Politechnika Warszawska
Wydział Matematyki i Nauk Informacyjnych

PRACA DYPLOMOWA MAGISTERSKA
NA KIERUNKU MATEMATYKA

PORÓWNANIE STATYSTYCZNYCH METOD RÓŻNICOWEJ ANALIZY DANYCH RNA-SEQ NA PRZYKŁADZIE PAKIETÓW *DESEQ* I *EDGER*

Autor:
Barbara Sozańska

Promotor:
dr hab. inż. Przemysław Biecek, prof. nzw.

Warszawa, wrzesień 2016

.....

podpis promotora

.....

podpis autora

Streszczenie

Niniejsza praca stanowi porównanie dwóch pakietów statystycznych – *DESeq* oraz *edgeR* — wykorzystywanych w różnicowej analizie sekwencji RNA. Wysokoprzepustowe sekwencjonowanie RNA (RNA-Seq) to technika biologii molekularnej pozwalająca na badanie pełnego transkryptomu próbek biologicznych. Dostarcza ilościowych informacji dotyczących poziomu ekspresji określonych genów w badanej próbce biologicznej w określonym momencie. Analiza statystyczna danych tego rodzaju może pozwolić na powiązanie danych genów (oraz kodowanych przez nie białek) ze stanem komórek oraz procesami w nich zachodzącymi. Niniejsza praca osadzona jest w kontekście poszukiwania powiązań między ekspresją danego zestawu genów w komórkach wyjściowych i komórkach macierzystych powstałych z komórek wyjściowych w procesie reprogramowania.

Pakiety *DESeq* i *edgeR* stanowią ogólnodostępne narzędzia pozwalające na zaawansowaną analizę różnicową sekwencji RNA w środowisku $\mathcal{R}$. Mają one pozwalać na wykrywanie statystycznie istotnych różnic w ekspresji genów pochodzących z próbek odmiennych biologicznie. Oznaczenie różnic pomiędzy transkryptomami komórek nowotworowych i zdrowych może pozwolić zarówno na odkrycie nowych markerów nowotworowych (diagnostyka), jak również potencjalnych metod selektywnego niszczenia komórek nowotworowych. Dzięki dynamicznemu rozwojowi biochemii w przeciągu ostatnich dwóch dekad, izolowanie RNA z próbek biologicznych i jego sekwencjonowanie może być obecnie traktowane jako stosunkowo rutynowa, szybka i dostępna procedura badawcza. Jednak ze względu na znaczną ilość oraz złożoność surowych danych generowanych w tego rodzaju eksperymentach, ilościowe oznaczanie poziomu ekspresji konkretnych genów oraz wyszukiwanie różnic pomiędzy próbkami nadal pozostaje nietrywialnym problemem. Analizowane w niniejszej pracy pakiety statystyczne stanowią istotny krok w kierunku jego rozwiązania.

W pierwszej części pracy zawarte zostało biologiczne wprowadzenie do omawianych zagadnień. W zwięzłej formie opisane zostały podstawy procesu ekspresji genów oraz sposoby regulacji ekspresji genów przez komórkę. Rozdział ten tłumaczy, w jaki sposób poziom określonych transkryptów RNA w komórce powiązany jest z ekspresją danych genów i produkcją białek przez komórkę. Przybliża również metodę RNA-Seq, która pozwala na precyzyjne i szybkie sekwencjonowanie RNA.

W kolejnym rozdziale pracy przedstawiono podobieństwa oraz różnice w metodach działania pakietów *DESeq* i *edgeR*. Matematyczne uzasadnienie wykorzystywanych w nich funkcji pomaga w dokładniejszym zrozumieniu problemów, które wiążą się z analizą różnicową. Przybliżono zagadnienia normalizacji próbek, nadwyżki dyspersji genów oraz zastosowania uogólnionych modeli liniowych do radzenia sobie z tą nadwyżką. Omawianie kolejnych zagadnień z podziałem na rozwiązania zastosowane w poszczególnych pakietach pozwala na łatwe dostrzeżenie podobieństw i różnic między nimi.

W ostatniej części przedstawiono symulacyjną analizę metod oraz analizę przykładowych danych pochodzących z rzeczywistych próbek biologicznych. Powtórzone 1000 razy obliczenia pozwalają ocenić powtarzalność uzyskiwanych wyników oraz stwierdzić, czy rezultaty otrzymywane za pomocą dwóch badanych pakietów są spójne. Pokazują też, który z pakietów jest bardziej czuły na niewielkie różnice w poziomie ekspresji. Testowanie dwóch próbek o różnej liczebności pomaga w odpowiedzi na pytanie, czy i jak wielkość zbioru wpływa na otrzymywane wyniki. Analizie poddano dane RNA-Seq wygenerowane w MD Anderson Cancer Center, będącym częścią University of Texas (Houston, USA). Kolejne etapy obliczeń pokazały, jak różnice między pakietami prowadzą do rozrzutu w otrzymywanych wynikach. Analiza

danych porównujących ekspresję w komórkach wyjściowych i komórkach macierzystych powstałych z komórek wyjściowych w procesie reprogramowania przedstawia, który z pakietów wykazuje silniejszą tendencję do wskazywania różnicy w ekspresji jako istotnej statystycznie. W Dodatku A opisano wykorzystane funkcje oraz parametry, które można w nich zastosować. W Dodatkach B i C przedstawiono kody, za pomocą których uzyskano zaprezentowane wyniki.

Uzyskane wyniki pokazują, że pakiet *edgeR* wyraźnie częściej wykazuje istotne różnice w ekspresji. Można podejrzewać, że geny wskazywane przez pakiet *DESeq* jako istotne statystycznie w rzeczywistości takie są, gdyż potwierdzają to również wskazania pakietu *edgeR*. Nie można jednak wykluczyć, iż takich genów jest więcej, mimo iż nie są wskazywane przez *DESeq*.

Słowa kluczowe

analiza ekspresji genów, rozkład ujemny dwumianowy, normalizacja danych, estymatory średniej i wariancji genu, uogólnione modele liniowe, testowanie różnic w ekspresji genów, False Discovery Rate, procedura Benjaminiego-Hochberga

Tytuł pracy w języku angielskim

Study of Statictical Methods for Differential Expression Analysis of RNA-Seq Data with Examples Based on *DESeq* and *edgeR* Packages.

Abstract

The following thesis compares two statistical tools used in differential expression analysis of RNA sequences: *DESeq* and *edgeR* packages. High-throughput RNA sequencing (RNA-Seq) is a molecular biology technique allowing for determination of full transcriptome of biological samples. It provides quantitative data on expression levels of given genes in a biological sample at a given moment. Statistical analysis of such data may lead to establishing connections between expression of certain genes (and production of proteins they encode) and the current state of cells as well as processes occurring therein. The context of this particular work is the search for connections between expression of the given set of genes in Primary human derminal fibroblasts and the stem cells created from Primary human derminal fibroblasts in reprogramming.

DESeq and *edgeR* packages are freely available tools for advanced differential expression data analysis developed for the $\mathcal{R}$ environment. They are designed to reveal statistically important differences in gene expression between different biological samples. Finding such differences between the transcriptomes of regular and cancer cells may allow to establish new cancer markers (for diagnostic purposes) as well as potential methods of selective inhibition of cancer cell proliferation. Thanks to the dynamic development of the biochemical methods over the last two decades, isolating and sequencing RNA from biological samples may now be considered as a routine, fast, and available procedure in cancer research. However, due to the vast quantity and complexity of data generated in such experiments, accurate determination of expression levels of given genes as well as establishing differences in expression patterns between samples still remains a nontrivial problem. The statistical packages discussed within this thesis are an important step towards its effective solution.

The first part of the thesis comprises a biological introduction to the discussed topic. The biochemical bases of the gene expression process are briefly presented, as well as the methods of gene expression regulation utilized by cells. The vital connection between the levels of RNA transcripts and expression of given genes and biosynthesis of proteins is clarified. Also, the basic ideas of the RNA-Seq method of rapid transcriptome sequencing are introduced.

The next chapter of the thesis summarizes the similarities and differences between the construction and methods applied in *DESeq* and *edgeR* packages. Formal mathematical description of the functions utilized within the packages allows for deeper understanding of the challenges of differential expression analysis of transcriptome sequence data. The issues of sample normalization and gene dispersion excess are discussed, as well as the application of generalized linear models to cope with this excess. Introduction of consecutive issues with parallel presentation of the way they are dealt with in the two investigated packages allows to pinpoint the crucial similarities and differences between them.

Within the last part of the thesis, the performance of the *DESeq* and *edgeR* packages is directly assessed using both simulations as well as real data from biological samples. The reproducibility of the results is verified within a 1000 repetitions of the data analysis cycle. The consistency between the results obtained using the two packages is assessed. The sensitivity of the two methods to minor differences in the expression profiles is compared. Tests performed on two differently populated samples allow to establish the influence of the size of the dataset on the quality of the obtained results. The analyzed RNA-Seq data were obtained from the MD Anderson Cancer Center, University of Texas (Houston, USA). Subsequent calculation and analysis stages demonstrated the discrepancies appearing in the output produces by the two packages. Comparison of the gene expression data from the original patient

cells and reprogrammed stem cells revealed, which of the packages tended to encounter more statistically significant mismatches.

In Appendix A, all the applied functions and parameters appearing therein are described. Appendices B and C comprise software codes written in order to obtain the results presented in the thesis.

The investigations described within this thesis lead to a conclusion that the *edgeR* package tends to reveal a higher number of statistically important expression differences. Since all the genes marked by the *DESeq* as differing in expression level are also returned by the *edgeR* package, we may expect all these matches to be relevant. However, it cannot be determined with certainty whether it is the *edgeR* that tends to return false positives, or rather the *DESeq* fails to find all the expression pattern mismatches.

Spis treści

Wstęp

Celem niniejszej pracy jest porównanie metod działania pakietów statystycznych *DESeq* i *edgeR*, które zostały napisane z myślą o różnicowej analizie danych RNA-Seq. Ekspresja genów to zagadnienie biologiczne, na które można patrzeć także od strony statystycznej. W niniejszej pracy przedstawione zostaną dwa narzędzia pomagające w ocenie występowania statystycznie istotnych różnic w ekspresji genów pochodzących z próbek odmiennych biologicznie. Pierwszy rozdział stanowić będzie biologiczne wprowadzenie do omawianych zagadnień. Przedstawi procesy prowadzące do zjawiska ekspresji genów oraz sposoby regulacji ekspresji genów przez komórkę. Przybliży również metodę RNA-Seq, która pozwala na precyzyjne sekwencjonowanie RNA. Ta stosunkowo nowa metoda zastąpiła popularną wcześniej metodę sekwencjonowania za pomocą mikromacierzy.

W kolejnym rozdziale przedstawione zostaną podobieństwa oraz różnice w metodach działania pakietów *DESeq* i *edgeR*. Matematyczne uzasadnienie wykorzystywanych w nich funkcji pozwoli na dokładniejsze zrozumienie problemów, które wiążą się z analizą różnicową. Przybliżone zostaną zagadnienia normalizacji próbek, nadwyżki dyspersji genów czy zastosowań uogólnionych modeli liniowych do radzenia sobie z tą nadwyżką. Omawianie kolejnych zagadnień z podziałem na rozwiązania zastosowane w poszczególnych pakietach pozwala na łatwe dostrzeżenie podobieństw i różnic między narzędziami.

W trzecim rozdziale przedstawiona zostanie symulacyjna analiza metod na podstawie danych rzeczywistych. Powtórzone 1000 razy obliczenia pomogą stwierdzić czy wyniki uzyskiwane za pomocą dwóch różnych pakietów są powtarzalne. Pokażą też, który z pakietów wykazuje większe różnice w ekspresji. Testowanie dwóch próbek o różnej liczebności może pomóc w odpowiedzi na pytanie, czy i jak wielkość zbioru wpływa na otrzymywane wyniki. W czwartym rozdziale zaprezentowany zostanie przykład analizy danych rzeczywistych za pomocą omawianych pakietów. Kolejne etapy obliczeń pokażą jak omawiane we wcześniejszym rozdziale różnice między pakietami prowadzą do rozrzutu w otrzymywanych wynikach. Analiza danych porównujących ekspresję w komórkach wyjściowych i komórkach macierzystych powstałych z komórek wyjściowych w procesie reprogramowania przedstawi, który z pakietów wykazuje silniejszą tendencję do wskazywania różnicy w ekspresji jako istotnej statystycznie.

Analiza różnicowa ekspresji genów wykorzystywana jest do poszukiwania powiązań między ilością wystąpienia danego genu a syntezą określonych białek w komórce w określonych warunkach. W szczególności, może pozwolić na określenie związku między występowaniem danego genu a prawdopodobieństwem rozwoju choroby nowotworowej określonego rodzaju. Nowotwory są drugą przyczyną zgonów w Polsce. W 2010 roku pozbawiły życia 93 tysiące osób [23]. Coraz tańsze i szybsze sposoby sekwencjonowania DNA mogą pozwolić na wczesne wykrycie, a w konsekwencji działania prewencyjne, które zapobiegną rozwojowi nowotworów. Przykładem takiego odkrycia, które zaczęło być wykorzystywane w diagnostyce na szeroką skalę jest

to, że kobiety posiadające mutację genu BRCA1 z 80% prawdopodobieństwem zachorują w przyszłości na raka piersi lub jajnika [31].

Rozdział 1

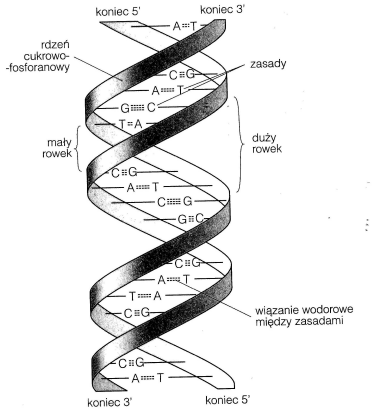
Wprowadzenie biologiczne

W pierwszym rozdziale opisane zostaną biologiczne procesy, które leżą u podstaw zjawiska ekspresji genów wokół którego skupiona jest niniejsza praca. Na funkcjonowanie ludzkiego organizmu składa się szereg powiązanych ze sobą procesów tworzących ciągi przyczynowo-skutkowe. Właśnie jako taki ciąg można przedstawić precyzyjnie kontrolowaną biosyntezę białek, których sekwencje i „instrukcje użycia" zapisane są w kodzie genetycznym – na niciach DNA. Opisane tu zjawiska transkrypcji i translacji to procesy uniwersalne i kluczowe dla funkcjonowania wszystkich znanych nauce organizmów żywych, w tym oczywiście organizmu ludzkiego. Opisane w ostatnim paragrafie narzędzie RNA-Seq służy do sekwencjonowania RNA, co pozwala na uzyskanie informacji o poziomach ekspresji określonych genów. To zaś umożliwia przejście od badania samego kodu genetycznego – identycznego dla wszystkich komórek danego organizmu – do poznawania mechanizmów różnicowania poszczególnych komórek i ich adaptacji do zmieniających się warunków.

1.1. DNA

Kwas deoksyrybonukleinowy (ang. *deoxyribonucleic acid* (DNA)) pełni rolę nośnika informacji genetycznej organizmów żywych. Zdolność ta jest bezpośrednio związana z jego strukturą. DNA jest polimerem zbudowanym z długich łańcuchów jednostek monomerycznych zwanych nukleotydami. Każdy nukleotyd zawiera grupę fosforanową, cukier oraz zasadę azotową (adeninę (A), cytozynę (C), guaninę (G) lub tyminę (T)). Związek cukru z zasadą nazywa się nukleozydem. Informacja genetyczna jest kodowana poprzez sekwencję zasad w polinukleotydach DNA [44]. Sekwencja ta jest zawsze przedstawiana w kierunku $5' \rightarrow 3'$ (gdzie $5'$ i $3'$ oznaczają końce łańcucha DNA, patrz Rysunek 1.1) [47]. Nie ma ograniczeń co do kolejności ułożenia oraz ilości nukleotydów w łańcuchu. Przykładowo, eukariotyczne chromosomy to pojedyncze, liniowe cząsteczki DNA, a chromosom 1 (największy) u człowieka to ok. $2{,}45 \cdot 10^8$ par zasad [16].

DNA składa się z dwóch wzajemnie się oplatających łańcuchów polinukleotydowych, które tworzą dwuniciową helisę (Rysunek 1.1). Na zewnątrz znajduje się część cukrowo-fosforanowa, która stanowi szkielet struktury. Do wnętrza helisy skierowane są pary zasad ułożone jedna obok drugiej. Łańcuchy polinukleotydowe są ułożone antyrównolegle, co oznacza, że każdy z łańcuchów „biegnie" w inną stronę: jeden od końca $3'$ do $5'$, a drugi $5' \rightarrow 3'$ [44]. Dwuniciowa struktura DNA jest regularna, ale można w niej dostrzec duży rowek (głęboka bruzda) i mały rowek (płytka bruzda). Zagłębienia te są istotne dla większości procesów, w których zachodzi



Rysunek 1.1: Dwuniciowa helisa DNA. Źródło: [47, str. 5]

wiązanie innych cząstek do DNA – na przykład podczas interkalacji, replikacji oraz ekspresji informacji genetycznej (proces opisany w Rozdziale 1.2) [45].

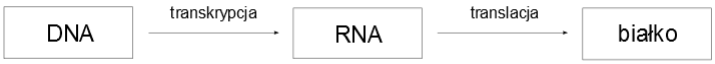
Dwie nici DNA splecione w helisę są połączone poprzez wiązania wodorowe występujące w każdej parze zasad. Tymina zawsze wiąże się z adeniną za pomocą dwóch wiązań, a guanina z cytozyną jest połączona trzema wiązaniami. Dlatego wiązania A-T są słabsze niż wiązania G-C. Sposób wiązania, który został przedstawiony powyżej nazywany jest komplementarnym parowaniem zasad. Ma on podstawowe znaczenie dla struktury i funkcji DNA. Pary A-T oraz G-C są znacznie bardziej korzystne energetycznie (więc bardziej stabilne) od innych konfiguracji [15]. Dzięki komplementarności nici DNA sekwencja jednej nici jednoznacznie określa sekwencję drugiej nici. Oznacza to, że za pomocą jednej nici można replikować strukturę drugiej. Jest to podstawowy mechanizm zachowania informacji genetycznej i jej dziedziczenia przez komórki potomne [44]. Komplementarność par zasad jest bardzo ważna dla procesu ekspresji informacji genetycznej. Dzięki niej sekwencja DNA może ulegać transkrypcji na mRNA, a ten translacji na białko funkcjonalne [22].

Informacja biologiczna niezbędna każdemu organizmowi do reprodukcji jest przechowywana w DNA. Jest ona zakodowana w sekwencji zasad i jest uporządkowana za pomocą dużej liczby genów[1], z których każdy posiada instrukcje dotyczące syntezy polipeptydu lub cząsteczki strukturalnego RNA. Gen można obrazowo przedstawić jako odcinek DNA o określonej sekwencji zasad, który koduje sekwencję aminokwasów. Długość genu może być znacząco różna i wahać się od mniej niż stu do kilku milionów par zasad [47]. W każdej komórce autosomalnej znajduje się jedna pełna kopia genomu. Genom ludzki zawiera około 20-25 tysięcy genów kodujących białka [24]. Co istotne, tylko ok. 1,5% genomu to eksony [25], czyli fragmenty kodujące białka. Reszta to introny, które mają kluczowe znaczenie dla regulacji ekspresji poszczególnych genów, jednak ich dokładna rola i sposób działania nadal w wielu wypadkach nie jest poznana [39]. Nić, z której odczytywana jest informacja biologiczna, nazywana jest nicią matrycową. Nić ta jest wykorzystywana do syntezy komplementarnej cząsteczki RNA. Druga z nici nazywana jest nicią kodującą, gdyż zsyntetyzowane RNA, które stanowi matrycę do syntezy białek, składa się z takiej samej sekwencji zasad, a zatem koduje układ aminokwasów w polipeptydzie [47].

[1] *gen* – Fragment sekwencji DNA (zazwyczaj długości kilku tysięcy par zasad) posiadający informacje potrzebne do syntezy RNA i białka. Jeden gen koduje jedno białko [20].

1.2. Ekspresja genów

Jak już zostało wcześniej wspomniane, informacja biologiczna jest zapisana w cząsteczce DNA za pomocą sekwencji zasad. Zostaje ona udostępniona komórce w procesie znanym jako ekspresja genów. Sposób wykorzystania informacji, zwany „centralnym dogmatem biologii molekularnej", polega na jej przepływie od DNA do RNA (transkrypcja genów) i dalej do białka (translacja genów). Proces ten został zaprezentowany na Rysunku 1.2. Dogmat sformułowany przez Francisa Cricka [10], jednego z odkrywców struktury DNA zakłada, że przepływ informacji może odbywać się tylko w jednym kierunku – od DNA przez RNA do białka. Istnieje jednak wyjątek od tej reguły – retrowirusy zawierają enzym nazywany odwrotną transkryptazą pozwalający na przepisywanie RNA na DNA. Informacja biologiczna zawarta w genach funkcjonuje jako zbiór instrukcji dotyczących syntezy białek w odpowiednim miejscu i o właściwym czasie. Z kolei skoordynowane działania wielu różnych białek są niezbędne do funkcjonowania komórek i całych organizmów.



Rysunek 1.2: Centralny dogmat biologii molekularnej. Źródło: Opracowanie własne na podstawie [10]

1.2.1. Transkrypcja genów

Komórki zawierają trzy podstawowe klasy RNA: transportujące (tRNA), rybosomowe (rRNA) i informacyjne (mRNA) [19]. tRNA i rRNA są molekularnymi narzędziami niezbędnymi w procesie biosyntezy (translacji) białka. mRNA działa podczas translacji jako matryca, na podstawie której syntetyzowane jest białko. Etap tworzenia tej matrycy na bazie oryginalnej informacji genetycznej zakodowanej w DNA nazywany jest transkrypcją. Cząsteczki mRNA są tworzone na nici matrycowej DNA przez enzymy zwane polimerazami RNA. Określenie „sekwencja genu" odnosi się zazwyczaj do nici niematrycowej, gdyż wyprodukowane mRNA składa się z takich samych sekwencji jak ta nić. Proces transkrypcji składa się z trzech etapów: inicjacji, elongacji i terminacji [47].

Transkrypcja nie rozpoczyna się w losowym miejscu na DNA – miejscem startu jest początek genu. Znakiem do zainicjowania transkrypcji jest sekwencja zasad promotora, który jest położony tuż przed sekwencją genu ulegającego transkrypcji i wskazuje miejsce rozpoczęcia działania polimerazy RNA. Rezultatem związania się enzymu z promotorem jest powstanie zamkniętego kompleksu promotorowego, w którym odcinek DNA będący promotorem występuje w postaci dwuniciowej helisy. Następnie nić pod wpływem polimerazy RNA ulega dysocjacji tworząc otwarty kompleks promotorowy. W nim dwa pierwsze rybonukleotydy wiążą się z DNA tworząc pierwsze wiązanie fosfodiestrowe jednocześnie inicjując transkrypcję [34].

Drugim etapem jest elongacja, podczas której polimeraza RNA przemieszcza się wzdłuż cząsteczki DNA jednocześnie topiąc i rozplatając dwuniciową helisę. Enzym dołącza rybonukleotydy do końca $3'$ wydłużającego się łańcucha RNA w kolejności narzuconej przez umiejscowienie zasad w matrycowej nici DNA. Zazwyczaj w pierwszej kolejności transkrypcji ulega sekwencja liderowa o różnej długości w zależności od genu, a dopiero po niej sekwencja kodująca genu. Na przeciwnym końcu sekwencji kodującej również zawarty jest odcinek, który

nie koduje aminokwasów, zwany niekodującą sekwencją 3′-końcową – po nim transkrypcja się kończy. RNA jest sparowane z fragmentem nici matrycowej DNA na długości około 12 zasad. Rozplatany jest niewielki rejon DNA, gdyż rozplatanie jednego fragmentu powoduje większą częstość skrętów we fragmencie przyległym, co z kolei prowadzi do tworzenia się napięć w cząsteczce DNA. Aby niwelować tę niedogodność, w trakcie trwania syntezy, powstałe RNA jest oddzielane od matrycy DNA, a DNA ulega ponownemu odtworzeniu dwuniciowej struktury helisy [17].

Ostatnim etapem procesu transkrypcji jest terminacja, podczas której transkrypt oddziela się od matrycy podobnie jak polimeraza RNA, która następnie przechodzi do kolejnej rundy transkrypcji [47].

W komórkach eukariotycznych DNA zgromadzone jest w jądrze komórkowym – i tam też przebiega transkrypcja. Zanim transkrypt mRNA zostanie wyeksportowany do cytoplazmy, gdzie zachodzi biosynteza białek, poddawany jest tzw. obróbce posttranskrypcyjnej. Polega ona przede wszystkim na splicingu, czyli wycinaniu z mRNA fragmentów niekodujących (intronów), pozostawiając jedynie fragmenty kodujące sekwencję aminokwasów w białku (eksony). Do gotowej cząsteczki mRNA bezpośrednio odpowiadającej jednemu białku dołączany jest kap na końcu 5′ oraz sekwencja poli(A) na końcu 3′. Struktury te zabezpieczają mRNA przed degradacją w cytoplazmie i promują jego rozpoznanie i wiązanie do rybosomu, na którym zachodzi translacja [7].

1.2.2. Translacja

Translacja to kluczowy etap biosyntezy białek w komórce. Odpowiada za nią złożony kompleks białkowy nazywany rybosomem. Podczas translacji informacja zaszyfrowana w cząsteczce mRNA jest wykorzystywana do wyznaczenia kolejności aminokwasów budujących nowo syntetyzowane białko. Kluczową rolę pełnią tutaj cząsteczki tRNA, gdyż dostarczają do rybosomu aminokwasy w porządku narzuconym przez sekwencję nukleotydową mRNA. Przed rozpoczęciem translacji aminokwasy łączą się kowalencyjnie z odpowiadającymi im cząsteczkami tRNA, które z kolei rozpoznają kodony mRNA oznaczające określone aminokwasy. Z czterech rodzajów zasad (adenina, guanina, cytozyna, uracyl) występujących w RNA można utworzyć 64 różne kodony (trójki zasad). Trzy z nich stanowią sygnał stop procesu translacji, pozostałe 61 koduje 20 aminokwasów występujących w białkach. Oznacza to, że większości aminokwasów odpowiada więcej niż jeden kodon. Zjawisko to zwane jest degeneracją kodu [7].

W przebiegu translacji wyróżnia się takie same etapy, jak w transkrypcji (inicjacja, elongacja i terminacja). Podczas pierwszego etapu mRNA związuje się z rybosomem. W trakcie elongacji powtarzany jest proces przyłączania kolejnych aminokwasów do wydłużającego się łańcucha polipeptydowego. Cząsteczki tRNA transportujące aminokwasy przyłączają się kodonem do antykodonu mRNA, jednocześnie „oddając” aminokwas powstającemu białku. Translacja kończy się w momencie, gdy rybosom odczyta kodon wyznaczający zakończenie sekwencji białka na łańcuchu mRNA (kodon stop). Nie istnieją tRNA zdolne do łączenia się z kodonami stop. W zastępstwie tRNA z polipeptydem wiąże się czynnik terminacyjny doprowadzający do jego uwolnienia. Na zakończenie terminacji rybosom uwalnia mRNA [47].

1.2.3. Regulacja ekspresji genów

Regulacja ekspresji genów komórek eukariotycznych jest bardzo złożona. Każda komórka danego organizmu zawiera pełną kopię jego genomu – w wypadku człowieka jest to 20-25 tysięcy genów kodujących białka. W przeciętnej komórce aktywnych jest tylko około 15% genów, przy czym w różnych komórkach ekspresji ulegają odmienne geny. Modyfikacje przejawiające się w charakterystycznych cechach komórki są następstwem zmian w zestawie genów ulegających ekspresji. Zaburzenia poziomu ekspresji genów mogą być powiązane między innymi z upośledzeniem kontroli dzielenia się i programowanej śmierci komórek, co leży u podstaw chorób nowotworowych [47].

Poziom biosyntezy poszczególnych białek w komórach eukariotycznych jest kontrolowany zarówno na poziome transkrypcji, jak i obróbki posttranskrypcyjnej oraz translacji. Ponieważ na pojedynczej nici mRNA może jednocześnie pracować wiele rybosomów, zmiana poziomu ekspresji danego genu na poziomie transkrypcji (tzn. ilości mRNA odpowiadającego danemu genowi na DNA) może być silnie amplifikowana na poziomie transkrypcji. Przekłada się to na szeroki zakres efektywnego działania całego systemu: komórka może utrzymywać stałą, bardzo niską ekspresję niektórych genów (co daje pojedyncze kopie danego białka w komórce), jak również dynamicznie odpowiadać na bodźce, zwiększając w krótkim czasie liczbę kopii danego białka o rzędy wielkości. Regulacja ekspresji genów odbywa przez przyłączanie się białek zwanych czynnikami transkrypcyjnymi do sekwencji promotorowych genów. Promotorem nazywany jest odcinek DNA znajdujący się przed sekwencją kodującą odpowiedzialny za regulowanie ekspresji genów. Polimeraza RNA i czynniki transkrypcyjne rozpoznają zachowawcze sekwencje promotora i łączą się z nimi, co prowadzi do transkrypcji i powstania RNA będącego kopią sekwencji danego genu. Zatem intensywność inicjacji transkrypcji zależy od siły oddziaływań między sekwencjami regulatorowymi znajdującymi się w okolicach promotora a białkowymi czynnikami transkrypcyjnymi. Zwiększająca się częstotliwość oddziaływań powoduje wzrost ekspresji danego genu w komórce [7].

Wydajność transkrypcji genu może być także kontrolowana poprzez fragmenty DNA zwane sekwencjami wzmacniającymi. Mogą się one znajdować w odległości tysięcy par zasad od miejsca, w którym inicjowana jest transkrypcja. Sekwencje wzmacniające składają się zazwyczaj z 100-200 par zasad i są zdolne wiązać czynniki transkrypcyjne jednocześnie stymulując transkrypcję określonego genu. Sekwencja wzmacniająca może być zlokalizowana zarówno powyżej, jak i poniżej genu, na który wpływa, działając równie efektywnie w obu orientacjach. Możliwa jest także regulacja ograniczająca ekspresję genu. Odpowiedzialne za nią są sekwencje wyciszające, które wiążą czynniki transkrypcyjne zmniejszając wydajność transkrypcji.

Ekspresja genów może być regulowana również przez hormony i cytokiny. Hormony, to związki chemiczne syntetyzowane przez określone komórki organizmu. Wpływają na inne komórki powodując zmianę ich funkcji i właściwości poprzez aktywację transkrypcji specyficznych genów. Cytokiny, to białka działające podobnie do hormonów. Komórkami, w których często działają cytokiny są komórki krwi. Oba związki wpływają na ekspresję genów w komórkach docelowych na kilka sposobów. Przykładem działania są rozpuszczalne w tłuszczach hormony steroidowe, które potrafią przechodzić przez błonę komórkową do cytoplazmy, gdzie następnie wiążą się z białkiem zwanym receptorem hormonu steroidowego. Związanie hormonu przez wspomniany receptor uwalnia receptor będący czynnikiem transkrypcyjnym z kompleksu z białkiem blokującym jego aktywność. Receptor następnie jest przemieszczany do jądra komórkowego, gdzie może aktywować transkrypcję genów poprzez związanie się z promotorem. Inaczej działają hormony polipeptydowe i cytokiny. Łączą się one z receptorami

na powierzchni błony komórkowej komórki docelowej. W procesie zwanym transdukcją sygnału kolejne kaskady białka ulegają aktywacji, co w konsekwencji prowadzi do stymulacji transkrypcji genów docelowych poprzez wiązanie czynników transkrypcyjnych z sekwencjami promotorowymi genów [47].

1.3. RNA-Seq

Badanie genomu organizmu daje jedynie informację o jego sekwencji DNA. Jak zostało już wspomniane w sekcji 1.1, sekwencja ta jest prawie identyczna dla każdej komórki organizmu i zawiera pełen zestaw genów oraz fragmenty niekodujące. Różnicowanie budowy oraz działania poszczególnych komórek, jak również cykl ich rozwoju czy odpowiedź na zewnętrzne bodźce, warunkowane są ekspresją poszczególnych genów. Jako że kluczowym etapem regulacji ekspresji jest transkrypcja DNA na mRNA, sekwencjonowanie i ilościowe oznaczanie mRNA dla danego rodzaju komórek może dostarczyć znacznie bardziej szczegółowych informacji niż sekwencjonowanie DNA. W transkryptomie (całości informacji przepisanej na mRNA) danej komórki występują bowiem jedynie te fragmenty kodu genetycznego, które są przez nią w danym momencie aktywnie wykorzystywane. Co więcej, ilość kopii danych fragmentów mRNA dostarcza informacji o poziomie ekspresji genu. Dzięki odczytowi mRNA jesteśmy więc w stanie stwierdzić, które geny są w komórkach aktywne i w jakim stopniu.

Obecnie wykorzystywany jest szereg różnych pod względem technicznym procedur sekwencjonowania mRNA [43], jednak wszystkie z nich opierają się na wspólnym schemacie. Punktem wyjścia jest stworzenie biblioteki fragmentów mRNA wyizolowanych z próbki, które kodują białka [12].

W ramach naturalnej obróbki posttranskrypcyjnej, do odcinków mRNA przyłączane są na końcu 3' polinukleotydy adeninowe – *poli(A)*. Ogon poli(A) chroni mRNA przed degradacją, jak również ma znaczenie dla jego transportu z jądra do cytoplazmy oraz wydajności późniejszej translacji [18]. Obecność ogona poli(A) na wszystkich fragmentach mRNA jest kluczowa dla ich izolacji na potrzeby RNA-Seq. Wykorzystuje się w tym celu np. mikrosfery magnetyczne funkcjonalizowane poli(T), które wiążą poliadenylowane mRNA, i które następnie można łatwo wydzielić z próbki i oczyścić dzięki zastosowaniu pola magnetycznego [32].

W większości procedur, kolejnym krokiem jest odwrotna transkrypcja mRNA na cDNA (DNA komplementarne – ang. *complementary DNA*) [46]. W zależności od techniki sekwencjonowania, może być wymagana dalsze cięcie enzymatyczne cDNA. Kiedy zostanie już przygotowana próbka zawierająca oczyszczone fragmenty cDNA o odpowiedniej długości, następuje właściwe sekwencjonowanie. Wykorzystuje się obecnie kilka systemów sekwencjonowania, takich jak 454 GS FLX (Roche), Genome Analyzer II (Illumina) czy SOLiD (Applied Biosystems) [46]. Wszystkie należą do klasy tzw. systemów następnej generacji (ang. *Next Generation Sequencing*) – charakteryzują się bardzo wysoką przepustowością, co związane jest z równoległym odczytem tysięcy, a nawet milionów sekwencji w tej samej próbce. Na podstawie odczytanych sekwencji cDNA, zgodnie z zasadą komplementarności, można bezpośrednio odtworzyć oryginalne sekwencje mRNA. Ze względu na możliwość wystąpienia artefaktów eksperymentalnych podczas odwrotnej transkrypcji [26], rozwijane są również techniki bezpośredniego sekwencjonowania mRNA.

Surowe dane eksperymentalne generowane w ramach procedury RNA-Seq to z reguły miliony względnie krótkich (o długości rzędu kilkudziesięciu nukleotydów) sekwencji. Aby umożliwić ich interpretację, niezbędne jest ich zmapowanie w ramach transkryptomu/genomu, to

znaczy określenie, jakim białkom/genom odpowiadają. W tym celu przeprowadza się analizę bioinformatyczną opartą o genom referencyjny bądź *de novo* [12]. Pierwsza z powyższych technik sprowadza się do alignmentu uzyskanych krótkich sekwencji wobec znanej sekwencji genomu danego organizmu. Choć ten rodzaj analizy niesie ze sobą pewne ryzyko (w genomie eukariotycznym występują introny, które nie znajdują odzwierciedlenia w sekwencji dojrzałego mRNA [47]), istnieje szereg wyspecjalizowanych, szybkich algorytmów zapewniających dopasowania wysokiej jakości [41]. Zasadniczą wadą tego podejścia jest wymaganie dotyczące znajomości genomu danego organizmu – dlatego też najlepiej sprawdza się ono w wypadku dobrze znanych organizmów modelowych.

Alternatywna metoda polega na rekonstrukcji transkryptomu *de novo*, wyłącznie na podstawie odczytanych sekwencji cDNA (mRNA). Jest to możliwe, ponieważ fragmentacja wykonywana przed sekwencjonowaniem przebiega w (do pewnego stopnia) losowych punktach. W związku z tym, przy odpowiedniej głębokości sekwencjonowania (analiza wielu kopii oryginalnego materiału – fragment odpowiadający danemu punktowi na mapie genomu jest odczytywany wielokrotnie), można wykorzystać nachodzące na siebie fragmenty odczytanych sekwencji do rekonstrukcji pełnej, oryginalnej sekwencji. Zasadnicza trudność związana jest z tym, że odczytywane fragmenty są krótkie, a więc również obszary na których się nakładają mają bardzo ograniczoną długość. To generuje niejednoznaczności przy łączeniu poszczególnych fragmentów ze sobą. Analiza *de novo* jest wymagająca pod względem mocy obliczeniowej [12]. Pozwala jednak ilościowo oceniać poziom ekspresji nawet w wypadku genów, których sekwencja nie jest znana *a priori*. Obecnie dostępny jest szereg algorytmów pozwalających na tego rodzaju analizę, takich jak Trinity, Trans-Abyss czy Oases [41], prowadzone są również intensywne prace nad ich dalszym rozwojem.

Jednym z podstawowych zastosowań sekwencjonowania mRNA jest analiza różnicowa ekspresji genów [43]. Jej celem jest ocena związku pomiędzy konkretnym czynnikiem, a poziomem ekspresji danych genów. Dotyczy to zarówno porównań między różnymi rodzajami komórek (różne organizmy/tkanki), jak i zmian w ekspresji genów w takich samych komórkach pod działaniem czynników zewnętrznych.

Rozdział 2

Matematyczne podstawy metod działania pakietów *DESeq* i *edgeR*

W niniejszym rozdziale przedstawione zostaną schematy działania pakietów DESeq i edgeR. Przybliżone zostanie jakimi założeniami kierowali się autorzy tworząc pakiet oraz jakie narzędzia matematyczne wykorzystali przy jego budowaniu. Wskazane zostaną również różnice w podejściu do analizowanych danych, które w konsekwencji wpływają na otrzymywane wyniki. Opisane będą kolejne działania, które pozwalają na otrzymanie listy statystycznie zmienionych genów wygenerowanej za pomocą dwóch różnych metod. Ostatecznie przybliżony zostaje problem wielokrotnego testowania oraz metoda Benjaminiego-Hochberga jako sposób na radzenie sobie z nim.

2.1. Dane wejściowe

Podstawowym obiektem jakiego wymaga zarówno pakiet *DESeq*[1] [1], jak i *edgeR*[2] [8, 11, 28, 30, 35–38, 48], jest tak zwana tabela zliczeń (ang. *count table*) przechowująca agregaty danych pochodzących z sekwencjonowania RNA (RNA-Seq). Jest to tabela zawierająca nieujemne, całkowite wartości. W i-tym wierszu i j-tej kolumnie umieszczona jest wartość K_{ij}, która określa ile razy i-ty odczyt został zmapowany (odczytany) w j-tym sekwencjonowaniu (próbce). Tabela składa się z G wierszy i n kolumn zawierających replikacje co najmniej dwóch warunków eksperymentalnych. Co ważne, każda z kolumn musi pochodzić z niezależnej biologicznie replikacji [3, 30]. W badanym w niniejszej pracy przypadku replikacje, to próbki pochodzące z sekwencjonowania RNA dla tego samego warunku eksperymentalnego. Analizowane będą replikacje pochodzące z dwóch warunków eksperymentalnych. W Tabeli

[1] *DESeq* – Pakiet statystyczny autorstwa Simona Andersa i Wolfganga Hubera upubliczniony w 2010 roku. Autorzy stworzyli narzędzie pomocne w analizie danych z RNA-Seq. Zaproponowana przez nich metoda zakłada ujemny dwumianowy rozkład danych, w którym wariancja ze średnią jest powiązana za pomocą regresji lokalnej [1]. Pakiet można pobrać pod adresem: https://bioconductor.org/packages/release/bioc/html/DESeq.html [dostęp na dzień: 23.07.2016 r.]

[2] *edgeR* – Pakiet statystyczny autorstwa Marka Robinsona, Davisa McCarthy'ego, Yunshuna Chena i Gordona K. Smytha upubliczniony w 2010 roku. Narzędzie to dostarcza metod statystycznych do określenia istotności różnic w danych dotyczących ekspresji genów. Autorzy proponują wykorzystanie do analiz danych z rozkładu ujemnego dwumianowego ich autorskiego testu *exactTest* [9]. Pakiet można pobrać pod adresem: https://bioconductor.org/packages/release/bioc/html/edgeR.html [dostęp na dzień: 23.07.2016 r.]

2.1 został pokazany przykładowy układ danych z zaznaczeniem przynależności próbek do poszczególnych warunków eksperymentalnych.

Tabela 2.1: Układ danych w tabeli zliczeń. Źródło: Opracowanie własne

warunek eksperymentalny	ρ_A				ρ_B			
ID próbki / ID genu	p_1	p_2	$\cdots$	p_k	p_{k+1}	p_{k+2}	$\cdots$	p_n
g_1	$k_{1,1}$	$k_{1,2}$	$\cdots$	$k_{1,k}$	$k_{1,k+1}$	$k_{1,k+2}$	$\cdots$	$k_{1,n}$
g_2	$k_{2,1}$	$k_{2,2}$	$\cdots$	$k_{2,k}$	$k_{2,k+1}$	$k_{2,k+2}$	$\cdots$	$k_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
g_G	$k_{G,1}$	$k_{G,2}$	$\cdots$	$k_{G,k}$	$k_{G,k+1}$	$k_{G,k+2}$	$\cdots$	$k_{G,n}$

Oznaczenia (w nawiasie podane są wartości odpowiadające analizowanym w tej pracy danym):

- G – liczba wszystkich genów (56 632),
- k – liczba replikacji (próbek) w warunku eksperymentalnym A (4),
- $n-k$ – liczba replikacji (próbek) w warunku eksperymentalnym B (9),
- n – liczba wszystkich replikacji (próbek) (13),
- ρ_A – warunek eksperymentalny A (Phdf),
- ρ_B – warunek eksperymentalny B (iPS),
- g_i – i-ty gen,
- p_j – j-ta próbka.

2.2. Rozkład ujemny dwumianowy

Celem analizy RNA-Seq jest wskazanie genów, w których występuje statystycznie istotna różnica w ekspresji genu pomiędzy warunkami eksperymentalnymi. Jeżeli odczyty są niezależną próbą z sekwencjonowania RNA, to mają one rozkład wielomianowy, który może zostać przybliżony rozkładem Poissona. Rozkład Poissona posiada jeden parametr jednoznacznie wyznaczony przez średnią. Wszystkie pozostałe właściwości są z nią funkcyjnie związane, w szczególności wariancja jest równa średniej. Zauważono [33], że empiryczne współczynniki rozkładu Poissona zachowują się inaczej niż teoretyczne: przewidywana wariancja jest mniejsza niż ta w rzeczywistości występująca w danych. Zaniżenie wartości tego parametru prowadzi do błędnego obliczania wartości statystyki testowej i p-wartości testu. Konsekwencją jest brak kontroli błędu I rodzaju (prawdopodobieństwa fałszywych odrzuceń hipotezy zerowej).

Sposobem na radzenie sobie z nadwyżką rozproszenia w analizie danych RNA-Seq jest wykorzystanie do modelowania rozkładu ujemnego dwumianowego (ang. *negative binomial distribution* (NB)). Rozwiązanie to zostało wykorzystane zarówno w pakiecie *DESeq*, jak i *edgeR* [1].

Zakładamy, że nieujemna i całkowita liczba odczytów z próbki j dla i-tego genu może być modelowana za pomocą rozkładu ujemnego dwumianowego,

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2). \tag{2.1}$$

Rozkład ma różne parametryzacje, w przypadku danych RNA-Seq wykorzystuje się parametryzację przez średnią μ_{ij} i wariancję σ_{ij}^2. Ogólniej, całkowita zmienna losowa K pochodzi

z rozkładu ujemnego dwumianowego z parametrami $p \in (0,1)$ i $r \in (1,\infty)$ oraz $r \in \mathbb{Z}$, gdy:

$$\Pr(K = k) = \binom{k+r-1}{r-1} p^r (1-p)^k. \tag{2.2}$$

Parametry p i r wyrażone w terminach średniej μ i wariancji σ^2 wynoszą [2]:

$$p = \frac{\mu}{\sigma^2} \text{ and } r = \frac{\mu^2}{\sigma^2 - \mu}.$$

Tradycyjnie wykorzystywana jest parametryzacja rozkładu ujemnego dwumianowego przez p i r, jednak w przypadku danych biologicznych wygodniej jest korzystać ze średniej μ i wariancji σ^2. W konsekwencji dopuszczalne jest otrzymanie niecałkowitej wartości parametru r.

Wartość oczekiwana liczby odczytów wyraża się wzorem [21]:

$$\mathbb{E}K = \frac{rp}{1-p} = \frac{\mu^3}{(\sigma^2 - \mu)^2}. \tag{2.3}$$

Rozkład ujemny dwumianowy posiada parametry jednoznacznie określone przez średnią μ i wariancję σ^2. Często jednak liczba replikacji w zbiorze danych jest zbyt mała by estymować oba parametry wiarygodnie dla każdego genu. W pakiecie *edgeR* poradzono sobie z tym problemem poprzez założenie, że średnia i wariancja są powiązane zależnością $\sigma^2 = \mu + \phi\mu^2$. ϕ jest nieznaną stałą proporcjonalności taką samą dla wszystkich genów, dlatego można ją wyestymować na podstawie danych. Dzięki temu zabiegowi konieczne jest wyznaczenie tylko jednego parametru, co pozwala na badanie niewielkich liczb replikacji. W pakiecie *DESeq* wykorzystano model rozszerzony o ogólniejsze, wynikające z danych relacje pomiędzy średnią a wariancją, które dają możliwość lepszego dopasowania modelu [1]. Więcej na temat relacji między parametrami w pakiecie *DESeq* w Rozdziale 2.4.

2.3. Normalizacja

DESeq

Celem wprowadzenia współczynników wielkości próbek s_j (ang. *size factor*) jest uczynienie odczytów pochodzących z różnych próbek, o być może odmiennej głębokości sekwencjonowania, porównywalnymi. Zatem proporcja $(\mathbb{E}K_{ij})/(\mathbb{E}K_{ij'})$ oczekiwanych wartości liczby odczytów dla tych samych genów, ale w różnych próbkach j i j' powinna być równa proporcji współczynników $s_j/s_{j'}$. Dzieje się tak w przypadku gdy w genie i nie jest obserwowana różnica ekspresji lub próbki j i j' są swoimi replikacjami. Niech k_{ij} oznacza zaobserwowaną liczbę odczytów genu i w próbce j. Całkowita liczba odczytów genu, $\sum_j k_{ij}$, jawi się jako dobra miara opisująca głębokość sekwencjonowania, co mogłoby sugerować wykorzystanie jej jako współczynnik s_j w próbkach. Jednak doświadczenia z rzeczywistymi danymi wykazały, że nie zawsze byłby to dobry wybór [1]. Czasami kilka genów z silną różnicą ekspresji potrafi mocno wpłynąć na całkowitą liczbę odczytów, co z kolei sprawia, że proporcja całkowitych liczb odczytów przestaje być dobrym estymatorem proporcji oczekiwanych odczytów. Z tego powodu do estymacji współczynników s_j wykorzystywana jest mediana proporcji obserwowanych odczytów. Zatem wartość estymowanych współczynników $\hat{s}_j$ opisana jest wzorem:

$$\hat{s}_j = \text{median}_i \frac{k_{ij}}{\left(\prod_{v=1}^{n} k_{iv}\right)^{1/n}}. \tag{2.4}$$

edgeR

Przyjmijmy, tak samo jak w poprzednim podrozdziale, że K_{ij}, to nieujemna i całkowita liczba odczytów z próbki j dla i-tego genu. Niech μ_{ij} oznacza prawdziwy i nieznany poziom ekspresji (liczbę transkryptów), L_i będzie długością genu i, a N_j – całkowitą liczbą odczytów w j-tej próbce. Wartość oczekiwaną K_{ij} możemy obliczyć jako:

$$\mathbb{E}K_{ij} = \frac{\mu_{ij} L_i}{D_j} N_j, \tag{2.5}$$

gdzie $D_j = \sum_{i=1}^{G} \mu_{ij} L_i$. D_j opisuje głębokość sekwencjonowania RNA w próbce. Zasadniczym problemem analizy RNA-Seq jest to, że podczas gdy znamy wartość N_j dla każdej próbki, to wartość D_j jest nieznana i może się znacząco wahać między próbkami w zależności od badanego fragmentu RNA. Jeżeli próbka wykazuje większą całkowitą ilość RNA, to analiza RNA-Seq może zaniżać ilość odczytów poszczególnych genów z tej próbki w porównaniu z innymi próbkami. W praktyce przyjmuje się, że parametr L_i, ponieważ jest stały dla genu, zawiera się w μ_{ij} i nie jest używany we wnioskowaniu [36].

Wartość D_j nie może być wprost estymowana dopóki nieznany jest poziom ekspresji i prawdziwa długość dla każdego genu. Jednak łatwiej jest wyznaczyć proporcję pomiędzy ilością RNA dla dwóch próbek $f_j = S_j / S'_j$. W pakiecie *edgeR* wykorzystano rozwiązanie polegające na porównywaniu poziomów ekspresji pomiędzy próbkami przy założeniu, że większość genów nie uległa ekspresji. Prostym, ale skutecznym sposobem na wyestymowanie f_j jest zastosowanie ważonej uciętej średniej dla logarytmów proporcji ekspresji między próbkami (ang. *trimmed mean of M values* (TMM)). Dla każdego genu definiujemy logarytm krotności zmiany (ang. *fold change*):

$$M_i = \log_2 \frac{K_{ij}/N_j}{K_{ij'}/N_{j'}} \tag{2.6}$$

oraz całkowity poziom ekspresji:

$$A_i = \frac{1}{2} \log_2 \left(\frac{K_{ij}}{N_j} \cdot \frac{K_{ij'}}{N_{j'}} \right), \; K_{i\bullet} \neq 0 \tag{2.7}$$

Ucięta średnia jest średnią obliczaną po usunięciu x% największych i najmniejszych wartości z danych. W procedurze TMM dane są przycinane dwukrotnie: logarytm krotności zmiany M_{ij}^r (próbka j w stosunku do próbki r dla i-tego genu) oraz całkowity poziom ekspresji A_i. W pakiecie *edgeR* domyślnie przycinanych jest 30% wartości M_i i 5% wartości A_i. Współczynnik normalizacyjny dla próbki j przy oznaczeniu próbki referencyjnej przez r wynosi:

$$\log_2(TMM_j^{(r)}) = \frac{\sum_{i \in G^*} w_{ij}^r M_{ij}^r}{\sum_{i \in G^*} w_{ij}^r}, \tag{2.8}$$

gdzie $M_{ij}^r = \frac{\log_2(K_{ij}/N_j)}{\log_2(K_{ir}/N_r)}$ i $w_{ij}^r = \frac{N_j - K_{ij}}{N_j K_{ij}} + \frac{N_r - K_{ir}}{N_r K_{ir}}$ dla $K_{ij}, K_{ir} > 0$, a G^* oznacza zbiór wszystkich genów.

Przypadki gdy $K_{ij} = 0$ lub $K_{ir} = 0$ są usuwane wcześniej, gdyż nie jest możliwe policzenie dla nich logarytmu krotności zmian. Oczywistym jest, że $TMM_r^{(r)} = 1$. W przypadku porównywania dwóch próbek, wystarczy obliczyć tylko jeden współczynnik normalizujący (f_j), który normalizuje próbkę referencyjną poprzez dzielenie jej przez $\sqrt{f_j}$, a próbkę j poprzez pomnożenie jej razy $\sqrt{f_j}$. Podobnie w przypadku analizy więcej niż dwóch próbek, jedna z nich jest wybierana jako próbka referencyjna, następnie obliczane są współczynniki normalizacyjne dla wszystkich pozostałych próbek [36].

2.4. Estymatory średniej i wariancji genu

DESeq

W praktyce wartości parametrów μ_{ij} i σ_{ij}^2 nie są znane, konieczne jest wyestymowanie ich na podstawie danych. Zazwyczaj liczba replikacji n jest mała przez co, aby możliwa była estymacja, potrzebne są dodatkowe założenia. W pracy Andersa i Hubera [1] przyjęto poniższe trzy założenia:

1. Średnia μ_{ij}, która stanowi wartość oczekiwaną zaobserwowanej liczby odczytów genu i w próbce j, jest iloczynem wartości $q_{i,\rho(j)}$ (gdzie $\rho(j)$ jest warunkiem eksperymentalnym dla próbki j) i współczynnika wielkości próbki s_j,

$$\mu_{ij} = q_{i,\rho(j)} s_j. \tag{2.9}$$

   $q_{i,\rho(j)}$ jest proporcjonalne do oczekiwanej, ale nieznanej wartości koncentracji fragmentu genu i pod warunkiem $\rho(j)$. W praktyce oznacza to, że $q_{i,\rho(j)}$ jest wyznaczane dla każdego genu i warunku eksperymentalnego, czyli w przypadku dwóch warunków eksperymentalnych może przyjmować $2 \cdot G$ różnych wartości. Współczynnik wielkości próbki s_j reprezentuje głębokość sekwencjonowania j-tej próbki. Termin *znormalizowana liczba odczytów* będzie oznaczał, że liczba odczytów każdego genu w próbce została podzielona przez współczynnik wielkości próbki odpowiadający tej próbce. Dzięki temu zabiegowi dane ze wszystkich próbek zostają sprowadzone do porównywalnych wielkości niezależnych od głębokości sekwencjonowania próbki.

2. Przyjmujemy, że wariancja σ_{ij}^2 jest sumą wariancji wynikającej z rozkładu Poissona (ang. *shot noise*) i składowej nadwyżki dyspersji (ang. *raw variance term*),

$$\sigma_{ij}^2 = \underbrace{\mu_{ij}}_{\text{wariancja Poissona}} + \underbrace{s_j^2 \upsilon_{i,\rho(j)}}_{\text{składowa nadwyżki dyspersji}} \tag{2.10}$$

3. Składowa nadwyżki dyspersji $\nu_{i,\rho}$ dla każdego genu jest gładką funkcją $q_{i,\rho}$,

$$\upsilon_{i,\rho(j)} = \upsilon_\rho(q_{i,\rho(j)}). \tag{2.11}$$

   Założenie to jest konieczne ze względu na zbyt małą liczbę replikacji powodującą problemy z uzyskaniem precyzyjnego estymatora wariancji genu i jedynie na podstawie danych dostępnych dla tego genu. Powyższe założenie pozwala na połączenie danych dla genów o podobnej ekspresji w celu dokładniejszej estymacji wariancji.

Dekompozycja wariancji przedstawiona w równaniu (2.10) jest umotywowana następującym modelem hierarchicznym: Wskaźnik $s_j r_{ij}$ odpowiada sytuacji gdy fragmenty genu i są sekwencjonowane, a bieżąca koncentracja fragmentów genu i z próbki j jest proporcjonalna do losowej zmiennej R_{ij}. Dla każdego genu i oraz wszystkich próbek j należących do warunku eksperymentalnego ρ, R_{ij} są niezależne, o tym samym rozkładzie, ze średnią $q_{i\rho}$ i wariancją $\upsilon_{i\rho}$. Stąd liczba odczytów K_{ij} pod warunkiem $R_{ij} = r_{ij}$ ma rozkład Poissona z parametrem $s_j r_{ij}$. Rozkład brzegowy K_{ij} ma średnią μ_{ij} i wariancję daną wzorem (2.11).

Model posiada trzy rodzaje parametrów:

(i) m współczynników wielkości próbki; wartości oczekiwane dla wszystkich odczytów w próbce j są proporcjonalne do s_j.

(ii) dla każdego warunku eksperymentalnego ρ, n parametrów siły ekspresji $q_{i\rho}$; opisują one oczekiwane pokrycie fragmentów i-tego genu pod warunkiem ρ, wartość oczekiwana liczby odczytów dla genu i jest proporcjonalna do $q_{i\rho}$.

(iii) gładkie funkcje $\nu_\rho : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$; dla każdego warunku eksperymentalnego ρ, ν_ρ modeluje zależność składowej nadwyżki dyspersji $\nu_{i\rho}$ od oczekiwanej średniej $q_{i\rho}$.

Do estymacji parametrów siły ekspresji $q_{i\rho}$ wykorzystujemy znormalizowane liczby odczytów:

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\hat{s}_j}, \tag{2.12}$$

gdzie m_ρ jest liczbą replikacji należących do warunku ρ, a suma jest liczona po tych replikacjach.

Estymator składowej nadwyżki dyspersji wyraża się wzorem [1]:

$$\hat{v}_\rho(\hat{q}_{i\rho}) = \hat{w}_\rho(\hat{q}_{i\rho}) - z_{i\rho}, \tag{2.13}$$

gdzie całkowitą wariancję próbkową obliczamy ze wzoru:

$$\hat{w}_{i\rho} = \frac{1}{m_\rho - 1} \sum_{j:\rho(j)=\rho} \left(\frac{k_{ij}}{\hat{s}_j} - \hat{q}_{i\rho} \right)^2, \tag{2.14}$$

a wariancja wynikająca z rozkładu Poissona dana jest wzorem:

$$\hat{z}_{i\rho} = \frac{\hat{q}_{i\rho}}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{1}{\hat{s}_j} \tag{2.15}$$

Pokażemy teraz, że różnica $w_{i\rho} - z_{i\rho}$ jest nieobciążonym estymatorem parametru $\nu_{i\rho}$ w równaniu (2.10). Ze wzoru (2.9) wiemy, że

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\hat{s}_j} \tag{2.16}$$

jest nieobciążonym estymatorem $q_{i\rho}$, gdyż $\mathbb{E}K_{ij} = s_j q_{i0}$. Z równania (2.14) mamy:

$$(m_\rho - 1)\hat{w}_{i\rho} = \sum_{j:\rho(j)=\rho} \left(\frac{k_{ij}}{\hat{s}_j} - \hat{q}_{i\rho} \right)^2. \tag{2.17}$$

Obliczając wartość oczekiwaną z obu stron równania otrzymujemy:

$$\begin{aligned}(m_\rho - 1)\mathbb{E}\hat{w}_{i\rho} &= \mathbb{E}\sum_{j:\rho(j)=\rho}\left(\frac{k_{ij}}{\hat{s}_j} - \hat{q}_{i\rho}\right)^2 = \\ &= \mathbb{E}\sum_{j:\rho(j)=\rho}\left(\frac{k_{ij}^2}{\hat{s}_j^2} - 2\frac{k_{ij}}{\hat{s}_j}\cdot\frac{1}{m_\rho}\sum_{l:\rho(l)=\rho}\frac{k_{il}}{\hat{s}_l} + \frac{1}{m_\rho^2}\left(\sum_{l:\rho(l)=\rho}\frac{k_{il}}{\hat{s}_l}\right)^2\right) = \\ &= \mathbb{E}\sum_{j:\rho(j)=\rho}\left(\frac{k_{ij}^2}{\hat{s}_j^2} - 2\frac{k_{ij}}{\hat{s}_j}\cdot\frac{1}{m_\rho}\sum_{l:\rho(l)=\rho}\frac{k_{il}}{\hat{s}_l} + \frac{1}{m_\rho^2}\left(\sum_{\substack{j:\rho(j)=\rho\\ l:\rho(l)=\rho\\ j=l}}\frac{k_{il}^2}{\hat{s}_l^2} + \sum_{\substack{j:\rho(j)=\rho\\ l:\rho(l)=\rho\\ j\neq l}}\frac{k_{ij}k_{il}}{\hat{s}_j\hat{s}_l}\right)\right) = \\ &= \left(1-\frac{1}{m_\rho}\right)\sum_{j:\rho(j)=\rho}\frac{\mathbb{E}K_{ij}^2}{\hat{s}_j^2} - \frac{1}{m_\rho}\sum_{\substack{j:\rho(j)=\rho\\ l:\rho(l)=\rho\\ j\neq l}}\frac{\mathbb{E}K_{ij}K_{il}}{\hat{s}_j\hat{s}_l}\end{aligned} \tag{2.18}$$

Korzystając z niezależności K_{ij} i K_{il} mamy:

$$\mathbb{E}K_{ij}K_{il} = \hat{s}_j\hat{s}_l q_{i\rho(j)}^2 \tag{2.19}$$

oraz

$$\mathbb{E}K_{ij}^2 = \left(\mathbb{E}K_{ij}\right)^2 + \mathrm{var}(K_{ij}) = \hat{s}_j^2 q_{i\rho}^2 + \hat{s}_j q_{i\rho} + \hat{s}_j^2 \upsilon_{i\rho(j)}. \tag{2.20}$$

Zatem:

$$(m_\rho - 1)\mathbb{E}\hat{w}_{i\rho} = \left(1-\frac{1}{m_\rho}\right)\sum_{j:\rho(j)=\rho}\left(q_{i\rho}^2 + \frac{q_{i\rho}}{\hat{s}_j} + \upsilon_{i\rho(j)}\right) - \frac{1}{m_\rho}\sum_{\substack{j:\rho(j)=\rho\\ l:\rho(l)=\rho\\ j\neq l}} q_{i\rho}^2 \tag{2.21}$$

Dzieląc obustronnie przez $(m_\rho - 1)$ otrzymujemy:

$$\mathbb{E}\hat{w}_{i\rho} = \frac{1}{m_\rho}\sum_{j:\rho(j)=\rho}\left(q_{i\rho}^2 + \frac{q_{i\rho}}{\hat{s}_j} + \upsilon_{i\rho(j)}\right) - q_{i\rho(j)}^2, \tag{2.22}$$

a stąd:

$$\begin{aligned}\mathbb{E}\hat{w}_{i\rho} &= q_{i\rho}^2 + \frac{q_{i\rho}}{m_\rho}\sum_{j:\rho(j)=\rho}\frac{1}{\hat{s}_j} + \upsilon_{i\rho(j)} - q_{i\rho}^2 = \\ &= \underbrace{\frac{q_{i\rho}}{m_\rho}\sum_{j:\rho(j)=\rho}\frac{1}{\hat{s}_j}}_{z_{i\rho}} + \upsilon_{i\rho(j)}.\end{aligned} \tag{2.23}$$

Zatem:

$$\hat{\upsilon}_\rho(\hat{q}_{i\rho}) = w_\rho(\hat{q}_{i\rho}) - z_{i\rho} \tag{2.24}$$

jest nieobciążonym estymatorem czystej wariancji [2].

edgeR

Niech π_{gi} będzie oczekiwaną frakcją wszystkich fragmentów w j-tej próbce pochodzących z genu i, wówczas dla każdej próbki $\sum_{i=1}^{G} \pi_{ij} = 1$. Niech $\sqrt{\phi_i}$ oznacza współczynnik zmienności π_{ij} pomiędzy replikacjami j (ang. *Coefficient of Variation* (CV)), liczony jako iloraz odchylenia standardowego i średniej liczby odczytów genu. Wówczas:

$$\mathbb{E}K_{ij} = \mu_{ij} = N_j \pi_{ij} \tag{2.25}$$

Zakładając, że K_{ij} pochodzi z rozkładu Poissona dla kolejnych sekwencjonowań tej samej próbki RNA, otrzymujemy wzór na wariancję K_{ij}:

$$\mathrm{var}(K_{ij}) = \mathbb{E}_\pi \left[\mathrm{var}(K \mid \pi)\right] + \mathrm{var}_\pi \left[\mathbb{E}(K \mid \pi)\right] = \mu_{ij} + \phi_i \mu_{ij}^2 \tag{2.26}$$

Dzieląc obustronnie przez μ_{ij}^2 dostajemy:

$$\mathrm{CV}^2(K_{ij}) = \frac{1}{\mu_{ij}} + \phi_i, \tag{2.27}$$

gdzie składnik $\frac{1}{\mu_{ij}}$ jest kwadratem współczynnika zmienności rozkładu Poissona zwanego technicznym współczynnikiem zmienności (ang. *Technical Coefficient of Variation* (TCV)), natomiast ϕ_i odpowiada kwadratowi współczynnika zmienności niezaobserwowanych wartości ekspresji. ϕ_i nazwiemy dyspersją, a $\sqrt{\phi_i}$ – biologicznym współczynnikiem zmienności (ang. *Biological Coefficient of Variation* (BCV)). W BCV zawiera się wariancja pomiędzy replikacjami, prawdziwa biologiczna wariancja oraz wahania wynikające z przygotowania próbki. TCV rośnie wraz ze wzrostem liczby odczytów w próbce. Inaczej jest z BCV, które może być dominującym źródłem niepewności dla genów z dużą liczbą odczytów. Dlatego też jak najdokładniejsze oszacowanie BCV ma kluczowe znaczenie dla realnej oceny ekspresji różnicowej w eksperymentach RNA-Seq [30].

2.5. Zastosowanie uogólnionych modeli liniowych

Ze zjawiskiem nadwyżki dyspersji (ang. *overdispersion*) mamy do czynienia, gdy w analizowanym zbiorze danych wariancja jest większa niż należałoby oczekiwać bazując na wariancji rozkładu z jakiego pochodzą dane. Z takim zjawiskiem spotykamy się w analizie danych RNA-Seq. Ze względu na to, że dyspersja często jest niedoszacowywana, a same wyniki analizy wykazują się brakiem kontroli występowania błędu I rodzaju, w obu pakietach zdecydowano się na zastosowanie uogólnionych modeli liniowych (ang. *Generalized Linear Models* (GLMs)) do estymacji rzeczywistej wartości dyspersji. Uogólnione modele liniowe są rozszerzeniem klasycznych liniowych modeli na zmienne odpowiedzi nie pochodzące z rozkładu normalnego. Pozwalają na to, aby zmienna odpowiedzi pochodziła z dowolnej rodziny wykładniczej rozkładów prawdopodobieństwa.

DESeq

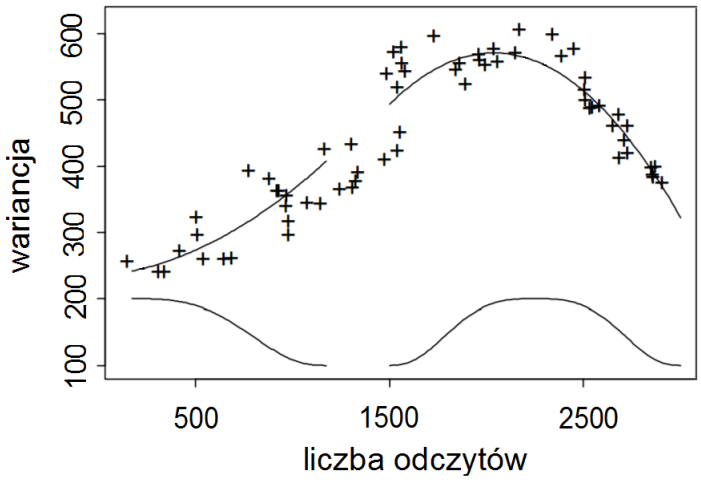
Okazuje się, że dyspersja wyestymowana na podstawie niewielkiej liczby replikacji oscyluje wokół swojej rzeczywistej wartości z tendencją do niedoszacowania. W pakiecie *DESeq* zaproponowane są dwie metody „doszacowywania” (estymowania nadwyżki) dyspersji: regresja

parametryczna i regresja lokalna. Pierwsza z nich jest zaimplementowana jako rozwiązanie domyślne. Obie metody oparte są na uogólnionych modelach liniowych z rodziny gamma.

Uogólniony model liniowy z rodziny gamma opisuje zależność postaci:

$$g(\mathbb{E}(Y_i)) = X_i'\beta, \tag{2.28}$$

gdzie Y_i jest zmienną objaśnianą z rozkładu gamma, X_i – wektorem zmiennych objaśniających, β – współczynnikami modelu, a $g()$ jest funkcją łączącą. Estymacja współczynników modelu odbywa się zazwyczaj metodą największej wiarogodności lub ważonych najmniejszych kwadratów [13].



Rysunek 2.1: Przykład działania regresji lokalnej. Wykres zależności wariancji liczby odczytów od liczby odczytów. Krzywe wyrysowane przy dolnej krawędzi wykresu wyznaczają przedziały zmiennej objaśniającej. Krzywe leżące pomiędzy wartościami zmiennej objaśniającej zostały wyestymowane lokalnie za pomocą metody najmniejszych kwadratów. Źródło: Schemat zmodyfikowany na podstawie [27, str. 17]

W przypadku wyboru regresji parametrycznej wykonywana jest funkcja `glm` z pakietu `stats`, która modeluje zależność postaci $g(\phi) \sim \mu$, przyjmując jako $g()$ funkcję odwrotną (ang. *inverse function*). Regresja lokalna wykonywana jest natomiast za pomocą funkcji `locfit` z pakietu `locfit`. Modeluje ona zależność postaci $\sigma^2 \sim \log(\mu)$. Cechą charakterystyczną regresji lokalnej jest to, że estymuje ona kilka oddzielnych krzywych regresji dla zadanych przedziałów zmiennej objaśniającej. Metoda ta pomaga zwiększyć dokładność dopasowania poprzez usunięcie wpływu różnic w zależnościach pomiędzy kolejnymi przedziałami zmiennej objaśniającej. Przykład regresji lokalnej, zaczerpnięty z [27], został przedstawiony na Rysunku 2.5.

Przyjmuje się, że dopasowana krzywa regresji odwzorowuje rzeczywistą dyspersję w próbce. Domyślnie, jeżeli wartość wariancji wyestymowana dla pojedynczego genu jest mniejsza niż odpowiadająca mu wartość na krzywej regresji, to uznaje się, że jest to wartość niedoszacowana. Wówczas do dalszych obliczeń wykorzystuje się wartość wskazaną przez krzywą regresji. W przeciwnym przypadku wariancja pozostaje bez zmian [3]. Dokładny opis sposobów wyboru wartości dyspersji wykorzystywanej do testowania różnic w ekspresji w pakiecie *DESeq* znajduje się w Dodatku A.

edgeR

W przypadku danych RNA-Seq związek pomiędzy średnią a wariancją jest opisany za pomocą równania (2.26) co wynika z założenia, że odczyty pochodzą z rozkładu ujemnego dwumianowego. Zakładając, że posiadamy wyestymowaną wartość ϕ_i, wariancję możemy oszacować dla każdej wartości μ_{ij}, a dopasowany model log-liniowy dla każdego genu przyjmuje postać:

$$\log \mu_{ij} = \mathbf{x}_j^T \beta_i + \log N_j, \tag{2.29}$$

gdzie $\mathbf{x}_j$ jest wektorem zmiennych towarzyszących (ang. *covariates*) specyfikujących warunek eksperymentalny próbki j, a β_i jest wektorem współczynników regresji dla i-tego genu.

Współczynniki regresji poszukiwane są z wykorzystaniem iteracyjnej metody Newtona-Raphsona [13]. Algorytm ten wykorzystuje maksimum funkcji wiarogodności ($U(\beta_i) = 0$) do znajdowania wektora kolejnych wartości β_i. W m-tej iteracji:

$$\beta_i^{(m)} = \beta_i^{(m-1)} - U_i^{(m-1)} \left(\mathcal{I}_i^{(m-1)}\right)^{-1}, \tag{2.30}$$

gdzie:

$\beta_i^{(m-1)}$ – wektor współczynników znalezionych w $(m-1)$ iteracji,

$U_i^{(m-1)} = \frac{dl}{d\beta_i^{(m-1)}}$ – pochodna logarytmu funkcji wiarogodności nazywana funkcją wynikową (ang. *score function*),

$\mathcal{I}_i^{(m-1)} = \mathbb{E}(-U')$ – macierz informacji Fishera (ang. *Fisher information matrix*).

Działanie algorytmu rozpoczyna inicjująca wartość $\beta_i^{(1)}$. Obliczenia wykonywane są do momentu gdy odległość pomiędzy współczynnikami znalezionymi w kolejnych iteracjach jest wystarczająco mała, tzn.:

$$\left|\beta_i^{(m)} - \beta_i^{(m-1)}\right| \leq \varepsilon. \tag{2.31}$$

Pochodną funkcji log-wiarogodności w przypadku modelu log-liniowego z uwzględnieniem współczynników β_i jest $X^T \mathbf{z}_i$, gdzie X jest macierzą planowania (ang. *design matrix*) o kolumnach $\mathbf{x}_j$ i $z_{ij} = (y_{ij} - \mu_{ij})/(1 + \phi_i \mu_{ij})$. Macierz informacji Fishera dla współczynników można zapisać jako $\mathcal{I}_i = X^T W_i X$, gdzie W_i jest diagonalną macierzą wag. Zatem w tym szczególnym przypadku:

$$\beta_i^{(m)} = \beta_i^{(m-1)} + (X^T W_i X)^{-1} X^T \mathbf{z}_i. \tag{2.32}$$

Opisana powyżej iteracyjna metoda poszukiwania maksimum funkcji wiarogodności jest przybliżeniem algorytmu Raphsona-Newtona z macierzą informacji Fishera odpowiadającą macierzy drugich pochodnych. Strategia wyszukiwania liniowego może być wykorzystywana do przybliżania dowolnej dodatnio określonej macierzy drugich pochodnych. W pakiecie *edgeR* wykorzystywane jest następujące uproszczenie nie powodujące straty ogólności: model liniowy jest parametryzowany w ten sposób, że $X^T X = I$. Jeśli ponadto μ_{ij} jest stałe dla genu i w próbce j, to macierz informacji upraszcza się do postaci iloczynu $\mu_i/(1 + \phi_i \mu_i)$ i macierzy jednostkowej I. Stosując powyższe przybliżenie dla macierzy informacji, iteracja z liniowo wyszukiwaną modyfikacją α jest postaci $\delta = \alpha X^T \mathbf{z}_i$, gdzie czynnik $\mu_i/(1+\phi_i \mu_i)$ jest zawarty w α. W przypadku stosowania powyższej metody każdy gen posiada indywidualną wartość α, która rośnie lub maleje w zależności od iteracji [30].

2.6. Testowanie różnicy w ekspresji

DESeq

Załóżmy, że posiadamy m_A replikacji próbek należących do warunku eksperymentalnego A i analogicznie m_B replikacji próbek należących do warunku eksperymentalnego B. Dla każdego genu i chcemy zmierzyć różnicę w ekspresji różnicowej dla tego genu pomiędzy warunkami eksperymentalnymi. W szczególności, chcemy testować hipotezę badawczą:

$$H_0 : q_{iA} = q_{iB}$$

$$H_1 : q_{iA} \neq q_{iB}$$

dla każdego genu $i = 1, 2, ..., G$, gdzie q_{iA} jest parametrem siły ekspresji dla próbek z warunku A, a q_{iB} dla warunku B. Definiujemy jako statystykę testową całkowitą liczbę odczytów dla każdego warunku,

$$K_{iA} = \sum_{j:\rho(j)=A} K_{ij},\ K_{iB} = \sum_{j:\rho(j)=B} K_{ij} \tag{2.33}$$

oraz ich całkowitą sumę $K_{iS} = K_{iA} + K_{iB}$. Pokażemy, że przy prawdziwości hipotezy zerowej możemy policzyć prawdopodobieństwo zdarzenia $K_{iA} = a$ i $K_{iB} = b$ dla dowolnej pary liczb a i b. Prawdopodobieństwo to oznaczymy przez $p(a, b)$. Wówczas p-wartość dla zaobserwowanej pary zliczeń (k_{iA}, k_{iB}) jest sumą wszystkich prawdopodobieństw mniejszych lub równych $p(k_{iA}, k_{iB})$. Wiedząc, że całkowita suma wynosi k_{iS} mamy:

$$p_i = \frac{\sum_{\substack{a+b=k_{iS} \\ p(a,b)\leqslant p(k_{iA},k_{iB})}} p(a, b)}{\sum_{a+b=k_{iS}} p(a, b)}. \tag{2.34}$$

Zmienne a i b w powyższych sumach przyjmują wartości $0, \ldots, k_{iS}$.

Załóżmy, że przy prawdziwości hipotezy zerowej liczby odczytów z różnych próbek są niezależne. Wówczas $p(a, b) = \mathbb{P}(K_{iA} = a)(K_{iB} = b)$. Problem polega na obliczeniu prawdopodobieństwa zdarzenia $K_{iA} = a$ i analogicznie $K_{iB} = b$.

Rozkład zmiennych K_{iA} i K_{iB} przybliżamy za pomocą rozkładu ujemnego dwumianowego z parametrami μ_{ij} oraz σ^2_{ij}. Na początek, ponieważ w hipotezie zerowej postulujemy $q_{iA} = q_{iB}$, estymujemy łączną średnią na podstawie liczby odczytów dla obu warunków eksperymentalnych,

$$\hat{q}_{i0} = \sum_{j:\rho(j)\in\{A,B\}} \frac{k_{ij}}{s_j}. \tag{2.35}$$

Zatem estymowane średnia i wariancja dla warunku eksperymentalnego A są postaci:

$$\hat{\mu}_{iA} = \sum_{j\in A} \hat{s}_j \hat{q}_{i0}, \tag{2.36}$$

$$\hat{\sigma}^2_{iA} = \sum_{j\in A} \hat{s}_j \hat{q}_{i0} + \hat{s}^2_j \hat{v}_A(\hat{q}_{i0}). \tag{2.37}$$

edgeR

Celem badania jest określenie czy istnieje statystycznie istotna ekspresja genu pomiędzy dwoma różnymi warunkami eksperymentalnymi. Hipoteza badawcza wyrażona jest w postaci:

$$H_0 : \pi_{iA} = \pi_{iB}$$

$$H_1 : \pi_{iA} \neq \pi_{iB}$$

dla każdego genu $i = 1, 2, ..., G$, gdzie A i B oznaczają replikacje należące do dwóch różnych warunków eksperymentalnych. Przybliżenia bazujące na testach asymptotycznych takich jak test Walda, score czy ilorazu wiarogodności, są adekwatne jedynie w przypadku posiadania dużych próbek. W związku z tym autorzy pakietu *edgeR* skonstruowali własny test do badania istotności różnic w ekspresji genów.

Dane znormalizowane za pomocą współczynników normalizacyjnych nazwiemy pseudodanymi. Jeśli współczynniki normalizacyjne stosujemy przy założeniu braku różnic między warunkami eksperymentalnymi, to pseudodane mają jednakowe rozkłady, co oznacza, że znane są również rozkłady całkowitej sumy pseudozliczeń dla każdego genu i warunku eksperymentalnego.

Niech Z_{iA} i Z_{iB} będą sumami pseudozliczeń odpowiednio dla warunku eksperymentalnego A i B przy liczbie replikacji należących do odpowiednich warunków wynoszącej m_A i m_B. Przy założeniu prawdziwości hipotezy zerowej, $Z_{ik} \sim \mathrm{NB}(m_k N^* \pi_i, \phi m_k^{-1})$, gdzie $k \in \{A, B\}$, a $N^* = \left(\prod_{j=1}^{n} N_j\right)^{\frac{1}{n}}$ jest średnią geometryczną rozmiarów próbek. Można skonstruować dokładny test podobny do dokładnego testu Fishera (ang. *Fisher's exact test*) dla tablic kontyngencji [14], zastępując prawdopodobieństwa hipergeometryczne ujemnymi dwumianowymi. Zakładając, że pseudosuma $Z_{iA} + Z_{iB}$ również jest zmienną losową z rozkładu ujemnego dwumianowego, możemy obliczyć prawdopodobieństwo wystąpienia wyższej ekspresji niż zaobserwowana. Innymi słowy, p-wartość dwustronnego testu jest zdefiniowana jako suma prawdopodobieństw nie występowania wyższej ekspresji genu w jednym z warunków eksperymentalnych [38].

2.7. Problem wielokrotnego testowania

W analizie RNA-Seq mamy do czynienia z problemem wielokrotnego testowania. Występuje on wtedy, gdy wykonywanych jest wiele testów na tym samym zbiorze danych. W tym przypadku testujemy geny pochodzące z tych samych próbek, dla każdego genu wykonując osobne testowanie. Takie działanie prowadzi do nadmiernej liczby wyników statystycznie istotnych, a w konsekwencji do wyciągania zbyt odważnych wniosków. W celu radzenia sobie z tym problemem, zarówno w pakiecie *DESeq*, jak i *edgeR*, wykorzystano współczynnik FDR (z ang. *False Discovery Rate*) [6] zaproponowany w 1995 roku przez Benjaminiego i Hochberga [3, 9].

Kontrola fałszywych odrzuceń

Rozważmy problem jednoczesnego testowania m hipotez zerowych, z których m_0 jest prawdziwych. **R** niech oznacza liczbę hipotez odrzuconych. W Tabeli 2.2 pokazano zestawienie otrzymanych wyników testowania:

Tabela 2.2: Liczba popełnionych błędów przy testowaniu m hipotez [5]

	Otrzymany wynik nieistotny statystycznie	Otrzymany wynik istotny statystycznie	Razem
H_0 prawdziwa	**U**	**V**	m_0
H_0 nieprawdziwa	**T**	**S**	$m - m_0$
	$m-$**R**	**R**	m

R jest zaobserwowaną zmienną losową, natomiast **U**, **V**, **S** i **T** są nieznanymi zmiennymi losowymi. Jeżeli każda pojedyncza hipoteza zerowa jest testowana oddzielnie na poziomie istotności α, to $\mathbf{R} = \mathbf{R}(\alpha)$ jest funkcją rosnącą. W dalszych rozważaniach przyjmiemy, że małe litery będą odpowiadały zaobserwowanym wartościom.

Przy tak oznaczonych zmiennych losowych wartość FWER (ang. *Family-wise error rate*) [6] wynosi $\mathbb{P}(\mathbf{V} \geqslant 1)$. Testując oddzielnie każdą z hipotez na poziomie istotności α/m mamy pewność, że $\mathbb{P}(\mathbf{V} \geqslant 1) \leqslant \alpha$.

Proporcja błędów polegających na fałszywym odrzuceniu prawdziwej hipotezy zerowej może być opisana za pomocą zmiennej losowej $\mathbf{Q} = \mathbf{V}/(\mathbf{V} + \mathbf{S})$. Przyjmuje się, że jeżeli $\mathbf{V} + \mathbf{S} = 0$, to $\mathbf{Q} = 0$ – nie występują błędy fałszywego odrzucenia hipotezy zerowej. **Q** jest nieznaną zmienną losową, ponieważ nie znamy v i s, a zatem nawet po eksperymencie i analizie danych nie znamy także $q = v/(v + s)$. Wskaźnik fałszywych odrzuceń Q_e definiujemy jako wartość oczekiwaną z **Q**,

$$Q_e = \mathbb{E}(\mathbf{Q}) = \mathbb{E}\left(\frac{\mathbf{V}}{\mathbf{V} + \mathbf{S}}\right) = \mathbb{E}\left(\frac{\mathbf{V}}{\mathbf{R}}\right) \tag{2.38}$$

Można łatwo pokazać dwie ważne właściwości Q_e:

(a) Jeżeli wszystkie hipotezy zerowe są prawdziwe, to FDR jest równoważne FWER: w tym przypadku $s = 0$ i $v = r$, więc jeżeli $v = 0$, to $\mathbf{Q} = 0$ i jeżeli $v > 0$, to $\mathbf{Q} = 1$. Wówczas $\mathbb{P}(\mathbf{V} \geqslant 1) \geqslant Q_e$. Zatem kontrola FDR implikuje kontrolę FWER w słabym sensie.

(b) Jeżeli $m_0 < m$, FDR jest mniejsze lub równe FWER: w tym przypadku jeżeli $v > 0$, to $v/r \leqslant 1$, co sprawia, że $\chi_{(\mathbf{V} \geqslant 1)} \geqslant \mathbf{Q}$. Biorąc wartość oczekiwaną z obu stron wyrażenia otrzymujemy $\mathbb{P}(\mathbf{V} \geqslant 1) \geqslant Q_e$. W rezultacie każda procedura kontrolująca FWER kontroluje także FDR. Jeżeli jednak procedura kontroluje jedynie FDR, to może być mniej rygorystyczna, co z kolei może powodować zysk na mocy testu. W szczególności, im większa liczba nieprawdziwych hipotez zerowych, tym większe może być **S** i stąd może wynikać różnica między wskaźnikami błędów. W związku z tym większy wzrost mocy testu jest możliwy przy przewadze nieprawdziwych hipotez zerowych.

Rozdział 3

Symulacyjna analiza metod za pomocą pakietów *DESeq* i *edgeR*

W tym rozdziale zostanie przedstawiona najważniejsza część niniejszej pracy, czyli symulacja oraz analiza uzyskanych wyników. Jej celem jest weryfikacja rezultatów, a zarazem ocena, który z pakietów lepiej radzi sobie ze wskazywaniem genów wykazujących istotną różnicę w ekspresji dla różnych rodzajów próbek eksperymentalnych.

3.1. Opis danych rzeczywistych

Analizowane dane RNA-Seq zostały wygenerowane w MD Anderson Cancer Center mieszczącym się na Uniwersytecie w Teksasie. Z danych korzystano dzięki uprzejmości Pracowni Terapii Genowej z Wielkopolskiego Centrum Onkologii w Poznaniu pod kierunkiem dr. n. med. Macieja Wiznerowicza. Do pierwszej grupy eksperymentalnej należą cztery replikacje komórek wyjściowych (ang. *Primary human derminal fibroblast* (Phdf)), a do drugiej dziewięć replikacji komórek macierzystych[1] wygenerowanych z komórek Phdf w procesie reprogramowania[2] (ang. *Induced Pluripotent Stem cells* (iPS)). Dane obejmują liczbę odczytów w próbkach dla 56 632 genów. Do wyznaczania współczynników normalizacyjnych oraz dyspersji wykorzystywane są wszystkie geny, następnie badana jest różnica w ekspresji dla wybranych 235 genów należących do pięciu grup (Autophagy, Clathrin independent, Lysosomes, RABs i SNARE).

3.2. Symulacyjna analiza metod

W celu próby rozstrzygnięcia, który z analizowanych pakietów jest bardziej wiarygodny w generowanych wynikach, wykonana została symulacja na danych opisanych w Rozdziale 3.1.

[1] *komórki macierzyste* – Niezróżnicowane komórki, które mogą różnicować się do wyspecjalizowanych komórek oraz w procesie podziału produkować nowe komórki macierzyste. Pluripotencjalne komórki macierzyste mogą dać początek każdemu typowi komórek dorosłego organizmu z wyjątkiem komórek łożyska.

[2] *reprogramowanie* – Proces odwrotny do różnicowania komórki, pozwalający na cofnięcie komórki do wcześniejszego etapu rozwoju (przekształcenie komórki funkcjonalnej do komórki macierzystej) [29].

Rozważono dwie ramki danych zawierające po 13 próbek oraz odpowiednio 25 lub 235 genów. Dane pochodzą z dwóch grup eksperymentalnych: komórek wyjściowych (cztery replikacje) i komórek wygenerowanych z komórek wyjściowych w procesie reprogramowania (dziewięć replikacji). W każdej z 1000 iteracji dla obu ramek danych wykonane zostały następujące kroki:

1. Losowy przydział próbek (wartości odczytów) do jednej z dwóch grup eksperymentalnych.
2. Utworzenie obiektów charakterystycznych dla pakietów: `CountDataSet` dla *DESeq* i`DGEList` dla *edgeR*.
3. Estymacja współczynników normalizacyjnych oraz dyspersji dla obu obiektów.
4. Testowanie istotności różnic w ekspresji funkcją `nbinomTest` dla pakietu *DESeq* oraz `exactTest` dla *edgeR*.
5. Zapisanie otrzymanych wartości FDR dla obu sposobów testowania.

Zadaniem symulacji jest wskazać jak często dana metoda testowania znajduje geny istotnie zróżnicowane w spermutowanych danych. Ponieważ próbki są losowo przydzielane do grup eksperymentalnych, to średnie liczby odczytów w obu grupach powinny być takie same. Zatem zakładamy, że analiza nie powinna wskazywać istotnych różnic w ekspresji genów pomiędzy warunkami eksperymentalnymi. Zatem zgodnie z oznaczeniami wprowadzonymi w Rozdziale 2.7: $\mathbf{V} = \mathbf{R}$. Testowanie dwóch, różniących się pod względem liczby genów, ramek danych może wskazać, jak wielkość zbioru wpływa na otrzymywane wyniki.

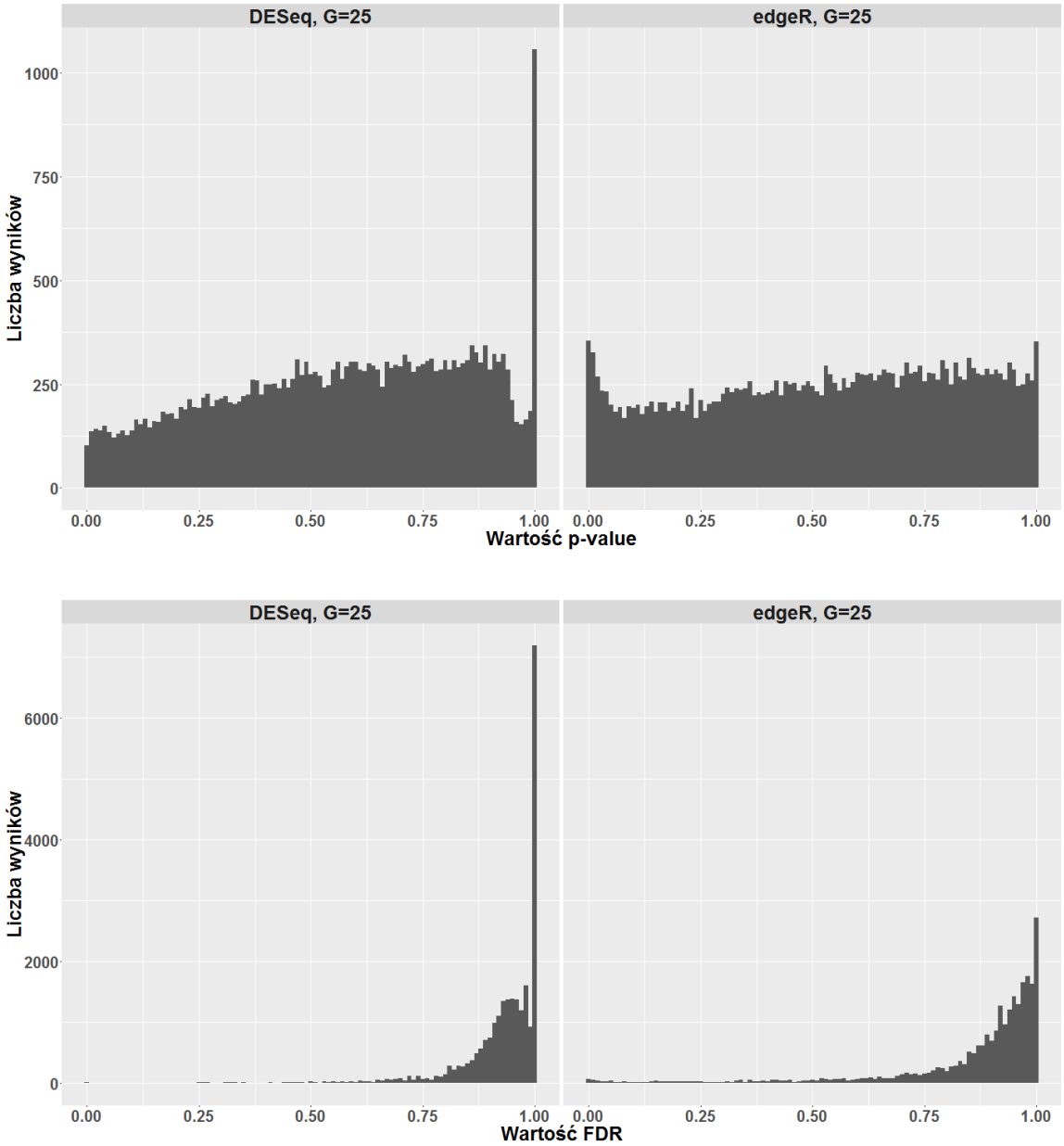
W Tabeli 3.1 znajduje się częściowe podsumowanie symulacji, czyli liczba istotnie zmienionych genów dla pierwszych 10 iteracji.

Tabela 3.1: Liczba istotnie zmienionych genów znalezionych za pomocą poszczególnych pakietów dla pierwszych 10 iteracji. Źródło: Opracowanie własne

Lp.	Liczba istotnie zmienionych genów			
	G = 25		G = 235	
	DESeq	edgeR	DESeq	edgeR
1	0	0	0	3
2	0	0	0	0
3	0	1	0	1
4	0	0	0	0
5	0	0	0	1
6	0	0	0	4
7	0	0	0	4
8	0	0	0	2
9	0	0	0	0
10	0	0	0	1

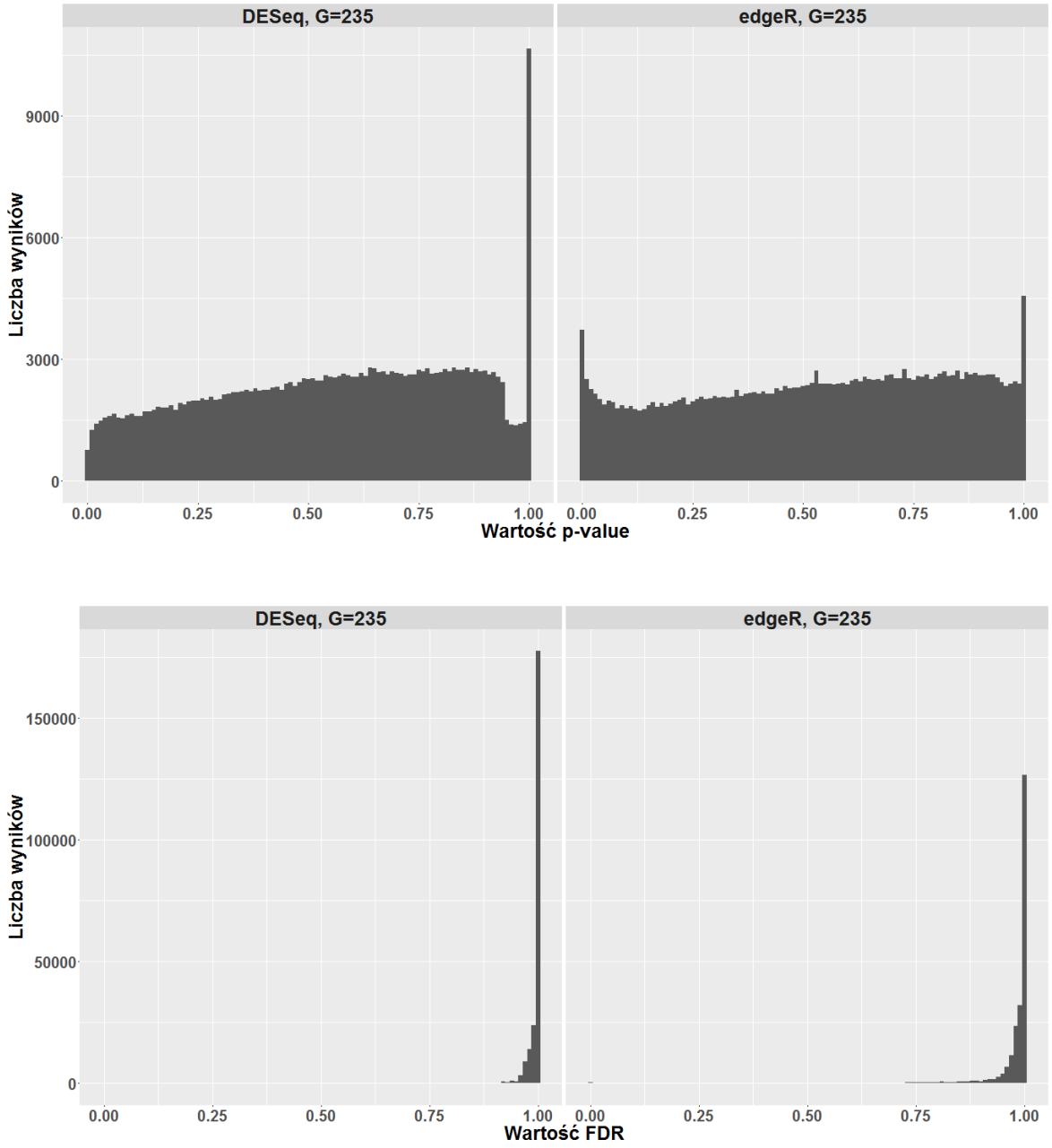
Podczas wykonywanej tysiąc razy symulacji 86 razy przy testowaniu 25 genów i 139 razy przy testowaniu 235 genów obliczenia pakietem *DESeq* wskazały na istotną różnicę w ekspresji genów przyjmując wskaźnik FDR na poziomie 10%. Takie same obliczenia wykonane za pomocą pakietu *edgeR* wskazały taką istotną różnicę odpowiednio 360 i 1988 razy. Na Rysunkach 3.1 oraz 3.2 przedstawiono histogramy wszystkich otrzymanych wartości p-value

i FDR z podziałem na liczbę testowanych genów i pakiet wykorzystany do testowania. Można wyraźnie zauważyć, że w przypadku obu pakietów przy mniejszej ilości testowanych genów stosunkowo więcej otrzymywanych wartości FDR jest mniejszych od 0.75. Natomiast p-value we wszystkich przypadkach rozkłada się mniej więcej jednostajnie, z wyjątkiem wyraźnego piku dla wartości 1 w pakiecie *DESeq*. Ponadto pakiet *DESeq* dla obu liczebności próbek zdecydowanie częściej wskazywał na kompletny brak różnic w ekspresji (FDR = 1).



Rysunek 3.1: Histogramy wartości p-value i FDR otrzymanych podczas symulacji dla pakietów *DESeq* i *edgeR* oraz liczebności próbek G=25. Źródło: Opracowanie własne

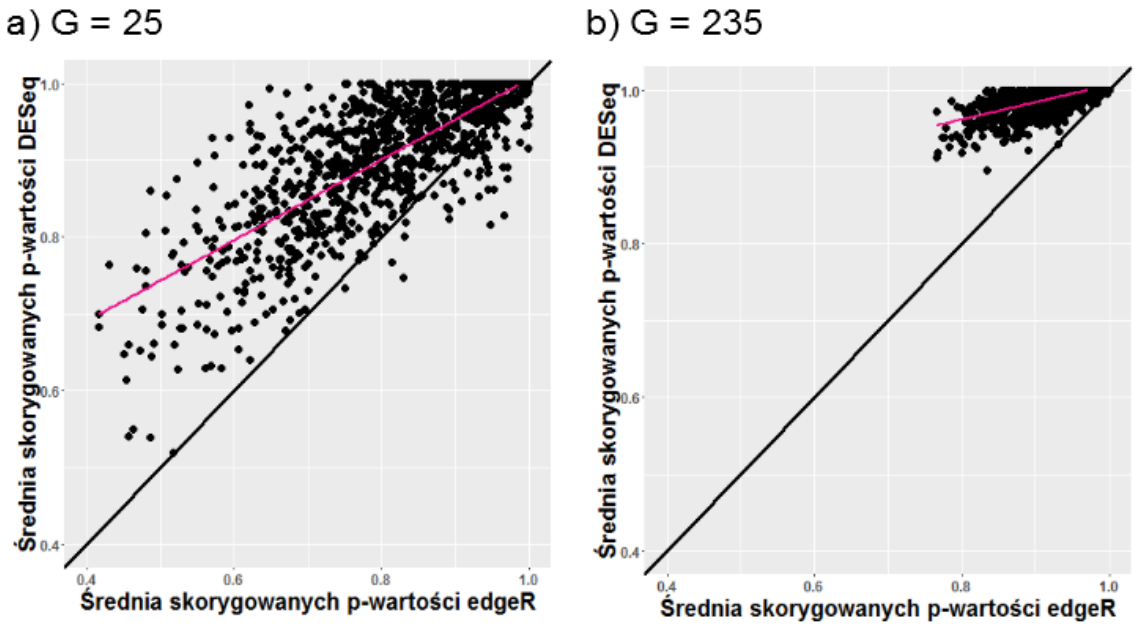
Maksymalnie liczba istotnych genów w iteracji wskazana przez pakiet *edgeR* wyniosła odpowiednio 4 i 10 razy (dla 25 i 235 genów), podczas gdy *DESeq* znalazł w tym przypadku odpowiednio 0 i 3 geny istotnie zmienione. Średnia wartość FDR dla obliczeń pakietem *DESeq* wyniosła 0.92 i 0.99 odpowiednio dla 25 i 235 genów. Za pomocą pakietu *edgeR* otrzymano średnie 0.86 i 0.95. Na podstawie uzyskanych średnich można stwierdzić, że oba pakiety lepiej



Rysunek 3.2: Histogramy wartości p-value i FDR otrzymanych podczas symulacji dla pakietów *DESeq* i *edgeR* oraz liczebności próbek G=235. Źródło: Opracowanie własne

szacowały przy większej liczbie analizowanych genów. Uzyskane wyniki pokazują, że obliczenia za pomocą pakietu *edgeR* zdecydowanie zbyt często wykazują istotną różnicę w ekspresji genów.

Graficzne zestawienie średnich ze wszystkich otrzymanych skorygowanych p-wartości znajduje się na Rysunku 3.3. Dla każdej iteracji wykresy przedstawiają zależność średniej wartości FDR otrzymanej w wyniku obliczeń pakietem *DESeq* od średniej wartości FDR pakietu *edgeR*. Średnia liczona była jako eksponenta ze średniej wartości logarytmów z wartości FDR w każdej iteracji. Czarna prosta odpowiada wartości $y = x$, czyli idealnej sytuacji, w której



Rysunek 3.3: Porównanie średnich wartości współczynników FDR otrzymanych za pomocą pakietów *DESeq* i *edgeR* dla 1000 symulacji dla a) G = 25 genów, b) G = 235 genów. Źródło: Opracowanie własne

otrzymane średnie są równe. Z kolei różowe proste zostały dopasowane za pomocą regresji liniowej. W przypadku obu liczebności zbiorów średnie wartości FDR dla pakietu *DESeq* w zdecydowanej większości były wyższe niż dla pakietu *edgeR*. Znalazło się jednak kilka odstępstw od tej reguły. Zdecydowanie więcej jest ich dla przypadku testowania próbki o mniejszej liczebności. Wartości *DESeq* są mniejsze od *edgeR* o nie więcej niż 0.002 podczas gdy w odwrotnej sytuacji taka różnica wynosi nawet 0.205 (Rysunek 3.3 a)). Dla przypadku b) tylko w trzech przypadkach średnia *edgeR* była większa niż *DESeq*. Rysunek 3.3 pokazuje także, że średnie ze skorygowanych p-wartości dla większej liczebności testowanego zbioru są skupione bliżej wartości 1. Można stąd wnioskować, że wielkość próbki ma wpływ na wiarygodność otrzymywanych wyników.

Uzyskane rezultaty pokazały, że pakiet *edgeR* niemal za każdym razem wskazuje na większe niż w rzeczywistości zróżnicowanie między grupami eksperymentalnymi. Zatem geny wskazywane przez pakiet *DESeq* jako istotnie zróżnicowane najczęściej rzeczywiście takie są, podczas gdy te wskazywane przez pakiet *edgeR* niekoniecznie. Pierwszy z pakietów daje mniej istotnych wyników, jednak są one bardziej wiarygodne. Ponadto większa liczba testowanych genów pozwala lepiej dopasować współczynniki normalizacyjne, co z kolei przekłada się na estymację rzeczywistej wartości dyspersji. W konsekwencji otrzymane wyniki testowania różnic w ekspresji genów mogą skutecznie wskazywać geny potencjalnie ciekawe do dalszych analiz.

Rozdział 4

Analiza danych rzeczywistych za pomocą pakietów *DESeq* i *edgeR*

W tym rozdziale Czytelnik zostanie przeprowadzony krok po kroku przez przykładową analizę danych rzeczywistych. Przedstawione zostaną kolejne etapy analizy oraz wyniki uzyskane za pomocą omawianych pakietów DESeq i edgeR. Liczne wykresy pozwolą na łatwiejsze zauważenie podstawowych różnic w otrzymanych wynikach. Zaprezentowana analiza jest wynikiem współpracy badawczej z zespołem Pracowni Terapii Genowej z Wielkopolskiego Centrum Onkologii w Poznaniu pod kierunkiem dr. n. med. Macieja Wiznerowicza.

4.1. Analiza danych rzeczywistych

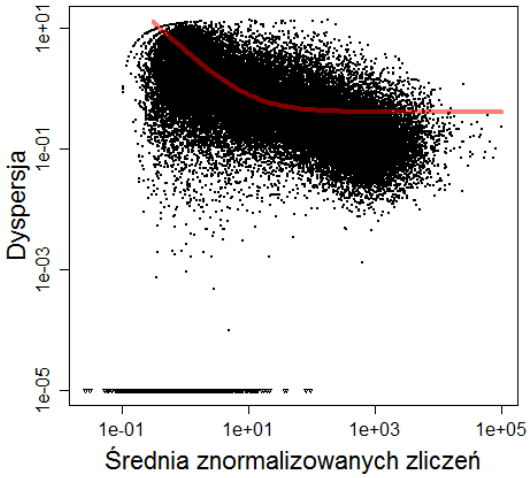
Celem badania jest porównanie ekspresji genów w dwóch warunkach eksperymentalnych. Opis analizowanych danych rzeczywistych znajduje się w Rozdziale 3.1.

Tabela 4.1: Współczynniki normalizacyjne obliczone za pomocą pakietów *DESeq* i *edgeR*. Źródło: Opracowanie własne

nazwa próbki	warunek eksperymentalny	wsp. normalizacyjny	
		DESeq	*edgeR*
Phdf 25	Phdf	0.4687508	0.8383620
Phdf 26	Phdf	1.1021224	0.8695647
Phdf 11	Phdf	0.6930038	0.8043354
Phdf X	Phdf	1.3234213	0.8526044
iPS 25.2	iPS	0.9064895	0.9870222
iPS 25.6	iPS	0.9507342	1.0660402
iPS 26.5	iPS	0.1558299	0.8807858
iPS 11.1	iPS	1.4241664	1.2056764
iPS X.7	iPS	2.8926552	1.2005784
iPS 26.6	iPS	1.4335421	1.2771690
iPS X.11	iPS	1.0836649	1.0407761
iPS 26.2	iPS	1.2783240	1.0634261
hES	iPS	2.4549015	1.0548158

W pierwszym kroku analizy obliczone współczynniki normalizacyjne sprowadzą dane ze wszystkich próbek do porównywalnej skali. Współczynniki normalizacyjne dla poszczególnych próbek i metod znajdują się w Tabeli 4.1.

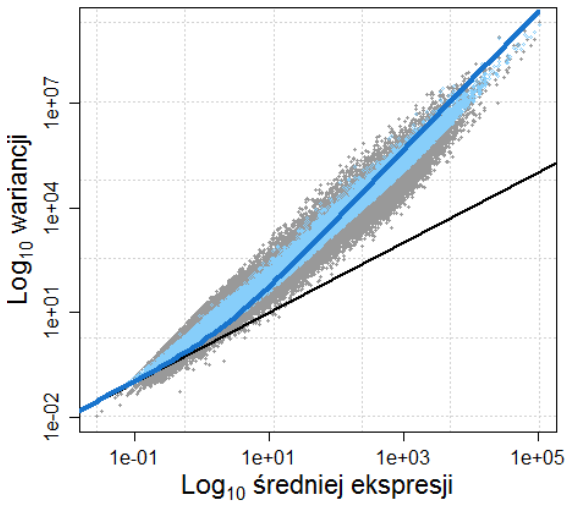
Kolejnym etapem działania w przypadku obu pakietów jest estymacja średniej i wariancji, a następnie zastosowanie uogólnionych modeli liniowych do estymacji dyspersji. Na Rysunku 4.1 przedstawiona jest wizualizacja tych etapów dla pakietu *DESeq*. Wykres przedstawia zależność dyspersji od średniej znormalizowanych zliczeń dla każdego genu. Czerwoną krzywą zaznaczona jest krzywa regresji dopasowana za pomocą uogólnionych modeli liniowych. Obliczenia zostały wykonane przy użyciu domyślnych opcji funkcji `estimateDispersions()`, czyli do estymacji dyspersji za pomocą uogólnionych modeli liniowych zastosowano metodę regresji parametrycznej. Ponadto wykorzystano metodę `pooled` (więcej o metodach w Dodatku A) polegającą na estymacji jednej wartości dyspersji dla każdego genu na podstawie wszystkich próbek z obu warunków eksperymentalnych. Zgodnie z opisem w Rozdziale 2.5, jeśli wyestymowana wartość dyspersji dla genu jest mniejsza niż ta wyznaczona przez krzywą regresji, to do testowania wykorzystywana jest wartość z krzywej regresji. W przeciwnym przypadku używana jest wartość dyspersji wyestymowana za pomocą metody `pooled` w funkcji `estimateDispersions()`.



Rysunek 4.1: Wykres zależności dyspersji od średniej znormalizowanych odczytów przy użyciu pakietu *DESeq*. Źródło: Opracowanie własne

Analogiczny wykres dla pakietu *edgeR* został przedstawiony na Rysunku 4.2. Każdy punkt odwzorowuje zależność logarytmu wariancji od logarytmu średniej ekspresji genu. Czarną prostą oznaczona jest teoretyczna wariancja wyznaczona na podstawie rozkładu Poissona, która jest równa średniej. Szarymi punktami oznaczone są wartości wariancji wyestymowane dla genów na podstawie wszystkich próbek, podobnie jak w pakiecie *DESeq*. Błękitne punkty odpowiadają dyspersji wyestymowanej za pomocą metody qCML (ang. *quantile adjusted Conditional Maximum Likelihood*) zaprezentowanej przez Robinsona i Smytha w [37]. Niebieska krzywa odpowiada krzywej regresji dopasowanej do otrzymanych wartości dyspersji.

Po wybraniu wartości dyspersji wykonywany jest test badający istotność różnic pomiędzy warunkami eksperymentalnymi dla każdego genu. W pakiecie *DESeq* wykonuje go funkcja `nbinomTest()`, a w pakiecie *edgeR* – `exactTest()`. W pierwszym z pakietów korekta otrzymanych p-wartości za pomocą procedury Benjaminiego-Hochberga jest wykonywana od razu



Rysunek 4.2: Wykres zależności wariancji od średniej ekspresji genu przy użyciu pakietu *edgeR*. Źródło: Opracowanie własne

przez funkcję testującą, natomiast w drugim zajmuje się tym funkcja `topTags()`. Tabela 4.2 zawiera zestawienie wyników dla czternastu pierwszych genów. Można zauważyć, że średnie liczności próbek są dość mocno zróżnicowane, jednak ten fakt nie wpływa na istotność różnic między warunkami eksperymentalnymi. Interesujące jest także to, że duża wartość krotności zmiany (np. dla genów ACP5 i AMBRA1) nie zawsze skutkuje wykazaniem istotnej różnicy przez wykonywane testy. Jest to wskazówka, że w analizie różnic w ekspresji genów nie należy kierować się jedynie wartością krotności zmiany.

Tabela 4.2: Zestawienie wyników uzyskanych za pomocą pakietów *DESeq* i *edgeR* dla wybranych genów. Źródło: Opracowanie własne

Nazwa genu	Grupa	Średnia	Średnia A	Średnia B	FC	$\log_2$FC	logCPM	p-wartość *DESeq*	p-wartość *edgeR*	FDR *DESeq*	FDR *edgeR*
ABCA2	Lysosomes	105,08	137,91	90,49	0,66	-0,61	2,83	0,42	0,24	0,60	0,37
ABCB9	Lysosomes	159,06	104,01	183,53	1,76	0,82	3,41	0,38	0,24	0,55	0,37
ACP2	Lysosomes	175,76	136,56	193,18	1,41	0,50	3,54	0,70	0,50	0,84	0,66
ACP5	Lysosomes	21,08	10,66	25,71	2,41	1,27	0,63	0,32	0,09	0,49	0,18
AGA	Lysosomes	320,65	566,10	211,56	0,37	-1,42	4,41	0,01	0,01	0,04	0,02
AMBRA1	Autophagy	317,38	165,56	384,85	2,32	1,22	4,38	0,17	0,07	0,32	0,15
ARHGAP26	Clathrin independent	255,22	222,79	269,64	1,21	0,28	4,07	0,67	0,73	0,81	0,82
ARSA	Lysosomes	45,39	64,32	36,97	0,57	-0,80	1,68	0,29	0,12	0,46	0,22
ARSB	Lysosomes	752,12	1649,70	353,20	0,21	-2,22	5,64	0,00	0,00	0,00	0,00
ARSG	Lysosomes	936,83	2121,36	410,37	0,19	-2,37	5,96	0,00	0,00	0,00	0,00
ASAH1	Lysosomes	1148,28	2303,76	634,73	0,28	-1,86	6,24	0,00	0,00	0,01	0,00
ATG10	Autophagy	1644,12	1583,35	1671,13	1,06	0,08	6,74	0,80	0,99	0,89	1,00
ATG12	Autophagy	1643,28	2661,27	1190,84	0,45	-1,16	6,76	0,07	0,03	0,15	0,07
ATG16L1	Autophagy	432,53	370,94	459,91	1,24	0,31	4,83	0,65	0,67	0,81	0,78

Oznaczenia:

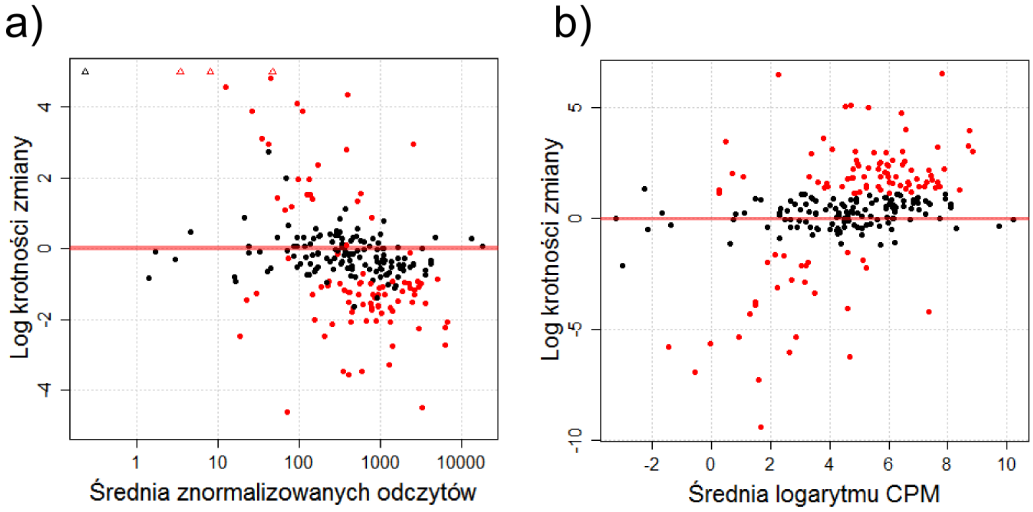
Średnia – średnia liczba odczytów genu obliczana na podstawie wszystkich n próbek
Średnia A – średnia liczba odczytów genu obliczana na podstawie k próbek należących do warunku eksperymentalnego A
Średnia B – średnia liczba odczytów genu obliczana na podstawie $n - k$ próbek należących do warunku eksperymentalnego B
FC – krotność zmiany obliczana jako iloraz średniej B przez średnią A
$\log_2$FC – logarytm przy podstawie 2 krotności zmiany
logCPM – logarytm naturalny CPM
p-wartość *DESeq* – p-wartość testu różnic w ekspresji wykonanego za pomocą funkcji `nbinomTest()` z pakietu *DESeq*
p-wartość *edgeR* – p-wartość testu różnic w ekspresji wykonanego za pomocą funkcji `exactTest()` z pakietu *edgeR*
FDR *DESeq* – p-wartość otrzymana za pomocą pakietu *DESeq* skorygowana za pomocą wskaźnika FDR
FDR *edgeR* – p-wartość otrzymana za pomocą pakietu *edgeR* skorygowana za pomocą wskaźnika FDR

W Tabeli 4.3 przedstawione jest podsumowanie wyników dla wszystkich grup analizowanych genów. Przyjmujemy współczynnik FDR na poziomie 10%, zatem istotną statystycznie różnicę w ekspresji wykazują te geny, dla których FDR < 0.1. Można zauważyć, że w każdej z grup testowanie za pomocą pakietu *edgeR* wskazywało średnio 15% więcej genów z istotną różnicą w ekspresji.

Tabela 4.3: Porównanie wyników otrzymanych za pomocą pakietów *DESeq* i *edgeR* w podziale na grupy. Źródło: Opracowanie własne

Grupa	Liczba genów testowanych	Liczba genów istotnych – *DESeq*	Liczba genów istotnych – *edgeR*
Autophagy	30	6	7
Clathrin independent	14	9	10
Lysosomes	94	47	51
RABs	61	23	25
SNARE	36	10	13

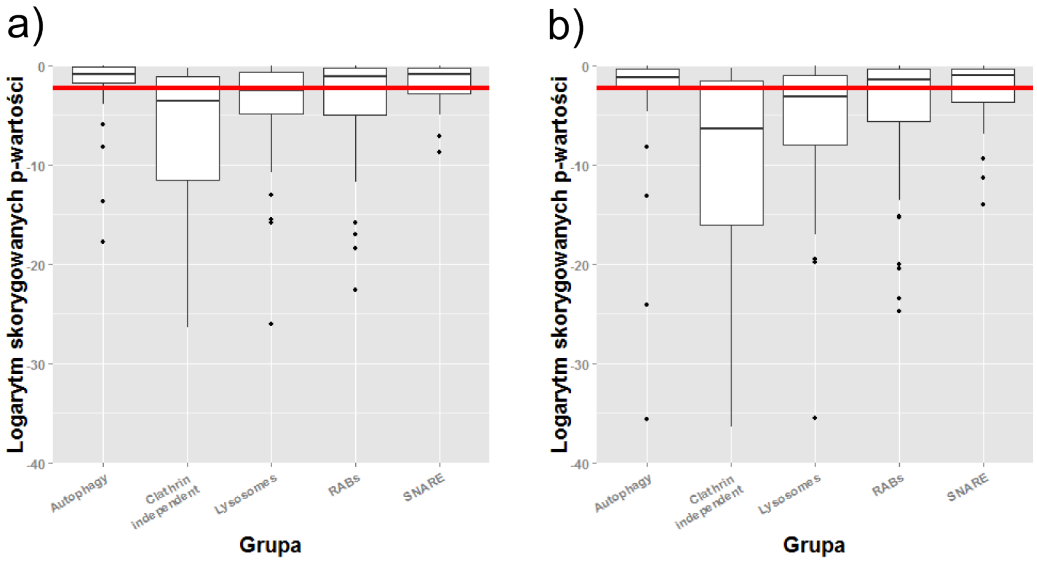
Podobny wniosek nasuwa się podczas porównywania wykresów na Rysunku 4.3: *edgeR* częściej uznaje różnicę w ekspresji genu za istotną. Ponadto można zauważyć, że zgodnie z intuicją, geny o dużych wartościach krotności zmiany (najbardziej oddalone od czerwonej prostej na wysokości zera) zazwyczaj w obu pakietach wskazywane są jako te z istotną różnicą w ekspresji. Zastosowanie na obu wykresach odmiennego sposobu normalizacji (wykres a) normalizacja za pomocą pakietu *DESeq*, b) normalizacja logarytmem z CPM[1]) wpłynęło na otrzymane różnych wartości logarytmu krotności zmiany. W przypadku wykresu dla pakietu *edgeR* pojawiło się kilka bardzo małych wartości (≤ -5), co oznacza, że średnie wartości liczby odczytów dla grupy eksperymentalnej Phdf były kilkukrotnie większe niż dla grupy iPS.



Rysunek 4.3: a) Wykres zależności logarytmu krotności zmiany od średniej znormalizowanych odczytów przy użyciu pakietu *DESeq*, b) Wykres zależności logarytmu krotności zmiany od średniej logarytmu CPM przy użyciu pakietu *edgeR*. Źródło: Opracowanie własne

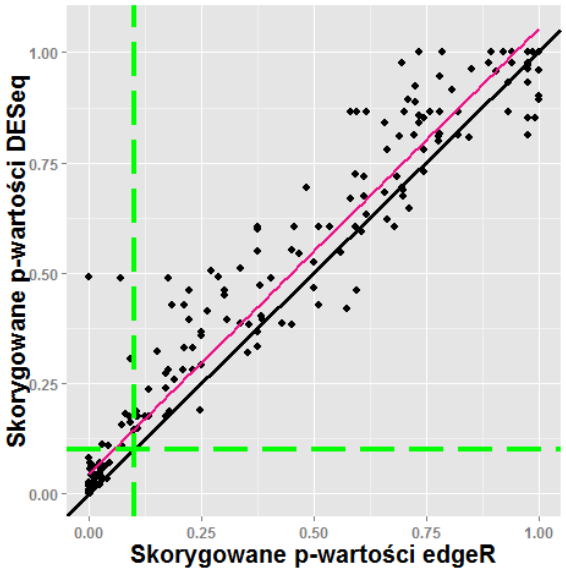
[1]CPM (ang. *Counts Per Million*)– jednostka znormalizowanej liczby odczytów genu obliczana jako iloraz liczby odczytów danego genu przez sumę odczytów wszystkich genów dla danej próbki przemnożony przez milion

Ciekawym zagadnieniem jest rozkład wartości FDR w podziale na grupy genów. Na Rysunku 4.4 czerwoną prostą został oznaczony poziom FDR wynoszący 0.1 – obserwacje leżące poniżej tej prostej wykazały wynik istotny statystycznie. Na wykresach pudełkowych wyraźnie widać, że proporcjonalnie najwięcej istotnych wyników zostało uzyskanych w grupach genów Clathrin independent oraz Lysosomes – mediana wyników jest mniejsza niż 0.1. Oznacza to, że ponad połowa genów w tych grupach charakteryzuje się istotną różnicą w ekspresji pomiędzy warunkami eksperymentalnymi. Wyniki w grupie Clathrin independent wykazują największy rozrzut, co może być spowodowane niewielką licznością grupy. Porównanie wykresów pudełkowych uzyskanych na podstawie wyników z dwóch różnych pakietów nie wskazuje na istotne różnice w rezultatach uzyskiwanych za pomocą tych pakietów.



Rysunek 4.4: Boxploty dla wartości FDR otrzymanych za pomocą pakietów a) *DESeq*, b) *edgeR*. Źródło: Opracowanie własne

Więcej informacji uzyskać można na podstawie Rysunku 4.5. Przedstawiono na nim zależność między skorygowanymi p-wartościami uzyskanymi za pomocą dwóch różnych testów statystycznych. Poziom FDR został oznaczony zielonymi prostymi. Gdyby testy dawały takie same wyniki, to punkty na wykresie leżałyby na zaznaczonej czarnej prostej. W rzeczywistości wartości FDR dla pakietu *DESeq* w wielu przypadkach są nieco większe niż te dla *edgeR*. W celu dokonania ilościowej oceny tej różnicy została dopasowana prosta uzyskana za pomocą regresji liniowej. Co ciekawe, otrzymana różowa prosta jest równoległa do czarnej prostej opisującej wyniki idealne. Można na tej podstawie wnioskować, że albo pakiet *DESeq* niedoszacowuje liczby istotnych różnic w genach, albo pakiet *edgeR* tę liczbę przeszacowuje. Punkty leżące poniżej czarnej prostej wskazują, że istnieją przypadki gdy pakiet *DESeq* wskazał mniejszą p-wartość. Dotyczy to jednak jedynie przypadków gdy uzyskany wynik i tak nie był istotny statystycznie. Wyniki wskazane jako istotne przez obie metody znajdują się w lewym dolnym kwadracie wykresu ograniczonym przez zielone proste. Zastanawiające są trzy geny, które są istotne statystycznie dla pakietu *edgeR* podczas gdy dla pakietu *DESeq* ich wartość FDR jest większa niż 0.25. Zestawienie wyników dla tych genów zostało przedstawione w Tabeli 4.4.



Rysunek 4.5: Porównanie wartości współczynników FDR otrzymanych za pomocą pakietów *DESeq* i *edgeR*. Źródło: Opracowanie własne

Tabela 4.4: Zestawienie wybranych wyników dla genów o znacznej różnicy w skorygowanych p-wartościach. Źródło: Opracowanie własne

Nazwa genu	Grupa	Średnia	Średnia A	Średnia B	FC	FDR *DESeq*	FDR *edgeR*
LAMP3	Lysosomes	68,34	12,58	93,12	7,40	0,49	0,00
RAB33A	RABs	16,56	28,44	11,28	0,40	0,49	0,07
TFEB	Lysosomes	16,04	25,98	11,62	0,45	0,3	0,09

W przypadku genu LAMP3 zmiana średniej pomiędzy warunkami eksperymentalnymi jest ponad siedmiokrotna, co wskazuje na bardzo istotną różnicę w ekspresji. Wartości dla pozostałych dwóch genów nie różnią się tak spektakularnie zatem trudno stwierdzić, dla którego z pakietów wyniki można uznać za bliższe prawdzie.

Na podstawie uzyskanych wyników nie można jednoznacznie określić, który z pakietów charakteryzuje się większą skutecznością w wychwytywaniu genów z istotną różnicą w ekspresji pomiędzy dwoma warunkami eksperymentalnymi. Nie ulega wątpliwości, że pakiet *edgeR* wskaże takich genów więcej, jednak nie wiadomo czy tym samym nie przeszacuje ich liczby. Można podejrzewać, że geny wskazane przez pakiet *DESeq* jako istotne w rzeczywistości takie są, gdyż potwierdzają to również wskazania pakietu *edgeR*, ale nie można wykluczyć, że takich genów jest więcej, a pakiet je pominął.

Podsumowanie

Celem niniejszej pracy było porównanie działania pakietów statystycznych *DESeq* i *edgeR* służących do analizy różnicowej danych z RNA-Seq. Badanie różnic w ekspresji pomiędzy próbkami odmiennymi biologicznie może być źródłem cennych informacji na temat sposobu zachowywania się komórek chociażby w ludzkim organizmie. W niniejszej pracy ukazane zostało podejście do tematu różnic w ekspresji od strony statystycznej. W pierwszym rozdziale przedstawiono ciąg przyczynowo-skutkowy prowadzący od nici DNA do pozyskiwania danych z sekwencjonowania RNA. Przybliżono również procesy wywołujące zjawisko ekspresji genów i jej regulację. Wyjaśniono zasadę działania procedur sekwencjonowania mRNA, które bazują na milionach krótkich sekwencji, które są następnie mapowane w ramach transkryptomu/genomu. Istnieje szereg wyspecjalizowanych algorytmów, które zapewniają dopasowanie wysokiej jakości.

W drugim rozdziale przedstawione zostały zasady działania pakietów *DESeq* i *edgeR*. Oba narzędzia zostały stworzone z myślą o analizie danych RNA-Seq, jednak wykorzystano w nich nieco odmienne rozwiązania. W niniejszej pracy przedstawiono podobieństwa i różnice w podejściu do normalizacji próbek, nadwyżki dyspersji czy zastosowania uogólnionych modeli liniowych do radzenia sobie z tą nadwyżką. Wyraźne różnice w zastosowanych rozwiązaniach pozwalają zrozumieć skąd biorą się rozbieżności w wynikach generowanych przez oba pakiety statystyczne.

W ostatnich dwóch rozdziałach przedstawiona została analiza przykładowych danych RNA-Seq. Najpierw za pomocą symulacji przeprowadzonej na danych rzeczywistych zostało pokazane, że pakiet *edgeR* regularnie wykazuje większe różnice w ekspresji. Sprawdzono również czy wyniki różnią się w zależności od wielkości testowanej próbki. Następnie w Rozdziale 4 kolejne etapy obliczeń za pomocą pakietów *DESeq* i *edgeR* pokazały, jak mogą różnić się od siebie otrzymywane wyniki w zależności od wykorzystanych metod analizy. Zestawienie danych rzeczywistych porównujących ekspresję w komórkach wyjściowych i komórkach macierzystych powstałych z komórek wyjściowych w procesie reprogramowania wykazało, że pakiet *edgeR* znacznie chętniej niż pakiet *DESeq* wskazuje różnicę w ekspresji między próbkami jako istotną statystycznie.

W 1995 roku zsekwencjonowano pierwszy pełen genom żywego organizmu, pałeczkę grypy (łac. *Haemophilus influenzae*) [40]. Od tego czasu mamy do czynienia z ciągłym wzrostem szybkości oraz dokładności w odczytywaniu informacji zapisanych w kodzie genetycznym. Podanie pełnej sekwencji ludzkiego genomu wymagało 15 lat współpracy szeregu grup badawczych (pod oryginalnym artykułem z 2001 roku podsumowującym Human Genome Project podpisanych jest 274 współautorów [42]). Współcześnie dostępna aparatura oferuje nieporównywalnie szybszy odczyt sekwencji. Na przykład pojedyncze urządzenie typu Illumina Genome Analyzer IIx jest w stanie wygenerować odczyty sekwencji o łącznej długości przekraczającej łączną długość sekwencji ludzkiego genomu w trakcie jednego, trwającego trzy

dni eksperymentu [4]. Technologia sekwencjonowania nadal dynamicznie się rozwija, a jego koszt spada. W związku z tym kluczowym problemem „technicznym" genomiki przestaje być pozyskiwanie surowych danych – najtrudniejszym zadaniem staje się efektywna analiza informacji. Przykładami współczesnych odpowiedzi na to wyzwanie są omawiane w niniejszej pracy pakiety statystyczne *DESeq* i *edgeR*. Wyznaczają one drogi badania różnic w poziomie ekspresji genów między próbkami. Analizy typu RNA-Seq, w których są wykorzystywane, wydają się niezwykle obiecującym krokiem w badaniu związków między informacją genetyczną zawartą w komórce a jej działaniem, zmiennością i zdolnością adaptacji. W szczególności, mogą okazać się kluczowe dla wczesnej diagnostyki i proaktywnej profilaktyki chorób nowotworowych.

Dodatek A

Podstawowe funkcje i ich parametry

W niniejszym załączniku przedstawione zostaną podstawowe funkcje pakietów DESeq i edgeR pozwalające na wykonanie analizy różnicowej genów. Będzie można zatem zobaczyć w jaki sposób stworzyć zbiór danych charakterystyczny dla danego pakietu, jak znormalizować dane, wyestymować dyspersję oraz wykonać test badający istotność różnic w ekspresji genów pomiędzy warunkami eksperymentalnymi. Przybliżenie najważniejszych parametrów funkcji pomoże zwrócić uwagę na niuanse w podejściu do analizy, ponieważ m.in. takie cechy zbioru danych jak sposób jego pozyskiwania oraz ilość posiadanych próbek, mogą mieć wpływ na otrzymywane wyniki.

DESeq

- `newCountDataSet` – funkcja tworząca z macierzy lub ramki danych obiekt klasy `CountDataSet`.
- `estimateSizeFactors` – funkcja obliczająca współczynniki wielkości próbek s_j i zapisująca je w obiekcie klasy `CountDataSet`.
- `estimateDispersions` – funkcja estymująca dyspersję na podstawie obiektu klasy `CountDataSet`. Dla każdego warunku eksperymentalnego (lub dla wszystkich warunków wspólnie w zależności od wyboru argumentu `method`) na początek obliczana jest empiryczna wartość dyspersji dla każdego genu. Następnie za pomocą regresji modelowana jest zależność pomiędzy dyspersją a średnią. Ostatnim krokiem działania funkcji jest wybór parametru dyspersji (spośród wartości empirycznej i dopasowanej krzywą regresji) wykorzystywanego w dalszych obliczeniach w zależności od argumentu `sharingMode`. Najważniejsze argumenty funkcji:
  - `method` – argument odpowiadający za wybór metody estymacji empirycznej dyspersji. Dostępne są cztery metody:
    * `pooled` – wykorzystuje wszystkie dostępne próbki z obu warunków eksperymentalnych do estymacji jednej empirycznej wartości dyspersji (ang. *pooled dispersion*) dla każdego genu. Wartość wyestymowanej dyspersji jest taka sama dla wszystkich próbek.

  * `pooled-CR` – dopasowuje model postaci zadanej przez argument `modelFormula`, następnie estymuje dyspersję za pomocą maksymalizacji skorygowanego profilu wiarogodności Coxa-Reida (CR-APL). Jest to metoda dużo wolniejsza niż `method=="pooled"`. Metoda ta jest wykorzystywana do estymacji dyspersji w pakiecie *edgeR*, jednak w przypadku pakietu *DESeq* nie jest wykorzystywana metoda ważonej maksymalizacji wiarogodności.

  * `per-condition` – empiryczna dyspersja estymowana jest oddzielnie dla każdego warunku eksperymentalnego na podstawie replikacji próbek należących do odpowiednich warunków. Dla warunków nie posiadających replikacji wykorzystywane jest maksimum wartości empirycznej dyspersji dla drugiego warunku zawierającego replikacje.

  * `blind` – ignoruje warunki z jakich pochodzą próbki i traktuje wszystkie próbki jakby pochodziły z jednego warunku eksperymentalnego, dla nich obliczając empiryczną wartość dyspersji. Metoda ta może być wykorzystywana również gdy w żadnym z warunków nie występują replikacje próbek. Metoda ta może prowadzić do obniżenia mocy testów.

- `sharingMode` – po obliczeniu empirycznej wartości dyspersji dla każdego genu, dopasowywana jest zależność między dyspersją a średnią. Działanie to ma na celu współdzielenie informacji pomiędzy genami, a w rezultacie redukcję zmienności estymowanej dyspersji. Po dopasowaniu krzywej regresji dostępne są dwie wartości dyspersji: empiryczna (pochodząca z danych o genach) i dopasowana (wartość dyspersji typowa dla genów o podobnej ekspresji). Argument `sharingMode` określa, która z dwóch powyższych wartości zostanie wykorzystana w funkcjach testujących różnice ekspresji.

  * `fit-only` – wykorzystuje jedynie wartość dopasowaną, wartość empiryczna jest wykorzystywana jedynie do estymacji krzywej regresji, następnie jest ignorowana. Użycie tego argumentu jest zalecane jedynie przy niewielkiej liczbie replikacji oraz w przypadku jeżeli użytkownikowi nie przeszkadza zwiększona liczba wyników fałszywie pozytywnych spowodowanych odstającymi wartościami dyspersji.

  * `maximum` – wykorzystuje maksimum wartości z dyspersji estymowanej i dopasowanej. Argument ten daje konserwatywne wyniki testów. Jego użycie zalecane jest przy posiadaniu co najmniej 3-4 replikacji w każdym z warunków eksperymentalnych.

  * `gene-est-only` – wykorzystuje jedynie wartość empiryczną. Metoda ta jest zalecana przy dużej liczbie replikacji oraz gdy wartość empiryczna dyspersji wydaje się być miarodajna. Jeżeli liczba replikacji jest niewielka, to wybór tego argumentu może prowadzić do wielu przypadków niedoszacowania dyspersji genów, a w konsekwencji do wzrostu wyników fałszywie pozytywnych w testowaniu ekspresji.

- `fitType` – parametr odpowiadający za metodę dopasowywania krzywej regresji przy estymacji dopasowanej wartości dyspersji. Może przyjmować dwie wartości:

  * `parametric` – za pomocą uogólnionych modeli liniowych modeluje zależność pomiędzy dyspersją a średnią

    * `local` – modeluje zależność pomiędzy dyspersją a średnią wykorzystując metodę regresji lokalnej zawartą w pakiecie `locfit` .

- `nbinomTest` – funkcja testująca różnicę w ekspresji bazowej średniej pomiędzy dwoma warunkami eksperymentalnymi. Wynikiem działania funkcji jest ramka danych zawierająca następujące kolumny:
  - `id` – nazwa genu
  - `baseMean` – bazowa średnia, czyli średnia ze wszystkich zliczeń dla danego genu podzielonych przez odpowiednie współczynniki wielkości próbek
  - `baseMeanA` – średnia ze wszystkich zliczeń dla danego genu pochodzących z warunku eksperymentalnego A podzielonych przez odpowiednie współczynniki wielkości próbek
  - `baseMeanB` – analogiczna średnia dla warunku eksperymentalnego B
  - `foldChange` – stosunek średniej dla warunku B do A (`baseMeanB`/`baseMeanA`)
  - `log2FoldChange` – logarytm o podstawie 2 z `foldChange`
  - `pval` – p-wartość dla hipotezy zerowej mówiącej, że średnia liczba odczytów dla warunku eksperymentalnego A jest równa średniej liczbie odczytów dla warunku eksperymentalnego B
  - `padj` – p-wartość skorygowana za pomocą procedury Benjaminiego-Hochberga.

edgeR

- `DGEList` – funkcja tworząca z macierzy lub ramki danych obiekt klasy `DGEList`.
- `calcNormFactors` – funkcja obliczająca współczynniki normalizacyjne próbek.
- `estimateCommonDisp` – funkcja estymująca wspólną dyspersję (ang. *common dispersion*) za pomocą warunkowego maksimum funkcji wiarogodności opisanego w [38].
- `estimateTagwiseDisp` – funkcja wykorzystuje ważoną warunkową funkcję wiarogodności do estymacji ujemnej dwumianowej dyspersji dla każdego genu. Metoda szczegółowo opisana w [37].
- `exactTest` – funkcja testująca różnicę w ekspresji bazowej średniej pomiędzy dwoma warunkami eksperymentalnymi za pomocą dokładnego testu. Wynikiem działania funkcji jest ramka danych zawierająca następujące kolumny:
  - `logFC` – logarytm o podstawie 2 z `foldChange`
  - `logCPM` – logarytm o podstawie 2 z parametru `counts-per-million` (liczba odczytów w przeliczeniu na milion odczytów)
  - `PValue` – p-wartość dla hipotezy zerowej mówiącej, że średnia liczba odczytów dla warunku eksperymentalnego A jest równa średniej liczbie odczytów dla warunku eksperymentalnego B.
- `topTags` – funkcja wyświetlająca n pierwszych obserwacji ze względu na rosnącą p-wartość lub malejący fold change. Wynikiem działania funkcji jest ramka danych taka

sama jak w przypadku funkcji `exactTest`, ale zawierająca dodatkowo kolumnę `FDR` z p-wartościami skorygowanymi domyślnie za pomocą procedury Benjaminiego-Hochberga.

Dodatek B

Symulacja – kody R

W niniejszym załączniku przedstawione zostaną kody wykorzystane do przeprowadzenia symulacji na danych rzeczywistych. W każdej z tysiąca wykonanych iteracji losowany jest przydział próbek do każdej z dwóch grup eksperymentalnych. Następnie wykonywane są kolejne etapy analizy różnic w ekspresji za pomocą pakietów DESeq i edgeR opisane w Rozdziale 2. Symulacja wykonywana jest dla zestawu 235 oraz 25 genów. Wynikiem działania są tabele zawierające wartości FDR uzyskane dla wszystkich genów i iteracji. Na końcu znajdują się kody odpowiadające za wykorzystane wykresy diagnostyczne.

```
# biblioteki
library(DESeq)
library(edgeR)
library(dplyr)
library(ggplot2)
library(reshape2)

# wczytanie danych
load(file = "phdf.Rda")
load(file = "iPS.Rda")
genes <- as.vector(read.table("https://www.dropbox.com/s/dffcgvav6l4en36/genes.txt"))

# utworzenie ramek danych
countTable <- cbind(phdf, iPS)
countTable235 <- countTable[genes, ]
countTable25 <- countTable235[1:25,]

# utworzenie macierzy wyników
pval_e235 <- matrix(0, nrow = 235, ncol = 1000)
pval_D235 <- matrix(0, nrow = 235, ncol = 1000)
pval_e25 <- matrix(0, nrow = 25, ncol = 1000)
pval_D25 <- matrix(0, nrow = 25, ncol = 1000)

fdr_e235 <- matrix(0, nrow = 235, ncol = 1000)
fdr_D235 <- matrix(0, nrow = 235, ncol = 1000)
fdr_e25 <- matrix(0, nrow = 25, ncol = 1000)
fdr_D25 <- matrix(0, nrow = 25, ncol = 1000)

group <- factor(c(rep("p", 4), rep("r", 9)))

# pętla iteracyjna
for (i in 1:1000){
```

```
  # losowanie próbek do grup eksperymentalnych
  perm235 <- t(apply(countTable235, 1, function(x){
    o <- sample(1:13, 13)
    x <- x[o]
  }))

  perm25 <- t(apply(countTable25, 1, function(x){
    o <- sample(1:13, 13)
    x <- x[o]
  }))

  # DESeq, G = 235
  cds235 <- newCountDataSet(perm235, c(rep("p", 4), rep("r", 9)))
  cds235 <- estimateSizeFactors(cds235)
  cds235 <- estimateDispersions(cds235)
  res235 <- nbinomTest(cds235, "p", "r")
  pval_D235[,i] <- res235$pval
  fdr_D235[,i] <- res235$padj
  #edgeR, G = 235
  y235 <- DGEList(counts = perm235, group = group)
  y235 <- calcNormFactors(y235)
  y235 <- estimateCommonDisp(y235)
  y235 <- estimateTagwiseDisp(y235)
  et235 <- exactTest(y235, dispersion = "common")
  et_padj235 <- topTags(et235, n = 235)
  pval_e235[,i] <- et_padj235@.Data[[1]]$PValue
  fdr_e235[,i] <- et_padj235@.Data[[1]]$FDR

  # DESeq, G = 25
  cds25 <- newCountDataSet(perm25, c(rep("p", 4), rep("r", 9)))
  cds25 <- estimateSizeFactors(cds25)
  cds25 <- estimateDispersions(cds25)
  res25 <- nbinomTest(cds25, "p", "r")
  pval_D25[,i] <- res25$pval
  fdr_D25[,i] <- res25$padj
  # edgeR, G = 25
  y25 <- DGEList(counts = perm25, group = group)
  y25 <- calcNormFactors(y25)
  y25 <- estimateCommonDisp(y25)
  y25 <- estimateTagwiseDisp(y25)
  et25 <- exactTest(y25, dispersion = "common")
  et_padj25 <- topTags(et25, n=25)
  pval_e25[,i] <- et_padj25@.Data[[1]]$PValue
  fdr_e25[,i] <- et_padj25@.Data[[1]]$FDR
}

# przekształcenie wyników na potrzeby wykresów
e25 <- melt(fdr_e25)
e25$group <- "edgeR, G=25"
pe25 <- melt(pval_e25)
pe25$group <- "edgeR, G=25"

D25 <- melt(fdr_D25)
D25$group <- "DESeq, G=25"
pD25 <- melt(pval_D25)
pD25$group <- "DESeq, G=25"

e235 <- melt(fdr_e235)
e235$group <- "edgeR, G=235"
pe235 <- melt(pval_e235)
pe235$group <- "edgeR, G=235"
```

```
D235 <- melt(fdr_D235)
D235$group <- "DESeq, G=235"
pD235 <- melt(pval_D235)
pD235$group <- "DESeq, G=235"
tot25 <- rbind(D25, e25)
tot235 <- rbind(D235, e235)

ptot25 <- rbind(pD25, pe25)
ptot235 <- rbind(pD235, pe235)

# Rysunek 3.1
ggplot(ptot25, aes(x=value)) + geom_histogram(binwidth = 0.01) +
  facet_wrap(~group) +
  xlab("Wartość p-value") +
  ylab("Liczba wyników") +
  theme(axis.text = element_text(size = 20, face = "bold"),
        axis.title = element_text(size = 25, face = "bold"),
        strip.text.x = element_text(size = 25, face="bold"))

ggplot(tot25, aes(x=value)) + geom_histogram(binwidth = 0.01) +
  facet_wrap(~group) +
  xlab("Wartość FDR") +
  ylab("Liczba wyników") +
  theme(axis.text = element_text(size = 20, face = "bold"),
        axis.title = element_text(size = 25, face = "bold"),
        strip.text.x = element_text(size = 25, face="bold"))

# Rysunek 3.2
ggplot(ptot235, aes(x=value)) + geom_histogram(binwidth = 0.01) +
  facet_wrap(~group) +
  xlab("Wartość p-value") +
  ylab("Liczba wyników") +
  theme(axis.text = element_text(size = 20, face = "bold"),
        axis.title = element_text(size = 25, face = "bold"),
        strip.text.x = element_text(size = 25, face="bold"))

ggplot(tot235, aes(x=value)) + geom_histogram(binwidth = 0.01) +
  facet_wrap(~group) +
  xlab("Wartość FDR") +
  ylab("Liczba wyników") +
  theme(axis.text = element_text(size = 20, face = "bold"),
        axis.title = element_text(size = 25, face = "bold"),
        strip.text.x = element_text(size = 25, face="bold"))

# Liczba istotnych wyników w każdej z iteracji
apply(fdr_D25[,1:10], 2, function(x){length(which(x<0.1))})
apply(fdr_e25[,1:10], 2, function(x){length(which(x<0.1))})
apply(fdr_D235[,1:10], 2, function(x){length(which(x<0.1))})
apply(fdr_e235[,1:10], 2, function(x){length(which(x<0.1))})

# Maksymalna liczba istotnych wyników w iteracji
max(apply(fdr_D25, 2, function(x){length(which(x<0.1))}))
max(apply(fdr_e25, 2, function(x){length(which(x<0.1))}))
max(apply(fdr_D235, 2, function(x){length(which(x<0.1))}))
max(apply(fdr_e235, 2, function(x){length(which(x<0.1))}))

# Średnia ze wszystkich wartości FDR
mean(fdr_D25)
mean(fdr_e25)
mean(fdr_D235)
```

```
mean(fdr_e235)

# Średnia wartość FDR dla każdej iteracji symulacji
lD235 <- log(fdr_D235)
le235 <- log(fdr_e235)
lD25 <- log(fdr_D25)
le25 <- log(fdr_e25)
mean_D235 <- as.data.frame(lD235) %>% summarize_each(funs(mean))
mean_e235 <- as.data.frame(le235) %>% summarize_each(funs(mean))
mean_D25 <- as.data.frame(lD25) %>% summarize_each(funs(mean))
mean_e25 <- as.data.frame(le25) %>% summarize_each(funs(mean))

mean_all <- as.data.frame(t(rbind(1:1000, mean_D235, mean_e235, mean_D25, mean_e25)))
colnames(mean_all) <- c("Lp", "mean_D235", "mean_e235", "mean_D25", "mean_e25")

#Rysunek 3.3 a)
ggplot(mean_all, aes(x=exp(mean_e25), y=exp(mean_D25))) + geom_point(size=3) +
  geom_smooth(se = FALSE, method = "lm", size = 1, col = "deeppink2") +
  geom_abline(intercept = 0, slope = 1, size = 1.25) +
  xlim(0.4, 1) + ylim(0.4, 1) +
  xlab("Średnia skorygowanych p-wartości edgeR") +
  ylab("Średnia skorygowanych p-wartości DESeq") +
  theme(axis.text = element_text(size = 12, face = "bold"),
        axis.title = element_text(size = 20, face = "bold"))

#Rysunek 3.3 b)
ggplot(mean_all, aes(x=exp(mean_e235), y=exp(mean_D235))) + geom_point(size=3) +
  geom_smooth(se = FALSE, method = "lm", size = 1, col = "deeppink2") +
  geom_abline(intercept = 0, slope = 1, size = 1.25) +
  xlim(0.4, 1) + ylim(0.4, 1) +
  xlab("Średnia skorygowanych p-wartości edgeR") +
  ylab("Średnia skorygowanych p-wartości DESeq") +
  theme(axis.text = element_text(size = 12, face = "bold"),
        axis.title = element_text(size = 20, face = "bold"))
```

Dodatek C

Analiza danych rzeczywistych – kody R

W niniejszym załączniku przedstawione zostaną kody wykorzystane do przeprowadzenia analizy danych rzeczywistych. Kolejne etapy analizy różnic w ekspresji za pomocą pakietów DESeq i edgeR wykonywane są w takiej kolejności, jak opisano w Rozdziale 2.

```
# biblioteki
library(DESeq)
library(edgeR)
library(ggplot2)
library(reshape2)

# wczytanie danych
load(file = "phdf.Rda")
load(file = "iPS.Rda")
genes <- as.vector(read.table("https://www.dropbox.com/s/dffcgvav6l4en36/genes.txt"))

countTable <- cbind(phdf, iPS)

# DESeq
cds <- newCountDataSet(countTable, c(rep("Phdf", 4), rep("iPS", 9)))
cds <- estimateSizeFactors(cds)
cds <- estimateDispersions(cds)

# Rysunek 4.1
plotDispEsts(cds, xlab = "Średnia znormalizowanych zliczeń", ylab = "",
             cex.lab = 1.8, cex.axis = 1.3)
title(ylab = "Dyspersja", line = 2.5, cex.lab = 1.8)

# wybranie interesujących genów do analizy
cds <- cds[genes, ]
countTable2 <- countTable[selected, ]

# dodanie nazw genow
gene_name <- as.vector(read.table("https://www.dropbox.com/s/finpwpg4tzx10se/gene_names.txt"))
rownames(countTable2) <- rownames(cds) <- gene_name

res <- nbinomTest(cds, "Phdf", "iPS")

# nadanie nazw grupom genow
res$group <- c(rep("Autophagy", 30), rep("Clathrin \n independent", 14),
```

```
              rep("Lysosomes", 94), rep("RABs", 61), rep("SNARE", 36))

# Rysunek 4.3.a)
plotMA2(res, ylab = "", xlab = "Średnia znormalizowanych odczytów",
        cex.lab = 1.8, cex = 0.9, cex.axis = 1.3, panel.first = grid(),
        col = ifelse(res$padj >= 0.1, "black", "red"))
title(ylab = "Log krotności zmiany", line = 2.5, cex.lab = 1.8)


# edgeR
group <- factor(c(rep("Phdf", 4), rep("iPS", 9)))
y <- DGEList(counts = countTable, group = group)
y <- calcNormFactors(y)
y <- estimateCommonDisp(y)
y <- estimateTagwiseDisp(y)

# Rysunek 4.2
plotMeanVar( y , show.raw.vars = TRUE, show.tagwise.vars = TRUE,
             show.binned.common.disp.vars = FALSE, show.ave.raw.vars = FALSE,
             dispersion.method = "qcml" , NBline = TRUE, nbins = 100,
             pch = 16, xlab = expression(Log[10]~średniej~ekspresji),
             ylab = "", cex.lab = 1.8, cex.axis = 1.3)
title(ylab = expression(Log[10]~wariancji), line = 2.5, cex.lab = 1.8)

# wybranie interesujących genów do analizy
y <- y[selected, ]
rownames(y) <- gene_name

et <- exactTest(y, dispersion = "common")
summary(de <- decideTestsDGE(et, p = 0.1, adjust = "BH"))
detags <- rownames(y)[as.logical(de)]

# Rysunek 4.3.b)
plotSmear(et, de.tags = detags, ylab = "",
          xlab = "Średnia logarytmu CPM", cex.lab = 1.8,
          cex.axis = 1.3, cex = 0.9)
title(ylab = "Log krotności zmiany", line = 2.5, cex.lab = 1.8)
abline(h = 0, col = "#ff000080", lwd = 3)

et_padj <- topTags(et, n = 235)


# połączenie wyników uzyskanych z DESeq i edgeR
total <- merge(res, et_padj, by.x = "id", by.y = 0)
colnames(total) <- c("Gene", "baseMean", "baseMeanA", "baseMeanB",
                     "foldChange",   "log2FC", "pval.D", "padj.D",
                     "group", "logFC", "logCPM", "pval.e", "padj.e")
total2 <- total[ , c(1,9,2,3,4,5,6,10,11,7,12,8,13)]
total3 <- cbind(total2[-174,1:2], round(total2[-174,-c(1,2)], 3))


# Rysunek 4.4.a)
ggplot(total2, aes(group, log(padj.D))) + geom_boxplot() +
        geom_hline(yintercept = -2.302585, col = "red", size = 1.5) +
        coord_cartesian(ylim = c(-40, 0)) + xlab("Grupa") +
  ylab("Logarytm skorygowanych p-wartości") +
        theme(axis.text.x =  element_text(angle = 30, hjust = 1),
            axis.text = element_text(size = 12, face = "bold"),
            axis.title = element_text(size = 20, face = "bold"))

# Rysunek 4.4.b)
```

```
ggplot(total2, aes(group, log(padj.e))) + geom_boxplot() +
        geom_hline(yintercept = -2.302585, col = "red", size = 1.5) +
        coord_cartesian(ylim = c(-40, 0)) + xlab("Grupa") +
  ylab("Logarytm skorygowanych p-wartości") +
        theme(axis.text.x = element_text(angle = 30, hjust = 1),
            axis.text = element_text(size = 12, face = "bold"),
            axis.title = element_text(size = 20, face = "bold"))

# Rysunek 4.5
ggplot(total2, aes(padj.e, padj.D)) + geom_point(size = 3) +
        geom_smooth(se = FALSE, method = "lm", size = 1, col = "deeppink2") +
        geom_abline(intercept = 0, slope = 1, size = 1.25) +
        geom_hline(yintercept = 0.1, colour = "green",
             linetype = "longdash", size = 1.5) +
        geom_vline(xintercept = 0.1, colour = "green",
             linetype = "longdash", size = 1.5) +
        xlab("Skorygowane p-wartości edgeR") +
  ylab("Skorygowane p-wartości DESeq") +
        theme(axis.text = element_text(size = 12, face = "bold"),
            axis.title = element_text(size = 20, face = "bold"))
```

Literatura

[1] S. Anders, W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[2] S. Anders, W. Huber. Differential expression analysis for sequence count data - Supplement. *Genome Biology*, 11:R106, 2010.

[3] S. Anders, W. Huber. Differential expression of RNA-Seq data at the gene level - the DESeq package. *Genome Biology*, 2010.

[4] M.W. Anderson, I. Schrijver. Next generation DNA sequencing and the future of genomic medicine. *Genes*, 1(1):38–69, 2010.

[5] Y. Benjamini, Y. Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

[6] P. Biecek. Testowanie zbioru hipotez z zadaną relacją hierarchii wraz z przykładami zastosowań w genetyce. *Rozprawa doktorska*, 2007.

[7] T.A. Brown. *Genomy*. Wydawnictwo Naukowe PWN, 2001.

[8] Y. Chen, A.T.L. Lun, G.K. Smyth. Differential expression analysis of complex RNA-seq experiments using edgeR. *Statistical analysis of next generation sequencing data*, strony 51–74. Springer, 2014.

[9] Y. Chen, D. McCarthy, M.D. Robinson, G.K. Smyth. edgeR: differential expression analysis of digital gene expression data. User's Guide. 2008.

[10] F.H.C. Crick. On protein synthesis. *Symp Soc Exp Biol*, 12(138-63):8, 1958.

[11] Z. Dai, J.M. Sheridan, L.J. Gearing, D.L. Moore, S. Su, S. Wormald, S. Wilcox, L. O'Connor, R.A. Dickins, M.E. Blewitt, et al. edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens. *F1000Research*, 3, 2014.

[12] P. De Wit, M.H. Pespeni, J.T. Ladner, D.J. Barshis, F. Seneca, H. Jaris, N. Overgaard Therkildsen, M. Morikawa, S.R. Palumbi. The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, 12(6):1058–1067, 2012.

[13] A.J. Dobson. *An Introduction to Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2010.

[14] R.A. Fisher. On the interpretation of chi-squares from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85:87–94, 1922.

[15] A. Ghosh, M. Bansa. A glossary of DNA structures from A to Z. *Acta Crystallographica Section D*, D59:620–626, 2003.

[16] S.G. Gregory, et al. The DNA sequence and biological annotation of human chromosome 1. *Nature*, 441:315–321, 2006.

[17] A.J.F. Griffiths, J.H. Miller, D.T. Suzuki, R.C. Lewontin, W.M. Gelbart. Population genetics. 2000.

[18] J. Guhaniyogi, G. Brewer. Regulation of mRNA stability in mammalian cells. *Gene*, 265(1):11–23, 2001.

[19] P.G. Higgs. RNA secondary structure: physical and computational aspects. *Quarterly reviews of biophysics*, 33(03):199–253, 2000.

[20] P.G. Higgs, T.K. Attwood, K. Murzyn, P. Liguziński, M. Kurdziel. *Bioinformatyka i ewolucja molekularna*. Informatyka - Zastosowania. Wydawnictwo Naukowe PWN, 2008.

[21] J.M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2011.

[22] L. Hood, D. Galas. The digital code of DNA. *Nature*, 421(6921):444–448, 2003.

[23] http://onkologia.org.pl/nowotwory-zlosliwe-ogolem 2/. Nowotwory złośliwe ogółem, dostęp na dzień: 05.09.2015.

[24] International Human Genome Sequencing Consortium, et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[25] E.S. Lander, L.M. Linton, B. Birren, Ch. 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*, 409(6822):860–921, 2001.

[26] D. Liu, J.H. Graber. Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation. *BMC bioinformatics*, 7(1):1, 2006.

[27] C. Loader. *Local Regression and Likelihood*. Statistics and Computing. Springer New York, 1999.

[28] A.T.L. Lun, Y. Chen, G.K. Smyth. It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. *Statistical Genomics: Methods and Protocols*, strony 391–416, 2016.

[29] N. Maherali, et al. Directly Reprogrammed Fibroblasts Show Global Epigenetic Remodeling and Widespread Tissue Contribution. *Cell Stem Cell*, 1:55–70, 2007.

[30] D.J. McCarthy, Y. Chen, G.K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40, 2012.

[31] Y. Miki, et al. A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1. *Science*, 266:66–71, 1994.

[32] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.

[33] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, M. Snyder. The transciptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320:1344–1349, 2008.

[34] A. Revyakin, Ch. Liu, R.H. Ebright, T.R. Strick. Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science*, 314(5802):1139–1143, 2006.

[35] M.D. Robinson, D.J. McCarthy, G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[36] M.D. Robinson, A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11:R25, 2010.

[37] M.D. Robinson, G.K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23:2881–2887, 2007.

[38] M.D. Robinson, G.K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9:321–322, 2008.

[39] J.H. Rogers. The role of introns in evolution. *FEBS letters*, 268(2):339–343, 1990.

[40] J. Soh, P.M.K. Gordon, C.W. Sensen. *Genome Annotation*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press, 2016.

[41] C. Trapnell, A. Roberts, L. Goff, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, 2012.

[42] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, Ch.A. Evans, R.A. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[43] Z. Wang, M. Gerstein, M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[44] J.D. Watson, F.H.C. Crick. A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738, 1953.

[45] M.H. Werner, A.M. Groenenborn, G.M. Clore. Intercalation, DNA kinking, and the control of transcription. *Science*, 271(5250):778, 1996.

[46] B.T. Wilhelm, J.R. Landry. RNA-Seq–quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48(3):249–257, 2009.

[47] P.C. Winter, G.I. Hickey, H.L. Fletcher, W. Prus-Głowacki, E. Chudzińska, Wydawnictwo Naukowe PWN. *Genetyka*. Krótkie Wykłady. Wydawnictwo Naukowe PWN, 2010.

[48] X. Zhou, H. Lindsay, M.D. Robinson. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic acids research*, 42(11):e91–e91, 2014.

Barbara Sozańska
Nr albumu 265178

Warszawa, 15 września 2016

Oświadczenie

Oświadczam, że pracę magisterską pod tytułem „Porównanie statystycznych metod różnicowej analizy danych RNA-Seq na przykładzie pakietów *DESeq* i *edgeR*", której promotorem jest dr hab. inż. Przemysław Biecek, prof. nzw. wykonałam samodzielnie, co poświadczam własnoręcznym podpisem.

.....
Barbara Sozańska