UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA BIOMEDICA

Comparison of single cell RNA sequencing data integration methods with application to breast cancer data

Relatore: Prof. Giacomo Baruzzo
Correlatrice: Dott.ssa Giulia Cesaro

Laureando/a: Arianna Zuanazzi
2001520

ANNO ACCADEMICO 2023 – 2024

Data di laurea: 19 luglio 2024

Index

Abstract - Italiano

La bioinformatica è una disciplina che utilizza metodi computazionali per raccogliere ed analizzare dati biologici. L'avvento delle tecnologie di sequenziamento dell'RNA a singola cellula ha aperto nuove possibilità di ricerca ed esplorazione nella biomedicina, consentendo in particolar modo lo studio sempre più approfondito di cancri e tumori e l'acquisizione di sempre più dati ad alta risoluzione. Tali dati vengono però spesso acquisiti in diversi contesti e con diverse metodologie e condizioni laboratoriali, esponendoli quindi a bias tecnici e ad elevati livelli di rumore. Questo richiede lo sviluppo di metodi standardizzati per l'integrazione.

Lo scopo di questo progetto di tesi è stato quello di confrontare quattro metodologie diverse di integrazione dati. I dataset utilizzati contengono l'espressione genica di campioni di cancro al seno, con diversi gradi di severità e metastasi. Al fine di fornire una valutazione oggettiva dei vari aspetti qualitativi dell'integrazione sono state implementate delle metriche statistiche e computazionali. È stata anche sviluppata una funzione il cui scopo sia di effettuare un benchmarking completo della procedura di integrazione e riportare visualizzazioni.

La tesi inizia con un'introduzione al cancro al seno e alle tecnologie di sequenziamento dell'RNA a singola cellula. A seguito verranno descritti nel dettaglio di dataset impiegati nello studio e le loro caratteristiche, nonché la metodologia di elaborazione impiegata. Successivamente saranno esplicate le quattro metodologie di integrazione dati implementate e le metriche impiegate nel processo di benchmarking. Infine verranno presentati i risultati, sia singolarmente che mettendo a confronto i quattro metodi.

Complessivamente, Harmony mostra le migliori prestazioni nel preservare le informazioni biologiche e, sebbene le sue prestazioni nella rimozione di bias tecnici siano le più basse tra i quattro algoritmi, i punteggi rilevanti restano comunque paragonabili a quelli ottenuti dagli altri algoritmi. FastMNN si dimostra il secondo migliore nel preservare la varianza biologica, sebbene le metriche indichino la persistenza di bias tecnici residui, in particolare specifici ai singoli campioni. RPCA mostra prestazioni equilibrate sia nella conservazione dell'informazione biologica sia nella rimozione di bias tecnici. LIGER mostra i migliori risultati nella rimozione dei bias tecnici, ottenendo tuttavia il punteggio peggiore per quanto riguarda la conservazione delle informazioni biologiche.

Abstract - English

Bioinformatics is a discipline that employs computational methods to collect and analyze biological data. The emergence of single-cell RNA sequencing technologies has opened up new avenues for research and enquiry in biomedicine, in particular by enabling the increasingly in-depth study of cancers and tumors and the acquisition of more and more high-resolution data. However, this data is often acquired in different contexts and under different protocols and laboratory conditions, thus exposing it to technical bias and high levels of noise. This issue requires the development of standardized methods for data integration.

The purpose of this thesis project is to compare four different data integration methodologies. The datasets used contain the gene expression profiles from different breast cancer samples, characterized by varying degrees of severity and metastasis. A panel of statistical and computational metrics was implemented, in order to provide an objective evaluation of the various qualitative aspects of integration. Additionally, a function was developed whose purpose is to comprehensively benchmark the integration procedure and provide useful visualizations.

The thesis begins with an introduction to breast cancer and single-cell RNA sequencing technologies. After that, a detailed description of datasets used in the study and of the processing methodology will be provided. Next, an elaboration will be provided on the four data integration methodologies implemented and the metrics employed in the benchmarking process. Finally, the results will be presented, both individually and by comparing the four methods.

Overall, Harmony displays the best performance at preserving biological information, and although its performance at removing batch effects ranks lowest among the four algorithms the relevant scores still mostly parallel those achieved by the other techniques. FastMNN demonstrates to be the second best at preserving biological variance, although the metrics point to a persistence of technical batches and specifically of sample-specific bias. RPCA exhibits a balanced performance in both biological information preservation and removal of batch effects. LIGER showcases the best results in removal of technical bias, scoring however the worst pertaining to preservation of biological information.

Introduction

Overview of breast cancer

The breast (also called mammary gland) is one of a pair of glandular organs located in the upper part of the chest. It is primarily composed of subcutaneous adipose tissue, which surrounds a network of lactiferous ducts lined by columnar epithelial cells in a matrix containing fibroblasts, adipocytes, endothelial, and immune cells[1]. These ducts end with lobules, clusters of alveoli where milk production takes place, and converge on the nipple. From these tissues breast cancer may arise through a process called carcinogenesis, a multifactorial and multistep process driven by genetic and environmental factors in which genetic alterations drive the progressive transformation of normal human cells into highly malignant derivatives[2,3].

Breast cancer ranks second in incidence and fourth in mortality among all cancers worldwide, and it is the most common cancer in women, with an estimated 2.296.840 new diagnoses and 666.103 deaths in 2022[4]. In Italy it is the most common kind of cancer both for women and in total, with an estimated 57.480 new diagnoses and 15.455 deaths in 2022, and women carry a 13.2% cumulative risk of developing breast cancer during their lifetime and an 8.0% risk of dying due to it[5]. Additionally, while the net survival ratio at 5 years after diagnosis is high, at around 88%, the high incidence means that there are around 834.200 women with a past breast cancer diagnosis[6].

Depending on molecular and histological profile breast cancer can be classified in 3 main groups: one expressing hormone receptors (either estrogen (ER+) or progesterone (PR+)), one expressing human epidermal receptor 2 (HER2+), and triple negative one (TNBC). This last group may be further divided in 6 subgroups: basal-like 1 (BL-1), basal-like 2 (BL-2), immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL) and luminal androgen receptor (LAR)[7]. Compared to other cancers, breast cancer has a comparatively high pattern of inheritance, with 5-10% of all cases following a Mendelian inheritance pattern and 15-20% a familial one[8]. The BRCA1 and BRCA2 mutations are particularly well studied since, though they only have a combined frequency of about 0.4%, have a high penetrance, comprising more than a third of all hereditary breast cancers and resulting in a lifetime risk of breast cancer for the carriers between 60% and 85%[9].

Depending on stage and on cancer subtype, treatment options may fall into three main groups: surgery, radiation, and medication. Surgical options involve the direct removal of cancerous tissue and are usually either breast-conserving surgery or mastectomy; these options may be applied alone in early stages of breast cancer or in conjunction with other treatments. Radiation therapy entails the irradiation with high levels of either X-rays or protons; it is often applied after surgery to ensure the complete elimination of cancerous cells and minimize the risk of recurrence. Medication options can be further divided in 3 more categories: chemotherapy, hormonal therapy and biological therapy. Chemotherapy is the use of cytotoxic drugs to kill cancer cells; while it contributes significantly to reduced mortality its side effects are often very significant[2]. Hormonal therapy entails the use of either hormones or drugs to lower hormone levels and block breast cancer cells to be stimulated by them, therefore making it most effective with ER+ and PR+ breast cancers. Biological therapy is the use of biological compounds to fight cancer, of which immunotherapy is for the purposes of this thesis its most interesting subtype: that involves the engineering of immune cells to attack breast cancer inside the body.

Breast cancer is a highly heterogeneous disease and both the tumor microenvironment and variations in mutation patterns within cells can significantly affect disease progression, prognosis, and treatment efficacy. Modern advancements in sequencing technology and computer science are now allowing increasingly higher resolution in the analysis of gene expression at lower costs, enabling a deeper understanding of tumorigenesis and making it possible to know the breast cancer gene expression specific for each patient. This is leading to important developments for personalized medicine in oncology and is assisting in bringing it into common clinical practice[10,11].

Single-cell RNA sequencing

In the life sciences, the word “sequencing” refers to the determination of the primary structure of a biopolymer, which could be either a protein or a string of DNA or RNA. Specifically, genomic sequencing refers to the determination of the sequences of bases of a nucleic acid, from which the complete set of genes (in case of DNA sequencing) and eventually of their transcription into proteins (in case of RNA sequencing) can be extracted.

The history of sequencing technologies starts in 1968, when Wu and Keiser first made use of primer extension to sequence 12 bases of the DNA of a bacteriophage[12]. 1977 saw the birth of Sanger sequencing, the first method that could be reliably used on a wide scale to reconstruct long DNA

sequences[13]. This technique, based on electrophoresis, remained the predominant one until the 1990s, when massive developments in life science technologies and increases in computational power gave rise to “Next Generation” (NGS) sequencing. They became commercially available in 2005 and, through enzymatic methods, enabled the massive parallel sequencing of samples and the execution of significantly larger-scale genomics projects[14]. In the year 2001 a landmark breakthrough occurred in genomics, when the Human Genome Project Consortium published the preliminary draft of the entire euchromatic human genome[15].

Single-cell omics emerged in 2009 with the publication by Tang et al., marking the first successful transcriptome analysis of a single cell and the introduction of scRNA-seq[16]. Although there may be variations in the specific materials and steps employed, the vast majority of techniques follow a similar procedure[17] (see Figure 1). First, cells are mechanically separated and isolated, using either micro-dissection, droplet-based, or microfluidics-based platforms. Subsequently, the cells are lysed in order to extract as much polyadenylated mRNA as possible. The mRNA is then reverse-transcribed into complementary DNA (cDNA). Polymerase chain reaction (PCR) is employed to amplify the cDNA, thereby generating sufficient material for downstream analysis. This material is subsequently associated with its own unique cell through a process referred to as "barcode-tagging"[18]. The cDNA is then sequenced using NGS technologies. Library generation techniques and genomic alignment tools are employed to convert the cDNA samples into fragments that can be sequenced by the instrumentation and map the sequencing reads to the reference genome. At this point the gene expression data is obtained and ready to be processed.

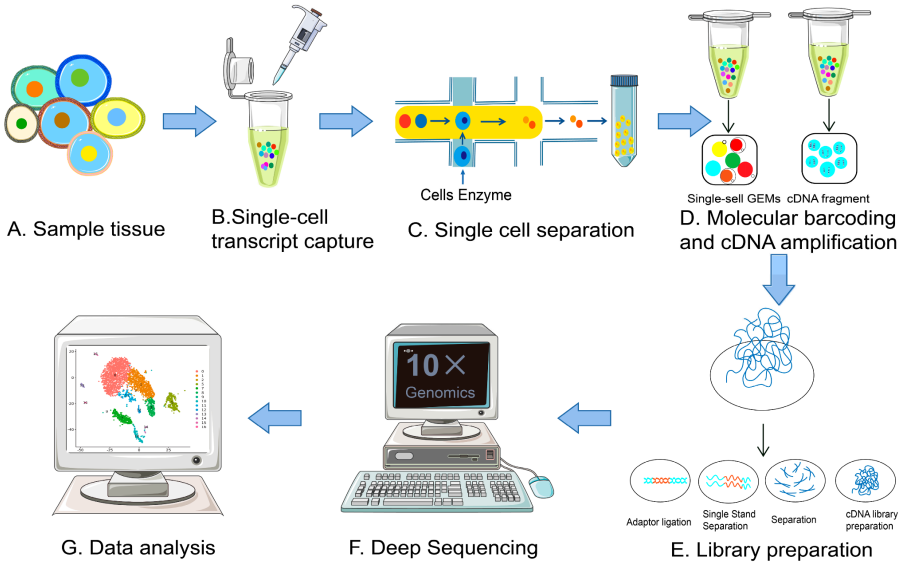

Figure 1: The figure shows the workflow of scRNA-Seq, which includes the following parts: single-cell isolation, reverse transcription, cDNA synthesis, single-cell library, high-throughput sequencing, and data analysis. Image taken from “The Evolution of Single-Cell RNA Sequencing Technology and Application: Progress and Perspectives.”, by Wang, S.; Sun, S.-T.; et al., Int. J. Mol. Sci. 2023

scRNA-seq offers several advantages over bulk sequencing. While bulk sequencing can only detect the average gene expression across the sample, scRNA-seq can detect the heterogeneity in gene expression across cells[19]. Most importantly, it enables the study of gene expression at a much higher resolution, allowing the identification of rare cell populations and the characterization of their unique transcriptional profiles. This makes it possible to reconstruct the complex cellular environment of a sample and identify internal variance[20]. The specific design of scRNA-seq allows for the sequencing of theoretically any kind of eukaryotic cell[21]. There are however some limitations. As it requires the isolation of every cell from each other, this technique is challenging to implement with mechanically hard or very elastic tissues. Additionally, it is less efficient in capturing RNA and provides a lower depth count of genes, requiring higher amplification and less ability to analyze weakly expressed genes. Furthermore, this technology is highly susceptible to dropout and reduplication events, as well as to minimal differences in protocol choice and environmental conditions, making the data very noisy[17].

scRNA-seq has been widely applied to all fields of medical research. For example, it has proven to be invaluable in cancer research. It is enabling the comprehensive characterization of the tumor microenvironment and the study of the genetic mechanisms behind the tumor evasion of the immune system and its drug response, as well as tumor growth and proliferation into the healthy tissue[22]. Furthermore, it is also contributing to the study of the pathogenesis of immune diseases and of the host immune response against bacterial and viral infections[19]. Moreover, scRNA-seq is enabling the identification of new molecular targets for drug development and precision medicine purposes, in particular for malignant cancers[23]. Finally, it is being used in embryonic and organ development research, by characterizing the development of the distinct cell lineages and types spatially and over time[24].

With the increase of data available it is now becoming possible to undertake large-scale projects. However, due to the characteristics of the technology and of the biological material, the scRNA-seq samples often present strong internal variations that are unrelated to the internal biological variance and are instead due to differences in technical protocols or environmental conditions at the time of sampling[25]. These unwanted variations, called batch effects, introduce technical bias that hinders the ability to conduct meaningful analysis. In order to merge the data together while eliminating batch effects, data integration algorithms have been developed, designed to remove the technical bias and cluster similar cells together.

Project activities and goals

My internship project took place from the 1st of March to the 31th of May 2023, through the Erasmus+ for Traineeship programme. It was carried out at the Digital Science Center (DiSC), with the Computational Biomedicine research group of the University of Innsbruck, under Professor Francesca Finotello. The group is highly interdisciplinary and focuses on the analysis of bulk and single-cell genomic data and on the development of computational methods for precision medicine, with a special focus on cancer immunology.

My project was in the field of bioinformatics, specifically in single cell omics, a branch of bioinformatics focussing on the detection and analysis of high-resolution biological molecules at a cellular level. My goals throughout this project were:

- To select and process labeled breast cancer scRNA-seq datasets and associated metadata from different human studies
- To integrate the datasets according to 4 different integration algorithms
- To benchmark on breast cancer data the integration quality of 4 different data integration techniques
- To document my scientific finding with a report and a powerpoint and present my work to my colleagues

The integration techniques applied were the following:

- Reciprocal Principal Component Analysis (RPCA)
- Fast Mutual Nearest Neighbour correction (FastMNN)
- Harmony
- Linked Inference of Genomic Experimental Relationships (LIGER).

The benchmarking was performed according to their fit in removing batch effects and in preserving biological variance. The metrics used were the following:

- k-nearest Neighbour Batch Effect Test (kBET)
- Principal Component Regression (PCR)
- Local Inverse Simpson's Index (LISI)
- Average Silhouette Width (ASW)
- Adjusted Rand Index (ARI)
- Normalised Mutual Information (NMI)
- Trajectory Conservation

- Ratio of Highly Variable Genes preserved
- Cell Cycle Conservation Score.

The project was implemented in the R programming language, using the Seurat framework for preprocessing and analysis. The deliverables consisted of a scientific report and a code repository, inclusive of a function designed to conduct the whole benchmarking process and complete with plots aiding to the interpretation of the results.

Data

Overview of the data processing

In the field of bioinformatics the data is often highly heterogeneous and prone to technical bias. The standardization of the data processing workflow is therefore essential to allow for the extraction of meaningful information and ensure the reproducibility of results. For this project the Seurat computational toolkit was used, which has become the standard in the subfield of scRNA-seq bioinformatics.

To prepare the data for downstream analysis, quality control and initial data preprocessing are applied. The dataset undergoes prefiltering to eliminate low-quality cells and genes, which may suggest the presence of droplets (genes with expression values inaccurately recorded as zero) or doublets (artifacts where the gene expression of multiple cells is inaccurately combined)[28]. This process establishes a minimum threshold of 3 cells per gene and 200 genes per cell. The percentage of mitochondrial and ribosomal genes is computed as a useful metric of cell quality. Then, in order to reduce the impact of technical biases and cell-specific noise, the data is normalized using log-normalization. Next, highly variable features are selected through variance-stabilizing transformation, in order to identify only genes that are informative of the variability in the data. Finally, the data is scaled to a mean of 0 and a variance of 1 to avoid uneven weighting of gene expression in downstream analysis.

Machine learning methods are employed to ease computational burden for subsequent analysis. This is first done by applying 3 different dimensionality reduction techniques: principal component analysis (PCA), uniform manifold approximation and projection (UMAP)[29] and t-distributed

stochastic neighbor embedding (t-SNE)[30]. PCA is particularly useful to ease computational burden and perform unsupervised machine learning, while UMAP and t-SNE allow for concise and informative visual representation of data[31]. After that, unsupervised clustering is obtained by calculating the K-nearest neighbor (KNN) graph in PCA space and applying the Louvain algorithm to iteratively cluster cells together. After that, celltype labels are standardized across the datasets and cell cycle scores are computed for each cell. Finally, a comprehensive round of quality control (QC) is applied to eliminate empty droplets, doublets, low-quality, and dying cells.

A panel of data visualization plots is then generated to explore the data. Initially, violin plots and scatter plots are produced for every numerical label (count of genes per cell, count of molecules per cell and percentage of mitochondrial genes) in order to visualize any internal discrepancies in data quality within the datasets. Next, cells are projected onto UMAP 2-dimensional plots classified by categorical labels (datasets, celltype and sample) to display the data's global structure and internal relations[32]. Bar plots with relative proportions of cell types and cell phases are produced as well. Finally, heat maps, dot plots and scatter plots are used to display feature expression and gene expression variance for highly expressed genes[33].

GEO GSE176078 - Wu Dataset

The first of the two datasets used in this project was curated by Sunny Z. Wu, Ghamdan Al-Eryani et al[26]. It originated from the study “A single-cell and spatially resolved atlas of human breast cancers”, published in Nature Genetics in 2021, and deposited on the GEO repository under the accession number GSE176078. The study aimed to provide a comprehensive atlas of human breast cancers and of their cellular architecture. By combining scRNA-seq data with spatial information, the researchers sought to provide a detailed view of breast cancer ecosystems.

The raw dataset was generated through the Illumina NextSeq 500 desktop sequencing technology and the 10x Chromium sequencing platform. It is approximately 4.3 GB in size and contains 100,064 cells, comprising 26 different tumor samples. It is composed of the following cell types: endothelial cells, fibroblasts, perivascular-like (PVL) cells, B cells, T cells, myeloid cells, epithelial cells, plasmablasts, and cancer epithelial cells. These cells are further classified into 49 subtypes, and the samples are categorized as HER2+, ER+, or TNBC based on their cancer subtype.

As shown in figures 2 and 3, prior to filtering and data preprocessing the dataset exhibits significant internal variance, characterized by a substantial presence of droplets and doublets. Furthermore, there is considerable variability in cell quality among distinct cell populations in the dataset. Specifically, cancer epithelial cells have on average a much higher amount of gene and RNA molecules, making them more likely to be doublets.

After filtering and quality control (QC) the composition of the dataset underwent substantial changes. Figures 4 and 5 are provided to aid in the visualization of the new composition and celltype proportions of the dataset. The dataset resized to 1.1 GB and now contains 31.977 cells. The cell populations experiencing the greatest loss were epithelial cells and cancer epithelial cells, of which only 8.36% and 13.64% remained after QC. B cells and T cells maintained the largest relative population at 54.70% and 57.38%.

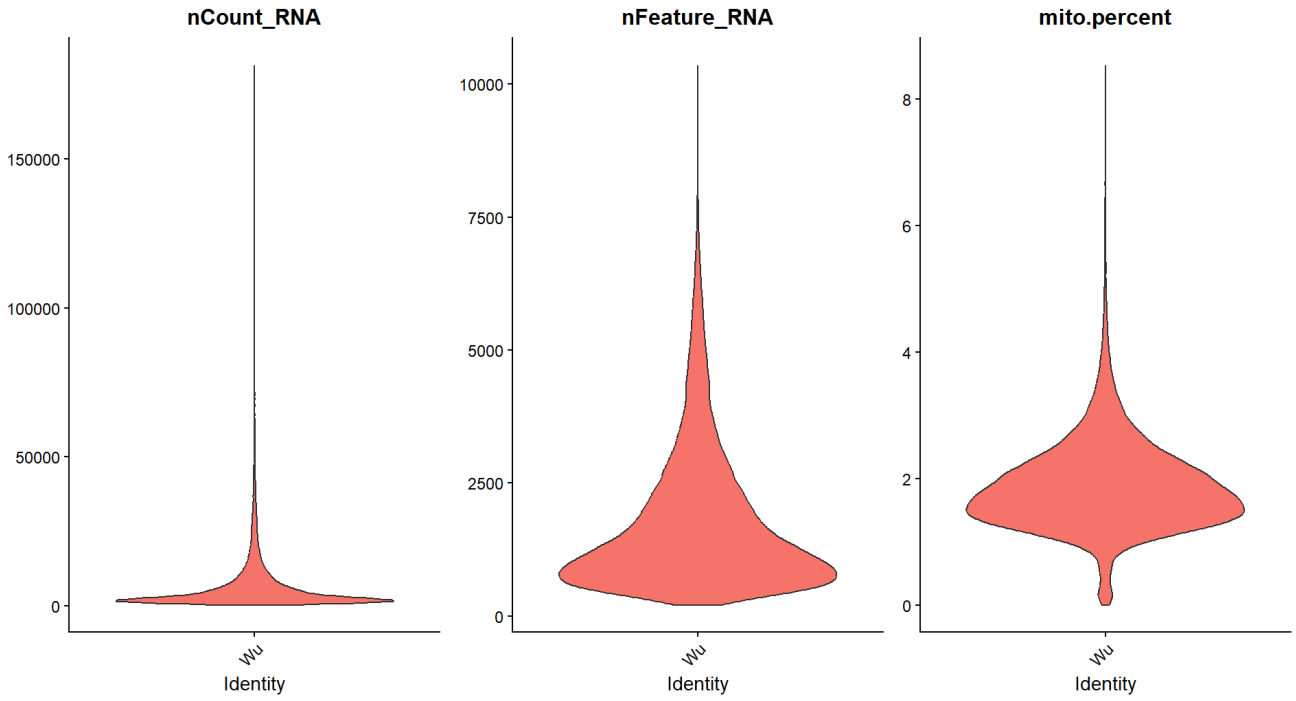

Figure 2: Violin plot of the raw Wu dataset, displaying the amount of cells according to the number of RNA molecules contained, the number of genes expressed and the percentage of mitochondrial genes expressed.

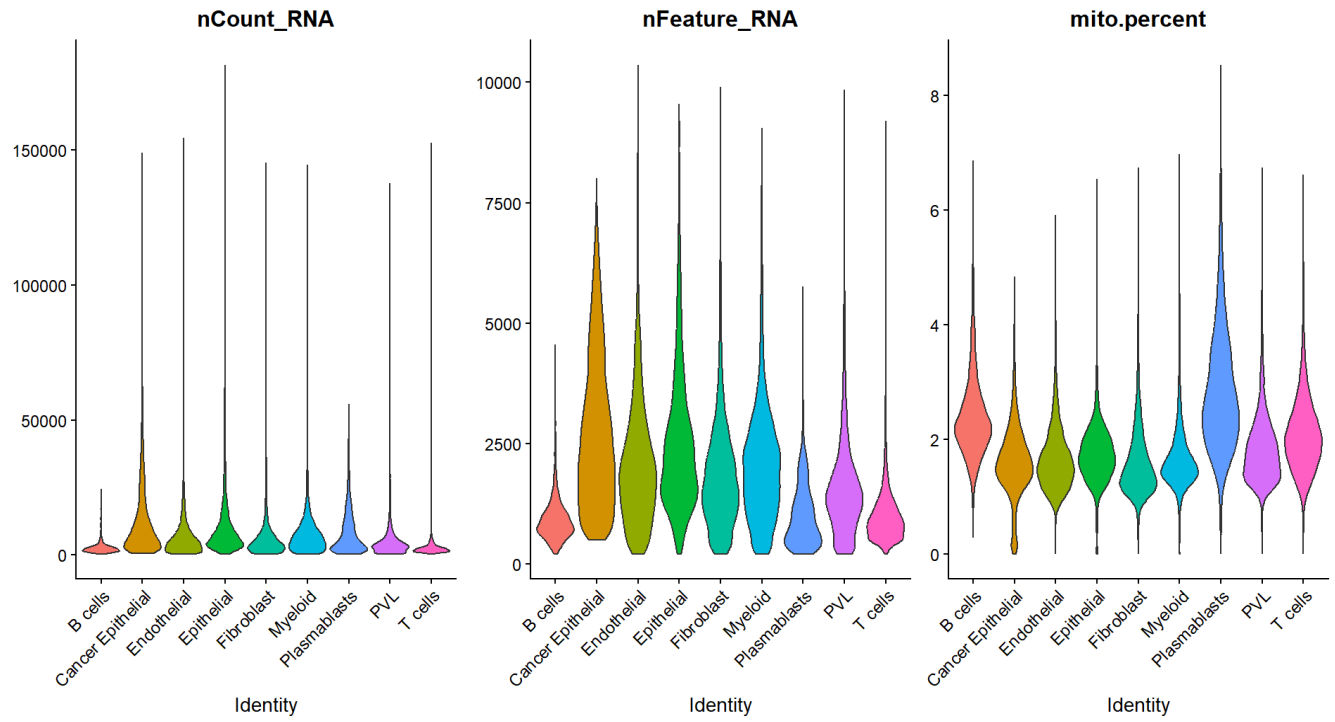

Figure 3: Violin plot of the raw Wu dataset, displaying the amount of cells according to the number of RNA molecules contained, the number of genes expressed and the percentage of mitochondrial genes expressed. Cells are split according to their celltype.

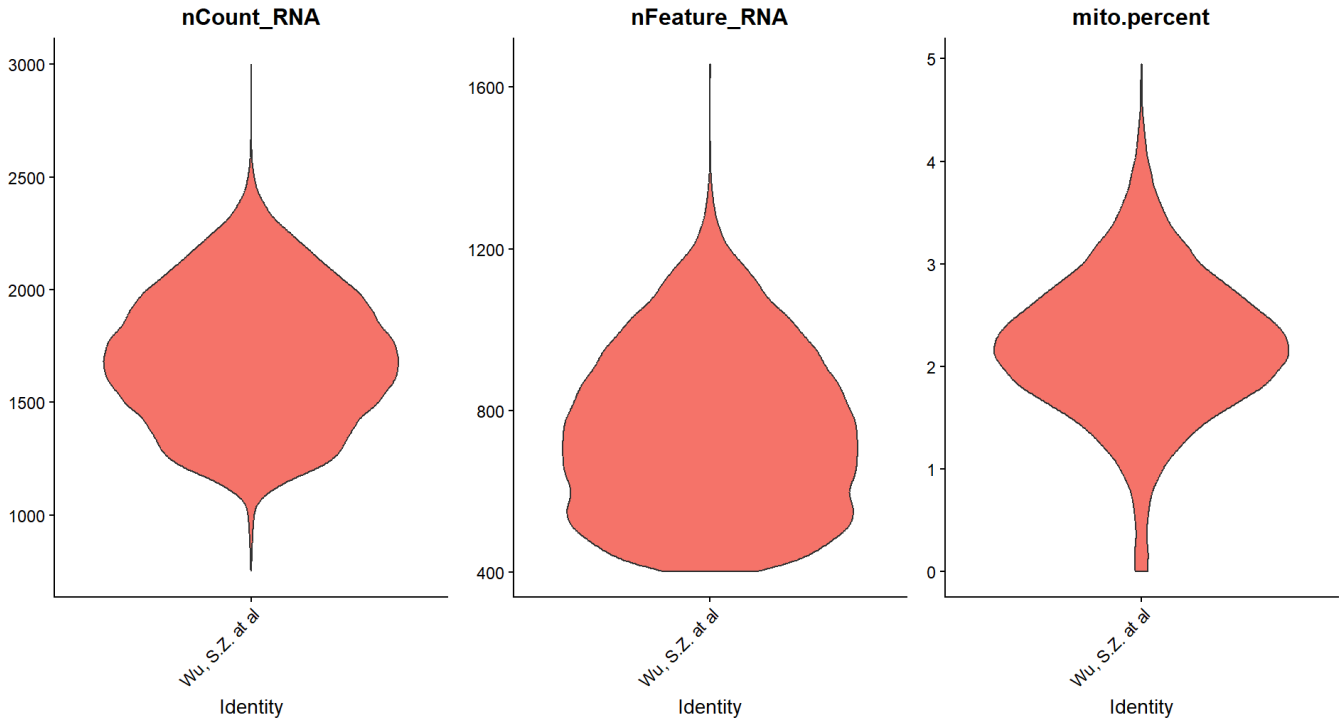

Figure 4: Violin plot of the filtered, processed Wu dataset, displaying the amount of cells according to the number of RNA molecules contained, the number of genes expressed and the percentage of mitochondrial genes expressed.

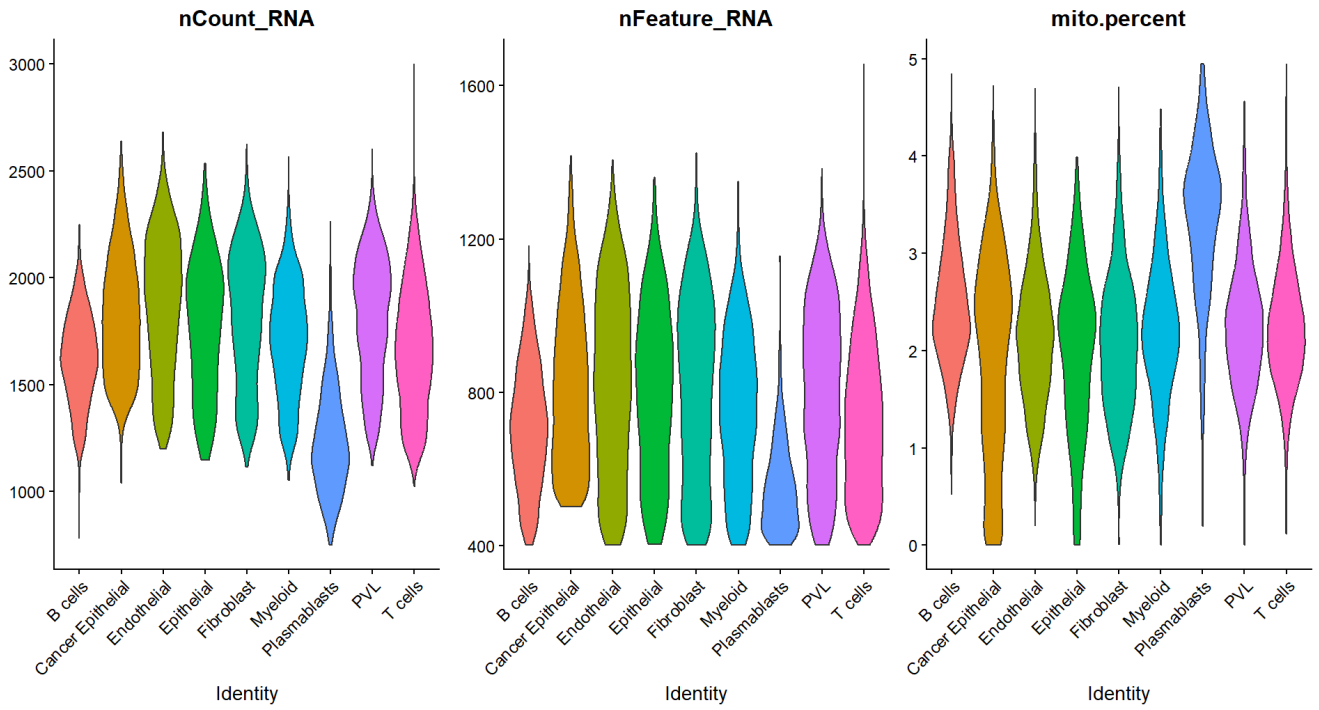

Figure 5: Violin plot of the filtered, processed Wu dataset, displaying the amount of cells according to the number of RNA molecules contained, the number of genes expressed and the percentage of mitochondrial genes expressed. Cells are split according to their celltype.

GEO GSE140819 - Slyper Dataset

The second of the two datasets used in this project was curated by Michal Slyper, Caroline B. M. Porte et al[27]. It stemmed from the study "A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors", which was published in Nature Medicine in 2020 and deposited in the GEO repository under accession number GSE140819. The study aimed to develop a systematic toolbox for profiling fresh and frozen clinical tumor samples. The original dataset contained samples from 9 different tumor types, of which I subsetted samples containing metastatic breast cancer data.

The raw data were obtained with HiSeq X Ten desktop sequencing technology and Chromium 10x sequencing platform. It is approximately 2.2 GB in size and contains 56.648 cells, comprising 9 different tumor samples. It contains the following cell types: macrophages, epithelial cells, T cells, B cells, NK lymphocytes, endothelial cells, fibroblasts, hepatocytes, oligodendrocytes and astrocytes.

As shown in figure 6, prior to data filtering and preprocessing the dataset exhibits some internal variance, characterized by a substantial presence of dying cells and considerable variability in cell quality among distinct cell populations. In figure 7 in particular it is clear that B cells on average show an abnormally high amount of RNA molecules (thus identifying them as doublets), while astrocytes and NK lymphocytes express very high ratios of mitochondrial genes, most of which exceed the 5% threshold.

After filtering and QC, the composition of the dataset changed substantially. Figures 8 and 9 are provided to aid in the visualization of the new composition and celltype proportions of the dataset. The data resized to 1.2 GB and is now composed of 28,483 cells. The cell population that suffered the greatest loss is B cells, of which only 8.62% were left after QC. The other cell populations maintained comparable proportions, ranging from 38 to 57%; the only exception is oligodendrocytes, with 78.22% of cells surviving QC.

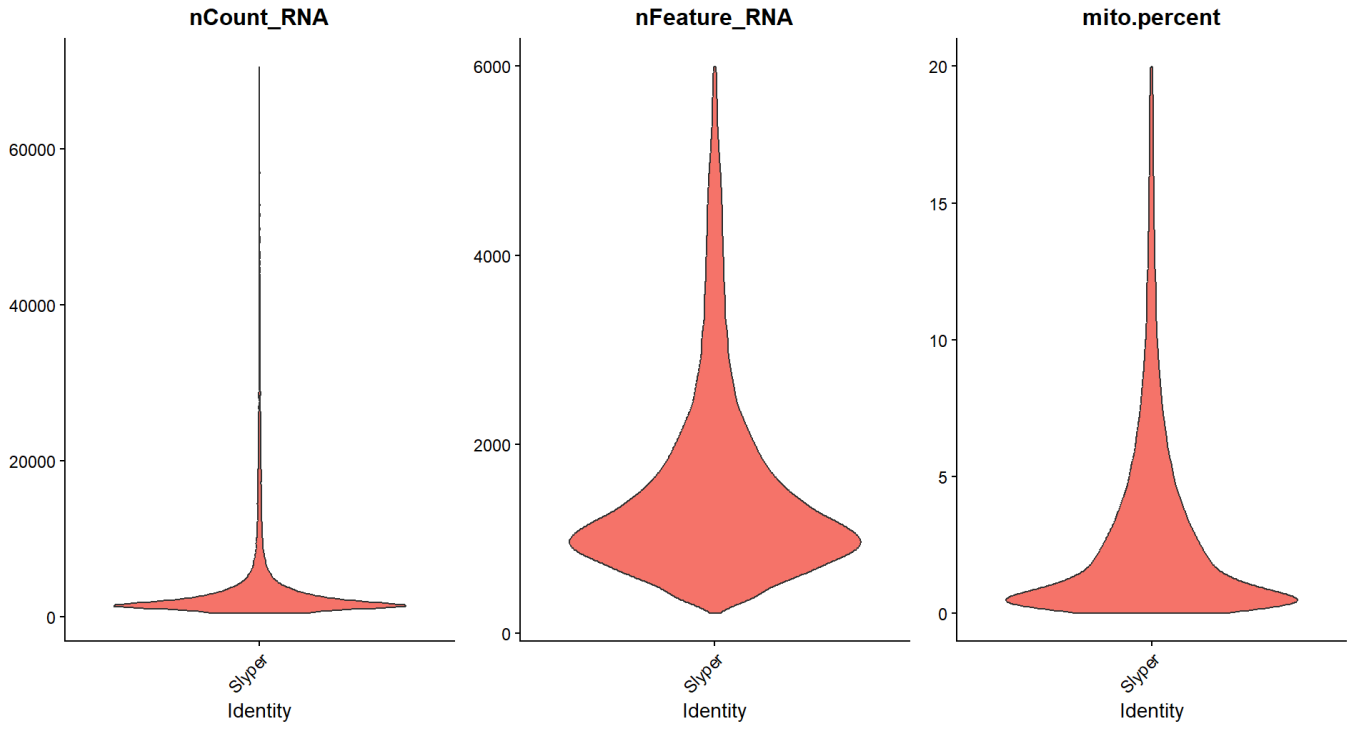

Figure 6: Violin plot of the raw Slyper dataset, displaying the amount of cells according to the number of RNA molecules contained, the number of genes expressed and the percentage of mitochondrial genes expressed.

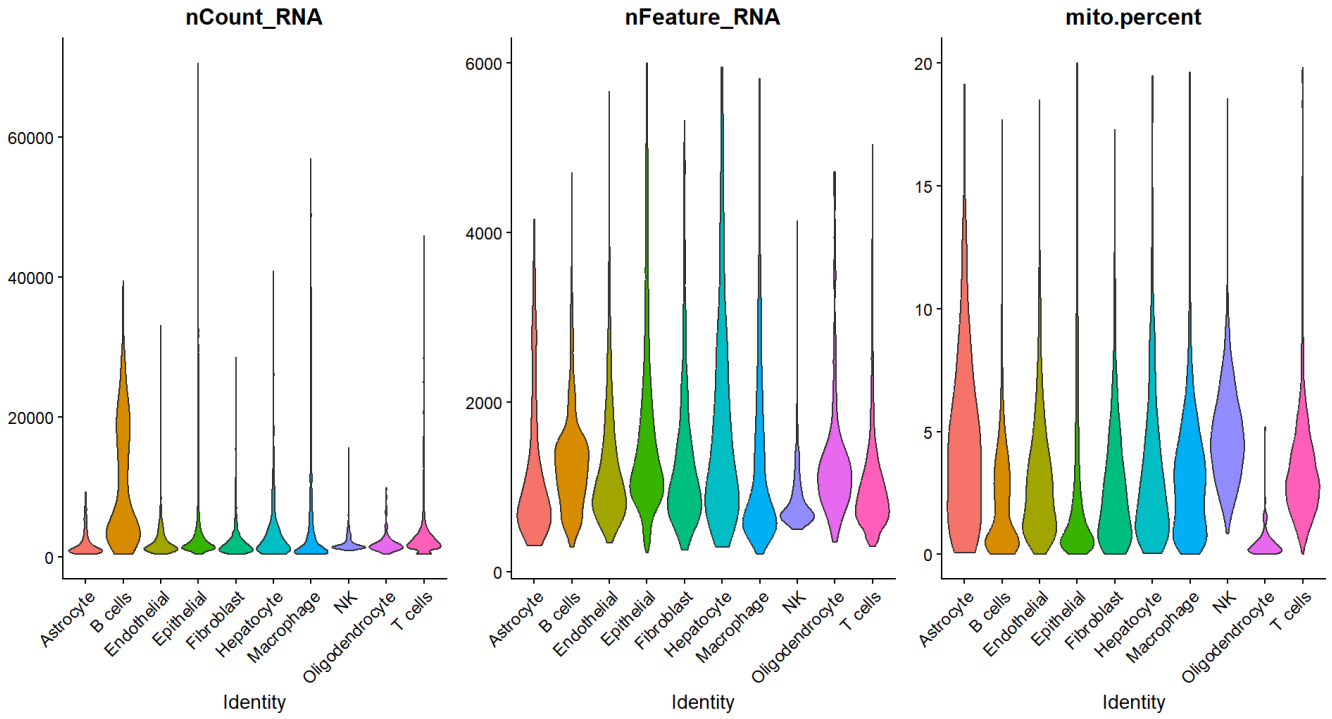

Figure 7: Violin plot of the raw Slyper dataset, displaying the amount of cells according to the number of RNA molecules contained, the number of genes expressed and the percentage of mitochondrial genes expressed. Cells are split according to their celltype.

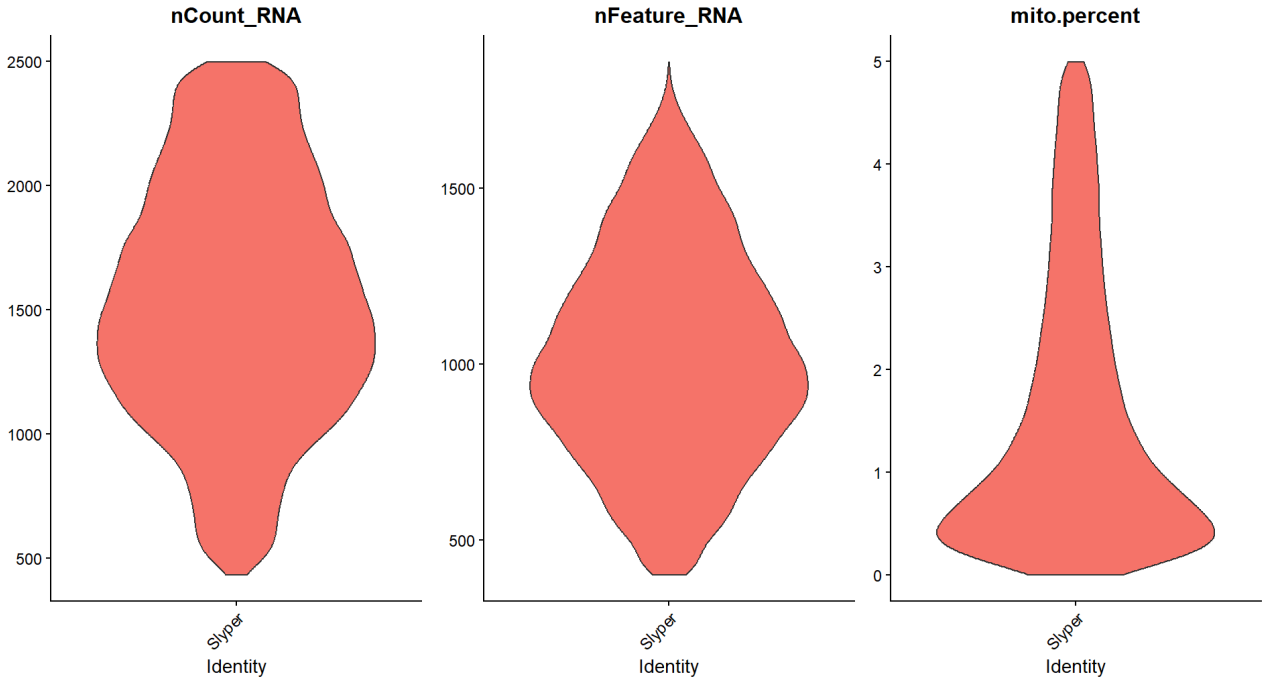

Figure 8: Violin plot of the filtered, processed, Slyper dataset, displaying the amount of cells according to the number of RNA molecules contained, the number of genes expressed and the percentage of mitochondrial genes expressed

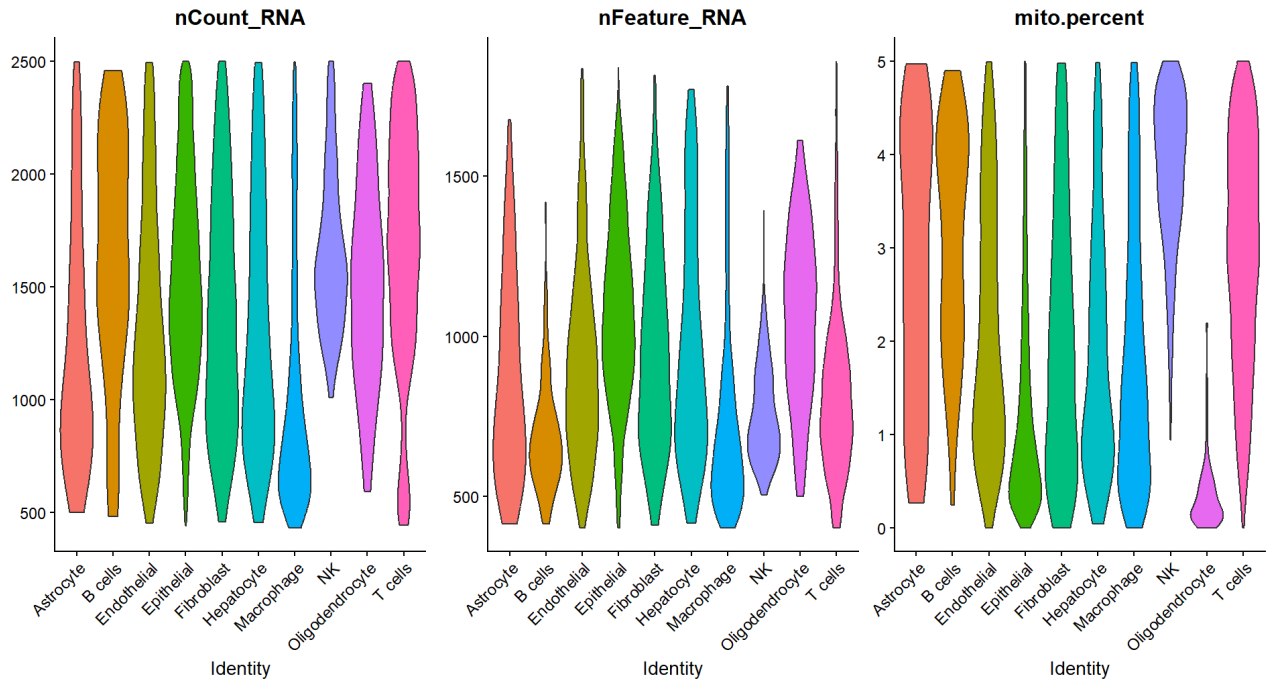

Figure 9: Violin plot of the filtered, processed Slyler dataset, displaying the amount of cells according to the number of RNA molecules contained, the number of genes expressed and the percentage of mitochondrial genes expressed. Cells are split according to their celltype.

Merging and preprocessing of the datasets

This processed merged dataset was obtained from the simple merging and preprocessing of the two preceding datasets, without the implementation of any further methodology. It served as a baseline and starting point for all subsequent project activities. Most importantly, it functioned as a benchmarking point for assessing the effectiveness of data integration techniques against the baseline scenario. It contains 60.460 cells, and weighs 2,4 GB.

As shown in figure 10, the dataset exhibits significant heterogeneity due to the large internal variance within the constituent datasets and the substantial imbalance in cell type composition. The overlap in cell population between the two datasets is relatively small and further biased by the differing relative proportions of cells belonging to each dataset. Specifically, the shared cell types include B cells, endothelial cells, epithelial cells, fibroblasts, and T cells. Notably, as seen in figure 10, the Slyper dataset predominantly contributes to epithelial cells (98.38%), whereas the Wu dataset predominantly contributes to B cells (94.45%), endothelial cells (69.36%), fibroblasts (75.45%), and T cells (88.20%). Upon visual examination of the data, the UMAP plots in figure 11

exhibit a clear separation between the two datasets, with cell type separation often aligning with dataset boundaries rather than true cell type distinctions.

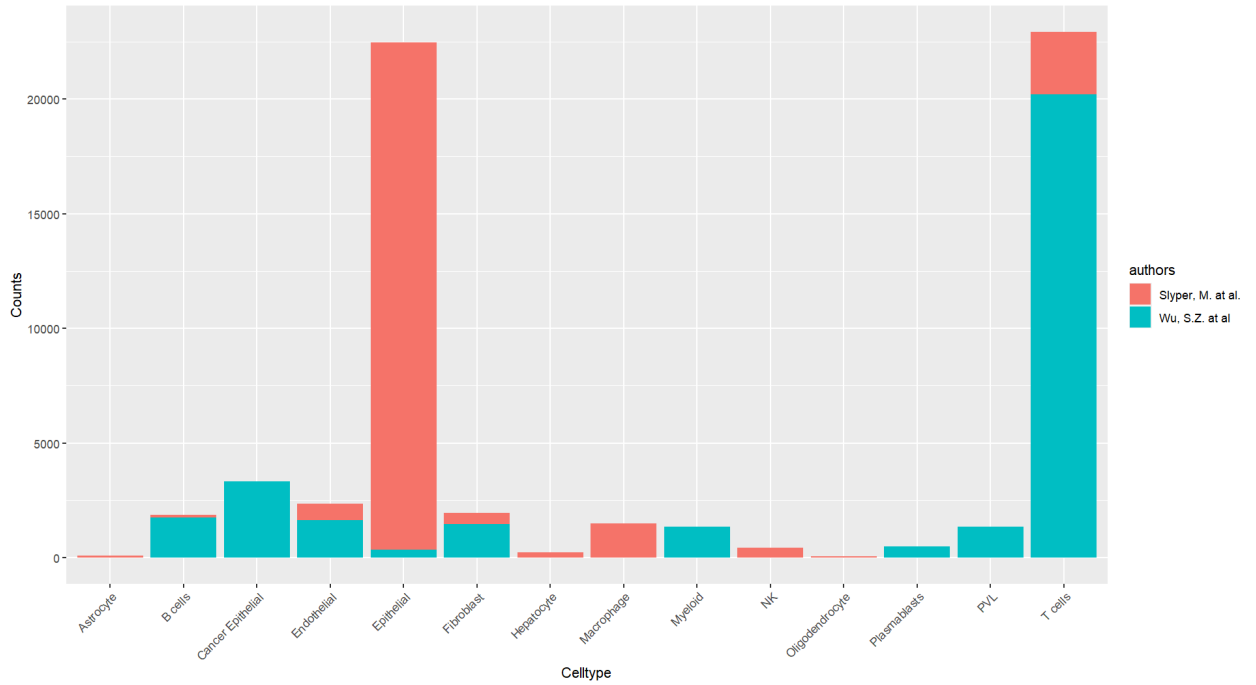

Figure 10: Barplot of the processed merged dataset, displaying the amount of cells for each celltype, colored by their source dataset.

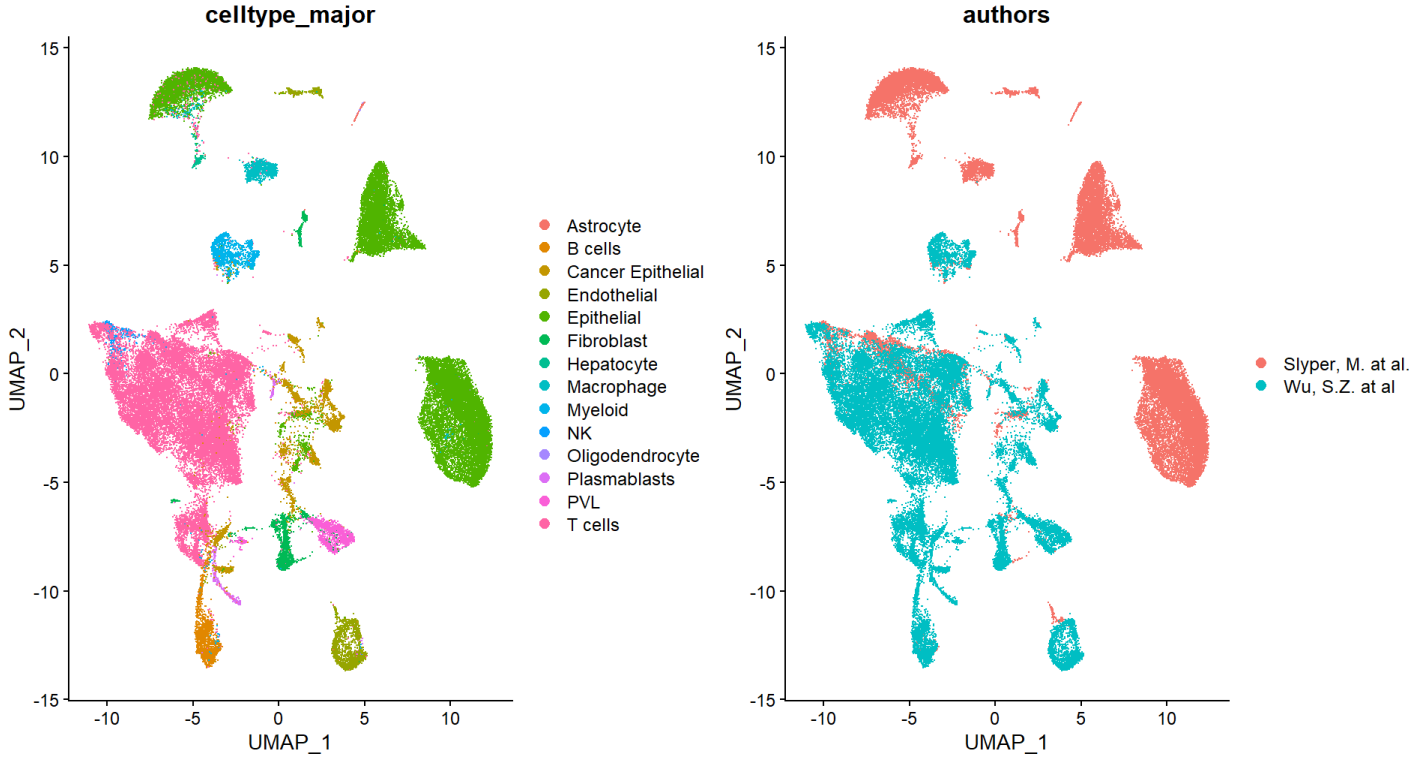

Figure 11: UMAP plot of the processed merged dataset. On the left cells are colored according to celltype, on the right according to the original dataset.

Integration methods

Thanks to the modern advancements in sequencing technology and to the decreasing costs associated it is now possible to obtain progressively larger amounts of data and to study it at increasingly better resolution. However, in order to extract meaningful information, it is often necessary to integrate data from multiple sources and acquired through a variety of different techniques, which inevitably exposes it to a multitude of different biases that are challenging to account for and to mitigate. This is especially pronounced for biological datasets in general and scRNA-seq data in particular, where biases may arise due to biological grounds, such as the vast diversity within cell samples, and laboratory methodologies, including specific techniques and environmental fluctuations during the data acquisition[34].

This issue highlights the importance of choosing appropriate ways to interface the different data into a single, unified view that addresses these concerns and allows the analysis of the data in a comprehensive manner. Different techniques have been developed for this purpose, each based on different concepts and approaches. In general, there are two primary challenges in the selection and implementation of effective data integration methodologies[31]:

- Arisal of batch effects: batch effects are apparent differences in data caused by technical factors that are not reflective of actual variations in biological data. The most common roots of this in scRNA-seq are the use of different sequencing technologies, instrumentation or lab protocols, and changes even minute in experimental and environmental conditions.
- Loss of biological variability: it is the presence of apparent similarities in data that do not reflect the variability in the biological data. This involves the technique mistaking true biological variations for technical bias and overcorrecting, inducing a loss of information.

Reciprocal Principal Component Analysis (RPCA)

Reciprocal Principal Component Analysis (RPCA) is a data integration technique introduced by the authors of a 2019 collaborative study in the journal Cell[35]. It is one of the first data integration techniques devised specifically for scRNA-seq data. It was designed to allow the construction of atlases of transcriptomic, epigenomic, proteomic, and spatially resolved single-cell data at tissue or organismal level. Unlike its predecessor, canonical correlation analysis (‘CCA’), it is intended to

have a conservative approach, assuming that the overlap in cell types between the constituent datasets is not very large and that cells in different biological states are less likely to 'align' after integration[36].

RPCA is based on the identification of pairwise mutually nearest cell correspondences (called "anchors") between single cells across datasets. The algorithm first reduces the dimensionality of the datasets using canonical correlation analysis. It then selects the integration features across overlapping cell populations and identifies the anchors between any possible pair of datasets in the shared low-dimensional representation. The cells are projected into the reduced PCA space using the aforementioned anchors. Two scores are then calculated: an anchor score for each anchor pair and a cell similarity score for each cell. These scores are weighted together to calculate a correction vector, which is then applied to every cell in the dataset to correct gene expression and finally integrate the data. Figure 12 is provided to aid in the visual understanding of the RPCA algorithm.

Figures 13 and14 are provided to offer an initial visual understanding of the results of the data integration process. An initial examination and comparison of the UMAP plots for the integrated dataset and the preprocessed one reveals a preliminary degree of mixing. This mixing occurs across dataset boundaries among cells of the same or similar cell types. However, the degree of mixing appears to be limited, with the datasets largely clustering separately.

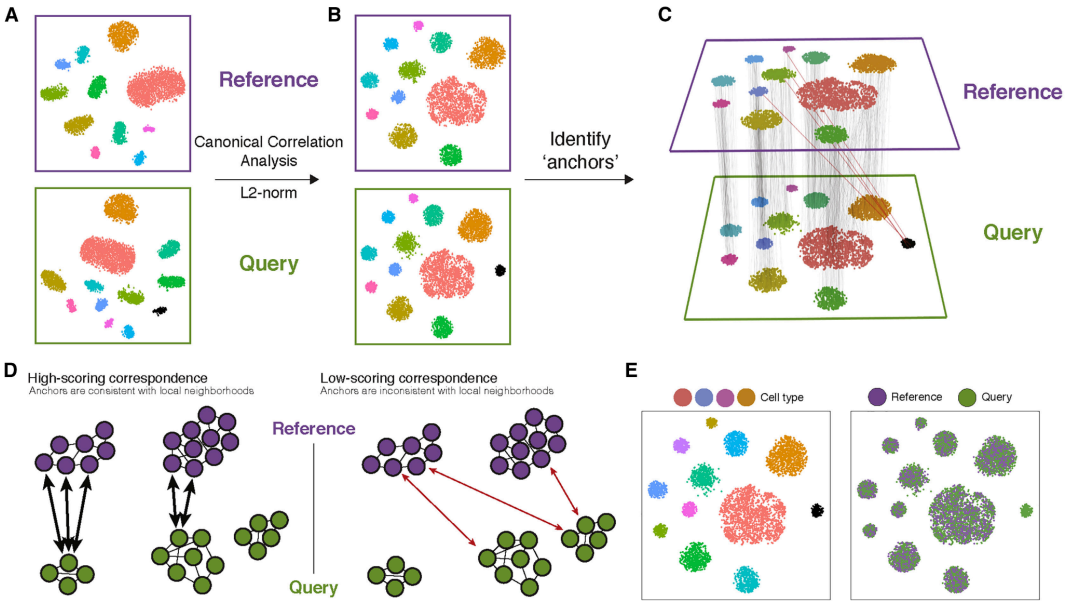

Figure 12: "In order to relate different experiments to each other, we assume that there are correspondences between datasets, and that at least a subset of cells represent a shared biological state. Inspired by the concept of mutual nearest neighbors (MNNs), we represent these correspondences as two cells (one from each dataset) that we expect to be defined by a common set of molecular features". Image taken from *T. Stuart et al., "Comprehensive integration of single cell data"*

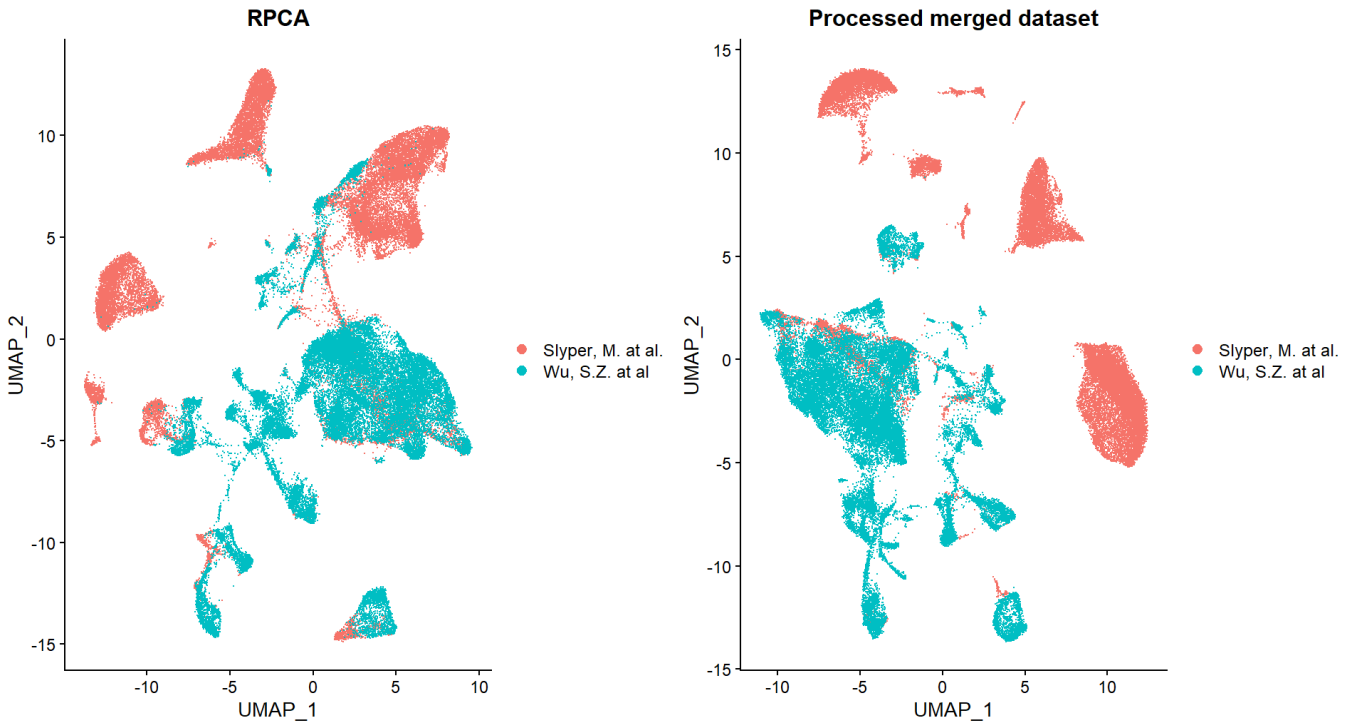

Figure 13: UMAP plots of the dataset integrated using RPCA (on the left) and of the processed merged dataset (on the right). Cells are colored according to source dataset.

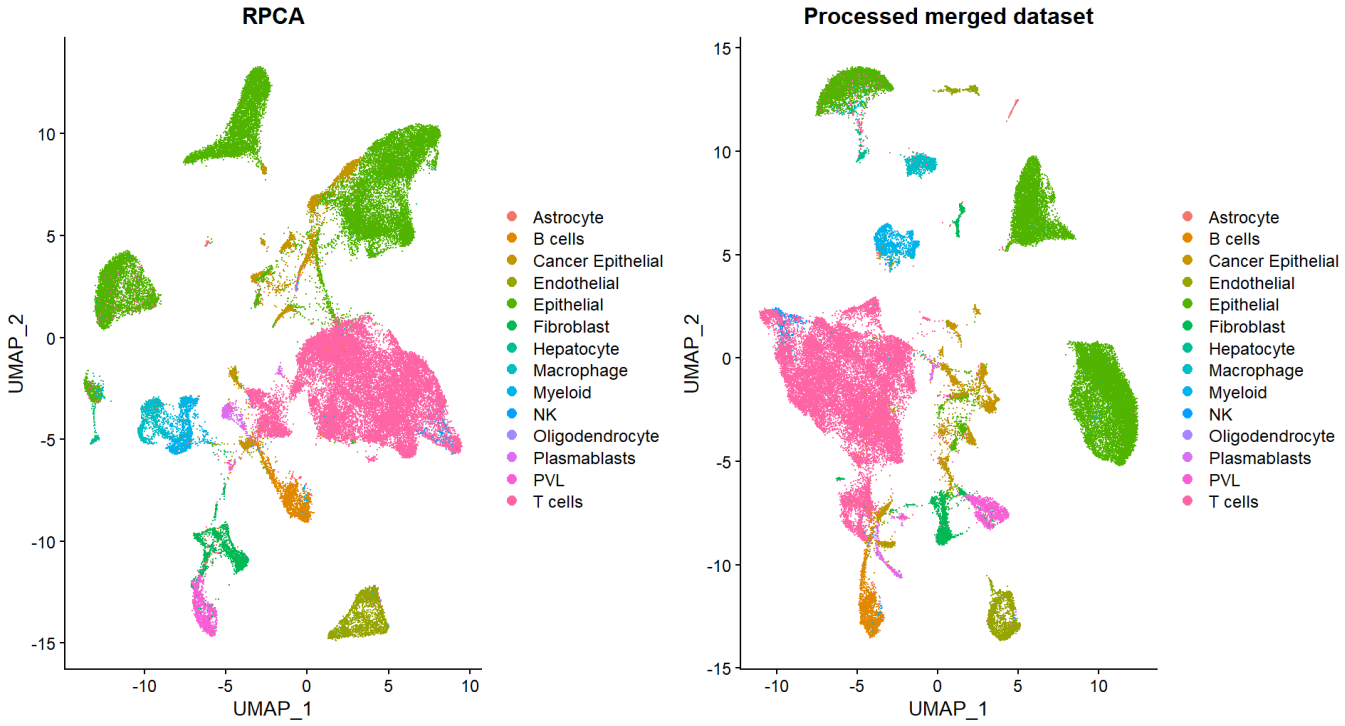

Figure 14: UMAP plots of the dataset integrated using RPCA (on the left) and of the processed merged dataset (on the right). Cells are colored according to celltype.

Fast Mutual Nearest Neighbour correction (FastMNN)

Fast Mutual Nearest Neighbors correction (FastMNN) is a data integration technique introduced by the authors of a 2018 collaborative study in the journal Nature Biotechnology[37]. It allows for integration of different datasets without requiring predefined or equal population compositions across batches. This method relies on the following assumptions: there is a shared population of cells between batches, the technical variations introduced by different batches are largely independent of the underlying biological differences, and they are much smaller than the biological-effect variation between different cell types. Unlike other techniques derived from bulk RNA-seq, it does not assume that the composition of the cell population within each batch is homogenous[37].

FastMNN is based on the identification of mutual nearest neighbors between batches and the calculation of their differences in gene expression[38]. First, the algorithm performs PCA across all cells in all batches, in order to reduce dimensionality and decrease technical noise. Then, it identifies mutually nearest neighbors between each reference batch and a target batch, which are pairs of cells across each pair of batches that resemble each other the most. The average batch vector is used to remove variation in both reference and target batches. Correction vectors are then computed for each paired cell in the target batch. Finally, locally weighted correction vectors are used to correct the target batch, which is then merged with the reference. The process is then iterated on all batches in the dataset. Figure 15 is provided to aid in the visual understanding of the FastMNN algorithm.

Figures 16 and 17 offer an initial visual understanding of the results of the data integration process. Upon initial examination and comparison of the UMAP plots for the integrated dataset and the preprocessed one, it becomes evident that there is a limited degree of mixing between the two datasets, with the majority of this occurring at the boundary between the two. It is evident that there is a greater degree of mixing across cell type boundaries, particularly in the case of the Wu dataset.

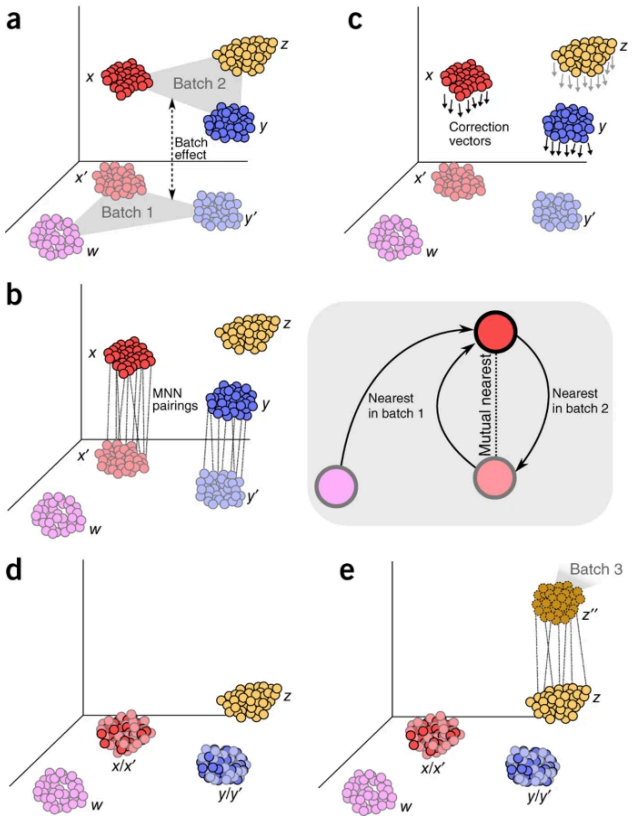

Figure 15: "The difference in expression values between cells in an MNN pair provides an estimate of the batch effect, which is made more precise by averaging across many such pairs. A correction vector is obtained from the estimated batch effect and applied to the expression values to perform batch correction.". Image taken from *L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni, "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors,"*

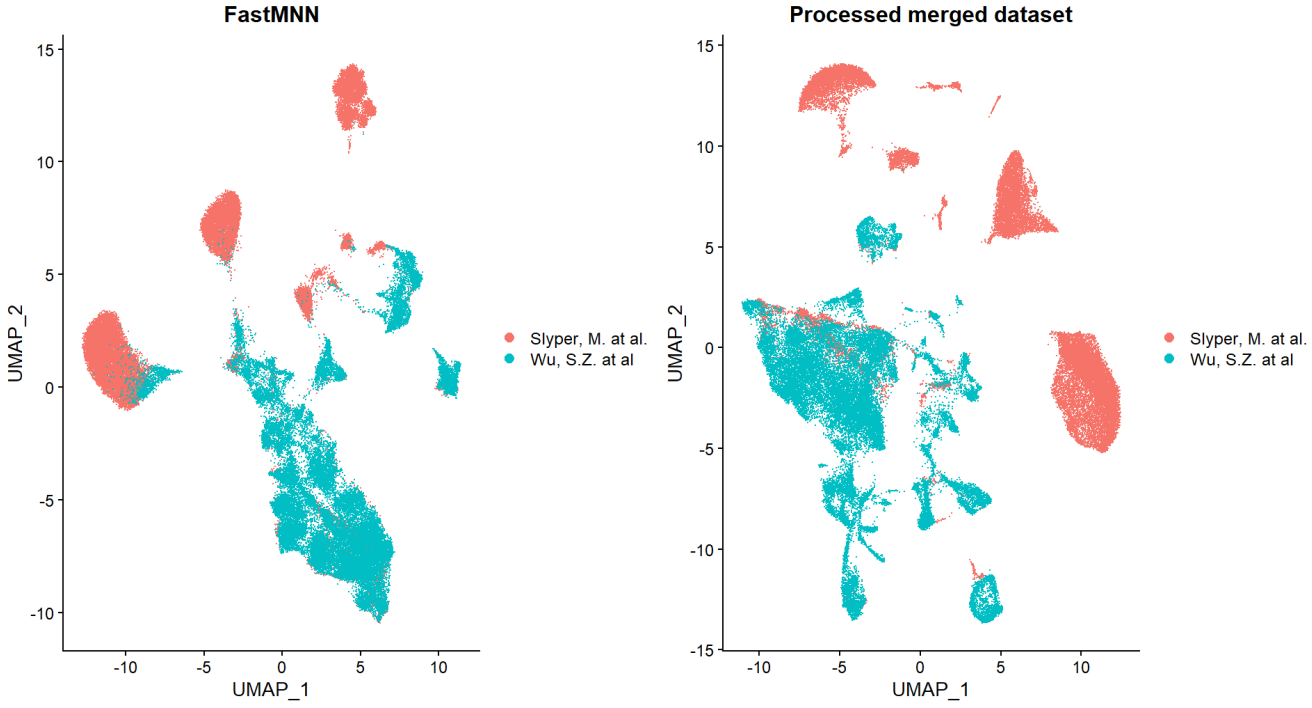

Figure 16: UMAP plots of the dataset integrated using FastMNN(on the left) and of the processed merged dataset (on the right). Cells are colored according to source dataset.

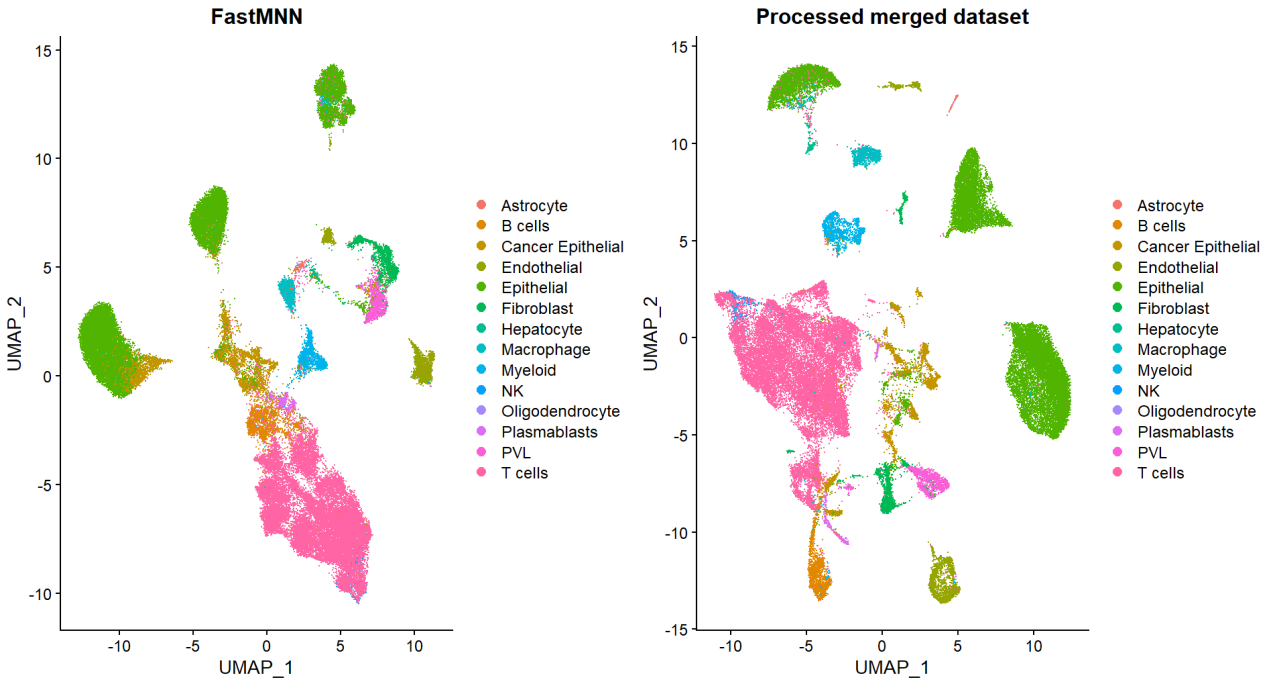

Figure 17: UMAP plots of the dataset integrated using FastMNN (on the left) and of the processed merged dataset (on the right). Cells are colored according to celltype.

Harmony

Harmony is a data integration technique introduced by the authors of a 2019 collaborative study in the journal Nature Methods[39]. It was designed to provide a computationally sound algorithm that scales well with dataset size, identifies both broad and fine-grained populations, and can easily accommodate complex experimental design across multimodal data.

Harmony is based on fuzzy clustering, an implementation of clustering in which each data point is initially assigned to more than one cluster[39]. The algorithm starts by projecting the cells onto the space obtained through PCA and grouping them into multi-dataset clusters. Initially every cell is assigned to multiple clusters, in order to account for smooth transitions between different cell states. Clusters containing disproportionate amounts of cells from a small subset of datasets are penalized. For each cluster, the algorithm computes a global centroid and cluster-specific linear correction factors, as well as correction factors specific to every identified cell-type and cell-state. Finally, each cell is assigned to a cluster-weighted average of these correction factors and a linear adjustment function is applied to correct it with its unique correction factor. The algorithm is then iterated until the assignment of every cell into a cluster becomes stable. Figure 18 is provided to aid in the visual understanding of the Harmony algorithm.

As shown in Figures 19 and 20, a preliminary visual overview suggests that, compared to the processed merged dataset, Harmony displays a limited but still comparatively higher degree of intermixing , forming clusters of cells belonging to similar cell types with limited relation with the originating dataset. Given the differing cell type proportions between the two datasets, epithelial cells appear to be particularly well-mixed.

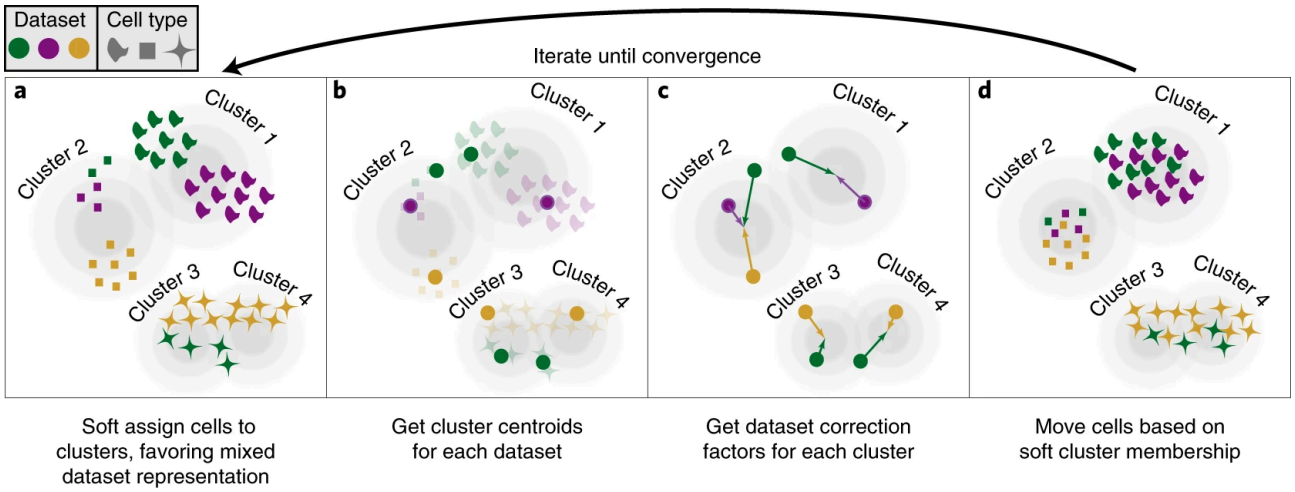

Figure 18: "PCA embeds cells into a space with reduced dimensionality. Harmony accepts the cell coordinates in this reduced space and runs an iterative algorithm to adjust for dataset specific effects." Schematic view of the batch correction process according to the Harmony method. Image taken from *I. Korsunsky et al., "Fast, sensitive and accurate integration of single-cell data with Harmony,"*

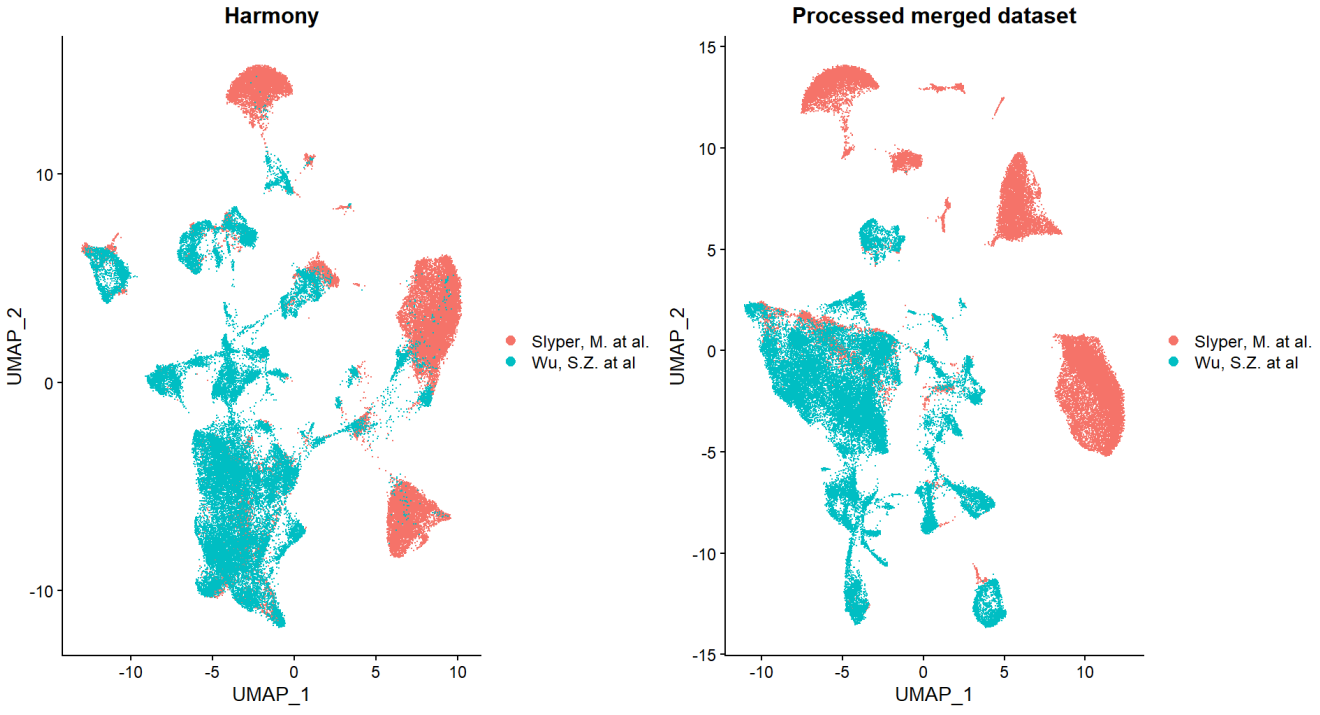

Figure 19: UMAP plots of the dataset integrated using Harmony (on the left) and of the processed merged dataset (on the right). Cells are colored according to source dataset.

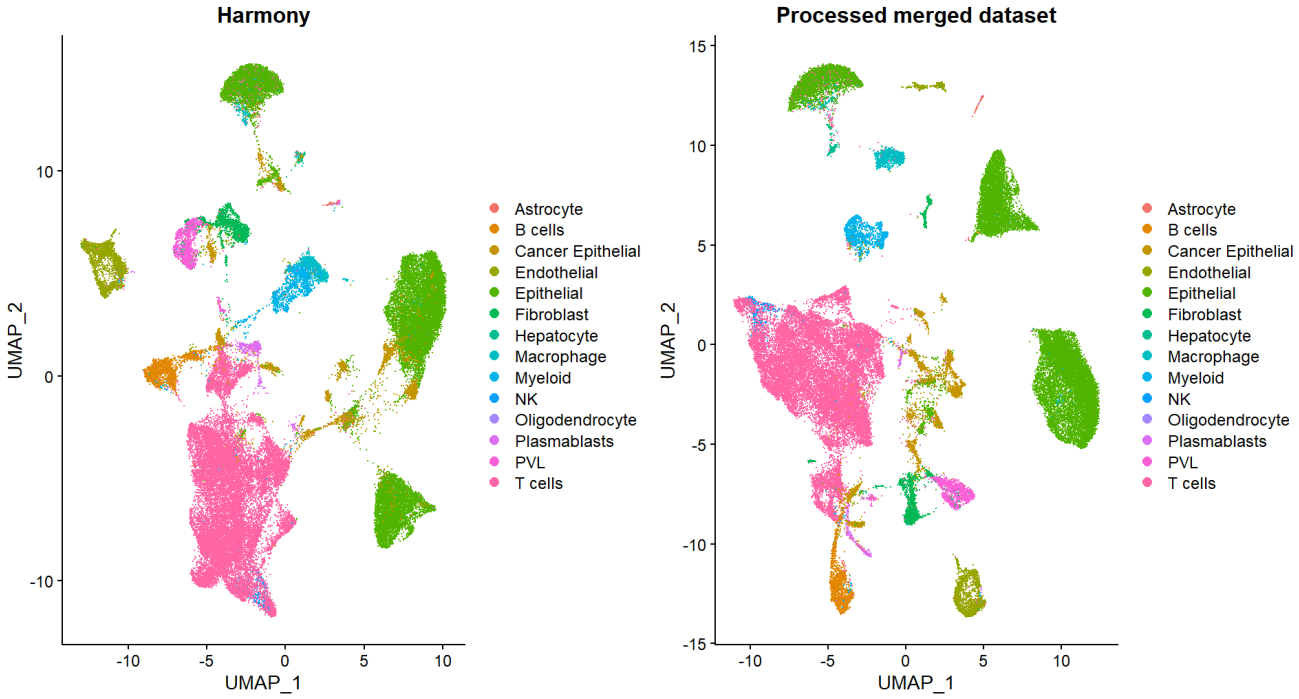

Figure 20: UMAP plots of the dataset integrated using Harmony (on the left) and of the processed merged dataset (on the right). Cells are colored according to celltype.

Linked Inference of Genomic Experimental Relationships (LIGER)

Linked Inference of Genomic Experimental Relationships (LIGER) is a data integration technique introduced by the authors of a 2019 collaborative study in the journal Cell[40]. It was designed to effectively integrate datasets that present high levels of internal and reciprocal heterogeneity in the data, with large differences in the number of cells and of features and in sequencing depth. Compared to other techniques, LIGER does not assume that differences within the data arise only due to batch effects and technical variations, but may also be reflective of true variance between widely differing cell populations. It is therefore devised to allow identification of shared and dataset-specific characteristics across diverse health statuses, individuals, species, and data modalities[40].

LIGER is based on iterative non-negative matrix factorization (iNMF), an algorithm performing the decomposition of the gene expression matrix into two constituent matrices, each containing only non-negative elements, therefore making them easier to inspect. LIGER uses iNMF to infer a set of dataset-specific underlying factors (referred as ‘metagenes’), which characterize every cell and often correspond to specific biological signals. After fine-tuning and normalization, cells are clustered together and labeled according to analogous maximum factor loadings. Finally, a shared

factor neighborhood graph is constructed, connecting cells with similar factor loading patterns. Figure 21 is provided to aid in the visual understanding of the LIGER algorithm.

For an initial visual assessment of the data integration results, figures X and Y are displayed in figures 22 and 23. The UMAP plots reveal extensive intermixing across both dataset and cell type boundaries. There is minimal discernible separation between the Slyper and Wu datasets, particularly among epithelial and cancer epithelial cells.

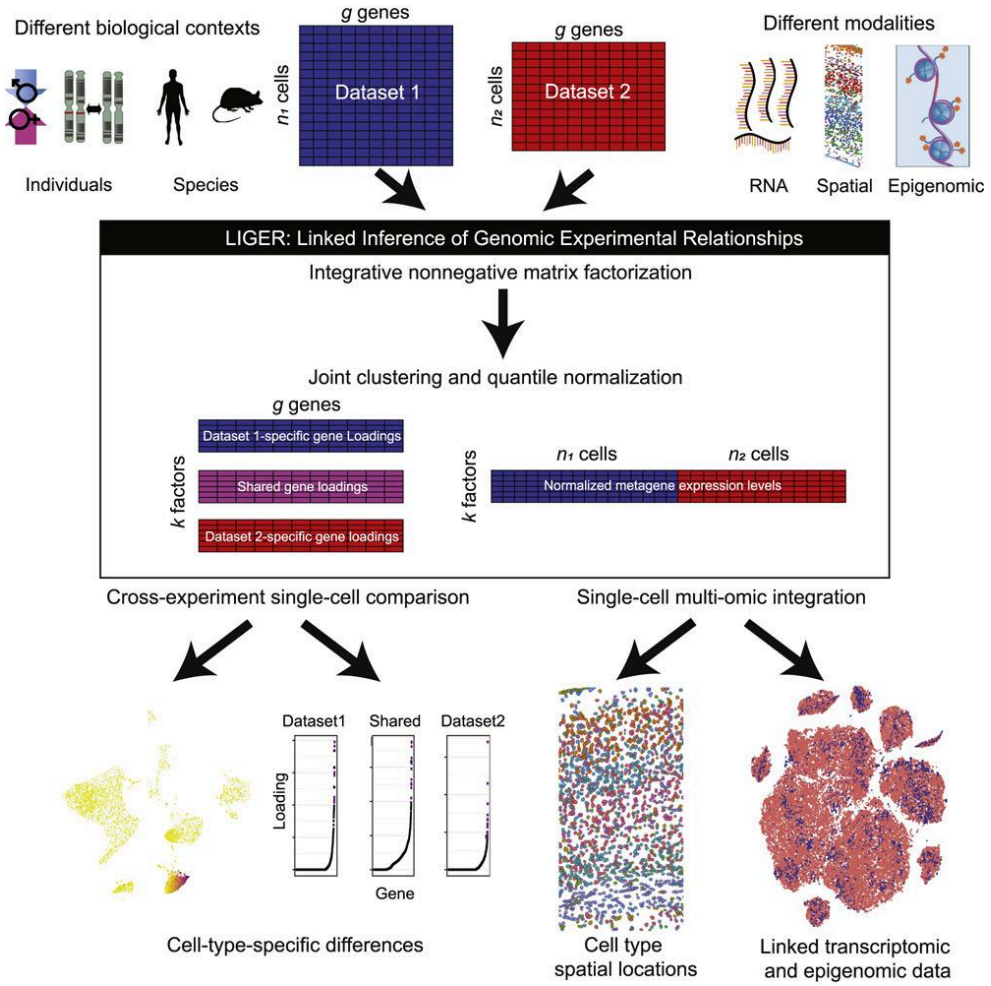

Figure 21: " The difference in expression values between cells in an MNN pair provides an estimate of the batch effect, which is made more precise by averaging across many such pairs. A correction vector is obtained from the estimated batch effect and applied to the expression values to perform batch correction.". Image taken from *J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko, "Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity,"*

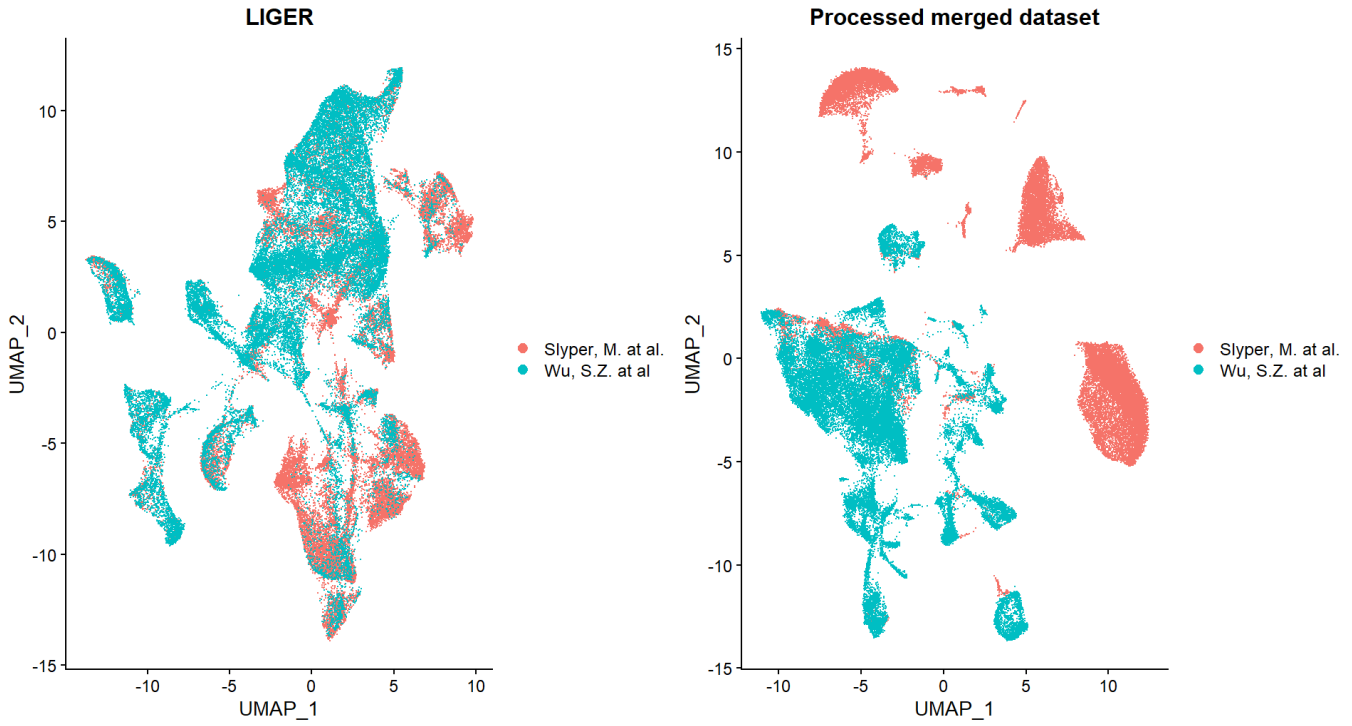

Figure 22: UMAP plots of the dataset integrated using LIGER (on the left) and of the processed merged dataset (on the right). Cells are colored according to source dataset.

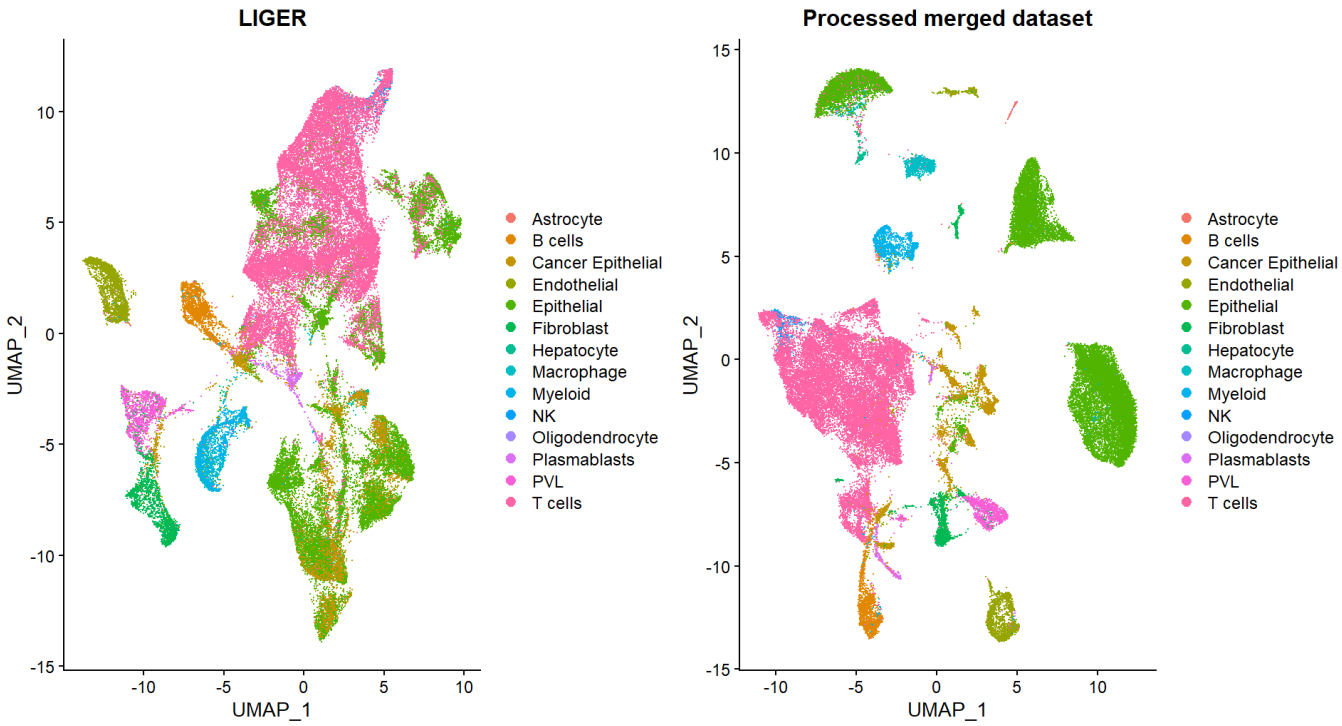

Figure 23: UMAP plots of the dataset integrated using LIGER (on the left) and of the processed merged dataset (on the right). Cells are colored according to celltype.

Metrics

k-nearest Neighbour Batch Effect Test (kBET)

The k-nearest neighbor batch-effect test (kBET) metric was introduced in 2019 in the journal Nature Methods in a collaborative study[41]. It was devised as a robust, sensitive and user-friendly test to objectively quantify batch effects It was the first metric designed specifically for scRNA-seq data.

The kBET metric is based on the repeated application of Pearson's χ^2 test to randomly selected neighbors[42]. The null-hypothesis is "the data is well-mixed". The algorithm takes an integrated dataset as an input. It first creates the k-nearest neighbor (kNN) matrix. Then, it selects 10% of all samples and rates the distribution of batch labels within samples according to the null-hypothesis. If the distribution of labels in the batches is sufficiently similar to that of the entire dataset, the null-hypothesis is accepted. The function returns a binary result for each of the tested samples, and finally it averages the test rejection rate across all samples.

Intuitively, kBET can be understood as a method for assessing whether the distribution of labels in subsamples of the dataset is similar to that of the entire dataset. The kBET score ranges from 0 to 1, with values closer to 0 indicating better mixing of labels throughout the dataset, and higher values suggesting a non-homogeneous distribution of labels within the dataset.

While useful, it is important to be mindful of a few issues when using this metric. First, it is highly sensitive to any kind of bias, which can potentially inflate the score. Second, kBET is sensitive to the quality of the data, and low-quality data may result in artificially high values. Third, the null hypothesis of kBET, which assumes that a homogeneous distribution of labels is expected, may not be applicable for all datasets. This is especially true for datasets with complex or subtle batch effects. Fourth, kBET is computationally intensive and does not scale well with large datasets.

In this project the kBET metric was applied to the "authors" label. For computational efficiency and to effectively capture the diversity of the data, PCA embeddings were used for the computations. Given the very high computational demands of kBET, every dataset was reduced in size to 250 cells through subsampling.

Principal Component Regression (PCR)

Principal Component Regression (PCR) is a regression analysis technique that models the relationship between a target variable and predictor variables. It uses principal components to estimate the unknown regression coefficients in a standard linear regression model.

As a metric, the algorithm takes as an input a dataset, a list of batches (used as predictor variables) and a target label (used as a dependent variable)[43]. First, PCA is used in order to generate principal components from the predictor variables. Then, the dependent variables are then regressed on the PCA loading obtained using the method of least squares. A vector of estimated regression coefficients is obtained, and then transformed back in order to obtain a final PCR estimator, which is then used to predict the target label based on the predictor variables.

This score of this metric represents the number of principal components needed to explain 100% of the variance in the target variable. Intuitively, it can be understood as a measure of how efficient the data integration technique is at using the predictor labels to calculate clusters. The predictor variables used in this project were “authors”, “sample” and “celltype”, while the target variable was the “cluster”.

Local Inverse Simpson's Index (LISI)

The Local Inverse Simpson's Index (LISI) is a metric introduced in 2019 by Korsunsky, I. et al in the journal Nature Methods. It was designed to provide good evaluation of batch effects in scRNA-seq data while also accounting for local distances between cells and data imbalance. It is an extension of Simpson’s diversity index, a metric used in ecology to measure diversity within populations.

First, a perplexity parameter k is chosen, providing a measure of uncertainty in the value of a sample from a discrete probability distribution[44]. After that, the cell embeddings are computed, and a kNN graph is built on the cell embeddings. Then, Simpson’s diversity index is computed using the formula $D_L(S) = \sum_{l \in L}(n_l/n)^2$. S is the kNN graph built from the cell embeddings, L is the specified set of labels, n_l is the number of cells in the graph carrying the label l and n is the total number of cells. The LISI score is finally computed as the inverse of the value just obtained.

Intuitively, the LISI score may be understood as the number of cells that need to be selected from a batch before the same label is observed twice[34]. Its value may range from 0 to N, where N is the number of different labels in the set chosen.

The project utilized the PCA reduction algorithm to calculate kNN graphs, simplifying the calculation and allowing good performance for data information capture. The dataset was reduced to 800 cells to improve computability. A perplexity value of 30 was chosen, due to its commons as a standard value in the field. Technical bias was assessed through the "authors" and "sample" labels, while the "celltype" and "cluster" labels were used to assess the conservation of biological variability and clustering quality.

Average Silhouette Width (ASW)

Average Silhouette Width (ASW) is a method of interpretation and validation of consistency within clusters of data, derived from unsupervised learning methodologies. It is based on the silhouette method introduced by Peter J. Rousseeuw in Journal of Computational and Applied Mathematics in 1987, and over the years has become one of the main metrics used in validation of scRNA-seq clustering.

The algorithm takes a dataset and a cluster label as an input. For each cell *i* in the dataset, the algorithm first calculates the average distance *a(i)* between the cell and every other cell in the same cluster[45]. The distance metric may be either Euclidean or Manhattan, in this thesis the Euclidean distance was used. Next, it computes the average distance *b(i)* between that cell and every other cell in the nearest cluster. Then, the algorithm computes the silhouette score for the cell using the following formula: $s(i)=(b(i)-a(i))/max\{a(i), b(i)\}$. Finally, all the values are averaged to obtain the ASW score for the specified label.

ASW scores evaluate the average similarity between components within a batch, compared to components in other batches[34]. The score ranges from -1 to +1, with values closer to 1 indicating well-separated clusters with high internal similarity, 0 indicating overlapping clusters, and -1 indicating strong dissimilarity within clusters and high levels of misclassification. The metric was developed in two versions: the standard one, and the isolated labels ASW metric. The latter is the

ASW metric applied to a subset of the integrated dataset containing only “isolated labels” (cells belonging to celltypes that are present only in a maximum of 5 different samples).

In this project, the ASW metric was used to evaluate both effectiveness in removal of batch effects and effectiveness against loss of biological information, with the isolated labels ASW metric assessing this specifically on outlier data. For the standard ASW metric, the data was downsampled down to 800 samples to ease computability. For computational efficiency and to effectively capture the diversity of the data, PCA embeddings were used for the computations. The evaluation of technical biases was conducted based on the ”authors” and “sample” labels, while the assessment of biological variability preservation and clustering quality utilized “celltype” and “cluster” labels. Due to the smaller size of the dataset, the isolated labels ASW metric is computed only on the “sample”, “celltype” and “cluster” labels.

Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) is a metric used to validate clustering performance. It was introduced by Hubert and Arabie in 1985, and it is based on Rand’s Index, developed by William Rand and published in the Journal of the American Statistical Association in 1971.

The algorithm takes as an input two labeled sets X and Y. It then computes the Rand index using the formula $RI = (a+b)/(a+b+c+d)$, where a represents the number of pairs of elements sharing the same label in both X and Y, b represents the number of pairs of elements that do not share a label in either X or Y, and c represents the number of pairs of elements that share a label in X but not in Y, and d represents the number of pairs of elements that share a label in Y but not in X[46]. Then, the ARI is computed using the formula $ARI = (RI - RI_{expected}) / (RI_{max} - RI_{expected})$. $RI_{expected}$ is the RI score computed using a contingency table[47] and RI_{max} is the maximum RI score, which is 1.

Intuitively, ARI can be understood as measuring the degree of agreement between an estimated clustering and a reference clustering, adjusted for the possibility that the agreement is due to chance[34]. The score can range from 0 to 1, where 0 indicates a complete lack of agreement between the two label sets and 1 indicates complete agreement. In this project, “cluster” was used as the reference label. Technical bias was measured using the “authors” and “sample” labels, while biological conservation was evaluated using the “celltype” label.

Normalised Mutual Information (NMI)

Normalized Mutual Information (NMI) is a metric used to evaluate network partitioning performed by community finding algorithms, measuring the degree of overlap between two clusterings. It is a normalization of the Mutual Information (MI) score, a metric introduced by Claude E. Shannon in 1949 in the journal Bell System Technical Journal[48].

First, the algorithm takes as an input two labeled sets X and Y[49]. It then computes the MI score as $MI(X, Y) = H(X) + H(Y) - H(X, Y)$, where $H(X)$ and $H(Y)$ are the marginal entropies of the two sets and $H(X, Y)$ is their joint entropy. There are multiple normalization methods used in literature to compute NMI, in this project normalization was performed by dividing the MI score by the maximum of the two marginal entropies, using the formula $NMI(X, Y) = MI(X, Y)/max\{H(X), H(Y)\}$.

Intuitively, it may be seen as expressing how much a chosen set of labels is informative about the true clustering of the cells[34]. The score may range from 0 to 1. An NMI score of 0 indicates that the two clusterings are completely independent, meaning that there is no mutual information between them. An NMI score of 1 indicates that the two clusterings are identical, meaning that there is perfect correlation between them. In this project the reference label used was “cluster”. Technical bias was measured using the “authors” and “sample” labels, while biological conservation was evaluated using the “celltype” label.

Trajectory Conservation

Trajectory conservation is a metric introduced by Malte D. Luecken, M. Büttner et al. in their 2021 in the journal Nature Methods. It measures the degree to which the developmental paths of cells, known as trajectories, are preserved during the process of integrating data from two datasets.

The algorithm first computes the pseudotime (a measure of a cell's progress along its developmental path[50]) of all cells between two datasets, a reference dataset and a comparison dataset[34]. The algorithm then calculates Spearman’s rank correlation coefficient between the two, defined as the Pearson correlation coefficient between the rank variables. Finally, the values are normalized between 0 and 1.The score may vary between 0 and 1, with 0 indicating lack of preservation on cell development proxy information and 1 total preservation of it.

Pseudotime may be computed in several different ways. In this project the first principal component was used, as it separates well cells by cell state while also being computationally efficient[50]. The referenced dataset used was the processed merged dataset, while the comparison dataset was the integrated dataset.

The trajectory conservation metric presents a few limitations. Firstly, it's very sensitive to the method used for the calculation of pseudotime, and results may not always be comparable. Secondly, it assumes that cells develop on a linear, natural developmental path, which may not always be the case for pathological populations[51]. Lastly, it assumes that biological information is more readily available before data integration, which may not always be the case.

Ratio of Highly Variable Genes preserved

The ratio of preserved highly variable genes (HVGs) is a metric introduced by Malte D. Luecken, M. Büttner et al. in their 2021 study 'Benchmarking atlas-level data integration in single-cell genomics' published in the journal Nature Methods. It serves as a proxy for the preservation of the most informative components of the biological signal.

The HVG ratio is calculated as $HVG(Y) = overlap(X,Y) = |X \cap Y| / min(|X|, |Y|)$, where X represents the highly variable genes of the reference non-integrated dataset (the processed merged dataset) and Y represents the highly variable genes of the integrated dataset[34]. In this project, a cutoff value of 500 HVGs was used for gene selection. This ensured that only the most highly variable genes were tested, which encapsulate most of the variance and are most representative of biological variability. The score may range from 0 to 1, with 0 indicating a complete loss of HVGs after integration and 1 indicating a complete preservation of them.

There are multiple methods in the literature to identify HVGs. In this project, the "dispersion" method provided by the Seurat platform was used. The algorithm selects genes based on the logarithm of the gene expression dispersion, which is defined as the variance-to-mean ratio $s_i^2 / \bar{X}_i$ Genes with larger expression dispersion are selected with higher priority[52]. The dispersion method was chosen due multiple reasons. Firstly, Seurat methods are the most widely used in literature. Secondly, among Seurat methods, dispersion provides the highest efficiency in terms of runtime[53]. Thirdly, it performs well with datasets of various sparsity (ratio of 0 values in the gene expression

matrix). Fourthly, this method is the least likely to result in errors when benchmarking diverse data integration techniques.

Cell Cycle Conservation Score

The cell cycle conservation (CCC) score is a metric that evaluates the ability to capture the cell-cycle effect before and after integration. It was introduced in the 2021 study 'Benchmarking atlas-level data integration in single-cell genomics' published in the journal Nature Methods by Malte D. Luecken, M. Büttner et al. The score is based on the cell cycle scoring method, which was introduced by Itay Tirosh, Benjamin Izar et al. in 2016.

The algorithm is divided into two main sections. The first section computes a score for the S phase (DNA synthesis phase) and the G2/M phase (DNA damage checkpoint before mitosis) for each cell[54]. These scores are calculated by aggregating the expression levels of a list of gene markers specific to each phase. The phase of each cell is determined based on its S score and G2M score. Cells with low S and G2/M scores are in the G1 phase, while cells with high S score and low G2/M score are in the S phase, and cells with high G2/M score and low S score are in the G2/M phase.

The second section computes the CCC score for each phase using the formula $CCC\ score = 1 - |(Var_{integrated} - Var_{merged}) / Var_{merged}|$[34]. $Var_{integrated}$ is the variance of the score of the specified phase in the integrated dataset, and Var_{merged} is the variance of the score of the specified phase in the merged dataset. The conservation score for the cell cycle was computed by averaging the two scores calculated for the S and G2/M phases. The score ranges from 0 to 1, with 0 indicating no conservation of variance explained by the cell cycle and 1 indicating full conservation of cell cycle-related variance.

Results

All metrics were classified in two categories: metrics informing on batch effect removal and metrics informing on preservation of biological variability. In green are the cells associated with the best results for the metric, while in red are the cells displaying the worst result. It is important to note

that red doesn't not necessarily denote a bad result, and green does not necessarily denote a good result.

Metrics pertaining batch effects:

Metrics			Integration Methods			
			RPCA	FastMNN	Harmony	LIGER
kBET		authors	1	0.478	0.478	0.173
LISI		authors	1.333	1.162	1.162	1.161
		sample	3.925	3.509	3.509	3.671
ASW	Non isolated	authors	0.065	0.140	0.140	0.090
		sample	-0.108	-0.044	-0.044	-0.046
	Isolated	sample	0.032	0.075	0.075	0.074
ARI		authors	0.085	0.120	0.212	0.029
		sample	0.314	0.403	0.289	0.127
NMI		authors	0.135	0.164	0.172	0.070
		sample	0.454	0.449	0.401	0.325

Metrics pertaining preservation of biological variability and model quality:

Metrics	Integration Methods			
	RPCA	FastMNN	Harmony	LIGER

PCR			29 dimensions	31 dimensions	16 dimensions	43 dimensions
LISI		celltype	1.401	1.328	1.328	1.331
		cluster	1.763	2.695	1.310	3.597
ASW	Non isolated	celltype	0.091	0.144	0.144	0.126
		cluster	0.070	0.088	0.158	-0.106
	Isolated	celltype	0.471	0.478	0.478	0.401
		cluster	0.175	-0.179	0.346	0.030
ARI		celltype	0.254	0.291	0.587	0.104
NMI		celltype	0.448	0.454	0.590	0.319
Trajectory conservation			0.438	1	1	0.348
HVG ratio			0.598	0.608	1	1
Cell Cycle Conservation			1	0.957	1	1

RPCA results

RPCA is one of the two top performers in regards to removal of batch effects, while also adequately preserving biological information.

When assessing the metrics related to the preservation of biological information, RPCA does not result as the top-perfomer method overall. Celltype-ARI and NMI scores are in the middle of the range of values assumed by this metric among the four algorithms, and indicate that celltype-related information is being used significantly to cluster cells together. This fact is further reinforced by the LISI scores for clusters and celltype, which point at a good separation between clusters. The ASW scores for isolated labels indicate that the algorithm performs well with outlier data, though the

lower results for non-isolated ASW hint at RPCA's limitations with evaluating the dataset comprehensively. This is supported by the value for the HVG ratio, which is the lowest among the four data integration techniques and indicates an incomplete consideration of gene expression variability. The low trajectory conservation score, although not necessarily all that meaningful on its own, also supports this conclusion. Additionally, the high value of the CCC score indicates that the cell cycle is well accounted for during the data integration process. The PCR value of 29 PCs points at a high degree of effectiveness in using the predictor labels to account for variance in the data.

Regarding batch effect removal, RCPA exhibits the best results for 50% of the metrics. The high kBET result is well explained by the size and unbalancedness of the dataset. The ASW scores for technical labels, as well as the authors- and sample-LISI scores, are the highest among the four data integration techniques, and demonstrate a successful removal of technical bias. The sample-NMI score is relatively high, ranking as the worst among the four algorithms and falling in the middle of the range of possible values, whereas the authors-NMI score is the second highest, indicating that the dataset-specific bias has been effectively mitigated, but that the sample-specific bias remains. The ARI scores follow the same pattern and also support this conclusion.

Overall, it achieves good performance in removing batch effects but does not stand out as a top-performing method in conserving biological variability. The residual presence of sample-specific bias is well explained by the conservative approach of the algorithm. Overall, RPCA demonstrates good and well-balanced performance, particularly given the unbalanced source datasets.

FastMNN results

FastMNN's performance is balanced across metrics related to both the preservation of biological information and the removal of batch effects. The algorithm prioritizes the preservation of biological information, being somewhat less effective at removing batch effects.

With respect to metrics on the preservation of biological information, FastMNN performs well. The HVG ratio is only around 60%, and, when combined with the good but not great CCC score, it suggests an incomplete accounting of the variance of gene expression. The NMI score for celltype is the second highest, but the corresponding ARI score is only in the middle of the range assumed

by this metric among the 4 data integration algorithms, suggesting that clustering has been effective in grouping together similar cellypes, but not to the same degree at separating clusters from each other. This view is supported by the value of the celltype-LISI score, which is one of the best two for this metric, and the comparatively high value of the cluster-LISI score, the second highest among the data integration techniques. This is further reinforced by the ASW scores for the celltype label, tying together with Harmony for best scores, and comparatively worse for the cluster label, close to 0 or displaying a negative value. In particular, the isolated-label celltype ASW score suggests that the algorithm struggles to cluster together outlier data. The PCR value is relatively low, indicating that FastMNN is quite efficient.

FastMNN does not result as a top-perfomer method in removing batch effects, with values that, while among the worst for the four algorithms, are still close to those achieved by the other methods. The low kBET score, at around the middle of possible values for this metric, proves that after the data integration process, the data is overall quite well-mixed. Interestingly, the sample scores for ASW, LISI, ARI and NMI are all among the worst across the four data integration methods, demonstrating a residual presence of sample-specific bias. However, the corresponding scores for the “authors” label are significantly better, indicating that the bias specific to the source datasets has been satisfactorily mitigated.

In conclusion, FastMNN displays the second best performance in preserving biological information, while not standing out regarding batch removal. Although there is a residual presence of sample-specific bias, it can be attributed to the assumptions behind the algorithm, particularly the assumption that there is always a shared population between batches.

Harmony results

Harmony seems to be the top performer in conserving biological variability. Although its results in regards to batch effects are overall the weakest, it is still comparable to the other methods.

When looking at the metrics regarding the preservation of biological information, Harmony is by far the best performer method across the 4 methods. As evidenced by celltype- ARI and NMI scores, it is particularly good at using celltype-specific information in clustering cells together. The ASW metrics point to a high degree of homogeneity in the way the gene expression data is adjusted

for clusters and celltypes, especially for the isolated labels. This fact is reinforced by the values of celltype- and cluster-LISI, with the former especially revealing a strong homogeneity and degree of separation in clustering of cells belonging to the same cell type. The HVG, CCC and trajectory conservation scores are all at 1, meaning that Harmony very accurately captures data related to gene expression variability and cell cycle stage. The PCR value of 16 PCs shows that the data integration method effectively utilizes the predictor labels to calculate clusters and account for variance in the data.

As for metrics pertaining batch effects, Harmony's performance is the weakest among the four algorithms, although even the worst scores still align with the higher-performing values achieved by other data integration techniques. The authors- and sample-ASW, although among the highest measures, are still close to 0 and suggest that, after data integration is performed, the two source datasets display greater similarity to each other than they do to one another. The corresponding LISI scores also indicate that the source datasets tend to cluster together. The authors-ARI and NMI values are close to the bottom of the range of possible values for this metric and point to a limited persistence of source-specific bias, while sample-NMI is close to the middle of the range of values, indicating that clustering was affected by sample-specific factors. However, the kBET score is less than half of its possible value, signifying that the data is overall quite well-mixed.

In conclusion, Harmony effectively preserves biological information, albeit at the expense of batch removal. The comparatively lower performance in removing technical bias may be due to the high degree of dissimilarity between the two source datasets. Overall, Harmony's performance aligns well with the underlying assumptions of the data integration algorithm.

LIGER results

LIGER proves to be one of the top performers at removing batch effects. However, its performance at handling the preservation of biological information is not very good.

When assessing the preservation of biological information, LIGER is the worst performer. The scores for celltype-ARI and NMI are the lowest among the four techniques,with a significant gap compared to the others, indicating that celltype-related information is scarcely used in clustering. All ASW scores, except the isolated celltype-ASW score, are very close to zero or display negative

values, suggesting inadequate and chance clustering of similar cells. The cluster-LISI score is markedly the highest and also supports the view that the clusters are poorly separated. The PCR score of 43 PCs shows that LIGER is very inefficient at using the predictor labels to calculate clusters and account for variance in the data. However, the HVG ratio and CCC scores of 1 demonstrates the algorithm's effectiveness in preserving cell cycle-related information and capturing highly variable genes.

LIGER's performance in terms of batch effect removal demonstrates to be the best one among the algorithms employed. The authors- and sample- scores for ARI and NMI are the lowest within the group of four data integration techniques, with the authors- ones being especially close to 0, indicating an almost complete absence of bias specific to source datasets and a low presence of sample-specific one. The LISI scores for both authors and samples are relatively close to the upper range of values, indicating a partial separation of clusters present due to source-specific factors. The kBET value is the lowest, indicating a very high degree of homogeneity between cells in the dataset after data integration. However, given the imbalanced nature of the source datasets, this warrants scrutiny, as it suggests a loss of biological information.

In conclusion, LIGER performs well at removing batch effects, but at the significant expense of not effectively preserving biological information. This result aligns with the assumptions of the algorithm, which was designed to integrate very highly diverse cell populations and may therefore not work as well with comparatively more homogeneous datasets.

Evaluation and comparison of the techniques

Harmony, FastMNN, RPCA and LIGER underwent a comprehensive analysis to benchmark their ability to integrate data well and in a useful fashion. Their performance was evaluated based on their effectiveness at clustering together similar cells, preserving gene expression variance, and removing batch effects. More weight was given to the preservation of biological information, as it is crucial for a successful bioinformatics analysis.

When evaluating the four data integration algorithms together with respect to the conservation of biological variance, all the methods perform similarly in regards to clustering similar cells together. The main differences lie in how well the algorithms preserve the variance of the gene expression and delineate distinct clusters. Harmony emerges as the top performer, exhibiting a robust

performance both comprehensively and with outlier data, while doing so in a computationally efficient way. FastMNN performs similarly, although it displays a higher degree of sample-specific bias. RPCA is computationally efficient and quite effective with outlier data, however, when looking at the data comprehensively, it seems to partially conflate biological information and technical bias. Finally, LIGER's performance is noticeably worse than that of the other methods, and the algorithm does not separate clusters effectively.

In regards to removal of batch effects, all methods seem to perform quite similarly at removing bias that is specific to the source dataset. The main differences lie in their ability to account for sample-specific bias. Harmony's approach is the most conservative one, resulting in the comparatively speaking poorest performance. Of the four algorithms it is the worst at removing dataset-specific technical bias, although it performs comparatively better at mitigating sample-specific bias. FastMNN slightly outperforms Harmony, demonstrating effectiveness in mitigating technical bias while maintaining a balanced approach. However, it is the least effective method at removing sample-specific bias. RPCA is the best performer, displaying minimal persistence of sample-specific bias. LIGER, despite achieving respectable ARI, NMI and kBET scores, displays non-optimal results in ASW and LISI, indicating an unbalanced and probably overly thorough removal of potential bias.

In the realm of overall performance, RPCA emerges as the most balanced approach among the evaluated techniques. It consistently performs well across the different source datasets and requires few assumptions regarding the source data, making it a good choice for performing data integration in scenarios with limited prior knowledge or significant dissimilarities between the source datasets. Harmony's methodology prioritizes the preservation of biological information, albeit at the partial cost of a less efficient removal of batch effects. Given its good overall performance it may be considered to be the top method of choice, especially with relatively homogeneous datasets. FastMNN, while comparable to Harmony in most aspects, underperforms slightly in biological conservation, probably due to its restrictive assumptions. Finally, LIGER is very prone to overcorrection of the data and focuses mostly on technical bias removal, at the expense of preservation of gene expression variance. However, it may be a good choice for extremely unbalanced and internally diverse datasets.

Figures 24 and 25 can offer a final aid in understanding the benchmarking results from a visual perspective. Visually the results seem to align with the benchmarking results. Higher degrees of intermixing across dataset boundaries seem to correlate with better results for the batch removal

metrics, while higher degrees of clusterization of cells belonging to similar celltypes seem to correlate with better scores in conservation of biological variability. LIGER displays the highest degree of intermixing across both dataset and celltype boundaries. RPCA exhibits a much lower but still noticeable level of mixing between datasets, with a comparatively higher extent of clusterization by similar celltype. Harmony demonstrates the highest degree of clustering of cells belonging to related celltypes, with a lower degree of dataset mixing that is present mostly among epithelial cells. FastMNN presents the lowest magnitude of intermixing between datasets, while having a higher degree of mixing between celltypes, particularly between cancer epithelial, B and T cells.

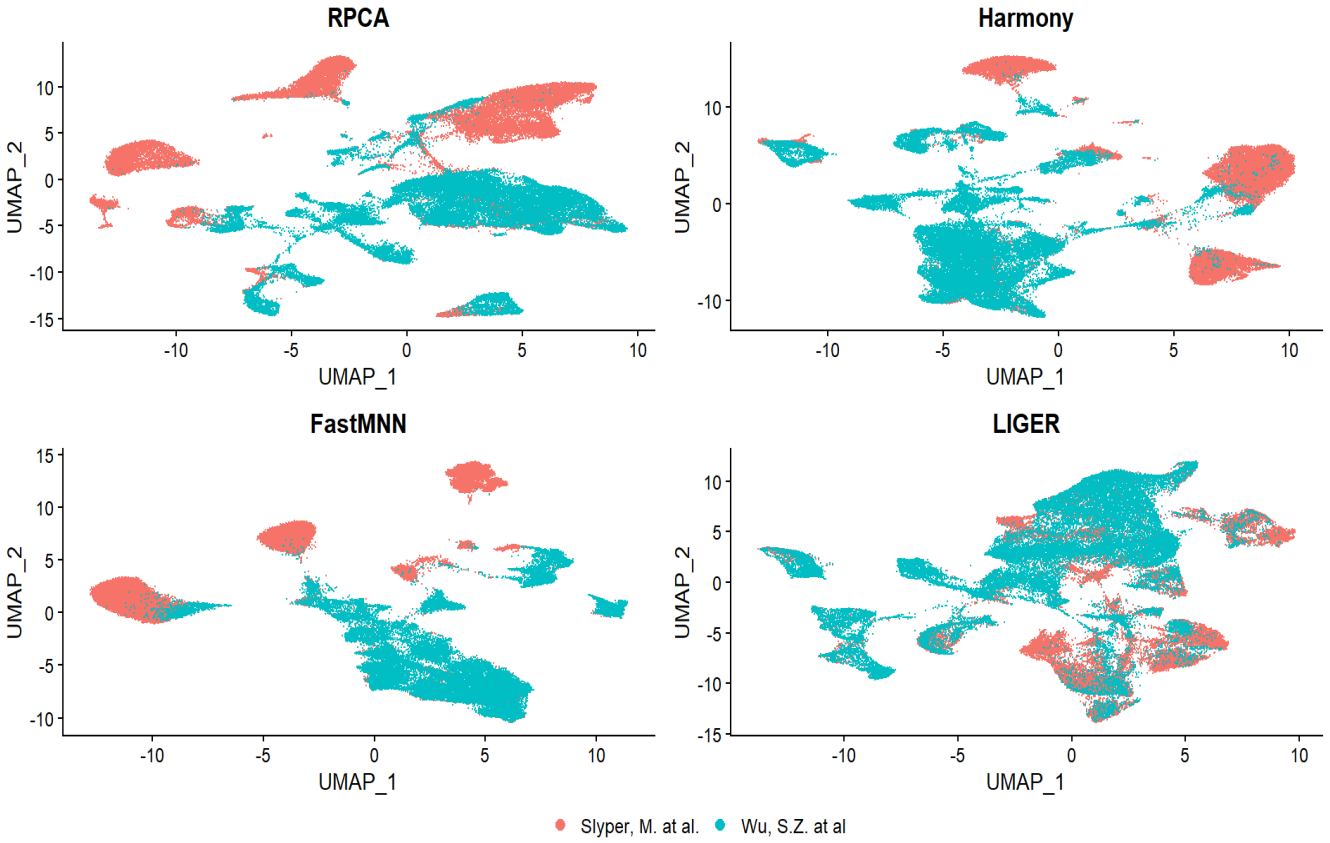

Figure 24: UMAP plots of the datasets integrated using 4 data integration techniques, grouped by source dataset.

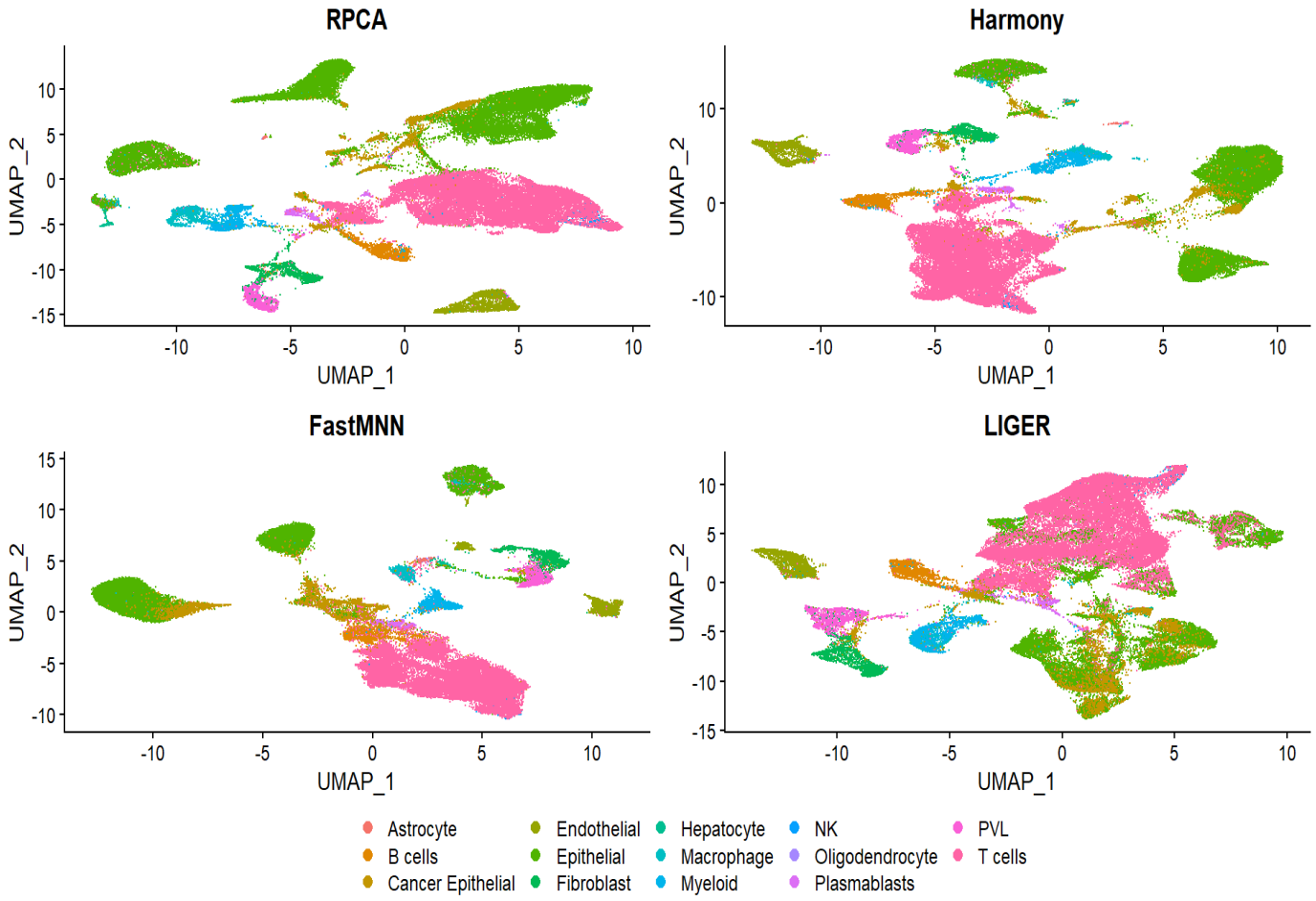

Figure 25: UMAP plots of the datasets integrated using 4 data integration techniques, grouped by celltype.

Conclusions and final considerations

In recent years, the field of bioinformatics has experienced a growth unprecedented in its history. The advent of new technological advances has enabled the collection of ever-increasing amounts of data and the execution of increasingly complex calculations and scientific simulations. This has allowed for the study of medical conditions in ways that were previously impossible. However, in order to have sufficient data to draw meaningful conclusions, it is often necessary to integrate together datasets originating from different studies and obtained under different environmental conditions and with different protocols. As single-cell omics is still a relatively novel field, the standardization of methods and procedures is still in many ways ongoing.

The main objective of this project was to analyze the performance of various integration tools, in order to determine the most effective ones in the context of breast cancer. Given the abundance of existing literature on Python-based methods, this project focused on R-based integration techniques, specifically RPCA, FastMNN, Harmony, and LIGER. To achieve this, the first step was to identify prelabeled breast cancer datasets, which were then loaded and processed. Due to the nascent nature of the field, there are as yet no established standards regarding data type and structure, which made this step especially challenging. Afterwards, I wrote the code for data integration and for the metrics. I experimented with different approaches to identify the most effective method, particularly in terms of computational efficiency, given the substantial data volume and the time required for computation. In addition to learning the R programming language, I had to gain familiarity with the statistical framework underlying the analysis, which relied heavily on machine learning methods. I also had to develop a working knowledge of cancer biology to evaluate the results and their biological implications.

The project presents several limitations. First, the availability of only two datasets limits the possibility for meaningful comparisons in larger studies. Second, time constrictions meant that the testing of only four data integration techniques in their standard forms was achievable, without extensive experimentation with algorithm parameters. A third limitation was the reliance on datasets that were prelabeled for celltype, which introduced an unknown variable in the benchmarking process, as the specific labeling methodology was not described in the studies and is not standardized in the field. Moving forward, a future expansion of this project could address these limitations by using a broader range of unlabeled datasets, which could be then labeled using a common, unified methodology. Additionally, more integration techniques could be tested, each using a variety of different parameter settings. The introduction of objective metrics to assess computational and time complexity would provide an additional measure for evaluating usability and efficiency in practical settings.

References

1. Xia, L. Heterogeneity of vascular cells in different breast cancer subtypes revealed by single-cell RNA-seq analysis. *Highlights Sci. Eng. Technol.* **36**, 1405–1414 (2023).
2. Łukasiewicz, S. *et al.* Breast Cancer-Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies-An Updated Review. *Cancers* **13**, (2021).
3. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).

4. Ferlay, J. *et al.* Cancer Observatory: Cancer Today. (2024).
5. Ferlay, J. *et al.* Global Cancer Observatory: Cancer Today. (2024).
6. Associazione Italiana di Oncologia Medica (AIOM) *et al.* I Numeri Del Cancro In Italia 2023. (2024).
7. Barzaman, K. *et al.* Breast cancer: Biology, biomarkers, and treatments. *Int. Immunopharmacol.* **84**, 106535 (2020).
8. Veronesi, P., Gentilini, O. D. & Leonardi, M. C. *Breast Cancer - Innovations in Research and Management*. (Springer International Publishing, 2017).
9. Lalloo, F. & Evans, D. G. Familial Breast Cancer. *Clin. Genet.* **82**, 105–114 (2012).
10. Pal, B. *et al.* A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *EMBO J.* **40**, e107333 (2021).
11. Ding, S., Chen, X. & Shen, K. Single-cell RNA sequencing in breast cancer: Understanding tumor heterogeneity and paving roads to individualized therapy. *Cancer Commun.* **40**, 329–344 (2020).
12. Wu, R. & Kaiser, A. D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.* **35**, 523–537 (1968).
13. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74**, 5463–5467 (1977).
14. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).
15. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
16. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
17. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).
18. Donati, G. The niche in single-cell technologies. *Immunol. Cell Biol.* **94**, 250–255 (2016).
19. Wang, S. *et al.* The Evolution of Single-Cell RNA Sequencing Technology and Application: Progress and Perspectives. *Int. J. Mol. Sci.* **24**, (2023).
20. Liu, N., Liu, L. & Pan, X. Single-cell analysis of the transcriptome and its application in the characterization of stem cells and early embryos. *Cell. Mol. Life Sci.* **71**, 2707–2715 (2014).
21. Regev, A. *et al.* The Human Cell Atlas. *bioRxiv* (2017) doi:10.1101/121202.
22. Jovic, D. *et al.* Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.* **12**, e694 (2022).

23. Van de Sande, B. *et al.* Applications of single-cell RNA sequencing in drug discovery and development. *Nat. Rev. Drug Discov.* **22**, 496–520 (2023).
24. Liang, P. *et al.* HelPredictor models single-cell transcriptome to predict human embryo lineage allocation. *Brief. Bioinform.* **22**, bbab196 (2021).
25. Ryu, Y., Han, G. H., Jung, E. & Hwang, D. Integration of Single-Cell RNA-Seq Datasets: A Review of Computational Methods. *Mol. Cells* **46**, 106–119 (2023).
26. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
27. Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).
28. Yu, X., Abbas-Aghababazadeh, F., Chen, Y. & Fridley, B. Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments. in *Methods in molecular biology (Clifton, N.J.)* vol. 2194 143–175 (2020).
29. McInnes, L., Healy, J., Saul, N. & Grossberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
30. Maaten, L. van der & Hinton, G. E. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
31. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
32. Seurat - Guided Clustering Tutorial. https://satijalab.org/seurat/articles/pbmc3k_tutorial.
33. Data visualization methods in Seurat. https://satijalab.org/seurat/articles/visualization_vignette.
34. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
35. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
36. Stuart, T. *et al.* Comprehensive integration of single cell data. *bioRxiv* (2018) doi:10.1101/460147.
37. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
38. Lun, A. A description of the theory behind the fastMNN algorithm. https://marionilab.github.io/FurtherMNN2018/theory/description.html.
39. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

40. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).
41. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
42. Büttner, M., Wolf, A. & Theis, F. kBET: k-nearest neighbour batch effect test. (2017).
43. Chapter 6 Principal Component Regression. https://bookdown.org/ssjackson300/Machine-Learning-Lecture-Notes/pcr.html.
44. Fouché, A., Chadoutaud, L., Delattre, O. & Zinovyev, A. Transmorph: a unifying computational framework for modular single-cell RNA-seq data integration. *NAR Genomics Bioinforma.* **5**, lqad069 (2023).
45. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
46. Adjusted Rand Index (ARI). https://oecd.ai/en/catalogue/metrics/adjusted-rand-index-%28ari%29.
47. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
48. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
49. Chiquet, J., Rigaill, G. & Sundqvist, M. aricode: Efficient Computations of Standard Clustering Comparison Measures. https://jchiquet.github.io/aricode/authors.html.
50. Sugihara, R., Kato, Y., Mori, T. & Kawahara, Y. Alignment of single-cell trajectory trees with CAPITAL. *Nat. Commun.* **13**, 5972 (2022).
51. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
52. Sheng, J. & Li, W. V. Selecting gene features for unsupervised analysis of single-cell gene expression data. *Brief. Bioinform.* **22**, bbab295 (2021).
53. Kasianchuk, N. *et al.* Scrutinised and compared : HVG identification methods in terms of common metrics. (2023).
54. Scialdone, A. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289–293 (2016).

Thanks

I'd like to take a moment to express my most sincere thanks to all of you who have listened, supported, encouraged, and inspired me along the way. The university journey is not always easy nor straightforward, especially under the shadow of the pandemic, and it has been full of challenges and surprises. These have been years of tremendous personal and academic growth, and I am lucky to have learnt from the many exceptional people I have met along the way.

I would like to express my deepest gratitude to Professor Giacomo Baruzzo and Dr. Giulia Cesaro for their invaluable guidance, support, and encouragement throughout the process of writing this thesis. I am also indebted to Professor Barbara Di Camillo for making my internship in Austria possible, as well as to Professor Francesca Finotello, for her mentorship and support while in Innsbruck. I would also like to extend my gratitude to my colleague Dr. Katharina Huber, for her collaboration and comradeship during the project in Austria.

I want to express my deepest thanks to my family: my mother Antonella, my father Roberto, and my only and favorite brother Kiko. Thank you for affection, your attentive listening, and your support, and for encouraging me to make my choices even when they weren't the natural or obvious ones.

The list of people to thank is extensive, so I extend my gratitude to everyone who has contributed. I would like to sincerely thank doctor Adam Chłopowiec, whose encouragement and support through my last two years of studies have proved to be invaluable. Your patience, fellowship and encouragement have played a huge role in my personal and academic growth. My final thanks go to my dear Graziella and to the city of Padova, for accompanying me along my academic journey and for being there at every special moment of these years.