

Politechnika Śląska w Gliwicach
Wydział Automatyki, Elektroniki i Informatyki



Analiza procesów regulacji ekspresji genów w komórkach
poddanych działaniu promieniowania jonizującego.

Roman Jaksik

Rozprawa doktorska

przygotowana pod kierunkiem:
Prof. dr hab. Joanny Rzeszowskiej-Wolny

GLIWICE 2013

Serdeczne podziękowania za zaangażowanie i pomoc udzieloną w trakcie pisania niniejszej pracy chciałbym złożyć Pani prof. Joannie Rzeszowskiej-Wolny.

Dziękuję także Pani prof. Joannie Polańskiej, dr Jerzemu Jurce, dr inż. Annie Lalik, mgr Robertowi Herokowi, mgr inż. Michałowi Marczykowi, mgr. inż. Krzysztofowi Biernackiemu, mgr inż. Dorocie Hudy oraz wszystkim kolegom i koleżankom, których nieoceniona pomoc i dobroć serca przyczyniła się do napisania niniejszej pracy doktorskiej.

Pragnę również podziękować Katarzynie Wuczyńskiej za ciągłe wsparcie duchowe w trakcie powstawania niniejszej pracy.

SPIS TREŚCI

1. STRESZCZENIE	3
2. WPROWADZENIE	5
2.1. MOTYWACJA PROWADZONYCH BADAŃ	5
2.2. HIPOTEZA	7
2.3. CELE PRACY	7
2.4. SŁOWNIK SKRÓTÓW STOSOWANYCH W PRACY	8
3. PODŁOŻE BIOLOGICZNE	10
3.1. PRZEPLYW INFORMACJI GENETYCZNEJ	10
3.2. BUDOWA GENU	11
3.3. REGULACJA EKSPRESJI GENÓW W KOMÓRKACH EUKARIOTYCZNYCH	12
3.4. PODSTAWOWE CZYNNIKI ODPOWIEDZIALNE ZA REGULACJĘ EKSPRESJI GENÓW	14
3.4.1. <i>Czynniki transkrypcyjne</i>	14
3.4.2. <i>Białka typu RBP</i>	15
3.4.3. <i>Interferencja RNA</i>	16
3.5. ZNACZENIE NIEKODUJĄCEGO OBSZARU KOŃCA 3' GENU	17
3.6. ROZKŁAD ELEMENTÓW REGULACYJNYCH W GENOMIE	18
3.7. WPŁYW PROMIENIOWANIA JONIZUJĄCEGO NA KOMÓRKĘ	20
3.8. WIELKOSKALOWE METODY ANALIZY FUNKCJI GENÓW	21
3.9. MIKROMACIERZE OLIGONUKLEOTYDOWE	22
3.9.1. <i>Typowe schematy eksperymentu mikromacierzowego</i>	23
3.9.2. <i>Budowa mikromacierzy</i>	24
3.9.3. <i>Podstawy biologiczne eksperymentu mikromacierzowego</i>	26
3.9.4. <i>Metody kontroli jakości danych</i>	31
3.9.5. <i>Metody wstępnego przetwarzania i problemy z nimi związane</i>	39
3.9.6. <i>Metody poszukiwania genów różnicujących</i>	41
3.10. POSZUKIWANIE WZORCÓW W SEKWENCJACH NUKLEOTYDOWYCH	42
3.10.1. <i>Metody i narzędzia do analizy sekwencji nukleotydowych</i>	43
3.10.2. <i>Publiczne bazy danych sekwencji nukleotydowych</i>	45
4. MATERIAŁY I METODY	47
4.1. ŹRÓDŁA DANYCH	47
4.1.1. <i>Dane mikromacierzowe – zbiór testowy</i>	47
4.1.2. <i>Dane mikromacierzowe – zbiór referencyjny</i>	48
4.1.3. <i>Sekwencje nukleotydowe i dane adnotacyjne</i>	51
4.2. ANALIZA SEKWENCJI NUKLEOTYDOWYCH	51
4.2.1. <i>Analiza składu nukleotydowego</i>	52

4.2.2.	<i>Macierze wag pozycji</i>	52
4.2.3.	<i>Miejsca wiążące miRNA</i>	53
4.3.	ANALIZA DANYCH MIKROMACIERZOWYCH	54
4.4.	EKSPERYMENTY RT-QPCR	56
5.	KONSTRUKCJA OPROGRAMOWANIA.....	58
5.1.	ANALIZA SEKWENCJI NUKLEOTYDOWYCH – NUCLEOSEQ	58
5.2.	ANALIZA CECH FUNKCJONALNYCH GENÓW – NUCLEOANNOT.....	61
5.3.	INNE PROGRAMY UŻYTKOWE	62
6.	WYNIKI ANALIZY	63
6.1.	ANALIZA DANYCH Z EKSPERYMENTU MIKROMACIERZOWEGO.....	63
6.1.1.	<i>Kontrola jakości i wstępne przetwarzanie danych</i>	63
6.1.2.	<i>Identyfikacja transkryptów różnicujących</i>	66
6.2.	IDENTYFIKACJA CECH TRANSKRYPTÓW RÓZNICUJĄCYCH.....	67
6.2.1.	<i>Podstawowe własności sekwencji</i>	67
6.3.	WERYFIKACJA UZYSKANYCH WYNIKÓW	69
6.3.1.	<i>Dodatkowe linie komórkowe</i>	69
6.3.2.	<i>Alternatywna platforma badawcza</i>	70
6.4.	ANALIZA WŁASNOŚCI SEKWENCJI O OKREŚLONYM SKŁADZIE NUKLEOTYDOWYM.....	71
6.4.1.	<i>Skład nukleotydowy genów i genomu</i>	71
6.4.2.	<i>Częstotliwość występowania motywów regulatorowych w DNA</i>	72
6.4.3.	<i>Częstotliwość występowania motywów regulatorowych w RNA</i>	73
6.5.	ŹRÓDŁA NIEDOKŁADNOŚCI POMIAROWYCH W EKSPERYMENTACH MIKROMACIERZOWYCH	76
6.5.1.	<i>Analiza wariancji sygnału sond</i>	77
6.5.2.	<i>Wariancja sygnału sond w zestawach</i>	79
6.5.3.	<i>Analiza znaczenia czynników wpływających na wariancję sygnału pomiędzy sondami w zestawach</i>	86
6.5.4.	<i>Różnice w poziomach sygnału sond o różnym składzie GC</i>	88
6.5.5.	<i>Wpływ obciążenia wynikającego ze składu GC na wyniki eksperymentu mikromacierzowego</i>	91
6.5.6.	<i>Wpływ składu GC na wyniki niezależnych eksperymentów</i>	95
6.5.7.	<i>Korekcja wpływu składu GC na poziomy sygnał sond</i>	97
6.6.	IDENTYFIKACJA CECH TRANSKRYPTÓW RÓZNICUJĄCYCH PO PRZETWORZENIU DANYCH METODĄ CSGC-RMA	102
6.6.1.	<i>Statystyki transkryptów różnicujących</i>	102
6.6.2.	<i>Mechanizmy regulacji ekspresji genów</i>	104
6.6.3.	<i>Odpowiedź komórek na promieniowanie</i>	109
7.	PODSUMOWANIE	114
8.	LITERATURA.....	120
9.	PUBLIKACJE AUTORA	133

1. Streszczenie

Promieniowanie jonizujące, jeden z powszechnie występujących w przyrodzie i używanych przez człowieka czynników genotoksycznych, indukuje szereg zmian w komórce powodując uszkodzenia większości jej elementów w tym materiału genetycznego a także wpływając na przebieg najróżniejszych procesów chemicznych zachodzących na poziomie transkryptów bądź pojedynczych białek. Badania mikromacierzowe komórek poddanych działaniu promieniowania pokazują, iż natychmiast po ekspozycji można obserwować zmiany poziomu setek różnych cząsteczek mRNA, które najprawdopodobniej są wynikiem zmian ich stabilności, a ta zależy od struktury nukleotydowej transkryptu w tym obecności specyficznych motywów sekwencyjnych wiążących regulatorowe białka i interferencyjne RNA.

Celem niniejszej pracy jest określenie cech sekwencji nukleotydowych transkryptów, które mogą wpływać na ich stabilność w warunkach stresu komórkowego wywołanego promieniowaniem oraz stworzenie zaplecza metodologicznego oraz oprogramowania, które pozwoliłoby na przeprowadzenie tego typu analizy w oparciu o dane z eksperymentów mikromacierzowych.

W pierwszych etapach pracy stworzono szereg aplikacji bioinformatycznych służących do automatycznego pobierania oraz analizowania sekwencji obszarów promotora genu oraz różnych obszarów transkryptu, do poszukiwania informacji o funkcji określonych genów w publicznie dostępnych bazach danych a także analizy statystycznej wyników badań mikromacierzowych. W oparciu o opracowane programy podjęto próbę scharakteryzowania motywów regulatorowych, których obecność różnicuje transkrypty stabilizowane lub destabilizowane przez promieniowanie oraz ustalenia, jakie mechanizmy regulacyjne odgrywają najistotniejszą rolę w odpowiedzi komórkowej na tego typu czynnik genotoksyczny. Badania przeprowadzono na komórkach czerniaka ludzkiego z linii Me45, komórkach białaczki limfoblastycznej K562 i raka okrężnicy HCT116.

Wyniki analizy, która objęła ponad 20 tysięcy transkryptów, pokazały istnienie wysokiej korelacji pomiędzy zmianą poziomu ekspresji oraz strukturą nukleotydową mRNA. Transkrypty, które wykazywały zwiększoną stabilność w warunkach stresu wywołanego promieniowaniem charakteryzowały się znacznie mniejszą zawartością nukleotydów GC oraz większą zawartością motywów regulatorowych wpływających na ich tempo produkcji a także czas półtrwania.

W ramach pracy zbadano również wpływ cech użytej do badań mikromacierzowych platformy Affymetrix, na uzyskiwane wyniki hybrydyzacji i wykazano, że przy użyciu zalecanego przez producenta oprogramowania do analizy wyników oraz powszechnie stosowanych metod normalizacyjnych uzyskuje się przeszacowanie poziomu ekspresji pewnej grupy transkryptów o skrajnych proporcjach nukleotydów GC. Efekt ten nazywany w literaturze obciążeniem wynikającym ze składu GC (z ang. GC content bias), został już wcześniej zaobserwowany w przypadku danych z głębokiego sekwencjonowania DNA lub RNA prowadząc do przeszacowania ilości odczytów o określonej zawartości nukleotydów GC.

Obciążenie danych wynikające ze składu GC ma swoje podłoże między innymi w procesie amplifikacji materiału genetycznego, przez co może mieć wpływ na wszystkie metody pomiarowe biologii molekularnej, które wykorzystują ten proces (sekwencjonowanie, RT-qPCR, mikromacierze). Ponieważ wpływ obciążenia nie jest stały pomiędzy próbkami to w przypadku mikromacierzy może on obniżać skuteczność algorytmów identyfikacji transkryptów różnicujących badane próbki, prowadząc do

nadrepresacji mRNA o skrajnych proporcjach składu GC w grupach transkryptów o zwiększonej lub zmniejszonej ekspresji.

W pracy pokazano, że częstotliwość występowania wszystkich znanych klas motywów regulatorowych jest bardzo silnie skorelowana ze składem nukleotydowym badanych sekwencji co w przypadku danych mikromacierzowych obciążonych składem GC może mieć dodatkowy wpływ na wyniki analizy częstotliwości występowania motywów regulatorowych w transkryptach różnicujących.

W ramach pracy zaproponowano nową metodę przetwarzania danych mikromacierzowych csGC-RMA, która pozwala na ograniczenie wpływu obciążenia wynikającego ze składu GC na wyniki eksperymentu mikromacierzowego i tym samym podwyższenie czułości oraz specyficzności algorytmów identyfikacji genów różnicujących.

Zaproponowaną metodę csGC-RMA wykorzystano do identyfikacji transkryptów różnicujących w napromieniowanych komórkach Me45 wykazując zależność pomiędzy wzrostem poziomu ekspresji genów a częstotliwością występowania motywów regulatorowych odpowiedzialnych za przyłączanie cząsteczek miRNA (mikro RNA). Pokazano, że transkrypty o wyższej stabilności w napromieniowanych komórkach Me45 posiadają znacznie więcej potencjalnych miejsc przyłączenia miRNA, sugerując, iż ekspozycja na promieniowanie prowadzi do relaksacji procesów związanych z degradacją mRNA poprzez mechanizm interferencji RNA. W ramach pracy zidentyfikowano także grupę genów o istotnym znaczeniu dla odpowiedzi komórkowej na promieniowanie, takich jak AUF1, CCNB1, PARP1, AGO2, DICER1, których zmiany poziomu ekspresji pod wpływem promieniowania dodatkowo zweryfikowano za pomocą techniki RT-qPCR.

Niniejsza rozprawa doktorska opisuje podstawowe elementy regulacji ekspresji genów w napromieniowanych komórkach a także prezentuje możliwości opracowanych narzędzi bioinformatycznych w analizie wewnątrzkomórkowych oddziaływań regulacyjnych. Zwraca także uwagę na nowe aspekty analizy i interpretacji danych z wielkoskalowych eksperymentów mikromacierzowych. Opracowane w ramach pracy programy mogą być wykorzystywane do testowania innych hipotez niż przedstawione a dzięki temu, że są publicznie dostępne i wyposażone w graficzny interfejs rozszerzają grono potencjalnych użytkowników o osoby bez zaawansowanej wiedzy z zakresu bioinformatyki.

2. Wprowadzenie

2.1. Motywacja prowadzonych badań

Promieniowanie jonizujące jest wszechobecne w środowisku człowieka, przez co stale narażeni jesteśmy na jego niskie dawki, w wyższych dawkach jest dodatkowo wykorzystywane w radioterapii, jako jedna z metod walki z chorobami nowotworowymi.

Ekspozycja na czynniki stresowe takie jak promieniowanie jonizujące wywołuje globalne zmiany w poziomach ekspresji genów, które bardzo silnie różnią się pomiędzy komórkami wyizolowanymi z różnych tkanek a nawet komórkami tego samego typu pobranymi od innych osobników. Różnorodność odpowiedzi komórkowej na promieniowanie jest bardzo istotnym problemem współczesnej radioterapii. W niektórych przypadkach komórki nowotworowe mogą wykazywać zwiększoną odporność na działanie określonych dawek promieniowania, czyniąc radioterapie nieefektywną. Z drugiej jednak strony obniżona odporność zdrowych komórek znajdujących się w okolicy napromieniowanej tkanki może prowadzić do powstania groźnych odczynów popromiennych, których niekorzystne skutki nieraz objawiają się dopiero po kilku latach po zakończeniu terapii.

Badania wpływu promieniowania na odpowiedź komórek w postaci zmiany profilu ekspresji genów mogą dostarczyć odpowiedzi na bardzo wiele pytań związanych z mechanizmami naprawy bądź programowej śmierci komórek. Pod wpływem promieniowania w komórkach aktywowanych jest szereg szlaków sygnałowych odpowiedzialnych m.in. za procesy naprawy DNA, wiele z nich ulega także wygaszeniu, co może być skutkiem destabilizacji cząsteczek mRNA uczestniczących w produkcji białek.

Badania zmian w profilach ekspresji genów napromieniowanych komórek pokazują, że nawet kilka minut po napromieniowaniu można zaobserwować znaczną zmianę poziomu ekspresji niektórych transkryptów sugerując istotny wpływ procesów opartych o stabilność RNA w mechanizmach regulacji ekspresji. Określone cechy sekwencji nukleotydowych transkryptów w tym obecność motywów sekwencyjnych w obszarze końca 3', odpowiedzialnych za interakcje z białkami cytoplazmatycznymi lub funkcjonalnym RNA, może istotnie wpływać na czas półtrwania mRNA jednak ich dokładna rola w odpowiedzi na promieniowanie wciąż pozostaje nieznana.

Sekwencja końca 3', pomimo, że nie zawiera informacji o strukturze białka, może wpływać na ilość transkryptów (ilość cząsteczek mRNA) poprzez ich stabilizację bądź destabilizację w wyniku poddania ich wpływom czynników genotoksycznych [1]. Różny skład nukleotydowy sekwencji końca 3' genów może determinować stabilność transkryptów poddanych wpływowi promieniowania jonizującego poprzez białka wiążące się z określonymi motywami sekwencyjnymi (RBP) lub poprzez miejsca oddziaływania z funkcjonalnymi cząsteczkami RNA (MRE), które obniżają ich stabilność.

Przeprowadzone w ramach pracy magisterskiej badania wskazały na istnienie istotnych różnic w budowie sekwencji nukleotydowej pomiędzy transkryptami o zwiększonej i zmniejszonej ekspresji na skutek promieniowania. Sugeruje to, że procesy odpowiedzialne za regulację poziomu ekspresji genów takie jak interakcje z białkami lub funkcjonalnym RNA mogą być zaburzone poprzez działanie promieniowania.

Istnieje jednak duża obawa, że różnice pomiędzy poziomami transkryptów wynikają nie tylko z działania biologicznych mechanizmów regulacji ekspresji, ale ze specyfiki badań mikromacierzowych, które na skutek nieznanych czynników prowadzą do powstawania różnic wynikających z aspektów technicznych samego eksperymentu. Analiza danych mikromacierzowych jest bardzo trudna ze względu na szereg czynników wpływających na dokładność metody na każdym z etapów eksperymentu. Kompensacja różnic technicznych pomiędzy próbkami w celu wydobycia różnic mających podłoże biologiczne wymaga stosowania skomplikowanych algorytmów standaryzacji danych, których skuteczność jest ograniczona specyfiką analizowanego zbioru danych, przez co powtarzalność samego eksperymentu jak i korelacja wyników uzyskanych za pomocą innych metod badawczych (RT-qPCR, sekwencjonowanie) jest często niezadowalająca.

Mikromacierze są jedną z najczęściej wykorzystywanych wielkoskalowych technik pomiarowych ekspresji genów i pomimo bardzo szybkiego rozwoju technik głębokiego sekwencjonowania, które uważane są za ich następcę, mikromacierze nadal są powszechnie wykorzystywane na całym świecie do badania zmian w profilach ekspresji genów. Publicznie dostępne zbiory danych mikromacierzowych dodatkowo oferują ogromne zasoby informacji zebranych przez ostatnie kilkanaście lat badań, które mogą być w dalszym ciągu wykorzystywane do uzyskania odpowiedzi na nowo postawione hipotezy. Nieustannie rozwijające się techniki standaryzacji i interpretacji danych dodatkowo wspomagają ten proces w stopniu, jaki nieraz był nieosiągalny w czasie gdy przeprowadzano eksperyment.

Wiele z czynników mogących mieć wpływ na dane uzyskane w eksperymentach mikromacierzowych zostało opisanych w literaturze naukowej, jednak ich dokładny wpływ na dokładność oszacowania zmian w profilach ekspresji genów jest nieznaną a większość problemów jest nieraz pomijana podczas typowej analizy. Konieczne jest zatem określenie potencjalnych źródeł różnic o podłożu technicznym pomiędzy transkryptami oraz uniezależnienie wyników od specyfiki procedury badawczej co może znacznie zwiększyć dokładność interpretacji danych mikromacierzowych.

Głównym problemem związanym z analizą danych z wielkoskalowych eksperymentów biologicznych jest dodatkowo nie tylko wysoki poziom skomplikowania algorytmów przetwarzania danych, ale sama ich ilość. W konsekwencji każda, nawet najmniej skomplikowana analiza wymaga odpowiedniego oprogramowania, które pozwoli na zrealizowanie opracowanej procedury badawczej w akceptowalnym przedziale czasowym, co stwarza potrzebę opracowania szeregu aplikacji bioinformatycznych. Pomimo bardzo szybkiego rozwoju bioinformatyki dostarczającej narzędzia służące do analizy i wizualizacji danych biologicznych brakuje szybkich i wygodnych w użyciu narzędzi służących do eksploracji danych w postaci sekwencji nukleotydowych w celu poszukiwania motywów mogących wpływać na zmiany w profilach ekspresji genów. Kolejnym problemem jest uzyskanie dodatkowych informacji o wybranych genach obejmujących ich funkcje biologiczne rozrzucone po najróżniejszych bazach danych i artykułach naukowych nieraz pod innym identyfikatorem. Opracowanie aplikacji wspomagającej pozyskiwanie informacji o sekwencjach nukleotydowych i obecnych w nich motywach a także roli poszczególnych genów jest bardzo istotne z punktu widzenia specyfiki zaproponowanych badań.

Identyfikacja czynników związanych z analizą danych mikromacierzowych, które mogą w istotny sposób wpływać na wyniki, oraz roli poszczególnych mechanizmów regulacji ekspresji genów w napromieniowanych komórkach eukariotycznych, może przyczynić się do lepszego zrozumienia mechanizmów odpowiedzi komórkowej na stres wywołany promieniowaniem jonizującym.

2.2. Hipoteza

Podstawowa hipoteza niniejszej pracy jest następująca:

Obserwowane za pomocą metod mikromacierzowych zmiany transkryptomu wywołane przez ekspozycję żywych komórek na promieniowanie jonizujące, są determinowane obecnością motywów określonego typu w sekwencjach nukleotydowych transkryptów.

Weryfikacja sformułowanej hipotezy składa się z następujących etapów:

- a) przegląd istniejącego oprogramowania oraz metod analizy danych mikromacierzowych i sekwencji nukleotydowych transkryptów
- b) wytworzenie nowego oprogramowania pozwalającego na analizę częstotliwości występowania motywów regulacyjnych w sekwencjach DNA i RNA oraz oprogramowania wspomagającego proces przeszukiwania literatury naukowej i baz danych na temat funkcji specyficznych genów
- c) określenie podstawowych czynników technicznych mogących wpływać na proces identyfikacji transkryptów różnicujących w eksperymentach opartych o mikromacierze Affymetrix oraz obniżenie ich wpływu na wyniki przeprowadzanej analizy
- d) wyodrębnienie i charakterystyka transkryptów, których ekspresja zmienia się pod wpływem promieniowania jonizującego w oparciu o dane mikromacierzowe uzyskane dla komórek czerniaka Me45
- e) określenie motywów regulatorowych charakterystycznych dla transkryptów o zmienionym poziomie ekspresji za pomocą opracowanych aplikacji
- f) charakterystyka odpowiedzi komórkowej na promieniowanie jonizujące na przykładzie komórek Me45 oraz określenie podstawowych mechanizmów regulacji ekspresji genów, które biorą udział w wyidukowaniu tej odpowiedzi

2.3. Cele Pracy

Rozwój narzędzi do wielkoskalowej analizy sekwencji nukleotydowych oraz danych mikromacierzowych

Określenie wpływu cech strukturalnych mRNA na ich stabilność wymaga udoskonalenia istniejących metod wstępnego przetwarzania i kontroli jakości danych pochodzących z wielkoskalowych eksperymentów mikromacierzowych. Umożliwi to dokładne określenie zmian w poziomach ekspresji genów, wynikających z poddania komórek działaniu promieniowania jonizującego. Rozwój oprogramowania umożliwiającego analizę ściśle określonych elementów sekwencji nukleotydowej genu jest także bardzo istotny ze względu na brak odpowiednich narzędzi pozwalających na przeprowadzanie wybranych schematów analizy na dużych zbiorach danych. Rozwój aplikacji komputerowych jest dodatkowo bardzo istotny w związku z nieustannie powiększającą się wiedzą na temat sekwencji nukleotydowych transkryptów oraz udoskonaleniami w technologii badań mikromacierzowych, które nie są na bieżąco uwzględniane w istniejących aplikacjach.

Analiza potencjalnych źródeł błędów systematycznych w eksperymentach mikromacierzowych

Dane mikromacierzowe znane są z niskiej dokładności wynikającej z wpływu czynników niezależnych od badanego materiału biologicznego. Identyfikacja potencjalnych źródeł błędów systematycznych w eksperymencie oraz ich wpływu na interpretacje danych jest zatem niezbędna w przypadku poszukiwania globalnych zmian w poziomach ekspresji.

Identyfikacja elementów regulacyjnych w obszarach sekwencji o różnym składzie nukleotydowym

Genom jest zbudowany z fragmentów sekwencji o różnych proporcjach składu GC, które mogą w istotny sposób wpływać na rozmieszczenie elementów regulacyjnych w sekwencjach genów, promotorów a także dojrzałych transkryptów. Określenie zależności pomiędzy proporcjami w składzie nukleotydowym a częstotliwością występowania motywów rozpoznawanych przez białka RBP, miRNA oraz czynniki transkrypcyjne jest zatem bardzo istotna z punktu widzenia badań mających na celu zidentyfikowanie tych czynników, które charakteryzują określone grupy transkryptów.

Analiza mechanizmów odpowiedzialnych za regulacje ekspresji w warunkach stresu komórkowego

Dodatkowym celem projektu jest określenie roli struktury nukleotydowej mRNA w stabilizacji bądź destabilizacji transkryptów w warunkach stresu komórkowego wywołanego promieniowaniem jonizującym. Wiadomo, że obecność niektórych motywów sekwencyjnych obecnych w obszarze niekodującego fragmentu końca 3' jest niezbędna do oddziaływania z białkami cytoplazmatycznymi i funkcjonalnym RNA wpływając na okres półtrwania cząsteczek mRNA, jednak ich dokładna rola w odpowiedzi na warunki stresu wywołane promieniowaniem wciąż pozostaje nieznana.

2.4. Słownik skrótów stosowanych w pracy

Skrót	Angielskie rozwinięcie	Znaczenie
ARE	AU-rich element	Motywy sekwencji nukleotydowych bogate w nukleotydy A i U, odpowiedzialne za przyłączenie specyficznych białek regulatorowych z rodziny RBP
CDF	Chip Definition File	Plik definiujący zestawy sond mikromacierzowych, określa identyfikator każdego zestawu oraz informacje o tym, które sondy do niego należą
CDS	Coding Sequence	Sekwencja kodująca – fragment sekwencji transkryptu określający kolejność aminokwasów w kodowanym białku
CEL	-	Format pliku przechowujący informacje o poziomie intensywności fluorescencji każdej z sond mikromacierzowych
DEG	Differentially Expressed Gene	Geny, których ekspresja ulega zmianie pomiędzy badanymi próbkami
LFC	Logarithm Fold-Change	Logarytm ilorazu dwóch porównywanych pomiarów
LOESS	LOcally Estimated Scatterplot Smoothing	Lokalnie estymowane wygładzanie wykresu rozrzutu - metoda dopasowania regresji liniowej
MRE	miRNA Response Elements	Motyw sekwencyjny pozwalający na przyłączenie cząsteczek miRNA
MM	MisMatch	Typ sondy mikromacierzowej, której środkowy nukleotyd celowo nie jest komplementarny do sekwencji badanego RNA

PCA	Principal Component Analysis	Analiza głównych składowych – metoda statystyczna służąca do określania podstawowych źródeł zróżnicowania pomiędzy próbkami
PM	Perfect-Match	Typ sondy mikromacierzowej o pełnej komplementarności do sekwencji badanego transkryptu
PVCA	Principal Variance Component Analysis	Analiza głównych składowych wariancji sygnału
RBP	RNA Binding Proteins	Białka wyposażone w domeny, które pozwalają im przyłączać się do RNA
RIN	RNA Integrity Number	Współczynnik integralności RNA – miara określająca stopień w jakim badane DNA jest zdegradowane (0-10 gdzie 0 to zdegradowane RNA)
RISC	RNA Induced Silencing Complex	Kompleks białkowy biorący udział w wyciszaniu ekspresji genów w procesie interferencji RNA
RMA	Robust Multi-array Average	Metoda wstępnego przetwarzania danych mikromacierzowych na którą składają się algorytmy korekcji tła, normalizacji oraz sumaryzacji.
ROC	Receiver Operating Characteristic	Metoda oceny jakości klasyfikatora, jest to wykres zależności czułości metody od poziomu jej specyficzności, gdzie parametrem jest próg wykorzystywany do podziału cech na klasy
SLGC	-	Parametr określający korelację Spearmana pomiędzy dwiema statystykami obliczonymi dla par mikromacierzy: zmianą współczynnika nachylenia prostej dopasowanej do zależności składu GC sond od ich poziomu sygnału oraz zależnością składu GC transkryptu od zmiany poziomu ekspresji danego zestawu sond (LFC).
SNP	Single Nucleotide Polymorphism	Polimorfizm pojedynczego nukleotydu – zamiana pojedynczego nukleotydu pomiędzy różnymi osobnikami określonego gatunku
SNR	Signal-to-Noise Ratio	Stosunek sygnału do szumu pomiarowego
TF	Transcription Factor	Czynnik transkrypcyjny – białko, które poprzez łączenie się z sekwencją DNA reguluje ekspresję położonego w pobliżu genu
TFBS	Transcription Factor Binding Site	Miejsce wiązania białka z rodziny czynników transkrypcyjnych w sekwencji DNA.
TSS	Transcription Start Site	Miejsce określające początek transkrypcji genu
UTR	Untranslated Region	Obszar niekodujący – fragment sekwencji transkryptu, który nie koduje struktury białkowej (nie ulega translacji)

3. Podłoże biologiczne

3.1. Przepływ informacji genetycznej

Białka produkowane we wszystkich żywych komórkach powstają w procesie translacji gdzie na podstawie matrycy zbudowanej z kwasu rybonukleinowego (RNA) powstaje łańcuch aminokwasów o strukturze określonej przez kolejność nukleotydów w matrycy, który następnie w procesie tzw. zwiłania białka (ang. protein folding) przyjmuje strukturę przestrzenną. Matrycą dla budowy białek jest tzw. przekaźnikowy RNA (ang. messenger RNA - mRNA) będący polimerem złożonym z rybonukleotydów, którego głównym zadaniem jest przekazanie krótkich fragmentów informacji zapisanych w DNA do rybosomów gdzie następuje proces jej odczytu i synteza białka. Częsteczka mRNA powstaje w jądrze komórkowym w procesie nazywanym transkrypcją gdzie na bazie łańcucha DNA powstaje jednoniciowa cząsteczka RNA w reakcji, w której kluczową rolę odgrywa polimeraza RNA - enzym przyłączający się do nici DNA o określonej sekwencji nukleotydów.

DNA zbudowane jest z dwóch nici wzajemnie do siebie komplementarnych, które pełnią kluczową rolę w procesach replikacji (powielania DNA) oraz naprawy uszkodzeń. Obie nici wiążą się ze sobą tworząc tzw. pary Watsona-Cricka poprzez podwójne lub potrójne wiązania wodorowe, odpowiednio między nukleotydami A i T oraz G i C. Zasada komplementarności sekwencji nukleotydowych ma kluczowe znaczenie nie tylko w przypadku budowy podwójnoniciowego DNA. Odgrywa ona istotną rolę w wielu mechanizmach opartych o oddziaływania pomiędzy cząsteczkami RNA gdzie obowiązują te same reguły za wyjątkiem pary A-T gdzie rolę tyminy pełni uracyl U.



Ryc. 1: Centralny dogmat biologii molekularnej

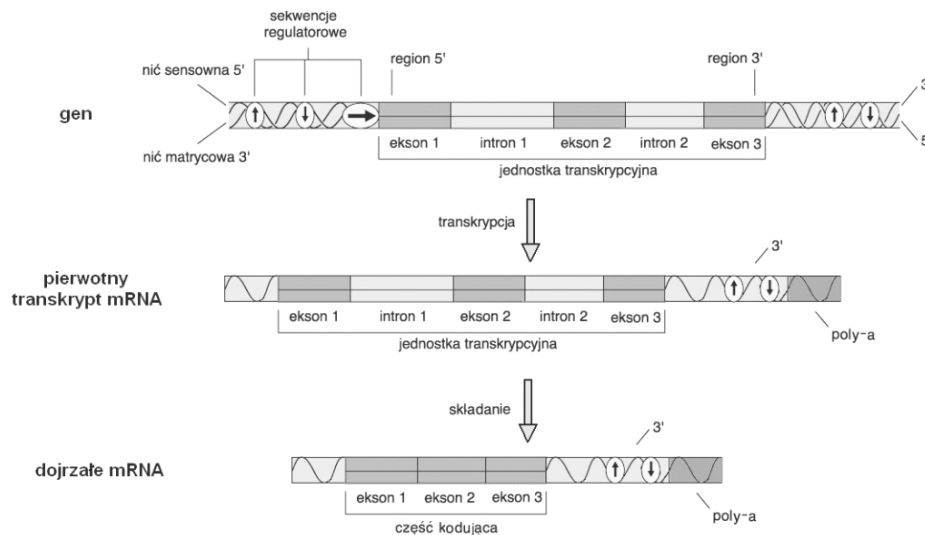
Sposób przekazywania informacji zapisanej w DNA za pośrednictwem matrycowego RNA, którego produktem są cząsteczki białka, określany jest centralnym dogmatem biologii molekularnej (Ryc. 1). Jego kluczowe znaczenie wynika z faktu, iż opisuje on uniwersalne reguły leżące u podstaw funkcjonowania wszystkich żywych organizmów.

W ujęciu technicznym komórka może być rozpatrywana jako fabryka związków chemicznych zbudowanych z pojedynczych cząsteczek białkowych lub całych kompleksów, które odpowiedzialne są za wszystkie podstawowe funkcje życiowe niezbędne dla komórki do przetrwania. Jądro komórkowe, w którym znajduje się materiał genetyczny w formie dwuniciowego DNA pełni rolę centrum, w którym przechowywane są wszystkie informacje niezbędne do przeprowadzania procesów biochemicznych oraz miejsca gdzie dochodzi do odczytu tej informacji i skopiowania na nośnik w postaci jednoniciowej cząsteczki mRNA. mRNA transportowane jest do cytoplazmy będącej magazynem związków chemicznych oraz jednocześnie miejscem ich syntezy i rozkładu. O prawidłowym działaniu tego systemu decydują

najróżniejsze mechanizmy automatycznej regulacji poziomów ekspresji genów oparte o specyficzne cechy ich struktury nukleotydowej.

3.2. Budowa genu

Geny są to niewielkie fragmenty DNA, które odczytywane są w procesach transkrypcji odpowiadając za strukturę cząsteczek RNA. U człowieka są one rozrzucone w postaci ok. 25 tys fragmentów o średniej długości 56 tys. par zasad na przestrzeni wszystkich chromosomów. Pozostałe elementy sekwencji nukleotydowej DNA odpowiadają głównie za kontrolowanie wydajności procesów zachodzących w jądrze komórkowym mających na celu przyłączenie polimerazy inicjującej proces transkrypcji oraz udostępnienie określonych fragmentów DNA do odczytu gdyż w normalnych warunkach jest ono silnie upakowane uniemożliwiając przyłączenie kompleksów białkowych.



Ryc. 2: Budowa genu oraz RNA matrycowego

Gen jednak nie składa się wyłącznie z informacji kodującej strukturę białka ale także z fragmentów niekodujących nazywanych sekwencjami końca 5' i 3' (Ryc. 2). Sam obszar kodujący zbudowany jest z fragmentów sekwencji – eksonów oddzielonych od siebie kilkukrotnie dłuższymi fragmentami sekwencji intronowych, które także ulegają transkrypcji jednak wycinane są na etapie składania mRNA. Odpowiednia regulacja procesu składania może dodatkowo prowadzić do powstawania tzw. alternatywnych form splicingowych danego genu w przypadku, gdy z jednego genu powstaje kilka różnych mRNA.

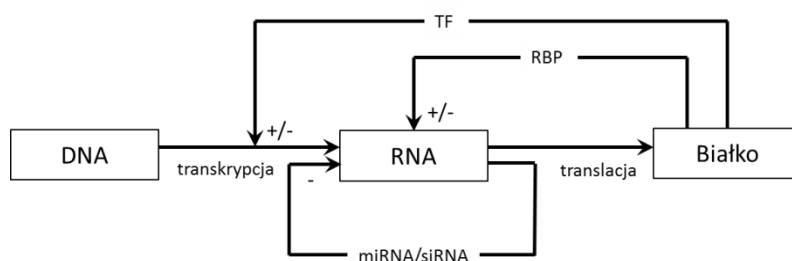
Dojrzała cząsteczka mRNA składa się z dwóch sekwencji niekodujących (z ang. 5'/3' untranslated regions - 5'/3'-UTRs) sekwencji kodującej (z ang. coding sequence - CDS) czapeczki guanylowej na końcu 5' oraz łańcucha nukleotydów adeniny na końcu 3' w nazywanego ogonem poli-A, który powstaje w procesie poliadenylacji. Ostatnie dwa elementy ułatwiają identyfikację cząsteczki wewnątrz komórki a ich brak prowadzi do szybkiej degradacji transkryptu, co jest jednym z głównych elementów mechanizmu rozpoznawania i usuwania obcych kwasów nukleinowych.

Dodatkowo translacja mRNA charakteryzującego się długą sekwencją adeniny na końcu 3' (poli-A) jest bardziej efektywna niż translacja cząsteczki o krótkim poli-A. Długi ogon poli-A zwiększa stabilność mRNA, podczas gdy jego brak może prowadzić do przyspieszonej degradacji transkryptów. [2]

3.3. Regulacja ekspresji genów w komórkach eukariotycznych

O tym ile określonych związków chemicznych jest produkowanych wewnątrz komórki decydują najróżniejsze mechanizmy regulacji procesu ekspresji genów, w którym informacja zapisana w genie wykorzystana jest do zsyntetyzowania białek.

Regulacja, której podstawy poznane zostały ponad 50 lat temu zachodzi na różnych etapach procesu ekspresji genów. Jest ona kontrolowana przez czynniki wpływające przede wszystkim na wydajność produkcji mRNA w procesie transkrypcji, tempo jego degradacji oraz wydajność procesu translacji [3].



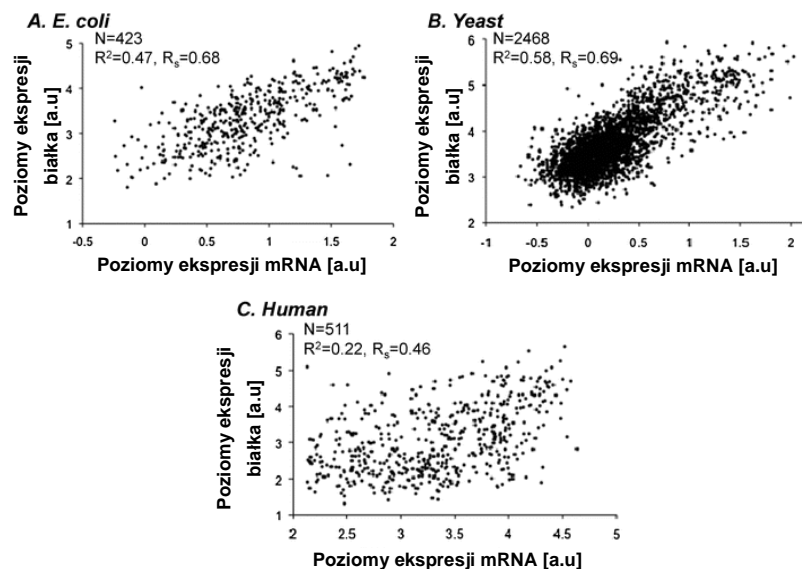
Ryc. 3: Podstawowe elementy procesu regulacji ekspresji genów

Główną zmienną w procesie regulacji jest poziom mRNA będący podstawowym elementem mechanizmu kontrolującego poziom określonych białek w komórce. Bilans określonego mRNA wewnątrz komórki określa tempo jego produkcji kontrolowane w dużej mierze przez białka z rodziny czynników transkrypcyjnych (ang. Transcription Factors - TF) oraz tempo jego degradacji uzależnione od czasu półtrwania samej cząsteczki, który w normalnych warunkach wynosi od kilku minut do kilkudziesięciu godzin [4]. Czynniki transkrypcyjne, które same powstają w procesach transkrypcji i translacji mogą albo stymulować translację genu docelowego albo blokować go w określonej sytuacji. Mogą one zatem pełnić funkcje zarówno dodatniej jak i ujemnej pętli sprzężenia zwrotnego w procesie wewnątrzkomórkowego przekazywania informacji (Ryc. 3). Głównymi czynnikami wpływającymi na czas półtrwania mRNA są tzw. funkcjonalne RNA, do których zaliczane są cząsteczki siRNA (small interfering RNA) oraz miRNA (micro RNA). W tym przypadku bezpośrednia regulacja polega wyłącznie na wygaszeniu ekspresji genu będącego ich celem, stanowią one zatem element ujemnej pętli sprzężenia zwrotnego. Białka poza funkcją bezpośredniego oddziaływania z DNA mogą także tworzyć kompleksy z RNA (ang. RNA Binding Proteins - RBP) zwiększając jego stabilność albo przyspieszając tempo degradacji.

Cechą wspólną opisanych mechanizmów regulacji jest sposób rozpoznawania cząsteczek będących celem regulacji oparty o wyszukiwanie określonych wzorców w sekwencjach nukleotydowych DNA lub RNA. Wzorce te nazywane potocznie motywami sekwencyjnymi różnią się zarówno długością jak i specyficznością wpływając na częstotliwość ich występowania co ma przełożenie na rozległość i poziom skomplikowania mechanizmów regulacyjnych, które nieraz obejmują tysiące genów.

Poziom złożoności mechanizmów kontrolujących ekspresję genów jest powiązany ze stopniem skomplikowania organizmu, który z uwagi na różnorodność pełnionych funkcji wymaga bardziej rozbudowanych systemów regulacyjnych [5]. Podkreśla to istotność tego typu mechanizmów w procesach niezbędnych dla prawidłowego funkcjonowania komórek jednocześnie czyniąc je układami o ogromnym poziomie skomplikowania, co utrudnia zarówno poznanie relacji pomiędzy określonymi elementami układów regulacyjnych a także ich pełnionej funkcji.

Ze względu na rozległość mechanizmów regulacyjnych, które obejmują wszystkie znane geny korelacja pomiędzy poziomem mRNA a kodowanymi przez nie białkami jest stosunkowo niska szczególnie w organizmach o wysokim poziomie złożoności (Ryc. 4). Wynika to między innymi ze specyfiki procesu transkrypcji, który ze względu na możliwość wyprodukowania wielu kopii białka w oparciu o jedną matrycę mRNA może kontrolować istotne zmiany w ich koncentracji, różniące się nieraz o kilka rzędów wielkości, poprzez niewielkie zmiany w ilości określonego mRNA [6, 7].



Ryc. 4: Wykres rozrzutu pomiędzy poziomem ekspresji mRNA i białka w 3 organizmach o różnym poziomie złożoności (A - bakterie E.Coli, B - drożdże, C - człowiek). (źródło: [5])

W ujęciu technicznym procesy zachodzące w komórce można rozpatrywać jak układ automatycznej regulacji, którego głównym celem jest utrzymanie stanu ustalonego w odpowiedzi na najróżniejsze czynniki zewnętrzne takie jak wysoka temperatura, promieniowanie czy też nagłe zmiany ciśnienia. Regulacja ma jednak na celu nie tylko utrzymanie pojedynczej komórki przy życiu, ale zapewnienie warunków do przeżycia całej ich populacji. Wiąże się to z przeprowadzaniem procesów prowadzących do rozmnażania się komórek (prolifracji), pozwalających na zachowanie interakcji pomiędzy sąsiadującymi komórkami oraz przekształcaniu w komórki pełniące określoną funkcję niezbędną dla przeżycia całej populacji jak np. tworzenie naczyń krwionośnych. W szczególnych warunkach regulacja prowadzi do zainicjowania zaprogramowanej śmierci zwanej apoptozą w przypadku, gdy komórka pełni nieprawidłową funkcję i jej dalszy podział stanowi zagrożenie dla funkcjonowania całego organizmu.

Główną niewiadomą w procesach regulacyjnych jest to, jakie sekwencje w DNA/RNA wpływają na proces ekspresji genu, jakie mechanizmy regulacyjne są odpowiedzialne za ich rozpoznawanie i co

wpływa na dynamikę ich zmian a także na zachowanie w warunkach stresu komórkowego gdy stan ustalony wielu procesów zostaje zachwiany. Stabilność procesów regulacyjnych ma kluczowe znaczenie dla prawidłowego funkcjonowania komórki a jej zaburzenia mogą nieraz prowadzić do bardzo groźnych chorób takich jak nieprawidłowe funkcjonowanie układu immunologicznego czy zmiany nowotworowe [8]. Przykładowo nadmierna stabilizacja transkryptów regulujących prawidłowy wzrost komórki takich jak c-fos, c-myc lub interleukina-3 prowadzi do transformacji nowotworowych, których jedną z cech jest niekontrolowany wzrost komórek [9-12].

3.4. Podstawowe czynniki odpowiedzialne za regulację ekspresji genów

3.4.1. Czynniki transkrypcyjne

Czynniki transkrypcyjne kontrolują transkrypcję genów znajdujących się w otoczeniu określonych motywów sekwencyjnych w DNA [13, 14] poprzez promowanie lub blokowanie przyłączenia polimerazy DNA lub innych białek wspomagających proces transkrypcji [15-17]. Są one kluczowe dla procesów ekspresji genów aktywując je w odpowiednim czasie i z określoną wydajnością w zależności od aktualnych potrzeb komórki i całego organizmu.

Czynniki transkrypcyjne przyłączają się do obszaru promotora genu znajdującego się kilkaset zasad przed miejscem startu transkrypcji (ang. transcription start site - TSS) lub zlokalizowanych w bardziej odległych obszarach, wewnątrz regionów niekodujących genu [18], intronów a nawet samej sekwencji kodującej [19]. W zależności od przyłączonego czynnika transkrypcyjnego ekspresja sąsiadującego genu może być albo wzmocniona albo wyciszona poprzez szereg najróżniejszych mechanizmów [20]. Mogą one stabilizować lub blokować proces przyłączenia polimerazy RNA, przyłączać enzymy z grupy acetylotransferaz/deacetylotransferaz odpowiedzialnych za interakcje DNA z histonami czyniąc go mniej lub bardziej dostępnym dla procesów transkrypcji [21] albo przyłączać białka pełniące funkcje koaktywatorów albo korepresorów do kompleksu DNA-czynnik transkrypcyjny [22].

Określone elementy mechanizmu oddziaływania z czynnikami transkrypcyjnymi mogą się różnić na przestrzeni organizmów, linii komórkowych a nawet wewnątrz pojedynczej komórki, dodatkowo tego typu czynniki regulacyjne nie przyłączają się jedynie do pojedynczej sekwencji nukleotydowej, ale do grupy sekwencji o zbliżonej budowie z różną siłą oddziaływania. Przykładowo większość genów kodujących struktury białkowe zawiera w obszarze promotora sekwencje TATA-box rozpoznawaną przez białko TBP. Sekwencja ta jest jednak niespecyficzna i pomimo, że najczęściej wiązany motywem jest TATATAA [23] to kompleks może zostać stworzony na podstawie sekwencji TATATATA, TATAAATA lub TATATAAA [24]. Utrudnia to analizę elementów regulacyjnych ponieważ nie istnieją uniwersalne kryteria pozwalające stwierdzić czy wariant określonej sekwencji jest odpowiedzialny za przyłączanie białek, czy tylko pojawia się z przyczyn losowych i nie ma wpływu na regulację procesu transkrypcji.

Czynniki transkrypcyjne są charakterystycznym elementem procesów regulacji ekspresji genów u wszystkich żywych organizmów w dodatku ich liczba jest proporcjonalna do rozmiaru całego genomu – większe genomy organizmów o większej złożoności zawierają więcej czynników transkrypcyjnych przypadających na jeden gen [25]. Szacuje się, że u człowieka istnieje ponad 2600 różnych białek

zawierających domeny pozwalające na przyłączanie DNA [26] oraz, że około 10% ludzkich genów koduje czynniki transkrypcyjne pozwalając w unikatowy sposób kontrolować ekspresję wszystkich genów znajdujących się w ludzkim genomie [27].

3.4.2. Białka typu RBP

Druga klasa mechanizmów, które odpowiadają za regulację procesów ekspresji genów oparta jest białka wyposażone w domeny pozwalające na tworzenie wiązań z cząsteczkami RNA (ang. RNA binding proteins - RBP). Białka te mogą zarówno stabilizować jak i destabilizować mRNA [6] poprzez mechanizm rozpoznawania określonych motywów sekwencyjnych. Większość z tego typu białek wiąże się z motywami w sekwencjach niekodujących, do których należą obszary 3' lub 5' UTR [28, 29], chociaż niektóre mogą wchodzić w interakcje także z sekwencją kodującą [30]. Przyłączenie białka z rodziny RBP może prowadzić do szybkiej degradacji mRNA [31] chociaż czasem jego stabilność może także zostać zwiększona [32], chroniąc transkrypt przed nadmierną temperaturą lub promieniowaniem [6]. Tego typu mechanizm reguluje głównie czas półtrwania mRNA kodujących cytokiny oraz inne białka biorące udział w procesach odpowiedzi na warunki stresu [33].

Jedną z najlepiej poznanych klas elementów sekwencji rozpoznawanych przez RBP są bogate w nukleotydy A i U sekwencje ARE (ang. AU-rich elements) zlokalizowane głównie w obszarze 3'-UTR. Motywy te mogą zarówno stabilizować jak i destabilizować transkrypt w zależności od przyłączonego białka. ARE zlokalizowane są głównie w transkryptach o niskiej stabilności, które pod wpływem oddziaływania stają się bardziej lub mniej stabilne [32]. Niektóre białka typu ARE takie jak HuR, TTP czy FRX1 mogą wpływać także bezpośrednio na proces translacji a nawet oddziaływać z funkcjonalnym RNA takim jak miRNA [34]. Gen c-Fos jest regulowany poprzez dwa rejony ARE [35], które są kluczowe dla jego stabilności wpływając na proces deadenylacji [36]. Delecje w tych obszarach mogą zwiększyć jego stabilność w ponadnaturalny sposób przetwarzając go w onkogen, co prowadzi do karcynogenezy [37].

Motywy typu ARE dzielą się na 3 klasy różniące się typem rozpoznawanej sekwencji oraz kinetyką degradacji mRNA [38]:

- *Klasa I to obszary bogate w nukleotydy A i U zawierające przynajmniej jeden pentamer AUUUA, są one charakterystyczne dla protoonkogenów takich jak c-myc i c-fos.*
- *Klasa II charakteryzuje się nakładającymi się powtórzeniami sekwencji (AUUU)nA, występują one głównie w transkryptach kodujących cytokiny*
- *Klasa III to najmniej specyficzne obszary, które nie zawierają żadnego określonego motywu jednak charakteryzują się obszarami bogatymi w nukleotydy A i U o długości około 13 zasad, ze względu na niską specyfikę można je znaleźć wśród genów każdej klasy*

W oparciu o doświadczalnie zidentyfikowane obszary ARE opracowano także kilka uniwersalnych motywów najczęściej rozpoznawanych przez białka z rodziny RBP, np.: WWWUAUUUAUWWW (gdzie W to U albo A) [39]. Motyw ten występuje u około 5% ludzkich genów, odpowiadając za ich przyspieszoną degradację.

Motywy typu ARE nie są jedynym regionem sekwencji, zdolnym do przyłączania białek, na przykład białko CP1 ma zdolność przyłączania się do obszarów 3'-UTRs, które są bogate w cytozynę zwiększając

stabilność transkryptu [40]. Znane są także białka, które oddziałują z powtórzeniami CNG (gdzie N = A, G, T, C), takie jak CUG-BP, ETR-3, CELF3 [41] lub DM1, które wiąże się z motywami sekwencyjnymi zbudowanymi z powtórzeń CUG [42].

Mechanizmy regulacyjne oparte o białka RBP są istotnym elementem odpowiedzi komórkowej na zewnętrzne bądź wewnętrzne czynniki wymuszające, na przykład przyłączanie białka KSRP aktywuje procesy rozpadu mRNA i jednocześnie wzmacnia ich przetwarzanie poprzez interakcje z prekursorami miRNA [43]. Z kolei przyłączanie RBP do mRNA kodującego inne elementy szlaków przekazywania informacji może stymulować jego rozpad albo stabilizację w odpowiedzi na czynniki środowiskowe takie jak na przykład niedobór surowicy niezbędnej do prawidłowego rozwoju komórek [34].

3.4.3. Interferencja RNA

Badania przeprowadzone na przestrzeni kilku ostatnich lat podkreślają znaczenie funkcjonalnych cząsteczek RNA, typu mikroRNA (miRNA) w modulowaniu poziomu mRNA poprzez albo zatrzymanie procesu translacji albo degradację mRNA poprzez trawienie endonukleazami. miRNA to małe niekodujące cząsteczki dwuniciowego RNA o długości około ~22 nukleotydów, które mogą hamować ekspresję określonych genów po-transkrypcyjnie. miRNA nie kodują struktury białek tak jak mRNA, wpływają one jednak na stężenie mRNA zawierających określone motywy w sekwencji nukleotydowej [44-47] typu MRE (z ang. miRNA Response Elements). Większość tego typu motywów położona jest w niekodującym obszarze końca 3' (3'-UTR), jednak w niektórych genach występują one także w sekwencji kodującej [48]. Wykorzystując komputerowe badania genomu mające na celu identyfikację motywów typu MRE pokazano, że w bardziej złożonych organizmach miRNA regulują setki różnych mRNA, co sugeruje, że znaczna część transkryptomu jest kontrolowana poprzez złożony system regulacyjny oparty o cząsteczki miRNA [48].

Cząsteczki miRNA są elementem kompleksu białkowego RISC (RNA Induced Silencing Complex), który poprzez białka z rodziny Ago prowadzi do degradacji lub zatrzymania translacji transkryptów o przynajmniej częściowej komplementarności do sekwencji miRNA. To czy mRNA zostanie zdegradowane, czy jedynie jego translacja zostanie zablokowana zależy od poziomu komplementarności pomiędzy mRNA i miRNA. Wysoka komplementarność sekwencji powoduje uruchomienie mechanizmu degradacji [49, 50], niska natomiast prowadzi jedynie do zatrzymania translacji [51-53], jednak w obu przypadkach ma to bezpośrednie przełożenie na wydajność procesu produkcji specyficznych białek w komórce [54]. Najsilniejsze wiązania pomiędzy miRNA i mRNA powstają poprzez idealne dopasowanie pozycji 2-8 miRNA nazywanych regionem *seed*. Dopasowanie tych siedmiu nukleotydów jest w większości przypadków warunkiem wystarczającym do powstania wiązania, chociaż dopasowanie na dodatkowych pozycjach może znacznie zwiększyć jego siłę [48, 55, 56].

Niektóre miRNA mają charakter onkogenów, podczas gdy inne wykazują aktywność supresorową, aberracje w ich budowie i poziomach ekspresji mogą zatem prowadzić do chorób nowotworowych [57]. Odkrycie to czyni je istotnym elementem procesu diagnostyki nowotworowej, co pokazuje, że oparte o miRNA terapie mogą mieć zastosowanie w klinice dzięki ich wysokiemu potencjałowi regulacyjnemu [58].

3.5. Znaczenie niekodującego obszaru końca 3' genu

Pomimo, że elementy regulacyjne znajdują się we wszystkich częściach mRNA, niekodujący obszar 3'-UTR jest dla nich najlepszym regionem sekwencji. Ten fragment nie jest odczytywany podczas procesu translacji umożliwiając przyłączanie się białek, co pozwala na kontrolowanie stabilności mRNA na różnych etapach jego przetwarzania [59]. Analizy bioinformatyczne pokazały, że obszar 3'-UTR jest średnio znacznie dłuższy niż 5'-UTR w genach większości kręgowców, co wskazuje na jego wysoki potencjał regulacyjny. Dodatkowo pokazano, że jego średnia długość wzrastała w toku ewolucji sugerując, iż może mieć on związek ze zwiększającą się złożonością organizmów [60].

Pomimo iż, informacje o budowie struktury białka przechowywane są poza obszarem 3'-UTR to mutacje w tym rejonie będące następstwem działania czynników genotoksycznych mogą być równie niebezpieczne jak w przypadku obszaru kodującego. Mimo, że budowa białka zostaje zachowana to zaburzenia mechanizmów regulacyjnych mogą być potencjalnym źródłem nieprawidłowości w funkcjonowaniu organizmu. Do powszechnie znanych chorób wywołanych zaburzeniami mechanizmów regulacji ekspresji genów należą różnego typu wady serca [61] chondrodysplazja [62], choroba Alzheimera [63], syndrom łamliwego chromosomu X [64] oraz nowotwory [65]. Najbardziej szkodliwe mutacje w tym obszarze obejmują przede wszystkim kodon terminacyjny wyznaczający koniec procesu translacji, sygnał poliadenylacyjny (motyw AATAAA) określający koniec procesu transkrypcji lub kształt struktury przestrzennej mRNA wpływający na oddziaływania z białkami regulatorowymi [66].

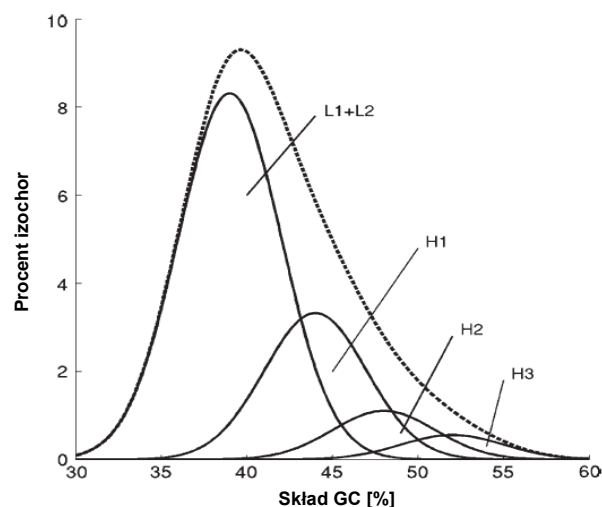
Długość końca 3'-UTR nie jest jednoznacznie określona i zmienia się pomiędzy transkryptami różnych genów a nawet pomiędzy różnymi transkryptami tego samego genu, w zależności od miejsca zatrzymania procesu transkrypcji oraz post-translacyjnej obróbce samego transkryptu. Ilość motywów sekwencyjnych jest proporcjonalna do długości sekwencji 3'-UTR jednak nie we wszystkich przypadkach gdyż ich liczba wystąpień może zależeć od dodatkowych czynników jak np. proporcji nukleotydów GC w sekwencji. Skład GC ma bardzo istotne znaczenie w przypadku częstotliwości występowania motywów regulatorowych. Motywy o określonych proporcjach nukleotydów GC występują znacznie częściej w regionach sekwencji o zbliżonym składzie ze względu na zwiększone prawdopodobieństwo ich losowego pojawiania się. Dotyczy to szczególnie motywów o niskim poziomie złożoności i specyficzności takich jak motywy ARE klasy III które zbudowane są z ciągu 13 nukleotydów A lub T.

Motywy ARE klasy III są przykładem na to, że nieraz fragment sekwencji nukleotydowej nie musi być jednoznacznie określony a warunkiem wystarczającym do przyłączenia się białka o określonej domenie rozpoznającej DNA lub RNA jest jedynie obszar o określonych proporcjach AT/GC. Obszary bogate w GC także mogą być elementem sekwencji rozpoznawanym przez białka regulatorowe. Przykładem takiego oddziaływania jest białko Sac7d, które u hipertermofilowych archeonów (organizmów odpornych na działanie wysokich temperatur) zwiększają stabilność DNA w skrajnie wysokich temperaturach przekraczających 100 °C poprzez przyłączanie się do obszarów bogatych w nukleotydy GC [67]. Wiele algorytmów predykcji miejsc wiązania miRNA pokazuje znamienne różnice w ich częstotliwości, w zależności od składu nukleotydowego badanych obszarów sekwencji [68, 69].

3.6. Rozkład elementów regulacyjnych w genomie

Skład GC genu jest bardzo silnie skorelowany ze składem GC obszaru genomu, w którym jest on położony, co dotyczy zarówno całej jego sekwencji, elementów niekodujących (5'/3'-UTR) jak i sekwencji kodującej (CDS) [70, 71]. Mimo to ludzkie geny są bardziej ekstremalne pod względem składu niż ich otoczenie (bardziej bogate w AT/GC) geny bogate w GC zwykle charakteryzują się znacznie większą częstotliwością nukleotydów GC niż ich sąsiadujące obszary [72].

Zawartość GC odgrywa istotną rolę w bardzo wielu procesach biologicznych a rozkład GC na przestrzeni sekwencji DNA wykazuje bardzo silne zróżnicowanie. Skład nukleotydowy jest podstawą jednej z najpopularniejszych teorii organizacji ludzkiego genomu nazywanej modelem izochorowym. Izochory to obszary genomu o długości ponad 300 kbp (ang. kilo base-pairs – kilo par zasad czyli 300 000 nukleotydów), które w sposób znamieny statystycznie różnią się pod względem składu GC od sąsiadujących obszarów genomu, należących do innych izochor [73]. Teoria ta wynika z obserwacji, wg. których genom jest mieszaniną obszarów o naprzemiennie wysokim i niskim średnim składzie GC podzielonym na pięć klas w zależności od procentowego składu GC. Izochory lekkie L1 i L2 charakteryzują się wysoką częstotliwością występowania nukleotydów AT, obszary H1, H2 i H3 są z kolei bogate w GC. Histogram fragmentów DNA o różnym składzie nukleotydowym oraz jego dekompozycja na poszczególne rozkłady Gaussa wyznaczające klasy izochor przedstawiono na Ryc. 5.



Ryc. 5: Histogram fragmentów ludzkiego DNA o różnym składzie nukleotydowym zdekomponowany na poszczególne rozkłady Gaussa odpowiadające klasom L1-2 oraz H1-3 izochor (źródło: [74])

Główną różnicą pomiędzy izochorami bogatymi i ubogimi w GC jest częstotliwość występowania w nich genów. Izochory o najwyższym składzie GC z grupy H3 zawierają średnio 20-razy więcej genów na 1Mbp (milion par zasad) niż izochory L1 i L2, pomimo, że stanowią najmniejszy procent całego genomu. Ze względu na duże zróżnicowanie ilości genów, obszary L1, L2 i H1 nazywane są potocznie pustą przestrzenią gdyż zawierają średnio jeden gen co 50-150kb w przeciwieństwie do H2+H3 gdzie geny rozmieszczone są co 5-15kb [75-77]. Pozostałe cechy izochor wynikające z różnic w składzie nukleotydowym obejmują:

- Izochory bogate w GC występują najczęściej w prążkach T chromosomów, które są bardziej odporne na denaturację termiczną [78-80]
- Czas trwania replikacji obszarów różniących się składem GC jest różny - najwcześniej replikowane izochory są krótkie i bogate w GC [81-83]
- Izochory różnią się pod względem obecności transpozonów (fragmentów DNA, które mogą się przemieszczać na inną pozycję w genomie). Transpozony typu SINE (z ang. Short Interspersed Nuclear Elements) najczęściej obecne są w obszarach bogatych w GC podczas gdy typ LINE (z ang. Long Interspersed Nuclear Elements) charakterystyczny jest dla obszarów ubogich w GC [70, 84-86]
- Obszary o wyższym składzie GC rekombinują znacznie częściej, częściej dochodzi w nich także do wymiany materiału genetycznego w wyniku czego powstają nowe genotypy [87, 88]

Skład nukleotydowy obszarów genomu jest bardzo silnie skorelowany ze składem nukleotydowym genu oraz wszystkimi jego składowymi elementami w tym introny, sekwencje kodujące i niekodujące obszary końca 3' i 5'. Skład nukleotydowy sekwencji kodującej może być kontrolowany przez wykorzystanie określonych kodonów (trójek nukleotydów kodujących pojedynczy aminokwas), które dzięki temu, że kod genetyczny jest zdegenerowany (kilka kodonów może kodować ten sam aminokwas) umożliwiają zmianę średniego składu nukleotydowego przy zachowaniu stałej struktury białka [89]. Skład nukleotydowy genów wynikający z położenia w izochorach o podobnym składzie jest powiązany z ich funkcją:

- Geny z izochorach bogatych w GC zawierają znacznie mniej intronów i kodują białka o krótszej sekwencji [90]
- Wysoki skład GC zwiększa stabilność struktury drugorzędowej RNA, co ma wpływ na jego okres półtrwania [91, 92]
- Skład GC części kodującej genu wynikający z wykorzystania określonych kodonów jest powiązany z jego poziomem ekspresji [93-96] oraz długością transkryptu [97-99]
- Częstotliwość mutacji jest większa w obszarach genów bogatych w nukleotydy GC [100, 101]

Pomimo wielu zidentyfikowanych cech izochor na przestrzeni ostatnich 25 lat teoria podziału genomu oparta o obszary o różnym składzie GC jest nadal kontrowersyjna [74]. Główną przyczyną jest to, iż niejasne są mechanizmy ich powstawania, tłumaczone silnymi ukierunkowanymi mutacjami na drodze ewolucji [100, 102, 103] zmianami wynikającymi z naturalnej selekcji [104] oraz zjawiskiem tzw. konwersji genów wynikającym z rekombinacji DNA obejmującego obszar genu [105, 106]. Brakuje niestety silnych dowodów na potwierdzenie tych hipotez. Wątpliwości wzbudzają także algorytmy identyfikacji izochor w genomie takie jak IsoFinder [107] czy GC-Profile [108]. Głównymi zarzutami są nieuzasadnione podejścia do ich wyznaczania (między innymi dotyczące arbitralnie przyjętej minimalnej długości fragmentów), które pomimo swojej znamienności statystycznej mogą być jedynie artefaktami przyjętej metodologii. Dodatkowo wciąż nie jest w pełni wyjaśnione dlaczego izochory występują wyłącznie u ciepłokrwistych kręgowców [73], kiedy powstały i z jakiego powodu wydają się powoli zanikać wraz z upływem czasu [109, 110].

3.7. Wpływ promieniowania jonizującego na komórkę

Promieniowaniem jonizującym określa się wszystkie rodzaje promieniowania, które wywołują jonizację ośrodka materialnego polegającą na oderwaniu się przynajmniej jednego elektronu od atomu lub cząsteczki, bądź też wybiciu go ze struktury krystalicznej. Powszechnie znanymi typami promieniowania jonizującego są promieniowanie rentgenowskie, strumienie cząstek alfa lub strumienie innych cząstek elementarnych, jakimi są elektrony, protony i neutrony.

Promieniowanie jonizujące indukuje szereg zmian w komórce powodując uszkodzenia większości jej elementów w tym materiału genetycznego, a także wpływając na przebieg najróżniejszych procesów chemicznych zachodzących na poziomie transkryptów bądź pojedynczych białek [111, 112]. Poza bezpośrednim oddziaływaniem na elementy struktury komórkowej działanie promieniowania może być także pośrednie poprzez indukcję wolnych rodników oraz reaktywnych form tlenu odpowiedzialnych za generowanie istotnych dla funkcjonowania komórki uszkodzeń w strukturze DNA [113]. Odpowiedź komórek na określoną dawkę promieniowania uzależniona jest w dużej mierze od charakteru procesów biochemicznych zachodzących w jej wnętrzu, przez co zachowanie komórek różnego typu, o różnych cechach morfologicznych, może być nieraz całkowicie odmienne. Odpowiedź komórek na promieniowanie może prowadzić do zwiększonej proliferacji, zatrzymania cyklu komórkowego lub uruchomienia kaskady sygnałowej prowadzącej do apoptozy, w przypadku zbyt rozległych uszkodzeń [114].

Dawka promieniowania jonizującego jest najczęściej wyrażana w grejach (Gy) zdefiniowanych, jako absorpcja jednego dżula energii przez jeden kilogram materii. Wpływ dawki promieniowania jonizującego, na liczbę uszkodzeń w DNA jest liniowy w szerokim zakresie wartości. Promieniowanie γ o dawce w wysokości 1 Gy prowadzi do około $5,8 \cdot 10^{-3}$ dwuniciowych pęknięć na 1Mbp (milion par zasad), co odpowiada około 40 pęknięciom przypadającym na jedną komórkę [129].

Promieniowanie indukuje szereg różnego typu uszkodzeń w DNA, spośród których najgroźniejsze są podwójnoniciowe pęknięcia, które blokują cykl komórkowy w jednym z punktów kontrolnych, takich jak koniec faz cyklu komórkowego G1, G2 lub S, dopóki poziom uszkodzeń nie zostanie obniżony [115]. Blokada cyklu komórkowego, która jest celem wielu terapii nowotworowych zatrzymuje procesy replikacji uszkodzonego DNA. Ten naturalny mechanizm jest bardzo często zaburzony w komórkach nowotworowych [116, 117] poprzez mutacje w genach odpowiedzialnych za procesy przekazywania informacji indukowane uszkodzeniami DNA, do których należą BRCA1, BRCA2, ATM, TP53, CDKN2A [118] NF- κ B [119] oraz PTEN [120].

Różnice w odpowiedzi komórkowej na promieniowanie jonizujące pomiędzy różnymi liniami komórkowymi wynikają głównie ze zmian w ścieżkach sygnałowych, w których opisane geny odgrywają istotną rolę. Jest to zjawisko obserwowane w badaniach klinicznych, wśród pacjentów, których nowotwory wykazują zwiększoną lub zmniejszoną radio-oporność [121]. Komórki mogą przeżyć ekspozycję na promieniowanie albo z powodu subletalnych dawek, które są ograniczane w związku z ryzykiem pojawiania się rozległych skutków ubocznych terapii lub też z powodu nieprawidłowo działających mechanizmów regulacji, które nawet w przypadku rozległych uszkodzeń DNA uniemożliwiają zainicjowanie procesów apoptozy.

Dodatkowym czynnikiem mającym wpływ na przeżywalność komórek pod wpływem promieniowania są odpowiednie warunki zachodzące w ich wnętrzu. Jednym z przykładów jest hipoksja

(niedobór tlenu), która obniża toksyczność promieniowania, co może mieć zarówno pozytywne jak i negatywne działanie np. w sytuacji, gdy hipoksja wywoływana jest przez samo promieniowanie może zmniejszać skuteczność radioterapii [122]. Duża zawartość tlenu w komórce może prowadzić do zwiększonych uszkodzeń DNA poprzez indukcję wolnych rodników pod wpływem promieniowania, które mogą być źródłem większych uszkodzeń niż te, które są skutkiem bezpośredniego wpływu promieniowania [113].

Pod wpływem promieniowania poziom wielu transkryptów zmienia się niezwykle szybko [111, 112], co pozwala na zainicjowanie mechanizmów odpowiedzi komórkowej na warunki stresu tuż po napromieniowaniu [123]. Zmiany równowagi pomiędzy szybkością degradacji mRNA a wydajnością procesów jego produkcji są jednym z kluczowych elementów regulacji poziomu ekspresji większości ludzkich genów. Ponadto zaburzenia tego typu mechanizmu leżą u podstaw znacznej części chorób genetycznych w tym nowotworów [124]. Oddziaływania poszczególnych elementów wewnątrzkomórkowych szlaków sygnałowych są także celem terapii łączących promieniowanie z odpowiednimi molekułami wpływającymi na ściśle określone elementy mechanizmu odpowiedzi na promieniowanie [125, 126].

Poznanie wpływu promieniowania jonizującego na mechanizmy stabilizacji poziomu ekspresji genów stanowi jeden z kluczowych elementów na drodze do poznania jego wpływu na komórki mające styczność z promieniowaniem nie tylko w wysokich dawkach, jakie stosowane są w medycynie, ale także tych bardzo niskich, na które organizmy żywe są stale narażone. Promieniowanie jonizujące jest wszechobecne w środowisku człowieka, ze względu na obecność w przyrodzie radioizotopów różnych pierwiastków oraz promieniowania kosmicznego. Jest ono uważane za główny czynnik mutacji w genach, niewielkich zmian w sekwencjach nukleotydowych, które po skumulowaniu odpowiadają za ewolucje organizmów. Może ono zatem mieć zarówno korzystny jak i negatywny wpływ na organizmy.

3.8. Wielkoskalowe metody analizy funkcji genów

Określenie zmian, jakie następują w mechanizmach regulacji ekspresji genów wymaga znajomości funkcji genów oraz zależności regulacyjnych pomiędzy nimi. Pozwala to określić jakie elementy mechanizmu regulacji mogą wpływać na zmianę ekspresji specyficznego genu w danych warunkach. Liczne modele szlaków sygnałowych są w stanie dostarczyć tego typu informacji mimo, iż w większości przypadków są jedynie silnie uproszczonym opisem zjawisk zachodzących w komórce. Do najpopularniejszych baz danych opisujących ścieżki sygnałowe należą KEGG-Pathway [127], Panther [128] i Reactome [129] będące zbiorem setek szlaków sygnałowych opracowanych przez indywidualnych użytkowników lub wprowadzonych przez pracowników instytucji odpowiedzialnych za określoną bazę danych. Wszystkie wprowadzone ścieżki podlegają procesowi recenzji co pozwala na zachowanie wysokiego poziomu jakości danych, jednak mimo to informacje o tych samych ścieżkach w różnych bazach w istotny sposób różnią się od siebie.

Jedną z najbardziej popularnych baz danych informacji o funkcji poszczególnych genów jest Gene Ontology (GO) [130]. W odróżnieniu od zbiorów ścieżek sygnałowych GO jest zestawem definicji przynależności poszczególnych genów do jednej z 3 grup funkcyjnych:

- komponenty komórkowe (ang. cellular component) – geny uczestniczące w procesach powstawania komórki oraz środowiska pozakomórkowego
- funkcje molekularne (ang. molecular function) – geny uczestniczące w procesach oddziaływań pomiędzy molekułami (np. enzymy)
- procesy biologiczne (ang. biological process) – geny uczestniczące w procesach odpowiedzialnych za prawidłowe funkcjonowanie procesów zachodzących w komórce

Każda z grup podzielona jest dodatkowo na ponad 10 podkategorii o różnym poziomie złożoności, do których przypisane są konkretne geny w zależności od tego jak dokładnie określona jest ich funkcja.

W literaturze istnieje bardzo wiele narzędzi ułatwiających procesy przetwarzania informacji o funkcji genów. Poza tradycyjnym przypisaniem właściwości do konkretnego genu bardzo popularne są testy na nadreprezentację określonych terminów w zdefiniowanej grupie genów np. genów zidentyfikowanych w określonym eksperymencie biologicznym. Różne wersje testów opartych o nadreprezentację terminów są podstawą działania takich aplikacji jak GStats [131] FatiGO [132] Ontologizer [133] czy Genelist Analyzer [134]. Każde z narzędzi dostępne jest za pośrednictwem strony internetowej, co pozwala na wyeliminowanie problemów z dostępnością narzędzia w przypadku różnego typu systemów operacyjnych, ich wspólną wadą jest jednak to, że bazują one na zewnętrznych bazach danych i nieraz są daleko w tyle za zmianami, jakie w tych bazach następują. Dodatkowo narzędzia tego typu skoncentrowane są głównie na analizie genów zidentyfikowanych na podstawie zmian poziomów ekspresji. Stwarza to potrzebę przanalizowania zmian profilu ekspresji dużych ilości genów w celu zachowania wysokiego poziomu znaczącości statystycznej, co umożliwia powszechnie stosowana technika badań oparta o mikromacierze oligonukleotydowe.

3.9. Mikromacierze oligonukleotydowe

Mikromacierze oligonukleotydowe są jedną z najczęściej stosowanych metod charakteryzowania zmian w profilu ekspresji genów wywołanych różnymi czynnikami fizycznymi lub chemicznymi a także różnic w profilu ekspresji pomiędzy komórkami wyizolowanymi z różnych typów tkanek lub od różnych pacjentów [135-137].

Mikromacierze posiadają szeroki zakres możliwości pozwalających między innymi na identyfikację genów o potencjalnie istotnej roli w odpowiedzi komórkowej na analizowane czynniki fizyczno-chemiczne lub zmian w profilach ekspresji genów charakterystycznych dla określonego stadium choroby. Mikromacierze wymagają jednak skomplikowanych metod statystycznych w celu odróżnienia zmian wywołanych przez analizowane czynniki doświadczalne od tych, które pochodzą od specyfiki metody badawczej i niedokładności samego pomiaru. Problem odpowiedniej analizy statystycznej jest ogromnym wyzwaniem i stanowi przedmiot bardzo wielu podręczników oraz artykułów naukowych, jednak pomimo faktu, że mikromacierze są używane od ponad dziesięciu lat wiele problemów związanych z analizą danych wciąż pozostaje nierozwiązane.

Najczęściej poruszane w literaturze problemy analizy danych dotyczą algorytmów normalizacji [138, 139], których celem jest wyeliminowanie różnic, pomiędzy analizowanymi próbkami, wynikających z aspektów technicznych nie związanych z analizowanymi różnicami biologicznymi. Podobne

zastosowanie mają techniki usuwania różnic technicznych pomiędzy grupami próbek (z ang. batch-effect) po to aby możliwe było porównywanie danych pochodzących z eksperymentów wykonanych w różnym czasie i w różnych laboratoriach [140]. Pozostałe z często poruszanych problemów obejmują proces identyfikacji genów różnicujących próbki [141, 142], ocenę poziomu szumu w eksperymencie co pozwala zidentyfikować geny, które ulegają ekspresji w danych warunkach [143], oraz identyfikację uszkodzeń i zanieczyszczeń na powierzchni mikromacierzy [144, 145].

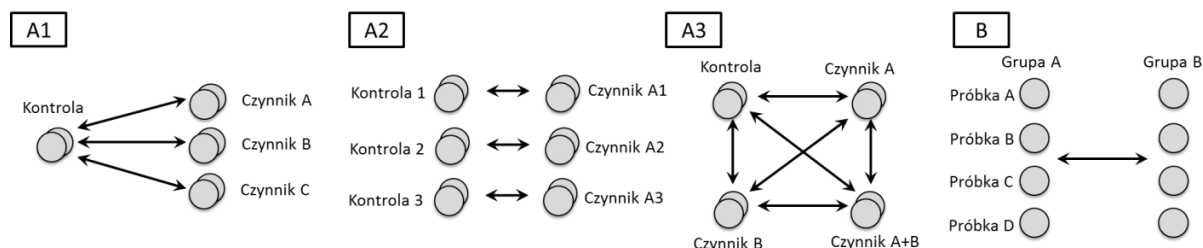
Najczęściej wykorzystywane mikromacierze firmy Affymetrix charakteryzują się dodatkowymi cechami wpływającymi na pomiar końcowy wynikającymi z ich specyficznej budowy. Obejmują one problemy związane z występowaniem kilku pomiarów poziomu ekspresji dla pojedynczego genu [146, 147], niewłaściwym przypisaniem sond do genów lub ich transkryptów [148, 149], oceną poziomu tła oraz niespecyficjnej hybrydyzacji sond [150] a także wpływu indywidualnych cech sond na algorytmy przetwarzania danych [151].

Do niewątpliwych wad technologii mikromacierzy należą m.in. wysoki koszt pojedynczego eksperymentu, duża liczba sond zaprojektowanych w oparciu o sekwencje nukleotydowe wyznaczone z niską dokładnością oraz w przypadku większości mikromacierzy brak wpływu na pule badanych transkryptów, gdyż zwykle na macierzy do dyspozycji jest jedynie zbiór sond jakie zaprojektowane zostały przez producenta. Dokładność eksperymentu także jest relatywnie niska, ponieważ pomiar jest bardzo wrażliwy na wiele czynników min. temperaturę, technologię pozyskiwania materiału genetycznego oraz jego czystość, co może mieć silny wpływ na dokładność oszacowania poziomów ekspresji genów. Niewielka różnica w stanie fizjologicznym badanych komórek lub środowisku, w jakim przeprowadzana jest reakcja hybrydyzacji a także zanieczyszczenia obecne w materiale genetycznym mogą prowadzić do spadku czułości eksperymentu poprzez niespecyficzną hybrydyzację z sondami mikromacierzy. Prowadzi to do zmniejszenia stosunku siły sygnału do szumu pomiarowego (z ang. Signal-to-Noise Ratio - SNR), czego wynikiem mogą być błędne pomiary poziomu ilości transkryptu szczególnie w przedziale niskich wartości sygnału.

Mikromacierze dostarczają ogromnych ilości informacji na temat poziomu ekspresji tysięcy genów jednak z powodu ich niskiej dokładności są one wykorzystywane jedynie w celu zidentyfikowania potencjalnie istotnych genów. W kolejnym etapie analizy poziom ekspresji wybranych genów jest oznaczany za pomocą bardziej dokładnych metod, takich jak np. reakcja łańcuchowa polimerazy DNA z analizą ilości produktu w czasie rzeczywistym (ang. real-time PCR), które z kolei nie nadają się do przeprowadzenia wielkoskalowych analiz obejmujących tysiące genów.

3.9.1. Typowe schematy eksperymentu mikromacierzowego

Mikromacierze oligonukleotydowe wykorzystywane są do badania zmian ekspresji tysięcy genów jednocześnie jednak pojedyncza wartość poziomu ekspresji określonego genu jest ze względu na arbitralne jednostki sygnału nieinformatywna. Użyteczność eksperymentu mikromacierzowego jest zatem uzależniona od możliwości porównywania wartości poziomu ekspresji genu pomiędzy różnymi mikromacierzami. Czynniki specyficzne dla danego eksperymentu, takie jak warunki, w jakich przeprowadzane były reakcje chemiczne, ograniczają możliwości dodawania kolejnych próbek w późniejszym czasie, z tego względu o jakości danych decyduje dobrze zaprojektowany eksperyment.



Ryc. 6: Podstawowe schematy eksperymentu z wykorzystaniem mikromacierzy oligonukleotydowych

Typowe schematy eksperymentu mikromacierzowego można podzielić na dwie zasadnicze grupy:

- Analiza wpływu czynników fizycznych bądź chemicznych na zmiany poziomu ekspresji – ten schemat zwykle charakteryzuje się niewielką liczbą próbek, z których część poddana została działaniu określonego czynnika lub grupy różnych czynników. Analiza porównawcza wykonywana jest względem oddzielnej grupy próbek kontrolnych, które zwykle stanowią komórki nie poddane działaniu żadnego czynnika lub czynnika wykorzystywanego jako punkt odniesienia np. komórki poddane działaniu różnych związków chemicznych w różnych dawkach względem komórek, które nie miały z nimi kontaktu (A1), komórki traktowane promieniowaniem i analizowane po różnym czasie łącznie z wykonywaniem dodatkowych próbek kontrolnych dla każdego punktu czasowego (A2), komórki traktowane np. specyficznym siRNA lub innym związkiem chemicznym analizowane razem lub oddzielnie, względem siebie oraz kontroli (A3)
- Poszukiwanie specyficznych cech profilu ekspresji genów charakterystycznych dla określonej grupy komórek – ten schemat charakteryzuje się zwykle bardzo dużą liczbą próbek jednak pozbawioną powtórzeń technicznych czy biologicznych. W typowym eksperymencie analizowane są komórki wyizolowane od dwóch grup pacjentów np. grupy chorej na określony typ nowotworu i zdrowej służącej jako kontrola. Tego typu dane zwykle służą do poszukiwania sygnatur genowych, za pomocą metod uczenia maszynowego, które charakteryzują określoną grupę.

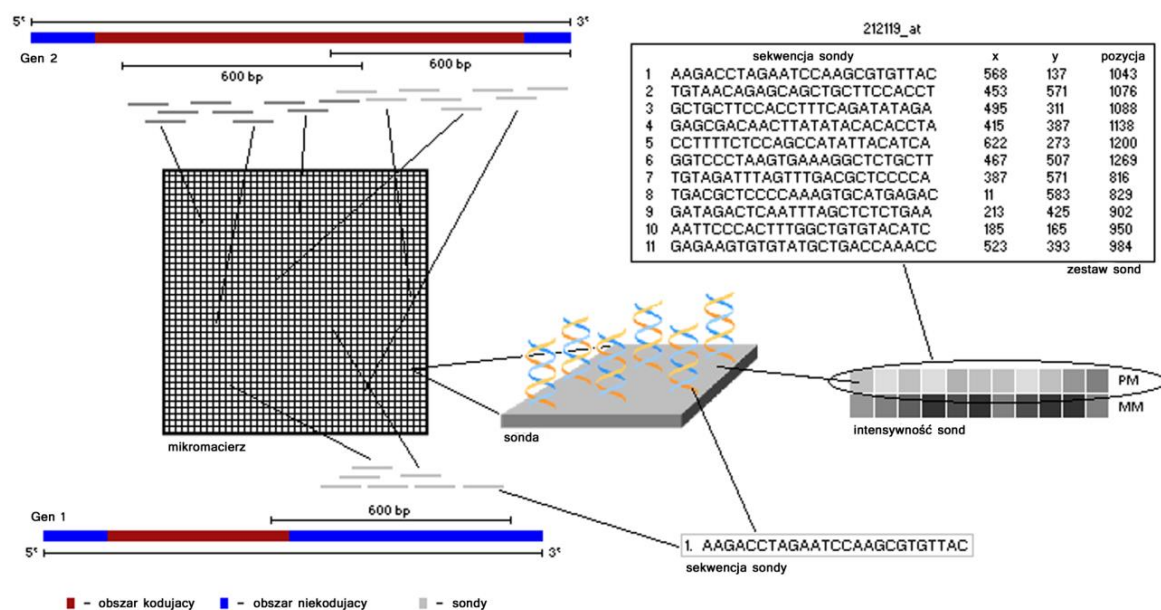
Pojedynczy, nawet najprostszy eksperyment mikromacierzowy dostarcza ogromnych ilości danych, które są niezwykle trudne w interpretacji, przez to rzadko analiza danych wykorzystuje ich pełen potencjał skupiając się jedynie na wybranym problemie. Z tego względu wyniki bardzo często są publikowane w bazach danych takich jak ArrayExpress [152] czy Gene Expression Omnibus [153] w formie zarówno przetworzonej jak i „surowej” nie poddanej żadnej wstępnej analizie, po to aby możliwe było przeanalizowanie tych samych danych przez inny zespół skupiając się na odmiennym problemie.

3.9.2. Budowa mikromacierzy

Mikromacierze zbudowane są z oligonukleotydów o długości kilkudziesięciu par zasad przytwierdzonych do powierzchni szklanej płytki. Technologia ich produkcji jest bardzo zbliżona do produkcji nowoczesnych procesorów wykorzystywanych w komputerach osobistych. Dzięki zastosowaniu odpowiednich masek fotolitograficznych możliwe jest przyłączanie pojedynczych nukleotydów określonego rodzaju w obszarach macierzy oświetlonych światłem UV. Pozwala to na zbudowanie mikromacierzy zawierającej setki tysięcy różnych oligonukleotydów specyficznych dla różnych fragmentów DNA lub RNA, zgrupowanych w punktach macierzy nazywanych sondami [154].

Podstawą działania mikromacierzy jest zasada komplementarności par Watsona-Cricka pomiędzy badanym RNA/DNA a kilkudziesięcio-nukleotydową sekwencją sondy. Materiał genetyczny wprowadzony na powierzchnie mikromacierzy hybryduje z oligonukleotydami odpowiednich sond, które umieszczone są w wielu egzemplarzach na określonych polach mikromacierzy. Miarą ilości zhybryzowanych transkryptów jest intensywność fluorescencji wyznakowanych nukleotydów, wzbudzona za pomocą lasera, która jest proporcjonalna do ilości cząsteczek mRNA określonego genu w badanej próbce. Podejście to pozwala na zmierzenie poziomu ilości transkryptu (ekspresji) tysięcy genów w relatywnie krótkim czasie, co dostarcza bezcennych informacji na temat procesów zachodzących w komórkach.

Najczęściej wykorzystywane mikromacierze Affymetrix typu 3'IVT (np. model HG-U133A) zbudowane są z zestawów 11 sond (PM - Perfect-Match) składających się z krótkich 25 nukleotydowych sekwencji, wybranych z fragmentu 600 nukleotydów, położonego blisko końca 3' transkryptu (Ryc. 7). Każdej sondzie typu PM odpowiada położona tuż obok sonda MM (Mismatch) o identycznej sekwencji, za wyjątkiem środkowego 13 nukleotydu, który jest zastąpiony nukleotydem do niego komplementarnym. Celem sond MM jest ocena stopnia niespecyficznej hybrydacji, która może zachodzić w sytuacji gdy sonda różni się pojedynczymi nukleotydami od określonego transkryptu [155].



Ryc. 7: Budowa mikromacierzy oligonukleotydowej

Najnowsza generacja mikromacierzy firmy Affymetrix (jak HuGene 1.0ST) składa się z podobnych sond typu PM jednak dopasowanych do poszczególnych eksonów danego transkryptu a nie jedynie do części niekodującej końca 3'. Sondy typu MM zostały w niej zastąpione zestawem ~1000 sond mierzących poziom intensywności tła (ang. Background Intensity Probes - BGP), o sekwencjach, które nie są komplementarne do żadnego ludzkiego genu. Sondy te różnią się proporcjami nukleotydów GC ze względu na to, że niespecyficzna hybrydacja zachodzi z różną wydajnością w zależności od składu nukleotydowego sondy. Podejście to pozwala na dokładniejszą ocenę stopnia niespecyficznej hybrydacji, która w przypadku starszych sond MM nie była dostatecznie wiarygodna, ze względu na to, że w przypadku sporej liczby genów sygnał niespecyficznych sond MM przewyższał sygnały sond PM [156]. Dodatkowo wykorzystanie mniejszej liczby sond służących do oceny

niespecyficznego hybrydyzacji, która w przypadku starszej generacji macierzy wymagała połowy jej powierzchni, pozwala na umieszczenie większej ilości sond typu PM. Zestawy sond w macierzach nowej generacji mają 2 poziomy – eksonu i genu. Eksonowe zestawy składają się średnio z 4 sond dopasowanych do konkretnego eksonu. Zestawy genowe z kolei łączą je w klastry (średnio 25 sond), które są specyficzne dla pojedynczego genu. Podejście to pozwala na określenie poziomu ekspresji różnych form splicingowych - transkryptów z alternatywnie wyciętymi eksonami lub różnym miejscem poliadenylacji.

Inną bardzo popularną platformą są mikromacierze firmy Agilent zbudowane w oparciu o technologię SurePrint, która wykorzystuje znacznie dłuższe w porównaniu do mikromacierzy Affymetrix sondy, zbudowane z 60 nukleotydów. Sond jest jednak znacznie mniej, średnio 8 na gen w przypadku najdroższych macierzy eksonowych (2x400k) lub 2 na gen w przypadku najtańszych (8x60k) zbudowanych zaledwie z 60 tysięcy sond. Podejście firmy Agilent ma zarówno swoje wady jak i zalety. Wykorzystanie dłuższych sond zwiększa ich specyficzność, jednak ich niewielka ilość na gen sprawia, że stają się one bardziej wrażliwe na pojedynczo-nukleotydowe polimorfizmy (z ang. Single Nucleotide Polymorphism - SNP) wynikające z cech badanego materiału lub pojawiające się w wyniku błędów amplifikacji materiału genetycznego, które w przypadku mikromacierzy Affymetrix wpłyną wyłącznie na sygnał pojedynczej sondy w zestawie. Pojedyncza mutacja nie zablokuje hybrydyzacji ale znacznie zmniejszy jej wydajność co w trakcie analizy genów różnicujących może być postrzegane jako obniżenie ekspresji.

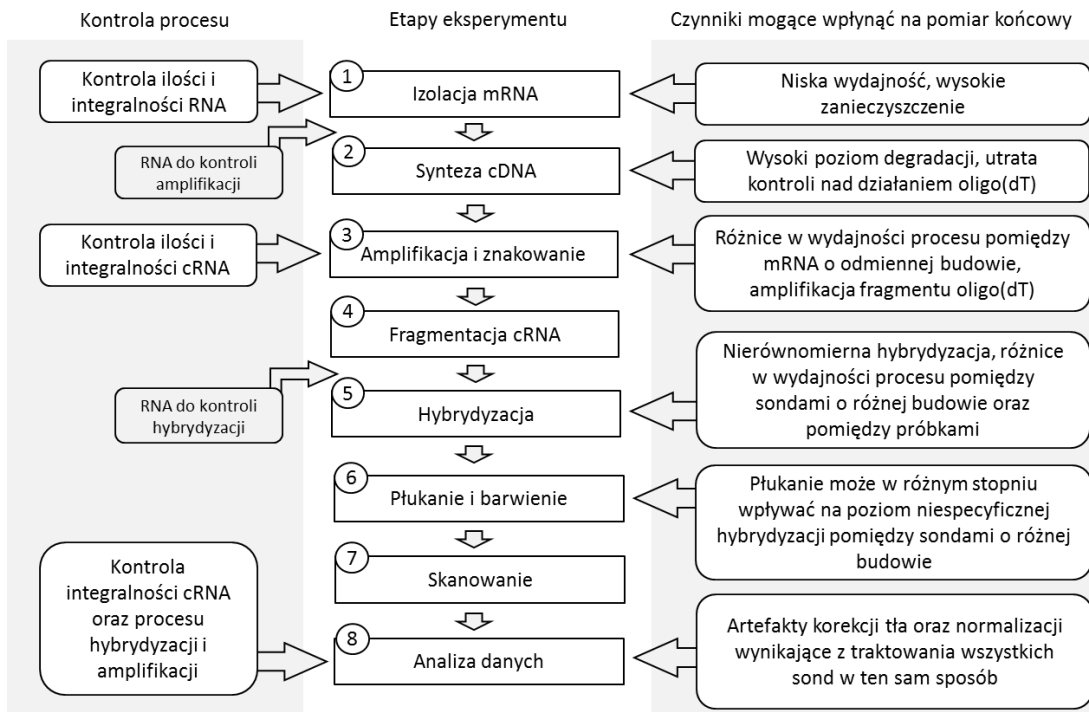
W przypadku mikromacierzy Affymetrix jedna źle zaprojektowana sonda w zestawie (np. stworzona w oparciu o niedokładne informacje w bazach danych sekwencji) pozwala na wyeliminowanie jej z analizy bez znacznych strat dla pomiaru ekspresji określanego genu, co w przypadku mikromacierzy Agilent wiąże się ze znacznie większymi konsekwencjami. Wykorzystanie pojedynczych sond pozwala jednak na uzyskanie bardziej stabilnego pomiaru niezależnego od różnic wynikających z budowy krótkich sond na mikromacierzach Affymetrix, wpływających na silną wariację w ramach zestawu. Różnice w wydajności hybrydyzacji, wynikające przykładowo z różnego składu GC, mają znacznie silniejszy wpływ na wyniki w przypadku 25 nukleotydowych sond, które bardzo ciężko dobrać w taki sposób aby różnice w proporcjach GC były minimalne, w przeciwieństwie do 60 nukleotydowych.

Wykorzystanie wielu sond na zestaw pozwala jednak obniżyć wpływ nierównomiernej hybrydyzacji na pomiary końcowe. Sondy Affymetrix rozmieszczone są w losowych punktach mikromacierzy, z tego względu zanieczyszczenia albo spadek wydajności hybrydyzacji określonego obszaru, wynikający np. z nierównomiernie rozprowadzonego materiału genetycznego, w mniejszym stopniu przekłada się na zmianę poziomu sygnału zestawu sond na przestrzeni różnych próbek, w przeciwieństwie do technologii pojedynczych sond firmy Agilent.

3.9.3. Podstawy biologiczne eksperymentu mikromacierzowego

Eksperyment przeprowadzony z wykorzystaniem mikromacierzy jest wieloetapowym procesem a dokładność wykonania każdego z poszczególnych etapów, może dramatycznie wpływać na wyniki końcowe. Dokładne zrozumienie procedury jest istotne nie tylko z punktu widzenia osoby odpowiedzialnej za wykonanie pracy laboratoryjnej, ale także podczas analizy danych pochodzących z eksperymentu.

W celu uniknięcia błędów powstałych podczas przeprowadzania eksperymentu dokładność i stan materiału biologicznego jest kontrolowana na kilku etapach, spośród, których najbardziej istotne przedstawiono na Ryc. 8.

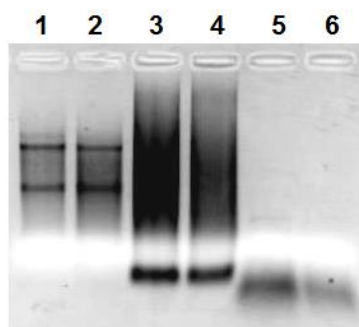


Ryc. 8: Uproszczony schemat przebiegu eksperymentu mikromacierzowego wraz z podstawowymi punktami kontrolnymi oraz kluczowymi czynnikami mogącymi wpłynąć na pomiar końcowy

I etap - Izolacja RNA: Eksperyment rozpoczyna izolacja RNA z komórek, którego stężenie i stopień degradacji analizowane są za pomocą spektrofotometru typu NanoDrop (analiza ilościowa) oraz bioanalyzera (analiza jakościowa). W RNA o wysokiej jakości rybosomalne RNA (rRNA) stanowi ponad 80% całkowitej jego puli mimo to nie jest ono przedmiotem analizy mikromacierzowej. Rybosomalne RNA jest bardzo dobrze widoczne na żelu agarozowym uzyskanym podczas elektroforezy gdzie, wyznakowane bromkiem etydydy, cząsteczki RNA rozdzielane pod względem ich długości. Dwie pierwsze ścieżki na Ryc. 9 pokazują wyizolowane RNA z dwoma wyraźnymi prążkami odpowiadającymi podjednostkom 18S i 28S rybosomów. Słabo widoczne prążki RNA rybosomalnego w tym badaniu świadczyłyby o wysokim poziomie degradacji RNA [157].

II etap – Synteza cDNA: Do wyizolowanego RNA dodawane jest RNA bakteryjne (tzw. polyA spike), które służy do oceny wydajności procesu syntezy cDNA gdyż odwrotna transkrypcja dołożonego RNA bakteryjnego przebiegać będzie niezależnie od stanu i objętości materiału wyjściowego. cDNA syntetyzowane jest dzięki tzw. starterom oligo-dT, które przyłączają się do ogona polyA (znajdującego się na końcu 3' mRNA) inicjując proces syntezy nici komplementarnej do mRNA dzięki zjawisku odwrotnej transkrypcji (Ryc. 10). Proces ten nie obejmuje rybosomalnego RNA gdyż w przeciwieństwie do mRNA nie zawiera ono ogona poliA. Dzięki temu nie jest konieczne oczyszczanie puli RNA z RNA rybosomalnego. Druga nić cDNA powstaje na matrycy pierwszej dzięki zhybrydyzowanemu do niej mRNA. Dodanie do roztworu rybonukleazy powoduje degradację mRNA w wielu niespecyficjnych miejscach pozostawiając

krótkie jego fragmenty przyłączone do cDNA (Ryc. 11). Te fragmenty służą za startery dla polimerazy, która syntetyzuje drugą nić cDNA z wielu miejsc jednocześnie usuwając napotkane na swojej drodze fragmenty RNA.



Ryc. 9: Analiza elektroforetyczna produktów różnych etapów eksperymentu mikromacierzowego 1 i 2 – wyizolowane RNA, 3 i 4 – oczyszczone cRNA, 5 i 6 – pocięte cRNA (Herok R., nieopublikowane dane)

Bardzo silny wpływ na ten etap ma stopień degradacji RNA powodując powstawanie skróconych cząsteczek cDNA (od strony końca 5') w przypadku gdy użyta matryca mRNA była zdegradowana [158]. W takiej sytuacji sondy położone w większej odległości od końca 3' mogą mieć niższy sygnał. Z tego powodu wszystkie sondy, w przypadku macierzy typu 3'IVT, wybierane są z obszaru położonego możliwie najbliżej końca 3' transkryptu. Degradacja w niewielkiej odległości od końca 3' całkowicie zatrzymuje proces syntezy cDNA ponieważ uniemożliwia to przyłączenie się startera oligo-dT i powielenie uszkodzonej cząsteczki (ze względu na brak ogona poli-A) nie wpływa jednak na różnice w poziomach sygnału między sondami pojedynczego transkryptu.



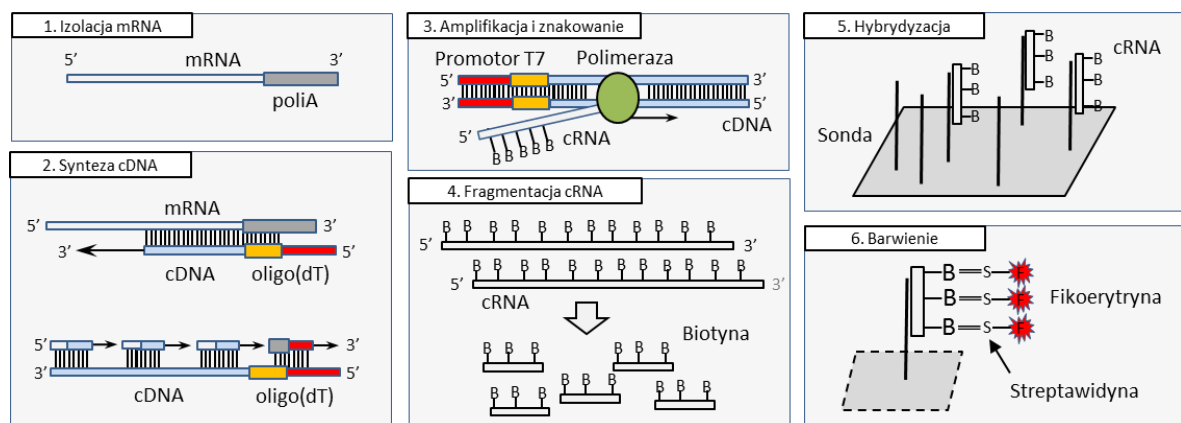
Ryc. 10: Synteza cDNA oparta o komercyjny starter T7-Oligo(dT) strzałka oznacza kierunek syntezy, czerwona czcionka oznacza sekwencje promotora wykorzystywaną w procesie amplifikacji

W założeniu starter oligo-dT powinien łączyć się z długim motywem (A)_n (ciąg n-nukleotydów adeniny), który jest częścią ogona poli-A transkryptu. Warunkiem koniecznym zainicjowania syntezy jest to aby koniec 3' startera był zhybrydyzowany z mRNA (tak jak na Ryc. 10). Pozycja przyłączania startera na jest losowa jednak jego ma bardzo istotne znaczenie, ponieważ niezhybrydyzowany koniec 3' oligo(dT) (wynikający z braku nukleotydów A na przeciwległej nici) uniemożliwia zainicjowanie procesu syntezy cDNA. Dodatkowo wraz ze wzrostem stężenia oligo-dT znacznie zwiększa się szansa na związanie więcej niż jednego startera [159]. W takiej sytuacji prawidłowa synteza może zajść tylko od startera położonego najbliżej końca 5' transkryptu, ponieważ w przypadku pozostałych synteza będzie zablokowana przez dodatkowe startery zhybrydyzowane bliżej końca 5' mRNA. Dodatkowo motywy znajdujące się wewnątrz transkryptu zawierające przynajmniej 8 nukleotydów A także mogą przyłączać do siebie starter oligo(dT) [159]. Wewnętrzne motywy (A)_n mogą inicjować proces syntezy cDNA, które w takiej sytuacji będą pozbawione fragmentu sekwencji pomiędzy motywem (A)_n a końcem 3' transkryptu. W praktyce jednak,

w związku z tym, że motywy te są znacznie krótsze niż kilkuset-nukleotydowy ogon poli-A synteza następuje bardzo rzadko gdyż szansa na przyłączenie startera w taki sposób, że jego koniec 3' będzie przylegał do transkryptu jest relatywnie niska. Konsekwencją tego jest powstawanie dużej liczby cząsteczek cDNA o sekwencji rozpoczynającej się w prawidłowym miejscu (na końcu 3' mRNA) jednak obciętych w miejscu pojawienia się wewnętrznego motywu (A)_n w związku z zablokowaną syntezą [159]. Ten etap procesu może mieć zatem kluczowe znaczenie dla wyników analizy w przypadku genów zawierających w swojej sekwencji długie motywy (A)_n.

III etap – Amplifikacja i znakowanie: Ten etap eksperymentu mikromacierzowego polega na powieleniu (amplifikowaniu) zsyntetyzowanego cDNA w procesie transkrypcji *in vitro*, której celem jest wyprodukowanie dużej ilości wyznakowanego biotyną cRNA (komplementarny RNA). Do tego celu wykorzystywany jest fragment startera oligo(dT) zaznaczony na Ryc. 10 czerwonym kolorem, który jest połączony z promotorem polimerazy bakteriofaga T7. Proces transkrypcji przebiega przy udziale wyznakowanych biotyną nukleotydów C i U [157].

Wydajność tego procesu ma duży wpływ na wyniki końcowe [160], z tego względu jest on kontrolowany przez RNA referencyjne dodane na pierwszym etapie procedury oraz niezależny pomiar wykonany za pomocą spektrofotometru i bioanalyzera, których celem jest określenie stężenia oraz integralności uzyskanego cRNA. Produkt uzyskany w tym etapie przedstawiony jest na pozycji 3 i 4 analizy elektroforetycznej (Ryc. 9). W tym przypadku rybosomalne RNA nie jest już widoczne (ponieważ jego matryca cDNA nie została zsyntetyzowana w związku z brakiem ogona poli-A a brak wyraźnych prążków świadczy o tym, że preparat składa się z wielu fragmentów cRNA o różnych długościach uzależnionych od długości wyizolowanych cząsteczek mRNA.



Ryc. 11: Przebieg eksperymentu mikromacierzowego z wyróżnieniem etapów syntezy cDNA i cRNA oraz procesu hybrydyzacji i barwienia cRNA

Produkty końcowe amplifikacji nie zawierają już sekwencji promotora T7 jednak pozostałością tego procesu jest krótki odcinek sekwencji startera oligo(dT) położony pomiędzy promotorem T7 a motywem (T)₂₄ zaznaczony na Ryc. 10 zielonym kolorem. Ponieważ ten fragment sekwencji znajduje się w każdej powstałej cząsteczce cRNA jego ilość jest bardzo duża co sprawia, że może on oddziaływać z sondami zawierającymi podobny fragment sekwencji, znacząco wpływając na ich poziom sygnału [161]. Uważa się, że proces amplifikacji może być źródłem wielu niedokładności, ze względu na to, że zachodzi on z różną

wydajnością w zależności od budowy transkryptów oraz ich potencjału do tworzenia struktur drugorzędowych [162], będąc motywacją do prowadzenia prac w dziedzinie wykorzystania nieamplifikowanego cDNA [163].

IV etap - Fragmentacja cRNA: Uzyskane cRNA jest cięte na krótkie fragmenty o długości 50-100nt. Ścieżki 5 i 6 na Ryc. 9 pokazują wynik analizy elektroforetycznej pociętego na fragmenty cRNA. Prążki w dolnej części wykresu wskazują na dużą liczbę krótkich odcinków RNA informując o prawidłowo przeprowadzonym etapie fragmentacji. Po tym etapie do preparatu dodawana jest kolejna mieszanina czterech cRNA bakterii *E.Coli* (tzw. bacterial spike) obecnych w mieszaninie w różnych stężeniach. Materiał ten służy jako kontrola wydajności następnego etapu jakim jest hybrydyzacja wyznakowanego cRNA do mikromacierzy [157].

V etap - Hybrydyzacja: Ten etap jest najbardziej czasochłonnym ze wszystkich, polega on na 16 godzinnej hybrydyzacji uzyskanego cRNA do sond mikromacierzowych. W zależności od struktury nukleotydu sondy hybrydują z różną dynamiką, wydajność tego etapu może zatem mieć znaczny wpływ na wyniki końcowe eksperymentu [164]. Zbyt długa hybrydyzacja może prowadzić do wysychania materiału w komorze mikromacierzy, co objawia się w postaci nierównomiernej hybrydyzacji. Odparowanie części wody z preparatu może dodatkowo zmienić stężenie obecnych w nim soli co wpływa znacząco na wydajność całego procesu [165].

VI etap - Płukanie i barwienie: Etap ten rozpoczyna proces odpłukania niespecyficznie związanych cząsteczek cRNA. Jest to kolejny etap, który może dramatycznie wpływać na wyniki końcowe. W zależności od warunków w jakich proces płukania został przeprowadzany (stężenie soli, jonów wapnia i magnezu w buforze płuczającym, temperatura) ilość niespecyficznie związanych cząsteczek jest odpłukiwana z różną wydajnością co znacząco wpływa na poziom tła. Dodatkowo indywidualne cechy sond wynikające z różnic w ich sekwencjach nukleotydu sprawiają, iż pomimo idealnego dopasowania do transkryptów wytworzone wiązania mają różną siłę [166], która dodatkowo uzależniona jest od temperatury w jakiej zachodzi reakcja. Po etapie płukania zhybrydyzowane fragmenty cRNA są wybarwiane za pomocą kompleksu fikoerytryny i streptawidyny (Ryc. 11). Fikoerytryna jest znacznikiem fluorescencyjnym natomiast streptawidyna jest białkiem posiadającym bardzo silne powinowactwo do biotyny, którą wyznakowane jest przygotowane wcześniej cRNA. Jakość użytego na tym etapie fluoroforu wpływa znacząco na poziom intensywności fluorescencji wszystkich sond obniżając czułość mikromacierzy w przypadku gdy substancja ta miała zbyt długi kontakt ze światłem dziennym.

VII etap - Skanowanie: W tym etapie, za pomocą światła lasera, wzbudzana jest fluorescencja fikoerytryny związanej z cRNA na każdej sondzie, za pomocą światła lasera. Następnie poziom fluorescencji odczytany zostaje poprzez detektor skanera mikromacierzowego. Czas trwania tego procesu uzależniony jest od rozmiaru mikromacierzy wahając się w okolicy 10minut. Na czas skanowania wszystkie macierze umieszczane są w komorze skanera po to aby ich fluorescencja nie była uzależniona od różnego czasu ekspozycji na światło dzienne co mogłoby zwiększać różnice pomiędzy badanymi próbkami. Zalecane jest przeprowadzenie wyłącznie jednego skanowania ponieważ każde kolejne obniża intensywność sygnału o 10-20% w związku z wyświecaniem się fluoroforu, jednocześnie obniżając czułość mikromacierzy [167].

VIII etap – Analiza danych: Ostatni etap polega na analizie danych, której pierwszym etapem jest kontrola jakości każdej macierzy w oparciu o zestaw sond kontrolnych określających wydajność poszczególnych procesów oraz jakości samego materiału wyjściowego. Kontrola jakości pozwala na wyodrębnienie mikromacierzy wysokiej jakości, które następnie przechodzą wstępne przetwarzanie. Polega ono głównie na odjęciu poziomu tła, którego celem jest ograniczenie wpływu niespecyficznego hybrydyzacji na poziomy sygnał. W następnym kroku przeprowadzana jest normalizacja w celu wyeliminowania różnic w poziomach sygnału wynikających z różnic w wydajności poszczególnych etapów, głównie procesów hybrydyzacji i amplifikacji. Wstępne przetwarzanie danych kończy sumaryzacja, która ogranicza różnice w poziomach ekspresji pomiędzy poszczególnymi sondami obliczając na ich podstawie jedną wartość ekspresji dla całego zestawu. Wstępne przetwarzanie ma bardzo duży wpływ na wyniki końcowe co uzależnione jest dodatkowo od zastosowanego algorytmu. Jego przeprowadzenie jest niezbędne, jednak wiąże się z kilkoma ograniczeniami, które wynikają z podstawowych założeń wykorzystanych metod. Główne z nich polega na tym, że średnia (lub określony kwantyl) poziomu ekspresji jest niezmienny pomiędzy próbkami, co skutkuje tym, że liczba genów różnicujących o zwiększonej i zmniejszonej ekspresji będzie zawsze podobna. Szczególnie jest to widoczne w przypadku metod bazujących na ujednoczeniu rozkładu intensywności sond pomiędzy próbkami (np. normalizacja kwantylowa). Założenia procedury eksperymentalnej a także algorytmów analizy uniemożliwiają zatem wykrycie globalnego spadku lub wzrostu poziomu ekspresji genów w wyniku zadziałania określonego czynnika wpływającego np. na tempo degradacji mRNA, sprawiając, iż w takiej sytuacji obserwowane trendy zmian mogą być niezgodne z rzeczywistością.

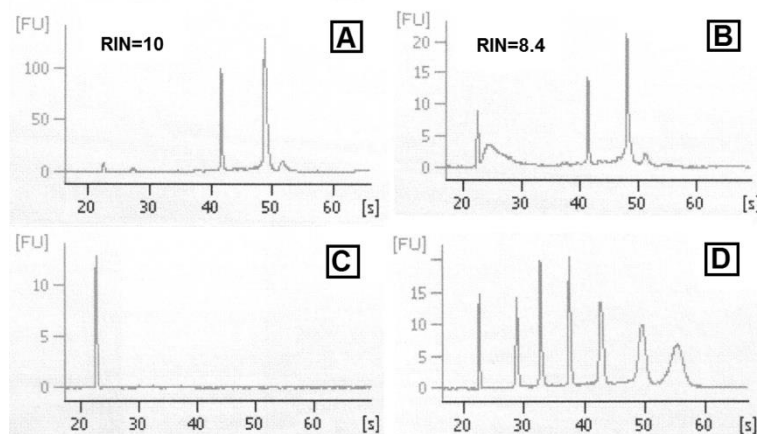
Przebieg procedury eksperymentalnej ulegał wielu zmianom od 2001 roku, kiedy pokazały się pierwsze mikromacierze firmy Affymetrix. Zmiany te dotyczyły głównie wykorzystania innego typu odczynników pozwalających na osiągnięcie większej wydajności przy krótszych czasach inkubacji. Najważniejsze zmiany wprowadzono w 2004 roku kiedy to dołożono dwuetapowy proces syntezy cDNA oraz w 2009 roku gdy zostały wprowadzone macierze z sondami specyficznymi dla eksonów (np. typu HuGene-1_0-st), które wykorzystują zestaw odczynników firmy Ambion. Starter oligo(dT) został w nich zastąpiony zestawem starterów o losowej sekwencji sześciu nukleotydów (N)₆ będących alternatywą dla specyficznego motywu (T)₂₄ wykorzystywanego przez oligo(dT). Dodatkowo znakowanie odbywa się po etapie fragmentacji z wykorzystaniem terminalnej transferazy deoksynukleotydowej (TdT), która dodaje wyznakowane nukleotydy wyłącznie na końcu 3' fragmentu cRNA. Zmiany w procedurze mogą być zatem przyczyną zmienności pomiędzy eksperymentami wykonanymi na przestrzeni kilku lat .

3.9.4. Metody kontroli jakości danych

Odpowiednia kontrola jakości i wybór próbek do dalszego przetwarzania jest jednym z kluczowych elementów procesu analizy danych mikromacierzowych [168]. Większość dostępnych algorytmów wstępnego przetwarzania danych, takich jak Robust Multi-array Average (RMA) [138] i jego odmiany [151, 169] oparte są o wiedzę dostarczoną przez wszystkie badane w eksperymencie próbki. Z tego względu wykorzystanie próbek o niskiej jakości, zawierających silne artefakty wynikające z niedokładności przeprowadzonej procedury eksperymentalnej wpływają na wyniki całego eksperymentu. Odpowiednia kontrola jakości jest zatem bardzo istotna z punktu widzenia całego procesu przetwarzania i zawsze powinna być przeprowadzona.

Ocena jakości RNA

RNA jest bardzo nietrwałym związkiem wrażliwym na szereg czynników zewnętrznych z tego względu jakość użytego RNA powinna być kontrolowana przed rozpoczęciem eksperymentu mikromacierzowego, w jego trakcie a także po zakończeniu jako element analizy danych mikromacierzowych. Przed naniesieniem materiału na powierzchnie mikromacierzy, jakość RNA oceniana jest za pomocą bioanalyzera, który działa na zasadzie podobnej do standardowej elektroforezy w żelu agarozowym. Wybarwione znacznikiem fluorescencyjnym RNA nanoszony jest na przystosowany do tego celu chip, przez który przepływa z różną prędkością w zależności od długości fragmentów RNA. Wynikiem jest wykres zależności poziomu fluorescencji w czasie, którego kształt uzależniony jest od stopnia degradacji naniesionego RNA. Dodatkowo jakość próbki oceniana jest poprzez współczynnik RIN (z ang. RNA Integrity Number) w skali 1-10 gdzie 10 oznacza RNA o najwyższej jakości, zawierające minimalną ilość produktów procesu degradacji. Ryc. 12 pokazuje przykładowe przebiegi uzyskane za pomocą bioanalyzera Agilent 2100.



Ryc. 12: Przykładowe wyniki analizy integralności RNA za pomocą bioanalyzera Agilent 2100. A- prawidłowe RNA, B- częściowo zdegradowane RNA, C-brak RNA w próbce, D-marker masy zbudowany z fragmentów RNA o długościach 0.1, 0.2, 0.3, 0.4, 0.5, 0.75 i 1kbp (Herok R., nieopublikowane dane)

RIN szacowany jest na podstawie indywidualnych cech wykresu przede wszystkim stosunku dwóch najwyższych pików widocznych na przebiegu A oznaczających proporcje jednostek 18S i 28S RNA rybosomalnego oraz piku markera w obszarze najniższych długości sekwencji widocznego szczególnie na wykresie B (w okolicy 22 sekundy). Niski poziom pików RNA rybosomalnego w stosunku do piku markera sugeruje, że znaczna część tego RNA uległa degradacji, co widoczne jest w postaci dodatkowego rozkładu w okolicy markera. Brak pików RNA rybosomalnego widoczny na wykresie C oznacza brak lub niedostatecznie wysokie stężenie RNA w badanej próbce. Wykres D pokazuje marker masy, dzięki któremu można określić długość RNA składającego się na poszczególne piki z pozostałych wykresów.

Po hybrydyzacji z mikromacierzą firmy Affymetrix poziom degradacji oraz zanieczyszczenie RNA rybosomalnym można określić poprzez analizę określonej grupy zestawów sond kontrolnych. Należą do nich zebrane w Tab. 1 sondy specyficzne dla genów referencyjnych (grupa 1) oraz rybosomalnego RNA (grupa 2). Podobnie jak w przypadku pozostałych zestawów sond pomiary uzyskane dzięki sondom kontrolnym nie pozwalają na stwierdzenie czy danego RNA było dużo czy mało gdyż jego wartość, wyrażona w jednostkach arbitralnych jest uzależniona od specyfiki samego eksperymentu oraz metody

wstępnego przetwarzania danych. Z tego względu arbitralne kryteria oparte o pojedyncze pomiary w tej sytuacji nie są wskazane. Wszystkie metody kontroli jakości bazujące na sondach kontrolnych oparte są o proporcje pomiarów albo na przestrzeni pozostałych mikromacierzy z danego eksperymentu albo w obrębie sygnału z pojedynczej macierzy (np. stosunek sygnału zestawu rRNA do średniej intensywności macierzy lub jednego z genów referencyjnych). Pozwala to na zidentyfikowanie potencjalnie problematycznych próbek, które znacznie różnią się od pozostałych pod względem jakości. Ze względu na to, że różne zestawy sond kontrolnych są podatne na wpływ określonych etapów eksperymentu pozwalają określić, który z nich jest źródłem zmienności i czy wobec tego określona próbka powinna być analizowana na dalszym etapie czy też odrzucona z eksperymentu.

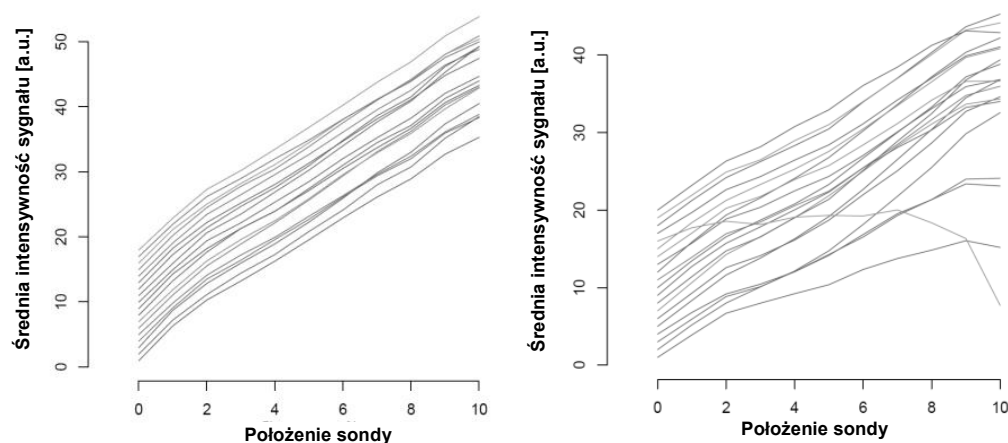
Grupa	Typ	Nazwa	Opis
1	Geny referencyjne	AFFX-HSAC07/X00351 AFFX-HUMGAPDH/M33197 AFFX-HUMISGF3A/M97935	ACTB - gen β -aktyny odpowiedzialnej za strukturę komórki GAPDH - enzym biorący udział w glikolizie STAT1 - czynnik transkrypcyjny
2	RNA rybosomalne	AFFX-HUMRGE/M10098 AFFX-M27830 AFFX-r2-Hs18SrRNA AFFX-r2-Hs28SrRNA	Gen kodujący jednostkę 18S rRNA Gen kodujący jednostkę 28S rRNA Gen kodujący jednostkę 18S rRNA - wersja 2 Gen kodujący jednostkę 28S rRNA - wersja 2
3	Kontrola amplifikacji (Polya spike)	AFFX-DapX / AFFX-r2-Bs-dap AFFX-ThrX / AFFX-r2-Bs-thr AFFX-PheX / AFFX-r2-Bs-phe AFFX-LysX / AFFX-r2-Bs-lys	Gen Dap bakterii <i>B.Subtilis</i> - proporcje 1:6,667 Gen Thr bakterii <i>B.Subtilis</i> - proporcje 1:25,000 Gen Phe bakterii <i>B.Subtilis</i> - proporcje 1:50:000 Gen Lys bakterii <i>B.Subtilis</i> - proporcje 1:100,000
4	Kontrola hybrydyzacji (Bac spike)	AFFX-BioB / AFFX-r2-Ec-bioB AFFX-BioC / AFFX-r2-Ec-bioC AFFX-BioDn / AFFX-r2-Ec-bioD AFFX-CreX / AFFX-r2-P1-cre	Gen BioB bakterii <i>E.Coli</i> – objętość 1.5 pM Gen BioC bakterii <i>E.Coli</i> – objętość 5 pM Gen BioD bakterii <i>E.Coli</i> – objętość 25 pM Gen Cre bakteriofaga P1 – objętość 100 pM

Tab. 1: Geny referencyjne mikromacierzy Affymetrix z serii 3'IVT

Poziom degradacji RNA także może być oceniony na podstawie zestawu sond kontrolnych. Każdy z zestawów znajdujących się w Tab. 1 składa się z 3 oddzielnych podgrup sond umieszczonych blisko końca 5', w środkowej części mRNA oraz na jego końcu 3'. Porównanie proporcji sygnału pomiędzy zestawami z końca 5' i 3' pozwala na oszacowanie stopnia degradacji, który z kolei jest porównywany z ustalonym przez producenta progiem odcięcia oraz z wartościami uzyskanymi dla pozostałych macierzy.

Inną formą oceny jakości RNA są wykresy stopnia degradacji, które przeciwieństwie do sond kontrolnych bazują na wszystkich sondach znajdujących się na macierzy [170]. Wykresy degradacji RNA pokazują średnią wartość ekspresji dla sond umieszczonych kolejno od końca 5' do 3' mRNA. Sondy PM pojedynczego zestawu, których w przypadku macierzy 3'IVT jest 11, ułożone są w kolejności odpowiadającej ich położeniu w sekwencji wiążanego mRNA. Ponieważ wszystkie sondy są wybrane z niewielkiego obszaru 600 nukleotydów różnica pomiędzy skrajnymi sondami jest niewielka, z tego względu zaobserwowanie określonego trendu jest możliwe dopiero po uwzględnieniu wszystkich zestawów. Średnia intensywność reprezentowana jest na wykresie osią Y natomiast kolejność wynikająca z położenia zaznaczona jest na osi X (Ryc. 13). W idealnych warunkach linie powinny być poziome – brak degradacji jednak tak nigdy nie jest gdyż RNA wyizolowane z komórek jest zawsze w pewnym stopniu zdegradowane. Z tego względu kąt nachylenia linii nie jest istotny, ważne jest jedynie aby był on

jednakowy dla wszystkich próbek. Nie jest zatem ważne w jakim stopniu RNA ulega degradacji (o ile mieści się w dopuszczalnym zakresie) a jedynie aby jego poziom był porównywalny dla wszystkich analizowanych mikromacierzy.



Ryc. 13: Przykładowe wykresy degradacji RNA pochodzące z eksperymentu o bardzo wysokiej jakości (lewy) oraz z eksperymentu zawierającego nieprawidłowo przygotowane mikromacierze (prawy). Poszczególne linie, wykonane dla oddzielnych mikromacierzy, reprezentują średnie poziomy sygnału dla sond położonych w różnej odległości od końca 5' transkryptu.

Poza oceną wzrokową opartą o przygotowany wykres, próbki porównywane są za pomocą testu na odchylenie wyznaczonej krzywej od linii regresji obliczonej na podstawie wszystkich mikromacierzy z danego eksperymentu. Pozwala to na zidentyfikowanie macierzy odstających, jednak należy pamiętać, że różnica w kącie nachylenia pojedynczej próbki wcale nie oznacza, iż wynika ona z różnego stopnia degradacji. Przykładem na to jest na próbka o przebiegu zbliżonym do poziomej linii widoczna w lewej części Ryc. 13, która wcale nie wynika z niskiego poziomu degradacji ale wysycenia sygnałów większości sond z tej pojedynczej mikromacierzy spowodowanej nałożeniem zbyt dużej ilości materiału biologicznego.

Ocena wydajności procesu amplifikacji i znakowania

Wydajność procesu amplifikacji i znakowania kontrolowana jest w przypadku technologii Affymetrix przez zestaw sond kontrolnych specyficznych dla transkryptów pięciu genów *Dap*, *Lys*, *Phe*, *Thr*, oraz *Trp* wyizolowanych z bakterii *B. subtilis*, nazywanych *polya spike* (grupa 3 w Tab. 1). Transkrypty te dodawane są w różnych proporcjach do wyizolowanego RNA a dzięki temu, że zawierają one koniec poli-A mogą brać udział w tych samych etapach procedury eksperymentalnej co ludzkie mRNA. Ilość dodawanego RNA genu *lys* jest najmniejsza, na granicy czułości mikromacierzy (1:100,000 część całkowitego RNA, czyli około 3 transkrypty na komórkę) z tego względu jest ono wykorzystywane jako kontrola czułości całego doświadczenia. Zdaniem producenta w prawidłowo przeprowadzonym eksperymencie poziom sygnału otrzymanego dla genu *Lys* powinien przewyższać poziom tła w przypadku przynajmniej połowy próbek. Pozostałe transkrypty dodawane są w większych objętościach tworząc zależność $Lys < Phe < Thr < Dap$ gdzie ilość *Dap* jest największa – na granicy nasycenia sond.

Analiza intensywności sygnału zestawów sond kontrolnych pozwala na monitorowanie procesu znakowania i amplifikacji niezależnie od typu ilości i jakości badanego RNA, chociaż zanieczyszczenia wewnątrz próbek RNA mogą wpływać na efektywność obu tych procesów. Niezachowane proporcje pomiarów w przedziale niskich lub wysokich wartości wskazują na problemy z czułością mikromacierzy, jeśli jednak sondy te pokazują prawidłowe wartości a mimo to istnieje problem z próbką to najprawdopodobniej przyczyną jest wtedy niska jakość wyjściowego materiału biologicznego.

Większość macierzy nowszego typu zawiera dwa typy zestawów sond kontrolnych o nieco bardziej informatywnych nazwach i przyrostkach „r2” (Tab. 1). Oba zestawy mają sondy specyficzne dla tego samego genu jednak wybrane z nieco innego fragmentu, który jest bardziej specyficzny. Z tego względu zestawy „r2” powinny być wykorzystane do analizy o ile są dostępne dla mikromacierzy określonego typu.

Ocena wydajności procesu hybrydyzacji

Podobnie jak w przypadku etapów znakowania i amplifikacji, proces hybrydyzacji kontroluje zestaw sond specyficznych dla transkryptów dodawanych do mieszaniny (grupa 4 w Tab. 1). *BioB*, *bioC* i *bioD* pochodzą od genów uczestniczących w syntezie biotyny u bakterii *E. Coli* natomiast *Cre* jest wyizolowane z bakteriofaga P1. Dodawane one są do mieszaniny dopiero przed procesem hybrydyzacji, po to aby możliwe było jego kontrolowanie niezależnie od poprzednich etapów eksperymentu takich jak znakowanie czy amplifikacja. *BioB* podobnie do *Lys* jest dodawany w ilości odpowiadającej stosunkowi 1:100,000 do RNA komórkowego, która jest na granicy czułości mikromacierzy. W związku z tym on także powinien charakteryzować się sygnałem hybrydyzacji powyżej poziomu tła w przypadku przynajmniej 50% próbek. Pozostałe zestawy kontrolne powinny pokazywać trend $bioB < bioC < bioD < Cre$, który z uwagi na to, że sygnały tych zestawów są niezależne od koncentracji oraz jakości RNA materiału początkowego, powinien być podobny na przestrzeni wszystkich badanych próbek.

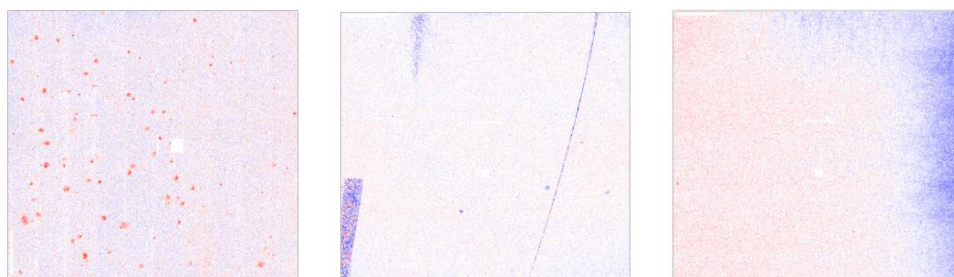
W przypadku, gdy poziomy ekspresji odczytane przez sondy należące do standardowych zestawów są prawidłowe, podczas gdy zestawy sond kontrolujących proces amplifikacji i znakowania znacznie odbiegają od pozostałych próbek, problem najprawdopodobniej nie leży po stronie procesu hybrydyzacji. Jednak jeśli oba typy sond kontrolnych wskazują na problemy to najbardziej prawdopodobną przyczyną jest zaburzenie wydajności procesu hybrydyzacji lub pęknięcia mikromacierzy. Problemy pojawiające się wyłącznie w przypadku sond kontrolnych typu *bac spike* sugerują błąd pipetowania podczas przygotowywania mieszaniny RNA kontrolnego (Tab. 2).

Geny referencyjne	polyA spike	Bac spike	Możliwa przyczyna
błąd	ok	ok	niska jakość wyjściowego RNA
błąd	błąd	ok	problemy na etapach amplifikacji/znakowania
błąd	błąd	błąd	problemy na etapie hybrydyzacji/pęknięcia
ok	ok	błąd	błąd pipetowania podczas dodawania bac spike
ok	błąd	ok	błąd pipetowania podczas dodawania polyA spike

Tab. 2: Problemy objawiające się różnicami w sygnale różnych grup sond kontrolnych oraz możliwe przyczyny

Innym sposobem na walidowanie przebiegu procesu hybrydyzacji w przypadku technologii Affymetrix są obrazy powierzchni mikromacierzy. Na podstawie plików z danymi (typu CEL) możliwe jest odtworzenie

obrazu odczytanego przez detektor skanera, który został przekonwertowany na dane liczbowe. Odzyskany obraz ma znacznie mniejszą dokładność niż oryginał głównie z tego względu, że pierwotnie sygnał każdej pojedynczej sondy reprezentowany był nie przez jeden ale 16 pikseli obrazu, których oryginalna intensywność (zapisana w plikach typu DAT) zwykle nie jest przechowywana po eksperymencie. Głównym założeniem podczas budowy mikromacierzy Affymetrix jest losowe rozmieszczenie sond na jej powierzchni, z tego względu silna wariancja poziomu sygnału wartości w określonych obszarach obrazu sugerują przyczyny inne niż zróżnicowanie biologiczne. Artefakty widoczne na obrazach powierzchni mikromacierzy to głównie drobne zanieczyszczenia, które powodują, że zwykle niewielkie obszary macierzy pokazują sztucznie większą bądź mniejszą intensywność fluorescencji [171]. Występują one bardzo często i o ile są niewielkie to stosunkowo dobrze radzą sobie z nimi metody sumaryzacji zestawów sond, w których pojedyncze sondy będące wielkościami odstającymi są odrzucane. Niektóre artefakty wynikające np. z nierównomiernej hybrydyzacji albo uszkodzeń powstałych w procesie płukania mogą zajmować bardzo duży obszar macierzy co ma bardzo silny wpływ na oszacowane poziomy ekspresji poprzez nieprawidłowo wyznaczony poziom tła albo niewłaściwe ujednoczenie rozkładów intensywności sond w procesie normalizacji kwantylowej.



Ryc. 14: Przykładowe obrazy różnicowe pomiędzy próbkami zawierającymi artefakty a obrazem referencyjnym (mediana sygnału ze wszystkich mikromacierzy). Niebieski kolor oznacza niską ekspresję w stosunku do macierzy referencyjnej, czerwony wysoką. Lewa strona – zanieczyszczenia powstałe podczas procesu płukania, środek – uszkodzenia mechaniczne macierzy, prawa strona – nierównomierna hybrydyzacja.

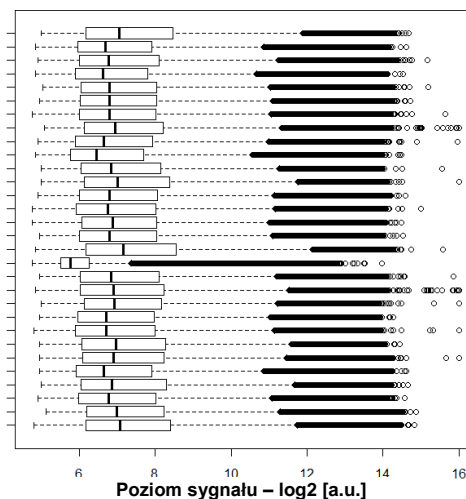
Artefakty obrazu powierzchni mikromacierzy można jednak stosunkowo łatwo zidentyfikować porównując obrazy uzyskane dla poszczególnych próbek względem obrazu referencyjnego zwykle stworzonego w oparciu o medianę intensywności wszystkich mikromacierzy z eksperymentu. Sondy należące do tego typu artefaktów mogą być pominięte podczas procesu analizy danych lub odtworzone na podstawie innych próbek lub innych sond należących do zestawu zawierającego sondy obarczone błędem [144, 145, 171]. W przypadku gdy artefakty są zbyt duże zaleca się odrzucenie danej mikromacierzy.

Rozkład intensywności sygnału

Analiza rozkładu intensywności surowych sygnałów ze wszystkich sond jest zwykle jednym z pierwszych etapów procesu kontroli jakości. Większość czynników wpływających na proces pomiarowy wpływa w istotny sposób na kształt rozkładu intensywności sond. Najczęściej rozkład intensywności przedstawiany jest w formie wykresów ramkowych (Ryc. 15) albo histogramów, których celem jest wyodrębnienie mikromacierzy o znacznie większej lub mniejszej intensywności. Uważa się, że różnice w poziomie intensywności pomiędzy próbkami nie są problematyczne gdyż kompensują to procedury normalizacji

jednak próbki o nienaturalnych rozkładach wymagają dokładniejszej analizy w celu określenia przyczyny obserwowanych różnic [170].

Różnice w poziomie intensywności mogą wynikać z niedokładności niemal każdego etapu procedury eksperymentalnej. Mimo, że wydajność niektórych etapów może być kompensowana w trakcie procedury np. ilość wyizolowanego mRNA, to dokładność pipetowania, od której uzależnione są właściwe warunki wszystkich reakcji oraz dokładność urządzeń wykorzystywanych do pomiaru stężenia RNA mogą znacząco wpływać na wyniki.

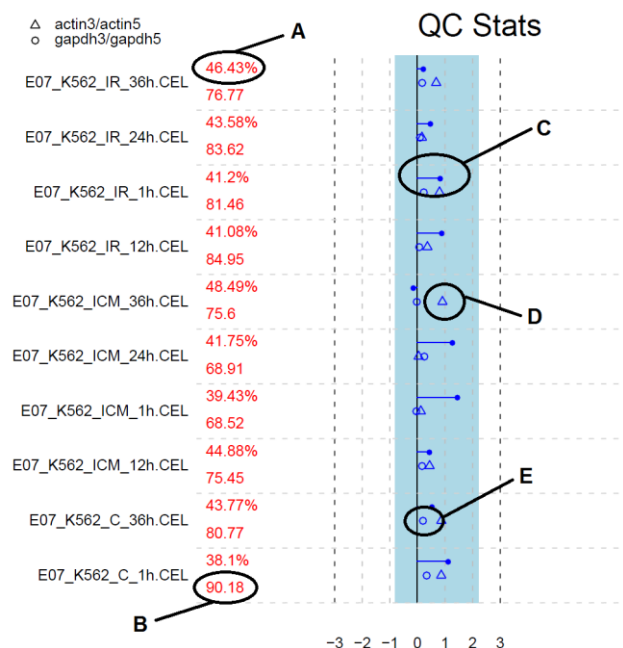


Ryc. 15: Przykładowe wykresy ramkowe z wyraźnie widoczną macierzą odstającą o nietypowo niższej medianie sygnałów surowych sond

Ze względu na to, iż większość transkryptów się nie zmienia a specyfika samej procedury doświadczalnej powinna zapewnić równy poziom intensywności pomiędzy próbkami to wszelkie zaburzenia widoczne na wykresach skupiających się na kształcie rozkładów intensywności są albo wynikiem spadku wydajności jednego z procesów (np. hybrydyzacji) albo wynikają z wysokiego stopnia degradacji badanego RNA. W przypadku znacznych różnic w średnim poziomie intensywności w dalszym ciągu można zastosować normalizację w celu efektywnego zmniejszenia różnic jednak należy pamiętać o tym, że zbyt niska intensywność przeskalowana do wysokich wartości wpłynie na zintensyfikowanie poziomu szumu w danych i dodatkowo, w przypadku, gdy liczba próbek w eksperymencie jest niewielka, doprowadzi do obniżenia czułości pozostałych mikromacierzy. Z tego względu nieraz lepszym rozwiązaniem jest odrzucenie próbki z dalszej analizy przed procesem normalizacji.

Kontrola jakości mikromacierzy wymaga nieraz przeanalizowania wielu parametrów w celu wyodrębnienia macierzy odstających, w tym bardzo pomocne są tzw. wykresy QC (Ryc. 16), które powstają poprzez połączenie kilku statystyk niezwykle pomocnych przy identyfikacji próbek o niskiej jakości [168]. Wykresy QC składają się z kilku elementów zaznaczonych niebieskim kolorem w przypadku gdy reprezentują wartości dopuszczalne zdefiniowane przez producenta mikromacierzy lub czerwonym gdy jest przeciwnie. Procenty obok nazw próbek (Ryc. 16 - A) reprezentują zestawy sond oznaczone jako „obecne” (z ang. „present”) tzn. takie, których ekspresja przewyższa poziom tła, wynikający z niespecyficznego hybrydyzacji. Zestaw sond jest typu „present” jeśli poziom sygnału jego sond PM jest znamienne silniejszy od sygnału sond MM. Niewielki procent tego typu zestawów sond oznacza, że albo znaczna część genów nie ulega ekspresji albo poziom niespecyficznego hybrydyzacji jest wysoki

(PM<=MM). W przypadku tych wykresów procent wykorzystywany jest głównie do identyfikacji macierzy odstających różniących się pod względem tej statystyki od pozostałych. Jego średnia wartość jest mało informatywna ze względu na fakt, że może się ona drastycznie różnić dla różnego typu komórek. Dopuszczalny zakres różnic tego współczynnika na przestrzeni wszystkich mikromacierzy to 10% jeśli założenie to nie jest spełnione to wszystkie macierze oznaczone są kolorem czerwonym. W przypadku macierzy eksonowych nowego typu gdzie nie ma już sond typu PM statystyka ta może być wyznaczona za pomocą algorytmu DABG (detection above background) bazującego na sondach intensywności tła.



Ryc. 16: Przykładowe wykresy QC dla mikromacierzy Affymetrix 3'IVT

Wartości pod procentami (B) to średnie tło, które także służy do identyfikacji macierzy odstających. Znacznie wyższa bądź niższa wartość tego współczynnika sygnalizuje różnice w poziomie niespecyficznej hybrydyzacji, na którą wpływ mogą mieć takie czynniki jak ilość cRNA i jego jakość, wydajność reakcji hybrydyzacji, płukania czy też barwienia.

Poziome linie (C) to współczynnik skalowania z normalizacji typu MAS 5.0. Tego typu normalizacja zakłada, że ekspresja większości genów nie ulega zmianie i średnia wartość sygnału na przestrzeni wszystkich eksperymentów powinna być taka sama. Dopuszczalna wartość tego współczynnika powinna się mieścić w zakresie zaznaczonym niebieskim prostokątem, wynikającym z dopuszczalnego odchylenia od wartości średniej wszystkich macierzy. Im wyższa jest wartość współczynnika tym mniejsza jest średnia intensywność macierzy i jednocześnie potrzebne jest silniejsze przeskalowanie w stronę większych wartości, co odbija się na czułości metody. Najwyższa wartość współczynnika skalowania nie powinna przekraczać trzykrotności najmniejszej wartości.

Znaczniki w kształcie koła i trójkątów to stosunki sygnałów sond umieszczonych na końcach 3' i 5' genów kontrolnych - β -aktyny (D) i GAPDH (E). Dzięki temu, że są to pojedyncze geny odporne na raptowne zmiany ekspresji to są one dobrym wyznacznikiem jakości użytego mRNA. GAPDH charakteryzuje się krótszym transkryptem (1310 nukleotydów) i zgodnie z danymi podawanymi przez firmę Affymetrix jego stosunek 3':5' nie powinien przekraczać 1.25. Wartości poniżej 1 są także

dopuszczalne w przypadku tych sond. Transkrypt β -aktyny (ACTB) jest dłuższy (1852nt) przez co jego stosunek 3' i 5' jest większy a dopuszczalna wartość to 3 .

Inne metody służące do oceny jakości próbek

W literaturze pojawia się bardzo dużo podejść do problemu identyfikacji macierzy odstających. skupiają się one jednak na zidentyfikowaniu skutków a nie przyczyny pojawiania się różnic pomiędzy próbkami. Najpopularniejsze metody obejmują:

- korelogramy służące do wyodrębnienia mikromacierzy nieskorelowanych z pozostałymi próbkami
- statystyki NUSE (Normalized Unscaled Standard Error) i RLE (Relative Log Expression) służące do oceny różnic w wartościach pomiędzy poszczególnymi sondami na przestrzeni całej macierzy [172]
- analiza PCA (Principal Component Analysis) i BGA (Between Group Analysis) [5] które skupiają się na identyfikacji czynników będących podstawową przyczyną zmienności pomiędzy próbkami
- metody klasteryzacji hierarchicznej określające podobieństwo pomiędzy próbkami w sposób pozwalający na podzielenie ich na różną liczbę klas

3.9.5. Metody wstępnego przetwarzania i problemy z nimi związane

Poza wnikliwą kontrolą jakości dane mikromacierzowe wymagają zastosowania dodatkowych metod wstępnego przetwarzania i standaryzacji danych w celu wyeliminowania różnic technicznych pomiędzy badanymi próbkami niezwiązanymi z badanymi różnicami biologicznymi. Typowy schemat procesu wstępnego przetwarzania danych z mikromacierzy Affymetrix składa się z trzech etapów:

- Korekcja tła – polegająca na oszacowaniu poziomu niespecyficznego hybrydyzacji oraz odjęciu poziomu tła od uzyskanego sygnału
- Normalizacja – polegająca na ujednoczeniu kształtu rozkładów sygnału sond pomiędzy próbkami lub określonych jego parametrów (np. średniej, mediany)
- Sumaryzacja – w przypadku zastosowania wielu sond zgrupowanych w zestawy, odpowiadające poszczególnym genom lub transkryptom, etap ten pozwala na zunifikowanie sygnału w celu otrzymania pojedynczej wartości dla danego zestawu sond

Istnieje wiele podejść do problemu standaryzacji danych jednak żadna z nich nie gwarantuje pełnej skuteczności ze względu na wciąż słabo poznane czynniki wpływające na różnice techniczne pomiędzy mikromacierzami. Z tego względu pomimo ponad dekady badań nad algorytmami wstępnego przetwarzania do dziś nie istnieje wypracowany standard procesu wstępnego przetwarzania danych mikromacierzowych. Najpopularniejsze metody wstępnego przetwarzania danych obejmują następujące algorytmy:

- MAS5 – metoda opracowana przez firmę Affymetrix, jej podstawową cechą jest wykorzystanie korekcji tła opartej o różnice sygnałów pomiędzy sondami PM i MM oraz normalizacji przeprowadzanej niezależnie od pozostałych próbek w eksperymencie [173]

- RMA (Robust Multi-array Average) – wykorzystuje korekcję tła opartą wyłącznie o sygnały sond PM, normalizację kwantylową silnie uzależnioną od pozostałych próbek z danego eksperymentu oraz sumaryzację typu *median polish* [138]. Metoda ta zawiera bardzo wiele odmian takich jak: GC-RMA uwzględniającą różnice w hybrydyzacji sond o różnym składzie GC na etapie sumaryzacji sygnałów [151], tRMA – ze zmodyfikowaną kolejnością odejmowania mediany od wierszy i kolumn podczas procesu sumaryzacji [169] oraz fRMA – która podczas przetwarzania nie bazuje na pozostałych próbkach danego eksperymentu ale na wartościach referencyjnych określonych na podstawie zbioru kilkuset innych eksperymentów mikromacierzowych [174]
- PLIER (Probe Logarithmic Intensity ERror) – algorytm opracowany przez firmę Affymetrix w 2004r jako następca metody MAS5. Uwzględnia ona niejednorodność zmian sygnału pomiędzy sondami o różnej budowie w przypadku zmian w koncentracji materiału biologicznego, które szacowane są na podstawie określonych sond rozmieszczonych w różnych częściach macierzy [175]
- FARMS (Factor Analysis for Robust Microarray Summarization) – wprowadza nowy algorytm sumaryzacji, w którym poziom RNA oszacowany jest na podstawie modelu analizy czynnikowej przy założeniu, że szum pomiarowy ma postać rozkładu Gaussa [176]
- MBEI (Model Based Expression Index) – znana także jako dChip od nazwy bazującego na niej oprogramowania lub metoda Li-Wong od nazwisk autorów [177]

Głównym założeniem metod normalizacji danych jest to, że ekspresja wszystkich badanych genów pomiędzy próbkami nie ulega silnym zmianom i jej sumaryczna wartość powinna być stała pomiędzy poszczególnymi próbkami. W przypadku normalizacji kwantylowej dodatkowo wymagane jest zachowanie kształtu rozkładów intensywności sond. Prowadzi to do tego, że liczba zidentyfikowanych genów o zwiększonej i zmniejszonej ekspresji pomiędzy próbkami zawsze będzie podobna.

Większość metod zakłada, że wszystkie sondy na macierzy są podobnej jakości, o zbliżonych parametrach dynamiki procesu hybrydyzacji, zbliżonej czułości oraz specyficzności [178], co jest znacznym uproszczeniem. Dodatkowo metody te nie biorą pod uwagę zmian jakie nastąpiły w dokładności opisu sekwencji nukleotydowych w publicznych bazach danych od momentu zaprojektowania mikromacierzy. Z tego względu potrzeba kontrolowania jakości dopasowania sond do genów bazująca na komplementarności sekwencji nukleotydowych jest jednym z czynników, które wymagają szczególnej uwagi podczas procesu wstępnego przetwarzania.

Problem nieprawidłowego dopasowania sond do genów w przypadku mikromacierzy Affymetrix był wielokrotnie podkreślany w literaturze [148, 149, 179], udowadniając potrzebę wykorzystania algorytmów tzw. re-adnotacji sond do genów. W przeciwieństwie do tzw. plików adnotacyjnych (ang. annotation files) zawierających opisy zestawów sond i ich powiązanie z istniejącymi genami, sama definicja zestawu z uwzględnieniem liczby i oraz sekwencji poszczególnych sond nie jest aktualizowana przez producenta mikromacierzy ze względu na potrzebę zachowania powtarzalności wyników dla określonej platformy pomiędzy eksperymentami.

Pomimo bardzo starannego procesu projektowania zestawów sond, różnice w sile sygnału pomiędzy sondami należącymi do tego samego zestawu są bardzo często obserwowane [177]. Najczęściej jest to związane z nieprawidłowościami w dopasowaniu sond do genów, które ujawnione zostały wraz z

udoskonalaniem metod sekwencjonowania i poprawą dokładności opisu sekwencji genomów [180]. W rezultacie znaczna część sond, szczególnie w przypadku starych platform jest specyficzna dla więcej niż jednego genu lub nie jest specyficzna dla żadnego z nich [181-185], będąc jednym ze źródeł wysokiej wariacji sond należących do tego samego zestawu.

Dodatkowym problemem jest występowanie kilku zestawów sond specyficznych dla pojedynczego genu [146], które ze względu na różnice w jakości dopasowania pojedynczych sond mogą nieraz pokazywać sprzeczne wartości, sprawiając wiele trudności podczas procesu interpretacji przetworzonych danych. Dodatkowe zestawy przypisane do pojedynczego genu wynikają najczęściej z błędów popełnionych podczas ich projektowania np. w sytuacji, gdy wcześniej zsekwencjonowane fragmenty DNA okazały się być elementem tego samego genu lub jego alternatywnej formy splicingowej. Wiele metod zostało zaproponowanych w celu zunifikowania sygnałów dla pojedynczego genu od bardzo prostych polegających na wybieraniu jednego z zestawów losowo [186] lub zestawu o największej ekspresji [187] po bardziej wyrafinowane oparte o najróżniejsze testy statystyczne [146, 147, 188].

Jednym z najczęściej wykorzystywanych podejść jest redefinicja zestawów sond polegająca na połączeniu sond w zestawy specyficzne dla pojedynczego genu lub transkryptu z dodatkową kontrolą jakości dopasowania poszczególnych sond. Podejście to pozwala na wyeliminowanie problemu błędnego dopasowania sond oraz występowania wielu zestawów specyficznych dla określonego genu. Pomimo, że redefinicja zestawów przynosi bardzo wiele korzyści zwiększając wiarygodność wyników [148, 149, 189] to także ma wiele wad, które sprawiają, iż nie jest wciąż ogólnie przyjętym standardem. Metody tego typu bazują całkowicie na strukturze sekwencji nukleotydowych publicznych baz danych, które w dalszym ciągu są rozwijane i wciąż zawierają wiele niedokładności, nie biorą one jednak pod uwagę rzeczywistego poziomu ekspresji sond, jakie są obserwowane w eksperymentach z wykorzystaniem danej platformy. Dodatkowo nowe zestawy nie zawierają stałej ilości sond ze względu na procesy filtrowania sond i łączenia kilku zestawów. Z tego względu ich liczba w zależności od przyjętego kryterium jest różna co jednocześnie przekłada się na różnice w jakości sygnału, który w przypadku niewielkiej ilości sond obarczony jest dużym szumem pomiarowym. Metody redefinicji zestawów pomijają także kryterium łączenia sond w zestawy pod warunkiem, że są one komplementarne dla sekwencji znajdujących się w wąskim przedziale 600 nukleotydów. Łączenie sond rozrzuconych w większej odległości od siebie może zatem być potencjalnym źródłem wysokiej wariacji sygnałów ze względu na proces degradacji RNA, który ma tym większy wpływ na sygnały im sondy położone są dalej od końca 3' genu [158].

Wybór odpowiedniego algorytmu wstępnego przetwarzania może mieć bardzo istotny wpływ na wyniki analizy szczególnie w przypadku niewielkich zbiorów danych [190]. Wybór odpowiedniej metody jest jednak niezwykle trudny gdyż każda z nich ma swoje wady i zalety, które uzależnione są od specyfiki analizowanego zbioru danych [138, 191].

3.9.6. Metody poszukiwania genów różnicujących

Jednym z podstawowych celów mikromacierzy jest identyfikacja genów które wykazują istotne zmiany w poziomach ekspresji na skutek oddziaływania określonego czynnika lub pomiędzy różnymi typami komórek. Tego typu porównanie wykonuje się pomiędzy parami próbek badanymi za pomocą oddzielnych mikromacierzy, grupami próbek w przypadku zastosowania powtórzeń technicznych/biologicznych lub pomiędzy różnymi kanałami fluorescencji w przypadku macierzy

dwukolorowych badających RNA z dwóch próbek wyznakowanych różnymi barwnikami fluorescencyjnymi.

Wybór odpowiednich kryteriów klasyfikacji genów jako różnicujące jest niezwykle trudny i stanowi przedmiot licznych dyskusji od wielu lat. Najprostsze metody poszukiwania genów różnicujących bazują na ilorazie dwóch wartości – FC (z ang. fold-change) lub ich odpowiednikowi w skali logarytmicznej – LFC (z ang. log-fold-change), który porównywany jest z określonym progiem odcięcia dobranym arbitralnie lub na podstawie specyficznych cech sygnału [192, 193].

Inną popularną metodą jest wykorzystanie testów parametrycznych takich jak test t-Studenta lub nieparametrycznych takich jak test Wilcoxa. Testy parametryczne mają zwykle większą moc w przypadku, gdy analizowany zbiór danych ma kształt rozkładu normalnego. Z kolei testy nieparametryczne nie wymagają aby to założenie było spełnione jednak w przypadku niewielkich zbiorów danych gdzie liczba powtórzeń technicznych/biologicznych jest ograniczona testy nieparametryczne mają bardzo niską moc [194]. Z tego względu zwykle stosuje się test t lub jego modyfikacje [195], nieraz także w połączeniu z innymi metodami takimi jak kryteria oparte o FC [196, 197]. Podstawowym ograniczeniem tego typu podejść jest jednak to, iż w przeciwieństwie do kryteriów opartych wyłącznie o FC testy statystyczne wymagają powtórzeń technicznych/biologicznych ze względu na swoją specyfikę.

Z punktu widzenia statystyki każdy gen analizowany za pomocą mikromacierzy jest niezależną zmienną i zwykle jego zmiana testowana jest jako oddzielna hipoteza w teście statystycznym. Z tego powodu konieczne jest zastosowanie korekty na wielokrotne testowanie w celu obniżenia poziomu fałszywie dodatnich wyników. Przykładowo, jeżeli testujemy pojedynczą hipotezę z poziomem istotności 5% to oczekujemy, że prawdopodobieństwo popełnienia błędu pierwszego rodzaju (prawdziwa hipoteza zostaje odrzucona) wynosi 0.05, jednak jeśli przedmiotem analizy jest 10 tysięcy genów i przez to testujemy tyle samo hipotez to możemy oczekiwać, że 5% z nich czyli 500 będzie obciążona błędem. Najczęściej wykorzystywane korekty na wielokrotne testowanie obejmują metodę Bonferroniego, Benjamini-Hochberga czy bardzo często wykorzystywaną w kontekście danych biologicznych metodę zaproponowaną przez Storeya w 2003 roku [198].

Najpopularniejsze programy służące do poszukiwania genów różnicujących obejmują pakiet Limma (Linear Models for Microarray Data) [142] oraz SAM (Significance Analysis of Microarrays) [199]. Oba programy przeprowadzają zmodyfikowaną wersję testu t dla każdego z genów oddzielnie a następnie wybraną korektę na wielokrotne testowanie. Limma dodatkowo tworzy prosty model liniowy dla każdego genu oraz przeprowadza zmodyfikowany test t, który bierze pod uwagę wariancję sygnału całej mikromacierzy a nie tylko pojedynczych genów. Obie metody cieszą się bardzo dużą popularnością posiadając kilkadziesiąt cytowań jednak w kontekście danych mikromacierzowych nie można o żadnej powiedzieć, że jest lepsza od drugiej.

3.10. Poszukiwanie wzorców w sekwencjach nukleotydowych

Od czasu rozpoczęcia projektu sekwencjonowania ludzkiego genomu liczba metod i narzędzi służących do analizy sekwencji nukleotydowych stale wzrasta w drastycznym tempie. Najpopularniejsze metody koncentrują się na poszukiwaniu podobieństw między sekwencją określonego motywu a zdefiniowanym obszarem genomu lub na poszukiwaniu podobieństw w sekwencjach nukleotydowych pomiędzy różnymi gatunkami [200]. Tego typu podejścia pozwoliły na poznanie funkcji i położenia najróżniejszych obszarów

sekwencji genomu w tym sekwencji genów [201] transpozonów [202], obszarów promotora genu [203], sekwencji prekursorów microRNA [204], motywów rozpoznawanych przez białka z domeną umożliwiającą przyłączanie RNA [205] lub motywów rozpoznawanych przez cząsteczki microRNA [206].

3.10.1. Metody i narzędzia do analizy sekwencji nukleotydowych

Poszukiwanie wzorców w sekwencjach nukleotydowych nie jest łatwym zadaniem głównie ze względu na często występującą niespecyficzną sekwencji lub niespecyficzną biologicznych mechanizmów rozpoznawania motywów. Dodatkowo częstym ograniczeniem jest konieczność przeanalizowania dużych zbiorów danych takich jak sekwencje genomów liczących od kilkuset tysięcy do kilkuset miliardów nukleotydów. Metody poszukiwania wzorców sekwencji nukleotydowych można podzielić na trzy podstawowe grupy:

- Metody oparte o specyficzne sekwencje i wiązanie wymagające pełnej komplementarności, np. w przypadku poszukiwania obszarów rozpoznawanych przez enzymy restrykcyjne wymagające pełnej komplementarności 4-8 nukleotydów
- Metody oparte o niespecyficzne motywy sekwencyjne, np. sekwencje rozpoznawane przez białka z rodziny czynników transkrypcyjnych, gdzie rozpoznawany motyw nie jest jednoznacznie określony ze względu na niespecyficzne oddziaływanie z białkami
- Metody oparte o niespecyficzne dopasowanie sekwencji, np. poszukiwanie miejsc rozpoznawanych przez miRNA, które pomimo, że mają specyficzną sekwencje to nie wymagają pełnej komplementarności

Większość metod bazuje na mechanizmach porównywania motywów sekwencyjnych reprezentowanych przez ciąg znaków w postaci liter lub wektorów wartości przypisanych określonym nukleotydom. Istnieje bardzo wiele narzędzi służących do tego typu analiz skupiających się na sekwencjach nukleotydowych DNA i RNA a także sekwencjach aminokwasów z jakich zbudowane są białka. Jednym z największych zbiorów aplikacji służących do przetwarzania tego typu danych jest pakiet EMBOSS [207], w którego skład wchodzi ponad 250 różnych narzędzi obejmujących wszystkie podstawowe schematy analizy. W celu usprawnienia ich efektywności oraz zapewnienia uniwersalności sposobu reprezentacji danych i formatu zapisu w skład pakietu wchodzi dodatkowo szereg reguł wyodrębniania i zapisu danych do plików, które są elementem każdego programu.

Specyficzne sekwencje, pełna komplementarność

W tym przypadku przeszukiwanie sekwencji ogranicza się zwykle do poszukiwania krótkich motywów wewnątrz długich fragmentów DNA za pomocą prostego porównywania ciągu znaków. W przypadku bardzo dużych zbiorów dodatkowo przeprowadza się indeksowanie sekwencji co znacznie przyspiesza proces przeszukiwania. Wszystkie niezbędne aplikacje do tego typu analiz można znaleźć w pakiecie EMBOSS.

Niespecyficzne motywy sekwencyjne

Tego typu motywy sekwencyjne pojawiają się najczęściej przy poszukiwaniu oddziaływań DNA/RNA z białkami. Bazują one na kombinacji oddziaływań elektrostatycznych oraz Van der Walsa, których siła

zależy od konformacji białka i jego zgodności dopasowania z sekwencją nukleotydową. Brak pełnej specyficzności sekwencji w tej sytuacji prowadzi jedynie do obniżenia sumarycznej siły oddziaływania jednak w dalszym ciągu może być ona na tyle mocna, że białko może spełniać swą rolę.

Niespecyficzne motywy sekwencyjne reprezentowane są albo w postaci kodu IUPAC (ang. International Union of Pure and Applied Chemistry) albo w postaci tzw. macierzy wag pozycji. Przykładowo, jednym z najpopularniejszych motywów rozpoznawanych przez białka wiążące DNA jest tzw. TATA-box. Jego sekwencja nie jest jednak jednoznaczna gdyż białka te mogą wiążąc jedną z jego 6 form (Ryc. 17a). Zgodnie z kodem IUPAC motyw ten można zapisać jako sekwencje „STATAWARRSSSS” gdzie S = G lub C; W = A lub T; R = A lub G. Motyw w tej formie może być bezpośrednio wykorzystany do przeszukiwania innych sekwencji jednak ze względu na to, że nie przechowuje on w sobie informacji o tym jakie kombinacje poszczególnych nukleotydów są dozwolone to podejście to nie jest odpowiednie dla wszystkich klas motywów [208].

(a)	GTATAAAAAGCGG CTATAAAAGGCC GTATAAAGGGCG GTATATAAGCGG CTATAAAGGGCC GTATAAAGGGGG	(b)	<table style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th></th> <th>1</th> <th>2</th> <th>3</th> <th>4</th> <th>5</th> <th>6</th> <th>7</th> <th>8</th> <th>9</th> <th>10</th> <th>11</th> <th>12</th> <th>13</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>0</td> <td>0</td> <td>6</td> <td>0</td> <td>6</td> <td>5</td> <td>6</td> <td>3</td> <td>1</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>C</td> <td>2</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>2</td> <td>2</td> <td>4</td> <td>2</td> </tr> <tr> <td>G</td> <td>4</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>3</td> <td>5</td> <td>4</td> <td>4</td> <td>2</td> <td>4</td> </tr> <tr> <td>T</td> <td>0</td> <td>6</td> <td>0</td> <td>6</td> <td>0</td> <td>1</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> </tbody> </table>		1	2	3	4	5	6	7	8	9	10	11	12	13	A	0	0	6	0	6	5	6	3	1	0	0	0	0	C	2	0	0	0	0	0	0	0	0	2	2	4	2	G	4	0	0	0	0	0	0	3	5	4	4	2	4	T	0	6	0	6	0	1	0	0	0	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13																																																												
A	0	0	6	0	6	5	6	3	1	0	0	0	0																																																												
C	2	0	0	0	0	0	0	0	0	2	2	4	2																																																												
G	4	0	0	0	0	0	0	3	5	4	4	2	4																																																												
T	0	6	0	6	0	1	0	0	0	0	0	0	0																																																												

Ryc. 17: Sekwencja motywu TATA-box w zapisie tradycyjnym (a) oraz w formie PWM (b).

Najbardziej popularną formą zapisu niespecyficznych motywów są macierze wag pozycji (ang. Position-Weight-Matrix - PWM). Jej najprostsza forma tworzona jest poprzez zliczenie wystąpień danego nukleotydu na określonej pozycji zbioru różnych sekwencji motywów, która przechowywana jest w tabeli (Ryc. 17b). Inne metody obejmują bardziej wyrafinowane sposoby określania częstotliwości wystąpień [209] lub nawet sieci neuronowe gdzie określone elementy macierzy reprezentują wagi poszczególnych neuronów [210].

Głównym problemem związanym z wykorzystaniem motywów w formie PWM nie jest jednak sposób ich tworzenia ale metody przeszukiwania sekwencji, od stosunkowo prostych bazujących na sumowaniu określonych elementów macierzy [211], po bardziej wyrafinowane bazujące na sieciach Bayesowskich [212] czy modelach Markowa [213]. Najpopularniejsze aplikacje obejmują narzędzia takie jak ConSite [203], ConTra [214], COTRASIF [215], rVista [216] oraz wiele innych. Narzędzia tego typu w większości zbudowane są w postaci serwisu internetowego i ograniczają się do analizy niewielkich zbiorów danych lub nawet pojedynczych par region promotora - czynnik transkrypcyjny co znacznie obniża ich użyteczność w przypadku wielkoskalowych analiz sekwencji.

Specyficzne sekwencje niespecyficzne dopasowanie

Niespecyficzność dopasowania pojawia się najczęściej podczas mapowania genów do obszarów genomu lub podczas poszukiwania miejsc oddziaływania pomiędzy cząsteczkami microRNA a mRNA. Ten typ analizy jest najbardziej czasochłonny ze względu na potrzebę porównywania specyficznego wzorca z różnymi obszarami sekwencji, które nie muszą być do niego w pełni komplementarne. Tego typu analizę przeprowadza się najczęściej w oparciu o tzw. algorytmy dopasowania (z ang. alignment). Algorytmy dopasowania są bardzo użyteczne w przypadku gdy przeszukiwana sekwencja może różnić się nieco od badanego wzorca np. na skutek mutacji, delecji a także insercji fragmentów sekwencji.

Najpopularniejsze z nich to algorytm Smitha-Watermana wykorzystywany do poszukiwania dopasowania lokalnego w przypadku gdy jedna z sekwencji jest częścią innej znacznie dłuższej. Drugim jest algorytm Nedelmana-Wunsha służący do poszukiwań globalnych, najbardziej użyteczny w przypadku porównywania dwóch sekwencji o zbliżonej długości lub gdy dopuszczamy występowanie dużych przerw w jednej z nich.

Działanie tego typu algorytmów polega na porównywaniu sekwencji wzorca z różnymi pozycjami badanej sekwencji. Dla każdej z nich algorytm oblicza współczynnik dopasowania, który zależy od ilości idealnie sparowanych nukleotydów, ilości niedopasowań oraz ilości i długości przerw jakie zostały wstawione przez program do jednej z sekwencji w celu zwiększenia stopnia dopasowania pozostałych jej fragmentów. O tym w jaki sposób obliczany jest współczynnik decyduje typ i parametry użytego algorytmu dobierane na podstawie specyfiki badanych sekwencji oraz samej hipotezy badawczej.

Najpopularniejszym programem do lokalnego dopasowania jest BLAST (Basic Local Alignment Search Tool) [217], który pozwala na porównywanie zarówno sekwencji aminokwasowych jak i nukleotydowych z dokładnością uzależnioną od wykorzystanej wersji algorytmu oraz użytych parametrów. BLAST a szczególnie jedna z jego odmian – BLAT (BLAST-Like Alignment Tool) [218] jest algorytmem na tyle szybkim, że idealnie nadaje się do przeszukiwania sekwencji genomów setki tysięcy razy w relatywnie krótkim czasie.

Innym bardzo często wykorzystywanym typem dopasowania jest tzw. dopasowanie typu MSA (ang. Multiple Sequence Alignment) polegające na porównywaniu kilku stosunkowo krótkich sekwencji pomiędzy sobą w celu zobrazowania różnic wynikających z niewielkich zmian w sekwencji. Najpopularniejszym algorytmem tego typu jest Clustal [219], którego niezwykła przydatność wynika z możliwości przeprowadzania bardzo szybkiego dopasowania pomiędzy dużą grupą sekwencji oraz tworzenia drzewa filogenetycznego obrazującego zależności ewolucyjne pomiędzy nimi.

3.10.2. Publiczne bazy danych sekwencji nukleotydowych

Pierwsze bazy danych sekwencji nukleotydowych pojawiały się w latach 70-tych ubiegłego wieku jeszcze przed erą komputerów osobistych, jednak ich największy rozwój zaczął się w latach 90-tych wraz z rozpoczęciem projektu badania ludzkiego genomu (Human Genome Project) polegającego na identyfikacji i opracowaniu wszystkich 20-25 tysięcy ludzkich genów oraz określeniu sekwencji DNA w ludzkim genomie. W związku z rozwojem technologii sekwencjonowania oraz innych wielkoskalowych metod biologii molekularnej potrzeba gromadzeniem przechowywania oraz udostępniania ogromnych zbiorów danych stała się niezbędna, przyczyniając się do rozwoju wielu metod i standardów przechowywania danych biologicznych, jakie stosowane są do dziś. Jednym z najpopularniejszych standardów jest FASTA służący do przechowywania sekwencji nukleotydowych i aminokwasowych najróżniejszego typu. Struktura zapisu FASTA jest bardzo prosta, składa się z nagłówka rozpoczynającego się znakiem większości „>”, który zawiera podstawowe informacje o identyfikatorze danej sekwencji oraz samej sekwencji umieszczonej w następnych liniach, kodowanych zgodnie ze wspomnianym wcześniej standardem IUPAC.

Pierwszą bazą danych sekwencji nukleotydowych DNA i RNA, która do dziś funkcjonuje jest GenBank [220]. Założony w 1983 GenBank jest obecnie największym zbiorem tego typu danych stanowiąc źródło dla innych zbiorów takich jak Reference Sequence (RefSeq) czy Protein Knowledge Database

(UniProtKB). GenBank przechowuje informacje o sekwencjach RNA oraz kodujących je fragmentach DNA stanowiąc bezcenne źródło informacji o budowie i położeniu wszystkich znanych genów. GenBank wykorzystuje kilka formatów zapisu w tym FASTA i własny bardziej skomplikowany format zapisu dodatkowych informacji o sekwencjach. Składa się on z identyfikatorów poszczególnych typów danych rozpoczynających nowe bloki informacji takie jak LOCUS, SOURCE, REFERENCE zawierających odpowiednio identyfikator sekwencji, nazwę organizmu, odnośniki literaturowe i wiele innych. Istotną wadą GenBanku jest jednak to, iż wprowadzone do bazy danych rekordy są własnością ich autorów i nie mogą być aktualizowane przez innych naukowców, przez co wszelkie zmiany wynikające z np. ze zwiększonej dokładności metod sekwencjonowania, które wymagają stworzenia nowego rekordu sprawiają, że baza danych jest wysoce redundantna z punktu widzenia poszczególnych genów.

Odpowiedzią na problem redundancji jest baza danych RefSeq [221], która podobnie jak GenBank prowadzona jest przez NCBI (National Center for Biotechnology Information). RefSeq jest zbiorem sekwencji genomów, transkryptów i białek połączonych z najróżniejszymi informacjami pochodzącymi z innych zbiorów dostarczając informacji o pełnych nazwach genów, domenach białkowych, właściwościach enzymatycznych, fenotypach, powiązanych chorobach, artykułach naukowych i wielu innych.

EntrezGene jest bazą danych genów, która z kolei stanowi pomost pomiędzy rekordami z RefSeq zawierającymi informacje o kodowanych przez dany gen transkryptach, produktach białkowych oraz obszarach genomu, z których dany gen pochodzi [222]. Wszystkie rekordy w bazie EntrezGene są podobnie jak w przypadku RefSeq na bieżąco aktualizowane zapewniając bardzo wysoką jakość danych i eliminując problem redundantnych lub nieaktualnych rekordów. EntrezGene będący własnością NCBI jest jedną z najczęściej wykorzystywanych baz danych informacji o genach dostarczając bezcennych informacji na temat ich roli i funkcji a także stanowiąc pomost do kilkudziesięciu innych baz danych uzupełniających te informacje.

Wszystkie przedstawione bazy danych są publicznie dostępne poprzez interfejs stron internetowych, zautomatyzowany system pobierania informacji SRS (ang. Sequence Retrieval System) [223] oraz w postaci pełnych plików z danymi dostępnymi za pośrednictwem kilku serwerów opartych o protokół FTP. Dodatkowo dane dostępne na serwerach NCBI są co 24 godziny synchronizowane z serwerami instytucji wchodzącymi w skład International Nucleotide Sequence Databases Collaboration (INSDC), do których poza NCBI należą European Molecular Biology Laboratory (EMBL) oraz DNA Data Bank of Japan (DDBJ). Ogólnodostępność danych biologicznych oraz duży wybór możliwości ich pobierania jest jedną z charakterystycznych cech INSDC, której jednym z podstawowych celów jest udostępnianie swoich zbiorów danych na potrzeby badań prowadzonych na całym świecie.

Innym bardzo cennym zbiorem informacji jest baza danych uniwersytetu w Kalifornii - UCSC (University of California, Santa Cruz) [224]. Jej główną cechą jest udostępnianie informacji o strukturze genomu kilku popularnych organizmów wraz z dokładnym położeniem najróżniejszych elementów sekwencji, poprzez bardzo wydajny mechanizm dostępny za pośrednictwem strony internetowej (z ang. Genome Browser). Dodatkowo z bazy danych UCSC można pobrać pełne sekwencje genomów oraz wszelkie informacje odnośnie położenia i pochodzenia zmapowanych elementów sekwencji w tym sekwencji genów, transpozonów, miejsc rozpoznawanych przez miRNA, położenie sekwencji rozpoznawanych przez sondy mikromacierzowe oraz wiele innych.

4. Materiały i metody

4.1. Źródła danych

4.1.1. Dane mikromacierzowe – zbiór testowy

Analiza odpowiedzi komórkowej na promieniowanie jonizujące przeprowadzona została przy wykorzystaniu mikromacierzy oligonukleotydowych firmy Affymetrix i Agilent. Eksperymenty wykonane zostały na komórkach Me45 (czerniak), K562 (białaczką) i 2 typach komórek HCT116 (rak jelita grubego) z prawidłową i znokautowaną wersją genu p53, badanych w różnym czasie po ekspozycji na promieniowanie jonizujące. Podstawowe cechy każdej linii komórkowej przedstawiono w Tab. 3.

Linia komórkowa	Opis	Białko p53	Mutacje	Zmiana cyklu kom. w odpowiedzi na promieniowanie
Me45	czerniak z przerzutami do węzłów chłonnych	produkowane, b.d. n. t. mutacji	b.d.	bez zmian
K562	białaczką limfoidalna	insercja C pomiędzy 406 a 407 nukleotydem części kodującej (utrata prawidłowej funkcji)	CDKN2A, TP53	zatrzymany w fazie G2
HCT116	nowotwór nabłonka jelita grubego	prawidłowa wersja	CDKN2A, CTNNB1, KRAS, MLH1, PIK3CA	zatrzymany w fazie G2
HCT116 p53 -/-	nowotwór nabłonka jelita grubego	brak	jw., brak danych o dodatkowych mutacjach po inaktywacji p53	zatrzymany w fazie G2

Tab. 3: Podstawowe cechy badanych linii komórkowych

Mikromacierze firmy Affymetrix wykorzystano do przeprowadzenia 9 eksperymentów (łącznie 95 próbek) w oparciu o platformy HG-U133A i HG-U133A_2. Wykonano także dodatkowe 2 eksperymenty (16 próbek) przy użyciu macierzy dwukanałowych Agilent typu SurePrint G3 Human GE 8x60K Microarray oraz jednokanałowych macierzy microRNA, Agilent Human_miRNA_V16.0.

Próbki oznaczono według następującego schematu: ID_LINIA_CZYNNIK_CZAS_POWTORZENIE, gdzie:

- ID - identyfikator eksperymentu
- LINIA - oznaczenie linii komórkowej
- CZYNNIK - oznaczenie badanego czynnika lub komórek kontrolnych
- CZAS - czas po zadziałaniu czynnika
- POWTORZENIE - numer powtórzenia technicznego/biologicznego

Przykładowo E01_Me45_IR_1h_1 to pierwsze powtórzenie techniczne próbki z eksperymentu E01 przeprowadzonego na komórkach Me45 poddanych działaniu promieniowania jonizującego i zbadanych po 1 godzinie.

Dokładny opis eksperymentów wraz z opisem badanych czasów po napromieniowaniu zebrano w Tab. 4.

Affymetrix - mRNA					
Identyfikator eksperymentu	Rok wykonania	Platforma	Ilość próbek	Komórki	Badane czasy po napromieniowaniu
E01	2003	HG-U133A	8	Me45	1h,12h, 24h
E02	2003		6	K562	36h
E03	2004		8	K562	1h,12h, 24h
E04	2004		6	Me45	36h
E05	2008		4	HCT116	1h
E06	2009		12	HCT116	1h, 36h
E07	2009		10	K562	1h,12h, 24h, 36h
E08	2009		20	HCT116	1h,12h, 24h, 36h
E09	2010	HG-U133A_2	20	HCT116	1h,12h, 24h
Agilent - mRNA					
A01	2012	SurePrint G3 Human GE	8	Me45, K562, HCT116	12h
A02	2012	8x60K Microarray	8	Me45, K562, HCT116	12h
Agilent - miRNA					
M01	2011	Human_miRNA_V16.0	8	Me45, K562, HCT116	12h
M02	2012		8	Me45, K562, HCT116	12h

Tab. 4: Opis wykonanych eksperymentów mikromacierzowych

Pierwsze eksperymenty wykonane na mikromacierzach Agilent (mRNA - A01 oraz miRNA - M01) nie spełniają wymagań stawianych przez kontrole jakości zdefiniowaną przez producenta. Z tego względu w pracy wykorzystano jedynie próbki z drugiego powtórzenia A02 i M02. Ze względu jednak na to, że pojedyncze powtórzenie jest niewystarczające do określenia znamienności statystycznej różnic, mikromacierze mRNA wykorzystano jedynie do porównania wyników uzyskanych z wykorzystaniem platformy Affymetrix.

Wszystkie eksperymenty wykonane zostały pod nadzorem prof. Joanny Rzeszowskiej Wolny, głównym wykonawcą eksperymentów E01-E09 był mgr inż. Robert Herok z Centrum Onkologii im. Marii Skłodowskiej Curie w Gliwicach, głównymi wykonawcami eksperymentów A01, A02, M01 i M02 byli dr inż. Anna Lalik oraz dr inż. Sebastian Student z Centrum Biotechnologii Politechniki Śląskiej.

4.1.2. Dane mikromacierzowe – zbiór referencyjny

Przetestowanie postawionej w niniejszej pracy hipotezy wymaga przeanalizowania dużego zbioru danych mikromacierzowych w celu uniezależnienia wyników badań od specyfiki danego eksperymentu oraz specyfiki określonej platformy. W tym celu zbudowano lokalną bazę danych eksperymentów mikromacierzowych składającą się z 10 różnych platform na podstawie danych opublikowanych w zbiorze ArrayExpress [152]. Algorytm budowy bazy danych składa się z następujących etapów:

- Za pomocą odpowiedniego zapytania do bazy danych stworzono listę dostępnych eksperymentów dla każdej platformy
- Z listy odrzucono eksperymenty, dla których nie są dostępne surowe dane oraz nieliczne eksperymenty składające się ponad 1000 próbek ze względu na ograniczenia sprzętowe. Tego typu eksperymenty zwykle są zbiorem próbek z innych eksperymentów także dostępnych w bazie ArrayExpress.

- Za pomocą serii zapytań wysyłanych za pośrednictwem protokołu FTP pobrano dane z serwera ArrayExpress dla każdego osobnego eksperymentu
- Dane z każdego pojedynczego eksperymentu rozpakowano i usunięto wszystkie pliki inne niż *.CEL
- Pozostałe pliki *.CEL przekonwertowano do formatu binarnego XDA w celu ograniczenia ich rozmiarów
- Z każdego pliku wyciągnięto nagłówki z opisem próbki a następnie na jego podstawie usunięto wszystkie pliki nienależące do wybranej platformy (w sytuacji gdy dany eksperyment obejmuje więcej niż jeden rodzaj mikromacierzy)

Całą procedurę realizuje skrypt napisany w języku Python dodatkowo korzystający z komend powłoki systemu operacyjnego Fedora.

Podczas tworzenia bazy danych wybrano 10 najpopularniejszych platform mikromacierzowych służących do badania ekspresji genów, różniących się założeniami konstrukcyjnymi oraz zaprojektowanymi na podstawie genomów różnych organizmów (Tab. 5). Przy wyborze platform dodatkowo kierowano się ilością dostępnych eksperymentów w bazie ArrayExpress. Użyte platformy to:

- HG_U95Av2 – to najstarsza z badanych mikromacierzy typu 3'IVT (sondy specyficzne dla końca 3' transkryptu) przeznaczona do badania ludzkiego transkryptomu, zaprojektowana na podstawie bazy danych UniGene w wersji 95 (luty 1999r.)
- HG-U133A, HG-U133B, HG-U133A_2, HG-U133_Plus_2 – najpopularniejsze macierze typu 3'IVT zaprojektowane na podstawie bazy UniGene w wersji 133 (kwiecień 2001r.). Macierze HG-U133A i HG-U133B pochodzą z 2001 roku i zawierają różne zbiory sond wzajemnie się uzupełniające. Macierz HG-U133A_2 zawiera te same sondy co HG-U133A (za wyjątkiem dodatkowych sond kontrolnych), zbudowana jest ona jednak w oparciu o nowszą technologię wykorzystującą sondy o mniejszej powierzchni. Macierz HG-U133_Plus_2 jest połączeniem macierzy HG-U133A i HG-U133B z dodatkowymi zestawami sond zaprojektowanymi w oparciu o bazę danych UniGene w wersji 159 (styczeń, 2003r.).
- RG_U34A i Rat230_2 to najpopularniejsze mikromacierze służące do badania szczurzego transkryptomu RG_U34A to najstarsza z użytych platform typu 3'IVT zaprojektowana na podstawie 34 wersji bazy danych UniGene (listopad 1998r.). Rat230_2 jest znacznie nowszą platformą zbudowana w oparciu o UniGene w wersji 99 (czerwiec 2002r.)
- Mouse430_2 jest odpowiednikiem macierzy Rat230_2 dla mysiego transkryptomu oparta o UniGene w wersji 107 (czerwiec 2002r.)
- HuGene-1_0-st i MoGene-1_0-st-v1 to znacznie nowsze platformy mikromacierzowe firmy Affymetrix do badania transkryptomu, zaprojektowane na podstawie wszystkich eksonów składających się na sekwencje genu a nie jedynie jego niekodującego końca 3'. Macierzy RaGene-1_0-st nie wykorzystano ze względu na niewielką liczbę opublikowanych eksperymentów (<50)

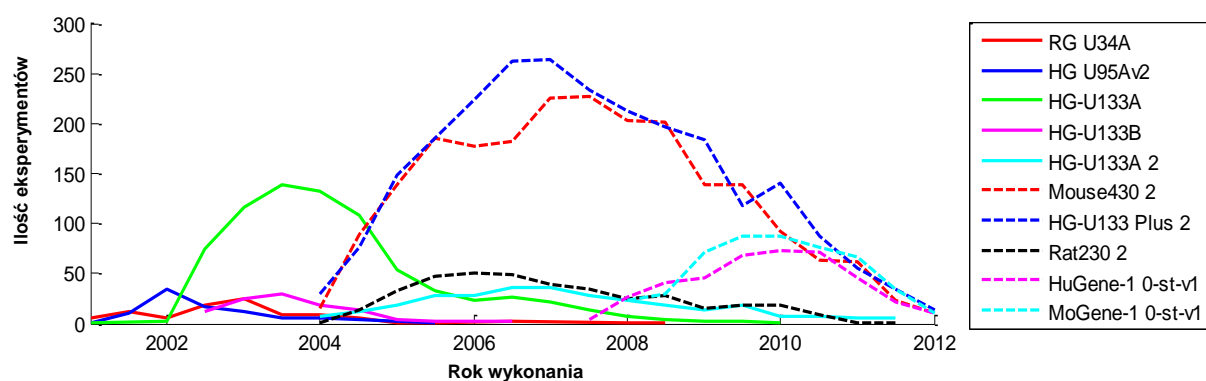
Wybór powyższych platform pozwala na testowanie wielu hipotez związanych z różnicami w typie mikromacierzy, gęstości upakowania sond, dokładności bazy danych sekwencji, na podstawie której

zaprojektowano sondy oraz różnic międzygatunkowych. Istotne jest, że wszystkie platformy są nadal w sprzedaży chociaż HG_U95Av2 i RG_U34A wyłącznie na indywidualne zamówienie bezpośrednio u producenta, ponadto większość z platform jest nadal często wykorzystywana w badaniach.

Platforma	Organizm	Typ	Informacje o sekwencjach	Rozmiar macierzy (sondy)	Statystyki lokalnej bazy danych		
					Eksperymenty	Próbki	Próbki na eksperyment
RG_U34A	szczur	3'IVT	1998	534 x 534	98	2845	29
HG_U95Av2	człowiek	3'IVT	1999	640 x 640	95	2651	28
HG-U133A	człowiek	3'IVT	2001	712 x 712	759	28202	37
HG-U133B	człowiek	3'IVT	2001	712 x 712	111	5151	46
HG-U133A_2	człowiek	3'IVT	2001	732 x 732	292	6442	22
Mouse430_2	mysz	3'IVT	2002	1002 x 1002	2174	33762	16
HG-U133_Plus_2	człowiek	3'IVT	2003	1164 x 1164	2466	63597	26
Rat230_2	szczur	3'IVT	2003	834 x 834	385	9948	26
HuGene-1_0-st	człowiek	eksonowa	2006	1050 x 1050	407	7679	19
MoGene-1_0-st-v1	mysz	eksonowa	2007	1050 x 1050	489	6229	13

Tab. 5: Informacje o wykorzystanych platformach mikromacierzowych oraz statystyki stworzonej bazy danych eksperymentów

Średnio jeden na sto eksperymentów zawierał uszkodzone pliki CEL wynikające z obciętej długości pliku, nieprawidłowych wartości czy struktury. Najczęstszym problemem były błędy powstałe prawdopodobnie podczas wysyłania danych na serwer ArrayExpress przez ich autorów. Tego typu nieprawidłowości wychwytywane były przez algorytm analizy danych odpowiedzialny za wyodrębnianie i usuwanie uszkodzonych plików. Ostatnia aktualizacja bazy danych eksperymentów przeprowadzona została 3 września 2012r. Łączny czas potrzebny na stworzenie pełnej bazy danych to około 6 dni, całkowity rozmiar plików bazy danych (po konwersji do formatu XDA) to 1.6TB.



Ryc. 18: Ilość eksperymentów wykonanych w poszczególnych latach dla danej platformy mikromacierzowej firmy Affymetrix

Dodatkowo w pracy wykorzystano cztery zbiory specjalne:

Affy-HuGene – to zestaw danych z mikromacierzy typu HuGene-1_0-st wykonanych w laboratorium firmy Affymetrix. Zestaw ten zawiera 33 mikromacierze, którymi zbadano RNA wyizolowane z 11 różnych tkanek, każda z nich posiada 3 powtórzenia biologiczne. Dane pobrano ze strony:

http://www.affymetrix.com/support/technical/sample_data/gene_1_0_array_data.affx

MAQC-133P2 – to zestaw bazujący na macierzach HG-U133_Plus_2 stworzony w ramach projektu MicroArray Quality Control (MAQC) [225]. Zestaw składa się ze 120 mikromacierzy którymi zbadano 4 próbki, każdą w 5 powtórzeniach technicznych, dodatkowo wykonanych w 6 różnych laboratoriach. Dane pobrano z bazy danych Gene Expression Omnibus na podstawie identyfikatora: GSE5350.

GoldenSpike – zestaw danych testowych opracowany przez grupę Marca Halfona w 2004 roku [226]. Składa się on z 6 mikromacierzy typu DrosGenome1 z sondami specyficznymi dla transkryptomu muszki owocowej (*Drosophila melanogaster*). Zbiór składa się z trzech powtórzeń technicznych przeprowadzonych na dwóch zestawach RNA przy czym do ich stworzenia wykorzystano tą samą pulę cDNA do jednej z nich dodano jednak mieszaninę kilkudziesięciu cDNA o różnych znanych proporcjach, co pozwala na określenie, które geny powinny różnić się poziomem ekspresji pomiędzy próbkami oraz jak silna powinna być ta różnica.

PlatinumSpike – jest to ulepszona wersja zbioru GoldenSpike opracowana w 2010r [227]. Zbiór ten zawiera dane z 18 mikromacierzy typu Drosophila_2 będących następcą mikromacierzy z sondami specyficznymi dla transkryptomu *Drosophila melanogaster* DrosGenome1. W tym przypadku cDNA w znanych ilościach dodano do obu grup próbek po to aby uzyskać zarówno geny o zmniejszonej jak i zwiększonej ekspresji co jest zgodne z założeniami metod standaryzacji danych. Ponadto poza 3 powtórzeniami technicznymi wykonano także po 3 powtórzenia biologiczne na każdej parze próbek.

4.1.3. Sekwencje nukleotydowe i dane adnotacyjne

Na potrzeby analizy sekwencji nukleotydowych transkryptów pobrano pełną bazę danych 38625 sekwencji z serwera EMBL przy wykorzystaniu aplikacji NucleoSeq [228]. Tylko transkrypty kodujące strukturę białka zostały wykorzystane do analizy (w sumie 30853) pomijając niekodujące RNA (ncRNA). Fragmenty sekwencji nukleotydowych o długości 5000 nukleotydów położone po obu stronach sekwencji genów wyciągnięte zostały za pomocą napisanego w tym celu skryptu bezpośrednio z genomu w wersji hg19/GRCh37 pobranego z bazy danych UCSC [224]. Do tego celu wykorzystano koordynaty transkryptów RefSeq także pobrane z bazy UCSC.

Sekwencje mikroRNA pobrane zostały z bazy danych miRBase [229] w wersji 18 wraz z informacjami o ich położeniu w genomie oraz budowie sekwencji prekursorowych miRNA (pre-miRNA). Sekwencje motywów rozpoznawanych przez ludzkie czynniki transkrypcyjne w formie macierzy wag-pozycji pobrano z bazy danych Jaspar [230]. Informacje o budowie znanych ścieżek sygnałowych pobrano z baz danych KEGG [127] oraz Panther [128] za pomocą aplikacji NucleoAnnot.

4.2. Analiza sekwencji nukleotydowych

Analizę sekwencji nukleotydowych, za wyjątkiem poszukiwania miejsc oddziaływania z miRNA, przeprowadzono przy wykorzystaniu aplikacji NucleoSeq. W celu zwizualizowania wyników posłużono się skryptami, napisanymi w środowisku Matlab, na potrzeby pracy. Poszczególne metody analizy opisano w punktach poniżej.

4.2.1. Analiza składu nukleotydowego

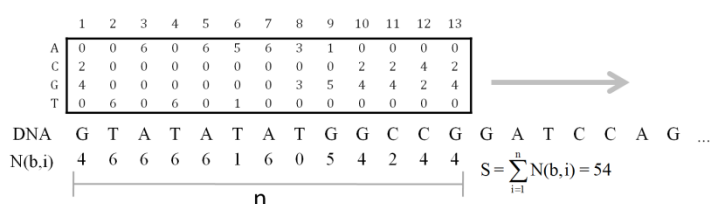
Skład nukleotydowy określany jest jako stosunek ilości nukleotydów GC do długości sekwencji (w skali procentowej), wyznaczonej z pominięciem nieznanymi nukleotydów, oznaczanych literą N w sekwencjach genomu. Znamienność różnic w proporcjach GC pomiędzy sekwencjami określano za pomocą standardowego testu na porównanie proporcji z korektą na wielokrotne testowanie Storey'a [198] i poziomem istotności 0.01. Korelacje w składzie nukleotydowym pomiędzy grupami sekwencji określano za pomocą testu Spearmana ze względu na brak zgodności badanych zbiorów danych z rozkładem normalnym, ocenianym za pomocą testu zgodności dopasowania Jarque-Bery.

Wykresy częstotliwości występowania nukleotydów GC wykonano poprzez zliczenie ilości nukleotydów na danej pozycji każdej z badanych sekwencji. Ze względu na brak zgodności w długości sekwencji należących do niektórych zbiorów, wszystkie wartości ustandaryzowano dzieląc je przez liczbę sekwencji, która osiąga określoną długość.

4.2.2. Macierze wag pozycji

Macierze wag pozycji (z ang. Position Weight Matrix - PWM) powstają na podstawie określonej liczby motywów sekwencyjnych, które odpowiedzialne są za wspólne procesy regulacyjne, np. oparte o przyłączanie białek z rodziny czynników transkrypcyjnych. Przykład macierzy wag pozycji oraz odpowiadający jej zbiór sekwencji przedstawiono na Ryc. 17.

Podejście wykorzystane w niniejszej pracy do poszukiwania sekwencji zbliżonych do motywu w formacie PWM oparte jest o jedną z najczęściej wykorzystywanych metod. Polega ona na przesuwaniu o 1 nukleotyd motywu wzdłuż sekwencji oraz wyznaczaniu współczynnika dopasowania określającego podobieństwo danego fragmentu sekwencji do zbioru sekwencji wykorzystanego do utworzenia macierzy PWM. Współczynnik ten jest wyznaczany poprzez sumowanie wartości z kolejnych kolumn macierzy PWM dla poszczególnych pozycji i określonych nukleotydów obecnych w badanym fragmencie sekwencji [203], co zilustrowano na Ryc. 19.



Ryc. 19: Przykład wyznaczania współczynnika dopasowania S na podstawie ilości zliczeń $N(b,i)$, gdzie i określa pozycje natomiast b typ nukleotydu. W przykładzie wykorzystano macierz PWM motywu TATA-box

Im wyższy jest współczynnik dopasowania tym podobieństwo pomiędzy badaną sekwencją a motywem jest większe. Macierze PWM są jednak różnej długości, z tego względu liczba motywów użyta ich stworzenia może być różna. Konieczne jest zatem ustandaryzowanie współczynnika dopasowania po to aby możliwe było wprowadzenie uniwersalnego progu odcięcia, powyżej którego dopasowanie uznawane jest za wystarczająco silne do powstania wiązania. W tym celu wykorzystano metodę zaproponowaną w pracy [211], polegającą na wyznaczeniu procentu danego współczynnika w stosunku do maksymalnej możliwej do uzyskania wartości:

$$S_{norm} = \frac{S - S_{min}}{S_{max} - S_{min}} \cdot 100 \quad (1)$$

gdzie S_{min} oraz S_{max} to odpowiednio najmniejsze i największe wartości współczynnika dopasowania dla określonej macierzy PWM.

Znormalizowana w ten sposób wartość porównywana jest z zdefiniowanym przez użytkownika progiem odcięcia (zwykle 80%) pozwalając na pozostawienie jedynie tych miejsc dopasowania, które charakteryzują się wysoką zgodnością z motywem. Główną zaletą tej metody jest jej prostota, co pozwala na przeanalizowanie ogromnych zbiorów sekwencji w krótkim czasie, bez potrzeby korzystania z komputerów o dużej mocy obliczeniowej.

Pomimo, że przedstawiona metoda sprawdza się bardzo dobrze dla różnych zbiorów danych kilka problemów matematycznych pojawia się w przypadku niektórych motywów PWM. Macierze PWM powstają nieraz z niewielkiej liczby motywów i często nie wszystkie typy nukleotydów występują na danej pozycji co prowadzi do pojawiania się zer w macierzy PWM. Powoduje to problemy w przypadku konwersji macierzy do skali logarytmicznej. Jednym z możliwych rozwiązań jest zastosowanie przekształcenia zaproponowanego w [231]:

$$W_{b,i} = \log \frac{P_m(b,i)}{P_b(b)} \quad (2)$$

gdzie $P_b(b)$ to prawdopodobieństwo pojawiania się zasady określonego typu w danej sekwencji (w większości przypadków $P_b(b)=0.25$ dla $b=1,\dots,4$) oraz $P_m(b,i)$ to skorygowane prawdopodobieństwo pojawiania się zasady b na pozycji i w motywie długości m , wyznaczone w następujący sposób:

$$P_m(b,i) = \frac{N(b,i)}{n} + \epsilon \quad (3)$$

gdzie $N(b,i)$ to zliczenia wynikające z macierzy PWM dla zasady b na pozycji i natomiast ϵ jest parametrem, który zapobiega problemowi związanemu z przekształceniem logarytmicznym ($\epsilon = 0.01$). Końcowa wartość współczynnika przed normalizacją będzie w takiej sytuacji sumą wartości $W_{b,i}$ dla każdej zasady w badanej sekwencji:

$$S = \sum_{i=1}^n W_{b,i} \quad (4)$$

Wpływ zastosowanej metodologii oraz jej warianty wraz z różnicami w wartościach parametrów przetestowano w ramach pracy [205].

4.2.3. Miejsca wiążące miRNA

Miejsca wiązania cząsteczek miRNA w transkryptach ludzkich genów wyznaczone zostały za pomocą programów miRanda [232] TargetScan [233] oraz PicTar [234]. Ze względu na ogromną ilość danych opisujących oddziaływanie między parami mRNA-miRNA przygotowano 2 bazy danych wyposażone w interfejs wspomagający wygodny dostęp do danych. Pierwsza z baz danych oparta jest o skrypt przygotowany w języku VBA (Visual Basic for Applications) dla Excela. Druga natomiast bazuje na bazie danej Microsoft Access, do której dostęp zapewnia aplikacja napisana w języku Delphi. Aplikację tą

wyposażono w wygodny edytor zapytań SQL pozwalający na bardzo szybki dostęp do wybranych informacji w oparciu o specyficzne kryteria jakości dopasowania oparte o wyniki z różnych metod poszukiwania miejsc oddziaływania z miRNA.

4.3. Analiza danych mikromacierzowych

Wstępne przetwarzanie i kontrola jakości

Dane z mikromacierzy Affymetrix w większości analizowano przy wykorzystaniu środowiska R. Kontrola jakości danych przeprowadzona została za pomocą programów *affyQCReport* oraz *arrayQualityMetrics* zaimplementowanych w Bioconductorze w oparciu o pakiet *simpleaffy* [168]. Metody wstępnego przetwarzania danych przeprowadzone zostały za pomocą algorytmów RMA [138], GCRMA [235], PLIER [175], MAS5 [173], FARMS [176] oraz MBEI [177] w oparciu o ich implementacje w Bioconductorze. Dane zostały przeanalizowane przy wykorzystaniu uaktualnionych wersji plików CDF zgodnie z metodologią opisaną w [148] z zestawami sond specyficznymi dla transkryptów z bazy Reference Sequence [221]. Kompensacja tzw. efektu partii (z ang. batch effect) wykonana została za pomocą oprogramowania ComBat [236]. Geny różnicujące wyznaczono za pomocą oprogramowania Limma [142] z korektą na wielokrotne testowanie Storey'a [198].

Dane z mikromacierzy Agilent przetworzono za pomocą metody AFE-TGE (Affymetrix Feature Extraction – Total Gene Signal) zaproponowanej przez producenta mikromacierzy. Dodatkowo jako porównanie wykorzystano metodę RMA (w przypadku mikromacierzy miRNA zaimplementowaną w pakiecie R *AgiMicroRna* [237]) oraz normalizację Loess zaproponowaną w [238, 239] jako alternatywę dla AFE-TGE.

Wielkoskalowa analiza wpływu składu GC na zmianę poziomu ekspresji

Cechy sekwencji nukleotydowych sond określono na podstawie informacji o ich sekwencjach pobranych z bazy danych plików CDF Brainarray [148]. Dla każdego zestawu danych mikromacierzowych przeprowadzono wstępne przetwarzanie zgodnie z jedną z 6 metod (RMA, GC-RMA, MAS5, FRAMS, PLIER, MBEI) z pominięciem etapu sumaryzacji. Na podstawie przetworzonych w ten sposób danych określono współczynnik nachylenia linii regresji dopasowanej do wartości określających zależność składu GC sondy od poziomu ich sygnału. Dodatkowo określono medianę poziomu ekspresji sond o różnych proporcjach GC oraz średni poziom sygnału całej mikromacierzy. W oparciu o uzyskane wartości wykonano kolejny etap analizy, w którym dla każdego zbioru danych przeprowadzono pełne przetwarzanie kolejno sześcioma wcześniej wybranymi metodami, łącznie z etapem sumaryzacji. Do wartości poziomu ekspresji każdego zestawu sond specyficznego dla określonego transkryptu z bazy danych RefSeq dopasowano wartości składu GC jego kompletnej sekwencji. Następnie dla każdej unikatowej pary dwóch próbek w danych z określonego eksperymentu wyznaczono logarytm stosunku każdych dwóch wartości poziomu ekspresji (LFC) zestawów sond oraz określono poziom korelacji LFC z wcześniej określonym składem GC transkryptu. Tego typu dane powiązано z różnicami w kącie nachylenia linii regresji pomiędzy tymi samymi próbkami obliczonych na podstawie danych uzyskanych w pierwszym etapie analizy. W ten sposób dla każdego zbioru danych i każdej unikatowej pary próbek do niego należących uzyskano dwie wartości: korelację pomiędzy zmianą poziomu ekspresji a składem GC transkryptu oraz różnicę w kącie

nachylenia linii regresji dopasowanej do przetworzonych danych. Wartości te wykorzystano do określenia współczynnika korelacji dla każdego zbioru danych z pojedynczego eksperymentu, który założono, że określa zależność wyników eksperymentu od różnic w proporcjach GC pomiędzy próbkami.

Z uwagi na bardzo dużą ilość danych mikromacierzowych wykorzystanych w tym etapie obliczenia przeprowadzono na klastrze obliczeniowym Ziemowit. Ponieważ dużym ograniczeniem obliczeń równoległych była w tym przypadku ilość pamięci RAM na każdym węźle obliczeniowym, niezbędna do przetworzenia każdego ze zbiorów danych (bardzo trudna do oszacowania przed analizą), poszczególne zbiory danych wymagające bardzo dużych ilości pamięci (zawierających kilkaset próbek) były pomijane podczas obliczeń równoległych w przypadku gdy pozostałe, aktualnie analizowane zbiory były na tyle duże, że dostępna ilość pamięci RAM była nie wystarczająca. Tego typu zestawy danych były następnie analizowane na samym końcu analizy, sekwencyjnie. Całkowita analiza danych przeprowadzona została na podstawie własnych skryptów w środowisku Python, dodatkowo korzystających z poleceń powłoki jądra systemu Linux oraz skryptów w języku R uruchamianych za pośrednictwem pakietu rpy2 dla języka Python.

Budowa drzewa decyzyjnego oraz określenie wpływu poszczególnych cech na jego strukturę

Drzewo decyzyjne zbudowano na podstawie pięciu cech zestawów sond oraz średniej wartości wariancji sygnału sond w zestawie na przestrzeni wszystkich analizowanych próbek. Cechy zestawów sond zdefiniowano w następujący sposób:

- Skład GC sond – im więcej jest sond o skrajnych proporcjach składu GC tym większa powinna być wariancja sygnału. Jako miarę zróżnicowania składu GC przyjęto wariancję składu GC sekwencji sondy w obrębie zestawów. Jednak podczas obliczania wariancji składu GC zamiast średniej wartości GC obliczonej dla określonego zestawu wykorzystano medianę składu GC wszystkich sond danej mikromacierzy.
- Położenie sekwencji rozpoznawanej przez sondy – wariancja sygnału powinna być tym wyższa im więcej sond położonych w dużej odległości od siebie. W celu określenia tej zależności wyznaczono odległości sond danego zestawu od sondy położonej najbliżej końca 3' transkryptu (odległości wyznaczono na podstawie sekwencji z bazy danych RefSeq). Następnie wyznaczono wariancję położenia sond jednak nie wokół średniej wartości położenia ale jej 25 percentyla, ponieważ im więcej jest sond położonych w dużej odległości od grupy sond w okolicy końca 3' tym większy powinien być wpływ degradacji RNA na wariancję sygnałów w zestawach sond.
- Dopasowanie do różnych grup form splicingowych transkryptów określonego genu – w tym przypadku wariancja powinna być największa, gdy w zestawie występuje dużo sond dopasowanych do wielu różnych grup form splicingowych. Przyjęta miara wpływu dopasowania do wielu grup form splicingowych (S_d) została zdefiniowana w następujący sposób:

$$S_d = \frac{K_m \cdot N_k}{N_s} \quad (1)$$

Gdzie K_m to ilość sond w największej grupie sond dopasowanej do wspólnych form splicingowych, N_k – ilość grup splicingowych, N_s – ilość sond w zestawie. Współczynnik jest tym wyższy im więcej jest grup form splicingowych zawierających dużą ilość sond.

- Obecność motywu (A)n w transkrypcie – zmienna binarna, 1-zestaw zawiera sondy po obu stronach motywu (A)n o długości przynajmniej 24nt, 0-zestaw nie zawiera tego typu sond
- Obecność motywu CCGCCTCCC (T7 spacer) w sekwencji sondy – zmienna binarna, 1-zestaw zawiera przynajmniej jedną tego typu sondę, 0-zestaw nie zawiera tego typu sond

Do budowy drzewa decyzyjnego oraz określania wpływu poszczególnych cech na jego kształt wykorzystano środowisko Matlab. Ocena procentowego wpływu parametrów na kształt drzewa przeprowadzona została zgodnie z metodologią opisaną w pracy [240].

Analiza skuteczności algorytmów poszukiwania genów różnicujących

Ocenę jakości algorytmów wstępnego przetwarzania danych przeprowadzono w oparciu o zbiory danych GoldenSipke i PlatinumSpike (dokładny opis danych zamieszczono w punkcie 4.1.2), które zaprojektowano w taki sposób, że znana jest dokładna lista genów różnicujących. Z nią porównano wyniki identyfikacji genów różnicujących oparte o test-t, które uzyskano po zastosowaniu wybranej metody przetwarzania danych. Różnice pomiędzy metodami przedstawiono w postaci krzywych ROC, do sporządzenia których wykorzystano statystykę t. Całkowita analiza przeprowadzona została w środowisku Matlab na podstawie własnych skryptów obliczeniowych.

4.4. Eksperymenty RT-qPCR

Startery do reakcji PCR zaprojektowane zostały za pomocą aplikacji Primer3 [241] a ich specyficzność dodatkowo potwierdzono poprzez wykonanie dopasowania ich sekwencji do bazy danych transkryptów Reference Sequence za pomocą programu BLAST [217]. W ten sposób sprawdzono czy startery nie są specyficzne nawet w niewielkim stopniu (minimum 10 nukleotydów) do innych transkryptów, oraz czy namnażana sekwencja (amplikon) nie zawiera eksonu, który może prowadzić do powstania alternatywnych form splicingowych i w ten sposób generować więcej niż jeden produkt w reakcji PCR. Sekwencje wszystkich wykorzystanych starterów zebrano w Tab. 17.

Symbol Genu	Identyfikatory transkryptów	Nazwa	Sekwencja
DICER1	NM_030621 NM_001195573 NM_177438	DICER1_F	TTAGACTTGTAGGCACTCTTC
		DICER1_R	ACCTCTACTGGTATGTTGATG
CCNB1	NM_031966	CCNB1_F	GAGCATCTAAGATTGGAGAG
		CCNB1_R	GGTAATGTTGTAGAGTTGGTG
HNRNPO (AUF1)	NM_031370 NM_031369 NM_002138 NM_001003810	AUF1_F	GAGTGTAGATAAGGTCATGGA
		AUF1_R	CCTCTTATTGGTCTTGTGT
EIF2C2 (AGO2)	NM_012154 NM_001164623	AGO2_F	GGTTCTCCACCACGAGTTGC
		AGO2_R	GGACGTGATAGTGC GAAGGC
PARP1	NM_001618	PARP_F	GTGTGGGTACGGTGATCGGTA
		PARP_R	GCCTGCACACTGTCTGCATT
SLC25A37	NM_016612	SLC25A37_F1	GTCCCTCCTCTCTCTAAGG
		SLC25A37_R1	AGCCTCCTCCCTGGATCCT
		SLC25A37_F2	GACACGCACAAGCACACACA
		SLC25A37_R2	GCTGTGTTTAGCCCTCCAG

Ryc. 20: Sekwencje wykorzystanych starterów do reakcji RT-qPCR (F-forward, R-reverse)

Wszystkie eksperymenty wykonane zostały w 4-6 powtórzeniach technicznych. Do syntezy cDNA wykorzystano odczynniki firmy EURx natomiast reakcje PCR wykonano przy użyciu zestawu odczynników A&A Biotechnology z barwnikiem fluorescencyjnym EvaGreen.

Analizę statystyczną wyników przeprowadzono za pomocą napisanych w tym celu skryptów w środowisku Matlab. Detekcje wielkości odstających przeprowadzono za pomocą testu Dixona, natomiast znamienność statystyczną różnic pomiędzy poziomami ekspresji określono na podstawie testu-t.

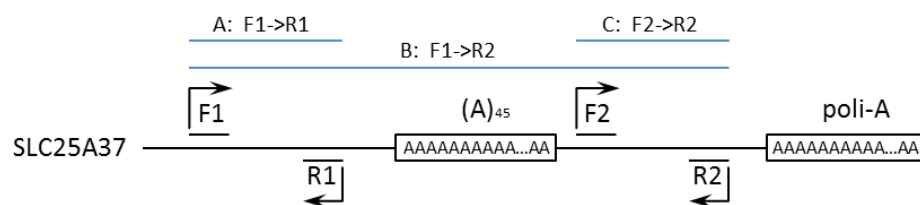
Analiza RT-qPCR obejmowała dwa oddzielne eksperymenty:

1. Walidacja wyników z eksperymentu mikromacierzowego

Poziomy ekspresji genów DICER1, CCNB1, AUF1, AGO2 oraz PARP1 określono w komórkach kontrolnych oraz w komórkach, z których RNA wyizolowano po 12 godzinach od poddanych ich działaniu promieniowania jonizującego. Poziomy ekspresji każdego z genów znormalizowano względem genu referencyjnego RPL41 (białko rybosomalne L41) a następnie wyznaczono poziomy ekspresji z wykorzystaniem metody Livaka [242]. Pomiar dla każdego indywidualnego genu znormalizowano względem komórek kontrolnych. Pomimo różnic w poziomach ekspresji pomiędzy kontrolami poszczególnych linii komórkowych standaryzacja do kontroli ułatwia interpretacje danych, szczególnie, że wszystkie pozostałe badania były wykonane w bezpośrednim odniesieniu do komórek kontrolnych.

2. Analiza produktów amplifikacji mRNA z motywem (A)_n

Startery zostały zaprojektowane dla genu SLC25A37, który zawiera w obszarze 3'-UTR motyw (A)_n składający się z 45 nukleotydów, w taki sposób aby ich parametry (temperatura topnienia, własności sekwencji amplikonów) były możliwie najbardziej zbliżone oraz aby jedna para obejmowała motyw (A)_n (produkt B) natomiast pozostałe dwie pary znajdowały się po jego obu stronach (produkty A i C), koncepcje tą ilustruje Ryc. 21.



Ryc. 21: Schemat rozmieszczenia starterów w eksperymencie z amplifikacją mRNA zawierającego motyw (A)_n. Sekwencje amplikonów powstałe w wyniku wykorzystania 3 kombinacji par starterów zaznaczono niebieskim kolorem.

Ponieważ analiza porównawcza wykonana została wyłącznie na podstawie jednego ekstraktu cDNA normalizacja w oparciu o gen referencyjny nie była wymagana. Poziomy ekspresji określono za pomocą metody Livaka [242] i odniesiono do ilości produktu B obejmującego motyw (A)_n. Doświadczenie wykonano na podstawie cDNA zsyntetyzowanego w oparciu o standardowy protokół oraz zmodyfikowaną wersją, w której użyto dwukrotnie większego stężenia starterów oligo-dT.

5. Konstrukcja oprogramowania

Ze względu na konieczność przeprowadzenia specyficznych obliczeń opracowano szereg aplikacji bioinformatycznych, które podzielić można na dwie zasadnicze grupy:

- Aplikacje zbudowane wyłącznie na potrzeby pracy – są to przede wszystkim zestawy skryptów napisanych w języku Python, R lub Matlab wykorzystywane do przeprowadzenia najbardziej czasochłonnnych obliczeń oraz do stworzenia wszystkich wykresów prezentowanych w niniejszej pracy. Tego typu programy wykorzystano do wstępnego przetwarzania danych mikromacierzowych, analizy sekwencji pod kątem miejsc wiązania miRNA, przeprowadzenia większości testów statystycznych, oraz gromadzenia i przetwarzania danych wykorzystywanych przez pozostałe programy.
- Aplikacje przeznaczone do wielu celów, publicznie dostępne – do tej grupy należą programy, które były napisane na potrzeby pracy jednak mogą być wykorzystane do innych celów, w dodatku wszystkie są wyposażone w wygodny interfejs graficzny. Programy z tej grupy napisane były w języku Delphi (wersja XE2). Pełne wersje programów dostępne są na stronie internetowej: www.bioinformatics.aei.polsl.pl. Do tej grupy należą programy służące do przetwarzania danych mikromacierzowych, analizy sekwencji nukleotydowych oraz poszukiwania informacji o funkcji genów i transkryptów.

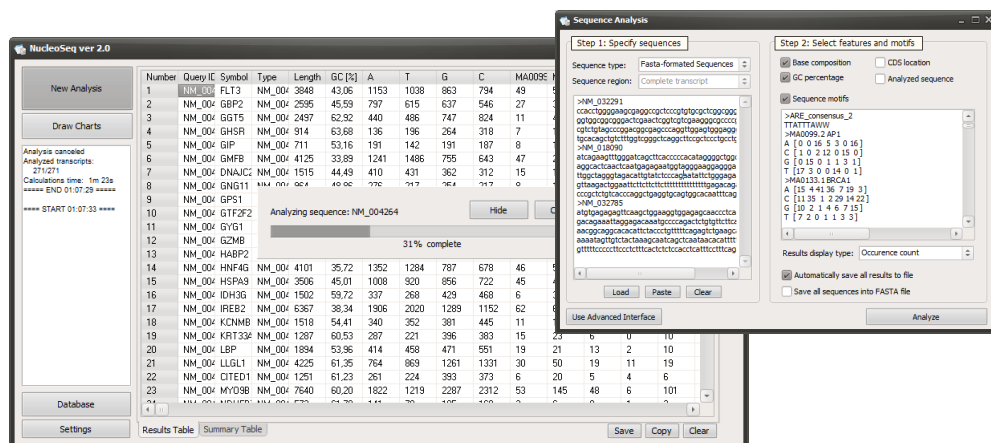
Głównym celem aplikacji z drugiej grupy jest nie tylko stworzenie funkcjonalnych i bardzo szybkich narzędzi bioinformatycznych pozwalających na przeprowadzanie skomplikowanych analiz na dużych zbiorach danych ale także ich propagowanie dzięki łatwemu i wygodnemu użytkowaniu. W przeciwieństwie do innych aplikacji tego typu nie wymagają one instalacji dodatkowych programów takich jak Matlab czy R lub środowisk uruchomieniowych typu .NET/Java-RE. Dodatkowo uruchamiane są one w pełni na komputerze użytkownika co eliminuje problemy związane z przesyłaniem dużych ilości danych na zewnętrzny serwer oraz pozwalają na wykorzystanie mocy obliczeniowej lokalnego komputera.

5.1. Analiza sekwencji nukleotydowych – NucleoSeq

NucleoSeq jest pierwszą ze stworzonych aplikacji pozwalającą na przeprowadzanie wielkoskalowych analiz sekwencji nukleotydowych DNA lub RNA [205]. To co wyróżnia NucleoSeq od innych podobnych programów to przede wszystkim możliwość automatycznego pobierania sekwencji nukleotydowych z bazy danych European Bioinformatics Institute (EBI) oraz University of California Santa Cruz (UCSC) za pośrednictwem Internetu a także ich bardzo szybka i wygodna analiza w oparciu o motywy sekwencyjne dowolnego formatu. NucleoSeq wyposażony jest w wygodny interfejs graficzny (Ryc. 22) ułatwiający proces analizy oraz pozwalający na natychmiastowe zwizualizowanie wyników jeszcze przed całkowitym zakończeniem obliczeń.

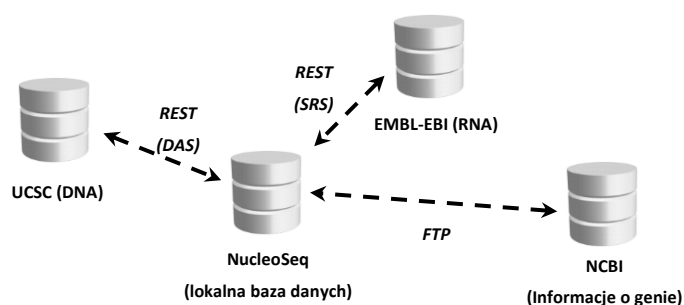
Sekwencje nukleotydowe będące przedmiotem analizy mogą być dostarczone przez użytkownika w formacie FASTA lub automatycznie pobrane z bazy danych Reference Sequence, publikowanej przez

EMBL, na podstawie identyfikatorów RefSeq, Ensembl lub EntrezGene. Powiązania między identyfikatorami zapewnia wewnętrzna baza danych symboli, która jest częścią programu, wraz z systemem automatycznej aktualizacji umieszczonych w niej danych za pośrednictwem Internetu. Pobrane sekwencje nukleotydowe transkryptów mogą być analizowane w całości lub z rozróżnieniem poszczególnych fragmentów takich jak sekwencje końca 3'/5' czy sekwencja kodująca. Dodatkowo wszystkie pobrane sekwencje transkryptów przechowywane są w lokalnej bazie danych programu na potrzeby przyszłych analiz co znacznie przyspiesza ich następane przetwarzanie. Baza danych sekwencji może być całkowicie bądź częściowo wyczyszczona w przypadku gdy sekwencje w niej przechowywane są już nieaktualne.



Ryc. 22: Interfejs programu NucleoSeq

NucleoSeq jest klientem usług internetowych (ang. web services client), który nie jest wyposażony w pełną bazę danych sekwencji nukleotydowych RNA i DNA a potrzebne informacje pobiera na bieżąco za pośrednictwem Internetu z zewnętrznych baz danych, działających w formie usług internetowych.



Ryc. 23: Bazy danych wykorzystywane przez NucleoSeq oraz protokoły użyte do przesyłania informacji

NucleoSeq wykorzystuje trzy bazy danych dwie z nich służą jako źródło sekwencji DNA i RNA natomiast trzecia dostarcza informacji na temat identyfikatorów sekwencji oraz położenia genów w sekwencji DNA (Ryc. 23). Sekwencje RNA pobierane są bezpośrednio z bazy danych Reference Sequence udostępnionej za pośrednictwem serwerów EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute). W tym celu wykorzystywany jest protokół SRS (Sequence Retrieval System) bazujący na formacie zapytań typu REST (z ang. Representational State Transfer). W oparciu o grupę identyfikatorów

transkryptów wysłanych do serwera zwracana jest odpowiedź zawierająca szczegółowe informacje na temat każdego z transkryptów, które z kolei przetwarzane są przez NucleoSeq i zapisywane w wewnętrznej bazie danych.

Sekwencje DNA pobierane są za pomocą protokołu DAS (Distributed Annotation System) z bazy danych UCSC (University of California Santa Cruz) na podstawie współrzędnych obszaru promotora genów obliczonych przez program w oparciu o położenia genu i jego orientacje na nici DNA. DAS jest jednym z najpopularniejszych formatów przesyłania danych biologicznych określającym ściśle reguły dotyczące samego zapytania wysłanego na serwer oraz format otrzymywanej odpowiedzi. Format zapytań podobnie jak w przypadku protokołu SRS bazuje na systemie REST jednak odpowiedź przesyłana z serwera ma w tym przypadku format XML, co ułatwia jej przetwarzanie na potrzeby analizy danych.

Informacje o identyfikatorach genów pobierane są jednorazowo z serwera NCBI (National Center for Biotechnology Information) za pośrednictwem protokołu FTP. Dane pobierane są automatycznie po uruchomieniu przez użytkownika procedury aktualizacji bazy danych a następnie przetwarzane i zapisywane w wewnętrznej bazie danych programu. Ponieważ baza danych jest indeksowana to po każdej aktualizacji wyszukiwanie w niej danych jest błyskawiczne, co daje znaczną przewagę nad wykorzystaniem zewnętrznego systemu konwersji symboli i informacji o położeniu genów, które nie są aktualizowane tak szybko jak sekwencje nukleotydowe RNA. W przeciwieństwie do sekwencji nukleotydowych baza danych symboli jest bardzo mała - kilka megabajtów.

Wykorzystanie serwisów internetowych pozwala na wyeliminowanie potrzeby rozprowadzania aplikacji razem z bardzo dużą bazą danych (ponad 4GB dla jednego organizmu) oraz eliminuje konieczność jej bezustannej aktualizacji, ponieważ sekwencje pobierane są bezpośrednio od źródła. Wadą tego rozwiązania jest jednak wydłużony czas analizy, który w przypadku sekwencji, jakie nie były wcześniej analizowane (nie zostały zapisane w wewnętrznej bazie danych programu) jest uzależniony od szybkości i niezawodności łącza internetowego a także parametrów samego serwisu internetowego (maksymalna liczba równoległych zapytań, minimalny czas odstępu pomiędzy zapytaniami etc.) Dodatkową wadą jest zależność aplikacji od dostępności samego serwisu internetowego jednak możliwość analizowania sekwencji zapisanej na komputerze w formacie FASTA pozwala ją częściowo wyeliminować.

Wszystkie sekwencje nukleotydowe mogą być analizowane w oparciu o specyficzne motywy sekwencyjne jak i te niespecyficzne zdefiniowane w postaci kodu IUPAC lub za pomocą macierzy wag pozycji w formacie bazy danych Jaspar [230]. Możliwość przeszukiwania sekwencji w oparciu o motywy typu PWM pozwala na odnajdywanie potencjalnych miejsc wiązania białek regulatorowych (w tym czynników transkrypcyjnych) zgodnie ze zdefiniowaną przez użytkownika czułością algorytmu.

Dodatkowe funkcje programu pozwalają na poszukiwanie motywów w zrandomizowanych sekwencjach, tj. sekwencjach z losowo wymieszanymi nukleotydami przy zachowaniu długości i częstotliwości występowania poszczególnych zasad a także w całkowicie losowych sekwencjach o zdefiniowanej przez użytkownika długości. Funkcja ta pozwala odpowiedzieć na pytanie czy liczba motywów w określonej sekwencji jest większa lub mniejsza od liczby wynikającej z losowego ułożenia nukleotydów lub różnic w składzie nukleotydowym pomiędzy określonymi grupami sekwencji.

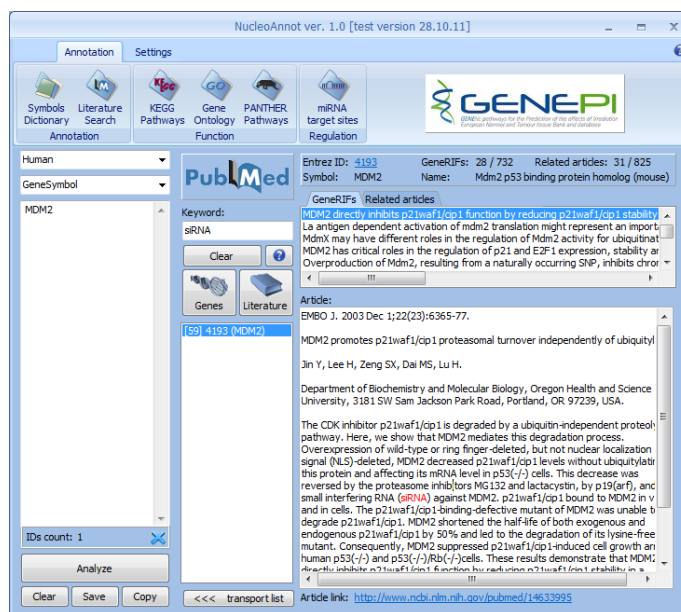
Wyniki analizy mogą być wyświetlone w formie całkowitej ilości motywów w badanej sekwencji, procentu sekwencji zawierającej określony motyw lub w formie pozycji każdego poszczególnego motywu, co pozwala na budowanie map częstotliwości występowania sekwencji na danej pozycji. Ponieważ

pobieranie i analiza sekwencji mogą być nieraz bardzo czasochłonne, częściowe wyniki mogą być na dowolnym etapie wyświetlone lub zapisane do pliku przez użytkownika. Zapisywane może się także odbywać automatycznie w celu uniknięcia utraty danych w sytuacji gdy komputer, na którym działa program uległ awarii.

Program NucleoSeq był wykorzystywany przy niemal wszystkich analizach sekwencji w niniejszej pracy.

5.2. Analiza cech funkcjonalnych genów – NucleoAnnot

NucleoAnnot jest kolejną aplikacją stworzoną w ramach niniejszej pracy, służącą do przetwarzania danych adnotacyjnych genów i ich transkryptów. Główną zaletą NucleoAnnot to możliwość przetwarzania danych adnotacyjnych na komputerze użytkownika przy jednoczesnym wykorzystaniu najnowszych wersji baz danych informacji o genach. Inne tego typu programy wymagają ręcznej aktualizacji baz danych przez ich twórców NucleoAnnot ma jednak wbudowane systemy aktualizacji, które pozwalają użytkownikowi zaktualizować wszystkie bazy danych w dowolnym momencie za pomocą odpowiedniego przycisku.



Ryc. 24: Interfejs programu NucleoAnnot

NucleoAnnot wyposażony jest w wygodny interfejs graficzny (Ryc. 24) ułatwiający wprowadzanie danych, ich analizę oraz eksport wyników. NucleoAnnot składa się z sześciu modułów realizujących następujące funkcje:

- Słownik oznaczeń genów – zbudowany na podstawie informacji z bazy danych EntrezGene słownik pozwala na zamianę identyfikatorów genów różnego rodzaju dodatkowo dostarczając informacji o jego typie oraz położeniu
- Przegląd literaturowy genów – oparty o informacje z bazy danych PubMed system przeszukiwania literatury pozwala na odnalezienie wszystkich artykułów naukowych opisujących funkcje wybranego genu lub listy genów, w których dodatkowo pojawiają się określone słowa

kluczowe. Dodatkowo moduł ten pozwala na odnajdywanie genów powiązanych ze zdefiniowanymi przez użytkownika słowami kluczowymi pojawiającymi się w literaturze naukowej

- Przypisanie do ścieżek sygnałowych KEGG – system pozwala na przypisywanie ścieżek sygnałowych ze zbioru KEGG do określonej listy genów, dodatkowo pozwala on na tworzenie raportów ilości genów przypisanych do każdej konkretnej ścieżki z dostarczonej listy w stosunku do wszystkich genów biorących w niej udział. Moduł ten pozwala także na wyszukiwanie innych genów powiązanych z genami znajdującymi się na dostarczonej liście poprzez wspólne ścieżki sygnałowe.
- Przypisanie do ścieżek sygnałowych Panther – moduł analogiczny do poprzedniego oparty o bazę danych ścieżek sygnałowych Panther.
- Opis funkcji genów na podstawie bazy Gene Ontology (GO) – moduł o funkcjonalności zbliżonej do modułów przypisania ścieżek sygnałowych bazujący jednak na bazie danych funkcji genów GO wg jednej z 3 klas terminów ontologicznych, rozszerzonej o funkcje określania poziomu skomplikowania danego terminu
- Identyfikacja miRNA regulujących określone geny – moduł wykorzystujący połączone bazy danych miRBase i microrna.org w celu stworzenia raportu powiązań pomiędzy zadaną listą genów a różnymi cząsteczkami miRNA. Podobnie jak w przypadku ścieżek sygnałowych możliwe jest tworzenie raportu ilości genów przypisanych do konkretnego miRNA a także graficzne wyświetlenie każdego dopasowania wraz z wyznaczonymi współczynnikami podobieństwa dla każdej pary mRNA-miRNA.

NucleoAnnot wyposażony jest w interfejs zarządzania wszystkimi wykorzystywanymi bazami danych pozwalający na automatyczne sprawdzanie dostępności nowej wersji bazy, przeprowadzanie aktualizacji oraz przeglądanie tabel z danymi.

Program NucleoAnnot wykorzystywany był podczas przeszukiwania literatury naukowej na potrzeby niniejszej pracy oraz do wykonania analizy funkcjonalnej genów znajdującej się w ostatnim rozdziale. Wykorzystany został także do wyszukiwania symboli genów i transkryptów a także ich pełnych nazw podczas analizy sekwencji i danych mikromacierzowych.

5.3. Inne programy użytkowe

Pozostałe z wykorzystanych programów obejmują aplikacje CelConverter wykorzystywaną do konwersji struktury zapisu plików CEL do formatu, który jest łatwiejszy do przetwarzania przez inne wykorzystywane programy. CelExplorer jest z kolei programem, który w bardzo wygodny sposób pozwala na wyciąganie surowych poziomów sygnału sond mikromacierzowych bezpośrednio z plików CEL na podstawie współrzędnych sond lub identyfikatora zestawu. Oba programy wykorzystano do przeprowadzenia części analiz w rozdziale 6.5 poświęconym źródłom niedokładności pomiarowych w eksperymentach z udziałem mikromacierzy Affymetrix.

6. Wyniki analizy

6.1. Analiza danych z eksperymentu mikromacierzowego

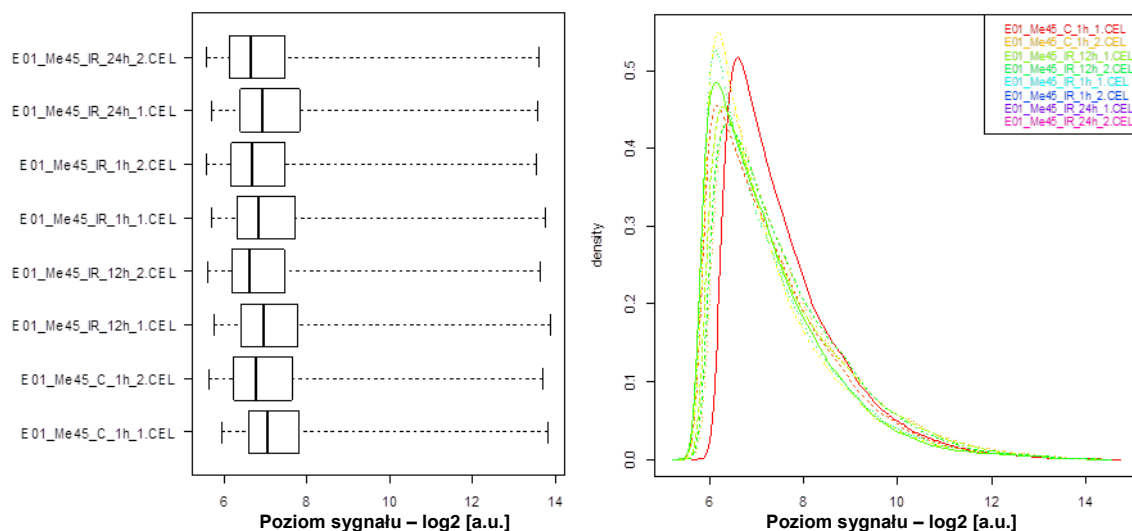
Jakie zmiany w poziomach ekspresji genów wywołuje promieniowanie jonizujące?

Prezentowane poniżej wyniki pochodzą z eksperymentu E01 opisanego w Tab. 4. Dane składają się z 8 mikromacierzy wykorzystanych do określenia poziomu ekspresji genów komórek czerniaka (Me45) poddanych działaniu promieniowania jonizującego w dawce 4 Gy. Poziomy ekspresji genów określono w komórkach niepoddanych działaniu żadnych czynników fizyko-chemicznych (kontrola) oraz po 1, 12 i 24 godzinach od napromieniowania komórek. Dla każdego punktu czasowego wykonano 2 powtórzenia biologiczne.

6.1.1. Kontrola jakości i wstępne przetwarzanie danych

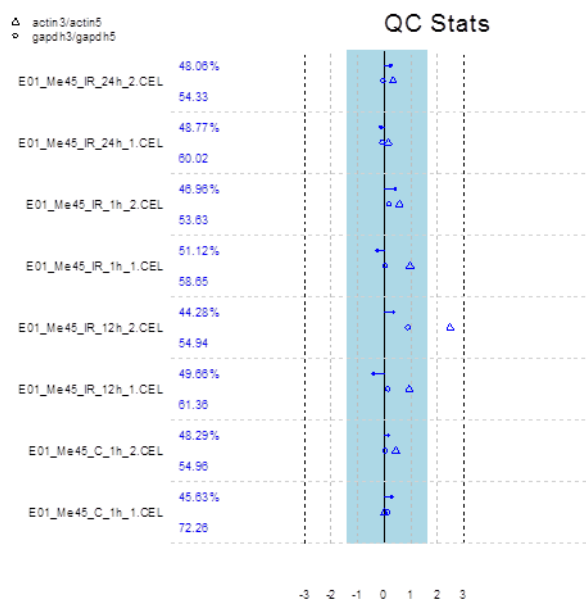
Czy analizowane mikromacierze są wysokiej jakości?

Kontrola jakości ma bardzo istotne znaczenie w eksperymentach opartych na niewielkiej liczbie próbek oraz stosunkowo małej liczbie powtórzeń biologicznych. Głównym jej celem jest określenie potencjalnych źródeł różnic pomiędzy porównywanymi próbkami, które mogą wynikać z niewielkich zmian w procedurze eksperymentalnej mogących wpłynąć na uzyskane w następnym etapie analizy wyniki. Z tego względu głównym celem kontroli jakości jest identyfikacja próbek istotnie odstających od pozostałych w eksperymencie oraz sprawdzenie czy różnice pomiędzy poszczególnymi grupami próbek w ramach powtórzeń technicznych są większe od różnic pomiędzy próbkami reprezentującymi wpływ promieniowania jonizującego, co może wskazywać na silny „efekt partii” (z ang. *batch effect*).



Ryc. 25: Analiza rozkładów surowych wartości poziomu ekspresji z eksperymentu E01 na komórkach Me45 (po lewej wykresy ramkowe, po prawej histogramy).

Kontrolę jakości rozpoczyna analiza rozkładów nieprzetworzonych sygnałów sond typu PM które najczęściej są w stanie wskazać wyraźne różnice pomiędzy próbkami wynikające z różnych czynników, nie wskazując jednak ich źródła (Ryc. 25). Wykresy ramkowe pokazują stosunkowo zbliżone do siebie wartości poszczególnych kwartyli rozkładu surowych intensywności sond, potwierdzają to histogramy chociaż w tym przypadku widoczne jest jednak niewielkie przesunięcie rozkładu próbek E01_Me45_C_1h_1. Mediana rozkładów nie wskazuje jednak na występowanie próbek istotnie różniących się od pozostałych w eksperymencie.

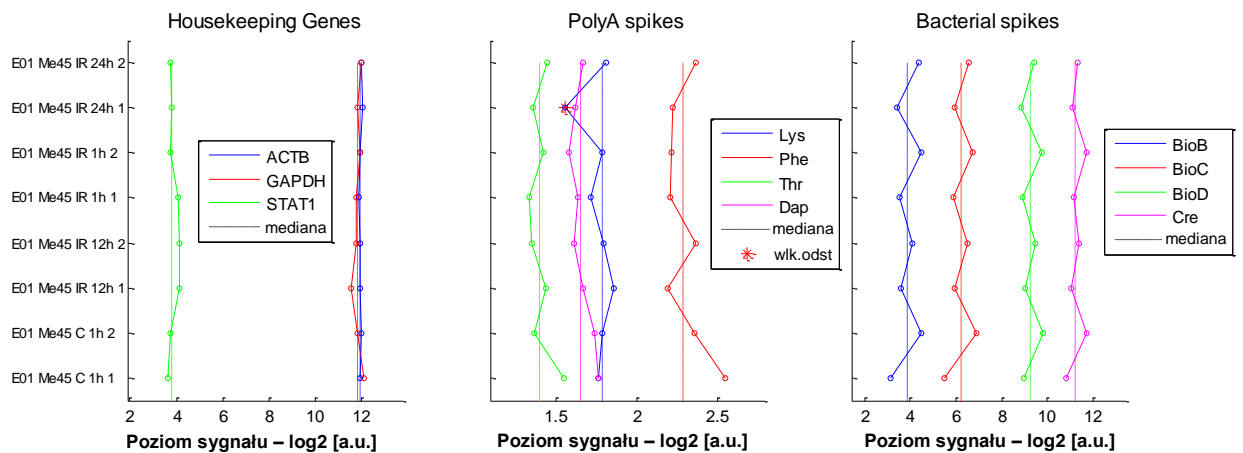


Ryc. 26: Wykres QC dla danych z eksperymentu E01 przeprowadzonego na komórkach Me45

Wykres QC (Ryc. 26) także nie wskazuje na występowanie próbek istotnie odbiegających od pozostałych. Poprzednio zidentyfikowana próbka E01_Me45_C_1h_1 charakteryzuje się znacznie wyższym współczynnikiem skalowania algorytmu MAS5, w porównaniu do całego zbioru, co jest bezpośrednio związane z przesunięciem rozkładu intensywności sond tej mikromacierzy. Różnica ta nie odbiega jednak od dopuszczalnych wartości określonych przez producenta mikromacierzy (opisanych w rozdziale 3.9.4).

Wykresy na Ryc. 27 pokazują zlogarytmowane wartości ekspresji sygnałów sond kontrolnych. Ważne jest, aby na tym etapie wykorzystać dane w formie jakiej będą analizowane w kolejnych krokach co pozwala nie tylko ocenić przebieg eksperymentu ale także skuteczność metod standaryzacji danych na potrzeby dalszych etapów analizy. Z tego względu Ryc. 27 bazuje na danych przetworzonych algorytmem GC-RMA. Przebiegi dla genów referencyjnych (*housekeeping genes*) oraz genów bakteryjnych odpowiedzialnych za kontrole procesu hybrydyzacji (*bacterial spikes*) wskazują na brak jakichkolwiek nieprawidłowości.

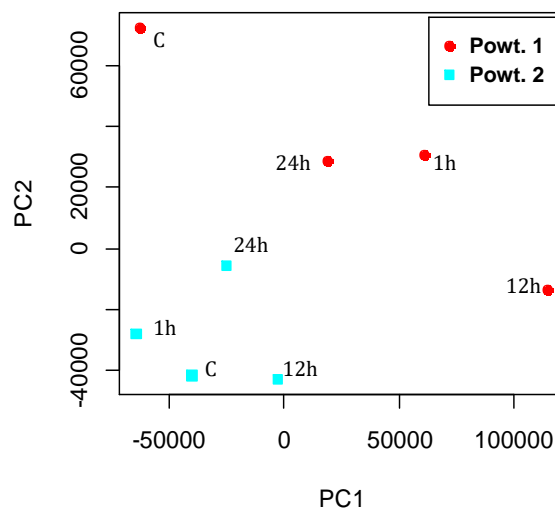
Niepokojące są jednak wartości sond kontrolnych polyA, które położone są poniżej granicy detekcji mikromacierzy (określoną przez poziom *BioB* widoczny na Ryc. 27 dla sond kontrolujących proces hybrydyzacji - *bacterial spikes*) oraz brak wymaganych relacji między pomiarami ($Lys < Phe < Thr < Dap$) wynikający z różnych proporcji dodawanego RNA. RNA specyficzne dla tej grupy sond dodawane jest na samym początku eksperymentu z tego względu poziomy sygnałów zestawów z tej grupy odzwierciedlają cały przebieg eksperymentu i ogólny stan badanego RNA. Ten sam efekt został zaobserwowany także w przypadku eksperymentów E02-E05.



Ryc. 27: Sygnały sond kontrolnych próbek z eksperymentu E01, po przetworzeniu algorytmem GCRMA

Brak jakiegokolwiek anomalii szczególnie w rozkładach intensywności pozostałych sond sugeruje nieprawidłowe przygotowanie mieszaniny *polyA spike* lub jej całkowity brak co jednak nie powinno mieć żadnego wpływu na jakość danych (przypadek ten opisano w Tab. 2 z rozdziału 3.9.4).

Powtórzenia biologiczne z eksperymentu E01 zostały wykonane w dwóch etapach oddzielonych od siebie kilkumiesięcznym okresem czasu. Istotnym źródłem zmienności pomiędzy próbkami mogą być zatem różnice w przygotowaniu i przebiegu eksperymentu. W celu zweryfikowania tej hipotezy wykorzystano analizę głównych składowych - PCA (Ryc. 28), która wskazuje podobieństwa oraz różnice pomiędzy analizowanymi próbkami wynikające z dwóch głównych składowych będących podstawowymi źródłami wariacji w eksperymencie.



Ryc. 28: Analiza PCA danych z eksperymentu E01 przeprowadzonego na komórkach Me45. Czerwony znacznik reprezentuje mikromacierze z pierwszego powtórzenia biologicznego, niebieski - drugie powtórzenie.

Analiza PCA wyraźnie wyróżnia dane z drugiego powtórzenia (niebieski znacznik), które wykonane zostały parę miesięcy po pierwszym. Potwierdza to potrzebę zastosowania algorytmu kompensacji *batch effect*, bez którego identyfikacja genów różnicujących może być utrudniona w związku ze zwiększoną wariancją sygnałów w powtórzeniach biologicznych, mającą źródło w różnicach technicznych pomiędzy próbkami.

6.1.2. Identyfikacja transkryptów różnicujących

Ekspresja, których transkryptów zmienia się w sposób znamiennej statystycznie na skutek promieniowania?

Głównym celem analizy w tym etapie jest odnalezienie transkryptów, których poziom ekspresji uległ zmianie na skutek działania promieniowania jonizującego. Przetestowane zostaną dwa podejścia, jedno oparte o ogólnie przyjęte standardy przetwarzania danych bazujące na prostych metodach z arbitralnie przyjętymi kryteriami oraz drugie bazujące na opisanych w literaturze naukowej algorytmach opartych o zaawansowane testy statystyczne.

Podejście bazujące na arbitralnym kryterium (LFC):

Metoda ta polega na uśrednieniu danych z dostępnych powtórzeń biologicznych a następnie na wyznaczeniu logarytmu stosunku wartości poziomu ekspresji transkryptów po napromieniowaniu do kontroli tzw. LFC (z ang. Log-Fold-Change). Wartości te są następnie porównywane z ustalonym arbitralnie progiem odcięcia (w tym przypadku użyto progę +/-0.5 LFC), powyżej którego przyjmuje się, że zmiana ekspresji jest znacząca.

Podejście tego typu często stosowane jest w badaniach gdzie nie ma do dyspozycji powtórzeń technicznych/biologicznych lub gdy ich liczba jest niewielka. Wiadomo jednak, że wszelkie metody oparte o arbitralne kryteria są silnie uzależnione od metod wstępnego przetwarzania danych i specyfiki badanego materiału biologicznego [243, 244]. Z tego powodu zalecane jest wykorzystanie bardziej stabilnych kryteriów, które pozwolą w lepszym stopniu oddzielić różnice wynikające z niskiego stosunku wartości sygnału do szumu pomiarowego od tych o podłożu biologicznym.

Metoda przetwarzania danych	Czas po napromieniowaniu	Liczba genów różnicujących		
		wzrost	brak zmian	spadek
LFC	1h	1404	21162	697
	12h	3579	17634	2050
	24h	1374	21155	734
Limma	1h	1322	20570	1371
	12h	4192	15896	3175
	24h	2010	19424	1829

Tab. 6: Liczba transkryptów różnicujących po zastosowaniu metod LFC (arbitralne kryterium) oraz Limma (metoda bazująca na wnioskowaniu statystycznym) dla danych z eksperymentu E01 przeprowadzonego na komórkach Me45

Podejście bazujące na wnioskowaniu statystycznym (Limma):

Metody tego typu wymagają przynajmniej 2 powtórzeń technicznych/biologicznych wykonanych w zbliżonych warunkach, które porównywane są pomiędzy sobą bez uśredniania danych [142, 199]. Metody te są jednak bardzo wrażliwe na różnice techniczne pomiędzy powtórzeniami, gdyż zwiększona wariancja sygnałów pomiędzy nimi znacznie obniża moc wykonywanych testów statystycznych [236]. Z tego względu przed wykonaniem testu zastosowano algorytm korekty *batch-effect* zaimplementowany w ramach oprogramowania ComBat [140]. Następnie dane przeanalizowano za pomocą oprogramowania Limma [142] z korektą na wielokrotne testowanie Storey'a [198] (poziom istotności 0.05).

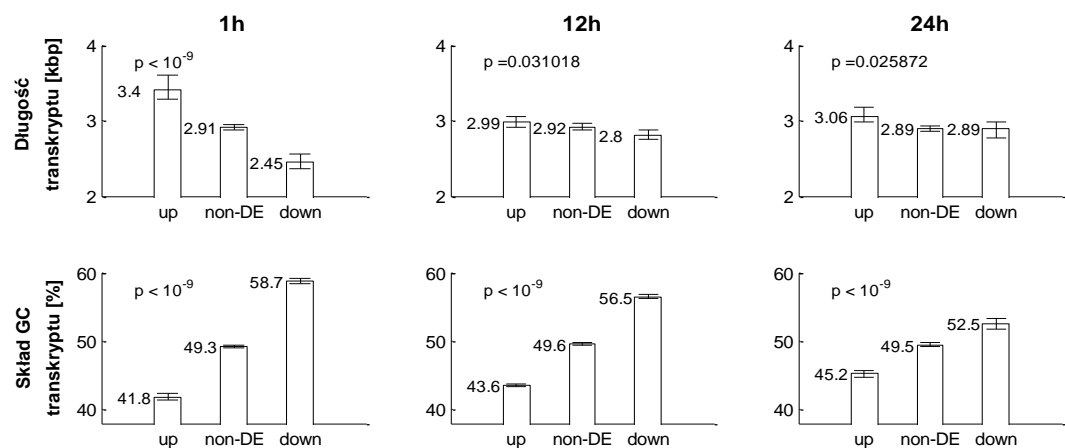
Tab. 6 pokazuje ilości genów różnicujących po zastosowaniu obu metod przetwarzania danych. W przypadku metody LFC ilość genów jest bardzo silnie uzależniona od przyjętego arbitralnie kryterium podziału jednak bez względu na wybrany próg odcięcia (z zakresu 0,25-2) po 12h od napromieniowania można zaobserwować największe zmiany w profilu ekspresji, na co wskazuje ponad dwukrotnie większa liczba genów różnicujących w porównaniu do pozostałych punktów czasowych. Podobne proporcje można zaobserwować w przypadku metody opartej o wnioskowanie statystyczne (Limma).

6.2. Identyfikacja cech transkryptów różnicujących

6.2.1. Podstawowe własności sekwencji

Czy transkrypty, których ekspresja ulega zmianie różnią się pod względem budowy?

Wyodrębnione za pomocą algorytmu Limma grupy transkryptów różnicujących poddano analizie w celu scharakteryzowania podstawowych cech struktury ich sekwencji nukleotydowej. Grupy transkryptów o zwiększonej, zmniejszonej i niezmienionej ekspresji opisano etykietami odpowiednio up, down, non-DE (z ang. non-differentially expressed) a następnie za pomocą aplikacji NucleoSeq zbadano ich sekwencje nukleotydowe. Dodatkowo geny każdego z transkryptów zmapowano do sekwencji ludzkiego genomu i otrzymane współrzędne położenia wykorzystano do scharakteryzowania własności obszaru genomu, z którego pochodzą.



Ryc. 29: Długości i skład nukleotydowy transkryptów różnicujących w eksperymencie E01 (przeprowadzonym na komórkach Me45) zidentyfikowanymi za pomocą algorytmu Limma. P-wartości nad wykresami pochodzą z testu Wilcoxon'a określającego znaczącość różnic badanych statystyk pomiędzy grupami up i down (odpowiednio genów o zwiększonej i zmniejszonej ekspresji). Słupki błędów reprezentują 95% przedziały ufności dla mediany.

Ryc. 29 pokazuje różnice w długości i składzie GC występujące pomiędzy poszczególnymi grupami transkryptów. Grupa o zwiększonej ekspresji charakteryzuje się dłuższymi sekwencjami transkryptu, co szczególnie wyraźnie widać w przypadku próbek zbadanych 1h po napromieniowaniu. Całkowita długość transkryptu jest bardzo silnie skorelowana zarówno z długością części kodującej (współczynnik korelacji Spearmana - Rho: 0.686) jak i długością sekwencji 3'-UTR (Rho: 0.761) co może wskazywać na istotne znaczenie obu tych obszarów. Skład nukleotydowy jest słabo negatywnie skorelowany z długością wszystkich znanych transkryptów (Rho: -0.201) jednak wzrasta w przypadku wyłącznie transkryptów zidentyfikowanych jako różnicujące (Rho: -0.309 dla genów zidentyfikowanych po czasie 1h od

napromieniowania). Nie można jednak jednoznacznie powiedzieć, że wyłącznie jedna z cech ma istotne znaczenie dla procesów regulacji ekspresji a druga wynika z naturalnych właściwości zidentyfikowanej grupy transkryptów. Wykresy na Ryc. 29 dodatkowo pokazują, że różnice zaobserwowane tuż po napromieniowaniu zmniejszają się wraz z upływem czasu, jednak po 24h różnice w składzie GC nadal są znamienne statystycznie.

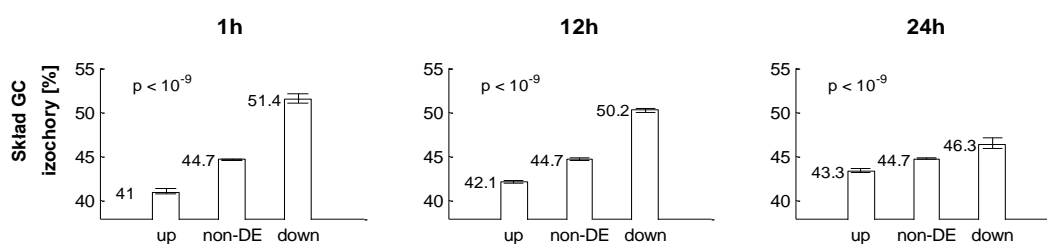
Stopień zależności pomiędzy składem GC a zmianą poziomu ekspresji wyrazić można za pomocą współczynnika korelacji Spearmana, który ze względu na to, że obejmuje wszystkie geny jest niezależny od metodyki poszukiwania genów różnicujących.

Metoda przetwarzania danych	Współczynnik korelacji Spearmana		
	1h	12h	24h
RMA	-0,440	-0,508	-0,181
GC-RMA + ComBat	-0,413	-0,477	-0,180

Tab. 7: Korelacja Spearmana pomiędzy składem GC transkryptu a zmianą ekspresji po różnym czasie od napromieniowania (LFC) w eksperymencie E01 (komórki Me45). Wszystkie korelacje są znamienne statystycznie (p -wartość $< 10^{-9}$)

Tab. 7 pokazuje jak silna jest korelacja uzyskana dla poszczególnych czasów po napromieniowaniu oraz w jakim stopniu wpływa na nią użyta metoda przetwarzania danych. Uzyskane wyniki pokazują, że metoda wstępnego przetwarzania ma nieznaczny wpływ na korelacje pomiędzy zmianą poziomu ekspresji badanych transkryptów (LFC) a ich składem nukleotydowym, pomimo istotnych różnic w algorytmie korekcji tła pomiędzy metodami RMA i GC-RMA. W przypadku GC-RMA dodatkowo kompensowane są różnice w poziomach niespecyficznego hybrydyzacji pomiędzy sondami wynikające z różnych proporcji nukleotydów G i C w ich sekwencjach.

Pomimo silnych różnic pomiędzy badanymi grupami transkryptów, szczególnie pod względem składu nukleotydowego, nie można jednoznacznie określić czy skład i długość sekwencji są cechami, które mają bezpośredni wpływ na zmiany w poziomach ekspresji czy jedynie charakteryzują geny zmienione na skutek zadziałania mechanizmu powiązanego ze strukturą nukleotydową. Długość transkryptu jest związana z długością obszaru 3'-UTR, w którym zlokalizowane są motywy sekwencyjne odpowiedzialne za regulację ekspresji genów. Dłuższy obszar 3'-UTR oznacza zatem zwiększone szanse na pojawianie się motywów, których częstotliwość występowania może być dodatkowo powiązana ze składem nukleotydowym. Istotne staje się zatem określenie motywów sekwencji nukleotydowej mogących wpływać na zmianę poziomu ekspresji pod wpływem promieniowania jonizującego.



Ryc. 30: Skład nukleotydowy izochor, w których położone są transkrypty różnicujące zidentyfikowane w eksperymencie E01 (przeprowadzonym na komórkach Me45) za pomocą algorytmu Limma. P-wartości nad wykresami pochodzą z testu Wilcoxon'a określającego znamienność różnic badanych statystyk pomiędzy grupami up i down (odpowiednio genów o zwiększonej i zmniejszonej ekspresji). Słupki błędów reprezentują 95% przedziały ufności dla mediany.

Ryc. 30 pokazuje różnice w składzie nukleotydowym obszarów genomu (izochor), w których położone są geny badanych transkryptów różnicujących komórki napromieniowane od komórek kontrolnych. Uzyskany wynik sugeruje, że geny o ekspresji zwiększonej na skutek promieniowania położone są najczęściej w obszarach genomu o niskiej zawartości GC (izochorach L1 i L2), co ze względu na różne właściwości tego typu obszarów sekwencji (omówione w pkt. 3.6 wstępu) może dodatkowo wpływać na wydajność procesu transkrypcji.

6.3. Weryfikacja uzyskanych wyników

Czy obserwowana korelacja pomiędzy zmianą poziomów ekspresji na skutek promieniowania a składem GC transkryptów jest możliwa do zaobserwowania wyłącznie w przypadku badanego eksperymentu na komórkach Me45?

W rozdziale 6.2.1 pokazano jak silna jest korelacja pomiędzy zmianą poziomu ekspresji a składem GC transkryptów i w jaki sposób zależność ta maleje wraz z upływem czasu w przypadku komórek Me45 badanych za pomocą mikromacierzy Affymetrix. Pomimo, że eksperyment wykonano w dwóch powtórzeniach biologicznych i oba niezależnie potwierdzają uzyskane zależności to obserwowane zjawisko może być charakterystyczne wyłącznie dla badanej linii komórkowej lub być uzależnione od charakterystyki wykorzystanej platformy badawczej. W celu zweryfikowania tej hipotezy przeprowadzono dodatkowe badania uwzględniające dane uzyskane przy wykorzystaniu innych linii komórkowych oraz alternatywnej platformy mikromacierzowej.

6.3.1. Dodatkowe linie komórkowe

Czy zależność pomiędzy składem GC a zmianą poziomu ekspresji genów pod wpływem promieniowania można zaobserwować w przypadku innych linii komórkowych?

Analizę zależności zmian poziomu ekspresji od składu nukleotydowego wykonano w oparciu o eksperymenty przeprowadzone na trzech dodatkowych liniach komórek: białaczki - K562 oraz raka okrężnicy - HCT116 z normalną i znokautowaną wersją genu odpowiedzialnego za produkcję białka p53 (p53 -/-). Białko p53 ma bardzo istotne znaczenie w przypadku odpowiedzi komórkowej na promieniowanie jonizujące będąc jednym z podstawowych elementów szlaku sygnałowego kontrolującego zahamowanie cyklu komórkowego, uruchomienie mechanizmów naprawy DNA oraz aktywowanie mechanizmu apoptotycznej śmierci komórki w przypadku gdy naprawa DNA jest nieefektywna.

Odpowiedź komórkowa na promieniowanie może zatem różnić się pomiędzy komórkami różnego typu jednak globalne zmiany poziomu ekspresji wynikające ze struktury nukleotydowej transkryptów powinny być podobne w różnych typach komórek o ile nie wynikają z globalnych zaburzeń mechanizmu regulacji ekspresji charakterystycznych dla danej linii.

Tab. 8 pokazuje jak silna jest korelacja pomiędzy składem nukleotydowym transkryptów a zmianą poziomu ekspresji pod wpływem promieniowania w różnych eksperymentach. Silną negatywną korelację można zaobserwować w eksperymentach E07 (K562) oraz E08 (HCT116 p53 -/-) jednak nie są one podobne do uzyskanych w dodatkowych eksperymentach wykonanych na tych samych komórkach (E03 i E09). Pomimo, że negatywna korelacja występuje w przypadku większości eksperymentów to brak

całkowitej zgodności we wszystkich grupach wyników sugeruje, że wysoka korelacja (lub jej brak) może wynikać ze specyficznych cech danego eksperymentu. Wysokie zróżnicowanie współczynnika korelacji pomiędzy eksperymentami może wynikać z właściwości wykorzystanej platformy badawczej opartej o technologię mikromacierzową firmy Affymetrix.

Linia komórkowa	Eksperyment	Współczynnik korelacji pomiędzy składem GC transkryptu a zmianą poziomu ekspresji w określonym czasie po napromieniowaniu			
		1h	12h	24h	36h
Me45	E01	-0,440	-0,508	-0,181	-
	E04	-	-	-	0,022
K562	E02	-	-	-	0,022
	E03	-0,067	0,131	-0,212	-
	E07	-0,358	-0,163	-0,196	-0,297
HCT116	E05	0,199	-	-	-
	E06	-0,127	-	-	-0,040
	E08	-0,011	-0,004	0,001	0,285
	E09	0,150	-0,099	0,084	-
HCT116 p53 -/-	E05	-0,052	-	-	-
	E06	-0,268	-	-	0,089
	E08	-0,469	-0,238	-0,361	-0,464
	E09	0,034	0,084	-0,454	-

Tab. 8: Współczynnik korelacji Spearmana pomiędzy zmianą ekspresji po napromieniowaniu (LFC) a składem GC transkryptu w eksperymentach E01-E09. Szarym kolorem zaznaczono wyniki uzyskane w rozdziale 6.2.1 dla metody RMA. Pola, w których brakuje wartości wynikają z niewykonania określonego czasu w danym eksperymencie.

Obserwowane zależności pomiędzy zmianą poziomu ekspresji a składem GC transkryptu mogą być cechą charakterystyczną komórek Me45 jednak ze względu na brak danych z oddzielnego eksperymentu przeprowadzonego na mikromacierzach Affymetrix nie można wykluczyć, że obserwowane zależności wynikają ze specyfiki samej procedury badawczej.

6.3.2. Alternatywna platforma badawcza

Czy uzyskane wyniki można potwierdzić za pomocą innej platformy mikromacierzowej?

Wykorzystanie alternatywnej platformy badawczej opartej o technologię mikromacierzową firmy Agilent ma na celu weryfikację wyników uzyskanych za pomocą mikromacierzy Affymetrix. Dane umieszczone w Tab. 8 wskazują na występowanie korelacji pomiędzy zmianą poziomu ekspresji transkryptów na skutek promieniowania a ich składem GC co szczególnie widoczne jest w przypadku linii komórkowej Me45 w 12 godzinie po ekspozycji na promieniowanie.

Linia komórkowa	Me45		K562		HCT116		HCT116 p53 -/-	
	A01	A02	A01	A02	A01	A02	A01	A02
Współczynnik korelacji (12h po napromieniowaniu)	0,007	0,028	-0,017	-0,014	-0,085	0,169	-0,015	0,004

Tab. 9: Korelacja pomiędzy zmianą ekspresji (LFC) a składem GC w eksperymentach A01 i A02

Tab. 9 zawiera dane uzyskane za pomocą mikromacierzy Agilent, 12 godzin po napromieniowaniu, które podobnie jak w przypadku mikromacierzy Affymetrix (Tab. 8) przetworzono algorytmem RMA. Dane pochodzą z dwóch powtórzeń biologicznych oznaczonych symbolami A01 i A02 jednak w obu przypadkach dla żadnej z linii komórkowej nie zaobserwowano korelacji podobnych do tych uzyskanych

w eksperymentach opartych o mikromacierze Affymetrix. Najwyższy współczynnik korelacji uzyskano dla linii komórkowej HCT116, w eksperymencie A02 jednak nawet w tym przypadku korelacja nie jest znamienna statystycznie charakteryzując się p-wartością 0.289.

6.4. Analiza własności sekwencji o określonym składzie nukleotydowym

Celem tego rozdziału jest określenie potencjalnych czynników biologicznych mogących tłumaczyć różnice w średnim składzie nukleotydowym pomiędzy transkryptami o zwiększonym i zmniejszonym poziomie ekspresji w napromieniowanych komórkach czerniaka (linia Me45). Skład nukleotydowy może wpływać na częstotliwość występowania motywów regulatorowych w sekwencjach transkryptów oraz w obszarach sąsiadujących genów co może być przyczyną tego, że geny o wysokim i niskim składzie GC są inaczej regulowane pod wpływem promieniowania.

6.4.1. Skład nukleotydowy genów i genomu

Czy skład GC genomu jest na tyle jednorodny w obrębie niewielkich obszarów sekwencji, że może tłumaczyć korelacje pomiędzy składem nukleotydowym transkryptu i obszaru genomu z którego on pochodzi?

Ludzki genom jest bardzo niejednorodny pod względem składu nukleotydowego do tego stopnia, iż wahania w procentowej ilości nukleotydów GC pomiędzy sąsiadującymi fragmentami sekwencji (powyżej 300kbp) mogą przekraczać 40%. W ramach fragmentu, nazywanego izochorą skład nukleotydowy nie jest całkowicie jednorodny jednak jego istotne wahania obejmują wyłącznie krótkie odcinki sekwencji. Z tego względu można zaobserwować bardzo silną korelację pomiędzy składem GC fragmentów sekwencji transkryptu oraz obszaru genomu, w którym są one położone oraz co jest tego następstwem wzajemną korelację pomiędzy składem GC wszystkich elementów transkryptu (5'/3'-UTR, CDS) oraz obszaru promotora o długości 1000 nukleotydów (Tab. 10).

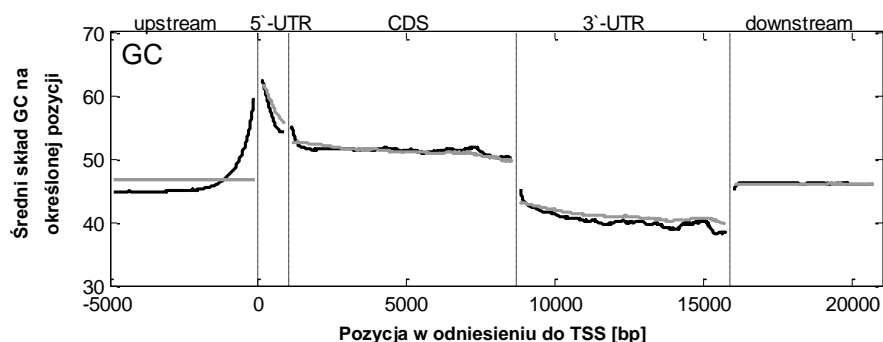
	Izochory	Transkrypt	Promotor	5'-UTR	CDS	3'-UTR
Izochory	1	0,791	0,508	0,352	0,754	0,737
Transkrypt	0,791	1	0,431	0,375	0,894	0,795
Promotor	0,508	0,431	1	0,494	0,401	0,346
5'-UTR	0,352	0,375	0,494	1	0,327	0,263
CDS	0,754	0,894	0,401	0,327	1	0,697
3'-UTR	0,737	0,795	0,346	0,263	0,697	1

Tab. 10: Korelacja pomiędzy składem GC wybranych fragmentów sekwencji nukleotydowych genów i promotorów

W obrębie sekwencji transkryptu wahania składu GC są stosunkowo duże. W celu ich zobrazowania przeprowadzono analizę częstotliwości występowania nukleotydów G lub C na poszczególnych pozycjach transkryptu i obszarów sąsiadujących genu o długości 5kbp, zgodnie z metodologią przedstawioną w punkcie 4.2.1.

Najwyższy średni skład GC zaobserwowano w bliskiej odległości od miejsca startu transkrypcji (Ryc. 31) co jest częściowo związane z występowaniem wysp CpG w tym obszarze sekwencji. Obszar tysiąca nukleotydów powyżej obszaru niekodującego końca 5' jest średnio bardziej bogaty w GC niż sam

transkrypt, jednak wraz z przesuwaniem się na większe odległości skład GC spada poniżej 50%. Obszary znajdujące się powyżej i poniżej transkryptu charakteryzują się składem GC mniejszym niż w sekwencji kodującej jednak znacznie wyższym niż średni skład GC ludzkiego genomu (40.91%) co spowodowane jest tym, że większość genów położona jest w obszarach genomu bogatych w GC [73, 245].



Ryc. 31: Średnia ruchoma skład GC w obszarach kodujących i niekodujących genu oraz w obszarach sąsiadujących o długości 5000 nukleotydów. Czarne linie oznaczają liczby wystąpień w prawdziwych sekwencjach, szare w sekwencjach randomizowanych.

Średni skład GC w obszarze 3'-UTR i sekwencji kodującej nie różni się znacząco po randomizacji sekwencji ze względu na równomierne rozmieszczenie G i C w oryginalnych sekwencjach. Silne nachylenie przebiegu w obszarze 5'-UTR zarówno przed jak i po randomizacji wynika z niewielkiej, negatywnej korelacji pomiędzy długością sekwencji i składem GC tego obszaru ($Rho: -0.18$; $p < 10^{-9}$ dla transkryptów o sekwencji 5'-UTR dłuższej niż 100 nukleotydów). Sekwencja kodująca charakteryzuje się wyższym średnim składem GC niż w pozostałych obszarach, za wyjątkiem najbliższego otoczenia TSS. Nie dotyczy to jednak wszystkich genów, gdyż niektóre charakteryzują się znacznie bardziej jednorodnym składem GC.

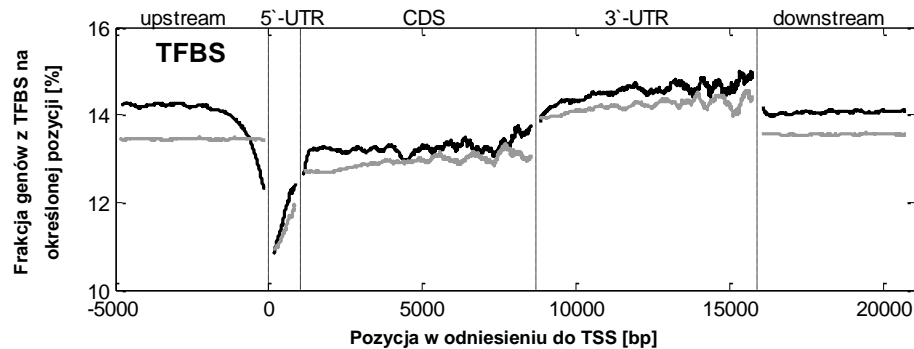
6.4.2. Częstotliwość występowania motywów regulatorowych w DNA

Czy skład nukleotydowy obszaru promotora genu jest skorelowany z częstotliwością występowania motywów rozpoznawanych przez czynniki transkrypcyjne?

W celu określenia częstotliwości występowania motywów rozpoznawanych przez czynniki transkrypcyjne (TFBS) w różnych obszarach sekwencji nukleotydowej wykorzystano podejście użyte podczas analizy składu GC. Ryc. 32 pokazuje wyraźne różnice w częstotliwości występowania motywów wiążących wszystkie znane czynniki transkrypcyjne pochodzące z bazy Jaspar [230]. Liczba motywów przypadająca na daną pozycję jest niemal na całej długości sekwencji wyższa niż w przypadku sekwencji randomizowanych za wyjątkiem najbliższego otoczenia miejsca startu transkrypcji (TSS), gdzie bardzo silnie spada on do poziomu nieobserwowanego nawet w przypadku innych elementów sekwencji. Średnia częstotliwość TFBS w regionach poza sekwencją genu jest podobna do częstotliwości z regionu 3'-UTR, oraz nieco większa niż w przypadku sekwencji kodującej.

Pomimo, że całkowita liczba wystąpień wszystkich czynników transkrypcyjnych jest wysoka ze względu na wykorzystane kryterium opartego o niepełną komplementarność motywów (85%) to pojedyncze czynniki transkrypcyjne występuje stosunkowo rzadko. W przypadku większości z nich odpowiadające im TFBS występują rzadziej niż raz na 1000bp na przestrzeni wszystkich analizowanych

sekwencji. Zmiany przyjętego progu odcięcia (85%) czy też metodologii poszukiwania motywów opisanej w pkt 4.2.2 nie wpływają znacząco na kształt wykresu z Ryc. 32 a jedynie na skalę na osi Y, co dokładnie opisano w pracy [205].



Ryc. 32: Częstotliwość występowania wszystkich badanych czynników transkrypcyjnych w obszarach kodujących i niekodujących genu oraz w obszarach sąsiadujących o długości 5000nt. Czarne linie oznaczają liczby wystąpień w prawdziwych sekwencjach, szare w sekwencjach randomizowanych.

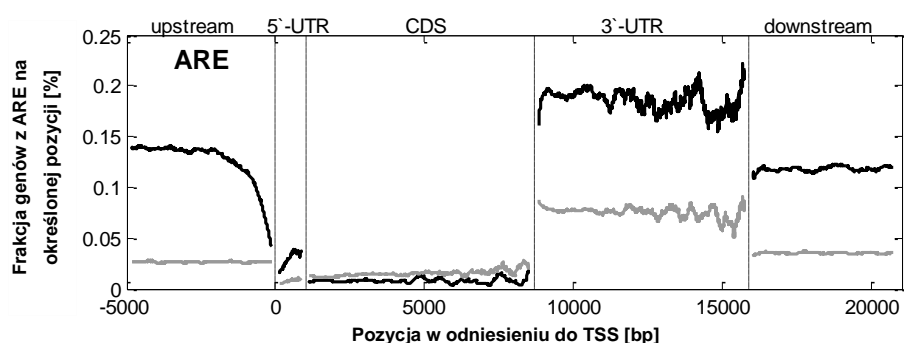
Kształt wykresów z Ryc. 31 oraz Ryc. 32 sugeruje, że częstotliwość występowania TFBS jest bardzo silnie negatywnie skorelowana ze składem GC. Potwierdza to analiza przeprowadzona na 30000 losowych sekwencjach o długości 1000 nukleotydów, w przypadku której uzyskano współczynnik korelacji -0.75 (p -wartość $<10^{-9}$) pomiędzy składem GC a ilością motywów TFBS. Pomimo, że fragmenty DNA z wyższą zawartością GC w sekwencji charakteryzuje niższa ilość TFBS to średni skład nukleotydowy samych motywów rozpoznawanych przez czynniki transkrypcyjne jest bliski 50%. Ponadto nie zaobserwowano korelacji pomiędzy składem GC motywu rozpoznawanego przez czynnik transkrypcyjny a jego długością (p -wartość $=0.262$), która ma często duży wpływ na jego złożoność i tym samym częstotliwość występowania. Sugeruje to, że zmniejszona częstotliwość występowania czynników transkrypcyjnych w obszarach bogatych w GC wynika z mniejszej specyficzności motywów bogatych w nukleotydy AT.

Częstotliwość występowania TFBS uzależniona jest od składu GC badanej sekwencji, ale ze względu na korelację pomiędzy składem GC obszaru promotora i sąsiadującego transkryptu (Tab. 10) wynikającą z jednolitego składu GC fragmentów genomu, częstotliwość występowania TFBS także powiązana jest ze składem GC transkryptu ($Rho = -0,236$ dla obszaru o długości 1kbp i $-0,460$ dla obszaru o długości 5kbp).

6.4.3. Częstotliwość występowania motywów regulatorowych w RNA

Czy skład nukleotydowy sekwencji 3'-UTR wpływa na częstotliwość występowania motywów regulatorowych?

RNA zawiera bardzo wiele motywów sekwencyjnych odpowiedzialnych za przyłączanie zarówno białek jak i funkcjonalnych RNA, które mogą wpływać na jego stabilność. Celem badań w tym rozdziale jest sprawdzenie czy podobnie jak w przypadku miejsc wiązania czynników transkrypcyjnych (TFBS), częstotliwość występowania motywów regulatorowych typu ARE (miejsca oddziaływania z białkami) i MRE (miejsca oddziaływania z miRNA) jest negatywnie skorelowana ze składem GC sekwencji transkryptu. Badania te dotyczą w szczególności sekwencji niekodującej 3'-UTR, w której zgodnie z doniesieniami literaturowymi obecność tego typu motywów najczęściej wpływa na stabilność RNA [246].



Ryc. 33: Częstotliwość występowania motywów typu ARE na danej pozycji w różnych obszarach sekwencji nukleotydowej genu. Czarne linie reprezentują rzeczywiste sekwencje, szare - randomizowane.

Motywy typu ARE zbudowane są wyłącznie z nukleotydów A i U, które dodatkowo w zależności od klasy mogą zawierać różne ilości powtórzeń pentameru AUUUA. Można zatem oczekiwać, że częstotliwość ich występowania powiązana będzie ze składem nukleotydowym. Motywy tego typu powinny pojawiając się tym częściej im więcej w badanej sekwencji jest nukleotydów A i U. Bardzo wyraźnie widać to w przypadku motywów klasy III zbudowanych z minimum 13 powtórzeń nukleotydów A i U (Ryc. 33). W przypadku sekwencji obszaru promotora genu, którą uwzględniono w analizie ze względu na specyficzny skład nukleotydowy (białka rozpoznające tego typu motywy nie wiążą się z DNA), można zaobserwować wyraźny spadek ilości motywów ARE wraz ze zwiększaniem się średniego składu GC na danej pozycji.

Sekwencja 3'-UTR zawiera najwięcej tego typu motywów w porównaniu do pozostałych regionów oraz znacznie więcej niż w przypadku sekwencji randomizowanych dla tego samego obszaru. Sekwencje 3'-UTR charakteryzują się najniższym składem GC spośród wszystkich badanych fragmentów transkryptu w przeciwieństwie do sekwencji kodującej (CDS) gdzie procent GC jest najwyższy za wyjątkiem obszaru znajdującego w bliskiej odległości od TSS. W przypadku sekwencji kodującej ARE występują bardzo rzadko zarówno w oryginalnych jak i randomizowanych sekwencjach.

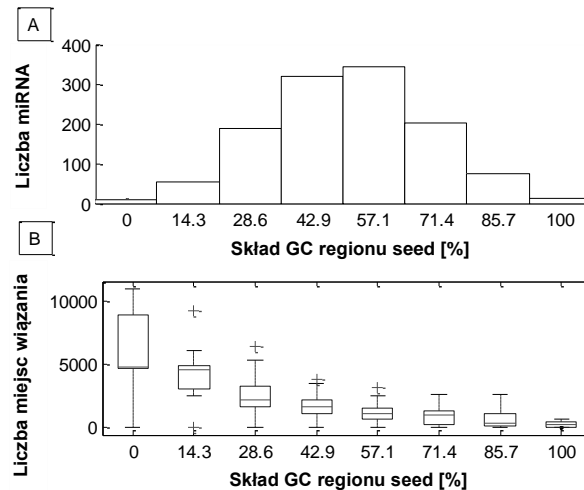
Częstotliwość występowania motywów ARE zgodnie z oczekiwaniami jest bardzo silnie, negatywnie skorelowana ze składem GC, co pokazuje współczynnik korelacji Spearmana w Tab. 11. Podobny rezultat uzyskano także dla motywów rozpoznawanych przez miRNA (mikro RNA), mimo, że ich sekwencje wykazują bardzo silne zróżnicowanie pod względem składu nukleotydowego w przeciwieństwie do motywów ARE.

	Korelacja z ilością wystąpień motywów sekwencyjnych w 3'-UTR przypadających na 1kbp	
	ARE	MRE
3'-UTR GC	-0.474	-0.291

Tab. 11: Korelacja Spearmana pomiędzy składem GC obszaru 3'-UTR transkryptu a częstotliwością występowania motywów sekwencyjnych typu ARE i MRE

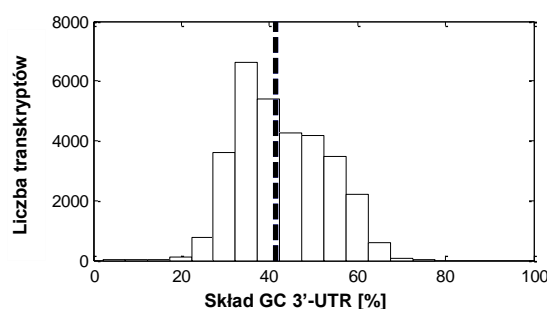
Podczas analizy korelacji wykorzystano 1200 znanych ludzkich miRNA, których miejsca wiązania określono za pomocą algorytmu miRanda, bardzo podobne wyniki uzyskano także przy wykorzystaniu algorytmów TargetScan i PicTar (wyników nie zamieszczono). Mediana składu GC 1201 badanych miRNA wynosi 50.0%, natomiast obszaru seed, który ma największy wpływ na siłę oddziaływania z mRNA 57.1%.

Jest to zatem sytuacja przeciwna do obserwowanej w przypadku motywów ARE, które bez względu na klasę charakteryzują się brakiem nukleotydów GC. Mimo to w przypadku obu grup motywów można zaobserwować korelację składu GC i częstotliwości ich występowania.



Ryc. 34: Statystyki regionu *seed* wszystkich miRNA: A - histogram ilości miRNA o określonym składzie GC w regionie *seed*; B - wykresy ramkowe dla ilości wystąpień miejsc wiązania miRNA o określonym składzie GC regionu *seed*.

Kształt histogramu składu GC obszaru *seed* jest zgodny z rozkładem normalnym ($p < 10^{-9}$ dla testu Jarque-Bera na zgodność dopasowania do rozkładu normalnego) o średniej w punkcie 51.1% i bardzo niskim współczynniku skośności ($-8.4 \cdot 10^{-4}$) - Ryc. 34A. Skład GC obszaru *seed* jest jednak ściśle powiązany z sumaryczną ilością miejsc wiązania w znanych transkryptach (Ryc. 34B). Im mniej nukleotydów GC znajduje się w regionie *seed* tym więcej miejsc wiązania danego miRNA występuje w transkryptach, jednak zależność ta nie ma charakteru liniowego.



Ryc. 35: Histogram składu GC obszarów 3'-UTR transkryptów. Przerzywaną linią zaznaczono medianę.

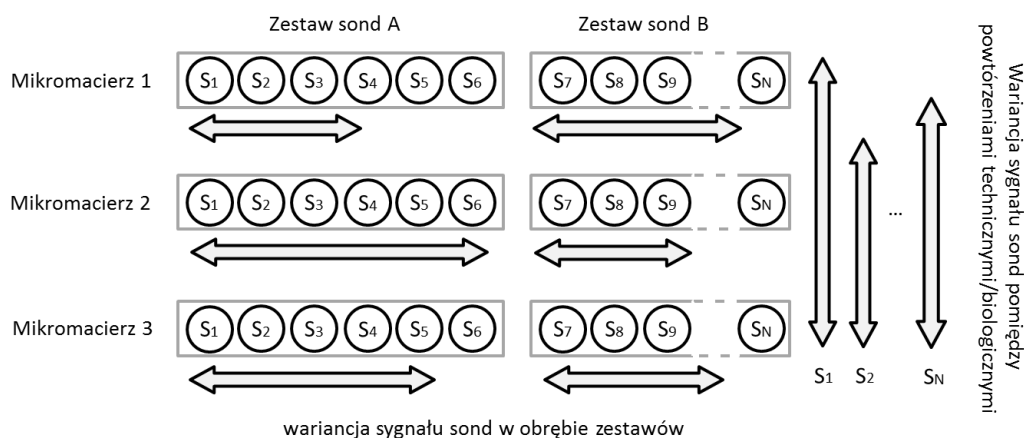
Zjawisko to jest bezpośrednio powiązane ze składem GC badanego obszaru transkryptów (3'-UTR). Większość z przebadanych sekwencji charakteryzuje się niską zawartością GC (mediana: 41.5%, skośność rozkładu: 0.281). W związku z tym obszary *seed* o niskim składzie GC charakteryzują się największą liczbą wiązań ze względu na ich kompatybilność składu GC do obszaru 3'-UTR przeważającej większości transkryptów.

6.5. Źródła niedokładności pomiarowych w eksperymentach mikromacierzowych

Czy korelacja pomiędzy zmianą ekspresji a składem GC transkryptu może wynikać ze specyfiki metody pomiarowej?

W przypadku dodatkowych eksperymentów opartych o platformę firmy Affymetrix powtarzalność korelacji pomiędzy zmianą poziomu ekspresji a składem GC transkryptów jest bardzo niska dodatkowo nie została ona zaobserwowana w przypadku eksperymentów wykonanych za pomocą platformy mikromacierzowej firmy Agilent. Wzbudza to wątpliwości względem wykorzystanych metod przetwarzania danych, które pomimo wielu udoskonaleń mogą prowadzić do powstawania błędów systematycznych skutkujących przeszacowaniem poziomów ekspresji transkryptów o skrajnych proporcjach GC. Dlatego istotne staje się zidentyfikowanie potencjalnych źródeł niedokładności pomiarowych w eksperymentach opartych o mikromacierze firmy Affymetrix, które mogą być powiązane ze składem GC transkryptów.

Pierwszym podejściem do rozwiązania tego problemu było wyznaczenie wariancji sygnału poszczególnych sond uzyskiwanego w powtórzeniach biologicznych eksperymentu E01, a następnie zbadanie właściwości sond, których sygnał charakteryzowała największa wariancja (co ilustrują pionowe strzałki na Ryc. 36).



Ryc. 36: Koncepcja analizy wariancji sygnału pomiędzy sondami mikromacierzowymi. Pionowe strzałki obrazują poziom wariancji sygnału sond na mikromacierzach w powtórzeniach eksperymentu (pierwsze podejście), poziome strzałki - wariancje sond w obrębie określonego zestawu (podejście drugie). Długość strzałek jest proporcjonalna do wartości wariancji.

Drugie podejście obejmuje analizę wariancji sygnału sond w obrębie zestawów, wykonywaną oddzielnie dla poszczególnych mikromacierzy i badanie cech zestawów o najwyższej wariancji, która zmienia się pomiędzy powtórzeniami technicznymi (co ilustrują poziome strzałki na Ryc. 36).

Pierwsze z zaproponowanych podejść jest trudne ze względu na:

- 1) Niewielką ilość mikromacierzy - w przypadku eksperymentu E01 do dyspozycji jest zaledwie 8 mikromacierzy i jedynie po 2 powtórzenia biologiczne na każdy punkt czasowy co sprawia, że

ocena wariancji sond pomiędzy powtórzeniami biologicznymi może być zbyt niedokładna w stosunku do potencjalnego wpływu analizowanych czynników

- 2) Drobne niedokładności w nanoszeniu RNA powodujące różnice w poziomie sygnału uzyskanego z poszczególnych sond w powtórzeniach technicznych/biologicznych oraz czynniki wpływające na niedokładność samego pomiaru fluorescencji, których wpływu nie da się oszacować.
- 3) Wpływ niewielkich różnic biologicznych pomiędzy powtórzeniami.

Trudność wynikającą z niewielkiej liczby mikromacierzy można ominąć poprzez przeprowadzenie analizy na innych zbiorach danych zawierających większą liczbę próbek, dodatkowo pochodzących z różnych źródeł. W tym celu wykorzystano wyniki eksperymentów wykonanych przez producenta mikromacierzy (zbiór Affy-HuGene), zespół badawczy odpowiedzialny za testowanie jakości mikromacierzy (MAQC-133P2) oraz dane z innych laboratoriów, stanowiące przegląd najróżniejszych własności jakie można obserwować w przypadku różnych platform mikromacierzowych firmy Affymetrix. Wykorzystanie wielu niezależnych zbiorów danych pozwala dodatkowo na wyciąganie bardziej ogólnych wniosków na temat obserwowanych zjawisk, które nie są charakterystyczne wyłącznie dla danych z pojedynczego eksperymentu lub pojedynczego laboratorium.

Wpływ cech opisanych w punkcie drugim i trzecim może być wyeliminowany poprzez zastosowanie drugiego podejścia, w którym analizowana jest wariancja sygnału sond w obrębie zestawu z poszczególnych mikromacierzy zamiast analizy wariancji sond pomiędzy powtórzeniami technicznymi. Podejście tego typu zakłada wyznaczenie wariancji sygnału sond w zestawach o określonych cechach budowy wykorzystywanych oligonukleotydów oraz cechach samych hybrydyzowanych transkryptów. Poziom wariancji poszczególnych zestawów został następnie porównany pomiędzy mikromacierzami, co pozwoliło na wyznaczenie cech, które mogą wpływać na niestabilność pomiaru obniżając skuteczność metod poszukiwania transkryptów różnicujących. Przejście z poziomu pojedynczych sond do zestawów pozwala na znaczne ograniczenie liczby analizowanych cech, zmniejszenie wpływu różnic w całkowitej ilości badanego mRNA oraz wyeliminowanie problemu związanego ze zmiennością biologiczną, ponieważ podstawowym założeniem sond połączonych w określone zestawy jest ich specyficzność dla określonego transkryptu.

6.5.1. Analiza wariancji sygnału sond

Jak wysoka jest wariancja sygnału sond mikromacierzowych pomiędzy powtórzeniami technicznymi oraz sond z pojedynczej mikromacierzy w obrębie zestawów specyficznych dla określonych transkryptów?

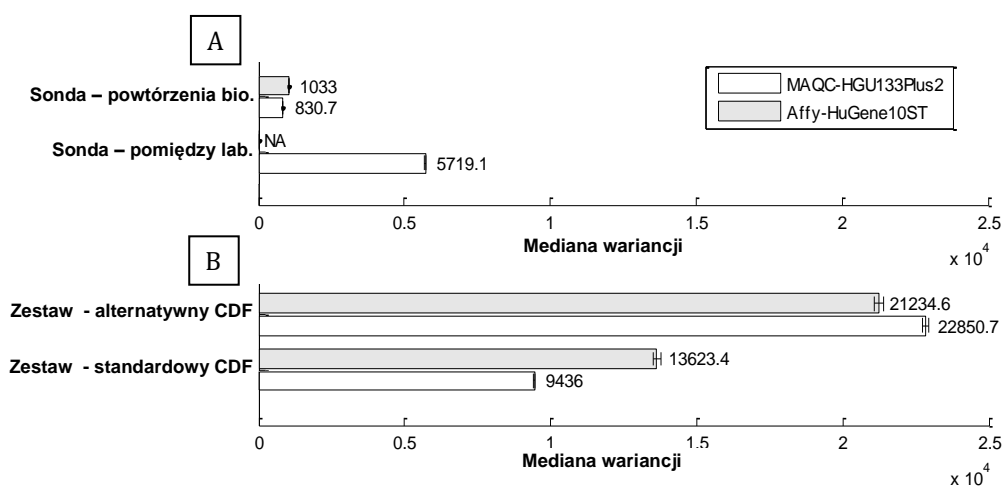
Poziom sygnału odczytanego z sond zależy od szeregu czynników mających wpływ na ilość związanego cRNA oraz poziom ich fluorescencji. Sondy różnią się od siebie pod względem jakości dopasowania do docelowych transkryptów, strukturą nukleotydową oraz, co się z tym wiąże, oddziaływaniami sonda-transkrypt, opisanymi w rozdziale 3.9.3. Wpływa to bezpośrednio na wariancje sygnału wewnątrz zestawu sond, której poziom może być dodatkowo uzależniony od specyficznych cech eksperymentu mikromacierzowego lub poszczególnych próbek. W praktyce wariancja sygnału sond w powtórzeniach technicznych może być różna dla sond o określonych cechach. Algorytmy normalizacji danych traktują

jednak wszystkie sondy w ten sam sposób, co może stanowić potencjalne źródło różnic pomiędzy badanymi próbkami.

Różnice pomiędzy sygnałami sond w obrębie zestawów zaprojektowanych dla określonych transkryptów mogą przekraczać różnice w sygnale pojedynczych sond w nieustandaryzowanych powtórzeniach technicznych lub biologicznych. Mimo, że sygnały różnych próbek mogą być nieraz obciążone stosunkowo dużym błędem systematycznym, wynikającym z różnic technicznych procesów będących częścią eksperymentu mikromacierzowego, to wariancja sygnału w obrębie pojedynczych zestawów sond jest znacznie większa. Podkreśla to istotny wpływ indywidualnych własności sondy na poziom zmierzonego sygnału.

Wysoki poziom wariancji wyników uzyskiwanych z sond należących do pojedynczego zestawu został zaobserwowany przez Cheng Li. Pokazał on, że wariancja sygnału sond PM (pomniejszona o sygnał sondy MM) na przestrzeni różnych mikromacierzy jest znacznie mniejsza niż wariancja pomiędzy sygnałami sond z tej samej mikromacierzy należących do tego samego zestawu [177].

Ryc. 37 pokazuje medianę wariancji sygnału sond w powtórzeniach eksperymentu, jednak nawet gdy rozpatrujemy nieprzetworzone dane to wariancja pomiędzy powtórzeniami jest znacząco mniejsza niż wariancja sond w ramach zestawu zaprojektowanego przez producenta (słupki standardowy CDF) czy zestawów zbudowanych w oparciu o aktualną wiedzę na temat budowy ludzkich genów i ich transkryptów [148] (słupki alternatywny CDF). Efekt ten jest bardzo podobny dla obu badanych platform różniących się od siebie podstawowymi założeniami odnośnie regionu hybrydyzacji sond z transkrypcją. Platforma HG-U133_Plus_2 zawiera sondy specyficzne wyłącznie dla obszaru 3'-UTR genu, podczas gdy zestawy HuGene-10ST składają się z sond specyficznych dla różnych eksonów.



Ryc. 37: Mediana wariancji sygnału sond w różnych eksperymentach mikromacierzowych. A - wariancja sygnałów sond w powtórzeniach biologicznych wykonanych w tym samym lub w różnych laboratoriach. B - mediana wariancji sygnału sond należących do pojedynczego zestawu, zdefiniowanego przez producenta (Zestaw - standardowy CDF) lub zdefiniowanego na podstawie aktualnej wiedzy (Zestaw - alternatywny CDF)

Wariancja sygnału sond obliczona w obrębie tych samych zestawów sond jest większa od wariancji sygnału wyznaczonej dla poszczególnych sond porównywanych w powtórzeniach technicznych eksperymentów wykonywanych w pojedynczym laboratorium jak i dla powtórzeń eksperymentów wykonanych w różnych placówkach badawczych. Wariancja sygnału sond uzyskiwanego z mikromacierzy pochodzących z różnych laboratoriów została zbadana wyłącznie dla danych MAQC-HGU133Plus2,

ponieważ eksperyment Affy-HuGene10ST nie zawierał tego typu danych. Wariancja sygnału sond przy porównaniu mikromacierzy wykonanych w różnych laboratoriach jest znacznie większa niż wariancja sygnałów obserwowanych w eksperymentach wykonanych w ramach jednego laboratorium (tzw. *batch effect*), co ma podłoże w niewielkich różnicach wynikających ze sposobu wykonywania eksperymentu. Zaskakującą obserwacją jest znacznie wyższa wariancja sygnału sond w przypadku zredefiniowanych plików CDF stworzonych poprzez ponowne dopasowanie wszystkich sond do najnowszych sekwencji genomu i transkryptomu. W przypadku mikromacierzy HuGene-10ST może to wynikać z łączenia w zestawy sond, które wcześniej należały do zestawów specyficznych dla różnych eksonów. Dotyczy to głównie genów, których produktem jest wiele różnych form splicingowych. Interesujące jest jednak to, że w przypadku macierzy HG-U133_Plus_2, która zawiera sondy dobrane w znacznej części do sekwencji końca 3'-UTR transkryptów wzrost wariancji po zastosowaniu zaktualizowanego pliku CDF jest jeszcze większy. Może to świadczyć o ciągle jeszcze niedoskonałych metodach tworzenia plików CDF.

W rozdziale tym pokazano, że wariancja sygnału sond w ramach pojedynczego zestawu jest znacznie wyższa niż pojedynczej sondy na przestrzeni powtórzeń technicznych (Ryc. 37). Wydaje się zatem, że zidentyfikowanie czynników wpływających na zróżnicowanie sygnału sond w obrębie zestawu a następnie sprawdzenie czy są one porównywalne, dla zestawów o określonych cechach pomiędzy mikromacierzami powinno być stosunkowo łatwe.

6.5.2. Wariancja sygnału sond w zestawach

Jakie są potencjalne przyczyny wysokiej wariancji sond należących do określonego zestawu, jaka obserwowana jest w przypadku mikromacierzy Affymetrix?

W punkcie 6.5.1 pokazano jak duża jest wariancja sygnałów sond należących do tego samego zestawu, który w założeniu łączy sondy specyficzne dla pojedynczego transkryptu. Wysoka wariancja może wynikać z cech zarówno specyficznych dla danej platformy takich jak nieprawidłowe dopasowanie sond do transkryptów (założono, że efekt ten jest minimalny po zastosowaniu zaktualizowanych plików CDF) jak i wynikających z cech badanego materiału biologicznego (np. obecności form polimorficznych zmieniających jakość dopasowania sond). Innym źródłem wysokiej wariancji może być niespecyficzna hybrydyzacja, która sprawia, że sonda o sekwencji podobnej do fragmentu innego mRNA może charakteryzować się silniejszym sygnałem uzależnionym od stopnia komplementarności sekwencji, ilości podobnych cząsteczek cRNA w mieszaninie oraz warunków w jakich przeprowadzana była hybrydyzacja i proces płukania, co może zależeć nie tylko od budowy mikromacierzy określonego typu ale w dużym stopniu od indywidualnych cech danego eksperymentu.

Na podstawie informacji zebranych w punkcie 3.9.3 określono 6 cech platformy badawczej firmy Affymetrix mogących odpowiadać za wysoką wariancję sygnału sond w obrębie zestawów, bez względu na specyfikę materiału biologicznego. Cechy podzielono na dwie podstawowe kategorie:

- 1) Różnice w budowie sond należących do określonego zestawu
 - a. Skład GC sond
 - b. Obecność motywów (G)₄
 - c. Obecność motywu spacer-T7 (CCGCCTCCC)
- 2) Specyficzne cechy regionu transkryptu komplementarnego do sond z zestawu
 - a. Obecność motywu (A)₂₄ w transkrypcie

- b. Położenie sekwencji komplementarnych do sondy w transkrypcie
- c. Komplementarność sond do różnych form splicingowych transkryptu

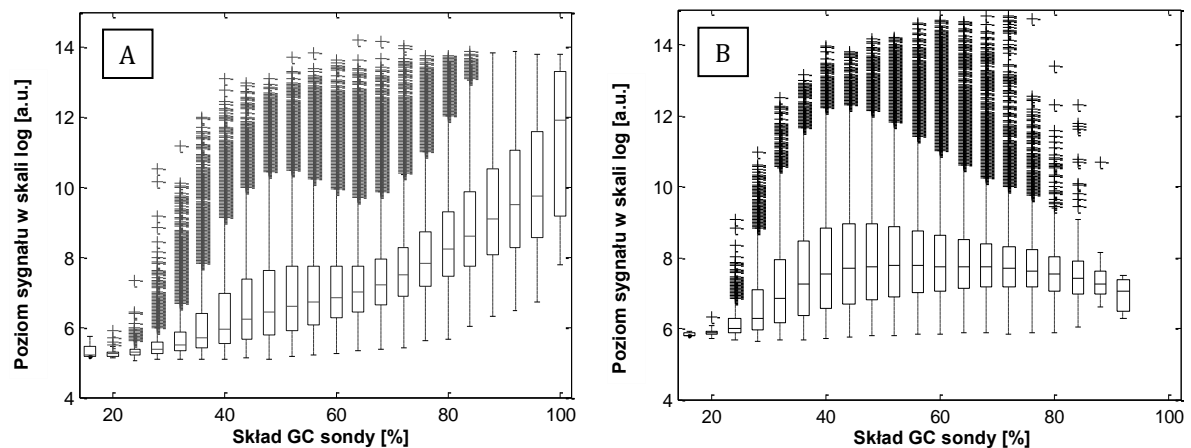
Analizę wpływu wymienionych czynników na poziomy sygnał poszczególnych sond przeprowadzono na dwóch niezależnych zbiorach danych MAQC-133P2 oraz Affy-HuGene. Eksperyment MAQC-133P2 został wykonany przez konsorcjum MicroArray Quality Control na jednej z najpopularniejszych platform HG-U133_Plus_2 z sondami specyficznymi do obszarów niekodujących genu (3'-UTR). Mikromacierze HG-U133_Plus_2 są technologicznie bardzo zbliżone do wykorzystanych w eksperymencie E01 macierzy HG-U133A, jednak składają się one ze znacznie większej ilości sond, z których część jest wspólna dla obu typów mikromacierzy.

Eksperyment *Affy-HuGene* został wykonany na platformie HuGene-1_0-st, która zawiera sondy specyficzne dla sekwencji różnych eksonów. Platforma HuGene-1_0-st która jest następcą HG-U133_Plus_2, wybrana została ze względu na wykorzystanie innych technologii projektowania sond oraz przygotowania materiału biologicznego wykorzystywanego w eksperymencie. Oba analizowane eksperymenty wykonano w oparciu o nieco inny protokół (podstawowe różnice opisano w ostatnim akapicie punktu 3.9.3), dokładniejsze informacje na temat obu eksperymentów znajdują się w opisie wykorzystanych zbiorów danych (pkt. 4.1.2).

1a. Skład GC sond

Sondy wykorzystywane w technologii mikromacierzowej firmy Affymetrix zbudowane są z 25 nukleotydów, komplementarnych do określonego fragmentu mRNA. Skład nukleotydowy sond jest powiązany ze składem nukleotydowym całej sekwencji transkryptu i nieraz niemożliwe jest wybranie zestawu kilkunastu sond o takiej samej proporcji nukleotydów GC i AT. Różnice w składzie GC pomiędzy sondami prowadzą do zwiększonej wariancji sygnału. Skład GC wpływa na siłę oddziaływania sondy z mRNA, im więcej jest nukleotydów G i C tym silniejsze jest oddziaływanie. Proporcje G i C wpływają w sposób liniowy (w szerokim zakresie wartości) na tzw. temperaturę topnienia, która określa temperaturę, przy której połowa wiązań wodorowych w dwuniciowej sekwencji nukleotydowej ulega rozerwaniu [247]. Temperatura topnienia sond Affymetrix zmienia się w zakresie 40-80 °C a ustalona optymalna temperatura, w jakiej przeprowadzana jest reakcja hybrydyzacji mikromacierzy to 60 °C. W obrębie sond należących do pojedynczego zestawu temperatura topnienia nie jest jednorodna, co prowadzi do znacznych różnic w poziomie sygnału. Jeśli temperatura topnienia sondy jest wyższa od temperatury w jakiej przebiega reakcja hybrydyzacji to wiązanie może nie wymagać pełnej komplementarności, pozwalając na powstawanie większej liczby niespecyficznych sparowań. Niska temperatura topnienia sprzyja z kolei rozrywaniu wiązań w procesie płukania zmniejszając w ten sposób poziom sygnału odczytanego z sond.

Zależność pomiędzy składem GC a poziomem sygnału sond ilustrują wykresy ramkowe przedstawione na Ryc. 38. Z uwagi na to, że maksymalna długość sondy to 25 nukleotydów, procent składu GC ma charakter dyskretny przyjmując jedynie 26 różnych wartości uzależnionych od ilości nukleotydów G i C (0-25).



Ryc. 38: Wykresy ramkowe sygnałów sond o różnym składzie GC (średni nieprzetworzony sygnał sond w skali logarytmicznej). A: dane eksperymentu Affy-HuGene, B: dane z eksperymentu MAQC-133P2. Symbolem + oznaczono wielkości odstające.

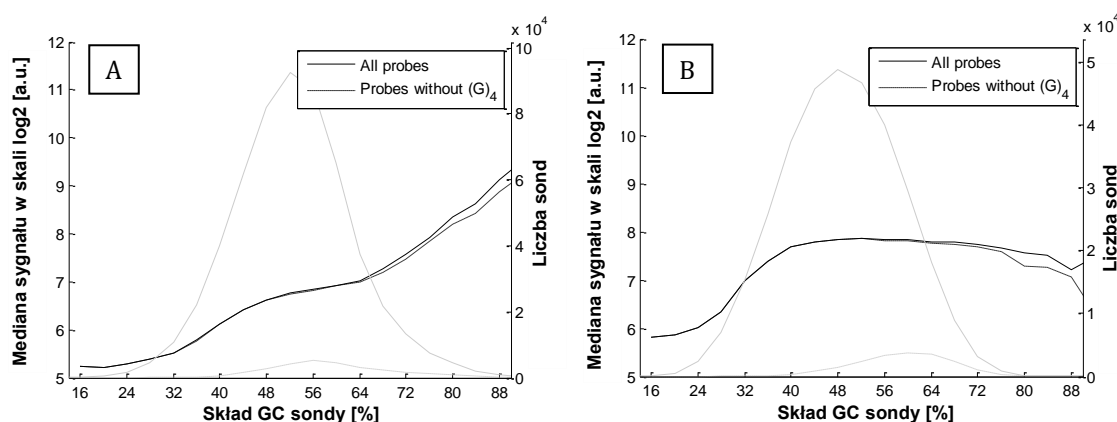
Wykres Ryc. 38A wykonany na podstawie danych z eksperymentu *Affy-HuGene* pokazuje, jak duża jest różnica w medianie sygnału sond różniących się składem GC, rosnąca wraz ze zwiększaniem się ilości nukleotydów G i C. Wzrost ten jest na tyle silny, że różnica pomiędzy skrajnymi proporcjami GC sond prowadzi do ponad 100-krotnego wzrostu poziomu sygnału w skali liniowej. Wykres ten ponadto charakteryzuje się nieznacznym obniżeniem sygnału w okolicy 60% GC, znacznie większym w przypadku danych z eksperymentu *MAQC-133P2* (Ryc. 38B). Wskazuje to na obecność dodatkowego czynnika, którego siła wzrasta wraz ze zwiększaniem się składu GC.

1b. Obecność motywów (G)₄ w sekwencji sond

Obecność motywu (G)₄ czyli czterech sąsiadujących guanin, w sekwencji sondy, wiele razy opisywana była w literaturze jako mająca istotny wpływ na poziom jednorodności sygnału wewnątrz zestawu sond [248-252]. Pomimo, iż jest to jeden z najlepiej opisanych mechanizmów mogących wpływać na wariację sygnału sond to dane literaturowe nie precyzują czy obecność motywu (G)₄ prowadzi do wzrostu czy też spadku poziomu sygnału uzasadniając to specyfiką warunków w jakich przeprowadzane jest doświadczenie. Zgodnie z danymi literaturowymi obecny w oligonukleotydach motyw (G)₄ może obniżać poziom sygnału przez tworzenie struktur zwanych kwadrupleksami-G z sąsiadującymi oligonukleotydami sondy, co uniemożliwia przyłączanie się cRNA. Z drugiej jednak strony tworzenie tego typu struktury może zwiększać odstęp między oligonukleotydami co sprzyja niespecyficzej hybrydyzacji [252]. W przypadku eksperymentu *Affy-HuGene* sondy z motywem G₄ charakteryzują się prawie dwukrotnie wyższym sygnałem od pozostałych (FC = 1.84). Z kolei w eksperymencie *MAQC-133P2* sytuacja jest odwrotna, tutaj sondy bez motywu G₄ mają około 50% większą ekspresję (FC = 1.51). Brak konsekwencji zmian poziomu sygnału wzbudza podejrzenia wobec motywu G₄, jako czynnika bezpośrednio odpowiedzialnego za wzrost wariacji wewnątrz zestawu sond. Ponieważ motyw ten składa się z nukleotydów G, to częściej będzie się pojawiał w sondach bogatych w GC.

Ryc. 39 porównuje histogramy sond o określonym składzie nukleotydowym wykonane dla wszystkich sond oraz sond z motywem (G)₄. Przesunięcie mediany sygnału sond z motywem (G)₄ w stronę wysokich wartości GC sugeruje, że sondy te mogą być odpowiedzialne za obniżenie poziomu sygnału w

obszarze ok. 60% GC co pokazano na Ryc. 38. Zaprzeczają temu jednak wykresy mediany sygnałów zaznaczone na Ryc. 39 przerywaną linią, pokazujące pomijalny wpływ sond $(G)_4$ na poziomy sygnał.



Ryc. 39: Mediana sygnału przed i po odrzuceniu sond z motywem $(G)_4$ oraz histogramy ilości sond A: dane eksperymentu *Affy-HuGene*, B: dane z eksperymentu *MAQC-133P2*. Ciągiem czarną linią zaznaczono poziomy sygnał wszystkich sond, przerywaną sond bez motywu $(G)_4$. Szara ciągła linia (oraz odpowiadająca jej skala po prawej stronie wykresu) reprezentuje histogram ilości sond o określonym składzie GC, przerywana szara linia to podobny histogram wykonany wyłącznie dla sond z motywem $(G)_4$.

Wobec braku dowodów potwierdzających niezależność analizowanego czynnika od różnic wynikających z wyższego, średniego składu nukleotydowego sond, cecha ta nie będzie poddana dalszej analizie.

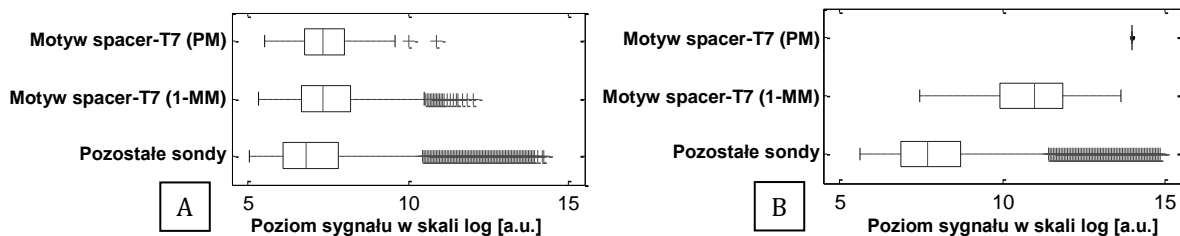
1c. Obecność motywu spacer-T7 (CCGCCTCCC) w sekwencji sond

Motyw CCGCCTCCC komplementarny do fragmentu startera oligo-dT odgrywa bardzo istotną rolę podczas procesu amplifikacji oddzielając sekwencje promotora polimerazy T7 od sekwencji $(T)_{24}$ związanej z mRNA. Ponieważ motyw ten jest częścią sekwencji, która ulega transkrypcji w procesie amplifikacji to zawiera go każda z wytworzonych cząsteczek cRNA sprawiając, że po fragmentacji ich ilość w badanej mieszaninie jest bardzo duża. W pracy [161] pokazano, że występowanie wewnątrz sekwencji sondy motywu CCGCCTCCC może prowadzić do silnego zwiększenia poziomu sygnału. Sondy tego typu są jednak bardzo rzadkie w przypadku mikromacierzy Affymetrix. Platforma HGU-133_Plus2, użyta w eksperymencie *MAQC-133P2* zawiera tylko jedną tego typu sondę (na ponad 330 tysięcy) natomiast w przypadku platformy HuGene-1_0-st, użytej w eksperymencie *Affy-HuGene* odnaleziono jedynie 69 tego typu sond spośród ponad 556 tysięcy.

Motyw CCGCCTCCC ma relatywnie niewielką długość 9 nukleotydów w stosunku do 25 nukleotydowej sekwencji sondy jednak mimo to pewna część fragmentów cRNA obecnych w mieszaninie może związać się z zawierającymi go sondami za sprawą niespecyficznego hybrydyzacji. Sondy nie w pełni komplementarne do motywu mogą także charakteryzować się wyższym poziomem sygnału na skutek niespecyficznego hybrydyzacji.

Wykresy na Ryc. 40 pokazują jak obecność motywu spacer-T7 o pełnej zgodności (PM – z ang. Perfect Match) oraz motywu z 1 niedopasowaniem (1-MM – z ang. Mismatch) w sekwencjach sond wpływa na poziom ich sygnału. Wyniki eksperymentu *MAQC-133P2* (Ryc. 40B) pokazują, że nawet sondy zawierające w sekwencji motyw spacer-T7 o niepełnej zgodności (z pojedynczym niedopasowaniem) charakteryzują się poziomem sygnału o około 10 rzędów wielkości większym od pozostałych sond na mikromacierzy. W znaczny sposób przewyższa to różnice w poziomie sygnału wynikające ze składu GC

jakie przedyskutowano podczas analizy motywów (G)₄. Wyniki uzyskane dla danych z eksperymentu Affy-HuGene wskazują na znikomy wpływ tego typu motywu na poziomy ekspresji (Ryc. 40A). Powodem jest wykorzystanie innego typu oligonukleotydów w procesie syntezy cDNA (bez motywu spacer-T7), które w protokole dla mikromacierzy najnowszego typu zastąpiły wykorzystywany w przypadku innych platform oligo-dT, co dodatkowo opisano w pkt 3.9.3 wstępu.



Ryc. 40: Wykresy ramkowe poziomu sygnału sond zawierających motyw CCGCCTCCC (spacer-T7) w swojej sekwencji – dokładny motyw (PM), motyw z jednym niedopasowaniem (1-MM), pozostałe sondy. A: dane eksperymentu *Affy-HuGene* (w którym nie wykorzystano starterów oligo-dT), B: dane z eksperymentu *MAQC-133P2*

Opisana cecha nie dotyczy najnowszych platform mikromacierzy eksonowych, starsze są jednak w dalszym ciągu powszechnie wykorzystywane, zgodnie z protokołem zakładającym wykorzystanie starterów oligo-dT. Pomimo, że problem dotyczy niewielkiej ilości sond (w przypadku mikromacierzy HG-U133_Plus2 motyw z maksymalnie jednym niedopasowaniem znaleziono w 187 sondach) to może on mieć istotny wpływ na poziomy sygnału zestawów sond. W przypadku dopuszczenia większej ilości niedopasowań liczba sond bardzo szybko się zwiększa maleje jednak wpływ samego motywu a próg pomiędzy specyficznością sekwencji a zmianą poziomów ekspresji jest kwestią indywidualną danego eksperymentu gdyż zależy od poziomu niespecyficznego hybrydyzacji określonej przez niewielkie zmiany w warunkach w jakich zachodzą reakcje (m.in. temperatury, stężenia soli).

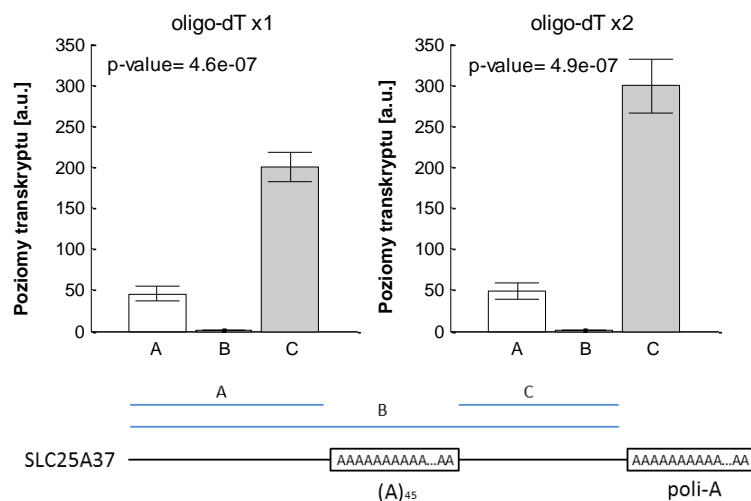
Druga klasa czynników znacznie odbiega od tych przedstawionych w pierwszym punkcie z tego względu, iż nie wpływają one jednoznacznie na globalny wzrost bądź spadek poziomu ekspresji w zależności od tego czy sonda posiada daną cechę czy też nie. Czynniki te działają inaczej w zależności od danego eksperymentu, platformy i przede wszystkim zestawu sond wpływając jednak na wariację sygnału pomiędzy sondami w zestawach.

2a. Obecność motywu (A)_n w transkrypcie

Długie motywy typu (A)_n wewnątrz sekwencji transkryptu mogą wpływać na różnice poziomu sygnału sond jeśli są one komplementarne do sekwencji położonych zarówno po jednej i drugiej stronie tego motywu. Poza naturalnie występującym ogonem poli-A motywy (A)_n o długości kilkudziesięciu nukleotydów, wewnątrz sekwencji transkryptu, mogą także przyłączać starter oligo-dT. W takiej sytuacji powstają dwie skrócone wersje kopii oryginalnego transkryptu, jedna od końca 5' do wewnętrznego motywu (A)_n oraz druga od motywu (A)_n do końca 3'. Występowanie tego drugiego wyniku z możliwości zatrzymania procesu syntezy nici cDNA w przypadku gdy w miejscu wewnętrznego motywu (A)_n przyłączy się dodatkowy starter oligo-dT [159].

Proporcje wszystkich powstałych nici cDNA (pełnej oraz dwóch skróconych) mogą być uzależnione od wielu czynników m.in. długości motywu (A)_n, stężenia oligo-dT, oraz warunków w jakich odbywa się reakcja syntezy cDNA. Problem ten może wpływać na wariację sygnału sond wyłącznie wtedy gdy

proporcje fragmentów cDNA są różne. Określenie stosunku ilości obu fragmentów na podstawie danych mikromacierzowych jest bardzo trudne ze względu na niewielką liczbę zestawów zawierających sondy po obu stronach motywu (A)_n (226 w przypadku platformy HG-U133_Plus_2), niewielką ilość sond w tego typu zestawach, oraz silną wariację pomiędzy sygnałami sond wynikającą z innych cech ich budowy.



Ryc. 41: Wyniki eksperymentu RT-qPCR dla transkryptów genu SLC25A37 przy użyciu starterów oligo-dT w standardowym stężeniu oraz dwukrotnie wyższym. Fragment A: sekwencja od końca 5' do wewnętrznego motywu (A)₄₅; fragment B: cała sekwencja transkryptu; fragment C: sekwencja od wewnętrznego motywu (A)₄₅ do końca 3'. P-wartości pochodzą z testu-t pokazującego różnice pomiędzy średnimi wartościami sygnału fragmentów A i C.

Z tego względu przeprowadzono dodatkowy eksperyment oparty o metodę RT-qPCR, którego celem było sprawdzenie czy transkrypty z długim wewnętrznym motywem (A)_n mogą prowadzić do powstawania skróconych form transkryptu po syntezie cDNA ze starterem oligo-dT w różnych proporcjach. Do analizy wybrano gen SLC25A37 (zawierający 45 nukleotydowy ciąg adeniny) dla którego zaprojektowano 4 startery w taki sposób aby możliwe było zbadanie proporcji fragmentów cDNA po procesie jego syntezy. Fragmenty obejmują sekwencje od końca 5' do wewnętrznego motywu (A)₄₅ (fragment A) całą sekwencje transkryptu (fragment B) oraz sekwencję od wewnętrznego motywu (A)₄₅ do końca 3' (fragment C). Dokładny opis eksperymentu znajduje się w punkcie 4.4. Różnice w ilości poszczególnych fragmentów prezentuje Ryc. 41.

Eksperyment pokazuje, że pełnych sekwencji cDNA jest niezwykle mało w mieszaninie (słupki B), w porównaniu do ilości obu skróconych sekwencji (słupki A i C). Słupki pełnej sekwencji - B jest w tym przypadku niewidoczny jednak oba wykresy przeskalowano tak aby jego wysokość równa była 1. Fragmentów C namnożonych od końca 3' do wewnętrznego motywu (A)₄₅ jest najwięcej w ilości znacznie przewyższającej druga skróconą formę - A, namnożoną od motywu (A)₄₅ do końca 5'. Dwukrotne zwiększenie stężenia oligo-dT dodatkowo powiększyło różnice w proporcji fragmentów A i C co sugeruje, że proporcje sygnału pomiędzy oboma fragmentami są uzależnione od warunków reakcji. Zjawisko to może w podobny sposób wpływać na proporcje sygnału sond mikromacierzowych, które mogą być różne pomiędzy badanymi próbkami. Wyższy sygnał fragmentu C położonego po stronie 3' transkryptu, w stosunku do fragmentu A może sugerować silny wpływ procesu degradacji RNA na wyniki. Jednak niewielka odległość pomiędzy fragmentami w transkrypcie (355 nukleotydów), wysoki współczynnik integralności badanego RNA (RIN=10) oraz zwiększanie się różnic pomiędzy ilością fragmentów wraz ze wzrostem stężenia starterów oligo-dT podważają tą hipotezę.

2b. Położenie sekwencji komplementarnych do sondy w transkrypcie

Sondy komplementarne do fragmentów mRNA położonych daleko od końca 3' mogą charakteryzować się słabszym sygnałem na skutek wysokiego poziomu degradacji tych cząsteczek mRNA, które w warunkach *in vivo* ma niską stabilność. Pierwszy etap syntezy znacznie bardziej stabilnego cDNA następuje od końca 3' transkryptu i zatrzymuje się po osiągnięciu końca 5' lub po napotkaniu miejsca, w którym cząsteczka mRNA została zdegradowana. Z tego względu im większy jest poziom degradacji tym większe są różnice w sygnale pomiędzy sondami rozpoznającymi fragmenty sekwencji położone w dużej odległości od siebie [158]. Silny wpływ procesu degradacji RNA na poziomy ekspresji jest podstawową przyczyną tworzenia przez producenta mikromacierzy zestawów sond specyficznych dla obszaru nie większego niż 600 nukleotydów. Dotyczy to jedynie mikromacierzy z sondami specyficznymi dla obszaru 3'-UTR, mikromacierze eksonowe (np. HuGene-1_0-st) zwykle nie spełniają tego warunku. Mimo to nawet w obrębie 600 nukleotydów degradacja następuje w takim stopniu, że można obserwować jej wpływ na wyniki, co bardzo dobrze ilustruje Ryc. 13.

Inaczej wygląda kwestia degradacji w przypadku powszechnie stosowanych metod polegających na ponownym przypisaniu sond do transkryptów i łączeniu ich w nowe zestawy. Tego typu proces zwykle pomija informacje o położeniu sond łącząc w pojedynczy zestaw sondy oddalone od siebie nieraz o kilka tysięcy par zasad. Przykładowo gen PRRC2C zawiera na mikromacierzy typu HG-U133A dwie grupy sond pierwsza oznaczona zestawem 211948_x_at położona jest w odległości 1385-1734bp od końca 5' transkryptu, druga 211946_s_at znacznie dalej 10067 – 10325bp, które połączone są w jeden zestaw w przypadku zaktualizowanych plików CDF. Łączenie w pojedynczy zestaw sond odpowiadających fragmentom położonym w dużej odległości od siebie w cząsteczkach RNA może być jedną z podstawowych przyczyn zwiększonej wariacji sond w obrębie uaktualnianych zestawów, co przedstawiono na Ryc. 37.

2c. Sondy komplementarne do różnych form splicingowych transkryptu

Poza zwiększonym wpływem degradacji RNA na wyniki, ponowne przypisanie sond do zestawów specyficznych dla genów lub transkryptów ma jeszcze jedną istotną wadę. Problem ten pojawia się w przypadku transkryptów charakteryzujących się więcej niż jedną formą splicingową, które na mikromacierzy posiadają sondy wspólne dla kilku form oraz takie, które są charakterystyczne wyłącznie dla niektórych z nich. Przykładowo gen CBS posiada 3 formy splicingowe transkryptów NM_001178009, NM_001178008 oraz NM_000071. Na mikromacierzy HG-U133_Plus_2 znajdują się 22 sondy specyficzne dla tego genu jednak 16 dla wszystkich 3 form oraz 6 specyficznych wyłącznie dla formy NM_001178008 oraz NM_000071. W przypadku połączenia sond w zestaw specyficzny dla genu CBS wariacja pomiędzy sondami z różnym dopasowaniem transkryptów może być wysoka jeśli w badanych komórkach transkrypt NM_001178009 ulega silnej ekspresji w porównaniu do pozostałych dwóch. Problemu nie rozwiązuje utworzenie zestawów sond specyficznych dla poszczególnych transkryptów. W przypadku zestawów NM_001178008 i NM_000071 wariacja sygnałów sond powinna być relatywnie niska jednak zestaw NM_001178009 będzie zawierał sondy, które są specyficzne wyłącznie do tej formy splicingowej oraz sondy, które są specyficzne dla wszystkich 3 form. Tego typu sytuacja powtarza się bardzo często szczególnie po ponownym dopasowaniu sond do zestawów. W przypadku platformy HG-U133_Plus_2 aż 6470 zestawów (18,6%), zaprojektowanych na podstawie bazy danych RefSeq, zawiera sondy, z których niektóre przypisane są do różnych grup form splicingowych transkryptów pojedynczego genu. Platforma

HuGene-1_0-st-v1 z sondami specyficznymi dla poszczególnych eksonów charakteryzuje się znacznie większą ilością tego typu zestawów 17881 (50,1%).

Łączenie w zestawy sond specyficznymi do alternatywnych form splicingowych ma jednak także dobre strony, daje pełniejszy obraz profilu ekspresji genów, które nieraz pomimo alternatywnych form (np. innych końców 3'-UTR) mają tę samą funkcję kodując białko o identycznej strukturze. Dodatkowo pozwala wykorzystać sygnały z większej ilości sond, dając dokładniejsze oszacowanie poziomu ekspresji genu, co może zmniejszać wariacje sygnału o podłożu technicznym [185]. Mimo w/w zalet niektóre źródła literaturowe sugerują, że lepszym rozwiązaniem jest przeprowadzanie analizy rozdzielającej sygnały dla poszczególnych form splicingowych [253].

6.5.3. Analiza znaczenia czynników wpływających na wariacje sygnału pomiędzy sondami w zestawach

Który z opisanych czynników ma kluczowy wpływ na wariacje sygnału sond, o podłożu innym niż biologiczny?

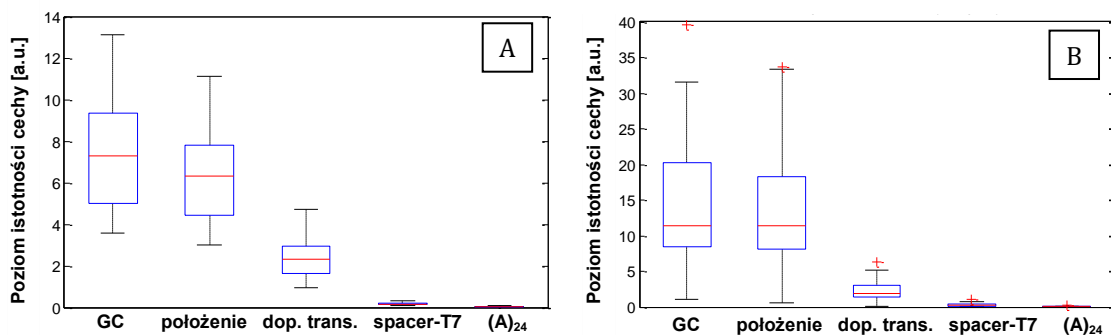
Wariacja sygnału sond określonego zestawu jest uzależniona od kilku czynników jednocześnie, które w różnym stopniu mogą wpływać na jej poziom. Z tego względu wykorzystanie standardowych testów parametrycznych nie pozwala na uzyskanie jednoznacznej odpowiedzi na pytanie która ze zidentyfikowanych cech ma największy wpływ na wysoki poziom wariacji sond w zestawach.

Idealnym narzędziem do tego celu wydaje się być metoda PVCA [254] (z ang. Principal Variance Component Analysis). Jest ona wykorzystywana między innymi do określania wpływu specyficznymi cech próbek na różnice pomiędzy mikromacierzami (np. różne dawki leku, typ materiału biologicznego, indywidualne cechy pacjentów lub dzień wykonania eksperymentu). Pozwala to określić wpływ wielu czynników na różnice pomiędzy próbkami w odniesieniu do różnic wynikających z badanych zmienności o podłożu biologicznym, co dostarcza bardzo cennych informacji szczególnie podczas analizy *batch effect*. Metoda ta może być także zaadoptowana do problemu poszukiwania różnic w poziomie sygnału pomiędzy poszczególnymi sondami (zamiast próbkami) oraz wpływu określonych cech sond na wariacje sygnału (zamiast poziomu ekspresji). W tym celu konieczne jest wyznaczenie wariacji wszystkich zestawów sond dla każdej badanej mikromacierzy a następnie przypisanie do każdego z nich cech charakteryzujących dany zestaw sond, wynikających ze zidentyfikowanych w punkcie 6.5.2 własności.

Trudność związana z zastosowaniem metody PVCA wynika z konieczności zdyskretyzowania wszystkich badanych cech gdyż metoda bazuje na porównywaniu grup pomiarów, które muszą być jasno zdefiniowane. Dodatkowo zdyskretyzowane cechy tracą relacje porządku. Oddzielnym problemem staje się zatem odpowiednie dobranie kryteriów podziału cech na grupy oraz samo określenie liczby grup. Z tego względu metoda ta okazała się nieefektywna w przypadku tego typu danych, pomimo licznych podejść do problemu doboru cech, których niewielka zmiana prowadziła do bardzo silnych różnic w wynikach analizy.

Bardziej elastyczną niż metoda PVCA jest metoda bazująca na wykorzystaniu drzew decyzyjnych zaproponowana przez Wei w 2008 [240]. Metodę tą wykorzystano do bardzo podobnego problemu polegającego na identyfikacji cech sond wpływających na poziom sygnału w mikromacierzach firmy NimbleGen istotnie różniących się pod względem budowy od platformy Affymetrix (przede wszystkim ze względu na zmienną długość sond). Metoda polega na budowie drzewa decyzyjnego o strukturze binarnej,

które podobnie jak w przypadku klasteryzacji rozdziela zbiór danych w każdym węźle na dwie grupy w oparciu o określone własności cech przypisane do zbioru danych, w taki sposób aby różnice pomiędzy obiema grupami były jak największe. W tym przypadku zmiennymi objaśniającymi (predyktorami) są poszczególne cechy zestawów natomiast atrybutami wariacje sygnału w ramach zestawów sond rozdzielone w każdym z węzłów na podstawie określonej wartości jednej z opisanych zmiennych objaśniających. Drzewo budowane jest iteracyjnie poprzez dokładanie nowych węzłów w każdym kroku na podstawie wartości jednej ze zdefiniowanych cech (np. średni skład GC sond w zestawie <50%), minimalizując różnice pomiędzy atrybutami w nowo powstałych grupach do momentu aż dalsze powiększanie drzewa prowadzi do zmniejszania sumarycznej różnicy pomiędzy atrybutami we wszystkich węzłach poniżej pewnego ustalonego poziomu. Drzewo decyzyjne nie wymaga dyskretnych wartości wektora cech dzięki czemu możliwe jest uniknięcie problemu związanego z ich podziałem na określone kategorie, co było konieczne w przypadku metody PVCA. Gotowe drzewo analizowane jest za pomocą algorytmu określającego wpływ każdej z cech na jego kształt. Dokładniejszy opis metodologii tworzenia drzewa decyzyjnego oraz wyznaczania wpływu poszczególnych cech opisano w punkcie 4.3.



Ryc. 42: Wykresy ramkowe pokazujące wpływ poszczególnych cech zestawów sond na wariacje ich sygnału wyznaczone dla różnych mikromacierzy. A: dane eksperymentu *Affy-HuGene* B: dane z eksperymentu *MAQC-133P2*

Ryc. 42 przedstawia wpływ poszczególnych cech, opisanych w punkcie 6.5.2, na strukturę drzewa decyzyjnego zbudowanego na podstawie wariacji sygnału sond wewnątrz zestawów. Założono, że poziom istotności danej cechy jest proporcjonalny do jej wpływu na wariacje sygnałów sond w zestawie, zgodnie z metodologią zaproponowaną w pracy [240]. Wykresy ramkowe tworzą wartości uzyskane dla poszczególnych mikromacierzy z danego eksperymentu. Wyniki pokazują, że skład GC sond i położenie sekwencji komplementarnej do sondy w transkrypcie (odległość od końca 3') mogą mieć porównywalny wpływ na wariację sygnału sond. Wpływ obu tych cech jest bardzo zróżnicowany w odniesieniu do różnych mikromacierzy mogą one mieć zatem istotny wpływ na różnice w poziomach sygnału pomiędzy badanymi próbkami.

Położenie w transkrypcie sekwencji komplementarnej do sondy ma bardzo istotne znaczenie w związku z różnym poziomem degradacji fragmentów bliskich i dalszych od końca 3' w trakcie eksperymentu. Silny wpływ tej cechy może być również konsekwencją łączenia sond w nowe zestawy z pominięciem kryterium budowania zestawów jedynie dla sond położonych w niewielkiej odległości od siebie. Pominięcie tego kryterium jest prawdopodobnie jedną z głównych przyczyn znacznego wzrostu wariacji w obrębie zestawów sond po zastosowaniu uaktualnionych wersji plików CDF (Ryc. 37). Wpływ readnotacji na wariację zestawów sond mógłby zostać określony za pomocą podobnej analizy wykonanej

na podstawie oryginalnych plików CDF, jednak dużą przeszkodą jest tutaj problem nieprawidłowego dopasowania znacznej części sond (w przypadku platformy HG-U133_Plus2 ponad 50%), przez którą niemożliwe byłoby określenie dokładnych wartości dla wszystkich zidentyfikowanych cech w tym przede wszystkim położenia w sekwencji transkryptu.

Wysoki poziom wariacji może być również powiązany z dopasowaniem poszczególnych sond do różnej grupy transkryptów, co może dotyczyć w większym stopniu platformy HuGene-1_0-st opartej o sondy specyficzne dla różnych eksonów. Motyw T7-spacer zgodnie z oczekiwaniami ma nieznaczny wpływ na wariacje co wynika z niewielkiej ilości sond, które go zawierają. Motyw (A)₂₄ w przypadku obu eksperymentów ma pomijalny wpływ na wariacje.

Podobna analiza została także wykonana dla wariacji poszczególnych sond pomiędzy powtórzeniami technicznymi w celu bezpośredniego określenia potencjalnych źródeł wysokiej wariacji pomiędzy różnymi mikromacierzami (podejście pierwsze opisane we wstępie do rozdziału 6.5). Jednak nawet po przeprowadzeniu procedur korekcji tła i normalizacji opisane cechy jedynie w nieznacznym stopniu tłumaczą wysoką wariację pomiędzy mikromacierzami. Może to wynikać albo ze słabości metody, która może nie być odpowiednia dla danych o zbyt dużej ilości atrybutów (ponad 10-krotnie więcej niż w przypadku analizy zestawów sond), lub zbyt dużym wpływem innych czynników, które nie zostały uwzględnione w analizie.

Przeprowadzona analiza pokazała, że różnice w składzie skład GC sond są bardzo silnie powiązane z wysoką wariacją sygnału sond w obrębie zestawu. Dodatkowo wpływ ten jest różny pomiędzy poszczególnymi mikromacierzami w eksperymencie. Konieczne jest jednak określenie jakie czynniki wpływają na sygnał sond o różnych proporcjach GC oraz czy mogą się one być przyczyną przeszacowania poziomów sygnału zestawów zawierających sondy o skrajnych proporcjach GC wpływając na proces identyfikacji transkryptów różnicujących.

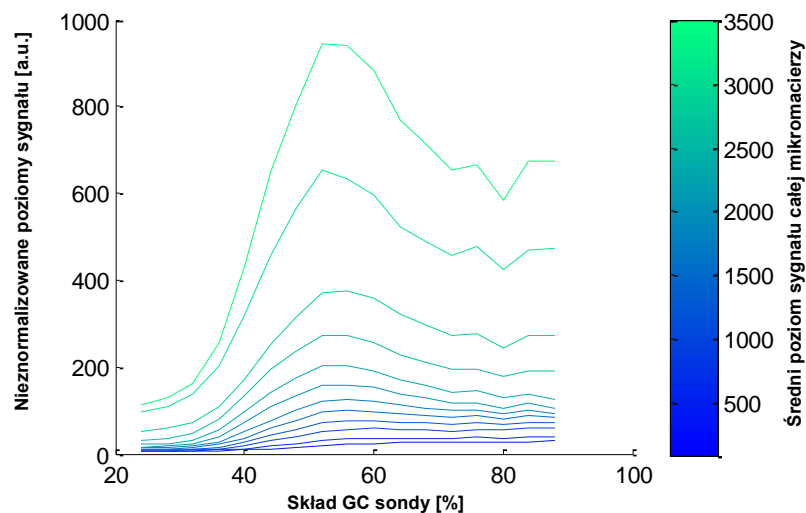
6.5.4. Różnice w poziomach sygnału sond o różnym składzie GC

Jakie są przyczyny różnic w sygnale pomiędzy sondami o wysokim i niskim składzie GC?

Wykresy ramkowe na Ryc. 38 pokazują mediany poziomu fluorescencji sond o różnym składzie GC, które silnie różnią się pomiędzy oboma badanymi eksperymentami. Różnica ta może nie wynikać jednak z wykorzystanej platformy a ze specyficznych cech eksperymentu mikromacierzowego.

Ryc. 43 przedstawia mediany wartości sygnału sond o różnym składzie GC wyznaczone dla mikromacierzy typu HG-U133A. Wykresy narysowano na podstawie danych z 28202 mikromacierzy, pochodzących z różnych eksperymentów, bez poddawania ich wstępnemu przetwarzaniu za pomocą algorytmów standaryzacji danych. Mikromacierze podzielono na 12 grup w oparciu o ich całkowity średni sygnał fluorescencji a następnie dla każdej grupy obliczono medianę przebiegów zależności sygnału sondy od jej składu GC, jakie uzyskano dla poszczególnych mikromacierzy.

Dla grup mikromacierzy o wysokim poziomie średniej fluorescencji sygnał uzyskiwany z sond o skrajnych, najniższych i najwyższych, zawartościach GC jest niższy niż dla sond o składzie GC w okolicy 50-60% GC. Kształt wykresu dla tych mikromacierzy prawdopodobnie oddaje cechy transkryptomu większości komórek, w których najwięcej jest transkryptów o średnim składzie GC w okolicy 50%.



Ryc. 43: Zależność pomiędzy składem GC sond a poziomem sygnału mikromacierzy o różnym średnim poziomie fluoroscencji (uśrednione dane z 28202 próbek wykonanych na platformie HG-U133A).

W przypadku wszystkich analizowanych platform mikromacierzowych skład nukleotydowy sond należących do konkretnych zestawów jest silnym odbiciem składu GC transkryptów, co wyrażają współczynniki korelacji w Tab. 12. Wysoka korelacja wynika z wysokiej jednorodności składu GC transkryptu powiązanego ze składem GC sekwencji genomu, w którym położony jest jego gen (co pokazano w punkcie 6.4.1), oraz tego, że sondy pokrywają spory obszar sekwencji transkryptu (średnio 15% w przypadku mikromacierzy HG-U133A).

Platforma	Rho	Platforma	Rho
Rat230_2	0.6584	MoGene-1_0-st-v1	0.7829
RG_U34A	0.6952	HG-U133A2	0.7915
Mouse430_2	0.7002	HG-U133A	0.7916
HG-U133B	0.7700	HG-U133_Plus_2	0.8096
HG_U95Av2	0.7825	HuGene-1_0-st-v1	0.8346

Tab. 12: Korelacja pomiędzy składem GC transkryptu a zestawu sond pochodzącego z różnej platformy mikromacierzowej

W grupach mikromacierzy o coraz niższej średniej jasności wszystkich sond, sygnał uzyskiwany z sond o skrajnie wysokich zawartościach GC ma coraz wyższy udział w całkowitym sygnale, co najprawdopodobniej wynika z przyczyn technicznych a nie z rzeczywistego wzrostu poziomu transkryptów w materiale biologicznym badanym na tych mikromacierzach. Zmiana kształtu wykresu wraz ze zmniejszaniem się średniej fluorescencji macierzy sugeruje wpływ dodatkowego czynnika.

Średnia fluorescencja wszystkich sond mikromacierzy zależy od procesu amplifikacji i wydajności znakowania cRNA. Wzrost udziału sygnału sond o wysokiej zawartości GC wraz ze zmniejszaniem się średniej fluorescencji mikromacierzy sugeruje, że przy słabszym wyznakowaniu cRNA lub jego mniejszej ilości w procesie hybrydyzacji, niespecyficzne wiązanie cRNA do sond o wysokim składzie GC odgrywa większą rolę zgodnie z sugestią Schuster i wsp. [255], którzy wykazali, że sondy bogate w GC tworzą silniejsze wiązania i pozwalają na częstsze powstawanie niespecyficznych oddziaływań sonda-cRNA.

W grupach mikromacierzy o wysokim poziomie fluorescencji z dobrze wyznakowanym cRNA, odczytywany z macierzy sygnał dla transkryptów bogatych w GC może także nie oddawać prawidłowo ich poziomu komórkowego wykazując niższą niż rzeczywista zawartość w wyjściowym preparacie RNA. Arezi i wsp. wykazali, że wydajność procesu amplifikacji jest uzależniona od składu GC amplifikowanego cDNA. Wysoki skład GC cDNA obniża wydajność polimerazy co spowodowane jest potrzebą wykorzystania większej ilości energii w celu rozerwania potrójnych wiązań wodorowych pomiędzy parami GC, podczas przejścia polimerazy wzdłuż dwuniciowej sekwencji cDNA [256]. Długość sekwencji mRNA, która jak pokazano w punkcie 6.2.1 jest skorelowana ze składem GC, także wpływa na wydajność amplifikacji – im dłuższa jest sekwencja tym mniejsza wydajność transkrypcji [256]. Dodatkowo, wydajność samego procesu hybrydyzacji także może być obniżona przez powstawanie struktur drugorzędowych w cRNA, które tworzone są znacznie częściej przez sekwencje o wysokim składzie GC [162]. Skład GC może więc mieć zarówno dodatni jak i ujemny wpływ na odczytywany z mikromacierzy poziom sygnału.

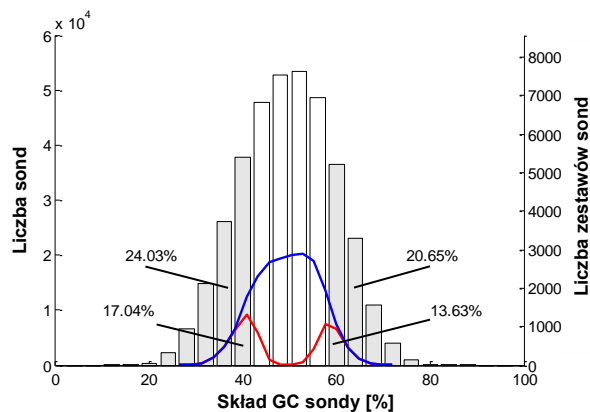
Czas po napromieniowaniu		1h	12h	24h
Współczynnik korelacji Spearmana pomiędzy zmianą ekspresji a składem GC	GC transkryptu	-0,440	-0,508	-0,181
	GC zestawu sond	-0,425	-0,479	-0,121
Test na równość współczynników korelacji (p-wartość)		0.0463	3.46e-05	3.3e-11

Tab. 13: Korelacja Spearmana pomiędzy składem GC transkryptu/zestawu sond a zmianą ekspresji po napromieniowaniu (LFC) w eksperymencie E01 przeprowadzonym na komórkach Me45. Wszystkie korelacje są znamienne statystycznie (p -wartość $< 10^{-9}$). Korelacje uzyskane dla różnych składów GC są od siebie różne (zgodnie z p -wartościami w tabeli). Szarym kolorem zaznaczono wynik wcześniej pokazany w Tab. 7 dla metody RMA

Istotny wpływ składu GC całego transkryptu na sygnały odczytane z sond o różnej budowie potwierdzają wyniki przedstawione w Tab. 13. Korelacja pomiędzy zmianą poziomu ekspresji po napromieniowaniu w eksperymencie E01 a składem GC transkryptu jest silniejsza niż korelacja ze średnim składem GC zestawu sond (test na równość współczynników korelacji). Sugeruje to, iż na sygnał pojedynczych sond wpływ mają nie tylko ich cechy ale także właściwości transkryptów w badanym materiale biologicznym, powiązane z ich składem GC.

Pojedyncze sondy różnią się od siebie składem nukleotydowym jednak różnice pomiędzy średnim składem GC sond danego zestawu powinny być stosunkowo niewielkie. Dodatkowo wpływ składu GC sond na różnice w sygnale jest wysoki dla sond o najbardziej skrajnych proporcjach składu GC, których jest stosunkowo niewiele. Istotne jest zatem określenie na ile jednorodne pod względem składu GC są zestawy sond oraz w przypadku jak wielu z nich dochodzi do nadreprezentacji sond o wysokim i niskim składzie GC.

Ryc. 44 pokazuje jak wiele jest sond oraz zestawów o określonym składzie GC (odpowiednio słupki oraz niebieska krzywa) oraz jaki procent zestawów sond charakteryzują się nadreprezentacją sond o skrajnych proporcjach GC (czerwone krzywe).



Ryc. 44: Histogram ilości sond o określonym składzie GC (słupki) wraz z zaznaczonymi szarym kolorem słupkami dla określonego procentu sond o najbardziej skrajnych proporcjach GC (wartości poniżej dolnego i powyżej górnego kwatyla). Niebieska krzywa to histogram ilości zestawów o określonym średnim składzie GC sond, które do niego należą powiązany z prawą osią Y układu współrzędnych. Czerwone krzywe to histogramy zestawów sond, w których nadreprezentowane są sondy o skrajnym składzie GC (należące do szarych słupków histogramu) na poziomie istotności 0,05.

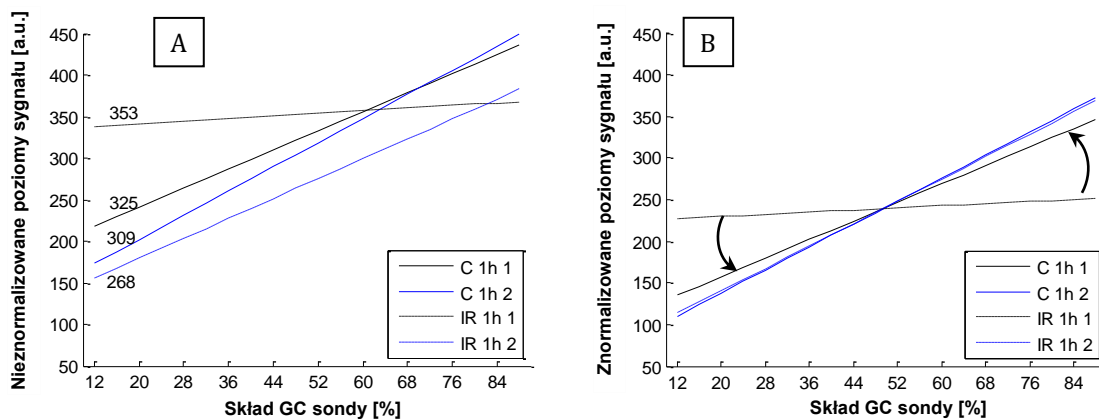
Brak jednorodności średniego składu GC sond w zestawach oraz aż ~30% zestawów, w których nadreprezentowane są sondy o wysokim lub niskim składzie GC wynika z dużych dysproporcji w składzie GC transkryptów, które wpływają na skład GC sond (zależność tą przedstawiono w Tab. 12). Wykres ten wyraźnie pokazuje, że sondy nie były dobierane w taki sposób aby różnice w składzie nukleotydowym były jak najmniejsze oraz aby średni skład GC sond w zestawach był porównywalny. W konsekwencji ~30% zestawów zawiera szczególnie dużo sond o bardzo niskim lub wysokim składzie GC, przez co mogą one być szczególnie mocno narażone na wpływ różnic pomiędzy próbkami wynikający z przeszacowania sygnału sond o skrajnych proporcjach GC, który pojawia się podczas identyfikacji genów różnicujących.

6.5.5. Wpływ obciążenia wynikającego ze składu GC na wyniki eksperymentu mikromacierzowego

Czy różnice pomiędzy mikromacierzami w wariancji sygnału wynikającej ze składu GC mogą wpływać na zmiany w poziomach sygnału odczytanych dla określonych zestawów sond?

Ryc. 43 pokazuje, że wraz ze zwiększaniem się średniego sygnału całej mikromacierzy zmienia się proporcja pomiędzy sygnałem sond o wysokim i niskim składzie GC. Różnice w średniej intensywności sond mikromacierzowych są bardzo powszechne nawet przy powtórzeniach technicznych eksperymentu, jednak są one niwelowane przez algorytmy normalizacji danych. Zmiana proporcji sygnału pomiędzy sondami o różnym składzie może jednak nie być skompensowana przez algorytmy nieuwzględniające informacji o składzie GC sond, będąc jednym ze źródeł wariancji pomiędzy znormalizowanymi próbkami.

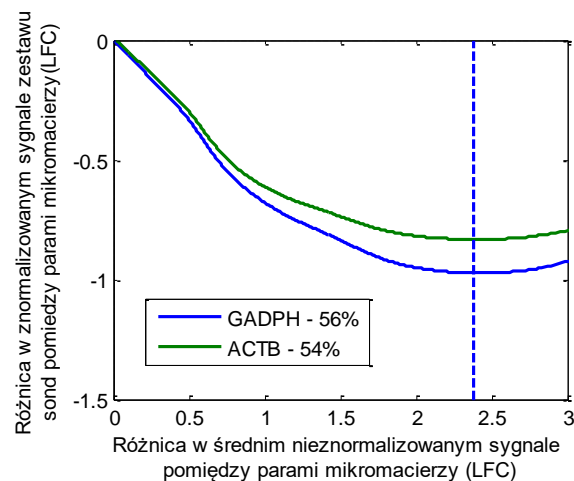
W celu sprawdzenia czy podobne zjawisko ma wpływ na poziomy sygnał w badanym eksperymencie E01 wykazującym dużą zależność zmian poziomu ekspresji po napromieniowaniu od składu GC transkryptów, porównano wpływ algorytmu normalizacji na proporcje poziomu sygnału pomiędzy sondami o różnym składzie GC. W celu ułatwienia interpretacji wyników zamiast mediany sygnału sond o różnym składzie (wykorzystaną na Ryc. 43) posłużono się linią regresji dopasowaną do danych uzyskanych dla wszystkich sond danej mikromacierzy.



Ryc. 45: Linia regresji wyznaczona na podstawie zależności składu GC od poziomu sygnału sond 4 oddzielnych próbek z eksperymentu E01. Wykres A przedstawia surowe dane (liczby po lewej stronie wykresu to średni poziom fluorescencji danej mikromacierzy), wykres B dane po korekcji tła algorytmem RMA i normalizacji kwantylowej.

Ryc. 45A pokazuje różnice w proporcji sygnału sond, o różnej zawartości GC (dla nieustandaryzowanych danych), które wyrażone są współczynnikiem nachylenia linii regresji. W wyniku wstępnego przetwarzania polegającego na odjęciu sygnału tła wyznaczonego za pomocą algorytmu RMA oraz przede wszystkim normalizacji kwantylowej, rozkłady sygnału wszystkich sond pomiędzy próbkami stają się jednakowe. Dzięki temu średni sygnał wszystkich sond mikromacierzowych jest równy we wszystkich analizowanych próbkach, jednak różnice w proporcji sygnałów sond o wysokim i niskim składzie GC nie są kompensowane gdyż w wyniku normalizacji zmianie ulega jedynie punkt przecięcia linii regresji z osią Y a nie kąt jej nachylenia (Ryc. 45B), wynikający z różnic technicznych pomiędzy próbkami jakiego omówiono w punkcie 6.5.4. Różnica ta jest szczególnie duża w przypadku pary próbek z pierwszego powtórzenia biologicznego (C_1h_1 oraz IR_1h_1), objawiając się przesunięciem linii regresji, zaznaczonym strzałkami, względem środka wykresu. Konsekwencją zmian w proporcji sygnału sond o wysokim i niskim składzie GC jest duża różnica w sygnale pomiędzy sondami z mikromacierzy kontrolnych (C) oraz uzyskanych po napromieniowaniu komórek (IR), która nie jest kompensowana przez normalizację kwantylową. Zastosowanie algorytmu sumaryzacji łączącego sygnały z pojedynczych sond, prowadzi do tego, że poziom ekspresji zestawu o średnio niskim składzie GC sond, będzie miał w próbkach kontrolnych zawyżony sygnał w porównaniu do próbek po napromieniowaniu i odwrotnie dla zestawów o wysokim składzie GC (powyżej 50%).

Efekt ten jest bardzo dobrze widoczny na przykładzie sond specyficznych dla genów referencyjnych (*housekeeping*), których poziomy ekspresji w założeniu nie powinny się zmieniać w próbkach [257]. Ryc. 46 przedstawia wyniki analizy 28202 mikromacierzy z 759 eksperymentów wykonanych na platformie HG-U133A, której celem było sprawdzenie jak różnice w średniej fluorescencji między dwoma mikromacierzami (które są negatywnie skorelowane ze zmianą współczynnika nachylenia linii regresji) wpływają na odczyt poziomu genu referencyjnego po normalizacji danych. Mikromacierze z pojedynczych eksperymentów łączone były w pary, w których oznaczano różnicę średniej fluorescencji obu mikromacierzy przed normalizacją i różnicę oznaczenia sygnału genów referencyjnych po normalizacji.

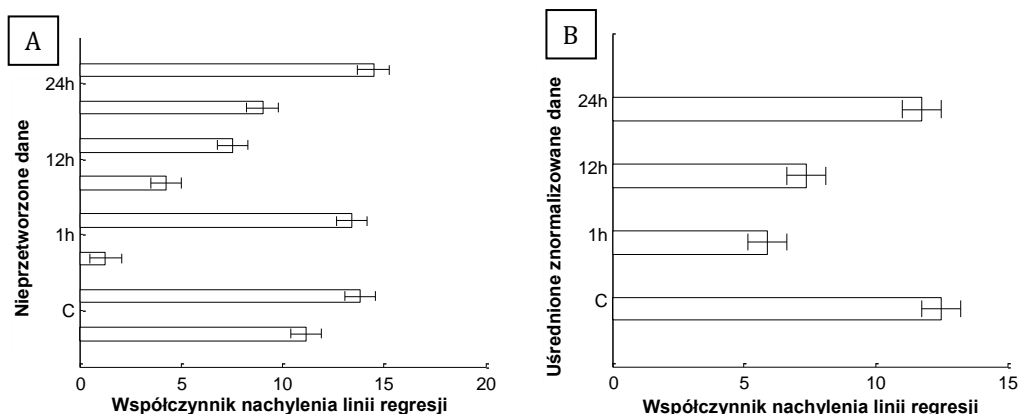


Ryc. 46: Zależność pomiędzy różnicą w średnim sygnale pomiędzy nieznormalizowanymi mikromacierzami oraz zmianą poziomu ekspresji po normalizacji, zestawów sond genów referencyjnych: GADPH i ACTB pomiędzy unikatowymi parami próbek w określonym eksperymencie. Wykres przygotowano na podstawie danych z 28202 mikromacierzy wykonanych na platformie HG-U133A, wygładzonych za pomocą regresji lokalnej LOESS. Pionowa przerywana linia określa medianę pary próbek o największej różnicy w średnim poziomie sygnału w każdym z eksperymentów.

Im większa jest różnica w średnim poziomie sygnału pomiędzy nieustandaryzowanymi mikromacierzami tym większa jest różnica w poziomie sygnału pomiędzy zestawami tych samych sond w ustandaryzowanych mikromacierzach, przy czym różnica ta jest tym większa im skład GC transkryptu bardziej odbiega od średniej wartości 50% (różnice w składzie GC genów GADPH i ACTB). Ryc. 46 pokazuje, że gdy porównujemy ze sobą dwie mikromacierze A oraz B i całkowity średni poziom sygnału mikromacierzy A jest większy od B to po normalizacji danych uzyskane poziomy ekspresji obu genów kontrolnych GADPH i ACTB są mniejsze w próbce A w stosunku do B. Zależność ta jest na tyle silna, że dwukrotnie wyższy poziom średniego sygnału jednej z mikromacierzy (LFC=1) prowadzi do tego, że sygnał genu referencyjnego po normalizacji staje się mniejszy o około 40% w mikromacierzy A w porównaniu z mikromacierzą B. W przypadku transkryptów o składzie GC poniżej 50% zależność ta jest najprawdopodobniej odwrotna.

Wysokie różnice w średnim poziomie sygnału pomiędzy mikromacierzami są bardzo powszechne w eksperymentach. Przerywaną pionową linią na Ryc. 46 zaznaczono medianę maksymalnej różnicy w średnim poziomie sygnału pomiędzy parami mikromacierzy z każdego z 759 analizowanych eksperymentów.

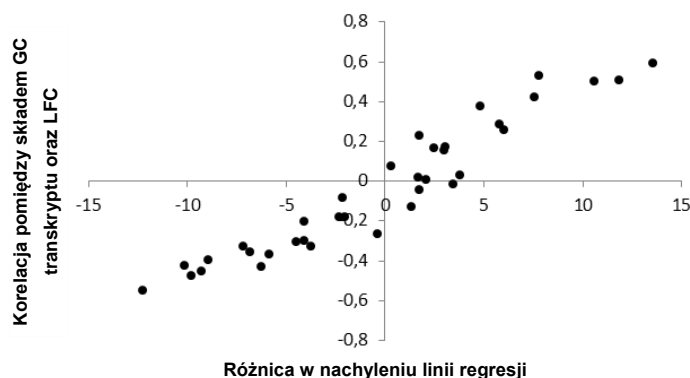
Wpływ tego zjawiska na różnice pomiędzy wszystkimi zestawami sond można oszacować porównując współczynnik nachylenia linii regresji dla surowych danych oraz danych po przetworzeniu i uśrednieniu powtórzeń biologicznych (Ryc. 47). Duże różnice współczynnika nachylenia pomiędzy danymi z mikromacierzy jakie uzyskano dla różnych powtórzeń biologicznych (Ryc. 47A) mogą zwiększać wariancję sygnałów zestawów sond obniżając w ten sposób skuteczność algorytmów poszukiwania genów różnicujących (czułość). Z kolei duże różnice pomiędzy uśrednionymi sygnałami z mikromacierzy dla różnych próbek z komórek kontrolnych i napromieniowanych (Ryc. 47B) mogą w konsekwencji prowadzić do identyfikacji fałszywych genów różnicujących, których zmiana nie ma podłoża biologicznego, obniżając tym sposobem specyficzność metod.



Ryc. 47: Współczynnik nachylenia linii regresji dla zależności składu GC od poziomu sygnału sond. Słupki błędów określają 95% przedziały ufności dla współczynnika nachylenia. Wykres wykonano dla mikromacierzy z eksperymentu E01 przed przetwarzaniem (A) oraz po przetwarzaniu metodą RMA i uśrednieniu powtórzeń biologicznych (B).

Ponieważ jak pokazano skład GC zestawu sond jest bardzo silnie skorelowany ze składem GC transkryptu, to różnice w sygnale zestawów o wysokim i niskim średnim składzie GC będą wpływać na różnice pomiędzy transkryptami o wysokiej i niskiej proporcji nukleotydów GC. Może to tłumaczyć różnice w składzie GC transkryptów o zmniejszonym i zwiększonym poziomie ekspresji, jaką zaobserwowano na Ryc. 29, oraz co się z tym wiąże silną negatywną korelację pomiędzy zmianą poziomu ekspresji po napromieniowaniu a składem nukleotydowym transkryptów, widoczną w danych z eksperymentu E01 (Tab. 7), najsilniejszą dla czasów 1 i 12 godziny po napromieniowaniu.

Sposób, w jaki różnice w sygnale sond o różnym składzie nukleotydowym (wyrażone za pomocą różnicy w kącie nachylenia linii regresji) wpływają na współczynnik korelacji pomiędzy logarytmem stosunków sygnału (z ang. log fold-change LFC) każdych dwóch unikatowych par próbek z eksperymentu E01 i składem nukleotydowym transkryptów ilustruje wykres rozrzutu na Ryc. 48.



Ryc. 48: Wykresy rozrzutu pomiędzy różnicą nachylenia linii regresji a korelacją pomiędzy składem GC transkryptu oraz zmianą poziomu ekspresji wykonany dla każdej unikatowej pary próbek z eksperymentu E01

Im większą jest różnica w nachyleniu linii regresji tym większą korelację zmiany poziomu ekspresji i składu GC transkryptu można zaobserwować. Zależność tych dwóch parametrów określa bardzo wysoki współczynnik korelacji Spearmana 0.9776, który nazwano **indeksem SLGC**. Założono, iż określa on siłę, z jaką obciążenie wynikające ze składu GC wpływa na różnice pomiędzy próbkami w danym eksperymencie,

co ułatwia porównanie wyników uzyskanych przy wykorzystaniu innych metod analizy danych oraz w różnych eksperymentach.

6.5.6. Wpływ składu GC na wyniki niezależnych eksperymentów

Czy obserwowany wpływ składu GC sond na wyniki dotyczy jedynie badanego eksperymentu, użytej metody przetwarzania danych lub określonej platformy mikromacierzowej?

Zmiana kąta nachylenia linii regresji na Ryc. 45, wynikająca ze zmian proporcji sygnału pomiędzy sondami o wysokim i niskim składzie GC, może tłumaczyć część różnic w poziomach sygnału pomiędzy próbkami w eksperymencie E01, w którym badano wpływ promieniowania na zmiany profilu ekspresji genów, komórek Me45. Nie dowodzi jednak tego, że zmiany te wynikają z różnic technicznych pomiędzy próbkami a nie ze zmian proporcji transkryptów o wysokim i niskim składzie GC jakie wyizolowano z komórek.

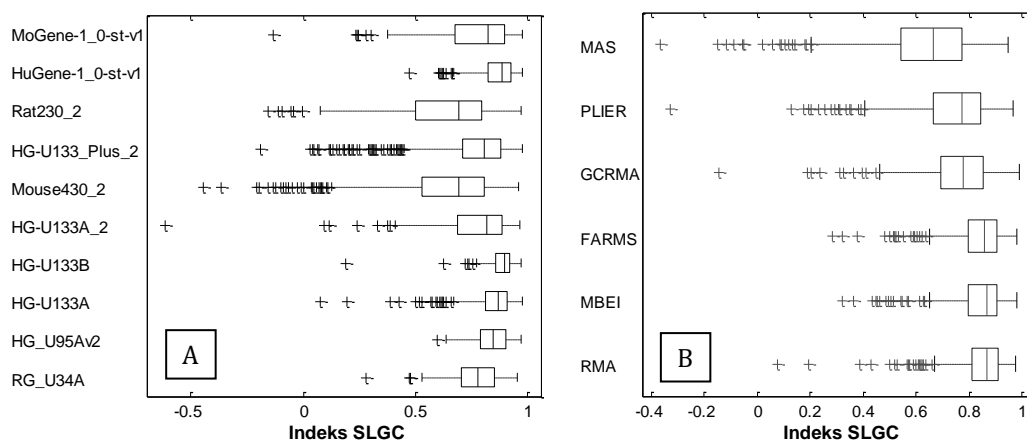
W przypadku danych z pozostałych eksperymentów wpływ składu GC na zmianę profilu ekspresji jest równie silny (Tab. 14) i pomimo, że wszystkie z nich badają wpływ promieniowania jonizującego na zmiany profilu ekspresji to wynik ten może być charakterystyczną cechą wykorzystanych metod przetwarzania danych, eksperymentów wykonanych w określonym laboratorium, danej platformy mikromacierzowej lub też cechą charakterystyczną dla wszystkich eksperymentów wykonanych na platformie mikromacierzowej firmy Affymetrix.

Eksperyment	E01	E02	E03	E04	E05	E06	E07	E08	E09
Linia komórkowa	Me45	K562	K562	Me45	HCT116	HCT116	K562	HCT116	HCT116
Ilość próbek	8	6	8	6	4	12	10	20	20
Indeks SLGC	0,98	0,94	0,90	0,95	1,00	0,74	0,92	0,85	0,50

Tab. 14: Indeks SLGC dla wszystkich naszych eksperymentów

Silny wpływ składu GC na zmiany profilu ekspresji może dodatkowo tłumaczyć brak powtarzalności wyników uzyskanych w różnych eksperymentach oraz przeciwne trendy zmian poziomu ekspresji niektórych genów np. pomiędzy próbkami z eksperymentu E05 i E06 dla komórek HCT116, co obrazuje przeciwny znak współczynnika korelacji pomiędzy zmianą poziomu ekspresji a składem GC (Tab. 8).

W celu sprawdzenia jak powszechne jest to zjawisko w przypadku danych uzyskanych w innych laboratoriach, na podstawie różnych platform, oraz jaki jest wpływ algorytmów przetwarzania danych, przeprowadzono analizę ponad 160 tysięcy mikromacierzy z 7276 eksperymentów wykonanych na 10 platformach mikromacierzowych firmy Affymetrix. Wykorzystanie dużego zbioru danych do tego typu analizy jest niezbędne ze względu na duże różnice pomiędzy pojedynczymi próbkami oraz pojedynczymi eksperymentami, które nie mogą być wykorzystane do wysuwania wniosków dotyczących określonej platformy badawczej bądź też wykorzystanej metody analizy danych. Dokładny opis danych mikromacierzowych wykorzystanych w tej analizie znajduje się w punkcie 4.1.2 materiałów i metod, metodologie analizy danych opisano w punkcie 4.3. Wykresy wykonano wyłącznie dla eksperymentów zawierających co najmniej 10 próbek.



Ryc. 49: Indeks SLGC dla danych pochodzących z różnych platform, przetworzonych metodą RMA (A) oraz przetworzonych różnymi algorytmami, dla platformy HG-U133A (B)

Ryc. 49 pokazuje, że zależność pomiędzy składem GC transkryptu a zmianą profilu ekspresji jest bardzo silna dla wszystkich analizowanych platform oraz najpopularniejszych metod wstępnego przetwarzania danych. Zależność zmian w poziomach sygnału transkryptów od ich składu GC można zaobserwować w przypadku niemal wszystkich badanych eksperymentów, przy czym szczególnie silne są one w przypadku eksperymentów wykonanych na platformach HG-U133A, HG-U133B oraz HuGene-1_0-st (Ryc. 49A). Metody wstępnego przetwarzania danych nieznacznie wpływają na wynik analizy (Ryc. 49B). Nawet metoda GC-RMA, która bazuje na algorytmie korekcji tła opartym o skład nukleotydowy indywidualnych sond charakteryzuje się wynikami podobnymi do uzyskanych standardową wersją algorytmu RMA. Najślabszą zależność zmian profilu ekspresji genów od składu GC transkryptów wykazuje metoda MAS5.0, która różni się od pozostałych tym, że nie uwzględnia informacji o wszystkich próbkach, podczas przetwarzania danych, standaryzując każdą mikromacierz oddzielnie. Najniższa mediana współczynnika SLGC eksperymentów przetworzonych tym algorytmem nie oznacza jednak, że jest on najlepszy gdyż powodem mogą być silne różnice w rozkładach poziomów ekspresji sond pomiędzy próbkami wynikające z zastosowanych algorytmów skalowania i korelacji tła, opartych o sondy PM i MM, nie mające podłoża biologicznego.

Wysoki wpływ składu GC na różnice pomiędzy próbkami nie oznacza, że określony eksperyment dostarcza jedynie informacji, które nie mają podłoża biologicznego, istnieje jednak duże ryzyko, że zmiana poziomu sygnału znacznej liczby transkryptów wynika z różnic technicznych, których żadna z metod wstępnego przetwarzania nie jest w stanie skutecznie ograniczyć. Wykorzystanie powtórzeń technicznych lub biologicznych także nie gwarantuje sukcesu, gdyż w przypadku podobnego trendu w ramach powtórzeń wpływ składu może być równie wysoki jak w przypadku danych z eksperymentu E01. Wykorzystanie większej liczby powtórzeń może obniżyć prawdopodobieństwo uzyskania fałszywie dodatnich wyników w postaci genów różnicujących jednak wysoka wariancja sygnału pomiędzy powtórzeniami, mająca podłoże w różnicach wynikających z obciążenia danych wynikającego ze składu GC, może znacznie obniżyć skuteczność algorytmów identyfikacji genów różnicujących.

Ryc. 49 pokazuje, że skład GC może silnie wpływać na wyniki niemal każdego eksperymentu jednak ważniejsze jest jak te różnice wpływają na wyniki eksperymentów, których celem jest określenie sygnatury genowej w postaci genów o zwiększonej i zmniejszonej ekspresji w badanych warunkach. W tym celu wykorzystano informacje pobrane z bazy danych MSigDB [258], będącej zbiorem najróżniejszych

sygnatur genowych opracowanych różnego typu metodami bazującymi na eksperymentach mikromacierzowych. Wykorzystany zbiór (oznaczony symbolem C6) zawiera 89 sygnatur w postaci genów o zwiększonej i zmniejszonej ekspresji. Spośród wszystkich zbadanych sygnatur genowych 50 na 89 (56.2%) charakteryzuje się znamiennej różnicą w składzie GC pomiędzy genami o zwiększonej i zmniejszonej ekspresji (q -wartość < 0.05 po korekcie na wielokrotne testowanie). Nie oznacza to, że są one nieprawidłowe jednak sugeruje, iż mogły one przynajmniej częściowo powstać w wyniku przetwarzania danych silnie obciążonych poprzez skład GC.

6.5.7. Korekcja wpływu składu GC na poziomy sygnał sond

Czy możliwe jest zmniejszenie wpływu różnic w składzie GC sond na proces identyfikacji genów różnicujących?

Różnice w proporcjach składu nukleotydowego pomiędzy sondami mogą być kompensowane długością oligonukleotydów z jakich są one zbudowane [259], podobnie jak się to odbywa w przypadku oligonukleotydów wykorzystywanych w reakcji PCR. Technologia mikromacierzowa firmy Affymetrix nie zakłada jednak sond o długości innej niż 25 nukleotydów w przypadku wszystkich dostępnych platform a sam problem różnic w składzie został pominięty przy założeniu, że mikromacierze przeznaczone są jedynie do porównywania zmierzonych wartości ekspresji pomiędzy próbkami a nie pomiędzy różnymi genami/transkryptami. Duże dysproporcje w składzie GC sond mogą sprawiać, że zmierzony za pomocą pojedynczej mikromacierzy poziom ekspresji dwóch genów nie jest ze sobą porównywalny. Właściwości sond są jednak stałe pomiędzy mikromacierzami z tego względu odczytane sygnały powinny być ze sobą porównywalne po zastosowaniu algorytmu normalizacji usuwającego różnice o podłożu technicznym pomiędzy poszczególnymi mikromacierzami. W punkcie 6.5.5 pokazano jednak, że najczęściej wykorzystywane algorytmy wstępnego przetwarzania danych traktują wszystkie sondy w ten sam sposób standaryzując je pomiędzy próbkami, pomimo, że różnice w sygnale powiązane są z własnościami sond i w inny sposób wpływają na wariancje sygnału.

Mikromacierze nieraz wykorzystywane są do określania proporcji sygnału pomiędzy sondami lub zestawami sond w przypadku niektórych metod uczenia maszynowego [260] lub niektórych platform mikromacierzowych takich jak mikromacierze promotorowe (ChIP-chip) [261] lub mikromacierze służące do identyfikacji miejsc metylacji DNA (MeDIP-chip) [262]. W takiej sytuacji konieczne wydaje się przeprowadzenie korekty poziomu sygnału opartej o skład GC sond [263], której celem jest uzyskanie porównywalnego poziomu ekspresji pomiędzy różnymi sondami o odmiennych proporcjach GC. Tego typu podejście wpływa jednak bardzo silnie na wartości ekspresji, co bez dokładnego poznania wywołujących je zjawisk fizycznych może prowadzić do utraty informacji. Niniejsza praca zakłada jedynie identyfikację genów różnicujących, przez co istotne nie jest kompensowanie różnic wynikających ze składu GC pomiędzy poszczególnymi sondami z danej macierzy, ale sond pomiędzy różnymi badanymi mikromacierzami.

Z tego względu konieczne jest zastosowanie korekty, która umożliwiłaby skorygowanie obciążenia danych, wynikającego ze składu GC sond, jakie może wpływać na identyfikację genów różnicujących. Algorytm tego typu powinien spełniać trzy następujące założenia:

- Musi być kompatybilny z innymi metodami wstępnego przetwarzania danych – różne metody sprawdzają się lepiej w przypadku danych określonego typu, istotna jest zatem możliwość

wykorzystania korekty w połączeniu z dowolnym typem algorytmu wstępnego przetwarzania danych (np. RMA, GC-RMA, MAS)

- Korekta nie może prowadzić do złamania założeń stawianych metodom normalizacji danych takich jak jednorodności kształtu rozkładu lub zachowanie określonych jego parametrów pomiędzy próbkami
- Algorytm korekty powinien w maksymalny sposób korygować obciążenie danych wynikające ze składu GC jednak przy zachowaniu wysokiej czułości i specyficzności metod detekcji genów różnicujących – zbyt silne wygładzenie danych może doprowadzić do wyeliminowania różnic biologicznych pomiędzy próbkami

Korektę obciążenia wynikającego ze składu GC przeprowadzono zgodnie z metodą zaproponowaną przez Benjaminiego w 2012r. [264], która podejmuje podobny problem obserwowany w przypadku danych z głębokiego sekwencjonowania RNA (RNA-Seq). W tym przypadku problem jest nieco inny i dotyczy dysproporcji w częstotliwości występowania krótkich kilkudziesięcio-nukleotydowych odczytów wynikających z zastosowania amplifikacji RNA opartej o PCR. Problem korekty obciążenia GC jest jednak podobny z tego względu przedstawioną metodę zaadoptowano na potrzeby danych mikromacierzowych. Metoda bazuje na skalowaniu danych w oparciu o współczynniki krzywej, dopasowanej do danych z każdej próbki za pomocą nieparametrycznej regresji lokalnej LOESS (z ang. LOcally Estimated Scatterplot Smoothing). Krzywa regresji jest bardzo zbliżona do mediany sygnału sond zaprezentowanej na Ryc. 38 jednak ma tą przewagę, iż w przypadku sond o bardzo niskim i wysokim składzie (których jest stosunkowo niewiele) jej kształt nie jest silnie uzależniony od poziomu fluorescencji pojedynczych sond, których zmiana pomiędzy próbkami może mieć podłoże biologiczne.

Skalowanie przeprowadzane jest po dopasowaniu krzywej regresji do każdej z próbek i przeprowadzone jest zgodnie z równaniem (1):

$$\hat{S}_{m,p} = S_{m,p} \frac{K_m(GC_p)}{\hat{K}(GC_p)} \quad (1)$$

gdzie $\hat{S}_{m,p}$ to sygnał sondy p odczytany z mikromacierzy m po korekcie, $S_{m,p}$ przed korektą, $K_m(GC_p)$ to współczynnik uzyskany dla sond p o określonej ilości nukleotydów GC poprzez dopasowanie krzywej regresji do danych z próbki m . \hat{K} to średni współczynnik obliczony na podstawie wszystkich linii regresji wyznaczonych dla N próbek z danego eksperymentu (2):

$$\hat{K}(GC) = \frac{1}{N} \sum_{m=1}^N K_m(GC) \quad (2)$$

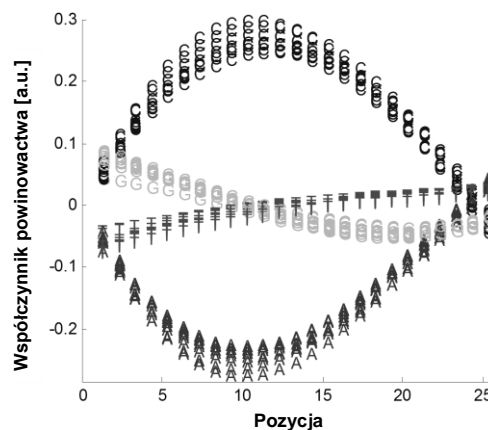
gdzie GC to ilość nukleotydów G i C w badanych sondach (0-25).

Bardzo istotne jest zastosowanie korekty na odpowiednim etapie przetwarzania danych. Można to zrobić przed korektą tła, po korekcji tła a przed normalizacją lub po normalizacji jednak przed sumaryzacją sond (wyznaczenie wartości sygnału dla zestawu). Zastosowanie korekty po normalizacji prowadzi do złamania założeń stawianych przez normalizację kwantylową gdyż na jej skutek rozkłady sygnału poszczególnych próbek mogłyby różnić się od siebie. Zastosowanie korekty przed korektą tła może z kolei zmniejszyć jej efektywność w przypadku gdy próbki silnie różnią się od siebie poziomem niespecyficznej hybrydyzacji (także powiązanego ze składem GC), którego oszacowanie jest oddzielnym

problemem. Z tego względu najlepszym etapem dla korekty jest przeprowadzenie jej po korekcji tła jednak przed normalizacją.

Metoda GC-RMA jest jedną z nielicznych wykorzystujących informację o strukturze nukleotydowej sondy, jednak jak pokazano w punkcie 6.2.1 nie kompensuje ona różnic wynikających ze składu GC pomiędzy próbkami. GC-RMA polega na zastosowaniu alternatywnej metody korekcji tła, która wyznacza poziom niespecyficznego hybrydowania w oparciu o macierz powinowactwa, zbudowaną dla określonych nukleotydów na określonych pozycjach sondy [151]. Macierz ta jest podstawą do wyznaczenia poziomu niespecyficznego hybrydowania poszczególnych sond w zależności od ich struktury nukleotydowej określonej przez proporcje poszczególnych zasad i ich pozycje wewnątrz sekwencji sondy [265]. Źródłem danych do stworzenia macierzy jest eksperyment wykonany przez autorów metody, w którym na pojedynczą mikromacierz nanoszony jest materiał genetyczny wyizolowany z organizmu, który nie jest kompatybilny z wybraną mikromacierzą (nie wykazuje specyficzności do sond umieszczonych na mikromacierzy). Poziom niespecyficznego hybrydowania różni się pomiędzy eksperymentami ze względu na niewielkie różnice w warunkach reakcji [255], dlatego dobrą alternatywą dla wykorzystania danych z eksperymentu autorów metody jest wykorzystanie sygnałów sond MM (Mismatch) określonej mikromacierzy do zbudowania macierzy powinowactwa charakterystycznej dla danego eksperymentu (zwykle w oparciu o pierwszą mikromacierz z wczytanego zbioru danych).

Ryc. 50 przedstawia wykres stworzony na podstawie macierzy powinowactwa dla różnych próbek z eksperymentu E01 pokazując jak silnie mogą różnić się wagi dla poszczególnych pozycji, pomiędzy próbkami, szczególnie w przypadku nukleotydów C i A.



Ryc. 50: Wykresy współczynników macierzy powinowactwa dla wszystkich 8 próbek z eksperymentu E01. Przebiegi dla poszczególnych nukleotydów zaznaczone różnym znacznikiem oraz kolorem określają współczynniki macierzy powinowactwa dla określonej pozycji w sekwencji sondy. Im wyższa jest sumaryczna wartość współczynnika dla wszystkich pozycji tym mocniejsze jest wiązanie sonda-transkrypt.

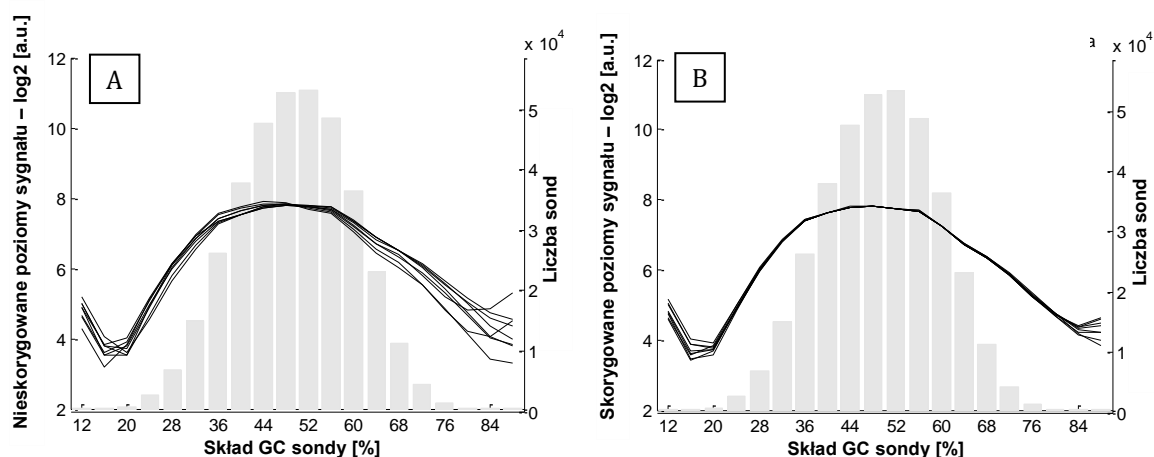
Wykorzystanie pojedynczej macierzy powinowactwa, wspólnej dla wszystkich próbek może być zatem nieodpowiednie, z tego względu poza skalowaniem opartym o proporcje GC wprowadzono dodatkową modyfikację do metody GC-RMA, która wyznacza oddzielną macierz powinowactwa dla każdej mikromacierzy i na jej podstawie dokonuje korekty tła dla danych z odpowiadającej jej próbki.

Zaproponowany algorytm przetwarzania mikromacierzy, który dla ułatwienia oznaczono akronimem **csGC-RMA** (corrected sample-based GC-RMA) składa się z następujących etapów:

- Korekcja tła oparta o zmodyfikowaną wersję algorytmu GC-RMA, która bazuje na sygnałach sond MM każdej indywidualnej próbki podczas oceny poziomu niespecyficznego hybrydyzacji
- Korekcja dysproporcji w sygnale sond o różnym składzie GC bazująca na dopasowaniu do danych regresji lokalnej LOESS
- Normalizacja kwantylowa (stosowana we wszystkich wariantach algorytmu RMA oraz w metodach PLIER i FARMS)
- Sumaryzacja typu median polish (stosowana we wszystkich wariantach algorytmu RMA)

Ocenę jakości korekcji przeprowadzono w dwóch etapach. Pierwszy z nich oparty jest o dane z eksperymentu E01 oceniając wpływ algorytmu na zmianę stopnia korelacji pomiędzy zmianą poziomu sygnału zestawów sond a składem GC transkryptów. Celem drugiego etapu jest ocena wpływu algorytmu na skuteczność identyfikacji genów różnicujących, bazująca na dwóch publicznie dostępnych zbiorach danych zaprojektowanych z myślą o tego typu badaniach.

Wykresy na Ryc. 51 pokazują linie regresji dopasowane do danych bez korekcji, przetworzone standardową wersją algorytmu GC-RMA (wykres A) oraz po korekcji i przetwarzaniu zaproponowanym algorytmem csGC-RMA (wykres B).



Ryc. 51: Linie regresji dopasowane do danych przetworzonych algorytmem GC-RMA (A) oraz po przetworzeniu zaproponowaną metodą wstępnego przetwarzania csGC-RMA (B). Szare histogramy oraz odpowiadająca im prawa oś Y pokazują ilości sond o określonym składzie GC

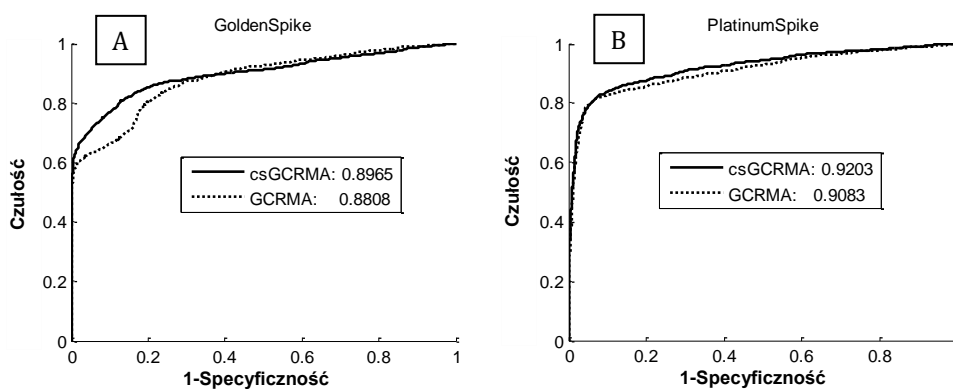
Linie regresji przed korektą charakteryzują się znacznym przesunięciem wynikającym z przetwarzania za pomocą normalizacji kwantylowej danych wykazujących różnice w proporcjach sygnału sond o wysokim i niskim GC. Po korekcji csGC-RMA przesunięcie jest znacznie zmniejszone, co obniża obserwowane wcześniej korelacje pomiędzy różnicą w poziomie sygnału pomiędzy próbkami (LFC) a składem GC transkryptów (Tab. 15).

Czas po napromieniowaniu		1h	12h	24h
Współczynnik korelacji Spearmana pomiędzy zmianą ekspresji a składem GC transkryptu	Przed zastosowaniem korekty	-0,440	-0,508	-0,181
	Po zastosowaniu korekty na skład GC sond	-0,031	-0,084	-0,021

Tab. 15: Korelacja Spearmana pomiędzy składem GC transkryptu a zmianą ekspresji po napromieniowaniu (LFC) w eksperymencie E01 przeprowadzonym na komórkach Me45. Wartości przed korektą prezentowane wcześniej w Tab. 7 zaznaczono szarym kolorem, poniżej wartości uzyskane po korekcji metoda csGC-RMA.

Samo zmniejszenie korelacji oraz zgodność linii regresji, która jest oczywistą konsekwencją skalowania opartego o ich kształt nie świadczy jednak o wysokiej jakości zaproponowanej metody. Zbyt silne spłaszczenie sygnału może dać podobny efekt zmniejszając jednak różnice biologiczne pomiędzy próbkami. Ocena wpływu korekty na identyfikację genów różnicujących jest jednak niemożliwa do przeprowadzenia w przypadku danych z eksperymentu E01, ponieważ geny, których ekspresja ulega zmianie na skutek promieniowania nie są znane dla tej linii komórkowej. Z tego względu niemożliwe jest ocenienie wpływu metody na czułość i specyficzną procesy identyfikacji genów różnicujących. W celu dokonania tego typu oceny wykorzystano dwa dodatkowe zbiory danych GoldenSpike [226] oraz PlatinumSpike [227], w których do określonej stałej mieszaniny RNA dodano kilkanaście dodatkowych transkryptów w znanych proporcjach będących zgodnie z założeniami jedynymi źródłami zmienności pomiędzy parami próbek A i B, o podłożu innym niż techniczne. Zbiór GoldenSpike zawiera wyłącznie transkrypty, których ekspresja w próbce B jest większa od A. Zbiór PlatinumSpike jest ulepszoną wersją, w której próbka B zawiera także geny o zmniejszonej ekspresji w stosunku do A co jest zgodne z założeniami algorytmów wstępnego przetwarzania danych (zbliżona liczba genów o zwiększonej i zmniejszonej ekspresji). Dokładny opis obu zbiorów danych znajduje się w punkcie 4.1.2.

Ryc. 52 przedstawia krzywe ROC (Receiver Operating Characteristic) dla obu zbiorów danych po zastosowaniu standardowej metody przetwarzania GC-RMA oraz zaproponowanej implementacji algorytmu csGC-RMA wykonane zgodnie z metodologią opisaną w punkcie 4.3.



Ryc. 52: Krzywe ROC dla danych ze zbioru GoldenSpike (A) oraz PlatinumSpike (B) po zastosowaniu standardowej wersji metody przetwarzania danych GC-RMA oraz zaproponowanej metody csGC-RMA. Wartości umieszczone w legendach to pola pod odpowiednimi krzywymi.

W przypadku obu zbiorów danych metoda csGC-RMA ma przewagę nad standardową wersją algorytmu GC-RMA charakteryzując się większą czułością i specyficzną podczas identyfikacji genów różnicujących, co szczególnie wyraźnie widoczne jest w przypadku zbioru PlatinumSpike. Świadczy o tym wzrost pola pod krzywą ROC, którego wartość umieszczono w legendzie wykresu (Ryc. 52B). Pokazuje to, że zastosowanie korekty wpływa na obniżenie liczby fałszywie dodatnich wyników (False-Positive) inaczej określanej jako 1-specyficznosc przy jednoczesnym zwiększeniu liczby prawdziwie dodatnich (True-Positive), poprawiając czułość algorytmu. Zastosowanie samej korekty na skład GC sond lub samej modyfikacji GCRMA także poprawia rezultaty, ale jedynie połączenie ich obu pozwala uzyskać tak dobre wyniki.

6.6. Identyfikacja cech transkryptów różnicujących po przetworzeniu danych metodą csGC-RMA

6.6.1. Statystyki transkryptów różnicujących

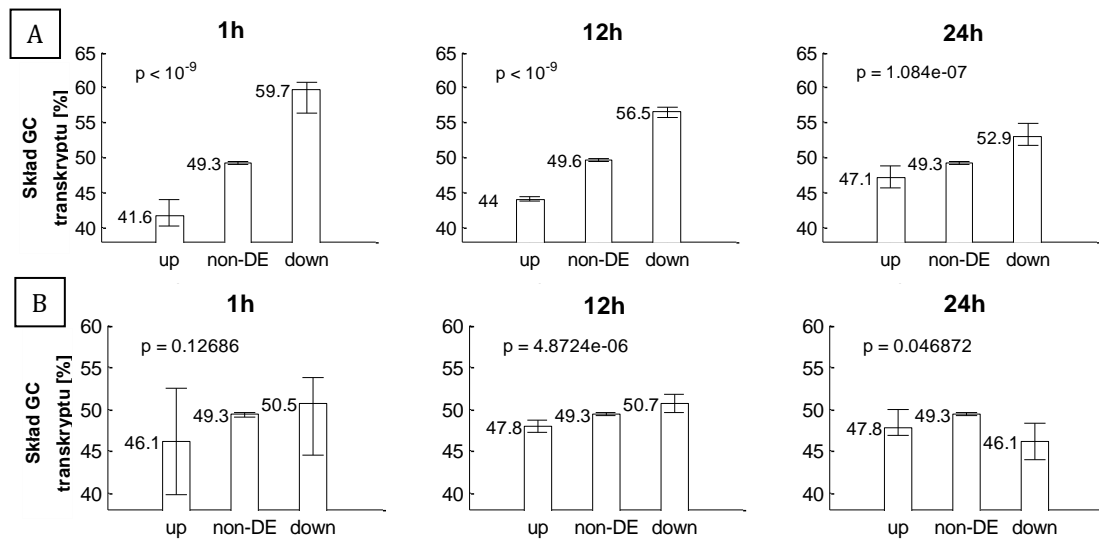
Zastosowanie zaproponowanej metody przetwarzania danych csGC-RMA znacznie obniżyło liczbę transkryptów różnicujących w komórkach Me45, zidentyfikowanych za pomocą algorytmu Limma (Tab. 16). Większość ze zidentyfikowanych transkryptów należy jednak do listy genów różnicujących zidentyfikowanych za pomocą standardowej metody przetwarzania (wyniki z rozdziału 6.1.2 dla metody WS – opartej o wnioskowanie statystyczne).

Czas po napromieniowaniu	Liczba genów różnicujących komórek Me45		
	wzrost	brak zmian	spadek
1h	134 (66%)	23026	93 (81%)
12h	1513 (94%)	21162	578 (94%)
24h	438 (87%)	22647	168 (90%)

Tab. 16: Liczba transkryptów różnicujących po przetworzeniu metodą csGC-RMA oraz procent genów należących do sygnatury zidentyfikowanej za pomocą metody przetwarzania danych opisanej w punkcie 6.1.2 (WS). Wyniki uzyskano dla danych z eksperymentu E01 przeprowadzonego na komórkach Me45.

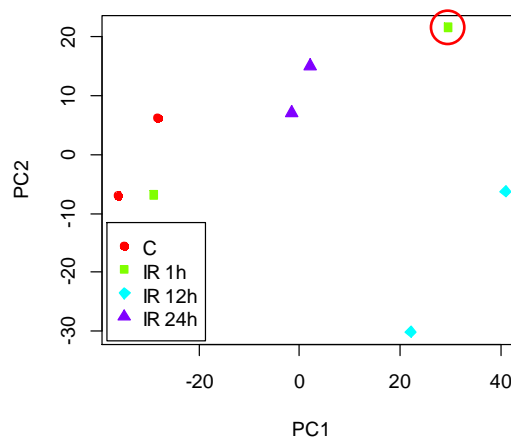
Pomimo, że liczba genów jest znacznie mniejsza niż uzyskana w rozdziale 6.1.2 to wydaje się, że zastosowanie algorytmu csGC-RMA wpłynęło korzystnie na identyfikację transkryptów różnicujących. Zmniejszenie wariancji pomiędzy powtórzeniami technicznymi, wynikające z obciążenia poprzez skład GC, pozwoliło na wyeliminowanie potrzeby stosowania algorytmu kompensacji *batch effect*, bez którego identyfikacja transkryptów różnicujących w eksperymencie E01 była znacznie utrudniona w związku z wysoką wariancją pomiędzy powtórzeniami technicznymi.

Ryc. 53 pokazuje procentowy skład GC transkryptów różnicujących zidentyfikowanych po zastosowaniu algorytmu csGC-RMA, który jest analogiczny do wykresów pokazanych na Ryc. 29 wykonanych na podstawie standardowego przetwarzania danych. Pomimo znacznego obniżenia liczby genów różnicujących i wyeliminowania korelacji składu GC ze zmianą ekspresji (Tab. 15), różnice w składzie GC pomiędzy próbkami są nadal bardzo wysokie (Ryc. 53A). Dopiero zastosowanie podobnej korekty opartej o skład GC transkryptu (zamiast składu nukleotydowego sond) pozwala dodatkowo zmniejszyć dysproporcje w składzie nukleotydowym genów różnicujących (Ryc. 53B), co sugeruje, iż w niektórych przypadkach różnice w intensywności sygnału mogą być silnie powiązane z czynnikami uzależnionymi od składu GC transkryptu (omówione w rozdziale 6.5.4) a nie pojedynczych sond. Korekta oparta o skład GC transkryptów nie może być jednak zastosowana ponieważ nie można założyć, że średni skład GC transkryptów powinien być jednakowy pomiędzy grupami genów różnicujących a tego typu korekta prowadziłyby do wyeliminowania podobnych różnic. Z tego względu dalsze obliczenia bazują wyłącznie na wynikach algorytmu csGC-RMA eliminującego różnice w poziomie sygnału pomiędzy sondami o różnym składzie GC.



Ryc. 53: Wykresy słupkowe mediany składu GC poszczególnych grup genów różnicujących komórek Me45 po 1, 12 i 24 h od napromieniowania (eksperyment E01). A: wykresy dla danych przetworzonych metodą csGC-RMA (korekta na podstawie składu GC sond). B: metoda analogiczna do csGC-RMA jednak bazująca na składzie GC transkryptów. P-wartości nad wykresami pochodzą z testu-t porównującego średni skład GC transkryptów pomiędzy grupami up i down, reprezentującymi transkrypty o odpowiednio zwiększonej i zmniejszonej ekspresji.

Po czasie 1h od napromieniowania zmianom ulega najmniej transkryptów (227) pochodzących od 115 różnych genów. Niewielka ich liczba może świadczyć o tym że jedna godzina to czas zbyt krótki na zmianę poziomu ekspresji znaczącej liczby genów w odpowiedzi na promieniowanie. Bardzo prawdopodobnym powodem mogą być jednak duże różnice pomiędzy oboma powtórzeniami technicznymi, których wysoka wariancja utrudnia identyfikację genów różnicujących. Duże różnice pomiędzy oboma powtórzeniami biologicznymi są bardzo dobrze widoczne w przypadku analizy PCA przeprowadzonej dla danych przetworzonych algorytmem csGC-RMA (Ryc. 54). W przypadku innych metod przetwarzania danych (standardowa wersja GC-RMA lub RMA) różnice są równie wysokie.



Ryc. 54: Analiza PCA dla danych z eksperymentu E01 przetworzonych algorytmem csGC-RMA. Czerwonym kółkiem oznaczono próbkę E01_Me45_IR_1h_1.

Standardowa kontrola jakości nie wykazała żadnych istotnych różnic pomiędzy obiema próbkami z czasu 1h, a niewielkie odchylenia w proporcjach sygnału sond kontrolnych, położonych w obszarach 3' i 5' (Ryc. 26), których celem jest ocena stopnia degradacji RNA, mieszczą się w dopuszczalnym zakresie określonym przez producenta mikromacierzy. Mimo to różnice mogą wynikać ze zwiększonego stopnia degradacji

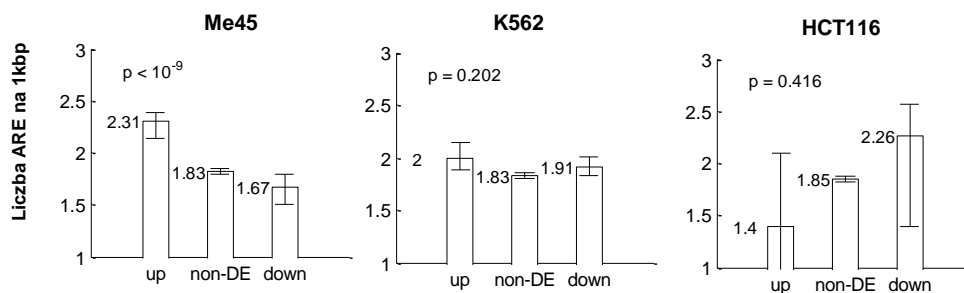
RNA w próbce E01_Me45_IR_1h_1. Prosta regresji przeprowadzona przez sygnały z tej próbki pogrupowane według odległości miejsca hybrydyzacji sondy od końca 3' transkryptu (analogicznie do krzywych na Ryc. 13) pokazuje, że próbka E01_Me45_IR_1h_1 charakteryzuje się wyższym współczynnikiem nachylenia linii regresji (2,616), istotnie większym od średniej dla wszystkich analizowanych próbek, który wynosi 1,786 (p-wartość=2.75e-11 testu na równość współczynników nachylenia linii regresji). Jest to najwyższa wartość ze wszystkich badanych próbek przy czym im większy współczynnik to tym większa jest różnica pomiędzy sygnałami sond z końca 5' i 3' co może być spowodowane wysokim stopniem degradacji badanego RNA. Badanie integralności RNA przeprowadzone na początku eksperymentu nie wykazało jednak żadnych nieprawidłowości (RIN dla wszystkich próbek >9.5).

Po czasie 12h zmianie ulega najwięcej transkryptów – 2091 (1011 genów) co sugeruje istotne znaczenie mechanizmów regulacji ekspresji genów na tym etapie odpowiedzi komórkowej. Z kolei po czasie 24h liczba genów różnicujących jest już o połowę mniejsza niż w przypadku 12 godziny, co może być przyczyną tego, że komórka zaczyna powoli wracać do stanu, przed napromieniowaniem.

6.6.2. Mechanizmy regulacji ekspresji genów

Regulacja poprzez motywy ARE

Geny, których ekspresja uległa zwiększeniu w komórkach Me45 po 12h od napromieniowania charakteryzują się większą liczbą motywów ARE klasy III (Ryc. 55), które mogą mieć zarówno pozytywny jak i negatywny wpływ na poziom ekspresji mRNA w zależności od typu przyłączonego białka. Zależność ta nie zanika po zastosowaniu korekty csGC-RMA opartej na składzie GC transkryptów. Podobnej zależności nie wykazują jednak pozostałe z badanych komórek – K562 i HCT116 gdzie różnice w średniej ilości motywów ARE pomiędzy genami o zwiększonej i zmniejszonej ekspresji nie są znamienne statystycznie.

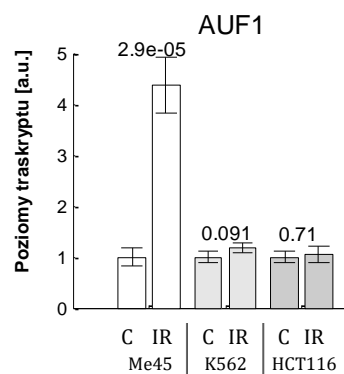


Ryc. 55: Średnia liczba motywów typu ARE (klasy III) w genach różnicujących komórek Me45, K562 i HCT116 po 12h od napromieniowania. P-wartości nad wykresami pochodzą z testu-t porównującego średnie ilości wystąpień motywów ARE pomiędzy grupami up i down, reprezentującymi transkrypty o odpowiednio zwiększonej i zmniejszonej ekspresji.

Spośród 10 znanych genów, które biorą udział w procesach regulacji ekspresji opartych o oddziaływanie z motywami ARE zmiana ekspresji wyłącznie jednego jest znamienne statystycznie w komórkach Me45. Gen AUF1 (znany też jako HNRNPD) charakteryzuje się znacznie wyższym poziomem ekspresji w komórkach Me45 po 12 godzinach od napromieniowania (FC=1.73, q-wartość=0.025). Silny wzrost poziomu ekspresji genu AUF1 potwierdzają też wyniki uzyskane na podstawie mikromacierzy Agilent (FC=1.56). Obie platformy jednoznacznie nie wykazują zmian w poziomach ekspresji genu AUF1 w żadnej

innej badanej linii komórkowej (K562, HCT116). Wzrost poziomu ekspresji genu AUF1 w komórkach Me45 dodatkowo potwierdzają niezależne eksperymenty oparte o technologie RT-qPCR (Ryc. 56).

Pomimo, że AUF1 zmienia swój poziom jedynie w przypadku napromieniowanych komórek Me45 nie oznacza to, że regulacja oparta o białko Auf1 może mieć większe znaczenie po napromieniowaniu tylko w tym przypadku. Ilość transkryptów nie zawsze jest proporcjonalna do ilości białka, dodatkowo aktywność genu AUF1 może być kontrolowana post-transkrypcyjnie poprzez procesy fosforylacji białka [266], czego badania mikromacierzowe nie są w stanie pokazać. Gen AUF1 dodatkowo zawiera 4 alternatywne formy splicingowe kodujące białka o różnych strukturach i masach 37, 40, 42 i 45 kDa (kilodaltonów). Białko p37 charakteryzuje się największym potencjałem regulacyjnym ze wszystkich 4 form [267], jednak mikromacierze nie są w stanie odróżnić poszczególnych form splicingowych gdyż 24 spośród 34 sond specyficznych dla genu AUF1 jest wspólna dla wszystkich 4 form splicingowych natomiast pozostałe 10 wiąże się z formami p40 oraz p42. Różne formy splicingowe mogą dodatkowo być regulowane w oddzielny sposób (z powodu różnic w strukturze obszaru 3'-UTR), generując różne proporcje białek p37-p45 w komórce, co jak pokazano odgrywa istotną rolę w przypadku roli białka Auf1 w komórkach K562 [268].



Ryc. 56: Wyniki z eksperymentu RT-qPCR dla genu AUF1 przeprowadzonego na komórkach kontrolnych (C) i po 12h od poddania komórek działaniu promieniowania (IR). P-wartości nad słupkami pochodzą z testu-t porównującego średnie poziomy ekspresji przed i po napromieniowaniu.

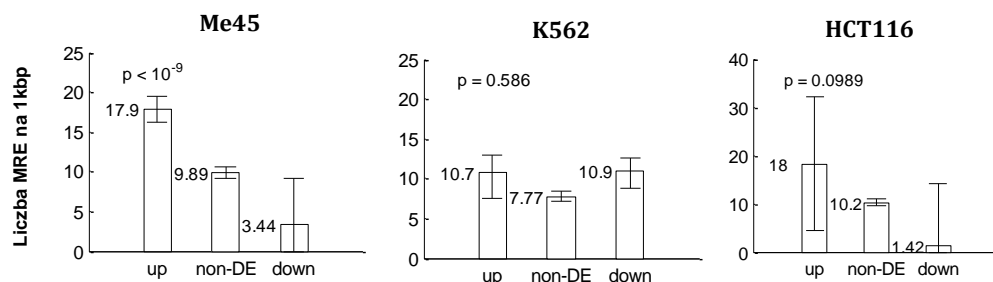
Białko kodowane przez gen AUF1 zwiększa poziom degradacji mRNA zawierających motywy typu ARE jednak określenie globalnych zmian w poziomach ekspresji genów regulowanych przez AUF1 jest bardzo trudne ze względu na to, że transkrypty zawierające motywy ARE mogą być także regulowane przez innego typu mechanizmy. Test na nadreprezentację motywów opisanych w pracach [39, 269] a także motywów, które doświadczalnie potwierdzono jako oddziałujące z białkiem Auf1 [270, 271] w genach o zmniejszonej ekspresji nie pokazał znamiennych różnic na poziomie istotności 0,01.

Regulacja oparta o ARE najprawdopodobniej zachodzi wyłącznie na poziomie pojedynczych transkryptów, przez co niemożliwe jest zaobserwowanie globalnych zmian w poziomach ekspresji pomimo różnic pokazanych na Ryc. 55. Mimo to białko Auf1 może odgrywać bardzo istotną rolę w odpowiedzi komórek Me45 na promieniowanie jonizujące.

Regulacja poprzez oddziaływania z miRNA

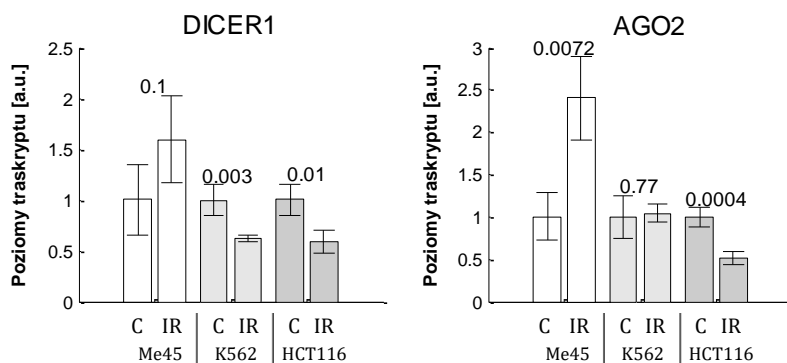
Poziomy dojrzałych form miRNA nie są możliwe do zbadania za pomocą standardowych mikromacierzy, z tego względu analizę uzupełniono o wyniki uzyskane z mikromacierzy Agilent dedykowanych do analizy miRNA. Zmiana profilu ekspresji miRNA może pociągać za sobą bardzo poważne konsekwencje dla genów,

których mechanizmy regulacji ekspresji są uzależnione od degradacji mRNA sterowanej przez cząsteczki miRNA. Transkrypty o zwiększonej ekspresji w komórkach Me45 zawierają znacznie więcej miejsc wiązania 262 miRNA, które ulegają ekspresji w tej linii komórkowej (Ryc. 57). Transkrypty o zwiększonej i zmniejszonej ekspresji, po napromieniowaniu komórek, zidentyfikowane przy użyciu danych przetworzonych algorytmem csGC-RMA nadal różnią się pod względem składu GC. Mimo to sumaryczna liczba motywów MRE wybranej grupy miRNA nie jest skorelowana ze składem nukleotydowym obszaru 3'-UTR, w którym były one badane ($Rho = -0.0351$).



Ryc. 57: Mediana liczby miejsc wiązania miRNA ulegających ekspresji w genach różnicujących o zwiększonej, zmniejszonej i niezmiętej ekspresji (odpowiednio up, down, non-DE) w komórkach Me45, K562 i HCT116 po 12h od napromieniowania. P-wartości nad wykresami pochodzą z testu-t porównującego średnie ilości wystąpień motywów pomiędzy grupami up i down, reprezentującymi transkrypty o odpowiednio zwiększonej i zmniejszonej ekspresji.

Korekta csGC-RMA bazująca na składzie nukleotydowym transkryptów, która w znaczny sposób obniża różnice w medianie składu nukleotydowego pomiędzy grupami genów różnicujących nie wpływa znacząco na różnice w proporcji MRE pomiędzy genami o zwiększonej i zmniejszonej ekspresji (mediana 17.5; 10.1; 1.73 odpowiednio dla grup up, non-DE, down). Podobna zależność nie jest obserwowana w przypadku komórek K562 i HCT116 (Ryc. 57).



Ryc. 58: Wyniki z eksperymentu RT-qPCR dla genów uczestniczących w procesie biogenezy miRNA - DICER1 i AGO2 przeprowadzonego na komórkach kontrolnych (C) i po 12h od poddania komórek działaniu promieniowania (IR). P-wartości nad słupkami pochodzą z testu-t porównującego średnie poziomy ekspresji przed i po napromieniowaniu.

Jednym z kluczowych genów regulowanych przez białko Auf1 jest DICER1 [272]. Auf1 może obniżać jego ekspresje poprzez liczne motywy ARE wszystkich znanych klas, znajdujące się w sekwencji 3'-UTR transkryptu [272]. Wyniki badań mikromacierzowych pokazują, że poziom ekspresji genu DICER1 spada w 12h po napromieniowaniu w komórkach Me45 ($FC = 0.56$, q -wartość = 0.041) oraz w K562 ($FC = 0.84$, q -

transfekcji komórek plazmidem zawierającym gen AUF1 (obniżający poziom DICER1) jednak wcześniejszych czasów nie opisano. Czas, po którym spada poziom miRNA może być znacznie krótszy w napromieniowanych komórkach, które regulowane są w naturalny sposób i dodatkowo charakteryzują się zwiększonym poziomem degradacji RNA indukowanym uszkodzeniami wynikającymi z poddania komórek wpływowi promieniowania.

Wzrost ekspresji pojedynczych miRNA jaki jest obserwowany w przypadku komórek Me45 (hsa-mir-362) oraz K562 (hsa-mir-494) jest możliwy nawet pomimo spadku poziomu ekspresji genu DICER1, który może być powiązany z obniżeniem poziomu jego białka. Niektóre miRNA mogą przechodzić proces biogenezy niezależny od genu DICER1, w którym kluczową rolę odgrywa białko Ago2 [280]. Mechanizm wyboru ścieżki biogenezy przez określony miRNA jest wciąż bardzo słabo poznany jednak najprawdopodobniej zależy od struktury drugorzędowej prekursora dojrzałej cząsteczki miRNA (pre-miRNA). Wysoki stopień dopasowania zasad na przeciwległych niciach pre-miRNA obniżający ilości i rozmiar wybrzuszeń (szczególnie u podstawy struktury) jest prawdopodobnie najistotniejszą cechą decydującą o przejściu do ścieżki biogenezy opartej wyłącznie o białko Ago2 [281], którego poziom transkrypty bardzo silnie wzrasta w napromieniowanych komórkach Me45 (Ryc. 58).

Ponieważ DICER1 kontroluje globalną ekspresję miRNA, które z kolei regulują ekspresję znacznej części genów to zmiana jego poziomu może silnie wpływać na globalne poziomy transkryptów w komórce. Obniżenie poziomu białek Ago2 lub Dicer może w skrajnych przypadkach prowadzić do śmierci komórkowej, pomimo, że brak miRNA nie wpływa na same mechanizmy naprawy DNA [274]. Obniżenie wydajności procesów produkcji miRNA może prowadzić do zaburzenia mechanizmów odpowiedzialnych za zatrzymanie cyklu komórkowego i uruchomienia procesów apoptozy w odpowiedzi na uszkodzenia DNA wywołane promieniowaniem, co zaobserwowano w pracy [282]. DICER1 dodatkowo może być degradowany przez kaspazy (enzymy kontrolujące apoptozę) regulując w ten sposób procesy uzależnione od degradacji poprzez cząsteczki miRNA [274]. Tego typu mechanizmy różnią się pomiędzy różnymi typami komórek i w niektórych przypadkach inaktywacja białek biorących udział w biogenezie miRNA (Ago2, Dicer, Drosha) ma znikomy wpływ na odpowiedź komórkową indukowaną promieniowaniem [283].

DICER1 jest bardzo silnie regulowany przez białko Auf1 jednak nie jest to jedyny z poznanych mechanizmów kontrolujących jego ekspresję. Mechanizmy regulacji genu DICER1 obejmują czynnik transkrypcyjny Sox4, który jak pokazano w pracy [284] na przykładzie komórek czerniaka zwiększa aktywność DICER1 poprzez oddziaływanie z sekwencją jego promotora. Ekspresja Sox4 silnie spada po napromieniowaniu w komórkach Me45 (FC=0.48; q-wartość=0.046) przez co obok białka Auf1 może być on jednym z czynników odpowiedzialnych za spadek ekspresji DICER1 w komórkach Me45. Poziomy transkrypty Sox4 nie zmieniają się w napromieniowanych komórkach K562 i HCT116.

DICER1 może być także regulowany poprzez miRNA z rodziny let-7, które przyspieszają jego degradację tworząc w ten sposób ujemną pętlę sprzężenia zwrotnego (DICER1 jest odpowiedzialny za biogenezę miRNA), utrzymującą określony poziom białka Dicer w komórce [285]. W przypadku zarówno komórek Me45 jak i K562 poziomy ekspresji let-7 bardzo silnie spadają razem z poziomem DICER1, co sugeruje, iż układ został wytrącony ze stanu ustalonego poprzez wpływ promieniowania.

Mechanizmy regulacji oparte o miRNA bardzo silnie różnią się pomiędzy różnymi typami komórek i dotychczas nie znaleziono żadnych wspólnych sygnatur miRNA charakterystycznych dla promieniowania we wszystkich badanych komórkach. Przykładowo miRNA z grupy hsa-let-7 uważane są za mające istotny

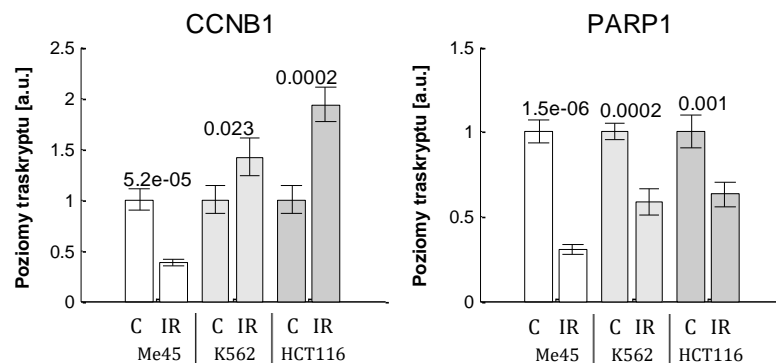
wpływ na odpowiedź komórek na stres oksydacyjny [286], jednak w niektórych komórkach obserwowany jest ich wzrost po napromieniowaniu [287] w innych silny spadek [288]. Dodatkowo pokazano, że kierunki zmian w profilach ekspresji miRNA są silnie uzależnione od dawki promieniowania [289].

6.6.3. Odpowiedź komórek na promieniowanie

Regulacja cyklu komórkowego

Komórki czerniaka przez wiele lat uważane były za odporne na promieniowanie najnowsze badania jednak temu zaprzeczają [290]. Radio-oporność jest często związana z zatrzymaniem cyklu komórkowego w fazie G2 [291-293], którego komórki Me45 w przeciwieństwie do komórek K562 nie wykazują [294, 295]. Może to być jedną z przyczyn zwiększonej wrażliwości linii komórkowej Me45 na czynniki genotoksyczne w tym promieniowanie jonizujące [296].

Zatrzymanie cyklu komórkowego zwykle odbywa się poprzez aktywację białka p53, które reguluje mechanizm zatrzymania cyklu w fazie G1-S. W komórkach z nieprawidłowo działającym szlakiem p53 zatrzymanie cyklu może zachodzić wyłącznie w fazie G2 poprzez mechanizm niezależny od szlaku sygnałowego p53 [297]. Białko p53 aktywuje transkrypcję licznych genów w tym PTEN i MDM2 w odpowiedzi na uszkodzenia DNA wywołane promieniowaniem [298]. Wyniki przeprowadzonych badań mikromacierzowych nie pokazują jednak istotnych zmian w poziomach ekspresji tych genów co dodatkowo potwierdza, że ścieżka sygnałowa p53 nie pełni swojej prawidłowej funkcji pomimo, że gen TP53 kodujący białko p53 ulega ekspresji zarówno w komórkach Me45 jak i K562 oraz HCT116. Utrata prawidłowego działania ścieżki sygnałowej p53 (powszechna w przypadku komórek nowotworowych) sprawia, że w przypadku żadnej z linii komórkowych nie zachodzi zatrzymanie cyklu komórkowego w fazie G1-S, po ich napromieniowaniu.



Ryc. 59: Wyniki z eksperymentu RT-qPCR dla genów CCNB1 i PARP1 przeprowadzonego na komórkach kontrolnych (C) i po 12h od poddania komórek działaniu promieniowania (IR). P-wartości nad słupkami pochodzą z testu-t porównującego średnie poziomy ekspresji przed i po napromieniowaniu.

Zatrzymanie cyklu komórkowego w fazie G2 w odpowiedzi na uszkodzenia DNA najczęściej powiązane jest z uruchomieniem procesów prowadzących do zatrzymania fosforylacji białka Cdc2 [299]. Zaobserwowano także, że zatrzymanie cyklu w fazie G2 jest poprzedzone zwiększoną ekspresją genu CCNB1 kodującego cyklinę B2 [291]. Cyklina B2 tworzy kompleksy z ufosforylowanym białkiem Cdc2 tworząc jeden z mechanizmów prowadzących do zatrzymania cyklu komórkowego w fazie G2 [299]. Badania mikromacierzowe nie wykazały zmiany w poziomach traskryptu genu CCNB1 w komórkach

Me45 jednak jego poziom bardzo silnie wzrasta w komórkach K562 (czasy 1, 12 i 24h po napromieniowaniu). Wzrost ekspresji w 12h po napromieniowaniu w przypadku komórek K562 potwierdza także eksperyment RT-qPCR (Ryc. 59). Podobny wynik uzyskano dla komórek HCT116, które podobnie jak K562 zatrzymują cykl komórkowy w odpowiedzi na promieniowanie.

Pomimo, że badania mikromacierzowe nie wykazały różnic w poziomie ekspresji genu CCNB1 dla komórek Me45 to jego spadek po 12h od napromieniowania, który pokazuje eksperyment RT-qPCR może być powiązany z zaburzeniem mechanizmu zatrzymania cyklu komórkowego w fazie G2, który jest charakterystyczny dla komórek K562 [295] i HCT116 [300].

Mechanizmy naprawy DNA

Mechanizmy naprawy DNA uruchamiane są zaraz po napromieniowaniu a o ich wysokiej wydajności świadczy bardzo szybki spadek poziomu indukowanych promieniowaniem pęknięć nici DNA, który przy dawce 4Gy już po około 2h osiąga poziom kontrolny [295, 301]. Ponieważ mechanizm ten musi działać bardzo szybko to korzysta on z białek, które stale produkowane są w komórce a w określonych warunkach są jedynie aktywowane poprzez kinazy ATM i ATR [302], które uczestniczą w procesach fosforylacji białek odpowiedzialnych za naprawę uszkodzeń DNA. Mimo to w przypadku transkryptów o zwiększonej ekspresji w komórkach K562 zaobserwowano silną nadreprezentację genów uczestniczących w procesach naprawy DNA, opisanych w pracy [303] (p-wartości = $3.82 \cdot 10^{-04}$; 0.017; 0.033 odpowiednio dla czasów 1, 12 i 24h po napromieniowaniu). Najsilniejsza nadreprezentacja po czasie 1h pokazuje, jak szybka może być odpowiedź komórkowa na stres związany z napromieniowaniem, który już po krótkim czasie może indukować odpowiedź na poziomie transkryptomu.

W przypadku komórek Me45 podobna zależność nie jest obserwowana a wzrost poziomu ekspresji genów uczestniczących w procesach naprawy DNA obserwowany jest jedynie po czasie 12h od napromieniowania. Mimo to, wśród transkryptów różnicujących zidentyfikowanych po 12h od napromieniowania komórek Me45, znajdują się transkrypty kodujące białka uczestniczące we wszystkich znanych mechanizmach naprawy DNA od naprawy niesparowanych zasad (gen MSH2, FC=2,9 q-wartość=0,014) po usuwanie zmodyfikowanych nukleotydów (gen TDG, FC=2,83; q-wartość=0.049).

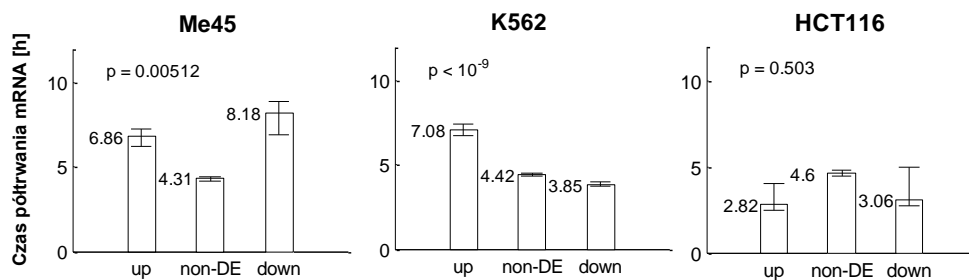
W 12h po napromieniowaniu komórki Me45 wykazują znaczne obniżenie poziomu ekspresji genu PARP1 (FC=0.69, q-wartość=0.022), znacznie silniejsze niż w przypadku pozostałych linii komórkowych, co dodatkowo potwierdzają wyniki z eksperymentu RT-qPCR (Ryc. 59). PARP1 jest odpowiedzialny za naprawę DNA, jednak ostatnie doniesienia literaturowe sugerują, że może on być także odpowiedzialny za obniżanie poziomu wolnych rodników w postaci reaktywnych form tlenu [301]. Obserwowany spadek poziomu mRNA PARP1 może być zatem jedną z przyczyn silnego wzrostu wolnych rodników w komórkach Me45 obserwowaną wiele godzin po napromieniowaniu [304].

Stabilność RNA

Stabilność RNA odgrywa bardzo istotną rolę w mechanizmach regulacji ekspresji. Transkrypty o krótkim czasie półtrwania (mniejszym niż 2h) są zwykle powiązane z procesami regulacyjnymi co w połączeniu z wydajnymi mechanizmami aktywacji procesów transkrypcji pozwala w bardzo szybki sposób wpływać na poziom określonych mRNA w komórce [305]. Transkrypty o długim czasie półtrwania są natomiast charakterystyczne dla genów typu *housekeeping* odpowiedzialnych za podstawowe procesy życiowe komórki [306].

W celu powiązania średniego czasu półtrwania cząsteczek mRNA z odpowiedzią komórek na promieniowanie wykorzystano bazę danych czasów półtrwania mRNA sporządzoną przez Tani w 2012r dla komórek HeLa (rak szyjki macicy) [4], którą porównano z oddzielną bazą wykonaną przez Yanga w 2003r dla komórek HepG2 (rak wątroby) [306]. Uzyskane czasy półtrwania są jedynie przybliżeniem gdyż pomimo, iż nie zależą one od tempa produkcji mRNA w danej komórce (eksperymenty wykonano w warunkach zahamowanych procesów ekspresji genów) to są uzależnione od mechanizmów regulacji tempa degradacji mRNA, które różnią się pomiędzy komórkami. Dodatkowo czasy półtrwania dostępne są jedynie dla części z badanych mRNA. W przypadku nowszej bazy Tani czasy półtrwania dopasowano do 10906 transkryptów (46,9%), w przypadku bazy Yanga do jedynie 7768 transkryptów (33,4%).

Czas półtrwania mRNA określony na podstawie bazy danych Tani [4] jest silnie skorelowany z odczytanym z mikromacierzy poziomem ekspresji transkryptów, w przypadku wszystkich badanych linii komórkowych ($Rho: 0,4-0,46$) co dodatkowo potwierdzają wcześniejsze badania [307]. Oznacza to, że transkrypty o długim czasie półtrwania są akumulowane w komórce, co może być powiązane z ich wysokim poziomem sygnału. Wniosków tego typu nie można jednak wysuwać dla pojedynczych transkryptów gdyż jak opisano w rozdziale 6.5.4 poziomy sygnału pomiędzy pojedynczymi zestawami sond dla różnych genów, odczytane z tej samej mikromacierzy nie są porównywalne. Transkrypty o długim czasie półtrwania dodatkowo są znacznie krótsze i zwykle zawierają mniej motywów sekwencyjnych typu MRE i ARE co sprzyja ich zwiększonej stabilności [307].

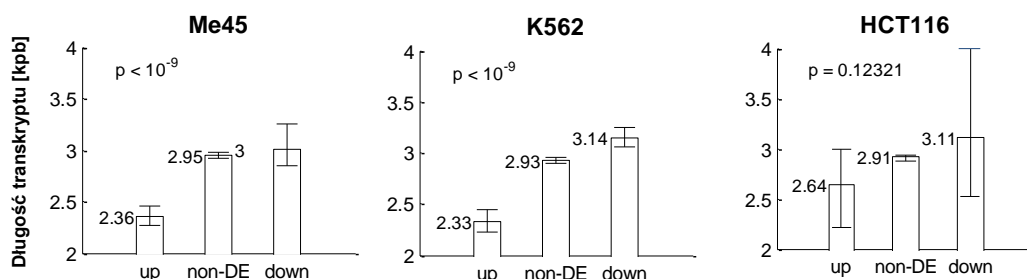


Ryc. 60: Mediana czasu półtrwania mRNA poszczególnych grup transkryptów różnicujących w 12h po napromieniowaniu dla komórek Me45, K562 i HCT116. Wykresy wykonano na podstawie danych z bazy Tani [4]. P-wartości nad wykresami pochodzą z testu-t porównującego średnie czasy półtrwania mRNA pomiędzy grupami up i down, reprezentującymi transkrypty o odpowiednio zwiększonej i zmniejszonej ekspresji.

Ryc. 60 pokazuje różnice w średnim czasie półtrwania transkryptów różnicujących zidentyfikowanych po 12h od napromieniowania dla komórek Me45 i K562 i HCT116. Wykresy te pokazują, że w przypadku komórek K562 bardzo silny wpływ na zmianę poziomu ekspresji może mieć czas półtrwania mRNA, pomimo, że pomiędzy grupami transkryptów o zwiększonej i zmniejszonej ekspresji nie zaobserwowano różnic w ilości miejsc oddziaływania z miRNA (Ryc. 57) oraz motywów ARE klasy III (Ryc. 55).

W przypadku komórek Me45 proporcje wyglądają zupełnie inaczej, co najprawdopodobniej wynika z relaksacji procesów wygaszania ekspresji za pomocą miRNA (co sugeruje Ryc. 57). Proporcje czasu półtrwania pomiędzy grupami transkryptów o zwiększonym i zmniejszonym poziomie ekspresji wyglądają bardzo podobnie dla danych przetworzonych standardowymi algorytmami (RMA, GCRMA) oraz przy wykorzystaniu korekty csGC-RMA bazującej na składzie GC transkryptów. Dodatkowo wyniki są porównywalne z uzyskanymi na podstawie starszej bazy danych Yanga z 2003 (wykresu nie zamieszczono w pracy) pomimo znacznie mniejszej liczby dostępnych czasów półtrwania, których

określono więcej w przypadku transkryptów o stabilniejszej strukturze [306]. Wyniki uzyskane dla komórek HCT116 nie pokazują żadnych zależności, co ponownie może wynikać z niskiej liczby zidentyfikowanych genów różnicujących wynikającej z dużych różnic pomiędzy powtórzeniami bioogicznymi.



Ryc. 61: Mediana długości sekwencji transkryptów poszczególnych grup transkryptów różnicujących w 12h po napromieniowaniu dla komórek Me45, K562 i HCT116. P-wartości nad wykresami pochodzą z testu-t porównującego średnie długości transkryptów pomiędzy grupami up i down, reprezentującymi transkrypty o odpowiednio zwiększonej i zmniejszonej ekspresji.

Pomimo bardzo istotnych różnic pomiędzy listami genów różnicujących komórek Me45 i K562 posiadają one wspólne cechy w postaci porównywalnej średniej długości transkryptu, które widoczne są na Ryc. 61. W przypadku linii komórkowych Me45 i K562 zmiana poziomu ekspresji genów jest negatywnie skorelowana z długością transkryptu, co sugeruje, że transkrypty o krótszej sekwencji (pochodzące od krótszych genów z mniejszą liczbą intronów) są szybciej produkowane niż degradowane po napromieniowaniu komórek. Z kolei w przypadku długich sekwencji, powyżej 3000bp produkcja mRNA jest mniej wydajna i niższa niż tempo degradacji, co widoczne jest w postaci spadku ilości dostępnych cząsteczek mRNA w komórce.

Identyfikator	Nazwa procesu	Liczba genów	Liczba genów różnicujących w komórkach Me45		
			1h	12h	24h
hsa03050	Degradacja białek	39	2	20	0
hsa03015	Kontrola jakości mRNA	69	0	17	0
hsa03018	Degradacja RNA	51	0	15	0
GO:0016070	Metabolizm RNA	191	4	49	16
GO:0034660	Metabolizm niekodujących RNA (ncRNA)	16	1	6	0
GO:0042981	Regulacja procesów apoptozy	154	2	33	19
GO:0006915	Apoptoza	564	7	81	32
GO:0042769	Detekcja uszkodzeń DNA	6	0	4	2
GO:0010467	Ekspresja genów	554	12	114	11
GO:0003723	Interakcje z RNA	427	4	77	19
GO:0003700	Interakcje z DNA (czynniki transkrypcyjne)	682	2	33	11

Tab. 18: Wybrane ścieżki sygnałowe KEGG i procesy GO dla genów różnicujących komórek Me45 (pogrubioną, podkreśloną czcionką zaznaczono nadreprezentacje na poziomie istotności 0.01)

W przypadku komórek M45 wśród transkryptów różnicujących zaobserwowano nadreprezentację transkryptów kodujących białka uczestniczące w procesach kontroli jakości RNA, detekcji uszkodzonych

cząsteczek mRNA oraz genów uczestniczących procesach związanych z degradacją RNA (Tab. 18). Aktywacja znacznej liczby genów uczestniczących w procesie apoptozy oraz genów odpowiedzialnych za degradację RNA i białek może sugerować, iż przeważająca część komórek uległa rozpadowi w procesie apoptozy, jednak niezależne badania pokazują, iż dawka 4Gy indukuje apoptozę u zaledwie 2-5% komórek z linii Me45 po 36-48h od napromieniowania [294, 304]. W przypadku komórek K562 nadreprezentacja genów różnicujących w podobnych procesach nie została zaobserwowana. Pomimo znacznej ilości genów uczestniczących w tego typu procesach identyfikacja nadreprezentacji jest w tym przypadku utrudniona przez bardzo dużą liczbę genów różnicujących, która w przypadku 12h od napromieniowania przekracza 50%. W przypadku komórek HCT116 jest natomiast odwrotnie, bardzo mała liczba genów różnicujących (najwięcej 110 w 12h po napromieniowaniu) uniemożliwia zaobserwowanie znamienych statystycznie nadreprezentacji.

7. Podsumowanie

Podstawowym celem niniejszej pracy było określenie czy następujące pod wpływem promieniowania jonizującego zmiany na poziomie transkryptomu zależą od obecności motywów określonego typu w sekwencjach nukleotydowych transkryptów. Hipoteza badawcza sformułowana została na podstawie obserwacji, które wskazały na istotne różnice w składzie nukleotydowym genów o zmniejszonej i zwiększonej ekspresji na skutek promieniowania w przypadku komórek Me45. Ponieważ uzyskane wyniki zmian w profilach ekspresji genów napromieniowanych komórek pochodzą z eksperymentów mikromacierzowych, konieczne było sprawdzenie czy zależność pomiędzy zmianą poziomu ekspresji genów a składem nukleotydowym transkryptów może być artefaktem samej procedury badawczej.

Czynniki odpowiedzialne za niedokładności eksperymentu mikromacierzowego

Pomimo, że mikromacierze mogą dostarczyć bardzo cennych informacji na temat genów o potencjalnie wysokim znaczeniu dla analizowanych procesów komórkowych to bardzo wiele czynników wpływa na niedokładność tego typu pomiaru obniżając w znaczny sposób jego powtarzalność. W rozdziale 6.5 opisano sześć niezależnych czynników o potencjalnie wysokim wpływie na interpretację danych mikromacierzowych oraz określono, które z nich najsilniej mogą wpływać na dokładność uzyskanych wyników. Sygnały rejestrowane na sondach w zestawach powinny być zbliżone i odpowiadać poziomowi transkryptu, dla którego zestaw był zaprojektowany. W praktyce poziomy sygnału sond należących do pojedynczych zestawów różnią się, podobnie jak sygnały mierzone na tej samej sondy w powtórzeniach eksperymentu. Czynniki, które mają największy wpływ na wariację poziomów sygnału sond są różnice w ich składzie nukleotydowym (zawartość GC) oraz położenie sekwencji komplementarnych do sondy wewnątrz sekwencji transkryptu (degradacja RNA). Czynniki te znacznie ograniczają możliwości porównywania poziomów ekspresji różnych genów w obrębie jednej mikromacierzy, wpływają też na wyniki oznaczeń przy badaniu zmian poziomu ekspresji pojedynczych genów w różnych próbkach. Pokazano także, że stosowana od pewnego czasu praktyka łączenie sond specyficznych dla różnych grup transkryptów pojedynczego genu w zestawy, prowadzi do znacznego wzrostu wariacji sygnału sond w obrębie zestawów.

Obecność w sekwencji sondy motywu (G)₄, opisanego w literaturze jako mającego istotny wpływ na różnice w poziomach sygnału pomiędzy sondami ma najprawdopodobniej znikomy wpływ na wariację sygnału sond w obrębie zestawów i przy powtórzeniach eksperymentu. Niewielkie różnice, jakie są obserwowane pojawiają się najprawdopodobniej wyłącznie ze względu na korelacji pomiędzy częstotliwością występowania tego motywu a składem GC sondy.

Obecność motywu CCGCCTCCC (spacer T7) wewnątrz sekwencji sond ma silny wpływ na ich poziomy sygnału w eksperymentach wykorzystujących starter oligo-dT do syntezy cDNA (Ryc. 40). Mimo, że niewielka ilość tego typu sond na mikromacierzy sprawia, że w skali całego eksperymentu wpływ tego motywu jest niewielki, to sondy tego typu powinny być odrzucone z analizy (np. na etapie budowy pliku CDF) w przypadku platform zakładających wykorzystanie starterów oligo-dT. Motyw (A)_n w sekwencji transkryptu może mieć istotny wpływ na wariację sygnału sond w zestawie dla tego transkryptu jeśli sekwencje sond zestawu są komplementarne do obszaru, w którym leży motyw, co potwierdzają

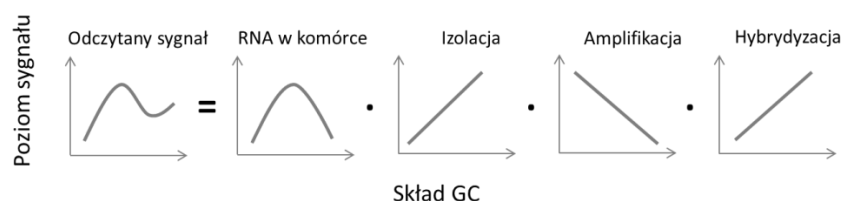
niezależne eksperymenty oparte o technologię RT-qPCR. Mimo to w skali całej mikromacierzy ich wpływ jest także znikomy w związku z niewielką ilością tego typu zestawów sond.

Wpływ składu GC sond na oszacowanie zmian w profilach ekspresji genów

Klasyczna analiza wyników uzyskanych w eksperymentach mikromacierzowych wskazuje na to, że promieniowanie jonizujące wpływa różnie na poziomy transkryptów o różnej zawartości nukleotydów GC. Przeprowadzone w niniejszej pracy badania pokazują jednak, że negatywna korelacja składu GC transkryptów ze zmianą odczytanego z mikromacierzy poziomu ekspresji może częściowo wynikać z różnic technicznych pomiędzy próbkami.

W rozdziale 6.5.4 pokazano jak duże są różnice w poziomach sygnału sond, należących do tych samych zestawów i różniących się składem GC oraz jak różnice tego typu zmieniają się pomiędzy poszczególnymi mikromacierzami. W rozdziale 6.5.6 pokazano, iż efekt ten dotyczy różnych platform mikromacierzowych firmy Affymetrix i nie jest kompensowany przez żaden z powszechnie wykorzystywanych algorytmów standaryzacji danych. Pokazano także, że zaobserwowany efekt może niekorzystnie wpływać na proces identyfikacji transkryptów różnicujących prowadząc do przeszacowania poziomów ekspresji w przypadku transkryptów o skrajnych proporcjach nukleotydów GC. Może też w znaczny sposób obniżać ilość zidentyfikowanych transkryptów różnicujących w związku ze zwiększoną wariancją powtórzeń technicznych lub biologicznych. Wykorzystanie większej ilości powtórzeń może obniżyć wpływ składu GC na odsetek fałszywie dodatnich wyników jednak mimo to duże różnice pomiędzy powtórzeniami obniżają czułość algorytmów identyfikacji genów różnicujących.

Na poziom sygnału sond o różnym składzie nukleotydowym może mieć wpływ wydajność procesu hybrydyzacji, która w różnym stopniu zachodzi dla sond o różnej temperaturze topnienia uzależnionej od proporcji nukleotydów GC w ich sekwencjach. Sondy o wysokim składzie GC mogą w większym stopniu przyłączać niespecyficzne fragmenty mRNA, ponieważ pary GC tworzą silniejsze, potrójne wiązania wodorowe. Z kolei sondy o niskim składzie GC charakteryzują się słabszym wiązaniem, które może być rozrywane w procesie odpłukiwania niespecyficznego związanego mRNA. Ponieważ jak pokazano w rozdziale 6.5.4 skład GC zestawu sond jest odbiciem składu GC transkryptów to czynniki wpływające na różnice w wydajności procesu syntezy cRNA także mogą określać poziom sygnału sond o różnym składzie GC. Różnice w poziomach sygnału sond o różnym składzie GC odzwierciedlają obecność mRNA o określonym składzie w badanej puli mRNA jednak mogą one także wynikać z trzech podstawowych cech, których symboliczny wpływ na kształt wykresu zależności mediany poziomu sygnału dla grup sond o różnych proporcjach GC zilustrowano na Ryc. 62:



Ryc. 62: Symboliczny potencjalny wpływ wybranych czynników na poziomy sygnału sond o różnej zawartości GC. Kształty wykresów są umowne, a przedstawione liniowe zależności są jedynie uproszczeniem nieliniowych procesów jakie zachodzą na różnych etapach eksperymentu.

- Izolacja – mRNA o wyższym składzie GC jest bardziej stabilne co w przypadku wysokiego poziomu degradacji mRNA może być przyczyną różnic w ilości wyizolowanego mRNA o odmiennym składzie GC
- Amplifikacja – cDNA o wysokiej zawartości GC wolniej ulega transkrypcji z powodu mniejszej wydajności polimerazy [256]
- Hybrydyzacja – sondy o wysokiej zawartości GC tworzą silniejsze wiązania z badanym cRNA co dodatkowo może mieć wpływ na ich poziom niespecyficznego hybrydyzacji [255]

Wymienione czynniki mogą wpływać na niezwiązane z poziomami transkryptu różnice w poziomach sygnału sond o różnym składzie GC a ponieważ są one podstawowymi czynnikami wpływającymi na różnice techniczne pomiędzy mikromacierzami to mogą prowadzić do znaczących różnic w poziomach sygnału odczytanych dla transkryptów w rzeczywistości nie różniących się stężeniem w różnych próbkach (szczególnie tych o skrajnych proporcjach nukleotydów GC).

Obserwowana korelacja pomiędzy zmianą poziomu ekspresji a składem GC transkryptu w znacznej mierze może wynikać z następujących cech eksperymentu mikromacierzowego:

- 1) Wysokiej korelacji pomiędzy średnim składem GC sond należących do zestawu oraz składem GC odpowiadających im transkryptów (co pokazano w Tab. 12)
- 2) Różnic w wydajności procesu amplifikacji dla transkryptów o różnym składzie spowodowanych różnicami w wydajności polimerazy [256] oraz różnicami w hybrydyzacji sond o różnych proporcjach GC [255] (co dokładnie opisano w rozdziale 6.5.4)
- 3) Normalizacji danych, która w jednakowy sposób traktuje sondy o różnych własnościach, przez co wzmacnia różnice pomiędzy zestawami o różnym, średnim składzie GC sond, co powoduje, iż geny o zmniejszonej i zwiększonej ekspresji różnią się poziomami sygnału (co opisano w rozdziale 6.5.5)

Tego typu korelacja nie była obserwowana w przypadku danych uzyskanych za pomocą mikromacierzy firmy Agilent, co może wynikać z różnic technicznych pomiędzy obiema platformami. Agilent wykorzystuje dłuższe oligonukleotydy, które, pomimo, że dają więcej możliwości w kwestii doboru proporcji GC to wariancja składu GC sond z całej mikromacierzy jest większa niż w przypadku platform Affymetrix. Ponieważ jednak Agilent wykorzystuje najczęściej 1-2 sondy dla danego transkryptu (w przeciwieństwie do >11 w przypadku platform Affymetrix) korelacja pomiędzy składem GC sond oraz transkryptów jest znacznie niższa (punkt 1). Dodatkowo wykorzystane mikromacierze Agilent są macierzami dwukanałowymi, co oznacza współzawodnictwo w procesie hybrydyzacji do sond mikromacierzowych pomiędzy dwiema próbkami wyznakowanych różnymi barwnikami. W znacznym stopniu zmniejsza to różnice w wydajności procesu hybrydyzacji (punkt 2). Mikromacierze dwukanałowe normalizowane są w inny sposób niż ich jednokanałowe odpowiedniki. Wykorzystują one dodatkowy etap przetwarzania, który standaryzuje pomiary w obrębie pojedynczej mikromacierzy pomiędzy oboma kanałami.

Mikromacierze różniące się od siebie proporcjami sygnału sond o wysokim i niskim składzie GC wynikającymi z różnic technicznych można wykryć prostą metodą opartą o współczynnik nachylenia linii

regresji dopasowanej do danych z każdej indywidualnej mikromacierzy (Ryc. 47), co w istotnym stopniu uzupełnia standardowe metody kontroli jakości. Dodatkowo zastosowanie korekty ze względu na skład GC sond pozwala na uzyskanie lepszych rezultatów podczas procesu identyfikacji genów różnicujących co pokazano na przykładzie dwóch niezależnych testowych zbiorów danych GoldenSpike [226] oraz PlatinumSpike [227].

Korekcja oparta o skład GC transkryptu (zamiast składu GC sondy) przeprowadzona na poziomie pojedynczych sond przed sumaryzacją lub na sygnałach zestawów sond po sumaryzacji pozwala dodatkowo zmniejszyć korelację pomiędzy zmianą poziomu ekspresji a składem GC transkryptu. Przeprowadzanie tego typu korekty może być jednak ryzykowne gdyż w ten sposób sztucznie obniżone zostają różnice w poziomach ekspresji pomiędzy próbkami wszystkich genów o wysokim i niskim składzie GC. Nie można też założyć, że skład GC genów o zwiększonej i zmniejszonej ekspresji powinien być zawsze zbliżony, chociaż podobne założenie stawiają metody normalizacji, np. odnośnie całkowitej liczba genów o zmniejszonej i zwiększonej ekspresji, która powinna być podobna.

Różnice w położeniu sekwencji rozpoznawanej przez sondy mikromacierzowe w transkrypcie, mogą również prowadzić do błędów systematycznych wynikających z degradacji RNA, jaka zachodzi w różnym stopniu pomiędzy próbkami. Jednak analiza przeprowadzona w niniejszej pracy jedynie w niewielki stopniu charakteryzuje ten problem, mimo, że skład GC i stopień degradacji RNA mogą być ze sobą powiązane ponieważ transkrypty o wysokich proporcjach GC mają tendencje do tworzenia struktur drugorzędowych, które zwiększają ich stabilność.

Charakterystyka mechanizmów regulacji ekspresji genów w napromieniowanych komórkach

Analiza odpowiedzi komórkowej na promieniowanie jonizujące dotyczy przede wszystkim komórek czerniaka (Me45), które fizjologicznie bardzo silnie różnią się od pozostałych badanych linii HCT116 oraz K562 i dodatkowo znane są z bardzo dużej zmienności odpowiedzi na promieniowanie. Potwierdzają to wyniki przeprowadzonych badań, w których komórki Me45 w istotny sposób różnią się odpowiedzią na promieniowanie od komórek K562 i HCT116, określoną dla dawki 4Gy, jaka jest porównywalna z dawkami stosowanymi podczas radioterapii.

Żadna z badanych linii komórkowych nie charakteryzowała się zatrzymaniem cyklu komórkowego w fazie G1-S, która jest typowa dla napromieniowanych komórek z prawidłowo działającą ścieżką sygnałową białka p53. Brak odpowiedzi na poziomie szlaku sygnałowego p53 potwierdzają wcześniejsze badania oraz wyniki eksperymentów mikromacierzowych, które nie pokazują żadnych zmian w poziomie ekspresji genów, kluczowych elementów tego szlaku (geny TP53, MDM2, PTEN [298]). Komórki z linii Me45 nie zatrzymują cyklu w odpowiedzi na promieniowanie [294] podczas gdy komórki K562 i HCT116 zatrzymują go dopiero w fazie G2 [295, 300], pomimo, że we wszystkich przypadkach wykryto znaczne uszkodzenia DNA w odpowiedzi na dawkę 4Gy promieniowania. Zmiany cyklu komórkowego w odpowiedzi na promieniowanie jonizujące powiązано z ekspresją genu CCNB1, który koduje cyklinę B2 uczestniczącą w procesie mogącym prowadzić do zatrzymania cyklu w fazie G2 [291]. W przypadku komórek Me45 ekspresja CCNB1 jest obniżona po napromieniowaniu, podczas gdy pozostałe dwie linie wykazują znaczny jego wzrost po 12h od napromieniowania. Nie odnaleziono jednak motywów, które mogłyby sugerować jak działa mechanizm regulacji prowadzący do obniżenia poziomu transkryptu genu CCNB1 w komórkach Me45. Jego zmiana może być powiązana z obniżoną ekspresją dwóch czynników transkrypcyjnych RELA i FOXO3, których liczne miejsca przyłączenia odnaleziono w obszarze promotora

genu CCNB1. Spadek poziomu ekspresji tych dwóch genów jest charakterystyczny wyłącznie dla komórek Me45, jednak potwierdzenie ich roli w regulacji genu CCNB1 po poddaniu komórek wpływowi promieniowania wymaga dodatkowych badań doświadczalnych, w związku z tym, że predykcja TFBS charakteryzuje się dużą liczbą fałszywie dodatnich wyników a sama regulacja uzależniona jest zwykle od kombinacji wielu czynników transkrypcyjnych.

W 12h po napromieniowaniu obserwowany jest silny spadek poziomu ekspresji genu PARP1, którego jednym z zadań, oprócz udziału w procesach naprawy DNA, jest utrzymanie stałego poziomu wolnych rodników w komórce. Jego spadek obserwowany był we wszystkich liniach komórkowych i może być powiązany z wtórnym wzrostem wolnych rodników obserwowanym po wielu godzinach od napromieniowania [301, 304].

Na poziomie mechanizmów regulatorowych komórki Me45 wyróżnia zmiana poziomu ekspresji genu AUF1 (spadek poziomu transkryptu po 12h od napromieniowania), którego białko odpowiedzialne jest za destabilizację transkryptów z motywami typu ARE [270, 271]. Dodatkowo komórki Me45 jako jedyne z badanych charakteryzują się różnicami w częstotliwości występowania motywów ARE wśród transkryptów różnicujących w 12h po napromieniowaniu. W ramach pracy podobną analizę wykonano także dla transkryptów białek, które według [308] mogą obniżać stabilność transkryptów z motywami GRE (GU-rich elements) takich jak MBNL1 czy CUGBP1 (CELF1). Nie stwierdzono jednak nadreprezentacji tego typu motywów w sekwencjach genów bogatych w GC o zmniejszonej ekspresji na skutek promieniowania a badania mikromacierzowe poziomów ekspresji genów MBNL1 oraz CUGBP1 na poziomie transkryptu nie wykazały zmian po napromieniowaniu w żadnych z badanych komórek (dokładnych wyników analizy częstotliwości występowania motywów GRE oraz zmian profilu ekspresji genów MBNL1 oraz CUGBP1 nie zamieszczono w pracy).

Czynnikiem, który może być powiązany z różnicami w składzie nukleotydowym genów o zmniejszonej i zwiększonej ekspresji w przypadku komórek Me45 jest obecność motywów sekwencyjnych odpowiedzialnych za interakcje z miRNA (MRE). Częstotliwość występowania motywów MRE specyficznych dla zidentyfikowanej podgrupy miRNA nie jest skorelowana ze składem GC i pomimo różnic w składzie nukleotydowym pomiędzy transkryptami zmieniającymi się na skutek promieniowania ich obecność nie może być traktowana jako konsekwencja różnic w składzie GC. Dodatkowo w komórkach Me45 obserwowany jest spadek poziomu ekspresji sporej grupy miRNA, który najprawdopodobniej wynika z obniżonej wydajności mechanizmów biogenezy miRNA w stosunku do tempa ich degradacji.

W przypadku dwóch niezależnych eksperymentów mikromacierzowych zaobserwowano spadek poziomu ekspresji genu DICER1 w komórkach Me45 i K562, który jest kluczowym elementem procesu biogenezy miRNA a jego obniżony poziom jest bardzo ściśle powiązany ze spadkiem ilości miRNA w komórce [272, 274]. Badania oparte o technikę RT-qPCR potwierdziły jednak spadek poziomów transkryptu wyłącznie w komórkach K562. Poziom mRNA genu DICER1 jest jednak bardzo niski w komórce, co znacznie utrudnia pokazanie znamienych statystycznie różnic w jego ekspresji.

Odpowiedź na promieniowanie jest bardzo silnie powiązana z mechanizmami regulacji opartymi o miRNA, które uważane są za kluczowe dla uruchomienia wielu procesów biologicznych niezbędnych dla prawidłowego funkcjonowania komórek w warunkach stresu indukowanego promieniowaniem [274].

W przypadku komórek K562 pomimo spadku poziomu miRNA nie zaobserwowano negatywnej korelacji pomiędzy zmianą ekspresji a występowaniem motywów MRE, jaka jest charakterystyczna dla komórek Me45. W przypadku komórek K562 zmiany na poziomie transkryptu są jednak dużo

silniejsze a wysoka korelacja pomiędzy spadkiem poziomu ekspresji a czasem półtrwania mRNA sugeruje, że w tym przypadku zmiany wywołane są przez inny czynnik destabilizacyjny, lub powiązane są one z obniżoną wydajnością produkcji mRNA, która jest kluczowa dla obserwacji relaksacji procesów opartych o degradację mRNA za pośrednictwem miRNA, jaka obserwowana jest w komórkach Me45.

Wnioski końcowe

W trakcie analizy wyników z eksperymentów wykorzystujących mikromacierze firmy Affymetrix, w których badano różnice poziomu transkryptów w komórkach kontrolnych i eksponowanych na promieniowanie jonizujące, zaobserwowano nieopisane dotychczas zależności pomiędzy sygnałem sond a ich składem nukleotydowym, prowadzące do zafałszowania wyników w przypadku porównywania mikromacierzy różniących się całkowitym, średnim poziomem fluorescencji. Zaobserwowane zależności w istotny sposób wpływają na proces identyfikacji genów różnicujących, których transkrypty charakteryzują się wysokim lub niskim składem GC.

Przeprowadzone w niniejszej pracy badania pokazują, że negatywna korelacja składu GC transkryptów ze zmianą odczytanego z mikromacierzy poziomu ekspresji, jaka obserwowana jest w przypadku napromieniowanych komórek Me45, może wynikać z różnic technicznych w procesie znakowania materiału i przeprowadzania eksperymentu mikromacierzowego, które nie są niwelowane, a w niektórych przypadkach są wynikiem użycia powszechnie stosowanych algorytmów normalizacji danych. W związku z tym zaproponowano nową metodę normalizacji csGC-RMA, która minimalizuje wpływ obciążenia danych jaki wynika z różnic w składzie nukleotydowym sond.

Opracowaną metodę przetwarzania danych wykorzystano do scharakteryzowania odpowiedzi komórek Me45 na promieniowanie jonizujące. Wykonana analiza wskazuje na istotne znaczenie procesu interferencji RNA w regulacji poziomów ekspresji genów w napromieniowanych komórkach. Zaobserwowany wzrost poziomu ekspresji genów bogatych w miejsca wiążące miRNA sugeruje relaksację procesów związanych z degradacją transkryptów w procesie interferencji RNA, co potwierdza zaobserwowany spadek poziomu miRNA w komórkach eksponowanych na promieniowanie.

Mikromacierze nie są jedyną techniką pomiarową, której rezultaty mogą być uzależnione od składu GC badanego materiału biologicznego. Podobny problem jak ten opisany dla mikromacierzy zachodzi także w eksperymentach opartych o sekwencjonowanie RNA (RNA-Seq) [264, 309] oraz w szczególności w eksperymentach RT-qPCR [310] a także wszędzie tam, gdzie wykorzystywane są techniki amplifikacji materiału biologicznego lub zjawisko hybrydyzacji, które z różną wydajnością zachodzą w zależności od składu GC badanych sekwencji [162, 256]. Sekwencje nukleotydowe o różnym składzie w istotny sposób różnią się mechanizmami regulacji poziomu ekspresji genów co nieraz może prowadzić do błędnych wniosków na temat roli mechanizmów regulacyjnych w odpowiedzi na badane czynniki. Świadomość tego zjawiska na etapie projektowania eksperymentu i w szczególności analizy danych może przyczynić się do lepszej interpretacji uzyskanych wyników.

Niniejsza praca ma w większości charakter obliczeniowy prezentując różne narzędzia i techniki analizy danych oraz określając, na co należy zwrócić uwagę podczas analizy i interpretacji danych z wielkoskalowych eksperymentów mikromacierzowych. Dodatkowo programy stworzone w ramach pracy mogą być wykorzystywane do testowania innych hipotez niż przedstawione a dzięki temu, że są publicznie dostępne i wyposażone w graficzny interfejs, mogą być wykorzystywane przez różne środowiska badawcze.

8. Literatura

1. Jasinska, A.J., P. Kozłowski oraz W.J. Krzyżosiak, *Expression characteristics of triplet repeat-containing RNAs and triplet repeat-interacting proteins in human tissues*. Acta Biochim Pol, 2008. **55**(1): p. 1-8.
2. Paillard, L. oraz H.B. Osborne, *East of EDEN was a poly(A) tail*. Biol Cell, 2003. **95**(3-4): p. 211-9.
3. Jacob, F. oraz J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins*. J Mol Biol, 1961. **3**: p. 318-56.
4. Tani, H., R. Mizutani, K.A. Salam, K. Tano, K. Ijiri, A. Wakamatsu, T. Isogai, Y. Suzuki oraz N. Akimitsu, *Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals*. Genome Res, 2012. **22**(5): p. 947-56.
5. de Sousa Abreu, R., L.O. Penalva, E.M. Marcotte oraz C. Vogel, *Global signatures of protein and mRNA expression levels*. Mol Biosyst, 2009. **5**(12): p. 1512-26.
6. Ross, J., *mRNA stability in mammalian cells*. Microbiol Rev, 1995. **59**(3): p. 423-50.
7. Rajagopalan, L.E. oraz J.S. Malter, *Regulation of eukaryotic messenger RNA turnover*. Prog Nucleic Acid Res Mol Biol, 1997. **56**: p. 257-86.
8. Guhaniyogi, J. oraz G. Brewer, *Regulation of mRNA stability in mammalian cells*. Gene, 2001. **265**(1-2): p. 11-23.
9. Aghib, D.F., J.M. Bishop, S. Ottolenghi, A. Guerrasio, A. Serra oraz G. Saglio, *A 3' truncation of MYC caused by chromosomal translocation in a human T-cell leukemia increases mRNA stability*. Oncogene, 1990. **5**(5): p. 707-11.
10. Algate, P.A. oraz J.A. McCubrey, *Autocrine transformation of hemopoietic cells resulting from cytokine message stabilization after intracisternal A particle transposition*. Oncogene, 1993. **8**(5): p. 1221-32.
11. Bernasconi, N.L., T.A. Wormhoudt oraz I.A. Laird-Offringa, *Post-transcriptional deregulation of myc genes in lung cancer cell lines*. Am J Respir Cell Mol Biol, 2000. **23**(4): p. 560-5.
12. Raymond, V., J.A. Atwater oraz I.M. Verma, *Removal of an mRNA destabilizing element correlates with the increased oncogenicity of proto-oncogene fos*. Oncogene Res, 1989. **5**(1): p. 1-12.
13. Latchman, D.S., *Transcription factors: an overview*. Int J Biochem Cell Biol, 1997. **29**(12): p. 1305-12.
14. Karin, M., *Too many transcription factors: positive and negative interactions*. New Biol, 1990. **2**(2): p. 126-31.
15. Roeder, R.G., *The role of general initiation factors in transcription by RNA polymerase II*. Trends Biochem Sci, 1996. **21**(9): p. 327-35.
16. Nikolov, D.B. oraz S.K. Burley, *RNA polymerase II transcription initiation: a structural view*. Proc Natl Acad Sci U S A, 1997. **94**(1): p. 15-22.
17. Lee, T.I. oraz R.A. Young, *Transcription of eukaryotic protein-coding genes*. Annu Rev Genet, 2000. **34**: p. 77-137.
18. Levine, M. oraz R. Tjian, *Transcription regulation and animal diversity*. Nature, 2003. **424**(6945): p. 147-51.
19. Barthel, K.K. oraz X. Liu, *A transcriptional enhancer from the coding region of ADAMTS5*. PLoS One, 2008. **3**(5): p. e2184.
20. Gill, G., *Regulation of the initiation of eukaryotic transcription*. Essays Biochem, 2001. **37**: p. 33-43.
21. Narlikar, G.J., H.Y. Fan oraz R.E. Kingston, *Cooperation between complexes that regulate chromatin structure and transcription*. Cell, 2002. **108**(4): p. 475-87.
22. Xu, L., C.K. Glass oraz M.G. Rosenfeld, *Coactivator and corepressor complexes in nuclear receptor function*. Curr Opin Genet Dev, 1999. **9**(2): p. 140-7.
23. Wong, J.M. oraz E. Bateman, *TBP-DNA interactions in the minor groove discriminate between A:T and T:A base pairs*. Nucleic Acids Res, 1994. **22**(10): p. 1890-6.
24. Mukumoto, F., S. Hirose, H. Imaseki oraz K. Yamazaki, *DNA sequence requirement of a TATA element-binding protein from Arabidopsis for transcription in vitro*. Plant Mol Biol, 1993. **23**(5): p. 995-1003.
25. van Nimwegen, E., *Scaling laws in the functional content of genomes*. Trends Genet, 2003. **19**(9): p. 479-84.
26. Babu, M.M., N.M. Luscombe, L. Aravind, M. Gerstein oraz S.A. Teichmann, *Structure and evolution of transcriptional regulatory networks*. Curr Opin Struct Biol, 2004. **14**(3): p. 283-91.

27. Brivanlou, A.H. oraz J.E. Darnell, Jr., *Signal transduction and the control of gene expression*. Science, 2002. **295**(5556): p. 813-8.
28. Wilusz, C.J., M. Wormington oraz S.W. Peltz, *The cap-to-tail guide to mRNA turnover*. Nat Rev Mol Cell Biol, 2001. **2**(4): p. 237-46.
29. Garneau, N.L., J. Wilusz oraz C.J. Wilusz, *The highways and byways of mRNA decay*. Nat Rev Mol Cell Biol, 2007. **8**(2): p. 113-26.
30. Grosset, C., C.Y. Chen, N. Xu, N. Sonenberg, H. Jacquemin-Sablon oraz A.B. Shyu, *A mechanism for translationally coupled mRNA turnover: interaction between the poly(A) tail and a c-fos RNA coding determinant via a protein complex*. Cell, 2000. **103**(1): p. 29-40.
31. Shim, J. oraz M. Karin, *The control of mRNA stability in response to extracellular stimuli*. Mol Cells, 2002. **14**(3): p. 323-31.
32. Derrigo, M., A. Cestelli, G. Savettieri oraz I. Di Liegro, *RNA-protein interactions in the control of stability and localization of messenger RNA (review)*. Int J Mol Med, 2000. **5**(2): p. 111-23.
33. Trzaska, D. oraz J. Dastyk, *[Role of AURE sequences in the regulation of mRNA stability]*. Postepy Biochem, 2005. **51**(1): p. 28-35.
34. Vasudevan, S. oraz J.A. Steitz, *AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2*. Cell, 2007. **128**(6): p. 1105-18.
35. Chen, C.Y. oraz A.B. Shyu, *Selective degradation of early-response-gene mRNAs: functional analyses of sequence features of the AU-rich elements*. Mol Cell Biol, 1994. **14**(12): p. 8471-82.
36. Shyu, A.B., J.G. Belasco oraz M.E. Greenberg, *Two distinct destabilizing elements in the c-fos message trigger deadenylation as a first step in rapid mRNA decay*. Genes Dev, 1991. **5**(2): p. 221-31.
37. Meijlink, F., T. Curran, A.D. Miller oraz I.M. Verma, *Removal of a 67-base-pair sequence in the noncoding region of protooncogene fos converts it to a transforming gene*. Proc Natl Acad Sci U S A, 1985. **82**(15): p. 4987-91.
38. Xu, N., C.Y. Chen oraz A.B. Shyu, *Modulation of the fate of cytoplasmic mRNA by AU-rich elements: key sequence features controlling mRNA deadenylation and decay*. Mol Cell Biol, 1997. **17**(8): p. 4611-21.
39. Bakheet, T., M. Frevel, B.R. Williams, W. Greer oraz K.S. Khabar, *ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins*. Nucleic Acids Res, 2001. **29**(1): p. 246-54.
40. Morris, B.J., D.J. Adams, D.J. Beveridge, L. van der Weyden, H. Mangs oraz P.J. Leedman, *cAMP controls human renin mRNA stability via specific RNA-binding proteins*. Acta Physiol Scand, 2004. **181**(4): p. 369-73.
41. Ladd, A.N., N. Charlet oraz T.A. Cooper, *The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing*. Mol Cell Biol, 2001. **21**(4): p. 1285-96.
42. Miller, J.W., C.R. Urbinati, P. Teng-Umuay, M.G. Stenberg, B.J. Byrne, C.A. Thornton oraz M.S. Swanson, *Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy*. EMBO J, 2000. **19**(17): p. 4439-48.
43. Trabucchi, M., P. Briata, M. Garcia-Mayoral, A.D. Haase, W. Filipowicz, A. Ramos, R. Gherzi oraz M.G. Rosenfeld, *The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs*. Nature, 2009. **459**(7249): p. 1010-4.
44. Bushati, N. oraz S.M. Cohen, *microRNA functions*. Annu Rev Cell Dev Biol, 2007. **23**: p. 175-205.
45. Eulalio, A., E. Huntzinger oraz E. Izaurralde, *Getting to the root of miRNA-mediated gene silencing*. Cell, 2008. **132**(1): p. 9-14.
46. Place, R.F., L.C. Li, D. Pookot, E.J. Noonan oraz R. Dahiya, *MicroRNA-373 induces expression of genes with complementary promoter sequences*. Proc Natl Acad Sci U S A, 2008. **105**(5): p. 1608-13.
47. Lynam-Lennon, N., S.G. Maher oraz J.V. Reynolds, *The roles of microRNA in cancer and apoptosis*. Biol Rev Camb Philos Soc, 2009. **84**(1): p. 55-71.
48. Lewis, B.P., C.B. Burge oraz D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets*. Cell, 2005. **120**(1): p. 15-20.
49. Engels, B.M. oraz G. Hutvagner, *Principles and effects of microRNA-mediated post-transcriptional gene regulation*. Oncogene, 2006. **25**(46): p. 6163-9.
50. Liu, J., M.A. Valencia-Sanchez, G.J. Hannon oraz R. Parker, *MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies*. Nat Cell Biol, 2005. **7**(7): p. 719-23.

51. Pillai, R.S., S.N. Bhattacharyya, C.G. Artus, T. Zoller, N. Cougot, E. Basyuk, E. Bertrand oraz W. Filipowicz, *Inhibition of translational initiation by Let-7 MicroRNA in human cells*. Science, 2005. **309**(5740): p. 1573-6.
52. Petersen, C.P., M.E. Bordeleau, J. Pelletier oraz P.A. Sharp, *Short RNAs repress translation after initiation in mammalian cells*. Mol Cell, 2006. **21**(4): p. 533-42.
53. Wang, X., *Systematic identification of microRNA functions by combining target prediction and expression profiling*. Nucleic Acids Res, 2006. **34**(5): p. 1646-52.
54. Guo, H., N.T. Ingolia, J.S. Weissman oraz D.P. Bartel, *Mammalian microRNAs predominantly act to decrease target mRNA levels*. Nature, 2010. **466**(7308): p. 835-40.
55. Berkhout, B. oraz K.T. Jeang, *RISCy business: MicroRNAs, pathogenesis, and viruses*. J Biol Chem, 2007. **282**(37): p. 26641-5.
56. Elbashir, S.M., J. Harborth, W. Lendeckel, A. Yalcin, K. Weber oraz T. Tuschl, *Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells*. Nature, 2001. **411**(6836): p. 494-8.
57. Croce, C.M., *Causes and consequences of microRNA dysregulation in cancer*. Nature Reviews Genetics, 2009. **10**(10): p. 704-14.
58. Ferracin, M., P. Querzoli, G.A. Calin oraz M. Negrini, *MicroRNAs: Toward the Clinic for Breast Cancer Patients*. Seminars in Oncology, 2011. **38**(6): p. 764-75.
59. Audic, Y. oraz R.S. Hartley, *Post-transcriptional regulation in cancer*. Biol Cell, 2004. **96**(7): p. 479-98.
60. Pesole, G., S. Liuni, G. Grillo, F. Licciulli, F. Mignone, C. Gissi oraz C. Saccone, *UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002*. Nucleic Acids Res, 2002. **30**(1): p. 335-40.
61. Reamon-Buettner, S.M., S.H. Cho oraz J. Borlak, *Mutations in the 3'-untranslated region of GATA4 as molecular hotspots for congenital heart disease (CHD)*. BMC Med Genet, 2007. **8**: p. 38.
62. Simon, D., B. Laloo, M. Barillot, T. Barnetche, C. Blanchard, C. Rooryck, M. Marche, I. Burgelin, I. Coupry, N. Chassaing, et al., *A mutation in the 3'-UTR of the HDAC6 gene abolishing the post-transcriptional regulation mediated by hsa-miR-433 is linked to a new form of dominant X-linked chondrodysplasia*. Hum Mol Genet, 2010. **19**(10): p. 2015-27.
63. Lambert, J.C., E. Luedeking-Zimmer, S. Merrot, A. Hayes, U. Thaker, P. Desai, A. Houzet, X. Hermant, D. Cottel, A. Pritchard, et al., *Association of 3'-UTR polymorphisms of the oxidised LDL receptor 1 (OLR1) gene with Alzheimer's disease*. J Med Genet, 2003. **40**(6): p. 424-30.
64. Bolognani, F. oraz N.I. Perrone-Bizzozero, *RNA-protein interactions and control of mRNA stability in neurons*. J Neurosci Res, 2008. **86**(3): p. 481-9.
65. Yuan, Z., J. Shin, A. Wilson, S. Goel, Y.H. Ling, N. Ahmed, H. Dopeso, M. Jhawer, S. Nasser, C. Montagna, et al., *An A13 repeat within the 3'-untranslated region of epidermal growth factor receptor (EGFR) is frequently mutated in microsatellite instability colon cancers and is associated with increased EGFR expression*. Cancer Res, 2009. **69**(19): p. 7811-8.
66. Chatterjee, S. oraz J.K. Pal, *Role of 5'- and 3'-untranslated regions of mRNAs in human diseases*. Biol Cell, 2009. **101**(5): p. 251-62.
67. Robinson, H., Y.G. Gao, B.S. McCrary, S.P. Edmondson, J.W. Shriver oraz A.H. Wang, *The hyperthermophile chromosomal protein Sac7d sharply kinks DNA*. Nature, 1998. **392**(6672): p. 202-5.
68. Davis, N., N. Biddlecom, D. Hecht oraz G.B. Fogel, *On the relationship between GC content and the number of predicted microRNA binding sites by MicroInspector*. Comput Biol Chem, 2008. **32**(3): p. 222-6.
69. Mishra, A.K., S. Agarwal, C.K. Jain oraz V. Rani, *High GC content: critical parameter for predicting stress regulated miRNAs in Arabidopsis thaliana*. Bioinformatics, 2009. **4**(4): p. 151-4.
70. Bernardi, G., *The human genome: organization and evolutionary history*. Annu Rev Genet, 1995. **29**: p. 445-76.
71. Clay, O., S. Caccio, S. Zoubak, D. Mouchiroud oraz G. Bernardi, *Human coding and noncoding DNA: compositional correlations*. Mol Phylogenet Evol, 1996. **5**(1): p. 2-12.
72. Press, W.H. oraz H. Robins, *Isochores exhibit evidence of genes interacting with the large-scale genomic environment*. Genetics, 2006. **174**(2): p. 1029-40.
73. Bernardi, G., *Isochores and the evolutionary genomics of vertebrates*. Gene, 2000. **241**(1): p. 3-17.
74. Cohen, N., T. Dagan, L. Stone oraz D. Graur, *GC composition of the human genome: in search of isochores*. Mol Biol Evol, 2005. **22**(5): p. 1260-72.

75. Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival oraz F. Rodier, *The mosaic genome of warm-blooded vertebrates*. Science, 1985. **228**(4702): p. 953-8.
76. Mouchiroud, D., G. D'Onofrio, B. Aissani, G. Macaya, C. Gautier oraz G. Bernardi, *The distribution of genes in the human genome*. Gene, 1991. **100**: p. 181-7.
77. Zoubak, S., O. Clay oraz G. Bernardi, *The gene distribution of the human genome*. Gene, 1996. **174**(1): p. 95-102.
78. Saccone, S., A. De Sario, J. Wiegant, A.K. Raap, G. Della Valle oraz G. Bernardi, *Correlations between isochores and chromosomal bands in the human genome*. Proc Natl Acad Sci U S A, 1993. **90**(24): p. 11929-33.
79. Saccone, S., S. Caccio, J. Kusuda, L. Andreozzi oraz G. Bernardi, *Identification of the gene-richest bands in human chromosomes*. Gene, 1996. **174**(1): p. 85-94.
80. Saccone, S., C. Federico, I. Solovei, M.F. Croquette, G. Della Valle oraz G. Bernardi, *Identification of the gene-richest bands in human prometaphase chromosomes*. Chromosome Res, 1999. **7**(5): p. 379-86.
81. Costantini, M. oraz G. Bernardi, *Replication timing, chromosomal bands, and isochores*. Proc Natl Acad Sci U S A, 2008. **105**(9): p. 3433-7.
82. Hiratani, I., A. Leskovaar oraz D.M. Gilbert, *Differentiation-induced replication-timing changes are restricted to AT-rich/long interspersed nuclear element (LINE)-rich isochores*. Proc Natl Acad Sci U S A, 2004. **101**(48): p. 16861-6.
83. Tenzen, T., T. Yamagata, T. Fukagawa, K. Sugaya, A. Ando, H. Inoko, T. Gojobori, A. Fujiyama, K. Okumura oraz T. Ikemura, *Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex*. Mol Cell Biol, 1997. **17**(7): p. 4043-50.
84. Jabbari, K. oraz G. Bernardi, *CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families*. Gene, 1998. **224**(1-2): p. 123-7.
85. Meunier-Rotival, M., P. Soriano, G. Cuny, F. Strauss oraz G. Bernardi, *Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA*. Proc Natl Acad Sci U S A, 1982. **79**(2): p. 355-9.
86. Soriano, P., M. Meunier-Rotival oraz G. Bernardi, *The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes*. Proc Natl Acad Sci U S A, 1983. **80**(7): p. 1816-20.
87. Fullerton, S.M., A. Bernardo Carvalho oraz A.G. Clark, *Local rates of recombination are positively correlated with GC content in the human genome*. Mol Biol Evol, 2001. **18**(6): p. 1139-42.
88. Eisenbarth, I., G. Vogel, W. Krone, W. Vogel oraz G. Assum, *An isochore transition in the NFI gene region coincides with a switch in the extent of linkage disequilibrium*. Am J Hum Genet, 2000. **67**(4): p. 873-80.
89. Costantini, M. oraz G. Bernardi, *The short-sequence designs of isochores from the human genome*. Proc Natl Acad Sci U S A, 2008. **105**(37): p. 13971-6.
90. Duret, L., D. Mouchiroud oraz C. Gautier, *Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores*. J Mol Evol, 1995. **40**(3): p. 308-17.
91. Wada, A. oraz A. Suyama, *Local stability of DNA and RNA secondary structure and its relation to biological functions*. Prog Biophys Mol Biol, 1986. **47**(2): p. 113-57.
92. Galtier, N. oraz J.R. Lobry, *Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes*. J Mol Evol, 1997. **44**(6): p. 632-6.
93. Gouy, M. oraz C. Gautier, *Codon usage in bacteria: correlation with gene expressivity*. Nucleic Acids Res, 1982. **10**(22): p. 7055-74.
94. Sharp, P.M. oraz W.H. Li, *An evolutionary perspective on synonymous codon usage in unicellular organisms*. J Mol Evol, 1986. **24**(1-2): p. 28-38.
95. Sharp, P.M., T.M. Tuohy oraz K.R. Mosurski, *Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes*. Nucleic Acids Res, 1986. **14**(13): p. 5125-43.
96. Sharp, P.M. oraz W.H. Li, *The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications*. Nucleic Acids Res, 1987. **15**(3): p. 1281-95.
97. Bains, W., *Codon distribution in vertebrate genes may be used to predict gene length*. J Mol Biol, 1987. **197**(3): p. 379-88.
98. Eyre-Walker, A., *Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy?* Mol Biol Evol, 1996. **13**(6): p. 864-72.
99. Wan, X.F., J. Zhou oraz D. Xu, *CodonO: a new informatics method for measuring synonymous codon usage bias within and across genomes*. International Journal of General Systems, 2006. **35**(1): p. 109-25.

100. Sueoka, N., *Directional mutation pressure, selective constraints, and genetic equilibria*. J Mol Evol, 1992. **34**(2): p. 95-114.
101. Sueoka, N., *Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C*. J Mol Evol, 1999. **49**(1): p. 49-62.
102. Filipiński, J., *Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells*. FEBS Lett, 1987. **217**(2): p. 184-6.
103. Wolfe, K.H., P.M. Sharp oraz W.H. Li, *Mutation-Rates Differ among Regions of the Mammalian Genome*. Nature, 1989. **337**(6204): p. 283-85.
104. Bernardi, G. oraz G. Bernardi, *Compositional Constraints and Genome Evolution*. Journal of Molecular Evolution, 1986. **24**(1-2): p. 1-11.
105. Holmquist, G.P., *Chromosome Bands, Their Chromatin Flavors, and Their Functional Features*. American Journal of Human Genetics, 1992. **51**(1): p. 17-37.
106. Eyre-Walker, A., *Recombination and mammalian genome evolution*. Proc Biol Sci, 1993. **252**(1335): p. 237-43.
107. Oliver, J.L., P. Carpena, M. Hackenberg oraz P. Bernaola-Galvan, *IsoFinder: computational prediction of isochores in genome sequences*. Nucleic Acids Res, 2004. **32**(Web Server issue): p. W287-92.
108. Gao, F. oraz C.T. Zhang, *GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W686-91.
109. Duret, L., M. Semon, G. Piganeau, D. Mouchiroud oraz N. Galtier, *Vanishing GC-rich isochores in mammalian genomes*. Genetics, 2002. **162**(4): p. 1837-47.
110. Belle, E.M., L. Duret, N. Galtier oraz A. Eyre-Walker, *The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny*. J Mol Evol, 2004. **58**(6): p. 653-60.
111. Falt, S., K. Holmberg, B. Lambert oraz A. Wennborg, *Long-term global gene expression patterns in irradiated human lymphocytes*. Carcinogenesis, 2003. **24**(11): p. 1837-45.
112. Kis, E., T. Szatmari, M. Keszei, R. Farkas, O. Esik, K. Lumniczky, A. Falus oraz G. Safrany, *Microarray analysis of radiation response genes in primary human fibroblasts*. Int J Radiat Oncol Biol Phys, 2006. **66**(5): p. 1506-14.
113. Chapman, J.D., D.L. Dugle, A.P. Reuvers, B.E. Meeker oraz J. Borsa, *Letter: Studies on the radiosensitizing effect of oxygen in Chinese hamster cells*. Int J Radiat Biol Relat Stud Phys Chem Med, 1974. **26**(4): p. 383-9.
114. Jeggo, P. oraz M.F. Lavin, *Cellular radiosensitivity: how much better do we understand it?* Int J Radiat Biol, 2009. **85**(12): p. 1061-81.
115. Lobrich, M. oraz P.A. Jeggo, *The impact of a negligent G2/M checkpoint on genomic instability and cancer induction*. Nat Rev Cancer, 2007. **7**(11): p. 861-9.
116. Jackson, S.P. oraz J. Bartek, *The DNA-damage response in human biology and disease*. Nature, 2009. **461**(7267): p. 1071-8.
117. Kastan, M.B., *DNA damage responses: mechanisms and roles in human disease: 2007 G.H.A. Clowes Memorial Award Lecture*. Mol Cancer Res, 2008. **6**(4): p. 517-24.
118. Negrini, S., V.G. Gorgoulis oraz T.D. Halazonetis, *Genomic instability--an evolving hallmark of cancer*. Nat Rev Mol Cell Biol, 2010. **11**(3): p. 220-8.
119. Ahmed, K.M. oraz J.J. Li, *NF-kappa B-mediated adaptive resistance to ionizing radiation*. Free Radic Biol Med, 2008. **44**(1): p. 1-13.
120. Pappas, G., L.A. Zumstein, A. Munshi, M. Hobbs oraz R.E. Meyn, *Adenoviral-mediated PTEN expression radiosensitizes non-small cell lung cancer cells by suppressing DNA repair capacity*. Cancer Gene Ther, 2007. **14**(6): p. 543-9.
121. Narayan, R.S., C.A. Fedrigo, L.J. Stalpers, B.G. Baumert oraz P. Sminia, *Targeting the Akt-pathway to Improve Radiosensitivity in Glioblastoma*. Curr Pharm Des, 2013. **19**(5): p. 951-7.
122. Koch, C.J., J. Kruuv oraz H.E. Frey, *Variation in radiation response of mammalian cells as a function of oxygen tension*. Radiat Res, 1973. **53**(1): p. 33-42.
123. Rzeszowska-Wolny, J., W.M. Przybyszewski oraz M. Widel, *Ionizing radiation-induced bystander effects, potential targets for modulation of radiotherapy*. Eur J Pharmacol, 2009. **625**(1-3): p. 156-64.
124. Benjamin, D. oraz C. Moroni, *mRNA stability and cancer: an emerging link?* Expert Opin Biol Ther, 2007. **7**(10): p. 1515-29.

125. Kargiotis, O., A. Geka, J.S. Rao oraz A.P. Kyritsis, *Effects of irradiation on tumor cell survival, invasion and angiogenesis*. J Neurooncol, 2010. **100**(3): p. 323-38.
126. Begg, A.C., F.A. Stewart oraz C. Vens, *Strategies to improve radiotherapy with targeted drugs*. Nat Rev Cancer, 2011. **11**(4): p. 239-53.
127. Kanehisa, M., S. Goto, Y. Sato, M. Furumichi oraz M. Tanabe, *KEGG for integration and interpretation of large-scale molecular data sets*. Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
128. Thomas, P.D., A. Kejariwal, M.J. Campbell, H. Mi, K. Diemer, N. Guo, I. Ladunga, B. Ulitsky-Lazareva, A. Muruganujan, S. Rabkin, et al., *PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification*. Nucleic Acids Res, 2003. **31**(1): p. 334-41.
129. Croft, D., G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, et al., *Reactome: a database of reactions, pathways and biological processes*. Nucleic Acids Res, 2011. **39**(Database issue): p. D691-7.
130. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*. Nat Genet, 2000. **25**(1): p. 25-9.
131. Falcon, S. oraz R. Gentleman, *Using GStats to test gene lists for GO term association*. Bioinformatics, 2007. **23**(2): p. 257-8.
132. Al-Shahrour, F., R. Diaz-Uriarte oraz J. Dopazo, *FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes*. Bioinformatics, 2004. **20**(4): p. 578-80.
133. Grossmann, S., S. Bauer, P.N. Robinson oraz M. Vingron, *Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis*. Bioinformatics, 2007. **23**(22): p. 3024-31.
134. He, X., M.S. Sarma, X. Ling, B. Chee, C. Zhai oraz B. Schatz, *Identifying overrepresented concepts in gene lists from literature: a statistical approach based on Poisson mixture model*. BMC Bioinformatics, 2010. **11**: p. 272.
135. Ramsay, G., *DNA chips: state-of-the art*. Nat Biotechnol, 1998. **16**(1): p. 40-4.
136. Stoughton, R.B., *Applications of DNA microarrays in biology*. Annu Rev Biochem, 2005. **74**: p. 53-82.
137. Lockhart, D.J., H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays*. Nat Biotechnol, 1996. **14**(13): p. 1675-80.
138. Bolstad, B.M., R.A. Irizarry, M. Astrand oraz T.P. Speed, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*. Bioinformatics, 2003. **19**(2): p. 185-93.
139. Li, C. oraz W. Hung Wong, *Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application*. Genome Biol, 2001. **2**(8): p. RESEARCH0032.
140. Johnson, W.E., C. Li oraz A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods*. Biostatistics, 2007. **8**(1): p. 118-27.
141. Vardhanabhuti, S., S.J. Blakemore, S.M. Clark, S. Ghosh, R.J. Stephens oraz D. Rajagopalan, *A comparison of statistical tests for detecting differential expression using Affymetrix oligonucleotide microarrays*. OMICS, 2006. **10**(4): p. 555-66.
142. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments*. Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
143. Calza, S., W. Raffelsberger, A. Ploner, J. Sahel, T. Leveillard oraz Y. Pawitan, *Filtering genes to improve sensitivity in oligonucleotide microarray data analysis*. Nucleic Acids Res, 2007. **35**(16): p. e102.
144. Suarez-Farinas, M., M. Pellegrino, K.M. Wittkowski oraz M.O. Magnasco, *Harshlight: a "corrective make-up" program for microarray chips*. BMC Bioinformatics, 2005. **6**: p. 294.
145. Moffitt, R.A., Q. Yin-Goen, T.H. Stokes, R.M. Parry, J.H. Torrance, J.H. Phan, A.N. Young oraz M.D. Wang, *caCORRECT2: Improving the accuracy and reliability of microarray data in the presence of artifacts*. BMC Bioinformatics, 2011. **12**: p. 383.
146. Jaksik, R., J. Polanska, R. Herok oraz J. Rzeszowska-Wolny, *Calculation of reliable transcript levels of annotated genes on the basis of multiple probe-sets in Affymetrix microarrays*. Acta Biochim Pol, 2009. **56**(2): p. 271-7.
147. Schneider, S., T. Smith oraz U. Hansen, *SCOREM: statistical consolidation of redundant expression measures*. Nucleic Acids Res, 2011.
148. Dai, M., P. Wang, A.D. Boyd, G. Kostov, B. Athey, E.G. Jones, W.E. Bunney, R.M. Myers, T.P. Speed, H. Akil, et al., *Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data*. Nucleic Acids Res, 2005. **33**(20): p. e175.

149. Ferrari, F., S. Bortoluzzi, A. Coppe, A. Sirota, M. Safran, M. Shmoish, S. Ferrari, D. Lancet, G.A. Danieli oraz S. Bicciato, *Novel definition files for human GeneChips based on GeneAnnot*. BMC Bioinformatics, 2007. **8**: p. 446.
150. Kroll, K.M., G.T. Barkema oraz E. Carlon, *Modeling background intensity in DNA microarrays*. Phys Rev E Stat Nonlin Soft Matter Phys, 2008. **77**(6 Pt 1): p. 061915.
151. Wu, Z.J., R.A. Irizarry, R. Gentleman, F. Martinez-Murillo oraz F. Spencer, *A model-based background adjustment for oligonucleotide expression arrays*. Journal of the American Statistical Association, 2004. **99**(468): p. 909-17.
152. Parkinson, H., U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, et al., *ArrayExpress update-an archive of microarray and high-throughput sequencing-based functional genomics experiments*. Nucleic Acids Research, 2011. **39**: p. D1002-D04.
153. Barrett, T., D.B. Troup, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, et al., *NCBI GEO: archive for functional genomics data sets-10 years on*. Nucleic Acids Research, 2011. **39**: p. D1005-D10.
154. Pease, A.C., D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes oraz S.P. Fodor, *Light-generated oligonucleotide arrays for rapid DNA sequence analysis*. Proc Natl Acad Sci U S A, 1994. **91**(11): p. 5022-6.
155. Affymetrix, *GeneChip Expression Analysis - Technical Manual*. 2004: p. 185.
156. Wang, Y., Z.H. Miao, Y. Pommier, E.S. Kawasaki oraz A. Player, *Characterization of mismatch and high-signal intensity probes associated with Affymetrix genechips*. Bioinformatics, 2007. **23**(16): p. 2088-95.
157. Affymetrix. *3' IVT Express Kit User Manual*. 2012; Adres: <http://www.affymetrix.com>.
158. Bemmo, A., D. Benovoy, T. Kwan, D.J. Gaffney, R.V. Jensen oraz J. Majewski, *Gene expression and isoform variation analysis using Affymetrix Exon Arrays*. BMC Genomics, 2008. **9**: p. 529.
159. Nam, D.K., S. Lee, G. Zhou, X. Cao, C. Wang, T. Clark, J. Chen, J.D. Rowley oraz S.M. Wang, *Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription*. Proc Natl Acad Sci U S A, 2002. **99**(9): p. 6152-6.
160. Wilson, C.L., S.D. Pepper, Y. Hey oraz C.J. Miller, *Amplification protocols introduce systematic but reproducible errors into gene expression studies*. Biotechniques, 2004. **36**(3): p. 498-506.
161. Kerkhoven, R.M., D. Sie, M. Nieuwland, M. Heimerikx, J. De Ronde, W. Brugman oraz A. Velds, *The T7-primer is a source of experimental bias and introduces variability between microarray platforms*. PLoS One, 2008. **3**(4): p. e1980.
162. Degrelle, S.A., C. Hennequet-Antier, H. Chiapello, K. Piot-Kaminski, F. Piumi, S. Robin, J.P. Renard oraz I. Hue, *Amplification biases: possible differences among deviating gene expressions*. BMC Genomics, 2008. **9**: p. 46.
163. Sudo, H., A. Mizoguchi, J. Kawauchi, H. Akiyama oraz S. Takizawa, *Use of non-amplified RNA samples for microarray analysis of gene expression*. PLoS One, 2012. **7**(2): p. e31397.
164. Sykacek, P., D.P. Kreil, L.A. Meadows, R.P. Auburn, B. Fischer, S. Russell oraz G. Micklem, *The impact of quantitative optimization of hybridization conditions on gene expression analysis*. BMC Bioinformatics, 2011. **12**: p. 73.
165. Affymetrix. *Gene Expression Assay and Data Analysis - Hybridization time*. 2012; Adres: http://www.affymetrix.com/support/help/faqs/ge_assays/faq_15.jsp.
166. Skvortsov, D., D. Abdueva, C. Curtis, B. Schaub oraz S. Tavaré, *Explaining differences in saturation levels for Affymetrix GeneChip arrays*. Nucleic Acids Res, 2007. **35**(12): p. 4154-63.
167. Affymetrix. *Gene Expression Assay and Data Analysis - Microarray scanning*. 2012; Adres: http://www.affymetrix.com/estore/support/help/faqs/ge_assays/faq_8.jsp.
168. Wilson, C.L. oraz C.J. Miller, *Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis*. Bioinformatics, 2005. **21**(18): p. 3683-5.
169. Giorgi, F.M., A.M. Bolger, M. Lohse oraz B. Usadel, *Algorithm-driven artifacts in median Polish summarization of microarray data*. BMC Bioinformatics, 2010. **11**: p. 553.
170. Gautier, L., L. Cope, B.M. Bolstad oraz R.A. Irizarry, *affy-analysis of Affymetrix GeneChip data at the probe level*. Bioinformatics, 2004. **20**(3): p. 307-15.
171. Song, J.S., K. Maghsoudi, W. Li, E. Fox, J. Quackenbush oraz X. Shirley Liu, *Microarray blob-defect removal improves array analysis*. Bioinformatics, 2007. **23**(8): p. 966-71.
172. McCall, M.N., P.N. Murakami, M. Lukk, W. Huber oraz R.A. Irizarry, *Assessing affymetrix GeneChip microarray quality*. BMC Bioinformatics, 2011. **12**: p. 137.

173. Hubbell, E., W.M. Liu oraz R. Mei, *Robust estimators for expression analysis*. Bioinformatics, 2002. **18**(12): p. 1585-92.
174. McCall, M.N., B.M. Bolstad oraz R.A. Irizarry, *Frozen robust multiarray analysis (fRMA)*. Biostatistics, 2010. **11**(2): p. 242-53.
175. Affymetrix. *Technical Note: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation*. 2005; Adres: http://media.affymetrix.com/support/technical/technotes/plier_technote.pdf.
176. Hochreiter, S., D.A. Clevert oraz K. Obermayer, *A new summarization method for Affymetrix probe level data*. Bioinformatics, 2006. **22**(8): p. 943-9.
177. Li, C. oraz W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection*. Proc Natl Acad Sci U S A, 2001. **98**(1): p. 31-6.
178. Binder, H., T. Kirsten, L. Markus oraz P.F. Stadler, *Sensitivity of Microarray Oligonucleotide Probes: Variability and Effect of Base Composition*. Journal of Physical Chemistry, 2004. **108**(46): p. 18003-14.
179. Lu, J., J.C. Lee, M.L. Salit oraz M.C. Cam, *Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays*. BMC Bioinformatics, 2007. **8**: p. 108.
180. Chalifa-Caspi, V., I. Yanai, R. Ophir, N. Rosen, M. Shmoish, H. Benjamin-Rodrig, M. Shklar, T.I. Stein, O. Shmueli, M. Safran, et al., *GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes*. Bioinformatics, 2004. **20**(9): p. 1457-8.
181. Gautier, L., M. Moller, L. Friis-Hansen oraz S. Knudsen, *Alternative mapping of probes to genes for Affymetrix chips*. BMC Bioinformatics, 2004. **5**: p. 111.
182. Mecham, B.H., G.T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D.Z. Wetmore, T.J. Mariani, I.S. Kohane oraz Z. Szallasi, *Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements*. Nucleic Acids Res, 2004. **32**(9): p. e74.
183. Harbig, J., R. Sprinkle oraz S.A. Enkemann, *A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array*. Nucleic Acids Res, 2005. **33**(3): p. e31.
184. Stalteri, M.A. oraz A.P. Harrison, *Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips*. BMC Bioinformatics, 2007. **8**: p. 13.
185. Yu, H., F. Wang, K. Tu, L. Xie, Y.Y. Li oraz Y.X. Li, *Transcript-level annotation of Affymetrix probesets improves the interpretation of gene expression data*. BMC Bioinformatics, 2007. **8**: p. 194.
186. Liao, B.Y. oraz J. Zhang, *Evolutionary conservation of expression profiles between human and mouse orthologous genes*. Mol Biol Evol, 2006. **23**(3): p. 530-40.
187. Jordan, I.K., L. Marino-Ramirez oraz E.V. Koonin, *Evolutionary significance of gene expression divergence*. Gene, 2005. **345**(1): p. 119-26.
188. Li, H., D. Zhu oraz M. Cook, *A statistical framework for consolidating "sibling" probe sets for Affymetrix GeneChip data*. BMC Genomics, 2008. **9**: p. 188.
189. Lu, X. oraz X. Zhang, *The effect of GeneChip gene definitions on the microarray study of cancers*. Bioessays, 2006. **28**(7): p. 739-46.
190. Verhaak, R.G., F.J. Staal, P.J. Valk, B. Lowenberg, M.J. Reinders oraz D. de Ridder, *The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies*. BMC Bioinformatics, 2006. **7**: p. 105.
191. Park, T., S.G. Yi, S.H. Kang, S. Lee, Y.S. Lee oraz R. Simon, *Evaluation of normalization methods for microarray data*. BMC Bioinformatics, 2003. **4**: p. 33.
192. Mutch, D.M., A. Berger, R. Mansourian, A. Rytz oraz M.A. Roberts, *The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data*. BMC Bioinformatics, 2002. **3**: p. 17.
193. Mariani, T.J., V. Budhreja, B.H. Mecham, C.C. Gu, M.A. Watson oraz Y. Sadovsky, *A variable fold change threshold determines significance for expression microarrays*. FASEB J, 2003. **17**(2): p. 321-3.
194. Hahne, F., *Bioconductor case studies*. Use R!2008, New York, NY: Springer. x, 283 p.
195. Cui, X.Q. oraz G.A. Churchill, *Statistical tests for differential expression in cDNA microarray experiments*. Genome Biology, 2003. **4**(4).
196. Jeffery, I.B., D.G. Higgins oraz A.C. Culhane, *Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data*. BMC Bioinformatics, 2006. **7**: p. 359.
197. McCarthy, D.J. oraz G.K. Smyth, *Testing significance relative to a fold-change threshold is a TREAT*. Bioinformatics, 2009. **25**(6): p. 765-71.

198. Storey, J.D. oraz R. Tibshirani, *Statistical significance for genomewide studies*. Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.
199. Tusher, V.G., R. Tibshirani oraz G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response*. Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
200. Margulies, E.H., M. Blanchette, D. Haussler oraz E.D. Green, *Identification and characterization of multi-species conserved sequences*. Genome Res, 2003. **13**(12): p. 2507-18.
201. Burge, C. oraz S. Karlin, *Prediction of complete gene structures in human genomic DNA*. Journal of Molecular Biology, 1997. **268**(1): p. 78-94.
202. Kohany, O., A.J. Gentles, L. Hankus oraz J. Jurka, *Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor*. BMC Bioinformatics, 2006. **7**: p. 474.
203. Sandelin, A., W.W. Wasserman oraz B. Lenhard, *ConSite: web-based prediction of regulatory elements using cross-species comparison*. Nucleic Acids Res, 2004. **32**(Web Server issue): p. W249-52.
204. John, B., A.J. Enright, A. Aravin, T. Tuschl, C. Sander oraz D.S. Marks, *Human MicroRNA targets*. PLoS Biol, 2004. **2**(11): p. e363.
205. Jaksik, R. oraz J. Rzeszowska-Wolny, *Position weight matrix model as a tool for the study of regulatory elements distribution across the DNA sequence* Archives of Control Sciences, 2010. **20**(4): p. 491-501.
206. Sewer, A., N. Paul, P. Landgraf, A. Aravin, S. Pfeffer, M.J. Brownstein, T. Tuschl, E. van Nimwegen oraz M. Zavolan, *Identification of clustered microRNAs using an ab initio prediction method*. BMC Bioinformatics, 2005. **6**: p. 267.
207. Rice, P., I. Longden oraz A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite*. Trends Genet, 2000. **16**(6): p. 276-7.
208. Day, W.H. oraz F.R. McMorris, *Critical comparison of consensus methods for molecular sequences*. Nucleic Acids Res, 1992. **20**(5): p. 1093-9.
209. Claverie, J.M. oraz S. Audic, *The statistical significance of nucleotide position-weight matrix matches*. Comput Appl Biosci, 1996. **12**(5): p. 431-9.
210. Stormo, G.D., T.D. Schneider oraz L.M. Gold, *Characterization of translational initiation sites in E. coli*. Nucleic Acids Res, 1982. **10**(9): p. 2971-96.
211. Bucher, P., *Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences*. J Mol Biol, 1990. **212**(4): p. 563-78.
212. Barash, Y., G. Elidan, N. Friedman oraz T. Kaplan, *Modeling dependencies in protein-DNA binding sites, in Proceedings of the seventh annual international conference on Research in computational molecular biology 2003*, ACM: Berlin, Germany. p. 28-37.
213. Zhao, X., H. Huang oraz T.P. Speed, *Finding short DNA motifs using permuted Markov models*. J Comput Biol, 2005. **12**(6): p. 894-906.
214. Broos, S., P. Hulpiau, J. Galle, B. Hooghe, F. Van Roy oraz P. De Bleser, *ConTra v2: a tool to identify transcription factor binding sites across species, update 2011*. Nucleic Acids Res, 2011. **39**(Web Server issue): p. W74-8.
215. Tokovenko, B., R. Golda, O. Protas, M. Obolenskaya oraz A. El'skaya, *COTRASIF: conservation-aided transcription-factor-binding site finder*. Nucleic Acids Res, 2009. **37**(7): p. e49.
216. Loots, G.G. oraz I. Ovcharenko, *rVISTA 2.0: evolutionary analysis of transcription factor binding sites*. Nucleic Acids Res, 2004. **32**(Web Server issue): p. W217-21.
217. Altschul, S.F., W. Gish, W. Miller, E.W. Myers oraz D.J. Lipman, *Basic local alignment search tool*. J Mol Biol, 1990. **215**(3): p. 403-10.
218. Kent, W.J., *BLAT - The BLAST-like alignment tool*. Genome Research, 2002. **12**(4): p. 656-64.
219. Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, et al., *Clustal W and Clustal X version 2.0*. Bioinformatics, 2007. **23**(21): p. 2947-8.
220. Benson, D.A., I. Karsch-Mizrachi, K. Clark, D.J. Lipman, J. Ostell oraz E.W. Sayers, *GenBank*. Nucleic Acids Research, 2012. **40**(D1): p. D48-D53.
221. Pruitt, K.D., T. Tatusova, G.R. Brown oraz D.R. Maglott, *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy*. Nucleic Acids Research, 2012. **40**(D1): p. D130-D35.
222. Maglott, D., J. Ostell, K.D. Pruitt oraz T. Tatusova, *Entrez Gene: gene-centered information at NCBI*. Nucleic Acids Research, 2011. **39**: p. D52-D57.

223. Etzold, T., A. Ulyanov oraz P. Argos, *SRS: Information retrieval system for molecular biology data banks*. Computer Methods for Macromolecular Sequence Analysis, 1996. **266**: p. 114-28.
224. Fujita, P.A., B. Rhead, A.S. Zweig, A.S. Hinrichs, D. Karolchik, M.S. Cline, M. Goldman, G.P. Barber, H. Clawson, A. Coelho, et al., *The UCSC Genome Browser database: update 2011*. Nucleic Acids Res, 2011. **39**(Database issue): p. D876-82.
225. Shi, L., L.H. Reid, W.D. Jones, R. Shippy, J.A. Warrington, S.C. Baker, P.J. Collins, F. de Longueville, E.S. Kawasaki, K.Y. Lee, et al., *The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements*. Nat Biotechnol, 2006. **24**(9): p. 1151-61.
226. Choe, S.E., M. Boutros, A.M. Michelson, G.M. Church oraz M.S. Halfon, *Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset*. Genome Biol, 2005. **6**(2): p. R16.
227. Zhu, Q., J.C. Miecznikowski oraz M.S. Halfon, *Preferred analysis methods for Affymetrix GeneChips. II. An expanded, balanced, wholly-defined spike-in dataset*. BMC Bioinformatics, 2010. **11**: p. 285.
228. Jaksik, R. oraz J. Rzeszowska-Wolny, *The distribution of GC nucleotides and regulatory sequence motifs in genes and their adjacent sequences*. Gene, 2012. **492**(2): p. 375-81.
229. Kozomara, A. oraz S. Griffiths-Jones, *miRBase: integrating microRNA annotation and deep-sequencing data*. Nucleic Acids Res, 2011. **39**(Database issue): p. D152-7.
230. Sandelin, A., W. Alkema, P. Engstrom, W.W. Wasserman oraz B. Lenhard, *JASPAR: an open-access database for eukaryotic transcription factor binding profiles*. Nucleic Acids Res, 2004. **32**(Database issue): p. D91-4.
231. Masuda, K., T. Werner, S. Maheshwari, M. Frisch, S. Oh, G. Petrovics, K. May, V. Srikantan, S. Srivastava oraz A. Dobi, *Androgen receptor binding sites identified by a GREF_GATA model*. J Mol Biol, 2005. **353**(4): p. 763-71.
232. Betel, D., A. Koppal, P. Agius, C. Sander oraz C. Leslie, *Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites*. Genome Biol, 2010. **11**(8): p. R90.
233. Friedman, R.C., K.K. Farh, C.B. Burge oraz D.P. Bartel, *Most mammalian mRNAs are conserved targets of microRNAs*. Genome Res, 2009. **19**(1): p. 92-105.
234. Krek, A., D. Grun, M.N. Poy, R. Wolf, L. Rosenberg, E.J. Epstein, P. MacMenamin, I. da Piedade, K.C. Gunsalus, M. Stoffel, et al., *Combinatorial microRNA target predictions*. Nat Genet, 2005. **37**(5): p. 495-500.
235. Wu, Z., A.R. Irizarry, R. Gentleman, F. Martinez-Murillo oraz F. Spencer, *A Model-Based Background Adjustment for Oligonucleotide Expression Arrays*. Journal of the American Statistical Association, 2004. **99**(468): p. 909-17.
236. Walker, W.L., I.H. Liao, D.L. Gilbert, B. Wong, K.S. Pollard, C.E. McCulloch, L. Lit oraz F.R. Sharp, *Empirical Bayes accomodation of batch-effects in microarray data using identical replicate reference samples: application to RNA expression profiling of blood from Duchenne muscular dystrophy patients*. BMC Genomics, 2008. **9**: p. 494.
237. Lopez-Romero, P., *Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library*. BMC Genomics, 2011. **12**.
238. Zahurak, M., G. Parmigiani, W. Yu, R.B. Scharpf, D. Berman, E. Schaeffer, S. Shabbeer oraz L. Cope, *Pre-processing Agilent microarray data*. BMC Bioinformatics, 2007. **8**.
239. Kerr, K.F., *Extended analysis of benchmark datasets for Agilent two-color microarrays*. BMC Bioinformatics, 2007. **8**: p. 371.
240. Wei, H.R., P.F. Kuan, S.L. Tian, C.H. Yang, J. Nie, S. Sengupta, V. Ruotti, G.A. Jonsdottir, S. Keles, J.A. Thomson, et al., *A study of the relationships between oligonucleotide properties and hybridization signal intensities from NimbleGen microarray datasets*. Nucleic Acids Research, 2008. **36**(9): p. 2926-38.
241. Rozen, S. oraz H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers*. Methods Mol Biol, 2000. **132**: p. 365-86.
242. Livak, K.J. oraz T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method*. Methods, 2001. **25**(4): p. 402-08.
243. Gyorffy, B., B. Molnar, H. Lage, Z. Szallasi oraz A.C. Eklund, *Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples*. PLoS One, 2009. **4**(5): p. e5645.
244. Stevens, J.R. oraz R.W. Doerge, *Combining Affymetrix microarray results*. BMC Bioinformatics, 2005. **6**: p. 57.
245. Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
246. Pillai, R.S., *MicroRNA function: multiple mechanisms for a tiny RNA?* RNA, 2005. **11**(12): p. 1753-61.
247. Marmur, J. oraz P. Doty, *Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature*. J Mol Biol, 1962. **5**: p. 109-18.

248. Memon, F.N., G.J. Upton oraz A.P. Harrison, *A Comparative Study of the Impact of G-Stack Probes on Various Affymetrix GeneChips of Mammalia*. J Nucleic Acids, 2010. **2010**.
249. Upton, G.J., W.B. Langdon oraz A.P. Harrison, *G-spots cause incorrect expression measurement in Affymetrix microarrays*. BMC Genomics, 2008. **9**: p. 613.
250. Fasold, M., P.F. Stadler oraz H. Binder, *G-stack modulated probe intensities on expression arrays - sequence corrections and signal calibration*. BMC Bioinformatics, 2010. **11**: p. 207.
251. Shanahan, H.P., F.N. Memon, G.J. Upton oraz A.P. Harrison, *Normalized Affymetrix expression data are biased by G-quadruplex formation*. Nucleic Acids Res, 2012. **40**(8): p. 3307-15.
252. Langdon, W.B., G.J. Upton oraz A.P. Harrison, *Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips*. Brief Bioinform, 2009. **10**(3): p. 259-77.
253. Robinson, T.J., M.A. Dinan, M. Dewhirst, M.A. Garcia-Blanco oraz J.L. Pearson, *SplicerAV: a tool for mining microarray expression data for changes in RNA processing*. BMC Bioinformatics, 2010. **11**: p. 108.
254. Boedigheimer, M.J., R.D. Wolfinger, M.B. Bass, P.R. Bushel, J.W. Chou, M. Cooper, J.C. Corton, J. Fostel, S. Hester, J.S. Lee, et al., *Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories*. BMC Genomics, 2008. **9**: p. 285.
255. Schuster, E.F., E. Blanc, L. Partridge oraz J.M. Thornton, *Estimation and correction of non-specific binding in a large-scale spike-in experiment*. Genome Biol, 2007. **8**(6): p. R126.
256. Arezi, B., W. Xing, J.A. Sorge oraz H.H. Hogrefe, *Amplification efficiency of thermostable DNA polymerases*. Anal Biochem, 2003. **321**(2): p. 226-35.
257. Thellin, O., W. Zorzi, B. Lakaye, B. De Borman, B. Coumans, G. Hennen, T. Grisar, A. Igout oraz E. Heinen, *Housekeeping genes as internal standards: use and limits*. Journal of Biotechnology, 1999. **75**(2-3): p. 291-95.
258. Subramanian, A., P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
259. Xia, X., *The Effect of Probe Length and GC% on Microarray Singal Intensity: Characterizing the Functional Relationship*. International Journal of Systems and Synthetic Biology, 2010. **1**(2): p. 171-83.
260. Kawano, S., K. Hashimoto, T. Miyama, S. Goto oraz M. Kanehisa, *Prediction of glycan structures from gene expression data based on glycosyltransferase reactions*. Bioinformatics, 2005. **21**(21): p. 3976-82.
261. Ren, B., F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, et al., *Genome-wide location and function of DNA binding proteins*. Science, 2000. **290**(5500): p. 2306-9.
262. Palmke, N., D. Santacruz oraz J. Walter, *Comprehensive analysis of DNA-methylation in mammalian tissues using MeDIP-chip*. Methods, 2011. **53**(2): p. 175-84.
263. Royce, T.E., J.S. Rozowsky oraz M.B. Gerstein, *Assessing the need for sequence-based normalization in tiling microarray experiments*. Bioinformatics, 2007. **23**(8): p. 988-97.
264. Benjamini, Y. oraz T.P. Speed, *Summarizing and correcting the GC content bias in high-throughput sequencing*. Nucleic Acids Res, 2012.
265. Naef, F. oraz M.O. Magnasco, *Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays*. Phys Rev E Stat Nonlin Soft Matter Phys, 2003. **68**(1 Pt 1): p. 011906.
266. Wilson, G.M., J.B. Lu, K. Sutphen, Y. Sun, Y. Huynh oraz G. Brewer, *Regulation of A+U-rich element-directed mRNA turnover involving reversible phosphorylation of AUF1*. Journal of Biological Chemistry, 2003. **278**(35): p. 33029-38.
267. Sarkar, B., Q.R. Xi, C. He oraz R.J. Schneider, *Selective degradation of AU-rich mRNAs promoted by the p37 AUF1 protein isoform*. Molecular and Cellular Biology, 2003. **23**(18): p. 6685-93.
268. Wilson, G.M., Y. Sun, J. Sellers, H.P. Lu, N. Penkar, G. Dillard oraz G. Brewer, *Regulation of AUF1 expression via conserved alternatively spliced elements in the 3' untranslated region*. Molecular and Cellular Biology, 1999. **19**(6): p. 4056-64.
269. Lagnado, C.A., C.Y. Brown oraz G.J. Goodall, *AUUUA is not sufficient to promote poly(A) shortening and degradation of an mRNA: the functional sequence within AU-rich elements may be UUAUUUA(U/A)(U/A)*. Mol Cell Biol, 1994. **14**(12): p. 7984-95.
270. Mazan-Mamczarz, K., Y. Kuwano, M. Zhan, E.J. White, J.L. Martindale, A. Lal oraz M. Gorospe, *Identification of a signature motif in target mRNAs of RNA-binding protein AUF1*. Nucleic Acids Research, 2009. **37**(1): p. 204-14.
271. Barker, A., M.R. Epis, C.J. Porter, B.R. Hopkins, M.C.J. Wilce, J.A. Wilce, K.M. Giles oraz P.J. Leedman, *Sequence requirements for RNA binding by HuR and AUF1*. Journal of Biochemistry, 2012. **151**(4): p. 423-37.

272. Abdelmohsen, K., K. Tominaga-Yamanaka, S. Srikantan, J.H. Yoon, M.J. Kang oraz M. Gorospe, *RNA-binding protein AUF1 represses Dicer expression*. *Nucleic Acids Res*, 2012. **40**(22): p. 11531-44.
273. Macrae, I.J., K. Zhou, F. Li, A. Repic, A.N. Brooks, W.Z. Cande, P.D. Adams oraz J.A. Doudna, *Structural basis for double-stranded RNA processing by Dicer*. *Science*, 2006. **311**(5758): p. 195-8.
274. Kraemer, A., N. Anastasov, M. Angermeier, K. Winkler, M.J. Atkinson oraz S. Moertl, *MicroRNA-mediated processes are essential for the cellular radiation response*. *Radiat Res*, 2011. **176**(5): p. 575-86.
275. Chiosea, S., E. Jelezcova, U. Chandran, M. Acquafondata, T. McHale, R.W. Sobol oraz R. Dhir, *Up-regulation of dicer, a component of the MicroRNA machinery, in prostate adenocarcinoma*. *Am J Pathol*, 2006. **169**(5): p. 1812-20.
276. Davoren, P.A., R.E. McNeill, A.J. Lowery, M.J. Kerin oraz N. Miller, *Identification of suitable endogenous control genes for microRNA gene expression analysis in human breast cancer*. *BMC Mol Biol*, 2008. **9**: p. 76.
277. Neilson, J.R., G.X. Zheng, C.B. Burge oraz P.A. Sharp, *Dynamic regulation of miRNA expression in ordered stages of cellular development*. *Genes Dev*, 2007. **21**(5): p. 578-89.
278. Pilotte, J., E.E. Dupont-Versteegden oraz P.W. Vanderklish, *Widespread Regulation of miRNA Biogenesis at the Dicer Step by the Cold-Inducible RNA-Binding Protein, RBM3*. *PLoS One*, 2011. **6**(12).
279. Gantier, M.P., C.E. McCoy, I. Rusinova, D. Saulep, D. Wang, D.K. Xu, A.T. Irving, M.A. Behlke, P.J. Hertzog, F. Mackay, et al., *Analysis of microRNA turnover in mammalian cells following Dicer1 ablation*. *Nucleic Acids Research*, 2011. **39**(13): p. 5692-703.
280. Yang, J.S. oraz E.C. Lai, *Dicer-independent, Ago2-mediated microRNA biogenesis in vertebrates*. *Cell Cycle*, 2010. **9**(22): p. 4455-60.
281. Yang, J.S., T. Maurin oraz E.C. Lai, *Functional parameters of Dicer-independent microRNA biogenesis*. *Rna-a Publication of the Rna Society*, 2012. **18**(5): p. 945-57.
282. Matskevich, A.A. oraz K. Moelling, *Stimuli-dependent cleavage of Dicer during apoptosis*. *Biochem J*, 2008. **412**(3): p. 527-34.
283. Surova, O., N.S. Akbar oraz B. Zhivotovsky, *Knock-Down of Core Proteins Regulating MicroRNA Biogenesis Has No Effect on Sensitivity of Lung Cancer Cells to Ionizing Radiation*. *PLoS One*, 2012. **7**(3).
284. Jafarnejad, S.M., G.S. Ardekani, M. Ghaffari, M. Martinka oraz G. Li, *Sox4-mediated Dicer expression is critical for suppression of melanoma cell invasion*. *Oncogene*, 2012.
285. Tokumaru, S., M. Suzuki, H. Yamada, M. Nagino oraz T. Takahashi, *let-7 regulates Dicer expression and constitutes a negative feedback loop*. *Carcinogenesis*, 2008. **29**(11): p. 2073-7.
286. Simone, N.L., B.P. Soule, D. Ly, A.D. Saleh, J.E. Savage, W. DeGraff, J. Cook, C.C. Harris, D. Gius oraz J.B. Mitchell, *Ionizing Radiation-Induced Oxidative Stress Alters miRNA Expression*. *PLoS One*, 2009. **4**(7).
287. Chaudhry, M.A. oraz R.A. Omaruddin, *Differential regulation of MicroRNA expression in irradiated and bystander cells*. *Molecular Biology*, 2012. **46**(4): p. 569-78.
288. Chaudhry, M.A., H. Sachdeva oraz R.A. Omaruddin, *Radiation-Induced Micro-RNA Modulation in Glioblastoma Cells Differing in DNA-Repair Pathways*. *DNA and Cell Biology*, 2010. **29**(9): p. 553-61.
289. Cha, H.J., S. Shin, H. Yoo, E.M. Lee, S. Bae, K.H. Yang, S.J. Lee, I.C. Park, Y.W. Jin oraz S. An, *Identification of ionizing radiation-responsive microRNAs in the IM9 human B lymphoblastic cell line*. *Int J Oncol*, 2009. **34**(6): p. 1661-8.
290. Cranmer, L.D., *Melanoma's radioresistant reputation challenged*. *Oncology (Williston Park)*, 2010. **24**(7): p. 656.
291. Tamamoto, T., K. Ohnishi, A. Takahashi, X. Wang, H. Yosimura, H. Ohishi, H. Uchida oraz T. Ohnishi, *Correlation between gamma-ray-induced G2 arrest and radioresistance in two human cancer cells*. *Int J Radiat Oncol Biol Phys*, 1999. **44**(4): p. 905-9.
292. Gogineni, V.R., A.K. Nalla, R. Gupta, D.H. Dinh, J.D. Klopfenstein oraz J.S. Rao, *Chk2-mediated G2/M cell cycle arrest maintains radiation resistance in malignant meningioma cells*. *Cancer Lett*, 2011. **313**(1): p. 64-75.
293. Teyssier, F., J.O. Bay, C. Dionet oraz P. Verrelle, *Cell cycle regulation after exposure to ionizing radiation*. *Bulletin Du Cancer*, 1999. **86**(4): p. 345-57.
294. Rzeszowska-Wolny, J., R. Herok, M. Widel oraz R. Hancock, *X-irradiation and bystander effects induce similar changes of transcript profiles in most functional pathways in human melanoma cells*. *DNA Repair (Amst)*, 2009. **8**(6): p. 732-8.
295. Herok, R., M. Konopacka, J. Polanska, A. Swierniak, J. Rogolinski, R. Jaksik, R. Hancock oraz J. Rzeszowska-Wolny, *Bystander effects induced by medium from irradiated cells: similar transcriptome responses in irradiated and bystander K562 cells*. *Int J Radiat Oncol Biol Phys*, 2010. **77**(1): p. 244-52.

296. Kumala, S., P. Niemiec, M. Widel, R. Hancock oraz J. Rzeszowska-Wolny, *Apoptosis and clonogenic survival in three tumour cell lines exposed to gamma rays or chemical genotoxic agents*. Cell Mol Biol Lett, 2003. **8**(3): p. 655-65.
297. Hwang, A. oraz R.J. Muschel, *Radiation and the G2 phase of the cell cycle*. Radiat Res, 1998. **150**(5 Suppl): p. S52-9.
298. Puszynski, K., B. Hat oraz T. Lipniacki, *Oscillations and bistability in the stochastic model of p53 regulation*. J Theor Biol, 2008. **254**(2): p. 452-65.
299. Pietenpol, J.A. oraz Z.A. Stewart, *Cell cycle checkpoint signaling: cell cycle arrest versus apoptosis*. Toxicology, 2002. **181-182**: p. 475-81.
300. Flatmark, K., R.V. Nome, S. Folkvord, A. Bratland, H. Rasmussen, M.S. Ellefsen, O. Fodstad oraz A.H. Ree, *Radiosensitization of colorectal carcinoma cell lines by histone deacetylase inhibition*. Radiat Oncol, 2006. **1**: p. 25.
301. Cieslar-Pobuda, A., Y. Saenko oraz J. Rzeszowska-Wolny, *PARP-1 inhibition induces a late increase in the level of reactive oxygen species in cells after ionizing radiation*. Mutat Res, 2012. **732**(1-2): p. 9-15.
302. Mannuss, A., O. Trapp oraz H. Puchta, *Gene regulation in response to DNA damage*. Biochim Biophys Acta, 2012. **1819**(2): p. 154-65.
303. Wood, R.D., M. Mitchell oraz T. Lindahl, *Human DNA repair genes, 2005*. Mutat Res, 2005. **577**(1-2): p. 275-83.
304. Widel, M., W.M. Przybyszewski, A. Cieslar-Pobuda, Y.V. Saenko oraz J. Rzeszowska-Wolny, *Bystander normal human fibroblasts reduce damage response in radiation targeted cancer cells through intercellular ROS level modulation*. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis, 2012. **731**(1-2): p. 117-24.
305. Tani, H. oraz N. Akimitsu, *Genome-wide technology for determining RNA stability in mammalian cells: Historical perspective and recent advantages based on modified nucleotide labeling*. RNA Biol, 2012. **9**(10).
306. Yang, E., E. van Nimwegen, M. Zavolan, N. Rajewsky, M. Schroeder, M. Magnasco oraz J.E. Darnell, Jr., *Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes*. Genome Res, 2003. **13**(8): p. 1863-72.
307. Pai, A.A., C.E. Cain, O. Mizrahi-Man, S. De Leon, N. Lewellen, J.B. Veyrieras, J.F. Degner, D.J. Gaffney, J.K. Pickrell, M. Stephens, et al., *The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels*. PLoS Genet, 2012. **8**(10): p. e1003000.
308. Masuda, A., H.S. Andersen, T.K. Doktor, T. Okamoto, M. Ito, B.S. Andresen oraz K. Ohno, *CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay*. Sci Rep, 2012. **2**: p. 209.
309. Risso, D., K. Schwartz, G. Sherlock oraz S. Dudoit, *GC-content normalization for RNA-Seq data*. BMC Bioinformatics, 2011. **12**: p. 480.
310. Aird, D., M.G. Ross, W.S. Chen, M. Danielsson, T. Fennell, C. Russ, D.B. Jaffe, C. Nusbaum oraz A. Gnirke, *Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries*. Genome Biol, 2011. **12**(2): p. R18.

9. Publikacje autora

Pełne artykuły naukowe:

1. Jaksik R, Polańska J, Herok R, Rzeszowska-Wolny J.: *Calculation of reliable transcript levels of annotated genes on the basis of multiple probe-sets in Affymetrix microarrays*. Acta Biochimica Polonica 2009; 56(2):271-7
2. Herok R, Konopacka M, Polanska J, Swierniak A, Rogolinski J, Jaksik R, Hancock R, Rzeszowska-Wolny J.: *Bystander effects induced by medium from irradiated cells: similar transcriptome responses in irradiated and bystander K562 cells*. Int J Radiat Oncol Biol Phys. 2010 May 1;77(1):244-52.
3. Jaksik R., Rzeszowska-Wolny J. *Position weight matrix model as a tool for the study of regulatory elements distribution across the DNA sequence*. Archives of Control Sciences 2010 20(4):491-501
4. Marczyk M, Jaksik R, Polański A, Polańska J. *Affymetrix Chip Definition Files Construction Based on Custom Probe Set Annotation Database*. Studies in Computational Intelligence 381 , pp. 135-144
5. Foszner P, Gruca A, Polanski A, Marczyk M, Jaksik R, Polanska J. *Efficient algorithm for microarray probes re-annotation*. Lecture Notes in Computer Science, 2011, Volume 6923/2011, pp. 281-289
6. Jaksik R, Rzeszowska-Wolny J.: *The distribution of GC nucleotides and regulatory sequence motifs in genes and their adjacent sequences*. Gene. 2012 Jan 25;492(2):375-81
7. Puszyński K., Jaksik R., Świerniak A., *Regulation of p53 by siRNA in radiation treated cells – simulation studies*. International Journal of Applied Mathematics & Computer Science. 2012; 22(4): 1011–1018
8. Chaabane, W., User, S.D., El-Gazzah, M., Jaksik, R., Sajjadi, E., Rzeszowska-Wolny, J., Łos, M.J. *Autophagy, Apoptosis, Mitoptosis and Necrosis: Interdependence Between Those Pathways and Effects on Cancer*. Arch Immunol Ther Exp. 2012 pp. 1-16
9. Marczyk M, Jaksik R, Polanski A, Polanska J.: *Adaptive filtering of microarray gene expression data based on Gaussian mixture decomposition*. BMC Bioinformatics 2013, 14:101

Doniesienia konferencyjne:

1. Jaksik R., Polańska J., Rzeszowska-Wolny J.: *Analiza niekodujących sekwencji nukleotydowych końca 3' genów przy użyciu nowego, stworzonego do tego celu narzędzia analizy bioinformatycznej*. (praca wyróżniona nagrodą) Konferencja BioMedTech Silesia, Zabrze, 2008
2. Jaksik R., Polańska J., Rzeszowska-Wolny J.: *Estimation of the noise level in microarray experiments with Affymetrix chips*. X International PHD Workshop, Wisła-Kopydło, 2008, pp.85-90
3. Jaksik R., Polańska J., Rzeszowska-Wolny J.: *The 3' non-coding nucleotide sequence influences RNA stability in irradiated cells*. Acta Biochimica Polonica, Supp. 3, Abstracts of the 43rd Meeting of the Polish Biochemical Society and 10th Conference of Polish Cell Biology Society, 7 – 11st September 2008, Olsztyn, Poland
4. Rzeszowska-Wolny J, Herok R, Jaksik R, Polańska J, Wideł M: *Transcription profile change after irradiation or in bystander cells; differences between cell lines*. Gliwice Scientific Meetings 2008, s. 32
5. Jaksik R., Rzeszowska-Wolny J.: *The elements of the systemic cellular response to stress: The regulation of transcript levels in cells exposed to ionizing radiation*. 21-Wilhelm Bernhard Nuclear Workshop, Ustroń, 2009, pp.62
6. Rzeszowska J, Łanuszewska J, Gdowicz A, Skonieczna M, Jaksik R: *Can blood cells be important in bystander effects? Bystander effects in lymphoblastoid cells*. Gliwice Scientific Meetings 2009, s. 49
7. Jaksik R., Habowski Ł., Rzeszowska-Wolny J.: *Regulation of gene expression in cells exposed to ionizing radiation*. XI International PHD Workshop, Wisła-Kopydło, 2009, pp. 164-169
8. Jaksik R., Rzeszowska-Wolny J.: *Nucleotide sequence analysis of transcripts exposed to ionizing radiation*. Acta Biochimica Polonica, Supp. 3, Abstracts of the 44th Meeting of the Polish Biochemical Society, 16-19th September 2009, Łódź, Poland pp.35
9. Jaksik R.: *Data mining in bioinformatics: a journey from raw datasets to biologically meaningful conclusions*. XII International PHD Workshop, Wisła-Kopydło, 2010, pp 265-270
10. Jaksik R, Rzeszowska-Wolny J.: *Prediction of regulatory elements in DNA using position weight matrix models*. Proceedings of the Sixteen National Conference on Applications of Mathematics in Biology and Medicine. Krynica, 14–18 September 2010, pp. 53-58

11. Skonieczna M, Student S, Cieślak-Pobuda A, Herok R, Jaksik R, Rzeszowska-Wolny J: *Oxidative RNA damage and its potential role in cellular responses to radiation*. Acta Biochimica Polonica, Abstracts of the 45th Meeting of the Polish Biochemical Society, 20-23th September 2010, Wisła, Poland, pp. 131
12. Jaksik R, Lalik A, Michalski A, Rzeszowska-Wolny J: *Prediction of glycan structures in cells exposed to ionizing radiation*. Automatyżacja Procesów Dyskretnych, Teoria i Zastosowania, Tom II. Gliwice 2010, pp. 73-80
13. Jaksik R, Swierniak A, Jurka J, Rzeszowska-Wolny J: *Transcripts from genes located in different isochores are differently regulated in cells exposed to ionizing radiation* EJC Supplements 2010 8(5):212
14. Polańska J, Marczyk M, Jaksik R: *A system for low level preprocessing of DNA microarray data*. III Convention of the Polish Bioinformatics Society in conjunction with 8th Workshop on Bioinformatics, Ustroń, October 1–3, 2010 pp. 9
15. Marczyk M, Jaksik R, Polańska J: *Discovery of genetic markers in patients over-responsive to low dose radiotherapy*. III Convention of the Polish Bioinformatics Society in conjunction with 8th Workshop on Bioinformatics, Ustroń, October 1–3, 2010 pp. 15
16. Jaksik R, Rzeszowska-Wolny J.: *Regulation of gene expression by nucleic acid binding factors evolved by gain or loss of binding sites*. Structural and Functional Diversity of the Eukaryotic Genome, Brno, 2010, pp. 53
17. Marczyk M, Jaksik R, Polańska J, Polański A.: *Gene Selection Problem in Identification of Patients Over-Responsive to Low Dose Radiotherapy*. Gliwice Scientific Meetings 2010, pp. 72
18. Rzeszowska-Wolny J, Cieślak-Pobuda A, Wideł M, Jaksik R, Łanuszewska J, Saenko Y, Skonieczna M, Herok R, Student S. *Oxidative stress and bystander effect*. Gliwice Scientific Meetings 2010, pp. 16
19. R. Herok. M. Śnietura, W. Pięglowski, R. Jaksik, A. Fiszer-Kierzkowska, E. Małusecka, G. Woźniak, M. Misiólek, B. Kolebach, A. Maciejewski, C. Szymczyk, R. Suwiński: *Cell Cycle gene expression analysis in Head & Neck cancer suggests the existence of patients subpopulations with different molecular profiles*. V Zjazd Polskiego Towarzystwa Radioterapii Onkologicznej, 17-18 May 2011, Tom 8, nr 2 (39)
20. Lalik A, Jaksik R, Rzeszowska-Wolny J. A bioinformatics tool for the prediction of changes in glycosylation of cells after exposure to various stress conditions. Glycoconjugate Journal, 2011, 28(4), p.336-337
21. Jaksik R, Marczyk M, Polańska J. *MicroImage as a tool for microarray image artifacts correction*. ECMTB Kraków June 28 - July 2 2011, pp. 450-451
22. Marczyk M, Polańska J, Jaksik J, Polański A. *Discriminative gene selection in low dose radiotherapy microarray data for radiosensitivity profile search*. ECMTB Kraków June 28 - July 2 2011, pp. 40-41
23. Foszner P, Jaksik R, Gruca A, Polańska J, Polański A. *Efficient reannotation system for verifying genomic targets of DNA microarray probes*. ECMTB Kraków June 28 - July 2 2011, pp. 26-27
24. Jaksik R, Rzeszowska-Wolny J.: *The journey towards understanding the compositional properties of vertebrate genomes*. Gliwice Scientific Meetings 2011, pp. 63
25. Rzeszowska-Wolny J, Jaksik R, Cieślak-Pobuda A, Skonieczna M, Lalik A, Student S: *Micro-RNA and Oxidative damage*. Gliwice Scientific Meetings 2011, pp.26
26. Jaksik R, Marczyk M, Polańska J, Polański A. *Microarray artifacts elimination algorithm based on image recognition techniques*. Proceedings of the Seventeenth National Conference on Applications of Mathematics in Biology and Medicine. Zakopane-Kościelisko, 1–6 September 2011, pp. 29-33
27. Puszyński K, Jaksik R: *Computational model of siRNA regulation in p53 signaling pathway*. Proceedings of the Seventeenth National Conference on Applications of Mathematics in Biology and Medicine. Zakopane-Kościelisko, 1–6 September 2011, pp. 85-90
28. Puszyński K, Jaksik R: *Model prostego sterowania ścieżką sygnałową p53 przy użyciu siRNA*. XVII Krajowa Konferencja Biocybernetyka i Inżynieria Biomedyczna, Gliwice/Tarnowskie Góry, 11–14 October 2011
29. Jaksik R, Lalik A, Student S, Swierniak A, Rzeszowska-Wolny J: *The role of RNA interference-based regulation of gene expression in cancer cells exposed to ionizing radiation* EJC Supplements 2012, 48(5), pp.160
30. Rzeszowska-Wolny J, Wideł M, Cieślak-Pobuda A, Lalik A, Skonieczna M, Jaksik R: *Reactive Oxygen Species May Play a Role in Ionizing Radiation-induced Bystander Effects and Regulation of mRNA Levels by MicroRNAs* EJC Supplements 2012, 48(5), pp.271
31. Janus, P., Kalinowska-Herok, M., Pięglowski, W., Szoltysek, K., Jaksik, R., Puszyński, K., Kimmel, M., Widlak, P. *Co-regulation of NF-kappaB signaling pathway by the active form of Heat Shock Factor 1*. FEBS JOURNAL, 2012, 279(SI-1), pp. 163
32. Marczyk M, Jaksik R, Polańska J: *Gene filtering in microarray data using adaptive filter type*. 10th Workshop on Bioinformatics and 5th Symposium of the Polish Bioinformatics Society, Gdańsk 25 – 27 May 2012, pp. 69

33. Jaksik R, Lalik A, Rzeszowska-Wolny J: *GlycoGene: a tool for the assessment of alterations in the glycosylation patterns*. 8-th International Symposium on Glycosyltransferases, Hannover, June 5-9 2012
34. Lalik A, Jaksik R, Rzeszowska-Wolny J: *From computational studies to biological relevance: a journey towards understanding the modulation of fucosylation reactions*. International Carbohydrate Symposium, Madrid, July 22-27 2012
35. Jaksik R. *NucleoSeq 2.0 an enchanted web services client applicable to complex nucleotide sequence analysis tasks*. Gliwice Scientific Meetings 2012, pp. 53
36. Puszyński K, Jaksik R, Świerniak A. *Regulation of p53 signaling pathway in malignant cells based on RNA interference*. Gliwice Scientific Meetings 2012, pp. 58
37. Janus P, Kalinowska-Herok M, Katarzyna Szołtysek K, Jaksik R, Kimmel M, Widłak P. *The influence of heat shock factor 1 on the NF- κ B signaling pathway*. Gliwice Scientific Meetings 2012, pp. 77
38. Cichońska A, Jaksik R, Polańska J. *Quality assessment of the probes placed on the Affymetrix Mouse430_2 microarray*. Gliwice Scientific Meetings 2012, pp.49
39. Cichońska A, Jaksik R, Polanska J, Widłak W.: *Comparative analysis of the annotation systems of Mus musculus 3 high density expression microarray*. International Work-Conference on Bioinformatics and Biomedical Engineering 2013, Granada, Hiszpania