INSTYTUT
GENETYKI CZŁOWIEKA
POLSKIEJ AKADEMII NAUK

ul. Strzeszyńska 32
60-479 Poznań
tel. +48/61/657 91 00
e-mail: igcz@man.poznan.pl
www.igcz.poznan.pl

Poznań, 26.08.2023 r.

prof. dr hab. n. med. inż. Andrzej Pławski

Recenzja rozprawy doktorskiej Pana mgr. Tomasza Kujawy pt. „*Skipping batch effect correction: clustering-based methods for analyzing confounded single-cell RNA-sequencing data*" w związku z postępowaniem o nadanie stopnia doktora w dziedzinie nauk inżynieryjno-technicznych, dyscyplina inżynieria biomedyczna. Poniższa recenzja została przygotowana na podstawie decyzji Rady Naukowej Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej w Zabrzu, która na posiedzeniu w dniu 22.06.2023 r. powołała mnie na recenzenta wyżej wymienionej dysertacji.

Ocena formalna

Rozprawa doktorska Pana mgr. Tomasza Kujawy pt. „Skipping batch effect correction: clustering-based methods for analyzing confounded single-cell RNA-sequencing data" została wykonana na Wydziale Automatyki, Elektroniki i Informatyki Politechniki Śląskiej w Gliwicach. Rozprawa liczy 110 stron, gdzie w poszczególnych rozdziałach zamieszczono: wstęp (Introduction), tło prowadzonych badań (Background), metody (Methods), wyniki (Results), dyskusje (Discussion), a także streszczenie w języku angielskim i polskim oraz bibliografie. Na końcu pracy zamieszczono listę skrótów oraz listy figur i tabel. Prace przedstawiono w języku angielskim. Układ pracy we wstępie (Introduction) zawiera informacje o przesłankach do wykonania badań, celu pracy oraz opisuje strukturę pracy. W mojej ocenie praca została przygotowana bardzo dobrze i co do układu pracy nie mam za zastrzeżeń

Ocena merytoryczna

Po pojawieniu się metod masowego sekwencjonowania, gdzie po raz pierwszy wprowadzono sekwencje identyfikujące tak zwane indeksy pozwalające w obróbce informatycznej identyfikować pochodzenie odczytów otwarła się zupełnie nowa era badań kwasów nukleinowych. Cztery różne nukleotydy

pozwalają na indywidulane oznakowywanie przy w sumie nie wielkich długościach sekwencji identyfikującej ponieważ liczba kombinacji sekwencji takich oligonukletydów jest czwartą potęgą długości takiej sekwencji. W przypadku kodonu liczba kombinacji to 64 ale już 10 nukleotydowy fragment pozwala na utworzenie 10`000 różnych kombinacji. Takie możliwości nieuchronnie musiały doprowadzić do utworzenia nowych metod opartych na znajomości pochodzenia zsekwencjonowanych masowo fragmentów kwasów nukleinowych, jak transkryptomika przestrzenna (spatial transcryptomics) oraz sekwencjonowanie RNA pojedynczej komórki (scRNAseq). Metody te opierają się na sekwencjonowaniu zestawów cDNA oznakowanego regionu preparatu w przypadku trankryptomiki przestrzennej, a w przypadku sekwencjonowania pojedynczych komórek wyznakowanie pozwala na określenie komórek z których pochodzi dany fragment cDNA. Na podstawie liczby odczytów poszczególnych sekwencji można porównywać poziomy ekspresji poszczególnych genów co jest kopalnią informacji o komórce. Dane te z punktu widzenia szeroko pojętych nauk biologicznych pozwalają na wnioskowanie jakie procesy zachodzą w badanej komórce czy to pod wpływem stanu patologicznego czy podania konkretnej substancji, np. leku. Aby wnioskowanie na podstawie tych danych było prawidłowe muszą one odzwierciedlać stan zgodny ze stanem faktycznym. Sama część laboratoryjna wykonywania takich analiz jest złożoną wieloetapową operacją, która może wpływać na rzetelność uzyskanych danych. Badania takie wykonywane jako wysokoprzepustowe, wielkoskalowe eksperymenty z powodów logistycznych niekiedy nie są wykonywane jednoczasowo. Powszechna dostępność uzyskanych rezultatów (udostepnienie wymagane przy publikacji) stwarza możliwości analiz retrospektywnych gdzie analizy wykonywane były w różnych laboratoriach na różnych platformach z różnymi partiami odczynników czy personelem. Wspólna analiza takich danych jest utrudniona właśnie powodu tak zwanego efektu paczki (batch effect). Korekta tego efektu może prowadzić do zniekształcenia dystrybucji danych. Doktorant w swej pracy podjął się opracowania podejścia złagodzenia negatywnego wpływu efektu paczki bez konieczności jego korekty. W moje ocenie podjęte badania są w pełni uzasadnione naukowo i maja duży potencjał zastosowania praktycznego.

We wstępie doktorant przedstawił przesłanki do podjęcia rozwiązania problemu batch efect (efektu paczki) oraz cel pracy, którym jest opracowanie podejścia umożliwiającego skonsolidowaną analizę danych pochodzących z różnych eksperymentów scRNAseq. W tym celu doktorant zaproponował podejście oparte na iteracyjnym grupowaniu komórek połączonym z selekcją cech oraz

oparte na iteracyjnym grupowaniu komórek połączonym z selekcją cech oraz analizą funkcjonalną zestawów genów. Doktorant przyjął następujące założenia:

- iteracyjne grupowanie komórek pozwoli niejako na rozcieńczenie efektu paczki w grupach otrzymanych w kolejnych iteracjach
- ścieżki sygnałowe składające się z wielu powiązanych ze sobą genów wykazują większą odporność na negatywny wpływ efektu paczki, niż pojedyncze geny.

Cel oraz przyjęte założenia w moje ocenie zostały sformułowane prawidłowo, jasno i precyzyjnie.

Kolejnym rozdziałem jest rozdział zatytułowany „Background", w którym doktorant opisuje teoretyczne podstawy techniki sekwencjonowania RNA pojedynczych komórek oraz przebieg eksperymentu wykonywanego w laboratorium. W tym rozdziale doktorant krytycznie scharakteryzował podejścia stosowane w analizie tego typu danych, dobrze opisał źródła ich zaburzeń oraz aktualne podejścia stosowane do redukcji szumu w danych z sekwencjonowań scRNAseq. W rozdziale zebrano aktualne grupy narzędzi do korygowania efektu paczki oraz opisano główne idee, na których się opierają. W moje ocenie rozdział ten jest napisany poprawnie i świadczy o dobrej znaności tematu przez doktoranta, a załączone ryciny dobrze uzupełnią tekst rozdziału. Figura 22 w wersji drukowanej jest mało czytelna z powodu doboru kolorów, ja jednak dysponowałem również wersją elektroniczna gdzie ten problem nie występuje.

W rozdziale Metody (Methods) doktorant przedstawił kluczowe aspekty dwóch eksperymentów, na podstawie których zostały wygenerowane dane do analizy. Oba eksperymenty dotyczyły badania wpływu navitoclaxu na transkryptom tej samej linii komórkowej potrójnie ujemnego raka piersi w celu identyfikacji mechanizmów lekoodporności w leczeniu raka piersi. W obu eksperymentach lek podawany był w trzech tych samych punktach czasowych. Biblioteki do obu eksperymentów przygotowano przy użyciu platformy firmy 10XGenomics i zsekwencjonowano przy użyciu tej samej platformy firmy Ilummina. W jednym eksperymencie użyto 6000 komórek co wygenerowało 25 000 odczytów na komórkę a w drugim użyto 1500 komórek co wygenerowało 200 000 odczytów na komórkę.

Kluczową różnicę między tymi dwoma eksperymentami stanowił aspekt techniczny związany z procesowaniem poszczególnych próbek. W eksperymencie zbalansowanym opisywanym przez doktoranta jako „*balanced*

study" próbki z poszczególnych punktów czasowych procesowane były na jednym czipie w tym samym czasie. Można zatem oczekiwać, że różnice w ekspresji wynikają wyłącznie z zastosowania leku. Ten eksperyment stanowił dla doktoranta wzorzec referencyjny. Natomiast w przypadku drugiego eksperymentu, lustrzanego pod kątem biologicznym, próbki z poszczególnych punktów czasowych procesowane były na różnych czipach i w różnym czasie. Ten eksperyment opisywany przez doktoranta jako zaburzony („*confounded study*") skutkował pojawieniem się silnego efektu paczki w danych, który uniemożliwił ich wiarygodną analizę biologiczną. W mojej ocenie układ doświadczalny jest dobrany właściwie dla realizacji celu pracy. Doktorant jasno precyzuje kryteria wyboru narzędzi bioinformatycznych służących do korekty efektu paczki oraz opisuje ogólny mechanizm ich działania. Rozdział ten doktorant finalizuje opisem proponowanej przez siebie metodyki, której główne etapy przedstawił na Figurze nr 30. Pierwszy etap proponowanego przez doktoranta protokołu obejmuje globalne jednorazowe filtrowanie danych, co pozwala na usunięcie znaczącej części szumu związanego z problemem zerowych pomiarów. Następnie stosowane jest filtrowanie lokalne w obrębie klastrów, na podstawie którego komórki są grupowane w kolejnej iteracji. Optymalna liczba klastrów jest określana na podstawie indeksu Caliński-Harabasz. Finalnie, klastry otrzymane w każdym eksperymencie poddawane są niezależnej analizie funkcjonalnej zestawów genów, a następnie łączone na podstawie podobieństwa ich profili funkcjonalnych. Do analizy funkcjonalnej doktorant wykorzystał zbiór 186 ścieżek sygnałowych z bazy KEGG. Po analizie funkcjonalnej otrzymane klastry były łączone między zestawami danych, tj. paczkami na podstawie ich funkcjonalnego podobieństwa. Oryginalnym propozycją doktoranta w kwestii łączenia klastrów było wykorzystanie jednej z miar wielkości efektu (effect size) w scenariuszu porównawczym "jeden-do-wszystkich" (one-to-all). Takie podejście daje możliwość identyfikacji w danym klastrze szlaków sygnałowych o silnym efekcie biologicznym.

Na samym początku sekcji wyników na figurze nr 31 doktorant zwizualizował oba eksperymenty na poziomie komórkowym za pomocą plotów UMAP. Wizualizacja ta stanowi dowód występowania silnego efektu paczki w danych pochodzących z eksperymentu zaburzonego „confounded study" oraz uzasadnia potrzebę jego korekty. W tym celu doktorant wykorzystał wybrane narzędzia bioinformatyczne, a dane po korekcie ponownie zwizualizował na poziomie komórkowym za pomocą plotów UMAP. Wizualizacje po korekcie przestawione

na figurze nr 32 dowodzą, że dostępne narzędzia nie są w stanie skorygować efektu paczki.

Doktorant ocenia również wpływ procesu korekty na oryginalny rozkład ekspresji genów. Na figurach 34 oraz 35 udowadnia, że proces korekty efektu paczki zaburza rozkład ekspresji genów, co uniemożliwia wykorzystanie skorygowanych danych do wszelkich analiz na poziomie genowym. Jednocześnie ta sekcja potwierdza teoretyczne podstawy podjęcia badań.

Na figurach 41-44 doktorant wykazał użyteczność zastosowanej przez niego metody lokalnej filtracji cech opartej na dekompozycji wariancji mieszaniny rozkładów Gaussowskich. Następnie doktorant przedstawił ścieżki sygnałowe specyficzne dla każdego otrzymanego klastra w danym eksperymencie. Do ich identyfikacji doktorant wykorzystał oryginalne podejście z wykorzystaniem miary wielkości efektu, którą to miarę użył w scenariuszu porównawczym "jeden-do-wszystkich" (one-to-all). Scenariusz ten polegał na tym, że w danym momencie analizowano konkretną ścieżkę sygnałową w danym klastrze *versus* wszystkie pozostałe klastry zidentyfikowane w danej próbce w ramach danego eksperymentu. Pozwoliło to na identyfikację ścieżek specyficznych dla każdego otrzymanego klastra i to takich ścieżek gdzie badany efekt biologiczny znacznie przewyższa efekt paczki. Ścieżki te doktorant przedstawia w tabelach nr 3 oraz 4. Na podstawie podanych tabel możemy zauważyć, że w obu zestawach danych na pierwszy plan wyłaniają się podobne ścieżki metaboliczne reagujące na podanie leku.

Do oceny funkcjonalnego podobieństwa między klastrami doktorant proponuje dwie miary: korelacyjną oraz opartą o typowy iloczyn wektorowy. Ta druga opisywana jest przez doktoranta jako tzw. „*similarity score*". W przypadku obu miar argumentami są wektory użytego indeksu wielkości efektu (Cliff's delta) dla porównywanych klastrów. Na figurach nr 50 oraz 52 stanowiących tzw. diagramy przepływu doktorant pokazuje, że dany klaster może wykazywać podobieństwo z wieloma innymi klastrami. Doktorant tłumaczy to heterogenicznością komórek oraz prawdopodobnie zbyt małą głębokością podziałów (komórki grupowane były tylko do drugiej iteracji). Mimo mnogości przepływów pokazanych na figurach 51 oraz 52 daje się jednak zauważyć, że w większości przypadków dany klaster wykazuje większe podobieństwo do jednego klastra i kilka mniejszych podobieństw do innych klastrów. Te maksymalne wartości podobieństwa dla danego klastra doktorant zbiera w formie tabel nr 5,6 oraz 7. Wzorce przepływów

zidentyfikowane w referencyjnym zbiorze danych, doktorant przenosi na zbiór testowy (zaburzony).

Przedstawione przez doktoranta wyniki wskazują, że zaproponowane przez niego oryginalne rozwiązanie otwiera nową drogę w dziedzinie podejścia do efektu paczki w danych scRNAseq. Doktorant miał do dyspozycji unikalny układ eksperymentalny, który pozwolił mu na rzetelne przetestowanie proponowanego podejścia bioinformatycznego, aczkolwiek w mojej opinii metodyka ta wymaga jeszcze dalszej weryfikacji. Dostępność podobnej pary lustrzanych zestawów eksperymentalnych możne stanowić istotną barierę w tej kwestii.

Doktorant jest świadomy ograniczeń proponowanej metodyki, o czym świadczy przeprowadzona przez niego dyskusja oraz co warte podkreślenia, ma w perspektywie pomysły na jej dalsze udoskonalania. W kwestiach edytorskich nie mam zastrzeżeń. Jednak co do oceny języka jako nie native speaker wolałbym nie zajmować stanowiska

Podsumowanie

Rozprawa doktorska Pana mgr. Tomasza Kujawy pt. „*Skipping batch effect correction: clustering-based methods for analyzing confounded single-cell RNA-sequencing data*" spełnia wymagania określone w artykule 187 ustawy z dnia 20 lipca 2018 r. „Prawo o szkolnictwie wyższym" i wnoszę do wysokiej Rady Naukowej Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej w Zabrzu o dopuszczenie rozprawy doktorskiej Pana mgr. Tomasza Kujawy do dalszych etapów przewodu doktorskiego. W związku z tym, że bardzo wysoko oceniam poziom rozprawy Pana mgr. Tomasza Kujawy wnoszę o wyróżnienie niniejszej rozprawy przez wysoką Radę Naukową Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej w Zabrzu

05420 Prof. dr hab. n. med. inż.
ANDRZEJ PŁAWSKI
DIAGNOSTA LABORATORYJNY
specjalista laboratoryjnej genetyki medycznej

Poznań, 2023.08.26

Andrzej Pławski

Andrzej Pławski