Katedra Biosensorów i Przetwarzania Sygnałów Biomedycznych

Wydział Inżynierii Biomedycznej

POLITECHNIKA ŚLĄSKA

PRACA DYPLOMOWA MAGISTERSKA

Narzędzie do analizy danych genomicznych

Kamil SUCHANEK

Kierunek studiów: *Inżynieria Biomedyczna*

Specjalność: *Przetwarzanie i Analiza Informacji Biomedycznej*

PROMOTOR

Dr inż. Anna Tamulewicz

ZABRZE – 2020

Wydział Inżynierii Biomedycznej
POLITECHNIKA ŚLĄSKA

OŚWIADCZENIE

Kamil SUCHANEK
Kierunek studiów: Inżynieria Biomedyczna
Specjalność: Przetwarzanie i Analiza Informacji Biomedycznej
Rok akademicki: 2019/2020

PROMOTOR: *Dr inż. Anna TAMULEWICZ* KONSULTANT: —-

Oświadczam, że niniejszą pracę wykonałem osobiście oraz, że przy jej realizacji nie naruszono praw osób trzecich, wynikających z przepisów prawa autorskiego.

Zabrze, dnia 10.11.2020 roku .....
(własnoręczny podpis studenta)

Zgodnie zaświadczamy, że wyrażamy zgodę*) nie wyrażamy zgody*) na udostępnienie niniejszej pracy dyplomowej.

....
(własnoręczny podpis studenta) *(własnoręczny podpis promtora)*

Zabrze, dnia 10 listopada 2020 roku

*)niepotrzebne skreślić

Spis ilustracji

Spis treści

Streszczenie

Niniejsza praca magisterska dotyczy utworzenia narzędzia do analizy danych genomicznych, dotyczących ekspretymentu mikromacierzowego. W ramach tej pracy poczyniono badania nauk bioinformatycznych, z dokładnym uwzględnieniem technologii mikromacierzowej oraz utworzono aplikację w środowisku Shiny języka programowania R do typowej analizy danych mikromacierzowych. Narzędzia użyto do przeprowadzenia przykładowej analizy na zbiorze danych z kohorty schizofrenii w celu odnalezienia istotnych wzorców i różnic w stosunku do grupy kontrolnej.

Abstract

This master's thesis concerns the creation of a tool for the analysis of genomic data related to the microarray experiment. As part of this work, research in the field of bioinformatics was carried out, with careful consideration of microarray technology, and an application was created in the Shiny environment of the R programming language for typical analysis of microarray data. The tool was used to perform an exemplary analysis on a dataset from the schizophrenia cohort to find significant patterns and differences from the control group.

1. Cel i założenia

Celem pracy jest stworzenie narzędzia do analizy danych, pochodzących z eksperymentów mikromacierzowych.

Założeniem jest stworzenie multiplatformowego oprogramowania, opartego o biblioteki języka C oraz interfejs graficzny i logikę stworzoną przy pomocy języka programowania Ruby. Utworzona aplikacja powinna być w stanie poprawnie odczytywać wynikowe pliki typowe dla tego eksperymentu oraz oferować podstawowe metody analizy danych względnej ekspresji genów. Narzędzie powstałe w ramach pracy powinno być również wykorzystane do analizy konkretnych danych mikromacierzowych.

2. Podstawy teoretyczne

Technologia mikromacierzowa jako narzędzie genomiki funkcjonalnej powstała dzięki gwałtownemu postępowi w zakresie nauk bioinformatycznych i eskperymentalnych, na które składają się wydajne techniki sekwencjonowania i analizy danych.

2.1. Sekwencjonowanie DNA

Sekwencjonowanie DNA jest zabiegiem polegającym na ustaleniu kolejności par zasad w badanym kwasie nukleinowym. Od momentu wprowadzenia technik sekwencjonowania przez między innymi Sangera, Maxama i Gilberta w 1977 roku nastąpił intensywny rozwój owych technik. Popularna stała się enzymatyczna metoda Sangera z użyciem dideoksynukleotydów, zwana metodą terminacji łańcucha. Metody z czasem ulegały wielu modyfikacjom, mającym na celu uczynienie procesu sekwencjonowania bardziej wydajnym, bezpiecznym i tańszym poprzez np. zastosowania znaczników fluoroforowych zamiast izotopowych. Po 10 latach od opublikowania metody sekwencjonowania Sangera pojawiły się pierwsze sekwenatory automatyczne, które zamiast rozdzielać produkty sekwencjonowania w żelach poliakrylamidowych zostały wyposażone w cienkie kapilary ku temu służące. Dalszy rozwój automatyzacji i miniaturyzacji sekwenatorów umożliwił równoczesne sekwencjonowanie kilkuset fragmentów DNA. Równolegle z rozwojem tych technik opracowywano nowe strategie tworzenia bibliotek genomów do ich sekwencjonowania. Istotnym postępem w tej dziedzinie było wprowadzenie nowej techniki sekwencjonowania dużych genomów - metody shotgun, polegającej na sekwencjonowaniu dużej liczby losowo pofragmentowanych odcinków DNA, które następnie są składane komputerowo. Podejście to wymogło wprowadzenia nowatorskich metod obliczeniowych składających setki tysięcy losowo uzyskanych fragmentów sekwencji DNA w dłuższe części. Powzięta strategia obniżyła czas i koszty sekwencjonowania. Metoda

ta posłużyła do zsekwencjonowania całego genomu ludzkiego, co opublikowano w 2001 roku [1].

Gwałtowny wzrost liczby kompletnie zsekwencjonowanych genomów oraz rozwój nowoczesnych technik sekwencjonowania dostarczył olbrzymich ilości danych użytecznych w dziedzinie z pogranicza biologii i informatyki - genomice [1].

2.2. Genomika

Genomika jest nauką o genomach, powstała ona na skutek intensywnego rozwoju technik biologii molekularnej, umożliwiających wydajne sekwencjonowanie genomów [2].

Dane pochodzące z sekwencjonowania deponowane są w postaci elektronicznej w bazach danych pierwotnych, najczęściej w GenBank (Stany Zjednoczone), EMBL (European Molecular Biology Laboratory Nucleotide Sequence Database, Wielka Brytania) lub DDBJ (DNA Data Bank of Japan, Mishima, Japonia), które tworzą jedno konsorcjum - The International Sequence Database Collaboration i wymieniają się na bieżąco danymi. Z tych baz danych korzystają tzw. bazy wtórne, które przetwarzają w różny sposób informacje o sekwencjach i strukturach. Osobną kategorię stanowią wyspecjalizowane bazy poświęcone określonym genomom i innym zagadnieniom genomicznym oraz bioinformatycznym [2].

Genom, na który składają się wszystkie sekwencje DNA zawarte w organizmie, podlega analizie, której celem może być rozpoznanie sekwencji kodujących, sekwencji regulatorowych i sekwencji powtórzonych oraz określenie ogólnej organizacji, np. zróżnicowania składu nukleotydowego w regionach chromosomu, rozmieszczenia genów na chromosomie, organizacji genów w operony. Zadania genomiki dotyczą również analizy transkryptomu, proteomu, lokalizomu, interaktomu i metabolomu, które również posiadają sprecyzowane ku sobie dziedziny nauki [2].

Wyróżnia się termin genomiki obliczeniowej, traktującej o skomplikowanych algorytmach i technikach obliczeniowych oraz genomiki funkcjonalnej dotyczącej badań eksperymentalnych związanych z analizą genomu. Eksperymenty mogą dotyczyć zbadania efektów fenotypowych, na skutek ingerencji w ekspresję genów, interakcji pomiędzy białkami, lokalizowania białek w komórce, analizy ekspresji genów przy pomocy chipów DNA, izolowania i charakterystyki elektroforetycznej

i strukturalnej białek oraz identyfikowania kompleksów białek za pomocą spektrometrii masowej. Rozgraniczenie poszczególnych zagadnień jest dyskusyjne z powodu dość płynnej granicy pomiędzy nimi. Genomika może również badać charakterystyczne regiony w sekwencjach, motywy, domeny, sekwencje homologiczne poprzez porównywanie genomów i poszczególnych sekwencji, co nazywa się genomiką porównawczą. Wyróżnia się również genomikę ewolucyjną, opisującą wszystkie aspekty w kontekście procesów ewolucyjnych, genomikę strukturalną zajmującą się określaniem struktur przestrzennych białek i farmakogenomikę dotyczącą projektowania chemioterapeutyków na podstawie informacji płynących z analiz genomowych [2].

2.3. Mikromacierze

Jedną z najczęściej stosowanych metod genomiki funkcjonalnej są badania mikromacierzowe, które umożliwiają szczegółowy wgląd w procesy komórkowe zaangażowane w regulację ekspresji genów. Celem eksperymentu mikromacierzowego jest pomiar poziomu ekspresji genów w komórce, tj. Zmierzenie stężeń mRNA tych genów. Roztwory ekstrahowane z próbek tkanek zawierają dużą liczbę mRNA wielu różnych typów, które akurat były obecne w komórkach w tym czasie. Nie można ich używać bezpośrednio na chipie [3].

Mikromacierz to szkiełko z umocowanymi w sposób regularny cząsteczkami DNA. Mocowane są one w określonych miejscach zwanymi plamkami (ang. spots). Plamki mają na sobie fragmenty identycznych DNA. Każda grupa cząsteczek jednoznacznie odpowiada danemu genowi. DNA na plamkach może być genomowym DNA albo krótkim fragmentem nici oligonukleotydowej. Za produkcję mikromacierzy odpowiada robot nadrukowujący plamki na szkiełku albo owe są syntetyzowane w procesie fotolitografii [4].

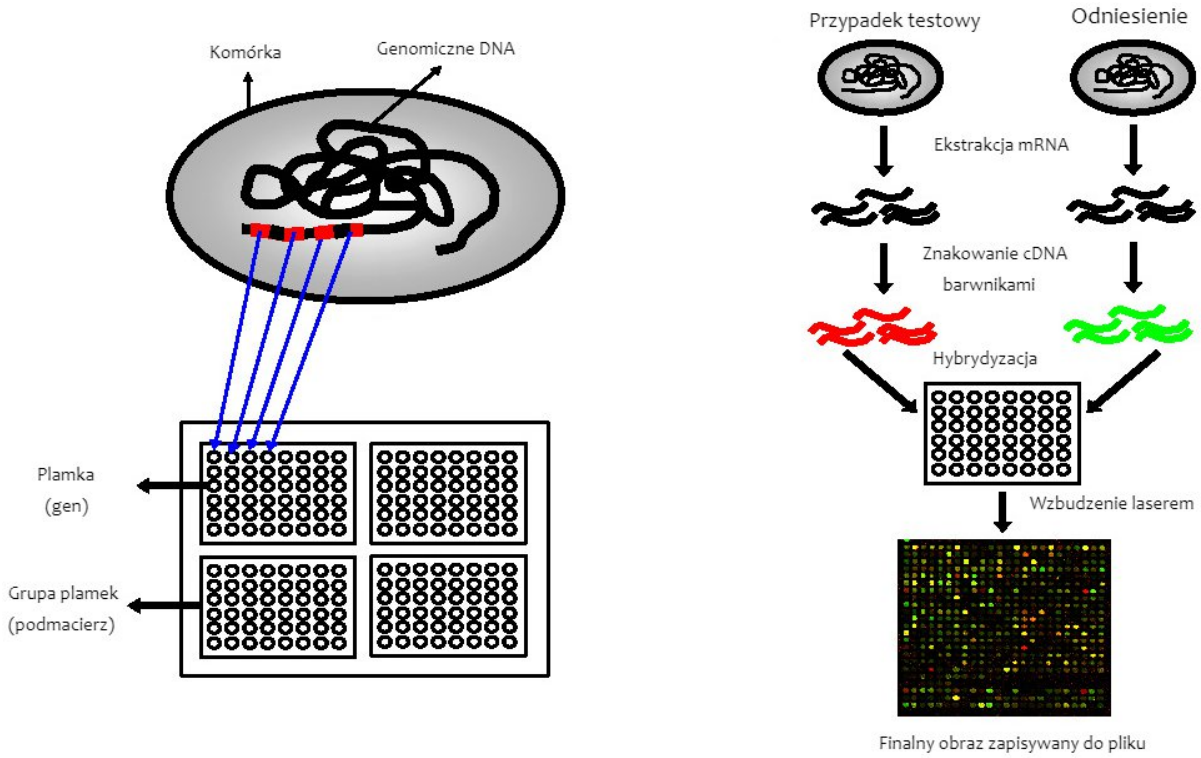
Istnieją dwa rodzaje mikromacierzy, które różnią się sposobem przygotowania zarówno macierzy jak i próbki. Pierwszy typ to macierze oligonukleotydowe, ponieważ DNA przyłączone do macierzy znajduje się w postaci krótkich oligonukleotydów, zwykle o długości 25 zasad. Tego typu tablice są produkowane komercyjnie przez firmę Affymetrix. Sekwencje użyte w każdej plamce są starannie wybierane z wyprzedzeniem w odniesieniu do genomu badanego organizmu. Każdy oligonukleotyd powinien hybrydyzować z określoną sekwencją genów organizmu. Jednak hybrydyzacja krzyżowa jest możliwa między genami z pokrewnymi

sekwencjami. Z tego powodu dla każdego genu wybiera się kilka oddzielnych sekwencji oligonukleotydowych, a ekspresję genu w próbce można wykryć tylko wtedy, gdy zachodzi hybrydyzacja z prawie wszystkimi z nich. Po wybraniu sekwencji oligonukleotydowych preparat przygotowuje się przez syntezę sekwencji, po jednej zasadzie na raz, w odpowiednim miejscu na szkiełku. Ten rodzaj macierzy jest czasem nazywany „chipem", ponieważ powierzchnia jest wykonana z krzemu. Stosowaną techniką jest fotolitografia. Do szkiełka dołączone są łączniki chemiczne, które zakotwiczą sekwencje. Te łączniki są chronione przez wrażliwą na światło grupę chemiczną. Na szkiełko nanoszone jest następnie roztwór zawierający jeden określony rodzaj nukleotydu - np. A. Światło jest następnie kierowane dokładnie na te plamki macierzy, w których wymagane jest A w sekwencji. To aktywuje wiązanie A z łącznikiem w tych pozycjach, ale nie gdzie indziej. Roztwór A jest następnie wypłukiwany i zastępowany każdym z pozostałych nukleotydów, jeden po drugim. Po wykonaniu tego cztery razy pierwsza zasada każdej sekwencji zostanie zsyntetyzowana. Proces jest następnie powtarzany dla każdej zasady w sekwencji. Po przygotowaniu procesu produkcyjnego dla danego chipa można zsyntetyzować wiele kopii zawierających dokładnie te same oligomery w powtarzalny sposób [3]. Tak powstały chip zanurza się w mieszaninie RNA znakowanych biotyną na kilka godzin, podczas których zachodzi hybrydyzacja z oligomerami DNA na chipie. Pozostałe niezhybrydyzowane sekwencje są następnie wypłukiwane. Ostatnim etapem jest dodanie połączonej z streptawidyną cząsteczki fluoroscencyjnej, która będzie wiązać się z biotyną wszędzie tam, gdzie na chipie znajduje się zhybrydyzowany RNA. Poziom fluorescencji można następnie zmierzyć w każdej z plamek przy pomocy mikroskopii optycznej [3].

Drugim typem mikromacierzy jest macierz cDNA. cDNA to nić DNA syntetyzowana przy pomocy enzymu odwrotnej transkryptazy, która tworzy sekwencję DNA komplementarną do matrycy RNA. Jest to odwrotność tego, co dzieje się w transkrypcji, gdzie nić RNA jest tworzona przy użyciu szablonu DNA. Istnieje możliwość syntezy cDNA z mRNA obecnych w komórkach. cDNA można następnie amplifikować do wysokich stężeń za pomocą PCR. Biolodzy stworzyli biblioteki cDNA zawierające duże zestawy sekwencji różnych genów, o których wiadomo, że ulegają ekspresji w określonych typach komórek. Biblioteki te można stosować jako sekwencje sond na mikromacierzach. Każdy cDNA jest dość długi (500–2 000 zasad) i zawiera znaczną część sekwencji genu, ale niekoniecznie cały gen. Hybrydyzacja do tych długich sekwencji jest znacznie bardziej specyficzna niż w przypadku oligonukleotydów - w związku z tym zwykle tylko jedno miejsce na macierzy cDNA odpowiada

jednemu genowi i jest to wystarczające do rozróżnienia genów. Macierz cDNA jest zwykle wykonana ze szkła. Jest wstępnie pokryta powierzchniową substancją chemiczną, która wiąże DNA. Roboty siatkowe z zestawami szpilek są następnie używane do przenoszenia niewielkich ilości cDNA na szkiełko. Ta technologia jest tańsza niż ta do przygotowania macierzy oligonukleotydów i dlatego jest bardziej dostępna dla wielu grup badawczych. Przygotowanie macierzy cDNA wymaga jednak dużej ilości wcześniejszej pracy laboratoryjnej w celu stworzenia cDNA. Podczas gdy w przypadku macierzy oligonukleotydowej znajomość sekwencji genów jest konieczna, najlepiej dla całego genomu, z macierzą cDNA, nie musimy znać całej sekwencji genomu, o ile już eksperymentalnie zidentyfikowaliśmy zestaw odpowiednich cDNA. Należy zauważyć, że techniki syntezy in situ stosowane do macierzy oligonukleotydów nie działają w przypadku sekwencji tak długich, jak cDNA, chociaż czasami macierze nakrapiane można przygotować z krótkimi sondami oligonukleotydowymi zamiast cDNA. Proces wytwarzania macierzy jest mniej powtarzalny w przypadku macierzy nakrapianych, a ilość DNA w każdej plamce nie jest tak łatwa do kontrolowania. W związku z tym zwykle nie jest możliwe porównanie bezwzględnej intensywności plamek z różnych slajdów. Problem ten można obejść stosując pomysłowy dwukolorowy fluorescencyjny system znakowania, który umożliwia porównanie dwóch próbek na jednym szkiełku. Zwykle mamy próbkę odniesienia, z którą chcemy porównać próbkę testową. Chcemy wiedzieć, które geny zwiększyły lub zmniejszyły poziom ekspresji w badanej próbce w stosunku do odniesienia. Ekstrakty RNA z tych dwóch próbek przygotowuje się oddzielnie. Następnie cDNA jest wytwarzany z próbki odniesienia przy użyciu nukleotydów znakowanych zielonym fluoroforem, a cDNA jest wytwarzany z próbki testowej przy użyciu nukleotydów znakowanych czerwonym fluoroforem. Te dwie oznakowane grupy sekwencji miesza się następnie i mieszaninę pozostawia do hybrydyzacji z macierzą. cDNA znakowane na czerwono i zielono z próbek powinny wiązać się z plamką proporcjonalnie do ich stężenia. Mierzono intensywność zarówno czerwonej, jak i zielonej fluorescencji z każdego punktu. Stosunek intensywności koloru czerwonego do zielonego nie powinien zależeć od rozmiaru plamki, więc eliminuje to ważne źródło błędu. W przypadkach, gdy istnieje wiele różnych warunków eksperymentalnych do porównania, każdy z nich można porównać z tą samą próbką odniesienia [3].

Typowym przykładem aplikacji mikromacierzy jest porównanie ekspresji zestawu genów z komórek utrzymywanych w określonych warunkach z komórkami stanowiącymi punkt odniesienia (Rys. 2.1) [4].



Rys. 2.1 - Zobrazowanie eksperymentu mikromacierzowego [4].

Na eksperyment (Rys. 2.1) składają się następujące etapy:

- ekstrahowanie RNA z komórek tkanki, w której ekspresja genów nas interesuje oraz tkanki odniesienia,
- odwrotna transkrypcja RNA do cDNA - przy pomocy odwrotnej transkryptazy oraz nukleotydów znakowanych różnymi barwnikami fluorescencyjnymi, inny barwnik dla tkanki wzorcowej, inny dla testowej,
- hybrydyzacja na plamkach zawierających sekwencje komplementarne, spowoduje to wiązanie się odcinków cDNA do plamek zawierających komplementarne nici,
- wzbudzenie laserem i skanowanie - ilość emitowanej fluorescencji odpowiada ilości związanego kwasu nukleinowego.

Wynik eksperymentu stanowi zestawienie względnego poziomu ekspresji genów, zapisany na obrazie w formacie TIFF [4].

3. Przetwarzanie wstępne danych mikromacierzowych

Na przebieg przetwarzania wstępnego danych mikromacierzowych składa się przetwarzanie obrazu, w celu wyodrębnienia poszczególnych intensywności sond, obliczenie współczynnika ekspresji i jego transformacja, aby jego wartość była przejrzysta oraz normalizacja, w celu podsumowania sond w ramach każdego genu i zniwelowania systematycznych błędów eksperymentu.

3.1. Proces przetwarzania obrazu

Uzyskany obraz nie stanowi danych spotykanych w typowych genomicznych bazach danych. Podlega on wstępnemu przetwarzaniu, w wyniku którego uzyskujemy pliki o specyfikacji określonej przez konkretnego producenta. Przetwarzanie to opiera się o następujące czynności:

- identyfikacja plamek i odróżnianie ich od fałszywych sygnałów,
- niektóre oprogramowanie wymaga od użytkownika podania lokalizacji grup plamek (podmacierzy) oraz innych istotnych parametrów,
- określenie obszaru plamki oraz regionu lokalnego w celu oszacowania hybrydyzacji,
- określenie sygnału dla punktów danej podmacierzy, oszacowanie natężenia tła - istnieją dwie metody określania sygnału punktowego:
  - zastosowanie obszaru o ustalonym rozmiarze, scentrowanym na środku masy plamki:
    - metoda o stosunkowo niedużej złożoności obliczeniowej,
    - metoda podatna na błędy w szacowaniu intensywności tła i plamki,
  - precyzyjne określenie obszaru plamki, uwzględnienie tylko pikseli sklasyfikowanych jako "plamka":
    - możliwe jest lepsze oszacowanie intensywności plamki,

- metoda bardziej złożona obliczeniowo.

❖ raportowanie statystyk podsumowujących i przypisywanie intensywności punktowej po odjęciu intensywności tła:

Raportowane są statystyki dla dwóch kanałów: zielonego i czerwonego (mianowicie: wartości średnie oraz mediany albo wartości całkowite natężenia), wyszczególniając:

- mediana plamki pomniejszona o medianę tła:
  - opcja niewrażliwa na kilka pikseli z anormalnymi wartościami w 1 lub 2 kanałach barw,
  - opcja wrażliwa na błędną identyfikację plamki i tła.
- zastosowanie wartości całkowitych intensywności:
  - niewrażliwe na błędną identyfikację tła - kilka dodatkowych pikseli z zerową wartością w tle nie wpłynie na całkowitą intensywność,
  - skłonność do przesunięcia przez kilka pikseli z ekstremalnymi wartościami intensywności

Podczas przetwarzania obrazu brana jest pod uwagę liczba pikseli, zazwyczaj piksel ma 10 µm lub 5 µm [4].

3.2. Współczynniki ekspresji, podstawowe porównanie

Współczynnikami ekspresji może być wartość całkowita albo mediana z odjętym tłem. Wartość całkowitą opisuje wzór:

$$T_k = \frac{R_k}{G_k}$$

gdzie:

- k - *określony gen,*
- R_k - *intensywność plamki dla tkanki testowej*
- G_k - intensywność plamki dla tkanki referencyjnej.

Natomiast dla mediany wzór dla poszczególnych genów wygląda następująco:

$$T_{mediana} = \frac{R_{mediana}^{plamka} - R_{mediana}^{tło}}{G_{mediana}^{plamka} - G_{mediana}^{tło}}$$

gdzie:

- $R_{mediana}$ *- mediana intensywności odpowiednio dla plamki i tła, oceniając tkankę testową,*
- $G_{mediana}$ *- mediana intensywności odpowiednio dla plamki i tła oceniając tkankę referencyjną [4].*

3.3. Transformacja współczynnika ekspresji

Otrzymany współczynnik ekspresji jest intuicyjny - geny o wartości T = 1 są takie same dla obu próbek, jednak nie zawsze taka prezentacja danych jest wygodna. Podstawowym problemem jest regulacja w górę i w dół, która nie jest równomierna. W górę wartość jest większa od 1 i może przyjąć teoretycznie dowolną liczbę powyżej, a w dół przedział wartości ujęty jest pomiędzy zerem a jedynką. Regulację w dół określa operacja dzielenia, a w górę mnożenia [4].

$$\frac{wartość\ testowa}{wartość\ referencyjna} \begin{cases} = 1 \\ > 1 \\ < 1 \end{cases}$$

Istnieją dwa podejścia w celu uniknięcia tego problemu:

- Transformacja odwrotna: polega na odwróceniu oraz pomnożeniu przez -1 współczynnika ekspresji mniejszego od 1; jeśli zaś współczynnik jest większy lub równy 1, nie podlega on żadnej operacji. Podejście to skutkuje dość równomiernym i bezpośrednim przedstawieniem współczynnika ekspresji, którego wadą jest nieciągłość pomiędzy -1 a 1 w zbiorze wartości współczynników wynikowych.
- Transformacja logarytmiczna: polega na zastosowaniu logarytmu o podstawie 2 na wartościach współczynnika ekspresji. Uzyskana prezentacja nie jest tak bezpośrednia jak w przypadku pierwszej metody, jednak uzyskane wartości prezentują zbiór ciągły.

Współczynniki ekspresji są dobrym narzędziem do obserwacji wzorców właściwych dla danych, jednak usuwają informacje o bezwzględnych poziomach ekspresji genów. Na przykład geny mające stosunek wartości

testowej i referencyjnej 400/100 oraz 4/1 posiadają ten sam współczynnik ekspresji równy 4, co może utrudnić interpretację wyników [4].

3.4. Normalizacja danych

Geny porządkowe (ang. Housekeeping genes), zwane też genami referencyjnymi powinny mieć współczynnik ekspresji T = 1. Geny te odpowiedzialne są za utrzymywanie podstawowych funkcji komórkowych, które są niezbędne do istnienia komórki. Cechują się względnie stałą ekspresją niezależnie od pewnych czynników, jednak obserwuje się zmiany z powodu różnej skuteczności znakowania barwnikami albo różną ilość znakowanego mRNA. Normalizacja ma na celu wyeliminowanie systematycznych zmian wpływających na pomiary poziomów ekspresji [4, 5].

Ogólnie proces normalizacji zakłada ustalenie zestawu genów, których ekspresja nie powinna się zmieniać w badanych warunkach (geny porządkowe). Z tego zestawu obliczany jest współczynnik normalizacji, który jest liczbą uwzględniającą zmienność obserwowaną w zestawie genów. Następnie stosowany jest on do innych genów w eksperymencie [4].

Należy zaznaczyć, iż operacja normalizacji przeprowadzana jest tylko dla wartości skorygowanych o intensywność tła [4].

4. Kontrola jakości danych mikromacierzowych

Kontrola jakości danych uzyskanych z eksperymentu mikromacierzowego jest istotnym etapem kwalifikacji poszczególnych prób (macierzy) do dalszej analizy.

Poprzez ocenę jakości należy rozumieć między innymi obliczenie i interpretację metryk. W przypadku mikromacierzy firmy Affymetrix metryki sugerowane przez producenta obejmują:

- średnie tło,
- współczynnik skali,
- liczba genów oznaczonych jako obecne,
- stosunek 3′ do 5′ dla Beta-aktyn i GAPDH,
- Spine-in RNA (kontrolne transkrypty o znanej sekwencji i ilości stosowane do kalibracji pomiarów w testach hybrydyzacji RNA) [6]

Średnie tło powinno być podobne w każdej macierzy eksperymentu, niestety istnieje wiele przyczyn możliwych znaczących różnic w tym parametrze. Sygnał odczytany z macierzy mógł być większy przez różne ilości cRNA obecnych w mieszaninach hybrydyzacyjnych albo proces hybrydyzacji był bardziej wydajny w jednej z reakcji, obejmując więcej znacznika i wytwarzając silniejszy sygnał [6].

Współczynnik skali jest zależny od algorytmu normalizacji. Przykładowo w algorytmie MAS 5.0 zakłada się, że ekspresja genów nie zmienia się znacząco w zdecydowanej większości transkryptów, w danym eksperymencie. W konsekwencji tego założenia intensywność dla każdej próbki jest skalowana tak, aby jej wartości średnie były równe. Stopień przeskalowania jest prezentowany przez współczynnik skali, który stanowi miarę ogólnego poziomu ekspresji macierzy i odzwierciedla stopień hybrydyzacji znakowanego RNA z próby. Duże różnice w skalach sygnalizują przypadki, w których założenia normalizacji mogą okazać się fałszywe, ze względu na problemy z jakością próby lub ilością materiału wyjściowego. Ewentualnie przyczyną kłopotliwych współczynników skali mogą być problemy z ekstrakcją RNA, znakowaniem, skanowaniem lub produkcją

matrycy. W przypadku firmy Affymetrix, zalecane jest, aby współczynniki skali nie były większe niż 3 [6].

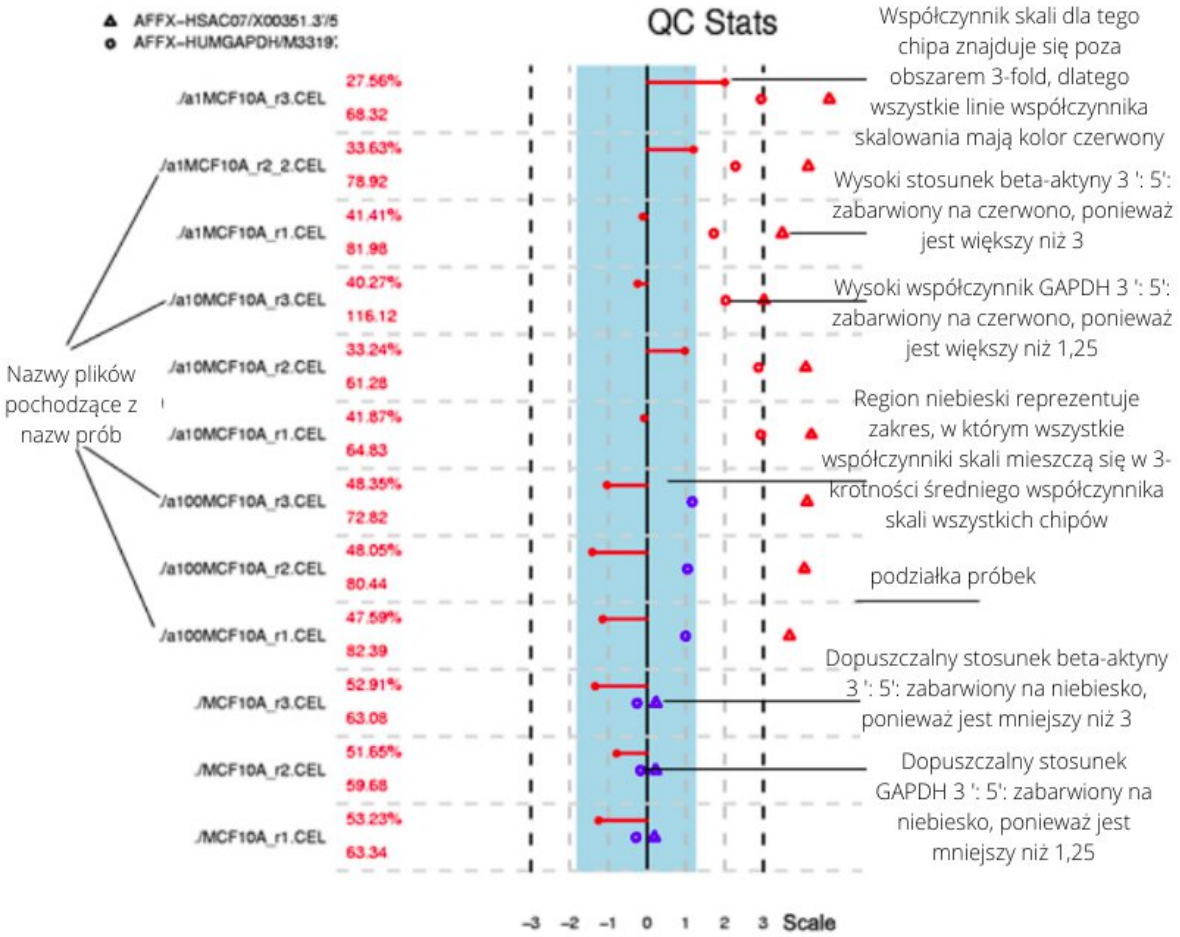
Liczba genów oznaczonych jako obecne pochodzi z podsumowania par sond MM/PM (ang. Mismatch/Perfect Match) w danych zestawach sond. Pary sond PM jest zaprojektowana tak, aby dokładnie dopasować sekwencję będącą przedmiotem zainteresowania, natomiast pary sond MM zaprojektowano tak, aby zawierały pojedynczą, niedopasowaną zasadę, do której wspomniana sekwencja nie może się dołączyć. Podsumowanie to powoduje oznaczenie genu jako obecnego, marginalnego albo nieobecnego (ang. Present/Marginal/Absent, P/M/A). Jeśli wartość sondy PM w danej parze nie będzie znacząco wyższa od sondy MM, to para zostaje podsumowana jako marginalna albo nieobecna. Podobnie jak w przypadku współczynników skali, duże różnice między liczbą genów oznaczonych jako obecne w różnych próbach mogą świadczyć o różnych ilościach zhybrydyzowanego RNA w poszczególnych matrycach lub innych problemach charakterystycznych dla tego eksperymentu. Stosuje się oznaczenie procentowe, które reprezentuje procent sond podsumowanych jako obecne w danej mikromacierzy [6].

W większości typów komórek ekspresji ulegają geny kodujące Beta-aktynę oraz GAPDH, będące genami referencyjnymi. Są to stosunkowo długie geny, w większości matryc firmy Affymetrix zawarte są zestawy sond skierowanych na regiony 5′, środkowy oraz 3′ transkryptów. Porównując intensywność sygnału z zestawu sond 3′ z środkowymi albo 5′, można uzyskać miarę jakości RNA zhybrydyzowanego z mikromacierzą. Jeśli uzyskany stosunek jest wysoki, oznacza to obecność skróconych transkryptów. Może się to zdarzyć, jeśli etap transkrypcji in vitro nie wypadł dobrze albo jeśli nastąpi ogólna degradacja RNA (dlatego stosunek sygnału 3′ do 5′ można uznać za miarę jakości RNA). W przypadku, gdy RNA zostało przygotowane wg protokołu Affymetrix Small Sample zamiast standardowego, zaleca się stosowanie stosunku 3′ do odcinka środkowego, ze względu na dodatkowy etap amplifikacji, który prawdopodobnie zwiększy częstotliwość krótkich transkryptów w roztworze i nieuchronnie wprowadzi błąd do sygnału 3′ znakowanych transkryptów. GAPDH jest mniejszym z dwóch genów, a stosunek 3′:5′ wynosi zazwyczaj około 1. Na podstawie doświadczeń [6] próg został ustanowiony na 1.25, natomiast w przypadku Beta-aktyny sugerowanym progiem przez firmę Affymetrix jest 3 [6].

Zestawy sond Spike-in zawierają sondy o znanej sekwencji i ilości. W celu zweryfikowania wydajności etapu hybrydyzacji, w późniejszych etapach przygotowania próby dodaje się znakowane cRNA. Owe transkrypty pochodzą o oznaczeniach BioB, BioC, BioD i CreX pochodzą z bakterii Bacillus subtillis, nic innego w mieszaninie nie powinno się wiązać z ich

zestawami sond. Są one dodawane do roztworu tuż przed naniesieniem go na matrycę [6].

Opisane metryki są przejrzyście prezentowane na jednej grafice, generowanej przez funkcję pakietu Bioconductor R - simpleaffy. Na Rys. 4.1 znajduje się przykład takiej grafiki.

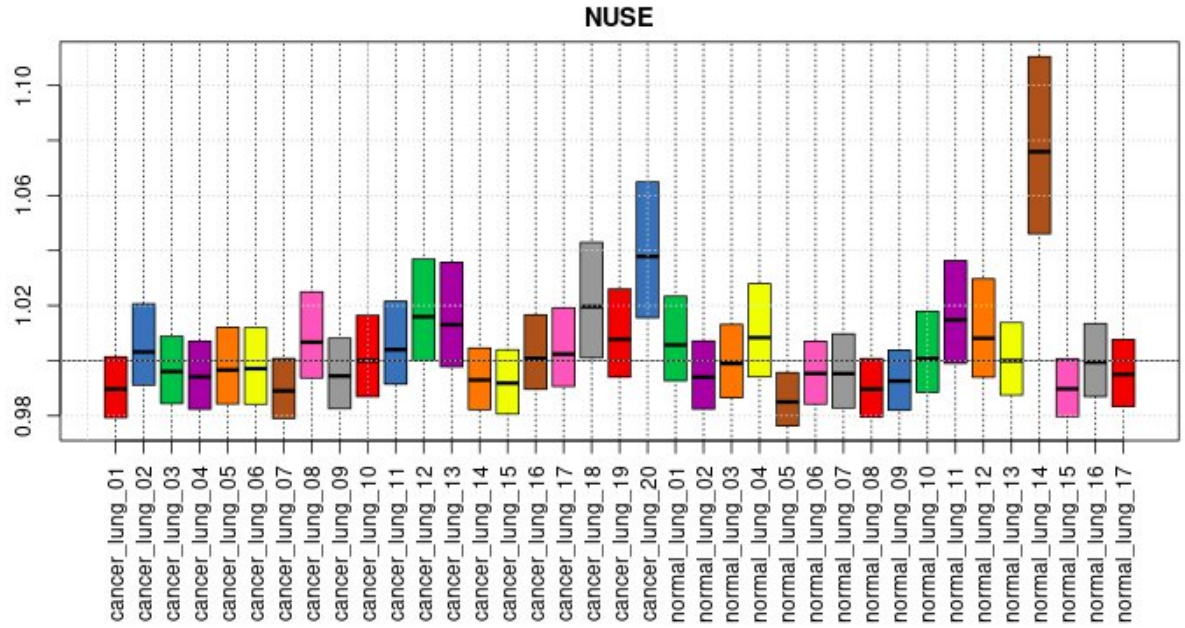


Rys. 4.1 - Grafika QC Stats dla stosunku 3':5' [6].

4.1. Wykresy NUSE i RLE

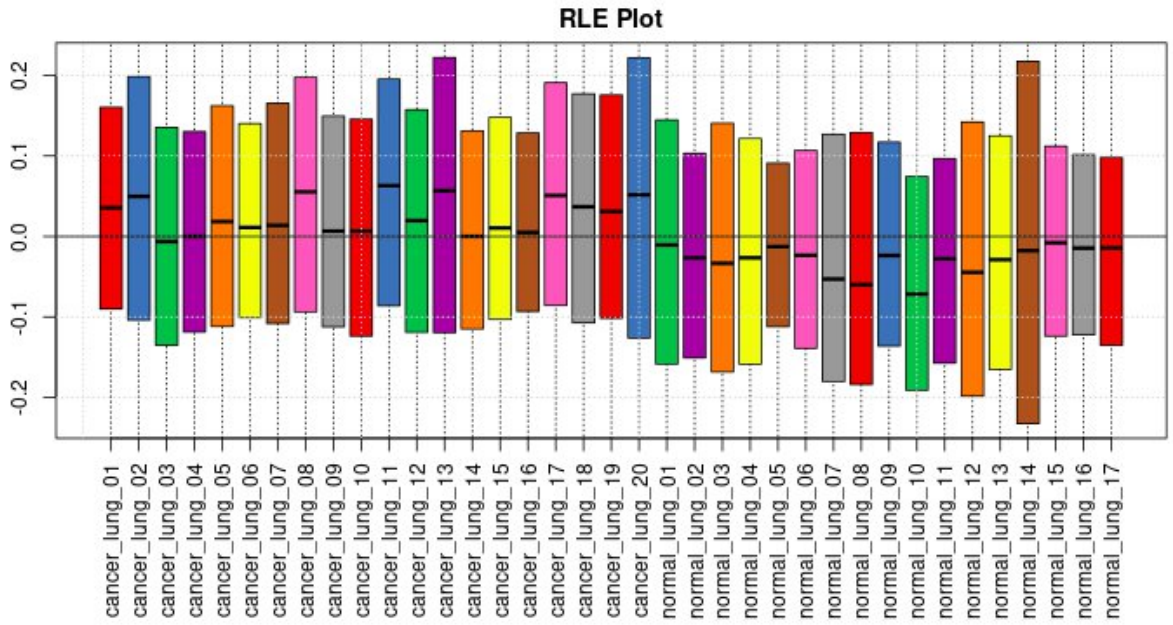
Przy ocenie jakości danych mikromacierzowych można również użyć znormalizowanych nieskalowanych błędów standardowych NUSE (ang. Normalized unscaled standard errors). Błędy są wizualizowane na wykresie pudełkowym (Rys. 4.2) tak, aby mediana błędu standardowego dla tych genów wynosiła 1 we wszystkich macierzach danego eksperymentu. Obserwuje się w ten sposób różnice w zmienności między genami. Macierz, w której występuje podwyższona wartość błędu standardowego w stosunku

do innych macierzy, jest zazwyczaj niższej jakości [7]. Należy zwracać uwagę na macierze z bardzo różnymi NUSE od pozostałych w eksperymencie oraz na macierze z szerszym rozkładem [8].



Rys. 4.2 - Przykładowy wykres pudełkowy NUSE [źródło własne].

Innym narzędziem oceny jakości mikromacierzy są wartości RLE (ang. Relative Log Expression), które obliczane są dla każdej matrycy poprzez porównanie wartości ekspresji w każdej z prób z medianą ekspresji dla wszystkich prób. Zakładając, że ekspresja większości genów nie zmienia się w obrębie matryc, oznacza to, że w idealnych warunkach, większość wartości RLE będzie bliska zera. Wartości RLE również przedstawia się przy pomocy wykresu pudełkowego (Rys 4.3). Należy skupić się na kształcie i położeniu pudełek , zwykle tablice o gorszej jakości nie są wyśrodkowane w pobliżu zera i są bardziej rozłożone względem pudełek odpowiadającym innym macierzom [7].



Rys. 4.3 - Przykładowy wykres pudełkowy RLE [źródło własne].

5. Analiza danych mikromacierzowych

Analiza danych oparta na mikromacierzach DNA umożliwia diagnozę jeszcze przed pojawieniem się wyraźnych symptomów choroby, identyfikację nowych jednostek chorobowych lub nowych wariantów znanych już schorzeń. Dane pozyskane z tego eksperymentu umożliwiają również ocenę predyspozycji pacjentów do zapadania na różnego typu choroby, precyzyjną ocenę ich stopnia zaawansowania, prognozowanie dalszego rozwoju schorzeń oraz dobór optymalnej terapii, w tym monitorowanie odpowiedzi na terapię i badanie metabolizmu leków, czy identyfikację potencjalnych celów terapeutycznych, jak geny, białka i nie tylko [6].

5.1. Macierze ekspresji genów

Dane po normalizacji można przedstawić w postaci macierzy ekspresji genów. Każdy wiersz takiej macierzy odpowiada określonemu genowi, a każda kolumna może odpowiadać warunkom eksperymentalnym lub danemu punktowi w czasie, w którym zmierzono ekspresję genów. Macierze ekspresji można przedstawiać z wartościami bezwzględnymi, porównawczymi oraz często po transformacji logarytmicznej [4].

Poziomy ekspresji genu w różnych warunkach eksperymentalnych nazywa się profilem ekspresji genów, natomiast poziomy ekspresji wszystkich genów w próbce dla danych warunków eksperymentalnych nazywa się profilem ekspresji próby. Tak złożone dane poddaje się adnotacji poprzez podawanie funkcji genów lub określanie stanów zdrowia/choroby określonej próby. W zależności od tego, czy zastosowano adnotację, czy też nie, analizę danych ekspresji genów można podzielić na uczenie z nadzorem i bez nadzoru [4].

Istotnym elementem przed dalszym przetwarzaniem danych mikromacierzowych jest sprawdzenie skuteczności procesu normalizacji

zestawu macierzy . Najłatwiej osiągnąć to wykonując wykresy pudełkowe przed i po procesie normalizacji, gdzie każde pudełko odpowiada jednej macierzy. Po normalizacji pudełka powinny być możliwie wyrównane, posiadając porównywalny średni poziom ekspresji i bardzo zbliżoną skalę wartości.

Dobrym sposobem wstępnego porównania konkretnych macierzy lub grup macierzy, np.: macierzy testujących i kontrolnych jest wykonanie wykresu MA. Wykres MA pokazuje, w jakim stopniu zmienność ekspresji zależy od poziomu ekspresji. Jest to wykres rozrzutu, na którym wykreśla się A względem M:

- M jest różnicą między intensywnością sondy/grupy sond na macierzy a medianą intensywności tej sondy w całym zbiorze matryc,
- A jest średnią intensywnością danej sondy w całym zbiorze.

W idealnym przypadku chmura punktów danych na wykresie MA powinna być wyśrodkowana w okół M = 0. Dzieje się tak, ponieważ zakładamy, że większość genów nie stanowi rozróżnienia pomiędzy badanymi warunkami eksperymentalnymi oraz liczba genów regulowanych w górę i w dół jest podobna. Przy pomocy wykresów MA można oceniać poszczególne matryce względem pozostałych, jak i grupy matryc odpowiadającym porównywanym w eksperymencie warunkom [9].

5.2. Analiza nadzorowana i nienadzorowana

Analiza bez nadzoru przeprowadzana jest w taki sposób aby zidentyfikować możliwe wzorce, mogące pogrupować geny/próbki w klastry bez użycia adnotacji. Np. geny o podobnych profilach ekspresji można pogrupować razem, można również uwzględnić adnotację w późniejszych etapach analizy danych w celu wyciągania wniosków biologicznych [4].

W przypadku analizy z nadzorem, na podstawie informacji o próbie, tworzymy klastry genów/próbek w celu określenia charakterystycznych wzorców dla danego klastra. Na przykład po podzieleniu profili ekspresji na wskazujące na stan chorobowy i nie, można uzyskać wzorce stojące na pograniczu zdefiniowanych stanów [4].

5.3. Klasteryzacja

Klasteryzacja, inaczej grupowanie, dotyczy analizy skupień i jest przeprowadzana w celu wyłonienia przypuszczalnie sensownych organizacji danych [10].

Metody grupowania można podzielić na hierarchiczne i niehierarchiczne. W przypadku tworzenia klastrów w sposób hierarchiczny określa się relacje pomiędzy obiektami w danej grupie, co przypomina drzewo filogenetyczne. Metody niehierarchiczne nie biorą pod uwagę relacji pomiędzy obiektami w klastrze [10].

Grupowanie hierarchiczne dzieli się na aglomeracyjne oraz deglomeracyjne. Pierwsze grupuje pojedyncze obiekty w określone klastry oraz stopniowo dodaje nowe grupy, aż pozostanie tylko jedna, natomiast deglomeracyjne grupuje obiekty zaczynając od uznania wszystkich obiektów jako jednej grupy i stopniowo wprowadza podziały [4].

Łączenie w grupy odbywa się na podstawie określonych metryk. W przypadku grupowania pojedynczego połączenia brana jest pod uwagę minimalna odległość pomiędzy obiektami lub klastrami. Ta metoda jest uważana za niewrażliwą na wartości odstające i nazywana jest również jako metoda najbliższego sąsiedztwa. Kolejną metodą hierarchicznego grupowania aglomeracyjnego jest grupowanie kompletnych połączeń, biorące pod uwagę wartości maksymalne odległości pomiędzy obiektami. Stosowane są również metody oparte o wartość średnią połączenia pomiędzy klastrami oraz odległość pomiędzy centroidami porównywanych grup [4].

Podział na grupy metodą deglomeracyjną odbywa się poprzez np. analizę głównego składnika, za pomocą której określa się wektor oddzielający obiekty pomiędzy nowe klastry [4].

Klastrowanie niehierarchiczne stanowi uzupełnienie wad grupowania hierarchicznego, dotyczących spotykanej niezasadności utworzonych struktur, które nie zawsze znajdują stosowne uzasadnienie dla podziału danej puli profili ekspresji. Podejście niehierarchiczne zakłada odgórne narzuconą ilość klastrów i im podporządkowuje dalsze grupowanie. Ta metoda może zostać przeprowadzona przy pomocy różnych wariantów algorytmu, na przykład przy pomocy K-średnich, w której na pierwszym etapie obiekty są dowolnie dzielone na przewidzianą liczbę klastrów. Liczbę klastrów można dobrać losowo albo oszacować przeprowadzając najpierw

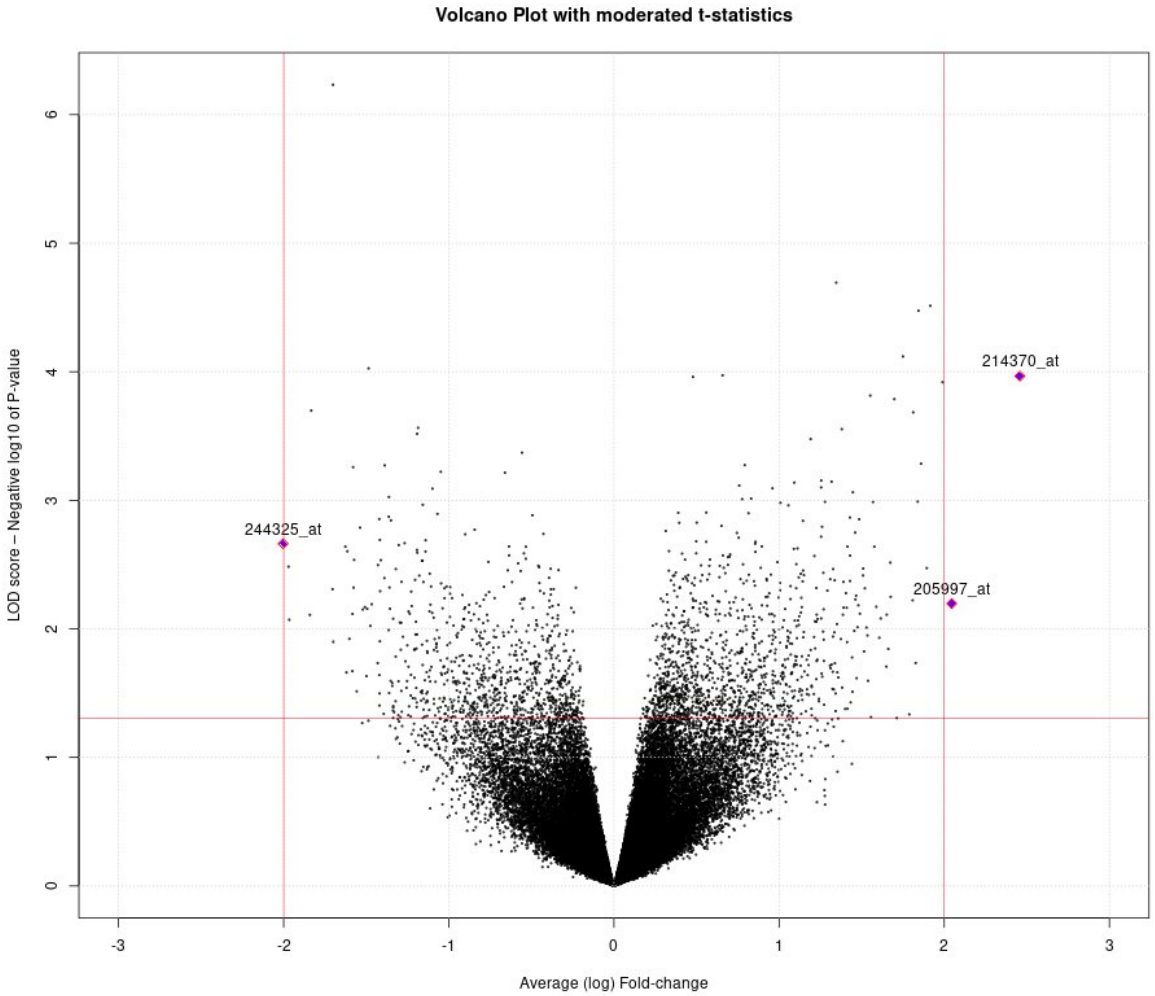
grupowanie hierarchiczne. Następnie dla każdej grupy obliczany jest średni profil ekspresji (np.: centroid), ten krok nazywa się inicjalizacją. W kolejnym kroku poszczególne obiekty są ponownie przypisywane z jednych grup do innych, w zależności od tego, który centroid jest bliżej genu/próbki. Proces ten jest powtarzany przez ustaloną liczbę iteracji albo do momentu konwergencji (brak zmian na przestrzeni kilku iteracji). Inną metodą grupowania niehierarchicznego są samo-organizujące się mapy - SOM (ang. Self Organizing Maps). W metodzie SOM pierwszym krokiem jest wybór liczby i orientacji klastrów względem siebie, np.: punktem wyjściowym może być siatka z punktami oznaczającymi określone klastry. Siatka ta jest rzutowana na przestrzeń danych ekspresji, gdzie każdy z obiektów przynależy do najbliższego sobie punktu siatki. W następnych krokach punkt siatki jest przesuwany bliżej losowo wybranego obiektu, pociągając za sobą pozostałe punkty w stopniu zależnym od ich odległości od przesuwanego punktu. W kolejnych krokach proces się powtarza, a początkowo dwuwymiarowa siatka dopasowuje się do rozkładu danych ekspresji. Metoda SOM w przeciwieństwie do K-średnich nie wymusza liczby klastrów, niektóre z obiektów mogą pozostać bez określonej grupy. Zaletą metody SOM jest dostarczenie informacji na temat podobieństwa pomiędzy grupami oraz wiarygodność nawet dla zaszumionych danych [4].

5.4. Analiza statystyczna

W celu wyłonienia genów mogących mieć wpływ na obserwowane różnice w warunkach eksperymentalnych (stan zdrowia/choroby), należy wybrać te o dostatecznie dużej ekspresji różnicowej oraz możliwie małym błędzie.

Wygodnym sposobem obserwacji tych wartości jest wykres wulkanu (Rys. 5.1). Przedstawiony jest na nim logarytmiczny wskaźnik zmian danego genu oraz odpowiadająca mu statystyka. Aby grafika przybrała pożądany kształt, wartości na osi pionowej wykresu wyświetlane są jako $-\log_{10}(\text{p-val})$, a na osi poziomej znajduje się średni współczynnik ekspresji różnicowej. Na podstawie wyglądu wykresu można wizualnie ocenić jak dobrać progi w celu filtracji genów do dalszej analizy.

W ten sposób wybrane geny można poddać grupowaniu i wizualizacji na wykresie mapy cieplnej, w celu obserwacji ewentualnych zależności, by w końcu sprawdzić ich biologiczne znaczenie dla badanych warunków eksperymentalnych.



Rys. 5.1 - Przykładowy wykres wulkanu [źródło własne].

6. Realizacja narzędzia do analizy danych mikromacierzowych - MicAff

W ramach pracy wykonano aplikację webową za pomocą frameworka do tworzenia stron internetowych Shiny, w języku programowania R. Aplikacja pomimo tego, że została stworzona przy pomocy technologii webowych, służy jako aplikacja desktopowa na lokalnym hoście.

Pliki zawierające dane mikromacierzowe zajmują często stosunkowo dużo pamięci i ich przesyłanie na serwer aplikacji w chmurze byłoby czasochłonne.

Aplikacja została nazwana MicAff i umożliwia załadowanie plików w formacie .CEL, ich normalizację, kontrolę jakości, kontrolę wyników normalizacji, przegląd wyników, wykonanie statystyk, progowanie, klasteryzację i wykonanie map cieplnych dla danych zawartych w załadowanych plikach. Poszczególne wyniki, jak znormalizowane współczynniki ekspresji i statystyki można pobrać w formacie .TXT. Uruchomienie aplikacji lokalnie umożliwia również dalszą pracę na obiektach środowiska R pozostawionych w pamięci sesji po zakończeniu działania, dzięki czemu nie trzeba polegać tylko na wartościach zwracanych w ramach interfejsu graficznego aplikacji przy szukaniu funkcji genów na podstawie ich adnotacji.

Język R posiada istotne biblioteki i narzędzia do zaawansowanej analizy danych i operacji na plikach, dzięki czemu wykonanie prototypu i wstępnej aplikacji zostało ułatwione o proces implementacji złożonych algorytmów i testowania istniejących implementacji w niestandardowych środowiskach. Jednak takie rozwiązanie nie zapewnia elastycznej logiki aplikacji do ewentualnego komercyjnego wdrożenia.

6.1. Specyfikacja wewnętrzna oprogramowania

MicAff stanowi zestaw narzędzi do przetwarzania i analizy danych zawartych w plikach .CEL osadzonych w ramach aplikacji webowej, możliwej do swobodnego użytkowania na lokalnym serwerze (dowolny komputer z wystarczającą pamięcią RAM).

Aplikacja została napisana i przetestowana w środowisku RStudio Server [wersja 1.3.959], przy pomocy bibliotek naukowych i nie tylko, kompatybilnych z następującą wersją języka R:

```
> sessionInfo()
R version 4.0.3 (2020-10-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.1 LTS
```

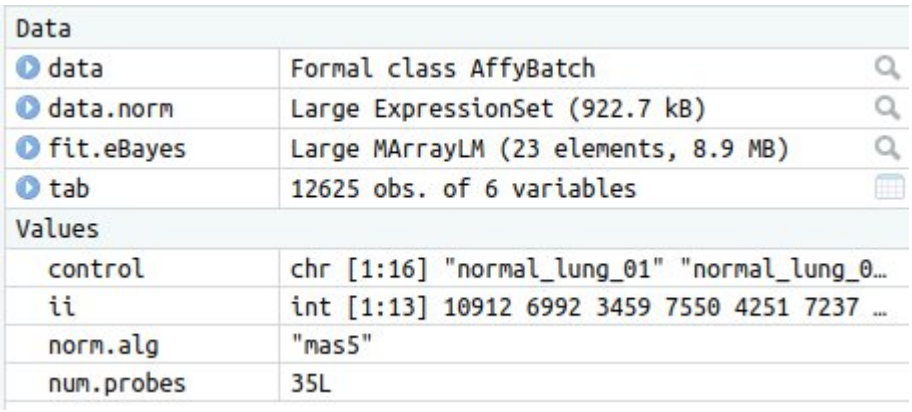
W celu poprawnego działania narzędzia, wymagane są następujące biblioteki:

```
library(BiocManager)
library(affy)
library(genefilter)
library(simpleaffy)
library(affyPLM)
library(limma)

library(shiny)
library(latticeExtra)
library(RColorBrewer)

library(pheatmap)
```

Aplikacja Shiny R funkcjonuje w obrębie jednej sesji na klienta i bazuje na 2 plikach źródłowych ui.R oraz server.R (możliwe jest umieszczenie całości programu w jednym pliku). Po zakończeniu działania aplikacji na lokalnym serwerze, można przystąpić do bardziej szczegółowej pracy na obiektach pozostawionych przez aplikację (Rys. 6.1). Do tych obiektów zaliczają się:

- data - jest to obiekt zawierający w sobie nieprzetworzone dane, załadowane z plików .CEL,
- data.norm - jest obiektem danych po procesie normalizacji,
- fit.eBayes - to obiekt, w którym zawarte są moderowane statystyki i inne obliczane wraz z nimi parametry,

- tab - zawiera tablicę przyporządkowującą każdemu genowi średnią wartość ekspresji, ekspresję różnicową, statystykę t wraz z wartością p oraz skorygowaną wartością p dla algorytmu FDR,
- control - jest tablicą zawierającą etykiety odpowiadające próbie kontrolnej eksperymentu,
- ii - jest tablicą indeksów dla N-najlepszych genów (o największej ekspresji różnicowej i najmniejszej wartości p przed lub po korekcji),
- norm.alg - zawiera w sobie nazwę algorytmu normalizacji,
- num.probes - zawiera w sobie liczbę prób eksperymentu.



Rys. 6.1 - Przegląd obiektów i zmiennych pozostawionych po wyłączeniu aplikacji MicAff [źródło własne].

Interfejs graficzny aplikacji zawarty jest w pliku ui.R i składa się z jednego elementu fluidPage, który podporządkowuje sobie elementy umieszczone wewnątrz niego domyślnie w wierszach, które można podzielić na poszczególne kolumny. W ten sposób wykonano podział na panel konfiguracji oraz panel raportu szerzej przedstawione w specyfikacji zewnętrznej.

Logika aplikacji znajduje się w pliku server.R. Okrojony kod programu, który przedstawia porządek pracy narzędzia, zawarty w tym pliku, został przedstawiony na Rys. 6.2.

```
shinyServer(function(input, output, session) {

  observeEvent(input$calculate.stats, {⟺})

  observeEvent(input$read.affymetrix.files, {⟺})

  observeEvent(input$update.statistics.plots, {⟺})

  observeEvent(input$update.statistics.for.thresholds,{⟺})

})
```

Rys. 6.2 - Funkcja główna w pliku server.R, zawierająca podstawowe akcje programu [źródło własne].

W funkcji shinyServer zawarte są 4 reakcje na zmiany wprowadzone przez inferfejs użytkownika:

- Obserwacja wejścia *input$read.affymetrix.files*, odpowiadającego za załadowanie plików .CEL na serwer aplikacji; wewnątrz tej obserwacji, na podstawie informacji zawartych w załadowanych plikach, powstaje główny obiekt danych mikromacierzowych, które poddawane są procesowi normalizacji, a wynik tej operacji zostaje udostępniony do pobrania,
- Obserwacja wejścia *input$calculate.stats*, odpowiadającego za uruchomienie obliczeń i generowanie raportu,

  Ta obserwacja zawiera w sobie większość obliczeń prowadzonych w narzędziu, które zawarte są w funkcji display.report.

- Obserwacja wejść *input$update** - odpowiadających za uaktualnienie poszczególnych elementów raportu zgodnie z zmianami wprowadzonymi w panelu konfiguracyjnym aplikacji.

Wprowadzenie zmian z poziomu interfejsu użytkownika spowoduje uruchomienie ciągu obliczeń zawartych w poszczególnych obserwacjach i zapis poszczególnych obiektów z danymi do pamięci danej sesji.

W celu uporządkowania zadań wewnątrz aplikacji powstało 12 funkcji:

- My_FDR_select <- function(input.checkbox) - funkcja ta przyjmuje odczytaną z interfejsu użytkownika wartość pola checkbox oraz decyduje o tym, czy w poszczególnych elementach raportu stosowana jest skorygowana wartość p,
- My_NUSE_data <- function(...) - zadaniem tej funkcji jest obliczanie danych NUSE i RLE w celu tworzenia wykresu pudełkowego. Pochodzi ona z przybornika affyPLM. Została ona dostosowana do formy aplikacji,
- My_NUSE_Plot <- function(dataPLM) - ta funkcja służy do tworzenia wykresu NUSE, również pochodzi z przybornika affyPLM i została zmodyfikowana tak, aby uczynić wykres czytelniejszym,
- My_RLE_Plot <- function(dataPLM) - ta funkcja służy do tworzenia wykresu RLE. Również pochodzi z przybornika affyPLM i została zmodyfikowana, aby uczynić wykres czytelniejszym,
- My_Box_Plot <- function(data, num.probes, ylab, main) - jest to funkcja służąca do kreślenia wykresu pudełkowego danych ekspresji przed i po normalizacji,

- My_check_and_normalise_data <- function(norm.alg, data) - ta funkcja służy do przeprowadzenia normalizacji danych, za pomocą wybranego algorytmu,
- My_MA_Plot <- function(data.norm, control) - funkcja ta służy do tworzenia wykresu MA w celu oceny danych grupy kontrolnej i testowej,
- My_fit.eBayes <- function(data.norm, control) - jest to funkcja odpowiedzialna za obliczenie moderowanych statystyk t,
- My_volcano_moderated <- function(...) - jest to funkcja odpowiedzialna za kreślenie wykresu wulkanu oraz odnalezienie indeksów N-najlepszych genów,
- My_volcano_moderated_threshold <- function(...) - ta funkcja odpowiada za utworzenie wykresu wulkanu dla genów znajdujących się wewnątrz zadanych progów,
- My_heatmap <- function(...) - jest to funkcja odpowiedzialna za klasteryzację i wykreślania mapy cieplnej dla wybranych genów w eksperymencie,
- display.report <- function(...) - jest to istotna funkcja, wewnątrz której wykorzystuje się wszystkie pozostałe w celu początkowego wygenerowania raportu. Poszczególne elementy raportu można zmienić, jednak jest to przeprowadzone za pośrednictwem obserwacji obiektów UI Shiny.

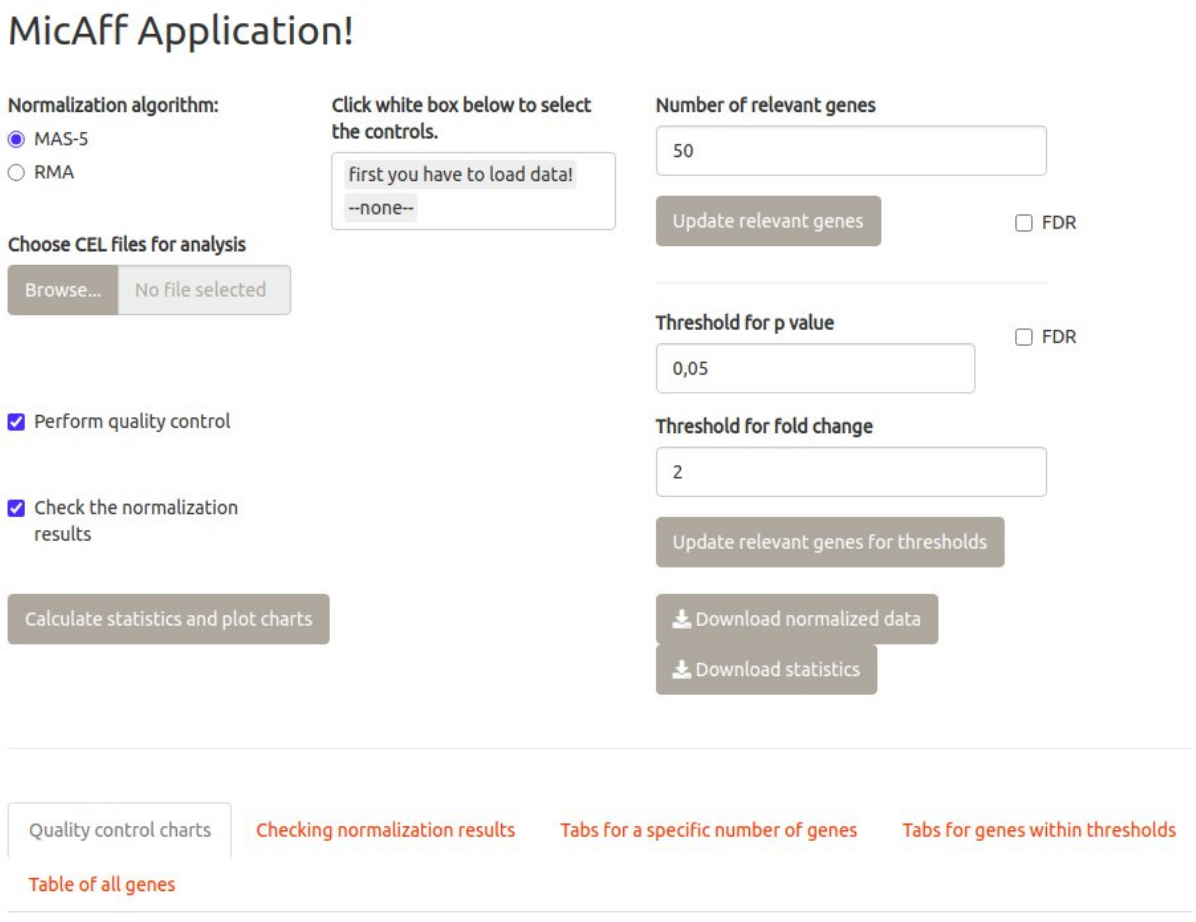
6.2. Specyfikacja zewnętrzna oprogramowania

Interfejs użytkownika aplikacji MicAff składa się z 2 części: panelu konfiguracji oraz panelu raportu, pierwszy znajduje się na samej górze okna aplikacji, a drugi rozmieszczony jest pomiędzy zakładki pod panelem pierwszym (Rys. 6.3).

Panel konfiguracji składa się z następujących elementów:

- Przyciski opcji pod etykietą „Normalization algorithm", wskazujące na algorytm normalizacji, którego zamierzamy użyć,
- Pole do załadowania plików .CEL pod etykietą „Choose CEL files for analysis",
- Pola wyboru z etykietami: „Perform quality control", „Check the normalization results", przy pomocy których można zdecydować, czy aplikacja ma wykonać części raportu odpowiedzialne za kontrolę jakości i sprawdzenie wyników normalizacji,

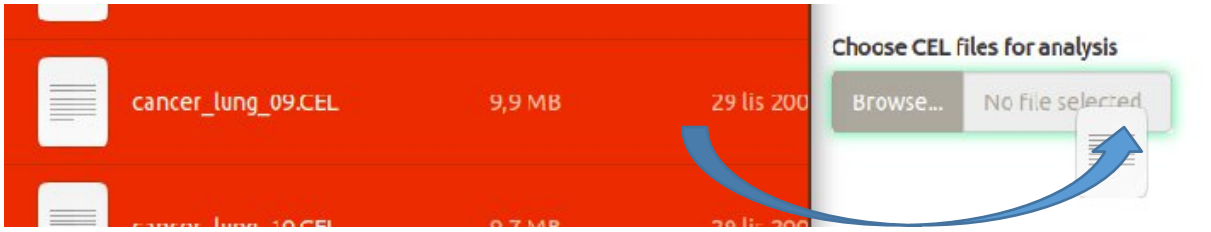
- Przycisk opisany jako „Calculate statistics and plot charts", służący do wygenerowania raportu w zakładkach,
- Lista wyboru pod etykietą „Click white box below to select the controls", która zawiera identyfikatory prób eksperymentu, z których należy wybrać te, odpowiedzialne za próbę kontrolną,
- Wejście numeryczne pod etykietą „Number of relevant genes", które służy do wprowadzania liczby „najlepszych" genów do oznaczenia w raporcie,
- Przycisk opisany jako „Update relevant genes", który służy do uaktualnienia raportu, w przypadku zmiany liczby genów lub zaznaczenia najbliższego pola wyboru „FDR",
- Pole wyboru „FDR", które należy zaznaczyć, jeśli N-najlepszych genów ma być oznaczonych z uwzględnieniem korekcji wartości p,
- Wejście numeryczne opisane etykietą „Threshold for p value", w którym należy umieścić próg dla wartości p,
- Ponownie pole wyboru „FDR", które należy zaznaczyć, jeśli próg dla wartości p, ma odnosić się do skorygowanej wartości p,
- Wejście numeryczne o etykiecie „Threshold for fold change", które zawiera próg dla ekspresji różnicowej danego genu,
- Przycisk opisany jako „Update relevant genes for thresholds", który służy do uaktualnienia raportu, w przypadku zmian wartości progów albo wybrania opcji „FDR", dla drugiego sposobu filtracji genów,
- Przyciski opisane jako „Download normalized data" oraz „Download statistics", które służą do pobierania wyników przetwarzania danych w formacie tekstowym,
- Na samym dole znajdują się zakładki, z czego jedna jest domyślnie otwarta - „Quality control charts". Każda z zakładek zawiera określone elementy raportu. Zakładki „Tabs for a specific number of genes" oraz „Tabs for genes within thresholds" zawierają dodatkowo wewnątrz po dwie kolejne zakładki odpowiadające za wyselekcjonowane geny przy pomocy progów albo poprzez wybranie N-najlepszych.



Rys. 6.3 - Panel konfiguracji aplikacji MicAff i puste zakładki raportu pod nim [źródło własne].

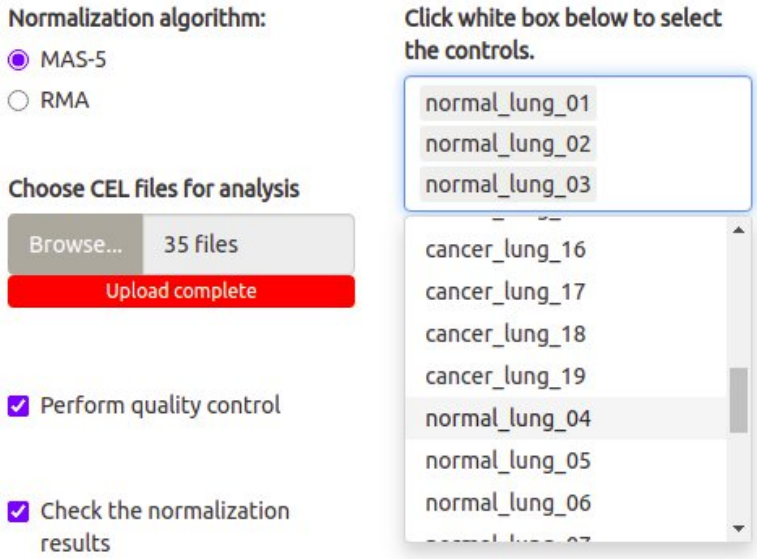
Aplikację poleca się uruchomić w środowisku RStudio, po zainstalowaniu potrzebnych zależności. Następnie należy wykonać następujące czynności w celu analizy danych zawartych w plikach .CEL:

- Wybranie algorytmu normalizacji poprzez zaznaczenie opcji RMA albo MAS-5,
- Wybranie przycisku „Browse...” w polu ładowania plików oraz odnalezienie folderu z plikami .CEL w menadżerze plików, zaznaczenie wszystkich i zatwierdzenie. Można również otworzyć folder poza aplikacją, zaznaczyć wszystkie wymagane pliki i przeciągnąć myszką na pole ładowania plików (Rys. 6.4),



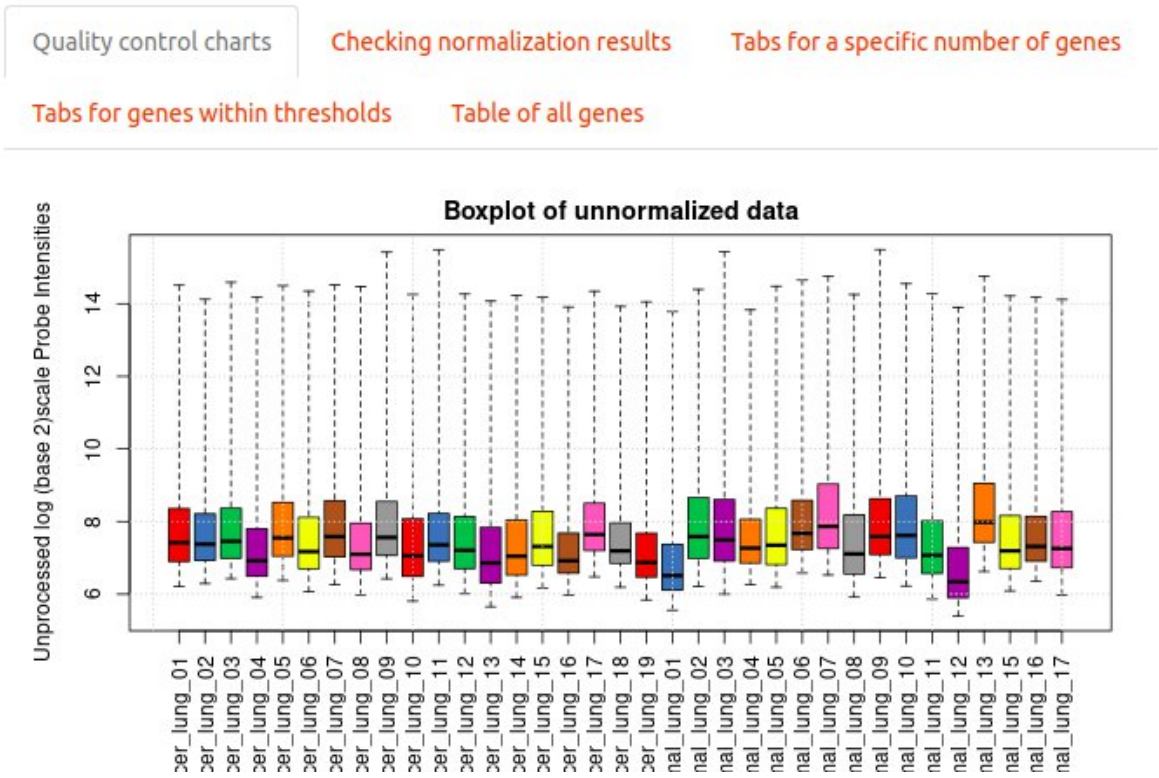
Rys. 6.4 - Przeniesienie zaznaczonych plików do pola ładowania danych [źródło własne].

- Po wykonaniu tej czynności należy odczekać, aż aplikacja załaduje wszystkie pliki i znormalizuje je przy pomocy wybranego wcześniej algorytmu,
- Oznaczenie opcji, czy chcemy aby aplikacja obliczyła metryki i wykonała wykresy w celu kontroli jakości prób oraz w celu sprawdzenia wyników normalizacji,
- Gdy już wszystkie pliki zostaną załadowane i znormalizowane, lista identyfikatorów prób pojawi się w liście rozwijanej, z której należy wybrać te, które odpowiadają próbie kontrolnej, poprzez kliknięcie właściwych z listy. W przypadku błędnego wyboru wystarczy kliknąć przy już wybranym identyfikatorze i usunąć go przy pomocy przycisku Backspace (Rys. 6.5),



Rys. 6.5 - Wybieranie prób kontrolnych [źródło własne].

- Następnie można wybrać przycisk „Calculate statistics and plot charts" w celu wykonania raportu. Może to zająć dłuższą chwilę, postęp można obserwować na paskach postępu pojawiających się w prawym dolnym rogu okna aplikacji,
- Pozostałe pola i opcje można zostawić w domyślnym stanie, a potem dostosowywać w miarę potrzeby,
- Po wykonaniu raportu, jego wyniki pojawią się w poszczególnych zakładkach w postaci wykresów i tabel z możliwością sortowania po wybranych kolumnach (Rys. 6.6).



Rys. 6.6 - Zakładki raportu z jednym wykresem [źródło własne].

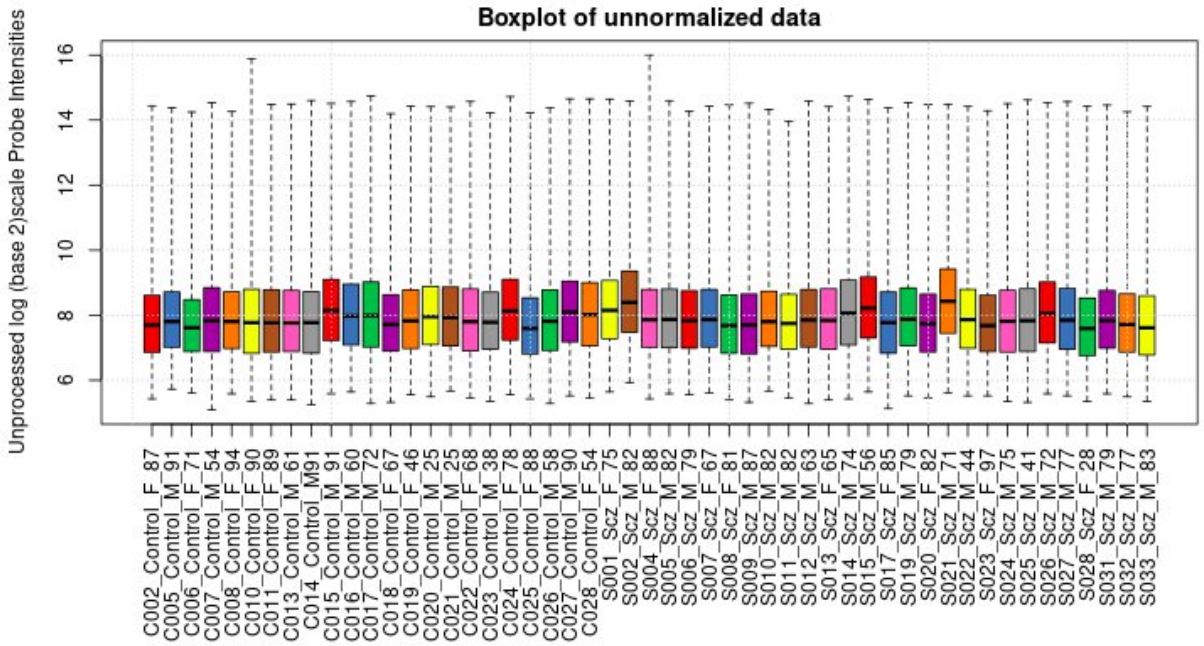
7. Analiza danych przy pomocy MicAff

Do analizy wybrano zestaw danych pozyskanych z eksperymentu mikromacierzowego na tkance pośmiertnej mózgu osób chorych na schizofrenię i osób zdrowych. Ten zestaw danych składa się z 28 próbek pochodzących od osób chorych i 23 prób kontrolnych. Autorzy publikacji sugerują w podsumowaniu, że analiza ekspresji genów wskazuje na liczne zmiany związane z funkcją zakończenia nerwowego [11].

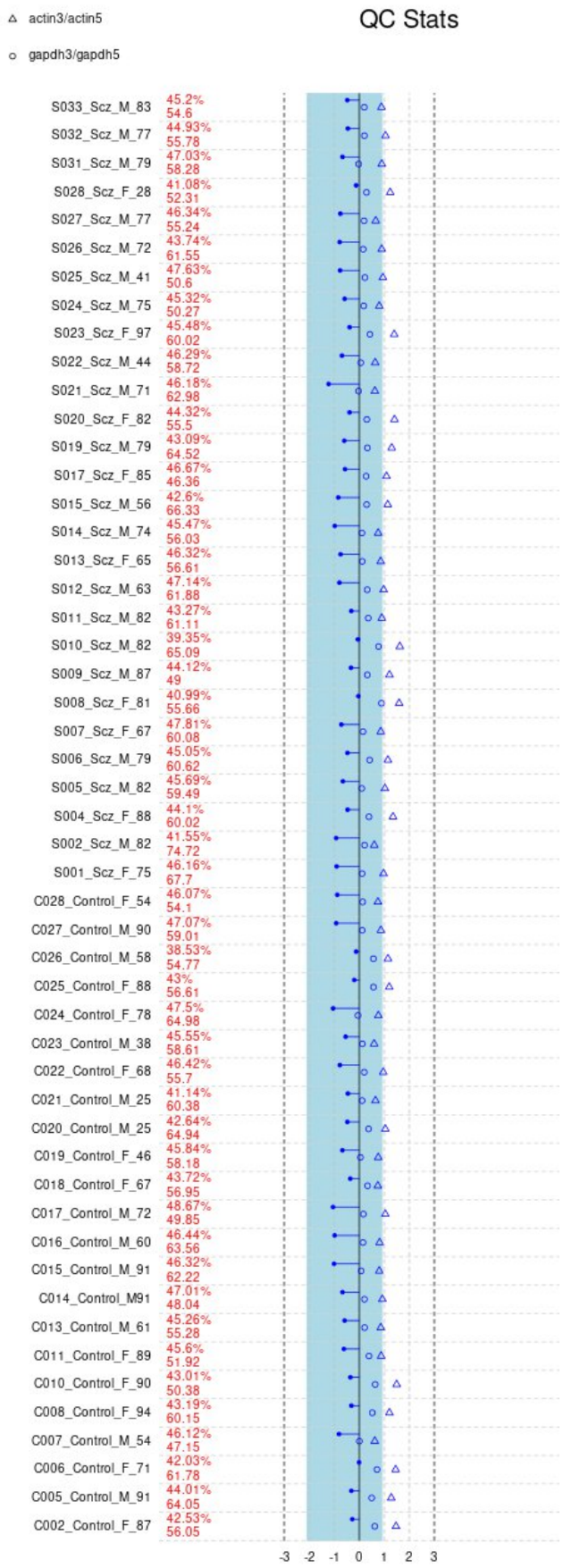
Schizofrenia jest poważnym zaburzeniem psychiatrycznym z częstością 1% na całym świecie. Patofizjologia choroby nie została dostatecznie poznana, jednak uważa się, że choroba ta ma silny komponent genetyczny, z pewnym wpływem środowiska na etiologię. Technologia mikromacierzy może okazać się przydatnym narzędziem do badania możliwych czynników sprzyjających takim schorzeniom. Jednak często nie można w pełni wykorzystać potencjału mikromacierzy ze względu na skomplikowany aspekt metrologiczny tego eksperymentu [11].

7.1. Kontrola jakości danych

Na Rys. 7.1, 7.2 i 7.3 przedstawiono wykresy zwrócone przez aplikację, w zakładce dotyczącej kontroli jakości. Wykres pudełkowy danych przed normalizacją pokazuje niewielkie zróżnicowanie w ekspresji pomiędzy próbami.

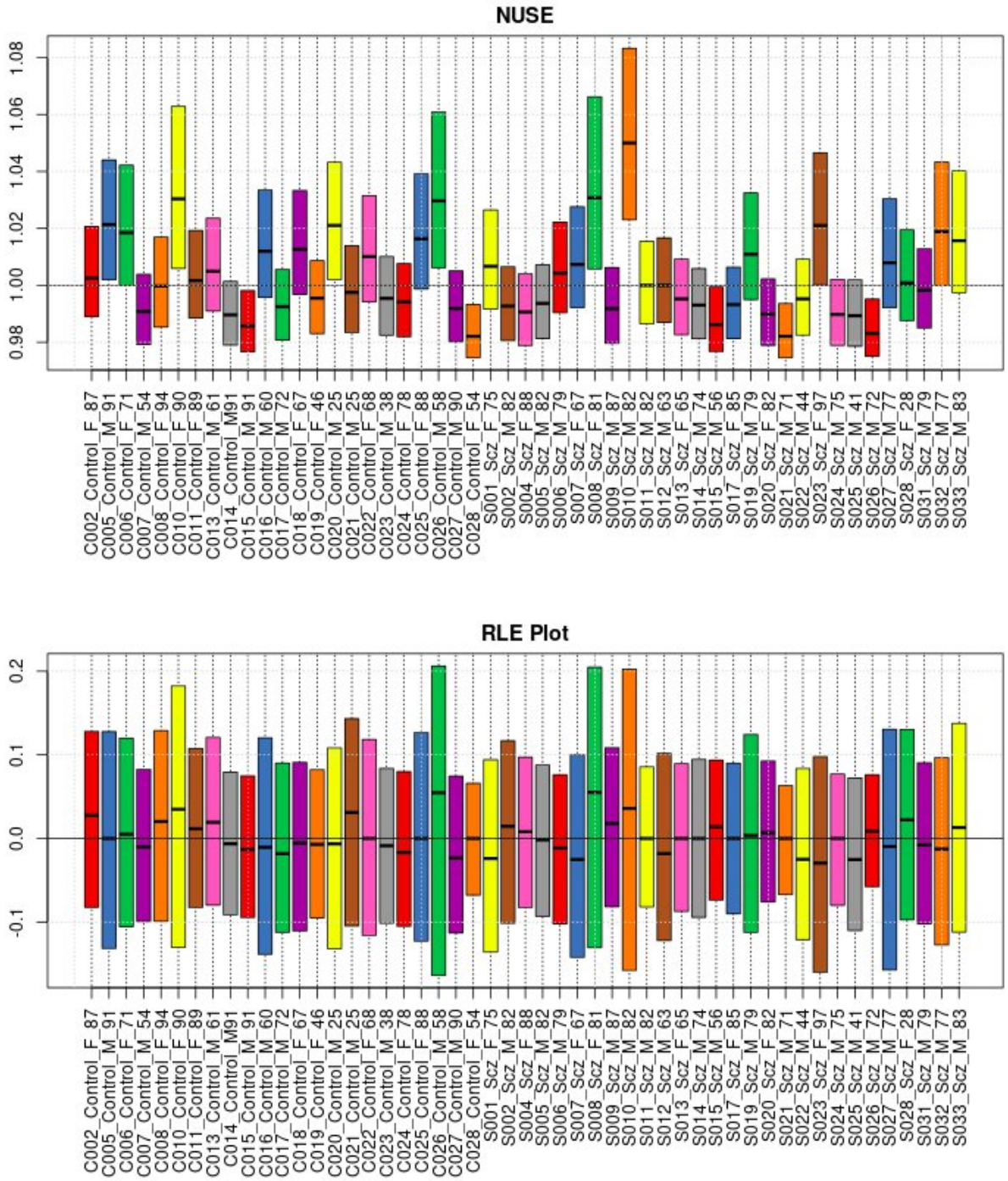


Rys. 7.1 - Wykres pudełkowy danych z prób przed normalizacją [źródło własne].



Rys. 7.2 - Wykres metryk Affymetrix [źródło własne].

Wyniki przedstawione na Rys. 7.2 sugerują, że większość obliczonych metryk nie wskazuje na niską jakość prób. W obrębie wszystkich prób występuje ponad 10% rozrzutu wartości procentowej obecnych genów, na co wskazuje czerwone zabarwienie metryk obok identyfikatorów prób ze znakiem procenta. Wartość podana bez znaku %, również zabarwiona na czerwono oznacza średnią intensywność tła, czerwona barwa po raz kolejny wskazuje na znaczący rozrzut w tej metryce pomiędzy próbami.

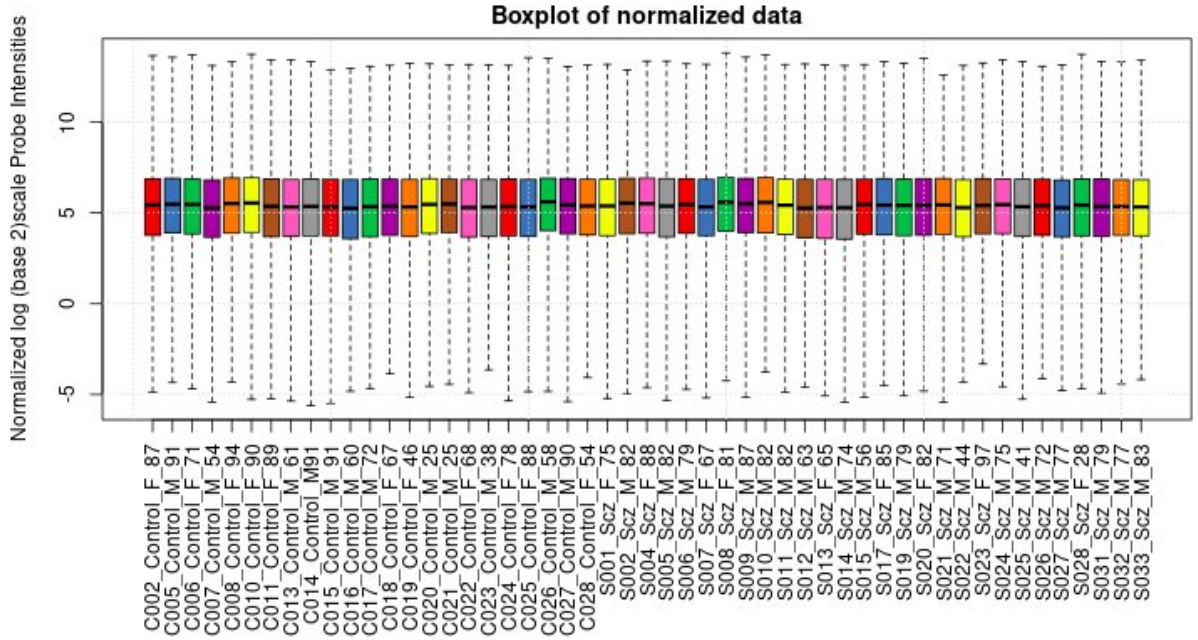


Rys. 7.3 - Wykresy NUSE oraz RLE [źródło własne].

Na wykresie NUSE widać kilka próbek (np. S010, C026, S008, C010), wyraźnie odstających od pozostałych. Na wykresie RLE widać nieco większy rozrzut wartości opisywanych tym wykresem. Próby wyróżniające się znacząco na tych wykresach warto obserwować w późniejszych etapach analizy oraz w przypadku napotkania wątpliwości co do wyników, spróbować wykluczyć wątpliwe próby i porównać wyniki przed oraz po redukcji zbioru.

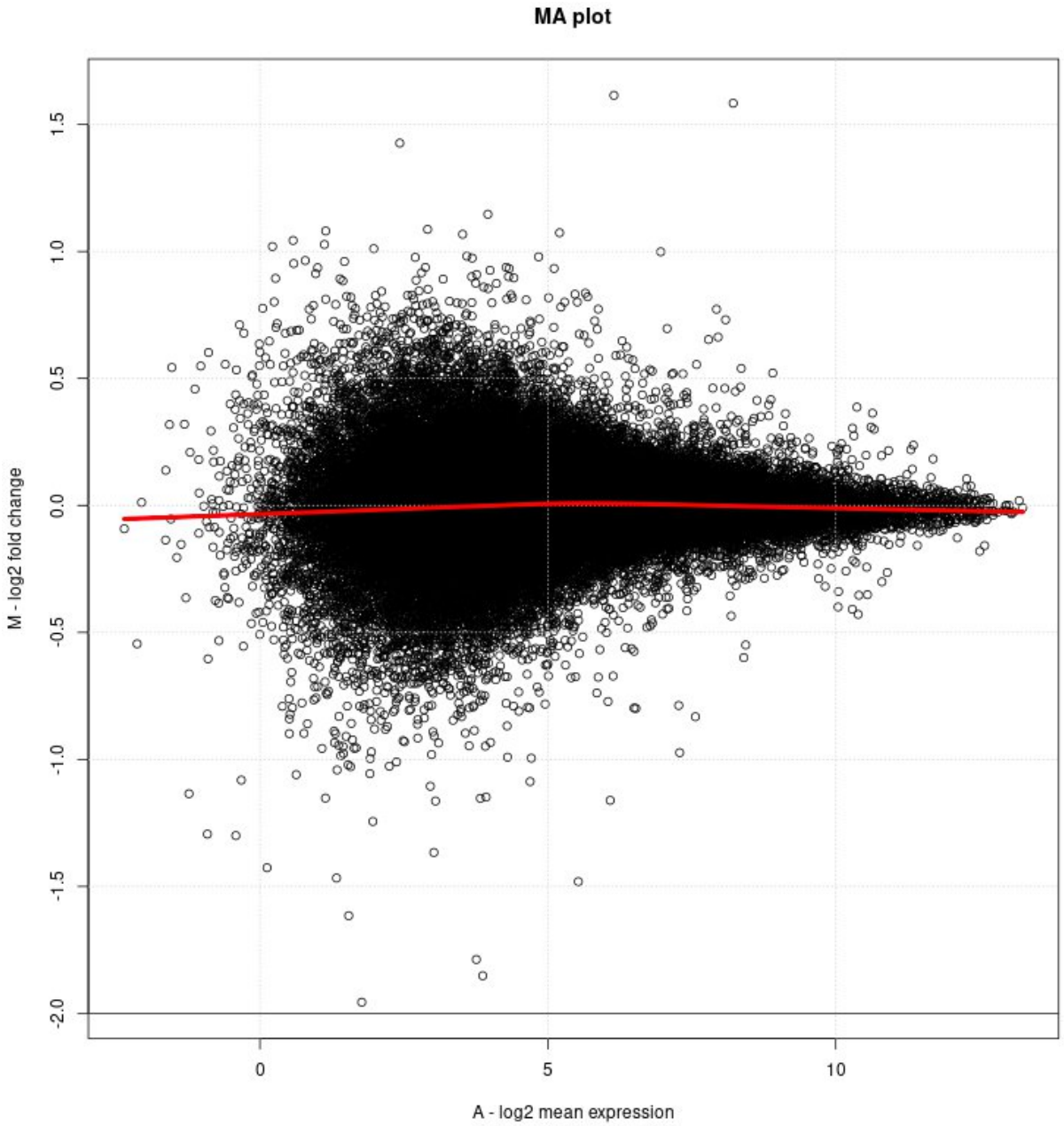
7.2. Sprawdzenie wyników normalizacji i ocena grup danych

Na Rys. 7.4 i 7.5 przedstawiono wykresy zwrócone w zakładce, odpowiadającej za sprawdzenie skuteczności normalizacji oraz porównanie pozyskanych danych po uwzględnieniu grupy kontrolnej i testowej. Wykres pudełkowy (Rys. 7.4) przedstawia dane prób z wyraźnie wyrównanymi rozkładami poziomów ekspresji.



Rys. 7.4 - Wykres pudełkowy danych po normalizacji [źródło własne].

Wykres MA natomiast (Rys. 7.5) prezentuje typowy kształt przez skupienie większości punktów wokół M = 0, ponieważ większość genów nie będzie istotnie różnicować przypadków eksperymentalnych. Z wykresu można wywnioskować, że tablice nie prezentują znaczących różnic, choć różnice te istnieją.



Rys. 7.5 - Wykres MA [źródło własne].

7.3. Typowanie genów i klasteryzacja

Wybierając metodę N-najlepszych, pozostawiono 50 genów w panelu konfiguracyjnym, przez co, w zakładce „Tabs for a specific number of genes" aplikacja wyselekcjonowała tylko 2 geny. Stało się tak, ponieważ w aplikacji sporządzono 2 listy: lista genów posortowanych malejąco względem ekspresji różnicowej oraz lista genów posortowanych rosnąco względem wartości skorygowanej p. Wspólne elementy N-pierwszych

pozycji obu tych list są wskazywane na wykresie wulkanu oraz zostają poddane klasteryzacji względem prób i genów. W aplikacji jest możliwość, aby lista dotycząca wartości p, brała pod uwagę wartości przed i po korekcji FDR.

Niestety po korekcji wartości p, żaden z genów w eksperymencie nie wykazuje się istotną różnicą pomiędzy grupą testową a kontrolną. Wyłonione 2 geny posiadają p-wartość na poziomie 48-49%, a stopień zmian obliczonych dla tych genów nie przekracza 2-krotności (Rys 7.6).

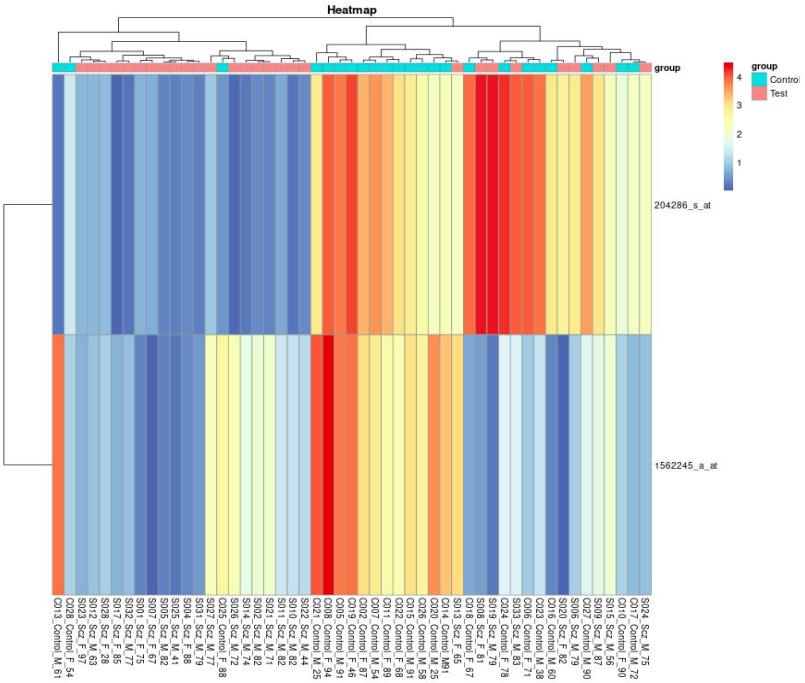
	logFC	AveExpr	t	P.Value	adj.P.Val
204286_s_at	-1.95607893334206	1.7682588080161	-4.73160688742655	0.0000178446684964593	0.487828625021957
1562245_a_at	-1.6154770249746	1.5414623982975	-4.9343277044358	0.00000886781059270857	0.484847544156341

Rys. 7.6 - Tabela zwrócona przez aplikację [źródło własne].

```
          1562245_a_at                                    204286_s_at
"zinc finger protein 578" "phorbol-12-myristate-13-acetate-induced protein 1"
```

Rys. 7.7 - Adnotacje wytypowanych genów [źródło własne].

Odszukane funkcje wyłonionych genów na podstawie adnotacji nie wydają się mieć bezpośredniego związku z schizofrenią (Rys. 7.7). W związku z uzyskanymi statystykami, wykres wulkanu dla skorygowanych wartości p jest pozbawiony kształtu, natomiast mapa cieplna wskazuje na wyraźnie większą ekspresję genu 204286_s_at dla próby kontrolnej (niebieskie kolumny), w tym klaster dla prób został podzielony w 76,5% zgodnie z adnotacją (Rys. 7.8).



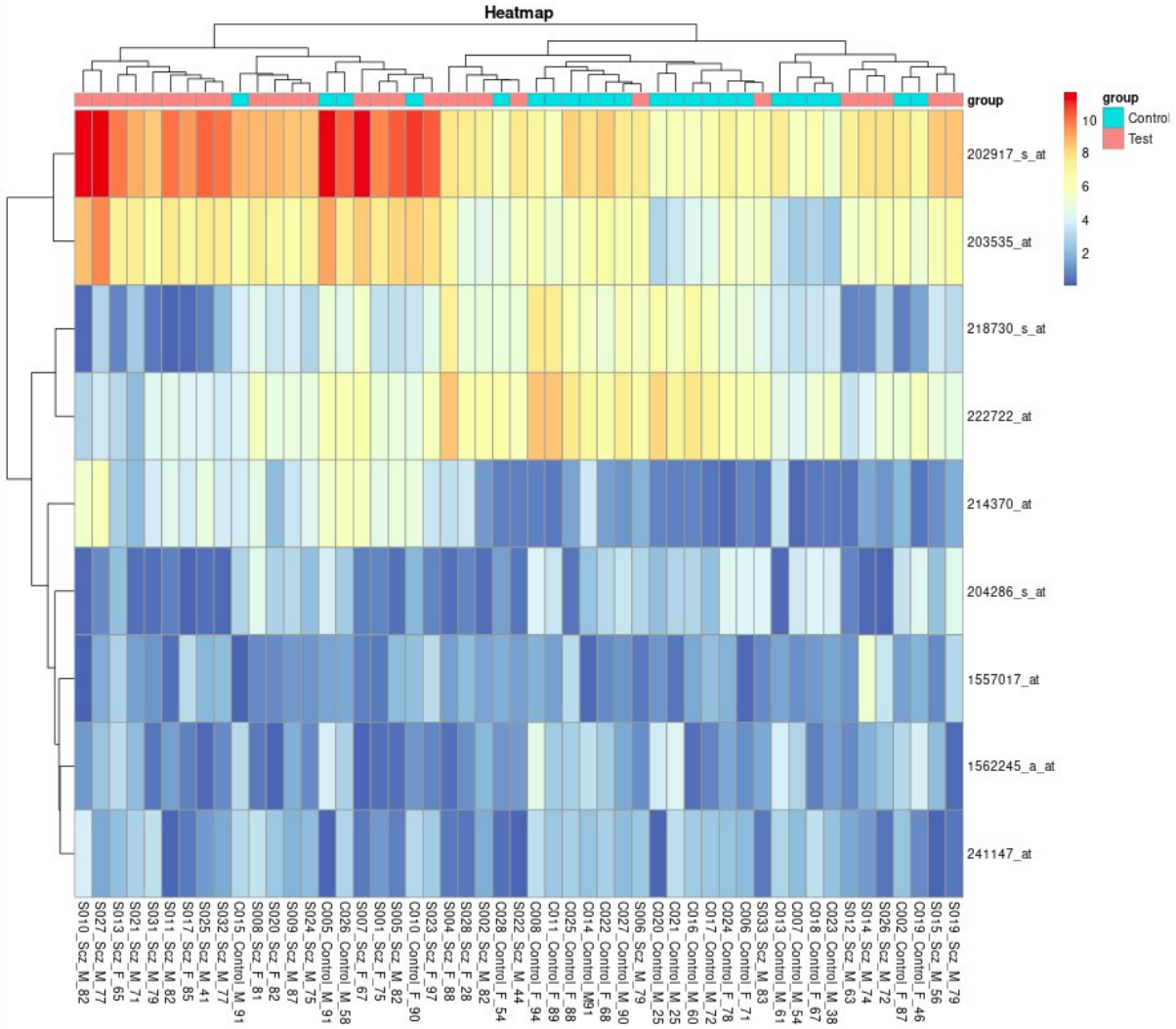
Rys. 7.8 - Mapa cieplna dla dwóch wytypowanych genów [źródło własne].

Przeprowadzono selekcję dla progów $p < 0.05$ i $FC > 1.4$ (ang. Fold Change), w której wyłoniono 9 genów, w tym również dwa poprzednie (Rys. 7.9).

```
> g[my_order,c("PROBEID","SYMBOL","GENENAME","P.Value","adj.P.Val","logFC")]
       PROBEID    SYMBOL                                                GENENAME      P.Value adj.P.Val     logFC
2 1562245_a_at    ZNF578                                  zinc finger protein 578 8.867811e-06 0.4848475 -1.615477
5  204286_s_at    PMAIP1 phorbol-12-myristate-13-acetate-induced protein 1 1.784467e-05 0.4878286 -1.956079
1   1557017_at LINC01990       long intergenic non-protein coding RNA 1990 3.248229e-03 0.9998189 -1.426597
3  202917_s_at    S100A8                    S100 calcium binding protein A8 4.546543e-04 0.9998189  1.583794
4    203535_at    S100A9                    S100 calcium binding protein A9 2.657500e-04 0.9998189  1.614823
6    214370_at    S100A8                    S100 calcium binding protein A8 4.794067e-03 0.9998189  1.426486
7  218730_s_at       OGN                                       osteoglycin 1.729640e-03 0.9998189 -1.787759
8    222722_at       OGN                                       osteoglycin 2.670769e-04 0.9998189 -1.481039
9    241147_at      <NA>                                              <NA> 6.762986e-04 0.9998189 -1.467165
```

Rys. 7.9 - Zestawienie adnotacji i statystyk wyłonionych genów [źródło własne].

Wynik grupowania sond w 70.6% pasuje do adnotacji grup (Rys. 7.10). W aktualnej skali, poprzednio wytypowane geny nie wydają się mieć widocznych różnic w ekspresji pomiędzy grupami. Trudno dostrzec wyraźne wzorce w wynikach tego grupowania, a funkcje wytypowanych genów nie wydają się mieć bezpośredniego związku z chorobą.



Rys. 7.10 - Mapa cieplna dla 9 najlepszych genów [źródło własne].

Usunięcie przypuszczalnie wadliwych matryc (sondy odstające na wykresach NUSE i RLE) nie przyczyniło się do poprawy wyników. Zaobserwowano jedynie mniejszy rozrzut w wartości procentowej obecnych genów na wykresie metryk Affymetrix, które przybrały kolor niebieski.

7.4. Analiza prób pochodzących od kobiet

W związku z tym, że usunięcie poszczególnych prób nie przyniosło widocznych rezultatów, przeanalizowano osobno dane dla matryc pochodzących tylko od mężczyzn oraz kobiet. W przypadku danych od kobiet udało się uzyskać w 100% dopasowane klastry do adnotacji prób (Rys. 7.13) oraz wyłoniono jeden statystycznie istotny gen, po korekcji FDR - PPARA (receptor aktywowany przez proliferatory peroksysomów alfa), przestawiony na Rys. 7.11 oraz w pierwszym wierszu na Rys.7.12.

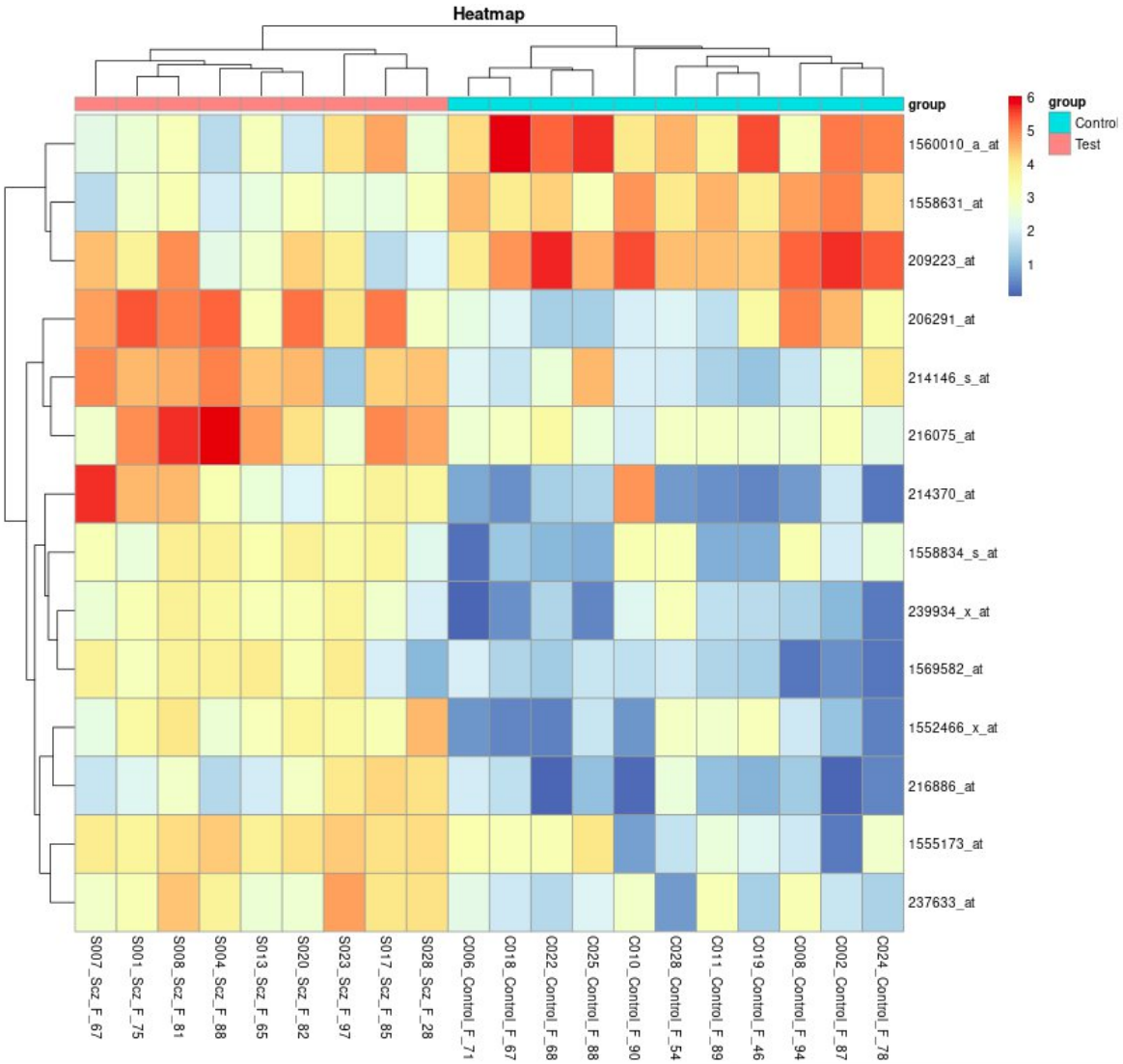
	logFC	AveExpr	t	P.Value	adj.P.Val
1558631_at	-1.70108144468068	3.50308934897067	-7.14138982831862	5.84989103208966e-7	0.0319842792179502

Rys. 7.11 - Tabelka zwrócona przez aplikację [źródło własne].

Uwzględniając funkcje biologiczne wytypowanych genów, nie znaleziono bezpośredniego związku z badaną chorobą. Szczegóły wytypowanych genów znajdują się na Rys. 7.12.

```
> g[my_order,c("PROBEID","SYMBOL","GENENAME","P.Value","adj.P.Val","logFC")]
        PROBEID    SYMBOL                                          GENENAME      P.Value  adj.P.Val     logFC
3    1558631_at     PPARA peroxisome proliferator activated receptor alpha 5.849891e-07 0.03198428 -1.701081
6    1569582_at   AADACP1          arylacetamide deacetylase pseudogene 1 3.355385e-05 0.45863924  1.843879
15  239934_x_at      <NA>                                            <NA> 3.071181e-05 0.45863924  1.915595
1  1552466_x_at LINC00161      long intergenic non-protein coding RNA 161 1.206492e-04 0.65964958  1.989655
11    214370_at    S100A8                 S100 calcium binding protein A8 1.080408e-04 0.65964958  2.458089
12    216075_at   ST3GAL5 ST3 beta-galactoside alpha-2,3-sialyltransferase 5 7.606496e-05 0.65964958  1.750014
2    1555173_at     STX19                                       syntaxin 19 1.632519e-04 0.74381626  1.698075
14    237633_at      <NA>                                            <NA> 1.532646e-04 0.74381626  1.551889
5  1560010_a_at SATB2-AS1                          SATB2 antisense RNA 1 2.005923e-04 0.80854858 -1.832985
13    216886_at    CHRNA4   cholinergic receptor nicotinic alpha 4 subunit 2.070358e-04 0.80854858  1.812036
4  1558834_s_at    AKNAD1                        AKNA domain containing 1 1.029487e-03 0.99999178  1.568316
7     206291_at       NTS                                      neurotensin 1.023513e-03 0.99999178  1.839684
8     209223_at        IK                                      IK cytokine 5.518166e-04 0.99999178 -1.579625
9     209223_at    NDUFA2     NADH:ubiquinone oxidoreductase subunit A2 5.518166e-04 0.99999178 -1.579625
10  214146_s_at      PPBP                      pro-platelet basic protein 5.197612e-04 0.99999178  1.859297
```

Rys. 7.12 - Zestawienie adnotacji i statystyk wytypowanych genów [źródło własne].



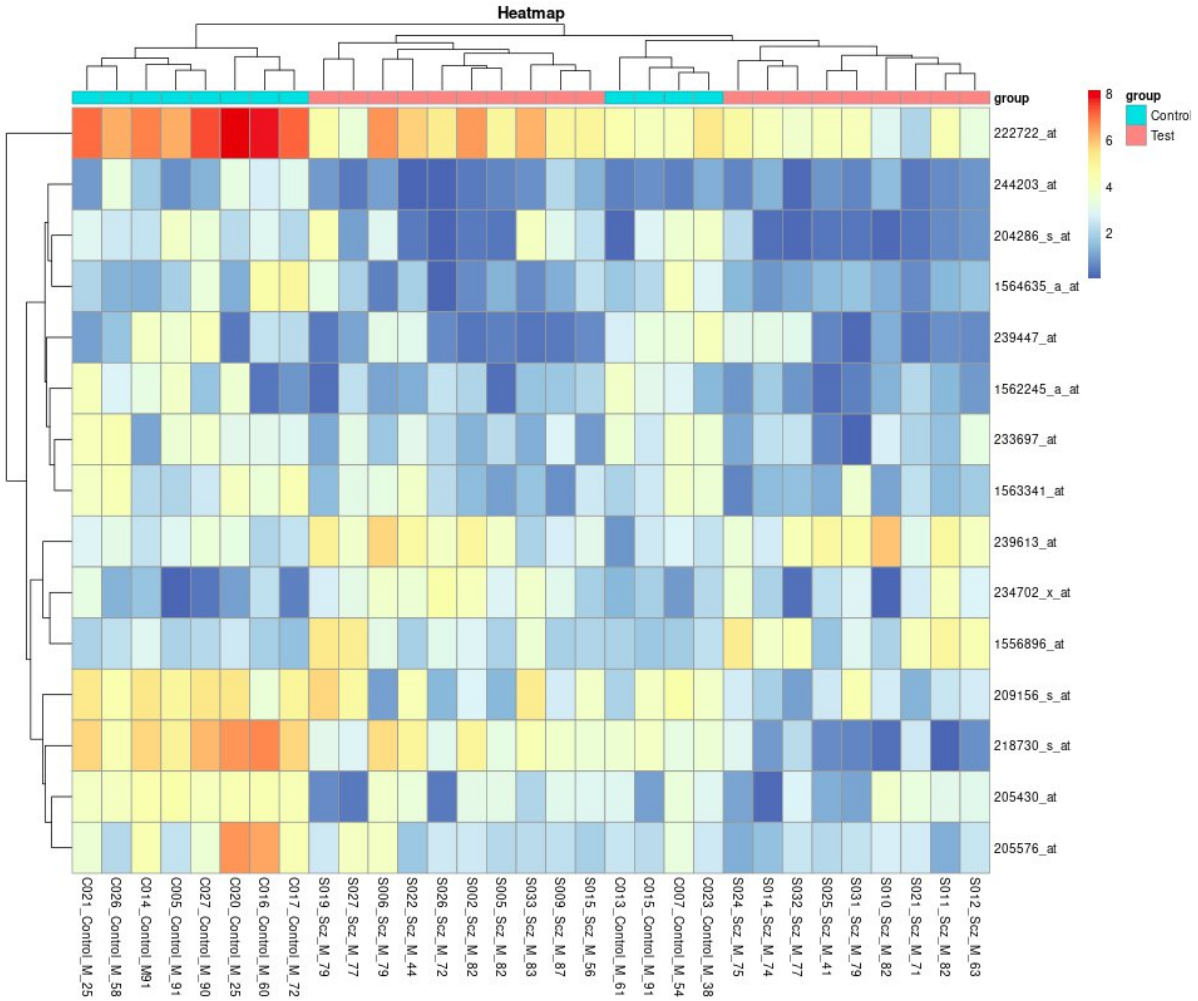
Rys. 7.13 - Mapa cieplna dla N-najlepszych genów [źródło własne].

7.5. Analiza prób pochodzących od mężczyzn

W przypadku analizy prób pochodzących od mężczyzn, wygląd wykresu metryk QC nie uległ istotnym zmianom. Nie uzyskano również żadnych istotnych różnic w ekspresji genów pomiędzy próbami, pochodzącymi od chorych a próbami kontrolnymi. Wykres mapy cieplnej dla wszystkich prób od mężczyzn przedstawiono na Rys. 7.15. W zestawieniu na Rys. 7.14 znajdują się szczegóły wytypowanych genów. Wynik klasteryzacji zawarty na mapie cieplnej zgadza się w 87,1% z danymi adnotacji.

```
> g[my_order,c("PROBEID","SYMBOL","GENENAME","P.Value","adj.P.Val","logFC")]
       PROBEID    SYMBOL                                                GENENAME      P.Value adj.P.Val     logFC
1    1556896_at LINC01270                long intergenic non-protein coding RNA 1270 0.0014370881 0.9999947  1.308445
2  1562245_a_at    ZNF578                                      zinc finger protein 578 0.0003665875 0.9999947 -1.774864
3    1563341_at      <NA>                                                        <NA> 0.0005720469 0.9999947 -1.295963
4  1564635_a_at     FHAD1 forkhead associated phosphopeptide binding domain 1 0.0009961184 0.9999947 -1.298860
5   204286_s_at    PMAIP1   phorbol-12-myristate-13-acetate-induced protein 1 0.0011994776 0.9999947 -1.887731
6     205430_at      BMP5                           bone morphogenetic protein 5 0.0016369798 0.9999947 -1.455323
7     205576_at  SERPIND1                               serpin family D member 1 0.0016848422 0.9999947 -1.345224
8   209156_s_at    COL6A2                         collagen type VI alpha 2 chain 0.0015594563 0.9999947 -1.628050
9   218730_s_at       OGN                                           osteoglycin 0.0001552544 0.9999947 -2.382595
10    222722_at       OGN                                           osteoglycin 0.0008162754 0.9999947 -1.706070
11    233697_at      <NA>                                                        <NA> 0.0001023940 0.9999947 -1.433361
12  234702_x_at      CFTR                 CF transmembrane conductance regulator 0.0007006533 0.9999947  1.512221
13    239447_at     TRA2B                           transformer 2 beta homolog 0.0014274548 0.9999947 -1.530312
14    239613_at     TMED3                transmembrane p24 trafficking protein 3 0.0001141768 0.9999947  1.500647
15    244203_at      <NA>                                                        <NA> 0.0003598345 0.9999947 -1.588021
```

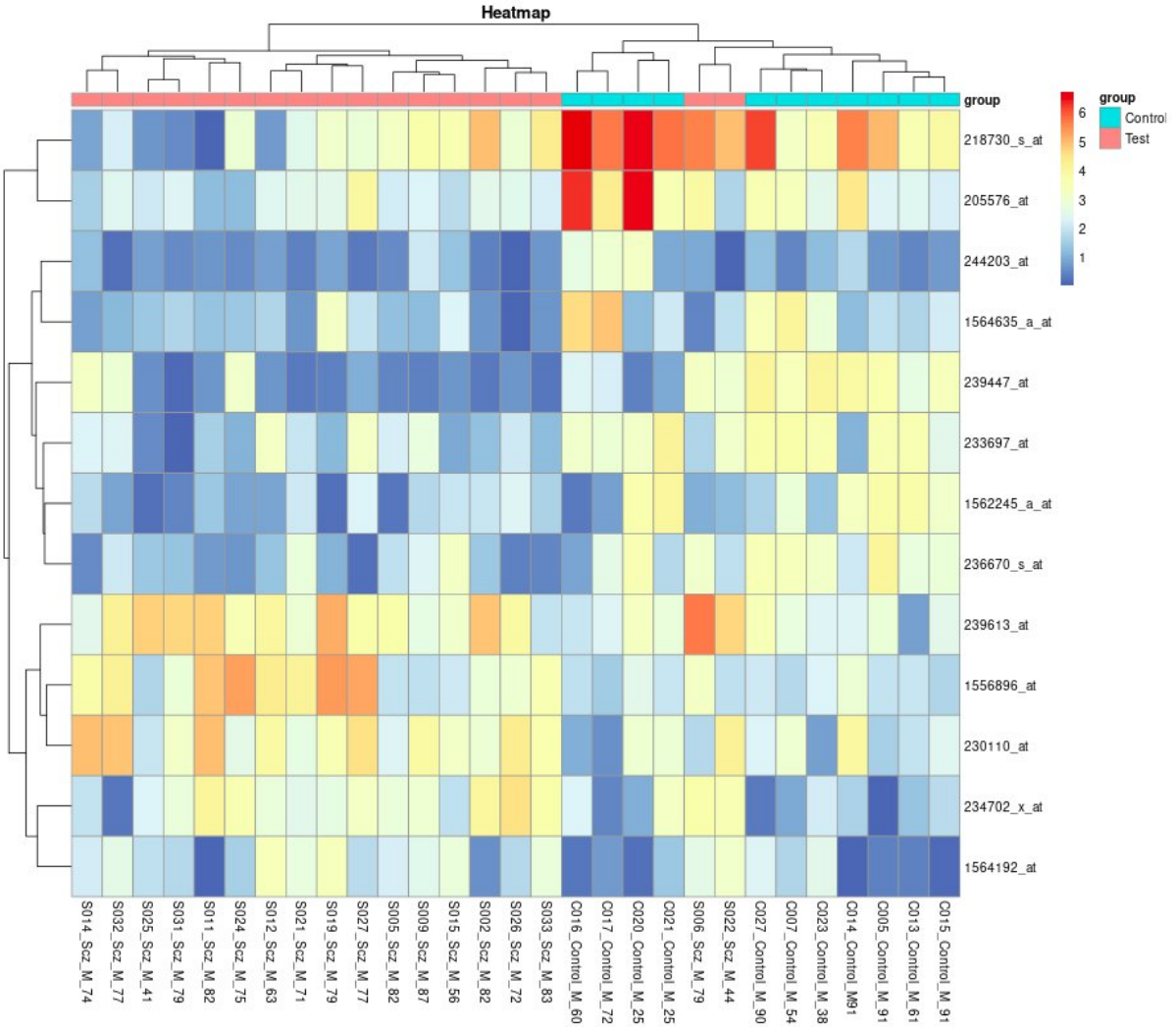
Rys. 7.14 - Tabela wytypowanych genów dla N-najlepszych genów [źródło własne].



Rys. 7.15 - Mapa cieplna dla N-najlepszych genów [źródło własne].

Na podstawie oceny wykresów NUSE i RLE wykluczono próby S010 oraz C026. W wyniki tej redukcji wykres metryk Affymetrix wskazał na dopuszczalny rozrzut wartości procentowej genów obecnych, przez co,

metryka zmieniła kolor na niebieski. Otrzymany w przypadku tego zbioru wykres mapy cieplnej wskazuje na 93,1% zgodności podziału na grupy z informacjami, pochodzącymi z adnotacji danych (Rys. 7.16).

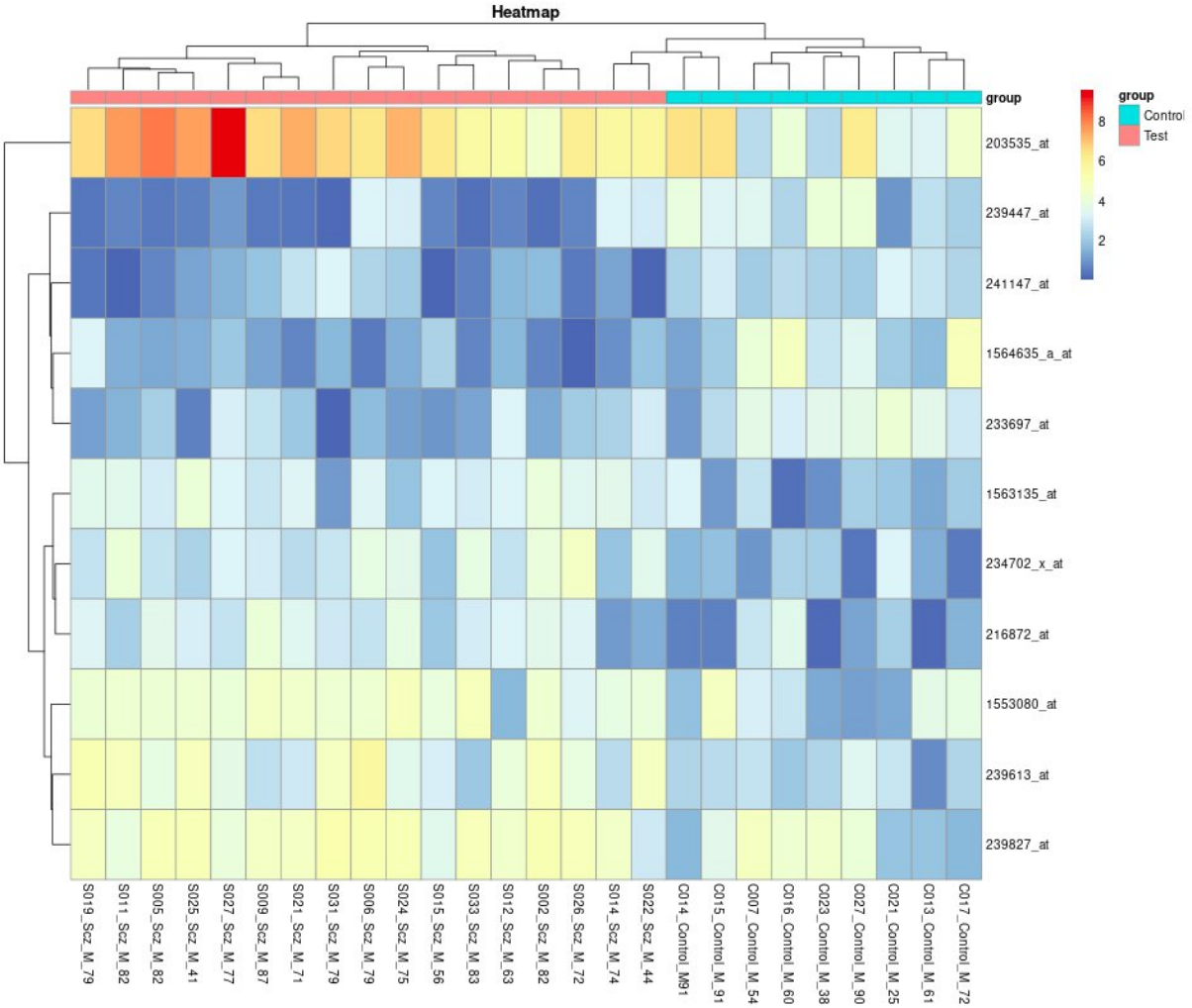


Rys. 7.16 - Mapa cieplna dla N-najlepszych genów [źródło własne].

Zestaw próbek zredukowano po raz drugi, tym razem o próby C020, C005 i S032. Dla zbioru zredukowanego o 5 prób uzyskano mapę cieplną z 100% poprawnie pogrupowanych próbek, zgodnie z adnotacją (Rys. 7.17). Szczegóły dotyczące wytypowanych genów przedstawiono na Rys. 7.18.

```
> g[my_order,c("PROBEID","SYMBOL","GENENAME","P.Value","adj.P.Val","logFC")]
        PROBEID    SYMBOL                                                GENENAME      P.Value adj.P.Val     logFC
2    1563135_at LINC02375        long intergenic non-protein coding RNA 2375 0.0002000481 0.8570635  1.463451
3  1564635_a_at     FHAD1 forkhead associated phosphopeptide binding domain 1 0.0002437951 0.8570635 -1.665885
4     203535_at     S100A9                     S100 calcium binding protein A9 0.0003002750 0.8570635  2.257632
7   234702_x_at      CFTR               CF transmembrane conductance regulator 0.0000744501 0.8570635  1.637768
8     239447_at     TRA2B                          transformer 2 beta homolog 0.0002941663 0.8570635 -1.911398
9     239613_at     TMED3           transmembrane p24 trafficking protein 3 0.0002854620 0.8570635  1.564335
5     216872_at      <NA>                                               <NA> 0.0005180059 0.9766198  1.530645
11    241147_at      <NA>                                               <NA> 0.0004894722 0.9766198 -2.039789
1    1553080_at  CSN1S2AP                casein alpha s2 like A, pseudogene 0.0009401498 0.9982299  1.414911
6     233697_at      <NA>                                               <NA> 0.0009992367 0.9982299 -1.383465
10    239827_at      RGCC                                regulator of cell cycle 0.0006588569 0.9982299  1.425050
```

Rys. 7.17 - Zestawienie szczegółów dla N-najlepszych genów [źródło własne].



Rys. 7.18 - Mapa cieplna dla N-najlepszych genów [źródło własne].

Niestety poszukując funkcji biologicznych dla wyszczególnionych genów, nie napotkano bezpośredniego powiązania z badanym schorzeniem.

7.6. Wnioski

Analiza danych z tego eksperymentu mikromacierzowego nie dostarczyła dowodu na istnienie istotnych różnic w ekspresji genów pomiędzy próbami pobranymi od ludzi chorych i zdrowych. W przypadku analizy prób pochodzących tylko od kobiet, pomimo znalezienia jednego genu o istotnie różnej ekspresji pomiędzy grupami, nie powiązano go funkcjonalnie ze schizofrenią. Natomiast w przypadku danych pochodzących tylko od mężczyzn, nie znaleziono żadnych istotnych różnic pomiędzy porównywanymi grupami. Porównanie otrzymanych wyników z wnioskami

autorów publikacji nie jest możliwe, ze względu na inny charakter wykonanej pracy. W przypadku tej analizy skupiono się na znalezieniu różnic pomiędzy grupami oraz istotnością tych zmian, natomiast autorzy publikacji skupili się na analizie innego typu oraz porównaniu uzyskanych danych z innym zbiorem dotyczącym schizofrenii.

8. Podsumowanie

W ramach pracy utworzono aplikację opartą o technologie webowe i przetestowane pakiety bioinformatyczne R, zamiast rozwiązania specyficznego, przy pomocy języków niskiego poziomu i języka Ruby. Pomimo tego, utworzone narzędzie spełnia swoją funkcję, umożliwiając zręczne przetwarzanie danych mikromacierzowych oraz dynamiczne typowanie poszczególnych genów w obrębie interfejsu użytkownika. W przypadku chęci dalszej pracy nad danymi, bądź zamiaru zmiany środowiska z uzyskanymi już danymi, możliwe jest pobranie danych w formacie tekstowym lub wyłączenie narzędzia i przystąpienie do dalszej obróbki danych z poziomu skryptu, ponieważ wszelkie obiekty danych pozostaną w pamięci sesji języka R.

Narzędzie MicAff można uznać za prototyp, biorąc pod uwagę trudności związane z wdrożeniem jako aplikację internetową. MicAff funkcjonuje dobrze jako aplikacja okienkowa w połączeniu z dynamicznym środowiskiem języka R, jednak posiada wiele ograniczeń, chociażby w stosunku do danych, które ma przetwarzać - eksperymenty muszą porównywać tylko 2 grupy, do wyboru są tylko dwa algorytmy normalizacji, a raz przerwana akcja aplikacji, nie może zostać wznowiona - pozostają tylko obiekty do manualnej manipulacji za pomocą skryptu.

Praca nad tym narzędziem dostarczyła wiedzy i doświadczenia, jak trudno zaprojektować rozwiązanie dla tak skomplikowanych i ciężkich danych. Tworzenie bibliotek dla języków programowania wydaje się więc jedyną słuszną drogą przy pracy nad tego typu narzędziami, a umiejętność posługiwania się skryptem niezbędną do pracy w takim charakterze. Ciekawą opcją jest tworzenie zaawansowanych środowisk programowania opartych o interfejs użytkownika wraz z maksymalnie prostym DSL-em (ang. Domain-specific language), którego doskonałym przykładem jest Matlab.

Bibliografia

1. M. Kotowska[1], J. Zakrzewska-Czerwińska[1,2], Prace przeglądowe, "Kurs szybkiego czytania DNA - nowoczesne techniki sekwencjonowania, [1]Instytut Immunologii i Terapii Doświadczalnej im. Ludwika Hirszfelda, Polska Akademia Nauk, [2]Wydział Biotechnologii, Uniwersytet Wrocławski, Wrocław. [www.pfb.info.pl/files/kwartalnik/4_2010/02.%20kotowska.pdf - dostęp: 10.02.2020]

2. P. Mackiewicz[1], J. Zakrzewska-Czerwińska[2], S. Cebrat[1], Prace przeglądowe, "Genomika - dziedzina wiedzy XXI wieku", [1]Zakład Genomiki, Instytut Genetyki i Mikrobiologii, Uniwersytet Wrocławski, Wrocław, [2]Zakład Mikrobiologii, Instytut Immunologii i Terapii Doświadczalnej im. L. Hirszfelda, Polska Akademia Nauk, Wrocław. [www.pfb.info.pl/files/kwartalnik/3_2005/Mackiewicz-Zakrzewska.pdf - dostęp: 10.02.2020]

3. P. G. Higgs, T. K. Attwood, Bioinformatics and Molecular Evolution - Chapter 13.2 „How do microarrays work?", Blackwell Science Ltd, 2005 [www.gen-info.osaka-u.ac.jp/~yonishi/bioinfo-new.pdf - dostęp: 7.11.2020]

4. M. Madan Babu - Chapter 11 - An Introduction to Microarray Data Analysis. [www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/chapter-final.pdf - dostęp: 10.02.2020]

5. www.genomics-online.com/resources/16/5049/housekeeping-genes - dostęp: 7.11.2020

6. C. Wilson[1], S. D Pepper[2], C. J Miller[1], QC and Affymetrix data, [1]Bioinformatics Group, [2]Molecular Biology Core Facility, Paterson Institute for Cancer Research. Christie Hospital NHS Trust, Wilmslow Road, Withington, Manchaster M20 4BX UK. [www.bioconductor.org-/packages//2.12/bioc/vignettes/simpleaffy/inst/doc/QCandSimpleaffy.pdf - dostęp: 31.10.2020]

7. B. Bolstad - affyPLM: Model Based QC Assessment of Affymetrix GeneChips, 27 Październik 2020r. [bioconductor.org/packages-/release/bioc/vignettes/affyPLM/inst/doc/QualityAssess.pdf - dostęp: 1.11.2020]

8. Blog poświęcony przetwarzaniu danych - datadeluge.blogs-pot.com/2010/10/nuse-and-rle-plots.html - dostęp 1.11.2020

9. Strona poświęcona analizie danych mikromacierzowych za pomocą narzędzia Bioconductor języka programowania R - wiki.bits.vi-b.be/index.php/Analyze_your_own_microarray_data_in_R/Bioconduc tor - dostęp 1.11.2020

10. L. Handschuh[1,2], T. Magacz[2], M. Figlerowicz[2], Prace przeglądowe, "Mikromacierze DNA w diagnostyce medycznej", [1]Katedra i Klinika Hematologii i Chorób Rozroztowych Układu Krwiotwórczego, Uniwersytet Medyczny im. Karola Marcinkowskiego, Poznań, [2]Instytut Chemii Bioorganicznej, Polska Akademia Nauk, Poznań. [www.pfb.info.pl/files/kwartalnik/2_2009/07.%20Handschuh.pdf - dostęp: 10.02.2020]

11. Maycox PR, Kelly F, Taylor A, Bates S et al. Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. Mol Psychiatry 2009 Dec;14(12):1083-94. PMID: 19255580 [ww-w.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17612 - dostęp: 3.11.2020]