PISCES: A pipeline for the Systematic, Protein Activity-based Analysis of Single Cell RNA Sequencing Data

Aleksandar Obradovic*, Lukas Vlahos*, Pasquale Laise,

Jeremy Worley, Xiangtian Tan, Alec Wang, Andrea Califano

Columbia University Department of Systems Biology, Herbert Irving Comprehensive Cancer Center

*these authors contributed equally

Abstract

While single-cell RNA sequencing provides a new window on physiologic and pathologic tissue biology and heterogeneity, it suffers from low signal-to-noise ratio and a high dropout rate at the individual gene level, thus challenging quantitative analyses. To address this problem, we introduce PISCES (**P**rotein-activity **I**nference for **S**ingle **Ce**ll **S**tudies), an integrated analytical framework for the protein activity-based analysis of single cell subpopulations. PISCES leverages the assembly of lineage-specific gene regulatory networks, to accurately measure activity of each protein based on the expression its transcriptional targets (regulon), using the ARACNe and metaVIPER algorithms, respectively. It implements novel analytical and visualization functions, including activity-based cluster analysis, identification of cell state repertoires, and elucidation of master regulators of cell state and cell state transitions, with full interoperability with Seurat's single-cell data format. Accuracy and reproducibility assessment, via technical and biological validation assays and by assessing concordance with antibody and CITE-Seq-based measurements, show dramatic improvement in the ability to identify rare subpopulations and to assess activity of key lineage markers, compared to gene expression analysis.

Introduction

High-throughput, droplet-based single-cell RNA Sequencing (scRNASeq) has recently emerged as a valuable tool to elucidate the diverse repertoire of cellular subpopulations comprising a broad range of mammalian tissues. Applications of this technology range from study of tissue development (He et. al., 2020) and tumor micro-environment (Qian et. al., 2020), to the elucidation of tissue heterogeneity (Zhao et. al., 2020) and even of tissue-level response to infectious diseases, such as COVID-19 (Xu et. al., 2020, Speranza et. al. 2021). More specifically, scRNASeq data allows identification of representative gene expression signatures for thousands of individual cells dissociated from a tissue sample (Zheng et al., 2017; Finak, et al., 2017), thus providing fine-grain characterization of the transcriptional state of individual cell types contributing to the emergence of complex phenotypes, which would be impossible from bulk profiles. This can help elucidate the role of rare populations, for instance, whose gene expression signature would be diluted below detection limits in bulk samples (Stuart et al., 2019). Moreover, in contrast to flow cytometry or CyTOF, scRNASeq generates genome-wide single cell profiles, without requiring *a priori* selection of a limited number of antibody-based markers. The value of scRNASeq in tumor biology has been broadly demonstrated in recent studies of melanoma (Sade-Feldman et al., 2018; Jerby-Arnon, et al., 2018), pancreatic cancer (Elayda et al., 2019), breast cancer (Chung et al., 2017), and renal cell carcinoma (Obradovic et. al., 2021).

The key drawback of scRNAseq technologies is that the total number of mRNA molecules per cell, combined with low capture efficiency, fundamentally limits the number of distinct mRNA molecules that can be detected in each single cell (UMI reads). As a result, scRNASeq profiles are extremely sparse, with as many as 90% of all genes producing no reads in any given cell and the majority of detected genes producing one or two reads. This phenomenon, commonly known as gene dropout, greatly hinders downstream analysis, making quantitative assessment

of differential gene expression extremely challenging. For instance, while broadly different cell types can be classified, a majority of biologically relevant genes, including the established lineage markers of specific cellular subpopulations, are undetected. As a result, cellular subpopulations presenting more subtle differences, such as different fibroblast or macrophage subpopulations, may be impossible to differentiate (Elyada et al., 2019, Obradovic et al. 2021). Even with cutting edge analysis tools such as the Seurat analysis pipeline (Butler et al., 2018), which can often identify individual subpopulations, scRNAseq gene expression data remains limited in its ability to elucidate fine-grain biological mechanisms due to its sparseness. Additionally, interrogation of individual genes of interest across cells is significantly impaired, particularly for transcription factors and signaling molecules, which do not need to be abundantly transcribed in order to fundamentally drive cell phenotype through their downstream effects on transcriptional state.

To address these limitations, we have shown that network-based analysis of protein activity, using the VIPER and metaVIPER algorithms (Ding et. al., 2018; Obradovic et al., 2021), can provide accurate, quantitative assessment for >6,000 proteins, including transcription factors, co-factors, chromatin remodeling enzymes, and signaling proteins. Moreover, we have shown that protein activity-based analysis can help identify rare subpopulations that are responsible for the presentation of key macroscopic phenotypes, ranging from immune evasion (Thorsson et. al., 2018) to relapse following surgery (Obradovic et. al., 2021). It can also help identify master regulator proteins representing mechanistic, causal determinants of cell state and cell state transitions, such as to de-differentiation to a pluripotent stem cell state (Kushwaha, 2015) or transdifferentiation between distinct tumor cell states (Laise et al., 2021). However, these analyses can be extremely complex because they require assembly of lineage specific regulatory networks and master regulator analyses that are challenging for biologists who are not trained in network biology.

To allow broad access to these methodologies to biologists with relatively limited network-based analyses expertise, we introduce a comprehensive pipeline for **P**rotein Activity **I**nference for **S**ingle **Ce**ll **S**tudies **(PISCES)**, which is made available to the research community via a general-use R package. The pipeline automates the optimal generation of lineage specific regulatory networks, via ARACNe (Algorithm for Reconstruction of Accurate Cellular Networks) (Basso et al., 2005; Lachmann et. al, 2017), measurement of protein activity via VIPER (Virtual Inference of Protein Activity by Enriched Regulon Analysis) (Alvarez et al, 2016), as well as the identification of molecularly distinct subpopulations via a variety of clustering methodologies, and the identification of Master Regulators of cell state and cell state transitions (**Figure 1A**).

ARACNe is an information theoretic algorithm for the inference of the direct transcriptional targets of transcriptional regulator proteins, as well as the least indirect targets of signal transduction proteins. This allows reconstructing the tissue specific repertoire of transcriptional targets (regulon) of ~6,500 regulatory and signaling proteins, including surface markers (SMs). VIPER computes the activity of each protein based on the differential expression of the genes in its regulon, as assessed by weighted gene set enrichment analysis. Since regulons are generally large, containing up to several hundred genes, we prune them to include the same number of the most likely targets (between 50 and 100), to avoid biasing the statistical significance of the gene set enrichment analysis, as discussed in (Alvarez et al., 2016). As a result, even when the specific gene encoding for a protein of interest is undetected, VIPER can still quantitatively assess its activity (**Figure 1B**).

Previous work in the Califano lab has shown the accuracy and reproducibility of these algorithms when used to analyze bulk data. Indeed, ARACNe and VIPER have been used extensively to identify master regulators (MRs) that were experimentally validated as mechanistic determinants of diverse biological states, many of which have been extensively validated, see (Rajbhandari et. al., 2018; Carro et al., 2010, Aytes et al., 2014; Alvarez et al,

2018), just to cite a few, and resulting in two CLIA-approved clinical tests to predict tumor drug sensitivity, including OncoTreat (Alvarez et. al., 2018) and OncoTarget (Zeleke et. al. 2020). Most critically, when comparing 30M read RNASeq profiles to down-sampled profiles with 10K to 50K reads (similar to typical scRNASeq profiles), VIPER-measured protein activity profiles retain high Spearman correlation ($\rho \geq 0.8$), while correlation of the raw gene expression profiles is extremely poor ($\rho \leq 0.3$) (Alvarez et al., 2016).

To adapt these tools to the analysis of scRNASeq profiles, PISCES implements three major modifications. First, an initial gene expression-based cluster analysis is used to identify molecularly distinct cellular subpopulations representing distinct sub-lineages. Fine grain cluster analysis is not necessary as we have shown that regulatory networks for closely lineage-related cells are virtually indistinguishable (Mani et al, 2010). ARACNe is then used to generate distinct regulatory networks for each cluster containing at least N = 500 cells. Second, to increase regulon coverage, cells within each sub-lineage-related cluster are combined into “meta-cells” using a K-nearest-neighbors graph analysis. This creates pseudo-bulk samples that can then be analyzed by ARACNe, producing networks with more accurate edges, larger regulons, and greater coverage of regulatory proteins. Finally, rather than using VIPER for protein activity measurement, we use its derivative metaVIPER (Ding et al., 2018), which is designed to optimally integrate protein activity inferences from multiple networks. This allows for the use of multiple single-cell and, when available, lineage-matched bulk-tissue-derived networks. Downstream of the ARACNe and metaVIPER analyses, PISCES provides access to a variety of novel protein-activity based clustering and data visualization algorithms, in addition to implementing interoperability with the popular Seurat single-cell data format.

In order to establish the efficacy of these tools and optimal parameters for future benchmarking and improvement, we have performed both technical and biological validation experiments, first by evaluating reproducibility of protein activity assessment from progressively downsampled

data, and then by assessing concordance of gene expression and protein activity to antibody-based measurements using multiplexed FACS (Cytek) and CITE-Seq (Stoeckius et. al., 2017).

Taken together, the results of these benchmarks show that the PISCES analytical pipeline dramatically outperforms gene expression-based analyses and even outperforms experimental assessment via selected antibodies, while allowing essentially proteome-wide activity quantitation. As such, these data suggest that PISCES provides a valuable and highly flexible tool for the analysis of scRNA-Seq datasets, which greatly improves the granularity of cell subpopulation detection, allowing detection of rare yet biologically relevant subpopulations that would be missed by gene expression analysis, due to gene dropout issues, and supports accurate assessment of Master Regulators of single-cell states.

Results

Analytical Pipeline Overview: The PISCES pipeline takes a single-cell Unique Molecular Identifier (UMI) count matrix as input, with genes organized by row and cells by column. Initial Quality Control filtering is adjustable, with user-defined parameters. By default it will remove cells with fewer than 1,000 UMIs or more than 25% mitochondrial gene UMIs. The gene expression matrix is then normalized and scaled to generate a matrix of gene expression signatures. By default, this is accomplished by converting counts to $Log_{10}(CPM + 1)$, where CPM indicates counts per million. However, it can also be implemented via the Seurat SCTransform algorithm (Stuart et al., Apr. 2018 Cell) or any other third-party methods of choice.

Following normalization, a first-pass clustering is performed on scaled gene expression using one of several clustering approaches implemented in the pipeline, including partition around medioids (PAM) (Teschendorf et. al., 2017) or Louvain clustering with resolution-optimization (Obradovic et. al., 2021). For each gene expression cluster with ≥ 500 cells, by default,

metaCells are computed by first selecting 250 unique cells at random and then transforming their scRNASeq profile into a metaCell by adding the UMI counts from the *k* nearest neighbors (*k* = 5 by default). Independent ARACNe networks are generated from each cluster using the $Log_{10}(CPM + 1)$ values of each metaCell.

In parallel, the normalized gene expression profile is transformed into a gene expression signature (GES). This can be done in a number of ways, either with an internal normalization against mean and standard deviation of all cells to query differences within the dataset or with an external reference to answer experiment-specific questions (i.e. the differences between cancerous and healthy cells). By default, PISCES will perform a standard internal normalization to generate the gene expression signature, which is then transformed into a matrix of protein activity using MetaVIPER. MetaVIPER takes as input the GES and the previously generated cluster-specific networks and identifies the best network matches to each sample by maximum regulon consensus. Enrichment scores from each matched network are then integrated using a weighted-average to produce a final enrichment value that can then be used for downstream visualization and analysis. The entire pipeline is visualized in Figure 1A.

Since every scRNAseq experiment is unique—depending on the specific cell types, the quality of the data, or the overarching question driving the research—PISCES allows users to fine tune the pipeline to match their specific requirements. For instance, since Seurat represents a widely used platform for scRNAseq analysis at the gene expression level, the Seurat batch-correction and SCTransform data scaling approach are incorporated as optional pre-processing steps to generate gene expression signatures before they are analyzed by PISCES. These may, however, be substituted by any user defined normalization and data scaling routine, such that effect of alternative normalization or pre-processing methods may be tested using PISCES's default technical and biological benchmarks. Output from the PISCES pipeline is converted to a Seurat object for convenient export into a variety of external visualization or processing tools,

and analyzed by other commonly used tools. In particular, cell type annotation is implemented in PISCES at the single-cell level using SingleR (Looney et. al., 2019), which infers cell types represented in the dataset by correlation of gene expression to expression of sorted bulk-RNASeq reference datasets and stores these labels as metadata for downstream analysis.

Technical Validation Shows Improved Recovery of Data Structure From Low-Depth Cells: To benchmark PISCES reproducibility relative to gene expression and to establish an optimal UMI depth for user-driven adjustment of metacell parameters, we executed the entire pipeline using progressively down-sampled profiles from relatively high-depth scRNAseq data. For this purpose, we used the SNU-16 cell line, a relatively homogenous stomach adenocarcinoma model that is transcriptionally complex and produces high UMI counts per cell (i.e., 40,000-50,000), on the high end of the typical yield for cell lines and significantly above the yield produced by clinical samples. Average UMI count in our dataset was 41,915 across 6157 single cells (**Figure S2A**). To create synthetic data with lower depth, we down-sampled this data by first drawing each cell's total UMI-count from a multinomial distribution with mean target depth manually specified and a uniform probability weight over all cells, then drawing the gene-specific counts from a second multinomial with probabilities given by the proportions of genes in the original, full depth profile for each cell. This procedure was applied with target depths between one and ten thousand UMIs at a step-size of 1,000 and between 10,000 and 40,000 UMIs at a step-size of 5,000. We then generated meta-cells using a consistent sub-set of 500 cells for each down-sampled matrix with depth of 10,000 UMIs or fewer. These data were used to generate 27 ARACNe networks in total; one for the full data, 16 from each of the down-sampled gene expression profiles, and 10 from each of the meta-cell matrices.

To generate gene expression signatures, we normalized each down-sampled matrix against the Cancer Cell Line Encyclopedia (CCLE) from The Broad. Because this data is from bulk-

sequencing, we first had to apply the previously described down-sampling scheme in order to generate depth-matched reference samples for each single-cell matrix. Gene expression profiles were then normalized gene-by-gene by subtracting the mean expression from CCLE, then dividing by the standard deviation of the expression in CCLE.

Finally, we generated VIPER matrices for all pairwise combinations of GES and regulatory networks, culminating in 459 VIPER matrices. A flowchart illustrating this experimental design is shown in Figure S5. To assess the reproducibility of gene expression and protein activity signatures at different depths, we computed the cell-by-cell Pearson correlation between each down-sampled matrix and the full depth data. In each cell, we subset the comparison to those genes or proteins with significantly different expression or activity (p-value < 0.05 with Bonferroni correction) in the full-depth data, then computed the correlation coefficients cell-by-cell between full-depth and down-sampled data using this subset. This reduction was performed in order to avoid inflation of correlation values based on non-significant data. In protein activity signatures generated fully from down-sampled data (down-sampled GEP as input to ARACNe, down-sampled GES as input to VIPER), we observe a statistically significant improvement in correlation to full-depth data relative to gene expression signature at all depths above 5,000 UMIs (Figure 2A; p-value < 0.05 by Wilcoxon signed rank test). Strikingly, when an ARACNe network generated from full-depth GEP is applied to down-sampled GES as input to VIPER, correlation to original full—depth VIPER signature is strongly conserved even at extremely low UMI counts, remaining above 0.75 on average at UMI depth of 1000, where average correlation of gene expression signature to full-depth data is below 0.1. This emphasizes the importance of constructing a high-quality ARACNe network in the VIPER inference pipeline, such that applying high-quality networks inferred for a given cell type from one dataset to a matched cell type in lower-quality data is likely to provide a significant boost to the power of protein activity inference even from very-low-depth data. Additionally, we find a significant improvement in correlation

values when constructing metaCell-based ARACNe networks from lower-depth data (Figure 2B), such that metaCell networks applied to run VIPER on GES matrices with mean UMI count of 3000 approach the inference quality seen when running ARACNe and VIPER on gene expression matrices with a mean UMI count of 20,000. However, at the very low mean depth of 1000 UMI/cell this breaks down, and metaCell ARACNe network inference no longer offers any statistically significant improvement over inference on low-depth data. Therefore, we strongly recommend applying the metaCell ARACNe network inference option in PISCES for any datasets with data quality between 1000 and 5000 mean UMIs/cell, which is common in clinical datasets.

Overall, these data show that the correlation between full-depth and down-sampled gene expression signatures is poor even at relatively high depth, and decays rapidly to a median value of less than 0.25 even at depths of 10,000 UMIs/cell (purple bars, Figure 2A). Protein activity, by comparison, is much more robust, significantly outperforming gene expression at all depths above 5,000 UMIs/cell. Interestingly, down-sampling only the gene expression signature input to VIPER while retaining a full-depth ARACNe network had little effect (red bars, Figure 2A) on protein activities robustness, while down-sampling either the data using to generate ARACNe networks or both ARACNe data and gene expression signature (green and blue bars respectively) had a much more significant effect on correlation to original full-depth VIPER matrix, which was partially rescued by metaCell ARACNe. The full heatmap showing mean correlation across cells comparing all VIPER matrices against full depth data is available in the supplement (Figure S3). These findings indicate that the quality of the ARACNe networks is the driving force behind protein activity signatures' ability to retain signal at low UMI depths and supports the idea of using metacells to rescue signal within the ARACNe network or use context-appropriate bulk networks where available.

Biological Validation Shows Improved Concordance with Antibody Profiling: To assess whether protein activities measured by PISCES effectively track with direct assessment of protein abundance in single cells, thus providing improved mechanistic understanding of single cell processes, we compared PISCES-measured protein activity to CITE-Seq single-cell measurements of protein abundance in a publicly available dataset of cord blood-derived mononuclear cells (CBMCs) (Stoeckius et. al., 2017).

Single cell clustering based on CITE-Seq measurements, using a pre-selected antibody panel, yields six major cell type clusters, including CD4 T-cells, CD8 T-cells, Monocytes, NK Cells, B-cells, and Hematopoietic Stem Cells (HSCs) (**Figure 3C**). In sharp contrast, gene expression-based clustering by Seurat identified only four distinct cell clusters, with NK cells and HSCs subsumed into the other major cell types (**Figure S1A**). Protein activity-based clustering by PISCES not only recapitulated all six clusters identified by antibody measurement (**Figure S2B**) but also identified many additional proteins representing established lineage markers of these sub-populations, which were completely missed by gene expression analysis. Indeed, the most differentially active proteins in each cluster present a highly cluster-specific activity pattern not visible by gene expression alone (**Figure S2C,D**).

Furthermore, when gene expression-based clustering was limited only to the genes encoding for the proteins in the CITE-Seq panel, the single-cell RNA-Seq dropout problem was so severe that cluster structure was completely lost (**Figure 3D**). This suggests that critical proteins, whose role in the biology of these populations is extremely well established, are completely missed in terms of their gene expression. In sharp contrast, PISCES analysis fully recapitulated the experimentally assessed cluster structure when the analysis was limited to the proteins represented on the CITE-Seq panel (**Figure 3E**).

Critically, the coefficients of variation (i.e., $COV = \sigma/\mu$), as computed for gene-expression, antibody-measured protein abundance, and VIPER-measured protein activity, shows that

VIPER-measured activity dramatically outperforms gene expression (p=0.0004 by paired t-test across the entire panel) and even antibody measurements for most proteins (p = 0.0083 across the entire panel), indicating a significant improvement in reproducibility and signal-to-noise ratio (**Figure 3A**). Finally, we assessed correlation between either gene expression or VIPER-measured protein activity against protein abundance as assessed by CITE-Seq. Across the board VIPER significantly outperformed gene expression (**Figure 3B**), with strong visual cluster-separation even on single genes (**Figure 3F**), and pairwise plots of VIPER activity vs paired CITE-Seq antibody staining resembling flow cytometry plots (**Figure 4**).

Furthermore, we would like to point out that protein abundance, as assessed by antibodies, is a poor proxy for protein activity. This is because, even after a protein is expressed, its activity is manifested only when it is effectively post-translationally modified, it is translocated into the appropriate sub-cellular compartment, and it has formed complexes with critical cognate binding partners. By measuring activity via expression of highly multiplexed gene reporter assay, VIPER can effectively report on the activity of proteins, which has been so far elusive, especially in single cells. In a separate analysis of CD45+ cells that were isolated from renal clear cell carcinoma, then split and profiled at the single cell level using both scRNA-Seq and a CyTEK high-throughput flow cytometry panel panel of 19 lymphoid and 19 myeloid antibodies (Obradovic et. al., 2021), the de-noising effect of PISCES was even more obvious. Not only did these results completely recapitulate the results obtained for the CITE-Seq comparison, but, given the larger number of experimentally assessed proteins, they provide further evidence of the dramatic improvement offered by PISCES analysis over both gene expression and antibody-measured protein abundance. This is reflected in three key findings. First, experimentally assessed protein abundance (e.g., using the 19 lymphoid markers) was unable to identify the clusters that could be identified by VIPER-based measurement of the same 19 proteins, including splitting of the myeloid cluster into monocytes and macrophages, the CD8 T cell

cluster into CD8 T cells and NK cells, and the CD4 T cell cluster into CD4 T cells and Regulatory T cells (Figure S4). Second, proteins not expressed on the surface of the cell, such as FOXP3, a canonical marker of regulatory T cells, could not be reliably detected by antibody measurements but were clearly detected in the correct sub-population by VIPER. Indeed, taken together, only 4 of 38 proteins assessed by VIPER and antibody measurement were not effectively and correctly detected by VIPER in the specific cellular sub-populations for which they represent an established lineage marker (NT5E/CD73, FCGR3B/CD16b, PTGDR2/CD294, CD33). In contrast, 9 of 38 proteins could not be consistently detected by antibody measurement or were not restricted to the associated sub-populations due to noisy background staining (CD14, CD127, FOXP3, CD38, CD25, CXCR3, CD161, CTLA4, CD39). Indeed, clustering on the full set of proteins identified by PISCES on this dataset (Obradovic et. al., 2021) led to identification of rare cellular subpopulations that play a critical role in post-surgical tumor recurrence, and for which PISCES-inferred markers were validated by immunohistochemistry.

This indicates amplification of biologically meaningful rather than artifactual signal from single cells by PISCES, and its ability to enable interrogation of individual genes of interest without data dropout. In fact, while CITE-Seq is limited by time-consuming antibody titration and panel optimization, ultimately profiling relatively few proteins in most experiments, PISCES typically captures several orders of magnitude more unique proteins, enabling interrogation of intracellular proteins which would otherwise be difficult to stain for without losing cellular RNA, as well as select surface markers of interest. Nevertheless, the cell-matched profiling of both gene expression and protein abundance by CITE-Seq enables direct comparison of PISCES inferences to measured protein abundance for a subset of proteins within the same cells, which may be used as a benchmark of the high concordance between PISCES and measured protein

abundance, and the degree to which PISCES improves signal-to-noise with respect to antibody-based measurements.

Discussion

The PISCES package for analysis of single-cell RNA-Sequencing data represents a comprehensive and highly generalizable pipeline for inference of protein activity to maximize utility of single-cell datasets. We have demonstrated its ability to mitigate the single-cell RNA-Seq data dropout problem and recapitulate high-depth data structure even from low UMI counts. We have also demonstrated its ability to recapitulate biological structure from CITE-Seq antibody-based protein profiling with much better gene-by-gene signal than gene expression. These technical and biological validations also serve as benchmarks for further refinement of the pipeline by which any changes can be comprehensively assessed.

For biological validation benchmarking, protein selection was based on pre-defined protein panels from CITE-Seq experiments. As a result, this represents a completely unbiased set of proteins that was not selected to skew performance in VIPER's favor. While we limited the comparison only to the CITE-Seq panel of proteins, PISCES produced activity profiles for 6,500 proteins. Thus, if these results are further confirmed in follow-up studies, PISCES would provide the equivalent of a single cell FACS with 6,500 antibodies, remedying the need to select and validate antibodies for specific cellular populations. Indeed, VIPER was originally developed for the analysis of proteins that directly control gene expression on the chromatin (i.e., TFs and co-TFs). As a result, accuracy and reproducibility of VIPER-based measurement of surface markers is likely to be significantly outperformed for TFs and co-TFs, which represent the most critical class of lineage markers.

In addition to the technical benchmarking of correlation between down-sampled and full-depth data, the extent of improvement by PISCES in coefficient of variation, number of genes

recovered, and gene-by-gene correlation to matched antibody profiling represent a critical biological benchmark for alternative workflows by PISCES users as new pre-processing methods are incorporated and existing algorithms are refined. The pipeline has been consciously designed to be highly modular, with customizable workflows and parameter optimization enabled by separate pre-processing, meta-cell, and clustering steps and interoperability with the popular Seurat workflow. We recommend targeting a median UMI depth / cell of no less than 5000, with the crucial step being inference of ARACNe network from high-depth data, applying the metaCell algorithm to improve sample depth for ARACNe network inference. Wherever a high-depth-derived ARACNe net is available, inference fidelity is high even on extremely low-depth datasets, so the increased availability of single-cell RNA-Seq datasets across a broad range of tissue contexts will continually allow construction of an expanding library of ARACNe networks which can be broadly applied to new data.

PISCES is chiefly limited by the fraction of 6,500+ total proteins recoverable at low UMI depth, although the number of proteins recovered nearly always compares favorably to CITE-Seq, which requires time-consuming antibody titration and is limited to predefined cell surface proteins, whereas PISCES captures proteins with the strongest signal-to-noise from the data and can infer both cell surface and intracellular protein activity. Applying metaCell ARACNe network inference addresses this to some degree, such that nearly 100% of all proteins recoverable at full depth in SNU-16 cell line sequencing data were recovered at a UMI depth of 10,000, where only half of the proteins inferred at full-depth were recoverable without metaCell, and over half of proteins remained recoverable with metaCell even at critically low UMI depth of 1,000 (Figure S2C). Future iterations of the pipeline will continue to improve on the fraction of recoverable proteins by integrating and testing novel pre-processing procedures and optimization of the ARACNe and VIPER inference steps. The development version of the PISCES R package will be continually available at https://github.com/califano-lab/PISCES.

Acknowledgements

This work was supported by an NCI Outstanding Investigator Award (R35 CA197745) and NIH Shared Instrumentation Grants (S10 OD012351 and S1 0OD021764), all to AC. Also, this research was funded in part through the NCI Cancer Center Support Grant (P30 CA013696).

Declarations of Interests

P.L. is Director of Single-Cell Systems Biology at DarwinHealth, Inc., a company that has licensed some of the algorithms used in this manuscript from Columbia University. .A.C. is founder, equity holder, consultant, and director of DarwinHealth Inc., which has licensed IP related to these algorithms from Columbia University. Columbia University is an equity holder in DarwinHealth Inc.

Online Methods

Quality Control, Normalization, and Scaling: As a pre-processing step, low quality cells and genes lacking enough data to be useful are removed from the analysis. Cell quality is determined by two primary factors – read depth and mitochondrial gene percentage. Samples with too many or too few reads are likely sequencing errors (doublets or empty droplets), while a high mitochondrial gene percentage is indicative of cell stress or damage. This latter group of cells will typically have a biased transcriptome not representative of the actual cell state. For most data sets, PISCES will simply remove genes with no reads at all. For larger data sets, genes that appear in less than 1% of the total cells will be removed in order to optimize computational complexity. Cells with fewer than 1000 total UMIs or mitochondrial transcript fraction greater than 25% are also removed in quality-control filtering. Filtered data are then normalized to log10(counts per million + 1). A gene expression signature is then generated from the normalized data using either double rank transformation or Seurat SCTransform scaling function.

Seurat Pre-Processing Workflow: Gene Expression UMI count matrices for each sample are processed in R using the Seurat SCTransform command to perform a regularized negative binomial regression based on the 3000 most variable genes. For datasets combining samples across multiple patients, normalized datasets may be integrated using the FindIntegrationAnchors and IntegrateData functions in Seurat. The resulting data are projected into their first 50 principal components, and further reduced into a 2-dimensional visualization space using the RunUMAP function with method umap-learn and Pearson correlation as the distance metric between cells. Differential Gene Expression between clusters is computed by the MAST hurdle model for single-cell gene expression modeling, as implemented in the Seurat FindAllMarkers command, with log fold change threshold of 0.5 and minimum fractional

expression threshold of 0.25, indicating that the resulting gene markers for each cluster are restricted to those with log fold change greater than 0 and non-zero expression in at least 25% of the cells in the cluster.

Initial Clustering and MetaCells: In order to generate accurate, robust networks, ARACNe requires data from a population that shares the majority of its transcriptional architecture. In the context of single cells, this requires separating the data into coarse cell type clusters before network generation. These clusters can be generated in a number of ways; any of the popular gene expression methods for clustering will work, as will a simple clustering based on the first 30 principle components in gene expression space. We have implemented clustering on gene expression signature by Partition Around Medioids (PAM), Multi-Way K-Means, and Louvain with Resolution Optimization. Once the data have been clustered, meta-cells can be generated for input to ARACNe. By pooling cells that are close together in either gene expression or VIPER space within a cluster, the number of interactions inferred using ARACNe can be increased. PISCES uses a simple K-nearest-neighbors approach to pool cells, then sums reads across neighbors and re-normalizing. This data then serves as the input to ARACNe.

ARACNe Network Generation: A full guide for utilizing ARACNe is available on the Califano Lab Github at https://github.com/califano-lab/PISCES. For each gene expression cluster, 250 metaCells are sampled to compute a regulatory network. All networks are reverse engineered by the ARACNe algorithm, run with 100 bootstrap iterations using 1785 transcription factors (genes annotated in gene ontology molecular function database as GO:0003700, “transcription factor activity”, or as GO:0003677, “DNA binding” and GO:0030528, “transcription regulator activity”, or as GO:0003677 and GO:0045449, “regulation of transcription”), 668 transcriptional cofactors (a manually curated list, not overlapping with the transcription factor list, built upon

genes annotated as GO:0003712, “transcription cofactor activity”, or GO:0030528 or GO:0045449), 3455 signaling pathway related genes (annotated in GO biological process database as GO:0007165, “signal transduction” and in GO cellular component database as GO:0005622, “intracellular” or GO:0005886, “plasma membrane”), and 3620 surface markers (annotated as GO:0005886 or as GO:0009986, “cell surface”). Each regulator set is run separately, as different types of proteins will have different mutual information thresholds. Once a set of regulons has been inferred for each group of regulators, the results are combined into a single network. ARACNe is only run on these gene sets so as to limit protein activity inference to proteins with biologically meaningful downstream regulatory targets, and we do not apply ARACNe to infer regulatory networks for proteins with no known signaling or transcriptional activity, for which protein activity may be difficult to biologically interpret. Parameters are set to zero DPI (Data Processing Inequality) tolerance and MI (Mutual Information) p-value threshold of 10^{-8}, computed by permuting the original dataset as a null model. Each gene list used to run ARACNe is available on github.

VIPER Analysis and Re-clustering: Once cluster-specific networks have been generated, they will serve as the input to a final VIPER run. More accurate networks will naturally lead to more accurate inferences of protein activity, which in turn allows for more robust downstream analyses. Bulk networks can also be incorporated to fill in any gaps present in the single-cell networks, as ARACNe will typically be unable to infer regulons for some proteins even with the implementation of MetaCells. These protein activities inferred from bulk should be considered less accurate, but they can be used to follow-up on previously known proteins of interest, for instance. Once a final VIPER matrix has been inferred, the data can be re-clustered. VIPER-space will typically allow for the parsing of smaller, more transcriptionally distinct populations. These classifications can then be used for a master regulator analysis that identifies the driving

regulators of the differential cell state. This can be done in several ways, with a Bootstrapped Mann Whitney-U test being the most robust. Cluster-specific Stouffer integration or a data-wide ANOVA or Kruskal-Wallis test are also viable alternatives and implemented within PISCES.

Weighted VIPER: Previously, MetaVIPER was developed as an initial adaptation of VIPER to single-cell data. By using multiple networks, MetaVIPER sought to accurately recapitulate protein activity in populations for which no context-specific network was available. To briefly explain this method, protein activity would be inferred from a given gene expression signature using multiple networks, which would then be integrated on a protein-by-protein basis using the square of the NES. Since a non-relevant network would generate a protein activity NES close to zero under the null model, networks that generate more extreme NES's can be interpreted to more accurately match the given biological context and were thus weighted more heavily. This approach has been improved on further in PISCES. Rather than relying on the square of the NES to integrate networks in a protein-by-protein manner, Weighted VIPER utilizes all the proteins in a given sample to determine network accuracy. For each sample, the NES's generated by the set of networks for each protein are ranked, and the ranks are totaled across proteins. Networks are then weighted based on their frequency as the most-accurate network. As an example, if network A generates the most extreme NES for 50% of the proteins in a sample and network B generates the most extreme NES for 25% of the proteins, network A will be weighted twice as heavily in the integration. This technique utilizes all proteins as a multiplexed reporter of network accuracy, allowing for more accurate matching of samples and the most-context specific network available.

Single Cell Visualization Functionality: Visualizing data with thousands of dimensions is a fundamental challenge of transcriptomics. PISCES has a number of pre-built plotting functions

to aid in the visualization of results. Scatter plots are based in UMAP coordinates, with the starting features filtered by the most significant proteins within each sample. Functions within PISCES allow for the visualization of clustering schemes, protein activity, or gene expression in UMAP space, along with density plotting. Additionally, we provide heatmap functionality for more tractable succinct visualization of a set of genes or proteins grouped by cluster, such as a set of known markers or a list of candidate master regulators.

Resolution-Optimized Louvain Clustering Algorithm: The default clustering method implemented in Seurat is Partitioning Around Medioids (PAM). However, for large datasets aggregating hundreds of thousands of single-cells, PAM is computationally slow, requiring more computational power than is available to the average user and computation of pairwise distance matrices exceeding the vector size limit in R. In such cases, it is preferable to run a network-based Louvain clustering, as implemented in Seurat, which optimizes network modularity score. However, practical implementations of Louvain clustering include a user-adjustable resolution parameter which allows over-clustering and under-clustering without an objective cluster quality metric. To solve this problem, we have implemented a hybrid clustering approach in PISCES which performs cluster assignment in two steps. First, Seurat Louvain clustering is performed with resolution values ranging from 0.01 to 1.0 at intervals of 0.01, then cluster quality is evaluated at each resolution value to select an optimum in this range. For each resolution value, clustered cells are subsampled to 1000, and silhouette score is computed for these 1000 cells and their corresponding cluster labels, with correlation distance metric. This procedure is repeated for 100 random samples to compute a mean and standard deviation of average silhouette score at each resolution value. The highest resolution value that maximizes mean silhouette score is selected as the optimal resolution at which to cluster the data.

Multi-Way K-Means Clustering Algorithm: In addition to PAM and Louvain with Resolution Optimization, PISCES further implements a Multi-Way K-Means Clustering approach. Transitioning populations, such as in a differentiation pathway, are extremely common, and such relationships will not be accurately characterized by a discrete clustering scheme. To handle this set of problems, we adapted the Multiway K-Means algorithm for use in biological settings, where samples can be thought of as linear combinations of related phenotypes rather than simply belonging to totally distinct populations. Originally developed for clustering speciating microbiome populations, Multiway K-Means technique has two primary advantages. First, it more accurately captures cluster center (in biological terms, a representative phenotype) for each population endpoint. Second, it places cells along a trajectory between cluster centers, providing a more accurate representation of cell state and allowing for additional inferences into the drivers of transitional populations.

Semi-Supervised Cell Type Calling: For each single cell gene expression sample, cell-by-cell identification of cell types is performed using the SingleR package and the preloaded Blueprint-ENCODE reference, which includes normalized expression values for 259 bulk RNASeq samples generated by Blueprint and ENCODE from 43 distinct cell types representing pure populations of stroma and immune cells (Martens et. al., 2013; ENCODE Project Consortium, 2012). The SingleR algorithm computes correlation between each individual cell and each of the 259 reference samples, and then assigns both a label of the cell type with highest average correlation to the individual cell and a p-value computed by wilcox test of correlation to that cell type compared to all other cell types. Cell-by-cell SingleR labels with $p<0.05$ are added as metadata and may be projected onto PISCES-generated UMAP space. Unsupervised clusters may then be labelled as a particular cell type based on the most-represented SingleR cell type label within that cluster.

Data Collection and Downsampling for Technical Validation: SNU-16, a stomach adenocarcinoma cell line, was dissociated into a single-cell suspension and scRNAseq was performed using 10X Genomics Chromium platform (3'v3). Libraries were sequenced on an illumina Novaseq 6000 according to 10X Genomics' protocol. In mid-log growth, SNU-16 is a transcriptionally complex cell line that will typically have 40,000-50,000 UMIs/cell with 134,000 reads sequencing. These data were then down-sampled to depths of 10-40,000 at 5,000 UMI intervals and 1-10,000 at 1,000 UMI intervals. Sample depths were first drawn from a uniformly distributed multinomial with $n = N*x$ and $p1,...,pn = 1 / N$, where N was the number of cells and x is the target mean depth. Once sample depths were drawn, UMI counts were drawn from a sample-specific multinomial with $n = di$ and p1...pg = 1 / G, where di is the sample depth and G is the number of genes detected in the original UMI matrix.

Biological Validation Analysis: A highly used public CITE-Seq dataset of cord blood mononuclear cells was downloaded from Gene Expression Omnibus (GEO), and subset to human cells only. RNA counts were processed by the standard PISCES workflow, and antibody dependent tags (ADTs) were concurrently analyzed. ADT matrix was normalized by Seurat Centered Log Ratio "CLR" function, and clustered by PISCES resolution-optimized Louvain algorithm. Two-dimensional data representation was computed by RunUMAP, and antibody staining of all markers was visualized in a heatmap, with cells grouped by ADT cluster. For single-cell sequencing data, both gene expression signature and PISCES-inferred VIPER matrix were subset to genes encoding proteins represented in the ADT panel, and data were re-clustered on those gene subsets. For genes shared across all three modalities, coefficient of variation was computed as standard deviation divided by mean across all cells, and Spearman

correlation was computed between gene expression or VIPER and corresponding protein-targeting antibody.

Data Availability: The PISCES pipeline is implemented as an R package with all dependencies listed and a usage tutorial available at https://github.com/califano-lab/PISCES. All data, ARACNe networks, and VIPER matrices referenced in this manuscript are also available at https://github.com/califano-lab/PISCES-validation.

Figures & Figure Legends

Figure 1: Graphical Representation of Analysis Pipeline

1A) Flowchart of overall analysis pipeline, showcasing sequential data transformations from original raw RNA-Seq gene expression counts matrix (blue) followed by Quality Control Filtering and Normalization (yellow) and data scaling (red), followed by cluster-specific ARACNe and final VIPER transformation to generate a single-cell VIPER-inferred protein protein activity matrix (green). **1B)** Graphical of the gene expression dropout mitigation effect. A theoretical ARACNe-inferred regulon of a proteomic master regulator of cell state (MR) and its downstream transcriptional targets (g1,g2,g3,g4,….) is shown, along with a matrix showing sparseness of expression for MR and each of its targets both in cells with high real activity of MR and cells with low activity. From MR expression alone, only a single sample with high MR-activity would be correctly identified. However, by integrating the expression values from each target gene, high protein activity of MR can be correctly inferred despite the high dropout rate of any single gene target.

Figure 2: Technical Benchmarking Shows Increased Recovery of Original Data Structure from Downsampled Matrices by VIPER vs Gene Expression

2A) Boxplot showing distribution across single cells of Pearson correlation between sub-sampled and original full-depth cells. Along the x-axis is the UMI/cell downsampling quotient. In purple, correlation between downsampled and original gene expression is shown to rapidly degrade, to a median consistently below 0.5, and below 0.25 even by the relatively high depth of 10,000 UMI/cell. In red, correlation is shown between VIPER inference on down-sampled gene expression signature with full-depth ARACNe network vs VIPER inference on full-depth gene expression signature using full-depth ARACNe network, such that correlation remains high

even at extremely low sample depth, with a median above 0.75 even at 1000 UMI/cell. In green, correlation is shown between VIPER inference on full-depth gene expression signature using ARACNe networks derived from full-depth vs down-sampled data, and in blue correlation is shown between full-depth VIPER inference using full-depth ARACNe networks and VIPER inference on down-sampled gene expression signature using down-sampled ARACNe network. In all cases protein activity improves on gene expression, and down-sampling of both VIPER and ARACNe simultaneously still improves correlation relative to gene expression down to a depth of 5000 UMI/cell, with Bonferroni-corrected p-values by paired Wilcox test < 0.05. **2B)** For UMI depths ranging from 1000 to 10000, correlation between full-depth VIPER matrix using full-depth ARACNe network and VIPER matrices computed on on down-sampled gene expression signatures with either full-depth or metaCell ARACNe. metaCell ARACNe significantly improves on correlation with full-depth data for all depths >1000 UMI/cell, by paired Wilcox test p-values < 0.05. Mean correlation at low-depth with metaCell ARACNe network approaches 0.75, seen only at UMI depths >20000 without applying the metaCell ARACNe inference approach.

Figure 3: Biological Benchmarking Shows Dramatically Increased Concordance with CITE-Seq Antibody Profiling by VIPER vs Gene Expression

3A) Coefficient of Variation (computed as σ/μ) for each gene profiled by the CITE-Seq antibody panel, shown for antibody staining (red), Gene Expression (green), and VIPER-inferred protein activity (blue), with higher Coefficient indicating lower signal-to-noise ratio. **3B)** Spearman Correlation between Gene Expression vs Antibody (red) and VIPER vs Antibody (blue) computed across cells for each gene profiled by the CITE-Seq antibody panel. **3C)** UMAP projection and clustering of CITE-Seq antibody staining panel, labelled with cell types inferred from SingleR and validated by staining for known markers. Row-scaled heatmap is shown

below with antibody staining intensity grouped by cluster. **3D)** UMAP projection and clustering of Gene Expression for the subset of genes concurrently profiled by CITE-Seq antibody staining panel. Row-scaled heatmap is shown below, with excessive noise for meaningful clustering due to single-cell RNA-Seq dropout effect. **3E)** UMAP projection and clustering of VIPER protein activity, labelled with cell types as in 3C. Row-scaled heatmap is shown below with VIPER activity grouped by cluster, for the subset of genes concurrently profiled by CITE-Seq antibody staining panel with activity inferred by VIPER. **3F)** Representative Correlation plots of Gene Expression vs Antibody and VIPER vs Antibody, showing greater concordance of CD3D VIPER activity with Antibody intensity, relative to CD3D Gene Expression.

Figure 4: Pairwise CITE-Seq Antibody vs VIPER Correlation Plots

4A) Correlation Plots of CD3D Gene Expression vs Antibody Intensity (left) and VIPER vs Antibody Intensity (right). **4B)** Correlation Plots of CD3E Gene Expression vs Antibody Intensity (left) and VIPER vs Antibody Intensity (right). **4C)** Correlation Plots of CD3G Gene Expression vs Antibody Intensity (left) and VIPER vs Antibody Intensity (right). **4D)** Correlation Plots of CD4 Gene Expression vs Antibody Intensity (left) and VIPER vs Antibody Intensity (right). **4E)** Correlation Plots of CD8B Gene Expression vs Antibody Intensity (left) and VIPER vs Antibody Intensity (right). **4F)** Correlation Plots of CD14 Gene Expression vs Antibody Intensity (left) and VIPER vs Antibody Intensity (right). **4G)** Correlation Plots of FCGR3A (CD16) Gene Expression vs Antibody Intensity (left) and VIPER vs Antibody Intensity (right). **4H)** Correlation Plots of PTPRC (CD45) Gene Expression vs Antibody Intensity (left) and VIPER vs Antibody Intensity (right).

Figure S1: Clustering of CITE-Seq Dataset on Full Gene Expression and VIPER matrices,

S1A) UMAP plot of the CITE-Seq CBMC RNA-Seq data clustered on the entire gene expression matrix, showing 4 distinct clusters, labelled according to the majority cell type represented in each cluster. **S1B)** UMAP plot of the corresponding VIPER inferences clustered on the entire set of inferred proteins, with clusters exactly recapitulating the cluster structure in Figure 3E, capturing all represented cell types. **S1C)** Gene Expression Heatmap of the top5 inferred master regulators of each VIPER cluster, scaled by row. **S1D)** VIPER Activity Heatmap of the top5 inferred master regulators of each VIPER cluster, scaled by row.

Figure S2: Technical Validation Dataset Quality Control and ARACNe Network Size

S2A) Distribution from original full-depth dataset of UMIs/cell (left, in red), number of genes with non-zero gene expression per cell (middle, in green), and percentage of mitochondrial transcripts (right, in blue). **S2B)** Fraction of Total ARACNe network regulons (y-axis) recovered at each down-sampling depth (x-axis) relative to full-depth data, such that fraction decreases log-linearly with down-sampling depth. **S2C)** Fraction of Total ARACNe network regulons relative to full-depth data (y-axis) recovered at each down-sampling depth from 1000 to 10000 UMI/cell, with metaCell approach (red) or without metaCell approach (black).

Figure S3: Pairwise Downsampling Correlation Matrix

Heatmap of mean correlation values compared to original full-depth VIPER matrix with full-depth ARACNe network for each combination of down-sampled ARACNe and VIPER gene expression signature depth. Each row corresponds to depth of gene expression signature input to VIPER, and each column corresponds to depth of gene expression input to ARACNe. Correlation is subset to proteins differentially up-regulated or down-regulated ($p<0.05$) within original full-depth

VIPER matrix, on a cell-by-cell basis, and mean correlation across all cells is plotted for each box on the heatmap corresponding to a particular down-sampling approach.

Figure S4: Comparison of VIPER Inferences and Gene Expression to Flow Cytometry in Renal Clear Cell Carcinoma Dataset

S4A) UMAP projection, clustering, and heatmap by flow cytometry proteins profiled in CyTEK Lymphoid Panel. **S4B)** UMAP and clustering by scRNASeq gene expression subset to the proteins profiled in S4A, showing noise-induced decrease in clustering resolution. **S4C**) UMAP and clustering by VIPER-inferred protein activity using PISCES, subset to the proteins profiled in S4A. **S4D**) UMAP and clustering by flow cytometry proteins profiled in CyTEK myeloid panel. **S4E)** UMAP and clustering by scRNA-Seq gene expression, subset to the proteins profiled in S4D. **S4F)** UMAP and clustering by VIPER-inferred protein activity using PISCES, subset to the proteins profiled in S4D. partially reproduced with permissions from Obradovic et. al., 2021.

Figure S5: Flowchart of Technical Validation Down-Sampling Approach

References

1. He P, Williams BA, Trout D, Marinov GK, Amrhein H, Berghella L, Goh ST, Plajzer-Frick I, Afzal V, Pennacchio LA, Dickel DE, Visel A, Ren B, Hardison RC, Zhang Y, Wold BJ. The changing mouse embryo transcriptome at whole tissue and single-cell resolution. Nature. 2020 Jul;583(7818):760-767. doi: 10.1038/s41586-020-2536-x. Epub 2020 Jul 29. PMID: 32728245; PMCID: PMC7410830.
2. Qian J, Olbrecht S, Boeckx B, Vos H, Laoui D, Etlioglu E, Wauters E, Pomella V, Verbandt S, Busschaert P, Bassez A, Franken A, Bempt MV, Xiong J, Weynand B, van Herck Y, Antoranz A, Bosisio FM, Thienpont B, Floris G, Vergote I, Smeets A, Tejpar S, Lambrechts D. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. Cell Res. 2020 Sep;30(9):745-762. doi: 10.1038/s41422-020-0355-0. Epub 2020 Jun 19. PMID: 32561858; PMCID: PMC7608385.
3. Xu G, Qi F, Li H, Yang Q, Wang H, Wang X, Liu X, Zhao J, Liao X, Liu Y, Liu L, Zhang S, Zhang Z. The differential immune responses to COVID-19 in peripheral and lung revealed by single-cell RNA sequencing. Cell Discov. 2020 Oct 20;6:73. doi: 10.1038/s41421-020-00225-2. PMID: 33101705; PMCID: PMC7574992.
4. Speranza E, Williamson BN, Feldmann F, Sturdevant GL, Pérez-Pérez L, Meade-White K, Smith BJ, Lovaglio J, Martens C, Munster VJ, Okumura A, Shaia C, Feldmann H, Best SM, de Wit E. Single-cell RNA sequencing reveals SARS-CoV-2 infection dynamics in lungs of African green monkeys. Sci Transl Med. 2021 Jan 27;13(578):eabe8146. doi: 10.1126/scitranslmed.abe8146. Epub 2021 Jan 11. PMID: 33431511; PMCID: PMC7875333.
5. Sade-Feldman M, Yizhak K, Bjorgaard SL, Ray JP, de Boer CG, Jenkins RW, Lieb DJ, Chen JH, Frederick DT, Barzily-Rokni M, Freeman SS, Reuben A, Hoover PJ, Villani

AC, Ivanova E, Portell A, Lizotte PH, Aref AR, Eliane JP, Hammond MR, Vitzthum H, Blackmon SM, Li B, Gopalakrishnan V, Reddy SM, Cooper ZA, Paweletz CP, Barbie DA, Stemmer-Rachamimov A, Flaherty KT, Wargo JA, Boland GM, Sullivan RJ, Getz G, Hacohen N. Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. Cell. 2018 Nov 1;175(4):998-1013.e20. doi: 10.1016/j.cell.2018.10.038. Erratum in: Cell. 2019 Jan 10;176(1-2):404. PMID: 30388456; PMCID: PMC6641984.

6. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, Leeson R, Kanodia A, Mei S, Lin JR, Wang S, Rabasha B, Liu D, Zhang G, Margolais C, Ashenberg O, Ott PA, Buchbinder EI, Haq R, Hodi FS, Boland GM, Sullivan RJ, Frederick DT, Miao B, Moll T, Flaherty KT, Herlyn M, Jenkins RW, Thummalapalli R, Kowalczyk MS, Cañadas I, Schilling B, Cartwright ANR, Luoma AM, Malu S, Hwu P, Bernatchez C, Forget MA, Barbie DA, Shalek AK, Tirosh I, Sorger PK, Wucherpfennig K, Van Allen EM, Schadendorf D, Johnson BE, Rotem A, Rozenblatt-Rosen O, Garraway LA, Yoon CH, Izar B, Regev A. A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. Cell. 2018 Nov 1;175(4):984-997.e24. doi: 10.1016/j.cell.2018.09.006. PMID: 30388455; PMCID: PMC6410377.
7. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, Ryu HS, Kim S, Lee JE, Park YH, Kan Z, Han W, Park WY. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun. 2017 May 5;8:15081. doi: 10.1038/ncomms15081. PMID: 28474673; PMCID: PMC5424158.
8. Obradovic A, Chowdhury N, Haake S, Ager C, Wang V, Vlahos L, Guo X, Aggen D, Rathmell K, Jonasch E, Johnson J, Roth M, Beckermann K, Rini B, McKiernan J, Califano A, Drake C. Single-Cell Protein Activity Analysis Identified Recurrence-Associated Renal Tumor Macrophages. Cell. 2021. In Press.

9. Lachmann A, Giorgi FM, Lopez G, Califano A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics. 2016 Jul 15;32(14):2233-5. doi: 10.1093/bioinformatics/btw216. Epub 2016 Apr 23. PMID: 27153652; PMCID: PMC4937200.
10. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, Califano A. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet. 2016 Aug;48(8):838-47. doi: 10.1038/ng.3593. Epub 2016 Jun 20. PMID: 27322546; PMCID: PMC5040167.
11. Alvarez MJ, Subramaniam PS, Tang LH, Grunn A, Aburi M, Rieckhof G, Komissarova EV, Hagan EA, Bodei L, Clemons PA, Dela Cruz FS, Dhall D, Diolaiti D, Fraker DA, Ghavami A, Kaemmerer D, Karan C, Kidd M, Kim KM, Kim HC, Kunju LP, Langel Ü, Li Z, Lee J, Li H, LiVolsi V, Pfragner R, Rainey AR, Realubit RB, Remotti H, Regberg J, Roses R, Rustgi A, Sepulveda AR, Serra S, Shi C, Yuan X, Barberis M, Bergamaschi R, Chinnaiyan AM, Detre T, Ezzat S, Frilling A, Hommann M, Jaeger D, Kim MK, Knudsen BS, Kung AL, Leahy E, Metz DC, Milsom JW, Park YS, Reidy-Lagunes D, Schreiber S, Washington K, Wiedenmann B, Modlin I, Califano A. A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. Nat Genet. 2018 Jul;50(7):979-989. doi: 10.1038/s41588-018-0138-4. Epub 2018 Jun 18. PMID: 29915428; PMCID: PMC6421579.
12. Zeleke, Tizita & Pan, Qingfei & Chiuzan, Cody & Onishi, Maika & Alvarez, Mariano & Honan, Erin & Yang, Min & Chia, Pei & Mukhopadhyay, Partha & Kelly, Sean & Wu, Ruby & Fenn, Kathleen & Trivedi, Meghna & Accordino, Melissa & Crew, Katherine & Hershman, Dawn & Maurer, Matthew & Jones, Simon & Califano, Andrea & Silva, José. (2020). Network-based assessment of HDAC6 activity is highly predictive of pre-clinical and clinical responses to the HDAC6 inhibitor ricolinostat. 10.1101/2020.04.23.20066928.

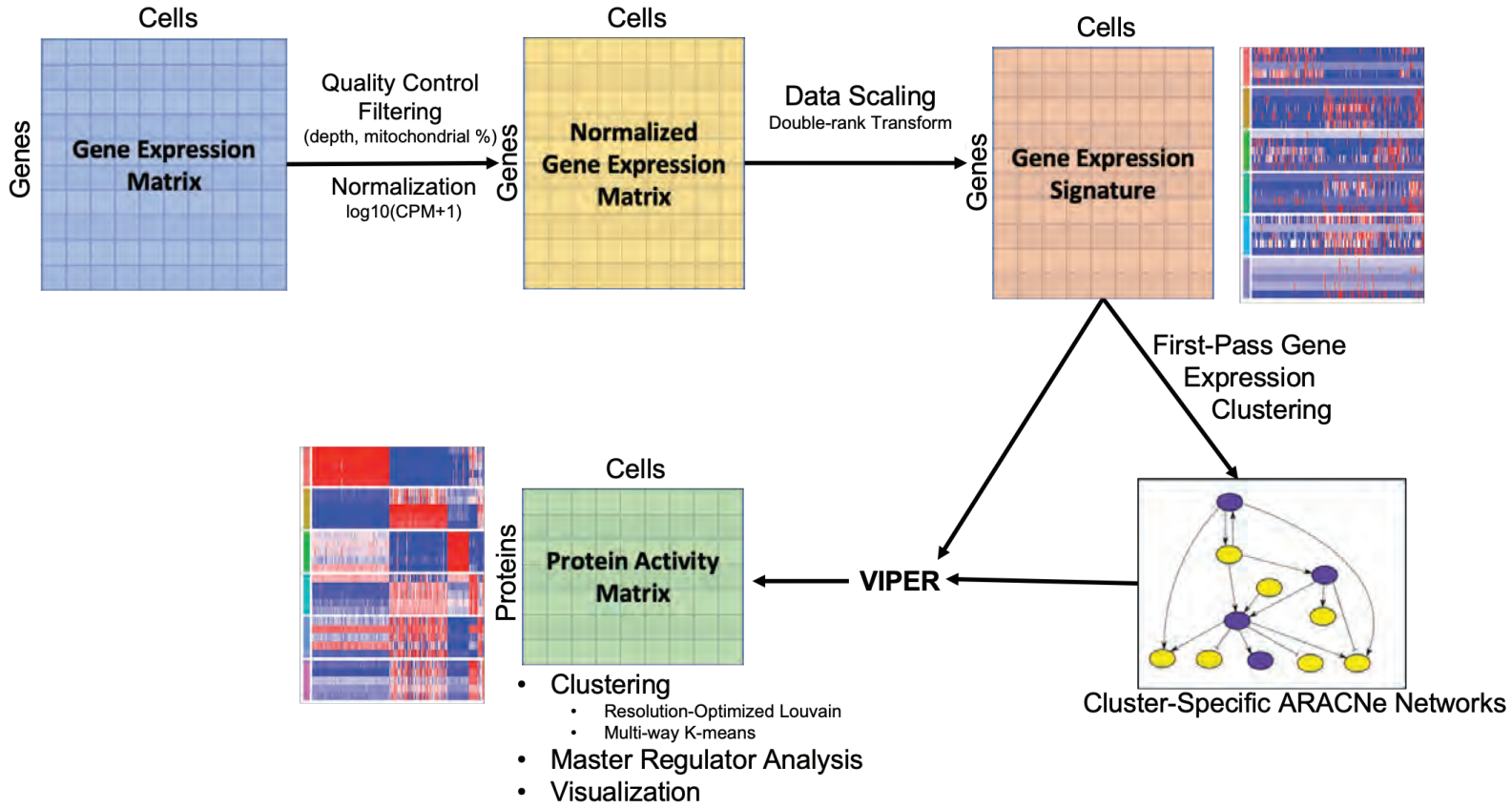
13. Rajbhandari P, Lopez G, Capdevila C, Salvatori B, Yu J, Rodriguez-Barrueco R, Martinez D, Yarmarkovich M, Weichert-Leahey N, Abraham BJ, Alvarez MJ, Iyer A, Harenza JL, Oldridge D, De Preter K, Koster J, Asgharzadeh S, Seeger RC, Wei JS, Khan J, Vandesompele J, Mestdagh P, Versteeg R, Look AT, Young RA, Iavarone A, Lasorella A, Silva JM, Maris JM, Califano A. Cross-Cohort Analysis Identifies a TEAD4-MYCN Positive Feedback Loop as the Core Regulatory Element of High-Risk Neuroblastoma. Cancer Discov. 2018 May;8(5):582-599. doi: 10.1158/2159-8290.CD-16-0861. Epub 2018 Mar 6. PMID: 29510988; PMCID: PMC5967627.
14. Ding H, Douglass EF Jr, Sonabend AM, Mela A, Bose S, Gonzalez C, Canoll PD, Sims PA, Alvarez MJ, Califano A. Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. Nat Commun. 2018 Apr 16;9(1):1471. doi: 10.1038/s41467-018-03843-3. PMID: 29662057; PMCID: PMC5902599.
15. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017 Sep;14(9):865-868. doi: 10.1038/nmeth.4380. Epub 2017 Jul 31. PMID: 28759029; PMCID: PMC5669064.
16. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019 Feb;20(2):163-172. doi: 10.1038/s41590-018-0276-y. Epub 2019 Jan 14. PMID: 30643263; PMCID: PMC6340744.
17. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun 13;177(7):1888-1902.e21. doi: 10.1016/j.cell.2019.05.031. Epub 2019 Jun 6. PMID: 31178118; PMCID: PMC6687398.

18. Martens JH, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*. 2013;98(10):1487-1489. doi:10.3324/haematol.2013.094243
19. The ENCODE Project Consortium., Overall coordination (data analysis coordination)., Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012). https://doi.org/10.1038/nature11247
20. Zhao, J., Zhang, S., Liu, Y. *et al.* Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human. *Cell Discov* **6,** 22 (2020). https://doi.org/10.1038/s41421-020-0157-z
21. Kushwaha R, Jagadish N, Kustagi M, Tomishima MJ, Mendiratta G, Bansal M, Kim HR, Sumazin P, Alvarez MJ, Lefebvre C, Villagrasa-Gonzalez P, Viale A, Korkola JE, Houldsworth J, Feldman DR, Bosl GJ, Califano A, Chaganti RS. Interrogation of a context-specific transcription factor network identifies novel regulators of pluripotency. Stem Cells. 2015 Feb;33(2):367-77. doi: 10.1002/stem.1870. PMID: 25336442; PMCID: PMC4305010.
22. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, Ziv E, Culhane AC, Paull EO, Sivakumar IKA, Gentles AJ, Malhotra R, Farshidfar F, Colaprico A, Parker JS, Mose LE, Vo NS, Liu J, Liu Y, Rader J, Dhankani V, Reynolds SM, Bowlby R, Califano A, Cherniack AD, Anastassiou D, Bedognetti D, Mokrab Y, Newman AM, Rao A, Chen K, Krasnitz A, Hu H, Malta TM, Noushmehr H, Pedamallu CS, Bullman S, Ojesina AI, Lamb A, Zhou W, Shen H, Choueiri TK, Weinstein JN, Guinney J, Saltz J, Holt RA, Rabkin CS; Cancer Genome Atlas Research Network, Lazar AJ, Serody JS, Demicco EG, Disis ML, Vincent BG, Shmulevich I. The Immune Landscape of Cancer. Immunity. 2018 Apr 17;48(4):812-830.e14. doi: 10.1016/j.immuni.2018.03.023. Epub 2018 Apr 5. Erratum in: Immunity. 2019 Aug 20;51(2):411-412. PMID: 29628290; PMCID: PMC5982584.

23. Teschendorff, A., Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat Commun* **8,** 15599 (2017). https://doi.org/10.1038/ncomms15599

Figure 1

A)


Cells
Genes
Gene Expression Matrix
Quality Control Filtering
(depth, mitochondrial %)
Normalization
log10(CPM+1)
Cells
Genes
Normalized Gene Expression Matrix
Data Scaling
Double-rank Transform
Cells
Genes
Gene Expression Signature
First-Pass Gene Expression Clustering
Cluster-Specific ARACNe Networks
VIPER
Cells
Proteins
Protein Activity Matrix
• Clustering
• Resolution-Optimized Louvain
• Multi-way K-means
• Master Regulator Analysis
• Visualization

B)

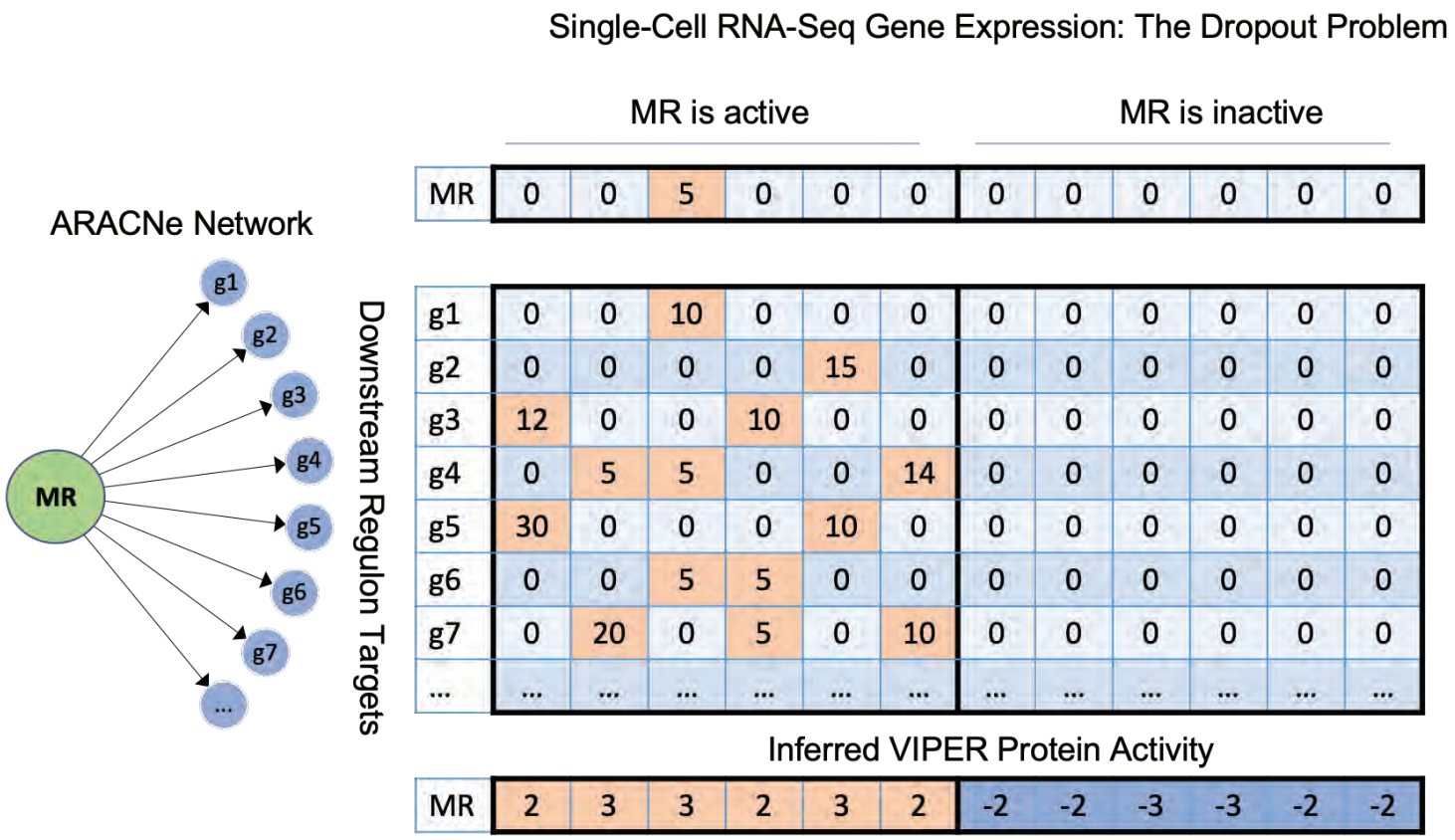
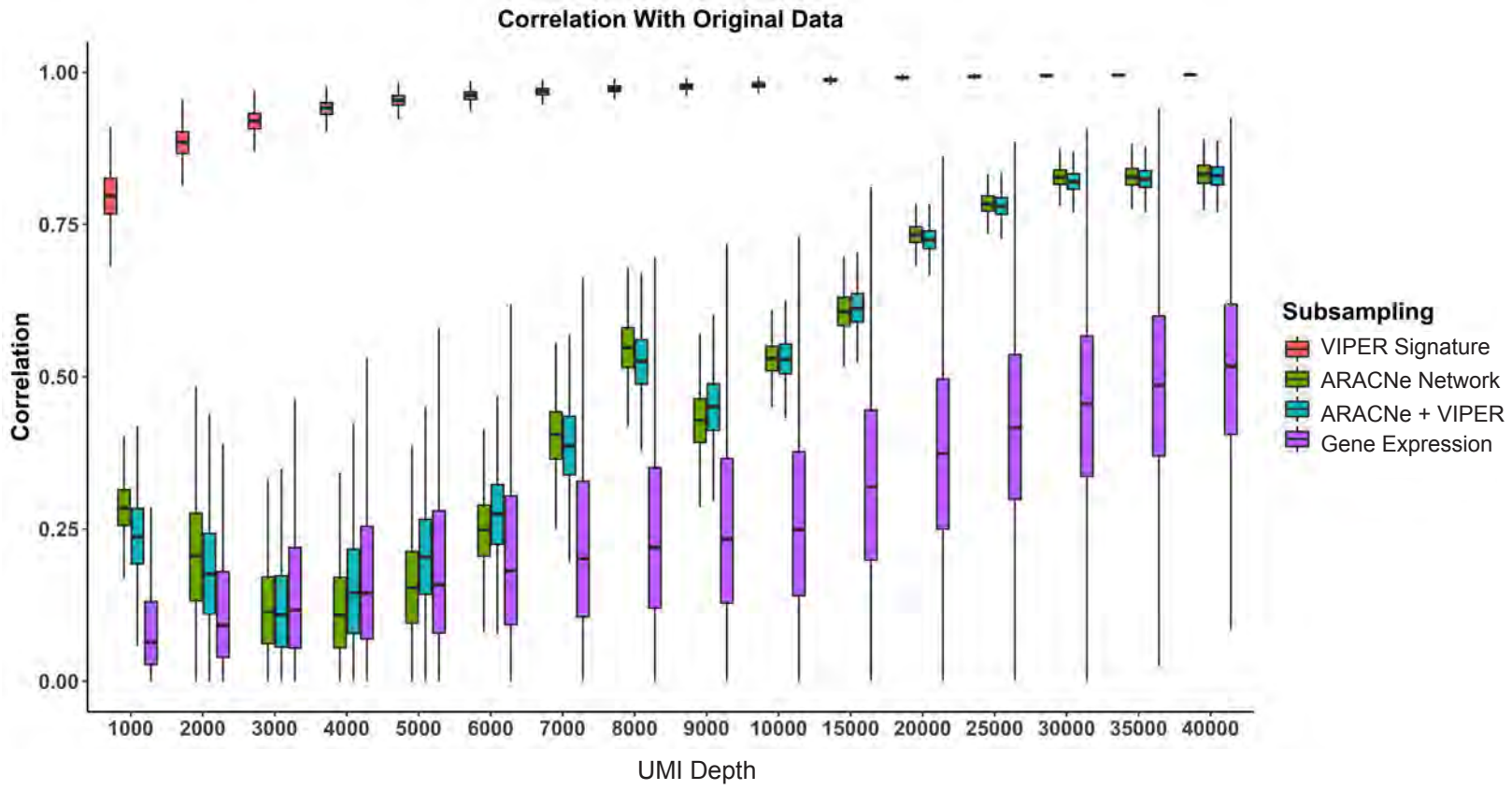

Single-Cell RNA-Seq Gene Expression: The Dropout Problem
MR is active
MR is inactive
ARACNe Network
MR
g1
g2
g3
g4
g5
g6
g7
...
Downstream Regulon Targets
MR 0 0 5 0 0 0 0 0 0 0 0 0
g1 0 0 10 0 0 0 0 0 0 0 0 0
g2 0 0 0 0 15 0 0 0 0 0 0 0
g3 12 0 0 10 0 0 0 0 0 0 0 0
g4 0 5 5 0 0 14 0 0 0 0 0 0
g5 30 0 0 0 10 0 0 0 0 0 0 0
g6 0 0 5 5 0 0 0 0 0 0 0 0
g7 0 20 0 5 0 10 0 0 0 0 0 0
... ...
Inferred VIPER Protein Activity
MR 2 3 3 2 3 2 -2 -2 -3 -3 -2 -2

Figure 2

A)


Correlation With Original Data
Correlation
1.00
0.75
0.50
0.25
0.00
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
15000
20000
25000
30000
35000
40000
UMI Depth
Subsampling
VIPER Signature
ARACNe Network
ARACNe + VIPER
Gene Expression

B)

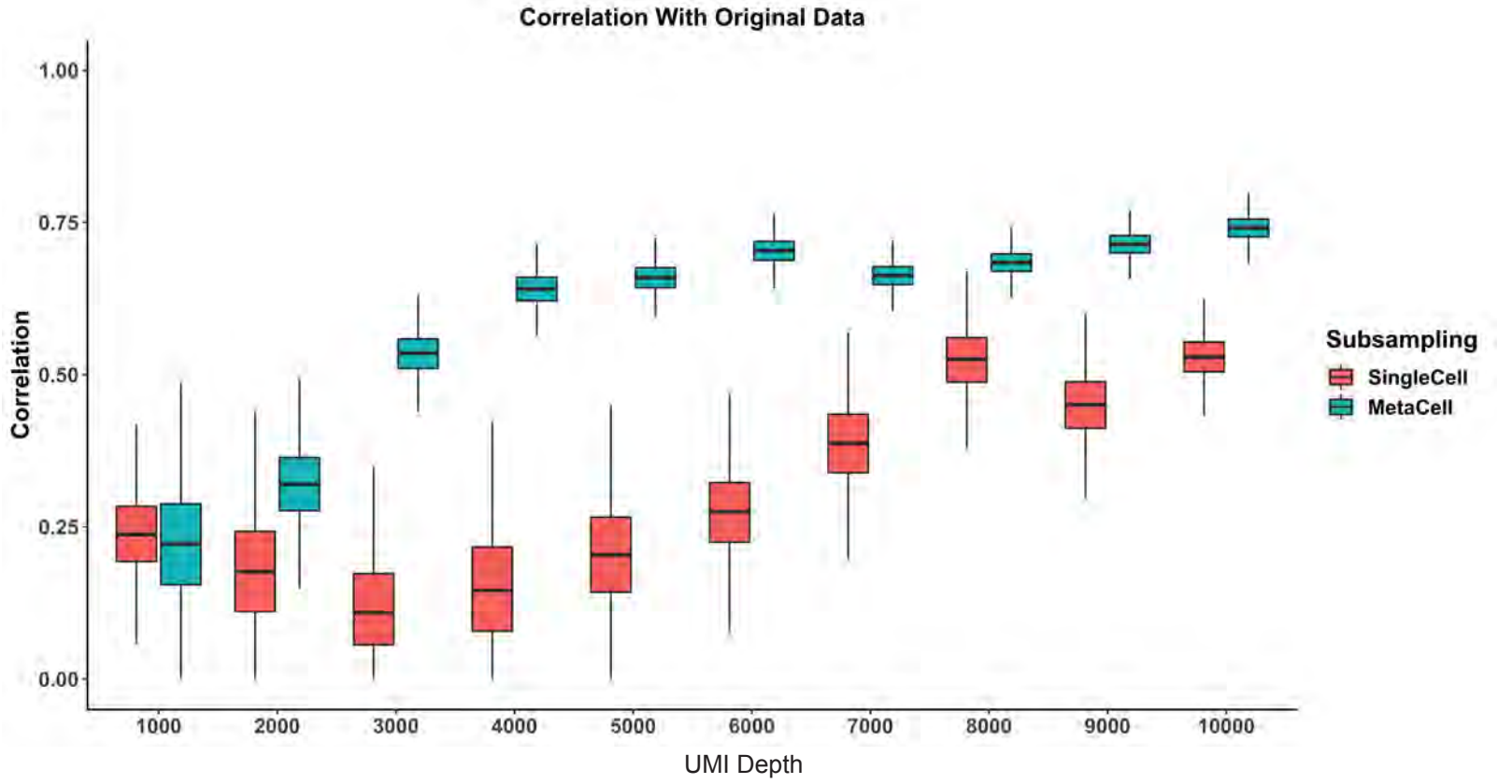

Correlation With Original Data
Correlation
1.00
0.75
0.50
0.25
0.00
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
UMI Depth
Subsampling
SingleCell
MetaCell

Figure 3

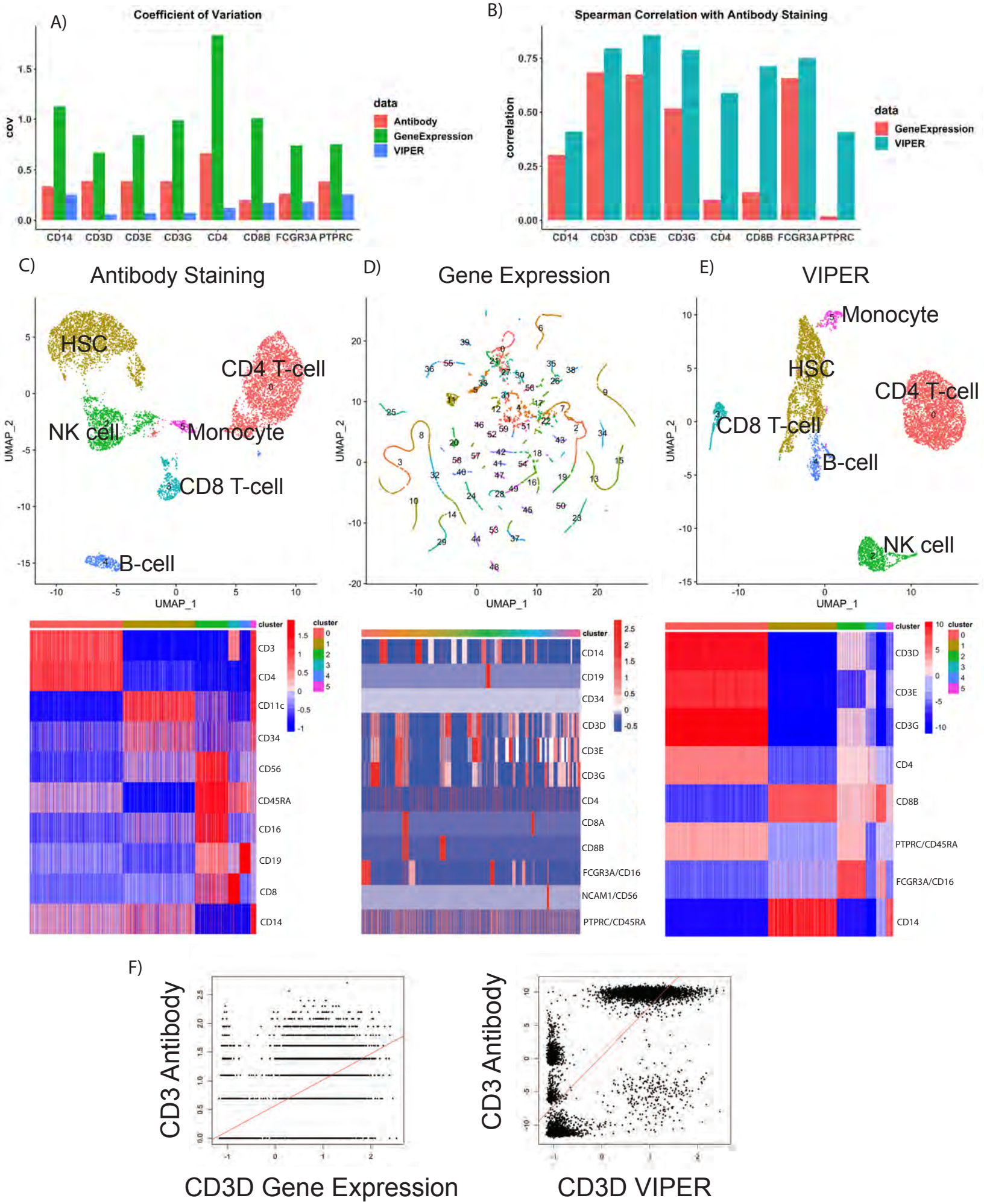

A)
Coefficient of Variation
cov
data
Antibody
GeneExpression
VIPER
CD14 CD3D CD3E CD3G CD4 CD8B FCGR3A PTPRC
B)
Spearman Correlation with Antibody Staining
correlation
data
GeneExpression
VIPER
CD14 CD3D CD3E CD3G CD4 CD8B FCGR3A PTPRC
C)
Antibody Staining
HSC
CD4 T-cell
NK cell
Monocyte
CD8 T-cell
B-cell
UMAP_1
UMAP_2
D)
Gene Expression
UMAP_1
UMAP_2
E)
VIPER
Monocyte
HSC
CD4 T-cell
CD8 T-cell
B-cell
NK cell
UMAP_1
UMAP_2
cluster
CD3
CD4
CD11c
CD34
CD56
CD45RA
CD16
CD19
CD8
CD14
CD3D
CD3E
CD3G
CD8A
CD8B
FCGR3A/CD16
NCAM1/CD56
PTPRC/CD45RA
F)
CD3 Antibody
CD3D Gene Expression
CD3 Antibody
CD3D VIPER

Figure 4

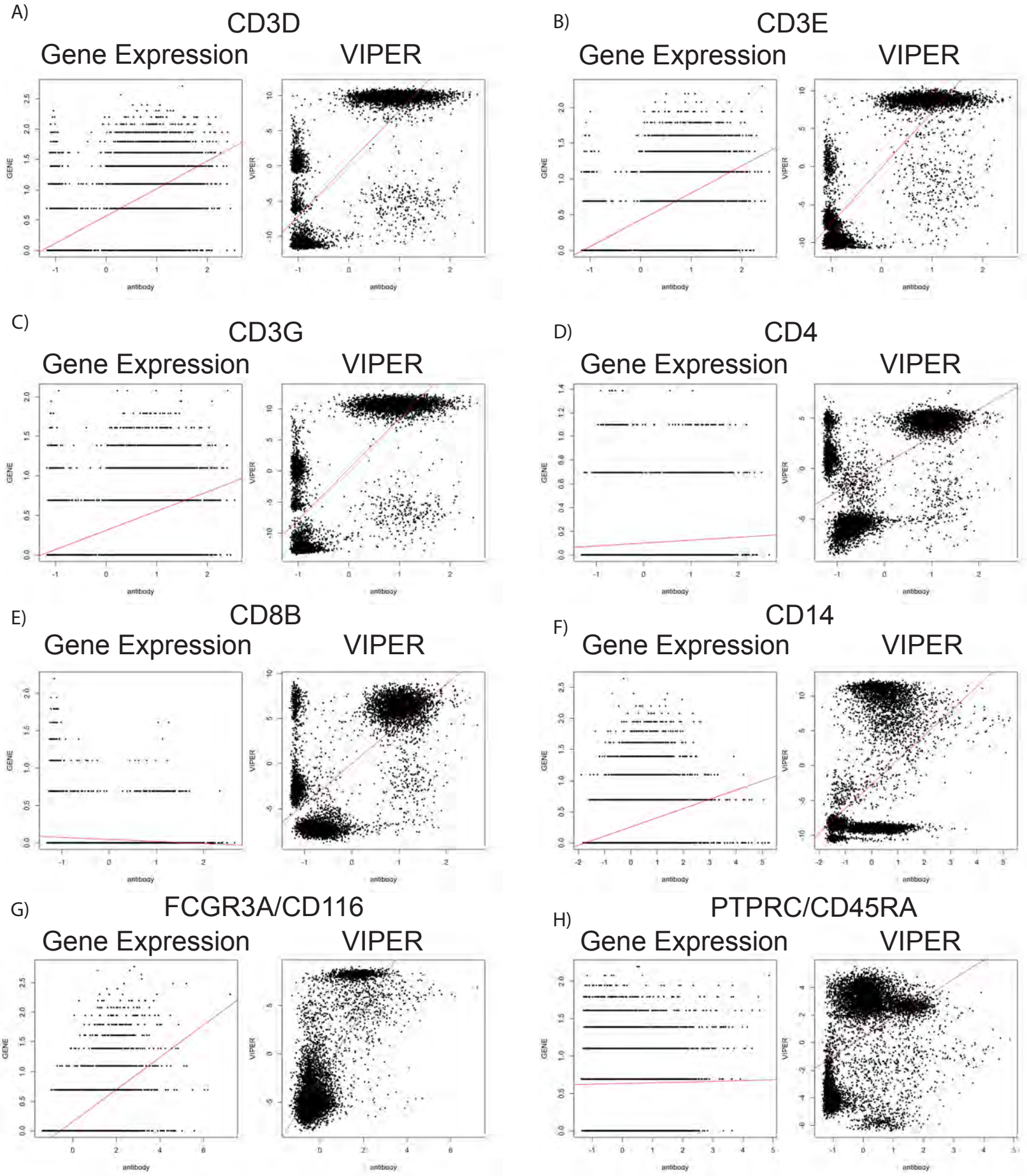

A)
CD3D
Gene Expression
VIPER
B)
CD3E
Gene Expression
VIPER
C)
CD3G
Gene Expression
VIPER
D)
CD4
Gene Expression
VIPER
E)
CD8B
Gene Expression
VIPER
F)
CD14
Gene Expression
VIPER
G)
FCGR3A/CD116
Gene Expression
VIPER
H)
PTPRC/CD45RA
Gene Expression
VIPER
GENE
VIPER
antibody

Figure S1

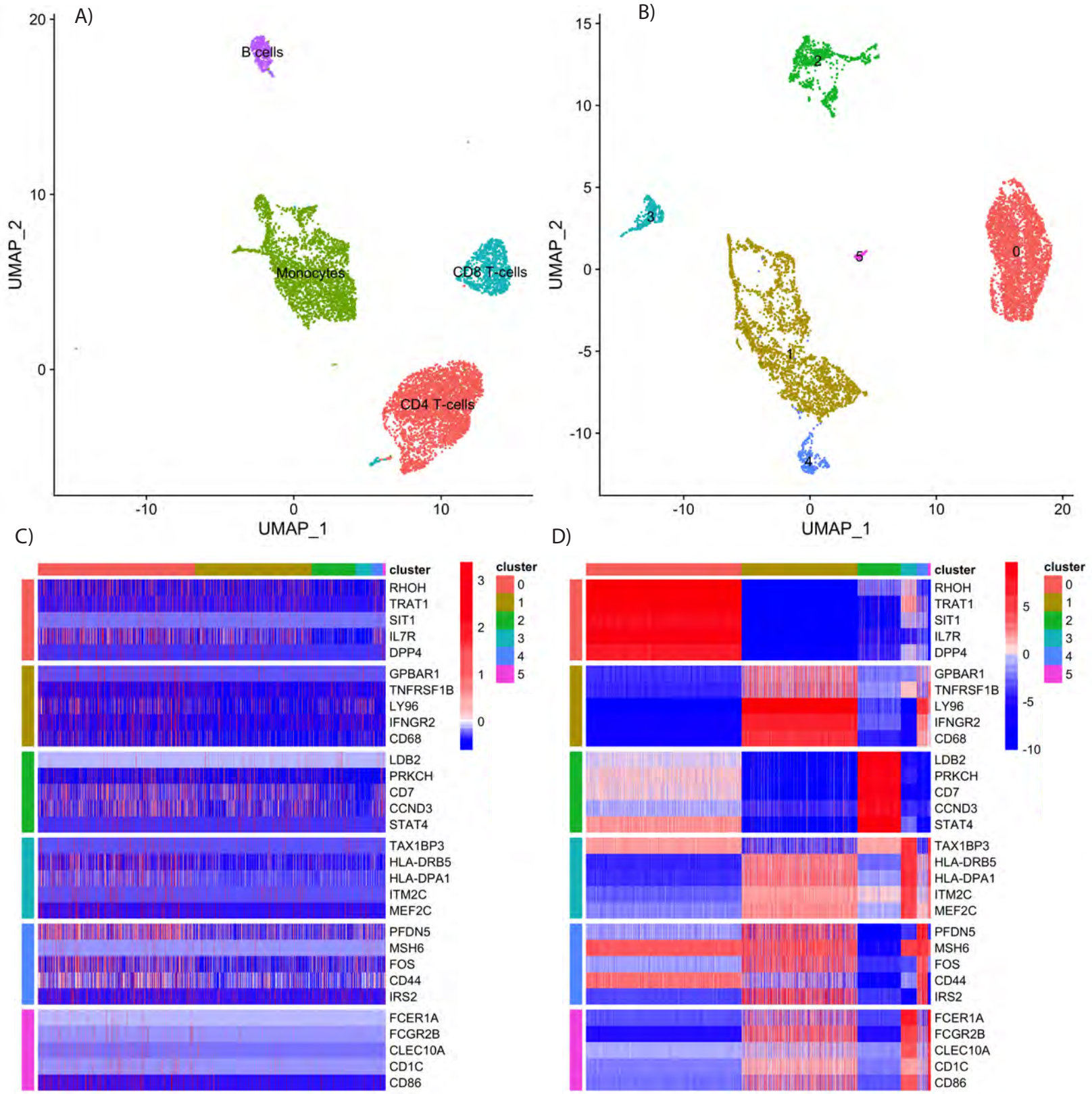
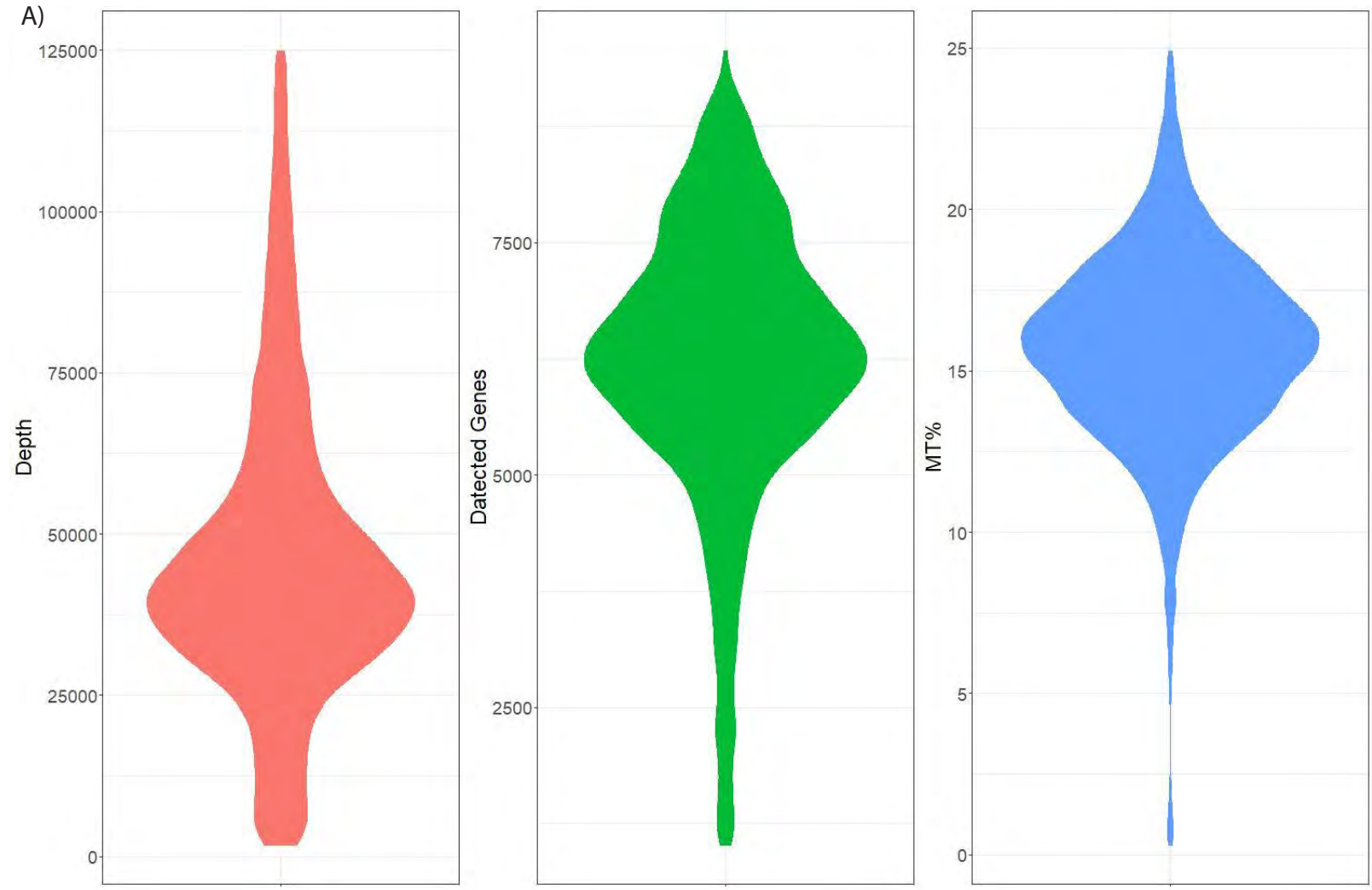
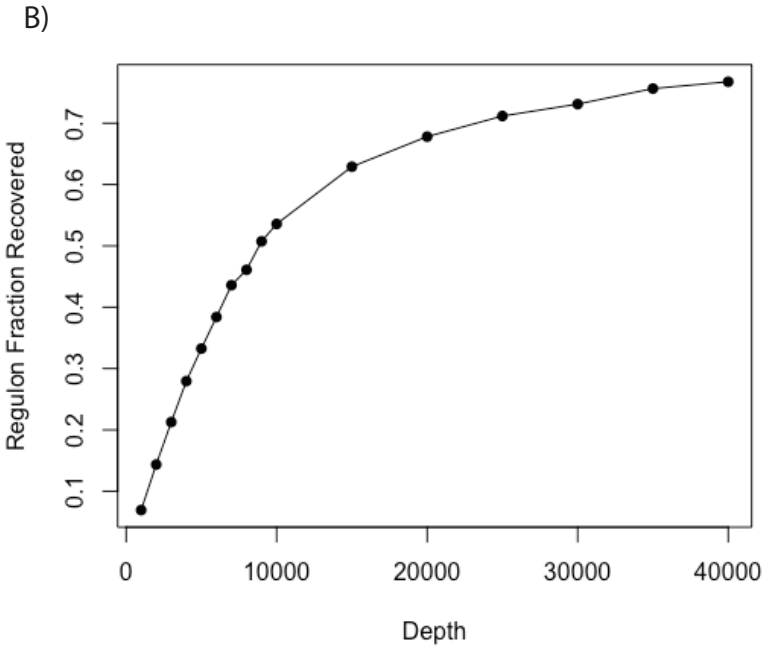

A)
B cells
Monocytes
CD8 T-cells
CD4 T-cells
UMAP_2
UMAP_1
B)
0
1
2
3
4
5
C)
D)
cluster
RHOH
TRAT1
SIT1
IL7R
DPP4
GPBAR1
TNFRSF1B
LY96
IFNGR2
CD68
LDB2
PRKCH
CD7
CCND3
STAT4
TAX1BP3
HLA-DRB5
HLA-DPA1
ITM2C
MEF2C
PFDN5
MSH6
FOS
CD44
IRS2
FCER1A
FCGR2B
CLEC10A
CD1C
CD86

Figure S2


A)
Depth
125000
100000
75000
50000
25000
0
Datected Genes
7500
5000
2500
MT%
25
20
15
10
5
0


B)
Regulon Fraction Recovered
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10000
20000
30000
40000
Depth

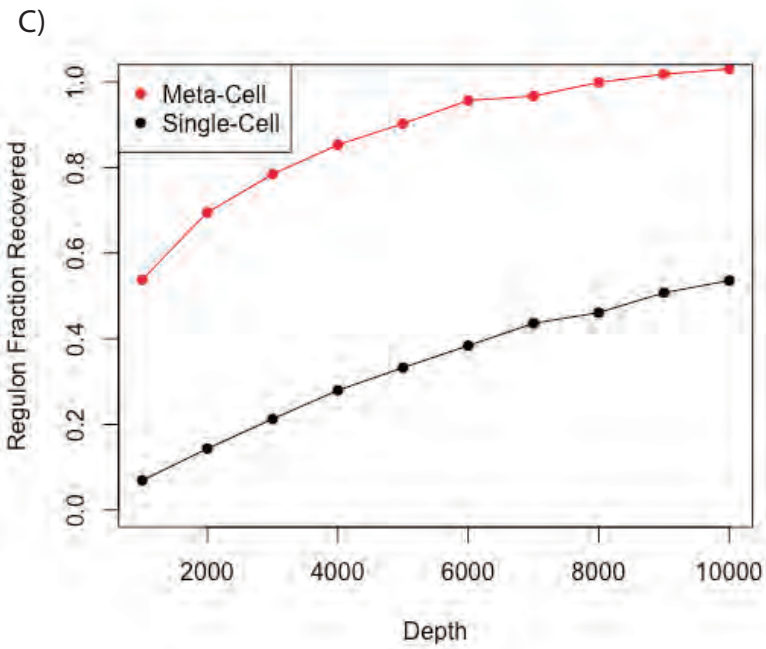

C)
Meta-Cell
Single-Cell
Regulon Fraction Recovered
0.0
0.2
0.4
0.6
0.8
1.0
2000
4000
6000
8000
10000
Depth

Figure S3

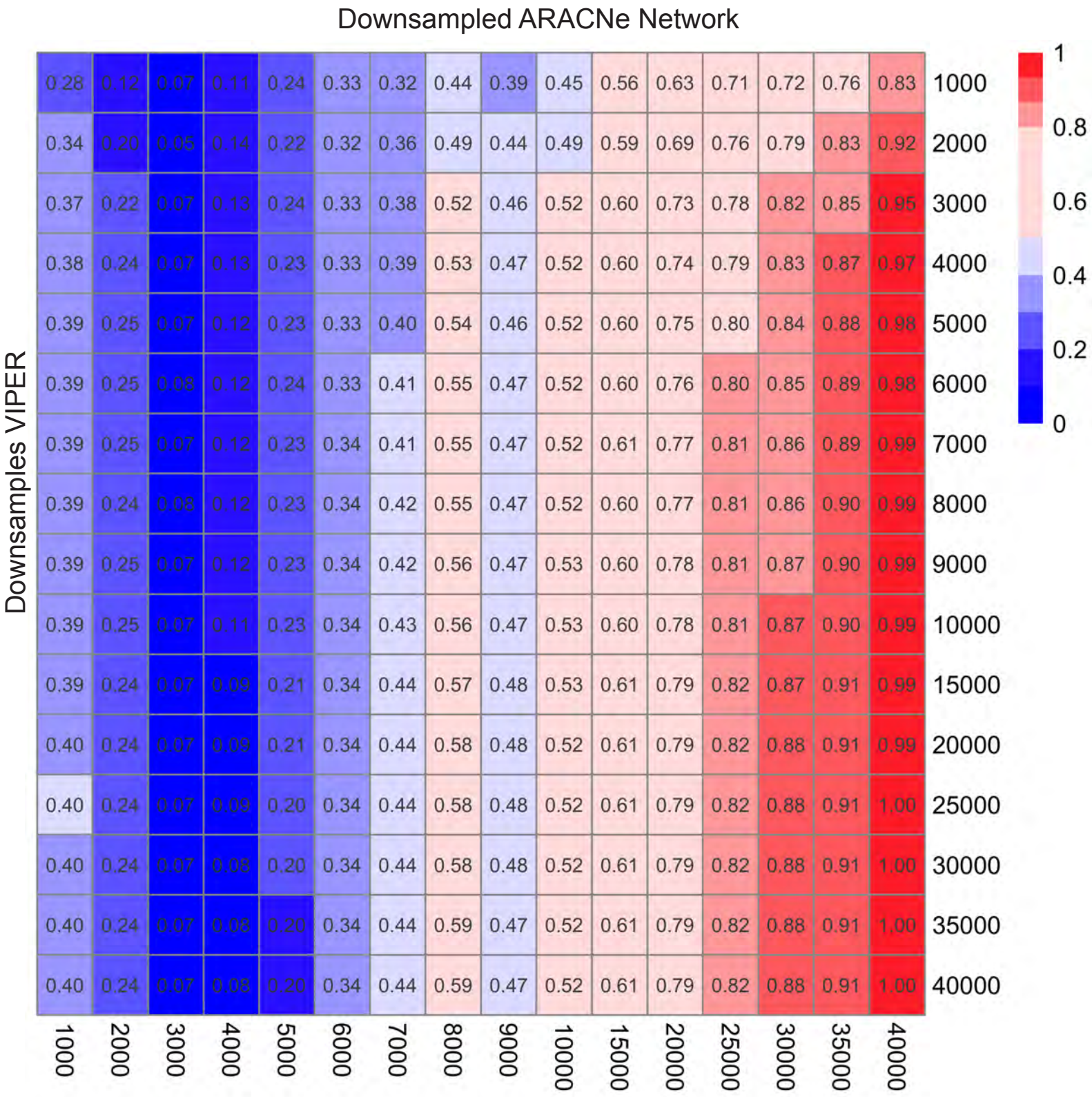

Downsampled ARACNe Network
Downsamples VIPER
1000 0.28 0.12 0.07 0.11 0.24 0.33 0.32 0.44 0.39 0.45 0.56 0.63 0.71 0.72 0.76 0.83
2000 0.34 0.20 0.05 0.14 0.22 0.32 0.36 0.49 0.44 0.49 0.59 0.69 0.76 0.79 0.83 0.92
3000 0.37 0.22 0.07 0.13 0.24 0.33 0.38 0.52 0.46 0.52 0.60 0.73 0.78 0.82 0.85 0.95
4000 0.38 0.24 0.07 0.13 0.23 0.33 0.39 0.53 0.47 0.52 0.60 0.74 0.79 0.83 0.87 0.97
5000 0.39 0.25 0.07 0.12 0.23 0.33 0.40 0.54 0.46 0.52 0.60 0.75 0.80 0.84 0.88 0.98
6000 0.39 0.25 0.08 0.12 0.24 0.33 0.41 0.55 0.47 0.52 0.60 0.76 0.80 0.85 0.89 0.98
7000 0.39 0.25 0.07 0.12 0.23 0.34 0.41 0.55 0.47 0.52 0.61 0.77 0.81 0.86 0.89 0.99
8000 0.39 0.24 0.08 0.12 0.23 0.34 0.42 0.55 0.47 0.52 0.60 0.77 0.81 0.86 0.90 0.99
9000 0.39 0.25 0.07 0.12 0.23 0.34 0.42 0.56 0.47 0.53 0.60 0.78 0.81 0.87 0.90 0.99
10000 0.39 0.25 0.07 0.11 0.23 0.34 0.43 0.56 0.47 0.53 0.60 0.78 0.81 0.87 0.90 0.99
15000 0.39 0.24 0.07 0.09 0.21 0.34 0.44 0.57 0.48 0.53 0.61 0.79 0.82 0.87 0.91 0.99
20000 0.40 0.24 0.07 0.09 0.21 0.34 0.44 0.58 0.48 0.52 0.61 0.79 0.82 0.88 0.91 0.99
25000 0.40 0.24 0.07 0.09 0.20 0.34 0.44 0.58 0.48 0.52 0.61 0.79 0.82 0.88 0.91 1.00
30000 0.40 0.24 0.07 0.08 0.20 0.34 0.44 0.58 0.48 0.52 0.61 0.79 0.82 0.88 0.91 1.00
35000 0.40 0.24 0.07 0.08 0.20 0.34 0.44 0.59 0.47 0.52 0.61 0.79 0.82 0.88 0.91 1.00
40000 0.40 0.24 0.07 0.08 0.20 0.34 0.44 0.59 0.47 0.52 0.61 0.79 0.82 0.88 0.91 1.00
1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 15000 20000 25000 30000 35000 40000
1 0.8 0.6 0.4 0.2 0

Figure S4

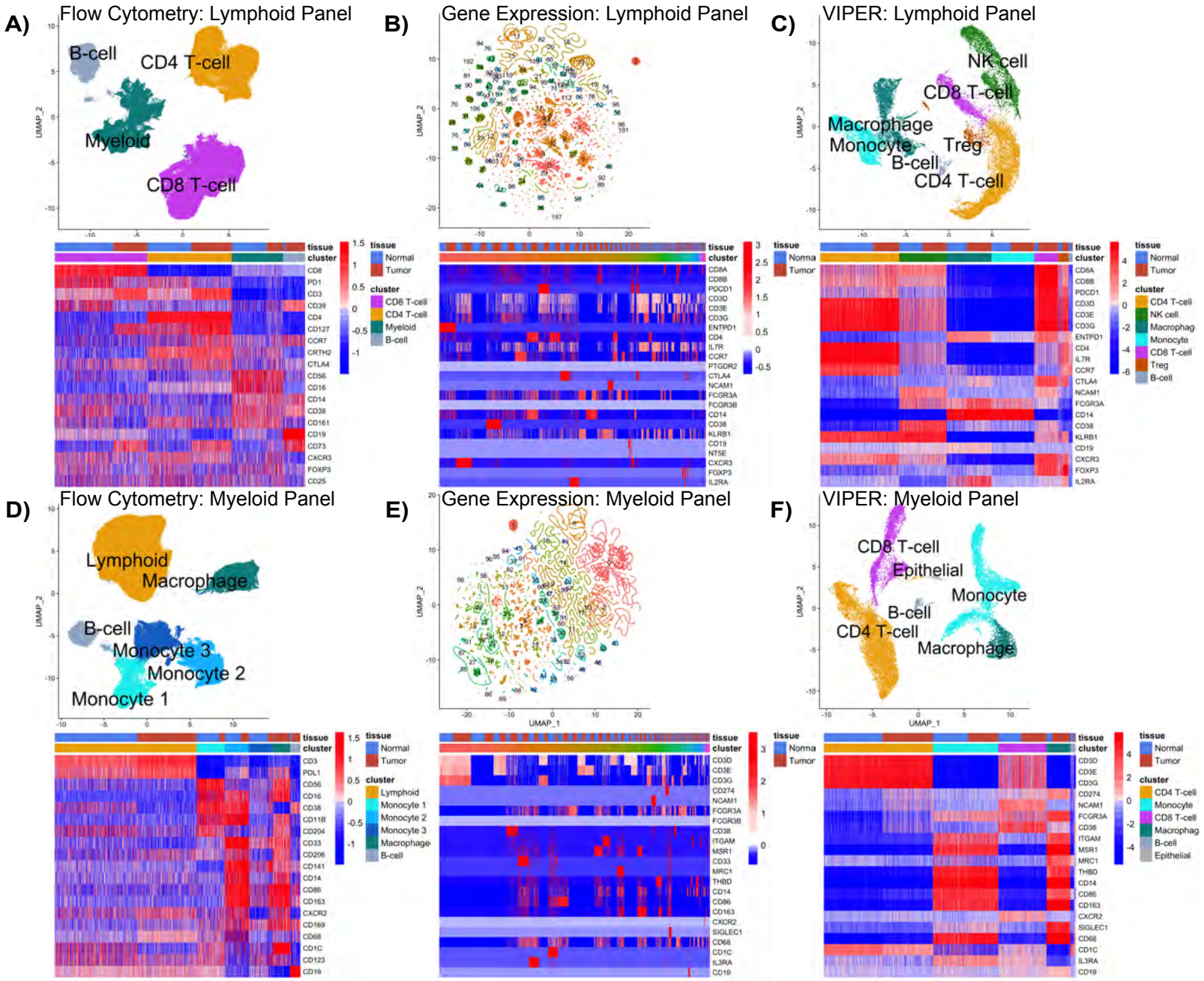
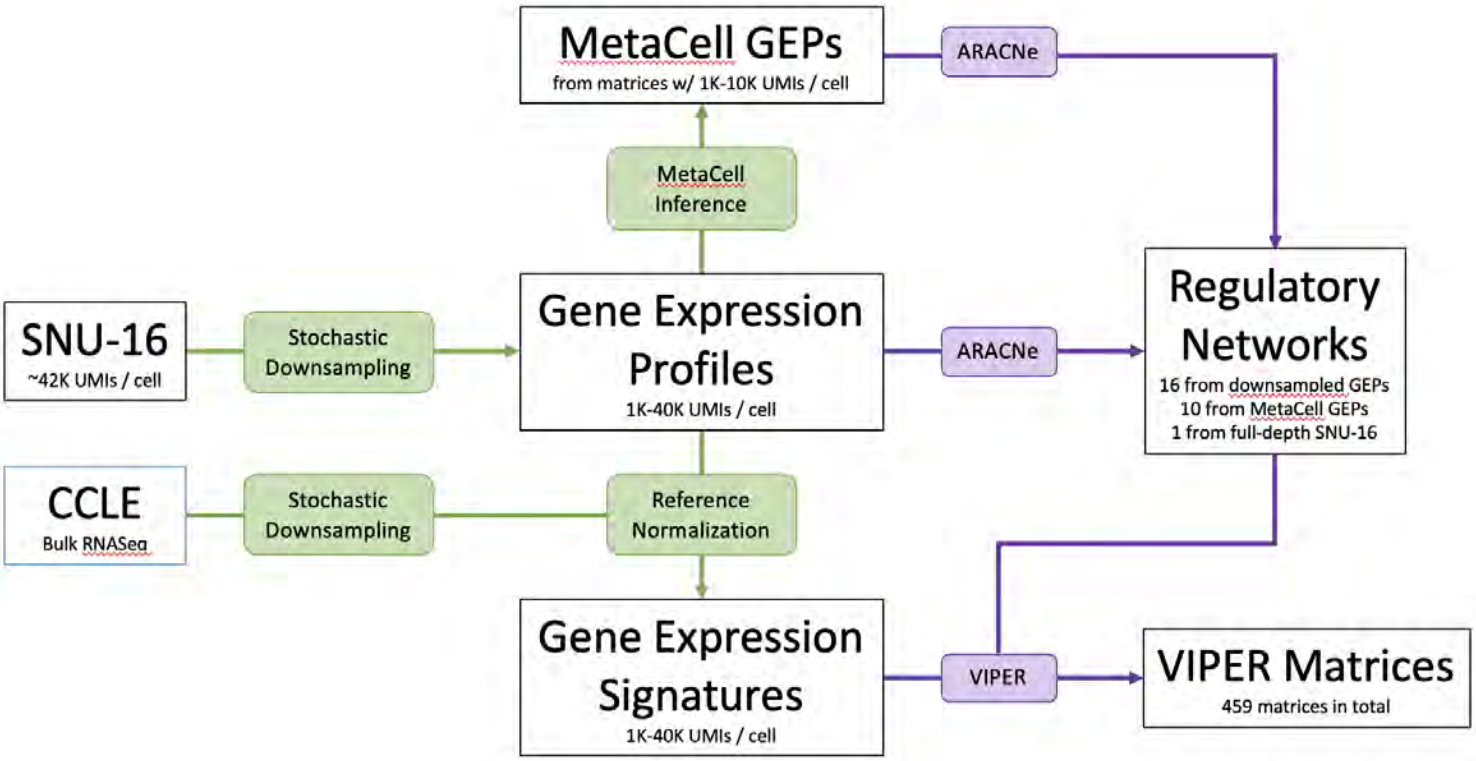

A)
Flow Cytometry: Lymphoid Panel
B-cell
CD4 T-cell
Myeloid
CD8 T-cell
B)
Gene Expression: Lymphoid Panel
C)
VIPER: Lymphoid Panel
NK cell
CD8 T-cell
Macrophage
Monocyte
Treg
B-cell
CD4 T-cell
D)
Flow Cytometry: Myeloid Panel
Lymphoid
Macrophage
B-cell
Monocyte 3
Monocyte 2
Monocyte 1
E)
Gene Expression: Myeloid Panel
F)
VIPER: Myeloid Panel
CD8 T-cell
Epithelial
Monocyte
B-cell
CD4 T-cell
Macrophage
UMAP_1
UMAP_2
tissue
cluster
Normal
Tumor
CD4 T-cell
NK cell
Macrophage
Monocyte
CD8 T-cell
Treg
B-cell
Lymphoid
Monocyte 1
Monocyte 2
Monocyte 3
Macrophage
Epithelial

Figure S5


MetaCell GEPs
from matrices w/ 1K-10K UMIs / cell
ARACNe
MetaCell Inference
SNU-16
~42K UMIs / cell
Stochastic Downsampling
Gene Expression Profiles
1K-40K UMIs / cell
ARACNe
Regulatory Networks
16 from downsampled GEPs
10 from MetaCell GEPs
1 from full-depth SNU-16
CCLE
Bulk RNASeq
Stochastic Downsampling
Reference Normalization
Gene Expression Signatures
1K-40K UMIs / cell
VIPER
VIPER Matrices
459 matrices in total

To-Do: Add Single-Cell Type Atlas