SingleViper: A Novel Protein-Based Cell-Type Annotation Workflow

Nathaniel Maretzki[1] Aleksandar Obradovic[2]

[1]Columbia University, New York, NY USA
[2]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY USA

August 13, 2025

Abstract

Accurate annotation of cell types from single-cell RNA sequencing (scRNA-seq) data remains challenging under conditions of low transcript depth and high technical noise. Traditional annotation pipelines rely on transcript abundance, which poorly reflects cellular state when genes are sparsely expressed. To address this, we present SingleViper, a workflow that integrates transcriptional regulatory network inference and protein activity estimation to classify cell types based on regulatory dynamics instead of simply gene expression alone. Using ARACNe3, we infer a regulon from reference expression matrices, which is then used with the VIPER algorithm to transform query and reference gene expression data into protein activity scores. The resulting activity matrices serves as inputs for supervised classification using either SingleR or a custom machine learning classifier. To test the pipeline's robustness, we downsample the query dataset and benchmark purity, accuracy, and per-class of SingleViper against standard gene-expression–based annotation methods. We hypothesize that VIPER-based annotations will consistently outperform baseline methods in low-depth settings. These results would suggest that protein activity signatures offer a stable, biologically informed representation for robust single-cell cell type classification going forward, which could serve as an alternative to gene-expression–based classification.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has transformed our ability to resolve cellular heterogeneity, enabling the identification of distinct cell types, states, and trajectories within complex tissue samples. A foundational step in scRNA-seq analysis is cell-type annotation, which typically relies on direct comparison of raw gene expression profiles to reference datasets. While effective and usable in many settings, expression-based annotation methods suffer from well-known limitations[1]: stochastic dropout, shallow coverage, and transcriptional noise. These obscure signals for lowly expressed or sparsely sampled transcripts. These challenges are then magnified in datasets with limited sequencing depth or highly heterogeneous populations, where subtle transcriptional differences distinguish closely related cell types[2].

To address these limitations, we propose a turn to regulatory network–based approaches, which infer the activity of transcription factors (TFs) based not on their own expression, but on the collective behavior of their downstream targets. The ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks)[3] algorithm leverages information-theoretic principles to reconstruct context-specific gene regulatory networks by identifying statistically significant TF–target interactions while eliminating indirect associations through data processing inequality filtering. These networks are then used for VIPER (Virtual Inference of Protein-activity by Enriched Regulon analysis)[4], a method that infers protein activity by integrating the expression of each TF's regulon—its predicted set of downstream targets—into a normalized activity score.

In this study, we adapt this framework to the problem of single-cell annotation. We present SingleViper, a hybrid pipeline that applies VIPER-based protein activity inference to both reference and query datasets, followed by modified SingleR-style correlation-based classification on the resulting protein activity matrices. By comparing this approach to standard expression-based SingleR[5] across multiple downsampling conditions and label sets, we assess the potential of protein-level annotation to improve robustness in low-depth or noisy single-cell experiments, such as single-nucleus data. Our results suggest that integrating regulatory network inference into the annotation pipeline can substantially enhance classification accuracy, particularly for transcriptionally similar immune cell subtypes.

2 Methods

2.1 Datasets

We evaluated our approach using publicly available single-cell RNA-seq datasets with well-annotated immune cell type (CITE-seq) data. As reference data, we used the Blueprint+ENCODE human hematopoietic atlas provided via the `celldex` R package, consisting of log-normalized gene expression profiles and curated main-label annotations across a range of blood-derived immune populations. As query data, we used a subset of the 10x Genomics CITE-seq PBMC dataset, which contains UMI counts and antibody-derived tag (ADT) labels for thousands of peripheral blood mononuclear cells.

To simulate low-coverage conditions, we generated downsampled query datasets across a range of UMI retentions using the `DropletUtils::downsampleMatrix()` function on raw UMI counts. All expression matrices were converted to counts per million, log-transformed, and Z-score normalized per gene prior to downstream use.

2.2 Regulatory Network Inference

We constructed transcriptional regulatory networks using the ARACNe3 algorithm. For each downsampled query dataset, we inferred cell-type–specific regulons by running ARACNe3 on Z-score normalized gene expression metacell matrices using a curated list of human transcription factors (TFs). Networks were pruned using a mutual information threshold and data processing inequality filtering, retaining only statistically significant TF–target edges. We excluded regulons with zero or invalid targets and filtered targets to those shared with the reference expression matrix to ensure compatibility across datasets. Each regulon was represented as a two-component list: a named vector of signed interaction weights (tfmode) and corresponding confidence scores (likelihood), normalized to unit scale when missing or inconsistent.

2.3 Protein Activity Inference

To infer transcription factor activity, we used the VIPER algorithm. VIPER evaluates the enrichment of each TF's regulon in a given gene expression profile using a rank-based, normalized enrichment-like scoring function. In this context, VIPER was applied to both the query and reference datasets using regulons generated on the query data, yielding a protein activity matrix where rows correspond to TFs and columns to cells or samples. This transformation shifts the feature space from direct gene expression to inferred regulatory activity, which has been shown to be more robust to dropout and transcriptional noise. Only transcription factors with at least one valid target were retained for scoring. All VIPER runs used consistent regulon lists and normalization parameters across reference and query to ensure comparability.

2.4 Cell-Type Annotation

We wrote a simplified version of the SingleR framework for cell-type classification that accepted its normal expression-based correlation input with the VIPER-derived protein activity matrices. In this modified setup, the test dataset consisted of per-cell TF activity vectors (from the downsampled query data), while the reference consisted of per-label averaged activity vectors (from the Blueprint data). Each query cell was annotated using a correlation-based similarity metric (Pearson), assigning the most similar reference label as the predicted cell type. To benchmark our method, we ran this SingleViper annotation pipeline in parallel with standard SingleR applied directly to the log-normalized gene expression matrices. Annotation performance was evaluated using both overall purity (fraction of correctly predicted cells per predicted cluster) and per-label accuracy (fraction of correct predictions within each true label class), computed across all downsampling conditions.

3 Results

At the time of writing, we were in the final stages of benchmarking the SingleViper pipeline against traditional expression-based annotation methods when an unexpected hardware failure on my local machine halted progress. The final computational run—comparing SingleR annotations derived from gene expression versus VIPER-inferred protein activity on downsampled PBMC query datasets—was actively executing when the system failure occurred. Unfortunately, this interrupted both the analysis and downstream visualizations, leaving the comparative results incomplete for the time being. Despite this setback, the full pipeline was implemented and validated across all components: ARACNe-based regulon construction, VIPER protein activity inference on reference and query datasets, and adapted SingleR classification on the resulting matrices. Preliminary quality checks confirmed that the regulons were successfully filtered, protein activity matrices generated without error, and cell-type predictions produced under both frameworks (SingleR and SingleViper). While the final quantitative results are pending, the setup was designed to assess annotation robustness across multiple levels of downsampling (from 10k to 500 UMI reads), using accuracy and purity metrics to evaluate classification performance across known immune subtypes. Prior literature and the theoretical basis for protein activity inference suggest that the SingleViper approach should outperform standard SingleR under conditions of reduced gene coverage, particularly for cell types where transcription factor expression is low but regulatory influence remains active. In future work, the final benchmarking will be completed with minimal effort given the modular structure of the implementation. Given the promise of network-based methods in handling noisy single-cell data, the SingleViper framework presents a scalable and reproducible approach for protein-level annotation across challenging datasets. Future work will also involve replacing the Pearson correlation used in our version of SingleR with a modular extension that supports the use of supervised machine learning classifiers for cell-type prediction. Specifically, we plan to replace the similarity scoring step with model-based label assignment using three classifiers: k-nearest neighbors (k-NN), support vector machines (SVM) with a radial basis function kernel, and random forests. These models will be trained on the reference matrix of VIPER-inferred transcription factor activity, using the known Blueprint labels as training targets. Predictions will be made on the VIPER-transformed query data at each downsampling level, allowing direct comparison to correlation-based SingleR and SingleViper. This approach enables evaluation of whether nonlinear or ensemble-based classifiers improve cell-type assignment in protein activity space, particularly under conditions of limited expression information.

4 Conclusion

This project introduced SingleViper, a protein activity–based pipeline for single-cell annotation that integrates ARACNe-inferred regulatory networks, VIPER activity inference, and a modified SingleR classification framework. While final benchmarking results were interrupted by hardware failure, the pipeline was fully developed, validated on real data, and structured to support reproducible comparisons between expression and activity-based annotation methods under various UMI depth conditions. Beyond benchmarking accuracy, one of the most promising implications of this work lies in its potential application to single-nucleus RNA sequencing (snRNA-seq) of cryopreserved or archival samples. Such samples—often obtained from biobanks, frozen tumors, or longitudinal clinical studies—cannot be processed immediately after extraction, and frequently suffer from degraded mRNA quality or transcript dropout. By leveraging inferred protein activity from VIPER, which integrates regulatory context rather than relying on direct gene expression, the SingleViper approach offers a pathway toward robust annotation even in low-quality or sparse data environments. This could enable high-fidelity cell-type identification in snRNA-seq datasets without requiring immediate processing or on-site analytical infrastructure—facilitating remote sample collection, long-term storage, and retrospective analysis in both clinical and research settings. Future work will explore integrating supervised learning approaches to further improve performance across diverse tissue types and conditions.

References

[1] Liu et al. Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. *PLOS ONE*, 8:1–10, 2013.

[2] Vogel & Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13:227–232, 2012.

[3] Basso et al. Reverse engineering of regulatory networks in human b cells. *Nature Genetics*, 37:382–390, 2005.

[4] Alvarez et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics*, 48:838–847, 2016.

[5] Aran et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20:163–172, 2019.

[6] Ding et al. Quantitative assessment of protein activity in orphan tissues and single cells using the metaviper algorithm. *Nature Communications*, 9:1471, 2018.