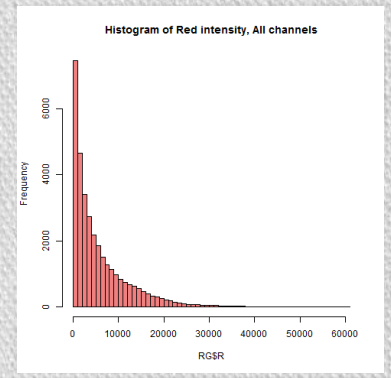
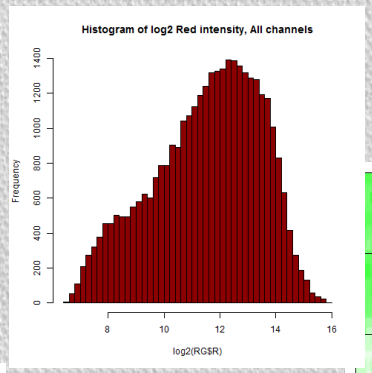
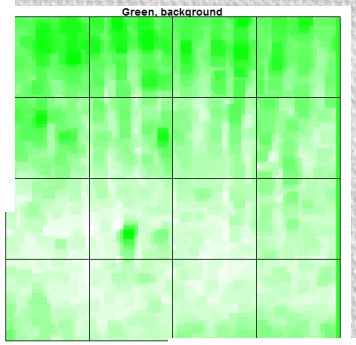
RÓŻNICOWA EKSPRESJA GENÓW

Na poprzednim wykładzie … skrót

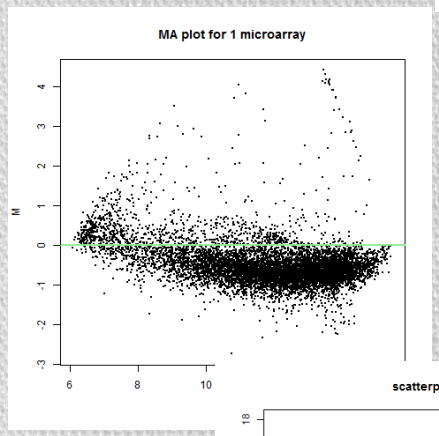
Wyniki eksperymentu nie są zgodne z oczekiwaniami

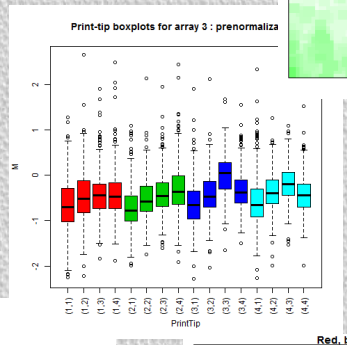


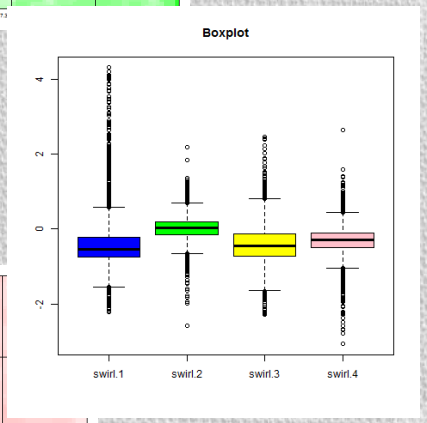


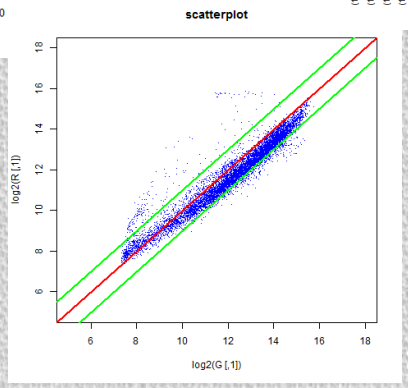


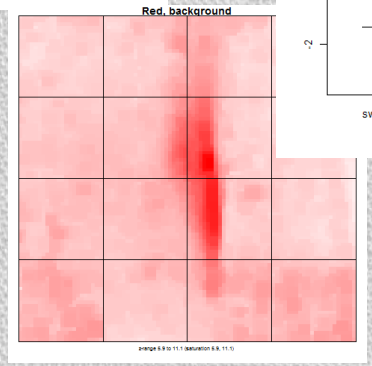
Normalizacja











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 nieznaczna 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

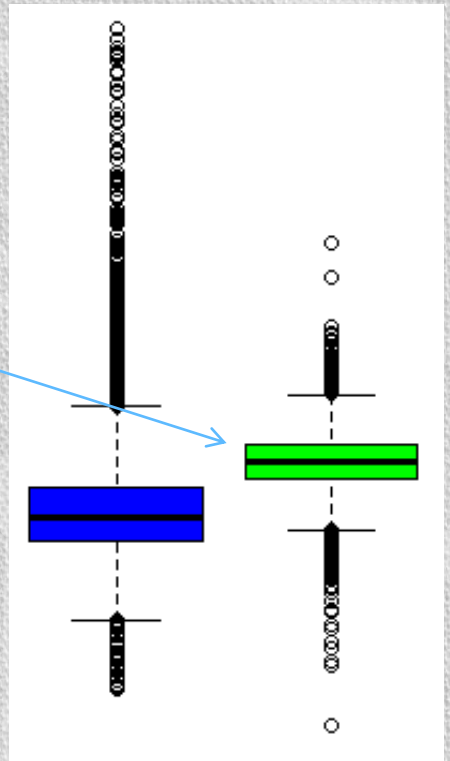
Lokalizacja i skalowanie to podstawowe pojęcia normalizacji

- ***Lokalizacja***

  Poprawia odchylenie przestrzenne lub od barwnika

- ***Skalowanie***

  Ujednolicenie różnorodności pomiędzy macierzami

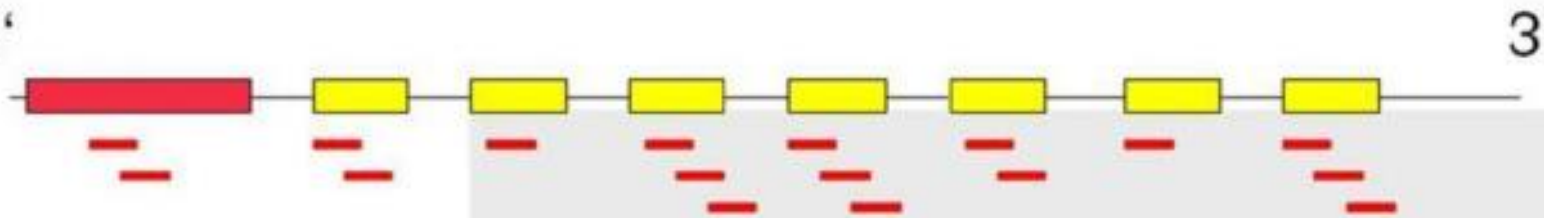


Znormalizowany stosunek log intensywności

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

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

Affymetrix Technology



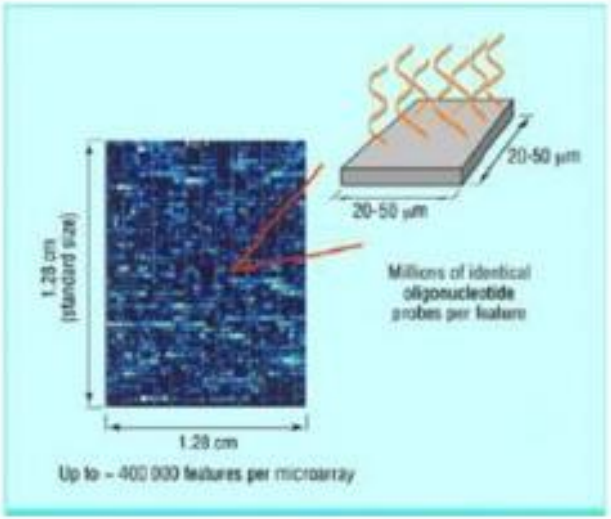
several *probe pairs*
(perfect match PM
and mismatch MM)
per *probeset*

PM: ATGAGCTGTACCAATGCCAACCTGG
MM: ATGAGCTGTACCTATGCCAACCTGG

64 pixels; Signal intensity is upper quartile of the 36 inner pixels

Stored in CEL file

16-20 probe pairs: HG-U95a
11 probe pairs: HG-U133



Affymetrix

- Miary PM_{ijg}, MM_{ijg}

  Absolutne intensywności dla *perfect match* i *mismatch*

  sonda *j* dla **genu *g*** na **mikromacierzy *i***

  $i = 1, \ldots, n$: od jednej do setek mikromacierzy (próbek)

  $j = 1, \ldots, J$: zazwyczaj 11-20 par sond dla 1 genu/sekwencji

  $g = 1, \ldots, G$: 10,000 - 50,000 genów na mikromacierz

- Sondy błędne (*MM*) powinny ocenić ilość niespecyficznej hybrydyzacji

- Użycie wielu sond dla jednej sekwencji/genu powinno zwiększyć jakość oceny RNA

Ocena jakości

Podobna jak dla macierzy dwukolorowych:

- Obraz macierzy (zanieczyszczenia)
- Boxplot z wartościami ekspresji
- Wykres MAplot
- Wykres degradacji RNA
- ...

MAplot

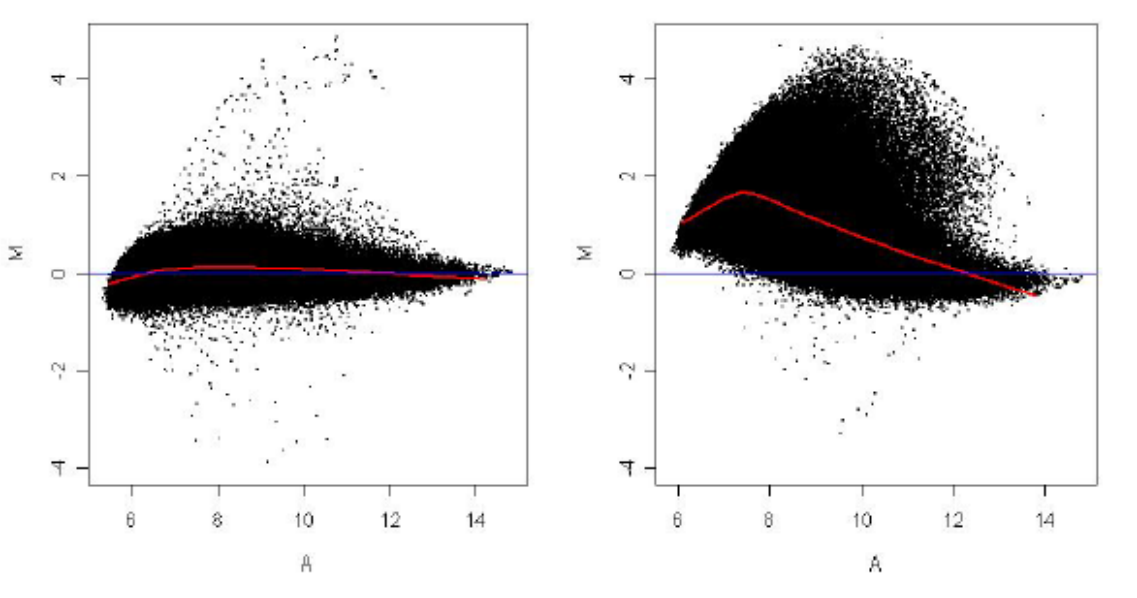
$M_{ijg} = \log_2(PM_{ijg}) - \log_2(PM_{*jg})$

Różnica pomiędzy macierzą i a macierzą referencyjną *

$A_{ijg} = [\log_2(PM_{ijg}) + \log_2(PM_{*jg})]/2$

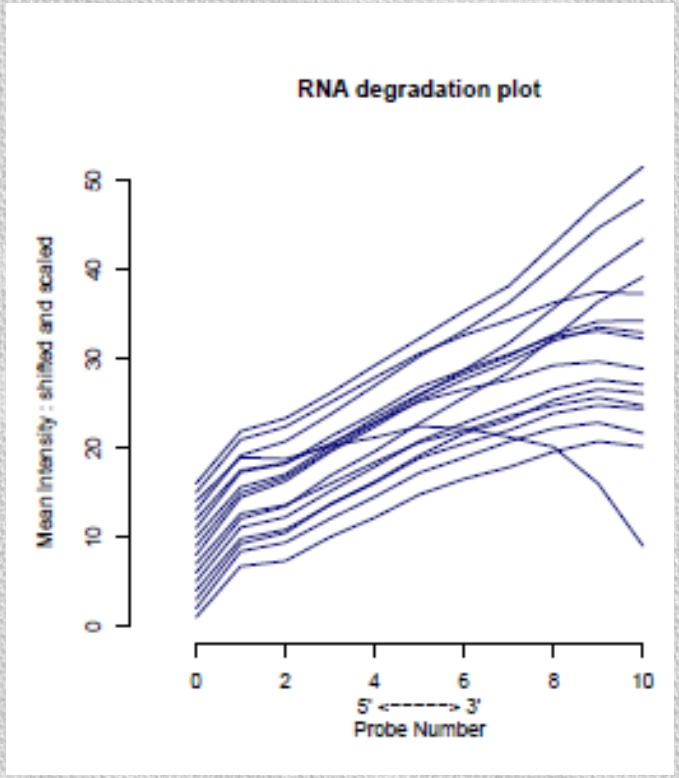
Średnia intensywność

PM_{*jg} jest to mediana dla sondy dla wszystkich mikromacierzy



Degradacja RNA

- RNA degraduje od końca 5’ sekwencji, dlatego sondy dla końca 3’ danego genu będą świeciły intensywniej niż dla 5’
- Wykres degradacji pokazuje średnią intensywność dla każdej pozycji zbioru sond (jednego genu)
- Duży spadek wskazuje na degradację
- Ważniejsze niż spadek jest zgodność pomiędzy macierzami



Preprocessing

- Korekcja tła
  - Przypadkowy szum
- Wartości PM i MM
  - Niespecyficzne wiązanie RNA
- Normalizacja
  - Kalibracja miary dla różnych mikromacierzy
  - Dla sondy, czy dla grupy sond?
  - Których sond/zbiorów używać do normalizacji
- Podsumowanie
  - Dla każdego zbioru sond jedna wartość ekspresji

Metody normalizacji

- **MAS** – oparte na *PM – MM*
- **dChip** – oparte na *PM – MM*
- **RMA** – korekcja tła oparta na PM, normalizacja quantile i wygładzanie mediany
- **GC-RMA** – RMA z dodatkiem korekcji niespecyficznej hybrydyzacji
- **Cyclic loess** – normalizacja loess na wykresach MAplot
- **vsn** – normalizacja stabilizacji wariancji (korekcja tła i normalizacja)

…

Nie ma najlepszej metody normalizacji

MAS

- Dla każdego zbioru sond (*A*) genu (*g*) na mikromacierzy *i* wyznaczamy

$$AvDiff = \frac{1}{|A|} \sum_{j \epsilon A} (PM_j - MM_j)$$

- Problem: dla około 1/3 sond MM > PM

Korekta

$$CT_j = \begin{cases} MM_j, & jeśli\ MM_j < PM_j \\ mniejsze\ niż\ PM_j, & jeśli\ MM_j \geq PM_j \end{cases}$$

sygnał = mean(log(PM_j - CT_j))

RMA

Robust multi-array average (Irizarry et al. 2003)

Motywacja:

- MM może wykryć prawdziwy sygnał dla niektórych próbek
- Różnica PM oraz „tła” wzrasta wraz z koncentracją (tak wykazały testy w eksperymentach spike-in)

RMA

$$PM_{ijk} = bg_{ijk} + s_{ijk}$$

sygnał dla sondy j z grupy sond k dla mikromacierzy i

tło spowodowane przez szum optyczny oraz niespecyficzne wiązanie

Korekcja tła jako transformacja $B()$

$B(PM_{ijk}) = E[s_{ijk} \mid Pm_{ijk}] > 0$

$s_{ijk} \sim \text{Exp}()$ $\quad$ $b_{ijk} \sim \text{normal}$

- Przy korekcji tła nie używamy sygnału MM
- Wartość sygnału jest pozytywna

RMA

- **Korekcja tła** jako transformacja B(.)
- **Log_2**
- Kalibracja pomiędzy macierzami poprzez **normalizację quantile**
- **Podsumowanie sond dla grupy sond** przez wygładzanie mediany (regresja)

Literatura

- Yang et al. (2002) „Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation" *Nucleic Acis Research* 30, 4 e15
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., Speed, T. P. (2003). „Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data" *Biostatistics* 4, 249-264
- Oshlack, A., Emslie, D., Corcoran, L., and Smyth, G. K. (2007). Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome Biology* **8**, R2
- Limma userguide

- Niektóre slajdy zapożyczone z kursu *NGFN-Courses in Practical DNA Microarray Analysis*

RÓŻNICOWA EKSPRESJA GENÓW

Szukamy odpowiedzi na pytania …

- Przeprowadzone zostały 2 eksperymenty
- Zbadaliśmy, że eksperymenty są dobrej jakości
  - *Histogram*, *imageplot*, …
- Usunięte zostały błędy systematyczne
  - Korekcja tła, normalizacja … *plotMA*, *boxplot*

- Które geny znajdujące się na mikromacierzy uległy *zróżnicowanej ekspresji?*

Czy wystarczy posortować po *log ratio*?

3 różne zbiory danych – zbiór A

Pobrano próbki od 20 pacjentów chorych na raka piersi przed i po 16-tygodniowym leczeniu chemioterapią. Zbadano je przy użyciu mikromacierzy.

Chcemy zidentyfikować geny, które uległy podwyższonej i obniżonej ekspresji związanej z leczeniem

Data are from the paper of Perou et al. (2000) and are available from the Stanford Microarray Database.

3 różne zbiory danych – zbiór B

Pobrany został szpik kostny od 27 pacjentów cierpiących na białaczkę ALL (*acute lymphoblastic leukemia*) i od 11 pacjetów cierpiących na białaczkę typu AML (*acute myleoid leukemia*). Zanalizowano próbki używając mikromacierzy Affymetrixowych

Chcemy zidentyfikować geny, które uległy podwyższonej i obniżonej ekspresji w ALL w porównaniu do AML

Data are from the paper of Golub et al. (1999) and are available from the Stanford Microarray Database

3 różne zbiory danych – zbiór C

Badano 4 typy nowotworów złośliwych drobnookrągłoniebiesko-komórkowych (small round blue cell tumors):

- Neuroblastoma (nerwiak płodowy) **NB**,
- non-Hodgkin lymphoma (chłonniak nieziarniczny) **NHL**,
- rhabdomyosarcoma (mięśniakomięsak prążkowanokomórkowy) **RMS**,
- Ewing tumors (mięsak Ewinga) **EWS**

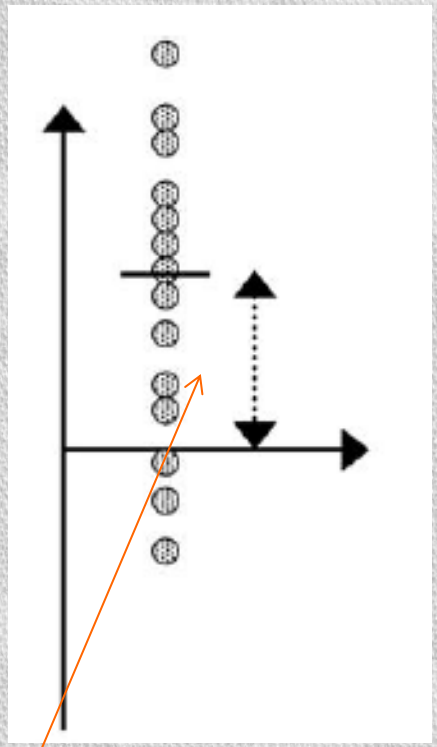
63 próbki z tych nowotworów (12, 8, 20, 23 z każdej z grup)

Chcemy zidentyfikować geny, które uległy zróżnicowanej ekspresji w jednej z tych 4 grup

The data are from the paper of Khan et al. (2001) and are available from the Stanford Microarray Database.

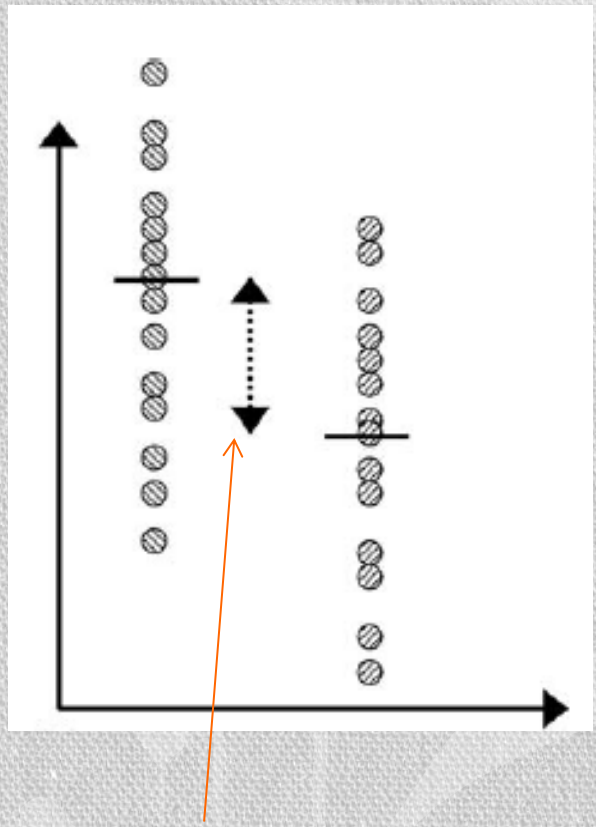
- W każdym z zestawów szukamy genów o zróżnicowanej ekspresji

W zestawie A mamy **dane sparowane**. Mamy 2 pomiary dla każdego pacjenta: przed i po chemii. Dane są ze sobą powiązane – to co nas interesuje to różnica pomiędzy dwoma pomiarami (*log ratio*), aby wykryć geny z podwyższoną i obniżoną ekspresją (*up-down-regulated*)



Pomiary dla każdego pacjenta są odjęte. Sprawdzamy czy mediana lub średnia są różne od 0

W zestawie B mamy **dane niesparowane**. Mamy 2 grupy pacjentów i chcemy zaobserwować jaki jest związek pomiędzy pacjentami w jednej grupie a pacjentami w drugiej grupie.



Mamy pomiary dla każdego pacjenta z obu grup. Sprawdzamy czy mediana lub średnia dla obu grup są różne.

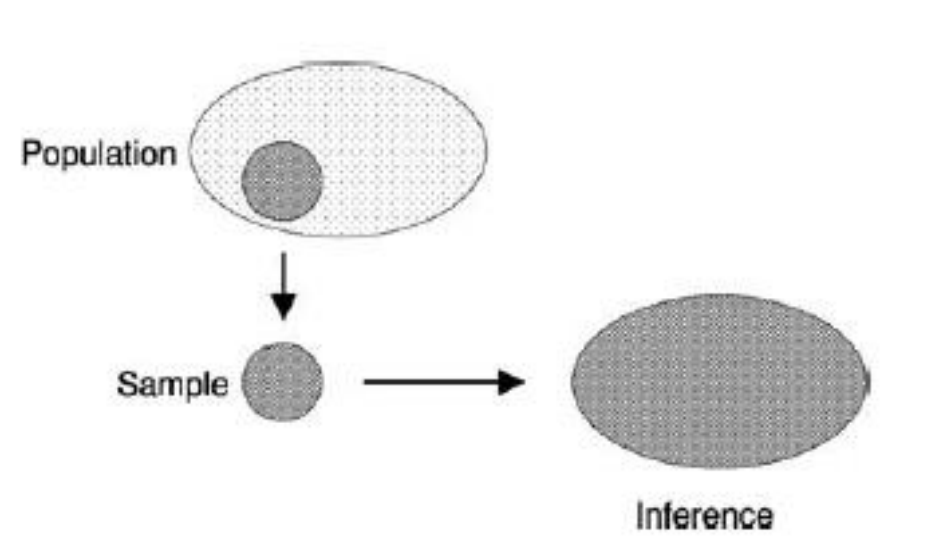
- Dane sparowane (zestaw A) i niesparowane (zestaw B) wymagają odmiennej analizy (sparowany i niesparowany test istotności t)

- Zestaw danych C składa się z 4 grup i wymaga bardziej skomplikowanej analizy, jak np. analizy wariancji (ANOVA)

Wiemy że w zestawie A dla pewnego genu różnica w ekspresji jest 2. Czy na tej podstawie możemy stwierdzić że ten gen uległ zróżnicowanej ekspresji?

Czy wystarczy posortować geny po *log ratio* i wyselekcjonować te, które spełniają założony próg?

Cała populacja vs wybrane osobniki

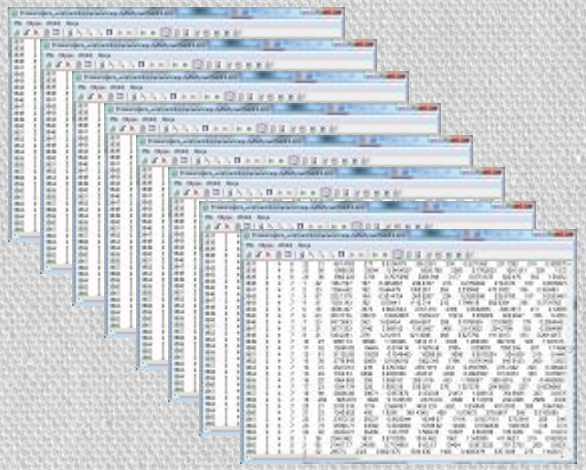


- Nie możemy zbadać całej populacji chorych na raka
- Wybieramy reprezentację kilku (dziesięciu) osobników
- Uogólniamy wyniki na całą populację pacjentów chorych na raka

Przeprowadzamy wnioskowanie statystyczne aby stwierdzić czy różnica w ekspresji genu pomiędzy jedną, a drugą grupą badanych osobników wynika z szumu w danych, z różnorodności pomiędzy osobnikami, czy rzeczywiście ze zróżnicowanej ekspresji

Dlatego potrzebujemy POWTÓRZEŃ eksperymentów

Grupa A



Grupa B

Im większa liczba powtórzeń, tym pewniejszy wynik

Standardowe odchylenie a standardowy błąd

Standardowe odchylenie (SD): zmienność pomiarów

Standardowy błąd (SE): zmienność średniej z kilku pomiarów

n powtórzeń

Dane z normalnym rozkładem

$$SE = \frac{1}{\sqrt{n}} SD$$

Hipoteza zerowa

H_0 zakłada że geny nie uległy zróżnicowanej ekspresji

Zbiór A – gen nie uległ zróżnicowanej ekspresji po chemii

Zbiór B – gen nie uległ zróżnicowanej ekspresji dla wszystkich pacjentów z białaczką typu ALL i AML

Jeśli H_0 jest prawdziwa, to znaczy że nie została stwierdzona istotna statystycznie zmiana w ekspresji.

Testy statystyczne

Każdy test hipotezy zerowej tworzy model, który wyznacza *prawdopodobieństwo* obserwowanej statystyki, np. średnie zróżnicowanie ekspresji genów.

To prawdopodobieństwo to ***p*-value** . Im mniejsze tym mniej prawdopodobne, że obserwowane dane pojawiły się przypadkowo i tym bardziej pewne wyniki.

Zakładamy że zróżnicowana ekspresja obserwowana dla genów z niską wartością *p*-value z małym prawdopodobieństwem pojawiła się przypadkowo, *a zatem jest skutkiem biologicznego efektu, który testujemy*.

p-value = 0.01 oznacza że jest 1% szansy na obserwowanie zróżnicowanej ekspresji przez przypadek

Niezależność zmiennych

Wszystkie testy statystyczne tu prezentowane wymagają, aby zmienne (pomiary) były niezależne.

Zbiór A – dane nie są niezależne, gdyż badamy pomiary poziomu ekspresji genów dla tych samych pacjentów (przed i po chemii). Dlatego dla każdego pacjenta badamy różnicę w ekspresji przed i po.

t-test sparowany – dla zbioru A (one-sample test)

Dla każdego pacjenta mamy 1 kolumnę liczb, odpowiadających wartościom *log ratio* (przed i po) dla każdego genu.

Dla każdego genu wyznaczamy statystykę $t = \frac{\bar{x}}{s/\sqrt{n}}$ SE

gdzie $\bar{x}$ to średnia log ratio (dla 1 genu) dla wszystkich pacjentów

s to standardowe odchylenie dla grupy pacjentów

n to liczba pacjentów

Wyznaczamy *p-value* poprzez porównanie t-statystyki z rozkładem T-studenta dla *n*-1 stopni swobody

t-test sparowany – dla zbioru A

Wykrycie genu ze zróżnicowaną ekspresją zależy od:

- średniej wartości *log ratio*
- odchylenia w populacji (czy jest rozkład normalny)
- liczby testowanych osobników

Można wykryć zmianę w ekspresji rzędu 1.5 (stosunek obu wartości) pod warunkiem że:

- różnorodność w populacji jest mała
- im większa badana próba osobników, tym łatwiej jest wykryć zmianę w ekspresji

t-test sparowany – algorytm dla **jednego** genu z zestawu A

1. Znormalizuj dane ze zbioru A. Wyznacz stosunek wartości (*log ratio)* dla próbki eksperymentalnej w stosunku do referencyjnej dla ‚przed' i ‚po' leczeniu (każdy eksperyment był dwu-kolorowy dlatego porównujemy do referencyjnej próbki)
2. Wyznacz pojedynczą wartość *log ratio* dla każdego pacjenta, czyli odejmij *log ratio* ‚przed' chemią od *log ratio* ‚po' chemii
3. Wykonaj statystykę *t* (wyznacz wartość wg wzoru). Wyznaczoną wartość *t* porównaj z rozkładem *t*-studenta dla 19 stopni swobody (20 pacjentów - 1). Odczytaj wartość parametru *p* dla dwustronnego sparowanego testu *t*

t-test niesparowany (two sample test) – zestaw B

Mamy dwie grupy pacjentów bez żadnego związku pomiędzy nimi. Sprawdzamy czy różnica pomiędzy średnimi w obu grupach jest równa zero.

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

$\bar{x}_1$ oraz $\bar{x}_2$ to średnie w obu grupach

s_1 oraz s_2 to odchylenia standardowe w grupach

n_1 oraz n_2 to liczność grup

t-test niesparowany przykład dla zestawu B

Chcemy znaleźć geny, które uległy podwyższonej lub obniżonej ekspresji AML w porównaniu do ALL.

1. Transformujemy dane do logarytmów
2. Dla każdej grupy pacjentów (AML, ALL) wyznaczamy średnią i odchylenie standardowe. Test statystyczny stwierdzi czy obie średnie są równe
3. Wyznaczamy parametr t wg wzoru
4. Odczytujemy wartość parametru *p* z tablic (rozkład t-studenta)

Dane dla genu Metallothinein IB

Patient	ALL Log	Patient	AML Log
1	8.60	28	8.42
2	7.85	29	8.35
3	8.85	30	9.58
4	8.20	31	9.18
5	7.60	32	9.41
6	8.21	33	8.96
7	8.47	34	8.81
8	8.51	35	9.55
9	8.75	36	8.18
10	6.75	37	8.71
11	7.93	38	9.46
12	7.71		
13	7.88		
14	7.55		
15	6.61		
16	8.75		
17	9.32		
18	8.40		
19	7.16		
20	8.41		
21	4.75		
22	7.92		
23	7.82		
24	8.42		
25	7.08		
26	7.38		
27	9.29		
Średnia	7.93		8.97
Odchylenie st.	0.94		0.51

$t = 4.35$

Porównanie z rozkładem t-studenta z 33 stopniami swobody daje wartość

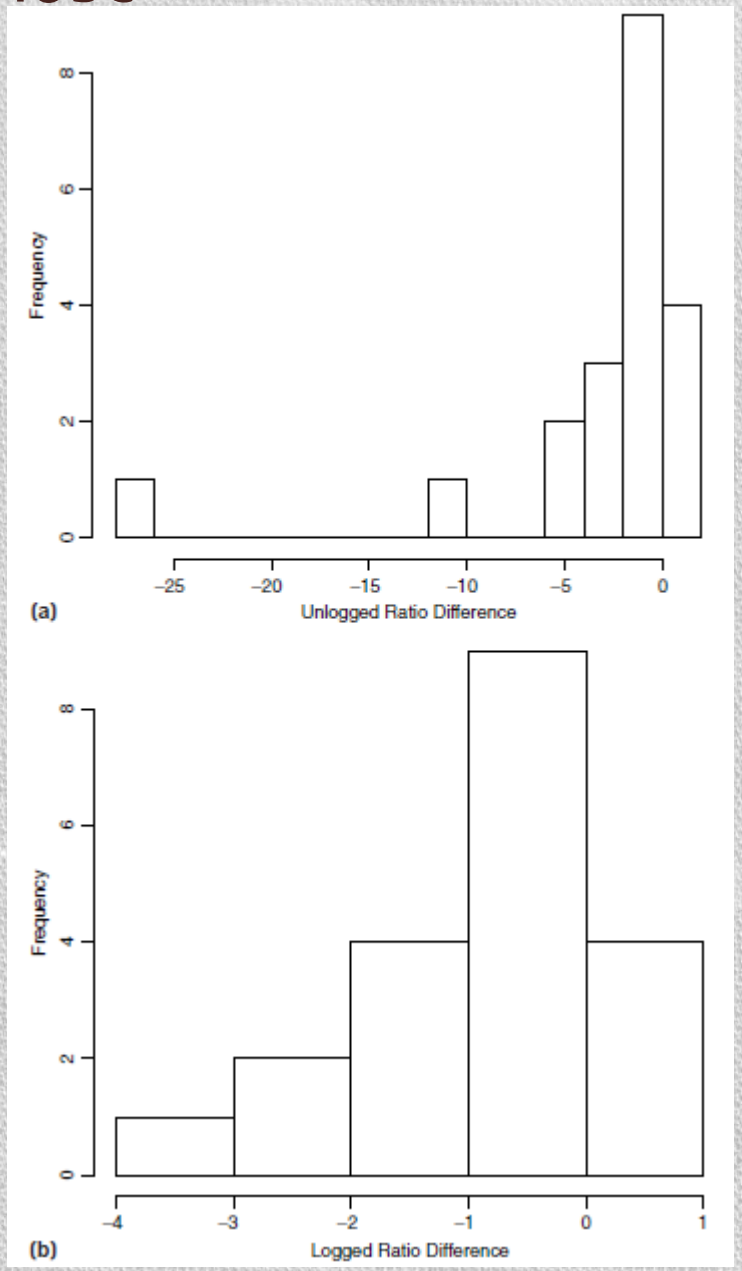
$p = 0.00012$

Możemy stwierdzić że gen uległ większej ekspresji u pacjentów z AML niż u pacjentów z ALL

Zestaw A – przykład na konieczność rozkładu normalnego

gen Diubiquitin

Patient	Unlogged Difference	Log Ratio	Fold Change
7	−10.08	−2.91	−7.54
10	0.85	0.62	+1.54
12	−0.28	−0.11	−1.08
14	0.04	0.08	+1.06
15	−0.68	−0.42	−1.34
18	0.17	0.12	+1.09
26	−4.93	−0.99	−1.99
27	−0.12	−0.16	−1.12
39	−1.67	−0.44	−1.35
41	−27.98	−1.64	−3.12
47	−0.92	−0.55	−1.46
48	−2.00	−0.99	−1.99
53	−3.04	−1.37	−2.58
61	−3.80	−2.05	−4.14
100	−3.53	−3.20	−9.18
101	−1.44	−1.12	−2.17
102	−0.62	−0.72	−1.64
104	−4.50	−1.19	−2.27
109	−0.23	−0.34	−1.27
112	0.10	0.12	+1.09



Dane surowe nie mają rozkładu normalnego, w przeciwieństwie do danych po transformacji logarytmicznej

Zestaw A

gen Diubiquitin

Data	Mean	Standard Deviation	Standard Error	Standard Error/Mean	*t*-statistic	*p*-value
Unlogged	−3.23	6.36	1.46	−0.45	−2.27	0.03
Logged	−0.86	1.00	0.23	−0.27	−3.86	0.001

Przykład jest mało intuicyjny …

- średnia z danych surowych bardziej wskazuje na obniżoną ekspresję
- Wartości odstające powodują wzrost błędu
- Dlatego obniżona jest wartość statystyki

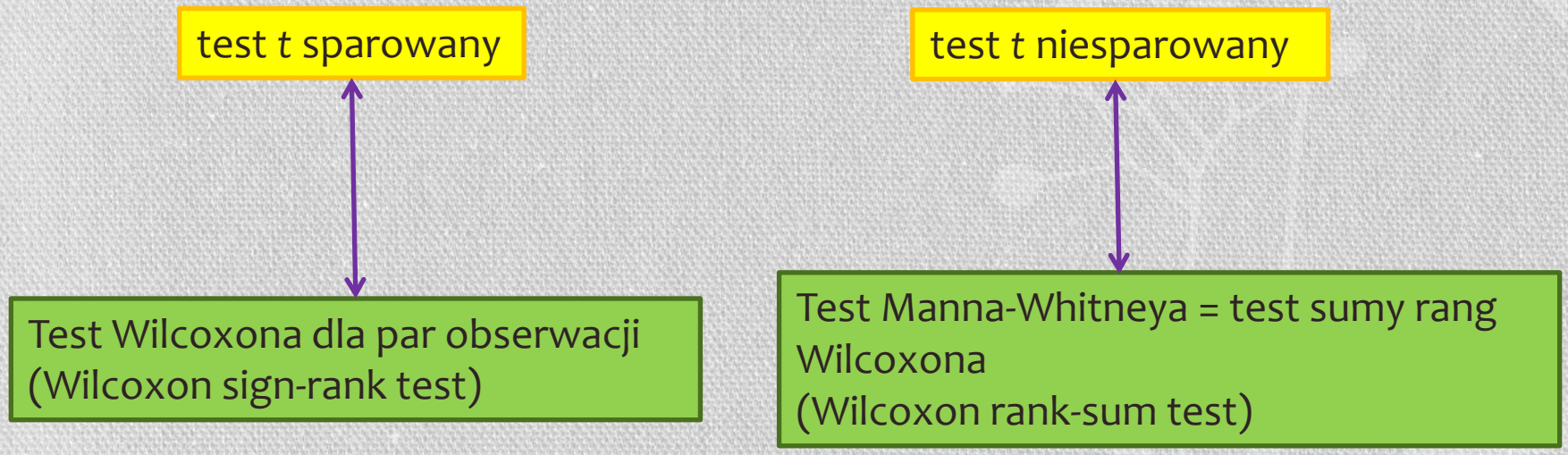
… ale wskazuje konieczność rozkładu normalnego danych

Co zrobić jeśli dane nie mają rozkładu normalnego?

- Dane mikromacierzowe zawierają szum – często zdarzają się wartości odstające
- Analiza danych mikromacierzowych jest wysokoprzepustowa – tysiące genów sprawdzane są jednocześnie.
  - Dla każdego genu sprawdzamy czy rozkład jest normalny
  - Geny zawierające wartości odstające (bez rozkładu normalnego) wymagają innej analizy
  - Przy braku rozkładu normalnego w danych lepiej stosować testy nieparametryczne -> dla wszystkich genów

Test nieparametryczne

Odpowiedniki testów sparowanych i niesparowanych



Test nieparametryczne

Test Wilcoxona dla par obserwacji

1. Sortujemy po bezwzględnej wartości różnic (przed i po chemii)
2. Dajemy rangi: 1 dla najmniejszego, 2 dla kolejnego, …
3. Wyznaczamy sumę rang dla obserwacji z różnicą dodatnią i ujemną (podwyższona i obniżona ekspresja)
4. Przeliczamy wartość statystyki: średnia/odchylenie standarowe
5. Porównujemy wynik z tablicą dla rozkładu normalnego, aby otrzymać wartość *p*

Test Manna-Whitneya

Działa podobnie, z tymże dane z dwóch grup są łączone i dopiero później rangowane

Test nieparametryczne - przykład

Dla przykładu z zestawu A – gen Diubiquitin

Data	Mean	Standard Deviation	Standard Error	Standard Error/Mean	*t*-statistic	*p*-value
Unlogged	−3.23	6.36	1.46	−0.45	−2.27	0.03
Logged	−0.86	1.00	0.23	−0.27	−3.86	0.001

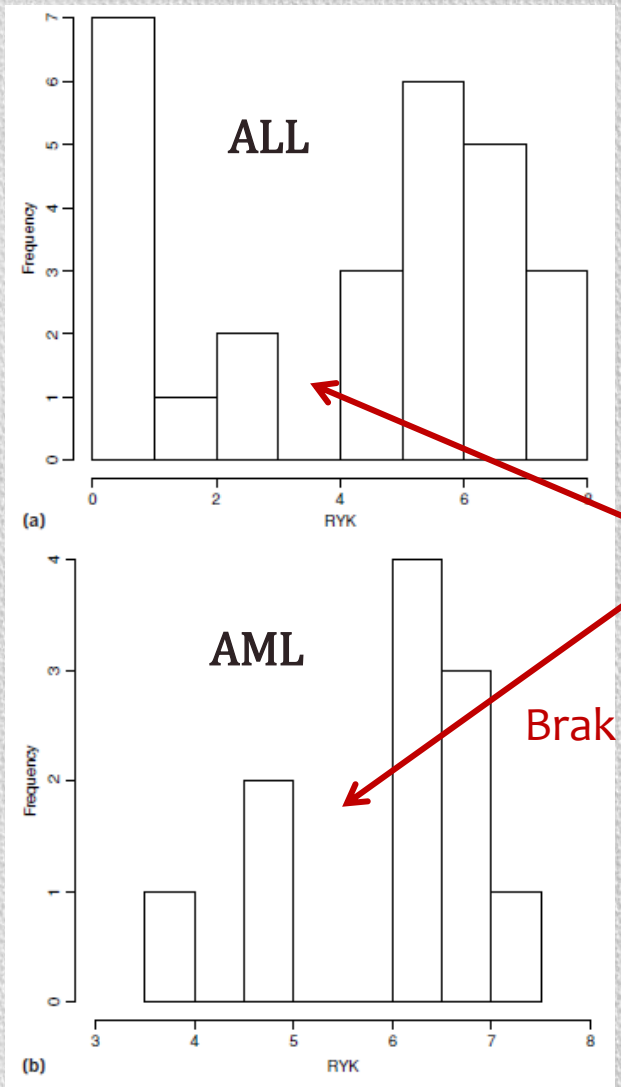
Otrzymujemy:

Unlogged: p-value = 0.00032

Logged p-value = 0.00048

Test Wilcoxona jest odporny na dane odstające (przed transformacją logarytmiczną)

Test nieparametryczne – przykład nr 2



Test Mann Whitneya:

p-value = 0.039

Test t niesparowany

p-value = 0.0032

Test t nie jest odpowiedni do analizy i dlatego jego wyniki nie są wiarygodne

Brak rozkładu normalnego

Test nieparametryczne

Zalety

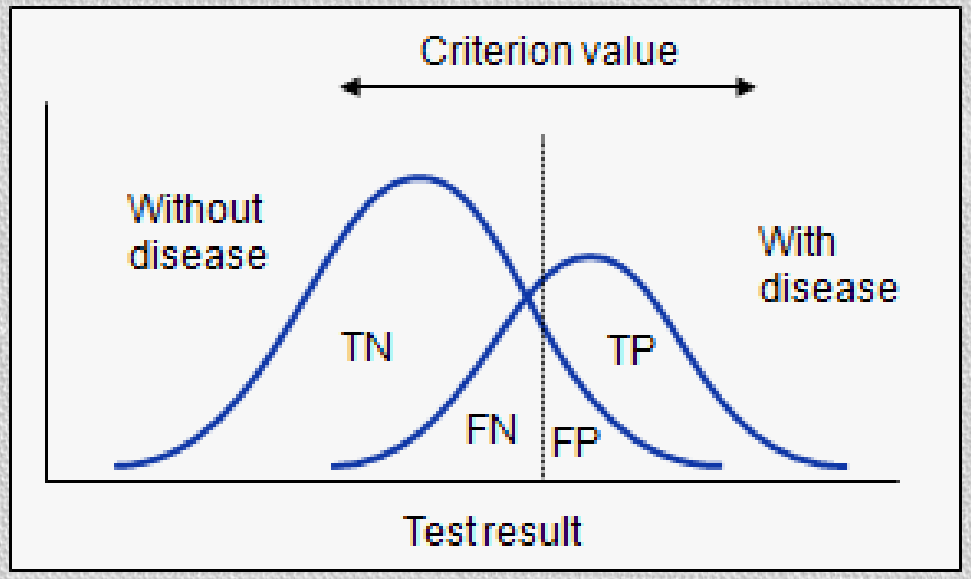
- Dane nie muszą mieć rozkładu normalnego
- Testy są odporne na wartości odstające

Wady

- Testy nie maja takiej mocy jak testy parametryczne

Moc testu – prawdopodobieństwo zaobserwowania pozytywnego wyniku, jeśli rzeczywiście jest pozytywny wynik

	Odrzucenie hipotezy zerowej	Nie-odrzucenie hipotezy zerowej
Hipoteza zerowa (H_0) jest prawdziwa	False positive **Błąd I typu**	True positive **poprawne**
Hipoteza zerowa (H_0) jest fałszywa	True negative **poprawne**	False negative **Błąd II typu**



H_0

Błąd typu I (false positive)

- Uznajemy że coś jest prawdziwe chociaż nie jest
- Odrzucona hipoteza zerowa, która mówi że geny nie uległy zróżnicowanej ekspresji
- Uznajemy zatem że geny uległy zróżnicowanej ekspresji, chociaż faktycznie tak nie było

- Współczynnik błędów I typu jest nazywany rozmiarem testu i oznaczany przez α. Poziom istotności testu (*significance*)
- (1- α) *swoistość (specificity)*

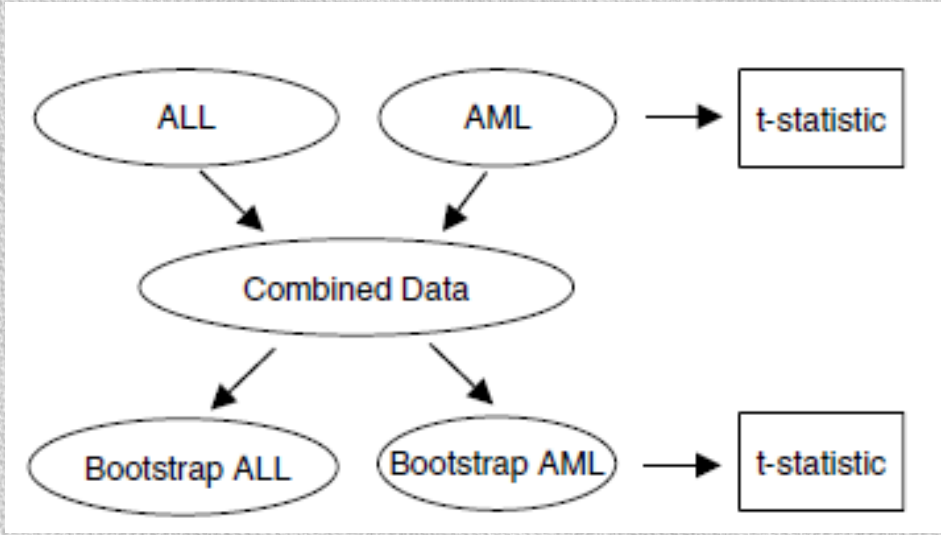
Błąd II typu (false negative)

- Nie udało się odrzucić hipotezy zerowej, chociaż faktycznie nie była spełniona
- Nie udało nam się znaleźć genów, które uległy zróżnicowanej ekspresji

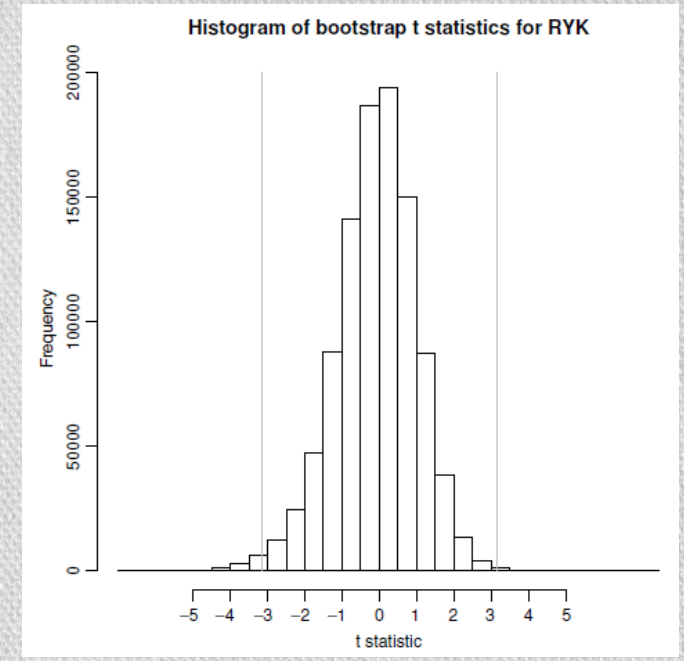
- Oszacowanie prawdopodobieństwa popełnienia błędu II typu oznaczamy przez β.
- (1- β) nazywamy *mocą testu* lub czułością (*sensitivity*)

Metoda bootstrap – „ciągnięcie za sznurówki”

- Utworzenie wystarczająco dużej populacji osobników, poprzez tworzenie grup osobników o takiej samej liczności jak zbiór oryginalny (reprezentacyjna próbka) z losowo wybranych (repróbkowanie) osobników ze zbioru
- Wyznaczamy statystykę (np. t-test dla niesparowanych) dla każdej z tak utworzonych grup
- Wyznaczamy statystykę dla zbioru oryginalnego
- Obliczamy p-value poprzez wyznaczenie proporcji statystyk bootstrapowych, które mają wartość bardziej ekstremalną niż wartość statystyki zbioru oryginalnego



Metoda bootstrap – przykład dla testu t niesparowanego – gen RYK



1. Wyznaczamy statystykę *t* dla genu RYK

   t=3.1596

2. Tworzymy 1 000 000 zbiorów bootstrapowych, każdy składający się z 27 ALL i 11 AML. Zbiory tworzone są przez losowanie wartości ze zbioru oryginalnego

3. Dla każdego bootstrapowego zbioru wyznaczamy statystykę t (1000 000 różnych wartości)
4. Z 1 000 000 wartości 9 750 ma absolutną wartość większą niż 3.1596, co daje nam wartość $p < 0.01$

Możemy stwierdzić, że gen uległ zróżnicowanej ekspresji

Wielokrotne testowanie - problem

- Z definicji p-value, każdy gen ma 1% szansy posiadania wartości $p<0.01$, czyli będzie znaczący przy poziomie istotności 1%
- Dla 10000 genów, oczekujemy że
  - 100 genów przejdzie próg $p<0.01$
  - 10 genów przejdzie próg $p<0.001$
  - 1 gen przejdzie próg $p<0.0001$
- Dla zestawu A mamy 9216 genów. Jeśli chemioterapia nie miałaby żadnego wpływu na zmianę ekspresji genów to i tak oczekiwalibyśmy, że dla 92 genów $p<0.01$

- Czy gen naprawdę uległ zróżnicowanej ekspresji, czy jest to wynik błędu I typu (false positive)?

Kontrolowanie false positives

- **Family-wise error rate** (**FWER**)
  - Prawdopodobieństwo co najmniej jednego błędu I typu pomiędzy genami wybranymi jako znaczące

$$FWER = \Pr(FP > 0)$$

- **False discovery rate** (**FDR**)
  - Oczekiwana proporcja błędów I typu spośród odrzuconych

$$FDR = E(Q), \text{gdzie } Q = \begin{cases} \frac{FP}{R}, & \text{jeśli } R > 0 \\ 0, & \text{jeśli } R = 0 \end{cases}$$

R – to suma False Positive i True Negative (czyli wszystkich z odrzuconą hipotezą)

Korekcja *p-value* Bonferroni

- Załóżmy że przeprowadziliśmy testowanie hipotezy dla każdego z *n* genów, wyznaczyliśmy:
  - statystykę t_i dla *i*-tego genu
  - wartość p_i dla *i*-tego genu

- Korekcja Bonferroni:

$$p_i' = \min(n * pi\,, 1)$$

- Wybierając geny $p_i' \leq \alpha$ kontrolujemy FWER na poziomie $\Pr(FP>0) \leq \alpha$

α- poziom istotności

Korekcja Bonferroni - wada

Przy dużej liczbie genów korekcja może spowodować, że dla żadnego genu nie będziemy mogli odrzucić hipotezy zerowej

TABLE 7.9: Significant Genes from the Breast Cancer Data Set

The unadjusted *p*-values are the proportion of the 100,000 bootstrap data sets that had *t*-statistics more extreme than the *t*-statistic from the real data. Thus the smallest possible *p*-value is 1/100,000 (or 10^{-5}). Because of the number of genes in the analysis, the Bonferroni corrected *p*-values are all too large to be significant, illustrating that this method is not applicable to most microarray data.

Accession	Description	*p*-Value	Bonferroni Adjusted *p*-Value
AA598794	connective tissue growth factor	10^{-5}	0.064
N23941	cyclin-dependent kinase inhibitor 1A	10^{-5}	0.064
AA478553	dopachrome tautomerase	10^{-5}	0.064
W96134	v-jun avian sarcoma virus 17 oncogene homolog	10^{-5}	0.064
AA044993	connective tissue growth factor	10^{-5}	0.064
AA040944	v-fos FBJ murine osteosarcoma viral oncogene homolog	10^{-5}	0.064
N95402	copine V	2×10^{-5}	0.13
R12840	v-fos FBJ murine osteosarcoma viral oncogene homolog	3×10^{-5}	0.19
AA442853	cyclin-dependent kinase 5, regulatory subunit 1 (p35)	$4.\times10^{-5}$	0.25
AA418077	GTP-binding protein overexpressed in skeletal muscle	5×10^{-5}	0.32
AA133129	transcription elongation factor B (SIII), polypeptide 3	5×10^{-5}	0.32
AA485377	v-fos FBJ murine osteosarcoma viral oncogene homolog	6×10^{-5}	0.38
AA134757	fibulin 1	6×10^{-5}	0.38
AI831083	dihydropyrimidinase-like 3	7×10^{-5}	0.45
AA004637	ESTs	9×10^{-5}	0.57
No Annotation		1.2×10^{-4}	0.76
AA025939	CD4 antigen (p55)	2×10^{-4}	1.3
H21041	activating transcription factor 3	2.3×10^{-4}	1.5
AA449463	KIAA0220 protein	2.6×10^{-4}	1.7
H05099	KIAA0182 protein	3.8×10^{-4}	2.4

Korekcja Benjamini – Hochberg

Kontrolowanie *FDR* = *E*(*FP*/*R*) na poziomie α

1. Sortowanie wartości p: $p_{r1} \leq p_{r2} \leq \cdots \leq pr_n$
2. Wyznaczenie:

$$j' = \max\{j: p_{rj} \leq (j/n)* \alpha\}$$

3. Odrzucenie hipotezy H_{rj} dla j=1,… , j’

ANOVA

One-way ANOVA

Dla zestawu C (cztery typy nowotworów złośliwych drobnookrągło-niebiesko-komórkowych) chcemy zidentyfikować geny, które uległy różnicowej ekspresji genów dla co najmniej jednego typu nowotworu

Możemy testować każdą parę grup pacjentów oddzielnie (wzrasta błąd pomiarowy – false positives)

Metoda ANOVA analizuje wszystkie grupy razem zwracając

tylko jedną wartość *p-value* dla każdego genu

ANOVA

multifactor ANOVA

- W *one way ANOVA* badaliśmy jeden czynnik – typ raka
- W *multifactor ANOVA* możemy sprawdzić czy na ekspresję genów oprócz typu raka ma wpływ także płeć

Metoda ANOVA analizuje wszystkie grupy razem zwracając jedną wartość *p-value* dla każdego badanego czynnika (płeć, typ raka)

Dla metody ANOVA możemy zbadać czy odpowiedź na czynniki jest addytywna (czynniki nie są od siebie zależne) czy multiplikatywna (jest interakcja pomiędzy czynnikami)

Złagodzona (moderated) statystyka *t*

- Dla statystyki t estymujemy wariancję dla każdego genu oddzielnie.
  - *Problem zaczyna się gdy mamy tylko kilka powtórzeń w każdej testowanej grupie (2-5) – nie możemy przewidzieć poprawnie wariancji*

- Dla złagodzonej statystyki t, zastępujemy wariancję – specyficzną dla genu s_g^2, wariancją globalną s_0^2 – czyli przewidzianą na podstawie tysiąca genów *typową* wariancję (metoda **empirical Bayes**)

$$T_g \sim \frac{\bar{X}_{g1} - \bar{X}_{g2}}{\sqrt{\mu s_g^2 + \lambda s_0^2}}$$

Metoda używana również do kwalifikowania w klientach pocztowych, które z otrzymanych wiadomości są SPAMem

Model liniowy

- Ekspresja genu y_j genu w próbce *j* jest zamodelowana jako *liniowo zależna* od kilku atrybutów (czynników, jak np. typ komórki, traktowanie)

$$y_j = a_1 xj_1 \; _+ a_2 xj_2 + \cdots + aj \; xj_k + \varepsilon_j$$

- Współczynniki a_j wyznaczane są metodą najmniejszych kwadratów (*least squares*), albo zachłanną (*robust*)
- Trzeba zdefiniować macierz projektową (*design matrix*) i kontrastów (*contrast*)
  - Projektowa – wskazuje na grupę próbki z mikromacierzy (x_{ji} ze wzoru). Wiersze odpowiadają macierzom, kolumny współczynnikom liniowego modelu
  - Kontrastowa – definiuje , które próbki z którymi należy porównać

Model liniowy

- Macierz kontrastów – definiuje grupy, które należy porównać (testować istotność (*t*-test))

Przykład: rozważmy badanie trzech różnych typów raka nerki. Dla każdego genu tworzymy model liniowy

$$y_j = a_1 xj_1 + a_2 xj_2 + a_3 xj_3 + \varepsilon_j$$

gdzie x_{ji}= 1 jeśli typ raka *j* jest rakiem typu *i* lub x_{ji}= 0 w p.p.

Współczynniki a_j estymowane przez metodę minimalnych kwadratów są średnią poziomu ekspresji w grupie.

Podsumowanie *eBayes* i *moderated t-test*

- Metoda ta pozwala na efektywne identyfikowanie genów ulegających zróżnicowanej ekspresji, szczególnie wtedy gdy liczba powtórzeń jest niewielka (estymacja wariancji dla pojedynczego genu na podstawie puli wszystkich genów)
- Moc testu wzrasta (czyli poprawne odrzucenie H_0) poprzez wykorzystanie wszystkich genów do estymacji

Podsumowanie

- Badamy różnicową ekspresję genów
- Zamiast badać log-ratio, badamy czy ta zmiana jest rzeczywiście statystycznie istotna
- Stawiamy hipotezę zerową (brak różnicy ekspresji genu pomiędzy badanymi grupami) – statystyki *t* sparowane i niesparowane
- Statystyki t wymagają rozkładu normalnego dla danych wejściowych
- Często wykorzystuje się statystyki nieparametryczne (Wilcoxon, Mann-Whitney) – mniejsza moc testów
- Abu zwiększyć moc testu wprowadza się metody bootstrapowe
- Testowanie wielu czynników – metoda ANOVA
- Zbyt mało powtórzeń powoduje że nie możemy przewidzieć rozkładu zmiennych – możemy zastosować korekcję Bayesowską i złagodzoną statystykę t (moderated t-test)
- False discovery rate – kontrola false positives, korekcja p-value