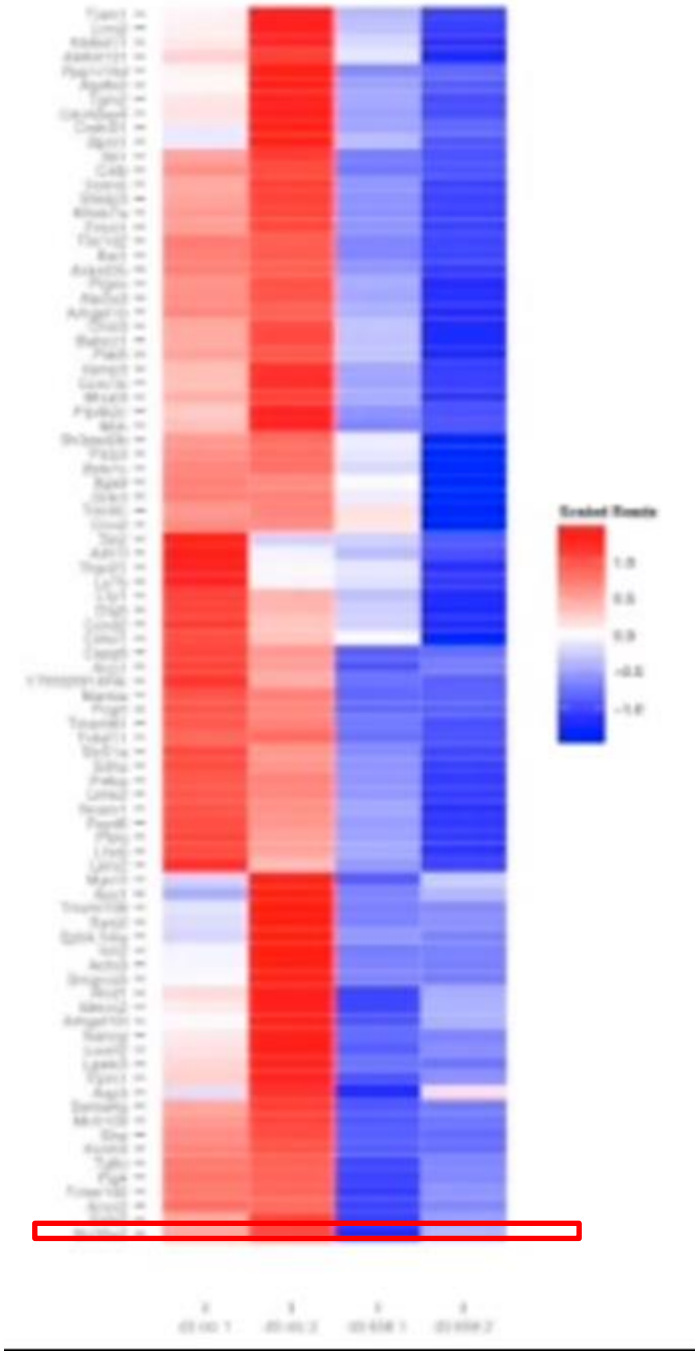
Heat Map, Skalowanie i Clustering

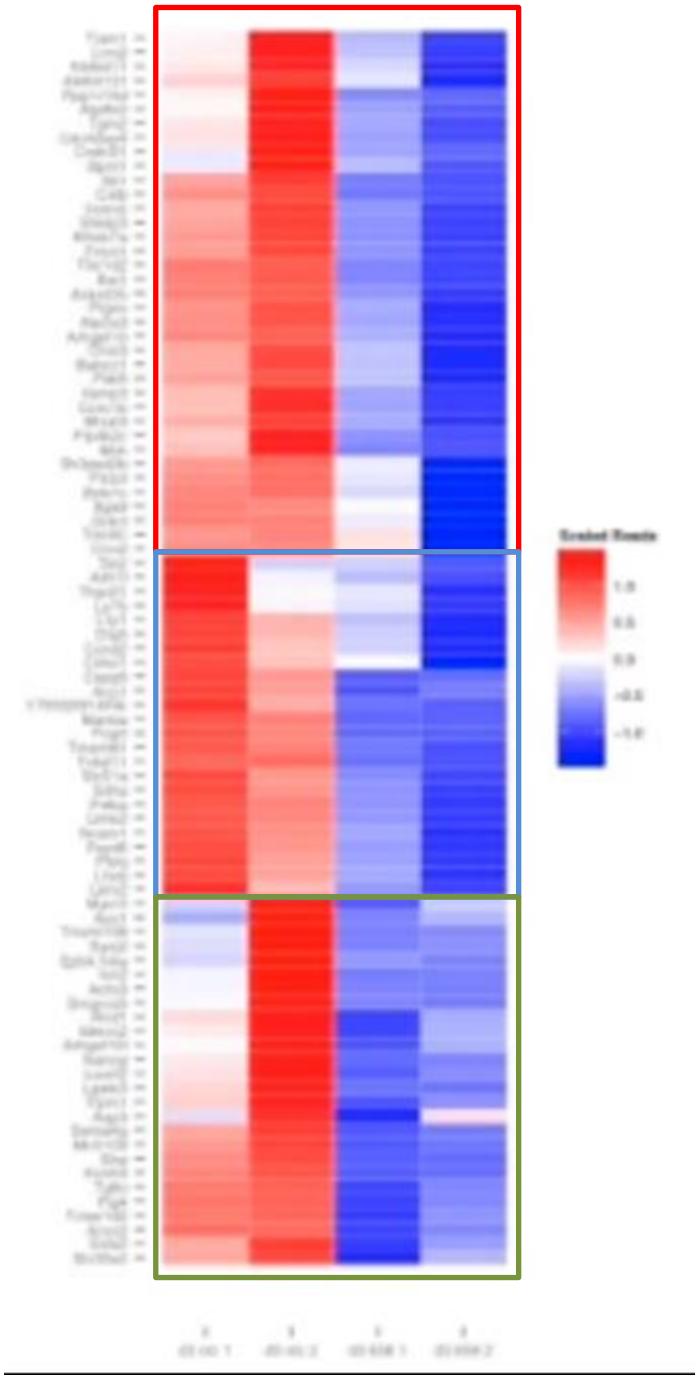


- Przykładowa heat mapa
- Wiersze to poszczególne geny
- Kolumny to poszczególne próbki

Dane prezentowane na tym wykresie zostały zmodyfikowane w dwojaki sposób, aby ułatwić analizę otrzymanych rezultatów

1. Względna ekspresja (ilość transkryptów) została wyskalowana, w tym przypadku skalowanie zostały wykonane osobno dla każdego genu. Dzięki takiemu podejściu łatwo możemy zobaczyć różnice w ekspresji konkretnego genu między próbką „X” i „Y”

Należy jednak pamiętać, że takie skalowanie (osobno skalowany każdy gen) uniemożliwia porównanie ekspresji pomiędzy różnymi genami. Przykładowo ciemny kwadrat dla próbki „X” i genu „G1” nie oznacza tego samego poziomu ekspresji co ciemny kwadrat dla próbki „X” i genu „G2”



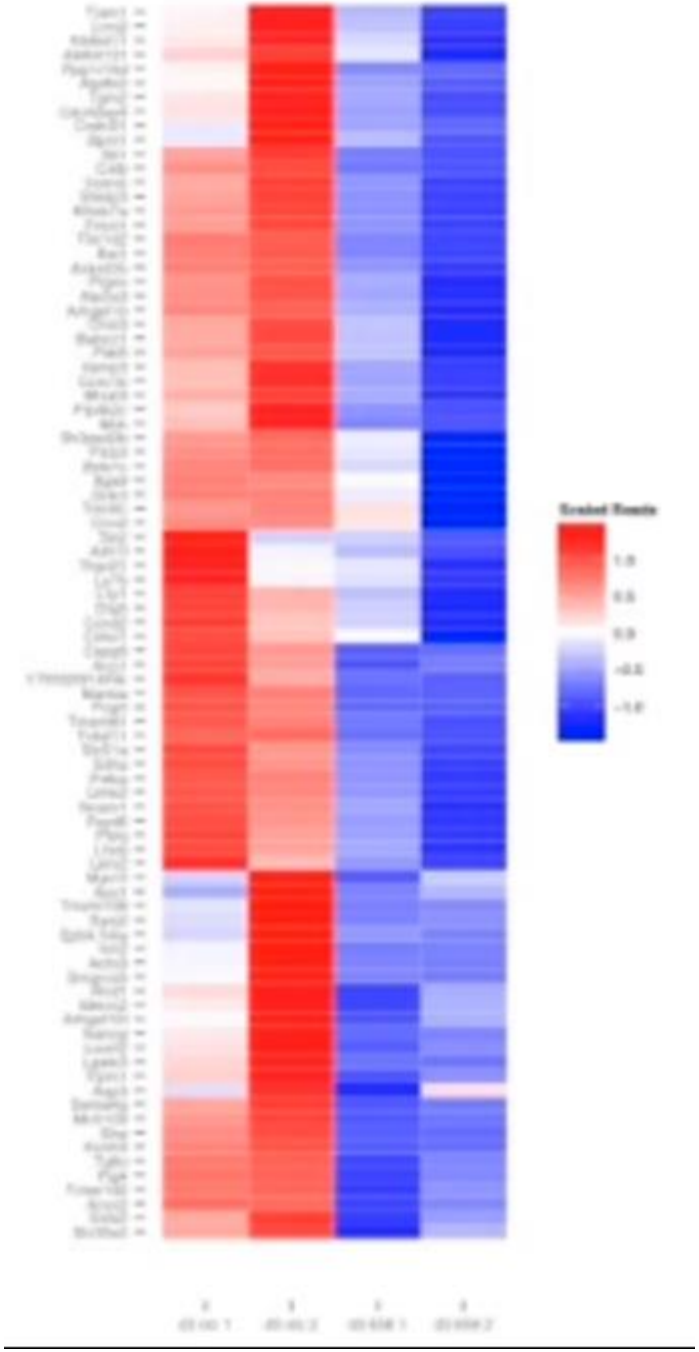
- Przykładowa heat mapa
- Wiersze to poszczególne geny
- Kolumny to poszczególne próbki

Dane prezentowane na tym wykresie zostały zmodyfikowane w dwojaki sposób, aby ułatwić analizę otrzymanych rezultatów
2. Geny o zostały zgrupowane na podstawie ich „podobieństwa”

Taka analiza pozwala zgrupować (i wskazać), geny które mają wysoką ekspresję w próbce 2 i niską w próbce 4

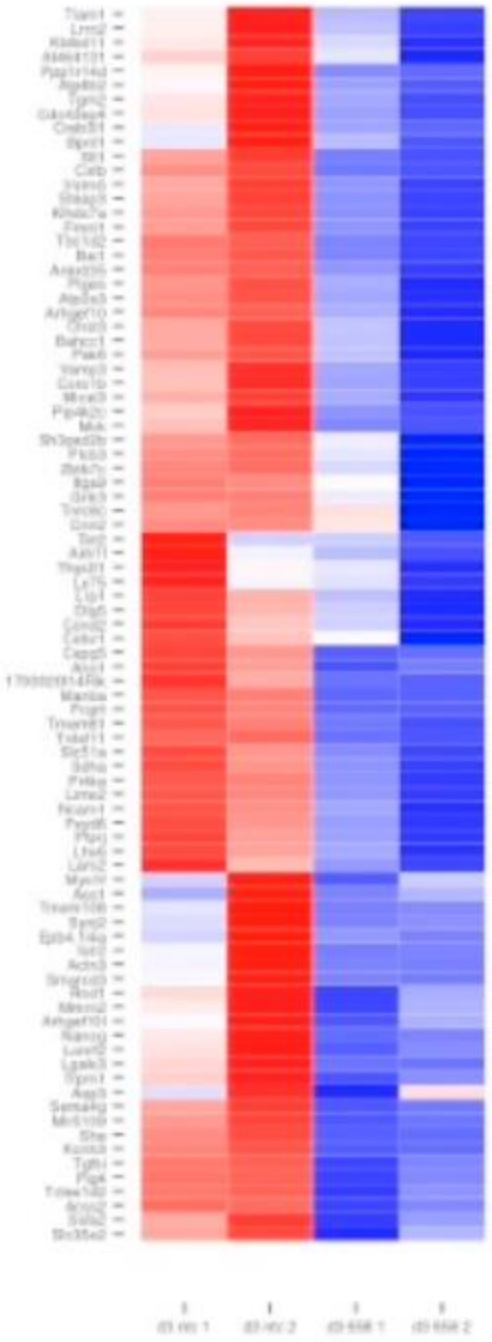
Kolejną grupę genów z wysoką ekspresją w próbce 1 i niską w próbce 4

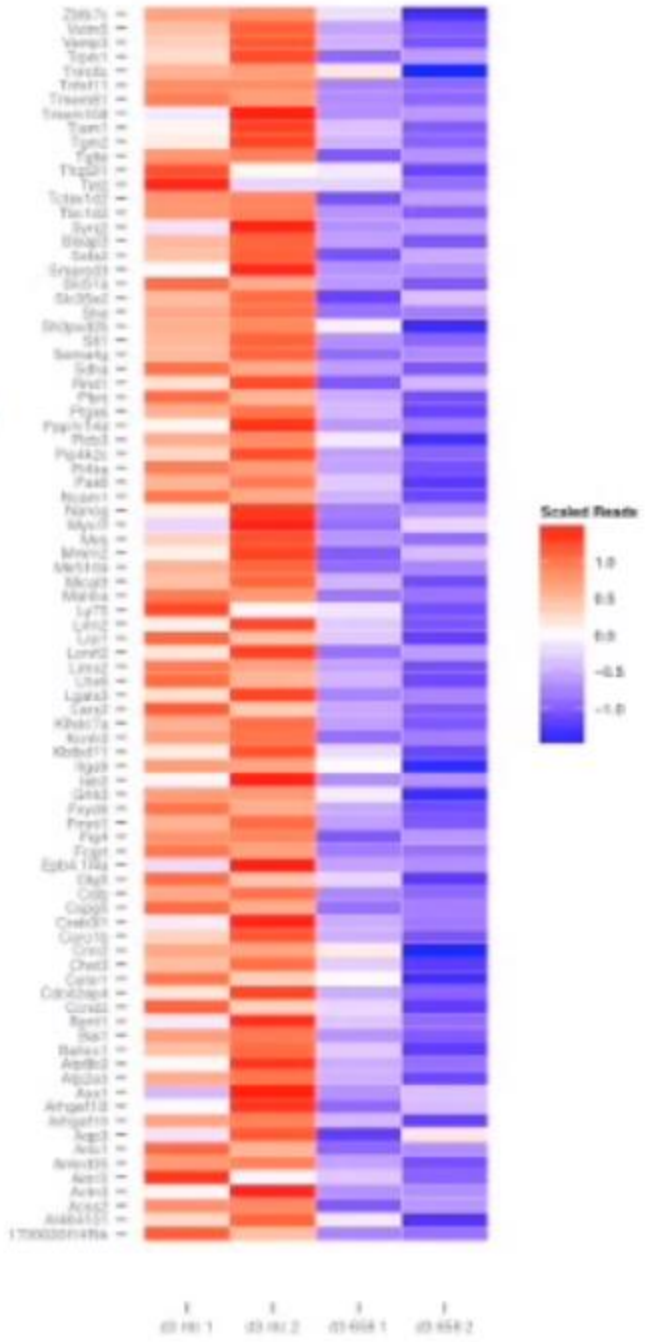
Oraz 3 grupę z genami o wysokiej ekspresji w próbce 2 i niskiej w próbce 3



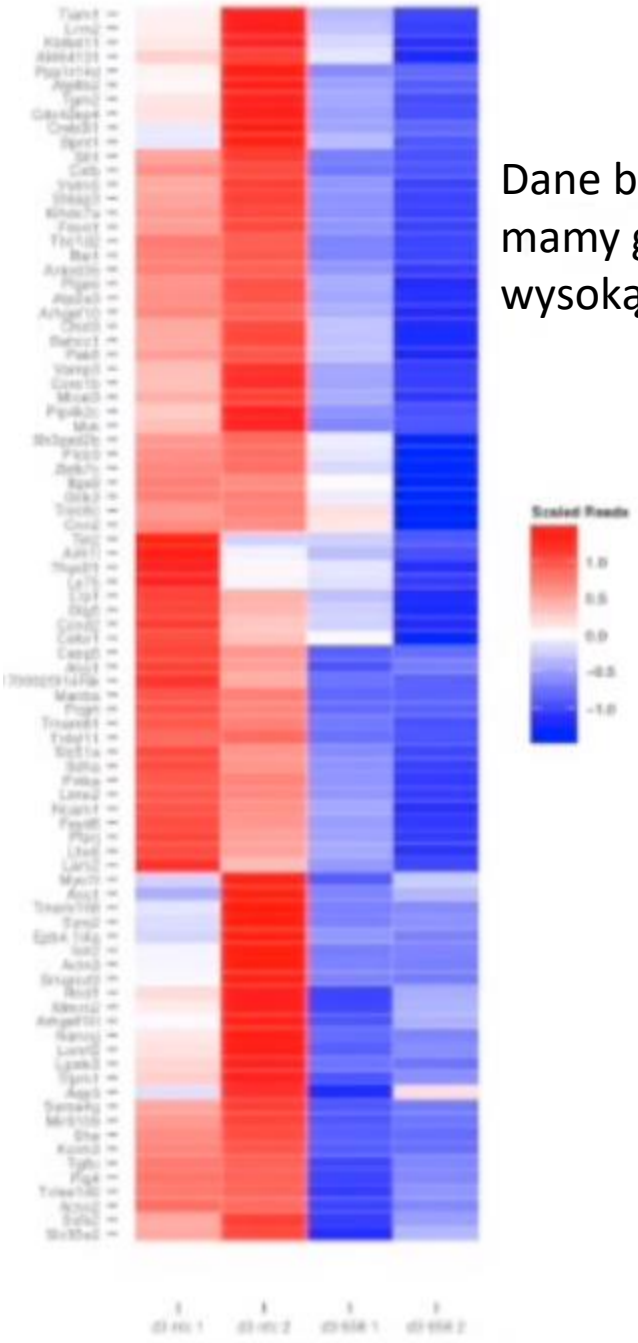
Taki układ nie jest wynikiem przypadku, a zastosowaniem przez program generujący wykres algorytmu który układa „podobne” obiekty w swoim sąsiedztwie

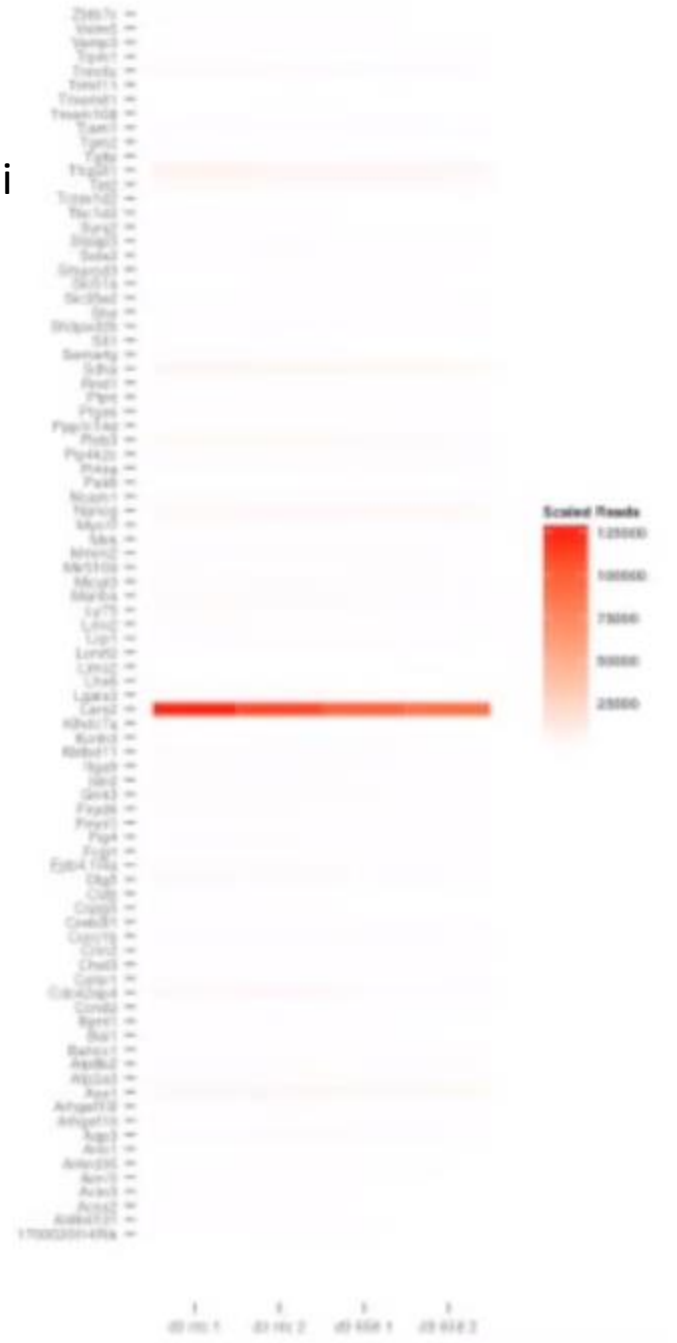
Dane bez clastrowania wyglądają tak…

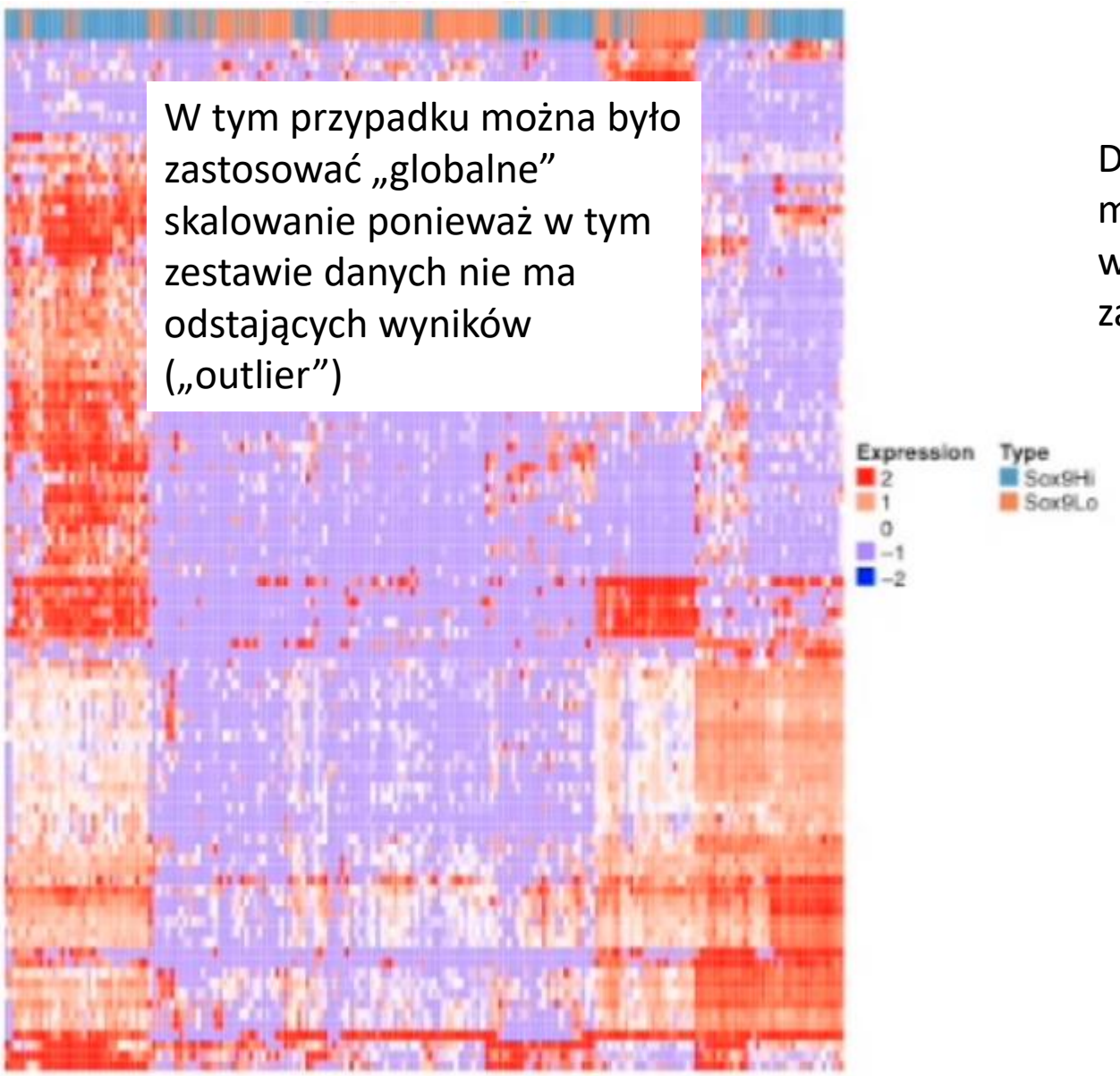




Dane bez skalowania, jeżeli mamy geny z bardzo wysoką ekspresją



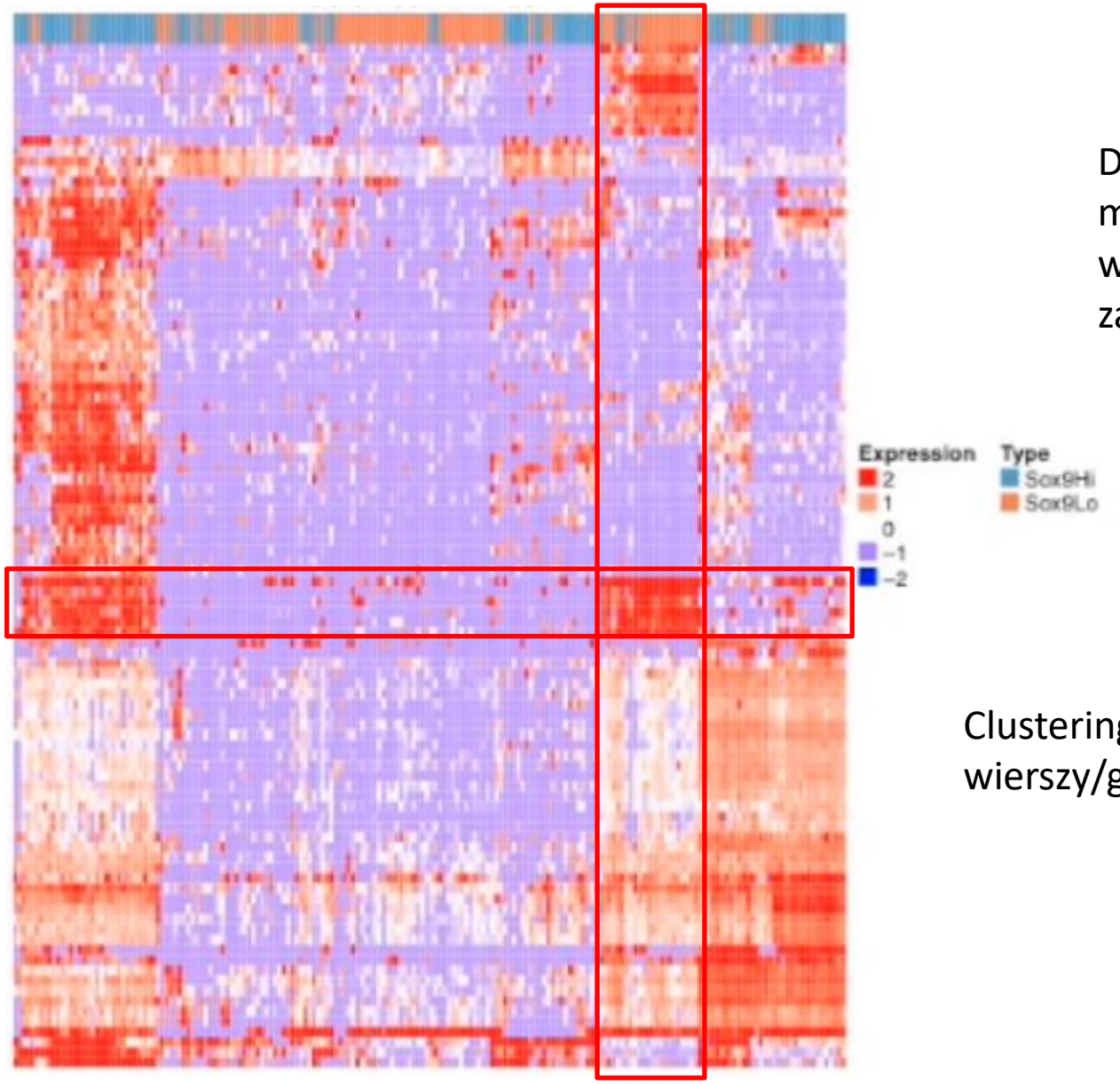




W tym przypadku można było zastosować „globalne” skalowanie ponieważ w tym zestawie danych nie ma odstających wyników („outlier”)

Do stworzenia tej heat mapy dane zostały wyskalowane, zastosowano clastering

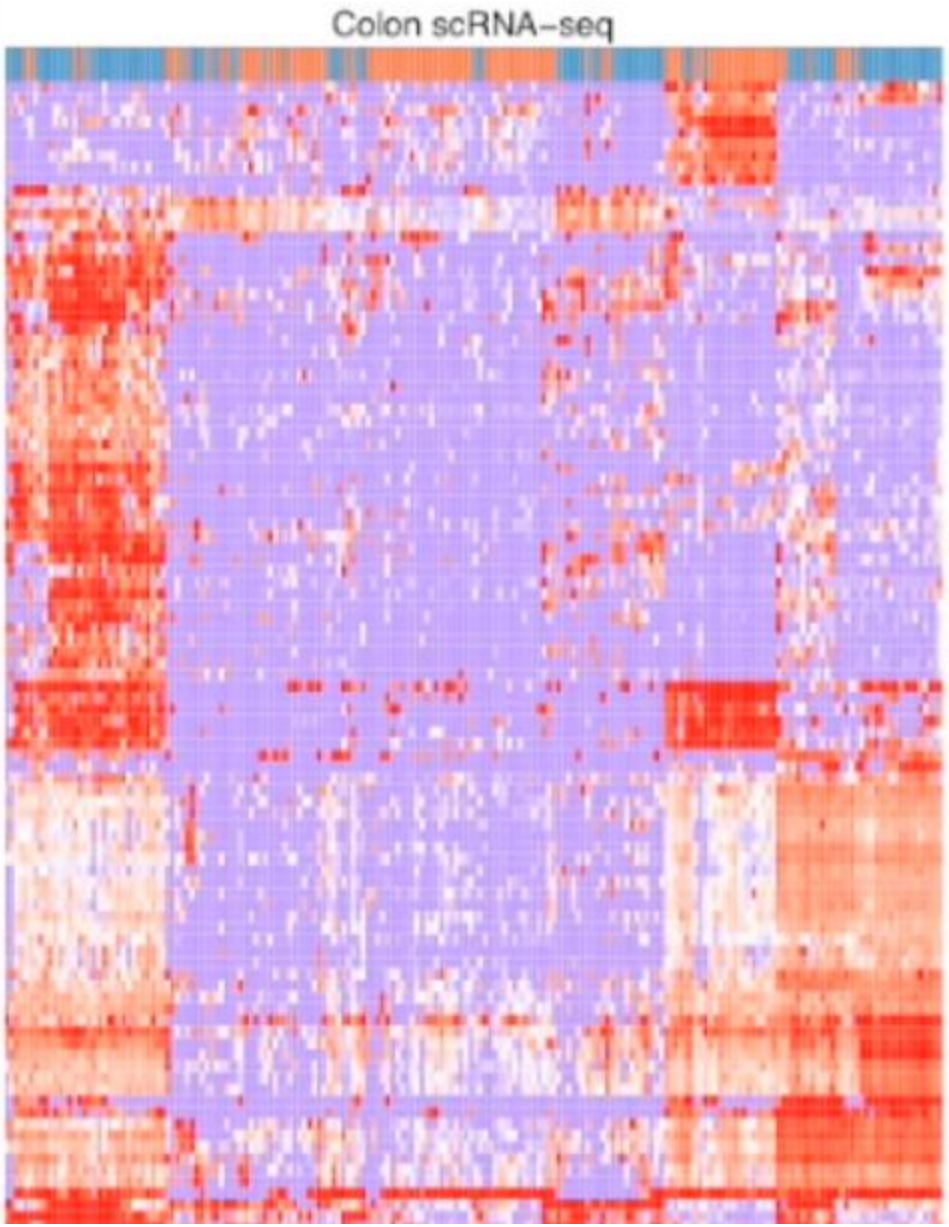
Zastosowano skalowanie „globalne”



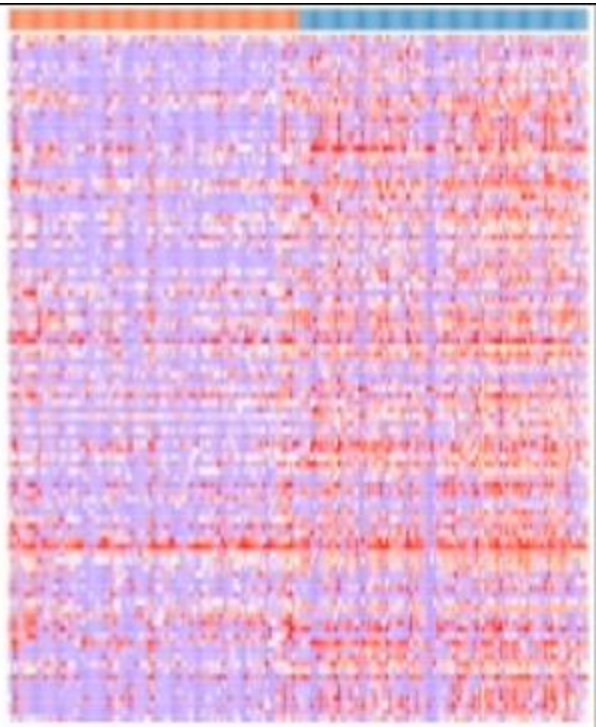
Do stworzenia tej heat mapy dane zostały wyskalowane, zastosowano clastering

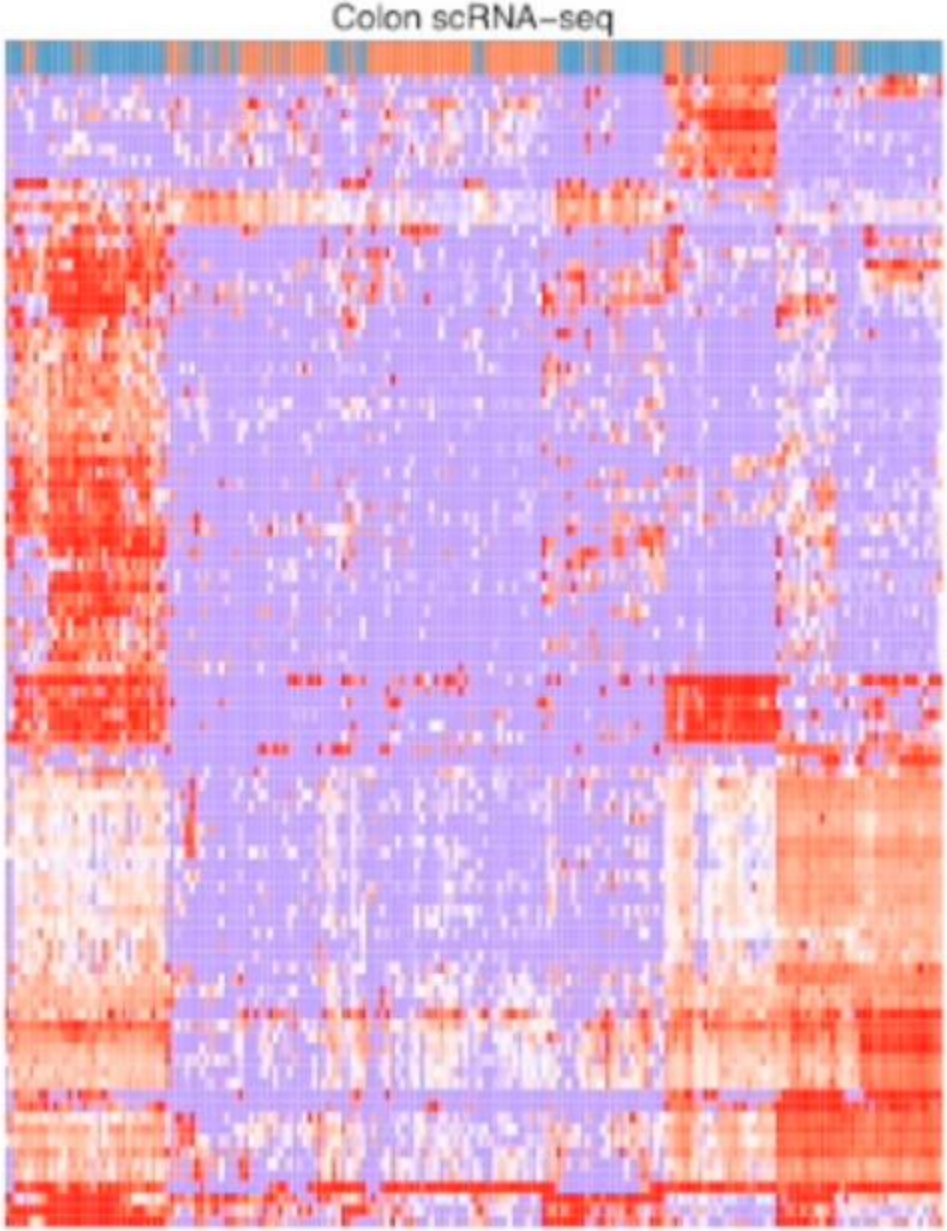
Zastosowano skalowanie „globalne”

Clustering zostały wykonane dla wierszy/genów oraz kolumn/próbek

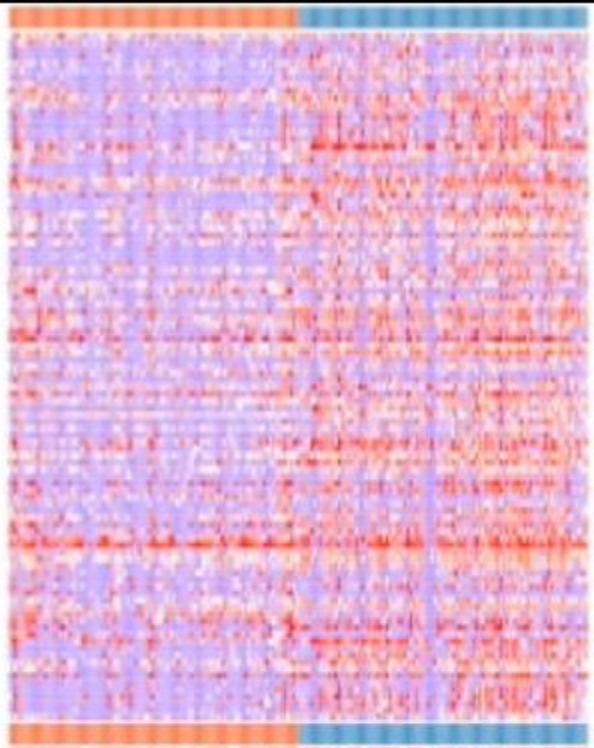


Bez klastrowania

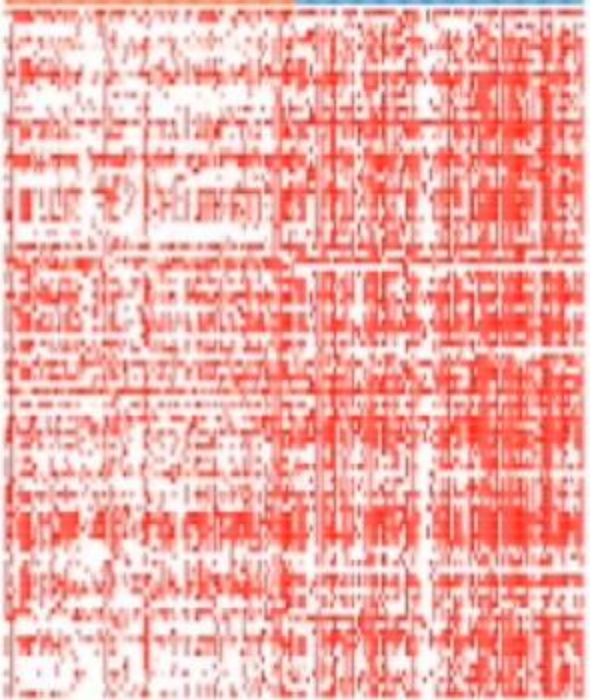


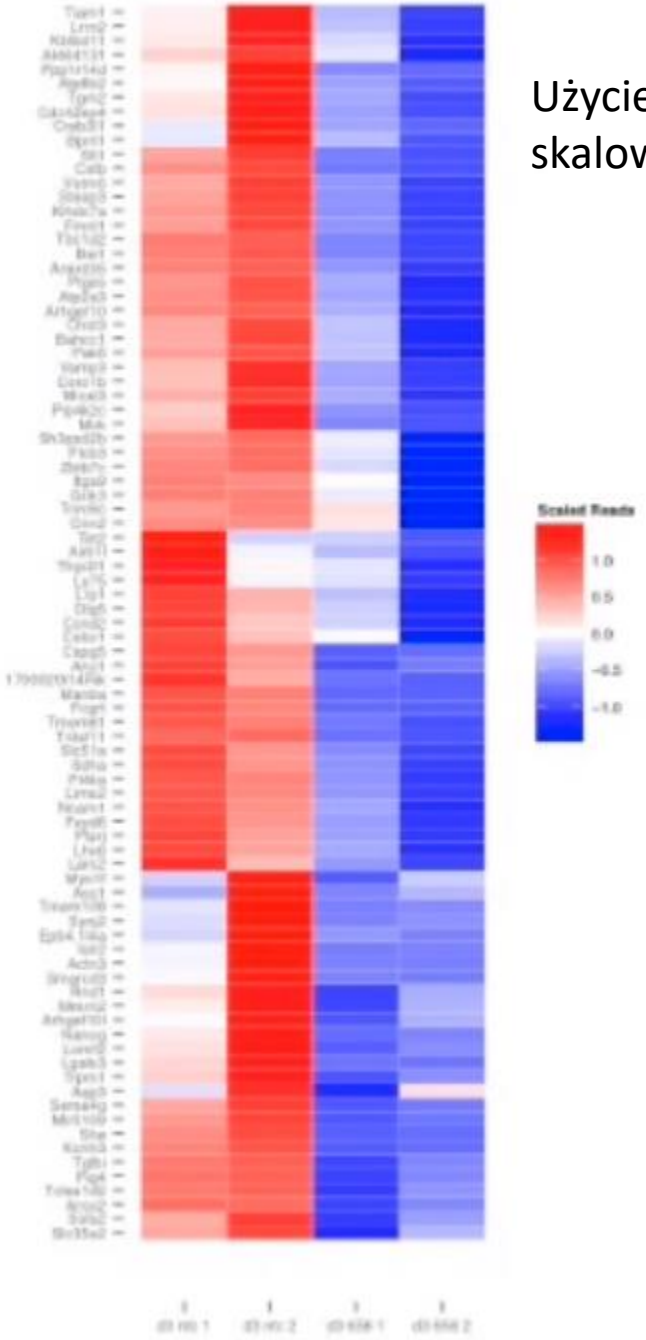


Bez klastrowania

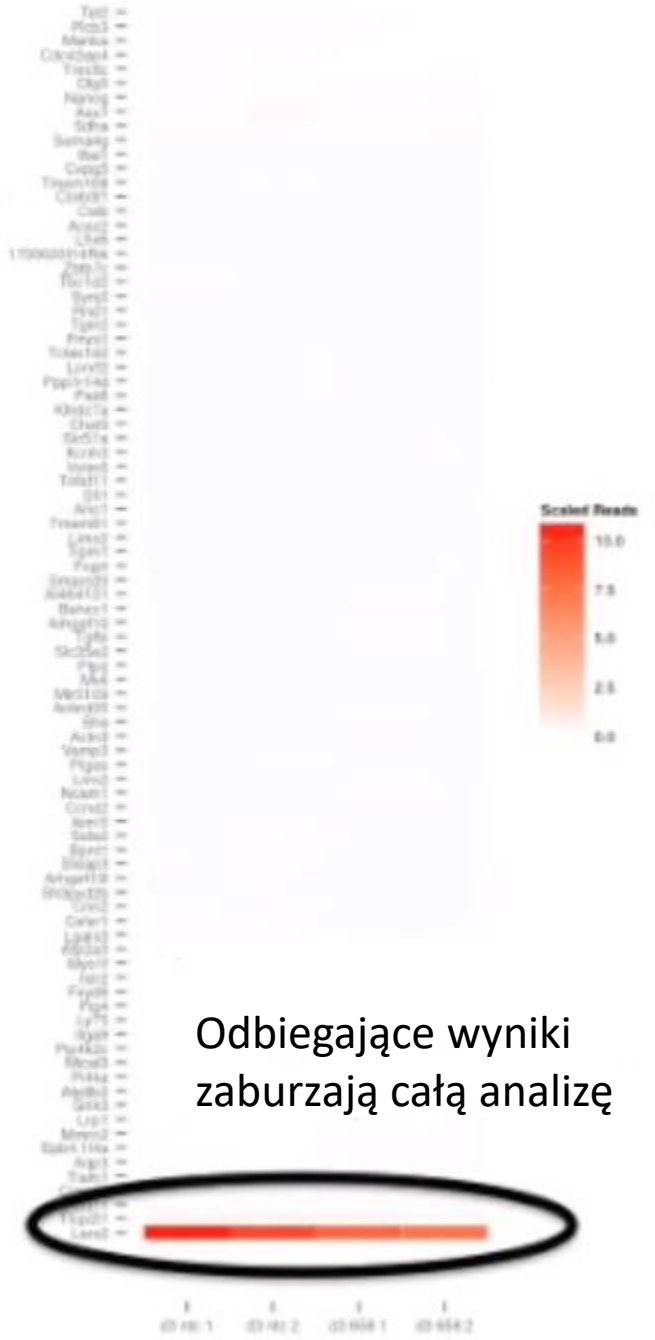


Bez klastrowania i bez skalowania

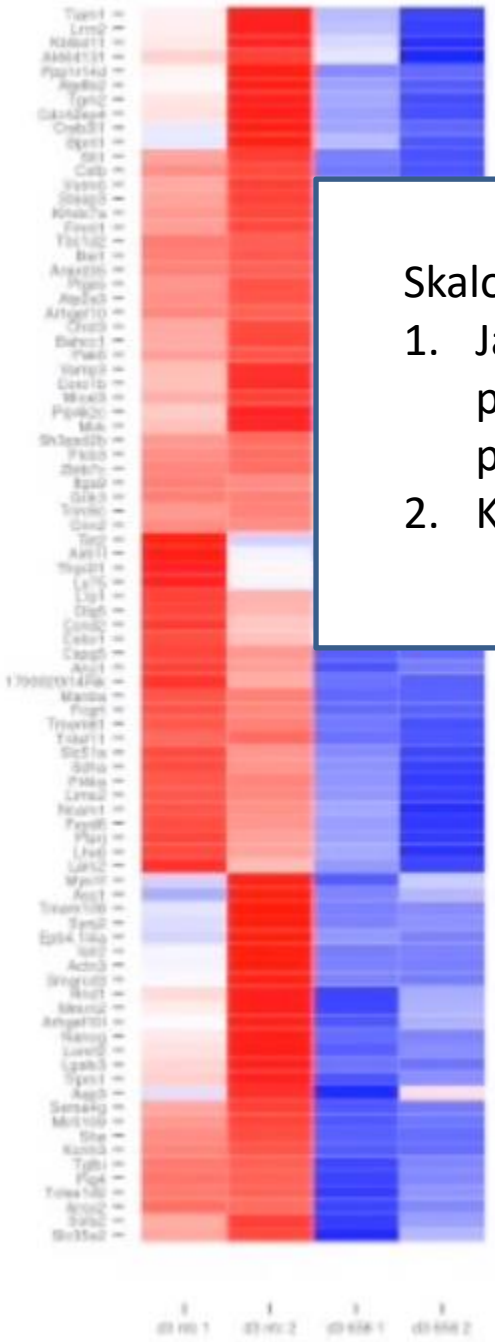




Użycie „globalnego” skalowania



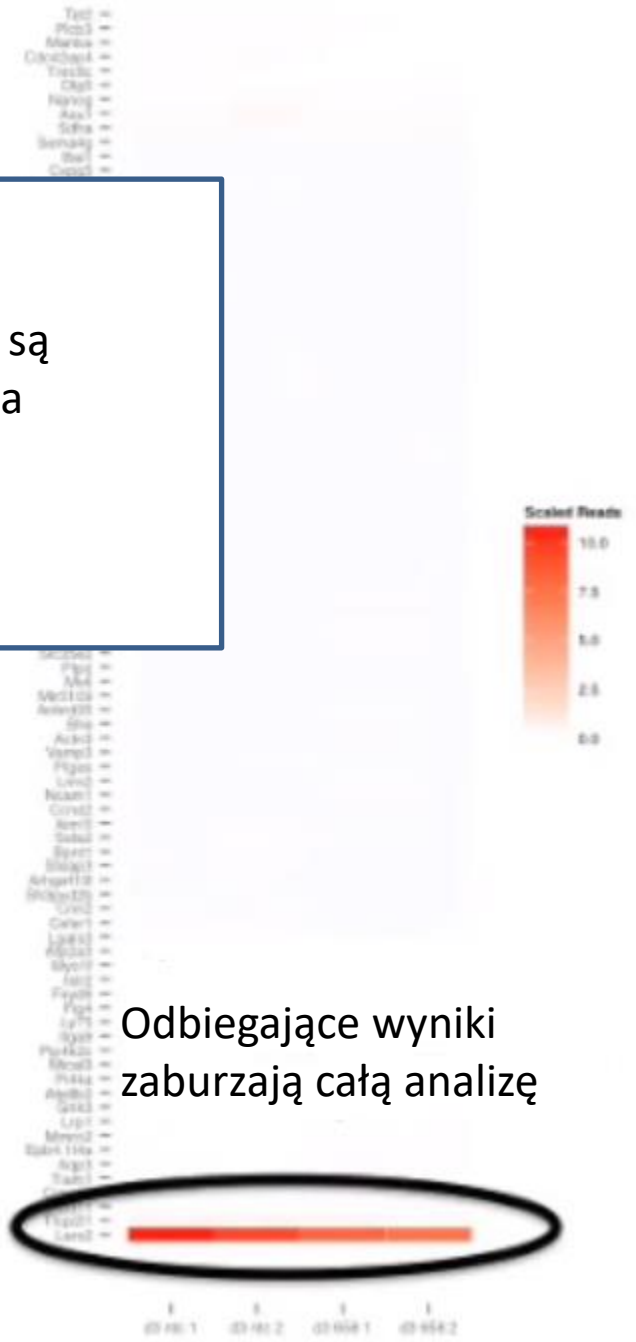
Odbiegające wyniki zaburzają całą analizę



Użycie „globalnego” skalowania

Skalowanie ma wpływ na:

1. Jak jasno/ciemno zaznaczone są poszczególne geny i czy można porównywać ich sekwencje
2. Klastrowanie



Odbiegające wyniki zaburzają całą analizę

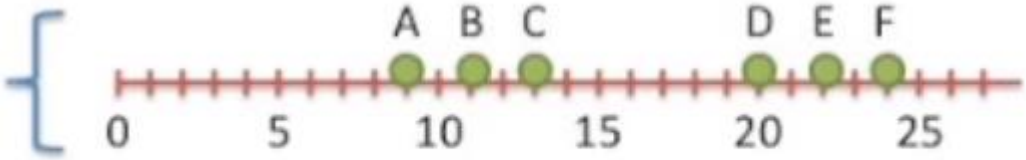
Jak wyskalować dane?

- Jaką metodę zastosować, nie zależnie od tego czy chcemy wykonać skalowanie globalne czy osobno dla każdego genu

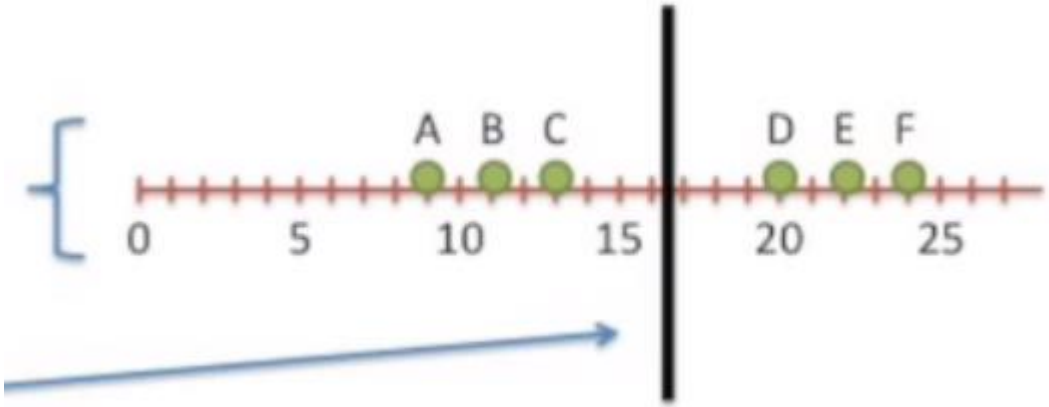
- Zastosowanie standaryzacji

Jak to działa?

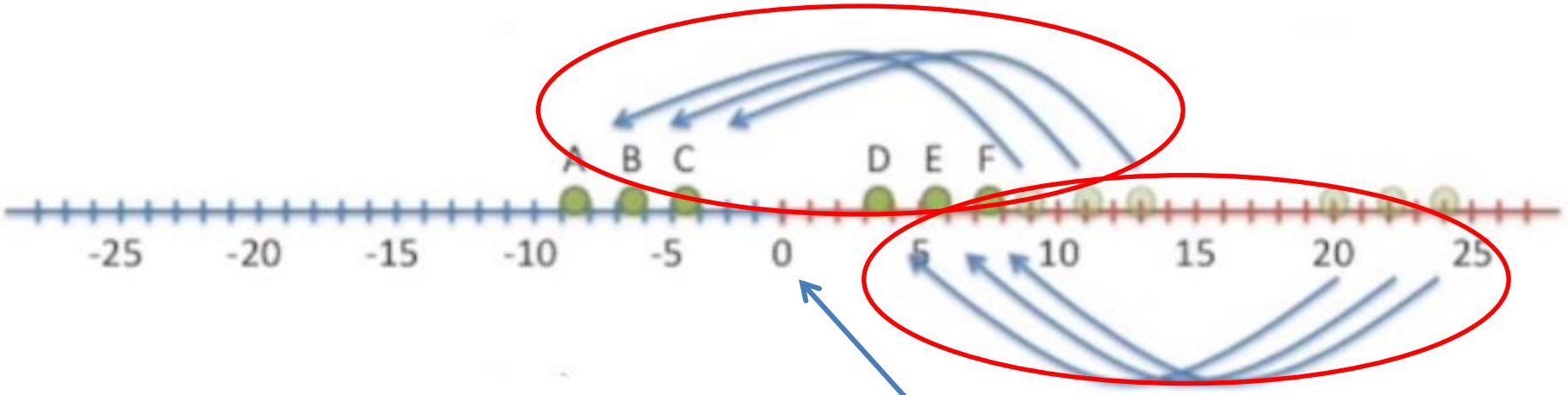
Liczby odczytów uzyskane w analize RNAseq dla pewnego genu i 6 różnych próbek



Liczby odczytów uzyskane w analize RNAseq dla pewnego genu i 6 różnych próbek

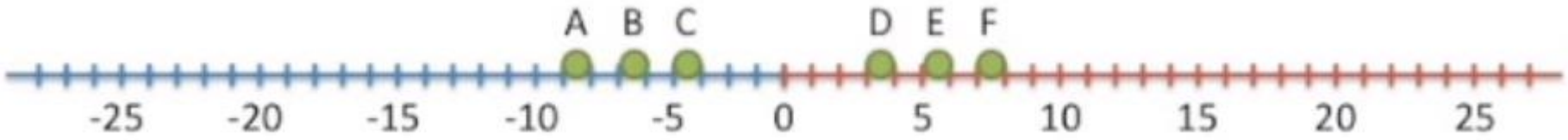


Krok 1 Obliczenie średniej (16,5)

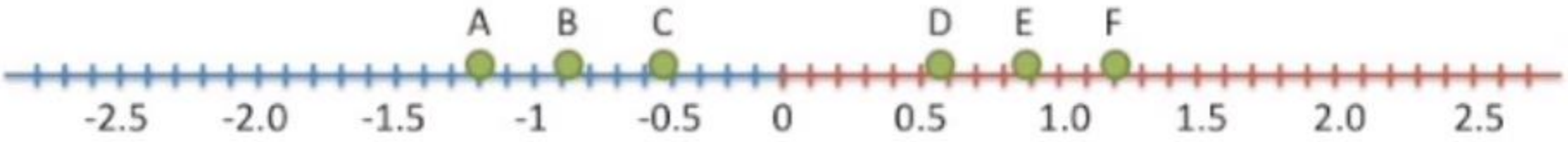


Krok 2 odjąć wartość średnią od każdego wyniku

Dzięki temu przesuwamy dane w okolice 0



Krok 3 Obliczamy
odchylenie standardowe
(6,28)

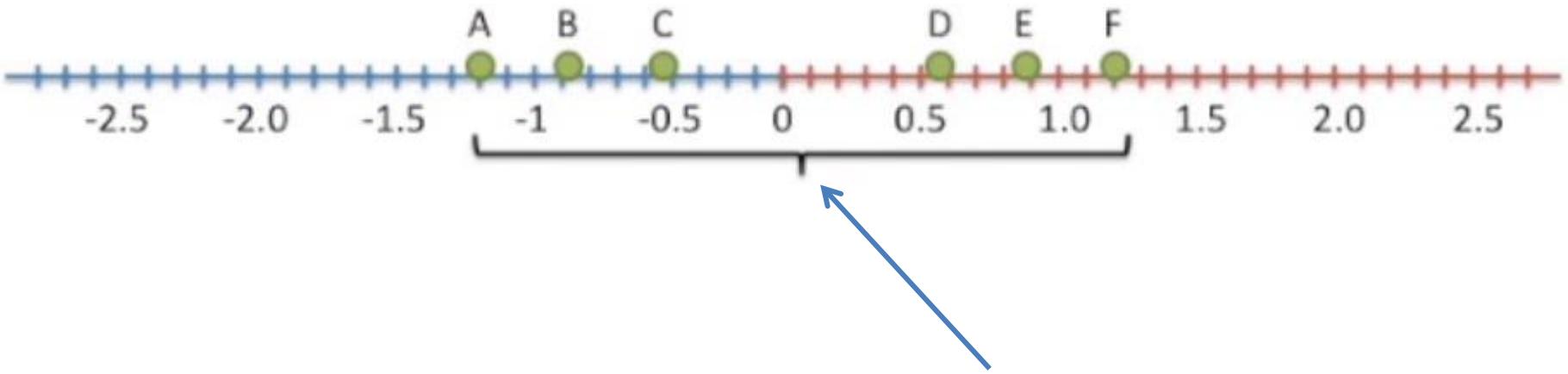


Krok 5 podzielić każdy wynik przez odchylenie standardowe

Powoduje to zmianę skali
Przed operacją uzyskane wartości były w przedziale -8 do 8
Po operacji są wprzedziale -1,2 do 1.2

$$z = \frac{x - \mu}{\sigma}$$

Gdzie:
z – wystandaryzowana wartość,
x – wartość,
μ – wartość średnia
σ – odchylenie standardowe

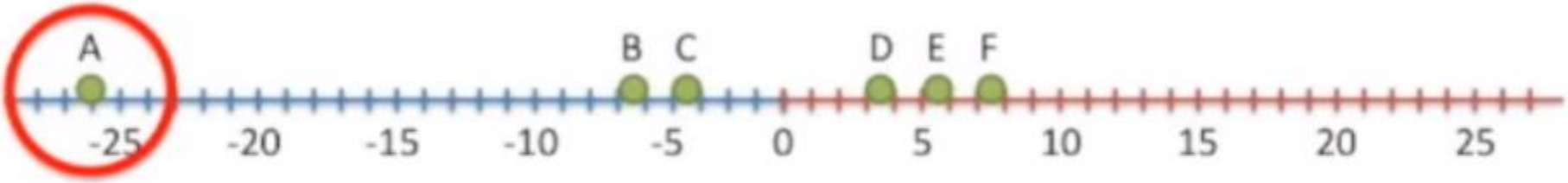


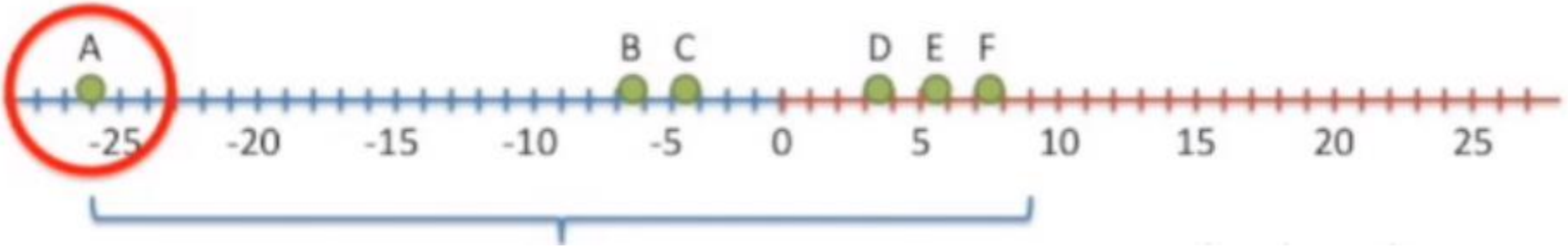
Niezależnie od zmienności (wariancji) występującej w oryginalnych danych, podział przez odchylenie standardowe zapewnia, że otrzymane dane po standaryzacji są zgrupowane blisko siebie (sd =1)

Jakie są tego korzyści?

- Potrafimy odróżniać ograniczoną liczbę barw
- Im węższy zakres występowania wartości cechy tym mniej kolorów potrzeba do stworzenia skali
- Dzięki temu zabiegowi łatwiej nam wizualnie oceniać dane

A co jeśli mamy odstający wynik (outlier)?

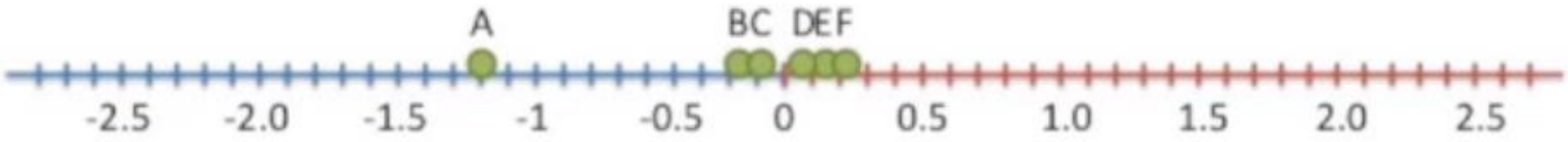

A
B
C
D
E
F
-25
-20
-15
-10
-5
0
5
10
15
20
25

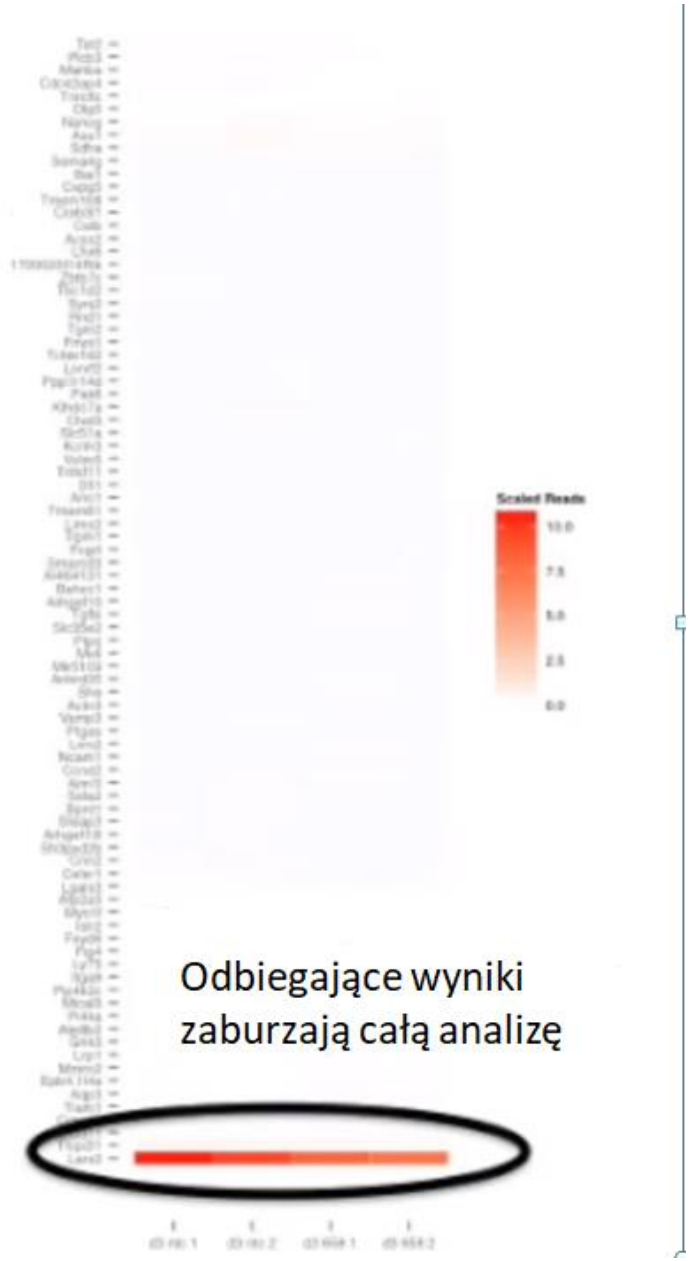


Wartość odchylenia standardowego będzie większa, czyli zwiększy się mianownik (wzór)

$$z = \frac{x - \mu}{\sigma}$$

Pozostałe wartości w okolicy 0 będą trudniejsze do odróżnienia, trudniej będzie obserwować te dane

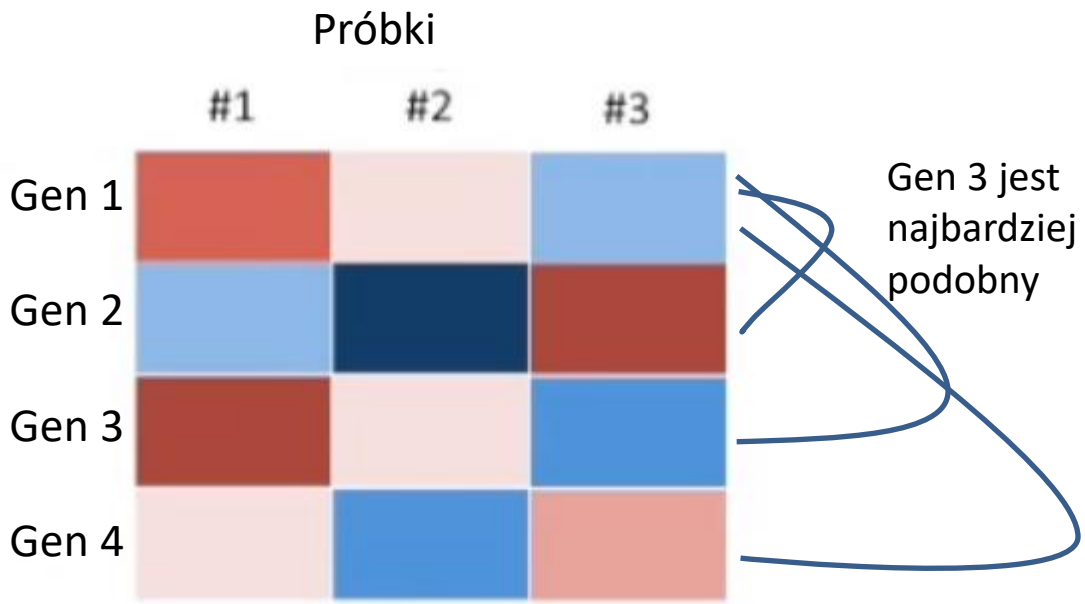



Scaled Reads
10.0
7.5
5.0
2.5
0.0
Odbiegające wyniki
zaburzają całą analizę

Clustering

- Dwa podstawowe algorytmy
  - Heierarhiczny
  - K-means (Algorytm centroidów)

Klastrowanie hierarchiczne



Gen 3 jest najbardziej podobny

Krok 1 Sprawdzamy który gen jest najbardziej podobny do Genu 1

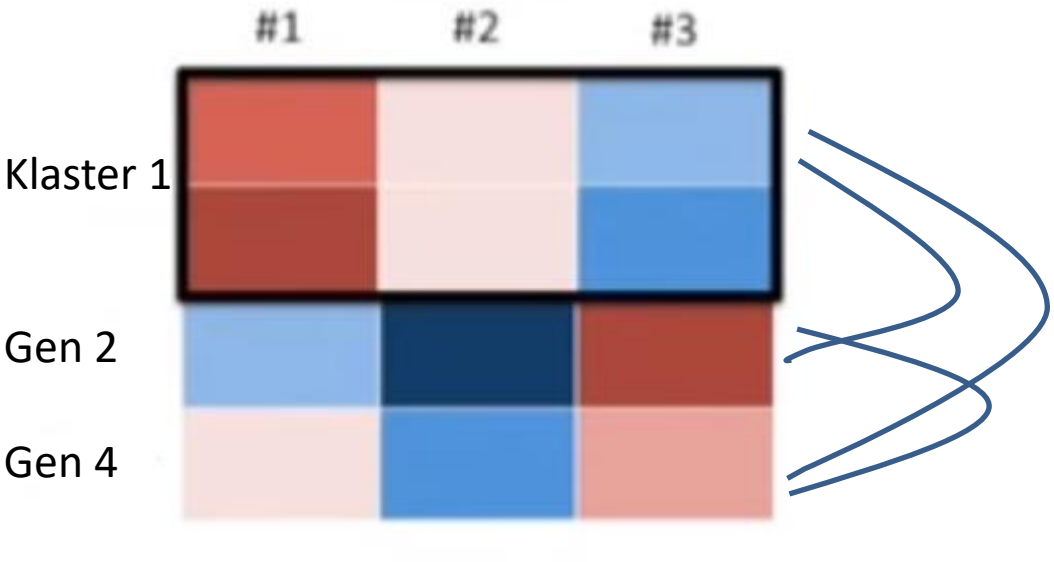
Krok 2 Sprawdzamy który gen jest najbardziej podobny do Genu 2… później do genu 3 i tak dalej

Klastrowanie hierarchiczne

Krok 3 Po ustaleniu dwóch najbardziej podobnych genów łączymy je w klastry

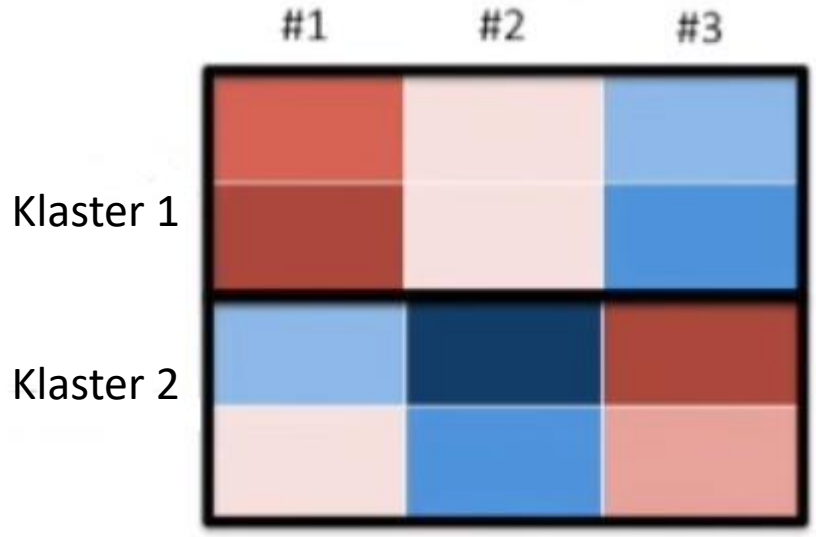
Próbki

Geny 1 i 3 tworzą pierwszy klaster



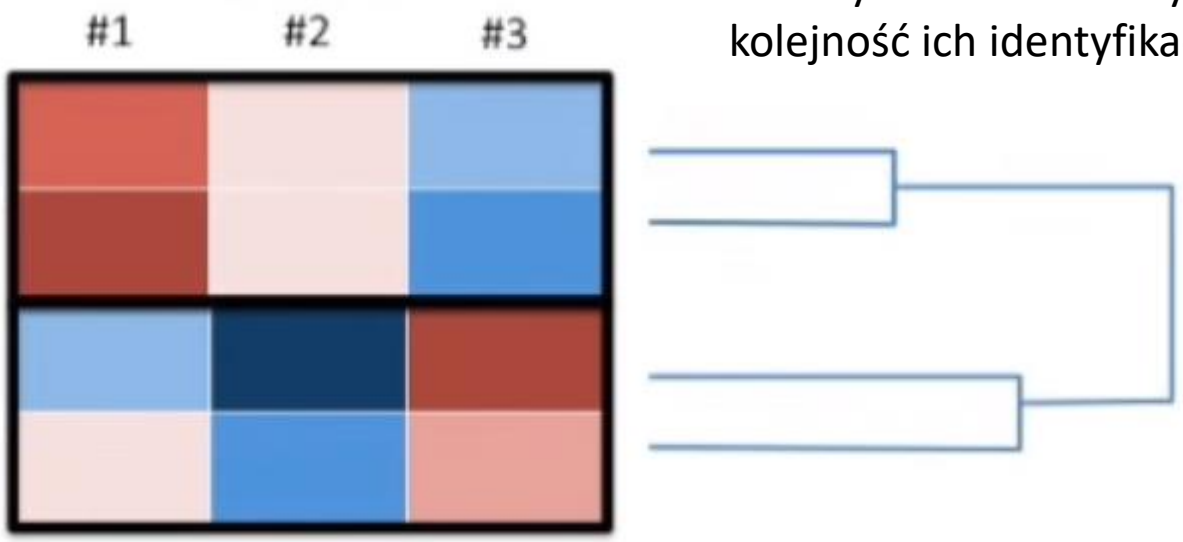
Krok 4 Wracamy do kroku 1, ale teraz szukamy genu podobnego do Klastra 1

Próbki



Klastrowanie hierarchiczne jest często obrazowane za pomocą dendrogramu, który jest dołączany do heat mapy

Dendrogram ilustruje zidentyfikowane klastry oraz kolejność ich identyfikacji



Jak określić które obiekty są najbardziej podobne?

- Metoda określania podobieństwa wybierana jest arbitralnie
- Najczęściej wykorzystuje się:
  - Dystans Euklidesowy

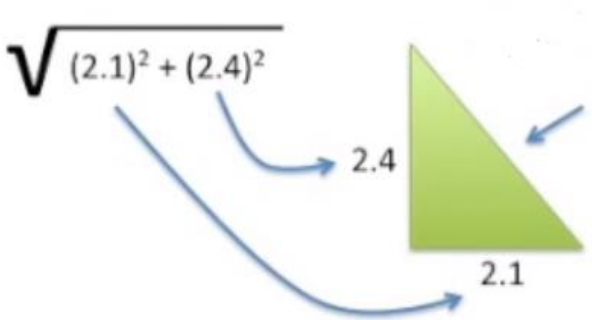
Dystans Euklidesowy

	Próbki	
	#1	#2
Gen 1	1,6	0,5
Gen 2	-0,5	-1,9

$$\sqrt{(1.6-(-0.5))^2+(0.5-(-1.9))^2}$$

Próbka pierwsza, różnica między genem 1 i 2

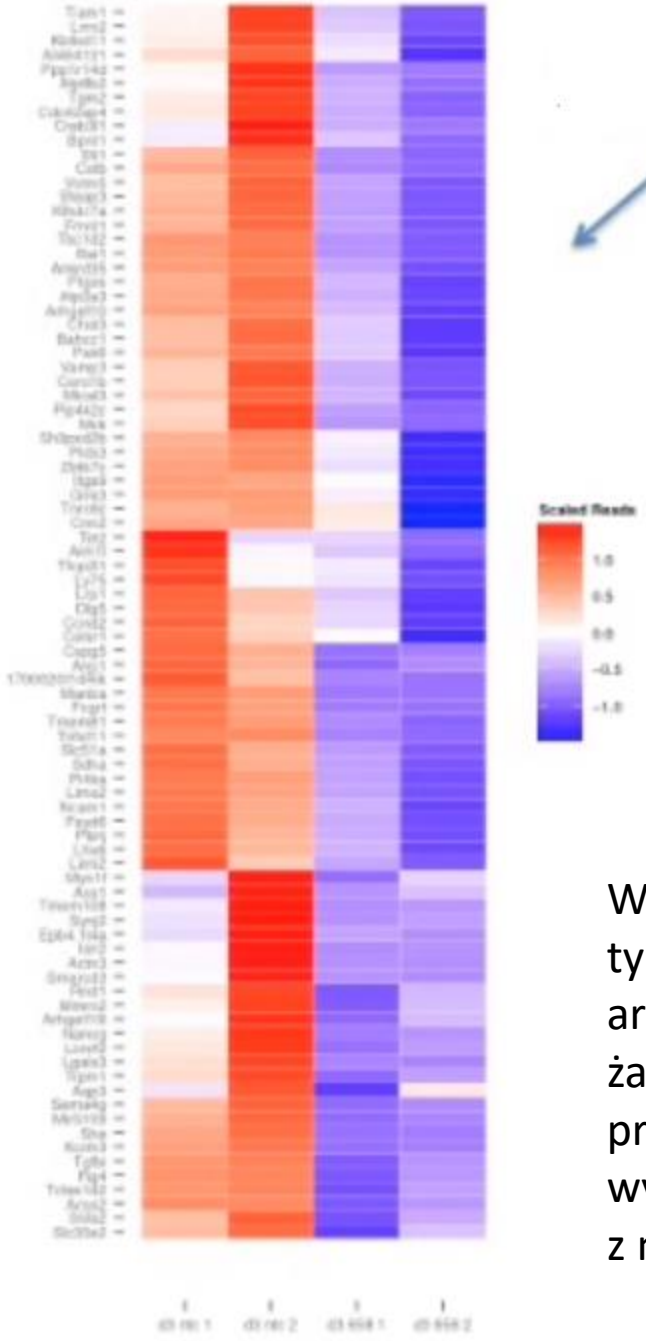
Próbka druga, różnica między genem 1 i 2



Dystans euklidesowy

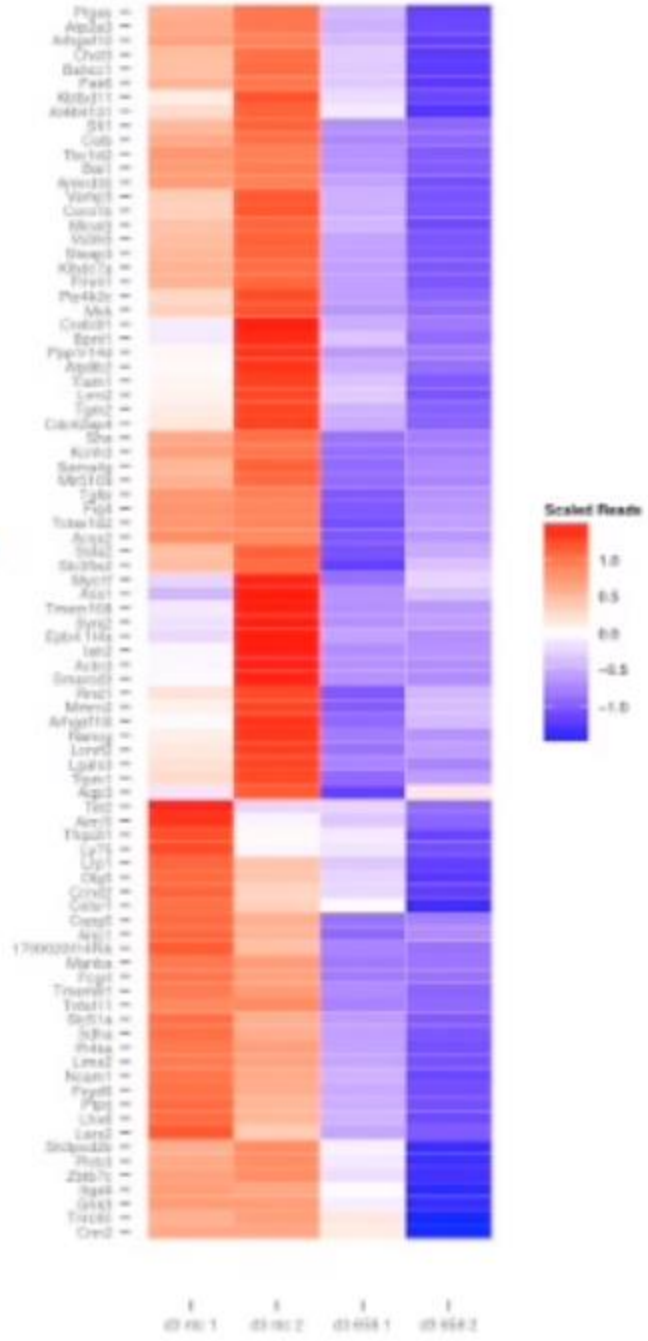
Jak określić które obiekty są najbardziej podobne?

- Metoda określania podobieństwa wybierana jest arbitralnie
- Najczęściej wykorzystuje się:
  - Dystans Euklidesowy
  - Dystans Malanobisa
  - Dystans Manhattan
  - i inne

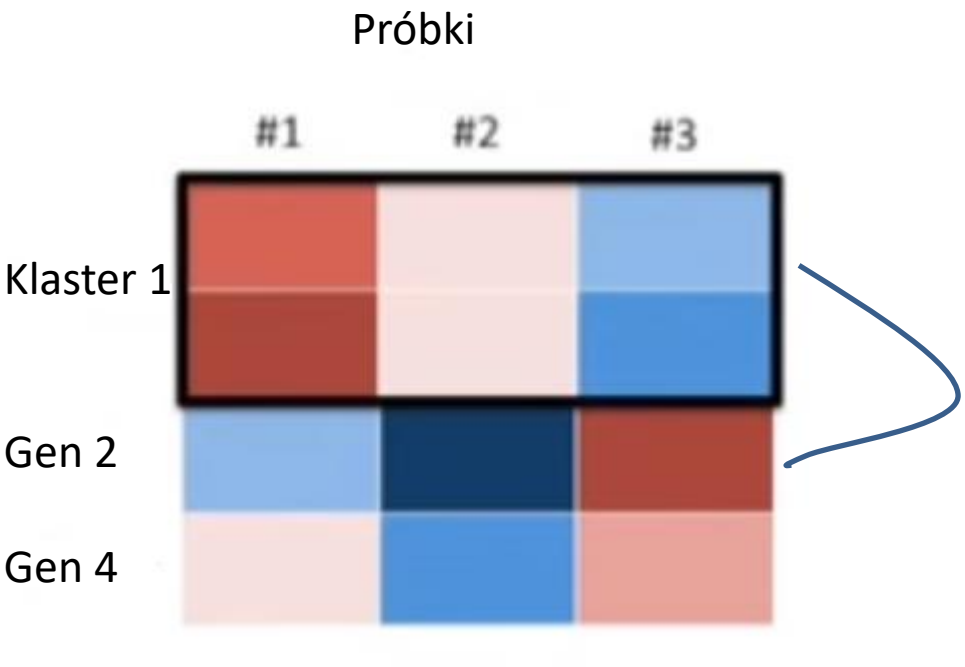


Dystans
Euklidesowy

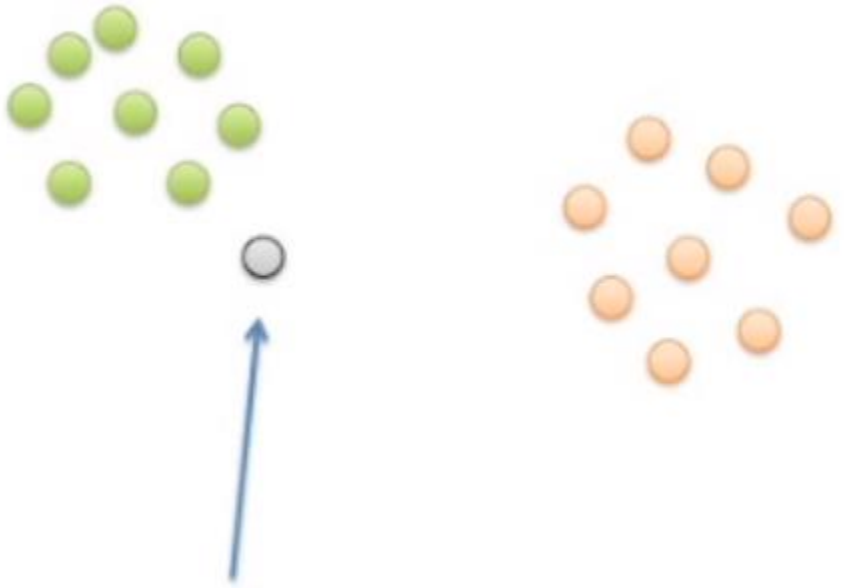
Dystans
Manhatan



Wykorzystanie danego typu dystansu jest arbitralne, nie ma żadnych biologicznych przesłanek za wykorzystaniem jednego z nich



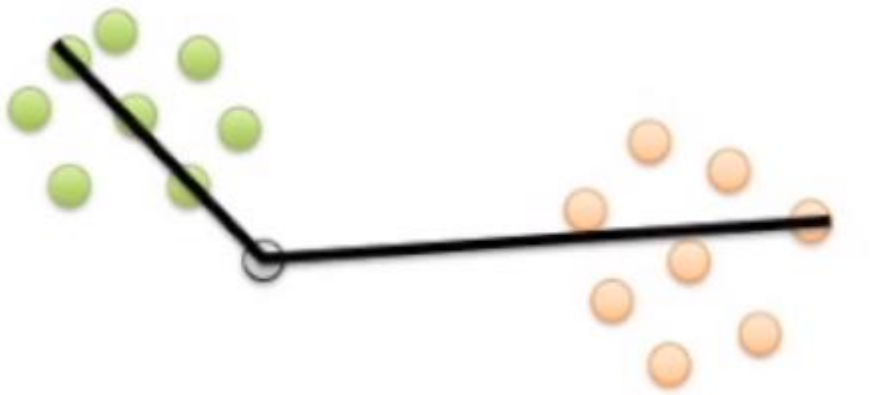
Gen 2 możemy porównać z klastrem 1 na kilka różnych sposobów ☹, każda zastosowana metoda będzie miała wpływ na końcowe rezultaty

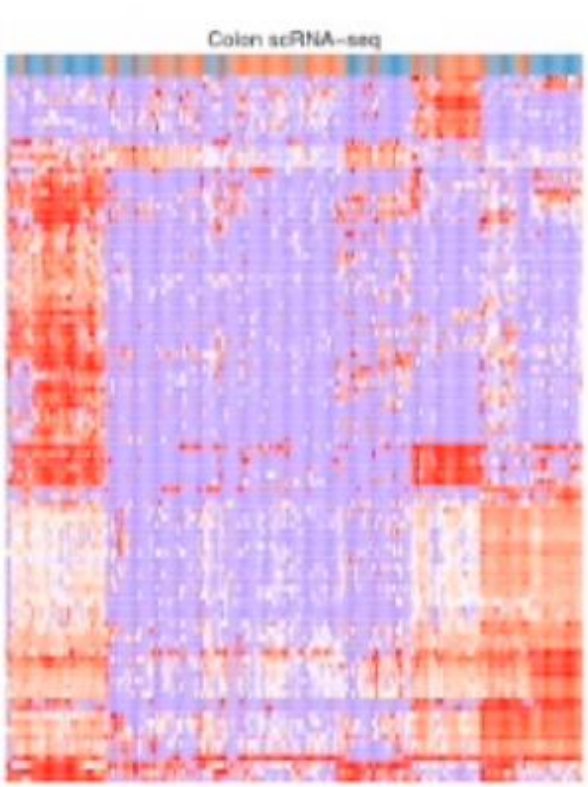


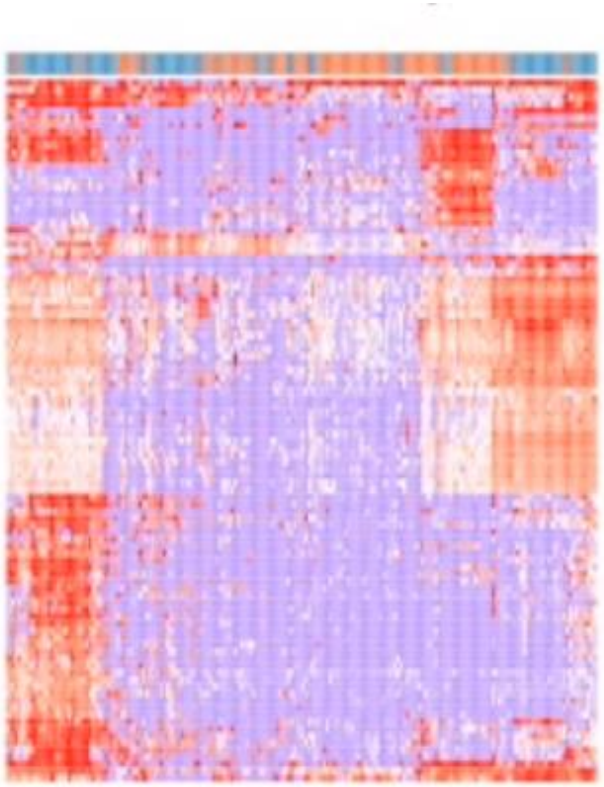
Średnia

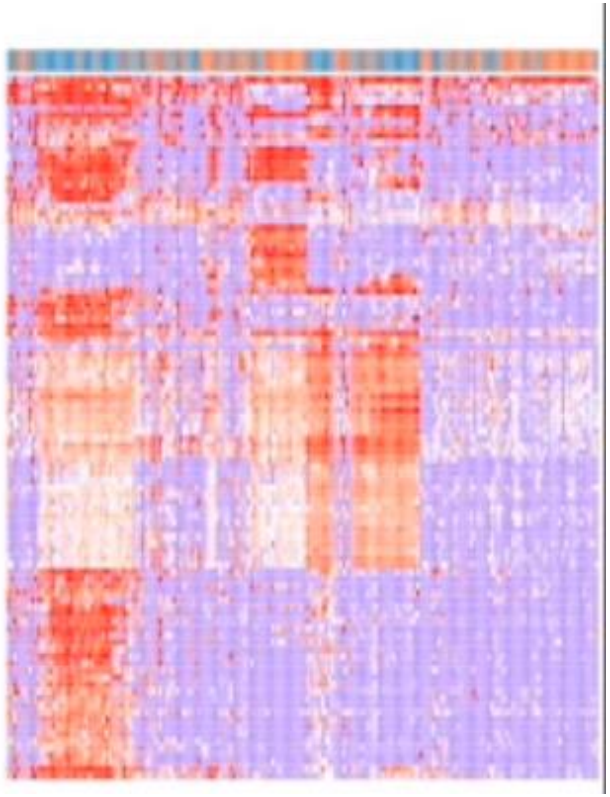
Najbliższy punkt

Najdalszy punkt – to jest standardowa metoda dla funkcji heatmap() w R!









Najdalszy punkt

Średnia

Najbliższy punkt

Etapy tworzenia heat mapy

- Skalowanie (normalizacja) danych (albo w obrębie genu, albo globalnie)
- Klastrowanie danych (albo dla genów, albo dla próbek, albo dla obu parametrów równocześnie)
  - Klastrowanie hierarhiczne, wymaga podjęcia arbitralnych decyzji
    - Na szczęście dla nas większość parametrów (dystans, metoda klastrowania mają swoje wartości domyślne w R
    - Możemy wykonać heat mapę, jeżeli „wygląda dobrze” to możemy ją zastosować bez zmieniania tych parametrów
  - Możemy zastosować metodę klastrowania K-means