka, metody analizy danych,
- programowanie,
- obsługa baz danych,
- modelowanie,
- korzystanie z pakietów obliczeniowych,
- umiejętności komunikacji ;).

Bioinformatycy używają baz danych! Pierwszorzędowych, drugorzędowych czy trzeciorzędowych.

□ GenBank/DDJB/EMBL	<a href="http://www.ncbi.nlm.nih.gov">www.ncbi.nlm.nih.gov</a>	Nucleotide sequences
□ Ensembl	<a href="http://www.ensembl.org">www.ensembl.org</a>	Human/mouse genome
□ PubMed	<a href="http://www.ncbi.nlm.nih.gov">www.ncbi.nlm.nih.gov</a>	Literature references
□ NR	<a href="http://www.ncbi.nlm.nih.gov">www.ncbi.nlm.nih.gov</a>	Protein sequences
□ UniProt	<a href="http://www.expasy.org">www.expasy.org</a>	Protein sequences
□ InterPro	<a href="http://www.ebi.ac.uk">www.ebi.ac.uk</a>	Protein domains
□ OMIM	<a href="http://www.ncbi.nlm.nih.gov">www.ncbi.nlm.nih.gov</a>	Genetic diseases
□ Enzymes	<a href="http://www.expasy.org">www.expasy.org</a>	Enzymes
□ PDB	<a href="http://www.rcsb.org/pdb/">www.rcsb.org/pdb/</a>	Protein structures
□ KEGG	<a href="http://www.genome.ad.jp">www.genome.ad.jp</a>	Metabolic pathways

# Homologia i podobieństwo

Homologi - dwie sekwencje które wyewoluowały od tego samego genu przodka. Oczekujemy, że homologi są do siebie podobne. Jednak, podobieństwo sekwencji nie jest tym samym co homologia.

#mutations

```
0  agtg+ccg+taag+cg+tc
1  agtg+ccg+ttatag+cg+tc
2  agtg+ccg+c+ttatag+cg+tc
4  agtg+ccg+c+ttaaagggcg+tc
8  agtg+ccg+c+ttcaaggggcg+
16 gggccg+ttcatggggg+
32 gcagggcg+cactgagggc+
```

#mutations

```
64  acag+ccg+tcgggctattg
128 cagagcactaccgc
256 cacgag+taagatatagc+
512 +aatcg+gata
1024 acc+ttatctact+tcctggag+
2048 agcgacctgccccaa
4096 caaac
```

Wyróżniamy tutaj Orotologi (powstały wskutek specjacji) oraz Paralogi (powstały wskutek duplikacji).

## Dopasowanie sekwencji

Dopasowanie pomiędzy sekwencjami określa które pozycje odpowiadają sobie.

acgtctag	acgtctag	acgtctag
actctag-	-actctag	ac-tctag

2 matches

5 mismatches

1 not aligned

5 matches

2 mismatches

1 not aligned

7 matches

0 mismatches

1 not aligned

Takie porównania mogą być użyte to:

- szukania relacji ewolucyjnych,
- zidentyfikowania konserwatywnych miejsc,
- zidentyfikowania odpowiadających sobie genów pomiędzy różnymi modelami (np ludzkimi czy mysimi).

**Indel:** insertion or deletion of a base with respect to the ancestor sequence

a	c	g	t	c	t	a	g
-	a	c	t	c	t	a	g

**Mismatch:** substitution (point mutation) of a single base

## Które dopasowanie jest najlepsze?

Na początku trzeba ustalić punktację za match/mismatch i gap. Poniżej przykładowe punkty. Łączny wynik dopasowania to suma punktów za każdą pozycję.

acgtctag

||

actctag-

2 matches

5 mismatches

1 not aligned

$$S = 2*1 + 5*(-1) + 1*(-2) = -5$$

acgtctag

|||||

-actctag

5 matches

2 mismatches

1 not aligned

$$S = 5*1 + 2*(-1) + 1*(-2) = 1$$

acgtctag

|| |||||

ac-tctag

7 matches

0 mismatches

1 not aligned

$$S = 7*1 + 0*(-1) + 1*(-2) = 5$$

# Sekwencje aminokwasowe

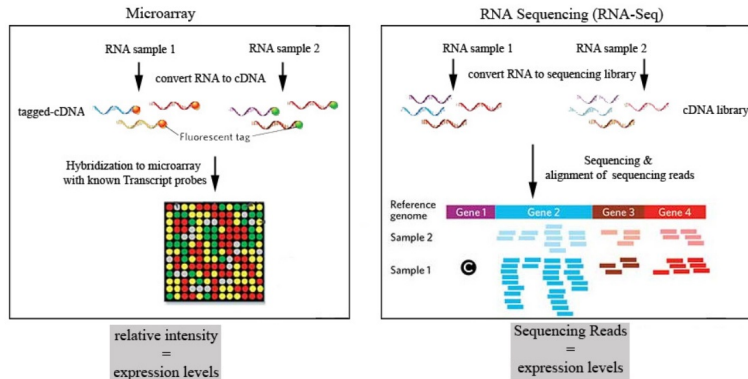
Do porównywania sekwencji aminokwasowych używamy odpowiednich macierzy podobieństwa (np BLOSUM)

A	4																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																				
---	---	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

AABCD...BBCDA  
 DABCD...A.BBCBB  
 BBBCDABA.BCCAA  
 AAACDAC.DCBCDB  
 CCBADAB.DBBDCC  
 AAACAA...BBCCC

# Jak zmierzyć poziom ekspresji genów w komórce?

- Reakcja łańcuchowa polimerazy w czasie rzeczywistym (real time PCR),
- Hybrydyzacja northern (RNA blot),
- Sekwencjonowanie (Sanger, Maxam-Gilbert),
- Mikromacierze,
- Sekwencjonowanie nowej generacji (NGS/RNA-Seq).





# Zastosowanie mikromacierzy

## Wybrane zastosowania mikromacierzy

- Wyznaczanie profili ekspresji genów (w różnych tkankach, stanach chorobowych...).
- Badanie polimorfizmu/detekcja SNP.
- Badanie oddziaływania DNA-białko.
- Badanie nowych leków.
- Identyfikacja procesów komórkowych w które zaangażowane są geny.

# Czy ktoś jeszcze korzysta z mikromacierzy?

[Advanced](#) [Create alert](#) [Create RSS](#)[Save](#)[Email](#)[Send to](#)

MY NCBI FILTERS

157,115 results

RESULTS BY YEAR



kilka tysięcy



## Analysis of Microarray and RNA-seq Expression Profiling Data.

1

Hung JH, Weng Z.

Cite

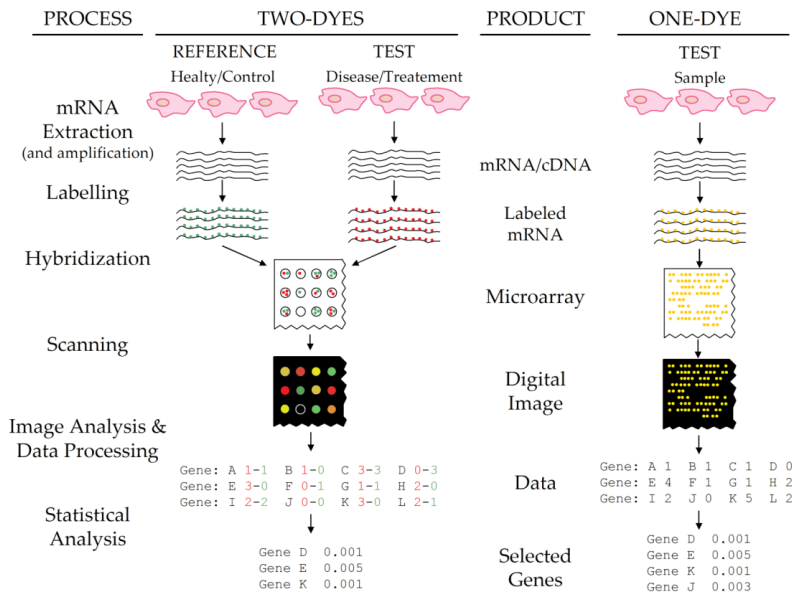
Cold Spring Harb Protoc. 2017 Mar 1;2017(3). doi: 10.1101/pdb.top093104.

PMID: 27574194

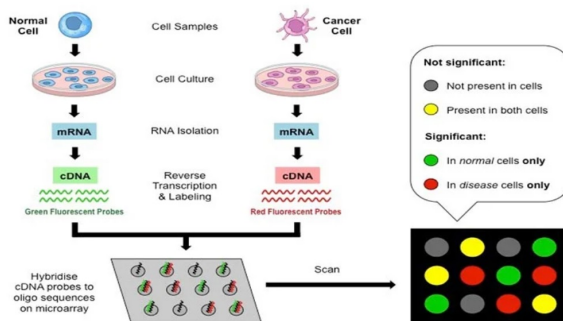
Share

Expression profiling is accomplished by assaying mRNA levels with **microarrays** or next-generation sequencing technologies (RNA-seq). This introduction describes normalization and **analysis** of data generated from **microarray** or RNA-seq experiments....

# Eksperyment mikromacierzowy



# Eksperyment mikromacierzowy



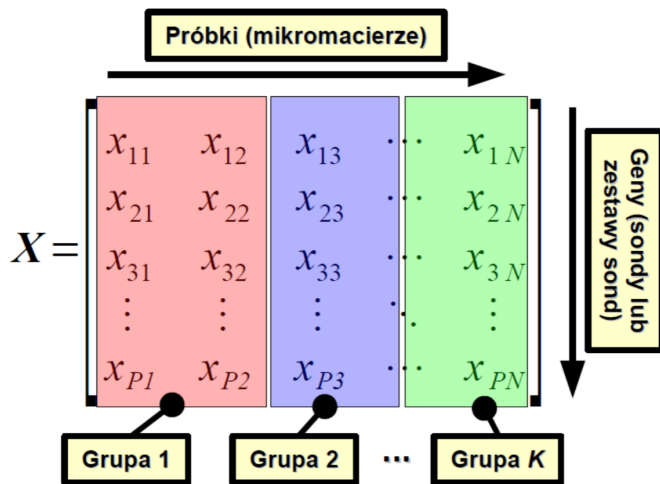
Wyznaczamy stosunek intensywności czerwonej i zielonej fluorescencji (czyli  $x_i = R_i/G_i$ ) dla każdego genu.

- Jeżeli  $x_i = 1$ , to oznacza to, że poziom  $i$ -tego genu był taki sam w próbce kontrolnej i badanej.
- Jeżeli  $x_i > 1$ , to oznacza to, że poziom  $i$ -tego genu był wyższy w próbce badanej w porównaniu do próby kontrolnej.
- Jeżeli  $x_i < 1$ , to oznacza to, że poziom  $i$ -tego genu był niższy w próbce badanej w porównaniu do próby kontrolnej.

## Eksperyment mikromacierzowy

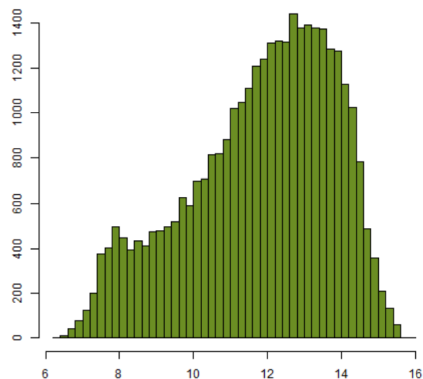
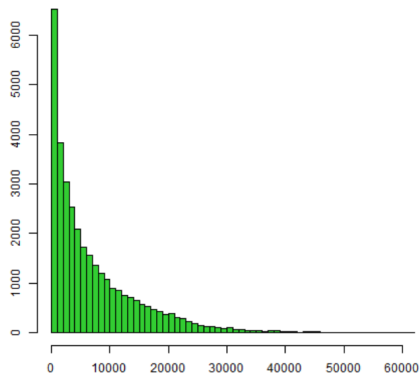
Niech  $N$  oznacza liczbę eksperymentów mikromacierzowych. Każda mikromacierz dostarcza informacje ilościowe o ekspresji  $P$  genów. Formalnie możemy więc rozważane dane przedstawić w postaci zbiorczej macierzy

$$X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{P \times N}.$$



# Transformacja logarytmiczna

Stosujemy transformację  $x_i = \log(x)$  aby wykres intensywności przybrał bardziej symetryczną (Gaussowską) postać.



# Normalizacja

Po co stosujemy normalizację?

- różna ilość materiału w kolejnych eksperymentach
- różna wydajność: ekstrakcji RNA, odwrotnej transkrypcji, znakowania, fotodetekcji

Rodzaje normalizacji


- globalna - wszystkie geny biorą udział w normalizacji
- lokalna - używamy niewielkiej puli genów (np. housekeeping genes)

# Gdzie można znaleźć dane z eksperymentów mikromacierzowych?

https://www.ncbi.nlm.nih.gov/gds

NCBI Resources How To Sign in to NCBI

GEO DataSets GEO DataSets Search Advanced Help

**COVID-19 Information** 

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

## GEO DataSets

This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.

### Getting Started

- [GEO Documentation](#)
- [GEO FAQ](#)
- [About GEO DataSets](#)
- [Construct a Query](#)
- [Download Options](#)

### GEO Tools

- [Submit to GEO](#)
- [Advanced Search](#)
- [DataSet Browser](#)
- [Programmatic Access](#)
- [GEO2R](#)

### More Resources

- [GEO Home](#)
- [GEO Profiles](#)
- [SRA](#)



## Specyfika danych z mikromacierzy

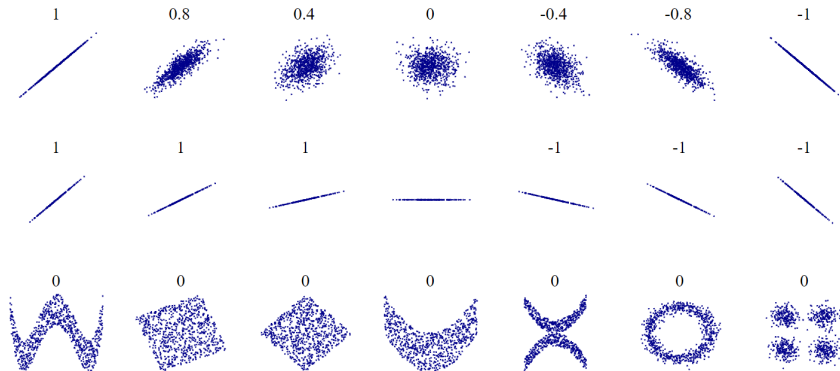
Przy analizie danych mikromacierzowych należy zwrócić szczególną uwagę na następujące zagadnienia:

- Zasadniczy problem:  $P \gg N$  (liczba genów znacznie większa niż liczba mikromacierzy).
- Duży zakres zmienności ekspresji genów (1% genów odpowiada za połowę masy mRNA w komórce).
- Możliwość braku danych - spowodowana m.in. lokalnymi defektami mikromacierzy.
- Duża podatność na zakłócenia i błędy w obróbce laboratoryjnej.

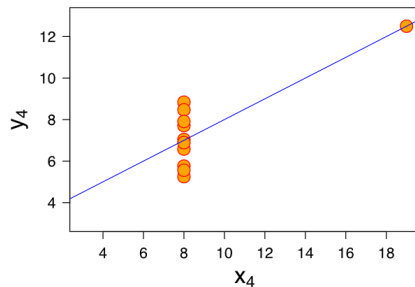
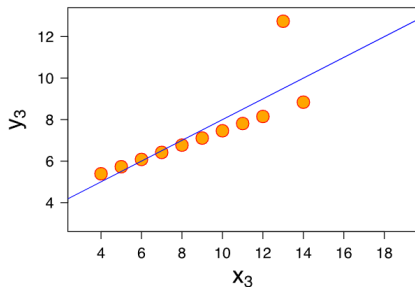
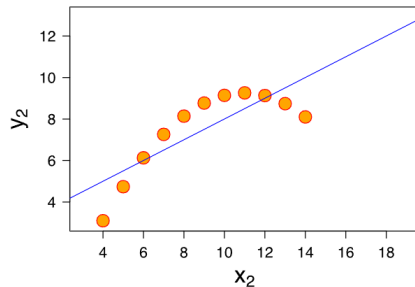
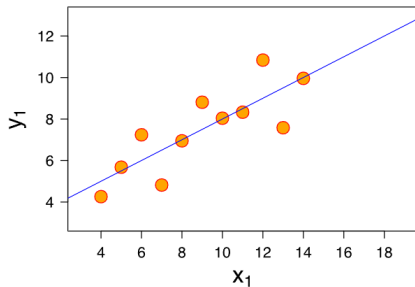
## Analiza korelacji

Jednym z celów przeprowadzenia analizy macierzowej może być wytypowanie genów o podobnym profilu ekspresji w kolejnych grupach (ta sama monotoniczność). Do oceny zależności liniowej między dwoma zmiennymi (tutaj: profilami genów) służy współczynnik korelacji. Najczęstsze stosowane to:

- współczynnik korelacji  $r$  Pearsona,
- współczynnik korelacji rang Spearmana.



# Kwartet Anscombe'a



## Problem testowania wielu hipotez

Założmy, że badamy poziom ekspresji 9000 genów aby sprawdzić skuteczność działania nowego leku. Mamy do dyspozycji grupę kontrolną i grupę badaną oraz wykonujemy test statystyczny dla każdego genu. Niech

- $H_0$  : gen nie uległ zróżnicowanej ekspresji
- $H_1$  : różnica w ekspresji genu jest znacząca

Przyjmijmy  $p_{value} = 0.01$  (jest 1% szansy na zaobserwowanie zróżnicowanej ekspresji przez przypadek). W przypadku badania 9000 genów, nawet gdyby lek nie miał żadnego wpływu, to spodziewamy się że dla 90 genów ich  $p_{value}$  będzie mniejsze niż 0.01.

## Problem testowania wielu hipotez

Przyjmijmy  $p_{value} = 0.01$  (jest 1% szansy na zaobserwowanie zróżnicowanej ekspresji przez przypadek). W przypadku badania 9000 genów, nawet gdyby lek nie miał żadnego wpływu, to spodziewamy się że dla 90 genów ich  $p_{value}$  będzie mniejsze niż 0.01.

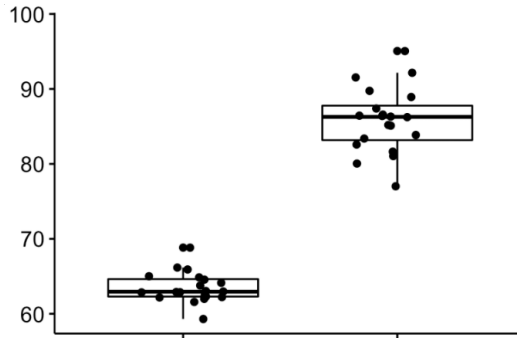
### Poprawka Bonferroniego

- wyznaczamy wartości  $p_i$  dla każdego genu,  $i = 1, 2, \dots, n$ .
- wyznaczamy  $p'_i = \min(np_i, 1)$
- wybieramy te geny, dla których  $p'_i < \alpha$ , gdzie  $\alpha$  - założony poziom istotności.

# Testy statystyczne

- Testy parametryczne (dla danych z rozkładem normalnym) - t test dla dwóch prób
- Testy nieparametryczne (mają mniej założeń) - test Wilcoxona dla par obserwacji, test Manna-Whitneya.
- Metoda Bootstrap - pozwala ominąć założenie o rozkładzie normalnym.
- eBayes - gdy mamy mało powtórzeń i nie można wyliczyć wariancji.
- Test Anova - gdy mamy więcej niż dwa warunki

## t-test



## Etapy:

- Dla każdej grupy i dla każdego genu wyznaczamy średnią i odchylenie standardowe.
- Wyznaczamy parametr  $t$ .

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

- Odczytujemy wartość parametru  $p$  z tablic (o rozkładzie t-studenta).

## Co dalej z wytypowanym zbiorem genów?

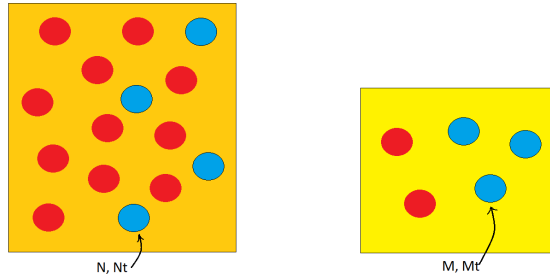
**Analiza funkcjonalna** - nadanie interpretacji biologicznej uzyskanym wynikom. W kontekście mikromacierzy takim wynikiem jest zbiór genów różnicujących. Przydatne w tym celu mogą być klasyczne bazy danych jak GenBank czy DDBJ czy te bardziej specjalistyczne (Gene Ontology, KEGG PATHWAY) zawierające informacje o wzajemnych powiązaniach między obiektami.





## Badanie nadreprezentacji grupy genów

Pytanie: Czy w wytypowanym przez nas zbiorze genów występuje większa grupa powiązana np z tym samym procesem biologicznym?



Porównujemy oczekiwane prawdopodobieństwo wystąpienia interesującej nas grupy genów z obserwowanym.

$$P(x = M_t) = \frac{\binom{N_t}{M_t} \binom{N - N_t}{M - M_t}}{\binom{N}{M}}$$