Silesian University of Technology

Faculty of Automatic Control, Electronics and Computer Science

Advanced data exploration techniques for augmented transcriptional landscape and its better quantification

PhD Thesis

Author: Agata Muszyńska
Supervisor: dr hab. inż. Paweł Łabaj
Co-supervisor: Assoc. Prof. Dr. David Kreil

Gliwice, February 2023

Zaawansowane techniki analizy danych ekspresyjnych dla lepszego zrozumienia transkryptomu i poprawy jego oceny ilościowej

Wydział Automatyki, Elektroniki i Informatyki
Politechnika Śląska

Streszczenie rozprawy doktorskiej
Autor: Agata Muszyńska
Promotor: dr hab. inż. Paweł Łabaj
Kopromotor: Assoc. Prof. Dr. David Kreil

W ostatnich latach badacze stają się coraz bardziej świadomi kryzysu powtarzalności, z jakim boryka się całe środowisko naukowe. Pomimo dostarczonych kompleksowych testów porównawczych i wytycznych, problem ten jest również ważnym zagadnieniem w przypadku analizy danych RNA-seq. Szybki rozwój technologii idzie w parze z napływem nowych narzędzi i podejść analitycznych, który przewyższa rozwój najlepszych praktyk w tym obszarze. Powodem tego jest bardzo często brak stosowania wskazówek w praktyce, a także korzystanie z narzędzi, które są przestarzałe, ale po prostu dobrze znane. Wyzwania powtarzalności mogą także wynikać z niedopracowanego projektu badawczego oraz po prostu z charakterystyki danych pomiarowych, takich, jak niskie poziomy sygnału czy czynniki zaburzające wyniki. Nawet przy starannej aplikacji najnowocześniejszych metod normalizacji, czynniki zakłócające nie zawsze mogą zostać zidentyfikowane i usunięte w celu przeprowadzenia sensownej analizy ilościowej. Moja praca pokazuje jednak, że mimo słabych wyników w analizie danych ilościowych, wciąż możemy uzyskać solidne i wartościowe wyniki jakościowe.

Innym, bardziej złożonym powodem braku powtarzalności w analizie RNA-seq jest złożoność transkryptomu. Chociaż genom ludzki jest najbardziej zbadany i kompletny, wciąż publikowane są nowe aktualizacje adnotacji, różniące się liczbą genów i transkryptów. Bliższe przyjrzenie się modelom referencyjnym dla człowieka i myszy ujawnia, że mimo iż oba zawierają zbliżoną liczbę genów, liczba transkryptów myszy jest mniejsza. Ta dysproporcja może wskazywać na niekompletne informacje dotyczące transkryptomu myszy, co może prowadzić do mylących dopasowań i tym samym problemów z powtarzalnością w miarę ewolucji adnotacji genów. Dlatego przy analizie danych RNA-Seq istotne jest również uwzględnienie składnika jakościowego w poszukiwaniu nowych alternatywnie złożonych transkryptów, które rozszerzają modele referencyjne genów, aby przyszłe eksperymenty mogły lepiej oddawać oszacowania profilu ekspresji obserwowanych transkryptów.

Celem niniejszej pracy jest zbadanie najlepszych praktyk w analizie danych RNA-seq oraz opracowanie podejścia mającego na celu poprawę powtarzalności i stabilności wyników. Rozwiązanie przedstawione tutaj to starannie zaprojektowany, kompleksowy proces analizy danych RNA-Seq, obejmujący analizę ilościową i jakościową, zweryfikowany na niezależnych zestawach danych zebranych w nowych eksperymentach. To, co wyróżnia to rozwiązanie, to integracja wielu etapów analizy danych RNA-seq, w przeciwieństwie do istniejących prac, które skupiają się na poszczególnych modułach (takich jak dopasowanie sekwencji). Jednocześnie moja metoda dostarcza różnorodnych możliwości wyboru dla każdego z kluczowych etapów analizy, obejmując kontrolę jakości, wstępną obróbkę, dopasowanie, kwantyfikację, różnicową ekspresję genów (DGE) oraz analizę wzbogacania zbiorów genów (GSEA), umożliwiając jednocześnie łatwą implementację alternatywnych opcji. Ponadto, wprowadza ona nowatorski element kompleksowej analizy jakościowej danych RNA-Seq.

Zaprezentowana analiza alternatywnego splicingu (AS) nie ogranicza się tylko do wykrywania zdarzeń alternatywnego splicingu, ale istotnie zapewnia również przegląd konsekwencji na poziomie transkryptów i białek, co jest niezbędne do funkcjonalnej interpretacji. Tego elementu brakuje w standardowych rozwiązaniach przetwarzania potokowego; jednak jest on kluczowy dla lepszego zrozumienia i adnotacji modeli genów. Zaprezentowane tu podejście wypełnia tę lukę, dostarczając kompleksowy system przetwarzania potokowego - od surowych odczytów, przez dopasowanie, analizę genów różnicujących i AS, aż po analizę skutków na poziomie białek. Doprowadziło to do potwierdzania przypuszczenia, że geny myszy są słabo zanotowane, ale również przyniosło nowe informacje, świadczące o tym, że dotyczy to zwłaszcza genów istotnych dla układu nerwowego. Zidentyfikowane nowe zdarzenia alternatywnego splicingu prowadzą do odkrycia nieznanych transkryptów, co wpływa na nowe funkcje w analizie GSEA. Ponadto stwierdziłam, że konkretne profile funkcjonalne związane są z różnymi rodzajami zdarzeń splicingu, co stanowi według mojej wiedzy pierwsze takie doniesienie. Co ciekawe, pomimo silnych efektów tzw. ‘paczki’ w analizie ilościowej, wyniki wykrywania zdarzeń alternatywnego splicingu były bardzo stabilne.

Chociaż główny nacisk w mojej pracy badawczej skupiał się na analizie danych RNA-seq, profilowanie ekspresji za pomocą mikromacierzy wciąż jest szeroko stosowane, istnieją również ogromne repozytoria publiczne i prywatne zawierające dane mikromacierzowe. Bazując na obserwacjach dotyczących uzupełniającego charakteru tych technologii o dużej przepustowości, mogłam zademonstrować, jak techniki analizy danych opracowane dla jednej technologii mogą być dostosowane do nowego kontekstu innej technologii, uwzględniając specyfikę źródła danych.

Praca przedstawia wyniki dla trzech różnych zbiorów danych: 'rzeczywistego' przykładu danych sekwencjonowania nowej generacji (NGS), danych NGS używanych do testowania oraz danych mikromacierzowych. Główny zbiór danych składał się z 88 próbek pobranych z części grzbietowej odcinka lędźwiowego rdzenia kręgowego myszy c57/BL6. Badanie profilowania transkryptomów na poziomie genomowym (RNA-seq) zostało przeprowadzone na trzech seriach kontrolnych (WTP) oraz trzech typach myszy z unieczynnieniem poszczególnych genów (knockout gene). W badaniu wykorzystano kilka linii myszy z warunkowym usunięciem receptorów opioidowych mu (MOR) i delta (DOR) oraz proenkafaliny (PENK) w określonych strukturach mózgu. Dla każdej grupy istniała podgrupa z wywołanym bólem neuropatycznym (PNSL) oraz odpowiednia grupa kontrolna, w której przeprowadzono operację pozorowaną (SHAM). Dla każdej grupy istniały cztery powtórzenia biologiczne. Dane te opisane są jako 'dane rzeczywiste', ponieważ nie były one przeznaczone jako zestaw do testów porównawczych, ale raczej jako sposób na znalezienie celów leczenia bólu neuropatycznego. W związku z tym niosły ze sobą pewne problemy 'rzeczywistego świata', które potencjalnie mogły wpływać na analizę. Mianowicie niskie wartości sygnału i złożone efekty grupowe.

Zbiór danych referencyjnych, pochodzących z eksperymentów *benchamarkingowych*, składał się z próbek RNA A i B pochodzących z konsorcjum SEQC2, gdzie A to mieszanina 10 różnych linii komórek nowotworowych, a B to zdrowa osoba. Próbki A i B zostały zmieszane w różnych proporcjach, co umożliwiło walidację wyników opartych na tytrowaniu. W ramach projektu próbki były badane za pomocą wielu paneli komercyjnych i niestandardowych. W części pracy przedstawionej tutaj wykorzystaliśmy dane uzyskane z wykorzystaniem następujących paneli docelowych:

- Panel komercyjny Agilent (A1) - panel komercyjny ukierunkowany na 1064 geny,
- Panel niestandardowy Agilent (A2) - projekt panelu wykonany przez konsorcjum SEQC, łączący różne cele z paneli komercyjnych (np. A1) + znane onkogeny, ukierunkowany na 2125 genów,

• Panel niestandardowy Roche (R1) - panel zaprojektowany przez Roche ukierunkowany na te same regiony genomowe co A2.

Każdą próbkę ukierunkowano tymi panelami, a następnie utworzono 4 niezależne biblioteki do sekwencjonowania krótkich odczytów przez Illuminę. Ponieważ był to zestaw danych zaprojektowany do testów porównawczych konsorcjum SEQC, jest dobrze opisany, a sygnał jest silny.

Dane mikromacierzowe uzyskano od siedmiu pacjentów cierpiących na chorobę Parkinsona oraz od siedmiu zdrowych wolontariuszy. Analiza została przeprowadzona przy użyciu mikromacierzy Illumina HumanHT-12 v4. Ponieważ mikromacierze te projektowane są dla określonych transkryptów, cała analiza została przeprowadzona na tym poziomie. Uzyskano czternaście próbek, ale ponieważ mikromacierze te składają się z dwunastu torów, dwie próbki zostały przeanalizowane na innej mikromacierzy (Zdrowy 6 i 7).

Każdy ze zbiorów danych dostarczył nowych spostrzeżeń i ulepszeń do budowy narzędzia przetwarzania potokowego, a jednocześnie uzasadnił wybór metod. Wszystkie potwierdziły, że nie ma rozwiązania uniwersalnego zarówno pod względem metody analizy danych, jak i wyboru technik laboratoryjnych. Dlatego bardzo korzystne jest dostarczanie różnych opcji w narzędziach analizy danych o wysokiej przepustowości, aby można było wybrać ścieżkę najlepiej dostosowaną do celu eksperymentu. Jest to szczególnie ważne podczas radzenia sobie z scenariuszami rzeczywistymi, gdy sygnał może być słaby, a rzeczywiste zakłócenia w danych nieznane.

Potok przetwarzania wykorzystuje system zarządzania przepływem pracy Snakemake do zebrania poleceń używanych dla różnych części analizy i automatycznego uruchamiania każdego etapu dla wybranych próbek. Jest szybszy niż zwykłe użycie basha, zapewnia lepszą kontrolę nad przepływem pracy i posiada zestaw dodatkowych zalet. Pod względem elastyczności i analizy eksploracyjnej jest lepszy niż przepływy pracy takie jak Galaxy, które są doskonałym narzędziem dla użytkowników chcących zautomatyzować pewne rutynowe analizy, ale mających niewielką wiedzę z zakresu nauk komputerowych. Prawie wszystkie zależności przepływu pracy są instalowane za pomocą pakietu i systemu zarządzania środowiskiem Conda. Ze wszystkimi wymaganiami zdefiniowanymi w pliku YAML, Conda automatycznie buduje nowe środowisko, rozwiązuje konflikty i pobiera wszystkie zależności. Dodatkowo przepływ pracy składa się z skryptu R Markdown do analizy danych mikromacierzowych oraz zestawu skryptów R Markdown do analizy danych RNA-seq. Ogólny schemat potoku przetwarzania danych przedstawiony jest na Figurze 1.

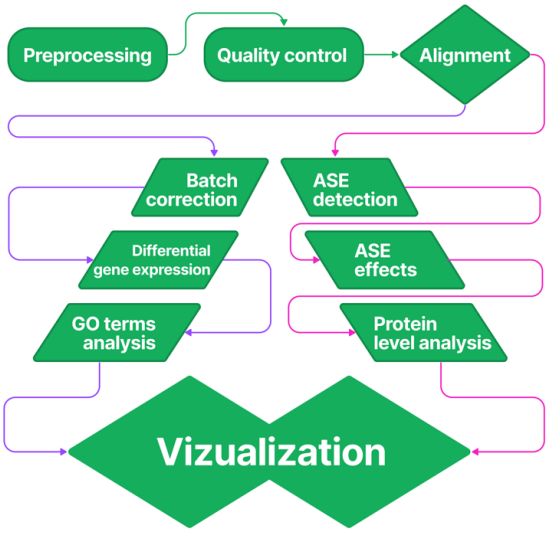

Figura 1: Schemat potoku przetwarzania danych.

Analiza danych mikromacierzowych w skrypcie R Markdown składa się z następujących kroków:

• Przetwarzanie wstępne

Akceptowane są pliki w formacie IDAT, które zawierają zsumowane intensywności dla każdego typu sondy na macierzy. Pierwsze podejście obejmuje wykorzystanie pliku BGX dostarczonego przez Illumina jako adnotacji oraz normalizację NEQC zaimplementowaną w pakiecie Limma. Druga i trzecia opcja wykorzystują pakiet illuminaHumanv4.db jako adnotację oraz normalizację VSN lub kwantylową dostarczaną przez pakiet beadarray.

• Kontrola jakości

Dostępne są wykresy MA, wykresy gęstości oraz wykresy pudełkowe do sprawdzania jakości danych.

• Korekcja czynników zakłócających

Aby uwzględnić czynniki zakłócające, użyto algorytmu SVA dla wszystkich trzech zestawów znormalizowanych danych.

• Analiza różnicowej ekspresji genów

Do wykrywania DEGs użyto pakietu Limma.

• Analiza terminów GO

Użyto algorytmu parentchild z testem Fishera oraz wartości odcięcia p-value 1%.

• Wizualizacja

Możliwe wizualizacje obejmują heatmapy, wykresy PCA, diagramy Venna i wykresy słupkowe dla najlepszych terminów GO.

Analiza danych RNA-seq jest podzielona na kilka skryptów R Markdown oraz skrypt Snakemake. Ponieważ pliki FASTQ zazwyczaj zajmują dużo miejsca na dysku, akceptowalne są pliki o trzech rozszerzeniach dla różnych metod kompresji. Istnieje możliwość dostarczenia plików nieskompresowanych, plików fastq.gz oraz plików fastq.dsrc.

Przepływ pracy składa się z następujących kroków:

• Kontrola jakości

Pierwsza kontrola jakości plików FASTQ jest wykonywana za pomocą narzędzia FastQC, następnie generowany jest również raport jakości dla procesu dopasowania za pomocą narzędzia MultiQC.

• Dopasowanie i kwantyfikacja

Dostępna jest opcja dopasowania do genomu za pomocą narzędzia HiSat2 lub pseudodopasowanie do transkryptomu za pomocą narzędzia Kallisto. Każdy wariant zawiera budowę indeksu (jeśli jest to konieczne). Wszystkie pliki pośrednie, które nie są potrzebne do dalszej analizy (nieskompresowane pliki FASTQ, pliki SAM, nieposortowane pliki BAM) są plikami tymczasowymi, usuwanymi po zakończeniu zadania. Opcja z narzędziem HiSat2 zawiera także kwantyfikację za pomocą StringTie2 oraz przygotowanie plików wejściowych do dalszej analizy w środowisku R.

• Wykrywanie alternatywnego splicingu za pomocą Spladder

Spladder przeprowadza analizę alternatywnego splicingu na plikach BAM uzyskanych z etapu dopasowania do genomu.

• Analiza konsekwencji na poziomie białek za pomocą Bisbee

Pliki wyjściowe Spladdera przygotowane są do analizy Bisbee, a następnie Bisbee raportuje efekty, peptydy oraz pliki FASTA z zmienionymi transkryptami dla wszystkich 6 ASE.

• Wspólna analiza Bisbee i Spladder

Narzędzie automatycznie uruchamia kolejny skrypt R Markdown do analizy wyników obu programów i dostarcza raport w formacie pdf, pliki csv i txt z wynikami dla interesujących zdarzeń oraz powiązanych terminów GO, a także pliki używane w kolejnym kroku przez InterProScan.

• Analiza konsekwencji na poziomie białek za pomocą InterProScan

Pliki FASTA z poprzedniego kroku są przeszukiwane pod kątem interesujących zdarzeń i dostarczane do InterProscan, aby uzyskać informacje o domenach białkowych.

• Wizualizacja

Również w formie skryptu R Markdown dostępna jest wizualizacja zmian wprowadzonych przez nowe zdarzenie.

Alternatywnie, po dopasowaniu można przeprowadzić analizę różnicowej ekspresji genów. Skrypt R Markdown, który wykonuje ten etap, składa się z następujących kroków:

• Przetwarzanie wstępne

Używane jest przetwarzanie DESeq2 lub edgeR TMM. Usuwane są również geny/transkrypty z niską liczbą zmapowanych odczytów.

• Korekta czynników zakłócających

Aby uwzględnić czynniki zakłócające, użyto algorytmu SVAseq dla wszystkich trzech zestawów znormalizowanych danych.

• Analiza różnicowej ekspresji genów

Dla analizy DGE dostępne są trzy opcje: limma, edgeR oraz DESeq2.

• Analiza terminów GO

Wykorzystano algorytm Parentchild z testem Fishera z p-value poniżej 1%.

• Wizualizacja

Możliwe wizualizacje obejmują heatmapy, wykresy skrzypcowe, wykresy PCA oraz diagramy Venna.

Dla zbioru danych dotyczących bólu neuropatycznego zostały uruchomione wszystkie etapy potoku przetwarzania. Analiza różnicowej ekspresji opiera się w dużej mierze na podejściu opracowanym dla danych mikromacierzowych. Celem było zbadanie, jak połączenie różnych wytycznych, opartych głównie na sztucznie stworzonych zestawach danych z silnymi sygnałami, może poprawić analizę problematycznego zbioru danych. Ponieważ projekt badania dotyczącego bólu neuropatycznego był dość złożony, moim celem było przeanalizowanie prostej różnicy między grupą kontrolną (WTP_SHAM) a grupą z wywołanym bólem neuropatycznym (WTP_PNSL) dla każdej serii. Celem było porównanie list genów różnicowo ekspresjonowanych uzyskanych dla każdej z 3 serii i kontynuowanie z metodą, która daje najlepsze wyniki pod względem powtarzalności. W pierwszej próbie większość narzędzi zastosowanych do wykrywania DEG (Limma, EdgeR, DeSeq2) nie dostarczyła żadnych wyników. Tylko Limma wykazał do 33 DEG, w zależności od serii, i dlatego też metoda ta została wybrana do dalszej analizy.

Ponieważ dane pochodzą z trzech serii i od różnych myszy, było oczywiste, że zostały one dotknięte zarówno znanymi, jak i ukrytymi czynnikami zakłócającymi. Dostosowanie do tego powinno skutkować znacznym poprawieniem powtarzalności między laboratoriami. W tym celu użyto algorytmu SVAseq, jednak jego właściwe zastosowanie wymagało gruntownego przemyślenia. Czy powinniśmy wskazać, że próbki SHAM i PNSL pochodzą z różnych serii (Figura 2a), czy też traktować je razem (Figura 2b)? Pierwsze podejście powoduje grupowanie według serii, a drugie według grupy (SHAM lub PNSL). Odpowiedź na to pytanie zależy od rodzaju analizy, którą chcielibyśmy przeprowadzić - czy chcemy porównać wszystkie próbki PNSL do wszystkich próbek SHAM, czy też

szukamy powtarzalności między seriami. Na Figurze 3 możemy zobaczyć, że zastosowanie algorytmu SVAseq znacząco zwiększyło liczbę wykrytych DEG, ale powtarzalność wciąż pozostaje na poziomie około 8% przy zastosowaniu SVAseq osobno (Figura 3a), a jeszcze niższym - 6% przy zastosowaniu razem (Figura 3b). Pomimo uzyskania list genów różnicowo ekspresjonowanych, wyniki powtarzalności pozostają niskie. Dlatego użyto kolejnej rekomendacji dotyczącej dodatkowego filtrowania wartości logFC (powyżej 1). Zgodnie z wcześniejszymi badaniami możliwe jest wtedy zwiększenie powtarzalności nawet do 95%. Jednak w tym konkretnym przypadku zmiana sygnału jest tak niewielka, że zastosowanie tego filtra spowodowało usunięcie dużej części genów i pogorszenie powtarzalności.

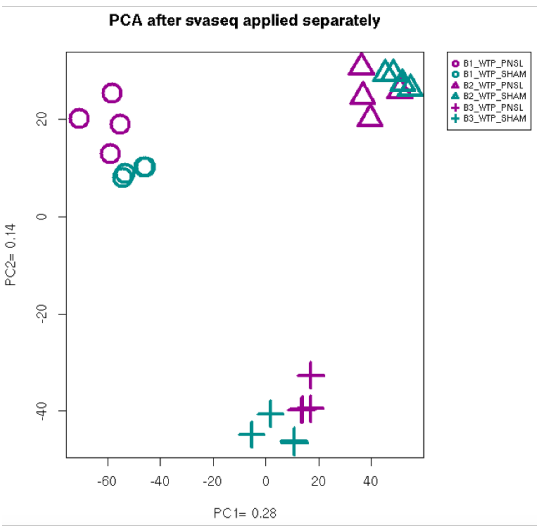

Figura 2a: PCA z usuniętymi 7 czynnikami , zastosowano osobno.

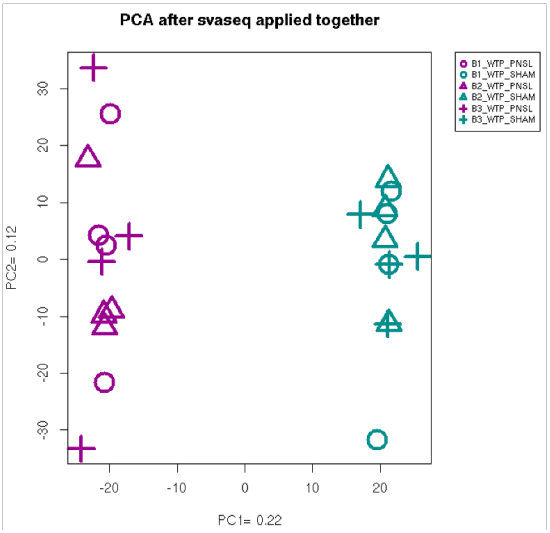

Figura 2b: PCA z usuniętymi 7 czynnikami , zastosowano razem.

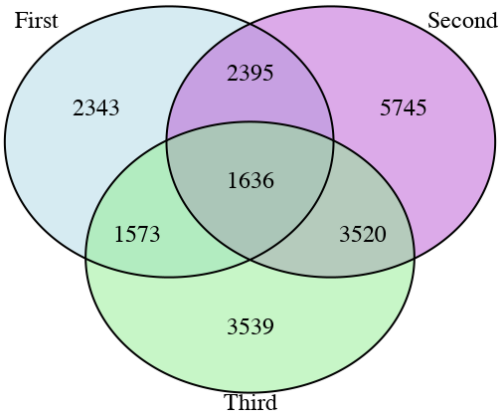

Figura 1a: Diagram Venna dla powtarzalności między seriami, SVAseq zastosowano osobno.

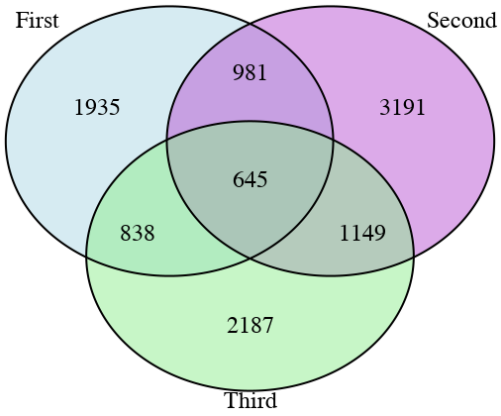

Figura 3b: Diagram Venna dla powtarzalności między seriami, SVAseq zastosowano razem.

Wyniki analizy różnicowej pokazały, że zalecenia oparte na badaniach z użyciem zestawów danych sztucznie stworzonych, gdzie zmiany sygnału są duże, powinny być stosowane ostrożnie wobec "rzeczywistych" problemów. W przypadku danych o niskich wartościach sygnału, powtarzalność może znacznie się różnić w zależności od wybranej metody i uzyskanie wyników tak wysokich, jak podane ponad 90% zgodność w listach DEG, nie zawsze jest możliwe. Ponadto, metody

filtracji powinny być dostosowane do wartości sygnału, a nie stosowane arbitralnie na podstawie artykułów benchmarkingowych.
Korzystne może być zastanowienie się nad możliwością wystąpienia efektów paczki w naszych danych i odpowiednie ich uwzględnienie. Pierwszym problemem do rozwiązania jest staranne przemyślenie pytań, na które musimy odpowiedzieć, oraz jakie porównania należy przeprowadzić. Dodatkowo, wybór odpowiedniej liczby czynników do skorygowania jest kluczowy.

Ponieważ analiza ilościowa dla danych dotyczących bólu neuropatycznego nie dostarczyła zadowalających wyników biologicznych, skupiono się na analizie jakościowej wszystkich 88 próbek, aby zbadać dokładniej transkryptom myszy. Poprzez uwzględnienie także próbek z knockout-ami genów, powinno być możliwe zidentyfikowanie nowych zdarzeń ASE, które są specyficzne dla rdzenia kręgowego, niezależnie od warunków stresowych. Zdarzenia retencji intronów zostały wykluczone z analizy, ponieważ protokół przygotowania biblioteki opierał się na rybodeplecji, co mogło spowodować dużą liczbę fałszywie pozytywnych nASEs tego typu ze względu na obecność niedojrzałego mRNA. Spladder raportuje wyniki dla dwóch izoform, jednej zawierającej dane zdarzenie, drugiej wykluczającej je. W związku z tym, po uwzględnieniu ASE już istniejących w adnotacji, nasze dane są podzielone na trzy grupy:

- nowa izoforma + znana izoforma (nowa+stara),
- obie nowe izoformy (nowa+nowa),
- obie znane izoformy (stara+stara).

Spladder raportuje, między innymi, wartość PSI (procentowy udział w splicingu) dla każdego zdarzenia i próbki. Jest to stosunek sygnału wspierającego dane zdarzenie do sumy sygnałów dla obu zdarzeń. Wykorzystując PSI i odpowiedni próg, liczba wyników fałszywie pozytywnych może być zmniejszona. Wybrano zestaw kryteriów do wyboru prawidłowych zdarzeń. Pierwsze, zastosowane dla wszystkich trzech grup, polegało na ustawieniu progu na odchylenie standardowe PSI. Zdarzenia, które miały odchylenie poniżej 0.2 we wszystkich 4 powtórzeniach, były traktowane jako prawidłowe. Ponieważ dalsza analiza narzędziem Bisbee jest obecnie dostępna tylko dla zdarzeń, które mają przynajmniej jedną izoformę już w adnotacji, przeprowadzono ją tylko dla grupy zdarzeń nowa+stara. Aby skupić się na wystarczająco silnych zdarzeniach, które mają większą szansę na niebycie fałszywymi pozytywnymi wynikami, zastosowano bardziej rygorystyczne podejście i dodano próg dla wartości PSI - powinna ona wynosić powyżej 0.2, co w istocie oznacza, że nowe ASE stanowi przynajmniej 1/4 już znanych ASE.
Mimo że powtarzalność analizy różnicowej ekspresji jest bardzo niska, analiza alternatywnego splicingu wykazała wiele zdarzeń AS spójnych dla wszystkich 88 próbek (Tabela 1). Wpływają one również na dużą liczbę genów (Tabela 2), ale dalsza analiza wykazała, że chociaż wydaje się, że wpływają na różne procesy i funkcje, są one głównie związane z układem nerwowym, ponieważ część wspólna dla terminów CC (Cellular Component) była wyższa niż dla innych kategorii, a także terminy te związane były z układem nerwowym.

Event	Both iso known			Both iso new			New and old					Total
		sd<0.2			sd<0.2			sd<0.2		PSI>0.2 & sd<0.2		
	All	Common	Percentage	All	Common	Percentage	All	Common	Percentage	Common	Percentage	
Alternative 3 prime	2692	828	33.76	1154	188	16.29	8256	4453	53.49	13	0.16	12102
Alternative 5 prime	2108	552	26.19	915	138	15.08	4996	2468	49.40	8	0.16	8019
Exon skip	5776	1294	22.40	3477	93	2.67	7269	2242	30.84	12	0.17	16522
Mutually exclusive exons	95	45	47.37	305	133	43.61	97	51	52.58	12	12.37	497
Multiple exon skip	383	164	42.82	1006	550	54.67	886	510	58.89	48	5.54	2255

Tabela 1: Liczba wykrytych ASEs w zależności od typu i grupy oraz wspólna liczba zdarzeń.

Event	Both iso known			Both iso new			New and old					Total
		sd<0.2			sd<0.2			sd<0.2		PSI>0.2 & sd<0.2		
	All	Common	Percentage	All	Common	Percentage	All	Common	Percentage	Common	Percentage	
Alternative 3 prime	2075	702	33.83	698	143	20.49	4364	2322	53.21	13	0.30	7137
Alternative 5 prime	1722	498	28.92	648	107	16.51	3295	1614	48.98	7	0.21	5665
Exon skip	3668	929	25.33	1885	61	3.24	4294	1419	33.05	11	0.26	9847
Mutually exclusive exons	86	42	48.84	131	45	34.35	89	46	51.69	11	12.36	306
Multiple exon skip	363	160	44.08	746	449	60.19	762	459	60.24	40	5.25	1871

Tabela 2: Liczba genów zawierających wykryte ASEs w zależności od typu i grupy oraz wspólna liczba genów.

Ponadto analiza funkcjonalna przy użyciu kombinacji zdarzeń z różnych grup wykazała, że nowe zdarzenia mogą dostarczyć nowych informacji do adnotacji. Rysunek 4 przedstawia 10 najważniejszych terminów Molecular Function (MF) dla zdarzeń *alternative 3 prime* oraz ich istotność statystyczną w porównaniu do innych grup. Zauważamy trzy terminy, które nie są istotne w grupie stary+stary i nowy+stary, jednak w przypadku grupy nowy+nowy, terminy te są istotnie adnotowane. Są to:

- structural constituent of postsynaptic intermediate filament cytoskeleton,
- phosphorylation-dependent protein binding,
- ATPase-coupled transmembrane transporter activity

Filamenty aktynowe, które budują cytoszkielet, mogą dynamicznie tworzyć różne struktury w odpowiedzi na nowe bodźce, co opisuje się jako plastyczność zależną od aktywności. Fosforylacja białek jest głównym czynnikiem w szlakach przewodzenia sygnałów. Aktywność transporterów transbłonowych może być związana z sygnalizacją za pośrednictwem neuroprzekaźników w układzie nerwowym.

Figura 4: Top 10 terminów MF GO dla zdarzenia alternative 3' end pokazana dla trzech grup.

Obserwacja, że różne typy zdarzeń ASE mogą potencjalnie być związane z różnymi typami funkcji, wydaje się być niewystarczająco zbadana. W dyskusji z kilkoma ekspertami wskazano to jako prawdopodobne, jednak informacja ta nie znalazła jeszcze potwierdzenia w artykułach.

Dalsza walidacja za pomocą, na przykład, sekwencjonowania długich odczytów, jest potrzebna do potwierdzenia, ale także do wizualizacji, jak dokładnie wyglądają całe transkrypty, ponieważ widzimy, że wiele zdarzeń może występować w tym samym genie i pokrywać się ze sobą, tutaj jednak możemy badać tylko krótkie fragmenty. Uzyskane wizualizacje (jedna z nich jest przedstawiona dalej) wydają się potwierdzać, przynajmniej dla grupy nowy+stary, że odczyty wspierają wykryte zdarzenia i

dodatkowo zdają się występować w genach związanych z układem nerwowym oraz wpływać nawet na ważne domeny białkowe.

Dla 93 wcześniej wybranych nASE różnych typów przeprowadzono analizę Bisbee i InterProscan dla grupy nowy+stary. Tabela 3 podsumowuje efekty ORF i aminokwasowe w białkach zawierających nową wersją transkryptu. Przedwczesne zakończenie było głównie spowodowane substytucją. Cztery zdarzenia spowodowały utratę białka, a siedem było milczących. Dla większości zdarzeń InterProscan był w stanie znaleźć i przypisać referencyjne domeny białkowe. W następnym kroku, zwizualizowano kilka wybranych zdarzeń.

Event	Premature Stop		In frame				Protein loss		Total	Assigned by InterProScan
	Insertion	Substitution	Deletion	Insertion	Substitution	Silent	Start loss	Stop loss		
Alternative 3 prime	0	3	3	4	2	1	0	0	13	12
Alternative 5 prime	0	2	2	2	1	0	1	0	8	7
Exon skip	0	3	4	4	1	0	0	0	12	12
Mutually exclusive exons	0	4	0	0	8	0	0	0	12	12
Multiple exon skip	1	21	3	12	2	6	0	3	48	41

Tabela 3: Tabela pokazująca typ zmiany wprowadzonej przez nASE dla grupy nowy+ stary oraz liczba zmodyfikowanych transkryptów, dla których zostały przypisane domeny przez InterProScan.

Przykład na Fig. 5 został stworzony dla zdarzenia pominięcia wielu eksonów w genie Nrcam, który między innymi jest zaangażowany w adhezję między neuronami i wspiera sygnały kierunkowe podczas wzrostu stożka aksonalnego. Może odgrywać ogólną rolę w komunikacji międzykomórkowej. Wykres jest podzielony na 5 części:

- zdarzenia dodatkowe - inne niż badane główne zdarzenia, również wybrane jako prawidłowe,
- ścieżka dopasowania - wsparcie odczytów dostarczone przez wszystkie 88 próbek,
- wykryte zdarzenie - transkrypt z włączonym nowym zdarzeniem,
- oryginalny - oryginalny transkrypt,
- domeny białkowe Interpro – domeny przypisane do oryginalnego transkryptu.

Część pokazująca wybrane zdarzenie oznaczona jest czerwonym prostokątnym polem. Widzimy, że piki dla eksonów uważanych za pominięte w badanym przypadku są faktycznie znacznie mniejsze niż piki dla pozostałych eksonów. Wykres dla domen białkowych pokazuje, że dwa obszary dotknięte są tym zdarzeniem. Opisane są w InterProScan jako neuronalna cząsteczka adhezji komórek.

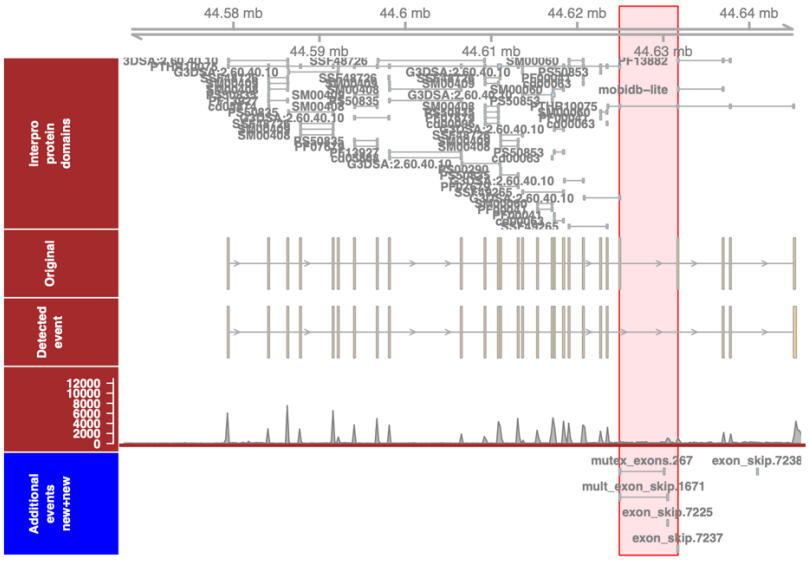

Figura 6: Wizualizacja zdarzenia multiple exon skip dla genu Nrcam.

Część poświęcona danym benchmarkingowym została wykorzystana, aby zbadać, jak Spladder będzie działać w przypadku silnego sygnału. Używaliśmy danych wygenerowanych poprzez sekwencjonowanie *targeted short reads*, gdzie wykorzystano panele A1, A2 i R1. Spladder był uruchamiany na tych danych, używając jako odniesienia zarówno adnotacji AceView, jak i AceView rozszerzonej przez konsorcjum SEQC2 (z narzędziem IsoQuant uruchomionym na długich odczytach). Rezultaty dostarczają solidnego dowodu na to, że wyniki Spladdera są wiarygodne i jest to dobra decyzja dla opracowanego procesu. Procent znanych intronów wykrytych przez Spladdera dla wybranych genów jest zgodny z frakcją transkryptomu, która powinna być ekspresjonowana w danym punkcie czasowym. Potwierdziło to także efektywność sekwencjonowania ukierunkowanego, ponieważ 87% wykrytych połączeń (junctions) pochodziło z genów obecnych na panelu.
Mimo rozszerzenia adnotacji referencyjnej, Spladder nadal raportuje prawie 95 tysięcy nowych połączeń. Trzeba pamiętać, że transkrypty rozszerzające adnotację zostały wybrane w bardzo rygorystyczny sposób, gdzie około 90% początkowo wybranych transkryptów zostało odrzuconych. Wyniki Spladdera mogą wskazywać, że połączenia te są poprawne, ale wymagają dalszej walidacji. Ponownie widzimy dowód na to, że nawet obszerne adnotacje są nadal niekompletne.

Analiza danych mikromacierzowych potwierdziła wyniki z analizy RNA-seq. Metody korekcji efektu paczki mogą być bardzo przydatne, jeśli są stosowane prawidłowo, ale dane uzyskane w eksperymentach medycznych mogą wciąż być bardzo problematyczne i mimo zastosowania najlepszych praktyk, możemy nadal uzyskać niską powtarzalność. Opierając się na wcześniejszych doświadczeniach z danymi RNA-seq, został zbudowany kompletny potok przetwarzania danych mikromacierzowych w stosunkowo krótkim czasie. Fakt, że technologie te mogą bazować na tych samych metodach, stanowi ogromną zaletę, ponieważ mikromacierze to technologia starsza, posiadająca wiele solidnych algorytmów preprocessingu. Zrozumienie obu podejść jest kluczowe, ponieważ nie powinny być traktowane jako metody konkurencyjne, ale raczej używane zamiennie, w zależności od problemu naukowego. Dlatego opracowanie i włączenie kroków do analizy danych mikromacierzy do narzędzia może być bardzo korzystne.

W trakcie realizacji projektów zdobyto głębsze zrozumienie wyzwań i ograniczeń napotykanych w analizie rzeczywistych zestawów danych biomedycznych. Odzwierciedlają one złożoność eksperymentów i niepożądaną zmienności. Krytyczna ocena procesu analizy i uzyskanych wyników pozwoliła na sformułowanie trzech głównych wniosków. Są one uzupełnieniem opracowanego procesu w postaci najlepszych praktyk/wskazówek i stanowią najważniejszy rezultat tej pracy.

1. Obecnie nie istnieją standardy złota w analizie danych o dużej przepustowości.
2. Rekomendacje dotyczące analiz oparte na badaniach z zestawami danych sztucznych powinny być stosowane do problemów rzeczywistych z ostrożnością.
3. Podejścia RNA-seq i mikromacierzowe mają swoje zalety i wady, powinny być używane zamiennie, w zależności od problemu naukowego.

Ta praca pokazała również, że złożoność genomu myszy, ale także człowieka, nie jest jeszcze w pełni zrozumiana. Pomimo obecności wielu czynników zakłócających, możliwe było wykrycie wspólnych wzorców w danych. Było oczywiste, że terminy Gene Ontology wspólne między różnymi typami alternatywnych zdarzeń splicingu (ASE) są związane z układem nerwowym. Wydaje się, że funkcje wykrytych ASE i procesy, w które są zaangażowane, są charakterystyczne dla danego typu zdarzenia. Wyniki wskazują, że istnieje wiele zdarzeń nieobecnych jeszcze w adnotacji i że wciąż można wprowadzić udoskonalenia w zakresie adnotacji referencyjnych transkryptów. Ta obserwacja jest

zgodna z kilkoma ostatnimi artykułami, które opisują wiele nowych zdarzeń alternatywnego splicingu występujących w różnych obszarach mózgu, a także w innych tkankach u różnych gatunków. Mimo to, wciąż brakuje wszechstronnego narzędzia do badania tego aspektu. W rezultacie przeprowadzonych badań opracowano narzędzie obejmujące różne etapy analizy i eliminujące tę lukę. Pomoże to poprawić powtarzalność, zapewniając większą kontrolę nad metodami i parametrami używanymi na różnych etapach analizy, ale także odpowiadając na potrzebę automatycznego wykrywania i analizy AS oraz jego konsekwencji. Ponieważ AS odgrywa również ogromną rolę w wielu chorobach, narzędzie to może dostarczyć nowych spostrzeżeń na temat zmian zachodzących w warunkach takich jak choroba Parkinsona, SMA czy różne typy nowotworów.