Silesian University of Technology
Faculty of Automatic Control, Electronics
and Computer Science

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

PhD Thesis

Author: **Tomasz Kujawa**
Supervisor: **prof. dr hab. inż. Joanna Polańska**
Co-supervisor: **dr inż. Michał Marczyk**

Gliwice, June 2023

STRESZCZENIE

Efekt paczki/partii jest nieuniknionym zjawiskiem w przypadku wysokoprzepustowych i wielkoskalowych eksperymentów, gdzie ograniczenia logistyczne wymagają generowania danych w różnym czasie i przy zaangażowaniu wielu laboratoriów, często wyposażonych w odmienne platformy sprzętowe, wykorzystujących różne partie odczynników i przy udziale zróżnicowanego personelu badawczego.
Efekt paczki wprowadza dodatkową warstwę szumu technicznego, który nie jest równomiernie rozłożony w obrębie badanych cech biologicznych. Tym samym szum ten nie podlega korekcie na etapie normalizacji danych i w konsekwencji prowadzi do fałszywych wniosków. W związku z tym konieczna jest komputerowa korekta danych z efektem paczki. Niestety korekta efektu paczki jest skomplikowanym zadaniem ze względu na trudności związane z rozróżnieniem zmienności wywołanej czynnikami natury czysto technicznej od heterogeniczności biologicznej.

Niestety sam proces korekty wiąże się z kilkoma negatywnymi konsekwencjami. Korekta zniekształca bowiem pierwotną naturę oraz dystrybucję danych. Ponadto brakuje miary do ilościowego szacowania niepewności tego procesu. Dlatego istnieje silna potrzeba rozwoju badań w dziedzinie usuwania lub korekcji efektu paczki przy użyciu nowych podejść i narzędzi bioinformatycznych.

Niniejsza praca ma celu dostarczenie protokołu umożliwiającego skonsolidowaną analizę danych wygenerowanych oddzielnie (partiami) z silnym efektem paczki. Protokół ten został opracowany w celu ominięcia konieczności zastosowania korekty przy jednoczesnym złagodzeniu negatywnych skutków powodowanych przez efekt paczki. Jest to możliwe dzięki wykorzystaniu procedury iteracyjnego grupowania w podprzestrzeni cech połączonego z analizą funkcjonalną ścieżek sygnałowych.
Oryginalna, a zarazem centralna idea pracy polega na zastosowaniu miary wielkości efektu do określenia ścieżek specyficznych dla klastra, a następnie procedury łączenia klastrów na podstawie ich podobieństwa funkcjonalnego. Innymi słowy, podejście to umożliwia łączenie klastrów pochodzących z różnych partii.

Wykorzystanie grupowania podprzestrzeni w połączeniu z analizą funkcjonalną ścieżek sygnałowych do łagodzenia efektów parti jak dotąd nie było badane, co stanowi nowy wkład w tę dziedzinę. Zaproponowane rozwiązanie stawia na prostotę, niski koszt obliczeniowy oraz łatwość interpretacji nie tylko dla statystyka, ale również dla biologa.
Podstawowym założeniem jest, że iteracyjne grupowanie podprzestrzeni może zmniejszyć efekty partii poprzez usuwanie większej ilości szumów z danych przy każdej kolejnej iteracji. W rezultacie komórki powinny tworzyć grupy na podstawie ich rzeczywistej biologii. Ponadto, oczekuje się, że zidentyfikowane ścieżki specyficzne dla klastra będą tymi o bardzo silnej manifestacji, a tym samym będą wykazywały swego rodzaju odporność na negatywny wpływ efektów partii, który jest zazwyczaj mniej wyraźny w odniesieniu do szlaków sygnałowych w porównaniu do pojedynczych genów.

Tezy tej rozprawy zostały sformułowane w następujący sposób:

1. Istniejące algorytmy korekcji efektu partii w scRNAseq często zniekształcają pierwotny rozkład ekspresji genów. W rezultacie analizy na poziomie genów, takie jak różnicowa ekspresja czy identyfikacja markerów, nie mogą być bezpiecznie stosowane na skorygowanym zbiorze danych.

2. Prosta strategia wyboru cech oparta na rozkładzie wariancji daje podobne wyniki do bardziej zaawansowanych i obliczeniowo kosztownych metod.

3. Korekta efektu paczki jest możliwa do pominięcia w danych pochodzących z scRNAseq. Zamiast tego można przeprowadzić wiarygodną analizę, identyfikując niezależnie podklastry komórek w każdej partii, a następnie łącząc je między partiami na podstawie podobieństwa ich profili funkcjonalnych w celu śledzenia podobnych komórek z różnych partii.

Ten projekt wyróżnia się ze względu na unikalne zestawienie zbiorów danych. Otóż zbiory te pochodzą z dwóch realnych eksperymentów identycznych pod kątem projektu biologicznego, które różniły się jedynie pod kątem aspektu technicznego, jak przedstawiono na **Figurze 1**.

Pierwszy eksperyment był częścią większego badania. Jednak po analizie okazało się, że eksperyment wykazuje silne efekty partii wynikające z różnic w procesowaniu grup biologicznych. W tym badaniu komórki zebrano w różnych punktach czasowych i

przetwarzano na oddzielnych chipach i w różnych dniach. Ten zbiór danych nazywany jest zbiorem zagłuszonym (ang. confounded study).

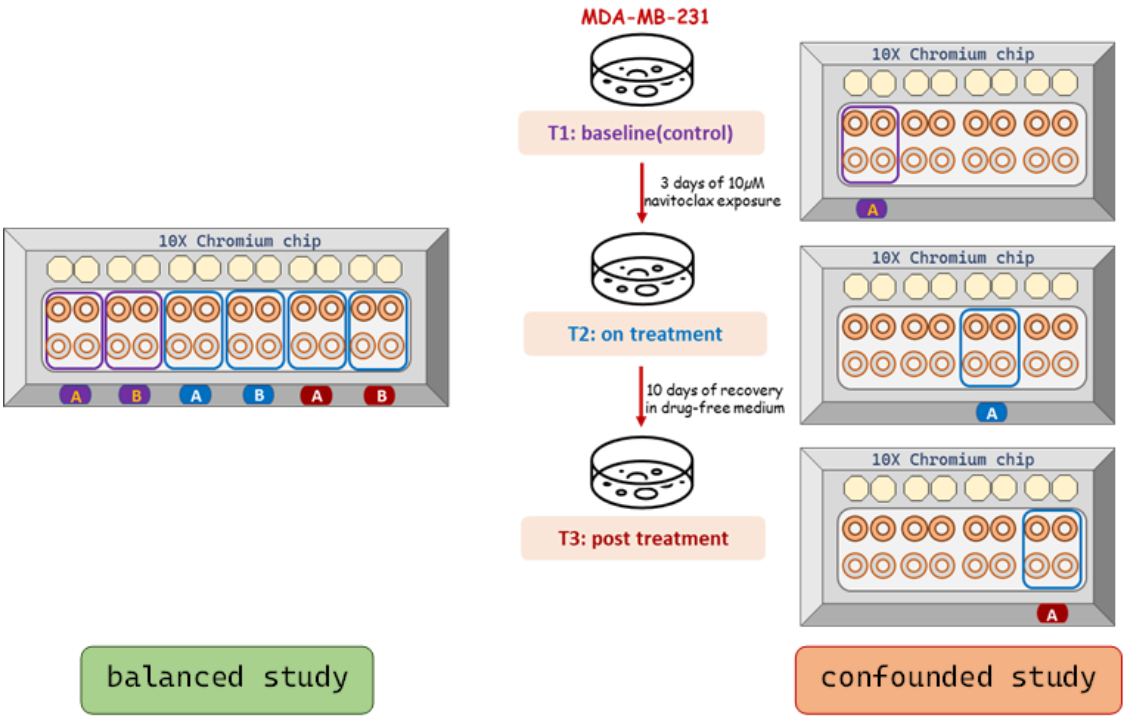

Figura 1. *Projekt eksperymentalny. Oba eksperymenty miały na celu zbadanie wpływu podaży naxitoclax'u na transkryptom linii komórkowej raka piersi (MDA-MB-231). W obu eksperymentach wykorzystano tę samą linię komórkową oraz dwa powtórzenia techniczne (A i B). Komórki zostały eksponowane na działanie 10 μM leku I zbierane do analizy w trzech punktach czasowych: przed podażą leku (kontrola – T1), po 3 dniach od ekspozycji (T2) oraz po 10 dniach od ekspozycji (T3).*

Drugi eksperyment, określany jako zbalansowany (ang. balanced study), został zaprojektowany w celu zminimalizowania zmienności technicznej. W tym badaniu komórki zebrane w różnych punktach czasowych były dzielone i przetwarzane na tym samym chipie i w tym samym dniu. Dążono do tego, aby wszystkie obserwowane różnice wynikały głównie z efektów leczenia farmakologicznego, a nie były wynikiem działania czynników technicznych.

To badanie służy jako odniesienie. Taki układ eksperymentalny różni się od istniejących ewaluacji, które często opierają się na symulacjach lub badaniach nadzorowanych, tzn. z etykietami komórkowymi.

Efekt paczki zwizualizowano za pomocą techniki UMAP (**Figura 2**). W zbiorze zbalansowanym komórki z obu powtórzeń technicznych grupują się zgodnie z badaną zmienną biologiczną, tj. punktem czasowym. W zbiorze zagłuszonym grupy tworzą się w odniesieniu do powtórzeń technicznych, co wskazuje na silny efekt paczki i jego całkowitą dominację nad badanym efektem biologicznym

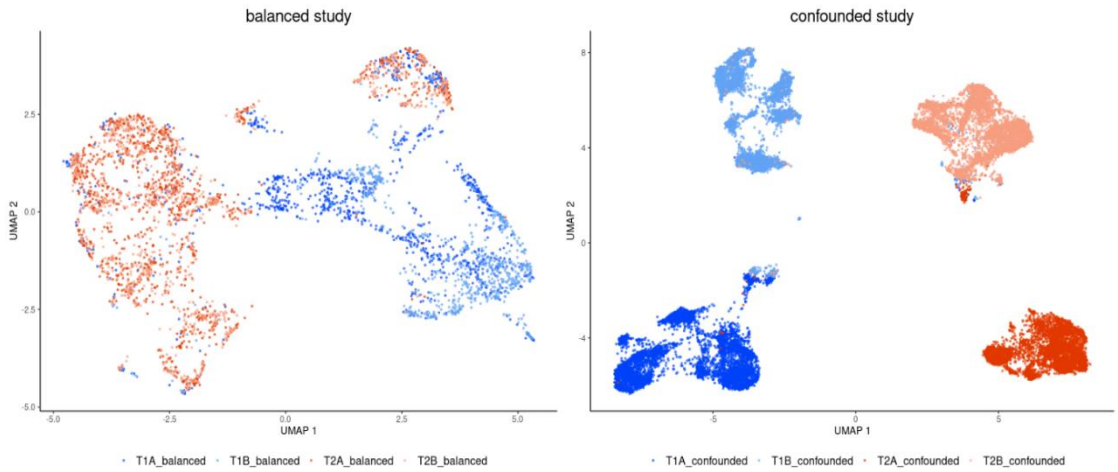

Figura 2. *Wizualizacja obu zbiorów danych za pomocą techniki UMAP*

Ponieważ nie jest możliwa wiarygodna analiza całkowicie zagłuszonego zbioru danych, zastosowano sześć narzędzi korekcji efektu partii w celu rozwiązania tego problemu. Jednak żadne z nich nie dawało zadowalających wyników. Ponadto, oceny wpływu korekty efektu partii ujawniły, że korekcja zniekształca pierwotny rozkład danych (**Figura 3**).

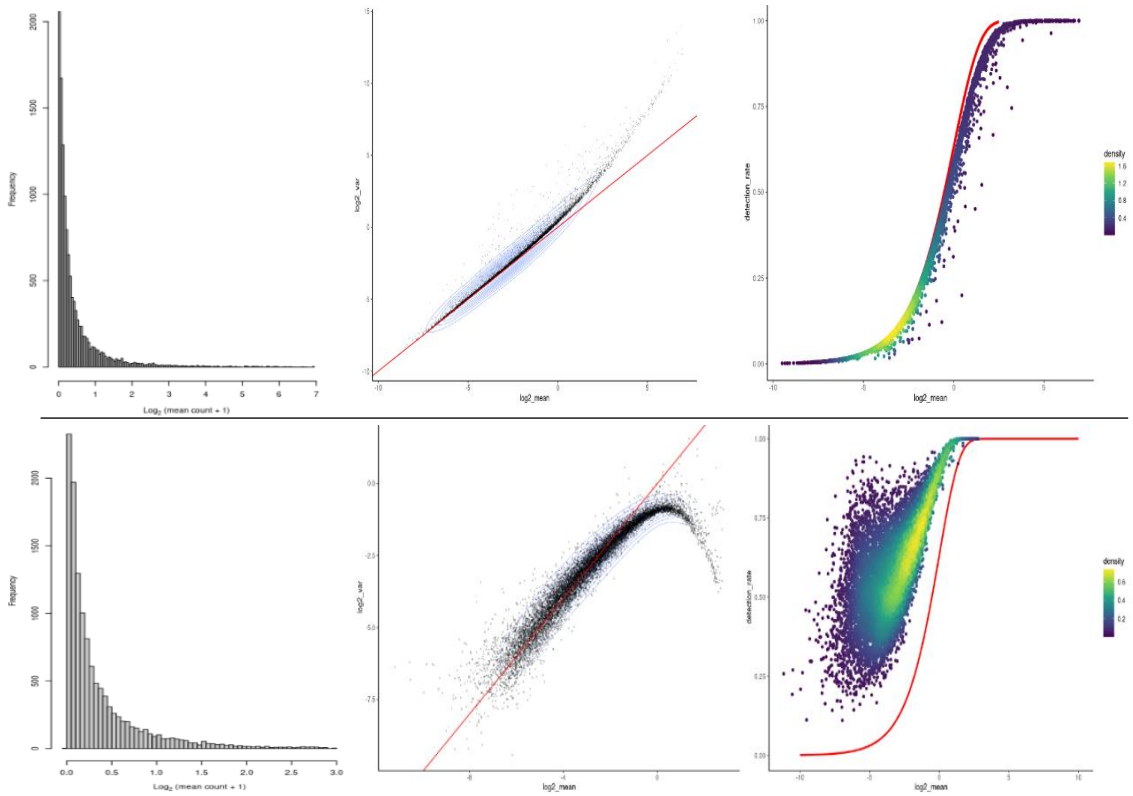

Figura 3*. Analiza dystrybucji ekspresji genów przed korektą (górny panel) oraz po korekcie efektu paczki (dolny panel) .*

Aby umożliwić skonsolidowaną analizę oddzielnie generowanych danych, zaproponowano protokół wykorzystujący iteracyjne grupowanie podprzestrzeni w połączeniu z analizą funkcjonalną (**Figura 4**).

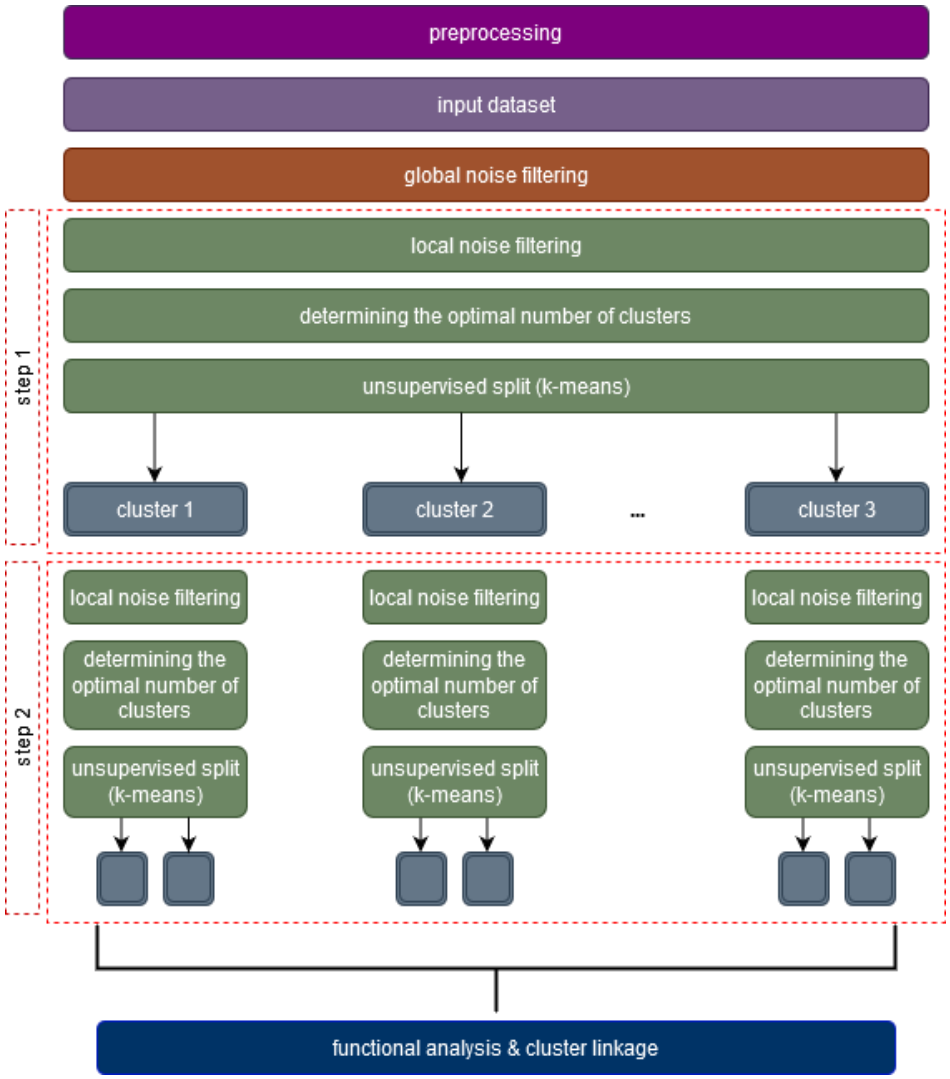

***Figura 4**. Proponowany protokół analizy danych z efektem paczki.*

Pierwszym jego krokiem jest procedura filtracji, która adresuje powszechny problem pomiarów zerowych w danych scRNAseq. Oba zbiory danych wykazywały znaczący poziom globalnego szumu – dropout rate na poziomie 90%. Aby zmniejszyć ten szum, zastosowano globalną filtrację, która polegała na użyciu hierarchicznego grupowania zbinaryzowanej macierzy ekspresji genów. W wyniku tego otrzymano zredukowaną przestrzeń genów (ang. reduced domain). Pełna przestrzeń przed filtracją określana będzie jako ‘full domain’.

Drugi krok filtracji, tzn. selekcja podprzestrzeni odbywała się lokalnie dla każdej zidentyfikowanej grupy. W tym celu zaproponowano dekompozycję wariancji na mieszaniny

gaussowskie (ang. GMM). To podejście zostało zweryfikowane przy użyciu algorytmu sparse k-means, który ma wbudowany mechanizm selekcji cech. Algorytm ten przypisuje większe wagi tym genom, które mają większe znaczenie dla jakości procesu grupowania. Grupowanie w pełnej przestrzeni genów skutkowało rozmytymi klastrami, natomiast grupowanie w podprzestrzeni wyznaczonej metodą GMM zdecydowanie poprawiło rozdzielczość.

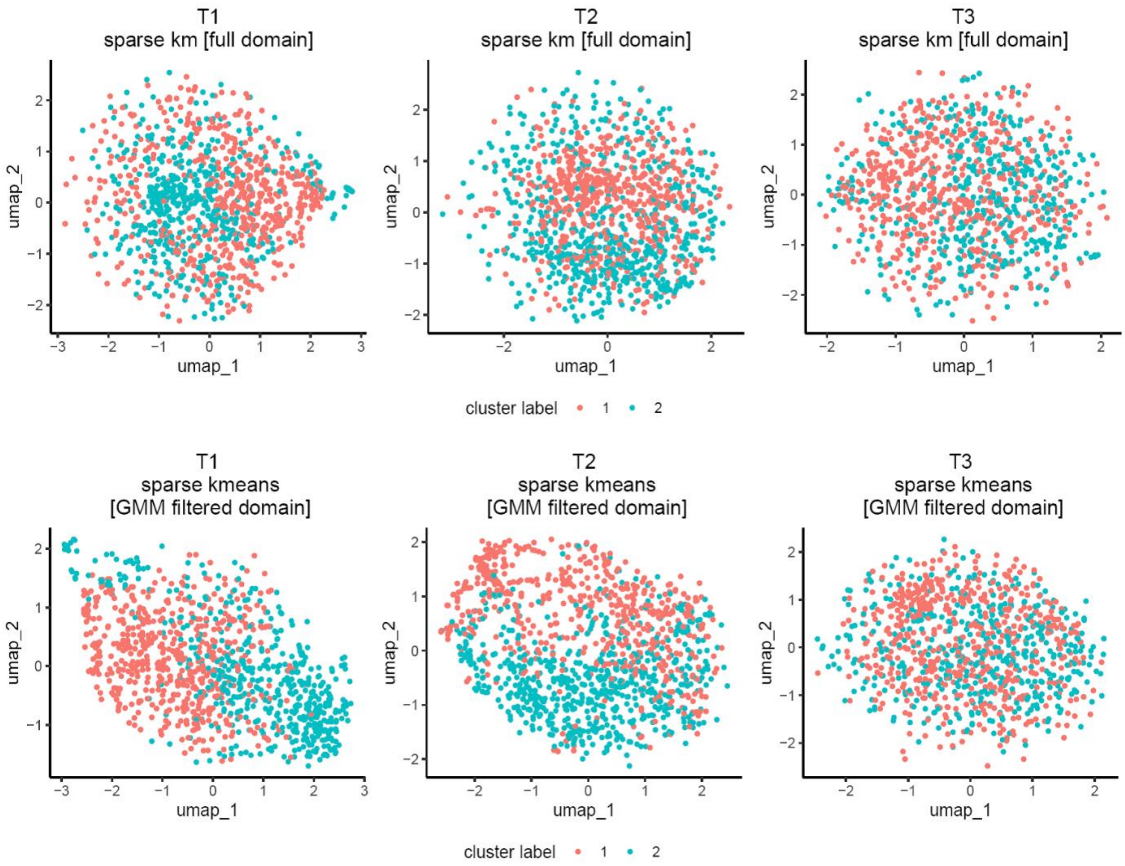

Figure 5. *Weryfikacja strategi selekcji cech opartej na filtrowaniu GMM. full domain – pełna przestrzeń cech przed filtracją, GMM filtered domain – podprzestrzeń wyznaczona metodą GMM*

Analiza dystrybucji wag wykazała większe wagi przypisane przez sparse k-means w grupie genów wyznaczonej metodą GMM (HVG group) w porównaniu z pozostałymi genami, co ostatecznie zweryfikowało zaproponowaną strategię (**Figura 6**)

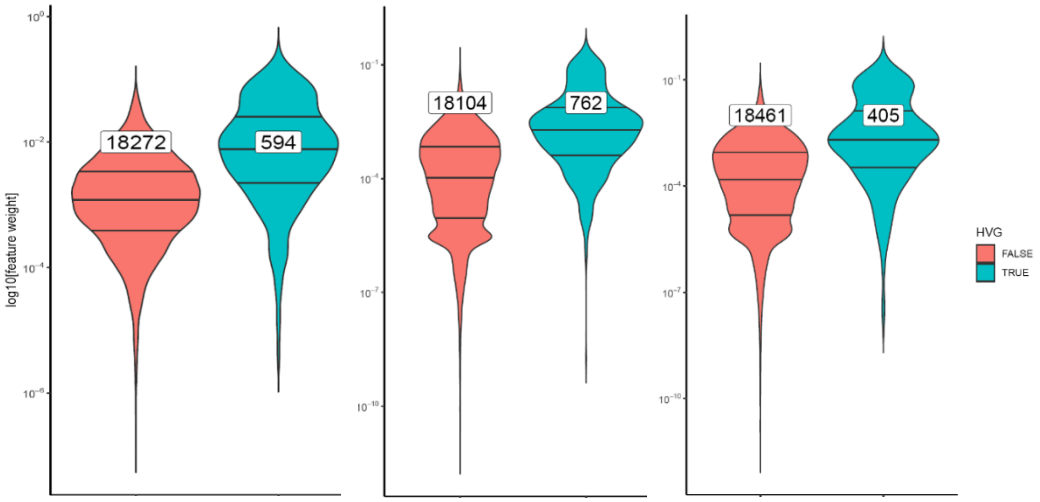

Figura 6. *Dystrybucja wag przypisanych przez sparse k-means w grupie genów wyznaczonej metodą GMM (HVG true) oraz poza nią (HVG false). Podane liczby oznaczają liczbę genów w danej grupie.*

Grupy komórek (klastry) zidentyfikowane w obrębie paczek poddano niezależnej analizie funkcjonalnej. W celu znalezienia ścieżek sygnałowych specyficznych dla danego klastra zaproponowano wykorzystanie miary wielkości efektu – Cliff's delta. Jest to statystyka, która mierzy jak bardzo jedna grupa dominuje nad drugą. Aczkolwiek metrykę tą zaimplementowano według scenariusza 'jeden-do-pozostałych', co oznacza profil funkcjonalny danego klastra porównuwano z „sumą" profili w pozostałych klastrach. Takie podejście miało na celu znalezienie tylko tych ścieżek sygnałowych o silnej manifestacji wykazujących niejako odporność na efekt paczki. Ścieżki zidentyfikowane dla kilku klastrów przedstawiono w **Tabeli 1**.

***Tabela. 1**. Ścieżki sygnałowe o największej (czerwone) i najmniejszej (czarne) wartości statystyki wielkości efektu (ES). Klastry zidentyfikowane w zbiorze zbalansowanym oznaczono na zielono, w zbiorze confounded na pomarańczowo.*

cluster	pathway	ES
T1_I_1	PROGESTERONE_MEDIATED_OOCYTE_MATURATION	0.236
	CELL_CYCLE	0.211
	OOCYTE_MEIOSIS	0.183
	PARKINSONS_DISEASE	-0.081
	SPLICEOSOME	-0.127
	SNARE_INTERACTIONS_IN_VESICULAR_TRANSPORT	-0.153
T1_I_2	CELL_CYCLE	0.188
	PROGESTERONE_MEDIATED_OOCYTE_MATURATION	0.162
	DNA_REPLICATION	0.156
	FATTY_ACID_METABOLISM	-0.153
	CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	-0.186
	ECM_RECEPTOR_INTERACTION	-0.202
T1_II_1	GLYCEROLIPID_METABOLISM	0.104
	SNARE_INTERACTIONS_IN_VESICULAR_TRANSPORT	0.103
	BASE_EXCISION_REPAIR	0.092
	OOCYTE_MEIOSIS	-0.117
	CELL_CYCLE	-0.145
	PROGESTERONE_MEDIATED_OOCYTE_MATURATION	-0.198
T1_II_2	ECM_RECEPTOR_INTERACTION	0.161
	NOTCH_SIGNALING_PATHWAY	0.118
	CELL_ADHESION_MOLECULES_CAMS	0.114
	DNA_REPLICATION	-0.136
	BASE_EXCISION_REPAIR	-0.158
	CELL_CYCLE	-0.171
T2_I_1	CELL_CYCLE	0.250
	PROGESTERONE_MEDIATED_OOCYTE_MATURATION	0.222
	SMALL_CELL_LUNG_CANCER	0.205
	CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	-0.140
	FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	-0.146
	DRUG_METABOLISM_OTHER_ENZYMES	-0.162

cluster	pathway	ES
T1_I_1	SMALL_CELL_LUNG_CANCER	0.122
	MAPK_SIGNALING_PATHWAY	0.092
	CHRONIC_MYELOID_LEUKEMIA	0.074
	OOCYTE_MEIOSIS	-0.114
	CELL_CYCLE	-0.117
	PROGESTERONE_MEDIATED_OOCYTE_MATURATION	-0.159
T1_I_2	HUNTINGTONS_DISEASE	0.113
	OXIDATIVE_PHOSPHORYLATION	0.109
	ALZHEIMERS_DISEASE	0.103
	UBIQUITIN_MEDIATED_PROTEOLYSIS	-0.093
	CHRONIC_MYELOID_LEUKEMIA	-0.109
	PROGESTERONE_MEDIATED_OOCYTE_MATURATION	-0.115
T1_II_1	PROGESTERONE_MEDIATED_OOCYTE_MATURATION	0.249
	OOCYTE_MEIOSIS	0.158
	CELL_CYCLE	0.154
	PARKINSONS_DISEASE	-0.114
	OXIDATIVE_PHOSPHORYLATION	-0.121
	HUNTINGTONS_DISEASE	-0.122
T1_II_2	TGF_BETA_SIGNALING_PATHWAY	0.239
	CELL_CYCLE	0.231
	PROGESTERONE_MEDIATED_OOCYTE_MATURATION	0.182
	VIBRIO_CHOLERAE_INFECTION	-0.119
	SPLICEOSOME	-0.125
	PROTEASOME	-0.155
T2_I_1	PROGESTERONE_MEDIATED_OOCYTE_MATURATION	0.214
	CELL_CYCLE	0.190
	TGF_BETA_SIGNALING_PATHWAY	0.149
	ANTIGEN_PROCESSING_AND_PRESENTATION	-0.106
	ENDOCYTOSIS	-0.111
	PROTEIN_EXPORT	-0.117

Następnie klastry z odpowiadających sobie punktów czasowych w obu zbiorach danych łączono wykorzystując jako miarę podobieństwa dwie metryki: współczynnik korelacji Pearsona oraz miarę określaną jako ‘similarity score’, która jest zwykłym iloczynem skalarnym dwóch wektorów. Im większa wartość tym większe podobieństwo.

W celu wizualnej oceny korespondencji między klastrami z obu zbiorów danych wygenerowano tzw. wykresy Sankeya (**Figura 7**)

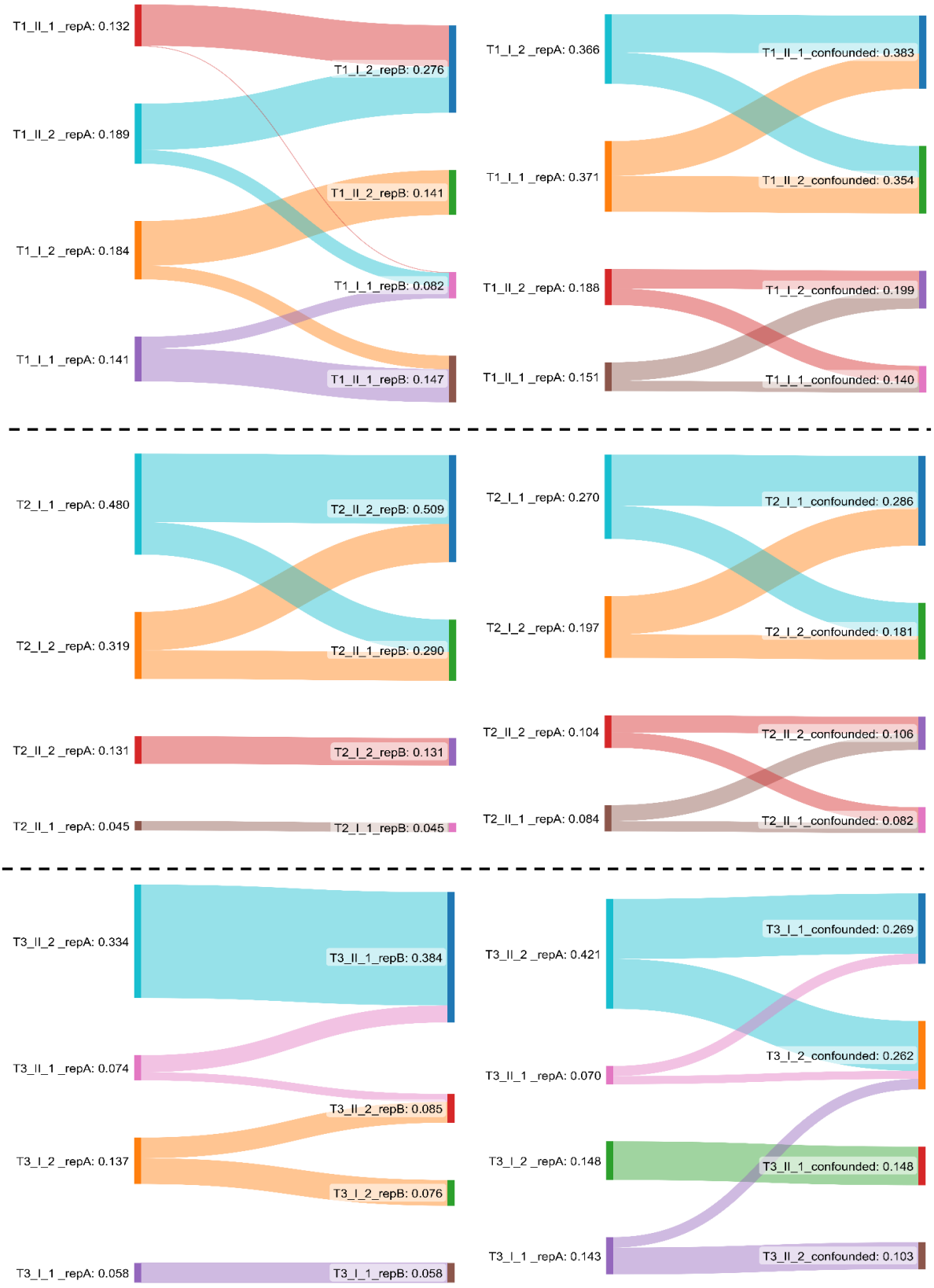

Figura 7. *Wykresy Sankey'a dla porównań między zbiorami danych. Lewa kolumna odpowiada porównaniu powtórzeń A i B balanced. Prawa kolumna odpowiada porównaniu powtórzenia A zbioru zbalansowanego i badania zagłusoznego. Wiersze odzwierciedlają odpowiednie punkty czasowe: góra – T1; środkowy – T2 i dolny – T3. Gęstość przepływów jest proporcjonalna do wartości podobieństwa. Etykieta po prawej stronie nazwy klastra reprezentuje sumę wartości wychodzących z danego klastra (dla węzłów źródłowych) lub sumę wartości wchodzących do danego klastra (dla węzłów docelowych).*

Następnie klastry z obu zbiorów sparowano w taki sposób, że dany klaster mógł utworzyć tylko jedną parę z innym klastrem na podstawie największej wartości miary podobieństwa funkcjonalnego (**Tabela 2**).

***Tabela. 2**. Pary najbardziej podobnych klastrów między zbiorami danych*

repA_vs_repB			repA_vs_confounded		
cluster_repA	cluster_repB	max_sim_score	cluster_repA	cluster_confounded	max_sim_score
T1_II_2	T1_I_2	0.146	T1_I_2	T1_II_1	0.201
T1_I_2	T1_II_2	0.141	T1_I_1	T1_II_2	0.189
T1_I_1	T1_II_1	0.104	T1_II_2	T1_I_2	0.101
T1_II_1	T1_I_1	0.002	T1_II_1	T1_I_1	0.053
T2_I_1	T2_II_2	0.326	T2_I_1	T2_I_1	0.164
T2_I_2	T2_II_1	0.136	T2_I_2	T2_I_2	0.075
T2_II_2	T2_I_2	0.131	T2_II_2	T2_II_2	0.055
T2_II_1	T2_I_1	0.045	T2_II_1	T2_II_1	0.033
T3_II_2	T3_II_1	0.334	T3_II_2	T3_I_1	0.23
T3_I_2	T3_I_2	0.076	T3_I_2	T3_II_1	0.148
T3_I_1	T3_I_1	0.058	T3_I_1	T3_II_2	0.103
T3_II_1	T3_II_2	0.024	T3_II_1	T3_I_2	0.031

Proponowany protokół stawia na prostotę, niski koszt obliczeniowy i łatwość interpretacji. Wszystkie metody tu zastosowane są dobrze ugruntowane i szeroko uznane. Warto jednak zauważyć, że celem, który nie został zdefiniowany w tej rozprawie, było przedstawienie podejścia łatwo zrozumiałego nie tylko dla statystyków, ale także dla biologów odpowiedzialnych za projektowanie eksperymentów.

Aby w pełni zweryfikować proponowane podejście, konieczne są dalsze badania dotyczące wyżej wymienionych zagadnień. Niemniej jednak należy zauważyć, że praca ta nie miała na celu dostarczenia gotowej do użycia metody, ale raczej utorowała drogę nowym kierunkom badań.