Framework for integrating scRNA-seq and scATAC-seq to reveal signatures and trajectories of immune cells

Research Proposal

September 2, 2022

Part A Applicant

Carlotta Schieler, 3737006
MSc Candidate Bioinformatics and Biocomplexity, Utrecht University

Daily Supervisors: Dr. Isabel Misteli Guerreiro and Pim Rullens, Hubrecht Institute
First Examiner: Dr. Jop Kind, Hubrecht Institute
Second Examiner: Dr. Onur Basak, UMC Utrecht

6725 words

Part B Scientific Proposal

1 Basic Details

1.1 Title

Framework for integrating scRNA-seq and scATAC-seq to reveal signatures and trajectories of immune cells

1.2 Abstract

Studying the dynamics and epigenetic signatures of immune cells is important to understand the differences in immune response of individuals. Vast amount of scRNA-seq and scATAC-seq data are available and computational tools exist to integrate these two data modalities. However, most computational tools are performing poorly when both measurements were done on parallel samples and lack keeping all the dynamics in the data after integration. Thus, to use the already existing data, we propose a framework designed for trajectory and comparison analysis. For integration we build on an already published neural network-based tool scDART and propose improvements in construction of the gene-activity matrix similar to MAESTRO to provide higher accuracy. In addition, we suggest implementation of further downstream analyses such as differential gene and accessibility analysis and gene set enrichment analysis specifically picked for immunological comparison studies.

1.3 Laymen Summary

In recent years much knowledge was gained in immunology due to new arising datatypes. These new measurements all provide vital information on their own, but even possess more informative value when they can be combined to build a holistic view. Combining or integrating of measurements is already a difficult computational task when they are measured at the same time.

However, currently most existing data was measured on similar samples, but not at the same time and in our case not in the same cell, but a similar cell in a parallel sample. This makes it difficult to combine the measurements together and we need to find reliable tools to find the similar cells from both measurements in order to combine the different measured properties for a complete picture. Computational tools are needed to tackle this problem and different tools have been developed, but all have their own advantages and disadvantages. There is not one single best tool, but the suitability of the tool depends on the biological question at hand.

In this proposal we focus on an immunological question. We want to reveal how the immune response differs between healthy and sick individuals after an event such as vaccination. To decide, which tool is the best one, we need to take the intrinsic values of the immune system into account. In particular, we want to keep the dynamic nature of the immune system in our measurements and not mask them by applying computational tools, who are not able to capture these dynamic and changing structures.

For this we look at already existing computational tools and propose how we can achieve better capturing of the dynamic structures by combining ideas from different tools together. Further, we propose a full framework of analysis steps to be done in order to answer the immunological questions at hand. By this we want to provide an addition to the existing tools in the bioinformatics community and provide a specialized tool for the vast amount of already existing data. Thus, we can take advantage of the data made available from researchers all around the world and hopefully provide new insights, which were not able to be captured before.

1.4 Keywords

Integration, neural networks, multi-model, immunology, single cell

2 Scientific Proposal

2.1 Research topic

Immunological studies rely on many different datatypes

Immunological studies try to elucidate the complex interplays present in our immune system. One main focus is how the immune response differs between individuals. In particular in current COVID-19 vaccine studies, it was seen that the immune response is often times significantly smaller in already immune compromised individuals such as cancer[1] or B-cell depleted patients[2]. Understanding how this immune response develops and how it interacts with the administration of the treatment medicine is important, so that optimal administration protocols can be implemented around the timing of the vaccine. In that way the immune response in risk patients can be maximized.

To understand how the immune response differs, recent studies have focused on using single-cell techniques. Some studies focus on combining transcriptomic analysis (scRNA-seq) and collecting T- or B-cell receptor sequencing (T-/ BCR-seq) from the same cells[3, 4, 5]. Other studies in the field of systems vaccinology focus on including epigenomic datatypes as well. The study by Wimmers et al.[6] did collect different data modalities from healthy subjects before and after receiving vaccination against influenza. By also collecting epigenomic data such as scATAC-seq, they revealed persisting epigenetic signatures and found persisting chromatin state changes even six months after vaccination. Another study compared healthy individuals against individuals, which recovered from SARS-CoV-2[7]. They found that in the innate as well as adaptive immune cells chromatin remodeling took place after a COVID-19 infection. The emerging chromatin profile in SARS-CoV-2-specific CD8+ T-cells shows promoting differentiation of effector or memory cells. This suggest that the immune memory is established by changes in the landscape of chromatin accessibility. Thus, to understand how our immune system adapts to pathogens and is able to be better prepared for the next invasion, it is vital to include epigenetic datatypes into the study.

Further, the immune system is also characterized by dynamic changes as immune cells maturate from the bone marrow to fully developed and specialized immune “fighters”. Therefore, studies have focused on understanding these trajectories[8]. Studying a human hematopoiesis scRNA-seq dataset, the differentiation and branching of human stem cells to megakaryocytic-erythroid progenitors and common myeloid progenitors was shown. However, full lineage analysis until full maturation of for example T-cells could not be tackled in this analysis as only bone marrow samples were included, and maturation of some immune cells occurs only after leaving the bone marrow[9].

The maturation of the immune cells such as Natural Killer (NK), CD8+ T- and T-regulatory (Treg) cells was shown in another study conducting a pseudo time cell trajectory analysis by using the clustering results of scRNA-seq analysis from the various cell types[10] . These trajectory analyses show also utility when predicting the outcome of potential therapeutic intervention such as seen in treating gastric cancer[11]. In this paper scRNA-seq was integrated with bulk RNA-seq from gastric cancer cells to find risk signatures of gastric cancer differentiation-related genes.

However, a trajectory analysis of the current datasets integrating not only analyzing the gene expression context but also integrating the chromatin accessibility landscape and thus immune memory is currently missing.

Looking at the computational tools currently available, it becomes clear that a tool is missing to use existing unmatched scRNA-seq and scATAC-seq datasets of immunological studies and study the trajectory of these immune cells.

Existing tools lack integration and trajectory analysis

There are existing tools to integrate scRNA-seq and scATAC-seq. Both are closely related datatypes and can be measured on the same platform such as 10X Genomics and thus share many steps of computational preparation and analysis such as removing non-biological variations by normalizing the count matrices, and a subsequent dimensionality reduction[12].

There are then different methods to integrate the data. One can use manifold alignment such as MATCHER[13], which shows good results while the different modalities stem from the same tissue. However, the distribution must match globally as well. This is a big limitation for analyzing different datatypes from different tissues and thus a limitation in the immunological context when integrating bone marrow with lymphocyte samples.

Other tools have explored the integration using matrix factorization such as LIGER[14] and coupled-NMF[15], whereas other tools such as Conos[16] and Seurat[17] focus on finding correlations to identify cells in each other neighborhood across the different datatypes. Both these approaches show good results for matched cells, meaning that scRNA-seq and scATAC-seq were measured in the same cell.

These measurements are becoming more and more common, but measuring all these quantities from a single cell at once is often limited due to experimental restrictions. Not only are these measurements significantly more expensive than single modal measurements, but also might require different sample preparation methods, thus not allowing simultaneous measurements.

And as we have seen in the studies mentioned above such as from Wimmers et al.[6] many datasets already exist, where different multi-omics profiles are measured not on the same cell, but on a similar cell from the same cell type in a parallel sample. To use this vast amount of already available data requires diagonal integration of unmatched cells, which is argued to be the most difficult integration[18].

For this diagonal integration neural networks have shown promising results. Most existing neural network approaches are based on autoencoders, but most of them require paired data to integrate cells. Two recently published methods scJoint[19] and scDART[20] are two neural network-based methods specifically developed for integrating unmatched cells from scRNA-seq and scATAC-seq. ScJoint transfers cell type labels from scRNA-seq to scATAC-seq by leveraging the many atlas-scale single-cell RNA data using a semi-supervised neural network. ScDART on the other hand, focuses on keeping the original cell trajectory structures in the data, thus it is defined for datasets with continuous trajectories in their data and not for discrete clusters. This thus is from particular interest for trajectory studies seen in immunological studies as described above.

The developers of both scJoint and scDART compared their new approach against Seurat and Signac[12], arguably the two most popular tools based on canonical correlation analysis. Seurat and Signac first perform a clustering of the cells based on their transcriptomics or accessibility information and then integrate the cells based on shared neighbors. In contrast, both mentioned neural network approaches first integrate scRNA-seq and scATAC-seq and then can perform clustering. Therefore, the advantages of clusters can still be utilized for downstream analysis.

In this research, we propose a method which allows the integration of scRNA-seq and scATAC-seq. We are aiming to explore the strengths of using a neural network approach for the latent embedding of both datatypes specifically for trajectory analysis. Combined with selected downstream analysis, this should be specifically designed for comparison studies in the immunological context, but also applicable to other contexts with a similar design in question. By this, we want to tackle the problem of lacking analysis tools for the existing vast amounts of data.

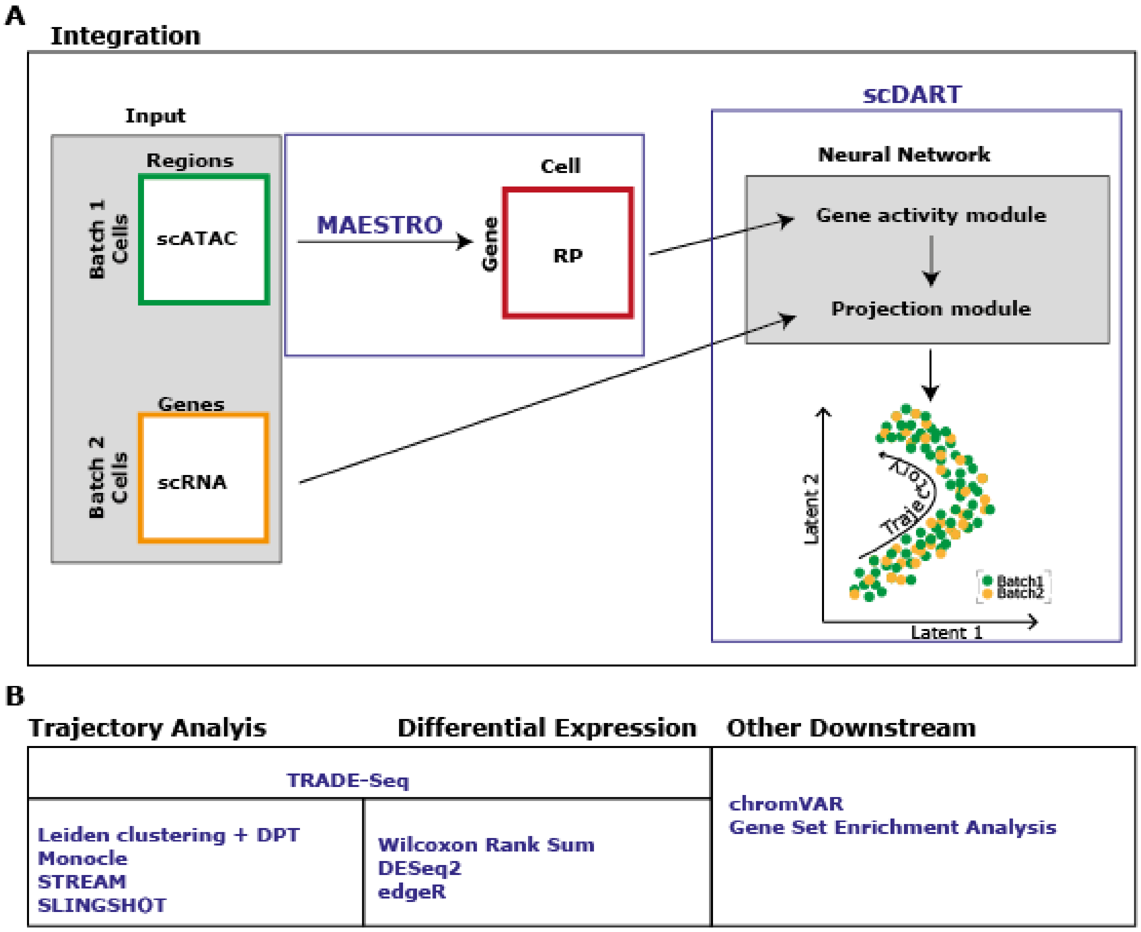

Figure 1: **Overview of framework.A** Overview of the proposed integration of scATAC-seq and scRNA-seq. A regulatory potential score per gene in each cell is calculated from scATAC-seq using an algorithm described in MAESTRO[21]. The three-layer neural network gene-activity module uses this binary matrix and scATAC-seq to create a pseudo RNA-seq count matrix in a non-linear fashion. The projection module then integrates the scRNA-seq to combine both into the same latent embedding. The overall structure of the neural network is based on scDART[20]. **B** Overview of the proposed downstream analyses.

2.2 Approach

We propose the development of a computational tool specifically designed for the analysis and integration of unmatched scRNA-seq and scATAC-seq data in regard to follow the dynamics of cells and compared between different individual groups. This tool thus is of specific relevance for long-term clinical immunological studies trying to understand the immune response and differences between patients and healthy individuals. But will also be applicable to a similar experimental design.

To develop such a tool, we want to implement principles and ideas from current neural network-based integration approaches as these methods have outperformed other existing methods in integrating unmatched cell types.

By optimizing specific input and biological assumptions, we hope to improve the algorithm further leading to even higher accuracy. The output of this tool should be ideal for further downstream analysis and in particular the comparison between at least two different groups.

The first step is optimizing the neural network integration of scRNA-seq and scATAC-seq. The recently developed tool scDART provides a good starting point for this integration.

ScDart's neural network structure requires three types of input: scRNA-seq, scATAC-seq and a pre-defined gene-activity matrix (GAM), which is constructed from the scATAC-seq data. Their neural network is designed with two modules. On the one hand the gene-activity module takes the scATAC-seq to transform the chromatin accessibility in a non-linear fashion into a pseudo-scRNA-seq count matrix. This pseudo-scRNA-seq count matrix is then combined with the scRNA-seq matrix in the projection module, to project both matrices into a shared latent space. The three-layer fully connected network, which is used, allows for non-linearity in the relationship between both datatypes. This provides an inherent advantage over integration methods, only using linear transformation since the complexity biological system can rarely be described with a linear relationship.

For the model training and latent embedding a stochastic gradient descent is used. Further, the post-processing of the embedding allows for a cleaner trajectory structure. This is done by constructing a mutual nearest neighbor graph of the latent embedding and constantly updating the embedding by the embedding of the neighbors. This approach is similar to the construction of the k-nearest neighbor graph in Signac clustering approach, however in this case it is used to detect the complex trajectory structures in a more accurate way. This embedding and integration of both datatypes can then be used by different trajectory inference methods to infer cell trajectories. For the selection for this method, see the discussion below.

ScDART further measures and minimizes the Maximum Mean Discrepancy (MMD) of the similarity of the trajectory structures to minimize the batch effect. In addition, by including knowledge of known matching cells a term is added to the overall loss function to optimize the matching in the latent space. By that this method is suitable for horizontal integration, but more importantly also unmatched cells and diagonal integration.

In its paper, scDART outperformed the other integration tools LIGER[14], Seurat[17] and UnionCom[22]. However, the researchers also note themselves that the accuracy of all methods is relatively low when it comes to predicting gene expression data from chromatin accessibility data. Even though they implement a non-linear relationship approach to infer the pseudo-scRNA-seq count matrix from scATAC-seq, they cannot capture all the elements and factors which can affect the level of gene expression.

The neural network to transform scATAC-seq into scRNA-seq is a three-layer fully connected neural network using leaky rectified linear unit (ReLU) for the activation function. The weight of each layer is calculated by using the pre-defined gene-activity matrix. They assume that this matrix includes all the potential regulators, however during the construction of the gene-activity matrix only regions around 2000 base-pairs upstream of the gene body on the genome are considered. However, it is known that other distal regulatory element can lie outside of these regions, some enhancers can even be located on different chromosomes and can have a significant impact of the activity of the respective gene[23].

Thus, to improve the neural network describing the nonlinear relationship between scATAC-seq and scRNA-seq, we propose an implementation of a more including approach in the construction of this gene activity matrix (Fig.1A).

This problem is not unknown for scATAC-seq. A popular tool targeting this issue is Cicero[24]). Cicero first quantifies correlation of putative regulatory elements and then uses them in an unsupervised machine learning approach to link them to target genes. By this it shows to also include distal regulatory elements. To do this, Cicero requires similar cell types either to be readily clustered or already along a trajectory to aggregate similar cells together and adjusting for technical confounders. Thus, it is not possible to use this approach with scDART, as scDART only performs clustering and trajectory analysis after integration.

Another method, MAESTRO, models gene activity by calculating a regulatory potential[21]. Their model assumes that the effect on gene expression by an accessibility peak is additive and independent and follows an exponential decay from peak to transcription start site. The regulatory potential calculation is a simple multiplication of the binary matrix output from the ATAC-seq peak calling step with a weight matrix. This weight matrix contains the regulatory potential of each gene calculated by the mentioned exponential decay function. The half-decay can be user customized, and the developers recommend a 1kb and 10kb for promoter-driven or enhancer-driven respectively. They even did further testing in their so-called "enhanced RP model" there they also include adjustments to the weights depending on if the peak is located in the exon region, then it is normalized by the total exon length and if the peak is present in the promoter region of a nearby gene it is excluded from the calculation of the other gene. They have shown that these assumptions improved the accuracy of their model. As scDART also uses a simple binary matrix for their calculation, the algorithm could also take in the results from MAESTRO's regulatory potential calculations. If this would improve the accuracy of the overall model must be tested.

Trajectory analysis

Following the integration of scATAC-seq and scRNA-seq in scDART one can conduct a trajectory inference. In their manuscript they first use Leiden clustering and minimum spanning tree (MST) to infer the backbone of the trajectory[25]. For pseudo time inference they use diffusion pseudo time (DPT), developed for reconstructing lineage branching[26]. This approach is similar to the partition-based graph abstraction (PAGA) developed for preserving continuous and disconnected structures of scRNA-seq at multiple resolutions[27]. We further want to evaluate other methods for this trajectory analysis.

One of the most popular tools for single cell omics is Monocle[28]. Monocle uses additive models to check the relationship between gene expression and pseudo time.

A method recently developed tackling the sparsity of scATAC-seq data is STREAM[29]. This method also focuses on density information thus studying how the subpopulation composition changes along the trajectory and whether there are significant changes at branch points.

As the goal of our study is to find differences between patient groups, we want to optimize our used methods including trajectory analysis for this means. TradeSeq is a method specifically designed for trajectory-based differential expression analysis[30]. This package identifies differential expression patterns along the trajectory.

By identifying differential expression patterns along the trajectories, it able to find "between-lineage DE" when comparing microvillous, sustentacular and neuronal lineages. We want to test this method on our immunological data and study if the trajectories are different between our studied individuals and disease groups. By this we can focus on the different in trajectories of immune cells in particular T and B cells, which was already studied in large scale trajectory analysis such as seen by He et al.[31].

Proposed downstream analysis

To analyze our immunological data for biological information, we also propose a framework for further downstream analysis (Fig.1B).

Described above is the method by TradeSeq trajectory-based differential expression analysis. These results we also want to compare with other current methods. Seurat and Signac use a simple two-group comparison Wilcoxon Rank-sum test. This method has however come under strong criticism for single cell modalities as it leads to many false positives[32]. Thus, we want to include also other methods for differential expression analysis. Most methods are based on creating pseudo-bulk RNA-seq datasets by aggregating counts together per cell type and then using established bulk RNA-seq methods such as DeSeq2[33] and edgeR[34]. We fear that these methods will lead to the loss of the gained information by our trajectory analysis since we are interested in the continuous trajectory in contrast to discrete clusters. But we do need to prove this hypothesis first and see if those methods are insufficient to cover the complexity of trajectory analyses.

Following this analysis, a gene set enrichment analysis can be conducted to find pathways, which are over-represented along the trajectories and between the two comparison groups[35].

The combination of scRNA-seq and scATAC-seq does not only allow the study of differential gene expression, but also differential accessibility study. We want to utilize chromVAR[36] to find motifs, which are differentially accessible in scATAC-seq along the trajectory.

This should allow for correlating differentially expressed genes with accessible regions. By this we can also check for known epigenetic signatures and how they persist along the trajectory of our cells. Thus, including scATAC-seq in the analysis allows for the study of the memory of the immune system. We can also test if those signatures allow us to make predictions of the immune response following another round of vaccination as epigenetic signatures are now commonly used for clinical outcome predictions[37].

Overall, this setup of the proposed framework including the neural network used to integrate scRNA-seq and scATAC-seq followed by selected downstream analyses allows good implementation of long-term immunological study. We envision analyzing samples from at least two different groups of individuals. By comparing healthy individuals with patients, we hope to elucidate underlying mechanism of the disease. Coupled with vaccination studies we want to study how these groups differ in their immune response. By utilizing scRNA-seq and scATAC-seq at different timepoints, we hope to study the gene expression and epigenetic signatures of our samples. In particular, for vaccination studies involving more than one dosage, we hope to be able to use the epigenetic

Tested Performace	**Evaluation Criteria**
Latent Embedding	Graph connectivity score Adjusted rand index
Trajectory	Pseudotime consistency score using Pearson correlation
Differential Expression Analysis	Area under the curve (AUCC)

Table 1: Overview of the different performance the framework will be tested and which criteria will be used for evaluation of the performance.

signatures to predict the vaccine response. That is why we focus in the proposal on choosing computational tools for keeping the trajectories of the different cell types in the data and selected further downstream analysis for differential gene and accessibility analysis, gene set enrichment analysis and pathway studies.

Workplan

Description of tools. The first step of this research is building up on the tool of scDART. ScDART is python-based, and their code is published online (https://github.com/PeterZZQ/scDART). After executing and reproducing their published results, we want to focus on improving the gene-activity matrix construction as explained in more detail above. For this, scDART provides an R-script "calc_gact.R” resulting simply in a csv file. Here, one can on the one hand play with the how many base pairs should be included up or downstream to see how the resulting matrix changes. One the other hand, as described, we want to test other means of construction and test MAESTRO. Their code is also accessible online (https://github.com/liulab-dfci/MAESTRO). Thus, we can integrate their way of constructing the regulatory potential as described in " ATAC-CalculateGenescore.R” as the input for scDART. We then will test if this enhances the accuracy of our model.

Next step is selecting the best trajectory analysis for our downstream analysis. As discussed, we first want to test TradeSeq (https://github.com/statOmics/tradeSeq) as we also want to test the differentially expression analysis included there. We want to compare these results to the trajectory analysis of DPT, Monocle and STREAM and then following differential expression analysis.

Description of data and evaluation criteria. All used methods must be thoroughly tested. For an overview of which performance will be tested using which criteria see table1. First, we want to test our new pipeline with the data provided by scDART and see if we can increase the accuracy. We can also simulate data by using Symsim (https://github.com/PeterZZQ/Symsim2), which simulates scRNA-seq and scATAC-seq specifically for trajectory analysis. The advantage is that by using simulated data, we can test our tool thoroughly for continuous trajectories and it provides easy ground-truths to test the results against other methods.

Further, Aragelaguet et al.[18] provides a list of benchmarking datasets for diagonal integration. In addition, we want to include datasets with matched cells such as from SNARE-seq[38], where scRNA-seq and scATAC-seq is measured simultaneously in the same cell. This will help us to benchmark our framework against existing methods of integration.

We believe in using a variety of different evaluation criteria as until now, no gold-standard for evaluation exists. For evaluation of the embedding of scRNA-seq and scATAC-seq we will use a graph connectivity score to assess the batch effect as demonstrated in benchmarking studies by Luecken et al.[39]. ScDART also used the adjusted rand index (AR) to evaluate how the identity of a cell was conserved during integration.

For the assessment of the trajectories, we can use a pseudo time consistency score calculated using Pearson calculation as seen in Zhang et al.[20] and Chen et al.[40].

For the evaluation of the differential expression analysis, we want to follow the many metrics such as area under the curve (AUCC) as proposed in a benchmarking study comparing many different DEA tools[32].

Lastly, we want to test our method on our proposed research goal of comparison studies. For example, the data from Wimmers et al.[6] can be used. In this study they collected scRNA-seq and scATAC-seq among other datatypes for a prolonged period of time after administration of the influenza vaccine and found "distinct subcluster of monocytes with reduced chromatin accessibility”. Other immunological datasets we want to investigate can be found in table2.

Test Data	Immunological studies of interest
Simulated Data using Symsim	Wimmers et al.[6]
Benchmarking data from Aragelaguet et al.[18]	Li et al[41]
SNARE-Seq[38]	Zheng et al.[42]
	Kartha et al.[43]

Table 2: Overview of immunological studies, which collected scRNA-seq and scATAC-seq data.

2.3 Feasibility/ Risk Assessment

Assumptions of our model

One critical assumption we are making in this proposal is that the latent embedding of both scRNA-seq and scATAC-seq exist and that it is to some degree shared by both datatypes. This is a common assumption and seen in Seurat and Signac as well. However, on top of that we also assume that the trajectories of both datatypes are also to some extent shared. If we have reason to believe that this is not necessarily the case, we have to overthink our integration approach and conduct a trajectory analysis before integrating both data modalities. This should be tested before designing the tool. One way to do this could be by using STREAM as this trajectory analysis also works for the unintegrated scATAC-seq as it can overcome the higher level of sparsity compared to scRNA-seq trajectory analysis tools.

Needs thorough testing on performance

This proposal only concerns computational methods. The proposed framework is however not only developed for future studies but can already be leveraged on current datasets. The big advantage of the integration is that unmatched cells can be integrated. Thus, different experiments can be combined for analysis.

This however raises the question how to test this method. Looking at other papers and studies the evaluation criteria are very diverse and rarely a consistent method is used.

Argelaguet et al. propose benchmarking datasets specifically for diagonal integration[18]. This allows to compare this method to other currently methods such as Seurat. Diagonal integration methods are in addition benchmarked with matched experiment to see whether the correct cells are matched.

It must be evaluated whether this method can outperform other computational methods. We do not suspect clearly better results for matched datasets or for data underlying distinct clusters. However, we also see no strong reason to believe that they will not be equal or worse than current methods. That is due to the fact that each proposed step was shown in their own development as equal or better than other standard methods. But we hope to propose a method, which can help finding new biological pathways for immunological or similar studies as described above.

Computational limitations and accessibility

As in all computational tools one point of concern remains the computational efficiency. Especially by leveraging neural networks as central step, we are aware that this method might require high computational power. We believe that typical research centers have access to some form of high-performance computing, but nevertheless this method must be developed with computational storage and memory in mind. Otherwise, the most accurate method still remains useless if the broader research community cannot access the analysis.

With the accessibility in mind, we want to develop this framework in the programming language of R. On the one hand, this allows easy integration with the above-mentioned tools as they are also mostly available as R package and their code will help to design this tool. On the other hand, by providing an open-source package other researchers will be easily able to make adaptations and we hope to reach the larger bioinformatics community.

2.4 Scientific and Societal Impact

The scientific impact of this research can be divided into short- and long-term effects.

In the short term, we hope to provide a framework for the bioinformatics community to gain more insights from comparison analysis. We also want to contribute to the discussion of which tools should be used when as bioinformaticians are aware of the fact the selection of the correct tool depends on the biological question at hand.

In the long-term we want to inspire the community for the development of new tools including neural networks to tackle the big data at hand. This tool can serve as an addition to the discussion of what limitations are present in our current methods. Integration of multiple datatypes will remain one of the biggest challenges of the bioinformatics community as more and more data is created. However, it remains unclear how this problem can be solved. Of course, in an ideal world all data modalities would be measured at the same time in a single cell without disturbing them in their environment. That would be the easiest and most precise integration, however even then many computational limitations such as the nature of sparse data would remain.

Although that ideal world does not exist yet, we can already measure many datatypes. By developing more and more integration methods, we will be able to provide a more and more holistic view of the underlying biology even with our existing datatypes. Our proposed method relies on a neural network, which does provide the option to also be trained with other data modalities. Thus, we could see in the long-term also the integration of more datatypes such as methylation patterns to provide a more detailed epigenetic signature. Therefore, this framework can be adapted to an even more including and holistic view of the immune system by including new arising data.

By using already existing datasets we also provide a societal impact by leveraging already funded research. Most datasets were collected during already published research. Using this data again and hopefully elucidating new biological information, we hope to provide an even more effective funding of public research.

2.5 Ethical Considerations

None. All proposed methods and data are publicly available and no new patient data must be collected for this study. Therefore, we see no ethical issues and concerns.

Bibliography

[1] Annika Fendler et al. "COVID-19 vaccines in patients with cancer: immunogenicity, efficacy and safety". In: *Nature Reviews Clinical Oncology* 19.6 (2022), pp. 385–401. ISSN: 1759-4774. DOI: 10.1038/s41571-022-00610-8.

[2] Anna Furlan et al. "COVID-19 in B Cell-Depleted Patients After Rituximab: A Diagnostic and Therapeutic Challenge". In: *Frontiers in Immunology* 12 (2021), p. 763412. DOI: 10.3389/fimmu.2021.763412.

[3] Qiqi Cao et al. "Integrated single-cell analysis revealed immune dynamics during Ad5-nCoV immunization". In: *Cell Discovery* 7.1 (2021), p. 64. ISSN: 2056-5968. DOI: 10.1038/s41421-021-00300-2.

[4] Kevin J. Kramer et al. "Single-cell profiling of the antigen-specific response to BNT162b2 SARS-CoV-2 RNA vaccine". In: *Nature Communications* 13.1 (2022), p. 3466. DOI: 10.1038/s41467-022-31142-5.

[5] Yi Wang et al. "Single[U+2010]cell transcriptomic atlas reveals distinct immunological responses between COVID[U+2010]19 vaccine and natural SARS[U+2010]CoV[U+2010]2 infection". In: *Journal of Medical Virology* (2022), 10.1002/jmv.28012. ISSN: 0146-6615. DOI: 10.1002/jmv.28012.

[6] Florian Wimmers et al. "The single-cell epigenomic and transcriptional landscape of immunity to influenza vaccination". In: *Cell* 184.15 (2021), 3915–3935.e21. ISSN: 0092-8674. DOI: 10.1016/j.cell.2021.05.039.

[7] Maojun You et al. "Single-cell epigenomic landscape of peripheral immune cells reveals establishment of trained immunity in individuals convalescing from COVID-19". In: *Nature Cell Biology* 23.6 (2021), pp. 620–630. ISSN: 1465-7392. DOI: 10.1038/s41556-021-00690-1.

[8] Daniel J. Kunz, Tomás Gomes, and Kylie R. James. "Immune Cell Dynamics Unfolded by Single-Cell Technologies". In: *Frontiers in Immunology* 9 (2018). can use to describe why sc and ATAC-seq is needed for immune cell trajectories, p. 1435. ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.01435.

[9] Lars Velten et al. “Human haematopoietic stem cell lineage commitment is a continuous process”. In: *Nature Cell Biology* 19.4 (2017), pp. 271–281. ISSN: 1465-7392. DOI: 10.1038/ncb3493.

[10] Daniel Wai-Hung Ho et al. “Single-cell RNA sequencing shows the immunosuppressive landscape and tumor heterogeneity of HBV-associated hepatocellular carcinoma”. In: *Nature Communications* 12.1 (2021), p. 3684. DOI: 10.1038/s41467-021-24010-1.

[11] Renshen Xiang et al. “Cell differentiation trajectory predicts patient potential immunotherapy response and prognosis in gastric cancer”. In: *Aging (Albany NY)* 13.4 (2021), pp. 5928–5945. DOI: 10.18632/aging.202515.

[12] Tim Stuart et al. “Single-cell chromatin state analysis with Signac”. In: *Nature Methods* 18.11 (2021), pp. 1333–1341. ISSN: 1548-7091. DOI: 10.1038/s41592-021-01282-5.

[13] Joshua D. Welch, Alexander J. Hartemink, and Jan F. Prins. “MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics”. In: *Genome Biology* 18.1 (2017), p. 138. ISSN: 1474-7596. DOI: 10.1186/s13059-017-1269-0.

[14] Joshua D. Welch et al. “Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity”. In: *Cell* 177.7 (2019). paper for LIGER, 1873–1887.e17. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.05.006.

[15] Zhana Duren et al. “Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations”. In: *Proceedings of the National Academy of Sciences* 115.30 (2018), pp. 7723–7728. ISSN: 0027-8424. DOI: 10.1073/pnas.1805681115.

[16] Nikolas Barkas et al. “Joint analysis of heterogeneous single-cell RNA-seq dataset collections”. In: *Nature Methods* 16.8 (2019), pp. 695–698. ISSN: 1548-7091. DOI: 10.1038/s41592-019-0466-z.

[17] Yuhan Hao et al. “Integrated analysis of multimodal single-cell data”. In: *Cell* 184.13 (2021), 3573–3587.e29. ISSN: 0092-8674. DOI: 10.1016/j.cell.2021.04.048.

[18] Ricard Argelaguet et al. “Computational principles and challenges in single-cell data integration”. In: *Nature Biotechnology* 39.10 (2021). nice overview article! see notes in word!, pp. 1202–1215. ISSN: 1087-0156. DOI: 10.1038/s41587-021-00895-7.

[19] Yingxin Lin et al. “scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning”. In: *Nature Biotechnology* 40.5 (2022). recent tool for integration using a semisupervised framework and a neural network to simultaneously train labeled and unlabeled data allowing label transfer and joint visualization in an integrative framework good information in the introduction about the other existing tools don’t have their own way of getting the gene activity matrix, pp. 703–710. ISSN: 1087-0156. DOI: 10.1038/s41587-021-01161-6.

[20] Ziqi Zhang, Chengkai Yang, and Xiuwei Zhang. “scDART: integrating unmatched scRNA-seq and scATAC-seq data and learning cross-modality relationship simultaneously”. In: *Genome Biology* 23.1 (2022). computational tool using a neural network for integration, p. 139. ISSN: 1474-7596. DOI: 10.1186/s13059-022-02706-x.

[21] Chenfei Wang et al. “Integrative analyses of single-cell transcriptome and regulome using MAESTRO”. In: *Genome Biology* 21.1 (2020), p. 198. ISSN: 1474-7596. DOI: 10.1186/s13059-020-02116-x.

[22] Kai Cao et al. “Unsupervised topological alignment for single-cell multi-omics integration”. In: *Bioinformatics* 36.Supplement_1 (2020), pp. i48–i56. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa443.

[23] Stavros Lomvardas et al. “Interchromosomal Interactions and Olfactory Receptor Choice”. In: *Cell* 126.2 (2006). Further complicating the issue, there has also been report that enhancers could activate target genes located on different chromosomes [45]., pp. 403–413. ISSN: 0092-8674. DOI: 10.1016/j.cell.2006.06.035.

[24] Hannah A. Pliner et al. “Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data”. In: *Molecular Cell* 71.5 (2018), 858–871.e8. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2018.06.044.

[25] V. A. Traag, L. Waltman, and N. J. van Eck. “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Scientific Reports* 9.1 (2019), p. 5233. DOI: 10.1038/s41598-019-41695-z. eprint: 1810.08473.

[26] Laleh Haghverdi et al. “Diffusion pseudotime robustly reconstructs lineage branching”. In: *Nature Methods* 13.10 (2016), pp. 845–848. ISSN: 1548-7091. DOI: 10.1038/nmeth.3971.

[27] F. Alexander Wolf et al. “PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells”. In: *Genome Biology* 20.1 (2019), p. 59. ISSN: 1474-7596. DOI: 10.1186/s13059-019-1663-x.

[28] Junyue Cao et al. “The single-cell transcriptional landscape of mammalian organogenesis”. In: *Nature* 566.7745 (2019). most details of monocle3, pp. 496–502. ISSN: 0028-0836. DOI: 10.1038/s41586-019-0969-x.

[29] Huidong Chen et al. “Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM”. In: *Nature Communications* 10.1 (2019). tool for trajectory analysis of scRNA-seq and scATAC-seq STREAM but with feature selection, p. 1903. DOI: 10.1038/s41467-019-09670-4.

[30] Koen Van den Berge et al. “Trajectory-based differential expression analysis for single-cell sequencing data”. In: *Nature Communications* 11.1 (2020). TradeSeq paper, p. 1201. DOI: 10.1038/s41467-020-14766-3.

[31] Shuai He et al. “Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs”. In: *Genome Biology* 21.1 (2020). maybe mention this for atlas integration integration of more data for more knowledge maybe mention this for atlas integration integration of more data for more knowledge not really a ATLAS level trajectory analysis, p. 294. ISSN: 1474-7596. DOI: 10.1186/s13059-020-02210-0.

[32] Jordan W. Squair et al. “Confronting false discoveries in single-cell differential expression”. In: *Nature Communications* 12.1 (2021), p. 5692. DOI: 10.1038/s41467-021-25960-2.

[33] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (2014), p. 550. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0550-8.

[34] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2010), pp. 139–140. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616.

[35] Aravind Subramanian et al. “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles”. In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550. ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102.

[36] Alicia N Schep et al. “chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data”. In: *Nature Methods* 14.10 (2017), pp. 975–978. ISSN: 1548-7091. DOI: 10.1038/nmeth.4401.

[37] Matteo Ferro et al. “Epigenetic Signature: A New Player as Predictor of Clinically Significant Prostate Cancer (PCa) in Patients on Active Surveillance (AS)”. In: *International Journal of Molecular Sciences* 18.6 (2017), p. 1146. DOI: 10.3390/ijms18061146.

[38] Song Chen, Blue B. Lake, and Kun Zhang. “High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell”. In: *Nature Biotechnology* 37.12 (2019), pp. 1452–1457. ISSN: 1087-0156. DOI: 10.1038/s41587-019-0290-0.

[39] Malte D. Luecken et al. “Benchmarking atlas-level data integration in single-cell genomics”. In: *Nature Methods* 19.1 (2022), pp. 41–50. ISSN: 1548-7091. DOI: 10.1038/s41592-021-01336-8.

[40] Huidong Chen et al. “Assessment of computational methods for the analysis of single-cell ATAC-seq data”. In: *Genome Biology* 20.1 (2019). Nice introduction explaining the difficulties of scATAC-seq, p. 241. ISSN: 1474-7596. DOI: 10.1186/s13059-019-1854-5.

[41] Shun Li et al. “Epigenetic Landscapes of Single-Cell Chromatin Accessibility and Transcriptomic Immune Profiles of T Cells in COVID-19 Patients”. In: *Frontiers in Immunology* 12 (2021), p. 625881. DOI: 10.3389/fimmu.2021.625881.

[42] Yingfeng Zheng et al. "A human circulating immune cell landscape in aging and COVID-19". In: *Protein & Cell* 11.10 (2020), pp. 740–770. ISSN: 1674-800X. DOI: 10.1007/s13238-020-00762-2.

[43] Vinay K. Kartha et al. "Functional Inference of Gene Regulation using Single-Cell Multi-Omics". In: *bioRxiv* (2021), p. 2021.07.28.453784. DOI: 10.1101/2021.07.28.453784.