Instytut Chemii Bioorganicznej PAN

Zakład Biologii Molekularnej i Systemowej

Zespół Genetyki i Genomiki Molekularnej

Optymalizacja ścieżek analizy niestandardowych danych uzyskiwanych przy użyciu mikromacierzy DNA

Rozprawa doktorska

Barbara Uszczyńska

Promotor: dr hab. Piotr Kozłowski, prof. IChB

Poznań 2013

Spis treści

Wykaz najważniejszych skrótów

Alexa 555	–	barwnik fluorescencyjny
Alexa647	–	barwnik fluorescencyjny
AML	–	ang. *acute myeloid leukemia*
AUC	–	ang. *area under curve*
Cy3	–	barwnik fluorescencyjny
Cy5	–	barwnik fluorescencyjny
GPR	–	ang. *GenePix Results* (rozszerzenie pliku komputerowego)
LNA	–	ang. *locked nucleic acid*
miRNA	–	ang. *microRNA*
SC	–	ang. *spot characteristics*
SDS	–	ang. *sodium dodecyl sulfate*
SSC	–	ang. *sodium saline citrate buffer*

I. Wprowadzenie

Ekspresja informacji zapisanej w genomowym DNA to wieloetapowy i precyzyjnie regulowany proces, którego prawidłowe funkcjonowanie decyduje o przeżyciu i sukcesie reprodukcyjnym organizmów. Komórki wchodzące w skład danego organizmu zawierają identyczną ilość DNA, jednakże aktywność poszczególnych genów może być różna. Poznanie zmian zachodzących w transkryptomie w odpowiedzi na różne bodźce zewnętrzne i wewnętrzne stanowi cenną informację o sposobie w jaki organizmy reagują na dane warunki biologiczne czy patologiczne. Analiza ekspresji genów w różnych sytuacjach (fizjologicznych, rozwojowych, stresogennych, środowiskowych, itd.) pozwala na powiązanie informacji na temat aktywności genów z fenotypem komórki, tkanki czy fenotypem całego organizmu. Dzięki temu istnieje możliwość lepszego zrozumienia funkcji poszczególnych genów. Takie podejście nazywane jest genomiką funkcjonalną.

Mikromacierze DNA, czyli regularnie rozmieszczone na podłożu stałym fragmenty DNA o różnej sekwencji to jedna z nielicznych w ostatnich czasach technik badania kompletnych transkryptomów. Wyniki uzyskane z użyciem mikromacierzy DNA w znaczący sposób poszerzyły naszą wiedzę na temat mechanizmów zgodnie z którymi funkcjonują układy żywe: ludzie, zwierzęta, rośliny i mikroorganizmy. O ogromnej popularności tej techniki świadczy duża liczba opublikowanych w ostatnich latach prac poświęconych analizom mikromacierzowym. W badaniach ekspresji genów mikromacierze DNA wykorzystywane są od ponad dekady. Przez ten czas rozwinięto i doskonale opracowano standardowe metody analizy dla mikromacierzy DNA, będących zazwyczaj produktami komercyjnymi wykorzystywanymi w badaniach całogenomowych. Standaryzacja metod analizy (standardy MIAME etc.) była jednym z głównych kierunków w jakich dokonywał się rozwój technik mikromacierzowych. Przetwarzanie i analiza uzyskiwanych danych wymaga zastosowania zaawansowanych metod bioinformatycznych oraz wiedzy z dziedziny matematyki i statystyki. Jednakże dostępność szerokiej gamy standardowych narzędzi spowodowała, że technika ta stała się dostępna także dla osób nie posiadających rozległej wiedzy w wyżej wymienionych dziedzinach.

Ograniczona oferta komercyjnych mikromacierzy DNA oraz kwestie dotyczące własności intelektualnej i komercjalizacji badań uniemożliwiają swobodne stosowanie tych mikromacierzy do specyficznych problemów badawczych. Z tej przyczyny wiele laboratoriów produkuje własne mikromacierze DNA. Jednym z takich miejsc jest Centrum Doskonałości CENAT, a obecnie Pracownia Mikromacierzy i Głębokiego Sekwencjonowania w Instytucie Chemii Bioorganicznej PAN. Samodzielna produkcja daje możliwość tworzenia

unikatowych mikromacierzy DNA, często w uproszczonej formie. Uproszczenie mikromacierzy DNA, na przykład poprzez ograniczenie liczby sond jedynie do takich, które reprezentują transkrypty o dużej lub niskiej aktywności w badanych warunkach, paradoksalnie znacznie komplikuje proces analizy danych. Dane tego typu w większości przypadków nie mogą być analizowane za pomocą standardowych procedur. Utrudnienia te często są jednak identyfikowane dopiero na etapie przetwarzania danych, po zakończonej części eksperymentalnej. Brak możliwości zastosowania standardowych procedur nie musi jednak oznaczać, że z niestandardowych danych nie można uzyskać informacji o znaczeniu biologicznym. Wymaga to jednak użycia nietypowych czy też rozszerzonych procedur, adekwatnych do rodzaju użytych mikromacierzy DNA.

Niniejsza praca doktorska poświęcona jest identyfikacji i charakterystyce najczęściej spotykanych czynników utrudniających proces analizy niestandardowych danych oraz modyfikacji standardowych metod analizy w celu zniwelowania efektu tych czynników.

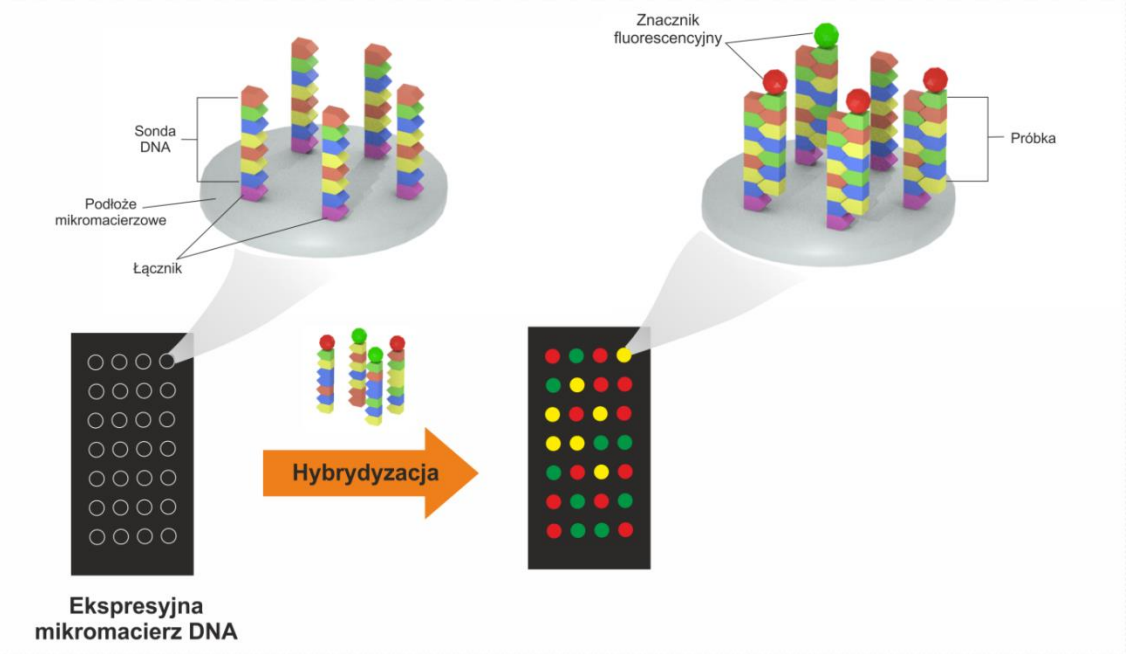
II. Wstęp

Dynamiczny rozwój genomiki na przestrzeni ostatnich 20 lat, diametralnie zmienił sposób prowadzenia badań we współczesnej biologii i medycynie. Obecnie w ramach realizacji pojedynczego eksperymentu możliwa jest analiza wielu tysięcy cząsteczek mRNA, miRNA, białek, metabolitów, oddziaływań białko-białko, mutacji genomowych, polimorfizmów oraz zmian epigenetycznych. Narzędzia do wysokoprzepustowych analiz, w szczególności mikromacierze (ang. *microarrays*) i technologia sekwencjonowania drugiej generacji (ang. *next-generation sequencing*), pozwoliły na zwiększenie szybkości generowania danych na temat systemów biologicznych. Wykorzystanie tych dwóch technologii umożliwia prowadzenie bardziej kompleksowych obserwacji na poziomie molekularnym, co wpływa także na rodzaj współcześnie zadawanych pytań w dziedzinie nauk biologicznych.

W biologii pod hasłem mikromacierze rozumie się najczęściej narzędzia powstałe w wyniku immobilizacji różnych biomolekuł na podłożu stałym. Wśród przyłączonych do powierzchni stałej cząsteczek można wyróżnić nie tylko fragmenty DNA, czy RNA, ale także białka, lipidy, węglowodany, a nawet chromosomy oraz tkanki (Simon 2004; Blalock 2003; Schena 2003; G. Hu i wsp. 2009; Rinaldi i wsp. 2009). Jednakże najpowszechniej stosowanym rodzajem mikromacierzy są mikromacierze DNA.

II.1. Mikromacierze DNA

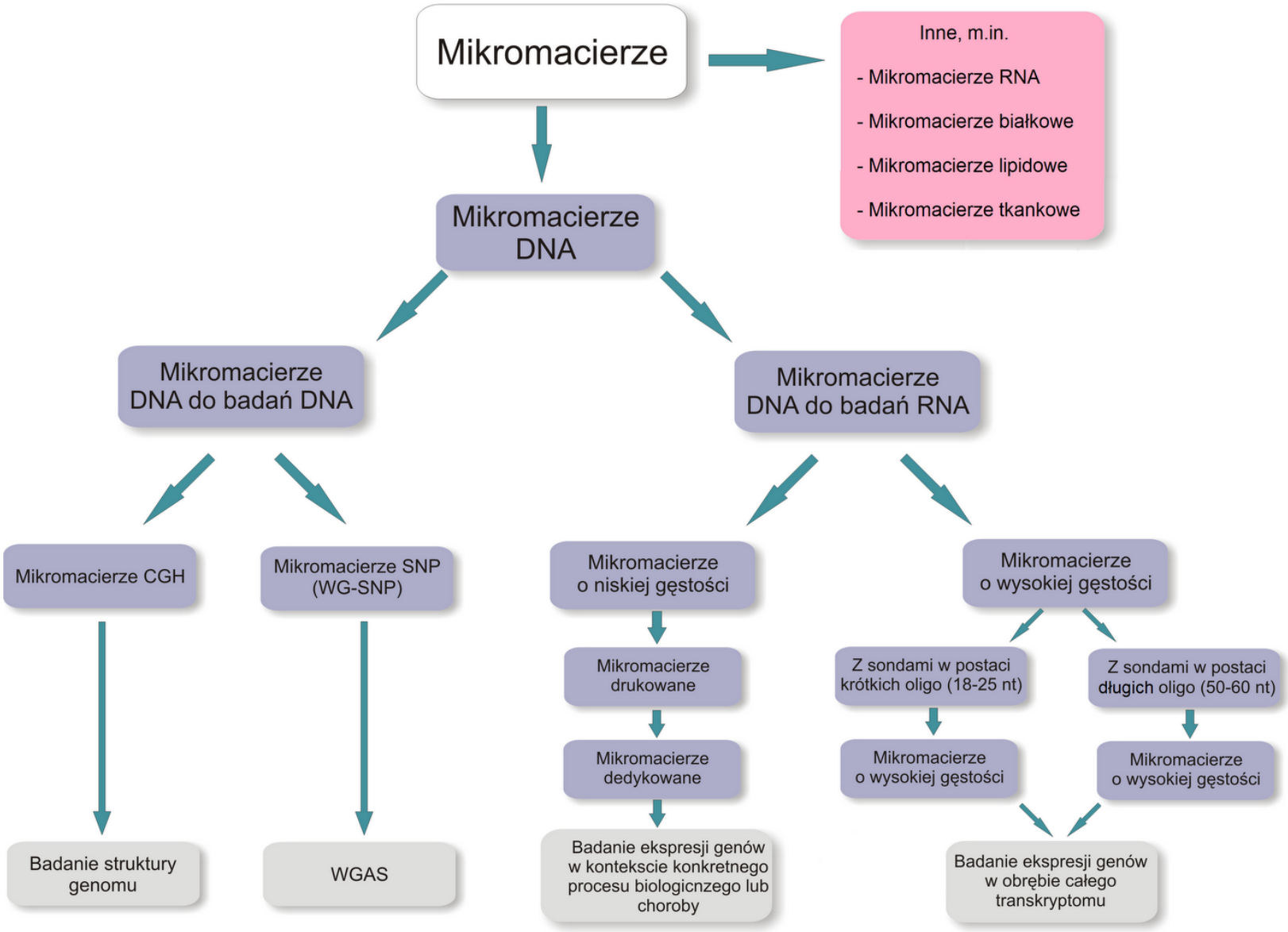
Podstawowa koncepcja mikromacierzy DNA opiera się na precyzyjnej lokalizacji na podłożu stałym (szkle, kwarcu, silikonowych ziarnach) fragmentów DNA (oligonukleotydów, fragmentów genomowych, cDNA) reprezentujących sondy (ang. *probes*) (Trevino i wsp. 2007; Venkatasubbarao 2004). Sekwencja każdej sondy jest komplementarna do specyficznego fragmentu danego genomu czy transkryptomu (Rysunek 1). Rodzaj stosowanych sond oraz typ docelowych cząsteczek (fragment genomu wyłapywany przez sondy danego rodzaju) (ang. *target*) jest różny w zależności od charakteru eksperymentu.



Rysunek 1. *Ogólny schemat działania mikromacierzy DNA.*

Sposób działania mikromacierzy DNA polega na zdolności jednoniciowych sond do selektywnej hybrydyzacji ze znakowanymi fluorescencyjnie jednoniciowymi cząsteczkami docelowymi. W wyniku reakcji hybrydyzacji powstają struktury dwuniciowe złożone z dwóch komplementarnych do siebie fragmentów, najczęściej DNA. Główną zaletą mikromacierzy DNA jest możliwość analizy wielu tysięcy fragmentów genomu lub transkryptomu w ramach pojedynczego eksperymentu. Mechanizm działania oraz format mikromacierzy DNA, obejmujący regularnie rozmieszczone sondy zapewniają niezwykłą wszechstronność tej technologii. Mikromacierze DNA posiadają szerokie spektrum zastosowań. Są one wykorzystywane do genotypowania polimorfizmów oraz mutacji (Hacia & Collins 1999), do badania odziaływań DNA-białko (Iyer i wsp. 2001), jak również do identyfikacji zmian strukturalnych przy wykorzystaniu porównawczej hybrydyzacji genomowej (ang. *comparative genome hybdridization*) (Pinkel i wsp. 1998). Jednakże przez długi czas to badania ekspresji genów kodujących białka były ich wiodącym zastosowaniem. Schemat prezentujący najbardziej popularne rodzaje obecnie stosowanych mikromacierzy DNA przedstawiony został na Rysunku 2.

Prezentowana praca doktorska poświęcona jest optymalizacji ścieżek analizy danych uzyskiwanych za pomocą mikromacierzy DNA do badania ekspresji genów kodujących białka. W związku z tym zagadnienia dotyczące tego rodzaju mikromacierzy DNA zostaną opisane szerzej w dalszej części tego rozdziału.



Rysunek 2. *Schemat przedstawiający najbardziej popularne rodzaje mikromacierzy DNA (kolor fioletowy). Kolorem różowym zaznaczono inne, dostępne rodzaje mikromacierzy, m.in. mikromacierze RNA, białkowe, lipidowe, tkankowe. Kolorem szarym zaznaczono zastosowanie każdego z wymienionych rodzajów mikromacierzy DNA.*

II.1.1 Mikromacierze DNA do badania ekspresji genów kodujących białka

Mikromacierze DNA do badania ekspresji genów (ang. *expression DNA microarrays*), umożliwiają jednoczesny pomiar ekspresji wielu tysięcy genów kodujących białka w ramach pojedynczego eksperymentu. Zawierają one na swojej powierzchni sondy w postaci oligonukleotydów DNA lub fragmentów cDNA, z których każdy komplementarny jest do specyficznego fragmentu danego genu. W wyniku reakcji hybrydyzacji do sond DNA na mikromacierzy przyłącza się materiał z próbki badanej lub kontrolnej. Najczęściej jest to znakowany fluorescencyjnie cDNA otrzymany w wyniku odwrotnej transkrypcji z mRNA wyizolowanego z tkanki badanej lub kontrolnej. Wyniki analiz prowadzonych z wykorzystaniem tego rodzaju mikromacierzy DNA umożliwiły korelację fizjologicznych stanów komórek z ich profilem ekspresji genów. Stosowanie mikromacierzy DNA do badania ekspresji genów kodujących białka pozwoliło m.in. na wykrycie zestawów genów (biomarkerów) wykazujących specyficzną ekspresję w ostrej białaczce limfoblastycznej (ang. *acute lymphoblast leukemia*)(Golub 1999), raku piersi (Veer i wsp. 2002), raku prostaty (Singh i wsp. 2002), raku płuca (Wang i wsp. 2000), indukcji apoptozy (Brachat i wsp. 2000) oraz w przypadku odpowiedzi na terapię farmakologiczną w leczeniu nowotworów (Brachat i wsp. 2002).

Dla przejrzystości tekstu mikromacierze DNA do badania ekspresji genów kodujących białka, w ślad za anglojęzyczną terminologią (ang. *expression DNA microarrays*), w dalszej części pracy nazwane będą ekspresyjnymi mikromacierzami DNA. Określenie badanie ekspresji genów w tekście zawsze będzie się odnosić do genów kodujących białka. W przypadku, gdy rozważania będą dotyczyć innych genów, kodujących np. miRNA, nazwa zostanie uściślona.

II.2 Rodzaje ekspresyjnych mikromacierzy DNA

Aktualnie dostępne są dwa rodzaje ekspresyjnych mikromacierzy DNA: o wysokiej oraz o niskiej gęstości. Ekspresyjne mikromacierze DNA o wysokiej gęstości charakteryzują się obecnością na powierzchni bardzo dużej liczby sond (nawet ponad 1 mln) i umożliwiają badanie ekspresji niemalże całego transkryptomu wybranego organizmu. Wykorzystywane są one głównie do ogólnej oceny aktywności transkryptomu w badanych warunkach. Natomiast ekspresyjne mikromacierze DNA o niskiej gęstości stosowane są do badania konkretnego procesu biologicznego lub chorobotwórczego i zawierają na swojej powierzchni głównie

sondy specyficzne jedynie dla genów, które mogą być potencjalnie zaangażowane w ten proces. Stąd też liczba sond wchodzących w skład tego rodzaju mikromacierzy jest znacznie niższa w porównaniu z mikromacierzami o wysokiej gęstości (średnio ok. 1000 sond).

II.2.1 Ekspresyjne mikromacierze DNA o wysokiej gęstości

W skład ekspresyjnych mikromacierzy DNA o wysokiej gęstości mogą wchodzić sondy w postaci krótkich (18-25 nt) lub długich (50-60 nt) oligonukleotydów DNA otrzymanych w wyniku syntezy *in situ*.

II.2.1.1 Ekspresyjne mikromacierze DNA z sondami w postaci krótkich oligonukleotydów

Ten rodzaj mikromacierzy DNA wprowadzony został na rynek przez firmę Affymetrix pod koniec lat 90-tych i pod taką nazwą funkcjonuje jako platforma. Mikromacierze Affymetrix były jednymi z pierwszych komercyjnych mikromacierzy DNA i to m.in. one w znaczący sposób przyczyniły się do rozpowszechnienia tej technologii. Mikromacierze Affymetrix posiadają szereg charakterystycznych cech. Jedną z nich jest projekt i długość sond (18-25 nt). W skład zestawu sond DNA ulokowanych na mikromacierzy mogą wchodzić dwa typy sond: PM (ang. *perfect match*) oraz MM (ang. *mis-match*). Sondy typu PM są całkowicie komplementarne do sekwencji docelowych znajdujących się w badanej próbce. Natomiast sondy typu MM nie wykazują całkowitej komplementarności do sekwencji docelowych ze względu na fakt, iż 13 nukleotyd (dla sond o długości 25 nt) został zamieniony na nukleotyd do niego komplementarny. Zadaniem sond typu PM jest pomiar ekspresji genów, natomiast sondy typu MM pozwalają określić poziom sygnału fluorescencji dla hybrydyzacji niespecyficznej (ang. *cross-hybridization*). Ze względu na niewielką długość sond, pojedynczy gen na platformie Affymertix reprezentowany jest przez cały zestaw sond. W skład pojedynczego zestawu wchodzi od 11-20 sond typu PM oraz od 11-20 sond typu MM. Występowanie dwóch różnych rodzajów sond wymaga stosowania specyficznej procedury hybrydyzacji oraz analizy danych. Jednym z poważniejszych ograniczeń stosowania tej platformy jest wysoki koszt pojedynczego eksperymentu. Koszt ten głównie wynika z metody syntezy sond, która jest połączeniem fotolitografii i chemii kombinatorycznej. Synteza sond prowadzona jest *in situ*. Eksperymenty z użyciem platformy Affymetrix prowadzone są jedynie z użyciem jednego barwnika fluorescencyjnego (eksperymenty jednokolorowe).

II.2.1.1 Ekspresyjne mikromacierze DNA z sondami w postaci długich oligonukleotydów

Obecnie na rynku dostępne są także ekspresyjne mikromacierze DNA z sondami w postaci długich oligonukleotydów DNA. Sondy najczęściej otrzymywane są w wyniku syntezy *in situ*, a sposób prowadzenia syntezy i długość sond ściśle zależą od producenta danej platformy. Do najbardziej znanych platform należą: NibmleGen (Roche), Agilent (Agilent Technologies) oraz Illumina (Illumina Inc). Platforma NimbleGen charakteryzuje się sondami o długości 50 nukleotydów, które powstają w wyniku syntezy *in situ* sterowanej przez tysiące aluminiowych lusterek, decydujących o wydłużaniu poszczególnych łańcuchów. Takie podejście skutkuje wysoką wydajnością reakcji (99,4%), krótkim czasem syntezy oraz pozwala otrzymać sondy o wysokiej jakości i ściśle kontrolowanym składzie nukleotydowym. Platforma Agilent łączy ze sobą cechy mikromacierzy drukowanych i tych syntetyzowanych *in situ*. Synteza sond odbywa się bezpośrednio na podłożu, ale w oparciu o technologię „InkJet", gdzie to drukarka kontroluje proces wydłużania poszczególnych łańcuchów oligonukleotydowych. Taka strategia pozwala otrzymać mikromacierze zawierające sondy o długości 60 nukleotydów z niemalże 100% wydajnością. Najbardziej nietypowym rodzajem ekspresyjnych mikromacierzy DNA są mikromacierze Illumina. Technologia wytwarzania tego typu mikromacierzy wykorzystuje silikonowe ziarna (ang. *beads*). Każdy rodzaj sondy (50 nt) przypisany jest do danego typu ziarna. Ziarna na podłożu umieszczone są losowo w taki sposób, że każdy typ ziarna występuje na mikromacierzy około 30 razy. W celu identyfikacji poszczególnych ziaren, a tym samym i sekwencji sond, konieczny jest dodatkowy proces dekodowania, który umożliwia poznanie lokalizacji danego typu ziarna. Podobnie jak w przypadku platformy Affymetix mikromacierze Illumina zawierają także sondy typu PM i MM, z tym, że sondy typu MM zawierają różną ilość niekomplementarnych nukleotydów (7-12). Obecność wielu ziaren tego samego typu wymaga również uwzględnienia w procesie analizy danych etapu uśredniania sygnałów fluorescencyjnych. Eksperymenty z użyciem opisywanych rodzajów mikromacierzy najczęściej prowadzone są z użyciem jednego barwnika fluorescencyjnego (Petersen i wsp. 2005). Wyjątkiem są mikromacierze produkowane przez firmy Agilent i Illumina.

II.2.2 Mikromacierze DNA o niskiej gęstości do badania ekspresji genów

II.2.2.1 Mikromacierze drukowane

Nazwa „mikromacierze drukowane" wywodzi się od metody tworzenia tego rodzaju mikromacierzy DNA. Powstają one w wyniku procesu drukowania, tj. równomiernego nanoszenia na podłoże (często za pomocą igieł) roztworu fragmentów cDNA lub długich oligonukleotydów DNA (50-70 nukleotydów) o ściśle zdefiniowanym stężeniu (25-40 μM). Proces drukowania umożliwia konstrukcję głównie ekspresyjnych mikromacierzy DNA o niskiej gęstości. Maksymalna liczba immobilizowanych sond wynosi ok. 20 tys. Liczba ta jest zmienna i ściśle zależy od właściwości stosowanej drukarki (Venkatasubbarao 2004). Drukowane mikromacierze DNA gwarantują dużą elastyczność w planowaniu eksperymentów, dzięki możliwości własnoręcznego projektowania i przygotowania mikromacierzy. Dzięki temu możliwe jest prowadzenie badań nad wybranymi procesami biologicznymi oraz gatunkami dla których dostępność komercyjnych mikromacierzy DNA jest ograniczona. Przyczyną popularności tego rodzaju mikromacierzy jest także przystępna cena pojedynczego eksperymentu, która jest kilkukrotnie niższa w porównaniu do eksperymentów prowadzonych z użyciem ekspresyjnych mikromacierzy DNA. Drukowane mikromacierze DNA najczęściej stosowane są w formie dedykowanych mikromacierzy DNA (ang. *custom arrays* lub *home made arrays*).

II.2.2.1.1 Dedykowane mikromacierze DNA

Dedykowane mikromacierze DNA służą do badania konkretnego procesu biologicznego lub choroby (Wilson i wsp. 2003; Held i wsp. 2004; Lu i wsp. 2005; Oshlack i wsp. 2007). Zawierają one na swojej powierzchni zestaw sond ograniczony jedynie do genów potencjalnie zaangażowanych w badany proces oraz niewielki zestaw sond kontrolnych. W praktyce pojedyncza dedykowana mikromacierz DNA zawiera od kilkuset do kilku tysięcy sond, co stanowi jedynie niewielki procent całkowitej liczby sond w przypadku ekspresyjnych mikromacierzy DNA o wysokiej gęstości. Pomimo szerokiej oferty gotowych do użycia komercyjnych mikromacierzy DNA, nadal istnieje potrzeba tworzenia specyficznych i mniejszych mikromacierzy. Dedykowane mikromacierze DNA często projektowane są w celu dokładniejszego poznania danego procesu biologicznego (Campanaro i wsp. 2002; Mcilroy i wsp. 2005; Ferrarini i wsp. 2008; Baron i wsp. 2011). Wynika to z faktu, iż w przypadku tego rodzaju mikromacierzy DNA liczba sond niezaangażowanych w badany proces jest

zredukowana do minimum. Stosowanie dedykowanych mikromacierzy DNA pozwala zwiększyć specyficzność eksperymentu, a tym samym zwiększyć siłę statystyczną analizy danych na skutek redukcji liczby testowanych hipotez. Mikromacierze dedykowane najczęściej wykorzystywane są w eksperymentach prowadzonych z użyciem dwóch barwników fluorescencyjnych (eksperymenty dwukolorowe).

II.3 Eksperyment badania ekspresji genów z użyciem mikromacierzy DNA

Eksperyment obejmujący badanie ekspresji genów z użyciem mikromacierzy DNA jest procesem złożonym (Rysunek 3). W klasycznym podejściu obejmuje on następujące etapy:

- Izolacja RNA
- Znakowanie próbek
- Hybrydyzacja
- Skanowanie
- Analiza ilościowa obrazu
- Analiza danych

II.3.1 Izolacja RNA

Pierwszy etap eksperymentu z użyciem mikromacierzy DNA obejmuje izolację RNA z wybranej tkanki albo hodowli komórkowej. Zwykle do przygotowania próbki wymagane jest około 0,5 µg mRNA, co odpowiada ilości ok. 20 µg całkowitego RNA (ang. *total RNA*). Wartości te mogą być różne, w zależności od rodzaju mikromacierzy DNA (platformy) (Zhu i wsp. 2006; Kaposi-Novak i wsp. 2004).

W przypadku, gdy ilość mRNA nie jest wystarczająca do przeprowadzenia eksperymentu, możliwe jest powielenie go za pomocą: procesu pulowania próbek (ang. *sample pooling*) lub amplifikacji. Pulowanie próbek polega na połączeniu całkowitego RNA uzyskanego z wybranych tkanek albo hodowli kmórkowych tego samego typu (badanych lub kontrolnych). Natomiast proces amplifikacji RNA najczęściej obejmuje dwa etapy. Pierwszym z nich jest synteza antysensownej nici cDNA z wykorzystaniem startera oligo(dT), komplementarnego do ogona poli(A). Starter oligo(dT) na końcu 5’ zawiera także dodatkową sekwencję zwykle jest to sekwencja promotora polimerazy T7. Amplifikacja cDNA najczęściej w takim układzie następuje w wyniku transkrypcji *in vitro* z użyciem

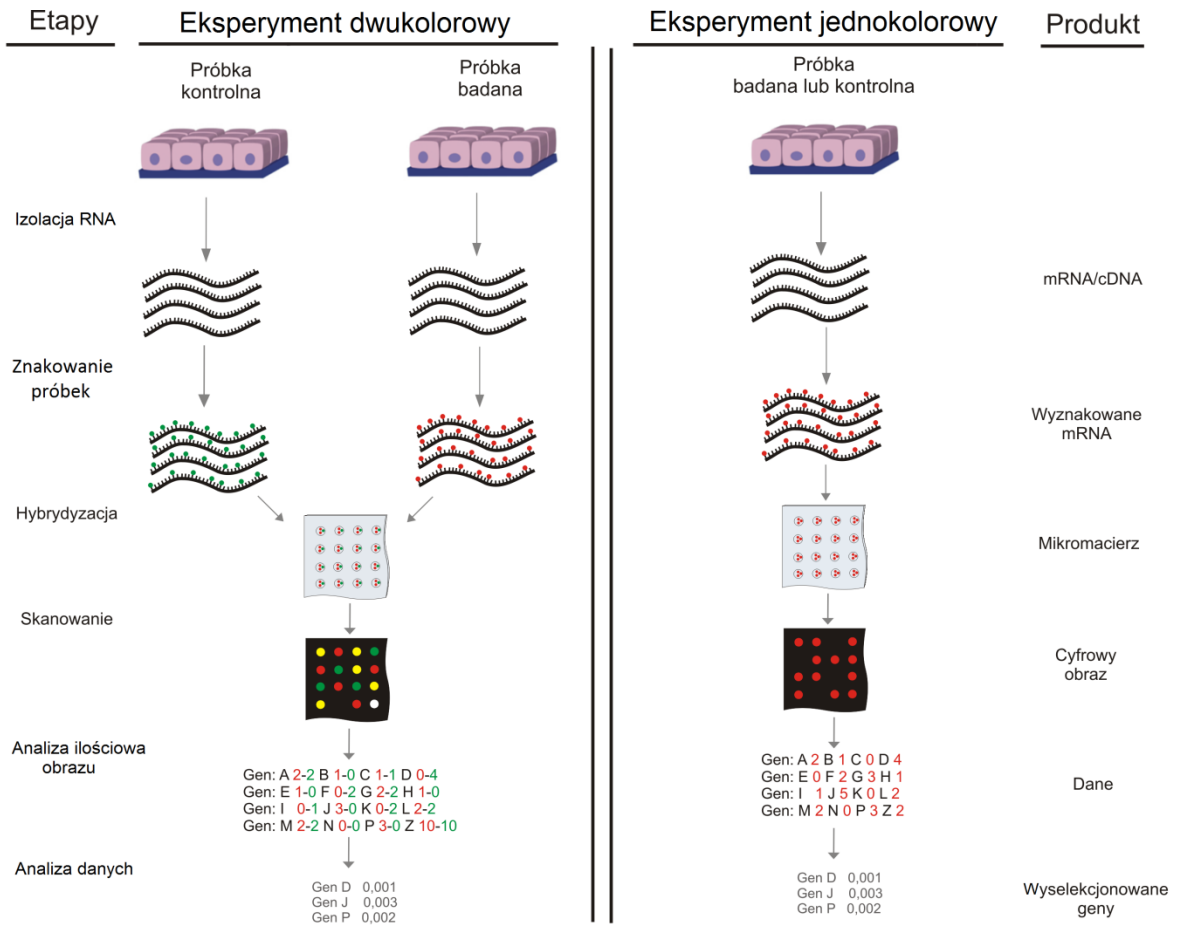
polimerazy T7. Otrzymane cRNA poddawane jest procesowi oczyszczania, a następnie znakowania.

Oba podejścia powielania materiału biologicznego mają nieco inny charakter i w różny sposób wpływają na wyniki końcowe i proces analizy danych (Jolly i wsp. 2005; Kainkaryam i wsp. 2010; Zhu i wsp. 2006; Kaposi-Novak i wsp. 2004). Dlatego też metoda powielania materiału biologicznego powinna być wykonana przed procesem znakowania próbki barwnikami fluorescencyjnymi, a jej wybór zależy od charakteru i celu eksperymentu (Coppola 2011).

II.3.2 Znakowanie próbek

Proces znakowania próbki barwnikami fluorescencyjnymi z reguły ma miejsce na etapie syntezy antysensownej nici cDNA lub cRNA, bezpośrednio z RNA lub z cDNA po procesie amplifikacji. Stosowanie znakowanego cDNA lub cRNA w reakcji hybrydyzacji często zależy od rodzaju eksperymentu i rodzaju użytej platformy. W praktyce funkcjonują dwa sposoby przyłączania barwników fluorescencyjnych do fragmentów kwasów nukleinowych: bezpośredni i pośredni. Metoda bezpośrednia polega na włączaniu na etapie syntezy antysensownej nici cDNA znakowanych fluorescencyjnie nukleotydów np. CTP-Cy5. Metoda pośrednia natomiast polega na wprowadzeniu do cząsteczki cDNA nukleotydów zawierających modyfikację typu amino-allylo, która stanowi podstawę i miejsce wiązania barwników fluorescencyjnych. W kolejnym etapie fluorofory przyłączane są do odpowiednich fragmentów cDNA. Niezwiązane cząsteczki barwników usuwane są za pomocą chromatografii kolumnowej lub wytrącania. Metoda pośrednia najczęściej stosowana jest do znakowania cDNA, natomiast metoda bezpośrednia do znakowania cRNA.

Eksperymenty z użyciem ekspresyjnych mikromacierzy DNA prowadzone są w oparciu o dwa schematy procesu znakowania: z wykorzystaniem jednego (eksperyment jednokolorowy) lub dwóch (eksperyment dwukolorowy) barwników fluorescencyjnych (Rysunek 3).

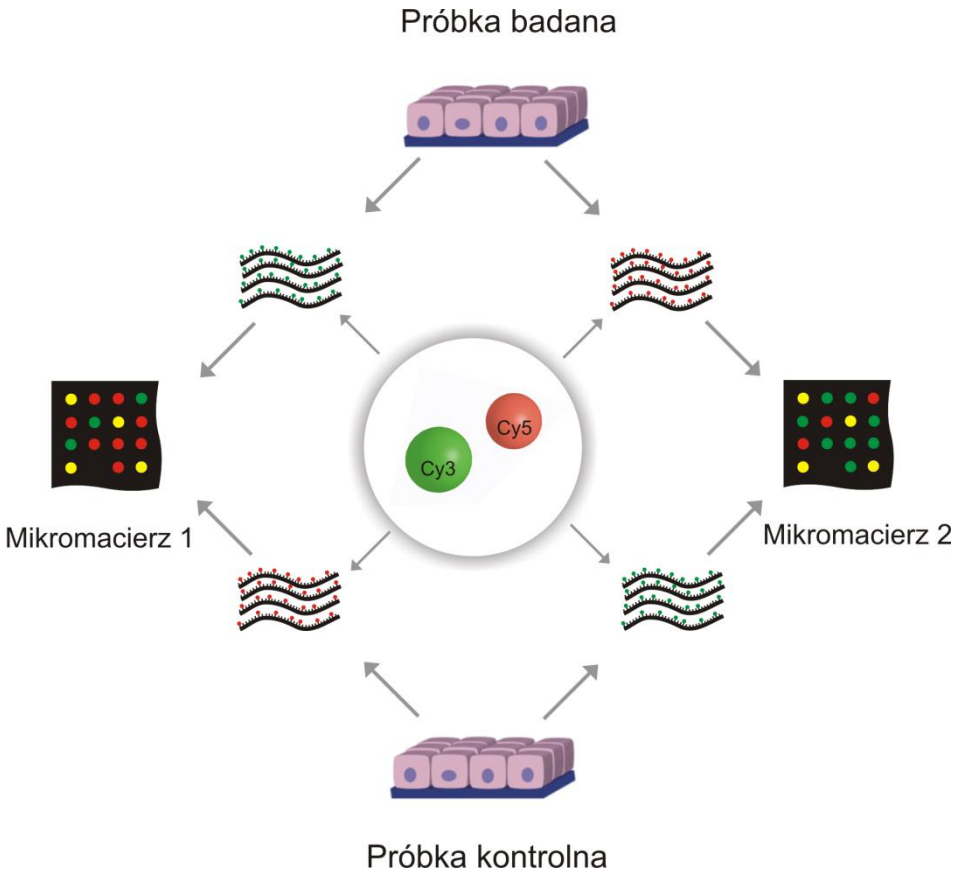


Rysunek 3. *Ogólny schemat eksperymentu z użyciem ekspresyjnych mikromacierzy DNA. Schemat uwzględnia podział na eksperymenty prowadzone z użyciem dwóch (eksperyment dwukolorowy) lub jednego (eksperyment jednokolorowy) barwnika fluorescencyjnego. W kolumnie „Etapy" przedstawione są poszczególne etapy wykonywane podczas realizacji eksperymentu z użyciem ekspresyjnych mikromacierzy DNA, natomiast w kolumnie „Produkt" przedstawione są wyniki poszczególnych etapów eksperymentu.*

Eksperyment jednokolorowy obejmuje znakowanie każdej z próbek tym samym barwnikiem fluorescencyjnym, np. Cy3 (Cyjanina 3). Następnie każda z próbek (badana i referencyjna) hybrydyzowana jest do odrębnej mikromacierzy, a ustalenie poziomu ekspresji genów w próbce badanej względem próbki referencyjnej obejmuje porównanie intensywności sygnałów fluorescencji pomiędzy mikromacierzami. W przypadku eksperymentu dwukolorowego próbki badane i próbki referencyjne znakowane są dwoma różnymi barwnikami fluorescencyjnymi (np. próbka referencyjna Cyjaniną 3-Cy3, a próbka badana Cyjaniną 5-Cy5). Eksperyment dwukolorowy obejmuje hybrydyzację mieszaniny próbek (próba referencyjna i badana) do jednej ekspresyjnej mikromacierzy DNA. Stosowanie dwóch barwników fluorescencyjnych o różnych parametrach wzbudzenia i emisji sygnału umożliwia niezależną ocenę poziomu ekspresji genów dla każdej z próbek. Selekcja genów potencjalnie różnicujących odbywa się w wyniku porównania intensywności sygnałów pomiędzy barwnikami. Eksperymenty dwukolorowe bazują na założeniu, że wpływ obu barwników na

charakterystykę hybrydyzacji jest jednakowy oraz, że stosunki ilościowe poszczególnych transkryptów pierwotnie istniejących w próbie badanej i referencyjnej są zachowane i pozostają niezależne od ilości RNA stosowanego podczas reakcji.

Stosowanie barwników fluorescencyjnych o różnych właściwościach chemicznych w ramach eksperymentu dwukolorowego wiąże się z wprowadzeniem dodatkowej zmienności pomiędzy próbkami badanymi, a referencyjnymi. W celu zminimalizowania udziału tej zmienności w przypadku eksperymentów dwukolorowych stosuje się system zamiany barwników (ang. *dye swap*). Strategia ta polega stworzeniu układu eksperymentalnego, gdzie dla każdej pary próbek badanej i referencyjnej tworzone są dwie mikromacierze: „podstawowa" i o odwróconym systemie znakowania próbek. Oznacza to, że jeżeli na „podstawowej" mikromacierzy próbka badana znakowana była np. barwnikiem Cy5, a próbka referencyjna np. barwnikiem Cy3 to mikromacierz o odwróconym systemie znakowania powinna zawierać próbkę badaną znakowaną barwnikiem Cy3 oraz próbkę referencyjną znakowaną barwnikiem Cy5 (Rysunek 4).



Rysunek 4. *Schemat obrazujący system zamiany barwników fluorescencyjnych na etapie znakowania próbek w ramach eksperymentu dwukolorowego.*

W literaturze często stosuje się uproszczenia oznaczeń poszczególnych barwników fluorescencyjnych, gdzie R (ang. *red*) oznacza czerwoną (np. Cy5), a G (ang. *green*) zieloną (np. Cy3) fluorescencję.

II.3.3 Hybrydyzacja

Po procesie znakowania próbek barwnikami fluorescencyjnymi, próbka (znakowana jednym barwnikiem fluorescencyjnym) lub mieszanina dwóch próbek (znakowana dwoma barwnikami fluorescencyjnymi) poddawana jest hybrydyzacji do mikromacierzy DNA. Warunki hybrydyzacji: siła jonowa roztworu, temperatura i stężenie próby, dostosowywane są do rodzaju prowadzonego eksperymentu i stosowanej platformy. Jako warunki hybrydyzacji, zwykle stosuje się temperatury od 42°C do 50°C i czas od kilku do kilkunastu godzin. Po każdej reakcji hybrydyzacji następuje odmycie niezwiązanych z sondami transkryptów za pomocą buforów o rosnącej sile odmywania.

II.3.4 Skanowanie

Podczas skanowania, światło lasera o odpowiedniej długości fali wzbudza fluorescencję wyznakowanych transkryptów związanych z sondami, a sygnał odczytywany jest przez detektor skanera. Intensywność sygnału jest wprost proporcjonalna do ilości cząsteczek transkryptu związanego przez daną sondę. Emitowany przez każdy barwnik sygnał fluorescencji zapisywany jest w formie obrazu, przechowywanego w postaci pliku graficznego (najczęściej w formacie TIFF).

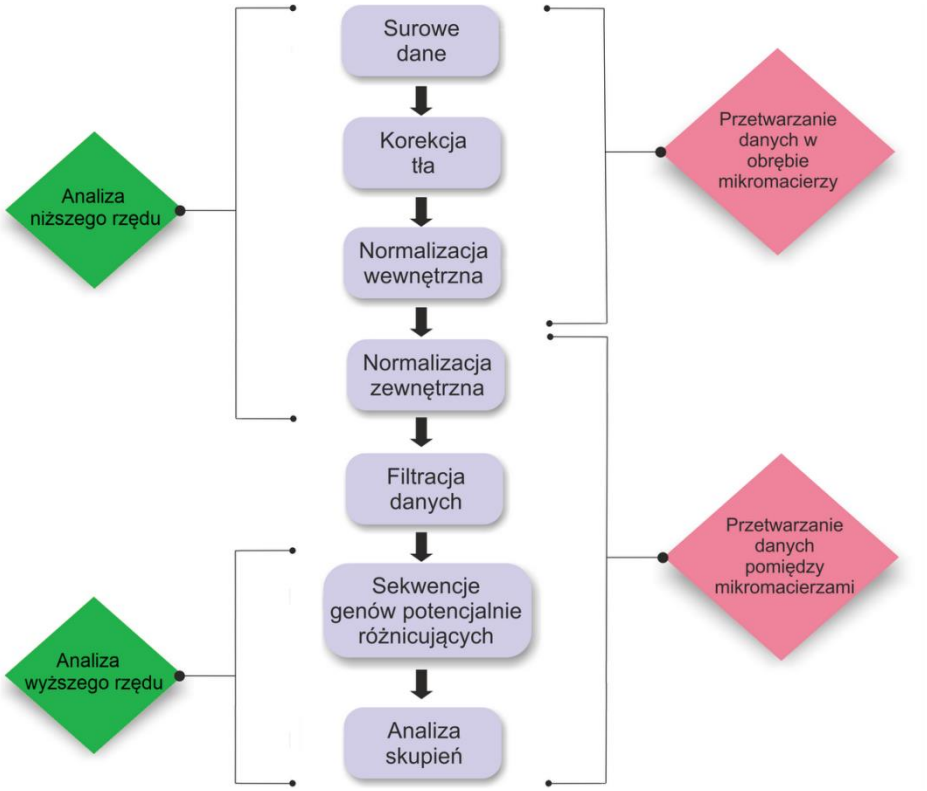
II.3.5 Analiza ilościowa obrazu

Uzyskane obrazy przetwarzane są na dane liczbowe dzięki zastosowaniu specjalistycznego oprogramowania do analizy ilościowej obrazu: np. GenePix, (Molecular Devices), ScanArray (PerkinElmer), TIGR-Spotfinder/TM4 (www.tigr.org). Wynikiem analizy ilościowej obrazu dla każdej mikromacierzy jest zestaw danych w postaci tabeli zawierającej wartości intensywności sygnału (eksperyment jednokolorowy) lub sygnałów (eksperyment dwukolorowy) fluorescencji dla każdej sondy oraz towarzyszące im wartości tła.

II.3.6 Analiza danych

Uzyskane w wyniku analizy ilościowej „surowe dane” poddawane są następnie analizie komputerowej. Proces analizy danych obejmuje zwykle dwa etapy: analizę niższego oraz wyższego rzędu (Rysunek 5). Analiza niższego rzędu ma na celu eliminację zmienności o podłożu technicznym (eksperymentalnym). Natomiast zadaniem analizy wyższego rzędu jest identyfikacja genów o specyficznej ekspresji w badanych warunkach oraz określenie występującej pomiędzy nimi zależności.

Proces analizy danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA stanowi główną część prezentowanej pracy doktorskiej. Stąd też wybrane etapy tego zagadnienia zostaną przedstawione szerzej w następnym rozdziale.



Rysunek 5. *Ogólny schemat analizy danych uzyskiwanych przy użyciu mikromacierzy DNA w badaniach analizy ekspresji genów.*

II.4 Elementy analizy danych

II.4.1 Analiza niższego rzędu

Analiza niższego rzędu zwana także wstępną analizą danych ma na celu przede wszystkim minimalizację zmienności pochodzenia technicznego, wynikającej z wieloetapowości eksperymentu. Czynnikami wprowadzającymi dodatkową zmienność w

układzie mogą być: proces drukowania mikromacierzy, proces znakowania i nierówna dystrybucja barwników fluorescencyjnych, a także przebieg reakcji hybrydyzacji, warunki płukania oraz wiele innych. Do etapów wstępnej analizy niższego rzędu należą głównie:

- korekcja tła
- normalizacja

Końcowym wynikiem wstępnej analizy danych dla każdej mikromacierzy jest tabela zawierająca znormalizowane wartości intensywności sygnału fluorescencji (eksperyment jednokolorowy) lub wartości M (eksperyment dwukolorowy) dla każdej sondy.

II.4.1.1 Korekcja tła

Analiza ilościowa obrazu, poza określeniem poziomu intensywności sygnału dla danego punktu na mikromacierzy, pozwala także na oszacowanie wartości otaczającego go szumu, tzw. tła. Korekcja tła jest jednym z alternatywnych etapów wstępnej analizy danych, umożliwiającym separację głównych wartości intensywności sygnału od wartości tła. Domyślna procedura korekcji tła w przypadku większości dostępnych algorytmów polega na prostej operacji odcięcia tła od głównej wartości intensywności sygnału dla danego punktu. Choć istnieje wiele metod eliminacji tła, właśnie ta prosta procedura korekcji tła jest szczególnie polecana przy analizie danych uzyskiwanych w ramach eksperymentów badania ekspresji genów (Ritchie i wsp. 2007). Z przyczyn technicznych zastosowanie korekcji tła nie zawsze jest korzystne i może skutkować zwiększeniem poziomu zmienności w układzie (Scharpf i wsp. 2007; Ritchie i wsp. 2007). Stąd też przed wykonaniem tego etapu analizy zalecana jest ocena konieczności jego stosowania, np. poprzez oszacowanie procentowego udziału tła w wartości intensywności głównej punktu oraz jego wpływu na rozkład tych wartości.

II.4.1.2 Normalizacja

Złożoność eksperymentu z użyciem mikromacierzy DNA sprzyja wprowadzaniu błędów systematycznych do układu. Głównym celem procesu normalizacji jest skorygowanie tych błędów przy zachowaniu informacji pochodzenia biologicznego. Istnieją dwa typy normalizacji: wewnętrzna i zewnętrzna.

II.4.1.2.1 Normalizacja wewnętrzna

Normalizacja wewnętrzna jest procesem normalizacji wartości intensywności sygnałów fluorescencji w obrębie danej mikromacierzy. Ten rodzaj normalizacji stosowany jest głównie do usuwania błędów systematycznych w zestawach danych otrzymanych w wyniku eksperymentów prowadzonych z użyciem dwóch barwników fluorescencyjnych. W celu lepszego zrozumienia mechanizmu funkcjonowania wewnętrznych metod normalizacji konieczne jest zdefiniowanie dwóch wartości: M i A, generowanych dla każdej sondy. Wartość M (1) jest to logarytm o podstawie 2 stosunku intensywności sygnału fluorescencyjnego próbki badanej (R) do kontrolnej (G). Natomiast, wartość A (2) stanowi średnią intensywność sygnału fluorescencyjnego próbki badanej (R) i kontrolnej (G) wyrażonych w skali logarytmicznej o podstawie 2.

$$M = \log_2 \frac{R}{G} \qquad (1)$$

$$A = \frac{1}{2}(\log_2 R + \log_2 G) \qquad (2)$$

Większość z opisywanych w literaturze metod normalizacji wewnętrznej bazuje na korekcji wartości M. Najprostszą formą normalizacji wartości intensywności sygnałów sond w obrębie mikromacierzy jest stosowanie poprawki w postaci średniej (lub mediany) wartości M ($\overline{M}$) liczonej dla zestawu genów nie wykazujących ekspresji różnicowej:

$$M_{\mathrm{k}} = M - c \qquad (3)$$

gdzie:

M_k- to skorygowana wartość M

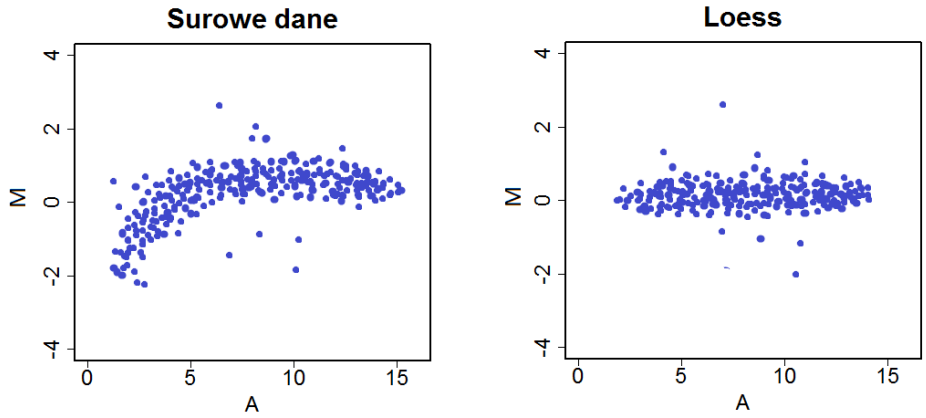
c-to stała określająca wartość różnicy pomiędzy kanałami, wyrażająca wartość $\overline{M}$ dla zestawu genów kontrolnych (nieulegających ekspresji różnicowej).

Przeprowadzona w ten sposób normalizacja danych nie jest odporna na efekty wynikające z różnic intensywności sygnałów pomiędzy danym rodzajem punktów, jak również i na efekty wynikające z różnej lokalizacji punktów na mikromacierzy. Aby zwiększyć dokładność procesu normalizacji możliwe jest zastosowanie metod normalizacji zakładających, iż wartość tych błędów (wartość stałej c) jest różna dla różnych punktów i w dużej mierze zależy od

wartości intensywności sygnałów fluorescencji (Rysunek 6). W takim przypadku wykreślana jest zależność c(A), gdzie wartość c odpowiada danej wartości A. Wartości M są korygowane na skutek eliminacji wartości wyznaczonych przez krzywą c(A):

$$M_{\mathrm{k}} = M - c(A) \qquad (4)$$

Wykres c(A) generowany jest za pomocą konkretnej metody wygładzania wykresu rozrzutu (ang. *robust scatter plot smoother*), np. ważonej lokalnej regresji liniowej-*loess* (ang. *local weighted regression*) (Yang i wsp. 2002).



Rysunek 6. *Dwa wykresy MA. W lewym panelu prezentowane są surowe dane, które wykazują trend zależny od intensywności sygnałów. W prawym panelu prezentowane są te same dane, ale po transformacji w wyniku zastosowania normalizacji wewnętrznej typu loess.*

Kolejnym sposobem poprawy jakości efektu normalizacji jest uodpornienie procedury normalizacji od efektów przestrzennych wynikających z różnej lokalizacji sond na mikromacierzy. Rozwiązaniem tej kwestii jest stosowanie innych krzywych c(A) do normalizacji różnych regionów mikromacierzy:

$$M_{\mathrm{k}} = M - \mathrm{c_i}(A) \qquad (5)$$

gdzie **i** oznacza indeks określony dla danego regionu mikromacierzy.

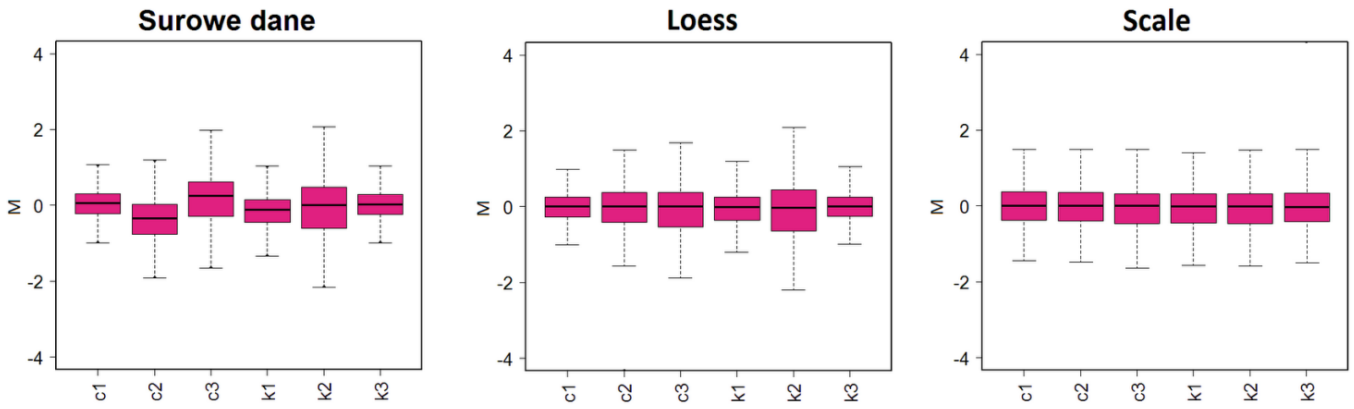
Takie podejście stosowane jest m.in. w metodzie normalizacji typu *print-tip loess*, gdzie inny rodzaj krzywej wykorzystywany jest do normalizacji różnych regionów mikromacierzy, odpowiadających grupom sond drukowanych danym rodzajem igły (ang. *print-tip groups*) (Smyth & Speed 2003). Normalizacja ta pozwala na eliminację zmienności systematycznej wynikającej z drukowania mikromacierzy różnymi igłami.

W wyniku działania normalizacji wewnętrznej średnia wartość M_k ($\overline{M_k}$) jest bliska 0.

II.4.1.2.2 Normalizacja zewnętrzna

Normalizacja zewnętrzna, tzw. normalizacja pomiędzy mikromacierzami powinna być stosowana w przypadku analiz, obejmujących przynajmniej dwie mikromacierze. Ma ona na celu zagwarantowanie tej samej skali dla pomiarów ekspresji pochodzących z różnych mikromacierzy oraz eliminację zmienności o podłożu technicznym. Zadaniem procesu normalizacji jest przekształcenie danych w taki sposób, aby wszystkie mikromacierze w ramach danej analizy charakteryzowały się podobnym rozkładem wartości. Dla zestawów danych otrzymanych w wyniku eksperymentów dwukolorowych ten etap analizy jest opcjonalny. Natomiast w przypadku eksperymentów wykonanych z użyciem jednego barwnika fluorescencyjnego normalizacja zewnętrzna stanowi podstawowy proces korygowania błędów systematycznych.

Najpopularniejszą procedurą normalizacji zewnętrznej dla danych uzyskanych w wyniku eksperymentów dwukolorowych jest normalizacja typu *scale* (Smyth & Speed 2003). Mechanizm działania tej metody obejmuje prosty proces skalowania wartości M_k (zwykle otrzymanych w wyniku normalizacji wewnętrznej), tak aby każda z nich wykazywała takie samo odchylenie od wartości $\overline{M_k}$ (Rysunek 7).



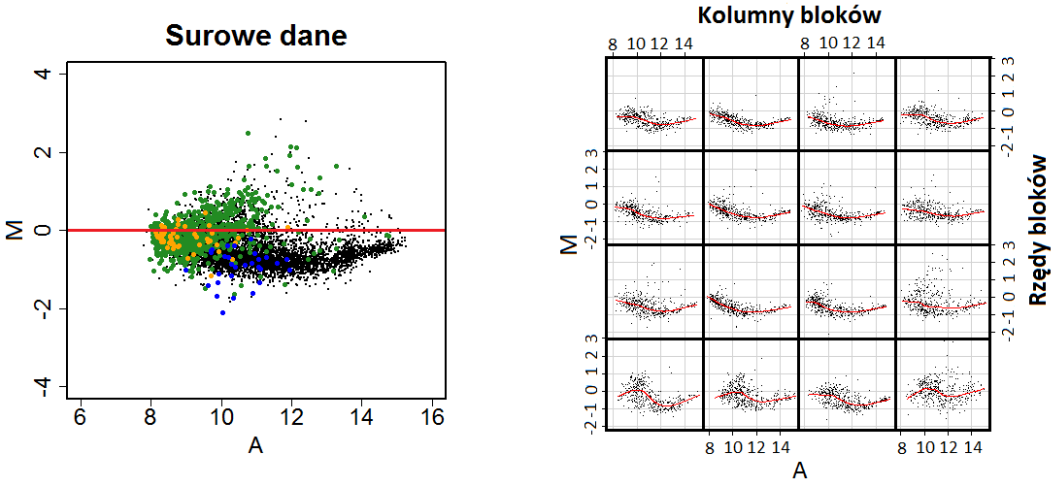
Rysunek 7. *Wykresy pudełkowe dla surowych danych, danych po normalizacji wewnętrznej typu loess (lewy panej) oraz po transformacji w wyniku normalizacji zewnętrznej typu scale. Na osi X prezentowane są rodzaje próbek stosowanych w ramach eskperymentu (c-próbka badana, k-próbka kontrolna), na osi Y prezentowane są wartości M dla każdego z typów próbek.*

W przypadku eksperymentów jednokolorowych powszechnie stosowaną metodą wyrównywania rozkładów wartości intensywności sygnałów jest normalizacja typu *quantile*. Mechanizm działania tej normalizacji polega na transformacji wartości intensywności sygnałów fluorescencji w oparciu o średnią wartość sygnału fluorescencji dla danych sond. Normalizacja metodą *quantile* polega na stworzeniu rankingu (ranking nr 1) wartości sygnału

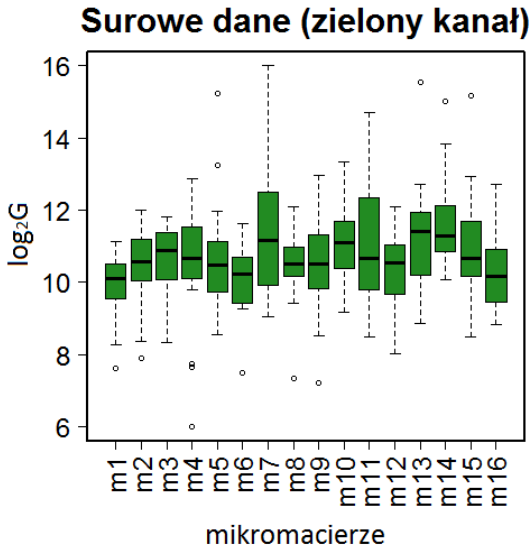
fluorescencji dla każdej z analizowanych mikromacierzy. W kolejnym etapie procesu normalizacji wartości intensywności sygnałów fluorescencji dla każdej z mikromacierzy sortowane są rosnąco. Następnie wartości sygnałów fluorescencji zajmujące te same pozycje po procesie sortowania są uśredniane pomiędzy mikromacierzami. Tworzony jest ranking uśrednionych wartości (ranking nr 2). Uśredniona wartość zajmująca daną pozycję w rankingu nr 2 podstawiana jest w miejsce tej samej pozycji z rankingu nr 1 dla każdej z mikromacierzy (Bolstad i wsp. 2003).

II.4.1.3 Ocena jakości w eksperymentach z użyciem mikromacierzy DNA

Ocena jakości danych (ang. *quality control*) powinna być prowadzona na każdym etapie analizy. Jest to szczególnie ważne na poziomie analizy niższego rzędu, w którym zestaw danych przechodzi najwięcej transformacji. Skuteczność poszczególnych etapów analizy danych jest najczęściej weryfikowana na podstawie wizualnej oceny tzw. wykresów diagnostycznych, w tym: wykresów MA (ang. *MA-plot*), wykresów pudełkowych (ang. *boxplot*) oraz wykresów typu *image plot* (Smyth & Speed 2003). Wykresy MA (Rysunek 8), jak i wykresy pudełkowe (Rysunek 9) stosowane są do analizy rozkładu intensywności sygnałów. Wykres pudełkowy pozwala na ocenę rozkładu wartości intensywności sygnału, a także na identyfikację artefaktów oraz ogólną ocenę analizowanego zestawu danych. Stosowanie wykresów MA w przypadku mikromacierzy drukowanych pozwala także na określenie poziomu zmienności pomiędzy blokami sond, tj. grupami sond drukowanymi innym rodzajem igły (ang. *print-tip groups*).

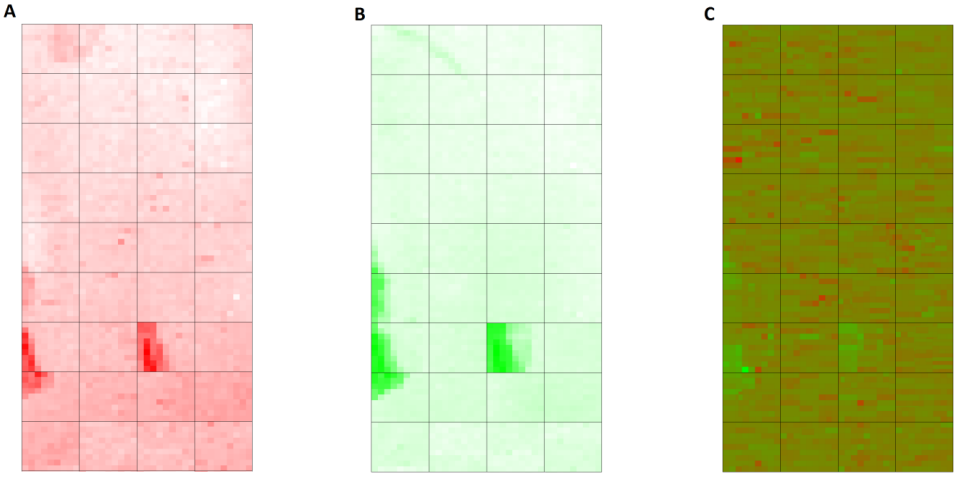


Rysunek 8. *Przykład wykresu MA. A. Wykres MA dla wszystkich sond ulokowanych na mikromacierzy. B. Zestaw wykresów MA z podziałem na poszczególne bloki sond, drukowanych danym rodzajem igły (ang. print-tip groups).*



Rysunek 9. *Wykres pudełkowy. A. Definicja poszczególnych elementów wykresu: zawartości bloku oraz jego granic. B. Przykładowy wykres pudełkowy dla zestawu mikromacierzy dedykowanych.*

Podstawowym celem stosowania wykresów typu *image plot* (Rysunek 10) jest ocena zmienności w obrębie wartości intensywności sygnałów fluorescencyjnych, jak i wartości tła, wynikającej z niespecyficznych efektów technicznych na podłożu.



Rysunek 10. *Przykład wykresów typu imageplot. A. Dla kanału czerwonego (próbka badana), B. Dla kanału zielonego (próbka kontrolna), C. Po złożeniu obrazów dla obu kanałów.*

II.4.2 Analiza wyższego rzędu

W przypadku eksperymentów z użyciem ekspresyjnych mikromacierzy DNA, podstawowym zadaniem analizy wyższego rzędu jest selekcja genów wykazujących znaczne różnice poziomów ekspresji pomiędzy dwoma lub większą liczbą grup badanych próbek. Często także celem analizy wyższego rzędu jest określenie zależności biologicznych pomiędzy badanymi genami i próbkami. Termin biologiczne zależności odnosi się tutaj

głównie do biomarkerów, genów wykazujących skorelowaną ekspresję (ang. *co-expressed genes*) oraz podobieństwem profilu ekspresji genów pomiędzy dwoma rodzajami próbek z danej grupy (np. podtypów danej choroby). Do głównych elementów analizy wyższego rzędu należą: etap selekcji genów różnicujących oraz proces analizy skupień.

II.4.2.1 Filtracja danych

Proces filtracji danych jest etapem pośrednim pomiędzy analizą niższego, a analizą wyższego rzędu. Etap ten powinien poprzedzać analizę wyższego rzędu. Filtracja danych jest często stosowaną praktyką, pozwalająca na minimalizację szumu w układzie. Ekspresyjna mikromacierz DNA, w zależności od rodzaju, może zawierać od kilkuset do kilkudziesięciu tysięcy genów, punktów lub sond. Jednakże jedynie niewielka część genów, spośród całkowitej liczby genów ulokowanych na ekspresyjnej mikromacierzy DNA, wykazuje faktyczną zmianę poziomu ekspresji w badanych warunkach. W praktyce zarządzanie tak dużymi zestawami danych może być niekiedy utrudnione oraz może obniżać siłę statystyczną eksperymentu. Stąd też powszechnie stosowanym podejściem jest redukcja zestawu danych w wyniku filtracji. Filtracja danych polega na eliminacji genów o bardzo niskim poziomie ekspresji. Dotyczy to zarówno genów dla których nie otrzymano kompletnej informacji (luki w zestawie danych) oraz tych charakteryzujących się bardzo niskimi wartościami ekspresji. Ponadto, proces filtracji danych może być także wykorzystywany w celu selekcji informacji na temat ekspresji genów pochodzących jedynie z danej grupy lub rodziny.

II.4.2.2 Selekcja genów różnicujących

Kluczowym elementem eksperymentów z użyciem ekspresyjnych mikromacierzy DNA jest identyfikacja genów, które wykazują znaczące różnice poziomów ekspresji pomiędzy dwoma lub większą liczbą grup próbek. Taki rodzaj analizy ma istotne znaczenie, ponieważ pozwala na otrzymanie charakterystyki próbek na poziomie molekularnym w badanych warunkach. Najprostszą formą selekcji genów różnicujących jest stworzenie listy rankingowej w oparciu o wartości określające poziom ekspresji badanych genów oraz ustalenie progów odcięcia powyżej których zmiany ekspresji mają istotny charakter. Powszechnie stosowaną metodą jest klasyfikacja względem wartości intensywności sygnału lub wartości M dla każdego punktu (posortowanych malejąco). Za geny znaczące z punktu widzenia prowadzonych badań uznawane są te wykazujące co najmniej dwukrotny wzrost lub spadek poziomu ekspresji w stosunku do próbek kontrolnych. Podejście to jednak nie jest

odporne na błędy I i II rodzaju. Błędy I rodzaju skutkują otrzymaniem wyników fałszywie pozytywnych (ang. *false positives*), poprzez klasyfikację jako różnicujące genów, które w rzeczywistości charakteryzują się niezmienionym poziomem ekspresji. Błędy II rodzaju natomiast oznaczają wyniki fałszywie negatywne (ang. *false negatives*). Rzetelna metoda identyfikacji genów różnicujących powinna uwzględniać zarówno wartości istotności statystycznych dla poszczególnych genów, jak i liczbę analizowanych genów (testowanych hipotez). Najczęściej stosowanym testem statystycznym na etapie selekcji genów różnicujących jest test t, zwłaszcza w przypadku zestawów danych obejmujących dwie grupy próbek (Trevino i wsp. 2007). Test t wykorzystywany jest zarówno w podstawowej wersji, jak i modyfikowanej np. moderowany test t (ang. *moderated t-test*)(Smyth 2005)(Smyth i wsp. 2003). Ze względu na ograniczenia stosowania testu t (Rensink & Hazen 2006), do identyfikacji genów różnicujących wykorzystywane są także inne metody statystyczne np. wieloetapowe procedury testowania (ang. *Multiple Testing Procedures*, MTP).

II.4.2.2 Analiza skupień

Proces analizy wyższego rzędu oprócz identyfikacji genów różnicujących ma na celu poznanie zależności występujących w badanym zestawie danych. Najprostszą i zarazem najpopularniejszą metodą poszukiwania zależności w zestawach danych uzyskiwanych za pomocą ekspresyjnych mikromacierzy DNA jest analiza skupień, zwana także klasyfikacją lub grupowaniem. Pozwala ona na łączenie elementów danego zbioru w grupy o wspólnym profilu w tzw. klastry lub skupiska. Reprezentację grupy stanowi wówczas pojedynczy profil, który jest uśrednieniem wszystkich elementów skupiska lub jednym z jego elementów, tzw. medoidem lub centroidem. W eksperymentach z użyciem ekspresyjnych mikromacierzy DNA analiza skupień może być prowadzona zarówno na poziomie genów w obrębie pojedynczej próbki, jak również na poziomie próbek danej grupy lub pomiędzy grupami. Podstawą klasyfikacji tego rodzaju danych jest nie tylko poziom ekspresji, ale również i inne indywidualne cechy próbek, np. wiek lub płeć pacjentów. Spośród dostępnych metod grupowania można wyróżnić dwa rodzaje metod: nadzorowane i nienadzorowane.

II.4.2.1 Metody nadzorowane

Nadzorowane metody analizy skupień wymagają definiowania liczby skupisk jaka ma powstać z danego zbioru elementów. Docelowa liczba klastrów najczęściej określana jest na podstawie hipotezy własnej lub oczekiwanego wyniku. Nadzorowane metody analizy skupień

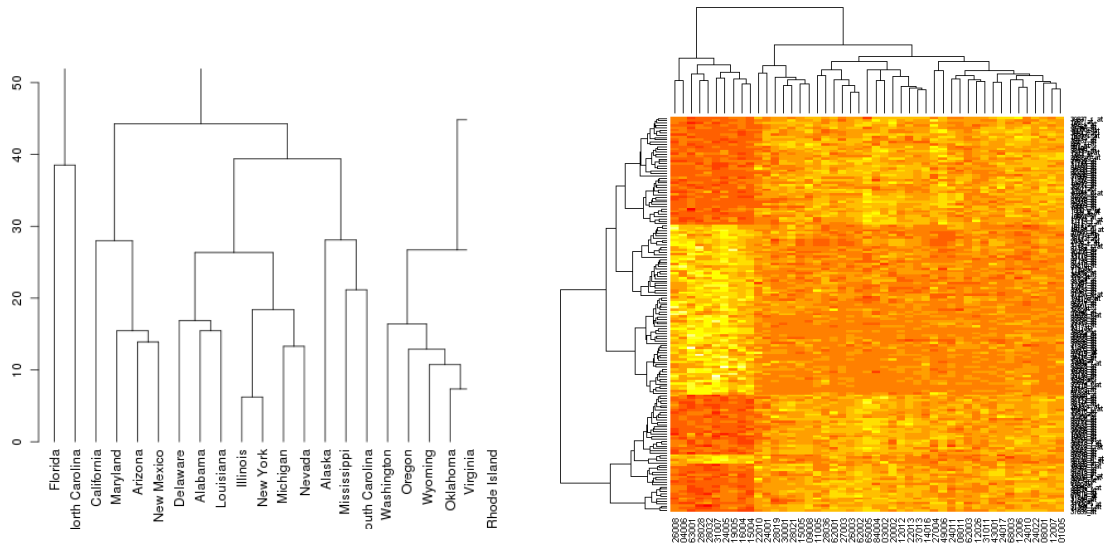
mogą działać także w oparciu o klasyfikatory pozwalające na przypisanie genów lub próbek do predefiniowanych klas. Ten rodzaj klasyfikacji w analizie danych uzyskanych z użyciem ekspresyjnych mikromacierzy DNA najczęściej wykorzystywany jest do identyfikacji markerów molekularnych, tzw. biomarkerów, które są wskaźnikiem danego stanu biologicznego, choroby, a także jej podtypu w przypadku chorób heterogenicznych. Fundamentalną różnicą pomiędzy identyfikacją genów różnicujących, a identyfikacją zestawu genów do diagnozowania lub prognozowania danej choroby jest fakt, iż biomarkery muszą posiadać wartość predykcyjną. Selekcja biomarkerów wymaga stworzenia na podstawie unikalnego zestawu genów sygnatury umożliwiającej identyfikację danego stanu biologicznego. W tym celu często stosowany jest klasyfikator przypisujący próbkę do danej grupy lub kategorii. Przykładowo, klasyfikatorem do identyfikacji cukrzycy jest poziom cukru w surowicy. W statystyce ten rodzaj klasyfikatora określany jest jako jednoczynnikowy. Oznacza to, iż identyfikację danego stanu biologicznego odbywa się na podstawie jednej zmiennej (poziom cukru). Niemniej jednak dla danych uzyskiwanych za pomocą ekspresyjnych mikromacierzy DNA powszechne jest otrzymanie długiej listy genów, które mogą być charakterystyczne dla danego stanu biologicznego. W przypadku analiz wielogenowych stosowane są klasyfikatory wieloczynnikowe, zwiększające stabilność klasyfikacji. Przykładowo, ryzyko wystąpienia danej choroby określane jest na podstawie poziomów ekspresji kilku lub kilkunastu wybranych genów z których każdy stanowi klasyfikator. Przykładem algorytmu do nadzorowanej klasyfikacji jest np. PAM (ang. *partioning around medoids*).

II.4.2.2 Metody hierarchiczne

Hierarchiczne metody grupowania w przypadku danych uzyskiwanych z użyciem mikromacierzy DNA są stosowane do identyfikacji genów o skorelowanej ekspresji (ang. *co-expressed genes*) i próbek wykazujących podobny profil ekspresji. Geny wykazujące skorelowaną ekspresję mogą być regulowane przez te same czynniki transkrypcyjne lub posiadać te same funkcje, np. wchodzić w skład tych samych szlaków metabolicznych lub sygnałowych. Identyfikacja takich genów może stanowić źródło odkryć nowych połączeń biologicznych pomiędzy genami, a także nowych cząsteczek o kluczowym znaczeniu z punktu terapii (ang. *potential clinical targets*). Podstawowym celem grupowania hierarchicznego w eksperymentach z użyciem ekspresyjnych mikromacierzy DNA jest oszacowanie podobieństwa pomiędzy próbkami biologicznymi w oparciu o profil ekspresji

genów. Tego rodzaju analiza ma na celu weryfikację, czy próbki o podobnych właściwościach biologicznych wykazują wspólne cechy na poziomie molekularnym. Niekiedy różnice w profilu ekspresji genów odzwierciedlają heterogenność choroby danego typu i stanowią podstawę identyfikacji nowych podtypów danej choroby. Grupowanie hierarchiczne może być wykorzystywane do identyfikacji próbek o nieznanej dotychczas klasyfikacji.

Podstawowa koncepcja metod hierarchicznej analizy skupień polega na konstrukcji klastrów poprzez stopniowe dodawanie jednego z elementów (genu, próbki lub mniejszego klastra). W ten sposób elementy zestawu danych wykazujące największe podobieństwo (względem danej cechy) są dodawane we wcześniejszej fazie do małych klastrów, a elementy wykazujące mniejsze podobieństwo do później tworzonych klastrów. Alternatywą dla opisywanej strategii grupowania hierarchicznego jest podejście polegające na podziale większych skupisk na mniejsze. Podobieństwo pomiędzy badanymi elementami szacowane jest za pomocą miary odległości między nimi. Wynikiem grupowania hierarchicznego jest drzewo klasyfikacji (ang. *dendrogram*) o takim ułożeniu, aby elementy wykazujące największe podobieństwo położone były blisko siebie (Rysunek 11). Drzewa klasyfikacji często prezentowane są w parze z dwuwymiarową mapą cieplną (ang. *heatmap*), na której gdzie poziom ekspresji genów w danej próbce przedstawiony jest za pomocą intensywności kolorów (Rysunek 11).



Rysunek 11. *Przykładowe wyniki analizy skupień. Lewy panel: Drzewo klasyfikacji, jako wynik hierarchicznej metody skupień. Prawy panel: Mapa cieplna.*

II.5 Programy do analizy danych

W miarę rozwoju technologii ekspresyjnych mikromacierzy DNA, wzrosła także zdolność do szybkiego i wydajnego prowadzania wysokoprzepustowych analiz (ang. *high-throughput*) na poziomie molekularnym. Mikromacierze DNA, które przed laty obejmowały jedynie niewielką liczbę sond, obecnie posiadają ich setki lub tysiące. Stąd też wyzwaniem na etapie stosowania ekspresyjnych mikromacierzy DNA nie jest sam eksperyment, ale etap zarządzania i przetwarzania danych w celu uzyskania znaczących statystycznie i biologicznie wyników. Problem analizy danych nie tyle wynika z ich rozmiaru, co raczej ze struktury. Przetwarzanie tego rodzaju danych często wymaga odpowiednich narzędzi. Obecnie dostępnych jest wiele programów umożliwiających analizę danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA. Każdy z nich ma swoje szczególne cechy, jednak największą popularnością cieszą się programy działające na zasadzie otwartej licencji (ang. *open source*). Wynika to głównie z ich dostępności oraz jawności kodu źródłowego. Dostęp do kodu źródłowego pozwala użytkownikom na modyfikację oprogramowania i dopasowanie go do specyficznych potrzeb. Obecnie spośród tego rodzaju programów, największą popularnością cieszą się programy: (I) R\Bioconductor (www.bioconductor.org) (R Development Core Team 2008; Gentleman i wsp. 2004), (II) program TM4 (Saeed i wsp. 2006; Saeed i wsp. 2003) oraz (III) BASE (ang. *BioArray Software Environment*) (Vallon-Christersson i wsp. 2009).

II.5.1 R/Bioconductor: narzędzie do statystycznej analizy danych

Projekt Bioconductor (Gentleman i wsp. 2004) jest ogólnodostępnym repozytorium dedykowanym biologii obliczeniowej. Aktualnie w skład zespołu pracującego nad jego stałym rozwojem wchodzi 24 najwyższej klasy specjalistów z dziedzin: biologii, bioinformatyki, statystyki i informatyki. Głównym celem projektu Bioconductor jest zapewnienie wysokiej jakości infrastruktury oraz narzędzi do analizy danych genomowych m.in. danych uzyskiwanych z użyciem mikromacierzy DNA, wyników sekwencjonowania drugiej generacji, analizy SNP, czy genotypowania CNV. Narzędzia tworzone w ramach projektu Bioconductor występują w postaci pakietów, tzn. bibliotek funkcji do analizy określonego rodzaju danych genomowych. Aktualnie w skład repozytorium Biocondcutor wchodzi 610 pakietów. Podstawowym systemem w oparciu o który funkcjonuje repozytorium Bioconductor jest środowisko do zaawansowanych analiz statystycznych- *R* (R Development Core Team 2008). Bioconductor jest w pełni komplementarny z podstawowym systemem

pakietów *R*, tzw. pakietów CRAN, co pozwala wykorzystywać funkcje zdeponowane w ramach tych pakietów w trakcie analizy z użyciem funkcji z pakietów Bioconductor. Dużą zaletą tego oprogramowania jest nie tylko jego rzetelność (każdy pakiet posiada krótki opis zawartych w nim funkcji i potencjalnych możliwości ich wykorzystania) oraz stały rozwój (przynajmniej 2 aktualizacje rocznie), ale możliwość uczestnictwa w rozwoju projektu (tworzenie nowych funkcji, pakietów oraz dokumentacji).

R\Bioconductor umożliwia analizę danych uzyskiwanych za pomocą wszystkich komercyjnie dostępnych mikromacierzy DNA (platform). W tym także tych pochodzących z mniejszych, tzw. dedykowanych mikromacierzy DNA (ang. *custom microarrays, boutique microarrays lub homemade microarrays*). Do przetwarzania danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA przeznaczonych jest kilkadziesiąt pakietów (blisko 60). Dzięki temu każdy użytkownik ma możliwość stworzenia specyficznego scenariusza analizy danych, dopasowanego do potrzeb eksperymentu oraz struktury analizowanych danych. R/Bioconductor umożliwia analizę danych uzyskanych w ramach eksperymentów dwu- i jednokolorowych. Głównym ograniczeniem środowiska R/Bioconductor jest brak interfejsu graficznego (ang. *graphical user interface*, GUI) i wykonywanie poleceń z linii komend, co wymusza konieczność posiadania przez użytkownika przynajmniej podstawowej wiedzy programistycznej.

II.5.2 TM4: oprogramowanie do analizy ekspresji genów

TM4 jest aplikacją stworzoną za pomocą języka Java i dzięki temu posiada przyjazny użytkownikowi interfejs graficzny (GUI). Struktura TM4 obejmuje 4 moduły: MADAM, TIGR Spotfinder, MIDAS, MeV oraz bazę danych MySQL. Każdy z tych modułów jest wyposażony w indywidualne cechy i może być stosowany niezależnie. Oprogramowanie TM4 powstało głównie z myślą o przetwarzaniu danych dwukolorowych, jednak z powodzeniem może być także stosowane do analizy eksperymentów jednokolorowych.

MADAM

Moduł MADAM (ang. *Microarray Data Manager*) ułatwia użytkownikowi wprowadzenie danych do relacyjnej bazy danych i prowadzi go przez cały proces analizy. MADAM korzystając z informacji na temat eksperymentu oferuje użytkownikowi prosty sposób uproszczenia procesu analizy, oferując pomoc w wyborze parametrów i interpretacji wyników.

TIGR Spotfinder

Aplikacja TIGR Spotfinder służy do szybkiej, wspomaganej komputerowo analizie jakościowej obrazu. Umożliwia ona odczytywanie sparowanych (pochodzących z eksperymentów dwukolorowych) 16 lub 8 bitowych obrazów w formacie TIFF. TIGR Spotfinder jest kompatybilny z większością dostępnych na rynku skanerów, a półautomatyczna konstrukcja siatki pozwala na identyfikację obszarów szkiełka, gdzie spodziewane są punkty. Wyniki analizy zapisywane są w formacie pliku (.tav) rozpoznawanym przez MIDAS lub eksportowane do bazy danych. Jako jedyny z modułów TM4, TIGR Spotfinder został stworzony w C++. Ponadto, TIGR Spotfinder jest jedynym z nielicznych ogólnodostępnych programów do analizy ilościowej danych (Saeed i wsp. 2003).

MIDAS

Moduł MIDAS (ang. *Microarray Data Analysis System*) umożliwia analizę niższego rzędu. Pozwala on na przeprowadzenie procesu normalizacji oraz filtracji danych, która ma na celu wyeliminowanie z zestawu danych elementów o niskiej jakości.

MeV

Aplikacja MeV (ang. *MultiExperiment Viewer*) jest najlepiej przygotowanym i najczęściej aktualizowanym modułem, umożliwiającym prowadzenie analizy wyższego rzędu. MeV za pomocą łatwego w obsłudze graficznego interfejsu daje użytkownikowi dostęp do szerokiego spektrum algorytmów włączając mi.n: analizę skupień k-średnich, grupowanie hierarchiczne, test t, SAM (ang. *Significance Analysis of Microarrays*), analizę głównych składowych (ang. *Principal Component Analysis*). Moduł ten wykazuje wysoki stopień kompatybilności z R/Bioconductor, zwiększając tym samym ilość dostępnych metod analizy.

Architektura blokowa oraz kompatybilność z innymi ogólnodostępnymi programami sprawia, iż TM4 jest elastycznym i łatwym w obsłudze oprogramowaniem do analizy danych.

II.5.3 BASE

BASE (ang. *Bioarray Software Enviroment*) jest oprogramowaniem dostępnym w postaci strony internetowej (ang. *Web-accesible system*). Takie rozwiązanie nie wymaga od użytkowników lokalnej instalacji i regularnej aktualizacji oprogramowania oraz daje dostęp do większej mocy obliczeniowej w postaci zewnętrznych serwerów. Program ten umożliwia

analizę danych pochodzących z różnych platform oraz analizę jedno- i dwukolorowych zestawów danych. Programy do przetwarzania danych uzyskiwanych za pomocą ekspresyjnych mikromacierzy DNA podlegają stałym modyfikacjom (Mehta & Rani 2011). Architektura tego oprogramowania oparta jest na systemie wtyczek (ang. *plug-in*), co pozwala szybkie dodawanie nowych modułów bez zbędnej ingerencji w rdzeń programu. BASE aktualnie wyposażony jest w trzy moduły do analizy danych umożliwiające: normalizację, wielowymiarowe skalowanie danych (w celu otrzymaniu danych w postaci dwu lub trójwymiarowych reprezentacji) oraz ich wizualizację.

II.6 Standardy jakości dla eksperymentów z użyciem mikromacierzy DNA

Specyficzne cechy dostępnych platform (np. długość sond) oraz programów do analizy danych skutkują obniżeniem powtarzalności wyników analiz prowadzonych z użyciem ekspresyjnych mikromacierzy DNA. W ciągu kilku ostatnich lat powołano szereg międzynarodowych konsorcjów oraz grup badawczych, których celem było zwiększenie jakości i powtarzalności uzyskanych rezultatów, jak również zwiększenie ich dostępności poprzez usprawnienie procesu deponowania tych danych. Jednym z takich konsorcjów było MAQC (MicroArray Quality Control), którego głównym zadaniem było ustalenie narzędzi kontroli jakości, które pozwalały na eliminację podstawowych błędów proceduralnych na etapie projektowania i prowadzenia (Shi i wsp. 2006; Shi i wsp. 2010; Patterson i wsp. 2006). Otrzymane przez MAQC wyniki pozwoliły na ocenę wydajności ekspresyjnych mikromacierzy DNA oraz opracowanie wytycznych dotyczących metod analizy uzyskiwanych danych. Kluczowe wyniki uzyskała także grupa NIST (The National Institutes of Standards and Technology), która skoncentrowała się na praktycznych problemach wynikających ze stosowania mikromacierzy DNA jako urządzeń pomiarowych, tj. różnic człuści, poziomu tła oraz zmienności w obrębie sygnałów dla różnych platform. Grupa NIST prowadziła także testy najczęściej wykorzystywanych skanerów do mikromacierzy DNA w celu opracowania metody walidacji i kalibracji skanerów pozwalającej na uzyskanie tych samych wyników niezależnie od rodzaju zastosowanego urządzenia. Wynikiem pracy ośrodków zrzeszonych w ramach konsorcjum ERCC (External RNA Controls Consortium) był zestaw dobrze scharakteryzowanych kontroli negatywnych, umożliwiających walidację eksperymentów prowadzonych z użyciem ekspresyjnych mikromacierzy DNA (Baker i wsp. 2005). Największy wkład w usprawnienie procesu deponowania i wymiany danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA miała grupa MGED

(Microarray Gene Expression Database Society). Dodatkowym sukcesem tej grupy było zdefiniowanie standardów MIAME (ang. *Minimum Information About a Microarray Experiment*), tj. podstawowych informacji jakie powinien zawierać manuskrypt dotyczący wyników badań prowadzonych z użyciem ekspresyjnych mikromacierzy DNA przed jego publikacją (Brazma i wsp. 2001). Grupa MGED, podobnie jak MAQC była także zaangażowana w opracowanie wytycznych dotyczących metod analizy danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA.

W ramach standardów jakości zdefiniowano szereg wytycznych dotyczących projektowania oraz prowadzenia eksperymentów z użyciem ekspresyjnych mikromacierzy DNA, jak również sposobu publikowania oraz deponowania otrzymanych wyników. Zgodnie z obowiązującymi standardami jakości, prawidłowo zaprojektowany eksperyment obejmuje:

- Rzetelny i starannie przygotowany opis eksperymentu, tak aby dany eksperyment mógł być powtórzony w innym laboratorium.
- Założenia eksperymentu sformułowane w sposób umożliwiający formułowanie dalszych hipotez badawczych.
- Odpowiednią ilość próbek, a także powtórzeń technicznych i biologicznych, w stosunku do rodzaju oraz skali prowadzonych badań.
- Możliwość wykonania kompleksowej analizy danych za pomocą dowolnie wybranego do tego celu oprogramowania.

II.7 Niestandardowe zestawy danych uzyskiwane przy użyciu mikromacierzy DNA

Wszystkie dane pochodzące z eksperymentów z użyciem ekspresyjnych mikromacierzy DNA dla których standardy jakości nie są spełnione, uznawane są za dane niestandardowe. Istnieje szereg czynników nadających zestawom danych niestandardowy charakter, stąd też analiza tego rodzaju danych jest specyficzna i wymaga indywidualnego podejścia do każdego zestawu danych. W praktyce najczęstszymi przykładami niestandardowych danych są dane uzyskiwane z użyciem dedykowanych mikromacieczy DNA.

II.7.1 Dane uzyskiwane z użyciem dedykowanych mikromacierzy DNA

Niestandardowy charakter danych uzyskiwanych z użyciem dedykowanych mikromacierzy DNA wynika z faktu, iż ten rodzaj mikromacierzy służy do badania

konkretnych procesów biologicznych lub chorobotwórczych, a nie ogólnych zmian ekspresji genów w obrębie transkryptomu. W związku z tym, iż dedykowane mikromacierze DNA zawierają na swojej powierzchni głównie sondy dla genów potencjalnie zaangażowanych w badany proces, ilość genów różnicujących dla tego rodzaju mikromacierzy jest znacznie większa niż dla mikromacierzy o wysokiej gęstości (<10%) i może sięgać nawet 50% (Oshlack i wsp. 2007; Campanaro i wsp. 2002; Mcilroy i wsp. 2005; Ferrarini i wsp. 2008; Baron i wsp. 2011). Co więcej, obecność tak dużej liczby genów różnicujących sprawia, iż rozkład intensywności sygnałów dla poszczególnych genów nie jest normalny i koncentruje się jedynie w określonych przedziałach zakresu intensywności. W przypadku danych uzyskanych z użyciem dedykowanych mikromacierzy DNA bardzo często zdarza się, iż równowaga pomiędzy genami ulegającymi podwyższonej lub obniżonej ekspresji jest przesunięta w jednym kierunku. Problem ten dotyczy zwłaszcza badań związanych z procesem nowotworzenia (Pelz i wsp. 2008). Ponadto, dedykowane mikromacierze DNA najczęściej otrzymywane są w wyniku drukowania mikromacierzy, co powoduje, iż mogą one posiadać szereg specyficznych cech nadanych im przez eksperymentatora. Wszystkie te cechy dedykowanych mikromacierzy DNA w znaczący sposób utrudniają proces analizy danych (Wyniki, Rozdział V.I oraz Rozdział V.III).

II.8 Metody sekwencjonowania drugiej generacji, a ekspresyjne mikromacierze DNA

Alternatywna metoda badania ekspresji genów, intensywnie (dynamicznie) rozwijana w ostatnich latach, wykorzystuje technologię sekwencjonowania drugiej generacji NGS (ang. *next-generation sequencing*), zwaną również sekwencjonowaniem nowej generacji, głębokim, masowym lub wysokoprzepustowym sekwencjonowaniem (ang. *high-throughput sequencing*). Terminy te odnoszą się do wszystkich metod sekwencjonowania nowszych niż automatyczna metoda Sangera (metoda pierwszej generacji). Technologia NGS obejmuje grupę metod umożliwiających generowanie jednorazowo ogromnej liczby nakładających się na siebie, zwykle krótkich, odczytów sekwencji DNA. Odczyty te powstają w wyniku syntezy nici komplementarnej do matrycy DNA, wcześniej losowo pofragmentowanej i poddanej liniowej amplifikacji. Głębokość pokrycia badanej sekwencji zależy od liczby generowanych odczytów. Sekwencjonowanie drugiej generacji daje możliwość uzyskiwania w pojedynczym eksperymencie sekwencji całych genomów czy transkryptomów. Badaniu ekspresji genów dedykowana jest metoda zwana RNA-seq (ang. *RNA sequencing*), choć sekwencjonowaniu nie poddaje się bezpośrednio RNA, lecz przepisane w procesie odwrotnej transkrypcji cDNA Pofragmentowane cDNA wyposaża się w uniwersalne adaptery, amplifikuje, a następnie

sekwencjonuje przez syntezę drugiej nici, z dodatkiem fluorescencyjnie znakowanych nukleotydów. Kolejnym krokiem jest mapowanie odczytów do sekwencji genomowej w celu uzyskania informacji zarówno o strukturze transkryptomu jak i poziomie ekspresji poszczególnych genów. Poziom ekspresji genów określany jest na podstawie liczby odczytów zmapowanych do danego genu lub transkryptu.

Technologia RNA-Seq coraz częściej zastępuje mikromacierze DNA (Bloom i wsp. 2009; S. Liu i wsp. 2011; Agarwal i wsp. 2010; Malone & Oliver 2011). Przyczyną takiego stanu rzeczy jest wszechstronność tej technologii, dzięki której oprócz analizy ekspresji genów możliwa jest także analiza niekodujących RNA oraz badanie procesu alternatywnego składania transkryptów (ang. *alternative splicing*). Główną zaletą RNA-seq jest fakt, iż jej zastosowanie nie jest ograniczone jedynie do badania transkryptów, które odpowiadają znanej sekwencji genomowej, jak to jest w przypadku mikromacierzy. Nie oznacza to jednak, iż technologia RNA-Seq może być swobodnie stosowana w przypadku organizmów o nieznanej lub fragmentarycznie poznanej sekwencji genomowej. Kompletna sekwencja genomowa wymagana jest bowiem na etapie analizy danych - w procesie mapowania odczytów, który stanowi podstawę dla ilościowej analizy transkryptów. Dodatkowym atutem RNA-Seq jest duży zakres dynamiki (ang. *dynamic range*), który pozwala na detekcję transkryptów wykazujących zmianę ekspresji nawet o kilka tysięcy razy (Z. Wang i wsp. 2009). Choć w przypadku mikromacierzy DNA zakres dynamiki jest znacznie mniejszy i umożliwia detekcję zmian ekspresji jedynie do kilkuset razy, większym problemem jest analiza genów o obniżonej aktywności transkrypcyjnej, jak również identyfikacja nowych genów. Metody analizy danych z RNA-Seq też nie oferują całkowitego rozwiązania tych kwestii. Identyfikacja nowych transkryptów za pomocą RNA-Seq nie zawsze jest możliwa nawet przy bardzo dużej głębokości sekwencjonowania analizowanych próbek (50-80 milionów) (Graveley i wsp. 2011). Powodem dla którego technologia RNA-Seq staje się coraz bardziej popularna jest także stale obniżający się koszt pojedynczego eksperymentu, który dzięki możliwości sekwencjonowania wielu próbek jednocześnie aktualnie jest już porównywalny z kosztem eksperymentu prowadzonego z użyciem mikromacierzy DNA. Poważnym ograniczeniem RNA-Seq jest natomiast proces analizy danych. Trudność przetwarzania danych wynika z ich formatu i rozmiaru, który dodatkowo zwiększa się w trakcie analizy ze względu na konieczność zachowywania wyników poszczególnych etapów. Kluczowym aspektem jest także wydajny sposób przechowywania danych i zarządzania nimi, co często wymaga znajomości obsługi systemów operacyjnych Unix i wiedzy programistycznej.

Analizę danych z RNA-seq utrudnia także fakt, iż struktura danych oraz metody ich przetwarzania, podobnie jak w przypadku danych mikromacierzowych, ściśle zależą od rodzaju użytej platformy. Ponadto, technologia RNA-Seq wciąż znajduje się w fazie rozwoju, optymalizacji i standaryzacji metod analizy danych.

Zarówno ekspresyjne mikromacierze DNA, jak i technologia RNA-seq pozwalają na wysokoprzepustową analizę transkryptomu. Jednakże istnieje kilka aspektów na etapie analizy transkryptomu, takich jak analiza niekodujących RNA, połączeń ekson-ekson oraz identyfikacja różnych izoform, gdzie to użycie RNA-seq będzie bardziej trafnym wyborem. Prawdopodobnie rozwój technologii RNA-seq w kierunku zwiększenia czułości, minimalizacji kosztów oraz wydajniejszej analizy danych, przyczyni się do ograniczenia stosowania mikromacierzy w badaniach ekspresji genów. Skutkiem tego może być specjalizacja technologii mikromacierzowych i stosowanie ich głównie w badaniach biomedycznych oraz klinicznych, w formie tzw. dedykowanych mikromacierzy DNA, np. zaprojektowanych do badania określonych procesów biologicznych lub przeznaczonych do wykonywania testów diagnostycznych, prognostycznych czy przesiewowych (Cernetich-Ott i wsp. 2012; Togawa i wsp. 2012; Lagorce i wsp. 2012).

III. Cel pracy

Głównym celem niniejszej pracy było opracowanie metod analizy najczęściej spotykanych rodzajów niestandardowych danych uzyskiwanych przy użyciu mikromacierzy DNA, które analizowane były w Centrum Doskonałości CENAT. Osiągnięcie tego celu wiązało się z realizacją następujących zadań szczegółowych, z których każde odpowiadało problemowi analizy konkretnego rodzaju niestandardowych danych:

1) Opracowanie metody analizy danych otrzymanych z użyciem ekspresyjnych mikromacierzy DNA o niestandardowym układzie sond.
2) Opracowanie metody analizy danych otrzymanych z użyciem mikromacierzy DNA do badania ekspresji miRNA.
3) Opracowanie uniwersalnej procedury wyboru metody normalizacji dla danych uzyskiwanych z użyciem dedykowanych mikromacierzy DNA.
4) Opracowanie metody analizy danych otrzymanych w wyniku hybrydyzacji międzygatunkowej przy zastosowaniu filtracji danych na podstawie morfologii punktów.

IV. Materiały i Metody

IV.I Materiały

W ramach realizacji niniejszej pracy doktorskiej wykorzystanych zostało pięć zestawów danych uzyskanych z użyciem ekspresyjnych mikromacierzy DNA. Należą do nich:

- **Zestaw AML:** zestaw danych do badania ekspresji genów u pacjentów z ostrą białaczką szpikową (ang. *acute myleoid leukemia*),
- **Zestaw AML miRNA:** zestaw danych do analizy ekspresji miRNA u pacjentów z białaczką szpikową,
- **Zestaw ALERGIA:** zestaw danych do badania ekspresji genów u dzieci z alergią krzyżową,
- **Zestaw ASTMA:** zestaw do badania ekspresji genów u dzieci z astmą,
- **Zestaw NT-CSH:** zestaw do badania ekspresji genów u *Nicotiana tabacum* pod wpływem stresu abiotycznego z zastosowaniem hybrydyzacji międzygatunkowej.
- **Zestaw OSHLACK:** zestaw do badania ekspresji genów na etapie późnego różnicowania limfocytów B u myszy.

IV.I.1 Zestaw AML: zestaw danych do badania ekspresji genów u pacjentów z ostrą białaczką szpikową

IV.I.1.1 Projekt eksperymentu

Badania dotyczące ustalenia profilu ekspresji genów u pacjentów z ostrą białaczką szpikową (AML) prowadzone były w Centrum Doskonałości CENAT Instytutu Chemii Bioorganicznej PAN w Poznaniu we współpracy z Katedrą i Kliniką Hematologii i Chorób Rozrostowych Układu Krwiotwórczego Uniwersytetu Medycznego im. Karola Marcinkowskiego w Poznaniu. Analizą objęto grupę 33 dorosłych pacjentów Kliniki, u których zdiagnozowano AML podtypu M1 lub M2 wg. klasyfikacji FAB (ang. *French-American-British classification*). Od wszystkich 33 pacjentów uzyskano próbki krwi i szpiku w momencie pierwszej diagnozy (czas T0), a od niektórych także po pierwszej serii chemioterapii (czas T1, 9 pacjentów) oraz w okresie wznowy choroby (czas T2, 11 pacjentów). Kontrolę eksperymentu stanowiło 14 próbek krwi i 1 próbka szpiku od zdrowych ochotników, natomiast jego referencję komórki z linii komórkowej HL60, wyprowadzonej z komórek ostrej białaczki promielocytowej typ M3. Z otrzymanych próbek krwi i szpiku pozyskano komórki jednojądrzaste, z których następnie izolowano całkowity RNA.

Całkowity RNA wyizolowano także z komórek HL60. Otrzymane RNA poddano analizie jakościowej (*BioAnalyzer 2100*, Agilent Technologies) i ilościowej (*NanoDrop*, ThermoScentific). Z wybranych próbek RNA uzyskano cDNA w wyniku reakcji odwrotnej transkrypcji z udziałem modyfikowanych nukleotydów (aminoallylo-dNTP wg *SuperScript Indirect cDNA Labeling System*, Invitrogen). Do znakowania otrzymanego cDNA użyto pary barwników fluorescencyjnych Alexa 555 oraz Alexa 647. Próbki od pacjentów z AML, jak i próbki zdrowych ochotników znakowane były barwnikiem Alexa 647. Natomiast próbki referencyjne z linii komórkowej HL60 znakowane były barwnikiem Alexa 555. Do pojedynczej mikromacierzy hybrydyzowana była mieszanina próbek z których jedna znakowana była barwnikiem Alexa 555, a druga barwnikiem Alexa 647. Reakcja hybrydyzacji prowadzona była metodą ręczną z użyciem komór do hybrydyzacji (*Hybridization Chamber*, Corning®) lub metodą automatyczną z użyciem hybrydyzatora (*HybArray 12*, PerkinElmer). Warunki reakcji hybrydyzacji przedstawiono w Tabela 1.

Tabela 1. *Opis warunków reakcji hybrydyzacji mieszaniny próbek znakowanego fluorescencyjnie cDNA do mikromacierzy DNA, stosowanych w ramach eksperymentu badania ekspresji genów u pacjentów z AML.*

Etap	T [°C]	Czas [h]		Uwagi
		automatyczna	ręczna	
reakcja hybrydyzacji: znakowany cDNA				
hybrydyzacja 1	50	5h	3h	Objętość próbki: 1.Hybrydyzacja automatyczna: 110µl 2.Hybrydyzacja ręczna: 40 µl
hybrydyzacja 2	45	5h	3h	
hybrydyzacja 3	40	5h	3h	
płukanie				
płukanie 1	40	5 min		Bufor: (2x SSC, 0,1% SDS)
płukanie 2	25	5 min.		Bufor: (2x SSC)
płukanie 3	25	5 min.		Bufor: (0,2x SSC)

Mikromacierze suszono wirując przez 2 min. przy 1500 rpm, a następnie poddano procesowi skanowania. Mikromacierze zostały przeskanowane z użyciem skanera do mikromacierzy *ScanArray Express* (Perkin Elmer) lub skanera *GenePix 4200AL* (Molecular Devices). Każda mikromacierz skanowana była przynajmniej dwukrotnie. Pierwsze, tzw. wstępne skanowanie wykonywane było przy rozdzielczości 50 µm. Natomiast ostateczne skanowanie wykonywano przy rozdzielczości 10 µm. Uzyskane obrazy w kolejnym etapie poddane były analizie ilościowej obrazu za pomocą programu *GenePixPro 6.0* (Molecular Devices). Wynikiem analizy ilościowej obrazu było przekształcenie obrazu graficznego w dane liczbowe, które zdeponowane zostały w plikach z rozszerzeniem .gpr (ang. *GenepixPro results*). Pliki z rozszerzeniem .gpr stanowiły dane wejściowe do dalszych etapów analizy, prowadzonych z

użyciem programu R/Bioconductor oraz Python.

W skład zestawu AML wchodzą 172 ekspresyjne mikromacierze DNA otrzymane w wyniku hybrydyzacji próbek pochodzących od 33 pacjentów z AML (151 hybrydyzacje) i 15 osób zdrowych (21 hybrydyzacji). Próbki od pacjentów z AML pochodziły zarówno z krwi jak i szpiku, pobranych w trzech punktach czasowych: T0 (120 hybrydyzacje, 33 pacjentów), T1 (21 hybrydyzacji, 9 pacjentów) i T2 (31 hybrydyzacji, 11 pacjentów).

Szczegółowe informacje na temat plików .gpr są dostępne w Załączniku 1.

IV.I.1.3 Projekt mikromacierzy

Dedykowane mikromacierze DNA użyte w ramach eksperymentu dotyczącego badań nad profilem ekspresji genów u pacjentów z AML zostały przygotowane w Centrum Doskonałości CENAT Instytutu Chemii Bioorganicznej PAN z użyciem unikalnego zestawu sond. W skład tego zestawu wchodziły sondy wyselekcjonowane w celu badania genów odpowiedzialnych za wystąpienie ostrej białaczki szpikowej oraz sondy kontrolne. Stosowne sondy miały długość 50-70 nukleotydów. Roztwory sond DNA w buforze 1xESB (*Epoxide spotting buffer*, Integrated DNA Technologies) o stężeniu 20 μM naniesiono na mikromacierz (*Epoxide Coated Slides*, Corning®) w procesie drukowania przy pomocy drukarki SpotArray24 (*PerkinElmer*). Sondy nadrukowywano przy użyciu 4 igieł (ułożonych w rzędzie). Każda sonda nadrukowana była w 3 powtórzeniach występujących obok siebie. W wyniku procesu drukowania otrzymano dedykowaną mikromacierz DNA zawierającą 3069 punktów w skład, których wchodziły:

- 938 unikalne wyselekcjonowane sondy (x 3 = 2814),
- 66 punktów, w których naniesiono tylko bufor (sondy oznaczone jako *buffer*),
- 96 punktów pustych (sondy oznaczone jako *empty*),
- 8 sond bakteryjnych (kontrola negatywna) (3x6 =18, 5x3=15),
- 3 sondy roślinne (kontrola negatywna) (x3=9)
- 1 sonda o losowej sekwencji (ang. *random oligo*) (kontrola negatywna) (x3=3)
- 8 sond kontrolnych typu *spike-in* (x 6 = 48).

Sondy na mikromacierzy tworzą układ 32 bloków (ang. *print-tip group*) rozmieszczonych w układzie 4 x 10 (kolumny x rzędy). W przypadku tej mikromacierzy bloki z rzędów 7 i 8 zostały zastąpione przez jeden, większy blok. Większość bloków zawiera sondy

rozmieszczone w układzie 9 x 9, natomiast blok zastępujący bloki z rzędów 7 i 8 zawiera sondy rozmieszczone w układzie 15 x 27 (Patrz. Rozdział V.I.1, Wyniki i Dyskusja).

IV.I.1.3 Zestaw AML II

W skład zestawu AML II wchodzi 40 (20 unikatowych hybrydyzacji i 20 powtórzeń technicznych) mikromacierzy wyselekcjonowanych z zestawu AML. Zestaw AML II obejmuje mikromacierze otrzymane w wyniku hybrydyzacji automatycznej użyciem hybrydyzatora (*HybArray 12,* PerkinElmer) oraz skanowania skanerem *ScanArray Express* (Perkin Elmer). Dane liczbowe dla zestawu AML II, podobnie jak dla zestawu AML otrzymano w wyniku analizy ilościowej obrazu z użyciem programu *GenePixPro 6.0* (Molecular Devices). Mikromacierze z zestawu AML II zostały otrzymane w wyniku hybrydyzacji próbek pochodzących od pacjentów z AML w czasie diagnozy T0 oraz próbek od zdrowych ochotników.

Szczegółowe informacje na temat plików .gpr z zestawu AML II są dostępne w Załączniku 2.

IV.I.2 Zestaw AML miRNA: zestaw danych do badania ekspresji ludzkich miRNA u pacjentów z ostrą białaczką szpikową

IV.I.2.1 Projekt eksperymentu

Badania dotyczące ustalenia profilu ekspresji miRNA u pacjentów z ostrą białaczką szpikową (AML) prowadzone były w Centrum Doskonałości CENAT Instytutu Chemii Bioorganicznej PAN w Poznaniu. Preparaty krwi i szpiku uzyskano dzięki współpracy z Katedrą Hematologii i Chorób Rozrostowych Układu Krwiotwórczego Uniwersytetu Medycznego w Poznaniu. Analizą objęto 24 dorosłych pacjentów, u których zdiagnozowano ostrą białaczkę szpikową (AML) podtypu M1 lub M2 wg. klasyfikacji FAB. Próbki pobrano od pacjentów w dwóch różnych stadiach choroby: T0- próbka pobierana po pierwszej diagnozie oraz T2– próbka pobrana od pacjenta podczas wznowy choroby. Referencję eksperymentu stanowiły komórki z linii komórkowej HL60, wyprowadzonej z komórek ostrej białaczki promielocytowej podtyp M3. Z otrzymanych próbek krwi i szpiku pozyskano komórki jednojądrzaste, z których wyizolowano frakcję niskocząsteczkowego RNA (*PureLink™ miRNA Isolation Kit,* Invitrogen**)**. Frakcję niskocząsteczkowego RNA wyizolowano także z komórek HL60. Otrzymane RNA poddano procesowi amplifikacji w wyniku transkrypcji *in vitro* (*NCode miRNA Amplification System,* Invitrogen). Otrzymane RNA poddano analizie ilościowej (*NanoDrop*, ThermoScentific) oraz jakościowej

(*BioAnalyzer 2100*, Agilent Technologies). Wybrane próbki RNA znakowane były metodą pośrednią, polegającą na ligacji do poliadenylowanego końca 3' RNA, dwóch adaptorów. Produkty ligacji hybrydyzowano do mikromacierzy, następnie po odmyciu niezwiązanych cząsteczek wykonywano drugą hybrydyzację z oligonukleotydem związanym z barwnikiem fluorescencyjnym (odpowiednio Alexa 546 lub Alexa 647), który był komplementarny do jednego z adaptorów dołączonych do cząsteczek RNA (*NCode™ miRNA Labeling System,* Invitrogen). Próbki od pacjentów z AML znakowane były barwnikiem Alexa 647, natomiast próbki referencyjne z linii komórkowej HL60 znakowane były barwnikiem Alexa 546. Do pojedynczej mikromacierzy hybrydyzowana była mieszanina próbek z których jedna znakowana była barwnikiem Alexa 546, a druga barwnikiem Alexa 647. Reakcja hybrydyzacji prowadzona była metodą ręczną z użyciem komór do hybrydyzacji (*Hybridization Chamber,* Corning®) lub metodą automatyczną z użyciem hybrydyzatora (*HybArray 12,* PerkinElmer). Warunki reakcji hybrydyzacji przedstawiono w Tabeli 2.

Tabela 2. *Opis warunków reakcji hybrydyzacji mieszaniny próbek znakowanego fluorescencyjnie cDNA do mikromacierzy DNA, stosowanych w ramach eksperymentu badania ekspresji miRNA u pacjentów z AML.*

Etap	T [^{o}C]	Czas [h]	Uwagi
I reakcja hybrydyzacji: znakowane miRNA			
Hybrydyzacja 1	52	6h	Objętość próbki:
Hybrydyzacja 2	46	6h	1.Hybrydyzacja automatyczna: 110µl
Hybrydyzacja 3	40	6h	2.Hybrydyzacja ręczna: 40 µl
płukanie			
Płukanie 1	50	10 min	Bufor: (2x SSC, 0,1% SDS)
Płukanie 2	25	10 min.	Bufor: (2x SSC)
Płukanie 3	25	10 min.	Bufor: (0,2x SSC)
II reakcja hybrydyzacji: barwniki Alexa 546 i Alexa 647			
Hybrydyzacja	58	4h	Objętość próbki: 1.Hybrydyzacja automatyczna: 110µl 2.Hybrydyzacja ręczna: 40 µl
płukanie			
Płukanie 1	50	10 min	Bufor: (2x SSC, 0,1% SDS)
Płukanie 2	25	10 min.	Bufor: (2x SSC)
Płukanie 3	25	10 min.	Bufor: (0,2x SSC)

Macierze suszono wirując przez 2 min. przy 1500 rpm, a następnie poddano procesowi skanowania. Mikromacierze zostały przeskanowane z użyciem skanera do mikromacierzy *ScanArray Express* (Perkin Elmer) lub skanera *GenePix 4200AL* (Molecular Devices). Uzyskane obrazy w kolejnym etapie poddane były analizie ilościowej obrazu za pomocą programu *GenePixPro 6.0* (Molecular Devices). Wynikiem analizy ilościowej obrazu było

przekształcenie obrazu graficznego w dane liczbowe, które zdeponowane zostały w plikach z rozszerzeniem .gpr. Pliki z rozszerzeniem .gpr stanowiły dane wejściowe do dalszych etapów analizy, prowadzonych z użyciem programu R/Bioconductor oraz Python.

W skład zestawu AML miRNA wchodzi 30 dedykowanych mikromacierzy DNA otrzymanych w wyniku hybrydyzacji próbek pochodzących od 24 pacjentów. Próbki od pacjentów pochodziły zarówno z krwi, jak i szpiku, pobranych w dwóch punktach czasowych: T0 (24 hybrydyzacje, 24 pacjentów), T2 (6 hybrydyzacji, 6 pacjentów).

Szczegółowe informacje na temat plików .gpr są dostępne w Załączniku 3.

IV.I.2.2 Projekt mikromacierzy

Mikromacierze użyte w eksperymencie analizy różnicowej miRNA u pacjentów z ostrą białaczką szpikową zostały wykonane w Centrum Doskonałości CENAT Instytutu Chemii Bioorganicznej PAN. Do przygotowania mikromacierzy wykorzystano komercyjne dostępny zestaw sond oligonukleotydowych (*NCode™ Mammalian miRNA Microarray Probe Set v. 1.0*, Invitrogen). Zestaw ten wyselekcjonowany został na podstawie sekwencji dojrzałych mikroRNA zawartych w bazie *Sanger miRBase 7.0* (http://microrna.sanger.ac.uk) dla człowieka (311 sond), myszy domowej (232 sondy) oraz szczura (185 sond). Wśród sond ulokowanych na mikromacierzy znajdują się 144 sondy dla przewidywanych sekwencji ludzkich miRNA (HMP_PREDICTED). Każda z sond składa się z dwukrotnie powtórzonej sekwencji odpowiednich miRNA, stąd długość sondy na poziomie 34-44 nukleotydów. Roztwory sond DNA w buforze Pronto! (Schott) o stężeniu 20 µM naniesiono na mikromacierz (*Epoxide Coated Slides*, Corning®) w procesie drukowania przy pomocy drukarki SpotArray24 (*PerkinElmer*). Sondy nadrukowywano przy użyciu 4 igieł (ułożonych w rzędzie). Każda sonda nadrukowana była w 3 powtórzeniach występujących obok siebie. W wyniku procesu drukowania otrzymano dedykowaną mikromacierz DNA zawierającą 2880 punktów w skład, których wchodziły:

- 728 unikalne sondy (*NCode™ Mammalian miRNA Microarray Probe Set v. 1.0*, Invitrogen) (x 3= 2184)
- 144 unikalne sondy dla przewidywanych sekwencji ludzkich miRNA(x 3=432)
- 10 sond kontrolnych *Alexa5 Test Feature* (kontrola pozytywna) (x3=30)
- 15 sond kontrolnych *NodeControl* (kontrola negatywna)(x3=45)
- 189 punktów pustych (*empty*)

Sondy na mikromacierzy tworzą układ 4 bloków, z których każdy zawiera 8 kolumn i 90 wierszy sond.

IV.I.3 Zestaw ALERGIA: Zestaw danych do badania ekspresji genów u dzieci z alergią krzyżową

IV.I.3.1 Projekt eksperymentu

Badania ekspresji genów u dzieci z alergią krzyżową i astmą prowadzone były w Centrum Doskonałości CENAT oraz w Centrum Badań Biokrystalograficznych Instytutu Chemii Bioorganicznej PAN w Poznaniu, we współpracy z III Kliniką Chorób Dziecięcych Akademii Medycznej w Białymstoku. Analizą objęto grupę 55 pacjentów w wieku 3,5 – 18 lat, niepalących (czynnie i biernie). Wykonano oznaczenia całkowitych i specyficznych IgE z krwi, testy skórne dla najczęściej uczulających alergenów (AW, AP, jabłko, soja, orzech, sezam, mleko, białko jaja, brzoza, drzewa, trawy, chwasty i roztocza). Dodatkowo wykonano test prowokacji pokarmowej oraz badanie poziomu eozynofilów. Ostatecznie do analiz z użyciem mikromacierzy zakwalifikowano 16 osób z alergia krzyżową oraz 15 zdrowych ochotników (niewykazujących atopii, wyłączono także osoby, u których poziom cIgE był wyższy niż 90 kU/l). Z otrzymanych próbek krwi obwodowej pozyskano komórki jednojądrzaste, z których następnie pozyskano całkowity RNA. Próbki RNA następnie poddano analizie ilościowej (*NanoDrop,* ThermoScientific) oraz jakościowej (*Bioanalyzer 2100,* Agilent Technologies). Wybrane próbki RNA zostały poddane reakcji odwrotnej transkrypcji z wykorzystaniem startera zawierającego sekwencję oligo-dT oraz fragment sekwencji wirusowej polimerazy T7. Otrzymane cDNA poddano amplifikacji. W kolejnym etapie procesu przygotowania próbek do reakcji hybrydyzacji, otrzymano antysensowną nić DNA. Podczas reakcji do tworzącej się nici włączane były modyfikowane nukleotydy (aminoallylo-dNTP wg *SuperScript Indirect cDNA Labeling System,* Invitrogen), które stanowiły miejsce wiązania barwników fluorescencyjnych. Do znakowania otrzymanego DNA użyto pary barwników fluorescencyjnych Alexa 555 oraz Alexa 647. Próbki od pacjentów z alergią znakowane były barwnikiem Alexa 647. Natomiast próbki kontrolne, pochodzące od zdrowych ochotników znakowane były barwnikiem Alexa 555. Do pojedynczej mikromacierzy hybrydyzowana była mieszanina próbek z których jedna znakowana była barwnikiem Alexa 555, a druga barwnikiem Alexa 647. Reakcja hybrydyzacji prowadzona była metodą automatyczną z użyciem hybrydyzatora (*HybArray 12,* PerkinElmer). Warunki reakcji hybrydyzacji przedstawiono w Tabeli 3.

Tabela 3. *Opis warunków reakcji hybrydyzacji mieszaniny próbek znakowanego fluorescencyjnie cDNA do mikromacierzy DNA, stosowanych w ramach eksperymentu badania ekspresji genów u dzieci z alergią krzyżową.*

Etap	T [^{o}C]	Czas [h]	Uwagi
reakcja hybrydyzacji			
Hybrydyzacja 1	50	5h	Hybrydyzacja automatyczna Objętość próbki:110µl
Hybrydyzacja 2	45	5h	
Hybrydyzacja 3	40	5h	
płukanie			
Płukanie 1	35	5 min	Bufor: (2x SSC, 0,1% SDS)
Płukanie 2	25	10 min.	Bufor: (2x SSC)
Płukanie 3	25	10 min.	Bufor: (0,2x SSC)

Macierze suszono wirując przez 2 min. przy 1500 rpm, a następnie poddano procesowi skanowania. Mikromacierze zostały przeskanowane z użyciem skanera do mikromacierzy *ScanArray Express* (Perkin Elmer). Uzyskane obrazy w kolejnym etapie poddane były analizie ilościowej obrazu za pomocą programu *GenePixPro 6.0* (Molecular Devices). Wynikiem analizy ilościowej obrazu było przekształcenie obrazu graficznego w dane liczbowe, które zdeponowane zostały w plikach z rozszerzeniem .gpr. Pliki .gpr stanowiły dane wejściowe do dalszych etapów analizy, prowadzonych z użyciem programu R/Bioconductor oraz Python.

W skład zestawu wchodzi 14 ekspresyjnych mikromacierzy DNA otrzymanych w wyniku hybrydyzacji próbek pochodzących od 12 pacjentów z alergią i 12 zdrowych ochotników oraz dwa powtórzenia techniczne.

Szczegółowe informacje na temat plików .gpr są dostępne w Załączniku 4.

IV.I.3.2 Projekt mikromacierzy

Mikromacierze DNA do badania ekspresji genów u dzieci z alergią krzyżową zostały wykonane w Centrum Doskonałości CENAT Instytutu Chemii Bioorganicznej PAN w Poznaniu. Projekt każdej z mikromacierzy obejmuje użycie sond dla 179 genów zaangażowanych głównie w reakcje zapalne (zależne od IgE) oraz genów powiązanych z rozwojem astmy i alergii (na podstawie danych literaturowych). Stosowne sondy miały długość 50 nukleotydów. Roztwory sond DNA w buforze 1xESB (*Epoxide spotting buffer*, Integrated DNA Technologies) o stężeniu 20 µM naniesiono na mikromacierz (*Epoxide Coated Slides*, Corning®) w procesie drukowania przy pomocy drukarki SpotArray24 (*PerkinElmer*). Sondy nadrukowywano przy użyciu 4 igieł (ułożonych w rzędzie). Każda sonda nadrukowana była w 6, następujących po sobie powtórzeniach. Sondy kontrolne

dodatkowo nadrukowywano w kilku miejscach na mikromacierzy. Łącznie otrzymano 1536 punktów obejmujących:

- 188 unikatowych sond (x 6= 1128)
- 180 punktów w których naniesiono tylko bufor (*buffer*)
- 24 punkty puste (*empty*)
- 8 sond kontrolnych typu *spike-in* (x 6=48)
- 4 sondy kontrolne (kontrola negatywna) (x 12=48)
- 9 sond kontrolnych (kontrola pozytywna) (x 12=108)

Sondy tworzą układ 12 bloków (4 kolumny x 3 rzędy). Bloki 1-8 zawierają 12 kolumn i 12 rzędów sond. Natomiast bloki 9-12 zawierają 12 kolumn i 8 rzędów sond.

IV.I.4 Zestaw ASTMA: zestaw do badania ekspresji genów u dzieci z alergią krzyżową astmą

IV.I.4.1 Projekt eksperymentu

Do eksperymentu zrekrutowano 11 pacjentów z astmą atopową uczulonych na roztocze kurzu domowego (*Dermatophagoides pteronyssinus*) oraz 10 pacjentów z rozwiniętą astmą oskrzelową, niewykazujących atopii. Ostatecznie do analiz z użyciem mikromacierzy zakwalifikowano 7 osób z grupy badanej (astmą atopowa) i 7 z grupy kontrolnej (astma nieatopowa). Od pacjentów pobrano po 4 ml krwi obwodowej. Z otrzymanych próbek krwi pozyskano komórki jednojądrzaste, z których następnie izolowano całkowity RNA. Otrzymany RNA poddano ocenie jakościowej (*BioAnalyzer 2100*, Agilent Technologies) i ilościowej (*NanoDrop*, ThermoScentific). Wybrane próbki RNA zostały poddane reakcji odwrotnej transkrypcji do cDNA z wykorzystaniem startera zawierającego sekwencję oligo-dT oraz fragment sekwencji wirusowej polimerazy T7. Otrzymane cDNA poddano amplifikacji. W kolejnym etapie procesu przygotowania próbek do reakcji hybrydyzacji, otrzymano antysensowną nić DNA. Podczas reakcji do tworzącej się nici włączane były modyfikowane nukleotydy (aminoallylo-dNTP wg *SuperScript Indirect cDNA Labeling System*, Invitrogen), które stanowiły miejsce wiązania barwników fluorescencyjnych. Do znakowania otrzymanego DNA użyto pary barwników fluorescencyjnych Alexa 555 oraz Alexa 647. Próbki od pacjentów z astmą atopową znakowane były barwnikiem Alexa 647. Natomiast próbki kontrolne, pochodzące od pacjentów z astmą nieatopową znakowane były barwnikiem Alexa 555. Do pojedynczej mikromacierzy hybrydyzowana była mieszanina

próbek z których jedna znakowana była barwnikiem Alexa 555, a druga barwnikiem Alexa 647. Reakcja hybrydyzacji próbek do mikromacierzy prowadzona była metodą automatyczną z użyciem hybrydyzatora (*HybArray 12,* PerkinElmer). Warunki reakcji hybrydyzacji przedstawiono w Tabeli 3. Macierze suszono wirując przez 2 min. przy 1500 rpm, następnie poddano procesowi skanowania za pomocą skanera do mikromacierzy *ScanArray Express* (Perkin Elmer). Analizę ilościową przeprowadzono za pomocą programu *GenePixPro 6.0* (Molecular Devices). Wynikiem analizy ilościowej obrazu było przekształcenie obrazu graficznego w dane liczbowe, które zdeponowane zostały w plikach z rozszerzeniem .gpr. Pliki .gpr stanowiły dane wejściowe do dalszych etapów analizy, prowadzonych z użyciem programu R/Bioconductor oraz Python.

W skład zestawu wchodzi 14 ekspresyjnych mikromacierzy DNA otrzymanych w wyniku hybrydyzacji próbek pochodzących od 7 osób z grupy badanej (astma atopowa) i 7 z grupy kontrolnej (astma nieatopowa) oraz 7 powtórzeń technicznych.

Szczegółowe informacje na temat plików .gpr są dostępne w Załączniku 5.

IV.I.4.1 Projekt mikromacierzy

Zestaw ASTMA otrzymano w oparciu o projekt mikromacierzy opisany w Rozdziale IV.I.3.2 (Materiały i Metody).

IV.I.5 Zestaw NT-CSH: zestaw do badania ekspresji genów u *Nicotiana tabacum* pod wpływem stresu abiotycznego z zastosowaniem hybrydyzacji międzygatunkowej

IV.I.5.1 Projekt eksperymentu

Zestaw NT-CSH został otrzymany w wyniku hybrydyzacji mieszaniny znakowanych fluorescencyjnie próbek z tytoniu pod wpływem stresu abiotycznego (NaCl lub $CdCl_2$) oraz roślin typu dzikiego (próba kontrolna) do trzech rodzajów ekspresyjnych mikromacierzy DNA:

- TIGR Potato cDNA Microarray (podzestaw POT)
- TOM1 (podzestaw TOM1)
- Arabidopsis Oligonucleotide Microarray (podzestaw ATH)

Materiał roślinny z korzeni oraz siewek tytoniu szlachetnego (*Nicotiana tabacum*) oraz rzodkiewnika (*Arabidopsis thaliana*) uzyskano dzięki współpracy Centrum Doskonałości

CENAT Instytutu Chemii Bioorganicznej PAN z Laboratorium Biochemii Roślin, Instytutu Biochemii i Biofizyki PAN. Rośliny zostały poddane działaniu czynników stresogennych w postaci: 100 μM chlorku kadmu (II) (CdCl2) lub 150mM chlorku sodu (NaCl) przez okres 6 godzin. Wszystkie rośliny uzyskano w wyniku hodowli metodą hydroponiczną. Materiał roślinny z korzeni i siewek został roztarty na proszek w moździerzu, w ciekłym azocie. Następnie z otrzymanego materiału (ok. 100 mg na jedną reakcję izolacji) izolowano całkowity RNA (*RNeasy Plant Mini Kit*, Qiagen). Otrzymane próbki całkowitego RNA zostały poddane procesowi usuwania DNA (*TURBO DNA-free™ Kit, Ambion*). Następnie RNA zostało poddane analizie ilościowej (*NanoDrop,* ThermoScientific) oraz jakościowej *(BioAnalyzer 2100,* Agilent Technologies). Procedura przygotowania wybranych próbek RNA do reakcji hybrydyzacji zależała od rodzaju ekspresyjnej mikromacierzy DNA stosowanej w ramach eksperymentu.

IV.I.5.1.1 Przygotowanie prób do hybrydyzacji

Wybrane próbki RNA stanowiły matryce w reakcji odwrotnej transkrypcji z udziałem modyfikowanych nukleotydów (aminoallylo-dNTP*, wg SuperScript Indirect cDNA Labeling System,* Invitrogen). Do znakowania materiału użyto pary barwników fluorescencyjnych Cyjanina 5 (Cy5) oraz Cyjanina 3 (Cy3). Próbki badane (rośliny tytoniu poddane działaniu NaCl lub $CdCl_2$) znakowane były Cy5. Próbki kontrolne (rośliny typu dzikiego) natomiast znakowane były barwnikiem Cy3. Proces znakowania próbek dla podzestawu POT obejmował także przygotowanie próbek typu *dye swap* dla każdego z czynników wywołujących stres abiotyczny (NaCl lub $CdCl_2$).

IV.I.5.1.2 Reakcja hybrydyzacji, skanowanie i analiza ilościowa obrazu

Do pojedynczej mikromacierzy hybrydyzowana była mieszanina próbek z których jedna znakowana była barwnikiem Cy3, a druga barwnikiem Cy5. Reakcja hybrydyzacji prowadzona była metodą automatyczną z użyciem hybrydyzatora *Hybarray 12,* PelkinElmer lub *Hs4800 Pro,* TECAN. Warunki reakcji hybrydyzacji przedstawiono w Tabeli 4.

Tabela 4. *Opis warunków reakcji hybrydyzacji mieszaniny próbek znakowanego fluorescencyjnie cDNA do mikromacierzy DNA, stosowanych w ramach eksperymentu badania ekspresji genów u Nicotiana tabacum pod wpływem stresu abiotycznego z zastosowaniem hybrydyzacji międzygatunkowej.*

Etap	T [^{o}C]	Czas [h]	Uwagi
Prehybrydyzacja (bez próbki)			
Prehybrydyzacja	42	45 min.	
płukanie			
Płukanie 1	25	5min.	H_2O miliQ
Płukanie 2	25	5 min.	H_2O miliQ
Płukanie 3	25	2 min.	izopropanol
Reakcja hybrydyzacji			
Hybrydyzacja	46	18h	Objętość próbki: 1.Hybrydyzacja automatyczna: 115µl
płukanie			
Płukanie 1	43	5 min	Bufor: (2x SSC, 0,1% SDS)
Płukanie 2	30	5 min.	Bufor: (0,5 x SSC)
Płukanie 3	25	5 min.	Bufor: (0,05 x SSC)

Mikromacierze suszono wirując przez 2 min. przy 1500 rpm, a następnie poddano procesowi skanowania. Mikromacierze zostały przeskanowane z użyciem skanera do mikromacierzy *ScanArray Express* (Perkin Elmer). Każda mikromacierz skanowana była przynajmniej dwukrotnie. Pierwsze, tzw. wstępne skanowanie wykonywane było przy rozdzielczości 50 µm. Natomiast ostateczne skanowanie wykonywano przy rozdzielczości 5 µm. Uzyskane obrazy w kolejnym etapie poddane były analizie ilościowej obrazu za pomocą programu *GenePixPro 6.0* (Molecular Devices). Wynikiem analizy ilościowej obrazu było przekształcenie obrazu graficznego w dane liczbowe, które zdeponowane zostały w plikach z rozszerzeniem .gpr. Dodatkowo analizę ilościową obrazu przeprowadzono za pomocą programu MAIA (Novikov & Barillot 2007) w celu otrzymania dodatkowych parametrów jakości punktów (SC, ang. *spot characteristics*). Pliki z rozszerzeniem .gpr stanowiły dane wejściowe do dalszych etapów analizy, prowadzonych z użyciem programu R/Bioconductor. Informacja otrzymana na temat wartości parametrów SC została dodana na etapie analizy danych za pomocą R/Bioconductor.

W skład zestawu NT-CSH wchodzą trzy podzestawu: POT, TOM1 oraz TOB. Każdy podzestaw składa się z 12 ekspresyjnych mikromacierzy DNA. W przypadku podzestawów POT i TOB 12 ekspresyjnych mikromacierzy DNA obejmuje 6 mikromacierzy (3 hybrydyzacje oraz 3 hybrydyzacje typu *dye swap*) dla roślin traktowanych NaCl i 6 mikromacierzy (3 hybrydyzacje oraz 3 hybrydyzacje typu *dye swap*) dla roślin traktowanych

$CdCl_2$. W skład podzestawu TOB wchodzi 6 mikromacierzy dla roślin traktowanych NaCl (brak hybrydyzacji typu *dye swap*) i 6 mikromacierzy (brak hybrydyzacji typu *dye swap*) dla roślin traktowanych $CdCl_2$.

IV.I.5.1 Projekt mikromacierzy

- **TIGR Potato cDNA Microarray,** mikromacierze cDNA dla *Symphytum tuberosum:*
  - Liczba sond: 32448
  - Długość sond: kilkaset-kilka tysięcy par zasad
  - Producent: The Institute for Genomic Research
  - Nazwa podzestawu: POT
- **TOM1** - mikromacierze cDNA dla *Lycopersicon esculentum:*
  - Liczba sond: 13440
  - Długość sond: kilkaset-kilka tysięcy par zasad
  - Producent: The Boyce Thompson Institute
  - Nazwa podzestawu: TOM1
- **Arabidopsis Oligonucleotide Microarray** - mikromacierze DNA dla *A.thaliana:*
  - Liczba sond: 26000
  - Długość sond: 60 nukleotydów
  - Producent: University of Arizona
  - Nazwa podzestawu: ATH

IV.I.6 Zestaw OSHLACK: zestaw do badania ekspresji genów na etapie późnego różnicowania limfocytów B u myszy

IV.I.6.1 Projekt eksperymentu

Zestaw danych Oshlack został opisanny przez zespół Oshlack i wsp. (Oshlack i wsp. 2007). Zestaw ten został otrzymany w ramach eksperymentów obejmujących badanie ekspresji genów na etapie późnego różnicowania limfocytów B. W reakcji hybrydyzacji stosowano cDNA syntetyzowany na podstawie RNA izolowanego z aktywowanych limfocytów B, otrzymanych *in vitro* z plasmoblastów lub RNA izolowanego z całkowicie zróżnicowanych komórek osocza, otrzymanych *ex vivo*. Komórki te pochodziły odpowiednio z myszy z wyciszonym genem OBF-1-/- (próbki badane) oraz myszy kontrolnych typu C57BL/6 (próbki kontrolne). Stosowane w reakcji hybrydyzacji próbki badane wyznakowane były barwnikiem fluorescencyjnym Cy5, natomiast próbki kontrolne wyznakowane były

barwnikiem Cy3. Mikromacierze skanowane były skanerem *GenePix 4200B* (Molecular Devices). Analizę ilościową przeprowadzona została za pomocą programu *GenePixPro 6.0* (Molecular Devices). Wynikiem analizy ilościowej obrazu było przekształcenie obrazu graficznego w dane liczbowe, które zdeponowane zostały w plikach z rozszerzeniem .gpr. Pliki z rozszerzeniem .gpr stanowiły dane wejściowe do dalszych etapów analizy, prowadzonych z użyciem programu R/Bioconductor.

W skład zestawu wchodzi 6 mikromacierzy otrzymanych w wyniku hybrydyzacji próbek pochodzących od myszy z wyciszonym genem OBF-1-/- (próbki badane) oraz próbek myszy kontrolnych typu C57BL/6 (próbki kontrolne).

Szczegółowe pliki .gpr są dostępne są pod adresem http://bioinf.wehi.edu.au/folders/boutique/

IV.I.6.2 Projekt mikromacierzy

Mikromacierze użyte w eksperymencie zostały przygotowane z użyciem dwóch rodzajów sond: sondy dla genów zaangażowanych w proces różnicowania limfocytów B oraz sondy kontrolne. Sondy dla genów zaangażowanych w proces różnicowania limfocytów B obejmują 109 sond, które stanowiły fragmenty PCR odpowiadające genom ulegającym ekspresji różnicującej w trakcie późnych etapów różnicowania limfocytów B. Sondy te zostały wyselekcjonowane na podstawie doniesień literaturowych lub na podstawie wyników półilościowego PCR (ang. *semi-quantitative PCR*). W skład zestawu sond kontrolnych wchodzą dwa typy sond: odpowiadające trzem genom o konstytutywnej ekspresji (ang. *housekeeping genes*) oraz zestaw sond MSP (ang. *microarray sample pool*). Zestaw MSP został otrzymany w wyniku połączenia klonów biblioteki cDNA NIA15K (Tanaka i wsp. 2000). Zestaw MSP został przygotowany z użyciem roztworów sond w różnych rozcieńczeniach, tak aby ostateczne stężenie sond wynosiło: 250 ng/µl, 120 ng/µl, 60 ng/µl, 30 ng/µl, 15ng/µl, 7 ng/µl, 4 ng/µl, 2 ng/µl and 1 ng/µl. Roztwór sondy o każdym rozcieńczeniu naniesiono na mikromacierz 32 razy, co łącznie daje 288 sond kontrolnych typu MSP. Sondy są fragmentami cDNA o długości od kilkuset do kilku tysięcy par zasad. Każda z sond została nadrukowana w 4 powtórzeniach.

IV.II Metody

Analiza poszczególnych zestawów danych opisanych w rozdziale IV.I przeprowadzona została z wykorzystaniem oprogramowania R/Bioconductor oraz środowiska programowania Pyton.

IV.II.1 R/Bioconductor

Analiza zestawów danych opisanych w rozdziale IV.I wykonana została z wykorzystaniem funkcji zaimplementowanych w pakiecie statystycznym R (R Development Core Team 2008) w wersji 2.14.1 z zestawem bibliotek projektu Bioconductor (Gentleman i wsp. 2004) w wersji 2.10.

Analiza danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA z użyciem R/Bioconductor obejmuje następujące etapy:

- wczytywanie danych wejściowych do R
- generowanie zestawu wykresów diagnostycznych
- korekta tła
- normalizacja wewnętrzna (normalizacja w obrębie mikromacierzy)
- normalizacja zewnętrzna (normalizacja pomiędzy mikromacierzami)
- uśrednianie powtórzeń technicznych (mikromacierzy)
- uśrednianie powtórzeń sond
- filtracja genów różnicujących

Tabela 5. *Wykaz ważniejszych pakietów funkcji projektu Bioconductor stosowanych na etapie analizy zestawów danych opisanych w rozdziale IV.I (Materiały i Metody).*

Etap analizy	Nazwa funkcji	Nazwa biblioteki
Wczytywanie danych	`read.GenePix()` `read.maimages()`	`marray` `limma`
Wykresy diagnostyczne	`maQualityPlots()`	`arrayQuality`
Korekcja tła	`backgroundCorrect()`	`limma`
Normalizacja wewnętrzna	`vsn2()` `normalizeWithinArrays()`	`vsn` `limma`
Normalizacja zewnętrzna	`normalizeBetweenArrays()`	`limma`
Uśrednianie powtórzeń technicznych	`duplicateCorrelation()`	`limma`

Identyfikacja genów różnicujących	`modelMatrix()`
	`avedups()`
	`lmFit()`
	`makeContrasts()`
	`contrasts.fit()`
	`eBayes()`
	`topTable()`
		`limma`

Uśrednianie powtórzeń sond	`avedups()` `avereps()`	`limma`

Filtracja genów	`decideTests()`
	`rowMeans()`
	`rowttests()`
	`read.AnnotatedDataFrame()`
		`limma`
		`base`
		`genefilter`
		`Biobase`

Analiza skupień	`pvclust()`
	`kmeans()`
	`em()`
	`heatmap()`
		`pvclust`
		`kmeans`
		`mclust`
		`stats`

Krzywe ROC	`roc()` `auc()`	 `pROC`

Szczegółowy opis funkcji, zdeponowanych w ramach repozytorium Bioconductor, stosowanych w ramach realizacji przedkładanej pracy doktorskiej dostępny jest w Załącznik 6.

Tabela 6. *Przegląd metod normalizacji zdeponowanych w repozytorium Bioconductor, stosowanych na etapie analizy danych zestawów AML II, ALERGIA, ASTMA i OSHLACK.*

Nazwa metody	Nazwa funkcji	Pakiet Bioconductor	Opis metody	Dodatkowe cechy	Referencje
Quantile1*	NormalizeBetweenArrays()	limma	Metoda normalizacji oparta na sortowaniu rozkładów próby badanej oraz próby kontrolnej. Rozkład badany oraz rozkład kontrolny powinny być tej samej długości.	Umożliwia normalizację danych zarówno jedno-, jak i dwukanałowych. Zewnętrzna metoda normalizacji.	(Smyth 2005)
Loess	NormalizeWithinArrays()	limma	Model normalizacji oparty na lokalnie ważonej regresji liniowej.	---	(Smyth 2005)
Spike	NormalizeWithinArrays()	limma	Model normalizacji oparty na lokalnie ważonej regresji liniowej.	Globalna metoda normalizacji bazująca na zestawie kontroli zewnętrznych typu spike-in.	(Smyth 2005)
Ploess	NormalizeWithinArrays()	limma	Model normalizacji oparty na lokalnie ważonej regresji liniowej.	Globalna metoda normalizacji stosowana oddzielnie dla każdej z grup sond drukowanych daną igłą	(Smyth 2005)
LoessM	maNorm()	marray	Model normalizacji oparty na lokalnie ważonej regresji liniowej.	Elastyczne podejście do normalizacji położenia oraz skali wartości M.	(Yang et al., 2009)
Vsn2* **Vsn1***	vsn2()	vsn	Model oparty na założeniu stałego współczynnika zmienności.	Umożliwia normalizację danych zarówno jedno-, jak i dwukanałowych.	(Huber 2002)
Nn	maNormNN()	nnNorm	Metoda normalizacji oparta na modelu sieci neuronowych. Oferuje obliczenie zależności stosunków intensywności w skali logarytmicznej (M) od głównej średniej intensywności (A). Dodatkowo pozwala na określenie przestrzennych koordynatów lokalizacji punktów na mikromacierzy (X,Y). Wartości A, jak również koordynaty X i Y są danymi wejściowymi do modelu sieci neuronowych i pozwalają uodpornić ten model na pojawienie się wartości odstających.	---	(Tarca et al. 2005)
Olin	olin()	olin	Obejmuje dwa schematy normalizacji oparte na iteracyjnej lokalnej regresji oraz selekcji modelu. Do optymalizacji parametrów stosowany jest uogólniony sprawdzian krzyżowy GCV (ang. generalized cross-validation). Dla lokalnej regresji stosowany jest algorytm LOCFIT.	Brak informacji o koordynatach lokalizacji punktów na mikromacierzy (X,Y).	(Futschik and Crompton, 2004)
Olin_c	olin()	olin		Wymagana informacja o koordynatach lokalizacji punktów na mikromacierzy (X,Y).	
Turbo	pspline()	TurboNorm	Normalizacja na podstawie średniej ważonej funkcji wygładzającej P-spline, która jest małą modyfikacją funkcji wygładzającej P-spline (Eilers i Marks, 1996), znanej jako połączenie funkcji wygładzającej (B-spline) i kary na współczynniki regresji.	---	(Iterson, in preparation)
Snm2* **Snm1***	snm()	snm	Nadzorowana metoda normalizacji, definiująca i modelująca różne źródła zmienności technicznej.	Umożliwia normalizację danych zarówno jedno-, jak i dwukanałowych.	(Mecham 2010)

IV.II.3 Python

Środowisko programistyczne Python2.7 (www.python.org) zostało wykorzystane w ramach przedstawianej pracy doktorskiej do przygotowania algorytmów:

- wymiar_macierzy.py (Załącznik 7)
- przyrównanie_sekwencji.py (Załącznik 8)
- filtracja.py (Załącznik 9)

Opisane wyżej algorytmy zostały przygotowane z wykorzystaniem modułów przedstawionych w Tabeli 7.

Tabela 7. *Opis modułów środowiska programistycznego Python stosowanych do analizy danych uzyskiwanych przy użyciu mikromacierzy DNA.*

Nazwa modułu	Funkcja	Wersja
os	Moduł ten zapewnia przenośny sposób wykorzystania funkcjonalności zależnej od operacyjnego systemu.	2.7
matplotlib	Biblioteka umożliwiająca wykonywanie różnego rodzaju wykresów 2D (histogramy, wykresy słupkowe, wykresy rozrzutu) spełniających kryteria wymagane podczas publikowania wyników. Wykresy mogą być generowane w różnych formatach plików (.jpg, .png, .pdf). Matplotlib może być wykorzystywany w skryptach pytona (pliki .py), jak również na etapie korzystania z aplikacji python lub ipython shell, serwery aplikacji internetowych oraz sześć bibliotek graficznego interfejsu graficznego.	1.1.0
NumPy	NumPy jest podstawowym modułem Pythona umożliwiającym obliczenia naukowe. Oprócz oczywistych naukowych zastosowań, NumPy może być stosowany do przechowywania wielowymiarowych danych rodzajowych. Istnieje możliwość zdefiniowana różnych typów danych. Takie podejście umożliwia temu modułowi na bezproblemową i szybką integrację z różnymi bazami danych.	1.6.2
Bio.Align.Applications	Moduł ten oferuje szereg metod dopasowania wielu sekwencji (ang. *multiple sequence aligment*). W module tym zawarte są m.in: • Clustalw • TCoffee • Muslce	6.92

IV.II.1 Błąd systematyczny i wariancja

W celu porównania dwukanałowych metod normalizacji zastosowano podejście zaproponowane przez Argyropoulos et al. (2006). Autorzy określili wartość błędu systematycznego (bs) oraz wariancji dla kontroli typu *spike-in* za pomocą następujących wzorów:

$$\mathrm{bs_i} = \sqrt{\sum_{\mathrm{j}}\sum_{\mathrm{k}}\left(\log_2\left(R_{i,j,k}/G_{i,j,k}\right)\right)^2/n} \qquad (1)$$

$$\mathrm{wariancja_i} = \sum_{\mathrm{j}}\sum_{k}\left(\log_2\left(R_{i,j,k}/G_{i,j,k}\right) - \left\langle\log_2\left(R_{i,j,k}/G_{i,j,k}\right)\right\rangle_i\right)^2/(n-1) \qquad (2)$$

gdzie $R_{i,j,k}$, $G_{i,j,k}$ stanowią intensywności sygnału fluorescencji dla kanału zielonego i czerwonego punktu k-tego powtórzenia, i-tej kontrolnej sondy na j-tej mikromacierzy, n=j*k, $\left\langle\log_2\left(R_{i,j,k}/G_{i,j,k}\right)\right\rangle_i$ odpowiada średniej wartości M dla i-tej sondy kontrolnej.

Podobna strategia została zastosowana do porównania jednokanałowych metod normalizacji. Założeniem są identyczne wartości intensywności sond dla kanału zielonego i czerwonego:

$$R_{i,j,k}/G_{i,j,k} \to 1 \Rightarrow \log_2\left(\frac{R_{i,j,k}}{G_{i,j,k}}\right) \to 0 \qquad (3)$$

Założeniem jest także stały poziom intensywności dla sond kontrolnych zarówno dla zielonego jak i czerwonego kanału. Stąd też: $\mathrm{R_{i,j,k}} \to \mathrm{const}$ and $\mathrm{G_{i,j,k}} \to \mathrm{const}$ k-tego powtórzenia oraz i-tej kontrolnej sondy na j-tej mikromacierzy. Z punktu widzenia analizy kluczowy jest kanał czerwony (próbki badane). Z założenia stałości poziomu intensywności dla sond kontrolnych wynika, że $\mathrm{R_{i,j,k}} \to \overline{\mathrm{R}}_\mathrm{i}$, gdzie $\bar{R}_i$ stanowi średnią wartość intensywności dla i-tej kontrolnej sondy dla czerwonego kanału. Korzystając z logarytmu umieszczonego w poprzednim wyrażeniu otrzymuje się następujące wyrażenie:

$$\log_2 R_{i,j,k} - \log_2 \bar{R}_i \to 0 \Rightarrow \log_2\left(\frac{R_{i,j,k}}{\bar{R}_i}\right) \to 0 \qquad (4)$$

Stąd też w celu obliczenia wartości błędu systematycznego oraz wariancji dla jednokanałowych metod normalizacji, we wzorach (1) i (2) wyrażenie $\log_2\left(R/G\right)_{(i,j,k)}$ zastąpione zostało wyrażeniem z (4). W wyniku opisanego podstawienia otrzymano następujące wyrażenia:

$$\mathrm{bs_i} = \sqrt{\sum_{\mathrm{j}}\sum_{\mathrm{k}}\left(\log_2\left(R_{i,j,k}/\overline{R}_i\right)^2\right)/n} \qquad (5)$$

$$\mathrm{wariancja_i} = \sum_{\mathrm{j}}\sum_{k}\left(\log_2\left(R_{i,j,k}/\overline{R}_i\right) - \left\langle\log_2\left(R_{i,j,k}/\overline{R}_i\right)\right\rangle_i\right)^2/(n-1) \qquad (6)$$

Taka sama procedura została powtórzona dla zielonego kanału.

V. Wyniki i Dyskusja

Rozdział „Wyniki i Dyskusja” podzielony został na cztery odrębne części. Każda część odpowiada potencjalnym problemom analizy niestandardowych danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA lub bezpośrednio procesom jakich dotyczą.

- Część I: *Ekspresyjne mikromacierze DNA o niestandardowym układzie sond*. Część ta dedykowana jest analizie danych uzyskanych w wyniku stosowania ekspresyjnych mikromacierzy DNA o niestandardowym układzie sond. W ramach tej części przedstawiono opis prostej metody standaryzacji formatu danych.
- Część II: *Analiza danych uzyskiwanych z wykorzystaniem dedykowanych mikromacierzy DNA do badania akumulacji różnicowej miRNA*. Część ta poświęcona jest specyfice analizy danych uzyskiwanych z wykorzystaniem dedykowanych mikromacierzy DNA do badania ekspresji miRNA. W ramach tej części zawarty został opis analizy danych otrzymanych w wyniku stosowania mikromacierzy DNA zawierających sondy o różnym stopniu komplementarności.
- Część III: *Normalizacja danych uzyskiwanych z użyciem dedykowanych mikromacierzy DNA*. Część ta poświęcona jest procesowi normalizacji danych otrzymywanych z użyciem dedykowanych mikromacierzy DNA. W ramach tej sekcji prezentowana jest siedmioetapowa, uniwersalna procedura wyboru optymalnej metody normalizacji dla danego zestawu danych.
- Część IV: *Analiza danych uzyskiwanych w wyniku hybrydyzacji międzygatunkowej*. Część ta dotyczy analizy danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA przy wykorzystaniu hybrydyzacji międzygatunkowej (CSH, ang. *cross-species hybridization*). W ramach tej części przedstawiono ocenę metody przekształcenia danych uzyskanych w ramach CSH w dane o cechach otrzymanych w ramach standardowej hybrydyzacji (SSH ang. *cross-species hybridization*) za pomocą filtracji danych opartej na morfologii punktów.

CZĘŚĆ I: Ekspresyjne mikromacierze DNA o niestandardowym układzie sond

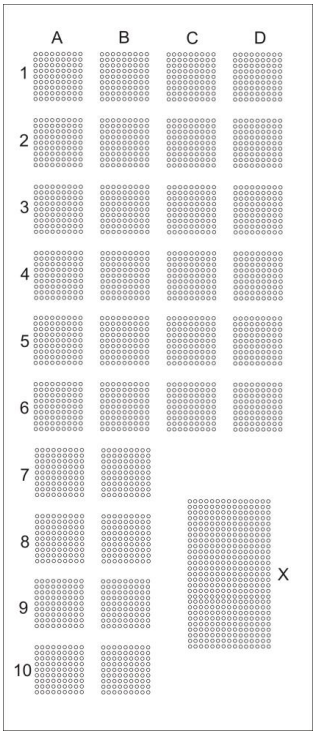
Drukowane mikromacierze DNA są rodzajem mikromacierzy najczęściej narażonym i podatnym na różnego rodzaju błędy konstrukcyjne (Liebert i wsp. 2002; Kooperberg i wsp. 2002; Tsai i wsp. 2005). Dostępny na rynku sprzęt laboratoryjny daje użytkownikowi pełną swobodę w projektowaniu eksperymentów z wykorzystaniem ekspresyjnych mikromacierzy DNA, jak i samych mikromacierzy. Najnowsze drukarki do mikromacierzy DNA pozwalają na przykład na produkcję mikromacierzy z dowolnym układem sond. Jednym z częstszych

błędów dla eksperymentów z użyciem drukowanych, ekspresyjnych mikromacierzy DNA jest właśnie problem niestandardowego układu sond, który wynika z nierównomiernego rozmieszczenia sond na powierzchni mikromacierzy. Z definicji mikromacierze powinny charakteryzować się równomiernym rozmieszczeniem sond w układzie blokowym o określonej liczbie kolumn i wierszy. Kwestia układu sond na ekspresyjnej mikromacierzy DNA nabiera znaczenia dopiero w kontekście analizy danych, gdyż przekłada się ona bezpośrednio na format całego zestawu danych. Dane uzyskiwane w ramach eksperymentów z użyciem ekspresyjnych mikromacierzy DNA, powinny mieć ściśle określoną strukturę, zdefiniowaną w postaci macierzy *m* o wymiarach [*i* x *j*], gdzie *i*- stanowi siatkę punktów i odpowiada liczbie sond zlokalizowanych na mikromacierzy, a *j*-liczbie mikromacierzy stosowanych w ramach eksperymentu. Siatka punktów jest ściśle zdefiniowana i zawiera informacje na temat układu sond na mikromacierzy. Układ sond zdefiniowany jest poprzez liczbę bloków i ich ułożenie (ilość kolumn i wierszy tworzona przez bloki na mikromacierzy) oraz wymiary każdego z bloków (ilość wierszy i kolumn tworzonych przez sondy). Zgodnie z ogólnie przyjętymi standardami jakości dla ekspresyjnych mikromacierzy DNA, bloki powinny być rozmieszczone w sposób regularny, a ich wymiary powinny być identyczne. Niestandardowy układ sond w przypadku ekspresyjnych mikromacierzy DNA ma znaczący wpływ na analizę danych, gdyż zaburza ogólnie przyjęty format macierzy *m*. Większość aktualnie dostępnych algorytmów do analizy danych jako dane wejściowe wymaga zestawu informacji w postaci macierzy *m*. Stosowanie ekspresyjnych mikromacierzy DNA o niestandardowym układzie sond, choć technicznie wykonalne, często uniemożliwia prowadzenie kompleksowej analizy danych z użyciem powszechnie dostępnych programów. Prowadzi to do niezgodności z opisywanymi we wstępie standardami jakości MIAME (Rozdział II.6.1)

V.I.1 Identyfikacja problemu

Problemem niestandardowego układu sond pojawił się w przypadku zestawu mikromacierzy DNA do badania ekspresji genów u pacjentów z ostrą białaczką szpikową (zestaw AML) (Materiały i Metody, rozdział IV.I.1.3). W skład zestawu wchodziły 172 ekspresyjne mikromacierze DNA, z których każda mikromacierz zawierała 32 bloki sond rozmieszczone w układzie 4 x 10 (kolumny x rzędy) (Rysunek 12). Prostokątny układ bloków o wymiarach 4 x 10 wskazuje na obecność 40 bloków sond. Jednakże w przypadku mikromacierzy z zestawu AML, bloki z rzędów 7-10 i kolumn C-D zostały zastąpione przez

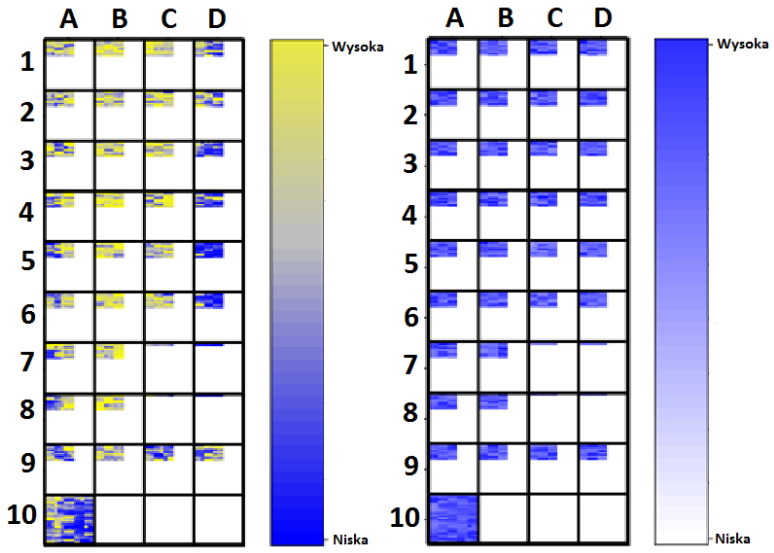
jeden duży blok X. Większość bloków zawiera 81 sond rozmieszczonych w układzie 9 x 9 (kolumny x rzędy). Natomiast blok X składa się z 405 sond rozmieszczonych w układzie 15 x 27.



Rysunek 12. *Projekt pojedynczej mikromacierzy DNA z zestawu AML. Kolumny i rzędy wyznaczają poszczególne bloki na mikromacierzy DNA. Blok X (15 x 27 sond.) znajduje się w miejscu, gdzie powinno być 8 bloków (każdy o wymiarach 9 x 9 sond).*

Niestandardowy układ sond skutkował otrzymaniem dla zestawu AML formatu danych, który ograniczał stosowanie większości dostępnych programów do analizy, w tym także R/Bioconductor. Chociaż program R/Bioconductor umożliwiał wczytanie „surowych danych" z zestawu AML, nie pozwalał on jednak na przeprowadzenie na tych danych podstawowych etapów analizy niższego rzędu. Przykładem ograniczeń wynikających z niestandardowego układu sond, był brak możliwości wykonania wykresów diagnostycznych typu *imageplot* dla „surowych danych" kanału zielonego i czerwonego oraz odpowiadających im wartości tła (R, pakiet `limma`, funkcja `imageplot()`). Ograniczenia pojawiły się także w przypadku wykresów diagnostycznych typu *imageplot* generowanych dla wartości M i A (R, pakiet `ArrayQualityMetrics`, funkcja `maQualityPlots()`). Pomimo możliwości wygenerowania, otrzymane wykresy typu *imageplot* dla wartości M i A miały nietypowy wygląd i były nieinformatywne (Rysunek 13). Kwadraty odpowiadające poszczególnym blokom nie zawierały kompletnej informacji na temat wartości M i A dla

poszczególnych sond. Wynika to z faktu, iż rozmiar bloków w formacie 9 x 9 (Rysunek 12) został przeskalowany względem wielkości bloku X. Ponadto, lokalizacja bloków na wykresie nie odpowiadała rzeczywistemu rozmieszczeniu bloków na mikromacierzy. Efekt ten jest rezultatem nietypowego, w stosunku do pozostałych bloków, położenia i rozmiaru bloku X, który na wykresie został błędnie umieszczony w pozycji A10 (Rysunek 13).



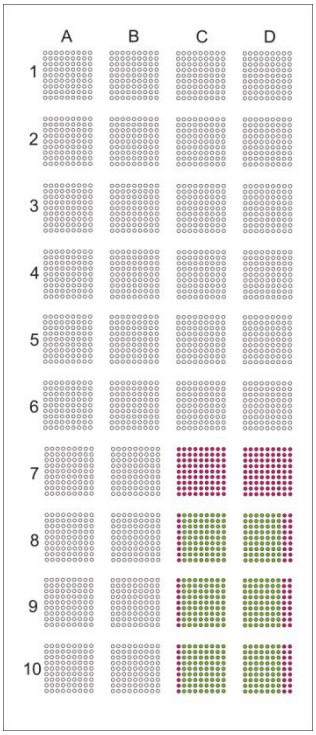
Rysunek 13. *Przykład wykresów diagnostycznych typu imageplot dla wartości M i A dla losowo wybranej mikromacierzy z zestawu AML przed wyrównaniem wymiaru. Lewy wykres w panelu przestawia obraz wartości M, a prawy wartości A.*

W przypadku analizy danych z zestawu AML pojawił się także problem z normalizacją danych. Największe trudności związane były z przeprowadzeniem normalizacji metodą *print-tip loess* oraz *quantile*. Metoda *print-tip loess* polega na normalizacji sond w obrębie bloków, które powstały poprzez nanoszenie sond daną igłą w procesie drukowania mikromacierzy (ang. *print-tip groups*). W związku z tym metoda ta wymaga stałego rozmiaru bloków. Procedura normalizacji typu *quantile* wymaga natomiast pełnej siatki punktów, która w przypadku zestawu AML odnosi się do 40 bloków z których każdy powinien zawierać 81 sond w układzie 9 x 9.

V.I.2 Rozwiązanie problemu

Niestandardowy format danych zestawu AML został skorygowany poprzez „wirtualny" podział bloku X (405 sond) na 6 mniejszych o wymiarze 9 x 9, tak aby na całej mikromacierzy zachowany był układ bloków 4 x 10. Jednakże blok X zawierał jedynie 405, a

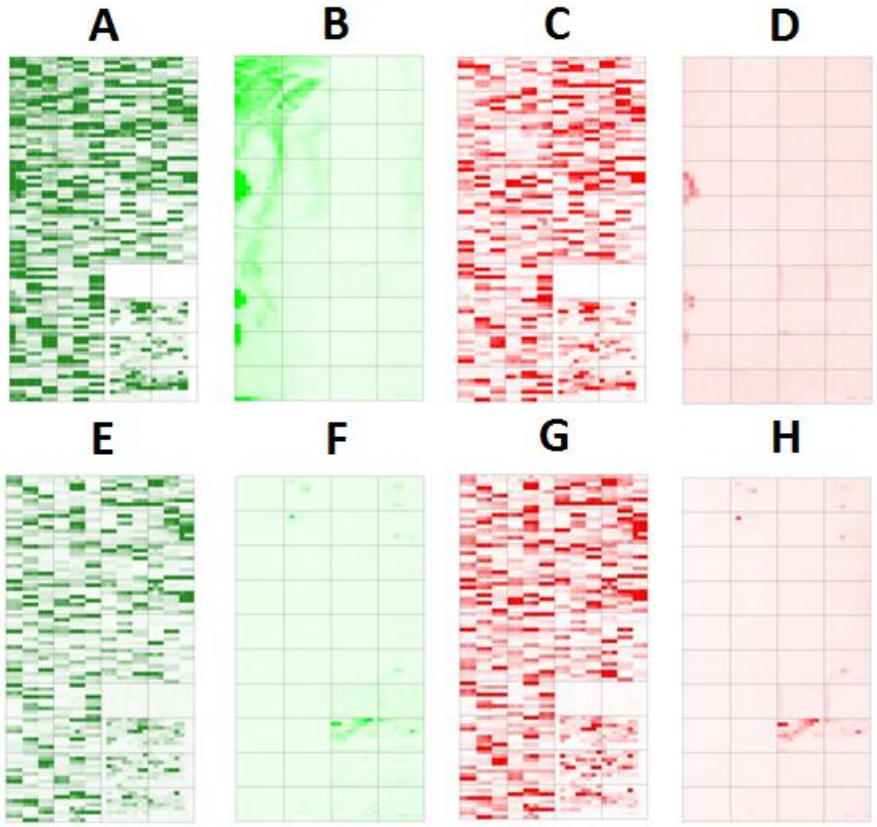
utworzenie 6 bloków sond o wymiarach 9 x 9, wymagało użycia 486 sond, co stanowiło ubytek 81 sond. Ponadto, przesunięcie bloku X w pozycję 6 brakujących bloków skutkowało powstaniem brakujących wartości dla dwóch bloków. Powstałe luki w strukturze danych uzupełnione zostały poprzez wstawienie do poszczególnych bloków średnich wartości tła (Rysunek 14). Średnie te zostały obliczone na podstawie informacji z plików .gpr (otrzymanych w wyniku analizy ilościowej obrazu za pomocą programu *GenePixPro 6.0*).



Rysunek 14. *Schemat obrazujący korektę formatu danych dla zestawu AML. Sondy z bloku X (kolor zielony) podzielono i przeniesiono do bloków C 8-10 i D 8-10. Powstałe luki w strukturze danych uzupełniono poprzez wypełnienie brakujących wartości średnią wartością tła (kolor różowy) obliczoną dla danego regionu mikromacierzy DNA (bloku).*

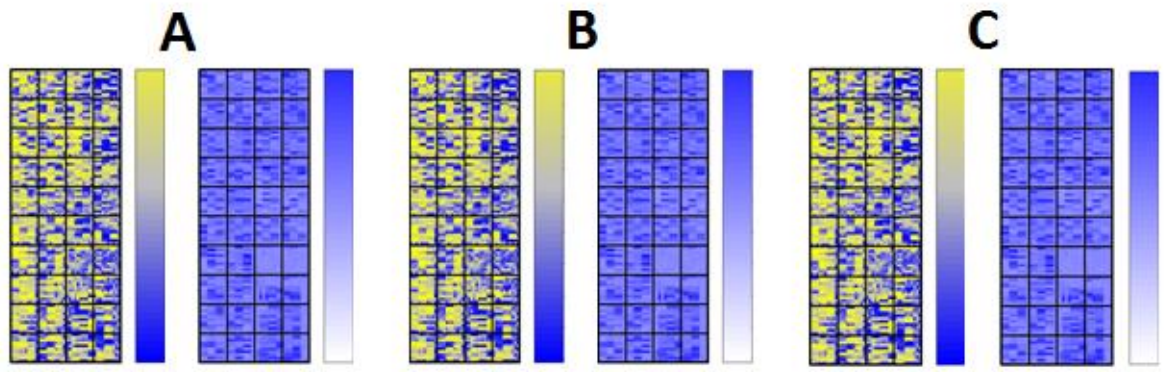
Natomiast całość procesu korekty formatu danych wykonana została na poziomie plików .gpr za pomocą przygotowanego algorytmu `wymiar_macierzy.py` (Załącznik 7). Algorytm ten przygotowany został w środowisku programistycznym Pyton, przy wykorzystaniu modłów *os* i *numpy*. Algorytm `wymiar_macierzy.py` po procesie korekty formatu danych do nazwy zmodyfikowanych plików .gpr wprowadzał ciąg znaków „fixed", który pozwalał na odróżnienie tych plików od plików źródłowych. Podejście to umożliwiło wykonanie operacji korekty formatu danych w pojedynczym katalogu przy zachowaniu oryginalnych plików źródłowych. Wynikiem działania algorytmu `wymiar_macierzy.py` był zestaw danych o standardowym formacie *m*.

Dzięki przedstawionemu procesowi korekty formatu danych możliwe było wykonanie, dla skorygowanych danych, wykresów diagnostycznych typu *imageplot* dla surowych danych kanału zielonego i czerwonego oraz wartości tła (R, pakiet `limma`, funkcja `imageplot`()) (Rysunek 15). Z rezultatów prezentowanych na Rysunku 16 wynika, iż proces korekty formatu danych pozwala na otrzymanie danych o strukturze zdefiniowanej przez standardy MIAME. Dodatkowo, jak wskazują wyniki, których przykłady przedstawione są na Rysunku 16, wszystkie modyfikacje plików .gpr wprowadzone przez algorytm `wymiar_macierzy.py` nie zaburzają charakteru i integralności danych. Wynika to z faktu, iż ilość wprowadzonych zmian do plików .gpr była minimalna, tzn. 6 bloków sond powstało w wyniku podziału istniejącego bloku X, a wprowadzone (w celu uzupełnienia brakujących wartości w zestawie danych) średnie wartości tła są na poziomie wartości tła obliczonych dla punktów w ramach danego bloku (Rysunek 15B oraz F, jak również Rysunek 15D oraz H).



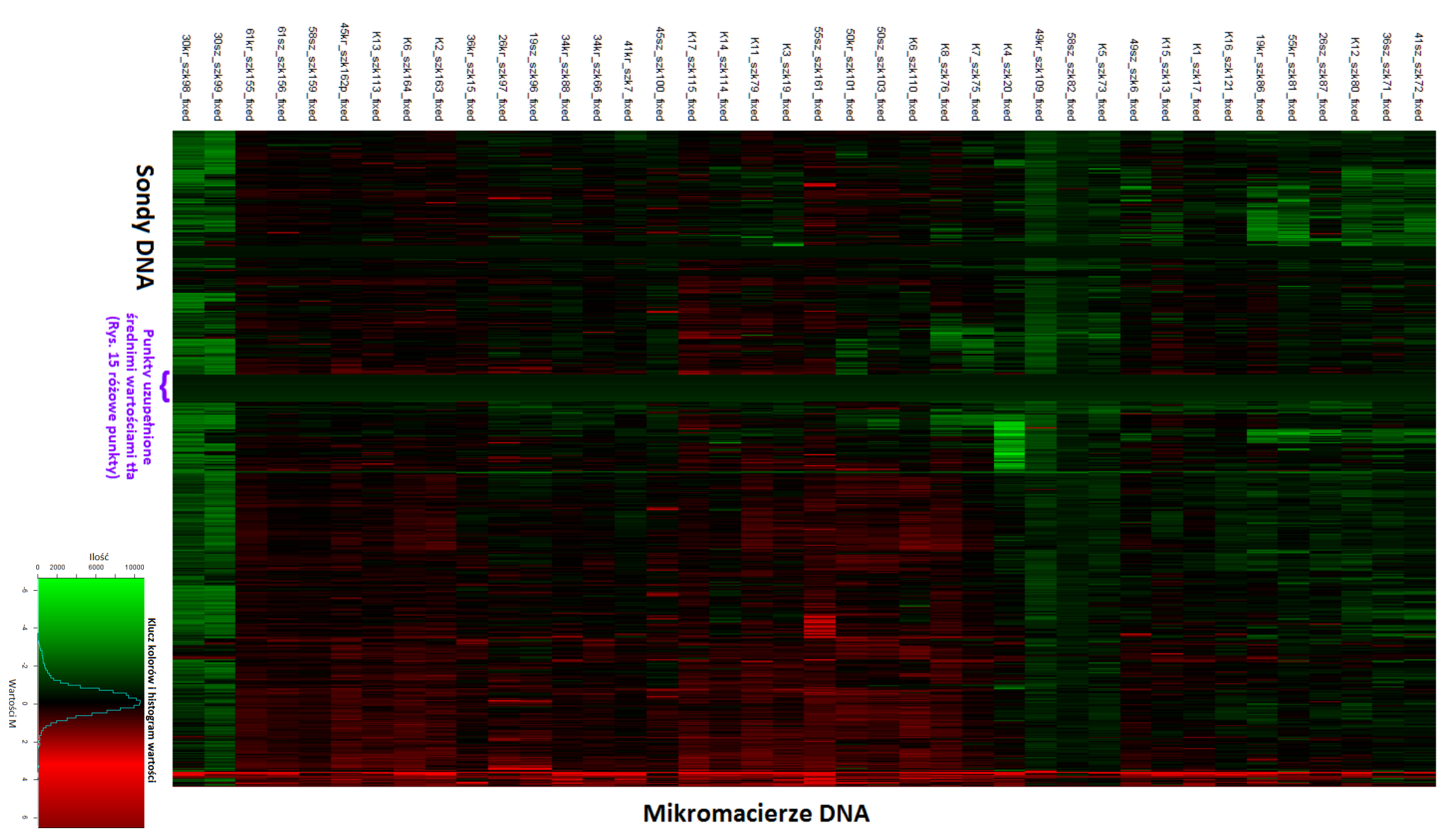
Rysunek 15. *Przykłady wykresów diagnostycznych dla 2 losowo wybranych mikromacierzy (odpowiednio A-D i E-H) z zestawu AML po procesie korekty formatu danych. Wykresy A oraz E przestawiają wartości intensywności sygnału fluorescencyjnego kanału zielonego. Wykresy B oraz F przestawiają wartości tła kanału zielonego. Wykresy C oraz G przestawiają wartości intensywności sygnału fluorescencyjnego kanału czerwonego. Wykresy D oraz H przestawiają wartości tła kanału czerwonego. Proszę zwrócić uwagę, iż dla wygresów intensywności sygnałów głównych (kanału zielonego: wykresy A oraz E; kanał czerwony: wykresy C oraz G) oraz wykresów tła (kanał zielony: wykresy B oraz F; kanał czerwony: wykresy D oraz H) zastosowany został inny kontrast.*

Proces korekty formatu danych pozwolił również na otrzymanie kompletnych wykresów typu *imageplot* dla wartości M i A (R, pakiet *ArrayQualityMetrics,* funkcja `maQualityPlots()`) (Rysunek 16 A-C). Z rezultatów prezentowanych na Rysunku 17 wynika, iż wszystkie bloki mają równy rozmiar oraz ściśle zdefiniowane położenie, zgodne z ich rzeczywistym układem, w przeciwieństwie do wyników prezentowanych dla danych przed korektą formatu (Rysunek 13 i Rysunek 16).



Rysunek 16. *Przykłady wykresów diagnostycznych dla trzech losowo wybranych mikromacierzy z zestawu AML po procesie korekty formatu danych (A-C). W każdym z paneli lewy wykres przestawia obraz wartości M, a prawy wartości A.*

Aby ocenić wpływ korekty formatu danych na strukturę całego zestawu AML wykonano analizę rozkładu danych za pomocą mapy cieplnej (R, pakiet `gplots`, funkcja `heatmap.2()`) (Rysunek 18). Mapę cieplną przygotowano dla wszystkich 172 mikromacierzy z zestawu AML w oparciu o wartości M. Ze względu na przejrzystość, w pracy przedstawiono wyniki dla 40 losowo wybranych mikromacierzy z zestawu AML (Rysunek 17). Wartości M wybrano zamiast „surowych danych", gdyż są to najbardziej charakterystyczne wartości dla analizy danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA. Rozkład wartości M otrzymany dla danych po korekcji formatu ma charakter rozkładu normalnego (Rysunek 17, histogram wartości M) i przypomina typowe wyniki dla eksperymentów z użyciem ekspresyjnych mikromacierzy DNA. Oznacza to, iż opisywany proces korekty formatu danych pozwala na otrzymanie zestawu danych spełniającego standardy MIAME przy wprowadzeniu minimalnej zmienności do układu eksperymentalnego.

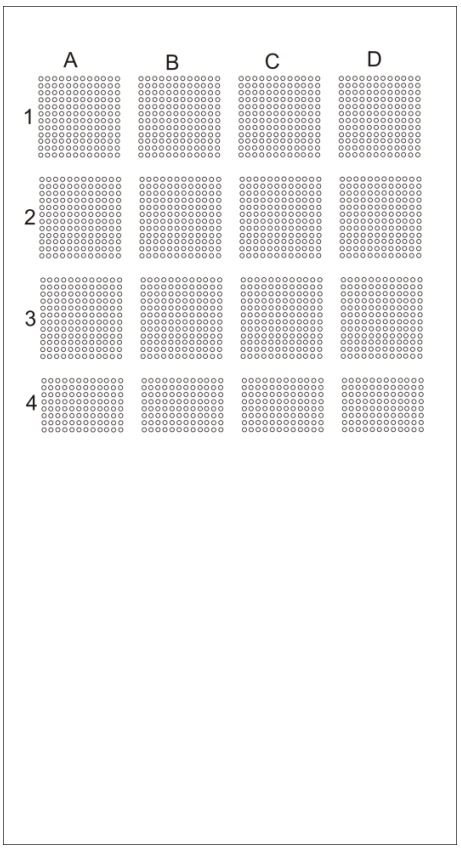


Rysunek 17. *Mapa cieplna obrazująca rozkład oraz wartości M dla 40 losowo wybranych mikromacierzy z zestawu AML. Na osi X znajdują się poszczególne mikromacierze DNA z zestawu AML po procesie wyrównania wymiaru. Na osi Y przedstawione są sondy DNA. Kolorami oznaczono wartości M. Mapa cieplna wykonana została dla danych po procesie korekty formatu.*

V.I.3 Inne przykłady zestawów danych o niestandardowym układzie sond

Podobny problem niestandardowego układu sond pojawił się także w przypadku dwóch innych zestawów mikromacierzy DNA: do badania ekspresji genów u dzieci z alergią krzyżową (zestaw ALERGIA) (Materiały i Metody, rozdział IV.I.3) oraz do badania ekspresji genów u dzieci z astmą (zestaw ASTMA) (Materiały i Metody, rozdział IV.I.4). Oba rodzaje danych zostały wygenerowane w eksperymentach z wykorzystaniem ekspresyjnej mikromacierzy DNA tego samego projektu. Mikromacierz DNA do badań nad alergią krzyżową i astmą u dzieci obejmuje sondy w układzie 12 bloków (4 kolumny x 3 rzędy). Przy czym bloki w rzędzie 1-3 zawierają 12 kolumn i 12 rzędów sond (12 x 12), natomiast bloki z rzędu 4 zawierają 12 kolumn i jedynie 8 rzędów sond (12 x 8) (Rysunek 18).

Niestandardowy format danych zestawów ALERGIA i ASTMA został z powodzeniem skorygowany przy wykorzystaniu prezentowanej metody standaryzacji formatu danych. Brakujące wartości w blokach z rzędu 4 zostały uzupełnione poprzez wprowadzenie średnich wartości tła, obliczonych dla danego regionu mikromacierzy.



Rysunek 18. *Projekt pojedynczej mikromacierzy DNA z zestawu ALERGIA i ASTMA.*

V.I.4 Przykłady wykorzystania wyników

Proces korekty formatu danych zarówno dla zestawu danych AML, jak i zestawów ALERGIA oraz ASTMA, pozwolił na przywrócenie struktury zdefiniowanej przez standardy jakości MIAME przyjęte dla eksperymentów z użyciem ekspresyjnych mikromacierzy DNA. Dzięki prezentowanej metodzie korekty formatu danych możliwe było przeprowadzenie analiz wyżej wymienionych zestawów danych i uzyskanie następujących wyników o znaczeniu biologicznym. Otrzymane wyniki ostatecznie potwierdziły skuteczność zastosowanej korekty. Analiza skrygowanych zestawów danych pozwoliła na:

1. Identyfikację genów potencjalnie zaangażowanych w proces rozwoju ostrej białaczki szpikowej.
2. Analizę różnic w profilu ekspresji genów pomiędzy podtypem M1 i M2 AML.
3. Wykazanie silnej zależności pomiędzy profilem ekspresji genów dla próbek krwi i szpiku pacjentów z AML. Wynik ten wskazuje na możliwość zastąpienia próbek szpiku próbkami krwi, znacznie łatwiejszymi do pobrania.
4. Identyfikację różnic w profilu ekspresji genów pomiędzy grupą zdrowych ochotników, a pacjentami z alergią krzyżową i astmą (każda z grup zawiera 12 pacjentów w wieku 3-17 lat).

Wymienione powyżej wyniki zostały uzyskane w ramach odrębnych projektów, które nie są przedmiotem prezentowanej pracy.

V.I.5 Omówienie wyników

Mikromacierze DNA są powszechnie stosowanym narzędziem do badania ekspresji genów. Należy jednak pamiętać, że eksperymenty z użyciem ekspresyjnych mikromacierzy DNA powinny być projektowane z najwyższą uwagą i precyzją oraz zgodnie z powszechnie obowiązującymi standardami jakości, np. MIAME. Wynikiem eksperymentów z użyciem ekspresyjnych mikromacierzy DNA jest złożony zestaw danych o określonym formacie (macierz *m*), który determinowany jest przez szereg parametrów, takich jak: liczba sond, ilość barwników użytych do znakowania próbek oraz liczba mikromacierzy użytych w ramach eksperymentu. Spośród wymienionych parametrów, najbardziej znaczący wpływ na format danych ma sam projekt mikromacierzy, w szczególności układ sond.

Niestandardowy układ sond często jest skutkiem projektu mikromacierzy, w którym zamiast bloków sond tworzone są tzw. podmacierze (ang. *sub-array*), w skład których wchodzą jedynie ściśle określone typy sond. Taki projekt mikromacierzy może wynikać np. z próby wykorzystania mikromacierzy do kilku niezależnych projektów. Podejście to sprawia jednak, iż dane podmacierze mogą się różnić między sobą zawartością sond, a tym samym i rozmiarem. Wiele niestandardowych układów sond jest także skutkiem stosowania zbyt małej liczby sond w procesie drukowania mikromacierzy, która nie jest wystarczająca do otrzymania pełnej siatki sond i skutkuje powstaniem bloków różniących się rozmiarem (np. zestawy ALERGIA i ASTMA). Przyczyną zaburzenia układu sond mogą być także trudności na etapie drukowania mikromacierzy, m.in. uszkodzenie igły, lepkość roztworu sondy, problemy na etapie przyłączania się sondy do podłoża (np. brak łącznika) oraz charakterystyka podłoża (np. ilość grup funkcyjnych). Dane otrzymane w wyniku stosowania ekspresyjnych mikromacierzy DNA o zaburzonym układzie sond mogą mieć inny format, tzn. inną strukturę macierzy *m*, niż ta zdefiniowana pośrednio przez standardy jakości MIAME. Standardy MIAME stanowią dodatkową kontrolę i pozwalają na zwiększenie jakości wyników otrzymywanych z użyciem ekspresyjnych mikromacierzy DNA. Jednakże w praktyce nie zawsze eksperymenty z wykorzystaniem ekspresyjnych mikromacierzy DNA spełniają te rygorystyczne wymagania, zwłaszcza w przypadku stosowania dedykowanych mikromacierzy DNA. Wynika to z faktu, iż niektóre wady mogą okazać się znaczące dopiero na etapie analizy danych. Ze względu na znaczne koszty eksperymentu oraz czas poświęcony na jego wykonanie, często pożądane jest otrzymanie informacji biologicznej także z niestandardowych zestawów danych.

W tej części pracy przedstawiono trzy przykłady danych o niestandardowym układzie sond: zestaw AML, ASTMA oraz ALERGIA. W przypadku wszystkich prezentowanych zestawów format macierzy *m* odbiegał od ogólnie przyjętych standardów. Poszczególne bloki sond, wyznaczone przez igły do drukowania mikromacierzy, różniły się między sobą wielkością, a w przypadku zestawu AML także i lokalizacją. Analiza tego rodzaju danych jest utrudniona nawet przy zastosowaniu programów posiadających cechy środowiska programistycznego. W przypadku zestawów AML, ASTMA oraz ALERGIA program R/Bioconductor pozwolił na wczytanie każdego z zestawów danych w oryginalnym formacie, jednakże dalsza analiza tych danych nie była możliwa. Wynika to z faktu, iż część funkcji dedykowanych analizie danych uzyskiwanych z użyciem mikromacierzy, zdeponowanych w ramach R/Bioconductor wymaga obiektów w postaci macierzy *m*. Dotyczy to zwłaszcza

procesu normalizacji danych. Większość metod normalizacji danych, które usuwają zmienność wynikającą z różnej lokalizacji sond, tzw. efekt przestrzenny (np. metoda *print-tip loess*) wymaga bloków sond o identycznym rozmiarze.

Potencjalnym rozwiązaniem kwestii normalizacji danych o niestandardowym układzie sond jest zastosowanie globalnej metody normalizacji. W przypadku globalnych metod normalizacji różnice w rozmiarze bloków oraz lokalizacja sond nie odgrywają kluczowej roli. Ograniczeniem stosowania tego rodzaju metod normalizacji jest jednak pominięcie w procesie normalizacji różnic intensywności sygnałów wynikających z różnej lokalizacji sond na mikromacierzy, co może w znacznym stopniu obniżać skuteczność tego etapu. Ponadto, zastosowanie globalnej metody normalizacji nie stanowi faktycznego rozwiązania kwestii niestandardowego układu sond. Utrudnienia w prowadzeniu analizy danych mogą się pojawić w przypadku realizacji innych etapów, dla których układ sond może mieć znaczenie, np. generowanie wykresów diagnostycznych.

Bradziej kompleksowym rozwiązaniem problemu analizy danych o niestandardowym układzie sond jest modyfikacja formatu tych danych. Istnieją dwa główne sposoby modyfikacji formatu danych, które pozwalają na zachowanie kompletnej informacji pochodzącej ze wszystkich sond ulokowanych na mikromacierzy: (I) podział mikromacierzy na symetryczne części i indywidualna analiza każdej z nich lub (II) korekta formatu danych. Choć oba z tych podjeść są poprawne statystycznie, korekta formatu danych pozwala na zintegrowaną analizę całego zestawu danych i nie powoduje obniżenia specyficzności tego procesu, jak to z reguły ma miejsce w przypadku podziału mikromacierzy na poszczególne części. W celu przywrócenia zestawom danych AML, ASTMA i ALERGIA standardowego układu zastosowana została metoda korekty formatu danych. W przypadku każdego z tych zestawów korekta formatu danych wiązała się z uzupełnieniem brakujących wartości (luk). Kwestia uzupełniania luk w zestawach danych uzyskanych z użyciem ekspresyjnych mikromacierzy DNA jest bardzo złożona. W takim przypadku kluczowe jest zachowanie spójności i integralności struktury danych (Tuikkala i wsp. 2008). Istnieje szereg algorytmów dedykowanych wypełnianiu brakującej informacji, m.in. metoda oszacowania największej wiarygodności (ang. *Maximum likelihood estimation*), szacowanie bayesowskie (ang. *Bayesian estimation*), czy metody wielokrotnego przypisania (ang. *Multiple imputation*). Choć metody te są bardzo zaawansowane statystycznie, żadna z nich nie gwarantuje otrzymania wyników wolnych od błędów systematycznych. W przypadku korekty formatu danych dla zestawów w których liczba brakujących wartości jest niewielka, liczy się sama

obecność danego punktu, a jego wartość intensywności sygnału fluorescencyjnego ma drugorzędne znaczenie. W związku z tym, do uzupełnienia luk w skorygowanych zestawach danych zastosowana została najprostsza metoda, polegająca na wypełnieniu brakujących wartości średnimi wartościami tła. Aby zachować spójność struktury danych na etapie ich przetwarzania, nowopowstałe punkty zostały wykluczone z dalszych etapów analizy poprzez nadanie im statusu (flagi) punktów o niskiej jakości. Podejście to sprawia, iż udział wprowadzonych, średnich wartości tła na przebieg procesu normalizacji i identyfikacji genów różnicujących jest minimalny.

Podsumowując, prezentowany sposób korekty formatu danych dla analizowanych zestawów danych AML, ASTMA oraz ALERGIA obejmował zmianę układu bloków (jeśli była niezbędna), wyrównanie ich rozmiaru i uzupełnienie brakujących wartości średnimi wartościami tła dla danego regionu mikromacierzy. Otrzymane wyniki wskazują, iż proponowany sposób korekty formatu danych pozwala na otrzymanie danych w postaci standardowej macierzy *m* przy wprowadzeniu minimalnej zmienności do układu (Rysunek 17). Analiza wyższego rzędu skorygowanych zestawów danych pozwoliła na uzyskanie istotnych biologicznie rezultatów. Część tych wyników, dotyczących identyfikacji genów różnicujących, została potwierdzona przy pomocy alternatywnej metody, ilościowego PCR.

V.I.6 Wnioski

Głównym wnioskiem wynikającym z prowadzonej analizy jest stwierdzenie, iż korekta formatu danych uzyskanych z użyciem ekspresyjnych mikromacierzy DNA o niestandardowym układzie sond może być przeprowadzona na etapie analizy danych. Dzięki temu możliwe jest przywrócenie tym danym formatu zdefiniowanego w ramach standardów jakości przy wprowadzeniu minimalnej zmienności do układu.

CZĘŚĆ II: Analiza danych uzyskiwanych z wykorzystaniem dedykowanych mikromacierzy DNA do badania ekspresji miRNA

MikroRNA (miRNA) są ważnymi regulatorami ekspresji genów, które kontrolują aktywność genów w fazie post-transkrypcyjnej. Monitorowanie poziomu ekspresji miRNA jest niezwykle ważnym elementem analizy zarówno w kontekście procesów fizjologicznych, jak i patologicznych. Badanie ekspresji miRNA jest wyzwaniem dla powszechnie stosowanych technik określania poziomu ekspresji genów z punktu widzenia specyficzności i dokładności. Wynika to głównie z niewielkiej długości sekwencji (ok. 22 nukleotydów) dojrzałych miRNA, jak również z faktu, iż niecałkowicie przetworzone formy miRNA, tzw. prekursory miRNA także zawierają sekwencje dojrzałych cząsteczek miRNA. Ponadto, znacznym utrudnieniem na etapie badania ekspresji miRNA jest występowanie blisko spokrewnionych członków danej rodziny miRNA, zwłaszcza w obrębie genomu ssaków, których sekwencje często różnią się jedynie pojedynczym nukleotydem (Roush & Slack 2008). Powyższe cechy miRNA wymagają, aby eksperymenty obejmujące analizę ekspresji tych cząsteczek z wykorzystaniem mikromacierzy DNA, projektowane były w specyficzny sposób. Stwierdzenie to w szczególności dotyczy sposobu projektowania sond. Temperatura topnienia sond dla fragmentu genu kodującego białka jest normalizowana poprzez odpowiedni wybór regionu genu oraz sterowanie długością sondy. Takie podejście nie jest możliwe w przypadku miRNA ze względu na niewielką długość ich sekwencji. Stąd też często w przypadku sond dla miRNA temperatura topnienia normalizowana jest w specyficzny sposób np. poprzez ligację odpowiednich sekwencji adaptora (ang. *adaptor sequence*) (Baskerville & Bartel 2005) czy stosowanie sond w postaci LNA (ang. *locked nucleic acid*) (Castoldi i wsp. 2006).

Niezależnie od sposobu projektowania sond, mikromacierze do badania ekspresji miRNA posiadają kilka cech wspólnych, takich jak: relatywnie duża ilość sond (do 60 000), występowanie na jednej mikromacierzy sond dla miRNA kilku gatunków oraz sond dla hipotetycznych sekwencji miRNA (C.-G. Liu, Calin, i wsp. 2008; W. Li & Ruan 2009; Goff i wsp. 2005). Powyższe cechy miRNA sprawiają, że procedury stosowane podczas badania ekspresji genów kodujących białka z użyciem mikromacierzy DNA nie mogą być w bezpośredni sposób wykorzystane do badania ekspresji miRNA. Fakt ten znacząco wpływa na proces analizy danych.

V.II.1 Identyfikacja problemu

Zestaw AML miRNA został otrzymany w wyniku eksperymentów obejmujących badanie ekspresji ludzkich miRNA u pacjentów z ostrą białaczką szpikową (Materiały i Metody, Rozdział IV.I.2). W skład zestawu AML miRNA wchodzi 30 ekspresyjnych mikromacierzy DNA, otrzymanych w wyniku dwukolorowego eksperymentu w którym wyjściowymi próbkami badanymi była niskocząsteczkowa frakcja RNA pochodząca od pacjentów z AML (kanał czerwony, barwnik Alexa 647), natomiast próbkami referencyjnymi były odpowiednie frakcje RNA z linii komórkowej HL60 (kanał zielony, barwnik Alexa 546).

Ekspresyjna mikromacierz DNA w oparciu o którą otrzymano zestaw AML miRNA, została przygotowana z użyciem komercyjnie dostępnego zestawu sond (*NCode™ Mammalian miRNA Microarray Probe Set v. 1.0,* Invitrogen). W skład tego zestawu wchodziły sondy przygotowane na podstawie sekwencji dojrzałych miRNA opisanych w bazie Sanger miRBase 7.0 (http://microrna.sanger.ac.uk) dla człowieka (*Homo sapiens*) (311 sond), myszy (*Mus musculus*) (232 sondy) oraz szczura (*Rattus norvegicus*) (185 sond). Ponadto, zestaw zawierał dodatkowe sondy dla sekwencji hipotetycznych ludzkich miRNA (sondy HMP_PREDICTED) (142 sondy).

Choć ze względu na wysoką konserwatywność, sekwencje miRNA trzech spokrewnionych gatunków: myszy, szczura i człowieka wykazują wysoką homologię względem siebie, to jednak nie zawsze musi być ona pełna (Sewer i wsp. 2005). W związku z powyższym pojawia się pytanie o możliwość wykorzystania ortologicznych sond dla miRNA myszy i szczura do analizy ekspresji ludzkich miRNA. Potencjalne sposoby użycia tych sond, w przypadku zestawu AML miRNA, obejmują:

1. Walidację wyników z użyciem ortologicznych sond jako replik,
2. Badanie ekspresji ludzkich miRNA z użyciem wszystkich sond, również tych, które nie znajdują odpowiednika u ludzi,
3. Identyfikację nowych ludzkich miRNA przy wykorzystaniu homologicznych sond dla myszy i szczura.

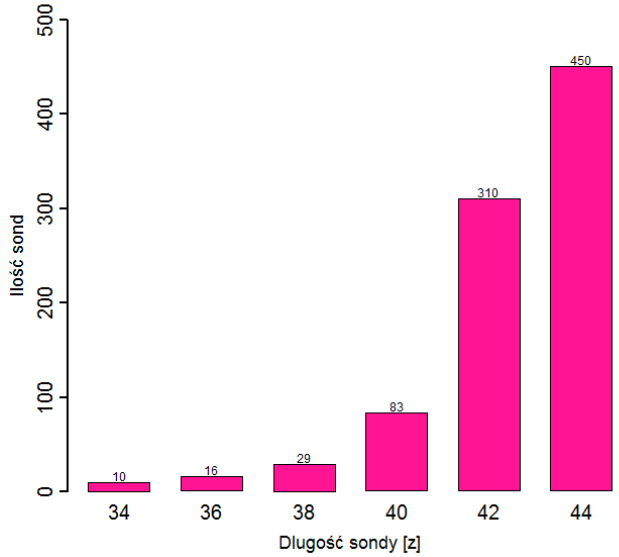
Analiza ludzkich miRNA z wykorzystaniem ortologicznych sond dla miRNA myszy i szczura jest możliwa jedynie w przypadku, gdy homologia pomiędzy sekwencjami sond, a sekwencjami docelowymi jest pełna lub gdy ewentualne różnice w sekwencjach mają nieznaczny wpływ na przebieg reakcji hybrydyzacji. Celem poniższej analizy było

sprawdzenie możliwości wykorzystania wszystkich sond ulokowanych na mikromacierzy DNA do badania ekspresji ludzkich miRNA u pacjentów z AML.

V.II.2 Rozwiązanie problemu

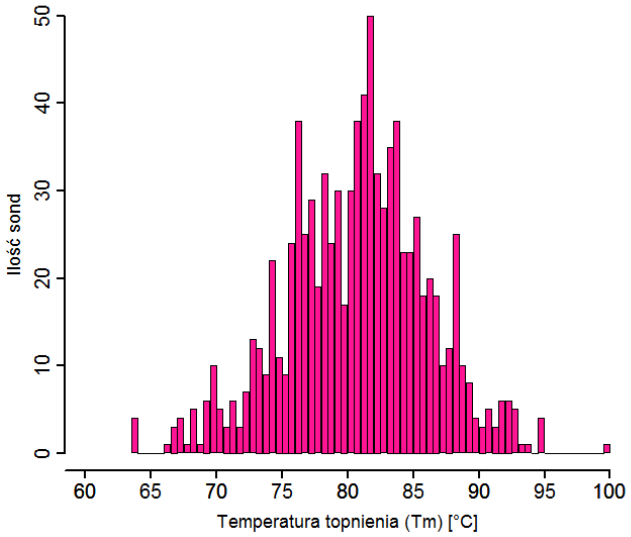
V.II.2.1 Wstępna charakterystyka sond

Pierwszym etapem sprawdzenia możliwości wykorzystania wszystkich sond ulokowanych na mikromacierzy DNA do badania ekspresji ludzkich miRNA u pacjentów z AML była wstępna charakterystyka użytych sond. Wstępną charakterystykę sond wykonano poprzez określenie zakresu i rozkładu długości ich sekwencji oraz temperatury topnienia. Jak wynika z wykresu przedstawionego na Rysunku 20, użyte w eksperymencie oligonukleotydowe sondy DNA różnią się między sobą długością (34 – 44 nukleotydów). Przy czym długość zdecydowanej większości sond (94%) zawiera się w zakresie 40 – 44 nukleotydów. Zarówno zakres długości sond, jak i parzyste wartości długości wynikają ze sposobu konstrukcji sond i odpowiadają podwójnej długości sekwencji poszczególnych miRNA. Każda sonda składa się z dwukrotnie powtórzonej sekwencji odpowiednich miRNA.



Rysunek 19. *Wykres ilustrujący zróżnicowanie sond względem ich długości.*

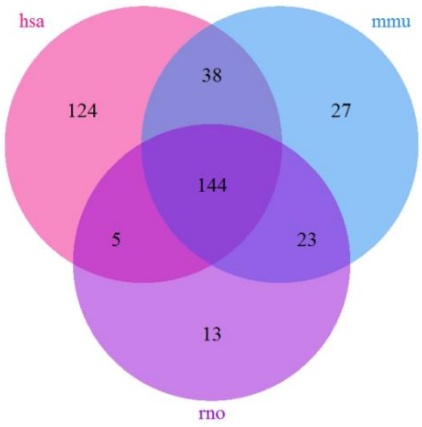
Duży poziom zmienności wśród badanych sond można także zaobserwować pod względem ich temperatury topnienia. Zakres temperatury topnienia dla badanych sond wynosi 63,7 – 99,9^{o}C (Rysunek 20), co oznacza, iż maksymalna różnica temperatury topnienia w zestawie wynosi 36,2 ^{o}C. Przy czym temperatura topnienia dla zdecydowanej większości (82%) sond mieści się w zakresie 75 – 90 ^{o}C. Średnia temperatura topnienia dla użytego zestawu sond wynosi 80,5 ^{o}C.



Rysunek 20. *Wykres ilustrujący zróżnicowanie sond względem ich temperatury topnienia.*

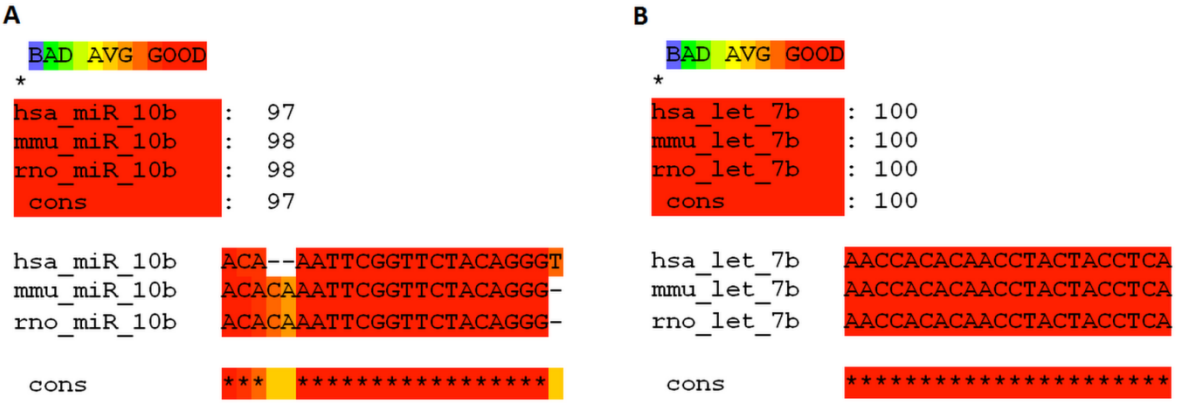
V.II.2.2 Identyfikacja i porównanie sond dla ortologicznych miRNA człowieka, myszy i szczura

Kolejnym etapem po procesie charakterystyki sond pod względem ich długości i temperatury topnienia była analiza homologii sekwencji sond miRNA dla trzech badanych gatunków. Etapem poprzedzającym tę analizę była selekcja podzestawu sond MIRNA3 w którym każda sonda posiadała trzy odpowiedniki, w postaci sond z zestawu człowieka, myszy i szczura. Selekcja sond wchodzących w skład podzestawu MIRNA3 spośród całkowitej liczby 728 sond (311 dla człowieka, 232 dla myszy oraz 185 dla szczura) obejmowała dwa etapy. W ramach pierwszego z nich porównane zostały nazwy sond, a następne wybrane te, które charakteryzowały się taką samą nazwą miRNA, np. hsa-miR7a, mmu-miR7a, rno-miR7a (hsa – człowiek, mmu – mysz, rno – szczur). Etap ten pozwolił na selekcję 142 trójek sond, posiadających odpowiedniki dla trzech badanych gatunków. W ramach drugiego etapu wykonana została analiza podobieństwa sekwencji pozostałych, różniących się nazwą sond. Analiza obejmowała kompleksowe, wzajemne porównanie 586 sekwencji sond. Drugi etap selekcji pozwolił na identyfikację 2 dodatkowych trójek. Sondy te choć posiadały różne nazwy, wykazywały pełną homologię pomiędzy porównywanymi gatunkami. W skład podzestawu MIRNA3 wchodziły, więc 144 trójki sond. Proces selekcji podzestawu MIRNA3 pozwolił także na określenie ilości sond wspólnych dla par poszczególnych gatunków. Dane te przedstawione zostały za pomocą Diagramu Venna (Rysunek 21).



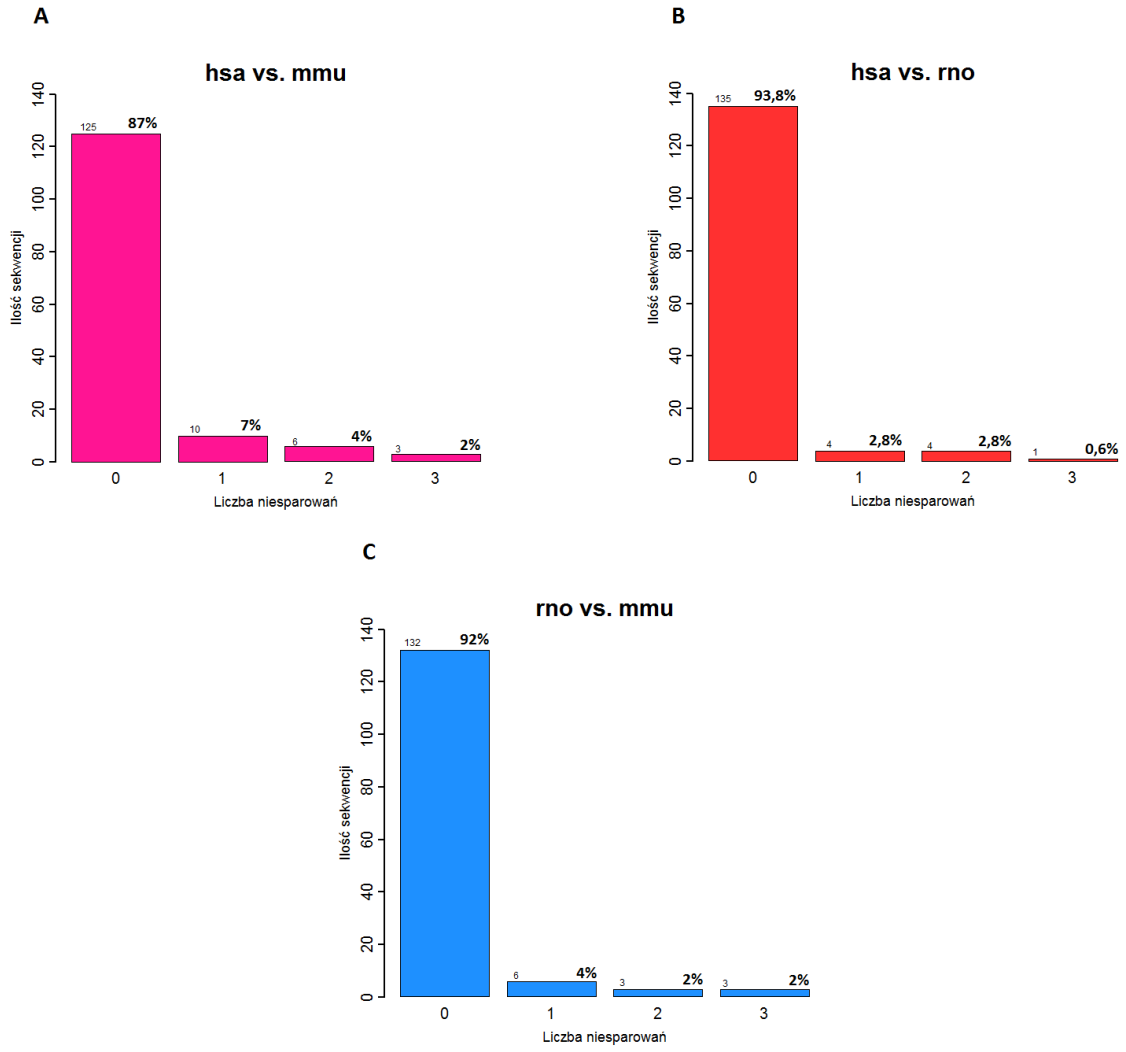
Rysunek 21. *Diagram Venna przestawiający porównanie sond dla trzech badanych gatunków: hsa – człowiek (korlor różowy), mmu – mysz (kolor niebieski), rno – szczur (kolor fioletowy). Połączenia poszczególnych zbiorów określają liczbę sond wspólnych dla danych gatunków. Dla wszystkich trzech gatunków zidentyfikowano 144 wspólne typy sond.*

W celu określenia rzeczywistego podobieństwa sekwencji dla miRNA, porównano sekwencje odpowiadających sobie sond z podzestawu MIRNA3. Porównanie sekwencji zostało wykonane za pomocą autorskiego algorytmu `przyrownanie_sekwencji.py` (Załącznik 8). Algorytm ten został przygotowany w środowisku programistycznym Python przy użyciu modułu *Biopython* (*Bio.Align.Applications*) oraz programu *TCoffee* umożliwiającego analizę porównawczą wielu sekwencji (ang. *multiple sequence aligment*). Przykład porównania sekwencji miRNA został przedstawiony na Rysunku 23. Jako, że sondy składały się z dwukrotnie powtórzonych tych samych sekwencji miRNA, analiza została przeprowadzona tylko na połówkach sekwencji sond, odpowiadających całkowitej sekwencji badanego miRNA.



Rysunek 22. *Wynik porównania sekwencji sond dla miRNA_10b (A) oraz let_7b (B) dla człowieka (hsa), myszy (mmu) i szczura (rno).*

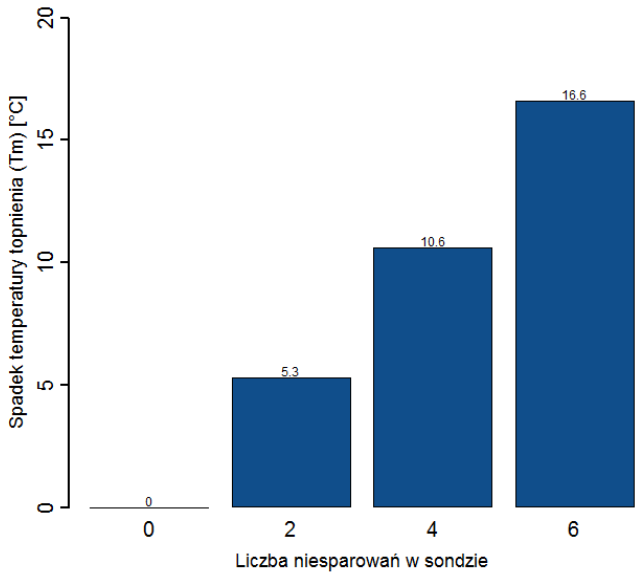
Wyniki analizy podobieństwa sekwencji ortologicznych sond dla miRNA wskazują, iż większość badanych sekwencji (87% – 93,8%) wykazuje pełną homologię między parami porównywanych gatunków (Rysunek 23 A-C). Różnice pomiędzy sekwencjami miRNA należących do poszczególnych trójek występują na poziomie od 1 do 3 nukleotydów, co odpowiada różnicy od 2 do 6 nukleotydów dla odpowiednich sond. Przy czym znaczna część sekwencji (ponad 80%), spośród tych zawierających niesparowania (ang. *mismatch*) różni się 1 – 2 nukleotydami (2 – 4 nukleotydów dla sond).



Rysunek 23. *Rozkład ilości sond miRNA pod względem liczby niesparowań (ang. mismatch), A. porównanie wykonane dla sekwencji sond dla miRNA człowieka oraz myszy; B. porównanie wykonane dla sekwencji sond dla miRNA człowieka oraz szczura; C. porównanie wykonane dla sekwencji sond dla miRNA szczura oraz myszy. Analizę podobieństwa sekwencji wykonano dla sond z podzestawu MIRNA3, zawierającego 144 rodzaje sond wspólnych dla trzech badanych gatunków.*

Aby w przybliżony sposób ocenić jak obecność niesparowań może wpływać na efektywność hybrydyzacji niecałkowicie komplementarnych sond, przeprowadzona została symulacja temperatury topnienia dla heterodupleksów sonda-sekwencja docelowa,

zawierających różną ilość niesparowań (Rysunek 24). Temperatura topnienia dla heteroduplеksów i homoduplеksów (sekwencja docelowa w pełni komplementarna do sondy) określona została za pomocą programu do obliczania temperatury topnienia na stronie Baker Lab Uniwersity of Washington (http://depts.washington.edu). Wyniki przedstawione na Rysunku 25 wskazują, iż heteroduplеksy charakteryzują się obniżoną temperaturą topnienia w stosunku do homoduplеksów. Spadek temperatury topnienia jest proporcjonalny do ilości niesparowań. Średni spadek temperatury topnienia wynosi 2,5 °C na jedno niesparowanie. Dla sond posiadających od 2 – 6 niesparowań, przewidywany poziom obniżenia temperatury topnienia wynosi 5,3 – 16,6 °C. Chociaż wartości te stanowią od kilku do kilkunastu procent średnich wartości temperatury topnienia dla analizowanych sond, to jednak stanowią one niewielki udział w zmienności temperatury topnienia wynikającej z różnic długości i sekwencji w pełni komplementarnych sond.



Rysunek 24. *Wykres ilustrujący wpływ obecności niesparowanych nukleotydów na temperaturę topnienia sekwencji. Analizę wykonano dla 40 sond wykazujących różnice nukleotydowe względem pozostałych sekwencji sond. Sondy pochodziły z podzestawu MIRNA3.*

V.II.2.3 Porównanie wartości intensywności sygnałów fluorescencji odpowiadających sobie sond dla miRNA myszy, szczura i człowieka na poziomie danych eksperymentalnych

Kolejnym etapem sprawdzenia możliwości wykorzystania ortologicznych sond do badania ekspresji ludzkich miRNA u pacjentów z AML jest porównanie wartości intensywności sygnału fluorescencji odpowiadających sobie sond dla miRNA człowieka,

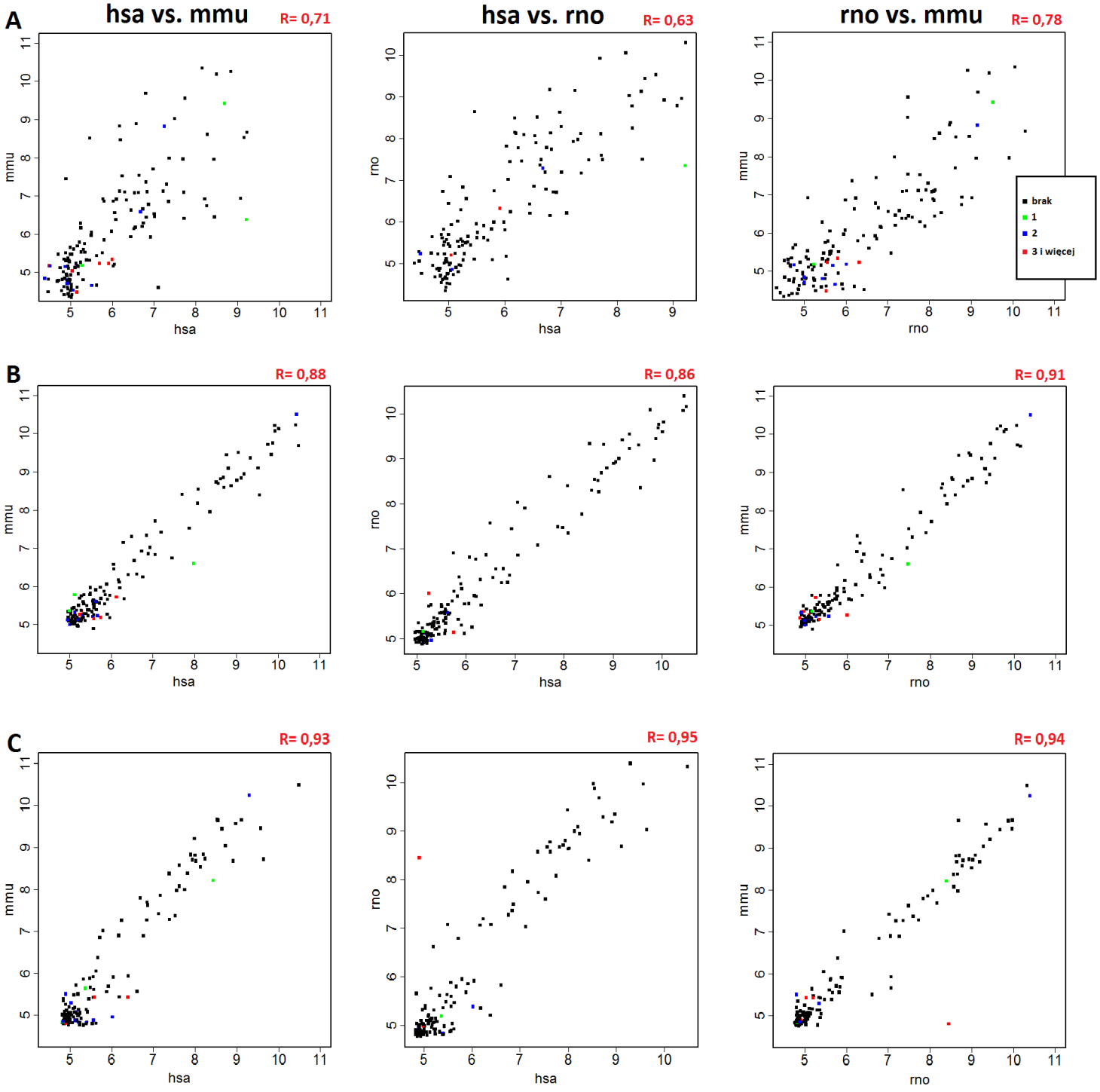
myszy i szczura. Etapem poprzedzającym tę analizę była selekcja ze źródłowych plików .gpr, sond wybranych w ramach podzestawu MIRNA3 wraz z odpowiadającymi im wartościami sygnałów fluorescencji dla kanału czerwonego (Alexa 647). Proces selekcji sond przeprowadzony został za pomocą algorytmu `filtracja.py`, przygotowanego w środowisku programistycznym Python. Wynikiem działania algorytmu `filtracja.py` (Załącznik 9) jest zestaw plików .txt, z których każdy zawiera 144 trójki sond wraz z odpowiadającymi im wartościami intensywności sygnałów fluorescencyjnych (Alexa 647) dla trzech badanych gatunków (Rysunek 25).

	A	B	C	D	E
1	ID sondy	hsa	rno	mmu	
2					
3	let_7a	5534,33	447	3046,33	
4	let_7b	2900	320	1335,33	
5	let_7c	3538	495	1990	
6	let_7d	1044,33	284,67	674	
7	let_7e	667,67	216,67	329,33	
8	let_7f	3653,67	385,33	1139,33	
9	let_7i	1022,33	212,67	823	
10	miR_100	237,33	264	248,33	
11	miR_103	478	237,67	184	
12	miR_106b	959,67	216	280,33	
13	miR_107	1622,67	218,67	174,33	
14	miR_10a	2841,33	2374,67	815,67	
15	miR_10b	343	156	219	
16	miR_122a	322,67	175,67	236	
17	miR_124a	440,67	171	305	
18	miR_125a	383,67	189,67	214,33	
19	miR_125b	457,33	150,67	253	
20	miR_127	422,67	157,67	285,33	
21	miR_128a	350,33	157,67	267,33	

Rysunek 25. *Przykład pliku .gpr po selekcji wspólnych typów sond dla miRNA człowieka (hsa), myszy (mmu) i szczura (rno). Plik jest wynikiem działania algorytmu filtracja.py przygotowanego w środowisku programistycznym Python.*

Porównanie wartości intensywności sygnału fluorescencji kanału czerwonego (Alexa 647) dla każdej mikromacierzy wykonane było dla trzech par sond: człowieka i myszy, człowieka i szczura oraz szczura i myszy. Podobieństwo sygnałów dla porównywanych zestawów sond określone zostało w wyniku analizy korelacji (program R, pakiet *stats*). Średnia wartość współczynnika korelacji R (dla wszystkich macierzy, N=30) wynosiła odpowiednio 0,826 dla sond człowieka i szczura, 0,743 dla sond człowieka i myszy oraz 0,792 dla sond myszy i szczura. Korelacja obliczona dla każdej mikromacierzy miRNA była istotna statystycznie (wartości p zawierały się w przedziale od $2,2 \times 10^{-16}$ do $9,47 \times 10^{-08}$). Średnie wartości współczynnika korelacji wyznaczone dla sond zawierających niesparowania wynosiły odpowiednio 0,678 dla sond człowieka i szczura, 0,645 dla sond człowieka i myszy oraz 0,632 dla sond myszy i szczura. Przy czym głównym składnikiem zmienności były sond

zawierające 3 niesparowania. Przykładowe wyniki korelacji zostały zobrazowane na Rysunku 27.



Rysunek 26. *Przykłady wykresów zależności wartości intensywności sygnałów fluorescencji kanału czerwonego (Alexa 647) dla trzech par zestawów 144 sond: człowieka i myszy, człowieka i szczura oraz szczura i myszy z trzech wybranych mikromacierzy. A, B i C zostały przygotowane na podstawie sygnałów dla reprezentatywnych mikromacierzy (A. 34bm_02; B. 34bp_01; C. 46bm_02). Kolor czarny obrazuje sekwencje sond o całkowitej homologii, zielony sekwencje sond różniące się jednym nukleotydem, niebieski sekwencje sond różniące się dwoma nukleotydami, a kolor czerwony sekwencje sond różniące się trzema i więcej nukleotydami.*

Wyniki analizy korelacji wskazują, iż sondy dla miRNA myszy i szczura w znacznym stopniu odzwierciedlają poziom ekspresji miRNA, uzyskany z użyciem sond dla miRNA człowieka. Ponadto, poziom korelacji dla ortologicznych sond w obrębie mikromacierzy DNA jest znacznie większy, niż poziom korelacji dla zestawu sond dla miRNA człowieka pomiędzy powtórzeniami technicznymi. Wynika to z faktu, iż większość badanych, ortologicznych miRNA człowieka, myszy i szczura ma taką samą sekwencję. Sekwencje docelowe z próbek AML wykazują zatem całkowitą komplementarność do niemalże wszystkich sond ulokowanych na mikromacierzy DNA. Sondy zawierające niesparowania charakteryzowały się niższym poziomem korelacji, jednakże wartości współczynnika korelacji dla tych sond nadal były wysokie. Eliminacja sond dla ortologicznych miRNA myszy i szczura w wyniku procesu filtracji w tym przypadku nie jest uzasadniona, gdyż wiązać się będzie z utratą znaczącej ilości danych, które mogą być z powodzeniem wykorzystane w dalszych etapach analizy.

Analiza sekwencji sond wspólnych dla miRNA trzech badanych gatunków (MIRNA3) z zestawu AML miRNA pozwoliła stwierdzić, iż uzyskane wartości ekspresji miRNA dla sond wszystkich badanych miRNA są bardzo zbliżone. W związku z tym analiza ekspresji ludzkich miRNA u pacjentów AML może być prowadzona z użyciem sond dla trzech badanych gatunków, bez konieczności uwzględniania procesu filtracji danych na etapie analizy. Uzyskane wyniki wskazują, iż jedynie sondy z 3 niesparowaniami, których liczba wynosi 4, powinny być usunięte z dalszej analizy.

V.II.3 Przykład wykorzystania wyników

Opisywane podejście umożliwiło analizę ekspresji miRNA u pacjentów z AML podtyp M1 i M2. Powyższa analiza została wykonana w ramach odrębnego projektu, który nie jest przedmiotem prezentowanej pracy.

V.II.4 Omówienie wyników

Na przestrzeni ostatnich lat pojawiło się kilka nowych sposobów wykorzystania ekspresyjnych mikromacierzy DNA. Jednym z nich jest możliwość stosowania technologii ekspresyjnych mikromacierzy DNA do badania niekodujących RNA (ncRNA, ang. *non coding RNA*) (Skreka i wsp. 2012). Większość mikromacierzy DNA dedykowanych jest w głównie miRNA, w szczególności w kontekście analizy ekspresji (Castoldi i wsp. 2007; C.-G. Liu, Spizzo, i wsp. 2008). Technologia ekspresyjnych mikromacierzy DNA została

zoptymalizowana jednak głównie pod kątem badania ekspresji genów kodujących białka i nie może być bezpośrednio stosowana do badania ekspresji miRNA. Cząsteczki miRNA posiadają zupełnie inną specyfikę, niż transkrypty genów kodujących białka, m.in. długość sekwencji. Stąd też proces planowania eksperymentów obejmujących analizę ekspresji miRNA, wymaga uwzględnienia specyficznych cech miRNA na etapie projektowania sond. Opisywane przykłady mikromacierzy do badania ekspresji miRNA obejmują nie tylko dedykowane mikromacierze DNA, ale także i LNA oraz mikromacierze DNA/LNA. Powszechnie stosowaną praktyką jest umieszczanie na powierzchni mikromacierzy DNA do badania ekspresji miRNA, sond dla miRNA kilku spokrewnionych gatunków lub sond dla potencjalnych i przewidywanych sekwencji miRNA. Podejście to umożliwia identyfikację nowych miRNA oraz ocenę stopnia konserwatywności sekwencji miRNA dla poszczególnych gatunków. Jednakże może też stanowić źródło dodatkowej zmienności wprowadzanej do układu, a tym samym i skutkować obniżeniem jakości otrzymanych wyników. Powodem są różnice nukleotydowe pomiędzy sekwencjami poszczególnych rodzajów miRNA dla różnych gatunków, co w efekcie sprawia, iż do badania ekspresji danego rodzaju miRNA stosowane są sondy o różnym stopniu komplementarności.

Stosowanie do badania ekspresji miRNA sond o różnym stopniu komplementarności nadaje otrzymanemu zestawowi niestandardowy charakter. Wynika to z faktu, iż w takim przypadku wartość sygnału fluorescencji nie tylko zależy od ilości sekwencji obecnych w próbce, ale również od stopnia ich dopasowania do sekwencji docelowych. Ogólnie przyjętym schematem analizy danych uzyskiwanych w ramach badania ekspresji miRNA z użyciem mikromacierzy jest eliminacja (w wyniku filtracji) z „surowych danych", informacji pochodzącej od sond ortologicznych. Przykładem narzędzia, które umożliwia taką filtrację jest pakiet *miChip* z repozytorium Bioconductor. Pakiet ten jest dedykowany analizie danych uzyskiwanych przy użyciu platformy *MiChip*, obejmującej mikromacierze z sondami w postaci LNA (Castoldi i wsp. 2007). Funkcja `removeUnwantedRows()`, wchodząca w skład tego pakietu, umożliwia eliminację rzędów z macierzy *m* (sond), które z punktu widzenia danego eksperymentu zawierają nieistotną informację. Stosowanie tego pakietu jest ograniczone jedynie do analizy jednokanałowych zestawów danych. Ponadto, eliminacja niektórych elementów z zestawu danych może skutkować zaburzeniem formatu danych, co w znacznym stopniu może utrudniać dalszą analizę danych (Wyniki, Rozdział V.I). Usunięcie sond ortologicznych może skutkować także pozbyciem się użytecznej informacji.

Zestaw AML miRNA jest przykładem zestawu, gdzie do badania ekspresji ludzkich miRNA u pacjentów z ostrą białaczką szpikową (AML) wykorzystano mikromacierz DNA zawierającą sondy komplementarne do miRNA trzech gatunków: człowieka, myszy i szczura. Choć sekwencje te wykazują dużą homologię względem siebie, to jednak nie jest ona pełna. Zgodnie z ogólnie przyjętym schematem analizy dla danych uzyskiwanych z użyciem mikromacierzy należałoby usunąć z zestawu danych sondy dla miRNA myszy i szczura. Sondy te stanowią 54% całkowitej liczby sond, zatem ich eliminacja w znaczący sposób zmniejszyłaby rozmiar zestawu danych AML miRNA. Zgodnie z doniesieniami literaturowymi, zbyt mały rozmiar danych może także zaburzać proces analizy i skutkować obniżeniem jakości otrzymanych wyników (G. K. Smyth i wsp. 2003). Celem przeprowadzonej analizy było sprawdzenie możliwości wykorzystania ortologicznych sond do badania ekspresji miRNA u pacjentów z AML. Możliwość uwzględnienia informacji pochodzącej od wszystkich sond w procesie analizy danych z zestawu AML pozwoliłaby na znaczne uproszczenie tego procesu poprzez eliminację procesu filtracji danych. Jednym z elementów przedstawionej analizy była ocena homologii sond dla miRNA myszy i szczura względem siebie oraz względem sekwencji dla ludzkich miRNA. Ze względu na fakt, iż poszczególne zbiory sond dla miRNA człowieka, myszy i szczura różniły się wzajemnie licznością, na potrzeby prowadzonej analizy wybrano mniejszy podzestaw sond (podzestaw MIRNA3), który zawierał 144 trójki odpowiadających sobie sond z trzech badanych gatunków. Wyniki analizy sekwencji wskazują, iż większość sond (87 – 93,8%) wykazuje pełną homologię (Rysunek 23). Natomiast pozostałe 6,2 – 13% sond posiada różnice w sekwencji na poziomie od 2 do 6 nukleotydów (1-3 niesparowań w obrębie pojedynczej sekwencji miRNA), z czego około 80% stanowią sekwencje posiadające jedynie od 2 do 4 niesparowań. W ramach dalszej części analizy w przybliżeniu określono wpływ obecności niesparowań na temperaturę topnienia dupleksów. Temperatura topnienia dupleksu jest parametrem, który ma kluczowe znaczenie w kontekście reakcji hybrydyzacji, a tym samym i określenia poziomu ekspresji miRNA. Z przeprowadzonej analizy wynika, iż spadek temperatury topnienia heterodupleksów spowodowany obecnością niesparowanych nukleotydów stanowi jedynie niewielki udział zmienności temperatury topnienia, wynikającej z różnic długości i składu nukleotydowego sond (Rysunek 20). Ustalone warunki reakcji hybrydyzacji (Materiały i Metody, Rozdział IV.I.2.1) sprzyjają możliwości wykorzystania ortologicznych sond do ustalenia ekspresji genów u pacjentów z AML. Wyniki wskazują, iż w podanych warunkach reakcji hybrydyzacji obecność niesparowań na poziomie 2 – 6 nukleotydów nie ma znaczącego wpływu na proces tworzenia dupleksu.

Potwierdzeniem powyższej sugestii są wyniki analizy korelacji wartości intensywności sygnałów fluorescencji kanału czerwonego (Alexa 647) odpowiadających sobie trójek sond (z podzestawu MIRNA3) dla miRNA człowieka, myszy i szczura. Otrzymane rezultaty wskazują, iż profil ekspresji miRNA określony za pomocą sond dla miRNA myszy i szczura w znacznym stopniu pokrywa się z profilem eskpresji miRNA uzyskanym z użyciem sond dla miRNA człowieka. Wynika to w znacznym stopniu z faktu, iż ok. 90% (87 – 93,8%) odpowiadających sobie sond dla miRNA myszy i szczura względem sond dla miRNA człowieka wykazuje pełną homologię, co sprawia, iż sondy te w istocie stanowią repliki sond dla miRNA człowieka. Filtracja tych sond skutkowałaby, więc pozbyciem się cennej informacji. Wyniki analizy korelacji dla sond zawierających niesparowania wykazują nieco niższe wartości współczynnika korelacji, jednakże wartości te nadal są bardzo wysokie.

Bardzo duże podobieństwo sekwencji badanych, ortologicznych miRNA trzech gatunków: człowieka, myszy i szczura pozwala na określenie poziomu ekspresji ludzkich miRNA przy użyciu informacji pochodzącej z wszystkich ortologicznych sond, także tych zawierających niesparowania. Wykorzystaniu sond zawierających niesparowania w dalszych etapach analizy sprzyja dobór warunków reakcji hybrydyzacji oraz wyniki analizy korelacji. Ponadto, podobieństwo sekwencji sond na tak wysokim poziomie pozwala z dużym zaufaniem wykorzystać sondy dla miRNA myszy i szczura do identyfikacji, nowych, ludzkich miRNA w ramach przeprowadzonego eksperymentu. Wyniki te jednak powinny być analizowane i interpretowane z zachowaniem szczególnej ostrożności.

V.II.5 Wnioski

Głównymi wnioskami wynikającymi z przeprowadzonych analiz w tej części pracy są:

- Większość ortologicznych miRNA spokrewnionych gatunków ma taką samą sekwencję lub wykazuje nieznaczne różnice w sekwencji;
- W związku z powyższym filtracja sond dla ortologicznych miRNA nie jest konieczna;
- Sondy dla ortologicznych miRNA często stanowią powtórzenia danego typu sondy i mogą być z powodzeniem wykorzystywane do walidacji otrzymanych wyników.
- Stosowanie sond dla miRNA spokrewnionych gatunków pozwala także na identyfikację nowych miRNA danego gatunku za pomocą homologicznych sond.

CZĘŚĆ III: Normalizacja danych uzyskiwanych z użyciem dedykowanych mikromacierzy DNA.

Ekspresyjne mikromacierze DNA dostarczają ogromnej ilości użytecznej informacji, która jednak obarczona jest licznymi błędami systematycznymi oraz zmiennością eksperymentalną. W praktyce, żaden eksperyment z użyciem ekspresyjnych mikromacierzy DNA nie jest wolny od efektów pochodzenia technicznego wprowadzonych na etapie przygotowania próbki, hybrydyzacji, płukania czy skanowania mikromacierzy. Drukowane mikromacierze DNA dodatkowo mogą posiadać wady techniczne, powstałe na etapie przygotowania (drukowania) mikromacierzy, np. różnice w wielkości i kształcie punktów (efekt zależny od igły), przesunięcia pojedynczych punktów, a nawet całych rzędów i bloków sond (ang. *print-tips*). Ponadto, drukowane mikromacierze DNA najczęściej wykorzystywane są w eksperymentach dwukolorowych, tzn. obejmujących użycie dwóch barwników fluorescencyjnych na etapie znakowania próbek. Stosowanie dwóch różnych barwników fluorescencyjnych wprowadza dodatkową zmienność pomiędzy próbkami badanymi i kontrolnymi, która wynika z różnych właściwości chemicznych barwników. Automatyzacja reakcji hybrydyzacji oraz procesu płukania mikromacierzy DNA znacząco podnosi jakość otrzymywanych danych. Nie jest to jednak wystarczający sposób minimalizacji zmienności w układzie i dane te wciąż muszą zostać odpowiednio przetworzone przed przeprowadzeniem analizy wyższego rzędu.

V. III.1 Identyfikacja problemu

Głównym wyzwaniem na etapie analizy danych uzyskiwanych z użyciem drukowanych mikromacierzy DNA jest proces normalizacji danych. Wynika to z faktu, iż drukowane mikromacierze DNA najczęściej stosowane są w formie dedykowanych mikromacierzy DNA. Ten rodzaj mikromacierzy posiada szereg specyficznych cech, opisanych bardziej szczegółowo we wstępie (Rozdział II.4). Do najbardziej charakterystycznych właściwości tego rodzaju mikromacierzy DNA należą: relatywnie niska całkowita liczba sond, wysoka proporcja sond dla genów ulegających ekspresji różnicowej w badanych warunkach (zwykle znacznie ponad 20%), zaburzona równowaga pomiędzy liczbą genów ulegających podwyższonej i obniżonej ekspresji oraz niska liczba sond kontrolnych w stosunku do całkowitej liczby sond. Przedstawiona wyżej specyfika dedykowanych mikromacierzy DNA sprawia, iż proces normalizacji tego rodzaju danych jest mało skuteczny,

a niekiedy nawet wprowadza dodatkową zmienność do analizowanego zestawu danych.

Obecnie istnieje wiele metod normalizacji, które pozwalają na eliminację zmienności różnego typu i zwykle w tym celu oferują różne podejścia. Większość metod normalizacji danych wymaga jednak spełnienia założeń, które prawdziwe są jedynie dla mikromacierzy o dużej gęstości (ang. *whole-genome arrays*). W badaniach ekspresji genów globalne metody normalizacji bazują na założeniu, iż liczba genów wykazujących ekspresję różnicową stanowi jedynie niewielki procent (< 10 %) w stosunku do genów ulokowanych na mikromacierzy DNA oraz że istnieje równowaga pomiędzy genami ulegającymi podwyższonej i obniżonej ekspresji. Dodatkowo proste globalne metody normalizacji, np. *median* czy *mean* zakładają, że wszystkie wartości intensywności na danej mikromacierzy DNA obarczone są tym samym błędem systematycznym. Powyższe założenia często nie są spełnione dla dedykowanych mikromacierzy DNA. Stąd też bardziej odpowiednim podejściem dla tego rodzaju danych jest stosowanie lokalnie ważonej regresji liniowej jako globalnej procedury normalizacji (*loess* lub *lowess*) lub metody uwzględniającej podział na grupy sond drukowanych tą samą igłą (*print-tip loess*). Alternatywą dla globalnej metody normalizacji *loess* jest podejście oparte na wybranych zestawach sond kontrolnych, które charakteryzują się jednolitym poziomem ekspresji w obrębie i między macierzami, np. sondy kontrolne typu *spike-in*. Strategia ta wymaga jednak obecności znacznej liczby sond kontrolnych. Różne metody normalizacji danych charakteryzują się odmiennym mechanizmem działania, a tym samym mają różny wpływ na otrzymane wyniki. Wybór odpowiedniej metody normalizacji zależy od cech danego zestawu i często w przypadku dedykowanych mikromacierzy DNA jest niezwykle trudny.

V.III.2 Rozwiązanie problemu

V. III.2.1 Cel

Celem tej części pracy doktorskiej było opracowanie uniwersalnej i zobiektywizowanej procedury wyboru metody normalizacji dla danego zestawu danych. W ramach realizacji tego zadania wybranych zostało 13 różnych metod normalizacji pochodzących z repozytorium Bioconductor: 10 metod normalizacji dla danych dwukanałowych (opisanych dalej jako dwukanałowe metody normalizacji) oraz 3 metody normalizacji dla danych jednokanałowym (opisanych dalej jako jednokanałowe metody normalizacji). Testowane metody normalizacji oceniane były na podstawie 5 kryteriów. Dwa pierwsze kryteria stanowiły wartości błędów systematycznych (I) oraz wariancji (II) liczone

dla sond kontrolnych. Następnie oceniana była czułość (III) i specyficzność (IV) analizy ekspresji różnicowej dla danych normalizowanych przy wykorzystaniu wybranych metod normalizacji. Ostatnim kryterium oceny procedur normalizacji była ocena zdolności klasyfikacji próbek określana w oparciu o profil ekspresji genów (V), który był wynikiem analizy danych przeprowadzonych z użyciem poszczególnych metod normalizacji spośród 13 wybranych. Zdolność klasyfikacji próbek określona została za pomocą krzywych ROC i wartości AUC. Dla każdego z wyżej opisanych 5 kryteriów, przygotowano ranking metod normalizacji na podstawie otrzymanych wyników. Optymalna metoda normalizacji pod względem danego kryterium zajmowała pierwsze miejsce w rankingu. Ostateczna selekcja odpowiedniej metody normalizacji prowadzona była w oparciu o ranking ostateczny otrzymany w wyniku uśrednienia wartości rankingów dla wszystkich 5 opisywanych wyżej kryteriów. Dwukanałowe i jednokanałowe procedury normalizacji analizowane były niezależnie.

V.III.2.2 Charakterystyka zestawów danych

Wybrane procedury normalizacji testowano na przykładzie czterech zestawów danych: zestawu do badania ekspresji genów u pacjentów z ostrą białaczką szpikową (zestaw AML II) oraz zestawów do analizy ekspresji genów u dzieci z alergią krzyżową (zestaw ALERGIA) i astmą (zestaw ASTMA). Dodatkowo wykorzystano zestaw danych prezentowany i opisany przez zespół Alicji Oshlack (Oshlack i wsp. 2007), zwany dalej zestawem OSHLACK. W skład zestawu danych AML II wchodziło 40 mikromacierzy o najlepszej jakości i odpowiadających sobie powtórzeniach technicznych, wybranych z zestawu AML (Materiały i Metody, rozdział IV.I.1.3). Ponadto, wszystkie mikromacierze wchodzące w skład zestawu AML II charakteryzowały się skorygowanym formatem danych, zgodnie z obowiązującymi standardami jakości dla ekspresyjnych mikromacierzy DNA (Wyniki, Rozdział V.I). W skład zestawów ASTMA i ALERGIA wchodziło jednakowo 14 mikromacierzy. Zestawy AML II, ALERGIA i ASTMA charakteryzowały się obecnością 8 sond kontrolnych typu *spike-in*. Natomiast zestaw OSHLACK obejmował 6 mikromacierzy, z których każda zawierała zestaw 288 sond kontrolnych typu MSP (ang. *microarray sample pool*). Zestaw sond typu MSP otrzymano w wyniku nanoszenia na podłoże 32 roztworów sond o 9 różnych stężeniach (1-250 ng/μl). Celem stosowania sond typu MSP przez Oshlack i wsp. było zwiększenie skuteczności normalizacji danych uzyskiwanych z użyciem dedykowanych mikromacierzy DNA, poprzez redukcję proporcji genów różnicujących w stosunku do całkowitej liczby sond.

Charakterystyka 4 zestawów danych wykorzystywanych na etapie przygotowania zobiektywizowanej procedury wyboru metody normalizacji przedstawiona została w Tabeli 8.

Tabela 8. *Charakterystyka czterech zestawów danych użytych do testowania metod normalizacji: AML II, ALERGIA, ASTMA oraz OSHLACK. Sondy kontrolne ArrayControl, Ambion są kontrolami typu spike-in. Natomiast sondy kontrolne typu MSP (ang. microarray sample pool) stanowią pulę sond występujących na mikromacierzy DNA w różnych stężeniach.*

Cechy	Zestaw danych			
	AML II	ALERGIA	ASTMA	OSHLACK
Liczba mikromacierzy	40	14		6
Rodzaj eksperymentu	dwukolorowy*	dwukolorowy		dwukolorowy
Próbka badana	pacjenci z AML zdrowi ochotnicy	dzieci z alergią krzyżową		myszy OBF-1-/-
Próbka referencyjna	HL60	zdrowi ochotnicy		myszy kontrolne C57BL/6
Liczba i rodzaj sond kontrolnych	8 (ArrayControl, Ambion)	8 (ArrayControl, Ambion)		288 sond MSP
Liczba sond w pojedynczym bloku	81	144		462
Całkowita liczba sond	3240	1728		11088

*w wyniku dwukolorowego eksperymentu otrzymano dwukanałowe zestawy danych, które na potrzeby tej części pracy przekształcone zostały także w jednokanałowe zestawy danych.

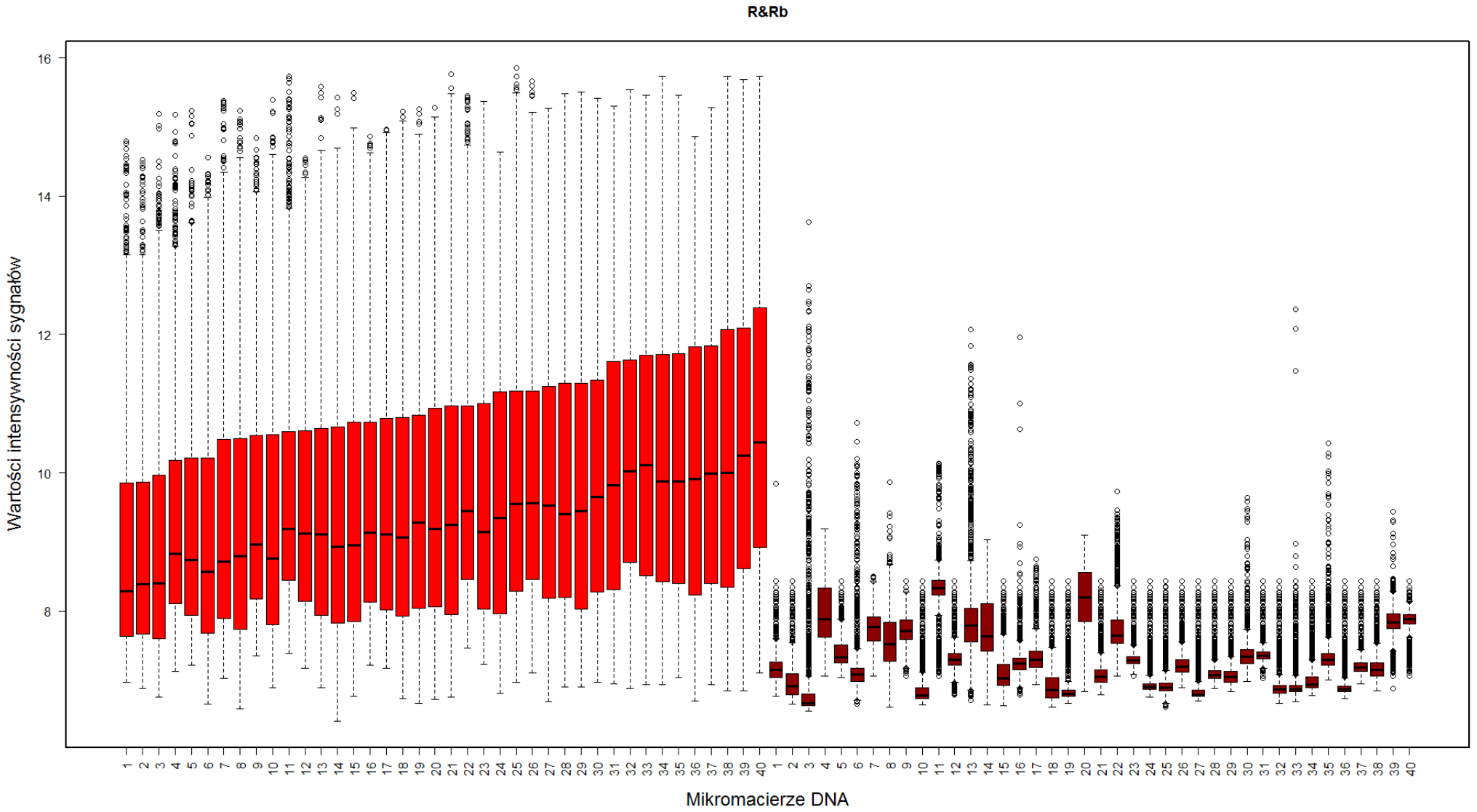
Każdy z analizowanych zestawów danych wczytany został do programu R/Bioconductor przy wykorzystaniu pakietu `limma`. Następnie dla każdego zestawu danych wykonano korekcję tła metodą *substract*, gdzie dla każdej sondy od wartości intensywności sygnałów odjęto wartości tła. Korekcja tła była konieczna, ponieważ procentowy udział tła w całkowitej wartości intensywności sygnału fluorescencji był duży. Rozkład wartości intensywności sygnałów fluorescencyjncych oraz wartości tła kanału czerwonego oraz zielonego dla danych z zestawu AML II przedstawione zostały za pomocą wykresów pudełkowych (Rysunek 28 oraz Rysunek 29).

V.III.2.3 Wybór i charakterystyka metod normalizacji

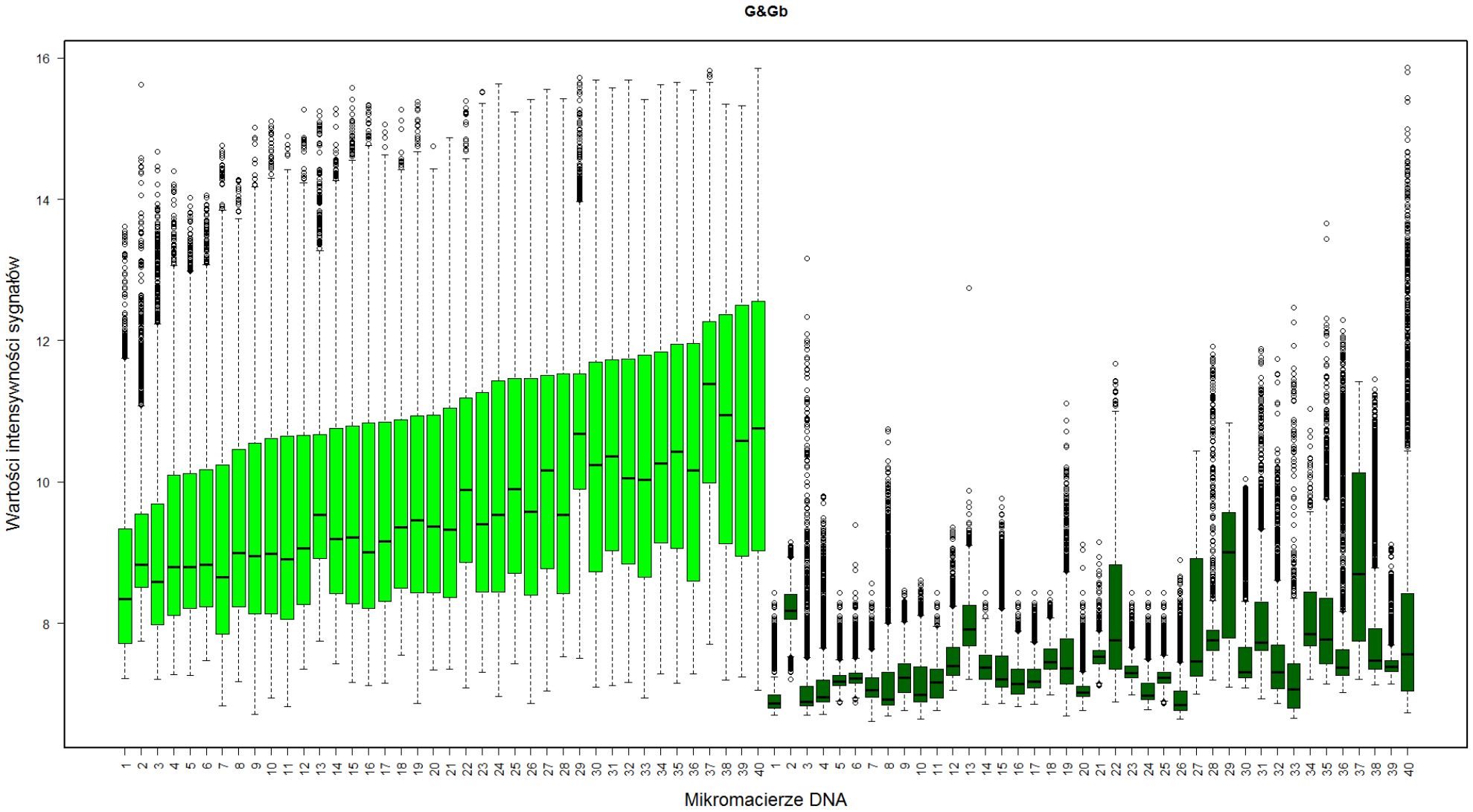
Analiza dwukanałowych metod normalizacji obejmowała 10 metod normalizacji zdeponowanych w ramach 7 pakietów Bioconductor: `limma`, `snm`, `vsn`, `nnNorm`, `OLIN`, `marray` oraz `TurboNorm`, z których każda testowana była na wszystkich 4 zestawach danych. Natomiast analiza jednokanałowych metod normalizacji wykonana była za pośrednictwem 3 procedur z 3 pakietów: `limma`, `vsn` oraz `snm`. Charakterystyka stosowanych metod normalizacji przedstawiona została w Tabeli 6 (Materiały i Metody, Rozdział IV.II.2).

Selekcja dwukanałowych metod normalizacji obejmowała głównie algorytmy do normalizacji wewnętrznej (w obrębie mikromacierzy) (Wstęp, Rozdział II.4.1.2.1). Przy wyborze tych metod szczególną uwagę zwrócono na najbardziej popularny rodzaj metod normalizacji: metody oparte na lokalnie ważonej regresji liniowej, najczęściej znane pod

nazwą *loess* lub *lowess* (ang. *locally weighted scatterplot smoothing*). Ta klasa metod normalizacji, należy do metod zależnych od intensywności sygnału i jest jedną z najczęściej stosowanych procedur normalizacji. Wśród testowanych procedur normalizacji znajduje się 7 metod typu *loess*: trzy metody z pakietu `limma`, jedna metoda z pakietu `marray`, dwie metody z pakietu `OLIN` oraz jedna metoda z pakietu `TurboNorm`. Pakiet `limma` oferuje globalną metodę normalizacji *loess* (zwaną dalej *Loess*), metodę *print-tip loess* uwzględniającą podział na grupy sond drukowanych tą samą igłą (zwaną dalej *Ploess*) oraz globalną metodę normalizacji opartą na zestawie sond kontrolnych typu *spike-in* (zwaną dalej *Spike*). Metoda normalizacji *loess* z pakietu `marray` (zwana dalej *LoessM*), podobnie jak odpowiadająca jej metoda z pakietu `limma` jest globalną metodą normalizacji. Jednak ze względu na różnice pomiędzy algorytmami funkcji, obie metody zostały włączone do analizy. Pakiet `OLIN` natomiast dostarcza metodę normalizacji danych: *OLIN* (ang. *Optimized Local Intensity-dependent Normalization*). Metoda OLIN została zaprezentowana w dwóch wersjach: z uwzględnieniem koordynatów X i Y, odpowiadających lokalizacji każdej z sond na mikromacierzy (metoda *Olin_c*) oraz z pominięciem koordynatów X i Y, gdzie algorytm normalizacji sam szacował położenie sond na mikromacierzy (metoda *Olin*). Pakiet `TurboNorm` oferuje bardzo dobrze zoptymalizowany algorytm normalizacji oparty na lokalnie ważonej regresji liniowej (metoda *Turbo*). Procedura normalizacji z pakietu `TurboNorm` jest bardzo szybka, a jej zastosowanie pozwala na znaczne skrócenie czasu prowadzonych obliczeń. Pozostałe 3 z 10 prezentowanych metod normalizacji dla zestawów dwukolorowych pochodzą z pakietów `vsn`, `snm` oraz `nnNorm`. Metoda normalizacji z pakietu `vsn` (zwana dalej *Vsn2*) zakłada stałość współczynnika wariancji i jest jedną z bardziej radykalnych metod normalizacji, często wykazującą skłonność do nadmiernej normalizacji (ang. *overfitting*) danych. Metoda normalizacji z pakietu `snm` (zwana dalej *Snm2*) należy do klasy nadzorowanych metod i na etapie normalizacji uwzględnia dodatkowe informacje na temat eksperymentu, takie jak: partia mikromacierzy, projekt eksperymentu, sprzęt. Zupełnie nowa strategia normalizacji danych, wykorzystująca algorytm sieci neuronowych, oferowana jest przez pakiet `nnNorm` (metoda *Nn*). Ilość dostępnych metod normalizacji dla danych jednokanałowych jest znacznie mniejsza, niż w przypadku dwukanałowych. Stąd też wybór ten obejmuje jedynie trzy metody: *quantile* z pakietu `limma` (zwana dalej *Q*), *vsn* z pakietu o tej samej nazwie (zwana dalej *Vsn1*) oraz *snm* z pakietu `snm` w wersji jednokolorowej (zwana dalej *Snm1*).



Rysunek 27. *Wykres pudełkowy prezentujący rozkład intensywności głównych (kolor jasnoczerwony) oraz wartości tła (kolor ciemnoczerwony) dla kanału czerwonego dla 40 wybranych mikromacierzy DNA z zestawu danych AML II. Mikromacierze DNA zostały uszeregowane względem rosnących wartości intensywności głównych w celu zobrazowania zależności pomiędzy wartościami intensywności głównych oraz wartościami tła. Oś X przedstawia numery poszczególnych mikromacierzy DNA z zestawu AML II, oś Y natomiast wartości intensywności sygnałów w skali logarymicznej (ln).*



Rysunek 28. *Wykres pudełkowy prezentujący rozkład intensywności głównych (kolor jasnozielony) oraz wartości tła (kolor ciemnozielony) dla kanału zielonego dla zestawu danych AML II. Mikromacierze DNA zostały uszeregowane względem rosnących wartości intensywności głównych w celu zobrazowania zależności pomiędzy wartościami intensywności głównych oraz wartościami tła. Oś X przedstawia numery poszczególnych mikromacierzy DNA z zestawu AML II, oś Y natomiast wartości intensywności sygnałów w skali logarymicznej (ln).*

V.III.2.4 Wstępna ocena efektu normalizacji przy pomocy wykresów MA

W praktyce wybór optymalnej metody normalizacji dla analizowanego zestawu danych najczęściej oparty jest na wizualnej ocenie rozkładu wartości M względem wartości A dla danej mikromacierzy (normalizacja wewnętrzna) lub rozkładu wartości M dla poszczególnych mikromacierzy eksperymentu (normalizacja zewnętrzna). Wykresy MA (Wstęp, Rozdział II.4.1.3.1) pozwalają ocenić zależność pomiędzy zmianą poziomu sygnału danej sondy dla próbki badanej względem próbki referencyjnej (M), a średnią wartością intensywności sygnału dla sondy analizowanego genu (A). Skuteczność procesu normalizacji określana jest na podstawie wizualnej oceny wykresów MA wykonanych dla zestawów danych przed i po normalizacji.

W celu wstępnej oceny efektu normalizacji, wykresy MA z użyciem pakietu `limma` wygenerowane zostały dla wszystkich 13 metod normalizacji dla 4 analizowanych zestawów danych. Przykład wykresów MA otrzymanych dla jednej mikromacierzy z zestawu AML II przestawiony został na Rysunku 30. Wykresy dla pozostałych mikromacierzy z zestawu AML II oraz dla wszystkich mikromacierzy zestawów ASTMA, ALERGIA oraz OSHLACK dostępne są w Załączniku 10. Wartości M i A dla wszystkich dwukanałowych metod normalizacji obliczone były na podstawie wartości intensywności sygnałów fluorescencji dla kanału czerwonego (badany) oraz zielonego (referencyjny). Natomiast w przypadku jednokanałowych metod normalizacji wartości M i A obliczone zostały z użyciem dwóch wartości intensywności sygnałów fluorescencji dla kanału czerwonego, pochodzących z odpowiadających sobie powtórzeń technicznych. Prezentowane na wykresach MA (Rysunek 29) „surowe wartości" dla danych dwukanałowych charakteryzują się typowym zakrzywieniem, które powinno być wyeliminowane w wyniku skutecznej normalizacji wartości M. Wykresy MA przedstawione na Rysunku 29 wskazują, że w przypadku wszystkich metod normalizacji, z wyjątkiem trzech (*Spike, Snm2* oraz *Snm1*) dane zostały skutecznie znormalizowane.

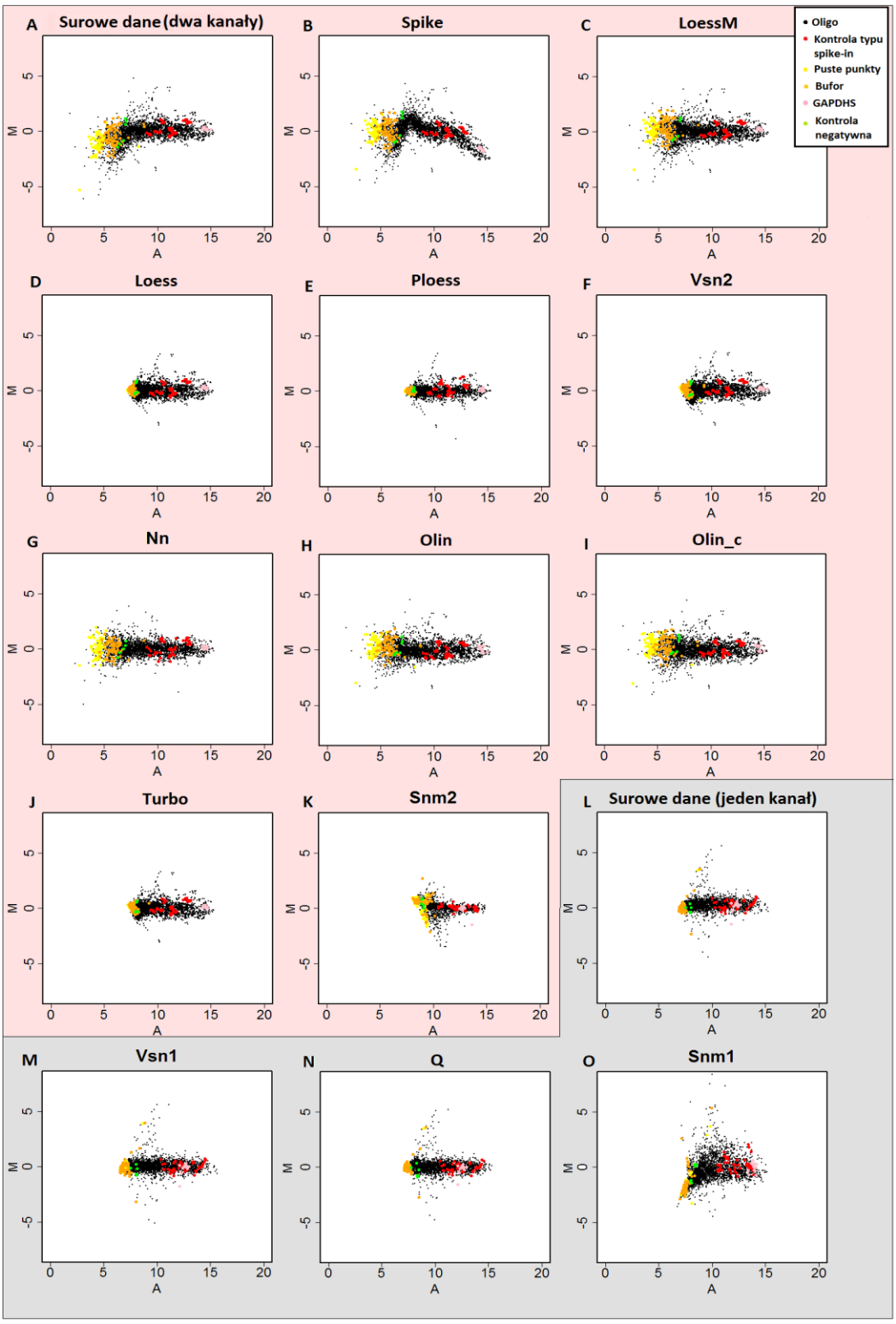
Wykresy MA są bardzo pomocne przy określaniu skuteczności normalizacji, jednak często jest to dość subiektywna i raczej intuicyjna ocena tego procesu. Ograniczeniem stosowania wykresów MA jest także ich liczba. W przypadku dużych zestawów danych ilość wykresów MA jest znacząca, co dodatkowo utrudnia ich rzetelną ocenę. Przykładowo przetwarzanie danych dla zestawu AML II obejmującego 40 mikromacierzy i 13 testowanych metod normalizacji wymagało przygotowania 520 wykresów MA. Wybór odpowiedniej

metody normalizacji jedynie na podstawie wykresów MA jest kłopotliwy, tym bardziej, iż niekiedy różnice pomiędzy wykresami MA dla poszczególnych metod są subtelne, np. *Loess* i *Turbo* (Rysunek 29). Znacznie lepszym podejściem jest porównanie testowanych metod normalizacji w sposób zobiektywizowany, np. w oparciu o kryteria liczbowe z uwzględnieniem wykresów MA jako pomocniczego kryterium wyboru procesu normalizacji. W tym celu przeprowadzonych został szereg analiz obejmujących porównanie metod na postawie wartości błędu systematycznego i wariancji, jak również czułości i specyficzności analizy ekspresji różnicowej danych normalizowanych przy wykorzystaniu wybranych metod oraz na podstawie krzywych ROC i wartości AUC.

V.III.2.5 Klasyfikacja metod normalizacji w oparciu o wartości błędu systematycznego i wariancji

Większość dedykowanych mikromacierzy DNA zawiera zestaw sond kontrolnych w postaci sond dla genów typu *housekeeping* lub kontroli zewnętrznych typu *spike-in*. Wartość intensywności sygnału tych sond może być z powodzeniem wykorzystana do oceny metod normalizacji. Wynika to z faktu, iż poziom sygnału dla sond kontrolnych powinien być taki sam zarówno dla próbki badanej, jak i kontrolnej (wartość M=0, brak zmiany ekspresji). Najprostszą formą oceny efektu normalizacji jest określenie wartości błędu systematycznego (ang. *bias*) oraz wartości wariancji dla sond kontrolnych (Argyropoulos i wsp. 2006). Wartość błędu systematycznego stanowi odchylenie wartości M dla danej sondy kontrolnej względem wartości M=0. Natomiast wariancja określa zmienność pomiędzy wartościami intensywności sygnałów fluorescencji dla powtórzeń każdej z sond kontrolnych (ang. *spot replicates*).

Wykorzystując wzory 1 oraz 2 (Materiały i Metody, Rozdział IV.II.1) dla dwukolorowych metod normalizacji oraz wzory 5 oraz 6 (Materiały i Metody, Rozdział IV.II.1) dla metod jednokanałowych, obliczone zostały wartości błędu systematycznego i wariancji dla 13 metod normalizacji testowanych na 4 zestawach danych. Na podstawie otrzymanych wyników analizy przygotowany został ranking wartości dla każdego z parametrów (Tabela 9 oraz Tabela 10). Za kryterium charakteryzujące najlepszą metodę normalizacji uznana została najniższa średnia wartość błędu systematycznego (Tabela 9) i wariancji (Tabela 10) obliczona dla sond kontrolnych każdego z zestawów danych. Metoda dla której wartości parametrów były najniższe otrzymała najwyższą pozycję w rankingu:1.



Rysunek 29. *Zestaw wykresów MA dla różnych metod normalizacji dla zestawu danych AML II. Na różowym tle przedstawione są wykresy MA dla dwukolorowych metod normalizacji: A. surowe dane, B. Spike-globalna metoda loess oparta na zestawie sond kontrolnych typu spike-in; C. LoessM-globalna metoda loess (pakiet marray); D. Loess-globalna metoda loess; E. Ploess-metoda loess uwzględniająca podział na grupy sond drukowane jednym rodzajem igły; F. Vsn2-globalna metoda normalizacji; G. Nn-metoda normalizacji wykorzystująca algorytm sieci neuronowych; H. Olin- globalna metoda loess; I. Olin_c-globalna metoda loess z uwzględnieniem koordynatów (X,Y) lokalizacji sond; J. Turbo-globalna metoda loess; K. Snm2-nadzorowana metoda normalizacji. Na tle szarym przedstawione są wykresy MA dla jednokolorowych metod normalizacji: L. surowe dane, M. Vsn1-globalna metoda normalizacji, N. Q-globalna metoda normalizacji typu quantile, O. Snm1-nadzorowana metoda normalizacji.*

Tabela 9. *Średnie wartości błędu systematycznego dla 13 testowanych metod normalizacji na przykładzie 4 zestawów danych: AML II, ALERGIA, ASTMA oraz OSHLACK. Wartości od 1-10 oznaczają pozycję w rankingu danej metody, natomiast wartości w nawiasie średnią wartość błędu systematycznego.*

Metoda normalizacji	Średnia wartość błędu systematycznego-ranking							
	zestaw danych AML II		zestaw danych ALERGIA		zestaw danych ASTMA		zestaw danych OSHLACK	
Metody normalizacji danych dwukanałowych								
Spike	**4,76**	10	**0,83**	9	**0,87**	8	**1,52**	10
LoessM	**0,76**	8	**0,76**	7	**0,92**	9	**0,53**	5
Loess	**0,56**	2	**0,40**	1	**0,65**	3	**0,21**	2
Ploess	**0,73**	7	**0,51**	4	**0,54**	1	**0,20**	1
Vsn2	**0,67**	3	**0,49**	3	**0,80**	5	**0,37**	4
Nn	**0,70**	6	**0,80**	8	**0,85**	7	**0,69**	8
Olin	**0,67**	3	**0,69**	5	**0,75**	4	**0,62**	7
Olin_c	**0,68**	5	**0,74**	6	**0,83**	6	**0,60**	6
Turbo	**0,55**	1	**0,42**	2	**0,63**	2	**0,26**	3
Snm2	**1,22**	9	**1,37**	10	**1,27**	10	**0,83**	9
Metody normalizacji danych jednokanałowych								
Vsn1	**0,79**	2	**0,55**	1	**0,55**	2	**0,58**	3
Q	**0,74**	1	**0,56**	2	**0,54**	1	**0,23**	1
Snm1	**1,03**	3	**1,13**	3	**1,20**	3	**0,39**	2

Tabela 10. *Średnie wartości wariancji dla 13 testowanych metod normalizacji na przykładzie 4 zestawów danych: AML II, ALERGIA, ASTMA oraz OSHLACK. Wartości od 1-10 oznaczają pozycję w rankingu danej metody, natomiast wartości w nawiasie średnią wartość wariancji.*

Metoda normalizacji	Średnia wartość wariancji -ranking							
	zestaw danych AML II		zestaw danych ALERGIA		zestaw danych ASTMA		zestaw danych OSHLACK	
Metody normalizacji danych dwukanałowych								
Spike	**46,41**	10	**1,49**	9	**1,08**	9	**4,60**	10
LoessM	**0,59**	7	**0,73**	7	**0,81**	8	**0,45**	5
Loess	**0,29**	2	**0,15**	1	**0,41**	3	**0,08**	2
Ploess	**2,43**	9	**0,25**	4	**0,29**	1	**0,07**	1
Vsn2	**0,39**	3	**0,19**	3	**0,56**	5	**0,14**	4
Nn	**0,52**	6	**0,84**	8	**0,72**	7	**0,57**	8
Olin	**0,43**	4	**0,60**	5	**0,56**	4	**0,46**	7
Olin_c	**0,46**	5	**0,66**	6	**0,64**	6	**0,46**	6
Turbo	**0,28**	1	**0,16**	2	**0,39**	2	**0,09**	3
Snm2	**1,54**	8	**2,04**	10	**1,57**	10	**0,69**	9
Metody normalizacji danych jednokanałowych								
Vsn1	**0,64**	2	**0,34**	1	**0,35**	2	**0,34**	3
Q	**0,56**	1	**0,35**	2	**0,34**	1	**0,06**	1
Snm1	**1,07**	3	**1,26**	3	**1,46**	3	**0,18**	2

Wyniki przedstawione w Tabeli 9 oraz Tabeli 10 wskazują, iż w kategorii błędu systematycznego i wariancji najlepszą grupą metod normalizacji są te oparte na lokalnie ważonej regresji liniowej: *Turbo* oraz *Loess* i *Ploess*. W przypadku zestawu danych OSHLACK, najlepszy wynik został osiągnięty przez metodę *Ploess*. Jednak metody *Loess* i *Turbo* charakteryzują się jedynie nieznacznie gorszymi wynikami. Metoda *LoessM*, globalna metoda *loess* z pakietu `marray`, prezentuje się znacznie gorzej w przypadku każdego z badanych zestawów danych, niż zbliżona do niej *Loess* - globalna metoda *loess* z pakietu `limma`. Najgorszy wynik w kategorii błędu systematycznego i wariancji dla każdego z

zestawów danych został osiągnięty przez metody S*pike oraz Snm2*. Otrzymane wyniki korespondują z rezultatami prezentowanymi na wykresach MA (Rysunek 30).

Analiza jednokanałowych metod normalizacji wykazała, iż stosując średnią wartość błędu systematycznego oraz wariancji jako kryterium wyboru, najlepszą procedurą normalizacji dla zestawów danych AML II i OSHLACK okazała się być metoda *Q* z pakietu `limma`. W przypadku zestawów ALERGIA i ASTMA różnice pomiędzy metodą *Q* i *Vsn1* są niewielkie. Najgorszą metodą normalizacji w przypadku zestawu OSHLACK jest *Vsn1*, a w przypadku pozostałych zestawów danych - *Snm1*. Jednakże różnice pomiędzy najlepszą i najgorszą metodą normalizacji nie są tak duże jak w przypadku dwukanałowych procedur normalizacji.

V.III.2.6 Analiza ekspresji różnicowej

Kolejnym kryterium oceny metod normalizacji była analiza ekspresji różnicowej (ang. *differential expression analysis*), której celem było określenie podobieństwa pomiędzy składem list genów różnicujących oraz określenie poziomu czułości i specyficzności analizy ekspresji różnicowej dla danych normalizowanych z wykorzystaniem testowanych procedur normalizacji. Analiza ekspresji różnicowej została wykonana na przykładzie zbioru danych AML II, który zawierał dwa rodzaje próbek: próbki uzyskane od pacjentów z ostrą białaczką szpikową oraz próbki od zdrowych ochotników. W celu określenia różnic ekspresji genów pomiędzy tymi dwoma grupami próbek stosowano test t z poprawką BH (*Benjamini-Hochberg*) dla testowań wielokrotnych ($\alpha<0,05$). Zgodnie z danymi dotyczącymi projektu eksperymentu, analiza danych z zestawu AML II powinna skutkować otrzymaniem blisko 200 genów różnicujących, co stanowi ok. 20% całkowitej liczby genów. W wyniku analizy danych AML II dla każdej z metod normalizacji otrzymane zostały inne listy genów różnicujących. Listy różniły się między sobą nie tylko licznością, ale także i składem genów (Tabela 11). Dla większości metod normalizacji, z wyjątkiem *Snm2* i *Spike*, listy genów różnicujących zawierały blisko 200 genów. Metoda *Spike* pozwoliła na wytypowanie jedynie 87 genów. Wyjątkowo niska liczba genów - tylko 35 została zidentyfikowana dla zestawu AML II po zastosowaniu metody *Snm2*.

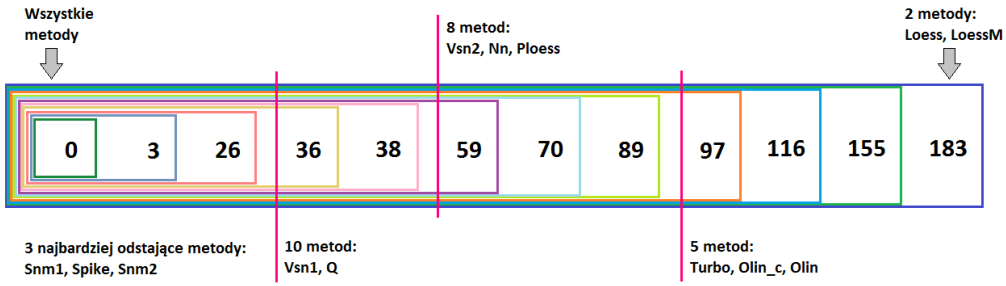
V.III.2.6.1 Analiza składu list genów różnicujących

W celu określenia stopnia podobieństwa pomiędzy listami genów różnicujących dla wszystkich 13 metod normalizacji, przeprowadzona została analiza porównawcza składu tych

list. Analiza ta została wykonana dla każdej pary metod normalizacji (Tabela 11). Kolejnym etapem tej analizy było pokazanie jak zmienia się liczba wspólnych genów różnicujących po dodaniu kolejnych metod normalizacji o malejącym stopniu podobieństwa (Rysunek 30).

Tabela 11. *Porównanie ekspresji genów zidentyfikowanych jako różnicujące w zestawie danych AML II przy zastosowaniu różnych metod normalizacji. Liczby w polu o kolorze zielonym oznaczają całkowitą liczbę genów różnicujących wytypowanych dla zestawu danych normalizowanych z użyciem danej metody. Powyżej i poniżej przekątnej przedstawiono liczbę genów wspólnych dla pary metod. Odsetek wspólnych genów obliczono względem całkowitej liczby genów zidentyfikowanych po zastosowaniu metody normalizacji wymienionej w pierwszym rzędzie. Liczba genów potencjalnie różnicujących była zidentyfikowana na podstawie testu t przy założonym poziomie istotności α <0,05.*

	Metody dwukanałowe										Metody jednokanałowe		
	Spike	**LoessM**	**Loess**	**Ploess**	**Vsn2**	**Nn**	**Olin**	**Olin_c**	**Turbo**	**Snm2**	**Vsn1**	**Q**	**Snm1**
Spike	87 (100%)	28 (32%)	21 (24%)	18 (21%)	24 (28%)	24 (28%)	21 (24%)	25 (29%)	23 (26%)	1 (1%)	25 (29%)	25 (29%)	16 (18%)
LoessM	28 (12%)	239 (100%)	183 (77%)	122 (51%)	147 (62%)	136 (57%)	153 (64%)	165 (69%)	173 (72%)	1 (0,4%)	95 (40%)	94 (39%)	82 (34%)
Loess	21 (10%)	183 (91%)	202 (100%)	114 (56%)	141 (70%)	116 (57%)	121 (60%)	130 (64%)	165 (82%)	0 (0%)	76 (38%)	77 (38%)	70 (35%)
Ploess	18 (10%)	122 (67%)	114 (63%)	181 (100%)	106 (59%)	115 (64%)	111 (61%)	113 (62%)	117 (65%)	3 (2%)	76 (42%)	74 (41%)	54 (30%)
Vsn2	24 (14%)	147 (89%)	141 (85%)	106 (64%)	166 (100%)	108 (65%)	116 (70%)	127 (77%)	148 (89%)	0 (0%)	93 (56%)	91 (55%)	76 (46%)
Nn	24 (10%)	136 (56%)	116 (48%)	115 (48%)	108 (45%)	242 (100%)	128 (53%)	132 (55%)	119 (49%)	5 (2%)	85 (35%)	87 (36%)	58 (24%)
Olin	21 (10%)	153 (73%)	121 (57%)	111 (53%)	116 (55%)	128 (61%)	211 (100%)	168 (80%)	119 (56%)	4 (2%)	92 (44%)	91 (43%)	64 (30%)
Olin_c	25 (12%)	165 (79%)	130 (62%)	113 (54%)	127 (61%)	132 (63%)	168 (80%)	209 (100%)	131 (63%)	4 (2%)	91 (44%)	92 (44%)	71 (34%)
Turbo	23 (11%)	173 (85%)	165 (81%)	117 (57%)	148 (72%)	119 (58%)	119 (58%)	131 (64%)	204 (100%)	0 (0%)	86 (42%)	87 (43%)	75 (37%)
Snm2	1 (3%)	1 (3%)	0 (0%)	3 (9%)	0 (0%)	5 (14%)	4 (11%)	4 (11%)	0 (0%)	35 (100%)	4 (11%)	6 (17%)	2 (6%)
Vsn1	25 (13%)	95 (48%)	76 (39%)	76 (39%)	93 (47%)	85 (43%)	92 (47%)	91 (46%)	86 (44%)	4 (2%)	196 (100%)	180 (92%)	78 (40%)
Q	25 (12%)	94 (46%)	77 (38%)	74 (36%)	91 (45%)	87 (43%)	91 (45%)	92 (45%)	87 (43%)	6 (3%)	180 (89%)	203 (100%)	79 (39%)
Snm1	16 (14%)	82 (69%)	70 (59%)	54 (46%)	76 (64%)	58 (49%)	64 (54%)	71 (60%)	75 (64%)	2 (2%)	78 (66%)	79 (67%)	118 (100%)



Rysunek 30. *Liczby genów dla zestawu AML II różnicujących współdzielonych między listami genów dla poszczególnych metod normalizacji. Dwie najbardziej zgodne metody to Loess i LoessM (183 wspólnych genów). Kolejne metody dodawano w zależności od stopnia podobieństwa: Turbo (155 genów wspólnych z Loess i LoessM), Olin_c (116 genów wspólnych z Loess, LoessM i Turbo), Olin (97 genów), Vsn2 (89 genów), Nn (70 genów), Ploess (59 genów), Vsn1 (38 genów), Q (36 genów), Snm1 (26 genów), Spike (3) i Snm2 (0 genów).*

Wyniki prezentowane w Tabeli 11 wskazują, iż najbardziej odmienne wyniki otrzymane zostały dla metody *Snm2*. Lista genów różnicujących dla tej metody nie tylko charakteryzowała się najmniejszą liczbą genów, ale także i najbardziej odmiennym składem

w stosunku do pozostałych procedur normalizacji. Jedynie 1-5 genów różnicujących (do 14%) współdzielonych jest z listami genów różnicujących otrzymanymi dla pozostałych metod normalizacji. Nieco lepsze wyniki otrzymane zostały dla metody normalizacji *Spike*, drugiej metody pod względem odmienności wyników. Spośród 87 genów wytypowanych jako różnicujące, do 32% genów było współdzielonych z wynikami dla pozostałych metod normalizacji. Dla pozostałych 11 metod normalizacji wyniki analizy podobieństwa składu list genów różnicujących były znacznie lepsze. Metoda *LoessM* wykazywała wysokie podobieństwo w stosunku do innych metod normalizacji. Największa zbieżność występowała pomiędzy metodą *LoessM*, a globalnymi metodami bazującymi na ważonej regresji liniowej: *Loess* i *Turbo* (odpowiednio 72% i 77% wspólnych genów). Metoda *Vsn2* umożliwiła otrzymanie listy genów różnicujących o zbliżonym składzie do metod: *Turbo*, *LoessM* oraz *Loess* (85-89% wspólnych genów). Stanowi to interesujący wynik, gdyż *Vsn2* opisywana jest w literaturze jako bardzo restrykcyjna metoda normalizacji, która w porównaniu z innymi dostępnymi metodami, pozwala na otrzymanie zwykle najniższej liczby genów różnicujących. Porównanie składu list genów różnicujących otrzymanych dla *Vsn2* z listami dla każdego z obu wariantów metody *Olin* (*Olin* oraz *Olin_c*) pozwoliło na otrzymanie 168 wspólnych genów różnicujących (80% wspólnych genów). Metoda *Ploess* charakteryzowała się wynikami zbliżonymi bardziej do metod *LoessM* i *Turbo*, niż do *Loess*, globalnej metody normalizacji z tego samego pakietu. Wyniki otrzymane dla metody *Nn* były najbardziej zgodne z wynikami dla *LoessM* i obu wersji metody *Olin*. W przypadku metod normalizacji dla danych jednokanałowych, bardzo zbliżone wyniki (89-92% wspólnych genów) otrzymane zostały dla metod *Q* i *Vsn1*.

V.III.2.6.2 Czułość i specyficzność analizy ekspresji różnicowej dla danych normalizowanych z wykorzystaniem wybranych metod

Kolejnym etapem było określenie czułości i specyficzności analizy ekspresji różnicowej dla danych normalizowanych z wykorzystaniem wybranych metod normalizacji. Czułość analizy ekspresji różnicowej dla każdej z testowanych metod normalizacji określona została na podstawie procentowego udziału kontroli pozytywnych (genów o potwierdzonej ekspresji różnicowej) w liście genów różnicujących w stosunku do całkowitej liczby kontroli pozytywnych (Tabela 12.). Natomiast specyficzność analizy różnicowej dla każdej z metod normalizacji oceniana jako procent negatywnych kontroli prawidłowo zaklasyfikowanych jako nieróżnicujące (Tabela 13). Za pozytywne kontrole uznane zostały geny, których ekspresja zweryfikowana została w wyniku analizy za pomocą ilościowego PCR lub geny

opisane w literaturze jako te ulegające nadekspresji w ostrej białaczce szpikowej lub niedojrzałych komórkach hematopoetycznych, np. *CD34*, *ENO1*, *AZU1* lub *HOX*. Geny te z dużym prawdopodobieństwem powinny ulegać ekspresji różnicowej. Natomiast negatywne kontrole stanowił zestaw genów metabolizmu podstawowego. Geny te powinny charakteryzować się stałym poziomem ekspresji w próbkach badanych oraz kontrolnych. Do tych genów należą m.in. *GAPDHS*, *VIM* oraz geny rodzin *PFK* oraz *RPL*. Podsumowując, wybrano 80 genów kontrolnych z czego 40 genów stanowi kontrole pozytywne, a pozostałe 40 kontrole negatywne eksperymentu.

Tabela 12. *Analiza czułości metod normalizacji dla zestawu AML II na podstawie weryfikacji obecności kontroli pozytywnych w listach genów różnicujących dla każdej z metod. Wartość „Tak" w polu oznaczonym kolorem niebieskim oznacza obecność danej kontroli, wartość „Nie" w polu oznaczonym kolorem białym oznacza brak obecności danej kontroli w liście genów różnicujących dla danej metody. Pola oznaczone kolorem fioletowym oznaczają nazwy 5 sond, odpowiadających 4 genom, których ekspresja różnicowa potwierdzona została za pomocą ilościowego PCR. Nazwy sond ułożono w kolejności alfabetycznej.*

Kontrole pozytywne	Metody normalizacji												
	Danych dwukanałowych										Danych jednokanałowych		
	Spike	LoessM	Loess	Ploess	Vsn2	Nn	Olin	Olin_c	Turbo	Snm2	Vsn1	Q	Snm1
AZU1	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Tak
BCL2	Nie	Nie	Nie	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie
BCL2L1	Nie	Tak	Tak	Tak	Nie	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Nie
BTG1	Nie	Nie	Nie	Nie	Tak	Tak	Nie	Nie	Tak	Nie	Tak	Tak	Tak
CD34	Nie	Tak	Tak	Tak	Nie	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Nie
CD34_O	Nie	Tak	Tak	Tak	Tak	Nie	Nie	Tak	Tak	Nie	Nie	Nie	Nie
CDK6	Nie	Tak	Tak	Nie	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Nie
CRYAA	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Nie	Nie	Nie	Nie
ENO1	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Tak
FTL3	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie
FLT3_O	Nie	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Tak	Nie	Tak	Tak	Nie
GATA2	Nie	Tak	Tak	Tak	Nie	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Nie
GJB1	Nie	Tak	Tak	Nie	Tak	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Tak
HCK	Nie	Tak	Nie	Nie	Nie	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Nie
HOXA10	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Nie
HOXA4	Nie	Tak	Tak	Nie	Nie	Nie	Tak	Nie	Tak	Nie	Nie	Nie	Nie
HOXA9	Nie	Tak	Nie	Nie	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Tak
HOXB2	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie
HOXB5	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Tak
HOXB6	Nie	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Tak	Nie	Tak	Tak	Tak
HRAS	Nie	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Tak	Nie	Nie	Tak	Tak
JUNB	Nie	Tak	Tak	Nie	Tak	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Nie
KIT	Nie	Tak	Nie	Nie	Nie	Nie	Tak	Tak	Nie	Nie	Tak	Tak	Nie
LTB	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Nie	Nie
MLLT1_O	Nie	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Tak	Nie	Tak	Nie	Tak
MLLT10	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Tak
MLLT4	Nie	Nie	Nie	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Tak	Tak	Nie
MN1	Nie	Tak	Tak	Nie	Tak	Nie	Nie	Nie	Tak	Tak	Nie	Nie	Tak
MPO	Nie	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Tak	Nie	Nie	Nie	Tak
NPM1	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Nie
PDE3B	Nie	Tak	Tak	Nie	Tak	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Nie
PF4	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Nie
PIM1	Nie	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Tak	Nie	Nie	Tak	Tak
PRG1	Nie	Tak	Tak	Nie	Tak	Nie	Nie	Tak	Tak	Nie	Tak	Tak	Tak
S1000A8_O	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Nie
S1000A9	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Nie	Nie	Nie
S1000A9_O	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Nie
SET	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Tak
STMN1	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Nie
TUBB	Nie	Tak	Nie	Tak	Tak	Tak	Tak	Tak	Nie	Nie	Tak	Tak	Tak
Ilość kontroli	*4*	*35*	*31*	*25*	*30*	*27*	*30*	*32*	*34*	*1*	*19*	*19*	*16*
Czułość [%]	*10*	*87,5*	*77,5*	*62,5*	*75*	*67,5*	*75*	*80*	*85*	*2,5*	*47,5*	*47,5*	*40*
Ranking	**9**	**1**	**4**	**8**	**5**	**7**	**5**	**3**	**2**	**10**	**1**	**1**	**3**

Tabela 13. *Analiza specyficzności metod normalizacji dla zestawu AML II na podstawie weryfikacji obecności kontroli negatywnych w listach genów różnicujących dla każdej z metod. Wartość „Tak" w polu oznaczonym kolorem pomarańczowym oznacza obecność danej kontroli, wartość „Nie" w polu oznaczonym kolorem białym oznacza brak obecności danej kontroli w liście genów potencjalnie różnicujących dla danej metody. Nazwy sond ułożono w kolejności alfabetycznej.*

Kontrole negatywne	Metody normalizacji													
	Danych dwukanałowych											Danych jednokanałowych		
	Spike	LoessM	Loess	Ploess	Vsn2	Nn	Olin	Olin_c	Turbo	Snm2		Vsn1	Q	Snm1
AAMP	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
ACTG1	Tak	Nie	Nie	Nie	Nie	Tak	Nie	Nie	Nie	Nie		Nie	Nie	Nie
ALDOC	Nie	Nie	Nie	Nie	Nie	Tak	Nie	Nie	Nie	Nie		Tak	Tak	Nie
ARF1	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
CANX	Nie	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie	Nie		Nie	Nie	Nie
CLU	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
FTL	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Tak	Tak	Tak
G6PD	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Tak	Nie	Nie		Nie	Nie	Nie
GAPDHS	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
H3F3A	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
H3F3A_O	Nie	Tak	Tak	Tak	Nie	Tak	Nie	Tak	Nie	Nie		Nie	Nie	Nie
HPRT1	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
HSP90AA1	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
LDHA	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Tak		Nie	Nie	Nie
LDHALGA	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
LDHC	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
MONO	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
MT2A	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Tak	Tak	Nie
NONO_O	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
PFKL	Nie	Nie	Nie	Nie	Nie	Tak	Nie	Nie	Nie	Nie		Nie	Nie	Nie
PFKM	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
PFKP	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
PGAM1	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
PGK1	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak	Nie		Nie	Nie	Tak
PGK2	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
RAC2	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
RPL0	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
RPL11	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
RPL19	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
RPL37A	Nie	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak		Tak	Tak	Nie
RPL5	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
*RPLP1_*	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Tak	Nie	Tak
RPS27A	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
RPS29	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
RPS3	Nie	Nie	Nie	Nie	Nie	Tak	Nie	Nie	Nie	Nie		Nie	Nie	Nie
TCEA1	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Tak
TCFL1	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
TMSB4X	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
TUBA1	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
TUBB	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie		Nie	Nie	Nie
Ilość kontroli	*38*	*37*	*38*	*37*	*39*	*34*	*39*	*37*	*39*	*38*		*35*	*36*	*36*
Specyficzność [%]	*95*	*92,5*	*95*	*92,5*	*97,5*	*85*	*97,5*	*92,5*	*97,5*	*95*		*87,5*	*90*	*90*
Ranking	**4**	**7**	**4**	**7**	**1**	**10**	**1**	**7**	**1**	**4**		**3**	**1**	**1**

Wyniki analizy ekspresji różnicowej (Rysunek 30, Tabela 11- Tabela 13) potwierdziły ogólne podobieństwo globalnych metod normalizacji opartych na ważonej regresji liniowej (*Loess, LoessM, Ploess, Spike, Turbo, Olin, Olin_c*), które charakteryzowały się najbardziej spójnym składem list genów różnicujących. Ponownie najbardziej odrębnymi wynikami charakteryzowała się metoda *Snm2*. Czułości i specyficzność analizy ekspresji różnicowej dla danych normalizowanych z użyciem tej metody była dość niska o czym świadczy prawidłowe zaklasyfikowanie jedynie 1 pozytywnej (Tabela 12) oraz błędne zaklasyfikowanie 2

negatywnych kontroli (Tabela 13). Podobny rezultat został uzyskany przy użyciu metody *Spike*. Najwyższą czułością i specyficznością analizy ekspresji różnicowej charakteryzują się dane normalizowane za pomocą metody *Turbo* oraz *LoessM*.

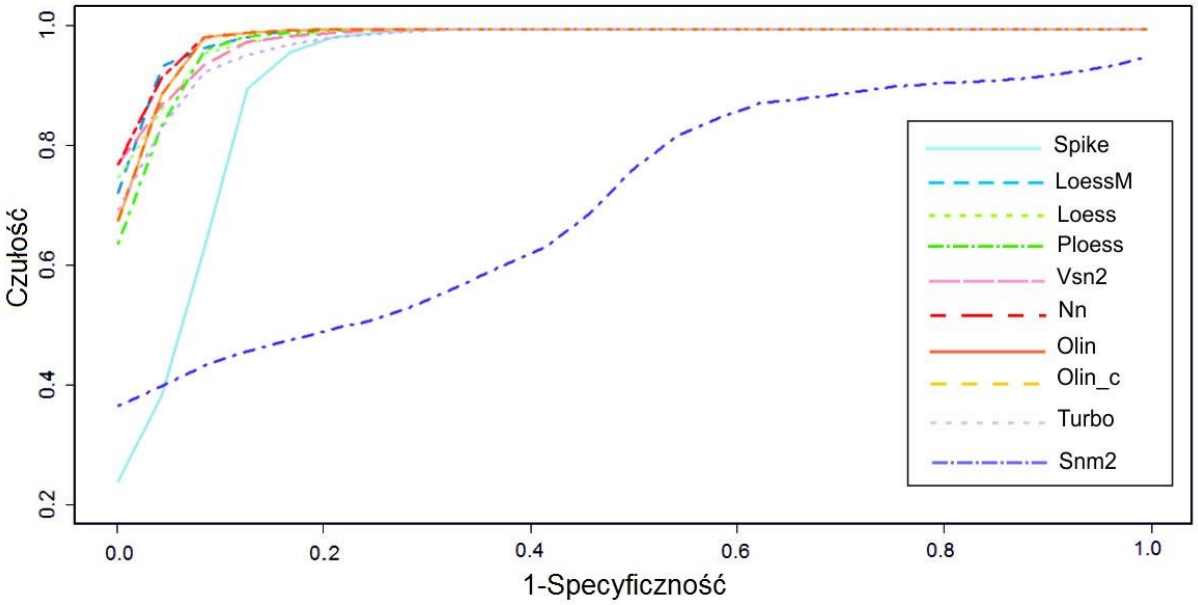
Jednokanałowe metody normalizacji charakteryzowały się nieco gorszym rezultatem. Analiza czułości dla danych normalizowanych z użyciem metod *Q* lub *Vsn2* wskazała na prawidłową klasyfikację 19 (47,5%), natomiast w przypadku metody Snm1 jedynie 16 kontroli pozytywnych (40%). Analiza specyficzności dla danych normalizowanych z wykorzystaniem metod jednokanałowych wykazała błędną klasyfikację 4 kontroli negatywnych (90%) dla metod *Q* oraz *Snm1* oraz 5 kontroli negatywnych (87,5%) dla metody *Vsn1*. Warto jednak podkreślić, iż w przypadku danych jednokanałowych różnice pomiędzy najlepszą, a najgorszą metodą normalizacji nie są tak duże jak w przypadku dwukanałowych metod normalizacji.

V.III.2.7 Krzywe ROC i wartości AUC

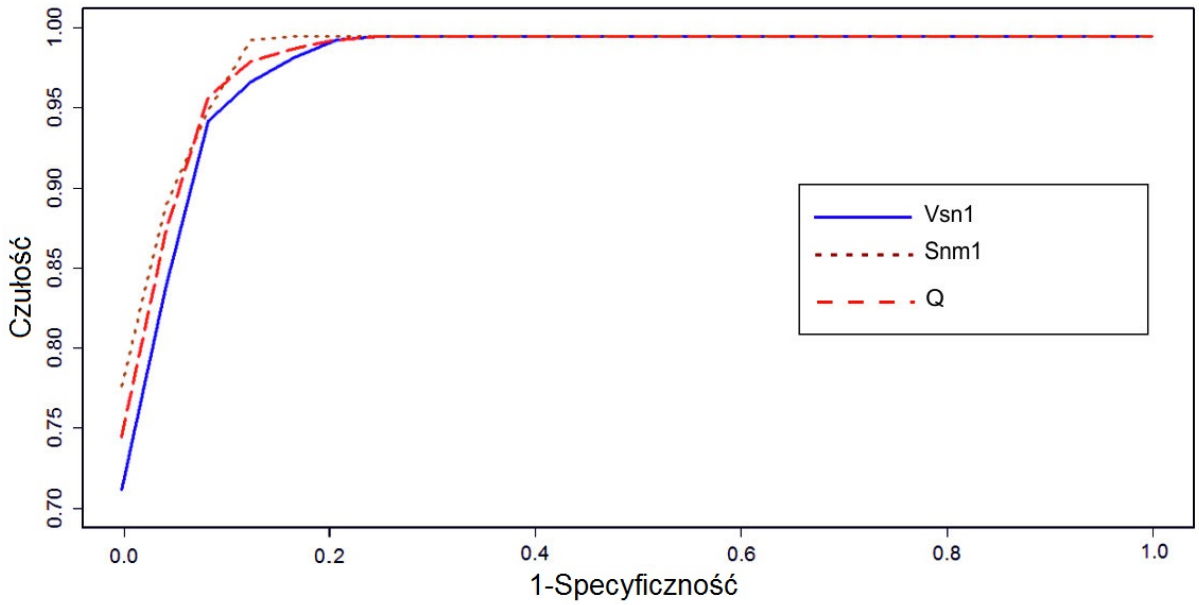
Ostatnim kryterium wyboru optymalnej metody normalizacji było sprawdzenie skuteczności klasyfikacji próbek (zdrowi ochotnicy i pacjenci z AML) za pomocą walidacji krzyżowej i krzywych ROC (ang. *receiver operating characteristics*) dla każdej z testowanych metod normalizacji. Klasyfikacja próbek wykonana została z wykorzystaniem łączonego klasyfikatora regresji stosowej (ang. *stacked regression*), zbudowanego na podstawie wartości ekspresji genów zidentyfikowanych jako najbardziej różnicujące dla zestawu AML II. Aby wyniki krzywych ROC były porównywalne dla każdej z metod normalizacji brana była pod uwagę taka sama liczba parametrów (genów różnicujących). Zgodnie z prezentowanymi wynikami analizy różnicowej (Tabela 11), najniższa liczba genów różnicujących wynosi 35 (metoda *Snm2*). Dlatego też do tworzenia klasyfikatorów wybrano 35 najbardziej różnicujących genów z każdej z list, czyli tych charakteryzujących się największymi zmianami poziomów ekspresji. Krzywe ROC generowane były w oparciu o walidację krzyżową typu *leave-one-out*, która jest szczególnym przypadkiem K-krotnej walidacji, gdzie K=40 jest jednocześnie wielkością analizowanej próby, jak i ilością podzbiorów na które jest ona dzielona. Oznacza to, że każda z próbek została użyta jako zbiór testowy klasyfikatora, który był budowany na podstawie pozostałych 39 próbek. Krzywe ROC dla dwu- i jednokanałowych metod normalizacji wygenerowane zostały niezależnie.

Otrzymane rezultaty (Rysunek 31-Rysunek 32) pokazują, iż krzywe ROC dla wszystkich testowanych dwukanałowych metod normalizacji, z wyjątkiem *Snm2* i *Spike*, mają

zbliżony kształt. Niekiedy nawet krzywe dla poszczególnych metod normalizacji wzajemnie się na siebie nakładają. Podobny wynik widoczny jest dla metod jednokanałowych. Na potrzeby oszacowania różnic pomiędzy krzywymi ROC wykorzystano wartości AUC (ang. *area under curve*). Wyniki przedstawiono w Tabeli 14.



Rysunek 31. *Krzywe ROC dla klasyfikatorów utworzonych na podstawie informacji dla 35 genów zidentyfikowanych w zestawie AML II dla testowanych dwukanałowych metod normalizacji.*



Rysunek 32. *Krzywe ROC dla klasyfikatorów utworzonych na podstawie informacji dla 35 genów zidentyfikowanych w zestawie AML II dla testowanych jednokanałowych metod normalizacji.*

Tabela 14. *Wartości AUC (zaokrąglone do dwóch miejsc po przecinku) dla każdej z metod normalizacji testowanych na zestawie AML II, obliczone dla klasyfikatorów utworzonych na podstawie informacji dla 35 genów potencjalnie różnicujących.*

Dwukanałowe metody normalizacji												Jednokanałowe metody normalizacji		
Metoda	**Spike**	**LoessM**	**Loess**	**Ploess**	**Vsn2**	**Nn**	**Olin**	**Olin_c**	**Turbo**	**Snm2**		**Vsn1**	**Q**	**Snm1**
Wartość AUC	0,91	0,99	0,99	1	0,99	1	1	1	0,98	0,58		0,99	1	1
Ranking	**9**	**5**	**5**	**1**	**5**	**1**	**1**	**1**	**8**	**10**		**3**	**1**	**1**

V.III.2.8 Ostateczny ranking metod normalizacji i ustalenie zobiektywizowanej procedury wyboru optymalnej metody normalizacji

W wyniku połączenia wszystkich kryteriów dla metod normalizacji opisywanych w poprzednich sekcjach można stwierdzić, która z testowanych metod normalizacji jest najbardziej odpowiednia do normalizacji zestawu danych AML II. Wyniki przedstawione w Tabeli 15 stanowią podsumowanie wszystkich prezentowanych wcześniej rankingów (błąd systematyczny, wariancja, czułość, specyficzność oraz klasyfikacja próbek na podstawie profilu ekspresji genów) zawartych odpowiednio w Tabelach 9, 10, 12, 13 oraz 14. Ostateczny ranking wykonano w oparciu o średnią arytmetyczną z pozycji zajmowanych przez daną metodę w poprzednich listach rankingowych.

Tabela 15. *Ostateczny ranking metod normalizacji dla zestawu danych AML II otrzymany na postawie informacji o odchyleniu, zmienności, biologicznej weryfikacji wyników oraz wartości AUC.*

Metoda normalizacji	Ranking					Średnia pozycja	Ranking ostateczny
	Błąd systematyczny	Wariancja	Analiza ekspresji różnicowej		AUC		
			Czułość	Specyficzność			
Metody dla danych dwukanałowych							
Spike	10	10	9	4	9	8,4	10
LoessM	8	7	1	7	5	5,6	6
Loess	2	2	4	4	5	3,4	3
Ploess	7	9	8	7	1	6,4	8
Vsn2	4	3	5	1	5	3,6	4
Nn	6	6	7	10	1	6	7
Olin	3	4	5	1	1	2,8	2
Olin_c	5	5	3	7	1	4,2	5
Turbo	1	1	2	1	8	2,6	1
Snm2	9	8	10	4	10	8,2	9
Metody dla danych jednokanałowych							
Vsn1	2	2	1	3	3	2,2	2
Q	1	1	1	1	1	1	1
Snm1	3	3	3	1	1	2,2	2

Według wyników zebranych w Tabela 15, najlepsze rezultaty wśród dwukanałowych metod normalizacji dla zestawu AML II osiągnęły globalne metody normalizacji oparte na lokalnie ważonej regresji liniowej. Metody *Turbo* i *Olin* zajmują kolejno pierwsze i drugie miejsce w ostatecznym rankingu. Tuż za nimi w rankingu klasują się metody *Loess* i *Olin_c*.

Dwa ostatnie miejsca natomiast zajmowane są przez metody *Spike* oraz *Snm2*. W przypadku metod jednokanałowych, najlepsza okazała się metoda *Q*. Jednakże różnice pomiędzy metodą *Q*, a pozostałymi metodami jednokanałowymi są subtelne. Wyniki porównania metod normalizacji mogą być różne dla różnych zestawów danych. W związku z tym zasadne wydaje się być zaproponowanie uniwersalnej i zobiektywizowanej procedury wyboru metody normalizacji dla danego zestawu. Procedura ta składa się z kilku kroków:

1. Normalizacja danych za pomocą kilku wybranych metod
2. Ocena wybranych metod normalizacji na podstawie 5 kryteriów:
    a) Średniej wartości błędu systematycznego dla sond kontrolnych znormalizowanego zestawu lub zestawów danych. Ranking metod normalizacji na podstawie średniej wartości błędu systematycznego.
    b) Średniej wartości wariancji pomiędzy powtórzeniami sond kontrolnych w obrębie znormalizowanego zestawu lub zestawów danych. Ranking metod normalizacji na podstawie średniej wartości wariancji.
    c) Czułość analizy ekspresji różnicowej w oparciu o liczbę kontroli pozytywnych prawidłowo zaklasyfikowanych jako różnicujące. Ranking metod normalizacji na podstawie wartości czułości.
    d) Specyficzność analizy ekspresji różnicowej na podstawie liczby kontroli negatywnych prawidłowo zaklasyfikowanych jako nieróżnicujące. Ranking metod normalizacji na podstawie wartości specyficzności.
    e) Zdolność klasyfikacji próbek określana w oparciu o profil ekspresji genów za pomocą krzywych ROC i wartości AUC. Ranking metod normalizacji na podstawie wartości AUC.
3. Ostateczny ranking metod normalizacji otrzymany w wyniku podsumowania list rankingowych dla wszystkich 5 opisywanych wyżej kryteriów. Pozycja danej metody w rankingu ostatecznym ustalana jest poprzez wyznaczenie średniej pozycji metody obliczonej na podstawie wszystkich list rankingowych wchodzących w skład rankigu ostatecznego. Metoda zajmująca pierwszą pozycję na liście jest uważana za najbardziej odpowiednią dla danego zestawu danych.

V.III.3 Przykład wykorzystania wyników

Prezentowana uniwersalna i zobiektywizowana procedura wyboru metody normalizacji wchodzi w skład procedur stosowanych w Zakładzie Biologii Molekularnej i

Systemowej na etapie analizy danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA. Opisywana procedura została zastosowana m.in. przy wyborze metody normalizacji danych uzyskanych w ramach:

- badania ekspresji genów u pacjentów z ostrą białaczką szpikową (zestaw AML II);
- badania ekspresji genów *Arabidopsis thaliana* w warunkach szoku cieplnego;
- badania zmian ekspresji genów (mRNA) w ludzkich komórkach śródbłonka żyły pępowinowej (HUVEC) hodowanych pod wpływem homocysteiny, tiolaktonu homocysteiny oraz N-homocysteinylowanych białek surowicy;
- Identyfikacji genów ulegających różnicowej ekspresji w ludzkich komórkach nabłonka jelita pod wpływem adhezji bakterii prebiotycznych.

V.III.4 Omówienie wyników

Siła technologii ekspresyjnych mikromacierzy DNA wynika z ich dostępności, szerokich możliwości wykorzystania, miniaturyzacji oraz stosunkowo niskich kosztów eksperymentów. Dedykowane mikromacierze DNA są znacznie tańsze, niż mikromacierze DNA o wysokiej gęstości. Warto jednak wspomnieć, iż dane uzyskiwane z użyciem dedykowanych mikromacierzy DNA wymagają większej uwagi na etapie analizy niższego rzędu. Dotyczy to głównie procesu normalizacji danych. Niebezpieczeństwo wynikające z zastosowania nieodpowiedniej metody normalizacji polega na usunięciu nie tylko zmienności technicznej, ale także i tej wynikającej z różnic biologicznych.

Pomimo, iż istnieje kilka metod normalizacji zaprojektowanych lub zmodyfikowanych w celu rozwiązania problemu normalizacji danych uzyskiwanych z użyciem dedykowanych mikromacierzy DNA, większość z nich jest ograniczona do zestawów danych konkretnego typu. Przykładem jest metoda *wloess* zaproponowana przez Oshlack i wsp. (2007). Procedura *wloess* polega na wprowadzeniu ilościowych wag dla oznaczenia efektów zależnych od intensywności sygnału i stanowi alternatywne podejście dla metody *composite* wprowadzonej przez Yang i wsp. (2002). Jednakże zarówno metoda *wloess*, jak i *composite* wymagają użycia dużego zestawu sond kontrolnych, najlepiej typu MSP (ang. *microarray sample pool*). Sondy typu MSP mogą być zastąpione innymi sondami kontrolnymi, np. typu *spike-in*. Warunkiem koniecznym jest jednak duża liczba sond kontrolnych występująca na mikromacierzy w kilku różnych stężeniach.

Każdy eksperyment z wykorzystaniem dedykowanych mikromacierzy DNA wymaga indywidualnego podejścia do kwestii analizy danych. Zazwyczaj wybór metody normalizacji wynika z oceny wykresów diagnostycznych: wykresy pudełkowe, wykresy MA. Jednakże stosowanie narzędzi graficznych do oceny skuteczności metody normalizacji jest często bardzo intuicyjne i opiera się na subiektywnym wrażeniu eksperymentatora. Znacznie bardziej obiektywną formą oceny normalizacji jest podejście zaproponowane przez Argyropoulos i wsp. (2006). Według autorów, najważniejszym aspektem wyboru algorytmu normalizacji jest dokładność, precyzja i efekt nadmiernej normalizacji. Nadmierna normalizacja zestawu danych pojawia się w przypadku zastosowania za bardzo restrykcyjnego modelu normalizacji i skutkuje usunięciem nie tylko zmienności technicznej, ale także i tej o podłożu biologicznym. Wszystkie te trzy aspekty (dokładność, precyzja i efekt nadmiernej normalizacji) mogą być zweryfikowane poprzez następujące kryteria: błąd systematyczny (ang. *bias*), wariancja oraz entropia względna (ang. *relative entropy*). Niska wartość błędu systematycznego oznacza większą dokładność normalizacji, natomiast niska wariancja równa się większej precyzji normalizacji. Metoda normalizacji dla której otrzymana wartość entropii względnej dla rozkładu logarytmicznego jest niska wykazuje mniejszą skłonność do nadmiernej normalizacji danych. Stąd też za pomocą parametrów ilościowych, możliwe jest bezpośrednie porównanie wyników procedur normalizacji i wybór optymalnej metody. Wartości błędu systematycznego oraz wariancji powinny być obliczone dla sond kontrolnych, które występują na mikromacierzy w dostatecznie dużej ilości. Oszacowanie zjawiska nadmiernej normalizacji jest możliwe jedynie, wówczas gdy zestaw danych zawiera tożsame hybrydyzacje (ang. *self-self hybridizations*). Wyrażenie tożsame hybrydyzacje odnosi się do sytuacji, gdzie mieszanina dwóch takich samych próbek, znakowanych różnymi barwnikami fluorescencyjnymi hybrydyzowana jest do jednej mikromacierzy DNA.

W praktyce, eksperymenty z użyciem mikromacierzy DNA nie zawsze przestrzegają rygorystycznych wymagań jakimi są standardy jakości. Stąd też wysokie wartości błędu systematycznego i wariancji mogą wynikać nie tylko z zastosowania mało optymalnej metody normalizacji, ale także z dużej zmienności technicznej wprowadzonej na etapie realizacji eksperymentu. Zestawy danych uzyskiwane z użyciem dedykowanych mikromacierzy DNA, często mają niestandardowy charakter, stąd też procedura wyboru metody normalizacji dla takich zestawów danych jest znacznie bardziej utrudniona i nie powinna być prowadzona jedynie w oparciu o wartości błędu systematycznego oraz wariancji. Zatem konieczne jest

uwzględnienie dodatkowych parametrów umożliwiających wybór odpowiedniej metody normalizacji.

Proponowana w tej części pracy procedura wyboru metody normalizacji dla danego zestawu danych pozwala ocenić efekt normalizacji na podstawie 5 kryteriów: błędu systematycznego, wariancji, czułości i specyficzności analizy ekspresji różnicowej oraz zdolności klasyfikacji próbek na podstawie profilu ekspresji genów za pomocą krzywych ROC. Przeprowadzenie pełnej procedury wyboru metody normalizacji możliwe jest dla danych dla których w wyniku analizy ekspresji różnicowej otrzymano zestaw genów różnicujących. W przypadku zestawów danych dla których analiza ekspresji różnicowej nie pozwoliła na selekcję genów różnicujących możliwa jest ocena metod normalizacji jedynie na podstawie wartości błędu systematycznego i wariancji. Ustalenie 3 spośród 5 stosowanych kryteriów dla rezultatów analizy ekspresji różnicowej wynika z faktu, iż rodzaj użytej metody normalizacji ma znaczący wpływ na przebieg procesu selekcji genów różnicujących. Proponowane kryteria pozwalają ustalić na podstawie wyniku analizy ekspresji różnicowej, która z testowanych metod normalizacji jest najbardziej odpowiednia dla danego zestawu danych.

Spośród 4 wybranych zestawów danych przeprowadzenie pełnej procedury możliwe było jedynie dla zestawu AML II. Wynikiem analizy ekspresji różnicowej danych z zestawów ASTMA i ALERGIA był brak genów różnicujących przy zastosowaniu każdej z wybranych metod normalizacji. Natomiast w przypadku zestawu OSHLACK informacje dotyczące sond umieszczonych na mikromacierzy DNA nie były wystarczające do oceny czułości i specyficzności analizy ekspresji różnicowej. Stąd też ocena metod normalizacji dla zestawów ASTMA, ALERGIA oraz OSHLACK prowadzona była jedynie w oparciu o wartości błędu systematycznego i wariancji.

Wszystkie stosowane w tej części pracy metody normalizacji pochodzą z repozytorium Bioconductor. R/Bioconductor jest aktualnie jednym z najbardziej popularnych programów do analizy danych uzyskiwanych z użyciem ekspresyjnych mikromacierzy DNA. Wynika to z jego dostępności (oprogramowanie typu *open source*) oraz faktu, iż jako jeden z nielicznych programów umożliwia kompletną analizę (niższego i wyższego rzędu) danych uzyskiwanych z użyciem dedykowanych mikromacierzy DNA.

Pełna procedura wyboru metody normalizacji, przygotowana w oparciu o własne kryteria i te opracowane przez Argyropoulos i wsp. (2006), obejmowała porównanie 13

różnych metod normalizacji i została przeprowadzona na przykładzie zestawu AML II. W skład testowanych metod wchodziło: 10 dwukanałowych metod normalizacji oraz 3 jednokanałowe. Niektóre z testowanych metod, np. *Loess* lub *Vsn* są dobrze znane, podczas gdy inne: *Turbo*, *Olin*, *Nn* oraz *Snm* nieco mniej. Z połączenia wszystkich badanych parametrów (błąd systematyczny, wariancja, czułość i specyficzność analizy ekspresji różnicowej danych znormalizowanych z użyciem wybranych metod, krzywe ROC) wynika, że spośród dwukanałowych metod normalizacji to metody globalne oparte na modelu ważonej regresji liniowej (metody *loess*) są najbardziej optymalnym rozwiązaniem dla normalizacji zestawu AML II. Dotyczy to w szczególności metody *Turbo* z pakietu `TurboNorm` oraz *Olin* z pakietu `olin`. Nieco zaskakujący wynik uzyskano dla metody *Spike*, która jest powszechnie stosowaną procedurą normalizacji dla dedykowanych mikromacierzy DNA (Dabney & Storey 2007). Niska skuteczność tej metody w przypadku danych AML II wynika najprawdopodobniej z niewielkiej liczby kontroli typu *spike-in*. Jednakże metoda *Spike* nie sprawdziła się także w przypadku zestawu OSHLACK, gdzie liczba kontroli MSP była znacznie większa, o czym świadczą wysokie wartości błędu systematycznego i wariancji. Gorszy wynik metody *Ploess* dla zestawu AML II jest spowodowany zbyt niską liczbą sond w obrębie danego bloku. Procedura *print-tip loess* nie może być stosowana do normalizacji danych, gdzie liczba sond w pojedynczym bloku jest niższa niż 150 (G. K. Smyth & T. Speed 2003). Zestaw AML II zawiera w każdym z bloków jedynie 81 sond. Stosowanie procedury *Ploess* jest także ryzykowne w przypadku zestawów danych, które charakteryzują się obecnością wielu brakujących wartości (luk). Ostatnie kryterium wyjaśnia słabą skuteczność metody *Ploess* w przypadku zestawów ALERGIA i ASTMA i wysokie wartości błędu systematycznego oraz wariancji, które zostały otrzymane dla tych zestawów. W skład pojedynczego bloku z zestawu OSHLACK wchodziły 462 sondy, z których większość charakteryzowała się dość wysokimi wartościami intensywności. Stąd też metoda *Ploess* okazała się być najlepszą procedurą normalizacji dla tego zestawu danych na podstawie wstępnej analizy błędu systematycznego i wariancji.

Projekt zestawu AML II umożliwiał przekształcenie go z zestawu dwukolorowego w jednokolorowy bez uszczerbku dla prowadzonej analizy (Tabela 8), której celem było porównanie ekspresji genów u pacjentów z AML względem zdrowych ochotników. Zastosowanie jednokanałowych metod normalizacji jest polecane w przypadku dwukanałowych zestawów, gdzie nie wykonano zamiany barwników fluorescencyjnych na etapie znakowania próbek (ang. *dye swaps*). W takim przypadku zmienność wynikająca z

różnic właściwości chemicznych pomiędzy użytymi barwnikami nie może zostać wyeliminowana z układu, co w znacznym stopniu zaburza końcowy pomiar poziomu ekspresji genów. Eliminacja tego rodzaju zmienności jest częściowo możliwa na etapie normalizacji metodami dwukanałowymi, jednakże w zależności od skali zjawiska, może przebiegać z różną wydajnością. Wybór rodzaju metody normalizacji (jedno- lub dwukanałowej), gdy obie są dostępne pozostaje do decyzji badacza. W przypadku zestawu AML II spośród wszystkich testowanych jednokanałowych metod normalizacji najbardziej optymalnym podejściem jest metoda *Q* z pakietu `limma`.

Podsumowując, celem analizy nie jest wskazanie najlepszej metody normalizacji, spośród tych dostępnych w ramach repozytorium Bioconductor, a jedynie wykazanie, że każdy zbiór danych może wymagać innego podejścia na etapie normalizacji. Wybór procedury normalizacji danych ma ogromny wpływ na wyniki końcowe analizy i powinien być dokładnie rozważony, zwłaszcza w przypadku danych uzyskiwanych z użyciem dedykowanych mikromacierzy DNA.

V.III.5 Wnioski

Na podstawie otrzymanych wyników możliwe było sformułowanie następujących wniosków:

- Proces normalizacji danych jest najbardziej kluczowym etapem w analizie danych uzyskiwanych z użyciem dedykowanych mikromacierzy DNA. Wynika to z faktu, iż w przypadku tych danych istnieje duże ryzyko nadmiernej normalizacji, czyli eliminacji oprócz różnic technicznych także tych o podłożu biologicznym. Ponadto, wybór metody normalizacji w znaczący sposób wpływa na wyniki analizy wyższego rzędu.
- Głównym utrudnieniem na etapie normalizacji danych uzyskiwanych z użyciem dedykowanych mikromacierzy był brak zobiektywizowanych i uniwersalnych kryteriów oceny efektu normalizacji.
- Klasą metod normalizacji, która pozwala na skuteczną normalizację danych uzyskiwanych z użyciem dedykowanych mikromacierzy DNA są metody oparte na lokalnie ważonej regresji liniowej (metody typu *loess*). Wynika to najprawdopodobniej z faktu, iż metody te są w stanie skutecznie znormalizować dane dla których liczba genów różnicujących stanowi nawet ok. 30% całkowitej liczby genów.

Część IV: Analiza danych uzyskiwanych w ramch hybrydyzacji międzygatunkowej

Hybrydyzacja międzygatunkowa (CSH, ang. *cross-species hybridization*) polega na hybrydyzacji RNA pochodzącego z organizmu jednego gatunku (gatunek badany) do mikromacierzy DNA dla innego gatunku (gatunek referencyjny). Choć CSH ma szereg zastosowań w badaniach ewolucyjnych, najczęściej stosowana jest w przypadkach braku dostępności komercyjnej mikromacierzy DNA do badania ekspresji genów dla danego gatunku. Pomimo swojej popularności, hybrydyzacja międzygatunkowa wciąż jest jednym z przykładów niestandardowego sposobu wykorzystania ekspresyjnych mikromacierzy DNA. Wynika to z faktu, iż dane uzyskiwane w ramach CSH posiadają szereg unikatowych cech. Niestandardowy charakter tych danych spowodowany jest głównie obecnością dużej ilości niskiej jakości sygnałów, powstałych z powodu różnic w komplementarności pomiędzy sekwencjami gatunku badanego i referencyjnego. Schemat analizy danych uzyskiwanych w ramach CSH, w przypadku dostępności informacji o sekwencji genomowej dla badanego gatunku, obejmuje proces filtracji danych na podstawie homologii sekwencji. W wyniku tego procesu z zestawu danych eliminowane są sondy o obniżonej komplementarności, co znacznie poprawia jakość otrzymywanych wyników (Bar-Or i wsp. 2006; Khaitovich i wsp. 2004). Jednakże, dla wielu gatunków badanych w ramach CSH dostępność informacji o sekwencji genomowej jest ograniczona. W takim przypadku proces analizy danych jest znacznie bardziej utrudniony. Rozwiązaniem może być filtracja danych na podstawie parametrów określających morfologię poszczególnych punktów na mikromacierzy (SC, ang. *spot characteristics*), tzw. filtracja morfologiczna (Bar-Or, Novikov, i wsp. 2007). Według Bar-Or i wsp. zaburzona morfologia (obniżona jakość) punktów może wskazywać na obniżoną homologię sond i ich sekwencji docelowych, a tym samym może być wykorzystywana do filtracji sond o obniżonej homologii bez konieczności znajomości sekwencji docelowej. Autorzy pracy badali wpływ homologii sekwencji na wartości 10 parametrów SC generowanych przez program do analizy ilościowej obrazu: MAIA (Novikov & Barillot 2007) (Novikov & Barillot 2005).

V.IV.1 Obiekt analizy

Celem tej części pracy doktorskiej była ocena filtracji morfologicznej, jako jednego z etapów analizy niższego rzędu dla danych uzyskiwanych w ramach CSH. Zestaw NT-CSH został otrzymany w ramach badania ekspresji genów tytoniu szlachetnego (*Nicotiana tabacum*) pod wpływem stresu abiotycznego z zastosowaniem hybrydyzacji

międzugatunkowej (Materiały i Metody, rozdział IV.I.5). W skład zestawu NT-CSH wchodzą trzy rodzaje ekspresyjnych mikromacierzy DNA: mikromacierze cDNA dla ziemniaka (*Solanum tuberosum*) (The Institute for Genomic Research), mikromacierze cDNA dla pomidora (*Lycopersicon esculentum*) (mikromacierze TOM1, The Boyce Thompson Institute) oraz mikromacierze DNA dla rzodkiewnika (*Arabidopsis thaliana*) (mikromacierze z sondami w postaci oligonukleotydów-60 mer, University of Arizona). Każda z trzech rodzajów ekspresyjnych mikromacierzy DNA tworzy podzestaw w ramach zestawu NT-CSH: podzestaw POT (mikromacierze cDNA dla ziemniaka), podzestaw TOM1 (mikromacierze cDNA dla pomidora) oraz referencyjny podzestaw ATH (mikromacierze DNA dla rzodkiewnika). Podzestawy POT i TOM1 otrzymane zostały w wyniku hybrydyzacji próbek uzyskiwanych z roślin tytoniu traktowanego odpowiednio NaCl lub $CdCl_2$ (próbka badana) oraz roślin tytoniu nie poddanych działaniu czynników stresogennych (próbka referencyjna). W skład zestawów POT i TOM1 wchodzi po 12 mikromacierzy DNA, z czego 6 mikromacierzy (3 hybrydyzacje i 3 hybrydyzacje typu *dye swap*) dla roślin traktowanych NaCl i 6 mikromacierzy DNA (3 hybrydyzacje i 3 hybrydyzacje typu *dye swap*) dla roślin traktowanych $CdCl_2$ (Tabela 16). Podzestaw ATH złożony jest z 8 mikromacierzy DNA. Zestaw ten otrzymano w wyniku hybrydyzacji próbek uzyskiwanych z roślin *A.thaliana* traktowanych odpowiednio NaCl lub $CdCl_2$ (próbka badana) oraz roślin *A.thaliana* nie poddanych działaniu czynników stresogennych. W skład 8 mikromacierzy DNA wchodzą 4 mikromacierze (2 hybrydyzacje i 2 hybrydyzacje typu *dye swap*) dla roślin traktowanych NaCl oraz 4 mikromacierze (2 hybrydyzacje i 2 hybrydyzacje typu *dye swap*) dla roślin traktowanych $CdCl_2$ (Tabela 16).

Tabela 16. *Skład zestawu NT-CSH z podziałem na poszczególne podzestawy: POT (kolor żółty), TOM1 (kolor pomarańczowy) oraz referencyjny ATH (kolor szary). Dany podzestaw wyznaczany jest przez rodzaj stosowanej ekspresyjnej mikromacierzy DNA. Oznaczenia Test oraz Ref oznaczają odpowiednio próbki uzyskane od roślin poddanych oraz niepoddanych działaniu czynników stresogennych (NaCl lub $CdCl_2$), znakowanych danym barwnikiem fluorescencyjnym (Cy5 lub Cy3).*

Zestaw NT-CSH						
Nazwa podzestawu	**POT**		**TOM1**		**ATH**	
Gatunek referencyjny	Ziemniak		Pomidor		Rzodkiewnik	
Gatunek badany	Tytoń		Tytoń		Rzodkiewnik	
Barwnik	Cy5	Cy3	Cy5	Cy3	Cy5	Cy3
Czynnik stresogenny	**NaCl**		**NaCl**		**NaCl**	
Hybrydyzacja	Test	Ref	Test	Ref		
	Test	Ref	Test	Ref	Test	Ref
	Test	Ref	Test	Ref	Test	Ref
Hybrydyzacja typu *dye swap*	Ref	Test	Ref	Test		
	Ref	Test	Ref	Test	Ref	Test
	Ref	Test	Ref	Test	Ref	Test
Czynnik stresogenny	**$CdCl_2$**		**$CdCl_2$**		**$CdCl_2$**	
Hybrydyzacja	Test	Ref	Test	Ref		
	Test	Ref	Test	Ref	Test	Ref
	Test	Ref	Test	Ref	Test	Ref
Hybrydyzacja typu *dye swap*	Ref	Test	Ref	Test		
	Ref	Test	Ref	Test	Ref	Test
	Ref	Test	Ref	Test	Ref	Test

V.IV.2 Parametry jakości punktów (SC)

Opis 10 parametrów SC charakteryzujących morfologię punktów, wykorzystywanych przez Bar-Or i wsp. przedstawiony został w Tabeli 17. W dalszej części pracy stosowane będą wyłącznie skrócone nazwy parametrów SC prezentowane w Tabeli 17. Według Bar-Or i wsp. spośród 10 badanych parametrów SC, tylko 4: CVR, Det, Dia, GSym mają związek z poziomem homologii sekwencji sond i sekwencji docelowych. Jednocześnie Bar-Or i wsp. w swojej pracy sugerują, iż ustalenie zależności pomiędzy wartościami parametrów SC i poziomem homologii sekwencji powinno być wykonane niezależnie dla każdego analizowanego zestawu danych. Wynika to ze specyficznych cech każdego zestawu danych uzyskiwanego w ramach CSH. Stąd też proces ustalania warunków filtracji morfologicznej dla zestawu NT-CSH obejmował analizę zależności poziomu homologii sekwencji względem wartości wszystkich 10 parametrów SC (Tabela 17).

Tabela 17. *Opis parametrów charakterystyki punktów (SC) stosowanych na etapie analizy ilościowej obrazu za pomocą programu MAIA.*

Skrócona nazwa parametru	Pełna nazwa parametru	Opis	Zakres wartości
Sig	Sygnał (ang. *signal*)	Wartość zdefiniowana jako S=min(S_{Cy5}-B_{Cy5}, S_{Cy3}-B_{Cy3}), gdzie: -S_{Cy5}(S_{Cy3}) jest średnią szacunkową intensywności w obrębie konturu danego punktu dla kanału Cy5(Cy3), -B_{Cy5}(B_{Cy3}) jest średnią szacunkową tła dla kanału Cy5(Cy3).	Zależy od mikromacierzy
CVR	Współczynnik zmienności dwóch stosunków intensywności (ang. *coefficient of variation of two ratio estimates*)	Wyliczane są dwie wartości określające stosunek intensywności kanału czerwonego do zielonego (Cy5/Cy3). Jedna wartość wyliczana jest za pomocą regresji linowej (RR), druga zaś za pomocą algorytmu segmentacji (RS, poprzez podział punktu na części). CVR określa zmienność pomiędzy wynikami otrzymanymi dla tych dwóch metod.	Średnio: 0-1 Wysokie wartości CVR określają punkty o niskiej jakości
Parametry szacowane na podstawie regresji liniowej			
Det	Współczynnik determinacji (ang. *coefficient of determination*)	Oznacza stopień liniowej zależności pomiędzy intensywnościami sygnału fluorescencyjnego dla kanału Cy3 i Cy5 (korelacja).	0-1 Wysokie wartości Det określają punkty o wysokiej jakości
DWS	Statystyka Durbina-Watsona (ang. *Durbin-Watson statistic*)	Kontroluje obecność autokorelacji pierwszego rzędu w pozostałości (resztach) dopasowania regresji liniowej.	0-4 Wartości DWS bliskie 2 określają punkty o wysokiej jakości
SPC	Zanieczyszczenie punktu (ang. *spot contamination*)	Oznacza liczbę nieprawidłowych pikseli w obrębie konturu punktu wyznaczonych przez procedurę filtracji zanieczyszczeń.	Wysokie wartości SC określają punkty o niskiej jakości
Parametry określane na podstawie konturu punktu			
Dia	Średnica punktu (ang. *spot diameter*)	Średnica punktu liczona na podstawie wzoru: $D=2(S/\pi)^{1/2}$, gdzie S oznacza liczbę pikseli w obrębie punktu.	Niskie wartości Dia określają punkty o niskiej jakości
GSym	Symetria geometryczna (ang. *geometrical symmetry*)	Mierzy odchylenie konturu punktu od idealnego okręgu.	Wysokie wartości GSym określają punkty o niskiej jakości
ISym	Symetria intensywności sygnału (ang. *intensity symmetry*)	Pozwala na określenie wpływu zanieczyszczeń (np. kurz) na całkowitą intensywność punktu.	Bardzo wysokie, jak i bardzo niskie wartości ISym określają punkty o niskiej jakości
Parametry określające jakość tła			
UB	Jednorodność tła w obrębie punktu (ang. *uniformity of background*)	Określa jednorodność tła w obrębie punktu na podstawie algorytmu segmentacji. Wysokie wartości UB mogą świadczyć o obecności zanieczyszczeń o stosunkowo wysokim poziomie intensywności, dużej zmienność na poziomie tła oraz o połączeniu się sąsiadujących punktów.	Wysokie wartości UB określają punkty o niskiej jakości
AB	Bezwzględny poziom tła (ang. *absolute level of background*)	Porównanie tła obliczonego dla danego punktu z typowym poziomem tła (średnia wartość tła obliczona z uwzględnieniem wszystkich punktów na mikromacierzy).	Wysokie wartości AB określają punkty o niskiej jakości

V.IV.3 Określenie stopnia homologii sekwencji za pomocą analizy BLAST

Pierwszym etapem analizy było określenie poziomu homologii sekwencji nukleotydowych pomiędzy gatunkami referencyjnymi z zestawu NT-CSH (ziemniak i pomidor), a gatunkiem badanym (tytoń). W tym celu przeprowadzona została analiza z

wykorzystaniem algorymu BLAST (ang. *Basic Local Alignment Search Tool*) dla sekwencji nukleotydowych (blastn, ang. *nucleotide blast*). Analiza blastn obejmowała porównanie sekwencji *UniGene*, którym odpowiadały sekwencje sond cDNA ulokowanych na mikromacierzy DNA dla ziemniaka lub pomidora, względem sekwencji genomowej tytoniu. Wynikiem analizy blastn były wartości *bit score*, określające poziom homologii sekwencji. Sekwencje wykazujące wyższe podobieństwo otrzymują wyższe wartości *bit score*. Charakterystyka otrzymanych wartości *bit score* dla ziemniaka i pomidora względem tytoniu przedstawiona została w Tabeli 18.

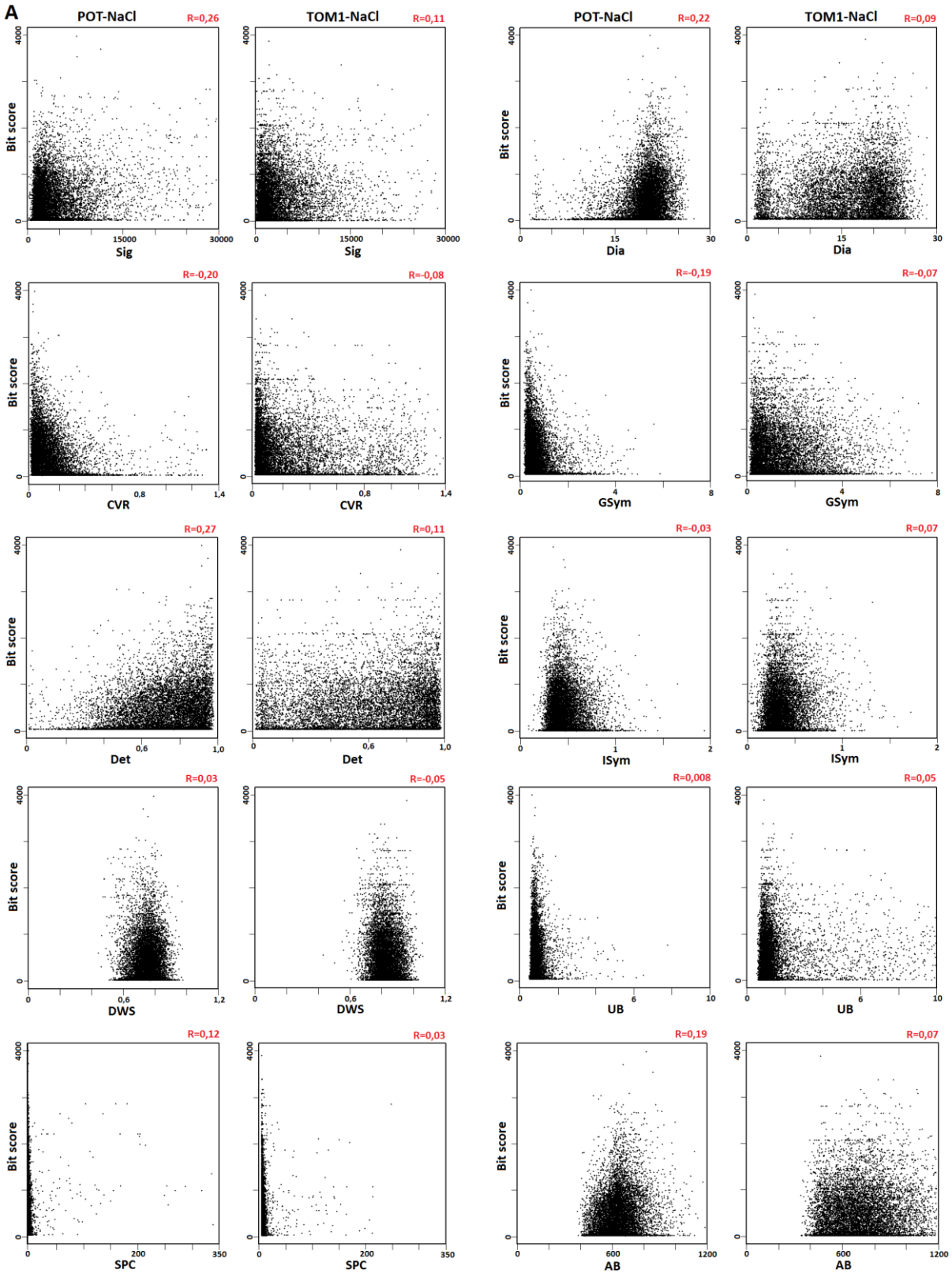
Tabela 18. *Wyniki analizy homologii sekwencji pomiędzy gatunkami referencyjnymi: ziemniak i pomidor oraz gatunkiem badanym: tytoń za pomocą programu BLAST.*

	Ziemniak	**Pomidor**
Liczba sond	26744	8951
Średnia wartość *bit score*	464,6	559,6
Mediana *bit score*	331	444
Maksymalna wartość *bit score*	3994	5551
Minimalna wartość *bit score*	30	32

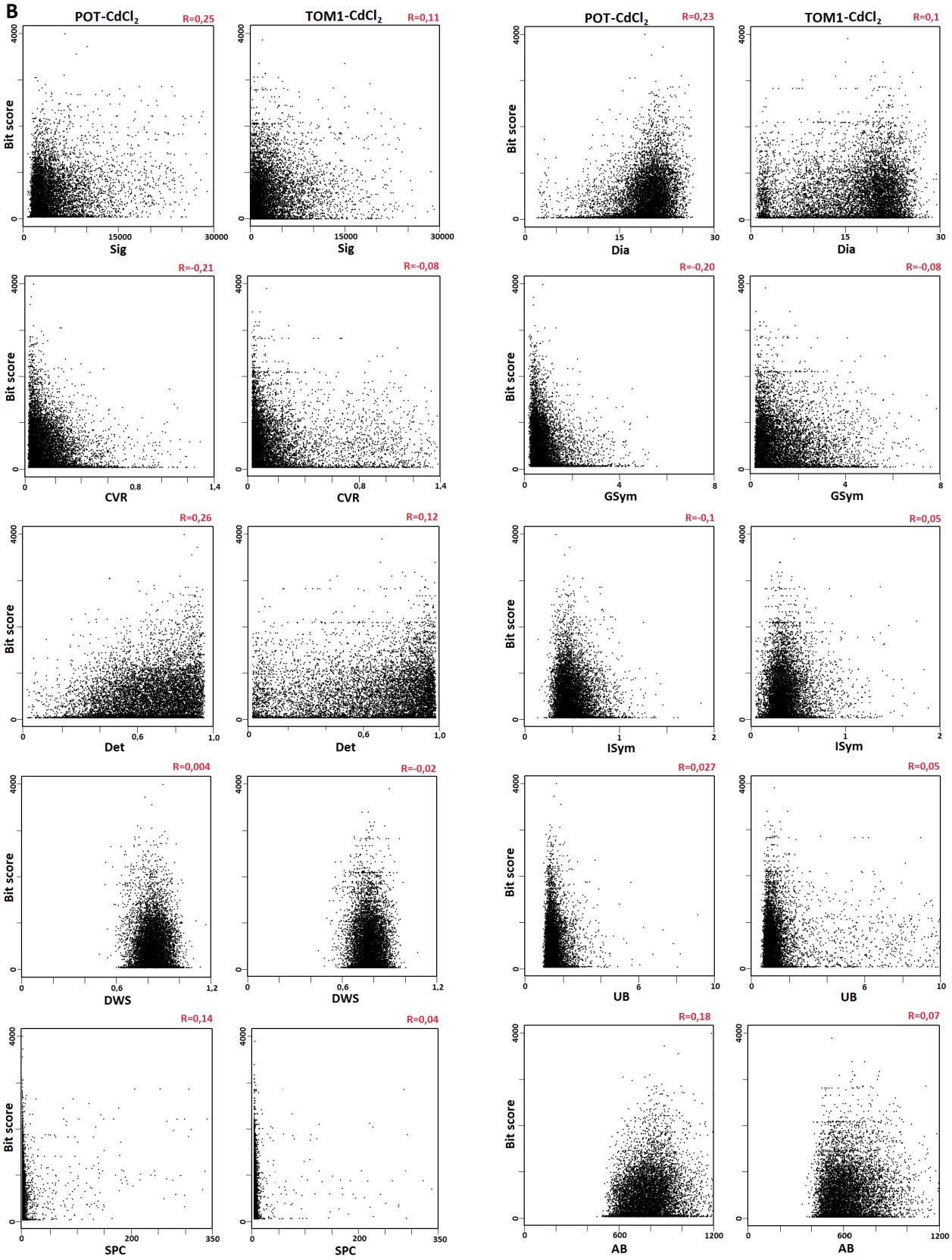
Wyniki prezentowane w Tabeli 18 wskazują, iż oba gatunki referencyjne (ziemniak i pomidor) wykazują zbliżony poziom homologii względem gatunku badanego (tytoń).

V.IV.4 Zależność wartości parametrów jakości punktów od homologii sekwencji

W kolejnym etapie sprawdzenia możliwości stosowania filtracji morfologicznej, jako części analizy niższego rzędu dla zestawu NT-CSH, określona została zależność pomiędzy wartościami parametrów SC oraz poziomem homologii sekwencji sond w stosunku do sekwencji docelowych. Celem tego etapu analizy była identyfikacja parametrów BS-SC (ang. *bit score correlated spot characteristics*), czyli parametrów SC, których wartości zależą od zmiany poziomu dopasowania sekwencji. W celu ustalenia parametrów BS-SC wykonano wykresy zależności dla każdego z 10 parametrów SC względem wartości *bit score* dla podzestawów POT i TOM1 z podziałem na rodzaj stosowanego czynnika stresogennego (NaCl lub $CdCl_2$). Wykresy zależności wykonane zostały z użyciem średnich wartości parametrów SC otrzymanych dla zestawu mikromacierzy, odpowiadających danemu rodzajowi platformy (POT lub TOM1) oraz danemu rodzajowi czynnika stresogennego (NaCl lub $CdCl_2$). Wyniki przedstawiono na Rysunku 34:



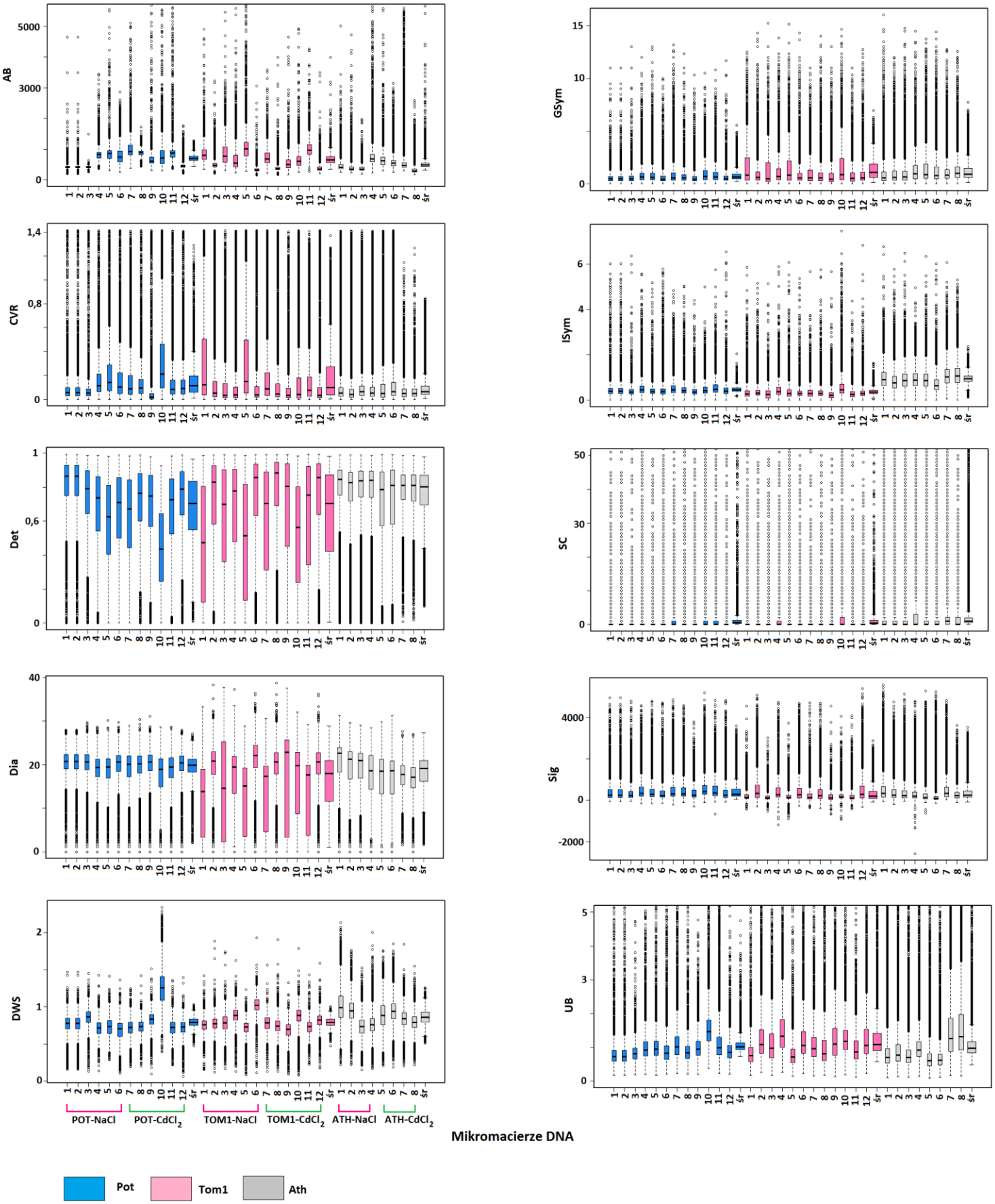
Rysunek 33. *Wykres zależności dla 10 parametrów SC względem wartości bit score. Wykresy wykonano dla podzestawów POT oraz TOM1 w zależności od stosowanego czynnika stresogennego NaCl (panel A) oraz $CdCl_2$ (panel B). Wykresy wykonane zostały dla średnich wartości parametrów SC dla każdego podzbioru mikromacierzy. Oznaczenia poszczególnych SC pokrywają się z tymi prezentowanymi w Tabeli 17. Wartości współczynnika korelacji (R) pomiędzy danym parametrem SC, a wartościami bit score prezentowane są w prawym górnym rogu każdego wykresu (kolor czerwony).*



(Kontynuacja Rysunku 34, opis znajduje się na poprzedniej stronie.)

Przeprowadzona analiza wskazuje na brak zależności pomiędzy wartościami *bit score*, a wartościami wszystkich 10 parametrów SC, w tym także parametrów: CVR, Det, Dia, GSym, uznanych za znaczące przez Bar-Or i wsp. (Rysunek 34A oraz 34B). Ponadto, wygląd wykresów zależności (Rysunek 34A oraz 34B) dla poszczególnych parametrów jest bardzo zbliżony, porównując pomiędzy rodzajami użytej mikromacierzy DNA (POT i TOM1) oraz stosowanymi czynnikami stresogennymi (NaCl lub $CdCl_2$). Fakt ten wskazuje, iż poziom dopasowania sond może nie mieć związku z morfologią punktów w obrębie analizowanych mikromacierzy. Stąd też wybór parametrów BS-SC dla uzyskanych danych eksperymentalnych na podstawie wykresów zależności (Rysunek 34A oraz 34B) nie był możliwy.

Inną próbą znalezienia związku pomiędzy parametrami morfologii punktów, a stopniem dopasowania sekwencji sondy i sekwencji docelowych była analiza rozkładu 10 parametrów SC dla podzestawów POT i TOM1 względem referencyjnego podzestawu ATH. W przypadku obecności zależności pomiędzy homologią sekwencji, a morfologią punktów, wartości parametrów SC uzyskanych dla podzestawów POT i TOM1 (otrzymanych w wyniku hybrydyzacji CSH) powinny być znacząco różne od wartości tych parametrów uzyskanych dla referencyjnego podzestawu ATH (otrzymanego w wyniku hybrydyzacji homologicznej (SSH ang. *single-species hybridization*)). Wyniki analizy przedstawione zostały w postaci wykresów pudełkowych na Rysunku 35.



Rysunek 34. *Wykresy pudełkowe dla wartości 10 parametrów SC: AB, CVR, Det, Dia, DWS, GSym, ISym, SC, Sig, UB. Wykresy wykonano dla podzestawów: POT (kolor niebieski), TOM1 (kolor różowy) oraz referencyjnego ATH (kolor szary).*

Otrzymane wyniki wskazują, iż wartości niemalże wszystkich 10 parametrów SC nie są zależnie od stopnia komplementarności sekwencji docelowych względem sond. Ponadto, wartości badanych parametrów SC wykazują duży rozrzut pomiędzy poszczególnymi mikromacierzami DNA, pochodzącymi nawet z tego samego podzestawu, np. wartości parametru Dia dla podzestawu TOM1. Podobny efekt widoczny jest dla parametrów CVR oraz Det. Jedynym z przedstawionych na wykresach pudełkowych parametrów SC, który wykazuje nieznaczne różnice w wartościach dla hybrydyzacji homologicznej (SSH) oraz hybrydyzacji heterologicznej (CSH) jest parametr ISym. Wartości ISym dla hybrydyzacji SSH są wyższe niż dla CSH, co oznacza iż jednorodność sygnałów fluorescencji, uzyskanych w wyniku hybrydyzacji homologicznej jest nieco większa.

Wyniki przeprowadzonych analiz wskazują na brak związku pomiędzy poziomem dopasowania sekwencji, a morfologią punktów i wykluczają możliwość stosowania filtracji danych na podstawie parametrów charakterystyki punktów (SC) na mikromacierzy w ramach analizy zestawu danych NT-CSH. Brak zależności pomiędzy dopasowaniem sekwencji, a parametrami morfologii punktów uniemożliwiał wybór parametrów BS-SC, które są podstawą do przeprowadzenia tej filtracji. Identyfikacja parametrów BS-SC nie była możliwa także ze względu na duży rozrzut wartości parametrów SC dla poszczególnych mikromacierzy DNA w ramach tego samego podzestawu w tym także i referencyjnego. W przypadku zestawu NT-CSH identyfikacja i eliminacja sond o obniżonej homologii względem sekwencji docelowych przeprowadzona była za pomocą filtracji homologicznej (filtracja oparta na homologii sekwencji).

Wykresy prezentowane na Rysunku 34A i 34B oraz Rysunku 35 otrzymano z użyciem programu R/Bioconductor oraz pakietu `limma`.

V.IV.5 Omówienie wyników

Większość komercyjnych mikromacierzy DNA do badania ekspresji genów dostępna jest jedynie dla niewielkiej liczby gatunków, w tym gatunków modelowych. Pozostała część gatunków zwierzęcych i roślinnych nie posiada dedykowanych im ekspresyjnych mikromacierzy DNA. Stąd też wysokoprzepustowa analiza ekspresji genów w obrębie całego transkryptomu dla tych gatunków jest mocno ograniczona. Potencjalnym rozwiązaniem tej kwestii jest stosowanie hybrydyzacji międzygatunkowej (CSH). Hybrydyzacja CSH jest aktualnie powszechnie wykorzystywanym podejściem do analizy ekspresji genów. Znajduje

ona zastosowanie głównie w badaniach porównawczych, ekologicznych i ewolucyjnych (Chalmers i wsp. 2005; Brodsky i wsp. 2005; Mitchell-Olds i wsp. 2003). Najnowsze doniesienia sugerują także wykorzystanie CSH w badaniach biomedycznych (E. S. Park i wsp. 2011).

Mikromacierze DNA zostały zaprojektowane z myślą o hybrydyzacji homologicznej (SSH), stąd też hybrydyzacja międzygatunkowa (CSH) jest formą niestandardowego wykorzystania mikromacierzy DNA. Otrzymanie znaczących biologicznie wyników z danych CSH wymaga przekształcenia tych danych w dane o cechach SSH. Przekształcenie to polega na wyeliminowaniu z zestawu, danych uzyskanych z użyciem sond o niskiej homologii względem sekwencji docelowych. Najpowszechniejszą formą tego przekształcenia jest filtracja homologiczna, bazująca na dostępnej informacji o sekwencji genomowej dla badanych i referencyjnych gatunków. W przypadku, gatunków dla których informacja ta nie jest dostępna, rozwiązaniem może być filtracja na podstawie parametrów morfologii punktów. Ten rodzaj filtracji pozwala na eliminację z zestawu danych informacji uzyskanej za pomocą sond o obniżonej komplementarności w oparciu o wartości wybranych parametrów charakterystyki punktów (SC). Metoda filtracji morfologicznej została zaprezentowana przez Bar-Or i wsp. (2007) i bazuje na 4 parametrach BS-SC (ang. *bit score correlated spot characteristics*), których wartości wykazują zależność względem poziomu homologii sekwencji. Opisywane przez Bar-Or i wsp. wartości parametrów BS-SC (CVR, Det, Dia oraz GSym) otrzymywane zostały w wyniku analizy ilościowej obrazu za pomocą programu MAIA.

Zgodnie z teorią zaproponowaną przez Bar-Or, związek pomiędzy morfologią punktów na mikromacierzy DNA, a stopniem dopasowania sekwencji docelowych do sekwencji sond wynika z faktu, iż dany punkt na ekspresyjnej mikromacierzy DNA składa się z wielu cząsteczek sond. Podczas CSH do sond przyłączają się zarówno transkrypty o pełnej, jak i niepełnej komplementarności. Na etapie płukania mikromacierzy, które następuje po reakcji hybrydyzacji, sekwencje wykazujące niską homologię dysocjują z utworzonego wraz z sondą dupleksu, pozostawiając wolne sondy. Proces skanowania wyraża sumę transkryptów związanych przez sondy wchodzące w skład pojedynczego punktu w pikselach. Stąd też obecność wolnych sond w zbiorze sond dla danego punktu skutkuje obniżeniem intensywności oraz jednorodności sygnału. Z racji faktu, iż dla CSH liczba sond niezwiązanych z żadnym transkryptem jest większa niż SSH, fragment punktu może być

nawet fałszywie zaklasyfikowany jako tło, co w znacznym stopniu może zaburzać morfologię punktu.

Celem tej części pracy doktorskiej była ocena możliwości zastosowania filtracji morfologicznej, jako jednego z etapów analizy niższego rzędu dla zestawu NT-CSH. W ramach realizacji tego celu wykonano dwie niezależne analizy pozwalające na określenie związku pomiędzy poziomem dopasowania sekwencji, a morfologią punktów: analizę zależności pomiędzy wartościami 10 parametrów SC, a wartościami *bit score* odzwierciedlającymi poziom homologii sekwencji (Rysunek 34A i 34B) oraz analizę rozkładu wartości parametrów SC dla danych uzyskanych w wyniku hybrydyzacji CSH (POT, TOM1) względem parametrów SC dla danych otrzymanych w wyniku hybrydyzacji SSH (ATH) (Rysunek 35). Analiza rozkładu wartości parametrów SC prowadzona była w oparciu o podzestaw referencyjny ATH, otrzymany w wyniku hybrydyzacji próbek uzyskiwanych z roślin *A.thaliana*, traktowanych odpowiednio NaCl lub $CdCl_2$, do ekspresyjnej mikromacierzy DNA dla *A.thaliana*. Zastosowanie zestawu ATH pozwoliło na określenie wartości parametrów SC dla punktów powstałych na skutek hybrydyzacji sekwencji docelowych o pełnej homologii względem sekwencji sond. Najbardziej optymalną kontrolą analizy rozkładu wartości parametrów SC byłby zestaw referencyjny otrzymany z użyciem mikromacierzy dla tytoniu. W trakcie prowadzonych badań dostępne były dwa rodzaje mikromacierzy DNA do badania ekspresji genów tytoniu szlachetnego. Jedna opracowana we współpracy z firmą Affymetrix (Edwards i wsp. 2010), a druga zaprojektowana przez firmę Agilent (Tobacco Gene Expression Microarray 4x44K, *Agilent*). Jednakże, zastosowanie każdej z tych dwóch mikromacierzy DNA do stworzenia zestawu referencyjnego w ramach podzestawu NT-CSH nie było możliwe ze względu na projekt eksperymentu (jednokolorowy, dla mikromacierzy z opracowanej udziałem firmy Affymetrix) lub projekt mikromacierzy (układ sond) uniemożliwiający przeprowadzenie analizy ilościowej obrazu za pomocą programu MAIA (Tobacco Gene Expression Microarray, Agielnt).

Wyniki obu przeprowadzonych analiz wskazują na brak związku pomiędzy dopasowaniem sekwencji, a morfologią punktów dla danych z zestawu NT-CSH. W przypadku analizy związku pomiędzy wartościami parametrów SC, a wartościami *bit score* otrzymane wyniki są powtarzalne i niezależne od rodzaju użytej mikromacierzy DNA (POT lub TOM1), badanego gatunku (ziemniak lub pomidor) oraz czynnika stresogennego (NaCl lub $CdCl_2$). Fakt ten może wskazywać, iż ewentualny wpływ zmiany poziomu dopasowania sekwencji na morfologię punktów jest bardzo subtelny i mógł zostać zniwelowany na skutek

różnic technicznych pomiędzy dwoma użytymi rodzajami mikromacierzy DNA oraz różnic poziomów ekspresji genów dla dwóch czynników stresogennych. Rezultaty analizy rozkładu natomiast, mają mniej powtarzalny charakter, który wskazuje na to, iż techniczny aspekt eksperymentu m.in. proces nanoszenia sond na podłoże, rodzaj podłoża, czy wydajność procesu immobilizacji, które mogą być specyficzne dla każdej mikromacierzy DNA (nawet tego samego rodzaju), ma znacznie większy wpływ na rozkład wartości parametrów SC, niż poziom dopasowania sekwencji docelowych do sekwencji sond.

Rozbieżność rezultatów otrzymanych dla zestawu NT-CSH i tych prezentowanych przez Bar-Or może wynikać ze znacznych różnic pomiędzy projektami analizowanych eksperymentów. Proponowany przez Bar-Or proces filtracji danych na podstawie parametrów morfologii punktów, zaprojektowany został na podstawie wyników modelowego eksperymentu uwzględniającego użycie tylko jednego rodzaju mikromacierzy DNA. Eksperyment obejmował badanie ekspresji genów 7 gatunków roślin (ziemniak, pomidor, bakłażan, pieprz, petunia oraz 2 gatunki tytoniu: *Nicotiana tabaccum* oraz *Nicotiana benthamiana*) za pomocą mikromacierzy DNA dla ziemniaka (mikromacierz cDNA, The Institute for Genomic Research). Natomiast zestaw NT-CSH zaprojektowany został z użyciem trzech różnych platform, z czego dwie (POT i TOM1) wykorzystano do badań międzygatunkowych w wyniku hybrydyzacji próbek tytoniu, a jedną (ATH) jako kontrolę eksperymentu. Zestaw NT-CSH zaprojektowany został w sposób umożliwiający wykorzystanie mikromacierzy DNA dla ziemniaka, stosowanej także przez Bar-Or. Podejście to miało na celu bezpośrednie porównanie wartości parametrów SC uzyskanych dla tego rodzaju mikromacierzy DNA z wartościami parametrów SC dla eksperymentu Bar-Or oraz innych platform (TOM1 i ATH). Rezultaty otrzymane dla zestawu NT-CSH, wskazujące na brak związku pomiędzy poziomem dopasowania sekwencji, a morfologią punktów mogą także wynikać z niewielkich różnic homologii pomiędzy pomidorem i ziemniakiem, a tytoniem. W przypadku analizy prezentowanej przez Bar-Or porównywano wyniki dla gatunku badanego wykazującego najwyższą homologię (pomidor), względem gatunku badanego o najniższej homologii (petunia) do gatunku referencyjnego (ziemniak).

Brak związku pomiędzy homologią sekwencji docelowych, a sekwencjami sond, zasadniczo wykluczał możliwość stosowania filtracji danych na podstawie parametrów morfologii punktów w przypadku zestawu NT-CSH. Duży rozrzut wartości parametrów SC widoczny na Rysunku 35 wskazuje, iż metoda filtracji morfologicznej zaproponowana przez Bar-Or i wsp. znajduje raczej zastosowanie w przypadku analizy wyników eksperymentów

modelowych, niż rzeczywistych lub gdy występują znaczne różnice w homologii pomiędzy badanymi gatunkami, a gatunkiem referencyjnym. Jednakże ten warunek jest sprzeczny z założeniami analizy ekspresji genów z wykorzystaniem hybrydyzacji międzygatunkowej w której gatunek badany i referencyjny są możliwie najbardziej spokrewnione.

V.IV.6 Wnioski

Na podstawie otrzymanych wyników możliwe było sformułowanie następujących wniosków:

- Brak zależności pomiędzy morfologią punktów, a homologią sekwencji w analizowanych zestawach danych.
- Rozbieżności pomiędzy uzyskanymi wynikami, a tymi prezntowanymi przez zespół Bar-Or mogą wynikać ze znaczących różnic w projektach eksperymentów. Eksperyment Bar-Or miał modelowy charakter i obejmował analizę gatunków, które wykazywały bardzo wysoki lub znacznie niższy poziom homologii względem gatunku referencyjnego.
- Brak zależności pomiędzy morfologią punktów, a homologią sekwencji może wynikać z różnic technicznych pomiędzy dwoma rodzajami użytych mikromacierzy DNA oraz niewielkich różnic w ekspresji genów dla dwóch czynników stresogennych. Ponadto, brak związku pomiędzy poziomem dopasowania sekwencji, a morfologią punktów może także wynikać z niewielkich różnic homologii pomiędzy gatunkiem badanym (tytoń), a gatunkami referencyjnymi (pomidor i ziemniak).

VI. Wnioski

Przeprowadzone badania pozwoliły na optymalizację ścieżek analizy dla czterech najbardziej powszechnych rodzajów niestandardowych danych. Na podstawie otrzymanych wyników możliwe było wyciągnięcie następujących wniosków:

1. Obecność standardów jakości wynika ze świadomości ograniczeń jakim podlegają dane uzyskiwane z użyciem ekspresyjnych mikromacierzy DNA. Mając na uwadze te ograniczenia możliwe jest uzyskanie z niestandardowych danych, informacji o znaczeniu biologicznym.
2. Analiza niestandardowych danych polega głównie na uwzględnieniu specyficznego etapu, którego celem jest eliminacja czynników nadających danym niestandardowy charakter. Etap ten powinien być realizowany na poziomie analizy niższego rzędu.
3. Rezultaty uzyskane w wyniku analizy niestandardowych danych powinny być interpretowane z zachowaniem szczególnej ostrożności oraz w miarę możliwości zweryfikowane przy zastosowaniu innych technik badania ekspresji genów, np. ilościowy PCR, czy sekwencjonowanie drugiej generacji.

VII. Literatura

Agarwal, A. i wsp., 2010. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC genomics*, 11, pp.383–399.

Argyropoulos, C. i wsp., 2006. Operational criteria for selecting a cDNA microarray data normalization algorithm. *Oncology reports*, 15, pp.983–96.

Bagnaresi, P. i wsp., 2008. Heterologous microarray experiments allow the identification of the early events associated with potato tuber cold sweetening. *BMC genomics*, 9, pp.176–199.

Baker, S.C. i wsp., 2005. The External RNA Controls Consortium: a progress report. *Nature methods*, 2(10), pp.731–734.

Baron, D. i wsp., 2011. Immune response and mitochondrial metabolism are commonly deregulated in DMD and aging skeletal muscle. *PloS one*, 6(11), p.e26952.

Bar-Or, C. i wsp., 2006. Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results. *BMC genomics*, 7, pp.110–123.

Bar-Or, C., Novikov, E., i wsp., 2007. Utilizing microarray spot characteristics to improve cross-species hybridization results. *Genomics*, 90(5), pp.636–645.

Bar-Or, C., Czosnek, H. & Koltai, H., 2007. Cross-species microarray hybridizations: a developing tool for studying species diversity. *Trends in genetics : TIG*, 23(4), pp.200–207.

Baskerville, S. & Bartel, D.P., 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA (New York, N.Y.)*, 11(3), pp.241–247.

Blalock, E.M., 2003. *A Beginner's Guide to Microarrays*,

Bloom, J.S. i wsp., 2009. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC genomics*, 10, pp.221–231.

Bolstad, B.M. i wsp., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, 19(2), pp.185–193.

Brachat, a i wsp., 2000. Comparative microarray analysis of gene expression during apoptosis-induction by growth factor deprivation or protein kinase C inhibition. *Oncogene*, 19(44), pp.5073–5082.

Brachat, A. i wsp., 2002. A microarray-based, integrated approach to identify novel regulators of cancer drug response and apoptosis. *Oncogene*, 21(54), pp.8361–8371.

Brazma, A. i wsp., 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics*, 29(4), pp.365–371.

Brodsky, L.I. i wsp., 2005. Evolutionary regulation of the blind subterranean mole rat, Spalax, revealed by genome-wide gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 102(47), pp.17047–17052.

Campanaro, S. i wsp., 2002. Gene expression profiling in dysferlinopathies using a dedicated muscle microarray. *Human molecular genetics*, 11(26), pp.3283–3298.

Castoldi, M. i wsp., 2006. A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA). *RNA (New York, N.Y.)*, 12(5), pp.913–920.

Castoldi, M. i wsp., 2007. miChip: a microarray platform for expression profiling of microRNAs based on locked nucleic acid (LNA) oligonucleotide capture probes. *Methods (San Diego, Calif.)*, 43(2), pp.146–152.

Cernetich-Ott, A. i wsp., 2012. Remarkable stability in patterns of blood-stage gene expression during episodes of non-lethal Plasmodium yoelii malaria. *Malaria journal*, 11(1), pp.265–288.

Chalmers, A.D. i wsp., 2005. A Xenopus tropicalis oligonucleotide microarray works across species using RNA from Xenopus laevis. *Mechanisms of development*, 122(3), pp.355–363.

Coppola, G., 2011. Designing, performing, and interpreting a microarray-based gene expression study. In G. Manfredi & H. Kawamata, eds. *Methods in Molecular Biology*. Humana Press, pp. 417–439.

Dabney, A.R. & Storey, J.D., 2007. A new approach to intensity-dependent normalization of two-channel microarrays. *Biostatistics Oxford England*, 8(1), pp.128–139.

Edwards, K.D. i wsp., 2010. TobEA: an atlas of tobacco gene expression from seed to senescence. *BMC genomics*, 11, pp.142–147.

Ferrarini, A. i wsp., 2008. Expression of Medicago truncatula genes responsive to nitric oxide in pathogenic and symbiotic conditions. *Molecular plant-microbe interactions : MPMI*, 21(6), pp.781–790.

Fraley, C. & Raftery, A.E., 2009. MCLUST Version 3 for R : Normal Mixture Modeling and Model-Based Clustering. *Technical Report*, pp.1–54.

Futschik, M.E., 2011. Introduction to OLIN package. *Technical Report*, pp.1–25.

Gentleman, R.C. i wsp., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), p.R80.

Goff, L.A. i wsp., 2005. Rational Probe Optimization and Enhanced Detection Strategy for MicroRNAs Using Microarrays. *RNA Biology*, 2(3), pp.93–100.

Golub, T. R., 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439), pp.531–537.

Graveley, B.R. i wsp., 2011. The developmental transcriptome of Drosophila melanogaster. *Nature*, 471(7339), pp.473–479.

Hacia, J.G. & Collins, F.S., 1999. Mutational analysis using oligonucleotide microarrays. *Journal of medical genetics*, 36(10), pp.730–736.

Hammond, J.P. i wsp., 2005. Using genomic DNA-based probe-selection to improve the sensitivity of high-density oligonucleotide arrays when applied to heterologous species. *Plant methods*, 1, pp.10–19.

Held, M., Gase, K. & Baldwin, I.T., 2004. Microarrays in ecological research: a case study of a cDNA microarray for plant-herbivore interactions. *BMC ecology*, 4, pp.13–24.

Hu, G. i wsp., 2009. MTDH activation by 8q22 genomic gain promotes chemoresistance and metastasis of poor-prognosis breast cancer. *Cancer cell*, 15(1), pp.9–20.

Huber, W. i wsp., 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics (Oxford, England)*, 18(1997), pp.S96–S104.

Irizarry, R.A. i wsp., 2005. Multiple-laboratory comparison of microarray platforms. *Nature methods*, 2(5), pp.345–350.

Iyer, V.R. i wsp., 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409(6819), pp.533–538.

Jolly, R.A. i wsp., 2005. Pooling samples within microarray studies: a comparative analysis of rat liver transcription response to prototypical toxicants. *Physiological Genomics*, 22(3), pp.346–355.

Kainkaryam, R.M. i wsp., 2010. poolMC: smart pooling of mRNA samples in microarray experiments. *BMC bioinformatics*, 11, pp.299–308.

Kaposi-Novak, P. i wsp., 2004. Oligonucleotide microarray analysis of aminoallyl-labeled cDNA targets from linear RNA amplification. *BioTechniques*, 37(4), pp.580–588.

Khaitovich, P. i wsp., 2004. A neutral model of transcriptome evolution. *PLoS biology*, 2(5), pp.0682–0689.

Kooperberg, C. i wsp., 2002. Improved background correction for spotted DNA microarrays. *Journal of computational biology a journal of computational molecular cell biology*, 9(1), pp.55–66.

Lagorce, A. i wsp., 2012. Genome-Wide Transcriptional Response of the Archaeon Thermococcus gammatolerans to Cadmium. *PloS one*, 7(7), p.e41935.

Li, W. & Ruan, K., 2009. MicroRNA detection by microarray. *Analytical and bioanalytical chemistry*, 394(4), pp.1117–1124.

Liebert, M.A. i wsp., 2002. Spotted DNA Microarrays. *Journal of Computational Biology*, 9(1), pp.55–66.

Liu, C.-G., Spizzo, R., i wsp., 2008. Expression profiling of microRNA using oligo DNA arrays. *Methods (San Diego, Calif.)*, 44(1), pp.22–30.

Liu, C.-G., Calin, G.A., i wsp., 2008. MicroRNA expression profiling using microarrays. *Nature protocols*, 3(4), pp.563–578.

Liu, S. i wsp., 2011. A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. *Nucleic acids research*, 39(2), pp.578–588.

Lu, T. i wsp., 2005. Can Zipf's law be adapted to normalize microarrays? *BMC bioinformatics*, 6, pp.37–50.

Malone, J.H. & Oliver, B., 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9, pp.34–43.

Mcilroy, D. i wsp., 2005. Profiling dendritic cell maturation with dedicated microarrays. *Journal of leukocyte biology*, 78(3), pp.794–803.

Mehta, J.P. & Rani, S., 2011. Software and Tools for Microarray Data Analysis. In L. O'Driscoll, ed. *Gene Expression Profiling: Methods and Protocols*. Totowa, NJ: Humana Press, pp. 41–53.

Mitchell-Olds, T., Feder, M. & Wray, G., 2003. Evolutionary and ecological functional genomics. *Nature Reviews Genetics*, 4, pp.649–655.

Novikov, E. & Barillot, E., 2005. An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments. *BMC bioinformatics*, 6, pp.293–311.

Novikov, E. & Barillot, E., 2007. Software package for automatic microarray image analysis (MAIA). *Bioinformatics (Oxford, England)*, 23(5), pp.639–640.

Oshlack, A. i wsp., 2007. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome biology*, 8, p.R2.

Paquet, A. & Yang, Jean Yee Hwa, 2010. arrayQuality: Assessing array quality on spotted arrays. R package version 1.32.0.

Park, E.S. i wsp., 2011. Cross-species hybridization of microarrays for studying tumor transcriptome of brain metastasis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(42), pp.17456–17461.

Patterson, T. a i wsp., 2006. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nature biotechnology*, 24(9), pp.1140–1150.

Pelz, C.R. i wsp., 2008. Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data. *BMC bioinformatics*, 9, pp.520–538.

Petersen, D. i wsp., 2005. Three microarray platforms: an analysis of their concordance in profiling gene expression. *BMC genomics*, 6, pp.63–71.

Pinkel, D. i wsp., 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, 20(2), pp.207–211.

R Development Core Team, R.F.F.S.C., 2008. R: A Language and Environment for Statistical Computing R Development Core Team, ed. *R Foundation for Statistical Computing*, 1(10), p.2673.

Renn, S.C.P., Aubin-Horth, N. & Hofmann, H., 2004. Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. *BMC genomics*, 5, pp.42–55.

Rensink, W. a & Hazen, S.P., 2006. Statistical issues in microarray data analysis. *Methods in molecular biology (Clifton, N.J.)*, 323, pp.359–366.

Rinaldi, S. i wsp., 2009. Analysis of lectin binding to glycolipid complexes using combinatorial glycoarrays. *Glycobiology*, 19(7), pp.789–796.

Ritchie, M.E. i wsp., 2007. A comparison of background correction methods for two-colour microarrays. *Bioinformatics (Oxford, England)*, 23(20), pp.2700–2707.

Roush, S. & Slack, F.J., 2008. The let-7 family of microRNAs. *Trends in cell biology*, 18(10), pp.505–516.

Saeed, A.I. i wsp., 2006. TM4 microarray software suite. *Methods in Enzymology*, 411, pp.134–193.

Saeed, A.I. i wsp., 2003. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2), pp.374–378.

Scharpf, R.B. i wsp., 2007. When should one subtract background fluorescence in 2-color microarrays? *Biostatistics (Oxford, England)*, 8(4), pp.695–707.

Schena, M., 2003. *Protein Microarrays*, Jones & Bartlett Publishing.

Schenk, P.M. i wsp., 2008. Identification of plant defence genes in canola using Arabidopsis cDNA microarrays. *Plant biology Stuttgart Germany*, 10(5), pp.539–547.

Sewer, A. i wsp., 2005. Identification of clustered microRNAs using an ab initio prediction method. *BMC bioinformatics*, 6, pp.267–282.

Shi, L. i wsp., 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9), pp.1151–1161.

Shi, L. i wsp., 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, 28(8), pp.827–838.

Simon, R.M., 2004. In Design and Analysis of DNA Microarray Investigations. *Springer-Verlag, New York*, 2nd edn, 2, pp.37–40.

Singh, D. i wsp., 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2), pp.203–209.

Skreka, K. i wsp., 2012. Expression Profiling of a Heterogeneous Population of ncRNAs Employing a Mixed DNA/LNA Microarray. *Journal of nucleic acids*, 2012, pp.283560–283570.

Smyth, G.K., 2005. Limma : Linear Models for Microarray Data. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, (2005), pp.397–420.

Smyth, G.K., Ritchie, M. & Thorne, N., 2010. Linear Models for Microarray Data User 's Guide. *Technical Report*, pp.1–99.

Smyth, G.K. & Speed, T., 2003. Normalization of cDNA microarray data. *Methods San Diego Calif*, 31(4), pp.265–273.

Smyth, G.K., Yang, Y.H. & Speed, T., 2003. Statistical issues in cDNA microarray data analysis. *Methods in molecular biology (Clifton, N.J.)*, 224, pp.111–136.

Suzuki, R. i wsp., 2012. Package pvclust version 1.2-2. *Technical Report*, pp.1–13.

Tan, P.K. i wsp., 2003. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, 31(19), pp.5676–5684.

Tanaka, T.S. i wsp., 2000. Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proceedings of the National Academy of Sciences of the United States of America*, 97(16), pp.9127–9132.

Tarca, a L., Cooke, J.E.K. & Mackay, J., 2005. A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data. *Bioinformatics (Oxford, England)*, 21(11), pp.2674–2683.

Togawa, N. i wsp., 2012. Gene expression analysis of the liver and skeletal muscle of psyllium-treated mice. *The British journal of nutrition*, pp.1–11.

Trevino, V., Falciani, F. & Barrera-saldaña, H.A., 2007. DNA Microarrays : a Powerful Genomic Tool for Biomedical and Clinical Research. *Molecular Medicine*, 13(October), pp.527–541.

Tsai, M.-H. i wsp., 2005. Evaluation of hybridization conditions for spotted oligonucleotide-based DNA microarrays. *Molecular Biotechnology*, 29(3), pp.221–224.

Tuikkala, J. i wsp., 2008. Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC bioinformatics*, 9, pp.202–216.

Vallon-Christersson, J. i wsp., 2009. BASE--2nd generation software for microarray data management and analysis. *BMC bioinformatics*, 10, pp.330–337.

Veer, L.J. Van i wsp., 2002. Gene expression profling predicts clinical outcome of breast cancer. *Nature*, 415(345), pp.530–536.

Venkatasubbarao, S., 2004. Microarrays--status and prospects. *Trends in biotechnology*, 22(12), pp.630–637.

Wang, T. i wsp., 2000. Identification of genes differentially over-expressed in lung squamous cell carcinoma using combination of cDNA subtraction and microarray analysis. *Oncogen*, 19, pp.1519–1528.

Wang, Z., Gerstein, M. & Snyder, Michael, 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), pp.57–63.

Wilson, D.L. i wsp., 2003. New normalization methods for cDNA microarray data. *Bioinformatics*, 19(11), pp.1325–1332.

Yang, Y.H. i wsp., 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research*, 30(4), p.e15.

Yang, Y.H. & Yang Y.H., P.A. and D.S., 2009. marray: Exploratory analysis for two-color spotted microarray data. R package version 1.32.0. , pp.1–4.

Yee, J.C. i wsp., 2008. Quality assessment of cross-species hybridization of CHO transcriptome on a mouse DNA oligo microarray. *Biotechnology and bioengineering*, 101(6), pp.1359–1365.

Zhu, B., Xu, F. & Baba, Y., 2006. An evaluation of linear RNA amplification in cDNA microarray gene expression analysis. *Molecular genetics and metabolism*, 87(1), pp.71–79.

VIII. Załączniki

Wszystkie załączniki znajdują się na płycie CD dołączonej do pracy:

- Załącznik 1- Charakterystyka plików .gpr dla zestawu AML
- Załącznik 2- Charakterystyka plików .gpr dla zestawu AML II
- Załącznik 3- Charakterystyka plików .gpr dla zestawu AML miRNA
- Załącznik 4- Charakterystyka plików .gpr dla zestawu ALERGIA
- Załącznik 5- Charakterystyja plików .gpr dla zestawu ASTMA
- Załącznik 6- Szczegółowy opis funkcji i ich parametrów użytych w trakcie analizy zestawów danych
- Załącznik 7- Kod algorytmu wymiar_macierzy.py
- Załącznik 8- Kod algorytmu przyrównanie_sekwencji.py
- Załącznik 9- Kod algorytmu filtracja.py
- Załącznik 10- Wykresy MA dla zestawów AML II, ALERGIA, ASTMA i OSHLACK normalizowanych z wykorzystaniem 13 wybranych metod normalizacji.