

ABSTRACT

Single cell RNAseq experiments are often conducted on a large scale, involving multiple laboratories or measurements taken at different times. Perfectly balanced experimental designs for such large projects may be infeasible, resulting in the need to conduct experiments in batches. Consequently, batch effects inevitably arise. They introduce an additional layer of technical noise to an already noisy scRNAseq data. However, this noise is not uniformly distributed across genomic data features, making it unsuitable to address during the normalization step. If left unaddressed, batch effects can result in misleading conclusions drawn from the analysis. Consequently, computational correction or removal becomes necessary, which is the objective of existing algorithms. Nevertheless, distinguishing batch effects from biological heterogeneity is a challenging task due to their differential origins.

Although batch effects have a detrimental impact on the data, the process of correction for them can also be harmful, particularly at the gene-level. The gene-level analyses are not safe to be performed on corrected data because in most cases correction distorts the original data distribution, and there is lack of a measure to quantify the uncertainty associated with the correction process. Therefore, there is a strong need to develop research in the field of batch effect removal, correction or mitigation employing new approaches and bioinformatic tools.

This work aims to provide a pipeline that utilizes iterative subspace clustering, combined with functional analysis of gene sets, to mitigate the negative impact of the batch effect on scRNAseq data. The novel and central idea was to employ effect size measure to determine cluster-specific pathways, followed by a linkage procedure that enables cluster tracking (linking) across different batches.

Therefore, the proposed workflow eliminates the need for applying batch-effect correction and enables consolidated analysis of batches that were generated separately. In contrast to existing complex and computationally demanding algorithms, this approach prioritizes simplicity, low computational cost, and ease of interpretation. The utilization of subspace clustering combined with functional analysis of gene pathways for mitigating batch effects has not been explored before, making this thesis a novel contribution to the field.

The underlying assumption is that iterative subspace clustering may diminish batch effects by removing more noise from the data with each subsequent iteration. As a result, cells should tend to form groups based on their true biology. Furthermore, the cluster-specific pathways identified are expected to exhibit robust manifestations and demonstrate resilience to the negative impact of batch effects, which is typically less pronounced compared to individual genes.

To address specific challenges encountered in scRNA-seq data, original adjustments were introduced, such as global noise filtration based on binarized gene expression matrix to handle dropouts.

The theses of this dissertation are formulated as follows:

1. Existing algorithms for batch effect correction in scRNAseq often distort the original distribution of gene expression data. Consequently, gene-level analyses such as differential expression or marker identification cannot be safely applied to the corrected dataset.
2. A simple feature selection strategy based on variance decomposition yields similar results to more sophisticated and computationally expensive methods.
3. In confounded scRNA-seq data, batch effect correction can be skipped. Instead, a reliable analysis can be performed by independently identifying subclusters of cells within each batch and then linking them between batches based on the similarity of their functional profiles to track similar cells from different batches.

This project stands out due to its unique experimental setup, which involves a pair of experimentally derived datasets. These datasets shared identical biological properties but differed only in technical study design, as illustrated in **Figure 1**.

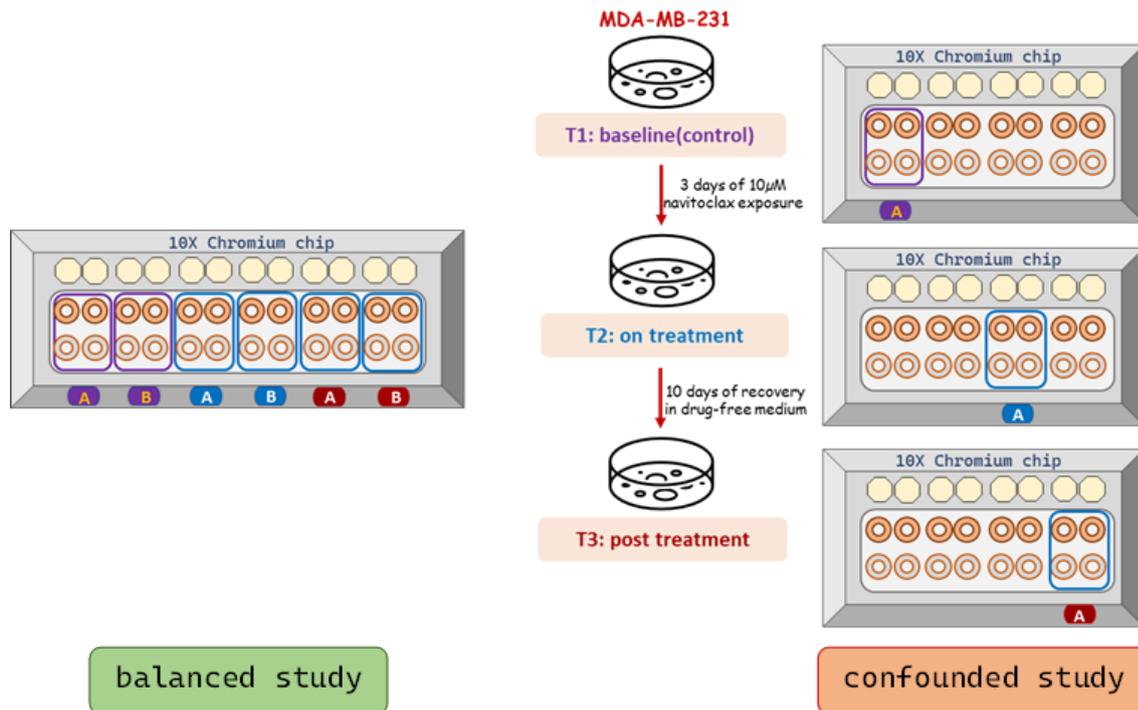


Figure 1. Experimental design. Both experiments conducted to explore the impact of navitoclax treatment on the transcriptome of a triple-negative breast cancer cell line. Both experiments utilized the MDA-MB-231 cancer cell line, and two biological replicates, labelled as A and B, were included. The cells were subjected to a 10 μM concentration of navitoclax and harvested at three specific time points: before the treatment (baseline; T1), after treatment (T2) and after recovery from the treatment (T3).

The first experiment was part of a larger study. However, upon analysis, it was discovered that the experiment exhibited strong batch effects resulting from variations in the experimental processing of the biological groups corresponding to the time of harvesting. In this study, cells collected at different time points were processed on separate chips and on various days. This dataset is referred to as a confounded study.

The second experiment, referred to as a balanced study, was designed to minimize technical variation. In this design, cells collected at different time points were split and processed on the same chip, all on the same day. This approach aimed to ensure that any observed differences were primarily due to the effects of the drug treatment and not influenced by technical factors. This study serves as a reference.

Such experimental setup is distinct from existing evaluations of batch effect correction, which often rely on simulation scenarios or include true cell identity labels. Both datasets (balanced and confounded) were visualized using UMAP plots (**Figure 2**). In the balanced dataset, cells from both repetitions group according to the biological variable of interest (timepoint). However, in the confounded dataset, each technical replicate forms its own cluster. This

indicates that the dataset is completely confounded, with batch effects overpowering the biological variable of interest.

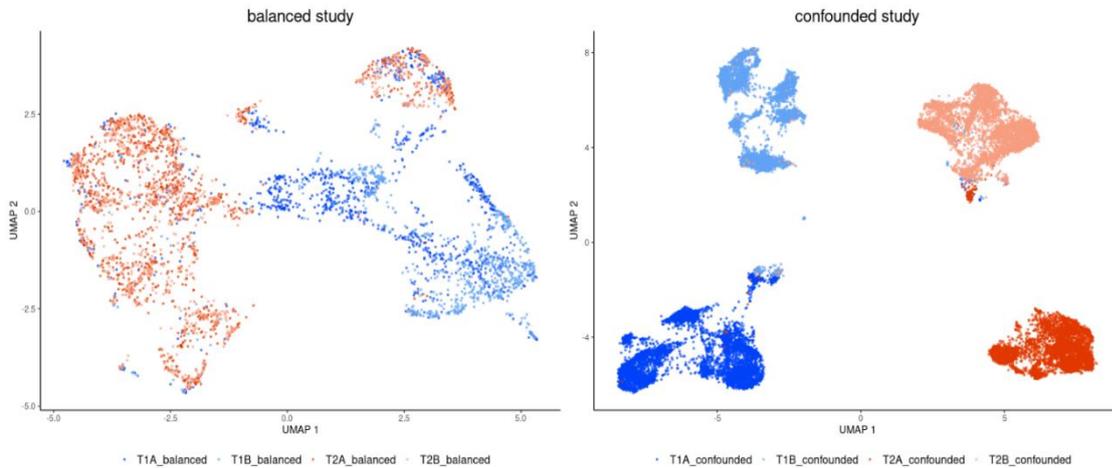


Figure 2. UMAP plots of balanced and confounded study.

Since reliable analysis of such a completely confounded dataset is not possible, six batch-effect correction tools were applied to address this issue. However, non of them gave satisfactory performance. Moreover, the evaluations of the impact of batch-effect correction on feature-level layer of the original data revealed that correction distorts the original data distribution (**Figure 3**).

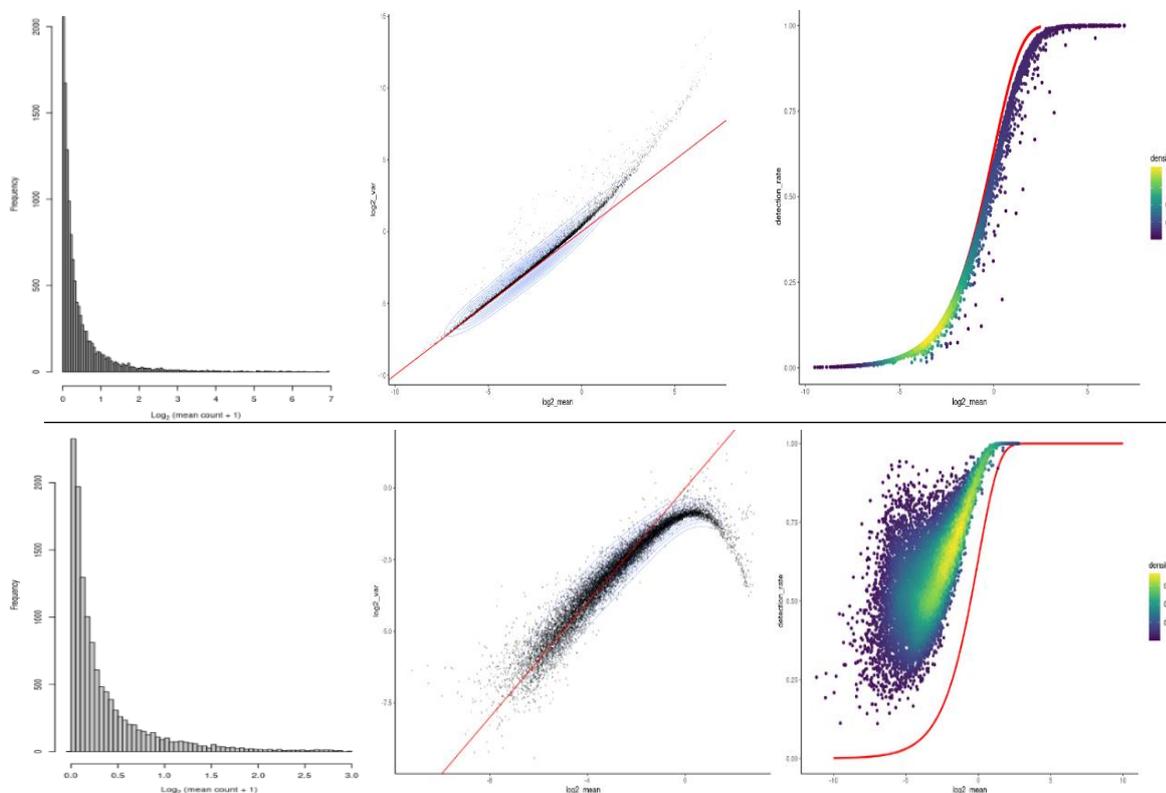


Figure 3. Feature characteristics of confounded study before (top panel) and after batch effect correction (bottom panel). From the left: (i) histogram of average gene expression, (ii) scatter plot of variance vs mean expression (red line with intercept = 0 and slope = 1) and (iii) detection rate vs average expression (red line indicates the expected distribution under a Poisson model). Individual points are colored by the number of neighboring points).

To facilitate the consolidated analysis of separately generated data, a pipeline utilizing iterative subspace clustering combined with functional analysis was proposed (**Figure 4**).

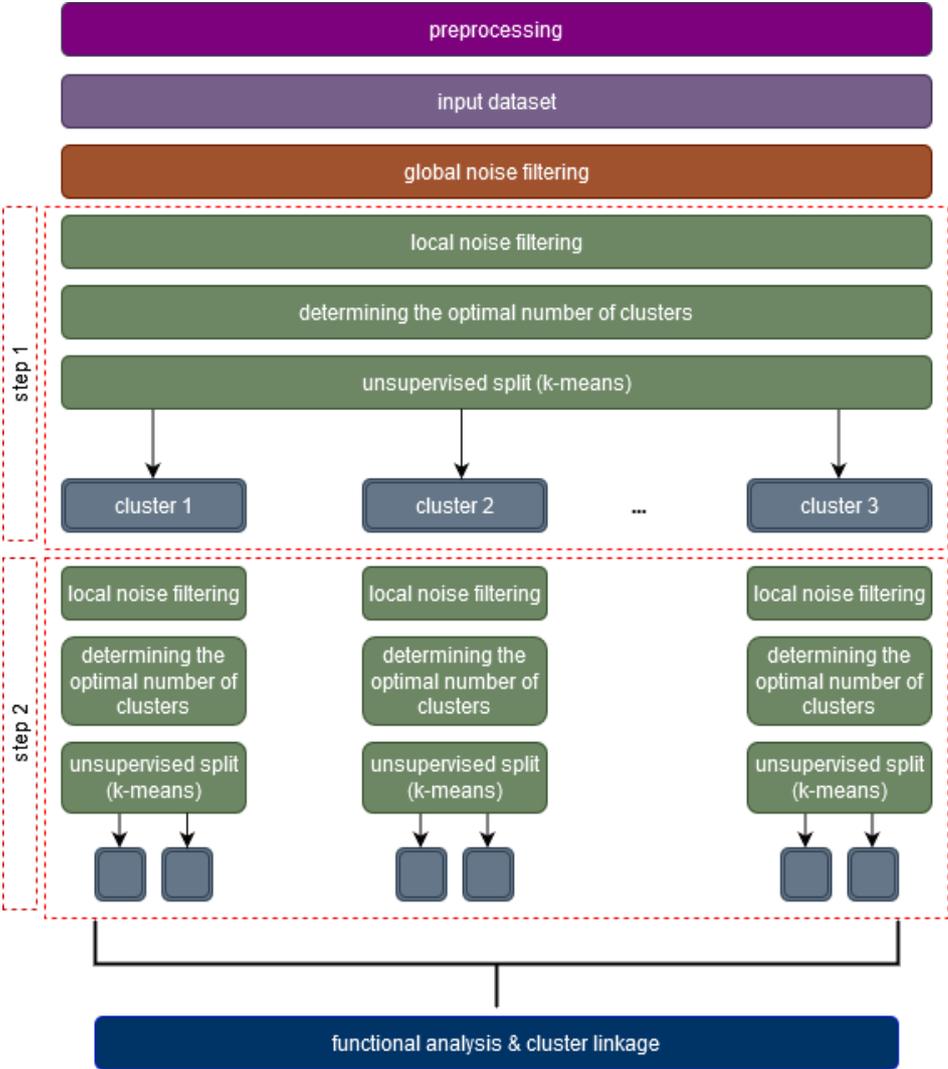


Figure 4. The proposed framework for analysis of confounded datasets.

The first step involves a filtering procedure that addresses the prevalent issue of zero measurements in scRNAseq data. All datasets exhibited a significant level of global noise, which was manifested in high dropout rates exceeding 90%. Many genes were rarely detected in any cell, with a dropout rate close to 1. To filter out noisy genes in a data-driven manner, a method based on hierarchical clustering was proposed. Global filtration resulted in the reduced domain of features. In contrast, the term 'full domain' is used to describe the original feature space before any filtration occurred.

In the second step of feature selection, a filtering strategy was applied locally to each cluster obtained at every clustering iteration. Local feature selection was performed through decomposition of gene variances into mixture of gaussian components.

The GMM filtration strategy was validated by sparse k-means clustering which incorporates a feature selection procedure. In other words, it internally assesses the importance of each gene in the clustering process by assigning higher weights to more significant genes. When the full domain is considered, the clusters are blurred due to the presence of many noise features that do not contribute to the clustering process (**Figure 5**). After GMM filtration, the quality of clustering substantially improved, and the clusters became more distinct.

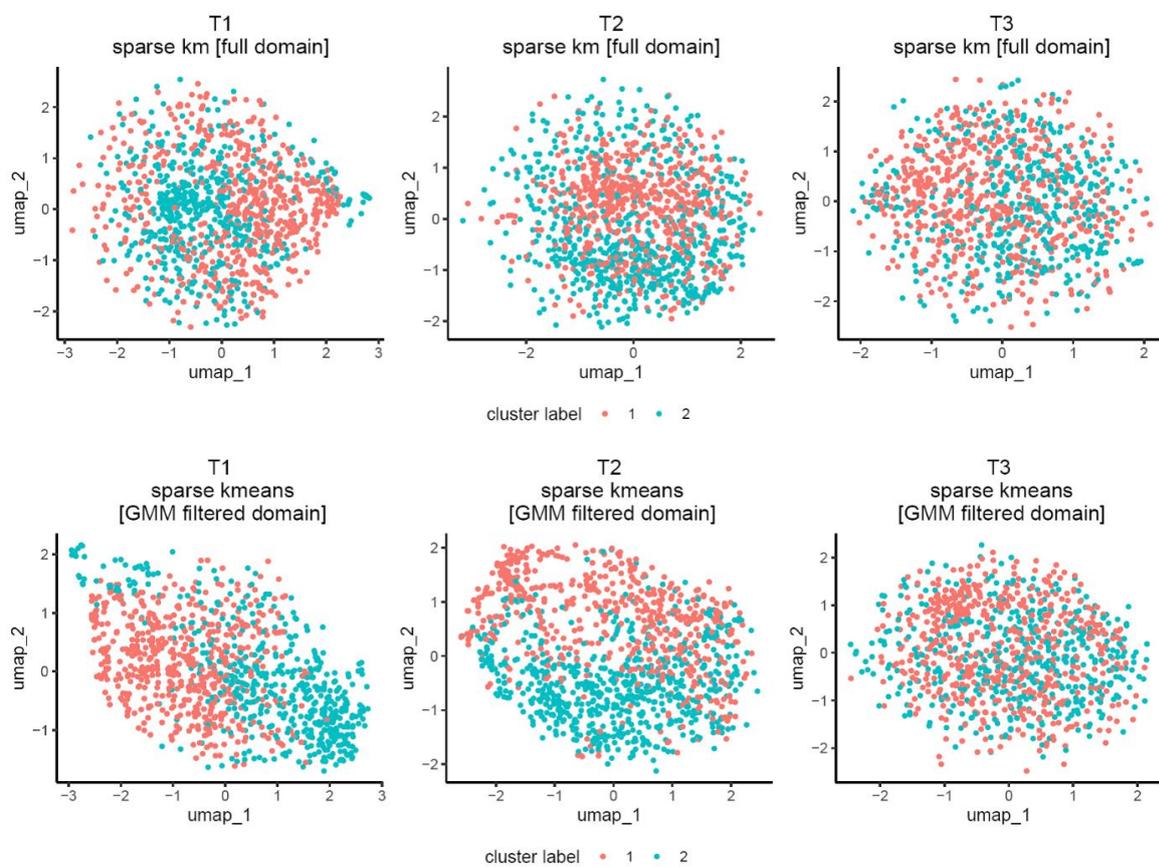


Figure 5. Sparse k-means clustering with different scenarios. full domain – corresponds to the original feature space before global filtration, GMM filtered domain – after variance decomposition by GMM

The evaluation of the distribution of weights assigned by sparse k-means revealed that the weights in the GMM filtered group of genes, called HVG group were substantially higher, despite comprising only a small number of genes, constituting 2-3% of the full domain (**Figure 6**). The above analyses proved that the proposed strategy of local feature selection yields similar results to more sophisticated and computationally expensive methods.

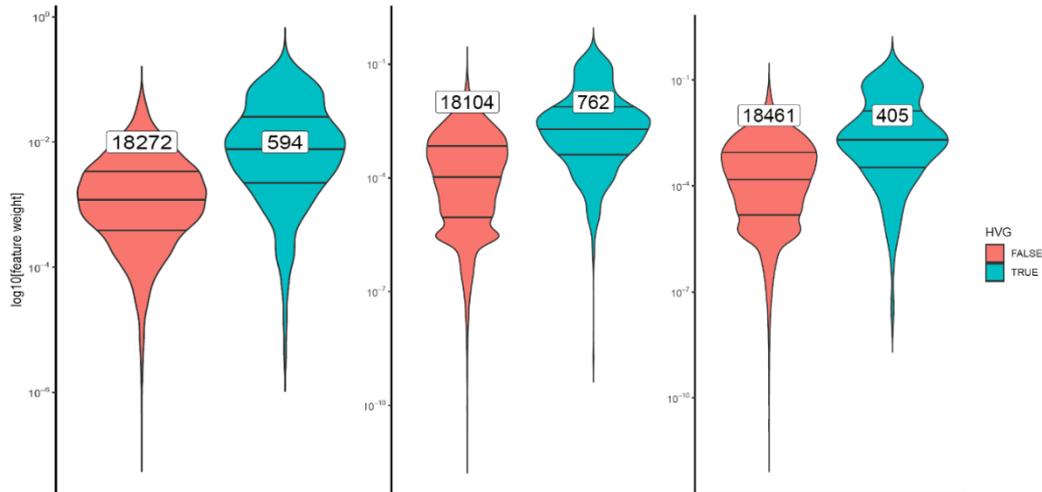


Figure 6. Distribution of feature weights assigned by sparse k -means. Two groups were considered: involving only HVGs obtained by GMM filtering (green), and non-HVG group (red). Weights were assigned automatically by the algorithm. There is a number of features depicted inside of each violin plot.

Clusters discovered within batches were subsequently subjected to independent functional analysis of gene sets. To find cluster specific pathways for each dataset, Cliff's delta effect size statistics was proposed. This metric quantifies the extent to which values in one group are larger (dominate) than the values in a second group. However, these groups were determined according to one-versus-others scenario, which was designed to identify pathways with a robust manifestation. These pathways were assumed to demonstrate resilience to the negative impact of batch effects, which is generally less pronounced compared to individual genes.

Table 1. Top three pathways with the highest (marked in red) and lowest effect size (ES) for sample clusters in balanced (green) and confounded study (orange)

| cluster | pathway | ES | cluster | pathway | ES |
|---------|---|--------|---------|---|--------|
| T1_I_1 | PROGESTERONE_MEDIATED_OOCYTE_MATURATION | 0.236 | T1_I_1 | SMALL_CELL_LUNG_CANCER | 0.122 |
| | CELL_CYCLE | 0.211 | | MAPK_SIGNALING_PATHWAY | 0.092 |
| | OOCYTE_MEIOSIS | 0.183 | | CHRONIC_MYELOID_LEUKEMIA | 0.074 |
| | PARKINSONS_DISEASE | -0.081 | | OOCYTE_MEIOSIS | -0.114 |
| | SNARE_INTERACTIONS_IN_VESICULAR_TRANSPORT | -0.153 | | CELL_CYCLE | -0.117 |
| | | | | PROGESTERONE_MEDIATED_OOCYTE_MATURATION | -0.159 |
| T1_I_2 | CELL_CYCLE | 0.188 | T1_I_2 | HUNTINGTONS_DISEASE | 0.113 |
| | PROGESTERONE_MEDIATED_OOCYTE_MATURATION | 0.162 | | OXIDATIVE_PHOSPHORYLATION | 0.109 |
| | DNA_REPLICATION | 0.156 | | ALZHEIMERS_DISEASE | 0.103 |
| | FATTY_ACID_METABOLISM | -0.153 | | UBIQUITIN_MEDIATED_PROTEOLYSIS | -0.093 |
| | CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION | -0.186 | | CHRONIC_MYELOID_LEUKEMIA | -0.109 |
| | | | | PROGESTERONE_MEDIATED_OOCYTE_MATURATION | -0.115 |
| T1_II_1 | GLYCEROLIPID_METABOLISM | 0.104 | T1_II_1 | PROGESTERONE_MEDIATED_OOCYTE_MATURATION | 0.249 |
| | SNARE_INTERACTIONS_IN_VESICULAR_TRANSPORT | 0.103 | | OOCYTE_MEIOSIS | 0.158 |
| | BASE_EXCISION_REPAIR | 0.092 | | CELL_CYCLE | 0.154 |
| | OOCYTE_MEIOSIS | -0.117 | | PARKINSONS_DISEASE | -0.114 |
| | CELL_CYCLE | -0.145 | | OXIDATIVE_PHOSPHORYLATION | -0.121 |
| | | | | HUNTINGTONS_DISEASE | -0.122 |
| T1_II_2 | ECM_RECEPTOR_INTERACTION | 0.161 | T1_II_2 | TGF_BETA_SIGNALING_PATHWAY | 0.239 |
| | NOTCH_SIGNALING_PATHWAY | 0.118 | | CELL_CYCLE | 0.231 |
| | CELL_ADHESION_MOLECULES_CAMS | 0.114 | | PROGESTERONE_MEDIATED_OOCYTE_MATURATION | 0.182 |
| | DNA_REPLICATION | -0.136 | | VIBRIO_CHOLERAE_INFECTION | -0.119 |
| | BASE_EXCISION_REPAIR | -0.158 | | SPLICEOSOME | -0.125 |
| | | | | PROTEASOME | -0.155 |
| T2_I_1 | CELL_CYCLE | 0.250 | T2_I_1 | PROGESTERONE_MEDIATED_OOCYTE_MATURATION | 0.214 |
| | PROGESTERONE_MEDIATED_OOCYTE_MATURATION | 0.222 | | CELL_CYCLE | 0.190 |
| | SMALL_CELL_LUNG_CANCER | 0.205 | | TGF_BETA_SIGNALING_PATHWAY | 0.149 |
| | CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION | -0.140 | | ANTIGEN_PROCESSING_AND_PRESENTATION | -0.106 |
| | FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS | -0.146 | | ENDOCYTOSIS | -0.111 |
| | | | | PROTEIN_EXPORT | -0.117 |
| | | | | | |

Following functional analysis, clusters from corresponding timepoints in the reference and confounded datasets are linked based on the similarity of their functional profiles.

Two variants of scoring function were utilized to enable cluster linkage: based on Pearson's correlation coefficient and based on metric called similarity score, which is simple dot product of two vectors. Both the similarity score and Pearson's correlation are related; however, the former focuses on representing the alignment between vectors, while the latter represents the strength and direction of the linear relationship between the variables. A larger dot product indicates a stronger alignment.

To visually track clusters across timepoints, Sankey diagrams were generated for each dataset based on the two similarity metrics under consideration. The Sankey plot effectively illustrates the flow of clusters between datasets (**Figure 7**).

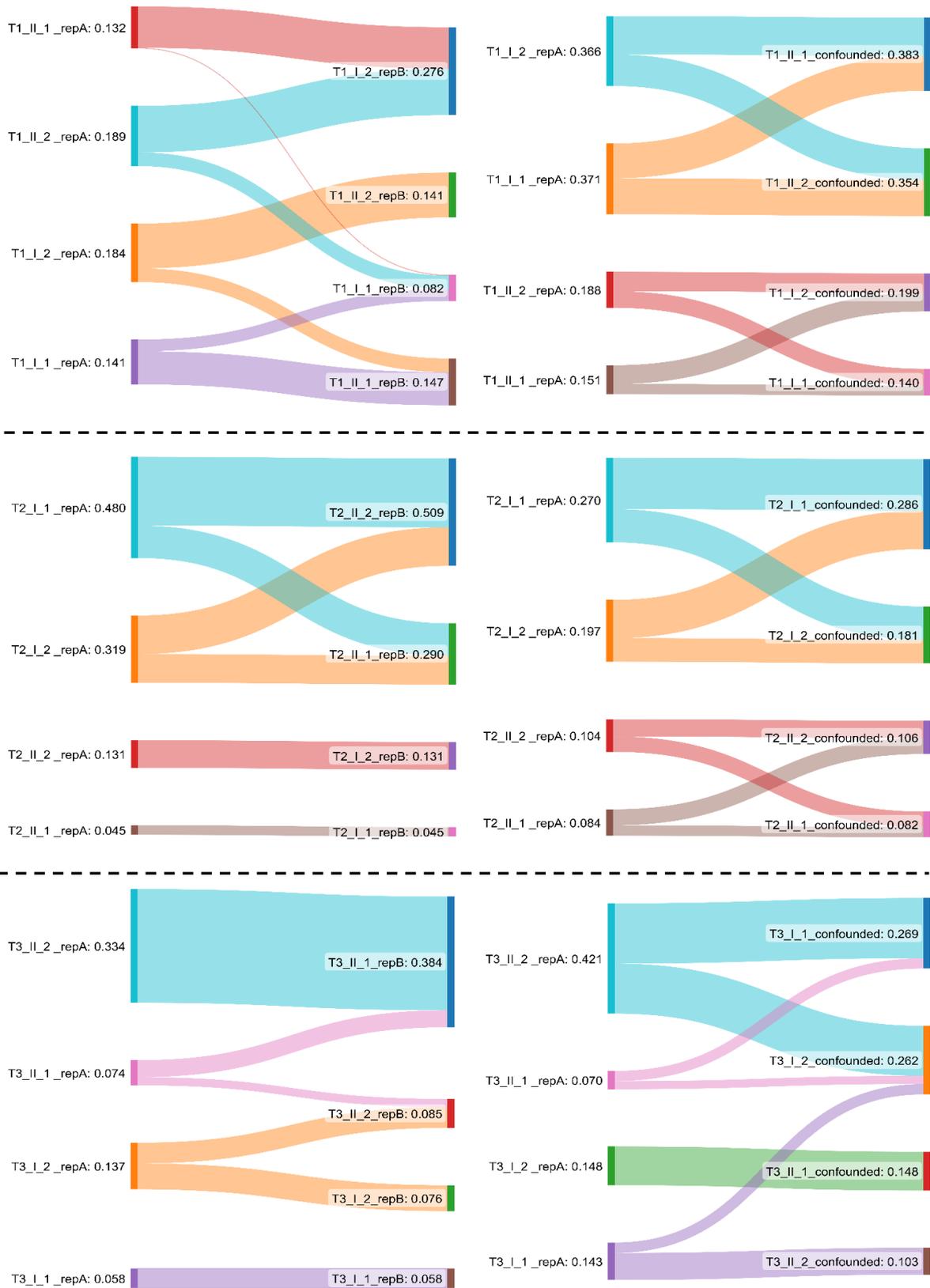


Figure 7. Sankey diagrams for between datasets comparisons. Left column corresponds to the comparison between repetition A and B of balanced study. Right column corresponds to the comparison between repetition A of balanced study and confounded study. Rows reflect the corresponding timepoints: top – T1; middle – T2 and bottom – T3. The thickness of flows is proportional to the value of similarity score. The label on the right of the cluster name represents the sum of the values coming out of the given cluster (for source nodes) or the sum of the values coming in the given cluster (for target nodes).

Based on maximization approach, only clusters with the highest positive similarity metric (separately for correlation and similarity score) were paired, allowing each cluster to form only one pair. (**Table 2**). These pairs of clusters can be considered the most similar between timepoints (batches).

Table 2. The most similar pair of clusters between batches

| 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 |

The proposed workflow prioritizes simplicity, low computational cost, and ease of interpretation. All the methods employed in this pipeline are well-established and widely recognized in the field. However, it is worth noting that the goal, which was not initially introduced in this dissertation, was to provide an approach that is easily understandable not only for statisticians or data analysts but also for biologists responsible for designing such experiments.

To fully validate the proposed approach, further research is necessary, addressing the aforementioned issues. Nonetheless, it is important to note that this work aimed not to provide a ready-to-use method but rather to pave the way for new directions in research