Bioinformatics
github.com/jonasjonker/Integrated-Bioinformatics-Project
Advance Access Publication Date: 14-12-2020
Integrated Bioinformatics project

KU LEUVEN

Novel tools for the Integration of Single-Cell Multi-Omics Data – A Case Study on Human Blood Cells

Jonas Jonker[1], Kristen Michelle Nader[1], James O'Reilly[1], Nina Rank[1]

[2]Faculty of Bioscience Engineering, KU Leuven, Leuven, 3000, Belgium.

Abstract

Recent developments in single-cell multi-omic sequencing technologies allow researchers to probe multiple molecular layers in the same cell simultaneously. A number of tools have been developed to integrate this complex multi-modal data and generate actionable insights. We use Multi-Omics Factor analysis (MOFA) to integrate and analyse scRNA-seq and scATAC-seq data from human peripheral blood mononuclear cells. This novel multi-omic assay is prohibitively expensive for many, and it is therefore practical to predict gene expression data from chromatin accessibility. Using this novel dataset, we also evaluate gene expression prediction in ArchR. We conclude that MOFA is a useful tool for in-depth analysis of single-cell multi-omic data and that prediction of gene expression from chromatin accessibility is a useful alternative for researchers on a budget.

1 Introduction & Background

The advent of high-throughput single-cell sequencing technologies in the last decade has enabled biological research at a resolution which was previously not possible[1]. The pipelines and tools necessary for processing and analysing single-cell omics data are now well-established, allowing for rapid analysis of single-cell genome, transcriptome, epigenome, and proteome data [2][3]. However, a holistic and robust study of complex biological processes at the single-cell level requires an integrative approach which combines these multi-omic data. This integrative analysis permits a better understanding of the interplay between molecular layers and ultimately their combined impact on cellular phenotype.

Drawing from advances in single-cell isolation and barcoding, a number of technologies have recently been developed for the high-throughput sequencing of single cell multi-omic data [4][5]. Some of these assays sequence both the epigenome and transcriptome, simultaneously probing gene expression alongside chromatin accessibility or DNA methylation [6][7]. Tools facilitating integrative analysis of transcriptome and epigenome data have been developed alongside these technologies [8]. Using PBMC data from 10X Genomics' Chromium scATAC-seq and scRNA-seq assay, we demonstrate the integration of multi-omic data using Multi-Omics Factor Analysis (MOFA) and investigate the efficacy of predicting gene expression from chromatin accessibility using ArchR.

1.1 Integration of Transcriptome and Epigenome Data

Understanding the regulatory relationship between epigenetic changes and gene expression at the single-cell level allows for comprehensive insight into the impact of epigenetics on cell-type specific gene regulation. This permits a detailed analysis of the epigenetic mechanisms underlying the behaviour and characteristics of cell populations such as cell differentiation, cell trajectories, and cellular heterogeneity. Simultaneously sequencing the epigenome and transcriptome in a single assay mitigates confounding factors which arise from cell-to-cell variation.

A number of tools are available for the integrative analysis of transcriptome and epigenome data, with varying underlying assumptions and applications [9]. Multi-omics factor analysis (MOFA) is an unsupervised approach that uses a probabilistic Bayesian model to disentangle the unique sources of heterogeneity from shared sources within the different data modalities [10]. To do this, MOFA infers a low-dimensional representation of the multi-omic data in terms of a small number of latent factors that capture the global sources of variability. The authors suggest this can be intuitively understood as a statistically rigorous generalization of principal component analysis (PCA). The relative contributions of specific features (eg. genes or ATAC peaks) to these latent factors can then be extracted to give insight into their impact on cellular heterogeneity. In this paper, we use MOFA to analyse data from human PBM cells generated by 10X Genomics' Chromium scATAC-seq and scRNA-seq assay.

1.2 Predicting Gene Expression from chromatin Accessibility

These new multi-omic assays are commercially available but expensive. For this reason, it is still useful and practical to predict data from one molecular layer in a cell given data from the other. This would also allow investigators to use single-omic data from previous experiments to gain

new insights. A number of papers have recently been published with this aim in mind [11][12][13].

Using single-cell chromatin accessibility data to predict gene expression data is useful for clustering, cell-type annotation, and differential expression analyses. ArchR, a recently developed R package with an extensive suite of tools for scATAC-seq data analysis, aims to impute gene expression from open chromatin using a distance-weighted accessibility model [14]. ArchR uses distance-weighted accessibility models to predict gene expression from ATAC-seq peaks in the vicinity of a given gene. The developers found the best predictor of gene expression to be local accessibility of the gene region which includes the promoter and gene body. The mathematics of gene-score imputation are quite dense and we refer the interested reader to the original paper for an in-depth explanation [14].

To evaluate the performance of their model, the developers compared predicted gene scores derived from the model to known gene expression data from previous methods integrating scATAC-seq with scRNA-seq. Using the novel single-cell multiome ATAC-seq and RNA-seq assay, we sought to evaluate the accuracy of gene expression prediction in ArchR.

2 Results

2.1 Multi-Omics Factor Analysis

The PBMC data was annotated with both broad and narrow cell types using a previously annotated reference dataset. A MOFA model was trained on the data to disentangle sources of variability across data modalities, determine factors which contribute to this variability, and cluster the cells using UMAP. The cell type annotations were used to evaluate cell clustering, both for broad and narrow cell type.

2.1.1 MOFA explains variability across data modalities

The MOFA model was trained with 15 factors. A correlation matrix indicated the factors were uncorrelated and therefore not redundant (see Fig. 1m). Across these factors, the RNA-seq component accounted for most of the variance in the dataset (see Fig. 1j, 1k). Two factors of note were factors 2 and 3, which used only ATAC peak and RNA-seq data to calculate their factor scores, respectively. (see Fig. 1c, 1e) Viewing the correlation between factors and feature counts showed that factor 2 and factor 3 were highly correlated to peak and sequence counts respectively (see Fig. 1l). Plotting the factor values and feature counts per cell confirmed that factor 2 has higher factor values with increased peak count, indicating that the increased factor value was a result of a high peak count and not due to a difference in the variance explained between data modalities. The swarmplot also shows that for factors 1 and 3, high sequence counts also had higher absolute factor scores, but that the sign of this factor score was dependent on cell-type. While factor 1 has high-correlation between feature counts and factor score, it is clear from Fig. 1a that factor 1 discriminates between broad cell-types particularly well.

2.1.2 Factor analysis reveals biomarkers for broad cell-type

Further analyses of each factor were performed using feature weight plots, which display the relative contributions of genes to the factor value. Features with no association with the corresponding factor are expected to have values close to zero, whereas features with strong association with the factor are expected to have large absolute values. The sign of the weights indicates the direction of the effect. As factor 1 discriminates between broad cell-types, we plotted the associated RNA weights to determine which genes can be used as biomarkers for lymphoid and myeloid cell-types (see Fig. 1i). Plotting the associated ATAC weights for factor 2 showed that the weights were not centered at zero for both proximal and distal regions, which again highlights that this factor is problematic in that any added feature will increase factor score (see Fig. 1b, 1d, and 1f).

2.1.3 UMAP generates clusters for broad and narrow cell-types

Non-linear dimensionality reduction was performed using the learned MOFA factors as input. The latent factors inferred by MOFA can replace principal components as input to algorithms such as UMAP that learn non-linear manifolds. This allows us to utilise information from both data modalities as well as characterise the useful factors and remove problematic factors arising from technical biases. Performing UMAP with different factor combinations showed that removing factor 2 improved the distance between clusters, but leaving out other factors decreased the number of clusters that can be distinguished. With factor 2 removed, MOFA can correctly cluster our annotated cells in clearly defined clusters and meta clusters that make biological sense (see Fig. 1o, 1p, 1q). There are three clearly defined super clusters each containing myeloid, B-cell or T-cell types. Within these clusters more specific cell types can be distinguished.

2.2 Evaluating Gene Score Imputation with ArchR

The accuracy of ArchR's gene score imputation (GSI) was evaluated by comparing gene scores imputed from PBMC ATAC-seq data with the corresponding expression data taken from the same cells. Pearson correlation was used as the similarity metric between gene-scores and gene-expression.

2.2.1 GSI accuracy varies with gene variability

The original publication reported an accuracy of 0.7 when evaluating GSI on genes which were differentially expressed or highly variable, though they did not present any data for other genes. To investigate if GSI accuracy varied between highly-variable genes and genes with relatively constant expression, we created two independent gene sets based on the impact of these genes on broad cell-type. From the MOFA analysis, we identified factor 1 as crucial in distinguishing between broad cell types. The feature weight decomposition of this factor was used to create two independent datasets for testing:

1. Broad cell-type (BCT) genes with a high positive or negative feature weight for factor 1 (see Fig. 1n).
2. Non-specific (NS) genes are genes with near-zero feature weight for factor 1. PBMC non-specific genes from BloodAtlas were also added to the gene set [15].

We found a significant difference in gene-score imputation accuracy between the BCT and NS gene sets, with a mean correlation of 0.53 and 0.24 respectively (see Fig. 1n left).

To test if the imputed gene scores differed from the actual gene expression in a systematic or reliable manner, we trained smoothing spline models for each gene using imputed gene scores as an input and gene expression as a target. The models were used to predict gene expression data from imputed gene scores, resulting in a higher accuracy of 0.58 and 0.27 for the BCT and NS gene sets, respectively (see Fig. 1n right).

2.2.2 GSI does not rank genes accurately

As the correlation between gene scores and gene expression was somewhat poor using absolute values, we hypothesised that gene score may instead serve as a proxy for gene expression and the qualitative properties of gene scores could mirror those of gene expression. To test this, we compared the rankings of gene scores and gene expression using Kendall's correlation coefficient [16]. The gene set was first pruned to include only the top 15% most highly expressed genes for both gene score and gene expression, as

many genes have zero or near-zero expression counts and would negatively influence the analysis, given our choice of rank metric.

The percentage of genes common to both genesets was first calculated for each cell, to give a rough metric of similarity. Fig. 1h shows the distribution of gene overlap percentage across all cells. We found this overlap was low ($24\% \pm 5\%$) for the majority of cells. Kendall's correlation coefficient τ was then calculated for the overlapping genes (see Fig. 1g). We found that overall the correlation was poor (0.11 ± 0.09), indicating that across cells the ranking of most highly expressed genes was not similar for imputed gene scores and gene expression.

2.2.3 GSI gives accurate clusters when compared with gene expression Lastly, we sought to determine if the poor results for both correlation and ranking affected the practical uses of gene score imputation advertised in the original ArchR publication (clustering and cell-type annotation). Dimensionality reduction was performed with the imputed gene scores using UMAP to visualise cell clusters. The projection was then visualised alongside the UMAP for gene expression and compared (see Fig. 1o). The UMAP for both gene scores and gene expression clustered PBM cells similarly, with some minor differences between the two projections (see Fig. 1p, 1q).

3 Methods

3.1 Datasets

This study uses scRNA-seq and scATAC-seq data from human peripheral blood mononuclear cells (PBMC), taken from a healthy 25 year old female donor. Sequencing was performed using 'Chromium Single Cell Muliome ATAC + Gene Expression Assay (10X Genomics)', which allows for simultaneous profiling of gene expression and open chromatin in the same cell. Raw data from this assay was processed by the 10X Genomics team using Cell Ranger Arc 1.0.0. The data is publicly available on the 10X Genomics support website.

3.2 Cell-type annotation

Cell-type annotation was performed using Seurat (3.2.2) [17]. To annotate cell-types, Seurat first clusters the cells using a graph-based clustering approach, and then identifies differentially expressed genes (DEGs) in each cluster. These DEGs serve as biomarkers which are used to assign cell-type identity to each cluster. Cell-types can be determined by comparing these biomarkers against manually curated databases and scientific literature, or by using a previously annotated dataset as a reference. A previously annotated PBMC dataset was used as a reference in this study. Further annotation based on broad cell-type was then included, as it is useful in downstream analyses. To ensure that the cell-type annotation was accurate, 10X Genomics LOUPE Browser (v4.1.0) was used to interactively view DEGs in each cluster.

3.3 Multi-Omics Factor Analysis: MOFA

We created Seurat objects from the 10X RNA-seq and ATAC-seq data using Seurat v3.2.2. The cell were then annotated with the cell type annotation we created. Cells which weren't annotated were dropped. Peaks within a promoter region were split from the distal peaks and stored as a different assay, within the Seurat object. Signac v1.0.0 and the JASPAR2020 database was to annotate the peak data with additional motif info. The RNA-seq data was log normalised and centered. Both distal and promoter ATACseq data was normalised using term frequency inverse document frequency (TF-IDF). The most variable RNA features where found using Seurat's variance stabilizing transformation (vst). The prepared Seurat object was then used to train a MOFA model with default settings of MOFA2 v1.1.6. We used MOFA2, Seurat and Signac to explore the trained MOFA model.

3.4 Gene Score Imputation with ArchR

Imputation of gene scores from chromatin accessibility data was performed performed using ArchR v1.0.0. [14] The standard ArchR pipeline was used with default parameters for creation of arrow files and for gene score imputation. Doublet cells were removed from the data using ArchR's built-in doublet inference and removal functions. We identified a doublet-filtering issue in ArchR's R v4.0.3 implementation and spoke with the developers to resolve this.

3.5 Single Gene Models

A number of statistical learning methods were used to model the relationship between imputed gene scores and gene expression data for a single gene. First, imputed gene score and gene expression is split into training and test data in a ratio of 80:20. Linear regression models and smoothing splines are then trained on this data and used to predict gene expression from imputed gene scores. Cross validation was used to determine the degrees of freedom for each smoothing spline. The single-gene models are evaluated by calculating the Pearson correlation coefficient between the actual and predicted expression data.

3.6 Rank Models

Similarity in ranks between gene scores and gene expression values was measured using Kendall's correlation coefficient τ. As the gene expression data is very sparse, the gene sets were pruned to keep the top 15% of genes to ensure this wouldn't negatively effect analysis. The gene overlap between the pruned gene sets was first calculated before ranks were compared using Kendall's τ.

3.7 Clustering

We rounded the predicted gene scores to the nearest integer and merged it with our previously trained Seurat object. Integration anchors were then found and used with Seurat to integrate the predicted and observed gene expression data. After PCA, UMAP was used to cluster and visualize the predicted and observed gene expression data. UMAP was used as well to cluster the latent factors found with MOFA.

4 Discussion

Multi-Omics Factor Analysis proved to be a useful approach for integrating scATAC-seq and scRNA-seq data. The factor analysis allowed us to determine which factors were useful in predicting broad cell type, as well as which factors were technical factors resulting from high peak or sequence counts. By plotting the relative weight contributions of the composite features in each factor, we determined a number of genes which were predictive of broad cell type and could be used as biomarkers (see Fig. 1i). MOFA+'s latent factors can be used as inputs to UMAP to accurately cluster the PBMC data for both broad and narrow cell type.

In conclusion, as a tool for integrative analysis, we found that MOFA can generate actionable insights and illustrated the how variance is spread across different data modalities. Our evaluation of ArchR's gene score imputation revealed that it performs well for genes which are differentially expressed (e.g. biomarkers of broad cell type), but performs poorly for genes which are not differentially expressed between cells and have no specific cell lineage. The original publication listed a median imputation accuracy of 0.72 on the top 1000 differentially expressed genes and 0.58 for the top 2000 most variable genes in their dataset.

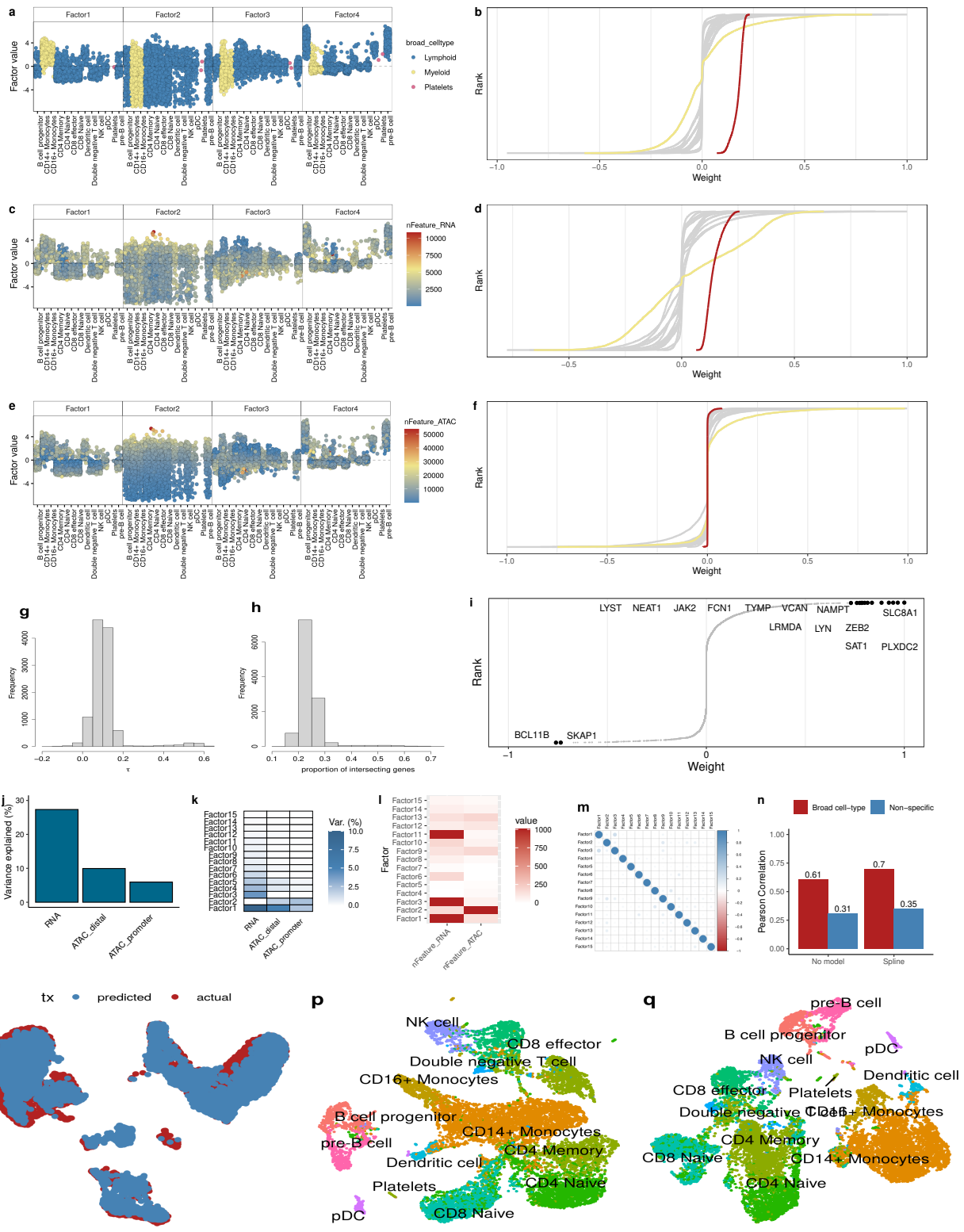

Fig. 1: **(a,c,e)** Swarm plots of factor scores per cell for factor 1-4. The same plot is shown thrice, colored by different variables. Each dot represents one cell. **(a)** colored by broad cell type, showing that factor 1 can distinguish between broad cell type. **(c)** colored by the total number of gene counts RNA sequence count. **(e)** colored by ATAC peak count, indicating that factor 2 mainly captures merely the number of ATAC peaks and thus gives little information about the epigenomic profile. **(b, d, f)** Gene weights distribution for factor 1 (yellow), factor 2 (red) and factors 3-15 (grey) separated by modalities; **(b)** ATAC peak in a promoter region; **(d)** ATAC peaks distal to promoter region(s); **(f)** gene expression. Well-balanced factors should be represented by a sigmoid curve centered around zero. For factor 1 gene weights are centered around zero but are not sigmoid, indicating that any gene will either increase or decrease the factor score. For factor 2, genes are skewed to the right, indicating that all peaks contribute positively to the factor score and is considered a technical factor. **(g)** Distribution of the intersecting gene ratio. Intersecting genes are defined as the genes that occurred in both, the 15% most highly expressed genes and the genes with the 15% highest imputed scores by ArchR, per cell respectively. **(h)** Kendall's correlation coefficient τ for the intersecting genes. **(i)** Gene weights for factor 1, with top 25 genes in absolute weight labeled. **(j)** Explained variance by modality. The RNA-seq component accounted for most of the variance in the dataset. **(k)** Variance decomposition plot revealing the explained variance for each modality by factor. Factor 1 explains around 8% of the total variance in the RNA-seq data. Factor 1 also explains the highest proportion of variance in distal ATAC regions. **(l)** Association between single factors and the total number of gene expression counts (left panel) and total number of peaks (right panel). Factor 1, 3, and 11 are highly associated with the total number of gene expression counts. Factor 2 is highly associated with the total number of ATAC peaks. **(m)** Correlation between latent factors. The factors are largely uncorrelated. **(n)** Pearson correlation for single genes between either the raw imputed gene scores ('No model') or the output of smoothing spline models ('Spline') and the gene expression counts, respectively. **(o)** UMAP clustering based on predicted gene scores (blue) and gene expression (red). Both clusterings were highly similar with some minor differences between the two projections. **(p, q)** UMAP clustering based on MOFA factors with **(p)** and without **(p)** factor 1 and 3, 11 show that these factors improve clustering even though they are correlated with RNA sequence count.

We obtained similar results for the highly-variable genes, using broad cell-type biomarkers as the gene set. The smoothing spline models which were trained using predicted and actual gene expression data gave improved performance, indicating that the imputed gene scores and actual gene expression data differed in a reliable manner. This might suggest room for improvement in ArchR's gene score imputation, as there is a relationship between gene expression and gene scores in our dataset which can be reliably modeled. This would need to be researched using different datasets across multiple species in order to establish if this is true.

Further analysis of both rankings and clustering indicated that while gene score imputation ranked genes poorly, a UMAP clustering using gene scores clustered the cells accurately, with some minor differences for narrow cell types. We conclude that the accuracy of gene score imputation is not adequate for analyses which require precise estimation of gene expression. However, the prediction for differentially expressed and highly-variable genes is sufficiently accurate to give an accurate clustering. Gene score imputation could therefore be used for standard downstream analyses of single-cell data investigating properties of cellular populations such as cellular heterogeneity, cell lineage, or cellular differentiation.

An important question for researchers to ask is whether the benefits of the multiome scATAC-seq and scRNA-seq assay justify the additional cost. Ultimately, the decision between single-omics ATAC-seq assay and the multi-omics assay depends on the research question at hand. If one wants to disentangle sources of variance across data modalities, or if the analysis requires accurate estimations of gene expression as well, then the multi-omics assay is necessary. If however, one is interested primarily in chromatin accessibility, and also wants to investigate properties of cellular populations through clustering, our research indicates a that a single ATAC-seq assay in combination with gene score imputation is sufficient.

4.1 Limitations and Practical Considerations

There are a number of limitations and practical considerations which we feel should be mentioned. Firstly, both MOFA+ and ArchR have a number of parameters which can be tuned, yielding different results. We did not have the computing power to systematically explore this parameter space, and so the default values recommended by the developers were chosen. This is particularly relevant for the analysis of gene score imputation in ArchR, as the model has a number of parameters such as tile size, window size, and a user-defined accessibility model. It is possible that tuning these parameters would improve gene score imputation.

In addition, cell-type annotation was done using a pre-annotated PBMC reference dataset. Our evaluation of clustering for both gene expression data and imputed gene scores is dependent on this clustering. Therefore different annotation methods or different reference datasets may give variable results. We did not have time to benchmark different clustering approaches and therefore different approaches may yield more well-defined clusters.

5 Acknowledgements

We would like to thank both Seppe and Swann for their help with the direction of the project, and also for providing us with a topic we found so interesting. We're also especially grateful to Ricard Argelaguet for his help with MOFA throughout the project and for the time he gave to all our questions. We want to also thank Prof. Aerts for granting us access to the VIB training which proved useful.

References

[1]Xiaoning Tang, Yongmei Huang, Jinli Lei, Hui Luo, and Xiao Zhu. The single-cell sequencing: new developments and medical applications. *Cell & Bioscience*, 9(1):53, 2019.

[2]Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

[3]Robert Salomon, Dominik Kaczorowski, Fatima Valdes-Mora, Robert E Nordon, Adrian Neild, Nona Farbehi, Nenad Bartonicek, and David Gallego-Ortega. Droplet-based single cell rnaseq tools: a practical guide. *Lab on a Chip*, 19(10):1706–1727, 2019.

[4]Lia Chappell, Andrew JC Russell, and Thierry Voet. Single-cell (multi) omics technologies. *Annual Review of Genomics and Human Genetics*, 19:15–41, 2018.

[5]Yukie Kashima, Yoshitaka Sakamoto, Keiya Kaneko, Masahide Seki, Yutaka Suzuki, and Ayako Suzuki. Single-cell sequencing techniques from individual to multiomics analyses. *Experimental & Molecular Medicine*, 52(9):1419–1427, 2020.

[6]Stephen J Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, et al. scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature communications*, 9(1):1–9, 2018.

[7]Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457, 2019.

[8]Jeongwoo Lee, Daehee Hwang, et al. Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine*, 52(9):1428–1442, 2020.

[9]Maria Colomé-Tatché and Fabian J Theis. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, 7:54–59, 2018.

[10]Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):1–17, 2020.

[11]Florian Schmidt, Fabian Kern, and Marcel H Schulz. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenetics & chromatin*, 13(1):4, 2020.

[12]Qiao Liu, Fei Xia, Qijin Yin, and Rui Jiang. Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics*, 34(5):732–738, 2018.

[13]Anirudh Natarajan, Galip Gürkan Yardımcı, Nathan C Sheffield, Gregory E Crawford, and Uwe Ohler. Predicting cell-type–specific gene expression from regions of open chromatin. *Genome research*, 22(9):1711–1722, 2012.

[14]Jeffrey M Granja, M Ryan Corces, Sarah E Pierce, S Tansu Bagdatli, Hani Choudhry, Howard Chang, and William Greenleaf. Archr: An integrative and scalable software package for single-cell chromatin accessibility analysis. *bioRxiv*, 2020.

[15]Mathias Uhlen, Max J Karlsson, Wen Zhong, Abdellah Tebani, Christian Pou, Jaromir Mikes, Tadepally Lakshmikanth, Björn Forsström, Fredrik Edfors, Jacob Odeberg, et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science*, 366(6472), 2019.

[16]M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 06 1938.

[17]Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[18]Jonas Jonker, Kristen M Nader, James O'Reilly, and Nina Rank. github.com/jonasjonker/Integrated-Bioinformatics-Project. *GitHub repository*, 2020.