Narodowy
Instytut
Onkologii
im. Marii Skłodowskiej-Curie
Państwowy Instytut Badawczy
Oddział w Gliwicach

ul. Wybrzeże Armii Krajowej 15, 44-101 Gliwice; tel. 32 278 98 88; fax. 32 232 35 12; e-mail: katarzyna.lisowska@io.gliwice.pl

Gliwice, 08.08.2023

RECENZJA ROZPRAWY DOKTORSKIEJ mgr Tomasza Kujawy

Tytuł rozprawy: Skipping batch effect correction: clustering-based methods for analyzing confounded single-cell RNA-sequencing data

Przedłożona mi do recenzji rozprawa doktorska Pana mgr. Tomasza Kujawy została wykonana w ramach projektu AIDA (Applied Integrative Data Analysis) realizowanego na Politechnice Śląskiej, pod kierunkiem prof. dr hab. inż. Joanny Polańskiej oraz promotora pomocniczego dr inż. Michała Marczyka.

Celem pracy było opracowanie protokołu postępowania (pipeline) pozwalającego zniwelować tzw. „efekt paczki" (batch effect), występujący często w danych z obszaru genomiki funkcjonalnej (w tym przypadku - sekwencjonowania RNA w pojedynczych komórkach; single cel RNA sequencing, scRNA-seq). Efekt paczki wprowadza zmienność niezwiązaną z badaną zmiennością biologiczną, co może prowadzić do błędnych wniosków z analizy danych.

Chociaż efekty paczki mają szkodliwy wpływ na dane, proces ich korekty może być również szkodliwy. Dlatego mgr Tomasz Kujawa podjął się opracowania nowej metody radzenia sobie z efektem paczki, bez stosowania standardowych metod korekcji.

Ocena formalna rozprawy doktorskiej

Praca liczy 110 stron i zawiera 141 pozycji bibliografii, w mojej opinii - dobranych właściwie. Praca ma układ, z którym często spotykam się w przypadku doktoratów przygotowywanych na Politechnice Śląskiej. Pierwszy rozdział, zatytułowany „Introduction" w zasadzie przedstawia cele pracy, zaś kolejny rozdział pt. „Background" zawiera wprowadzenie do obszaru badawczego, w którym mieści się projekt doktorski. Kolejne rozdziały wymienione w spisie treści to „Methods", „Results" oraz „Discussion" i „Bibliography". Ostatnią część dyskusji stanowi (pod)rozdział zatytułowany Summary, który jednak nie jest wymieniony w spisie treści. Spis treści nie wymienia również streszczeń: w jęz. angielskim na str. 99 i w jęz. polskim na str. 100. Na str. 101 znajdują się jeszcze podziękowania (Acknowlegments), które również nie są wymienione w spisie treści. Jest to o tyle niezrozumiałe, że za tymi elementami znajduje się spis cytowań (Bibliography, str. 102), który jest wymieniony w spisie treści. Czytelnik nie ma więc szans dowiedzieć się o istnieniu streszczeń, podziękowań czy podsumowania, o ile nie dotrze z lekturą rozprawy niemal do samego końca. Co prawda, podziękowania zazwyczaj nie miewają charakteru osobnego rozdziału, jednak zwykle znajdują się na początku rozprawy. Podobnie streszczenia – nawet jeżeli nie są numerowane jako osobny rozdział, znajdują się na początku, a jeżeli znajdują się na końcu, to są wymienione w spisie treści. Również podsumowanie (Summary) zasługuje na rangę osobnego rozdziału, tym bardziej, że jest ono elementem rozprawy poszukiwanym przez czytelnika, który pragnie zapoznać się w skrótowy sposób z wynikami pracy. Nie są to z mojej strony uwagi krytyczne pod adresem Doktoranta, gdyż domyślam się, że ten nieco dziwny układ pracy nie jest dziełem jego inwencji, a raczej został narzucony przez Uczelnię.

W mojej ocenie rozprawa doktorska jest napisana jasno, czytelnie i dobrym językiem, oraz jest starannie opracowana pod względem edytorskim. Dobrze się ją czyta, szczególnie rozdział pt.

„Background". Rozdział ten zawiera też wiele właściwie dobranych rycin, dobrze ilustrujących omawiane zagadnienia. Jedynym wyjątkiem jest rycina 27 na str. 38 - dla mnie słabo opisana. Tym niemniej uważam, że niewielkim nakładem pracy można ten rozdział rozbudować do postaci pracy poglądowej, omawiającej metodykę sekwencjonowania transkryptomu pojedynczych komórek i analizy danych scRNA-seq, w tym metody wstępnej obróbki danych, odfiltrowywania komórek o niskiej jakości, różne metody normalizacji, selekcji cech, redukcji wymiarowości danych, w tym klasteryzacji. Do tego zachęcam Doktoranta.

Do tej części pracy mam następujące pytanie – o ile dobrze rozumiem, na str. 36, przy omawianiu procesu selekcji cech (feature selection, highly variable genes, HVG) – pojawia się założenie, że istotne różnice biologiczne będą się przejawiać poprzez wyższy poziom zmienności (wariancji) ekspresji, a szum techniczny to niewielkie zmiany poziomu ekspresji. Czy to jest założenie zawsze prawdziwe?

Z obowiązku recenzenta muszę wymienić drobne błędy edytorskie, które udało mi się wyśledzić w pracy:

Str. 31 - opis Fig. 22 – dodatkowe myślniki w słowach highlighted, conducted. Figura w słabej rozdzielczości

Str. 32, Rozdział II.4 – błąd redakcyjny: … to **reduce improve** signal to noise ratio… (albo: reduce, albo: improve)

Str. 34 - … the expression matrix still not accurately reflect … (powinno być: reflects)

Str. 36 - Urwane zdanie (bez początku): basic approach is based on modeling of the mean-variance relationship.

Str. 37 - brak kropki po "dimensionality reduction", niepotrzebna kropka przed (Figure 27C).

Str. 39 – brak przecinka przed "as" w zdaniu: "Clustering plays a crucial role as many subsequent biological analyses rely on its results."

Str. 41 – odsyłacz do Fig. 35, która ma przedstawiać różne wzory klasteryzacji w zależności od metody. Fig. 35 przedstawia co innego

Str. 64 – słowo „compared" powtórzone niepotrzebnie dwa razy

Str. 79 - podpis „Table 3 continued" – powinien być na str. 80

Dane scRNA-seq

Sekwencjonowanie transkryptomu pojedynczych komórek (sc RNA-seq) jest często wykonywane w ramach wielkoskalowych i wieloośrodkowych projektów, co skutkuje koniecznością przeprowadzania eksperymentów partiami, w różnym czasie i w różnych laboratoriach. W konsekwencji, w danych z scRNA-seq, nieuchronnie pojawia się efekt paczki.

Mocną stroną projektu doktorskiego jest wykorzystanie unikalnego zbioru danych scRNA-seq, pochodzącego z dwóch różnych publikacji, których współautorem jest promotor pomocniczy, dr inż. Michał Marczyk. Oba te zestawy danych miały identyczne właściwości biologiczne, a różniły się jedynie aspektami technicznymi, co powodowało silnie wyrażony efekt paczki w jednym zbiorze danych i praktyczny brak takiego efektu w drugim zestawie. Pierwszy zestaw określono roboczo jako „confounded" (zaburzony), drugi zaś „balanced" (zrównoważony).

W obu przypadkach dane pochodziły z doświadczeń in vitro na komórkach potrójnie ujemnego raka piersi MDA-MB-231, traktowanych eksperymentalnym lekiem o nazwie Navitoclax (inhibitor antyapoptotycznych białek z rodziny BCL2); sekwencjonowanie pojedynczych komórek wykonano za pomocą platformy firmy 10x Genomics. Celem prowadzonych badań było poszukiwanie molekularnych mechanizmów rozwoju oporności na Navitoclax.

W pracy z 2021 roku analizowano ponad 500 wariantów traktowania komórek MDA-MB-231 samym Navitoclaxem lub w kombinacji z Kryzotynibem (inhibitor ALK i ROS1). Z tego zestawu danych Doktorant wybrał i) próbki traktowane Navitoclaxem przez 3 dni (próbki

oznaczone jako T2), ii) próbki traktowane jw. i hodowane przez kolejnych 10 dni w medium bez leku (recovery phase; próbki T3) oraz iii) próbki kontrolne (nie traktowane lekiem; próbki T1); wszystkie warianty były zsekwencjonowane w dwóch powtórzeniach (A i B).

Takie same punkty eksperymentalne analizowano za pomocą scRNA-seq także w pracy z 2020 roku. Różnica między zestawami danych scRNA-seq z tych dwóch prac wynikała ze schematu eksperymentalnej obróbki materiału biologicznego (izolacja RNA, amplifikacja, synteza bibliotek, sekwencjonowanie). W pracy z 2021 roku, ze względu na ogromne rozmiary eksperymentu in vitro oraz związane z tym trudności logistyczne, popełniono zasadnicze błędy w projektowaniu schematu eksperymentalnego, a mianowicie, materiał biologiczny był obrabiany eksperymentalnie i sekwencjonowany w transzach odpowiadających jednemu punktowi czasowemu/jednemu powtórzeniu. Efektem tego był niezwykle silnie wyrażony efekt paczki (co widać doskonale na Rycinie 31), natomiast zmienność biologiczna pomiędzy punktami eksperymentalnymi, była znacznie trudniejsza do zaobserwowania.

Zrównoważony zbiór danych, składający się z powtórzeń technicznych oznaczonych jako A i B, został wykorzystany jako zbiór walidacyjny.

Ocena pracy doktorskiej

W pierwszej kolejności przeprowadzono kontrolę jakości danych scRNA-seq, wykorzystując dedykowane oprogramowanie firmy 10x Genomics (Cell Ranger software), które umożliwia m.in. detekcję i filtrowanie odczytów niskiej jakości (pochodzących z martwych i pękniętych komórek, dubletów i pustych kropli). Oceniono rozmiar biblioteki, liczbę wykrywanych genów na komórkę, proporcję genów zmapowanych w stosunku do genomu mitochondrialnego.

Następnie zastosowano kolejno 6 algorytmów korekcji efektu paczki (dla punktów T1 i T2): ComBat-seq, Limma, Mutual Nearest Neighbour, scMerge, Seurat v4, i Scanorama. Wizualne porównanie efektów zastosowania poszczególnych algorytmów przedstawia rycina 32. Najlepsze wyniki otrzymano za pomocą programu scMerge wykorzystującego algorytm k-średnich, przy wartościach kmeansK= (4,4,3,3), który skutkował najlepszym grupowaniem komórek zgodnie z właściwym punktem czasowym (T1, T2) oraz dobrym wymieszaniem powtórzeń A i B. Z kolei algorytm Scanorama sprawował się najgorzej. W tym punkcie Doktorant wykazał, że korekta efektu paczki z wykorzystaniem niektórych narzędzi bioinformatycznych, przy głęboko zaburzonym układzie eksperymentalnym (confounded dataset), może być praktycznie niemożliwa. Ponadto, w kolejnym kroku, Doktorant wykazał, że algorytmy korekty efektu paczki mogą zniekształcać pierwotną naturę i oryginalną dystrybucję danych, wpływając potencjalnie na wiarygodność dalszej analizy. Istniejące algorytmy często bowiem traktują priorytetowo osiągnięcie pełnego wymieszania komórek (danych transkryptomicznych) między partiami, kosztem zatarcia podstawowej struktury danych. Doktorant podkreśla, że brakuje miary do ilościowego szacowania niepewności związanej z procesem korekcji.

Co ważne, na podstawie wykonanych analiz Doktorant wykazał, że nie ma jednej, najlepszej metody korekty, którą można by zastosować do wszystkich zestawów danych i konfiguracji eksperymentalnych; algorytm korekty efektu paczki należy zatem dobierać empirycznie, w zależności od charakteru analizowanych danych.

Oryginalnym i kluczowym aspektem projektu doktorskiego jest próba zmierzenia się z wyzwaniem analizy głęboko zaburzonych danych scRNA-seq, bez stosowania wymienionych wyżej bioinformatycznych algorytmów do korekty efektu paczki. W tym celu zaproponowano podejście oparte na iteracyjnej metodzie grupowania (klasteryzacji) z selekcją cech, połączonej z analizą funkcjonalną zestawów genów. Do analizy funkcjonalnej wykorzystano zbiór 186 ścieżek sygnałowych z bazy KEGG. Założono, że szlaki sygnałowe, czyli zestawy funkcjonalnie powiązanych genów, wykazują większą odporność na negatywny wpływ efektów paczki, niż pojedyncze geny.

Do klasteryzacji wykorzystano procedurę analizy danych w oparciu o algorytm Divisive Intelligent K-Means (DiviK), który został oryginalnie zaprojektowany dla analizy danych ze

spektrometrii masowej, jednak tutaj został odpowiednio zmodyfikowany i dostosowany do analizy danych z scRNA-seq. Po analizie funkcjonalnej otrzymane klastry były łączone między paczkami na podstawie ich funkcjonalnego podobieństwa. Oryginalnym podejściem w zakresie łączenia klastrów było wykorzystanie miary wielkości efektu (effect size) w scenariuszu "jeden-do-wszystkich" (one-to-all). Takie podejście pozwoliło wyodrębnić szlaki sygnałowe o silnym efekcie biologicznym. To podejście pozwoliło na złagodzenie negatywnego wpływu efektu paczki, bez użycia algorytmów korekcyjnych, czyli unikając negatywnych konsekwencji stosowania tych algorytmów.

Zbilansowany zbiór danych, składający się z powtórzeń technicznych oznaczonych jako A i B, służył jako dwupoziomowy zbiór walidacyjny. Na początkowym etapie ten zbiór danych wykorzystano do walidacji strategii selekcji cech opartej na rozkładzie wariancji genów na mieszaniny składowych gaussowskich. Wykazano, że zastosowany algorytm k-średnich, w połączeniu z selekcją cech, przypisuje większe znaczenie wysoce zmiennym genom zidentyfikowanym za pomocą strategii Gaussian Mixture Model (GMM). Wydajność grupowania była znacznie lepsza przy wykorzystaniu genów charakteryzujących się dużą zawartością informacji. Jednak nawet po odrzuceniu większości „nieistotnych" genów podczas etapu globalnej filtracji, stosunek sygnału do szumu pozostał niezadowalający, powodując duże nakładanie się składowych Gaussa. Dodatkowo, zliczenia bliskie zeru obserwowane w macierzy ekspresji miały znaczący wpływ na dokładność obliczania wariancji.

Na początkowym etapie przeprowadzono śledzenie klastrów między zbiorami próbek odpowiadającymi różnym punktom czasowym. Analiza ujawniła skomplikowane wzorce przepływu między powtórzeniami technicznymi w zrównoważonym zbiorze danych. Zaobserwowano, że określony klaster może wykazywać podobieństwa z wieloma innymi klastrami. Doktorant tłumaczy to heterogenią komórek, która może występować nawet w jednorodnej hodowli komórkowej.

Drugi scenariusz obejmował ustalenie powiązania między tymi samymi punktami czasowymi w zbiorze zaburzonym i zrównoważonym, traktowanym jako referencyjny. Wzorce przepływu wykryte w referencyjnym zbiorze danych, zostały przeniesione do zaburzonego zbioru danych, ujawniając bardziej jednolite struktury przepływu. Wartości wyników podobieństwa są stosunkowo niskie, jednak należy pamiętać, że analiza dotyczyła wybitnie zaburzonego zbioru danych, stanowiącego skrajny przykład efektu paczki. Uzyskane wyniki sugerują, że proponowane podejście może być obiecującym narzędziem do analizy takich skrajnie zaburzonych danych z eksperymentów scRNAseq.

Brakuje mi natomiast szerszego omówienia wyników analizy funkcjonalnej. Komórkowe szlaki sygnałowe zostały wykorzystane wyłącznie jako techniczne narzędzie umożliwiające wykrywanie podobnych klastrów w danych z różnych transz eksperymentalnych. Nie zostały natomiast przeanalizowane pod kątem celu, w jakim wykonywane były obydwa eksperymenty scRNA-seq [Marczyk M, et al., 2020; Patwardhan, GA, et al., 2021], czyli poszukiwania molekularnych mechanizmów oporności komórek raka piersi na Navitoclax. Chętnie usłyszałabym podczas obrony pracy doktorskiej więcej informacji na temat. Drugie pytanie, które mnie nurtuje, to czy zastosowanie zaproponowanego w pracy doktorskiej postępowania obliczeniowego rzeczywiście doprowadziło do wykrycia w zaburzonym zbiorze danych podobnych szlaków sygnałowych związanych z odpowiedzią na traktowanie komórek Navitoclaxem, co w zbiorze referencyjnym? W ramach pracy doktorskiej przeprowadzono dwie rundy tworzenia klastrów, co okazało się wystarczające na potrzeby łączenia klastrów. Czy Doktorant próbował większej liczby rund i czy znalazł tam szlaki sygnałowe związane powstawaniem lekooporności, wspólne dla obu zbiorów danych? Wykrycie w zaburzonym zbiorze danych podobnych wzorców jak w zbiorze zrównoważonym, byłoby rodzajem walidacji zaproponowanego postępowania obliczeniowego służącego do niwelacji efektu paczki. Warto również w przyszłości (co podkreśla sam Doktorant) wykorzystać inne bazy danych szlaków sygnałowych lub zastosować wybraną kolekcję szlaków istotnych dla badanego systemu biologicznego.

Podsumowując, w ramach swojej pracy doktorskiej, mgr Tomasz Kujawa zaproponował podejście obliczeniowe (pipeline), które wykorzystuje iteracyjne grupowanie podprzestrzeni w połączeniu z analizą funkcjonalną zestawów genów, aby złagodzić negatywny wpływ efektu paczki na dane scRNAseq. Kluczowym aspektem analizy funkcjonalnej jest identyfikacja ścieżek specyficznych dla klastrów i ustalenie ich powiązań między partiami eksperymentalnymi. Proponowany algorytm umożliwia skonsolidowaną analizę zestawów danych, które zostały wygenerowane oddzielnie, bez potrzeby korekcji efektu paczki. Algorytm bazuje na dobrze ugruntowanych i uznanych metodach, charakteryzuje się prostotą, niskim kosztem obliczeniowym i stosunkowa łatwością interpretacji. Jak rozumiem, algorytm jest proponowany jako lepszy dla bardzo silnie zaburzonych zbiorów danych, w których standardowa korekta efektu paczki może skutkować poważnym zaburzeniem pierwotnej struktury danych i otrzymaniem zupełnie fałszywych wyników analizy.

W literaturze są opisane przykłady wykorzystania klasteryzacji do łagodzenia szumu w danych scRNA-seq. Jednak podejście polegające na wykorzystaniu tej techniki specjalnie do łagodzenia efektu paczki nie było dotąd badane. Nowatorskim pomysłem było także zastosowanie miary wielkości efektu do określenia ścieżek specyficznych dla klastrów, a następnie procedury łączenia, która umożliwia śledzenie klastrów w różnych partiach danych. Statystyki wielkości efektu delta Cliffa są często używane, jednak ich zastosowanie do określania ścieżek specyficznych dla klastra nie było wcześniej stosowane.

Wprowadzono także oryginalne poprawki, takie jak globalna filtracja szumu, oparta na zbinaryzowanej macierzy ekspresji genów, aby poradzić sobie z nadmiarem odczytów zerowych, typowym w danych scRNA-seq.

Główne wnioski:

- korekta efektu paczki ma szkodliwy wpływ na pierwotną dystrybucję danych scRNAseq, co może negatywnie wpływać na wiarygodność wyników dalszej analizy
- prosta metoda selekcji cech oparta na dekompozycji wariancji może zastąpić bardziej wyrafinowane algorytmy
- połączenie analizy funkcjonalnej, w oparciu o komórkowe szlaki sygnałowe, i procedury łączenia klastrów pozwala na pominięcie etapu korekty efektu paczki

W mojej opinii, praca spełnia warunki określone w ustawie z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce, z późniejszymi zmianami (Dz. U. z 2023 r. poz. 742, 1088 i 1234; ogłoszone dnia 30 kwietnia 2023 r.), dlatego wnioskuję do Rady Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej o dopuszczenie mgr Tomasza Kujawy do dalszych etapów przewodu doktorskiego.

Podpisała Katarzyna Lisowska (podpis odręczny)

Narodowy Instytut Onkologii im. M. Skłodowskiej-Curie PIB, Oddział w Gliwicach