Charles University

Faculty of Science

Molecular biology and biochemistry of organisms

Zuzana Ševčovičová

Algorithms for integration of single-cell transcriptomic data
Algoritmy pro integraci dat z transkriptomiky jednotlivých buněk

Bachelor's thesis

Supervisor: Mgr. Michal Kolář, Ph.D.

Co-advisor: Rajendra Kumar Labala, M.Sc.

Prague, 2024

I declare that I have prepared the thesis independently and that I have listed all information sources and literature used. Neither this thesis nor any substantial part of it has been submitted for another or the same academic degree.

In Prague April 29, 2024 Zuzana Ševčovičová

Acknowledgement

I would like to express my sincere gratitude to my supervisor Mgr. Michal Kolář, Ph.D. and co-advisor Rajendra Kumar Labala, M.Sc. for teaching and guiding me through the process of writing this thesis. I am truly grateful for the knowledge and insights they've shared with me. Their support, along with the time and effort they've invested in helping me understand the subject matter and navigate the thesis writing process, mean a lot to me. Working under their guidance has been a rewarding experience, and I really appreciate the opportunity. Furthermore, I would like to thank Lucie Pfeiferová, M.Sc., for providing me with the data I was able to use for this work.

I also want to extend my gratitude to my entire family for their unwavering support during my studies. Lastly, I am thankful to my classmates for their support throughout our academic journey and the thesis writing process.

Abstract

Single-cell transcriptomics represents an innovative technique that allows for the examination of gene expression of individual cells in tissues or other complex biological samples. The data generated through this approach offer crucial insights into various biological domains, such as embryology, oncology, and immunology. Understanding the functioning and mutual interactions of individual cells within tissues indeed requires knowledge of their individual expression profiles.

Given the novelty of this method and technical challenges faced during data generation, comparison of different samples may be difficult and affected by various batch effects. Effectively reconciling the information obtained regarding gene expression in distinct biological replicates or technical replicates generated by different methodologies may be onerous. The development of bioinformatics approaches to address this challenge is an ongoing process. The objective of this thesis is to:

1. survey the existing solutions for integration of single-cell transcriptomic data,
2. benchmark these solutions using publicly available datasets, specifically peripheral blood mononuclear cells, and
3. select the most suitable integration method to apply to a tumour microenvironment dataset generated by our laboratory.

This bachelor project encompasses the aforementioned steps, ultimately aiming to enhance our understanding of single-cell transcriptomics data integration.

Keywords: single-cell RNA sequencing, integration algorithms, batch correction, peripheral blood mononuclear cells, vestibular schwannoma

Abstrakt

Transkriptomika jednotlivých buněk je inovativní technika, která umožňuje zkoumat genovou expresi jednotlivých buněk v tkáních či jiných komplexních biologických vzorcích. Data získaná tímto přístupem přinášejí zásadní poznatky v různých biologických oblastech, jako jsou například embryologie, onkologie či imunologie. Skutečné pochopení fungování a vzájemných interakcí jednotlivých buněk v rámci tkání vyžaduje znalost jejich individuálních expresních profilů.

Vzhledem k novosti této metodiky a technickým výzvám při generování dat může být porovnání různých vzorků obtížné a ovlivněno různými dávkovými efekty. Efektivní srovnání informací získaných ohledně genové exprese v odlišných biologických replikátech nebo v technických replikátech generovaných různými metodami může být náročné. Vývoj bioinformatických přístupů k vyřešení tohoto problému stále probíhá.

Cílem této práce je:

1. Provést průzkum stávajících řešení pro integraci dat z transkriptomiky jednotlivých buněk,
2. porovnat tato řešení s použitím veřejně dostupných souborů dat, konkrétně mononukleárních buněk periferní krve, a
3. vybrat nejvhodnější metodu integrace, která bude použita pro soubor dat z nádorového mikroprostředí vytvořený naší laboratoří.

Tato bakalářská práce zahrnuje výše uvedené kroky a v konečném důsledku má za cíl zlepšit naše postupy při integraci dat z transkriptomiky jednotlivých buněk.

Klíčová slova: sekvenovanání RNA jednotlivých buněk, integrační algoritmy, dávková korekce, mononukleární buňky periferní krve, vestibulární schwannom

Contents

1 Introduction

The advent of single-cell RNA sequencing (scRNA-seq) heralded a significant breakthrough in 2009 with the publication demonstrating the sequencing of mRNA at the individual cellular level [1], just two years after the application of RNA sequencing techniques for bulk cell population analysis [2]. While bulk RNA sequencing provides insights into average expression levels across cell populations, the emergence of scRNA-seq enables exploration of gene expression variations among individual cells, uncovering rare populations that might otherwise go unnoticed in pooled cell analyses.

Experimental design employing scRNA-seq may involve sequencing of multiple distinct batches. Whether these batches are technical replicates or comparisons of samples from different conditions (e.g. from a healthy or diseased conditions, samples with different time intervals after exposure to a particular agens etc.). Analysing separate datasets independently poses the risk of misinterpreting results, as unintended technical and biological variations may confound the analysis. Given the high sensitivity of scRNA-seq technology, the integration of scRNA-seq datasets ensures more robust and reliable results, facilitating downstream analysis and the evaluation of meaningful outcomes.

Differences between datasets may stem not only from biological factors, such as tissue type and cell state, but also from sequencing protocols, including sample preservation methods, single-cell suspension preparation, cell sorting, number of captured and successfully sequenced cells, and library preparation methods. The stochasticity of gene expression values is induced by the success of cell capture or the success of subsequent PCR amplification. These variations introduce unwanted batch effects to the datasets, arising from both technical and biological differences, which should be addressed before drawing any conclusions. Therefore, it is essential to develop and compare mathematical statistical models tailored to tackle this challenge.

In this thesis, we focus on benchmarking the integration algorithms such as Seurat [3], Harmony [4], FastMNN [5], Stacas [6], Scanorama [7], scVI [8] and CanSig [9], on publicly available scRNA-seq datasets, specifically peripheral blood mononuclear cells. After evaluation of selected approaches, we will integrate four distinct datasets originating from scRNA-seq of vestibular schwannomas. Application of scRNA-seq in investigation of the tumour microenviroment enables characterisation of the cellular compartments of tumours as well as identification of the expression heterogeneity within tumour microenvirment.

Through this thesis, we aim to contribute to the comparison of scRNA-seq integration methods and its application in understanding complex biological systems.

2 Vestibular schwannoma

Vestibular schwannomas (VSs), also known as *acoustic neuromas*, are benign brain tumours originating from the Schwann cells [1] of the vestibulocochlear nerve. VSs are the most prevalent benign tumours found in the cerebellopontine angle and internal auditory canal [10]. These tumours exhibit slow but progressive growth and are typically well-circumscribed and encapsulated.

VSs characteristically arise within the internal auditory canal and affects vestibular portion of the vestibulocochlear nerve [11]. The vestibulocochlear nerve, also known as cranial nerve eight (CN VIII), comprises the vestibular and cochlear nerves. These nerve roots originate in the brainstem and extend to the inner ear through the internal auditory canal. Upon reaching the inner ear, the vestibulocochlear nerve splits into the vestibular and cochlear branches. The vestibular nerve innervates the vestibular system of the inner ear and relays information related to motion and body position, while the cochlear nerve extends to the cochlea, forming the spiral ganglia that facilitate the sense of hearing [12].

In the context of the surrounding microanatomy of the posterior cranial fossa, VSs present a wide spectrum of symptoms. Approximately 90% of patients experience ipsilateral sensorineural hearing loss [13], while 61% report dizziness or imbalance [13], [14]. Additionally, asymmetric tinnitus affects 55% of patients. In the initial stages, the prevalence of tinnitus is very low, affecting less than 1% of patients. In the later stages of the disease, especially after the onset of asymmetric hearing loss, the incidence of asymmetric tinnitus increases [15]. The loss of binaural hearing in these cases can lead to increasing difficulties with sound localisation and speech comprehension, particularly in the presence of background noise.

VSs are categorised based on their cause of their occurrence. They are divided into sporadic and familial VS. Familial VS typically arise in individuals with tumour predisposing genetic disorders, such as neurofibromatosis type 2 [16]. Biallelic mutations in the neurofibromin 2 (NF2), which acts as a tumour suppressor, can lead to the development of bilateral vestibular schwannomas. Typically, the biallelic deactivation of the NF2 gene is a result of a combination of a point mutation or a multiexon deletion [17]. NF2 exhibits an autosomal dominant inheritance pattern, with approximately half of cases inherited from affected parents and slightly over 45% resulting from de novo mutations [18]. In contrast, sporadic VS occurs independently of genetic predispositions, often arising from spontaneous mutations in genes, and in most cases gives rise to unilateral VS. Sporadic unilateral VSs are the most prevalent, constituting the majority of cases, while bilateral tumours are rarer, accounting for only 5% of cases [18].

One characteristic of VS is their slow progression and indolent tumour growth. Remarkably, up to 75% of cases exhibit no growth during the 3.6 years of observation. In cases where tumour growth is observed, the average growth rate of sporadic VS is approximately 1.1 mm per year in diameter [19].

[1] Schwann cells, which are a type of glial cells, are responsible for creating the myelin sheath around axons in the peripheral nervous system. Each Schwann cell wraps around a segment of the axon, forming multiple layers of myelin sheath. These cells are named in honour of Theodore Schwann (1810–1882), who proposed the cell theory stating that all living organisms are made up of cells.

There are three primary therapeutic approaches for the treatment of vestibular schwannomas. In cases where the tumour does not exhibit signs of progressive growth and its diameter remains under 2 cm, a "wait-and-scan" approach is typically employed. However, for growing VS tumours, treatment options include stereotactic radiotherapy or microsurgical resection, both of which carry risks of increased mortality or facial nerve damage [20].

2.1 Molecular biomarker of VSs

Extensively researched molecular biomarker for VSs is the mutation of the neurofibromin 2 gene (NF2). As previously stated, around 5% of VS cases are attributed to a genetic predisposition involving a tumour suppressor mutation in the NF2 gene, while approximately 50% of sporadic VSs cases result from a spontaneous mutation in the NF2 gene. Loss-of-function mutations in the NF2 gene are recognised as pivotal drivers of VS formation.

Situated on chromosome 22q12, the NF2 gene encodes product merlin, also referred to as the moesin-ezrin-radixin-like protein [21]. NF2 consists of 17 exons that are alternatively splices to form variety of isoforms. Two common isoforms are isoform 1 and 2, which differ in exon 16 retention [22], [23]. Merlin shares significant sequence homology (about 64%) with members of the ERM protein family. These proteins are vital for connecting cytoskeletal elements with membrane proteins. The N-terminal domain contains a highly conserved FERM domain (residues 1 - 335), an α-helical domain (residues 336 - 505), and a C-terminal tail domain (residues 506 - 595) [24].

The merlin is regulated through post-translational modifications. Phosphorylation, mediated by p21-activated kinase or PKA, at the critical regulatory residue Ser^{518} converts protein from its active to inactive state. When merlin is dephosphorylated, it functions as a growth suppressor [25]. Interestingly, studies on the product of the mutant NF2 gene have revealed a correlation between protein degree of closure and functionality. It has been observed that the dephosphorylated mutant protein usually tends to adopt a more closed conformation, impairing its proper functioning [26].

Tumourigenic Schwann cells are unable to monitor the cell count in their surroundings and continue to proliferate. Merlin inhibits cell growth after interacting with the hylauronic acid transmembrane receptor CD44, which monitors cell density in the surrounding area [27]. By the process of contact mediated inhibition of proliferation, activated merlin regulates intracellular signalling cascades associated with tumour formation, such as the Ras/Raf/MEK/ERK pathway [28], the PI3K/Akt pathway [29], the Rac/p21-PAK/c-Jun Kinase pathway [30], and the Hippo signaling pathway [31].

In 1986, Dvorak proposed a theory: tumours are wounds that never heal [32]. This concept also applies for VSs, where cells lacking the NF2 gene trigger a perpetual cascade of wound response signals. VS adopt transcriptional states similar to repair type Schwann cells by upregulating genes crucial for acute nerve repair and immune cell recruitment. This observation indicates a significant involvement of immune cells in tumour progression, underscoring the importance of further exploring the tumour microenvironment of VS [33].

2.2 Tumour microenviroment (TME)

TME usually consists of immune cells, fibroblasts, platelets, stem cells, and other types of cells. The non-cellular components are cytokines, chemokines and growth factors that effect tumour progression [34]. The main challenge in determining the most suitable treatment strategy lies in the understanding of variability observed in the growth rates of these tumours. Given the diverse disease progression observed in VS patients, it is essential to investigate the intratumoural composition at the molecular level.

Employing scRNA-seq technology enables the study of molecular patterns in VSs, facilitates investigation of the TME, and offers new insights into this complex issue. Understanding these will allow development of new treatment modalities to arrest or reverse tumour growth. It is one of the most widespread theories that future tumour growth is induced by its microenviroment [34].

VSs are characteristic for the presence of tumour-associated macrophages (TAMs), pivotal in regulating tumour progression. TAMs contribute significantly to various facets of tumour development, including supporting angiogenesis, tumour cell proliferation, invasion, metastasis, and mechanisms of resistance to treatment. TAMs are typically categorised into two main types: M1-type macrophages, also known as classically activated macrophages, and M2-type macrophages, referred to as alternatively activated macrophages [35].

M1-type macrophages are primarily responsible for producing pro-inflammatory cytokines like IFN-γ, TNF-α, and IL-18, which are crucial for initiating and maintaining host defence mechanisms. On the contrary, M2-type macrophages dampen anti-tumour inflammatory responses in order to protect surrounding healthy tissue from the collateral damage often caused by M1 macrophages. They are typically encountered during later stages of the inflammatory response [36].

Within the TME, M2-type macrophages exert various effects, including the suppression of anti-tumour immunity, facilitation of tumour matrix remodelling, and stimulation of angiogenesis. The polarisation of macrophages into M2-type within the TME is predominantly driven by the production of cytokines such as IL-4, IL-13, IL-33 or macrophage colony-stimulating factor (M-CSF1, produced by T_H1 lymphocytes, basophils, and innate lymphoid cells [37], [38].

One of the possible indicators of future tumour growth is a increase in the expression of CSF1 (colony stimulating factor 1) by VS Schwann cells. Via CSF1-CSFlR signaling, VS promote myeloid cells to migrate and proliferate [38].

Research focusing on TAMs has revealed that M2-type macrophages are more prevalent in rapidly growing VS tumours. Specifically, analyses of CD163 (the hemoglobin scavenger receptor) expression, a specific marker for M2-type macrophages, which is involved in anti-inflammatory pathways to protect oxidative stress and tissue damage, have shown significantly higher levels in fast-growing VS tumours [36].

Another marker of M2-type macrophages, whose role is not fully understood, is the upregulation of MS4A4A. Expression of MS4A4A is not directly responsible for tumour growth but rather plays role in enhancing NK cell mediated resistance to metastasis [39].

3 Summary of data analysis

In this study, our focus was on benchmarking integration strategies on scRNA-seq datasets from various conditions and then perform integration on four datasets from scRNA-seq of VSs.

For comparative analysis, we selected eight distinct datasets obtained from single-cell RNA sequencing of peripheral blood mononuclear cells (PBMCs). These datasets encompass a diverse array of immune system cell types, such as monocytes, dendritic cells, T cells, B cells, and natural killer cells. PBMCs are frequently used in comparative studies due to their well-defined characterization and annotation, rendering them an ideal model for such analyses [40].

Our selection criteria for choosing datasets aimed to encompass variations in three primary aspects: the number of cells sequenced, the preservation method employed, and whether the sequencing was performed on whole cells or isolated nuclei, see Table 1.

	Datasets	**Cell or Nuclei**	**Cell count**	**Preservation method**	**Year**
1	10k human PBMCs stained with TotalSeq™-B human universal cocktail, singleplex sample	scRNA-seq	8958	Fixed	2022
2	20k human PBMCs, 3' HT v3.1, Chromium X	scRNA-seq	23837	Fresh	2021
3	Human PBMC from a healthy donor, 10k cells - multi (v2)	scRNA-seq	10548	Fresh	2020
4	Human PBMC from a healthy donor, 1k cells (v2)	scRNA-seq	972	Fresh	2020
5	Integrated GEX, TotalSeq™-C, and BCR analysis of Chromium Connect generated library from 10k human PBMCs	scRNA-seq	11075	Fresh	2022
6	PBMC from a healthy donor - granulocytes removed through cell sorting	snRNA-seq	2711	Cryopreserved	2021
7	PBMC from a healthy donor - no cell sorting	snRNA-seq	3009	Cryopreserved	2021
8	Peripheral blood mononuclear cells from a healthy donor - Chromium Connect	scRNA-seq	3363	Cryopreserved	2020

Table 1: Datasets used for comparative analysis of integration algorithms. All are publicly available on the 10× Genomics portal, see *https://www.10xgenomics.com/datasets*.

3.1 Dataset variations

The preservation method ScRNA-seq is a highly sensitive technique, and improper sample handling can invalidate the results. While working with fresh cells is optimal, it's not always feasible. This has led to a pressing need for developing methods for cell fixations without altering their transcriptional profile. Therefore, researchers have explored various cell preservation methods, such as dimethyl sulfoxide (DMSO)-based cryopreservation, methanol, or glyoxal fixations for scRNA-seq. These approaches have gained popularity for their ability to preserve cells while enabling downstream analysis.

DMSO cryopreservation, methanol or glyoxal fixation have been found to have minimal impact on the transcriptional profile of cells, with only a slight increase in mitochondrial transcript abun-

dance, which may reflect cell permeabilization. Despite this, the overall results are encouraging, as fixed samples maintain a high-quality transcriptome with impressive purity and library complexity [41], [42].

Cells vs. Nucleus ScRNA-seq can be performed using either whole cells or isolated nuclei, and the choice between the two methods depends on the research objectives and the biological questions being addressed. Both options can be used for cell type identification.

The concept of single-nucleus RNA sequencing (snRNA-seq) is to assess the transcriptome of individual nuclei instead of individual cells. This is a significant approach that is applicable to tissues that cannot be easily broken down into single-cell suspensions and preserved samples (such as in the cells of central and peripheral nervous system), and also reduces the potential alteration of gene expression that may be caused by disintegration. SnRNA-seq has been shown to perform well for sensitivity and classification of cell types [43], [44].

The main advantage of snRNA-seq is the ability to avoid the presence of artificial gene expression. Through a simple freezing process, it is possible to gain access to the nucleus for sequencing. This is a capability that is lacking in scRNA-seq because usually it requires the use of enzymes that enable dissociation of cells, which can induce transcriptome changes [45]. However, a disadvantage of snRNA-seq is the limited amount of RNA accessible in the nucleus for sequencing compared to the amount obtained by sequencing the whole cell.

The number of cells sequenced Integrating multiple datasets can offer a potent means of identifying distinct cell populations and shedding light on the inherent heterogeneity within these populations. However, different numbers of cells in individual datasets may pose a challenge, unlike larger datasets, smaller datasets might not encompass all cell types. It's crucial to preserve the heterogeneity between datasets during integration. Therefore, algorithms must be carefully designed to avoid over-integrating of the datasets with various sizes.

Cell sorting To initiate scRNA-seq, cells should be broken down into a single cell solution. Certain procedures exist that can selectively isolate specific cell subpopulations from a single cell solutions. FACS (fluorescence-activated cell sorting) [46] is one such protocol, which was also employed in the preparation of 6th sample (PBMC_6_3k_Nuclei).

FACS, a specialised type of flow cytometry, operates by utilising fluorescent markers. Designed monoclonal antibodies marked with fluorescent tags attach to surface proteins on the cells of interest. The cell mixture is then oscillated at an ideal frequency to create droplets, with attention paid to dilution preparation to reduce the possibility of doublet creation within each droplet. As these droplets traverse a laser and a detector, those with labelled cells are recognised, and the computer selectively redirects them using charge application. Electric pulses, called sorting triggers, charge fluid exactly when desired cell is forming. This technique allows for the targeted sorting of specific cells from a heterogeneous population, determined by the expression of their cell surface markers [47].

However, this method necessitates a substantial sample volume, and it relies on prior knowledge of the protein structure of interest for the design of the monoclonal antibodies.

This technique was used to selectively remove granulocytes from the PBMC_6_3k_Nuclei dataset. The absence of granulocytes were confirmed by expression analysis. The absence of a subpopulation of cells could present challenge for integration algorithms. It's essential that no method attempts to align datasets forcefully, and preserve this difference instead.

3.2 Data analysis

Preprocessing of data ScRNA-seq technologies generate vast amounts of raw data, often ranging from 10^6 to 10^{10} reads obtained from 10^3 to 10^6 cells in a single experiment [48]. To prepare this data for downstream analysis, such as normalisation, clustering, trajectory analysis, cell type identification, and data integration, it must undergo several preprocessing steps.

For UMI (unique molecular identifier) [49] based scRNA-seq protocols these preprocessing steps typically include cell barcode detection, alignment of cDNA reads to reference genomes, and UMI correction (deduplication). These processes result in the generation of cell-by-gene count matrices. Once these matrices are obtained, the data must undergo quality control steps to filter out low-quality cells. In the following, we describe the 10× Genomics technology, which has been employed for the generation of all respective datasets.

Each cDNA read obtained from scRNA-seq carries two crucial sequences incorporated during library preparation. First, there are cell barcodes [50], consisting of unique 12nt sequences that aid in assigning sequence reads to their respective cell. Reads sharing identical barcodes are grouped together, representing transcriptomes of individual cells. Second, UMIs are employed to mitigate possible bias caused by PCR amplification during library construction. This ensures that each transcript is quantified only once during the process[51].

Depending on the selected preprocessing software, each read is aligned to a reference genome to generate information about expressed genes. In our study, we choose eight different datasets published between 2020 and 2022 that were preprocessed using the Cell Ranger software [52], specifically designed for the 10× Genomics Chromium platform. Cell Ranger provides pre-built GENCODE reference packages used for gene mapping in the Ensembl project [53]. Ensembl offers an annotated and continuously updated database of human genome. The pre-built packages are occasionally updated to ensure accurate annotation of genes. The outcome of preprocessing steps are cell-by-gene count matrices.

Quality control and selecting cells for further analysis While scRNA-seq technology is experiencing a surge in popularity, it still presents several limitations that must be addressed during data analysis. The inherent challenges include the low levels of transcripts within individual cells, inefficient mRNA capture, and technical losses during reverse transcription (RT). These factors contribute to the creation of noisy, high-dimensional, and sparse gene expression matrices. In fact, scRNA-seq data are known for their characteristic sparsity and noise.

Raw preprocessed data still contain many empty droplets or low-quality cells, which can cause

numerous problems during downstream analysis. One of them is the risk of forming similar patterns between cells with low expression values and forming misleading clusters of low-quality cells that may resemble novel cell types [54]. For these reasons, applying thresholds for filtering low-quality cells is essential for further analysis.

For quality control, we evaluated the following three parameters to determine the quality of the cells in the datasets and set thresholds for data filtering: *Count Depth* (The number of UMI counts per barcode) and *The number of genes per barcode*: These parameters reflect the number of transcripts sequenced within a single cell. Low values might signify subpar sequencing, empty droplets or nonviable cells, while an unusually high count could indicate doublets. *The fraction of mitochondrial genes* (number of mitochondrial genes per barcode): Often presented as the proportion of transcripts assigned to mitochondrial genes. A high mitochondrial fraction suggests apoptotic cells or cells with membrane damage during sample preparation.

Thresholds for these three parameters should be set with consideration for the biological variability of cell types and the specific tissue being sequenced. For instance, cells with a high fraction of mitochondrial counts may be associated with respiratory processes, while cells with higher counts may simply be larger or actively proliferating.

Data normalisation and variance stabilisation As mentioned earlier, due to the different transcripts capture efficiency and PCR amplifications, data obtained from scRNA-seq contains technical bias across all cells. The sequencing depth (the number of genes or molecules detected per cell) may exhibit considerable variation among individual cells within a single experiment.

For purposes of our analysis, we applied standard normalisation method found in popular scRNA-seq packages such as Seurat [3] and Scanpy [55]. This is a two step method, referred as log-normalisation, which belongs to a group of global-scaling normalisation methods. Feature counts for each cell are divided by the total counts for that cell and multiplied by the `scale.factor`, which is by default set to 10 000 and represent expected number of all transcripts in a cell. This is then natural-log transformed [3]. Transformation of the expression value matrix ensure accurate comparison of gene expression levels among cells and obtain correct relative gene expression values for further analysis [56].

Feature selection ScRNA-seq provides invaluable insights into the transcriptional profiles of thousands of cells. However, the high-dimensional data generated poses a significant challenge for downstream analysis, known as the curse of dimensionality. To address this challenge, researchers often turn to the selection of highly variable features.

Each tissue has its own gene activity profile. While some genes show consistent expression across all cells (like housekeeping genes), others exhibit distinct expression patterns that differentiate tissues and cell types from each other. In order to understand cellular heterogeneity, it's essential to identify these highly variable features that carry the most informative content for downstream analysis.

By retaining only a chosen number of highly variable features, the computation time for subse-

quent calculations of large-scale scRNA-seq data is significantly accelerated [57], [58]. Depending on the task and complexity of the dataset, typically are selected 1000 - 5000 highly variable genes (HVGs). The method for choosing highly variable features, a key component of the well-known Seurat, operates under the presumption that genes (features) exhibiting high relative variance compared to the average expression are influenced by biological factors rather than being a product of technical noise [59].

Dimension reduction As previously stated, the information obtained from sequencing the transcriptome of single cells is multi-dimensional. Each dataset comprises thousands of genes and thousands of cells (after quality control and feature selection). Visualisation aims to represent scRNA-seq data in two or three dimensions while preserving the biological structure as accurately as possible. The most commonly used techniques for reducing the dimensionality of the expression matrix include linear and nonlinear methods.

Principal Component Analysis (PCA) [60] belongs to linear approaches used to visualise high-dimensional data to low-dimensional space and forms the foundation for certain nonlinear approaches. This method first examines the expression matrix of a dataset to identify a combination of genes that are significant for capturing the variability of gene expression. This gene expression vector called a principal component (PC) is derived from the gene covariance matrix. PCs are weighted based on their importance in capturing the variability of gene expression [61].

PCs serve as latent variables that capture the largest variance within a dataset. Typically, the first PCs capture the most variance. A specific number of PCs can often adequately represent the variation in the data. The optimal number of selected PCs can be determined through methods such as the Elbow plot or Jack Straw plot.

Since PCs are derived from features that capture the largest variance within the dataset, analysing the genes associated with these components allows us to identify the significance of feature variability in the datasets.

This straightforward linear method offers numerous benefits. The distances among cells are uniform and accurately represent the actual distances. However, a drawback is that PCA does not portray the dataset structure as effectively as non-linear methods.

t-distributed Stochastic Neighbour Embedding (t-SNE) [62] is a non-linear technique for reducing dimensionality, derived from SNE (stochastic neighbour embedding). It was developed to address a limitation of the original SNE method, which had a tendency to cluster points towards the map's centre [63].

t-SNE prioritise preserving local relationships over global ones. t-SNE is proficient at maintaining proximity among similar cell populations, although the cluster-cluster distances may not accurately reflect actual distances. So using t-SNE for identifying connections between cell populations is not recommended. Another limitation is the significant computational time. So researchers are now opting for alternatives like UMAP.

Uniform Manifold Approximation and Projection (UMAP) [64] is a non-linear technique for reducing dimensionality. UMAP, introduced in 2018, revolutionised the landscape of data visu-

alisation methods by offering a unique blend of efficiency and accuracy. Unlike other non-linear approaches, like t-SNE, UMAP ensures the preservation of both local and global structures while significantly reducing computation time. This innovation addresses a critical limitation in previous non-linear techniques, where cluster-cluster distances may not accurately reflect true distances, hampering the identification of connections between cell populations. UMAP has emerged as the go-to tool for visualising scRNA-seq data, empowering researchers with unparalleled insights into cellular dynamics and interactions [65].

Clustering ScRNA-seq datasets consist of groups of cells of different types. Cell groups should be determined by the similarity of their transcriptomes (based on their expression values) using various algorithms including unsupervised clustering.

The k-means clustering algorithm [66] is used to group cells into clusters using projection of the dataset into a low-dimensional space. Each cluster is associated with a centroid, which helps assign other cells to clusters by minimising their distances from the centroid. The centroids are located in dense areas where cell clusters are expected, and their selection is adjusted iteratively until a stable state is reached.

The widely-used Seurat package employs the Louvain algorithm [67] on single-cell k-nearest neighbours graphs, which identifies cell groups with stronger connections among themselves than with other cells. Louvain community detection algorithm is an unsupervised algorithm that does not require number of expected communities before running. Advantage of the Louvain algorithm is its resolution parameter, which enables the specification of the clustering scale for the cells [3]. It has been shown, that Louvain algorithm performed optimally for gene expression single cell data with high accuracy and fast performance [68].

Another prominent approach is the Leiden clustering method, predominantly implemented in Python programming language packages, such as Scanpy [55]. The Leiden clustering algorithm can be conceptualised as an enhancement of the Louvain algorithm. The Leiden surpasses the Louvain algorithm in both speed and the quality of outcomes [69].

We must note that efficiency and accuracy of clustering algorithm strongly depends on method and number of selected variable features.

4 Tools for scRNA-seq data integration

To compare algorithms for integration of single-cell transcriptomic data, we selected seven pipelines from R and Python programming languages. This comparison aims to highlight the disparities and advantages of each pipeline within their respective environments.

The R environment's packages include Seurat v5 [3], Harmony [4], STACAS [6] and FastMNN [5]. The Python programming language is represented by packages scVI [8], Scanorama [7], and CanSig [9].

Seurat, within the R programming language, and Scanpy, within Python, stand out as the predominant platforms for general analysing scRNA-seq data. Across various integration algorithms, these two platforms play pivotal roles in data preprocessing and comprehensive analysis. It's important to note that Seurat and Scanpy employ distinct formats for storing scRNA-seq data. Seurat generates the Seurat Object, while Scanpy utilizes the AnnData object format. Transferring data between these platforms demands a deep understanding of their respective object structures, in order to transfer data from Python to R and reverse.

To analyse all eight chosen datasets, we applied methods that we described earlier. The datasets were already processed by the submitters using Cell Ranger software [52] and been made publicly available on 10× Genomics portal. We applied quality control thresholds for each dataset separately, to filter out low quality cells (Table 2). Then we log normalised, scaled the count matrices and run PCA on the 2 000 selected highly variable features. After that we could construct the k-nearest neighbour graph using 30 principal components to cluster cells. All datasets were merged into one single object and projected in 2 dimensions using PCA and UMAP.

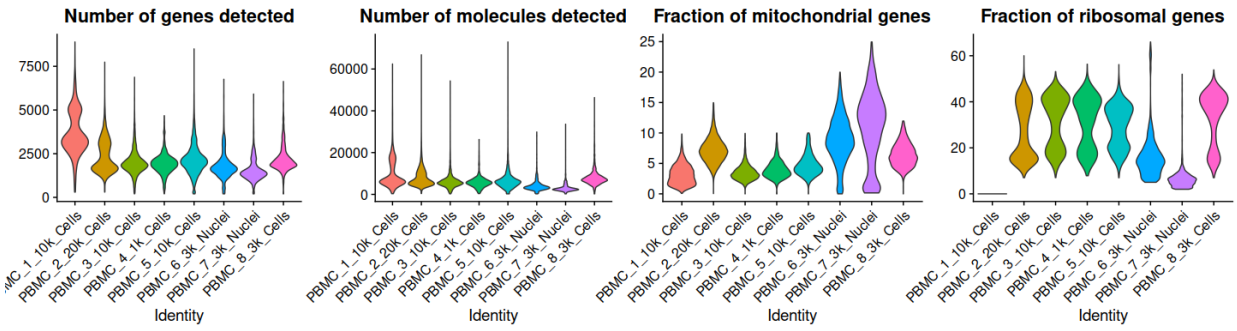


Figure 1: Violin plot showing the distribution of the cells on the preprocessed data. The first plot illustrates the distribution of expressed number of genes (features) detected in each cell. The second plot shows the number of detected molecules per cell. The third and fourth plots displays fraction of mitochondrial and ribosomal reads. Notably, datasets 6 and 7 from snRNA-seq exhibit higher fraction of mitochondrial transcripts.

On Figure 1, it is evident that the datasets exhibit a high degree of similarity. However, datasets PBMC_6_3k_Nuclei and PBMC_7_3k_Nuclei stands out due to their higher proportions of mitochondrial transcripts. This anomaly can be attributed to the origin of these datasets, as they were obtained from the snRNA-seq. In such cases, the proportion of mitochondrial genes to the total number of captured genes may be higher. When sequencing nuclei, read counts tend to be lower compared to scRNA-seq, resulting in a heightened representation of mitochondrial genes within the datasets.

Variations in the datasets also exist in the percentages of ribosomal genes. The dataset PBMC_1_10k_Cells had removed ribosomal transcripts prior to being published in the repository. Notably PBMC_6_3k_Nuclei and PBMC_7_3k_Nuclei datasets have smaller proportions of ribosomal genes, since they originate from snRNA-seq.

	Original data		**QC Thresholds**			**Filtered data**	
Dataset local ID	Genes	Cells	Genes per cell	% mito	% ribo	Genes	Cells
1_PBMC_10k_Cells	14 912	9 078	>300	<10	xxx	14 773	8 982
2_PBMC_20k_Cells	24 019	23 972	>300	<12	>7	23 274	22 939
3_PBMC_10k_Cells	20 233	10 691	>200	<10	>7	19 543	10 096
4_PBMC_1k_Cells	14 854	975	>200	<10	>7	14 205	902
5_PBMC_10k_Cells	20 985	11 318	>200	<10	>7	20 164	9 303
6_PBMC_3k_Nuclei	21 296	2 801	>200	<20	>5	20 259	1 983
7_PBMC_3k_Nuclei	21 524	3 047	>200	<25	>2	20 406	2 232
8_PBMC_3k_Cells	17 614	3 335	>200	<12	>7	16 986	2 904

Table 2: Table of quality control thresholds for preprocessing of PBMC scRNA-seq datasets.

Mutual nearest neighbours A mutual nearest neighbours (MNNs) search method [5], provides a solution for removing batch effects between related datasets. This approach is incorporated into a number of integration algorithms, such as Seurat, FastMNN, STACAS and others, that will be discussed subsequently.

The algorithm for finding MNNs between datasets first scales the data globally via cosine normalisation. Subsequently, the Euclidean distance between the normalised batches is computed. The results represent the cosine distances of the actual distances between the batches. Cosine distances are a more appropriate approach for algorithms because they are not dependent on scaling i.e. on technical variations between batches.

After computing the Euclidean distances are identified MNNs cell pairs. These pairs indicate cells with comparable expression values, with the discrepancy being attributed to the batch effect. A specific batch correction vector is then calculated for each MNN pair. The overall batch correction vector is determined by weighted average of the pair-specific vectors, which is then employed to correct the batch effect across all cells.

It is reasonable to expect that pairs of MMNs should match the same cell type, even if they are produced in separate batches. Variations in the levels of expression of MMNs pairs are attributed to batch effects. The primary requirement for MNN correction across datasets is the presence of at least one shared cell population to calculate a batch correction vector [5].

4.1 Seurat $_R$

Package Seurat is designed for the analysis and integration of scRNA-seq datasets in the programming language R [3]. The most recent version of Seurat, version 5 [70], has been available since November 2023. In contrast to its predecessor [71], Seurat v5 provides an enhanced and simplified framework for executing various integration algorithms.

As mentioned earlier, Seurat uses a data storage format known as Seurat Object, which belongs to R's S4 object-oriented system. The architecture condense single-cell genomics data (metadata) with additional information such as dimensionality reduction embeddings. Within the Seurat Object, layers are maintained to retain the original dataset's identity information. Raw counts are stored in the layer `counts`, normalised data are stored in the layer `data`, or z-scored/variance stabilised data are stored in layer `scale.data` [72].

Introducing new layers offers the benefit of integrating individual datasets within a single object, allowing for the execution of integration analysis within it. Unlike the previous version of Seurat v4, where datasets had to be kept in distinct objects. The ability to split and merge layers as required simplifies the analysis process considerably.

Anchor-based Canonical correlation analysis (CCA) integration Seurat's approach to data integration uses Canonical correlation analysis [73], which enables the condensation of relationships into a reduced dimensional space while retaining the key aspects across datasets. Initially, the aim is to identify linear combinations of genes that exhibit the highest correlation between the datasets. Through the application of nonlinear warping, these vectors are adjusted to account for variations in population density. The nonlinear warping locally compresses or elongates the vectors during the alignment process, resulting in a reduced-dimensional subspace encompassing all input data [3].

Next Seurat identifies MNN in created subspace, referred to as anchors by Seurat. By using the identified anchors, it becomes possible to align two datasets and correct the differences between them. The process of finding anchors is carried out in pairs of datasets until all datasets are aligned. The order is established by Seurat based on the similarities between the datasets, although it is possible to determine own order.

In Seurat v5 is possible to integrate datasets with a single command.

```
Int <- IntegrateLayers(
  object = merged_seurat, method = CCAIntegration,
  orig.reduction = "pca", new.reduction = "integrated.cca",
  verbose = FALSE)
```

Seurat has been demonstrated to possess one of the most effective algorithms for integrating datasets with diverse disparities. It excels in integrating large datasets, datasets with non-identical cell types or datasets with technical variations. Seurat demonstrates efficient integration across all these scenarios with a small number of incorrect matches [74].

4.2 Harmony $_R$

Harmony was first introduced in 2019 as a new, fast and sensitive technique for integrating scRNA-seq datasets [4]. Its unique algorithm allows for quick and easy exploration of multiple datasets integration.

The first step of integration using Harmony, is to reduce the dimensionality of all datasets using PCA. This step is also known as identifying low-dimensional embeddings of the cells. Harmony receives the coordinates of each cell, which it then uses in further analysis.

Each cell is assigned a potential cluster using the *soft k-means clustering technique* created by Harmony. Then, a centroid is chosen from each cluster that is created. The coordinates of this centroid are used to correct the coordinates of the remaining cells within the clusters. Harmony iterates through these steps until the cell clusters reach convergence. During each iteration, it clusters together similar cells from different batches, aiming to enhance the variety of batches within each cluster [4].

The Harmony integration algorithm can be executed within Seurat enviroment by utilising the `RunHarmony` function from the Harmony library on Seurat Object. By default, harmony will access the PCA cell embeddings, which have done earlier, and utilise them to run harmony integration algorithm. Hence, it is necessary to normalise the datasets in advance. This enables the creation of a combination of highly variable features that are used for generating principal components. By enabling the plot convergence argument, harmony will produce a convergence plot that illustrates the integration process, see Figure 2.

```
Int <- RunHarmony(merged_harmony, group.by.vars ="orig.ident",
    plot_convergence = TRUE, nclust = 50, max_iter = 10, early_stop = T)
```

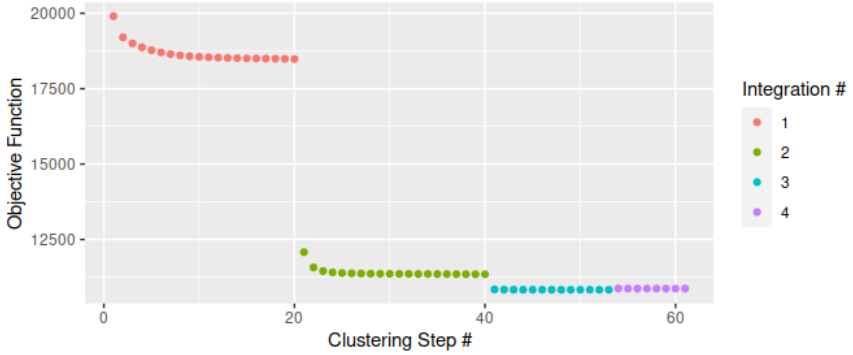


Figure 2: Integration process of harmony. Colours represent different iterations.

It is important to mention that Harmony removes the expression matrix at the initial stage, unlike Seurat, which retains these values. Harmony takes as input a matrix of coordinates and covariate labels for each cell. It outputs a matrix of modified coordinates that maintains the same dimensions as the original input matrix. Consequently, the integrated outcome will not contain the expression values of the individual genes.

By the fact that the integration algorithm does not need to compute corrected expression values, the time required for the analysis is shorter. The Harmony's low runtime also makes it

well-suited for the initial exploration of large datasets. [74]. However, we might lose biologically relevant data that might be required in the downstream analysis.

4.3 FastMNN $_R$

One commonly used method in various approaches is finding MNNs. As mentioned earlier on page 12, the algorithm starts by detecting MNNs pairs of cells between datasets to create links between them. The list of MNNs obtained is then used to calculate translation vectors for aligning the datasets in a common space.

FastMNN presents a more efficient computational method in relation to CPU runtime and memory consumption than MNNs. It employs the strategy of identifying MNNs within a subspace generated through PCA. This technique has significantly improved both the speed of execution and the precision of aligning datasets [5].

To run FastMNN integration algorithm on Seurat Object within Seurat interface, we can function `RunFastMNN`.

```
Int <- RunFastMNN(object.list = SplitObject(merged_pbmc, split.by =
    "orig.ident"))
```

4.4 STACAS $_R$

STACAS is a package that can be also incorporated into Seurat enviroment. It is specifically designed to compete with the powerful Seurat algorithms. Its uniqueness lies in new methods that seek to reduce the risks of over-clustering.

The STACAS algorithm belongs to linear embedding models, where a MNN search method is used to find anchors across all datasets. In contrast to the Seurat CCA algorithm, STACAS does not inherently rescale expression values to zero mean and unit variance as a default setting. Such rescaling could impact the preservation of biological variability across datasets [6]. Initially, the algorithm computes the level of similarities between datasets and return a cluster dendrogram (Figure 3). Based on this information, the alignment starts with datasets exhibiting the highest similarity score.

After identifying cell types, STACAS provides the option to incorporate cell type annotations for alignment guidance. In the cases where there are some inconsistency in cell types among anchors, STACAS imposes penalties, operating on the assumption that cell clusters should exhibit uniform cell type composition.

However, research has indicated that semi-supervised integration with STACAS using cell type annotations does not always provide a significant advantage. In scenarios where datasets exhibit increasing levels of cell type imbalance, unsupervised STACAS demonstrated comparable performance to semi-supervised integration. Conversely, when integrating datasets with similar cell types but significant batch effects, the semi-supervised STACAS integration algorithm outperformed the unsupervised approach. Additionally, when compared with Seurat CCA integrations,

semi-supervised STACAS integration outperformed Seurat, which intend to overcorrect batch effects [75].

The STACAS integration algorithm can be implemented in the Seurat interface by utilising the `FindAnchors.STACAS` and `IntegrateData.STACAS` functions from the STACAS library on Seurat Object.

```
anchors <- FindAnchors.STACAS(obj.list, anchor.features = 2000, dims =
    1:ndim)
st1 <- SampleTree.STACAS(anchorset = stacas_anchors, obj.names =
    names(obj.list))
```

```
Int <- IntegrateData.STACAS(stacas_anchors, sample.tree = st1,
    dims=1:ndim)
```

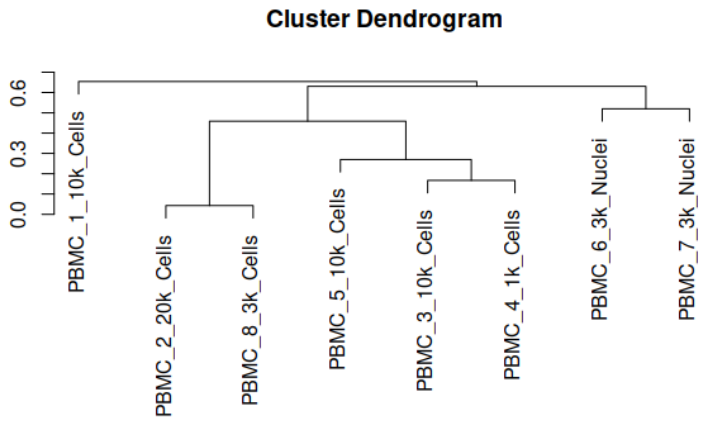


Figure 3: The dendrogram provides visual information on how datasets are grouped based on their similarity levels. This information is utilised in dataset integration, where datasets exhibiting the highest similarity levels are integrated first.

4.5 Scanorama *Python*

Scanorama's algorithm [7] for integration of scRNA-seq datasets is inspired from techniques used in image processing to stitch together overlapping images into a single panoramic view. Developed to address the challenge of integrating multiple scRNA-seq datasets with distinct transcriptional profiles, Scanorama stands out for its ability to integrate datasets without requiring shared cell populations across all samples [7].

Unlike other integration methods, Scanorama is agnostic to the order in which datasets are integrated, as it performs alignment among all pairs of datasets.

The Scanorama integration algorithm can be executed within Scanpy interference by utilising the `scanorama.correct_scanpy` function from the Scanorama library on divided AnnData object.

```
Int = scanorama.correct_scanpy(adata, return_dimred = True, knn = 40)
```

In comparison to Seurat, Harmony, fastMNN, and scVI, the Scanorama algorithm falls short in performance, when integrating un-annotated homogeneous datasets. Studies have revealed that Scanorama often overlooks homologous cell pairs and struggles to effectively integrate data [76]. However, Scanorama was specifically designed to compete with these robust tools in situations

where datasets display significant compositional differences, demonstrating better performance in such scenarios [7].

4.6 scVI *Python*

Single cell variational interference (scVI) [8] belongs to a deep learning model that uses a probabilistic framework to integrate single-cell transcriptomics data.

Deep Learning (DL) models [77] form a subset of machine learning (ML) models renowned for their utilisation of deep neural networks to analyse large and complex datasets. Inspired by the human brain's learning mechanisms, deep neural networks consist of multiple preprocessing layers, referred to as artificial neurons. These models excel at extracting biologically significant features from scRNA-seq data.

ScVI, grounded in a hierarchical Bayesian model, employs deep neural networks to define conditional distributions, presuming a zero-inflated negative binomial distribution [8].

In comparison to prior methodologies, scVI stands out for its ability to effectively integrate datasets, particularly those characterised by their size and complexity. Notably, scVI achieves this integration while faithfully preserving biological variations across multiple datasets, all without introducing any spurious artefacts or false signals [78].

The scVI integration algorithm can be performed on AnnData format preprocessed with package Scanpy. To run scVI integration, we design model and then execute function `model.train()`.

```
scvi.model.SCVI.setup_anndata(
    adata, layer="counts", batch_key = "batch",
    categorical_covariate_keys=["orig.ident"],
    continuous_covariate_keys=["percent_mito", "nCount_RNA"])
model = scvi.model.SCVI(adata, n_layers=2, n_latent=40,
    gene_likelihood="zinb", dispersion = "gene-batch")
```

```
model.train()
```

4.7 CanSig *Python*

CanSig, a novel integration method tailored for the analysis of tumour scRNA-seq data, was introduced in 2022 [9]. This innovative approach offers a fully automated computational solution for analysing transcriptional states across different scRNA-seq data obtained from multiple samples, especially on tumour scRNA-seq samples.

CanSig consists of four distinct modules for data analysis. The initial module focuses on data preprocessing, during which individual cells are categorised as either malignant or non-malignant. This step sets the foundation for subsequent data analysis. The second module is responsible for integrating the data in a manner that preserves the distinctions between malignant and non-malignant cells. CanSig employs a deep probabilistic generative model from scVI to implement cell integration. In the third module, the integrated data is clustered in a latent space using the Leiden clustering method. This is followed by differential gene expression analysis for each cluster

versus the rest of the cells. And in the fourth and last module are identified shared transcriptional states in the cells of interest [9].

Since the CanSig integration algorithm implement scVI cell integration, the commands are similar. Firstly is created model configuration and than is performed integration using function `integrate_adata()` .

```
integration_config = cansig.models.scvi.SCVIConfig(
    batch="batch", n_latent=40, random_seed=0,
    train=cansig.models.scvi.TrainConfig(max_epochs=135),
    discrete_covariates=None)
```

```
representations = integrate_adata(data=data, config=integration_config)
```

4.7.1 Comparison of integration algorithms on eight PBMCs datasets

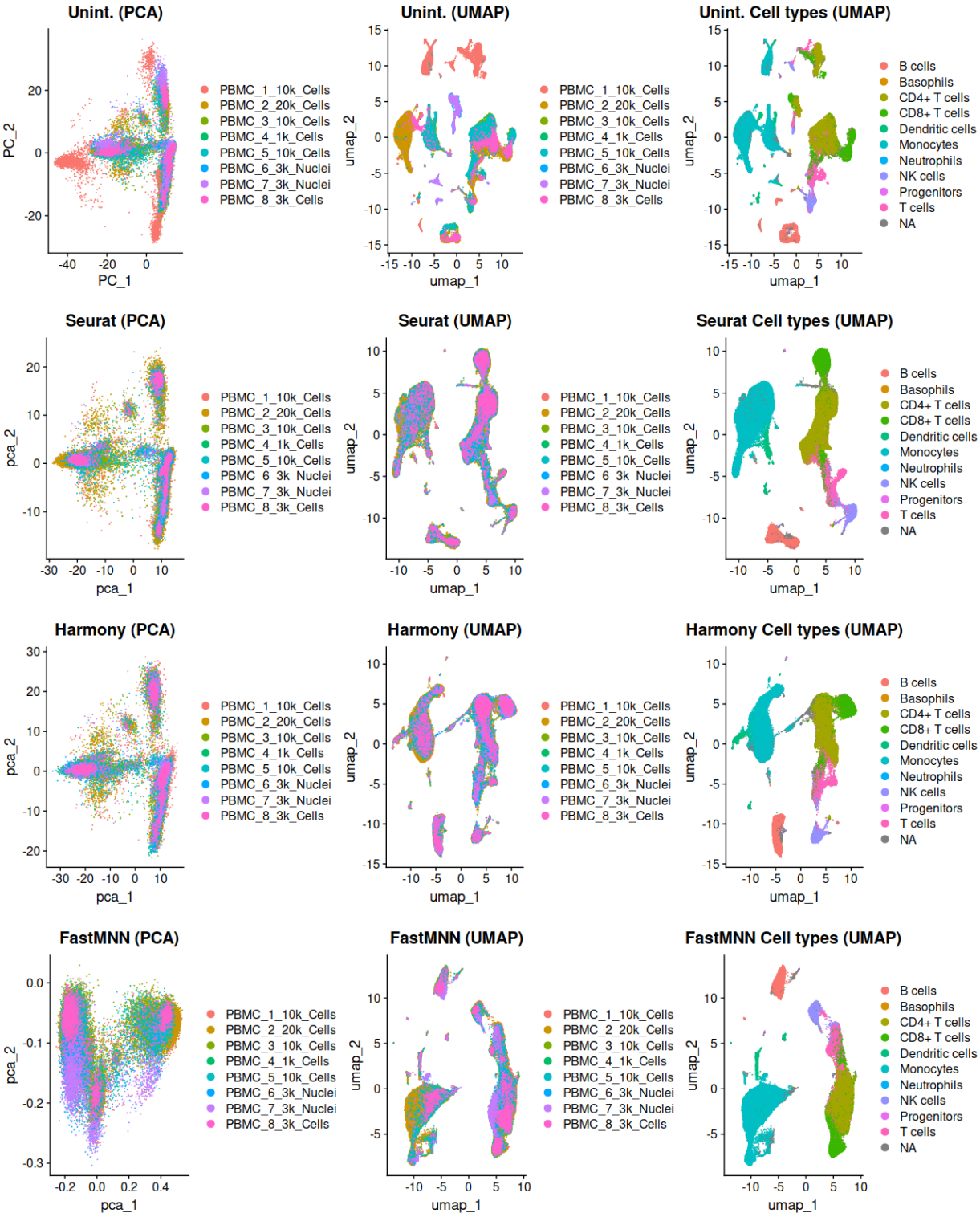
Figure 5 presents the results of integration analysis conducted on eight scRNA-seq dataset of PBMCs. The top row displays the structure of the unintegrated datasets. Then, each integration algorithm is depicted in subsequent rows, using PCA and UMAP dimensional reductions (only scVI does not use PCA and instead employs MDE (Minimum-distortion embedding) [79], an alternative method that offers a faster solution to conventional dimensional reduction techniques). The third column illustrates cell type annotation on UMAP, utilising the Single R package [80]. After performing integration analysis, AnnData objects were converted into Seurat Objects. A technical challenge arose with CanSig integration results, as conversion was not feasible, resulting in a visualisation through Python.

Integration outcome from Scanorama is subject to discussion, as it leaves parts of dataset PBMC_2_20k_Cells unaligned. Based on visual comparison, it is unclear whether this effect is beneficial or detrimental, yet it is evident that the different sizes of the batches plays a role in Scanorama's performance. On the other hand, the cell type annotation verifies that this particular subset of cells is classified as monocytes as well as the neighbourhood of unaligned cells from dataset 2.

The dataset PBMC_6_3k_Nuclei had removed granulocytes by FACS cell sorting, which was confirmed by evaluation of typical markers of granulocytes. The integration analysis showed that each integration algorithm dispersed cells in a way preserving this difference.

Despite variations among the datasets, including differences in the number of sequenced cells, the sequencing method employed, cell sorting techniques, and whether whole cells or only nuclei were sequenced, each algorithm adeptly handled batch effects and achieved optimal alignment across all datasets. These findings undoubtedly underscore the effectiveness and widespread adoption of these tools. All algorithms demonstrate their suitability for integrating high-quality data, such as scRNA-seq PBMC data.

4.7.2 Visual comparison of integration algorithms



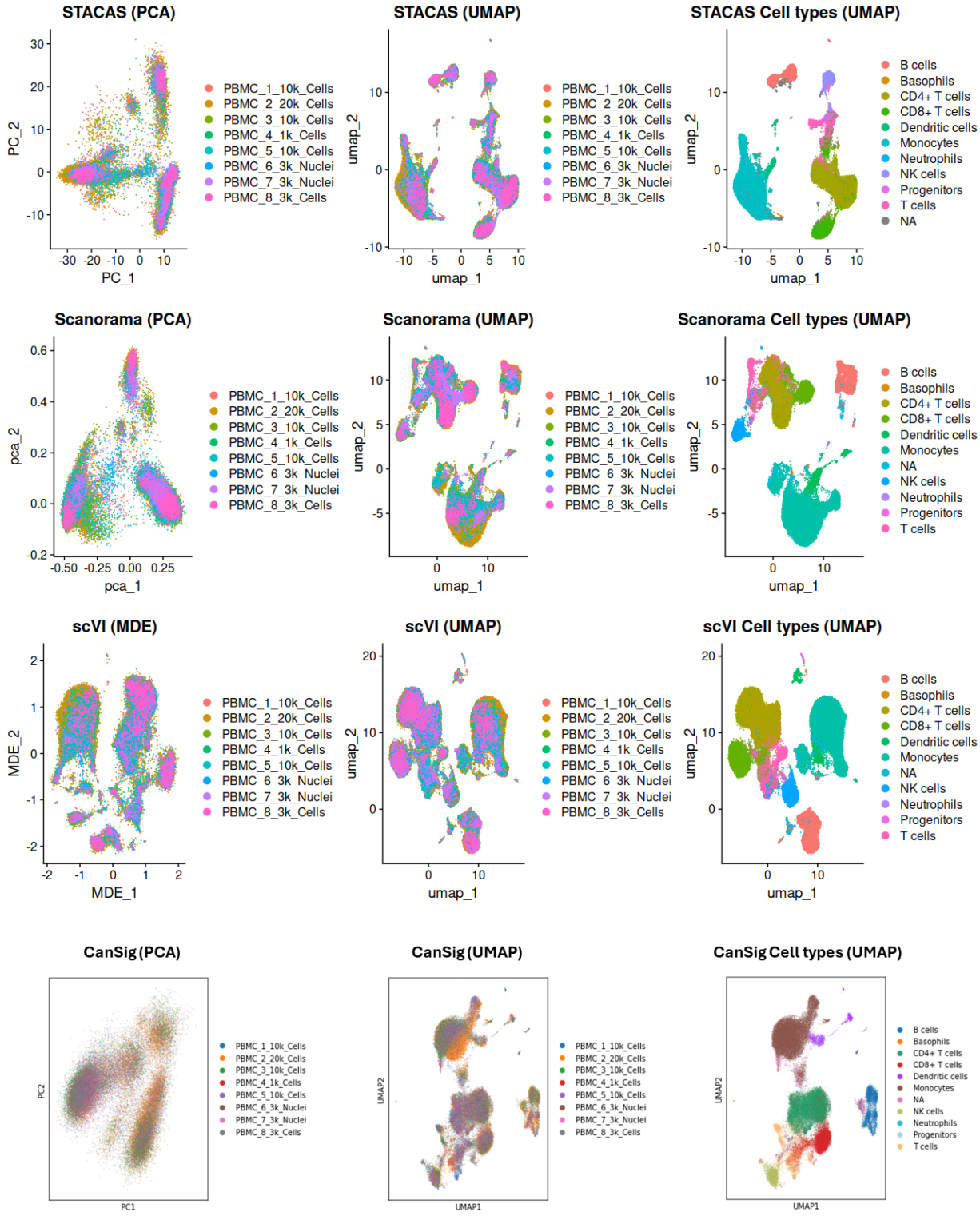


Figure 5: Evaluation of integration results of eight PBMCs scRNA-seq datasets. Each row represents an outcome from a one integration algorithm using PCA, UMAP visualisation (only scVI employs MDE instead of PCA). Integration algorithms are arranged in order: Seurat v5.0.0, Harmony v1.1.1, FastMNN v1.12.3, STACAS v1.12.3, Scanorama v2.2.2, scVI v0.20.3, CanSig v0.3.1. Cell type annotation was made by SingleR.

5 Integration of VSs scRNA-seq datasets

5.1 Samples

We obtained four datasets (local ID: VS_39, VS_40, VS_58, VS_59) derived from VSs scRNA-seq from the Laboratory of genomics and bioinformatics from Institute of Molecular Genetics of the Czech Academy of Sciences. These datasets exhibit high variations among themselves, making them ideally suited for benchmarking integration algorithms.

All samples were prepared from freshly collected tumours by enzymatic digestion of the tissues and sorted on FACS to remove nonviable cells and extracellular matrix debris. The samples VS_58 and VS_59 were also depleted of $CD45^{-}/NCAM1^{-}$ cells, retaining immune and neuron cells.

5.2 VSs scRNA-seq data

Raw data stored in FASTQ formats must undergo preprocessing using Cell Ranger to obtain gene expression matrices. One notable variation was seen in the sample VS_59 barcode rank plot (Figure 6), generated by plotting the number of UMIs associated with the barcodes. In these plots, blue colour represents captures where the number of detected UMIs is high enough to consider them as cells. The first three samples exhibit a typical "cliff and knee" structure. This structure is expected in high quality samples, where the barcodes linked to cells are more distinctly differentiated from empty droplets. However, the fourth sample, VS_59, displays a round curve, indicating compromised sample quality. This round curve suggests low sample viability during preparation, leading to a less single-cell behaviour [52].

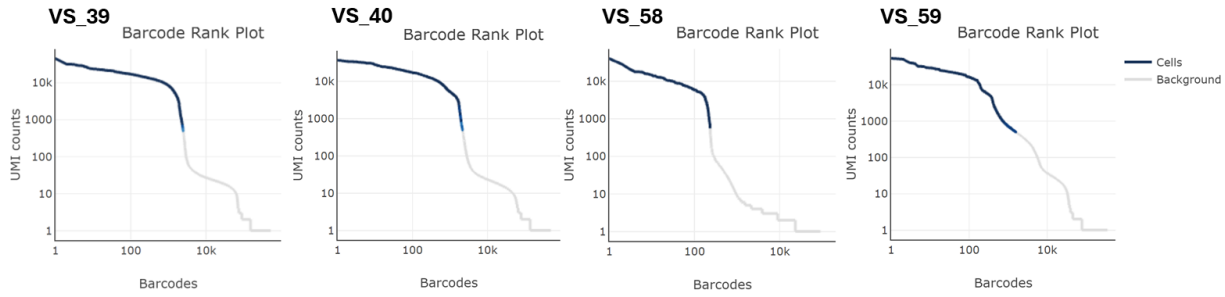


Figure 6: Bacode rank plots of VS_39, VS_40, VS_58, VS_59, showing the number of assigned reads to specific barcode. Barcodes are plotted against decreased number of UMIs linked with particular barcode. The blue colour indicates barcodes with a reasonable count of associated UMIs to signify cells.

Using Cell Ranger v6, datasets were aligned to the reference genome, resulting in raw feature matrices saved in the Matrix Market Exchange format (.mtx). These formats are divided into three files: Barcodes.tsv.gz and Features.tsv.gz containing cell identifiers, and Matrix.mtx.gz containing expression levels of cell features. The .mtx efficiently stores sparse matrices, common for scRNA-seq data.

Each dataset was preprocessed separately using Seurat to preserve relevant biological insights. Our strategy was to eliminate low quality cells, that bring technical noise rather than biological meaning. The selected thresholds are in the Table 3 where the last two columns represent the number of retained cells and the number of genes.

	Original data		QC Thresholds			Filtered data	
Dataset local ID	Genes	Cells	Genes per cell	% mito	% ribo	Genes	Cells
VS_39	14 786	2 204	>200	<55	>5	14 466	1 971
VS_40	14 390	2 140	>200	<60	>5	14 039	1 853
VS_58	13 110	197	>200	<40	>5	12 255	189
VS_59	18 014	3 033	>200	<40	>5	17 070	2 966

Table 3: Table of quality control thresholds for scRNA-seq data derived from four vestibular schwannomas samples.

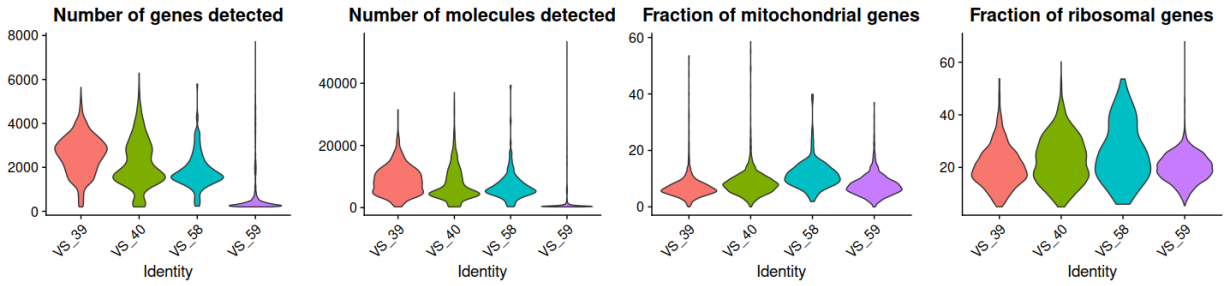


Figure 7: Violin plot showing the distribution of the cells on the preprocessed scRNA-seq data from VSs. Despite the VS_59 dataset having a substantial quantity of captured cells, its quality is compromised, exhibiting a notably low assigned genes and molecules per cell.

Several differences are apparent among the datasets illustrated in Table 3 and Figure 7. Notably, dataset VS_58 is relatively small with 189 cells after quality control and filtering, and it exhibits the highest fraction of mitochondrial transcripts, indicating potential cellular stress. The table also highlights VS_59 with high number of detected genes, but has a lower average gene count per cell (Figure 7), consistent with Cell Ranger results (Figure 6), suggesting poor input sample quality only for VS_59.

With the processing steps the data undergoes, it acquires several technical variances. Hence before starting with integration analysis, we should also take a look into biological variances within the datasets. The biological variances can be attributed to the unique characteristics of the tissue from which they are derived. VSs, situated in the peripheral nervous system, present challenges in preparation of single-cell dilutions due to tissue complexity. It is clear from the morphology of Schwann cells that the success rate of obtaining a separate viable Schwann cell is very low, leading to predominantly TME cells in our datasets (Figure 8). However, as discussed earlier, the TME plays a crucial role in the future growth characteristics of a tumour, highlighting the importance of molecular-level analysis of TME.

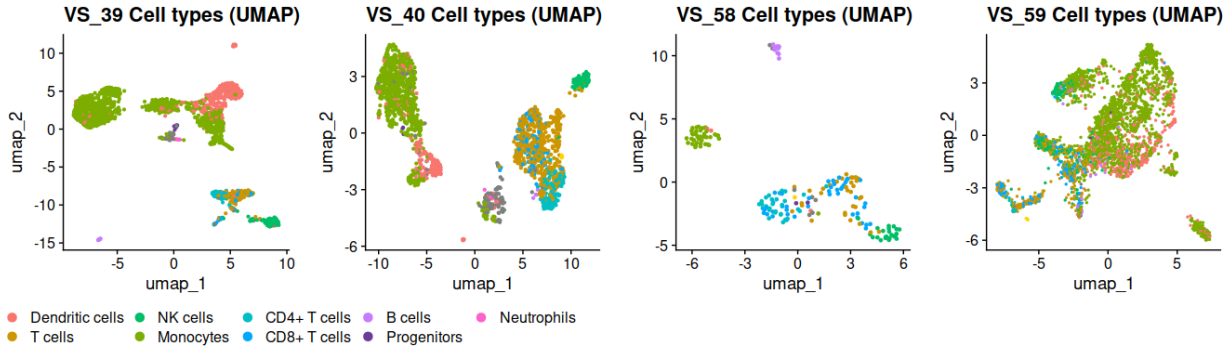


Figure 8: Structure of datasets before integration analysis. Four dimensional reduction UMAP plots represent the expression values of each cell. The datasets exhibit considerable heterogeneity, with variations in the representation of cell types. The same legend applies to all four plots.

Since we were not able to capture either Schwann cells or M2-type macrophages using SingleR cell annotation (Figure 8), we decided to analyse marker genes in individual datasets using dot plot (Figure 9) and feature plot (Figure 10), to depict their average expression. Moreover, analysis revealed cell heterogeneity after FACS sorting, suggesting effective preservation of immune cells in the TME. Viable neuron cells were possibly forming a rare subpopulation prior to FACS sorting, leading to noise confusion and subsequent removal from samples [81].

As discussed in Chapter 2.2, key factors driving tumour growth are TME elements. Marker genes important for myeloid cell proliferation and infiltration into the TME include CD163 and MS4A4 (associated with M2-type TAMs), and CSF1R (myeloid cell receptor for VSs ligands). Furthermore, CD68 expression indicates M1-type macrophages [33]. Schwann cell presence is suggested by EGR2, with increased ERB2 expression noted in Schwann cells undergoing changes due to nerve damage and need of myeloid formation [82]. However, the common VSs marker, S100B [83], was not significantly detected in any dataset.

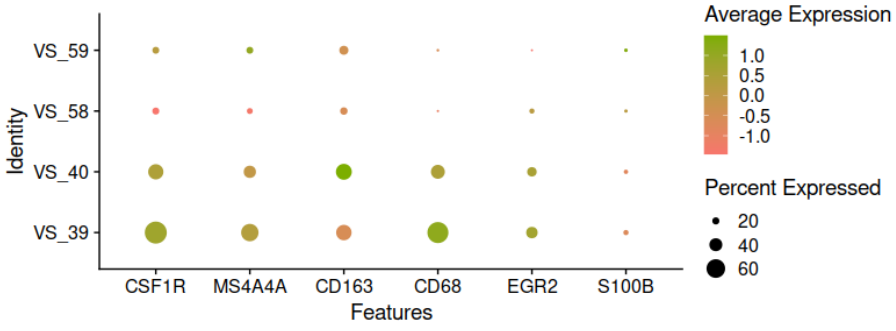


Figure 9: The dot plot illustrates the mean expression of selected marker genes identified in VSs scRNA-seq. CDF1R, MS4A4A, CD163 are indicative of myeloid cells that contribute to tumour progression. CD68 is a marker for M1-type macrophages. EGR2 is markers for myelinating Schwann cells. S100B is expressed in VSs.

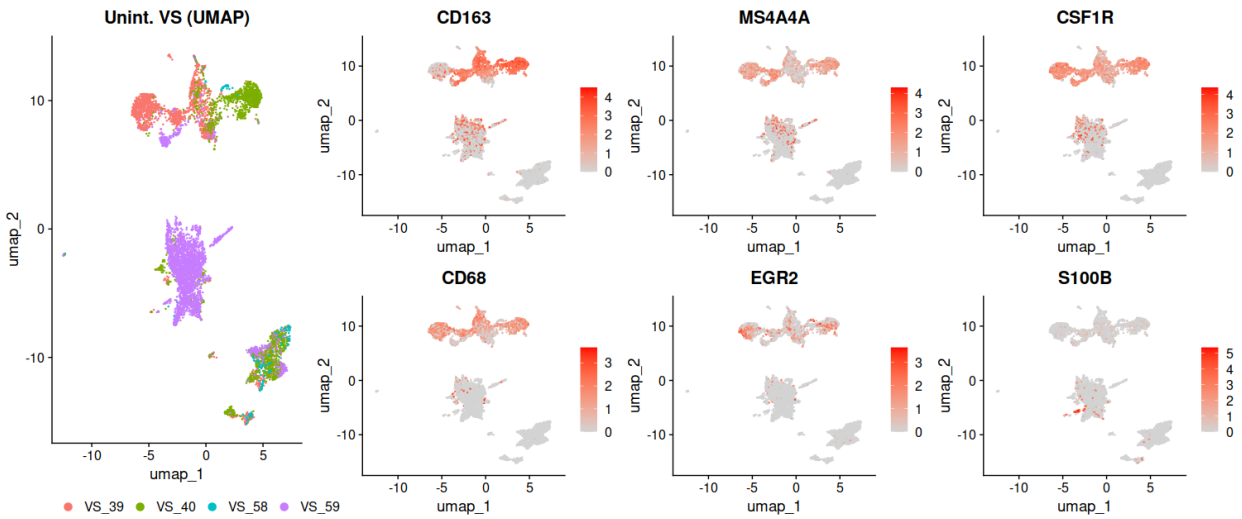


Figure 10: The feature plots illustrate the expression of selected marker genes identified in TME and VSs on dimensional reduction UMAP plot of unintegrated datasets. Most of the VS_59 dataset creates a large cluster of cells that do not show a TME or VSs transcriptional profile.

Upon assessing the outcomes from Figures 9 and 10, it's necessary to mention that VS_58 and VS_59 aren't particularly effective for making definitive conclusions. The extremely low expression levels of our marker genes of interest suggested an inconclusive result. On the other hand, the datasets VS_39 and VS_40 exhibit encouraging outcomes from the initial data processing and analysis, encouraging further exploration of TME.

Datasets exhibiting such high biological and technical variability could present challenges for integration algorithms, thus making them perfect for comparative study of statistical models developed for integration of scRNA-seq data.

5.2.1 Algorithms

The eight algorithms described in Chapter 4 were tested on the preprocessed sample from vestibular schwannomas single-cell RNA sequencing. All commands were run with default parameters.

Evaluation of runtime and maximum memory consumption All the analyses was performed using in-house server of IMG CAS named ORCA, equiped with Intel(R) Xeon(R) CPU E7-8890 v4 @ 2.20GHz processor (24 CPU cores and 250.8 GB total memory).

Figure 11 presents a comparison of system CPU runtime (s) and peak memory usage (MB) of integration of PBMCs and VSs scRNA-seq datasets. Following the combining of PBMCs, the resulting dataset comprises 59 335 cells (red), while the VSs dataset includes 6 979 cells (blue). According to the measurements, Harmony exhibits the minimal memory usage and runtime in both scenarious. The STACAS algorithm allocates anchors to a distinct object, resulting in the greatest memory consumption. ScVI and CanSig, being deep generative models, require extended runtime relative to other approaches.

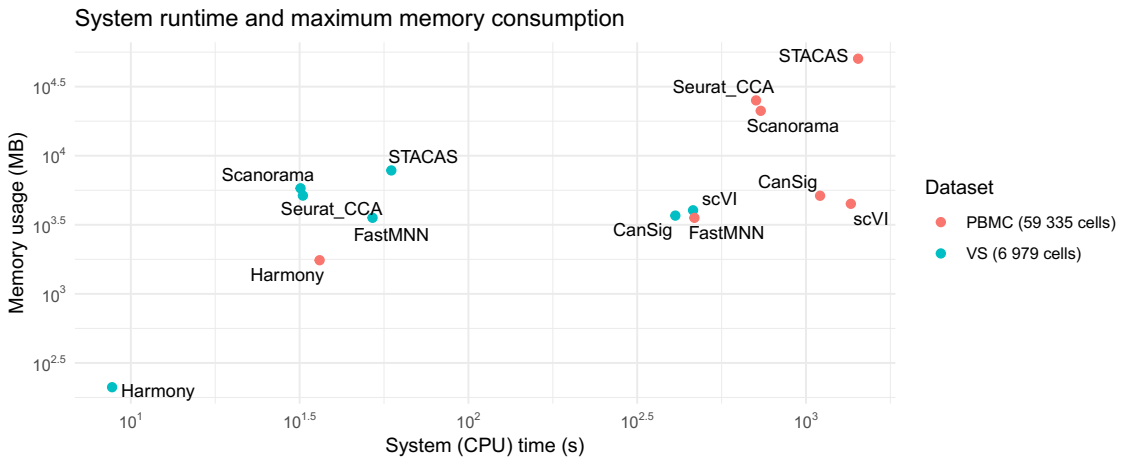


Figure 11: Benchmark of the system CPU runtime and maximum memory usage of integration algorithms. Blue corresponds to the integration of VSs scRNA-seq data comprising 6 979 cells. Red represents the integration of PBMC scRNA-seq involving 59 335 cells in the combined dataset. Harmony achieves the quickest integration time and the minimal memory usage. Scanorama, Seurat CCA, and FastMNN display similar performance. STACAS records the greatest memory consumption. ScVI and CanSig are observed to have the longest runtime.

5.2.2 Comparison of integration algorithms on four VSs datasets

Integration of diverse data, such as scRNA-seq samples from VSs, has provided valuable insights into comparison algorithms. Our primary focus was on ensuring that integration do not overestimate the differences between datasets but rather preserve the true disparities (Figure 13).

Seurat, employing its canonical correlation analysis method, is among the most widely used approaches. However, our data suggests that it tends to overestimate the dataset differences. Our findings agree with other comparative studies [75], indicating a substantial risk of over-integration and misinterpretation of results with the Seurat algorithm. However, the results are promising.

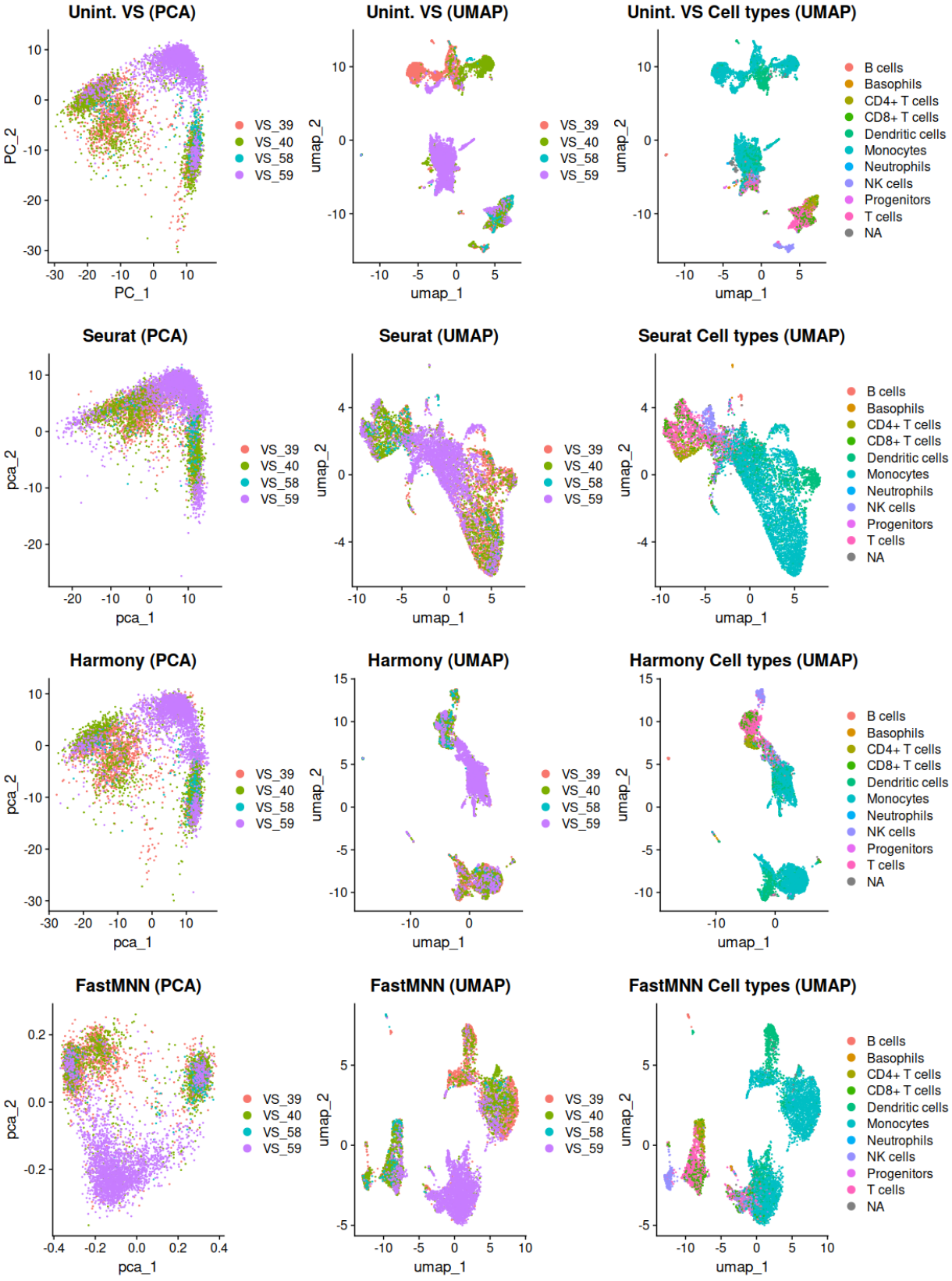
It can be seen in our, but also in other comparative studies, that Scanorama's integration algorithm might be influenced by batches [84]. It requires separate datasets labelled with batch IDs and utilises this information during alignment [7]. While Scanorama under-performed with our datasets, we anticipate better performance in a more heterogeneous samples.

The common use of the scVI algorithm leads to comparable results of scVI and CanSig and distinguishes it from other methods. ScVI reveals distinct clusters of cells originating from single dataset, whereas CanSig shows a tendency to heavily mix datasets. These findings suggest that neither of these approaches may be the most optimal choice for our specific datasets.

In terms of system runtime and memory usage, Harmony emerged as the top performer. Its straightforward and intuitive approach removes batch effects without excessive strain on the GPU. Harmony shows promise for initial analysis, but more robust integration algorithms are needed to complement its capabilities.

For our scRNA-seq VS datasets, FastMNN proved to be an optimal choice. It effectively preserved biologically relevant differences between datasets without forcibly merging individual cell clusters, thereby revealing differences in cellular heterogeneity. Similarly to FastMNN, STACAS emerged as the optimal method for batch effect removal in our data, preserving true differences and aligning with our expectations.

5.2.3 Visual comparison of integration algorithms



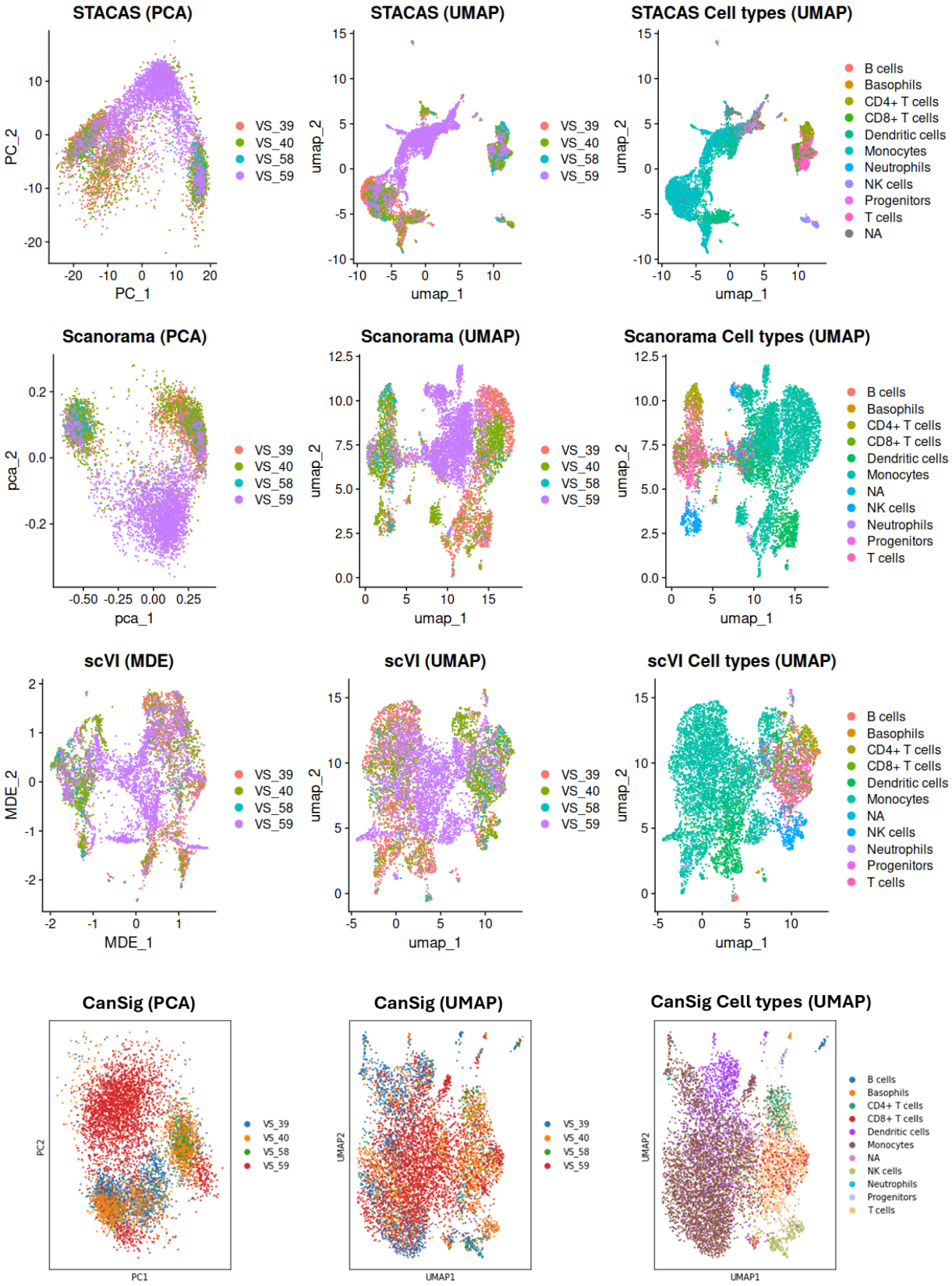


Figure 13: Evaluation of integration results of four VS scRNA-seq datasets. Each row represents an outcome from a one integration algorithm using PCA, UMAP visualisation (only scVI employs MDE instead of PCA). Integration algorithms are arranged in order: Seurat v5.0.0, Harmony v1.1.1, FastMNN v1.12.3, STACAS v1.12.3, Scanorama v2.2.2, scVI v0.20.3, CanSig v0.3.1. Cell type annotation was made by SingleR.

6 Discussion

The integration of peripheral blood mononuclear cell and vestibular schwannomas single-cell RNA sequencing data unquestionably underscores the capabilities of each compared method for integrating transcriptomics data. Every statistical approach successfully eliminates unwanted batch effects while preserving the true differences between datasets. Factors such as missing cell subpopulations, sample heterogeneity, preservation methods, or the number of sequenced cells do not impact the results.

In two out of four samples from the scRNA-seq of vestibular schwannomas, we identified marker genes representing the myeloid lineage of immune cells within the tumour microenvironment. Our findings mirror those of previous studies [38], which reported the upregulation of CD163 in fast-growing tumours, suggesting the potential involvement of M2 tumour associates macrophages in tumour progression. Additionally, we also observed the expression of MS4A4A, which has been linked to NK cell-mediated resistance to metastasis [39].

Our comparative study of integration algorithms yielded results similar to other benchmarking studies [78]. We confirm that Harmony is suitable for initial analysis but needs to be complemented with other methods. Consistent with previous research, we found a balance between dataset mixing and biological conservation in the scVI integration result [76]. Our observations align with statements made in a recent cancer study [85], where authors compared the integration of several methods on tumour-associated samples and concluded that FastMNN and STACAS were the best performers.

Integrating the aforementioned scRNA-seq datasets from vestibular schwannomas with samples from healthy tissue may reveal intriguing results, shedding light on different transcriptional profiles in the tumour microenvironment. The use of batch correction techniques may uncover therapeutically significant cell types and states within vestibular schwannomas. We foresee the future continuation of this comparative study.

7 Conclusion

This thesis undertook a comprehensive comparison of seven statistical approaches for the batch effects removal and the integration of single-cell RNA sequencing datasets. We assessed the performance of Seurat, Harmony, FastMNN, STACAS, Scanorama, scVI, and CanSig, specifically designed for handling transcriptomic data, in two scenarious. First scenario encompassed integration of scRNA-seq of peripheral blood mononuclear cells from 10× dataset repositories. These datasets are widely used in comparative studies due their good annotation and high sample quality. In the second scenario we introduced our own datasets of scRNA-seq derived from vestibular schwannomas of individual patients, which exhibit large variations and poses challenge for integration methods. Before integration we summarised the molecular patterns in vestibular schwannomas and analysed the scRNA-seq data. The findings from the summary were taken into account in the subsequent analysis.

The comparison analysis began with integration of eight distinct PBMCs scRNA-seq datasets, which varied in several key aspects including the number of sequenced cells, the preservation method utilised, and whether the sequencing was conducted on whole cells or isolated nuclei. Results revealed that each method integrated all datasets and effectively eliminate batch effects while preserving biological variability.

Next we evaluated the computational capabilities of these methods by integrating datasets characterised by substantial technical and biological variances. Preprocessing of the scRNA-seq datasets derived from VSs highlighted significant differences between batches. Through the analysis of reported marker genes associated with VSs and components of the TME, we meticulously examined and assessed each dataset individually. Even when considering samples as technological and biological replicates, disparities in sample viability, capture efficiency, and library preparation were observed. However, such variations between scRNA-seq results are not unusual, and there is a need for modern computational methods to be built to work with such data.

The capabilities of each integration algorithm on VS datasets revealed that FastMNN yielded the most promising results. However, it cannot be definitively stated that this method is the best fit for all samples, as the structure of the data must always be taken into account.

We anticipate further enhancements in each method compared, given the crucial role of data integration in future scRNA-seq data analysis. Integration facilitates a systematic comparison of samples under varying conditions, thereby enabling a comprehensive understanding of the functional roles of different cell types and states. This approach is applicable across a wide range of research fields, including developmental biology, immunology, and cancer research. We believe that our review has aided in understanding the fundamentals of scRNA-seq data analysis and has contributed to the comparison of integration algorithms for future applications.

References

[1] F. Tang, C. Barbacioru, Y. Wang, *et al.*, "Mrna-seq whole-transcriptome analysis of a single cell," *Nature methods*, vol. 6, no. 5, pp. 377–382, 2009.

[2] M. N. Bainbridge, R. L. Warren, M. Hirst, *et al.*, "Analysis of the prostate cancer cell line lncap transcriptome using a sequencing-by-synthesis approach," *BMC genomics*, vol. 7, pp. 1–11, 2006.

[3] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature biotechnology*, vol. 36, no. 5, pp. 411–420, 2018.

[4] I. Korsunsky, N. Millard, J. Fan, *et al.*, "Fast, sensitive and accurate integration of single-cell data with harmony," *Nature methods*, vol. 16, no. 12, pp. 1289–1296, 2019.

[5] L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni, "Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors," *Nature biotechnology*, vol. 36, no. 5, pp. 421–427, 2018.

[6] M. Andreatta and S. J. Carmona, "Stacas: Sub-type anchor correction for alignment in seurat to integrate single-cell rna-seq data," *Bioinformatics*, vol. 37, no. 6, pp. 882–884, 2021.

[7] B. Hie, B. Bryson, and B. Berger, "Efficient integration of heterogeneous single-cell transcriptomes using scanorama," *Nature biotechnology*, vol. 37, no. 6, pp. 685–691, 2019.

[8] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature methods*, vol. 15, no. 12, pp. 1053–1058, 2018.

[9] J. Yates, F. Barkmann, P. Czyz, *et al.*, "Cansig: Discovery of shared transcriptional states across cancer patients from single-cell rna sequencing data," *bioRxiv*, pp. 2022–04, 2022.

[10] N. Koen, C. Shapiro, E. D. Kozin, *et al.*, "Location of small intracanalicular vestibular schwannomas based on magnetic resonance imaging," *Otolaryngology–Head and Neck Surgery*, vol. 162, no. 2, pp. 211–214, 2020.

[11] F. Constanzo, B. C. d. A. Teixeira, P. Sens, D. Escuissato, and R. Ramina, "Cerebellopontine angle schwannomas arising from the intermediate nerve: A scoping review," *Neurosurgical Review*, vol. 43, no. 6, pp. 1431–1441, 2020.

[12] B. De Foer, C. Kenis, D. Van Melkebeke, *et al.*, "Pathology of the vestibulocochlear nerve," *European journal of radiology*, vol. 74, no. 2, pp. 349–358, 2010.

[13] C. Matthies and M. Samii, "Management of 1000 vestibular schwannomas (acoustic neuromas): Clinical presentation," *Neurosurgery*, vol. 40, no. 1, pp. 1–10, 1997.

[14] M. L. Carlson, Ø. V. Tveiten, C. L. Driscoll, *et al.*, "Long-term dizziness handicap in patients with vestibular schwannoma: A multicenter cross-sectional study," *Otolaryngology–Head and Neck Surgery*, vol. 151, no. 6, pp. 1028–1037, 2014.

[15] A. D. Sweeney, M. L. Carlson, N. T. Shepard, *et al.*, "Congress of neurological surgeons systematic review and evidence-based guidelines on otologic and audiologic screening for patients with vestibular schwannomas," *Neurosurgery*, vol. 82, no. 2, E29–E31, 2018.

[16] D. G. R. Evans, S. Huson, D. Donnai, *et al.*, "A clinical study of type 2 neurofibromatosis," *QJM: An International Journal of Medicine*, vol. 84, no. 1, pp. 603–618, 1992.

[17] K. Hadfield, M. Smith, J. Urquhart, *et al.*, "Rates of loss of heterozygosity and mitotic recombination in nf2 schwannomas, sporadic vestibular schwannomas and schwannomatosis schwannomas," *Oncogene*, vol. 29, no. 47, pp. 6216–6221, 2010.

[18] D. G. R. Evans, A. Moran, A. King, S. Saeed, N. Gurusinghe, and R. Ramsden, "Incidence of vestibular schwannoma and neurofibromatosis 2 in the north west of england over a 10-year period: Higher incidence than previously thought," *Otology & neurotology*, vol. 26, no. 1, pp. 93–97, 2005.

[19] A. Herwadker, E. A. Vokurka, D. G. R. Evans, R. T. Ramsden, and A. Jackson, "Size and growth rate of sporadic vestibular schwannoma: Predictive value of information available at presentation," *Otology & Neurotology*, vol. 26, no. 1, pp. 86–92, 2005.

[20] R. Goldbrunner, M. Weller, J. Regis, *et al.*, "Eano guideline on the diagnosis and treatment of vestibular schwannoma," *Neuro-oncology*, vol. 22, no. 1, pp. 31–45, 2020.

[21] J. A. Trofatter, M. M. MacCollin, J. L. Rutter, *et al.*, "A novel moesin-, ezrin-, radixin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor," *Cell*, vol. 72, no. 5, pp. 791–800, 1993.

[22] A. B. Bianchi, T. Hara, V. Ramesh, *et al.*, "Mutations in transcript isoforms of the neurofibromatosis 2 gene in multiple human tumour types," *Nature genetics*, vol. 6, no. 2, pp. 185–192, 1994.

[23] M. J. Pykett, M. Murphy, P. R. Harnish, and D. L. George, "The neurofibromatosis 2 (nf2) tumor suppressor gene encodes multiple alternatively spliced transcripts," *Human molecular genetics*, vol. 3, no. 4, pp. 559–564, 1994.

[24] T. Shimizu, A. Seto, N. Maita, *et al.*, "Structural basis for neurofibromatosis type 2: Crystal structure of the merlin ferm domain," *Journal of Biological Chemistry*, vol. 277, no. 12, pp. 10 332–10 336, 2002.

[25] A. M. Petrilli and C. Fernández-Valle, "Role of merlin/nf2 inactivation in tumor biology," *Oncogene*, vol. 35, no. 5, pp. 537–548, 2016.

[26] I. Sher, C. O. Hanemann, P. A. Karplus, and A. Bretscher, "The tumor suppressor merlin controls growth in its open state, and phosphorylation converts it to a less-active more-closed state," *Developmental cell*, vol. 22, no. 4, pp. 703–705, 2012.

[27] H. Morrison, L. S. Sherman, J. Legg, *et al.*, "The nf2 tumor suppressor gene product, merlin, mediates contact inhibition of growth through interactions with cd44," *Genes & development*, vol. 15, no. 8, pp. 968–980, 2001.

[28] Y. Cui, S. Groth, S. Troutman, *et al.*, "The nf2 tumor suppressor merlin interacts with ras and rasgap, which may modulate ras signaling," *Oncogene*, vol. 38, no. 36, pp. 6370–6381, 2019.

[29] K. J. Blair, A. Kiang, J. Wang-Rodriguez, M. A. Yu, J. K. Doherty, and W. M. Ongkeko, "Egf and bfgf promote invasion that is modulated by pi3/akt kinase and erk in vestibular schwannoma," *Otology & Neurotology*, vol. 32, no. 2, pp. 308–314, 2011.

[30] K. Kaempchen, K. Mielke, T. Utermark, S. Langmesser, and C. O. Hanemann, "Upregulation of the rac1/jnk signaling pathway in primary human schwannoma cells," *Human molecular genetics*, vol. 12, no. 11, pp. 1211–1221, 2003.

[31] D. Pan, "The hippo signaling pathway in development and cancer," *Developmental cell*, vol. 19, no. 4, pp. 491–505, 2010.

[32] H. F. Dvorak, "Tumors: Wounds that do not heal," *New England Journal of Medicine*, vol. 315, no. 26, pp. 1650–1659, 1986.

[33] T. F. Barrett, B. Patel, S. M. Khan, *et al.*, "Single-cell multi-omic analysis of the vestibular schwannoma ecosystem uncovers a nerve injury-like state," *Nature communications*, vol. 15, no. 1, p. 478, 2024.

[34] C. J. Hannan, D. Lewis, C. O'leary, *et al.*, "The inflammatory microenvironment in vestibular schwannoma," *Neuro-Oncology Advances*, vol. 2, no. 1, vdaa023, 2020.

[35] M. de Vries, I. Briaire-de Bruijn, M. J. Malessy, S. F. de Brune, A. G. van der Mey, and P. C. Hogendoorn, "Tumor-associated macrophages are related to volumetric growth of vestibular schwannomas," *Otology & Neurotology*, vol. 34, no. 2, pp. 347–352, 2013.

[36] A. Etzerodt and S. K. Moestrup, "Cd163 and inflammation: Biological, diagnostic, and therapeutic aspects," *Antioxidants & redox signaling*, vol. 18, no. 17, pp. 2352–2363, 2013.

[37] S. K. Biswas and A. Mantovani, "Macrophage plasticity and interaction with lymphocyte subsets: Cancer as a paradigm," *Nature immunology*, vol. 11, no. 10, pp. 889–896, 2010.

[38] W. De Vries, I. Briaire-de Bruijn, P. Van Benthem, A. Van Der Mey, and P. Hogendoorn, "M-csf and il-34 expression as indicators for growth in sporadic vestibular schwannoma," *Virchows Archiv*, vol. 474, pp. 375–381, 2019.

[39] I. Mattiola, F. Tomay, M. De Pizzol, *et al.*, "The macrophage tetraspan ms4a4a enhances dectin-1-dependent nk cell–mediated resistance to metastasis," *Nature immunology*, vol. 20, no. 8, pp. 1012–1022, 2019.

[40] M. D. Luecken and F. J. Theis, "Current best practices in single-cell rna-seq analysis: A tutorial," *Molecular systems biology*, vol. 15, no. 6, e8746, 2019.

[41] J. Bageritz, N. Krausse, S. Yousefian, S. Leible, E. Valentini, and M. Boutros, "Glyoxal as an alternative fixative for single-cell rna sequencing," *G3: Genes, Genomes, Genetics*, vol. 13, no. 10, jkad160, 2023.

[42] C. T. Wohnhaas, G. G. Leparc, F. Fernandez-Albert, *et al.*, "Dmso cryopreservation is the method of choice to preserve cells for droplet-based single-cell rna sequencing," *Scientific reports*, vol. 9, no. 1, p. 10 699, 2019.

[43] N. Habib, I. Avraham-Davidi, A. Basu, *et al.*, "Massively parallel single-nucleus rna-seq with dronc-seq," *Nature methods*, vol. 14, no. 10, pp. 955–958, 2017.

[44] J. Ding, X. Adiconis, S. K. Simmons, *et al.*, "Systematic comparison of single-cell and single-nucleus rna-sequencing methods," *Nature biotechnology*, vol. 38, no. 6, pp. 737–746, 2020.

[45] S. C. van den Brink, F. Sage, Á. Vértesy, *et al.*, "Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations," *Nature methods*, vol. 14, no. 10, pp. 935–936, 2017.

[46] W. Bonner, H. Hulett, R. Sweet, and L. Herzenberg, "Fluorescence activated cell sorting," *Review of Scientific Instruments*, vol. 43, no. 3, pp. 404–409, 1972.

[47] L. A. Herzenberg, R. G. Sweet, and L. A. Herzenberg, "Fluorescence-activated cell sorting," *Scientific American*, vol. 234, no. 3, pp. 108–118, 1976.

[48] V. Svensson, R. Vento-Tormo, and S. A. Teichmann, "Exponential scaling of single-cell rna-seq in the past decade," *Nature protocols*, vol. 13, no. 4, pp. 599–604, 2018.

[49] T. Kivioja, A. Vähärautio, K. Karlsson, *et al.*, "Counting absolute numbers of molecules using unique molecular identifiers," *Nature methods*, vol. 9, no. 1, pp. 72–74, 2012.

[50] E. Z. Macosko, A. Basu, R. Satija, *et al.*, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015.

[51] S. Islam, A. Zeisel, S. Joost, *et al.*, "Quantitative single-cell rna-seq with unique molecular identifiers," *Nature methods*, vol. 11, no. 2, pp. 163–166, 2014.

[52] G. X. Zheng, J. M. Terry, P. Belgrader, *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature communications*, vol. 8, no. 1, p. 14 049, 2017.

[53] T. Hubbard, D. Barker, E. Birney, *et al.*, "The ensembl genome database project," *Nucleic acids research*, vol. 30, no. 1, pp. 38–41, 2002.

[54] Z. Zhang, F. Cui, C. Wang, L. Zhao, and Q. Zou, "Goals and approaches for each processing step for single-cell rna sequencing data," *Briefings in Bioinformatics*, vol. 22, no. 4, bbaa314, 2021.

[55] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: Large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, pp. 1–5, 2018.

[56] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of rna-seq data," *Genome biology*, vol. 11, pp. 1–9, 2010.

[57] D. Ramsköld, E. T. Wang, C. B. Burge, and R. Sandberg, "An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data," *PLoS computational biology*, vol. 5, no. 12, e1000598, 2009.

[58] D. Ramsköld, E. T. Wang, C. B. Burge, and R. Sandberg, "An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data," *PLoS computational biology*, vol. 5, no. 12, e1000598, 2009.

[59] P. Brennecke, S. Anders, J. K. Kim, *et al.*, "Accounting for technical noise in single-cell rna-seq experiments," *Nature methods*, vol. 10, no. 11, pp. 1093–1095, 2013.

[60] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[61] G. Heimberg, R. Bhatnagar, H. El-Samad, and M. Thomson, "Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing," *Cell systems*, vol. 2, no. 4, pp. 239–250, 2016.

[62] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[63] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[64] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[65] E. Becht, L. McInnes, J. Healy, *et al.*, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[66] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol. 1, 1967, pp. 281–297.

[67] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, P10008, 2008.

[68] C. Feng, S. Liu, H. Zhang, *et al.*, "Dimension reduction and clustering models for single-cell rna sequencing data: A comparative study," *International journal of molecular sciences*, vol. 21, no. 6, p. 2181, 2020.

[69] V. A. Traag, L. Waltman, and N. J. Van Eck, "From louvain to leiden: Guaranteeing well-connected communities," *Scientific reports*, vol. 9, no. 1, p. 5233, 2019.

[70] Y. Hao, T. Stuart, M. H. Kowalski, *et al.*, "Dictionary learning for integrative, multimodal and scalable single-cell analysis," *Nature Biotechnology*, 2023. DOI: 10.1038/s41587-023-01767-y. [Online]. Available: https://doi.org/10.1038/s41587-023-01767-y.

[71] Y. Hao, S. Hao, E. Andersen-Nissen, *et al.*, "Integrated analysis of multimodal single-cell data," *Cell*, 2021. DOI: 10.1016/j.cell.2021.04.048. [Online]. Available: https://doi.org/10.1016/j.cell.2021.04.048.

[72] R. Satija, P. Hoffman, Y. Hao, *et al.*, *Seuratobject: Data structures for single cell data*, R package version 5.0.0, 2023. [Online]. Available: https://CRAN.R-project.org/package=SeuratObject.

[73] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics: methodology and distribution*, Springer, 1992, pp. 162–190.

[74] H. T. N. Tran, K. S. Ang, M. Chevrier, *et al.*, "A benchmark of batch-effect correction methods for single-cell rna sequencing data," *Genome biology*, vol. 21, pp. 1–32, 2020.

[75] M. Andreatta, L. Hérault, P. Gueguen, D. Gfeller, A. J. Berenstein, and S. J. Carmona, "Semi-supervised integration of single-cell transcriptomics data," *Nature Communications*, vol. 15, no. 1, p. 872, 2024.

[76] Y. Song, Z. Miao, A. Brazma, and I. Papatheodorou, "Benchmarking strategies for cross-species integration of single-cell rna sequencing data," *Nature Communications*, vol. 14, no. 1, p. 6495, 2023.

[77] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[78] M. D. Luecken, M. Büttner, K. Chaichoompu, *et al.*, "Benchmarking atlas-level data integration in single-cell genomics," *Nature methods*, vol. 19, no. 1, pp. 41–50, 2022.

[79] A. Agrawal, A. Ali, S. Boyd, *et al.*, "Minimum-distortion embedding," *Foundations and Trends in Machine Learning*, vol. 14, no. 3, pp. 211–378, 2021.

[80] D. Aran, A. P. Looney, L. Liu, *et al.*, "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage," *Nat. Immunol.*, vol. 20, pp. 163–172, 2019. DOI: 10.1038/s41590-018-0276-y.

[81] B. A. Sutermaster and E. M. Darling, "Considerations for high-yield, high-throughput cell enrichment: Fluorescence versus magnetic sorting," *Scientific reports*, vol. 9, no. 1, p. 227, 2019.

[82] R. Srinivasan, G. Sun, S. Keles, *et al.*, "Genome-wide analysis of egr2/sox10 binding in myelinating peripheral nerve," *Nucleic acids research*, vol. 40, no. 14, pp. 6449–6460, 2012.

[83] G. E. Gregory, A. P. Jones, M. J. Haley, *et al.*, "The comparable tumour microenvironment in sporadic and nf2-related schwannomatosis vestibular schwannoma," *Brain Communications*, vol. 5, no. 4, fcad197, 2023.

[84] J. Li, C. Yu, L. Ma, J. Wang, and G. Guo, "Comparison of scanpy-based algorithms to remove the batch effect from single-cell rna-seq data," *Cell Regeneration*, vol. 9, pp. 1–8, 2020.

[85] L. M. Richards, M. Riverin, S. Mohanraj, *et al.*, "A comparison of data integration methods for single-cell rna sequencing of cancer samples," *bioRxiv*, pp. 2021–08, 2021.

Grammar correction resources

The grammar corrections were made using:

Writefull (Writefull, https://writefull.com/, Accessed 22.4.2024, 2022)

OpenAI (OpenAI, ChatGPT (version 3.5), https://openai.com/chatgpt, Accessed 22.4.2024, 2022)

Funding

This research was funded in part by Ministry of Education, Youth, and Sports of the Czech Republic under the project LX22NPO5102, which is financed by the European Union – Next Generation EU as part of the Czech Recovery Plan.