Integrating single-cell multi-omics data to improve gene regulatory network reconstruction

Dilyara Grafova[1], Martijn van der Werff [1], Karina Díaz-Barba [1], Roya Gharehbeiklou[1], Dan Kaptijn [2], Maryna Korshevniuk [2], Roy Oelen [2], Monique van der Wijst [2]*

[1] Hanzehogeschool Groningen, Zernikeplein, 9747 AA Groningen, Netherlands.

[2] University Medical Center Groningen, Department of Genetics Hanzeplein 1, 9713 GZ Groningen, Netherlands.

* Corresponding author.

Abstract

In the last 10 years single-cell technologies have grabbed the attention of scientists since such techniques provide more insights into gene regulation. However, one of the main struggles with such technology is the sparsity of the data. By simultaneously profiling multiple molecular layers, these technologies provide a comprehensive understanding of cellular processes, provide more data and more features available to the analysis. The integration of such multi-omics data allows us to unravel hidden regulatory mechanisms, view biological data from the perspective of multiple levels. For this purpose, the single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq) data of the human peripheral blood mononuclear cells was analyzed. We preprocessed the available sequencing data with R tools, Seurat and ArchR. As a goal of our research, we set to examine the effect from the integration of both gene expression and chromatin accessibility, and for that reason we performed data integration with Portal. To get a broader idea of how gene regulation works and uncover key regulators we decided to generate a gene regulatory network (GRNs). For this purpose, we compared two different tools, FigR and SCENIC+. Using multiomics data allowed us to connect transcription factors (TFs) to their binding sites and target genes, which in its turn provides great possibility of looking into potential malfunctioning of such linkages and target drug development. For FigR we show that topic modelling has a beneficial effect on GRN construction. Comparing both tools, FigR and SCENIC+, with topic modelling step resulted in a drastic difference.

Key words: scRNA-seq, scATAC-seq, GRN, single-cell, Seurat, ArchR, Portal, FigR, SCENIC+

Introduction

Nowadays it has become clear that genetic variations (primarily single nucleotide polymorphism (SNP)) play a key role in the diversity of the population. Such variations subsequently contribute to the appearance of diseases that can alter gene expression. However, there is a problem of distinguishing causal variants that contribute to a risk of a disease and discovered SNPs provide little effect on the susceptibility [1]. To derive causal variants with some level of certainty, it is beneficial to look at it at the cellular level, since bulk sequencing gives very averaged results and miss out cellular variability [2]. The latest advances in single-cell analysis have made it possible for the profiling of genetic variation specific to cell type and individual along with many observations per individual [3].

ATAC sequencing allows us to get information on the transposase accessible open chromatin regions. Cis-regulatory elements such as enhancers, promoters or silencers, which in their turn play a vital role in gene regulation, are in such open chromatin regions. Combining two single-cell tools, scRNA-seq and scATAC-seq, can assist in uncovering the gene regulation mechanisms by building gene regulatory networks (GRNs) [4].

The drawbacks from using multi-omics data are the experimental complexity and data being sparse and noisy. ATAC-seq data especially contributes to the sparsity due to its binarity [5]. To tackle this issue multiple software tools, mainly including deep learning algorithms, are being developed. In the review article, Brombacher et al., 2022, various multi-omics integration tools were assessed and evaluated based on their performance in terms of capacity of adjusting to technical issues while maintaining biological signals [6]. Based on this step we decided to pursue a goal of comparing the effect of integration of these two types of data using a tool that was among the top-performing ones, Portal [7]. To evaluate the effect of integration we aimed to reconstruct gene regulatory network. Different tools for GRN reconstruction aim to solve a certain aspect, for example FigR [8] focuses on chromatin interactions and derives Domains of Regulatory Chromatin (DORCs) and novel SCENIC+ [9] method aims to accurately predict enhancers along with transcription factors and link them to the target genes. Furthermore, both GRN methods make a use of topic modelling, however that is not a mandatory step for FigR, for that reason we would also like to evaluate if there is place for better reconstruction based on the use of cisTopic object.

Methods

Single-cell multi-omics dataset

The data collection used includes details of 73,0821 human peripheral blood mononuclear cells (PBMCs) of a healthy donor, which is publicly available from 10x Genomics (https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0). This dataset was analyzed by integrating two single-cell omics layers: chromatin accessibility (scATAC-seq) and gene expression (scRNA-seq).

scRNA-seq preprocessing and quality control (QC)

The QC of scRNA-seq data, as well as the identification of sources of heterogeneity of the samples was done using R package Seurat (v4.0) [10] We used the 10X raw unfiltered feature-barcode matrix to create the Seruat object by keeping the entries with a minimum number of 3 cells and a minimum number of 200 features (default parameters). The following QC parameters were used to filter the cells: we removed the

number of molecules detected within a cell (nCount) with nUMIs ≤ MAD_2, as well as the cells with percentage of reads that map mitochondrial genome (%mtRNA) ≥ MAD_3.

scATAC-seq preprocessing and QC

ArchR package (1.0.2) [11] was used to analyze ATAC per fragment information from the 10x data, cell barcodes were aligned to hg38 reference genome (available in http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.chrom.sizes). The following QC parameters were used to filter the cells: we kept the cells that had number of unique fragments mapped to ref genome (nFrags) >= 1000, reads aligned to transcription starting sites vs reads aligned to background regions (TSSEnrichment) >= 4, fraction of reads in peaks after pseudo-bulking (FRiP) > 0.5, as well as extra filter to perform a doublet removal.

Matching preprocessed scRNA-seq and scATAC-seq

Both scRNA-seq and scATAC-seq datasets contain a unique barcode sequence which includes a suffix (a nucleotide sequence) with a dash separator followed by a number. We used these cell-associated barcodes to filter the cells that were present in both datasets, that is, the intersection of both. After this, we ended up with two different objects of the same size and with the same cells, but one with chromatin accessibility information (ArchR object) and the other with gene expression information (Seurat object). We did this step because we need the same number of cells to use it as an input for FigR [8].

Cell type annotation

To automatically annotate the cell types in the scRNA-seq data, R package Azimuth (v.0.4.6) [10] which is a part of Human Biomolecular Atlas Program was used. Specifically, the reference dataset used was the 'celltype.l2' [10] which consists of 161,764 PBMCs (human peripheral blood mononuclear cells) of 8 volunteers. Azimuth returns the cell type annotations for the Seurat object created before (that is for the scRNA-seq data), based on the cell labels described in the 'celltype.l2' reference dataset.

To also gain a lower resolution of cell types, subtypes of cells were regrouped to their major cell type and re-annotated (Table S1*)*, yielding 8 new cell categories. This lower resolution annotation object was the one that was used as an input for further steps. Apart from that, we used the cell barcodes of the scRNA-seq data to transfer the cell type annotations to the scATAC-seq data.

Multimodal data integration

After that, we combined the outcome from the preprocessing steps in a data integration phase. By using an individual omics layer methodology, Portal (v1.0.2) [7] was used as a tool to identify information that would have been missed. The integration, dimensionality reduction, clustering, and visualization steps were performed using the Portal tool.

GRN reconstruction

CisTopic (v2.1.0) [12] was used for topic modeling to group co-accessible regions into regulatory topics and to cluster cells based on regulatory topic contributions. These topics (or peak clusters), which are generated using Latant Dirichlet allocation, are used for motif discovery to predict combinations of TFs [13]. All of this enables summarization of variability of the observed chromatin accessibility counts. To assess the topics CisTopic uses log-likelihood, when it does not increase anymore for additional topics it

means that there are no more main topics to explain the data. Therefore, the topics selection is made after the “flattening” of the log-likelihood. We tested 2, 4, 10, 16, 20, 30 and 35 topics to find the optimal model. After selecting this model all input data could be loaded into the GRN tools.

The last step involved the reconstruction of the GRN by using known transcription factors (TF) and motifs. This was achieved by using two tools: SCENIC+ [9] and FigR [8]. Both tools handle single cell chromatin accessibility and gene expression data as input but use different approaches to select TFs for generating the networks. FigR uses scATAC-seq peak counts as input in the form of a ‘SummarizedExperiment’ which we obtained from ArchR [11]. scRNA-seq data can be provided in the form of a sparse matrix of the log normalized gene expression levels. Additionally, the cisTopic object is loaded, in the form of a matrix containing probability scores based on the dimensionality reduction (topic generation). In the case of SCENIC+ the annotated Seurat object was converted into AnnData format (n_obs × n_vars). The first part of SCENIC+ is topic modeling, for which pseudobulk profiles and consensus peaks set were generated from the initial scATAC data and then filtered out based on the barcodes for scATAC data retrieved after QC in ArchR. Using pycisTopic package [9], a faster alternative to the cisTopic, the pyCistopicObject was created (n_cells × n_regions).

Results

Preprocessing

QC of scRNA-seq data, as well as the identification of sources of heterogeneity of the samples was done by Seurat. After creating the Seurat object, we obtained the gene expression of 26, 412 genes in 12,391 cells. After performing the QC, a matrix of 26,412 genes x 11,939 cells was kept. Extra analyses performed can be seen in (Figure S1 and S2). From 11,898 after the QC filtering of the ATAC data, 10,051 cells were kept, several QC metrics can be seen in (Figure S3). After running Azimuth to create annotations we obtained an object of 26,412 genes across 11,939 cells (Figure S4). With the lower resolution we obtained: 26,412 genes across 11,853 cells (Figure S5). Due to manually assigning lower resolution cell types, 86 cells were lost. In-between scRNA-seq and scATAC-seq data 9,763 barcodes were equal and kept for downstream analysis.

Portal data integration

For the integration of scATAC-seq and scRNA-seq data, we have employed Portal and evaluated Portal's capacity to integrate across data types when no cell types or only a small number of cell types are shared [7]. The findings showed that Portal preserved discrete cell types while appropriately aligning cells of shared cell. The single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) and scRNA-seq datasets were then aligned using Portal. Chromatin accessibility is measured using the epigenomic profiling technique scATAC-seq, which offers an alternative perspective to scRNA-seq. scRNA-seq and scATAC-seq integrated analysis [7]. As input, we used final input included 11853 cells from the RNA assay and 9763 cells from the ATAC assay, where 9763 cell barcodes intersected. The integrated visualization enabled it to be easier to explore multi-omics data and provided insights into cellular heterogeneity and regulatory dynamics. This is due to the two-dimensional representation of our UMAP plots and other visualizations [7]. According to Brombacher, E. et. al article, the ideal values to utilize for Portal are at lambdacos=50.0, n_latent=30, npcs=20, and hvg_num=12000. In other words, these numbers indicated the best overlap [6]. By using

a unified domain translation framework with an adversarial mechanism, Portal achieves good data alignment performance. In a latent space with no domain-specific effects, it learns representations of cells. Portal can take into consideration intricate domain-specific effects thanks to the nonlinearity of domain translation networks.

cisTopic modeling

Topic modelling of the 9,763 PMBC cells was performed by cisTopic seeking to find topics that summarize variability in ATAC-seq accessibility counts. First on our predefined set of different topics we tested which number of topics is optimal to summarize the data into. The number of topics chosen was 35 (Figure 1). For each model the log-likelihood is visualized (y-axis). The curve flattens most when the model is trained on 35 topics.

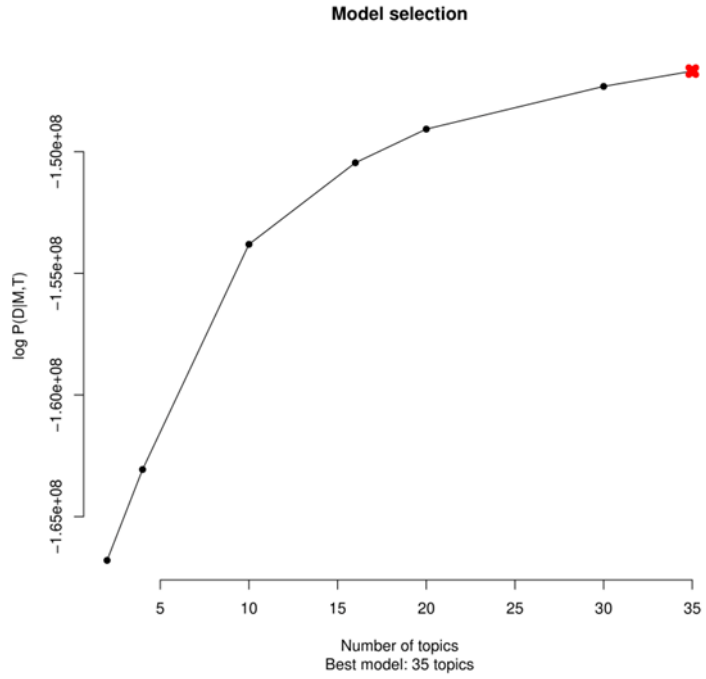

Figure 1. cisTopic model selection testing several models. *Log-likelihood is calculated and depicted for each number of topics. Flattening of the curve means that extra topics do not contain extra information of the cells. 2, 4, 10, 16, 30 and 35 number of topics were tested. The curve flattens most at 35 topics, which is marked with a red cross (top-right).*

We visualized all cells in two-dimensional space based on their contribution (probability) to that specific topic. Furthermore, we added cell type annotations to show cell type contributions for a certain topic (Figure S8). We found that different cell types (in total 6) could be described among the generated topics, e.g., B-cells (yellow colored) in topic 29 and NK-cells (light-blue colored) in topic 2. Looking more closely to the topics, we observe a gradient emerges within the monocytes (orange colored) (Figure 2).

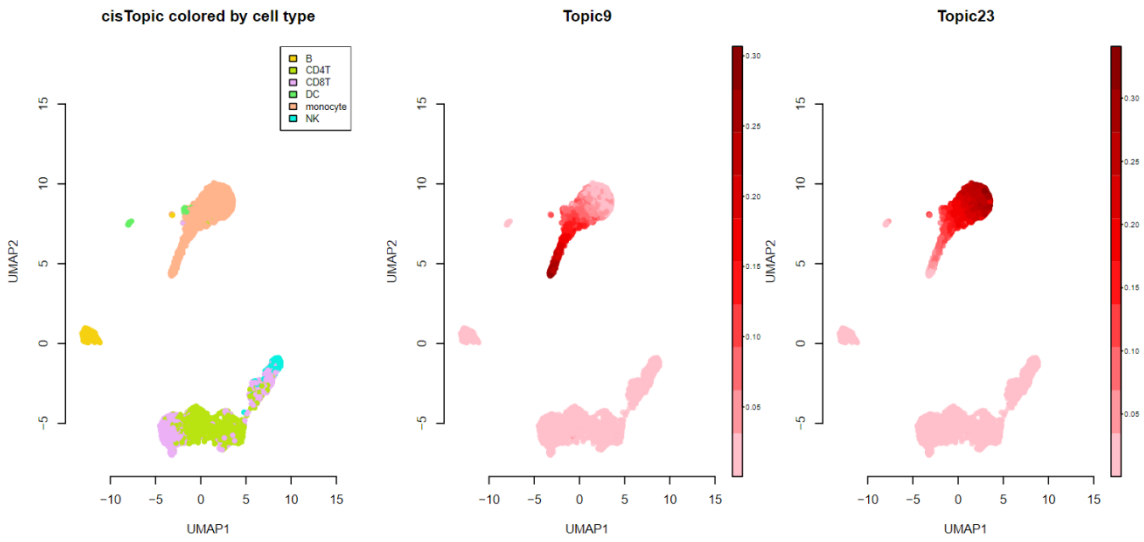

Figure 2. UMAPs of topics 9 (middle) and 23 (right) alongside the cell types (left). *Monocyte cells forming a gradient captured by topic 9 and 23. For both topic plots the calculated probability is depicted on a continuous scale. The higher er probability a cell contributes to that topic the more intense the red color becomes.*

FigR

Peak-gene associations were tested, calculating Spearman correlations across all cells between peak counts and RNA expression. Based on these results, domains of regulatory chromatin (DORCs) could be identified. To call a gene a DORC, we used a cutoff value of 3 significant genes peaks per gene to call it a DORC. In total 221 genes have been labeled as a DORC, with gene TRG-ASI having the highest number (10) of significant correlated peaks (Figure 3).

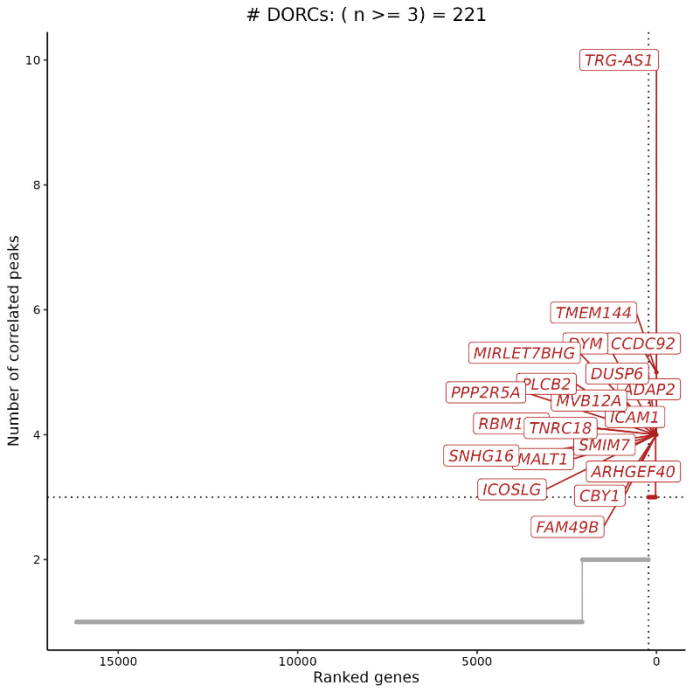

Figure 3. DORCs identified by FigR. *In total 221 DORCs are found with three or more significant peaks per gene. Most DORCs in the figure have a value of four peaks. TRG-AS1 has a interestingly higher number of significant correlated peaks of 10.*

Comparison FigR GRN reconstruction with and without cisTopic

FigR assumes that the DORC and RNA matrices are smoothed based on cisTopic. We compared GRN construction with and without the use of cisTopic. The main objective of FigR is to determine TFs that are regulators (activators or repressors) of parts of the genome labeled as DORC. Enriched DORCs for different TF binding motifs accompanied by their correlation to TF RNA expression show equal scores among both methods (Figure 4). The GRN without cisTopic contains 32 DORCs associated with 17 TFs (Figure 4A). CisTopic generated input data results in 34 DORCs associated with 21 TFs (Figure 4B). Top 10 regulatory TF motifs can show that TET3 and CXXC5 are the top regulators of DORCs within both methods of GRN construction (Table S2).

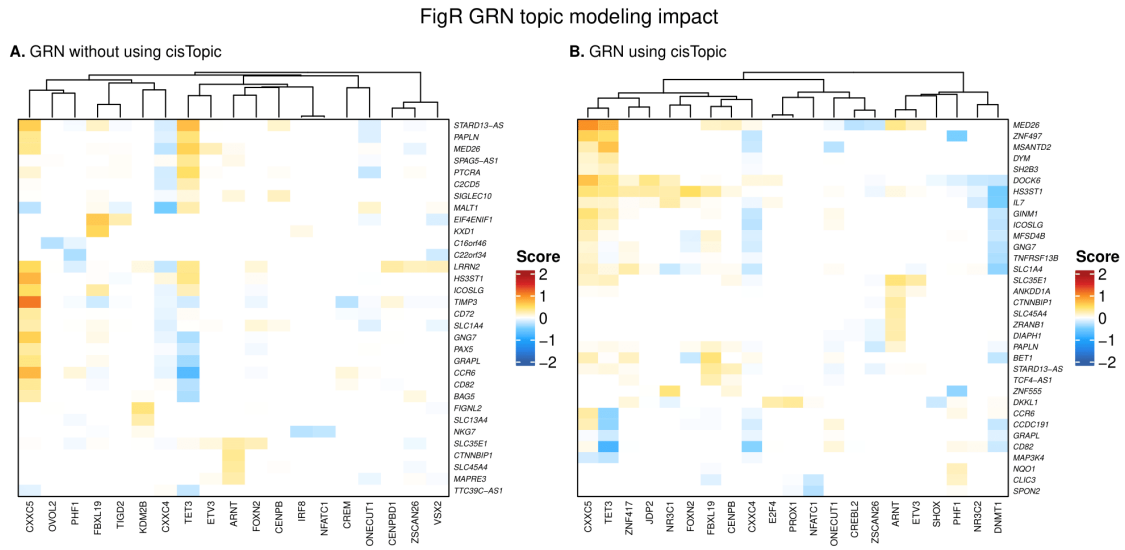

Figure 4. GRN construction with (B) and without cisTopic (A). *TF-DORC associations that meet an absolute regulation score cut-off of 0.2 (calculated by FigR). TFs are depicted on the x-axis and DORCs are found on the y-axis. (A) 32 DORCs, 17 TFs; (B) 34 DORCs, 21 TFs. Scores above zero show TFs that are activators and scores below zero depict TFs that serve as repressors.*

SCENIC+

For SCENIC+ we used annotated Seurat object with the references ‘celltype.l2’ and raw scATAC-seq data. By performing scATAC preprocessing and filtering it out with the barcodes that were received from the QC step in ArchR we have generated a cisTopic object (n_cells × n_regions = 9763 × 350257). Running topic modelling and analyzing the results concluded in optimal number of topic equal to 10.

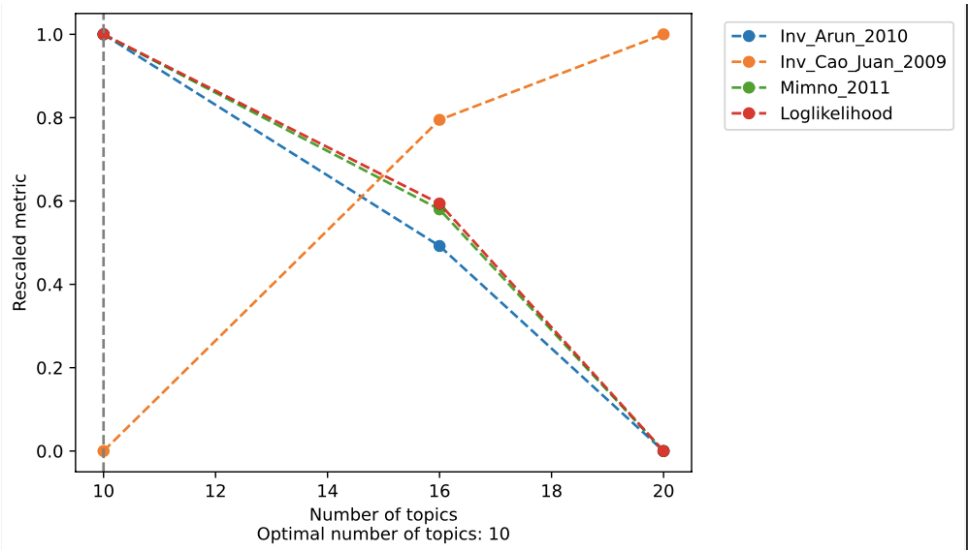

***Figure 5. cisTopic model selection testing several models for SCENIC+**. Several evaluating metrics were calculated and depicted for each number of topics.*

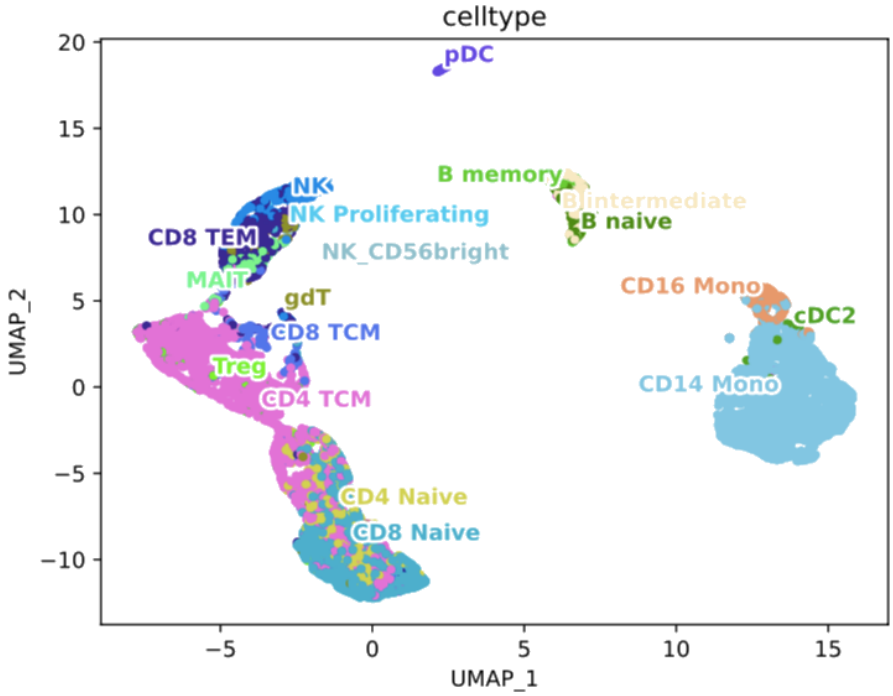

Figure 6. UMAP of SCENIC+ cisTopic object with 10 topics.

UMAPs for cell-topic probabilities were generated (Figure S9).

Discussion

In this study we proposed a method to compare tools to find improve the interpretation of genetic variation-driven gene regulation. This could contribute to more insights into disease-causing variants. During the scope of this study, we were able to show that 10X PBMC data can be used in FigR to construct GRNs. Using topic modelling we show that identification of TF-DORC associations improves.

We demonstrated that the original Azimuth-derived cell type annotations that were projected on the scRNA-seq data modality were extremely well aligned with the Portal-integration results, after integrating scRNA-seq and scATAC-seq data. The original Azimuth cell type annotations are largely reliable, so we could rely on them to understand the topics generated by cisTopic. We were unable to use Portal results directly as FigR input because Portal yield one dataset in latent space, for which we were unable to adjust the output to input for FigR. SCENIC+ could not be utilized to produce GRN results due to time limitations. However, when seeking a new method of incorporating Portal-integrated data in FigR (and potentially SCENIC+), this could potentially lead to new findings among both data modalities. Portal results have shown that the tool is able to impressively integrate scRNA-seq and scATAC-seq data. The previous version of SCENIC+, SCENIC [14], did a study where they used the same 10k PBMC dataset, but without the granulocytes filtered out [15]. This allows for a comparison between FigR and SCENIC results. SCENIC uses pyCisTopic for topic modelling. FigR found 35 topics to be the most suitable number of topics, the number of topics stabilized at 16 for pyCisTopic. This is counterintuitive, since the data used in the latter contains more cell types.

In future research pathway analysis can be included. Results shown in this study, e.g., the TET3 TF-DORC-target gene connection can be further investigated. Using topic modelling output and cell type annotation in combination with pathway analysis could provide valuable information about processes being regulated by the TF-DORC association. By using integrated data new underlying mechanisms could be identified which will have profound impact on the identification of disease-causal genetic variant identification.

Acknowledgements

We would like to thank our coordinators, Maryna Korshevniuk, Dan Kaptijn, Roy Oelen and Monique van der Wijst for guiding us through this project and sharing essential knowledge. We also would like to express gratitude to Hanze UAS for the opportunity to participate in such a project as part of our learning experience.

Author contributions

DG, M. Werff, KDB, and RG analyzed the data, interpreted the results, and drafted the manuscript. DK, MK, RO and MW contributed to the project design, and provided feedback and supervision for the analyses. All authors reviewed the manuscript. DG, M. Werff, KDB, and RG made the last changes and approved the final manuscript.

Data and code availability

Code is available at github repository, which is here, https://github.com/Roya-Gharehbeiklou/Integrated_single_cell_multiomics

References

[1] J. H. Moore, F. W. Asselbergs, and S. M. Williams, “Bioinformatics challenges for genome-wide association studies,” *Bioinformatics*, vol. 26, no. 4. 2010. doi: 10.1093/bioinformatics/btp713.

[2] K. A. Jagadeesh *et al.*, “Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics,” *Nat Genet*, vol. 54, no. 10, 2022, doi: 10.1038/s41588-022-01187-9.

[3] B. Hwang, J. H. Lee, and D. Bang, “Single-cell RNA sequencing technologies and bioinformatics pipelines,” *Experimental and Molecular Medicine*, vol. 50, no. 8. 2018. doi: 10.1038/s12276-018-0071-8.

[4] G. Li *et al.*, “A deep generative model for multi-view profiling of single-cell RNA-seq and ATAC-seq data,” *Genome Biol*, vol. 23, no. 1, pp. 1–23, Dec. 2022, doi: 10.1186/S13059-021-02595-6/FIGURES/5.

[5] S. Baek and I. Lee, “Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation,” *Computational and Structural Biotechnology Journal*, vol. 18. 2020. doi: 10.1016/j.csbj.2020.06.012.

[6] E. Brombacher, M. Hackenberg, C. Kreutz, H. Binder, and M. Treppner, “The performance of deep generative models for learning joint embeddings of single-cell multi-omics data,” *Front Mol Biosci*, vol. 9, p. 1192, Oct. 2022, doi: 10.3389/FMOLB.2022.962644/BIBTEX.

[7] J. Zhao *et al.*, “Adversarial domain translation networks for fast and accurate integration of large-scale atlas-level single-cell datasets,” *bioRxiv*, p. 2021.11.16.468892, Mar. 2022, doi: 10.1101/2021.11.16.468892.

[8] V. K. Kartha *et al.*, “Functional inference of gene regulation using single-cell multi-omics,” *Cell Genomics*, vol. 2, no. 9, p. 100166, Sep. 2022, doi: 10.1016/J.XGEN.2022.100166.

[9] C. B. González-Blas *et al.*, *SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks*, vol. 2022. 2022.

[10] Y. Hao *et al.*, “Integrated analysis of multimodal single-cell data,” *Cell*, vol. 184, no. 13, pp. 3573-3587.e29, Jun. 2021, doi: 10.1016/J.CELL.2021.04.048.

[11] J. M. Granja *et al.*, “ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis,” *Nature Genetics 2021 53:3*, vol. 53, no. 3, pp. 403–411, Feb. 2021, doi: 10.1038/s41588-021-00790-6.

[12] C. Bravo González-Blas *et al.*, “cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data,” *Nature Methods 2019 16:5*, vol. 16, no. 5, pp. 397–400, Apr. 2019, doi: 10.1038/s41592-019-0367-1.

[13] Q. Yang, Z. Xu, W. Zhou, P. Wang, Q. Jiang, and L. Juan, “An interpretable single-cell RNA sequencing data clustering method based on latent Dirichlet allocation,” *Brief Bioinform*, May 2023, doi: 10.1093/BIB/BBAD199.

[14] S. Aibar *et al.*, “SCENIC: Single-cell regulatory network inference and clustering,” *Nat Methods*, vol. 14, no. 11, p. 1083, Oct. 2017, doi: 10.1038/NMETH.4463.

[15] B. Van de Sande *et al.*, “A scalable SCENIC workflow for single-cell gene regulatory network analysis,” *Nature Protocols 2020 15:7*, vol. 15, no. 7, pp. 2247–2276, Jun. 2020, doi: 10.1038/s41596-020-0336-2.

Supplementary material

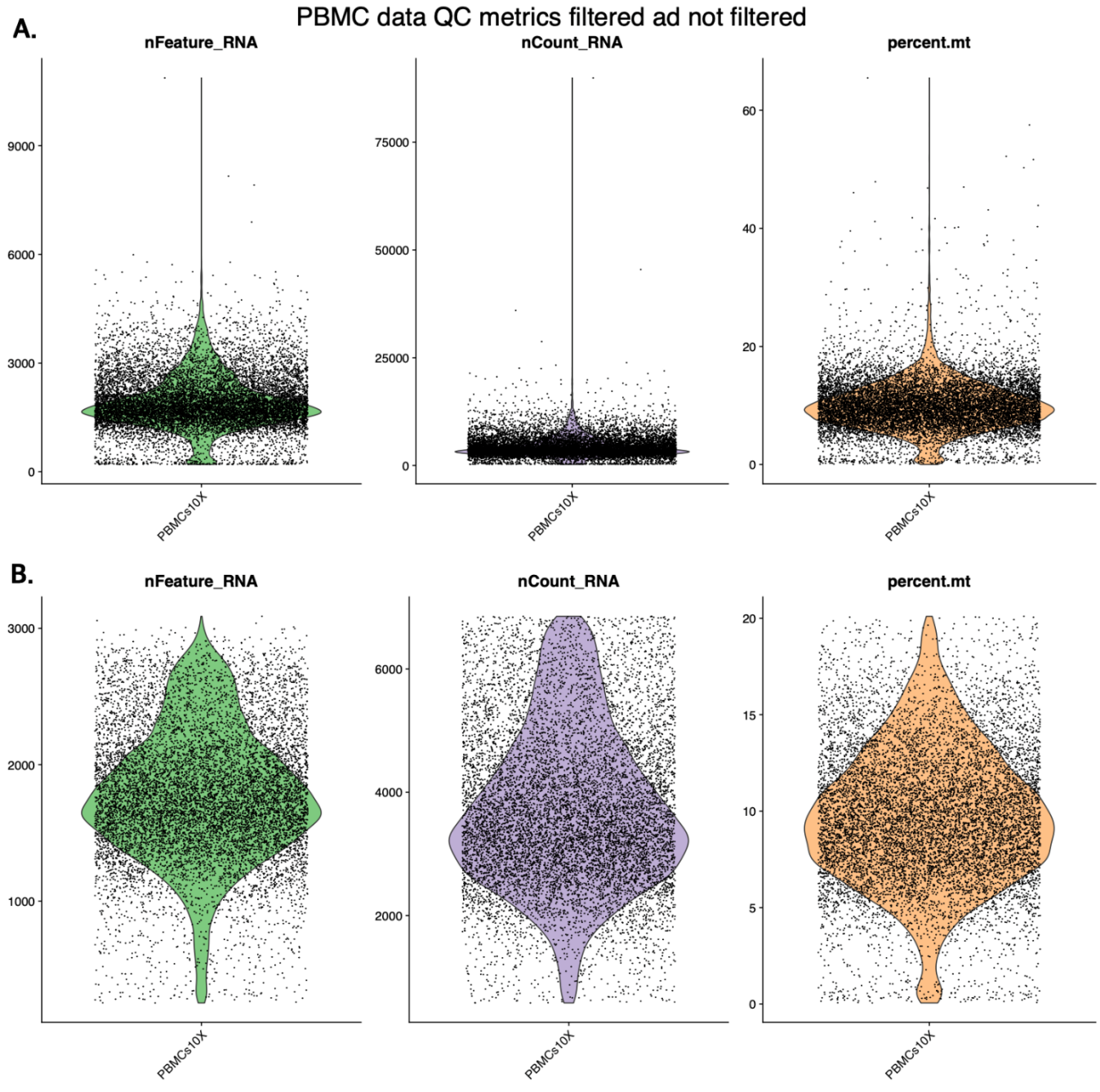

Fig.S1. Violin plot of the cells before and after applying the QC metrics in Seurat. *Panel A. Violin plots before applying QC metrics to the different features (nFeature, nCount and %mtRNA). Panel B. Violin plots after before applying QC metrics (nUMIs: =< MAD_2, %mtRNA: >= MAD_3) to the different features (nFeature, nCount and %mtRNA).*

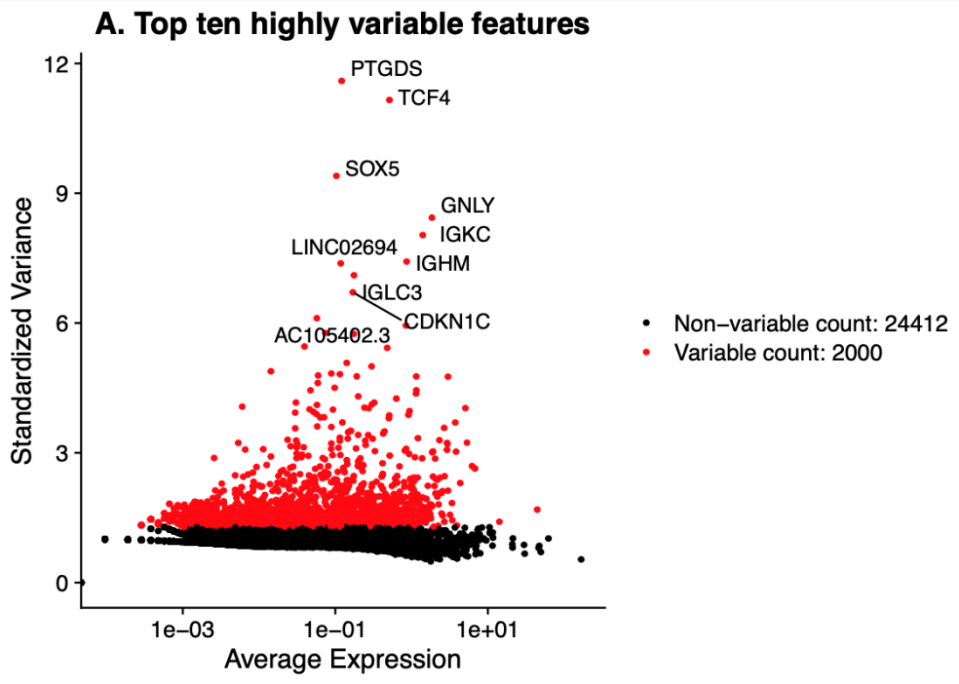

Figure S2. Top ten highly variable features in filtered 10X data scRNA-seq data. *A. Subset of genes that exhibit high cell-to-cell variation in the dataset (highly expressed in some cells, and lowly expressed in others).*

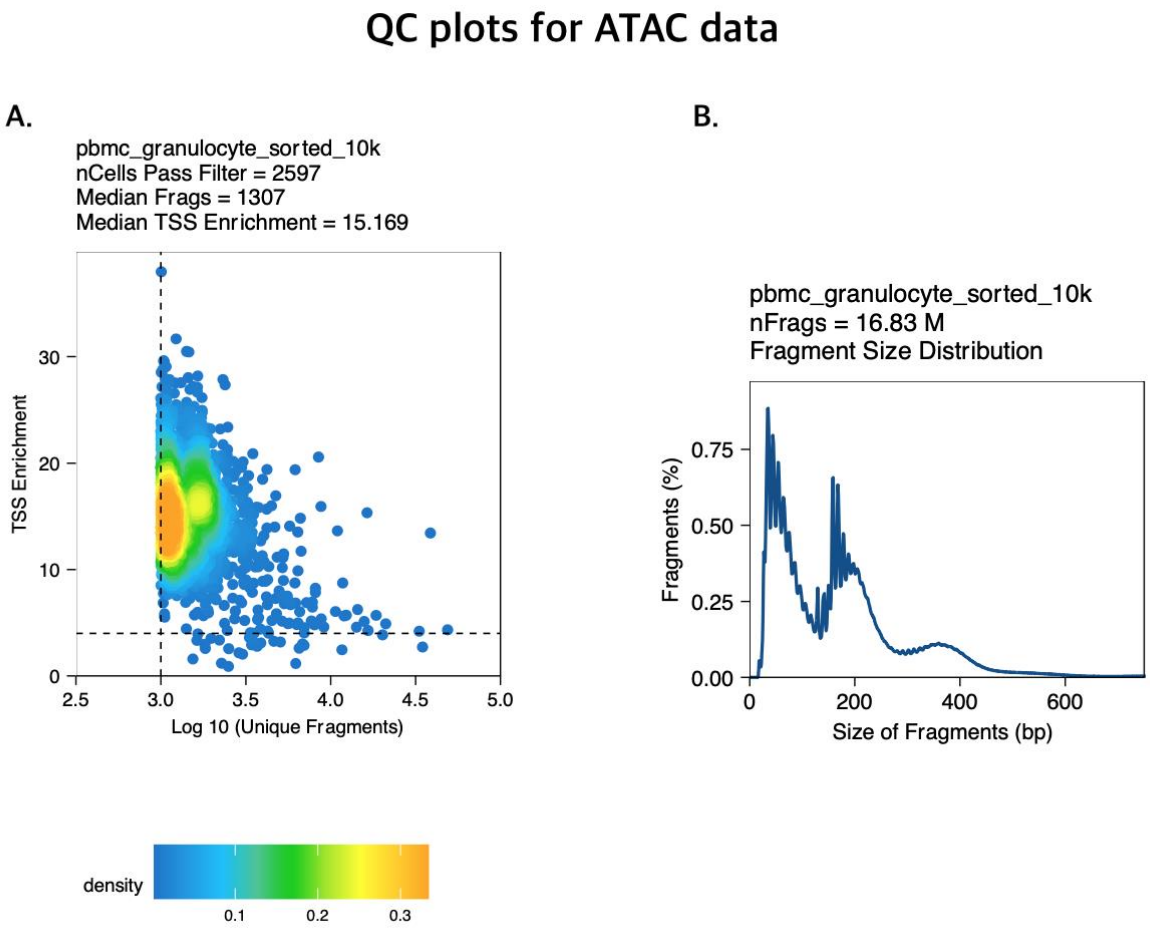

Figure S3. Quality control metrics of 10X scATAC-seq data. *A. Number of unique nuclear fragments shown as log10(Unique Fragments) vs TSS enrichment score. B. The fragment size distribution in the dataset.*

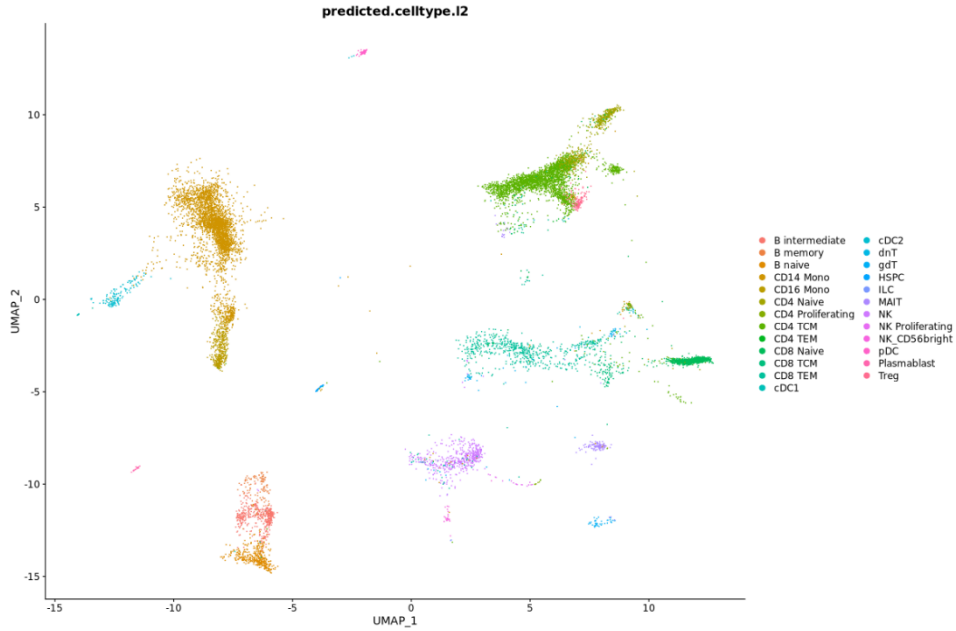

Figure S4. *UMAP of cell type annotations predicted by Azimuth*

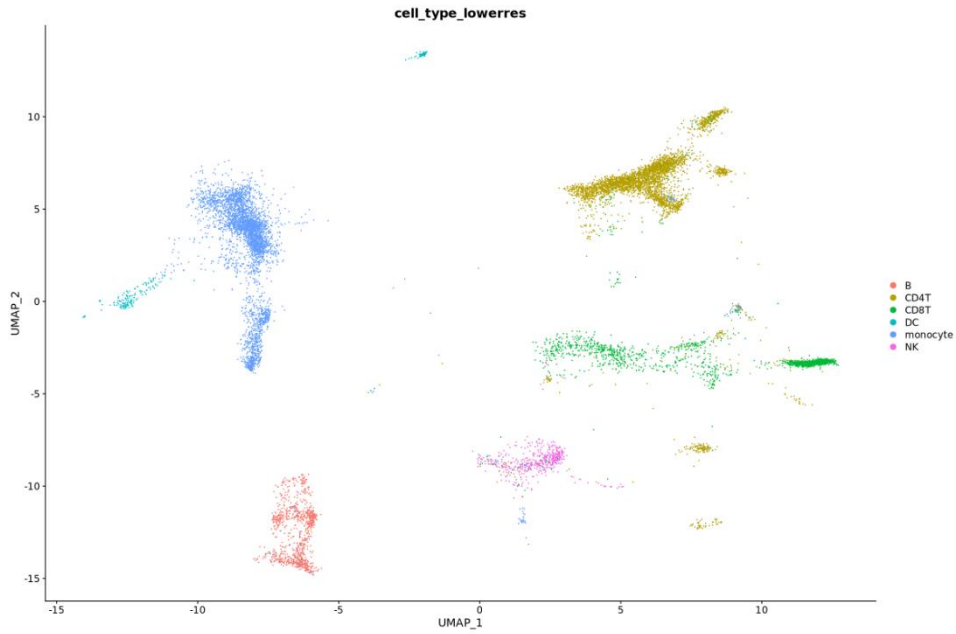

Figure S5*. UMAP cell type annotations Azimuth on reduced cell type resolution*

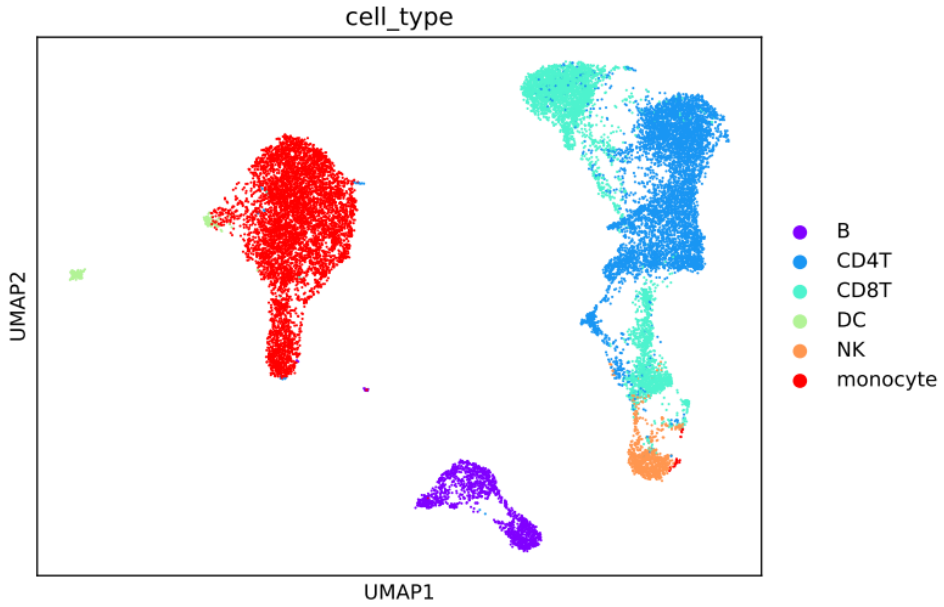

Figure S6. *UMAP plot of portal's integration in the 10X space colored by profiling methos and cell types after integration. latent cell type using Portal is illustrated here.*

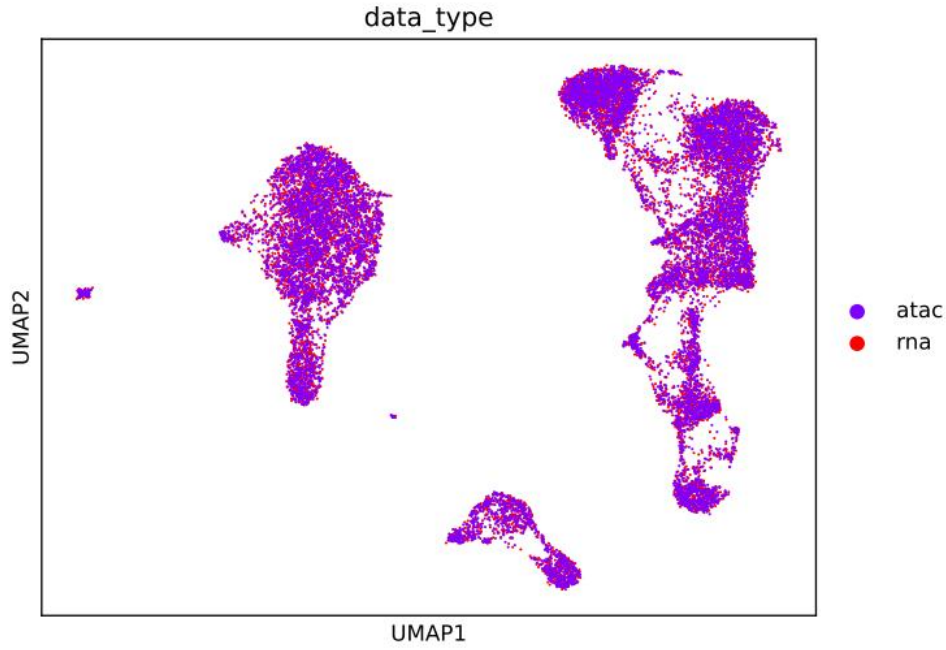

Figure S7. *UMAP plots of Portal's integration in the shared latent space for single-cell same-cell RNA/ATAC-seq integration*

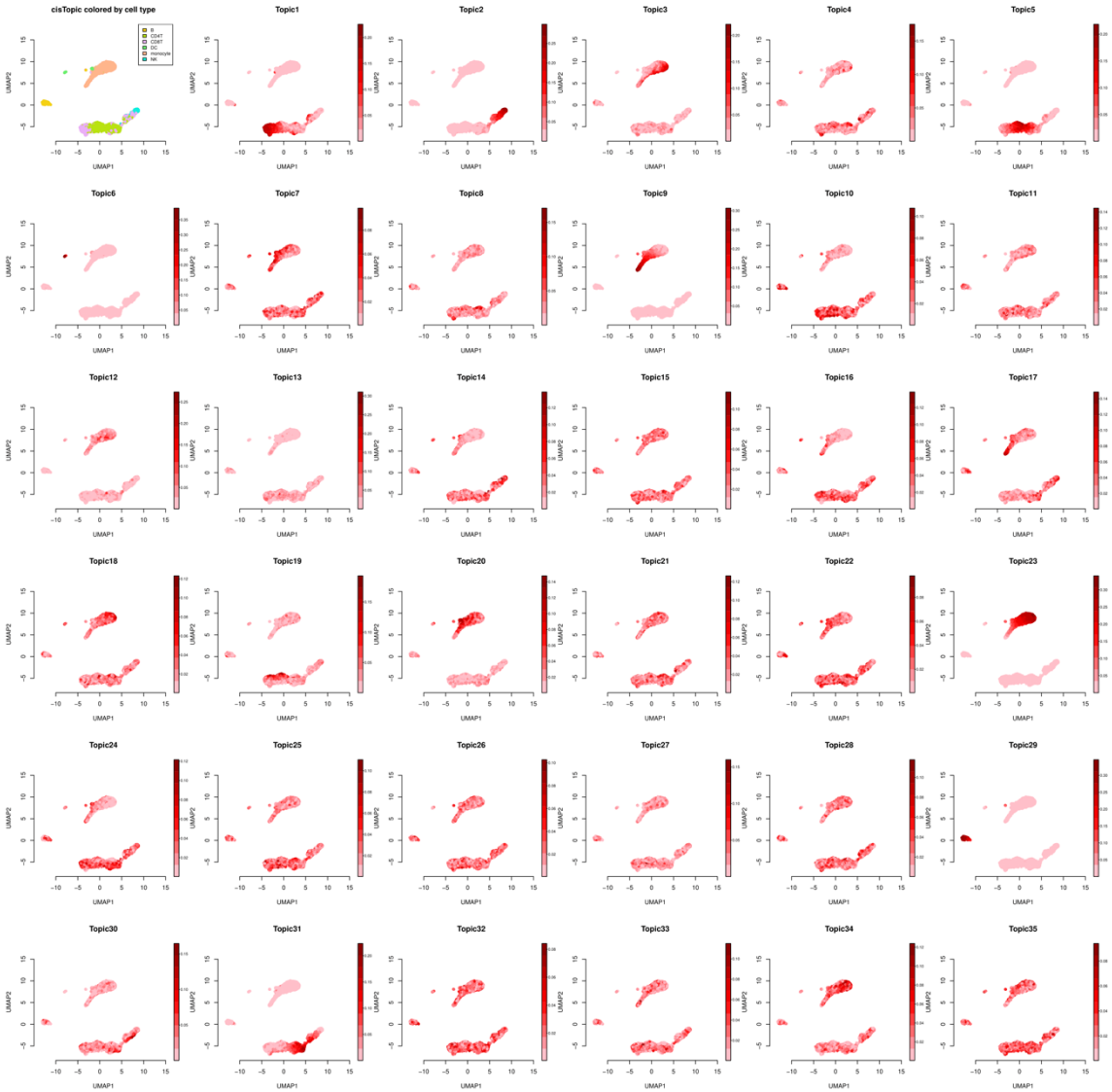

Figure S8. Topic contributions cell types. *All 35 topics represented in two-dimensional space (UMAP), accompanied by a UMAP coloured by cell type. Topic 29 contains information about B-cells, whereas topic 5 describes CD4T cell information. A gradient is visible among monocyte cells (topic 9 and 23).*

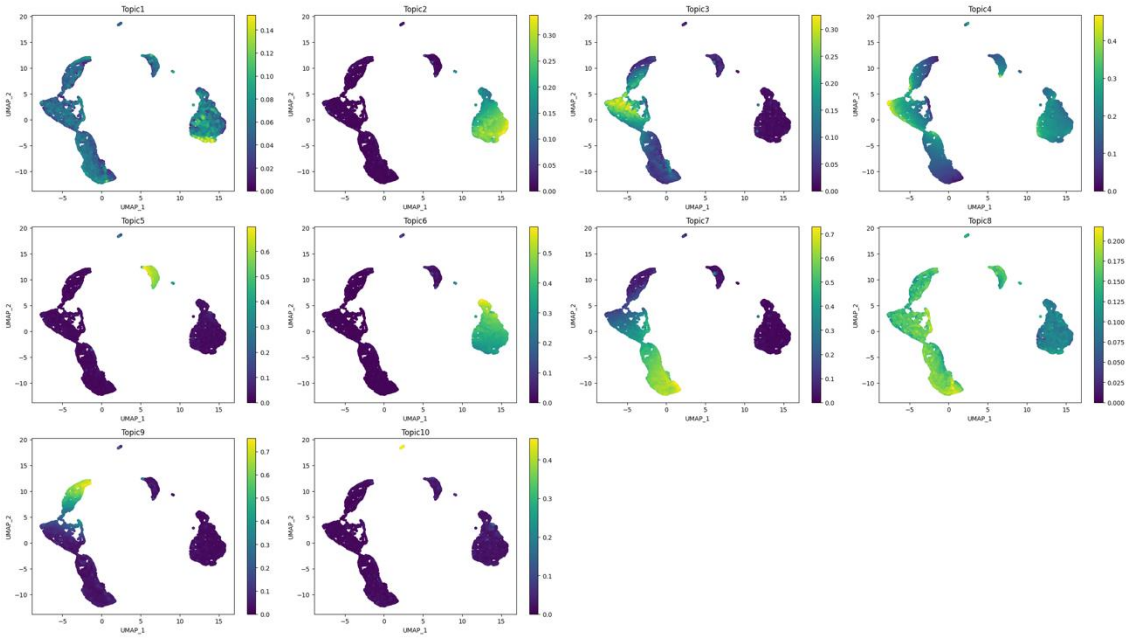

***Figure S9. Topic modelling with SCENIC+**. UMAPs of cell-topic probabilities help to interpret the cell type specificity.*

Labels in Azimuth PBMC reference dataset 'celltype.l2'	Low resolution label
Plasma	plasma
B, B intermediate, B naive, B memory	b
CD4 TCM, CD4 Naive, Treg, MAIT, gdT, CD4 CTL	cd4t
CD8+ T, CD8 TEM, CD8 Naive, CD8 TCM	cd8t
pDC, mDC, cDC2, cDC1	dc
cMonocyte, ncMonocyte, CD14 Mono, CD16 Mono, NK_CD56bright	monocyte
NK, CD56(dim) NK, CD56(bright) NK, NK Proliferating	nk
Megakaryocyte	megakaryocyte

***Table S1.** Remapping of Azimuth cell subtypes into a major cell type.*

With cisTopic			**Without cisTopic**		
DORC	***Motif***	***Score***	***DORC***	***Motif***	***Score***
MED26	CXXC5	0.975619504082458	TIMP3	CXXC5	1.12479506918549
CD82	TET3	-0.814122275719975	HS3ST1	CXXC5	0.752459745445124
MED26	TET3	0.77366790973536	CCR6	CXXC5	0.746248599670619
DOCK6	CXXC5	0.680283351287069	CCR6	TET3	-0.741355034390808
MSANTD2.	TET3	0.677887285532567	STARD13-AS	TET3	0.710289495690136
ZNF497	CXXC5	0.586939593957642	EIF4ENIF1	FBXL19	0.630955928019182
HS3ST1	DNMT1	-0.506494700056907	MALT1	CXXC4	-0.512547479207889
IL7	DNMT1	-0.498272511627022	GRAPL	TET3	-0.368040182217966
ZNF497	PHF1	-0.489697189005266	BAG5	TET3	-0.33078556684496
CD82	CXXC4	-0.456867256148232	C22orf34	PHF1	-0.324246382134412

Table S2. Top 10 DORC – TF motif regulation scores. *Positive scores means that the associated TF motif serves as an activator, where a negative value serves as a repressor of the DORC. Depicted for smoothed data (left) and unsmoothed data (right).*