Dr hab. inż. Paweł Piotr Łabaj
Małopolskie Centrum Biotechnologii
Uniwersytet Jagielloński
Gronostajowa 7a, 30-387 Kraków
pawel.labaj@uj.edu.pl

Kraków, 16.08.2023

Recenzja rozprawy doktorskiej

Tytuł rozprawy:	Skipping batch effect correction: clustering-based methods for analyzing confounded single-cell RNA-sequencing data
Autor rozprawy:	mgr inż. Tomasz Kujawa
Promotor rozprawy:	Prof. dr hab. inż. Joanna Polańska
Kopromotor rozprawy:	dr inż. Michał Marczyk
Dziedzina:	nauki inżynieryjno-techniczne
Dyscyplina:	Inżynieria Biomedyczna

Rozwój wysokoprzepustowych biomedycznych technik pomiarowych pozwolił w ostatnich latach na znaczny postęp w naukach biologicznych i medycznych. Szczególne przyspieszenie nastąpiło w związku z rozwojem Next Generation Sequencing (NGS) i technik pochodnych. Jednym z przykładów jest zastosowanie NGS do ilościowej oceny ekspresji genów i ich alternatywnych transkryptów. Domyślną strategią jest charakteryzacja transkryptomu w populacjach komórek pochodzących z określonej tkanki (ang. bulk RNAseq). To podejście jest motywowane założeniem, że komórki z tego samego typu tkanki są jednorodne i jednocześnie takie podejście jest historycznie odziedziczone po technikach nisko-przepustowych oraz mikromacierzowych. W rezultacie uzyskane wyniki są swego rodzaju średnią ważoną z różnych profili ekspresji z badanej populacji komórek. Podejście to bardzo dobrze sprawdza się w ogólnej charakteryzacji tkanki, jednak w szczególnych przypadkach, gdzie interesujący nas sygnał nie pochodzi od dominującej grupy komórek, jest niewystarczające. Alternatywnym podejściem jest sekwencjonowania RNA pojedynczych komórek (scRNA-Seq), które pod względem technologicznym dynamicznie rozwija się w ostatnich latach. W tym podejściu co do zasady uzyskujemy profile ekspresji dla każdej komórki z osobna, co pozwala na rozróżnienie nie tylko pod względem typu komórki ale także ich fazy „cyklu życia".

Jednak analiza danych pochodzących z eksperymentów scRNA-Seq jest trudnym zadaniem, znacznie trudniejszym niż z „klasycznego" RNA-Seq (bulk RNA-Seq). Z jednej strony musimy mierzyć się z wysoką wymiarowością danych, a z drugiej z szumem o złożonej etiologii. W typowym eksperymencie scRNA-Seq dla każdej próbki dla dziesiątek a nawet setek tysięcy komórek mierzymy poziomy ekspresji dla setek tysięcy alternatywnych transkryptów genów (często dla uproszczenia sumowanych do poziomu

genów, co wiąże się z utratą części informacji). Jednocześnie cały proces generacji danych jest złożony z wielu etapów, gdzie na każdym sygnał może ulec zaszumieniu. Do tego oczywiście dochodzą ograniczenia i charakterystyka samej technologii sekwencjonowania. Na całość nakłada się także charakterystyka biologiczna: z naturalną, stochastyczną różnicą w sygnałach pomiędzy komórkami tego samego typu, różnicą między typami komórek, jak i różnicą wynikająca z fazy „cyklu życia" danej komórki. Większość z tych składowych jest też obecna w danych z eksperymentów typu „bulk RNA-Seq", jednak większa złożoność przygotowania próbki jak i dodatkowy wymiar złożoności w postaci dziesiątek/setek tysięcy komórek powoduje, że analiza tego typów danych jest trudniejsza zarówno ze względu na bardziej złożony szum jak i większą wymiarowość danych w stosunku do klasycznego RNA-Seq.

Nałożenie tych dwóch kwestii powoduje jeszcze jedną trudność. O ile w klasycznym podejściu, czy też w eksperymentach z użyciem technologii mikromacierzowej, kwestia efektu paczki, lub w szerszym kontekście, czynników zakłócających jest dobrze poznana, a zaproponowane narzędzia nieźle radzą sobie z modelowaniem tych czynników, o tyle w przypadku scRNA-Seq jest to ciągle wyzwanie. Na sile nabiera tutaj kwestia znana z RNA-Seq dotycząca odróżnienia zmienności biologicznej od technicznej, co często skutkuje nadmierną korektą. W podejściu nienadzorowanym może ona doprowadzić do usunięcia różnic biologicznych, natomiast w podejściach „częściowo" nadzorowanych, może doprowadzić do usunięcia również tych składowych zmienności technicznej, które leżą u podstaw zastosowanego modelu rozkładu danych do celu analizie różnicowej, co w konsekwencji powoduje znaczny wzrost wyników fałszywie pozytywnych.

Pomimo zaproponowania wielu podejść do tego wyzwania, dotychczas nie wykształcił się konsensus odnośnie standardów postępowania. Wydaje się, że konieczna jest dalsza eksploracja rozwiązań oraz dalsze badania w obszarze modelowania, korekcji, czy łagodzenia efektu paczki (czynników zakłócających) dla danych z eksperymentów scRNA-Seq. W przedstawionej rozprawie doktorskiej Autor podejmuje to wyzwanie poprzez innowacyjną adaptację koncepcji stosowanej w narzędziu DiviK, które z sukcesem jest rozwijane w grupie prowadzonej przez promotora rozprawy. W realizacji zaproponowanej w niniejszej pracy wykorzystywane jest iteracyjne grupowanie podprzestrzeni w połączeniu z analizą funkcjonalną zestawów genów, aby złagodzić negatywny wpływ efektu paczki/partii na dane scRNA-Seq. Jak podkreśla doktorant, kluczowym aspektem analizy funkcjonalnej jest identyfikacja ścieżek specyficznych dla klastrów komórek i ustalenie ich powiązań między partiami. Założeniem jest, że iteracyjne grupowanie podprzestrzeni może zmniejszyć siłę czynników zaburzających poprzez usuwanie większej ilości szumu z danych przy każdej kolejnej iteracji. W rezultacie komórki powinny mieć tendencję do tworzenia grup w oparciu o ich prawdziwą biologię. Takie podejście powoduje „obejście" konieczności stosowania korekty efektu paczki poprzez skupienie się bezpośrednio na integracji zestawów danych, które zostały wygenerowane osobno. Autor wykazuje, że takie podejście nie było wcześniej badane, co nadaje pracy wymagany wymiar nowatorskiego wkładu w naukę.

Układ rozprawy jest typowy dla tego typu opracowań. Jednak powstała rozbieżności między spisem treści a strukturą pracy zaprezentowaną w I.3. W kolejnych rozdziałach Autor:

- przybliża charakterystykę eksperymentów scRNA-Seq wraz z identyfikacją źródeł szumów oraz prezentacją obecnych podejść do radzenia sobie z tym wyzwaniem (Rozdział 2),
- przedstawia projekt eksperymentu, zestaw danych użytych w analizie jak i przybliża metody stosowane do generowania i analizy danych (Rozdział 3),
- przedstawia wyniki stosowania różnych podejść korekcji efektu paczki jak również przykład użycia i wyniki zaproponowanego podejścia (Rozdział 4),
- prezentuje dyskusję zaproponowanego podejścia w kontekście uzyskanych wyników jak również podsumowuje pracę w odniesieniu do założonych celów (Rozdział 5)

Autor bardzo dokładnie omawia obecny stan wiedzy co świadczy o dużym oczytaniu i dobrym przygotowaniu do podjęcia zagadnień poruszanych w dalszych częściach pracy. Jednak jest jedna kwestia, której pominięcie, zwłaszcza w kontekście tak dogłębnej analizy stanu wiedzy, jest dla mnie niewytłumaczalna. W technologii RNA-Seq nie mierzymy bezpośrednio poziomu ekspresji genów gdyż wykorzystanie NGS powoduje, że w sposób ilościowy budujemy profil występowania różnych form RNA. To co nas głównie interesuje to ekspresja mRNA oraz lncRNA. Dlatego należy mocno podkreślić, że poprzez RNA-Seq określamy ilościowy profil alternatywnych transkryptów genów. Ekspresja genów jest tutaj wartością pochodną powstałą poprzez zsumowanie ekspresji ich alternatywnych transkryptów. Jest ona bardzo często stosowana przy analizie danych RNA-Seq ponieważ analiza na poziomie alternatywnych transkryptów jest trudniejsza. Należy jednak podkreślić, że decydując się na analizę na poziomie genów upraszczamy proces kosztem pozbawienia naszej analizy pełnej mocy wynikającej z technologii NGS.

Uproszczenie analizy ma oczywiście swoje konsekwencje. W kontekście kwestii związanych ze zmianą poziomu ekspresji (strony 8-9) czy kinetyki transkrypcji (strona 11) należy zwrócić uwagę, że przy analizie na poziomie alternatywnych transkryptów możliwe do zaobserwowania są zmiany niewidoczne na poziomie genów: i) „włączenie" jednego alternatywnego transkryptu może być kompensowane przez „wyłączenie" innego, więc na poziomie genu nic się nie zmienia, ii) zmiany mogą następować dla alternatywnych transkryptów na drugim czy trzecim poziomie ekspresji, a więc są maskowane przez niezmienną ekspresję dominującego alternatywnego transkryptu.

Powyższa kwestia jest głównym merytorycznym niedociągnięciem pracy, która ogólnie jest dobrze napisana a narracja jest przejrzysta. Niestety Autor nie ustrzegł się też błędów mniejszego kalibru dotyczących:

i) zbyt dużych skrótów myślowych, np.:

- strona 5 „… As a result, scRNA-seq data exhibits a high fraction of zero measurements, often referred to as sparsity." – to stwierdzenie jest poprawne, ale trzeba dużego doświadczenia w analizie danych, żeby móc taki wniosek wysnuć na podstawie zdań poprzedzających,
- strona 28 – brakuje dokładniejszego poprowadzenie czytelnika przez koncepcję radzenia sobie z efektem paczki poprzez integrację danych. Tutaj kwestia integracji pojawia się nagle i znowu osoba bez doświadczenia może mieć problem z jasnym powiązaniem tych kwestii,

- strona 29 – następuje mocne „przyspieszenie” pod względem wymaganej wiedzy koniecznej do zrozumienia omawianych koncepcji.

ii) czy stawiania kontrowersyjnych tez, które są „rozbrajane” kilka stron dalej, ale pierwsze negatywne wrażenie pozostaje z czytelnikiem, np.:

- strona 16 *“To be detected by the sequencer, cDNA needs to be duplicated (amplified) millions of times.”* – prawda dla technologii Illumina czy IonTorrent ale już nie dla PacBio czy ONT, doprecyzowane później w tekście,
- strona 18 *“However, full-length methodologies often introduce a bias towards long genes, as long transcripts tend to have a higher number of reads mapped to them compared to short genes with similar expression levels [51, 52].”* – korekcja tego jest dostępna w limma czy edgeR poprzez algorytm TMM czy natywnie w DESeq2, złagodzenie wydźwięku następuje później w pracy,
- strona 27 *“These cells should be intermingled and indistinguishable even if they originate from different batches.”* – naturalny szum stochastyczny powoduje, że zawsze będą różnice.

Powyższe kwestie dotyczą szeroko rozumianego kontekstu pracy, który jest niezmiernie istotny, bo to w nim osadzone jest proponowane nowe podejście. Jednak kontekst ten nie ma wpływu na samą jakość wykonanej pracy, a tą należy ocenić wysoko. Koncepcja wynikająca z algorytmu DiviK jest szeroko i twórczo eksplorowana przez zespół promotor pracy i przykład przedstawionej do oceny pracy doktorskiej jest tego dowodem. Merytorycznie takie podejście jest oczywiście poprawne, a ponieważ sama koncepcja ma solidne podstawy w innych zastosowaniach to zaproponowana innowacyjna adaptacja DiviK-a nie mogła się nie udać.

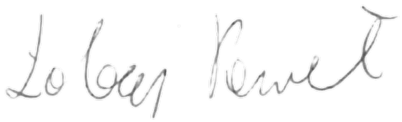
Tezy rozprawy są sformułowane jasno i przystępnie oraz są w pełni poparte danymi zawartymi w poszczególnych rozdziałach rozprawy. Podsumowanie rozprawy jest syntetycznym wykazaniem, że założone w pracy cele zostały osiągnięte. W zakończeniu podsumowania doktorant dla każdej z tez wskazuje, która część rozprawy się do niej odnosi.

Rozprawa zawiera 52 szczegółowe ryciny oraz 7 tabel, które co do zasady są jasne i czytelne. Piśmiennictwo obejmuje 141 co do zasady dobrze dobranych i aktualnych pozycji, chociaż w kontekście ogólnego podejścia do łagodzenia wpływu czynników zaburzających kilka prac konsorcjum SEQC byłoby mile widzianych.

Podsumowując, przedstawiona do oceny praca doktorska stanowi bardzo wartościowe uzupełnienie obecnego stanu wiedzy odnośnie łagodzenia wpływu efektu paczki w analiza danych scRNA-Seq. Zaproponowano podejście umożliwiające skonsolidowaną analizę zestawów danych pochodzących z eksperymentów scRNA-Seq i prezentujących silny efekt paczki. Proponowane rozwiązanie opiera się na iteracyjnej metodzie grupowania z selekcją cech połączonej z analizą funkcjonalną zestawów genów. Po analizie funkcjonalnej otrzymane klastry są łączone między paczkami na podstawie ich funkcjonalnego podobieństwa. Jest to innowacyjna adaptacha algorytmu klasteryzującego (Divisive Intelligent K-Means) do zupełnie nowego zagadnienia i typu danych. W konsekwencji należy stwierdzić, że praca ta w pełni odpowiada warunkom stawianym rozprawom doktorskim oraz wypełnia istotną lukę w obecnym stanie wiedzy.

Na podstawie powyższej oceny stwierdzam, że wymieniona rozprawa doktorska w pełni odpowiada warunkom stawianym w ustawie Prawo o szkolnictwie wyższym i nauce / Dz. U. z 2022 r. poz. 574, w zakresie nadawania stopni naukowych i na tej podstawie wnoszę do Wysokiej Rady Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej o dopuszczenie mgr inż. Tomasza Kujawę do dalszych etapów przewodu doktorskiego.

Nie mam wątpliwości, że doświadczenie zgromadzone przez Autora stawia cały zespół badawczy w doskonałej pozycji wśród międzynarodowych grup zajmujących się tą tematyką.



Dr hab. inż. Paweł Piotr Łabaj