Briefings in Bioinformatics, 2024, **25(6)**, bbae457
https://doi.org/10.1093/bib/bbae457
Problem Solving Protocol

COFFEE: consensus single cell-type specific inference for gene regulatory networks

Musaddiq K Lodi[1], Anna Chernikov[2], Preetam Ghosh[3]*

[1]Integrative Life Sciences, Virginia Commonwealth University, 1000 W Cary St, Richmond, VA 23284, United States
[2]Center for Biological Data Science, Virginia Commonwealth University, 1015 Floyd Ave, Richmond, VA 23284, United States
[3]Department of Computer Science, Virginia Commonwealth University, 401 W Main St, Richmond, VA 23284, United States
*Corresponding author. Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, United States.
E-mail: pghosh@vcu.edu

Abstract

The inference of gene regulatory networks (GRNs) is crucial to understanding the regulatory mechanisms that govern biological processes. GRNs may be represented as edges in a graph, and hence, it have been inferred computationally for scRNA-seq data. A wisdom of crowds approach to integrate edges from several GRNs to create one composite GRN has demonstrated improved performance when compared with individual algorithm implementations on bulk RNA-seq and microarray data. In an effort to extend this approach to scRNA-seq data, we present COFFEE (COnsensus single cell-type speciFic inFerence for gEnE regulatory networks), a Borda voting-based consensus algorithm that integrates information from 10 established GRN inference methods. We conclude that COFFEE has improved performance across synthetic, curated, and experimental datasets when compared with baseline methods. Additionally, we show that a modified version of COFFEE can be leveraged to improve performance on newer cell-type specific GRN inference methods. Overall, our results demonstrate that consensus-based methods with pertinent modifications continue to be valuable for GRN inference at the single cell level. While COFFEE is benchmarked on 10 algorithms, it is a flexible strategy that can incorporate any set of GRN inference algorithms according to user preference. A Python implementation of COFFEE may be found on GitHub: https://github.com/lodimk2/coffee

Keywords: single-cell biology; gene regulatory networks; transcriptional mechanisms; wisdom-of-crowds

Introduction

The study of biological systems is being conducted in several different ways. One popular way of analyzing the relationship between chromatin, transcription factors, and genes is to represent them as a complex network known as the gene regulatory network (GRN). GRNs are crucial to the understanding of how cellular identity is established and corrupted in disease. The popular abstraction for analyzing GRNs is in the form of graphs, in which the relationship between any two genes is quantified by an edge score. The goal of GRN inference is to better understand the gene expression patterns that connect transcription factors and signaling proteins to target genes [1]. Several algorithms have been developed to infer GRNs from bulk RNA-sequencing data [2–4]. However, single cell transcription data present the opportunity to observe cell-type specific gene expression patterns and potentially gain further insight into the regulation of cells [5]. The noise present in scRNA-seq datasets makes it difficult to determine if results are biologically significant. Hence, validation of constructed GRNs through pathway enrichment and literature review become paramount [6].

Algorithms that infer GRNs from bulk RNA-sequencing data have been adapted for single cell transcriptome data, to varying degrees of success. The algorithms available to the community vary in architecture and approach. Some are correlation-based methods, such as LEAP and PPCOR [7, 8]. More recent algorithms rely on linear regression and non-linear ordinary differential equations (ODEs) to make GRN predictions. Additionally, some algorithms require the input of time-point expression data, while others do not. Often times, scRNA-seq experiments do not collect this information, so a common accepted practice has been to generate pseudotime point data with methods such as Slingshot [9]. The increasing number of algorithms for this purpose becomes an insurmountable task for researchers; how should a scientist choose the best algorithm to construct a GRN from single cell transcriptomics data? As scRNA-seq data have become more accessible, Pratapa *et al.* sought to solve this problem by creating an evaluation framework for 12 prominent GRN inference algorithms [2]. Based on the performance of these algorithms on synthetic, curated, and experimental single cell transcriptomic datasets, the authors were able to make the recommendation that the algorithms PIDC, GENIE3, and GRNBoost2 are the methods of choice for researchers seeking to use a GRN inference algorithm [2]. The robustness of these algorithms was measured using the early precision ratio (EPR) score. EPR is a network, which essentially measures the number of true positive interactions in a network [10]. Rather than choosing select algorithms based on an evaluation criteria, another approach to constructing GRNs is to leverage information from all of the present algorithms, using the wisdom of crowds theory.

Received: June 3, 2024. **Revised:** July 22, 2024. **Accepted:** September 2, 2024

The wisdom of crowds theory states that the collective knowledge of a community is greater than the knowledge of any individual. This theory has broad practical applications to a number of fields [11]. Its implementation in GRN inference is not a new concept; a study in 2012 by the DREAM5 Consortium *et al.* used a consensus network approach for bulk RNA-sequencing data [12]. This consensus approach used the Borda count method, which is a ranked choice voting algorithm invented by John Charles de Borda in 1770. The way the system works is that candidates are ranked by the choice of the crowd and are assigned a score accordingly (the candidate ranked in first receives the maximum number of points, and so on). The candidate with the highest average score then wins the election [13]. Building on this platform, we implemented a normalized version of the Borda count system to generate a consensus network approach applied to the inference of miRNA–miRNA networks [1, 14]. This current study seeks to improve this implementation by making it more specific to GRN inference for single cell transcriptomics data.

The consensus approach for GRN inference has demonstrated improved performance for microarray and bulk RNA-seq data [12]. Our motivation for this study was two-fold: we wanted to test if a similar wisdom of crowds approach is effective for scRNA-seq data. GRN inference for scRNA-seq data presents a unique set of challenges. The noise present in pseudotime data, as well as gene/cell dropouts, requires more sensitive algorithms to predict high-quality interactions [15]. With suboptimal performance shown in several GRN inference algorithms, integrating them to achieve increased performance may not necessarily be as intuitive as inference for bulk-RNA sequencing data [2]. Additionally, with the advent of higher quality scRNA-seq and scATAC-seq data, inferring GRNs by individual cell-type has become a paramount task. Determining the individual regulatory interactions per cell-type can lead to the identification of novel cell-types and disease progression mechanisms [16]. We sought to determine if a consensus-based approach can lead to higher performance in GRN inference for specific cell-types. To this end, we present COFFEE, a consensus algorithm for scRNA-seq data, for both cell-type specific and general scRNA-seq data. Since the consensus approach works on the individual networks predicted by each algorithm, it can readily integrate algorithms that leverage scRNA-seq data or even multi-omics datasets (e.g. scRNA-seq and scATAC-seq data). COFFEE differs from previous consensus GRN inference methods as it is the first one to be applied to scRNAseq data, as well as providing a way for individual algorithms or genes to be prioritized in the weighting system.

Overall, our study makes the following contributions:

- Establish that Borda count-based consensus GRN inference is superior to individual algorithms, reaffirming the wisdom of crowds approach in network inference.
- Create a robust framework for consensus inference of GRNs from scRNAseq data, available on GitHub implemented in Python.
- Illimunate the value of a weighted consensus approach for cell-type specific GRN inference.

Materials and methods

Selection of algorithms

The algorithms selected for this study that form the basis for the consensus network construction were from the BEELINE framework proposed by Pratapa *et al.* A summary of these algorithms is presented in Table 1.

We implemented the BEELINE evaluation framework and ran each algorithm through the pipeline provided [2]. Once these networks for each individual algorithm were constructed, we used our proposed COFFEE framework to integrate them into one consensus network.

Borda count implementation

Each algorithm from the BEELINE implementation outputs a ranked edge list with a confidence score attached to each edge. Since each algorithm is normalized in a different way, the edge weights are not distributed equally, so the ranking of each edge within the list across algorithms is considered. For example, consider the ranking of four edges for three different algorithms, each inferring a distinct ranked edge list.

The Borda count method allocates points to each rank, where the highest ranked interaction receives the maximum number of points, and the lowest ranked interaction receives zero points. To receive a final rank between 0 and 1, the resulting weighted ranks are normalized. In the example in Tables 2 and 3, there are four example interactions; the interaction I4 is ranked at the first position for Algorithm 1. Therefore, it receives the maximum of three Borda points and a normalized score of 1. The resulting Borda rank used is the normalized number of points received for each algorithm.

Modifications were made to the original Borda count method in order to make it more applicable for gene-gene relationships. Self loops in the graph were removed, so a gene interacting with itself was removed from consideration in the final consensus algorithm. Additionally, due to high variance amongst algorithms, there were several cases where an edge would appear in some algorithms with a high ranking, and not appear at all in others. To handle this, 0 was inserted where an edge was not present into the calculation and was considered for the Final Rank. We implemented a version of the Borda count algorithm in Python, which takes a directory of ranked lists as inputs, and will output the consensus network for a user specified threshold.

Single cell-type specific GRN inference using multimodal data: scMTNI

A promising direction of GRN inference for single cell transcriptomics data is identifying unique GRN's by cell type. The primary challenge of this approach is identifying accurate cell lineages, and the varying transcription factors that define them [16]. One way to infer high-quality cell lineages, and therefore the important transcription factors by cell type, is to integrate scRNA-seq data with scATAC-seq data [5]. Single Cell Multi-Task Inference (scMTNI) is a recently published multi-task learning framework that integrates scATAC-seq with scRNA-seq data to infer cell-type specific GRNs. It has demonstrated improved performance over existing cell-type specific inference models. [5].

scMTNI is a multi-task learning framework that uses a probabilistic graphical model-based method to infer GRN dynamics from a cell lineage tree. The method defines a cell type as a group of cells with similar transcriptome and accessibility levels. Each cell type is treated as a task, and the goal of the method is to infer a GRN for each task, as well as the ideal parameters. scMTNI calculates the probability of each gene and its set of regulators for each cell type, and uses this probability calculation to inform the predicted network. For our comparative analysis, we used the inferred consensus networks for each cell type in the coarse Human Fetal Hematopoiesis dataset, with four cell types [25]. scMTNI was executed on several subsamples of each cell type, and the final network prediction was informed through a confidence

Table 1. Summary of algorithms used in the BEELINE framework that were leveraged in our proposed consensus network approach

Algorithm Name	Description
GENIE3 [17]	Tree-based ensemble methods such as Random Forest that predict the expression profile of target genes from all other genes. Interaction weights stem from how important an input gene on a target's expression data. GENIE3 has been demonstrated to be a consistent performer across various datasets [2].
PPCOR [8]	Computes partial and semi-partial correlation coefficients for every pair of genes. Ranks are scaled between -1 and 1, supporting inhibitory and activating network inference. PPCOR results in an undirected network. The negative and positive weights (-1 to 1) are meant to signify an activating or inhibitory interaction. PPCOR was demonstrated to be a consistent performer across various types of datasets [2].
LEAP [7]	Lag-based expression association for pseudotime series (LEAP). Calculates Pearson correlation of normalized mapped read counts; final weightage per edge is the maximum Pearson correlation across all lag values. LEAP also contains a permutation based test that assists in decreasing false discovery rates. LEAP outputs a directed network.
SCODE	Implements ODEs for regulatory network representation from gene expression dynamics. Combines linear regression and dimension reduction to improve algorithm efficiency.
PIDC [18]	Partial Information Decomposition and Context (PIDC). Computes pairwise mutual information between two genes. From here, PIDC calculates per-gene thresholds that identify the most important interactions for each gene. PIDC outputs an undirected network.
SINCERITIES [19]	SINgle CEll Regularized Inference using TIme-stamped Expression profileS (SINCERITIES). Linear regression-based model to recover directed regulatory relationships between genes. SINCERITIES uses Granger casualty, which infers the relationship between change in gene expression of TFs from one window of time and its target genes in another window of time. The edges are inferred through partial correlation analyses.
GRNVBEM [20]	GRN Variational Bayesian Expectation-Maximization (GRNVBEM). Implements a Bayesian network model using a first-order autoregressive system to estimate gene fold change at specific times. From here, GRNVBEM uses a Bayesian framework and produces a directed graph with associated signs.
SCRIBE [21]	Uses Restricted Directed Information (cRDI) to measure mutual information between the past state and current state of a target gene based on time-stamped single-cell gene expression data. SCRIBE is further made efficient for larger datasets by using a context likelihood of relatedness algorithm, which removes edges that do not correspond to direct effects from a TF and a target.
GRNBoost2 [22]	Based on GENIE3 framework, improves efficiency through stochastic Gradient Boosting Machine regression. GRNBOOST2 trains a regression model to infer its edges for each gene in the dataset. GRNBOOST2 has been demonstrated to have consistent performance across various types of datasets [2].
GRISLI [23]	Gene Regulation Inference for Single-cell with Linear differential equations and velocity interference (GRISLI). Estimates cell velocity based on changing gene expression data. From here, GRISLI computes the GRN by solving a sparse regression problem that relates the gene expression of each cell.
SINGE [24]	Single-cell Inference of Network using Granger Ensembles (SINGE). Uses kernel-based Granger Causality regression to solve irregularities in time-stamped single-cell genomics data. The inspiration from SINGE came from the fact that pseudotime data for each cell do not take into account the over cell system's dynamic processes.

Table 2. Ranked individual predictions for three example algorithms, e.g. interactions denoted I

Normalized Borda Points (Borda Ranking)	Borda Points	Rank	Alg. 1	Alg. 2	Alg. 3
1	3	1	*I4*	*I2*	*I2*
0.667	2	2	*I2*	*I3*	*I3*
0.334	1	3	*I1*	*I4*	*I1*
0	0	4	*I3*	*I1*	*I4*

Table 3. Final Borda ranks for each example interaction

Interaction	Average of Borda Ranking	Final Rank
I2	(0.667+1+1) / 3	0.889
I3	(0+0.667+0.667) / 3	0.445
I4	(1+0.334+0) / 3	0.445
I1	(0.334+0+0.334) / 3	0.222

score from the subsampling. We used a confidence cut off of 0.8 to calculate the evaluation metrics, as recommended by the authors of scMTNI [5].

Cell-type specific COFFEE framework

The algorithms used for COFFEE have not been specifically optimized for cell-type GRN inference. However, interactions that are predicted from these algorithms may still be valuable in predicting cell-type specific GRNs. To test this theory, we used an adapted version of COFFEE and compared its performance to scMTNI on a cell-type specific dataset.

For cell-type specific GRN inference, we included four algorithms that do not require pseudotime data as input: PPCOR, GENIE3, GRNBOOST2, and PIDC [8, 17, 18, 22]. The reasons for this were two-fold. Firstly, these algorithms were shown to be the top performing and had high stability on experimental datasets [2]. Additionally, since they do not require pseudotime data as input, these algorithms were less likely to be sensitive to poor pseudotime calculation and remove an element of uncertainty from the calculation.

When using COFFEE for cell-type specific GRN inference, we prioritized the information from the well-established cell-type specific GRN inference algorithm, such as scMTNI. scMTNI is superior in inferring cell-type specific GRNs, as it incorporates scRNA-seq and scATAC-seq data [5]. We added scMTNI to the COFFEE pipeline and calculated a consensus network with scMTNI, PPCOR, GENIE3, GRNBOOST2 and PIDC. For this analysis, we modified COFFEE to provide an initial score of 1 for any edges found in the scMTNI pipeline, thus prioritizing the edges found in scMTNI higher than that in the other four algorithms. Then, we evaluated COFFEE+scMTNI on the baseline inferred scMTNI network. Conceptually, this framework is similar to earlier consensus methods for GRN inference where expert knowledge on possible edges is prioritized a priori.

Filtering highly varying genes with Slingshot

The primary goal of Slingshot is to reconstruct pseudotime data based on cell lineages for scRNA-seq datasets. Slingshot organizes the cell into clusters and defines cell lineages based on the potential ordering or changing of cell states. We chose Slingshot to determine the pseudotime data to maintain consistency with the BEELINE evaluation framework; Pratapa *et al.* calculated the pseudotime data for the synthetic, curated, and experimental datasets used in our benchmarking, and we did not deviate from this calculation.

We used Slingshot to filter highly varying genes for cell-type specific GRN inference. We first loaded the gene expression matrix into the BEELINE package and performed the standard pre-processing and dimensionality reduction recommended by the package authors. Then, we calculated the highly varying genes across pseudotime points for one cell lineage, as we were computing varying genes per cell cluster. From here, we selected the top 500 highly varying genes; all *P*-values were <0.01 [9].

Evaluation

To evaluate our COFFEE framework against baseline algorithms, we used precision, recall, and F-score when comparing to the gold standard network.

Precision

Precision is defined as the number of correctly predicted interactions divided by the total number of predicted interactions. It measures how accurate the positive predictions of the algorithm are and calculated as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall

Recall is defined as the ratio of all correctly predicted interactions to all actual positives. It measures how completely the algorithm predicts true interactions as follows:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F-score

F-score is defined as the harmonic mean between precision and recall. It provides a balanced representation of the relationship between precision and recall and calculated as follows:

$$F\ \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

We used F-score to determine the ideal threshold for COFFEE when applied to various dataset sizes.

Data

Synthetic datasets

The primary benefit of using Synthetic datasets is to have a known GRN that has a comparable ground truth for reliable evaluation. Previous methods have used GeneNetWeaver; however, Pratapa *et al.* described limitations to this method [26]. Therefore, they created BoolODE. BoolODE converts a Boolean model to nonlinear ODEs, which assists in capturing the logical relationship between regulators [2]. We directly used the Synthetic datasets provided by the authors of BEELINE without any additional preprocessing. The BEELINE authors created six datasets based on different network structures: Linear (LI), Cycle (CY), LL (Linear Long), BF (Bifurcating), BFC (Bifurcating Converging), and TF (Trifurcating). For our evaluation of COFFEE, we aggregated the datasets by number of cell-types, rather than the network structure. The cell-group sizes were 100, 200, 500, 2000, and 5000 cells. Each group contained 50 individual datasets with a varying number of genes, each with their own ground-truth network. We used the Synthetic datasets to determine the ideal threshold for the consensus algorithm implemented in COFFEE.

Curated datasets

Curated datasets are published Boolean models for GRNs that capture the specific regulatory processes of a given developmental process. For our evaluation, we selected three of the four Curated datasets used in the BEELINE evaluation framework: ventral spinal cord (VSC) development, hematopoietic stem cell (HSC) differentiation, and gonadal sex determination (GSD) [27–29]. The authors of the BEELINE framework used BoolODE to create 10 simulated datasets with 2000 cells based on the cell trajectories and gene expression patterns of the original Boolean models. We used the gene expression matrix and pseudotime data as is from the BEELINE data download, without any additional preprocessing.

The BEELINE authors also used the mammalian cortical area development (mCAD) Boolean model dataset in their evaluation framework; however, the authors noted that the algorithm results for this model were outliers, with poor performance results across all algorithms, which differed for the other Curated datasets [2, 30]. Therefore, we opted not to include mCAD in our evaluation process.

Experimental datasets

To evaluate COFFEE on experimental methods, we chose two from human cells and two from mice cells. Similar to the Synthetic and Curated datasets, we used the preprocessed data from the BEELINE framework [2].

To compute the GRN inference using the important genes, we used the gene ordering file computed by the GAM R package to select the 500 top genes varying across pseudotime points, as detailed in the BEELINE evaluation protocol. This gene ordering file was provided in the dataset download from BEELINE. [2].

We also utilized the ground truth networks provided by the authors of BEELINE per experimental dataset cell type [2]. These ground truth sets were obtained from the ENCODE, ChIP-Atlas, and ESCAPE databases for ChIP–seq data from the same or similar cell type. For our evaluation, we only considered interactions that contained genes present in the ground truth networks.

Table 4. Experimental datasets used to evaluate COFFEE

Dataset	Description
hHEP [31]	scRNA-seq experiment on induced pluripotent stem cells (iPSC). Contains 425 scRNA-seq measurements across various timepoints. Pseudotime was calculated using Slingshot with Day 0 as the starting cluster and Day 21 as the ending cluster [9].
hESCs [32]	Timecourse scRNA-seq experiment from 758 cells using the differentiation protocol in order to produce definitive endoderm cells from human embryonic stem cells, measured at 0, 12, 24, 36, 72, and 96 h. Using slingshot, the pseudotime was calculated using 0 h as the starting cluster and 96 h as the ending cluster [9].
mHSCs [33]	Contains normalized expression data for 1656 HSPCs across 4773 genes. Pseudotime was computed using Slingshot across three lineages, which were erythroid, granulocyte-monocyte, and lymphoid [9]. The GRN for each lineage was separately inferred.
Mouse embryonic stem cells (mESC) [34]	scRNA-seq expression measurements for 421 primitive endoderm (PrE) cells differentiated from mESCs, from five diferent time points: 0, 12, 24, 48, and 72 h. Pseudotime was calculated using Slingshot, with the starting cluster being 0 h and the ending cluster being 72 h [9].

Table 5. Number of cells per cell type in Human Fetal Hematopoiesis Cell dataset

Cell-Type	Number of Cells
GPs-Granulocytes	443
HSC-MPP	1367
LMP	1522
MEMP	1522

Table 6. Optimal threshold by F-Score for Synthetic dataset sizes

Dataset Size	Optimal Threshold
100 Cells	0.75
200 Cells	0.75
500 Cells	0.65
2000 Cells	0.65
5000 Cells	0.65

Cell-type–specific dataset

To evaluate COFFEE's performance on cell-type–specific lineages, we used a dataset that scMTNI was benchmarked on. We compared the performance of COFFEE, scMTNI, and COFFEE+scMTNI with a published scRNA-seq and scATAC-seq dataset of human fetal hematopoiesis cells. This study captured specifications for various blood lineages. We considered the coarse resolution of study to test the model on larger datasets, which contained four cell-types: g HSC, multipotent lymphoid-myeloid progenitors (LMPs), MK-erythroid-mast progenitors (MEMPs), and granulocytic progenitors (GPs). The authors of scMTNI provided their preprocessed data per cell-type, in addition to the networks inferred by their method.

To use the data on COFFEE, we followed a similar process to our Experimental dataset evaluation by selecting the top 500 genes across pseudotime points using Slingshot [9]. To compare to scMTNI, we used the networks inferred by scMTNI provided in the author's dataset download, using a confidence cutoff of 0.8, which is the recommendation of the authors of scMTNI [5].

Results

To evaluate the performance of the consensus algorithm using the Borda algorithm, we tested it on four different kinds of datasets: Synthetic, Curated, Experimental, and Cell-Type–Specific inference. In each case, we demonstrate that the wisdom of crowds approach leads to better performance across datasets. To evaluate, we used precision, recall, and F-score.

Synthetic datasets

The Synthetic datasets were obtained from the Beeline evaluation framework [2]. We grouped the datasets by size, to evaluate the performance of the consensus algorithm as more genes and cells are present in a given dataset. There were five size groups present in the Synthetic datasets: 100, 200, 500, 2000, and 5000 cells. The number of genes varied depending on a specific dataset within the size group.

A key component of the consensus algorithm is determining a threshold at which to keep high confidence edges. A similar Borda-based method for miRNA networks, miRsig used a default threshold cut off of 90%, which results in keeping the top 10% of predicted edges [1]. However, due to the cell to cell gene variation in gene expression present in single cell genomics data, we tested the algorithm on lower thresholds and evaluated its performance [35]. We used the mean F-Score to determine the ideal threshold value by dataset size [1].

In Fig. 1, we see that a different consensus threshold is appropriate depending on the size of the dataset. For the smaller datasets (100 or 200 cells), a threshold value of 0.75 leads to the best F-Score performance for COFFEE. For larger datasets (500, 2000, and 5000 cells), a threshold value of 0.65 leads to the best F-Score. Users may decide to maximize precision or recall rather than F-Score, which would lead to a different threshold being used. We found that increasing the threshold value increases the precision.

Table 6 shows the optimal threshold by F-Score for each dataset size. These thresholds were used for the evaluation of COFFEE against the baseline algorithms, as well as reporting the performance on Curated, Experimental, and Cell-Type–Specific datasets.

To evaluate the performance of COFFEE against the baseline algorithms, we primarily used F-Score, precision, and recall as the metrics.

Figure 2 depicts the performance of COFFEE against the baseline algorithms by F-Score. We observe that across dataset sizes, COFFEE demonstrates a better performance. It is also significant to note that algorithms perform differently based on the data sizes. For example, SINCERITIES has a comparitively weaker F-Score for smaller datasets containing 100 or 200 cells than it does with 500, 2000, or 5000 cells. A consensus-based approach such as COFFEE mitigates this variance by integrating information

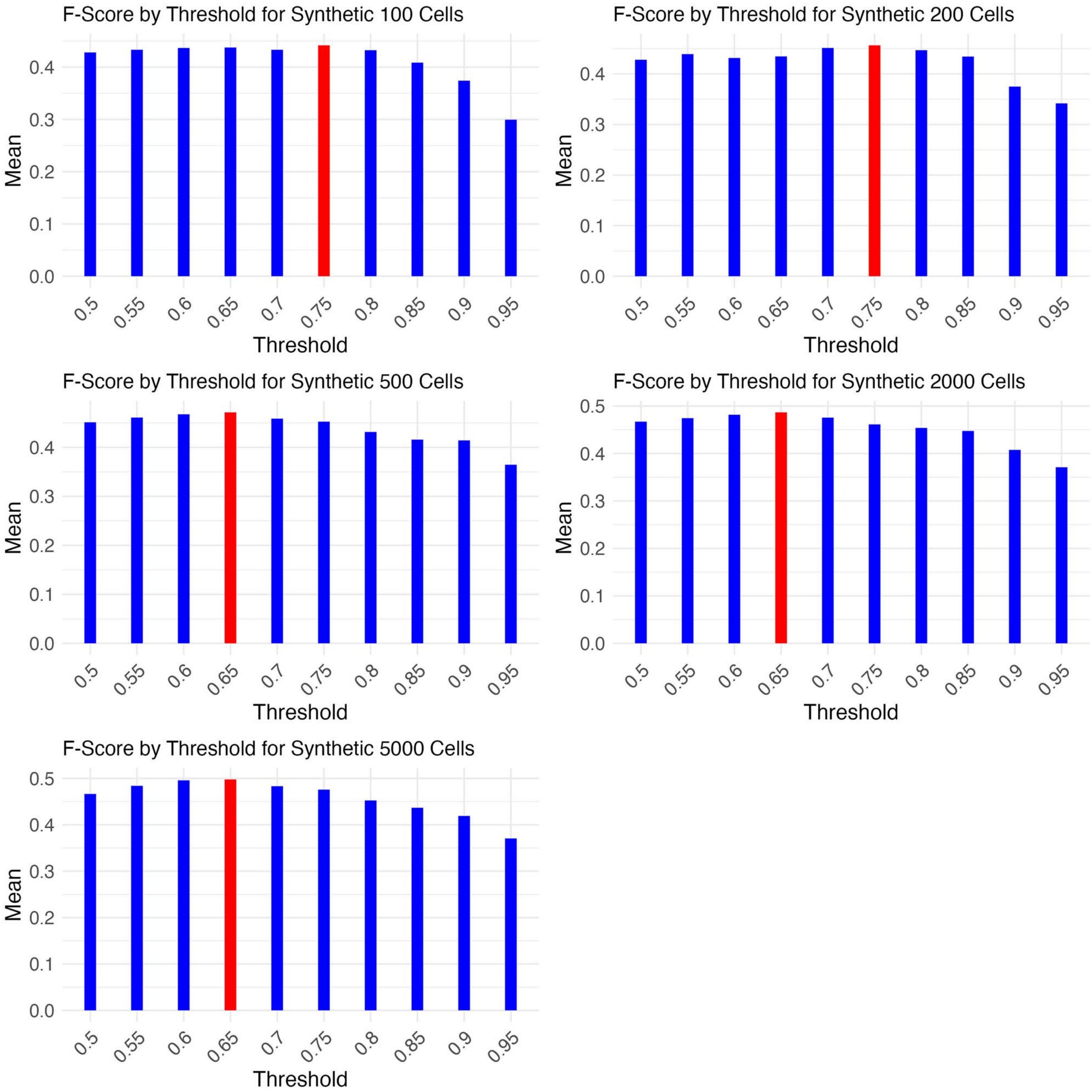


Figure 1. Optimal threshold for COFFEE based on dataset size, by mean F-Score; across Synthetic datasets, the threshold yielding the best mean F-score for each size grouping was selected as the ideal threshold for future COFFEE experiments, and the highest mean F-Score is colored in red.

from the top performing algorithms despite dataset size. In short, COFFEE is less susceptible to variation in its performance based on differences present in a dataset.

To analyze the performance of the algorithms in further detail, we also looked at the mean precision and recall for each dataset size group across algorithms. From the analysis in Fig. 3, we see that the precision in COFFEE is much higher than in any of the other base line algorithms, while its recall is much lower. In Supplementary Figs S4 and S5, we evaluated COFFEE using AUPRC and AUROC as well.

Curated datasets

We next evaluated the performance of COFFEE on the Curated datasets from the BEELINE evaluation framework. Each dataset contained 2000 cells, so the threshold used for COFFEE was the previously identified optimal one of 0.65. We evaluated COFFEE's performance using precision, recall, and F-Score.

The precision-recall analysis for Curated datasets can be seen in Fig. 4. Similar to the Synthetic sets, COFFEE showcases higher precision compared with the baseline algorithms in three of the four datasets. This demonstrates COFFEE's stability across differing datasets; even with Curated data, the consensus approach is able to capture high confidence edges when compared with the ground truth data.

We further explored COFFEE's performance by F-score against the baseline algorithms. These results are visualized in Fig. 5. We can observe that COFFEE performs very well compared with the baseline algorithms in terms of F-score in three of the four datasets. Much like the Synthetic datasets, we see high variation in the baseline algorithms performance, even when applied to

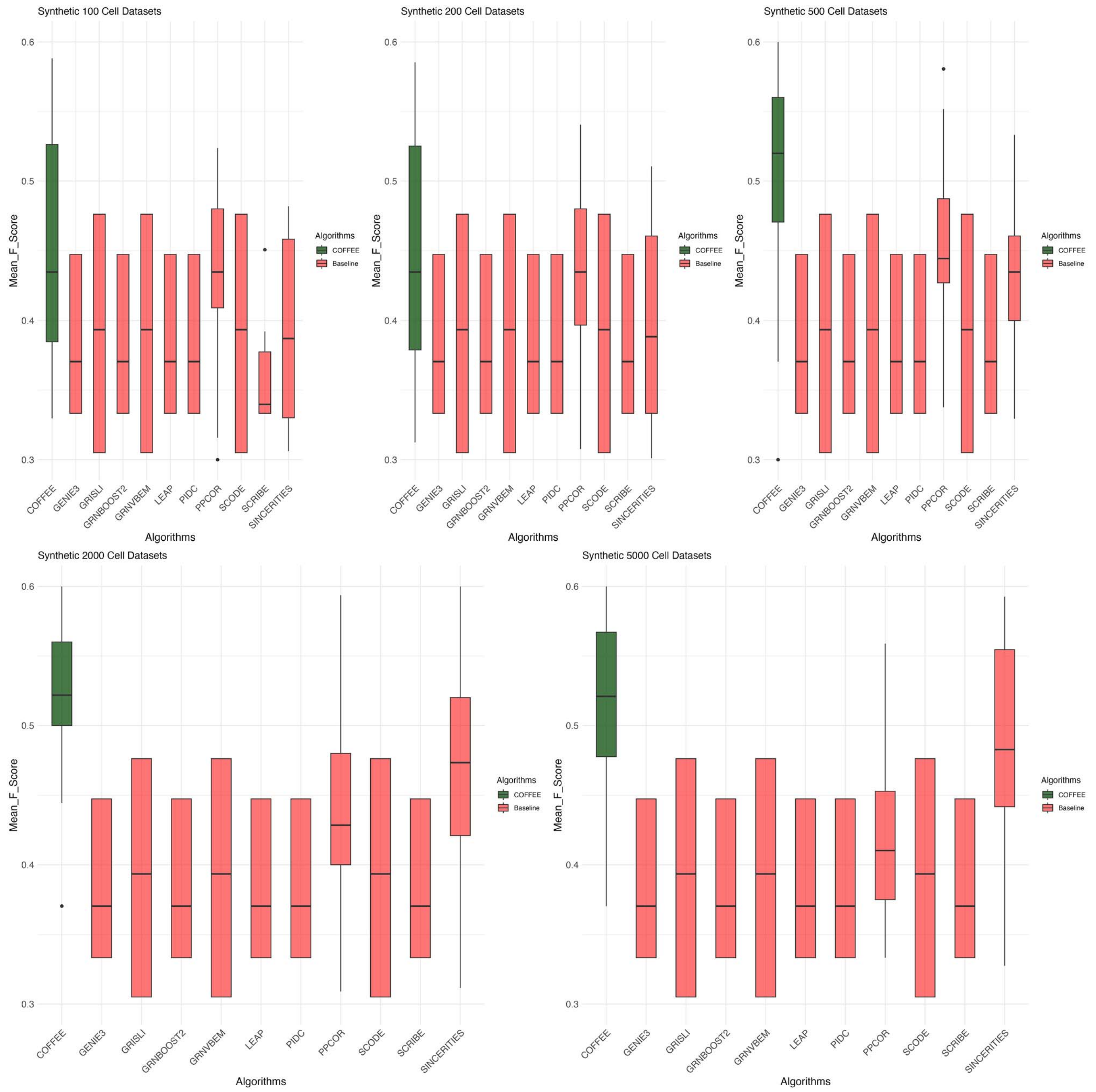


Figure 2. Boxplot of F-Score performance on Synthetic datasets grouped by size; from this analysis, it is clear that COFFEE has superior performance when compared with the individual algorithms making up the consensus, across dataset sizes, and COFFEE is colored in green.

Curated sets. For example, SCODE performed better than most other algorithms in the GSD Curated dataset but performed the worst in the VSC set. COFFEE was able to perform better than most other algorithms in three of the four datasets. In Supplementary Figs S6 and S7, we evaluated COFFEE using AUPRC and AUROC.

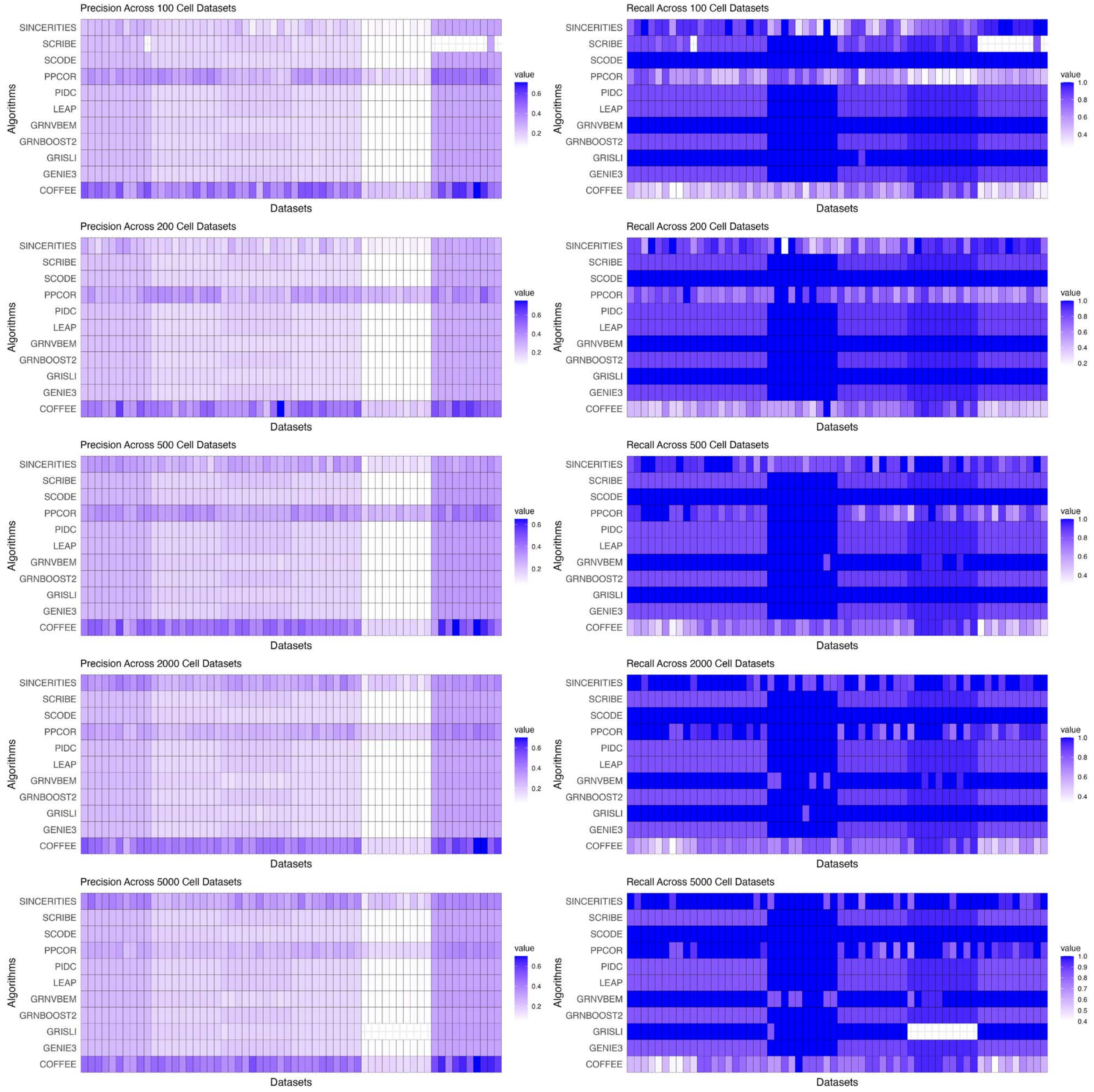
Experimental datasets

Furthermore, we evaluated the performance of COFFEE on four of the Experimental datasets from the BEELINE evaluation framework. Each dataset contained a variable number of cells, and so we used a threshold of 0.65. We evaluated COFFEE's performance using precision, recall, and F-score.

To be consistent with BEELINE's evaluation framework, we evaluated the Experimental datasets on cell-type–specific and non-specific networks. All datasets were collected from Chip-Seq protocol, as outlined in BEELINE [2].

Figures 6 and 7 display the results of COFFEE compared with the baseline algorithms. We noticed that some algorithms did not predict any edges for certain datasets, while they did predict some edges for the others. Therefore, each COFFEE run contained a differing number of algorithms across datasets.

Across the majority of datasets and both the cell-type–specific ground truth and non-specific ground truths, we see that COFFEE performs better than the baseline algorithms in terms of F-score. However, for the mDC dataset, COFFEE performed poorly; the BEELINE evaluation framework also noted that mDC was an outlier in their performance evaluations, with several algorithms demonstrating worse than normal performance [2]. This could be due to mDC having a much higher density in their


Precision Across 100 Cell Datasets

Recall Across 100 Cell Datasets

Precision Across 200 Cell Datasets

Recall Across 200 Cell Datasets

Precision Across 500 Cell Datasets

Recall Across 500 Cell Datasets

Precision Across 2000 Cell Datasets

Recall Across 2000 Cell Datasets

Precision Across 5000 Cell Datasets

Recall Across 5000 Cell Datasets

Figure 3. Precision and recall performance on Synthetic datasets grouped by size; across dataset sizes, COFFEE demonstrates improved precision, but less comparable recall than the individual algorithms.

gold-standard network, making it difficult for algorithms to identify high confidence edges.

We evaluated COFFEE against the baseline algorithms by the EPR metric, and these results are shown in Supplementary Fig. S2. Furthermore, we evaluated COFFEE on the Experimental datasets using AUPRC and AUROC, and these results are shown in Supplementary Tables S1 and S2, respectively.

Contribution of individual algorithms to COFFEE

When working with ensemble algorithms such as COFFEE, an important step in the workflow is deciding which, or how many, algorithms to include in the final consensus aggregation. For consensus GRNs on Bulk RNA-Sequencing data, including as many algorithms as possible lead to the best results [12]. By default, COFFEE includes all algorithms fed into the pipeline and weighs them equally.

To test if other combinations of algorithms would outperform the case of including all possible algorithms into the pipeline as done currently, we performed an experiment by dropping one algorithm from COFFEE and evaluate its performance by F1 score on the Synthetic datasets from BEELINE [2]. The results of this experiment are shown in Fig. 8. From this analysis, we see that the performance is generally even across dataset sizes. For the larger dataset sizes, such as 2000 and 5000 cells, excluding algorithms yielded better performance when compared with including all 10 available algorithms. This result demonstrates that it is possible to have improved results when including less algorithms, which was not the case in Bulk RNA-Seq consensus GRNs [12].

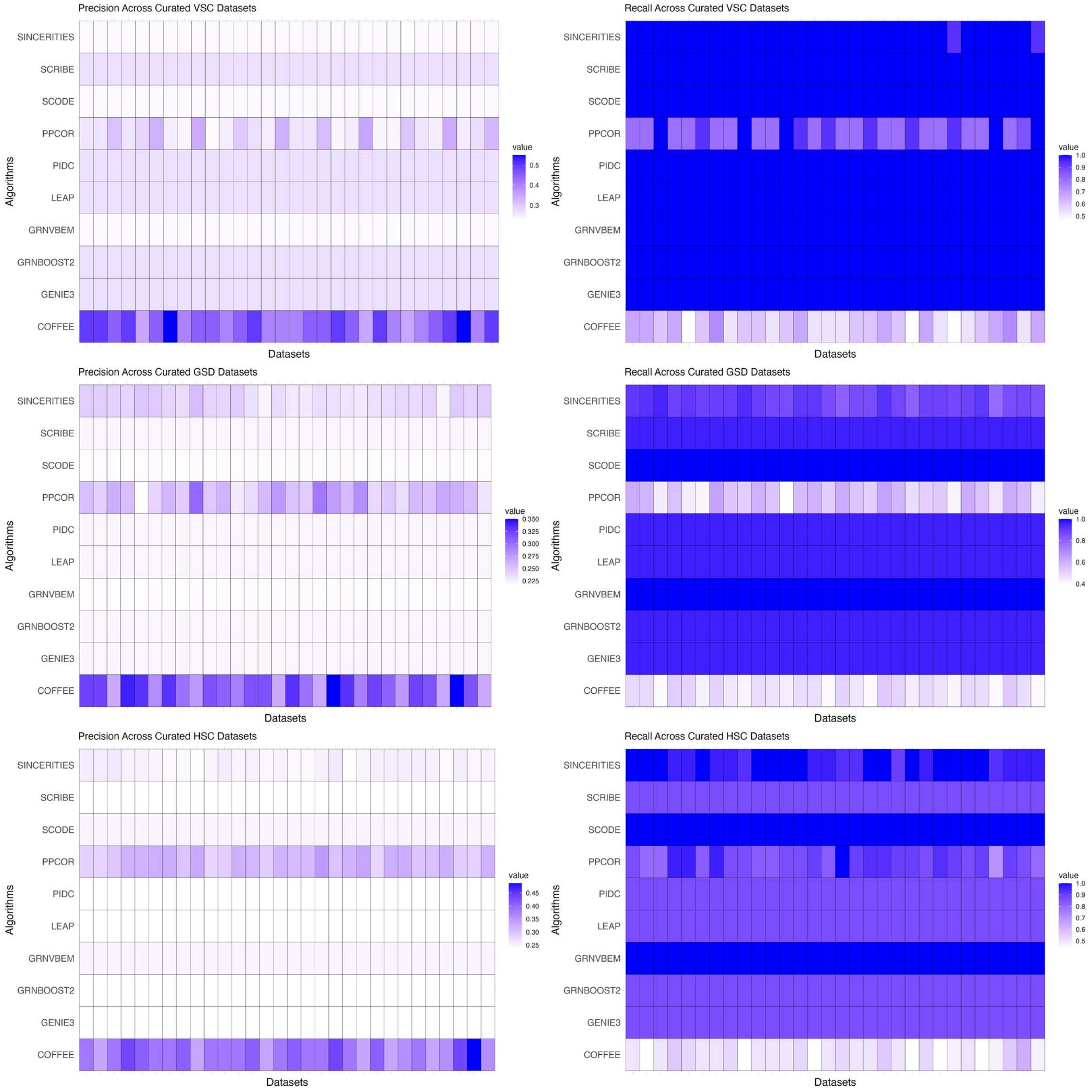


Figure 4. Precision and recall performance on Curated datasets; across Curated datasets, COFFEE demonstrated improved precision but less comparable recall.

Another key question is to identify which algorithms should be included in the final consensus pipeline, if not all. To determine this, we evaluated the effect of dropping an algorithm on the Fscore performance, when compared with including all algorithms. The result of this experiment is shown in Fig. 9. From this analysis, we see that across groups of dataset sizes, the impact of removing a singular algorithm is consistent. There is no conclusive combination of fewer algorithms that would improve the full consensus results with 10 algorithms included considering different datasets and their sizes. In Supplementary Fig. S1, we demonstrate the variation of performance when considering an 8 algorithm combination (i.e. dropping two algorithms rather than just one).

Cell-type–specific inference

Existing GRN inference methods for scRNA-seq data have not been specifically designed for cell-type specific datasets and networks. These methods do not consider the global regulatory structure of cell-types within tissue and therefore may not be representative of key regulatory interactions [16]. To test the consensus method for cell-type–specific inference, we first tried a two-level consensus approach.

On each of the Synthetic datasets used to evaluate the performance of COFFEE, we partitioned the dataset into cell types using Seurat clustering and filtered the gene expression matrices and pseudotime data accordingly [36]. From here, we ran COFFEE with all 10 baseline algorithms; once the networks were obtained

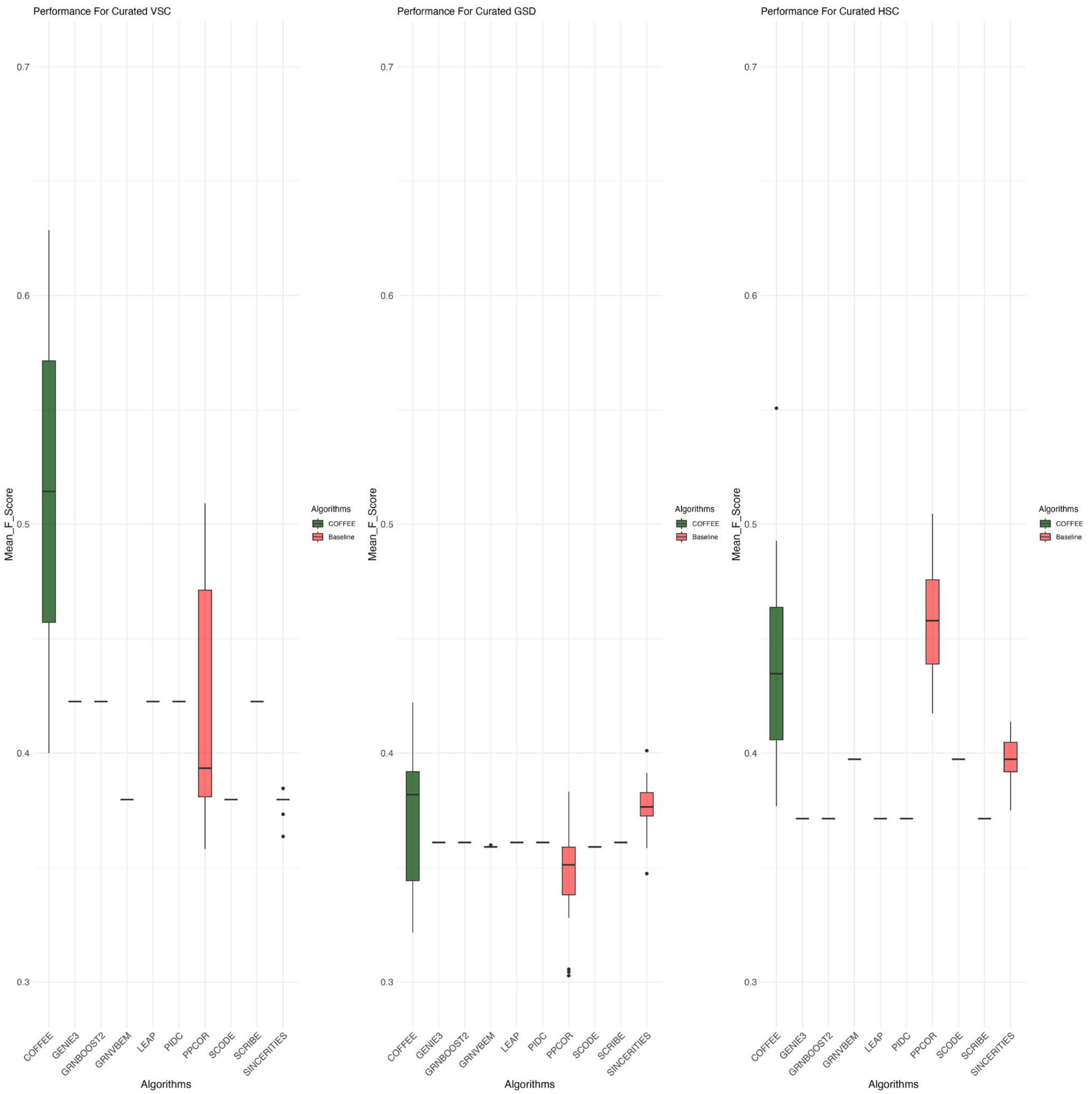


Figure 5. F-score performance on Curated datasets, and COFFEE demonstrated improved performance across two of the three Curated datasets in terms of F-score.

for each cell type, we ran the Borda point algorithm again, this time integrating edges from each cell type to have one composite network for the dataset. We then compared this second level consensus network to the COFFEE algorithm ran on the whole Synthetic datasets without cell-type partitioning.

We see in Fig. 10 that using the second level consensus approach drastically decreases the performance of COFFEE. We conclude from this analysis that the networks from individual cell-types are not entirely representative of the true global regulatory structure; therefore, a different approach is required for cell-type specific inference that takes into account cell-lineage information [5].

To analyze the performance of COFFEE against scMTNI for cell-type–specific performance, we used a dataset that scMTNI was initially benchmarked on [5]. The dataset is from a Human Fetal Hematopoietic cell study that studied the regulatory dynamics during human development for multiple human blood cell types [25]. In concordance with scMTNI's benchmarking process, we used the annotated lineages clusters, which were: HSC and Multi-Potent Progenitor (HSC-MPP), MKerythroid-mast progenitors combined with cycling MEMPs, GPs, and LMPs [5].

We used the single cell expression matrix provided by the authors of scMTNI to run the four algorithms in COFFEE. Similar to the BEELINE framework, we calculated the 500 most varying genes across pseudotime points using Slingshot per cell type [9], using one lineage. These 500 genes were used to infer the GRNs for COFFEE.

For evaluation, we used the Cus-KO gold standard network described in scMTNI [5]. We chose this dataset to be gold standard since it was the network that scMTNI had the best performance

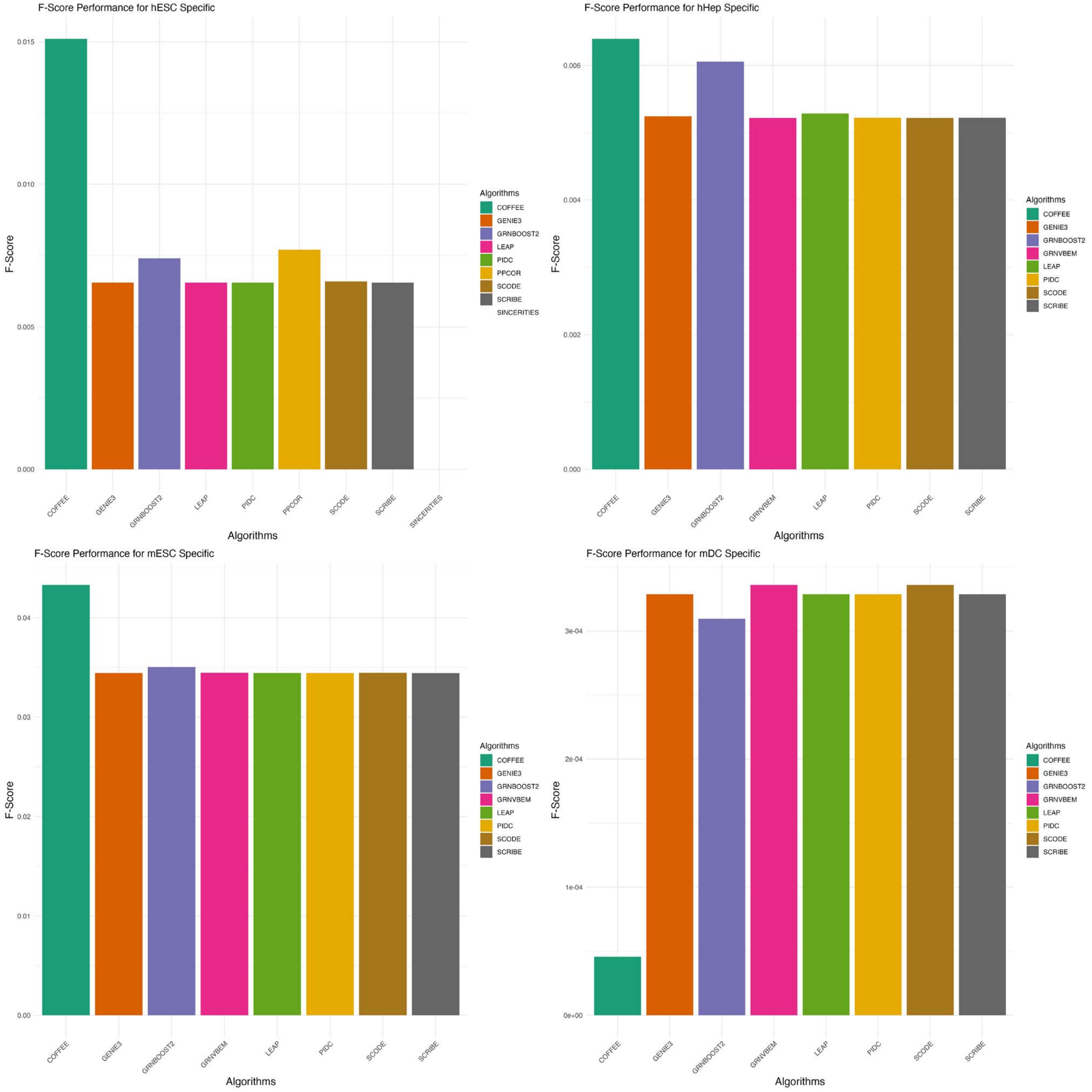


Figure 6. F-score performance on Experimental datasets using Cell-Type–Specific Ground Truth data; COFFEE demonstrated improved performance in terms of F-score in three of the four Experimental datasets.

metrics with [5]. Cus-KO contains interactions from the knockdown-based GM12878 lymphoblastoid cell line downloaded from Cusanovich *et al.* [37]. We filtered the gold-standard network to only contain interactions with a *P*-value <0.01. Additionally, we filtered the inferred networks from both scMTNI and COFFEE to contain only genes present in the gold standard network. Finally, we selected the top 1000 edges from the scMTNI inferred networks and performed a sensitivity analysis for the best COFFEE threshold.

In Fig. 11, we see the performance of COFFEE for various thresholds when compared with scMTNI. In two cell types, LMP and MEMP, COFFEE showcases a better performance by F-score. However, scMTNI performes significantly better on the GPs and HSC-MPP cell types. We also noted that there is little to no effect with the COFFEE threshold on the performance of the algorithm.

With the cell-type-specific prioritized algorithm, which we call COFFEE+scMTNI, we see that a consensus approach is able to improve the performance of scMTNI on all cell types. Thus, we can determine that the individual four algorithms predict true interactions that were not initially learned by scMTNI. We evaluated COFFEE+scMTNI on the Human Fetal Hematopoietic as described earlier [25]. The results of this analysis are shown in Fig. 12.

Case study: COFFEE uncovers key regulatory interactions in breast cancer T cells

In order to demonstrate the practical usability of COFFEE, we analyzed a dataset containing 6311 T Cells isolated from human breast cancer tissue. The study by Savas *et al.* revealed heterogeneity within the breast cancer T Cell population. Subclustering

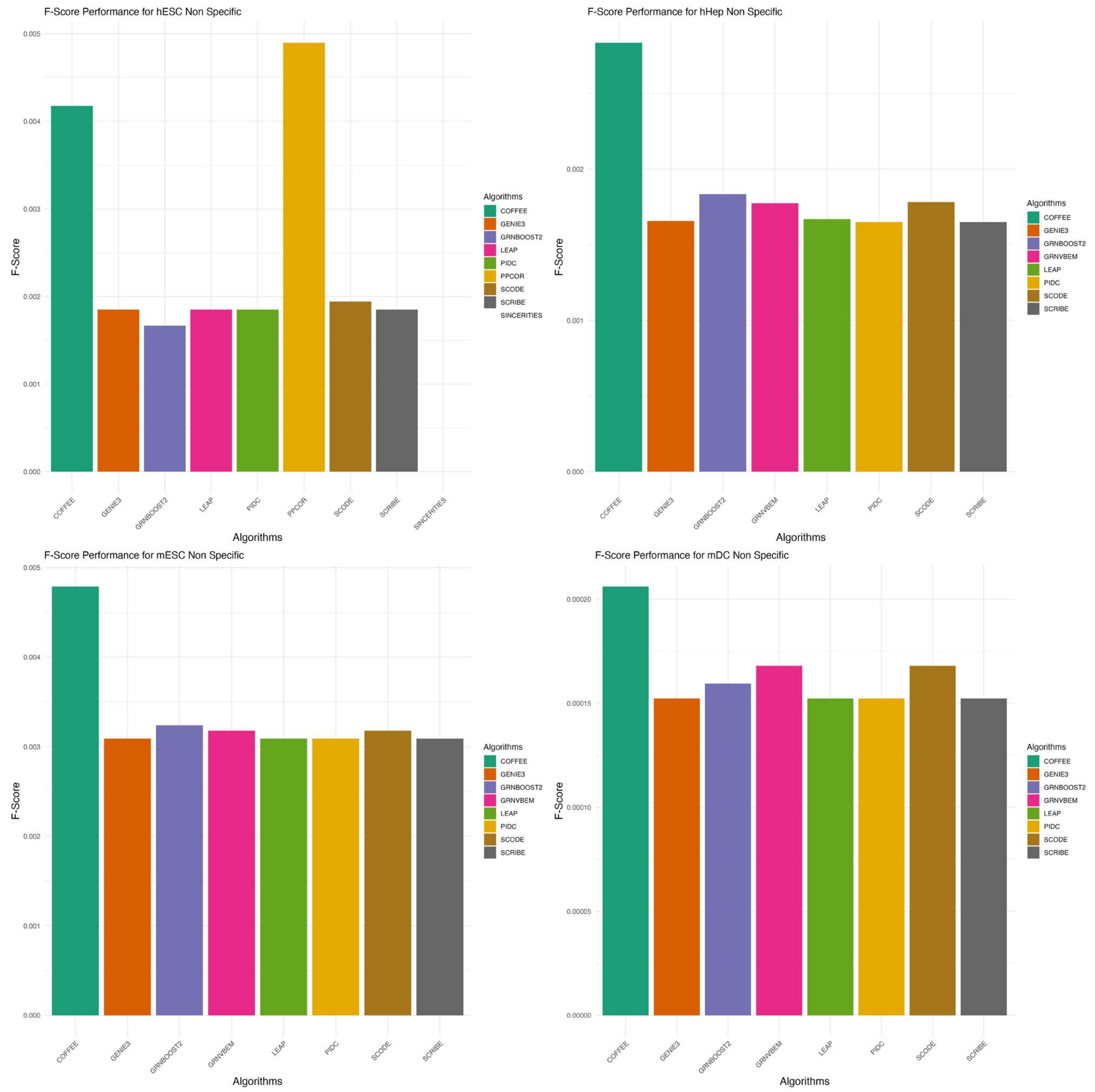


Figure 7. F-score performance on Experimental datasets using non-Cell-Type–Specific Ground Truth; COFFEE demonstrated improved performance in terms of F-score in all of the Experimental datasets for the non-Cell-Type–specific ground truth.

revealed the presence of tissue resident memory cells (named as CD8+ TRM cells) [38]. Tissue resident memory cells are crucial to immune response and defense, particularly from an oncology perspective [39]. Further analysis from Savas *et al.* revelaed improved patient prognosis by increased presence of the CD8+ TRM cells [38].

The transcription factor STAT1 is linked to immune response pathways [38]. STAT1 is also demonstrated to be imperative for CD8+ T cell proliferation. To better understand the role of STAT1 within CD8+ TRM cells specifically, we inferred a GRN specific to the CD8+ TRM cells using COFFEE. We filtered from the 6311 cells down to 685 using Seurat [36]. From there, we followed a similar process as the Experimental datasets, by selecting the top 500 genes varying by pseudotime using Slingshot [9]. To ensure analysis of STAT1, we manually included STAT1 into the resulting gene list. We ran COFFEE using PPCOR, GENIE3, GRNBOOST2, and PIDC, since they demonstrated the best performance on experimental datasets, using a threshold of 0.65. Once the network was generated, the network was filtered by selecting interactions only where STAT1 was the regulator. This resulted in 137 total significant interactions of STAT1 within the CD8+ TRM cells.

To determine which interactions were relevant to cancer, we used the Network of Cancer Genes and Healthy Drivers [40]. Of the 137 total interactions with STAT1, 30 of them were demonstrated to be cancer associated by the Network of Caner Genes and Healthy Drivers. We then performed a GO Biological Process (BP) Overrepresentation Analysis using the ClusterProfiler R package to determine the functional role of these cancer associated genes [41].

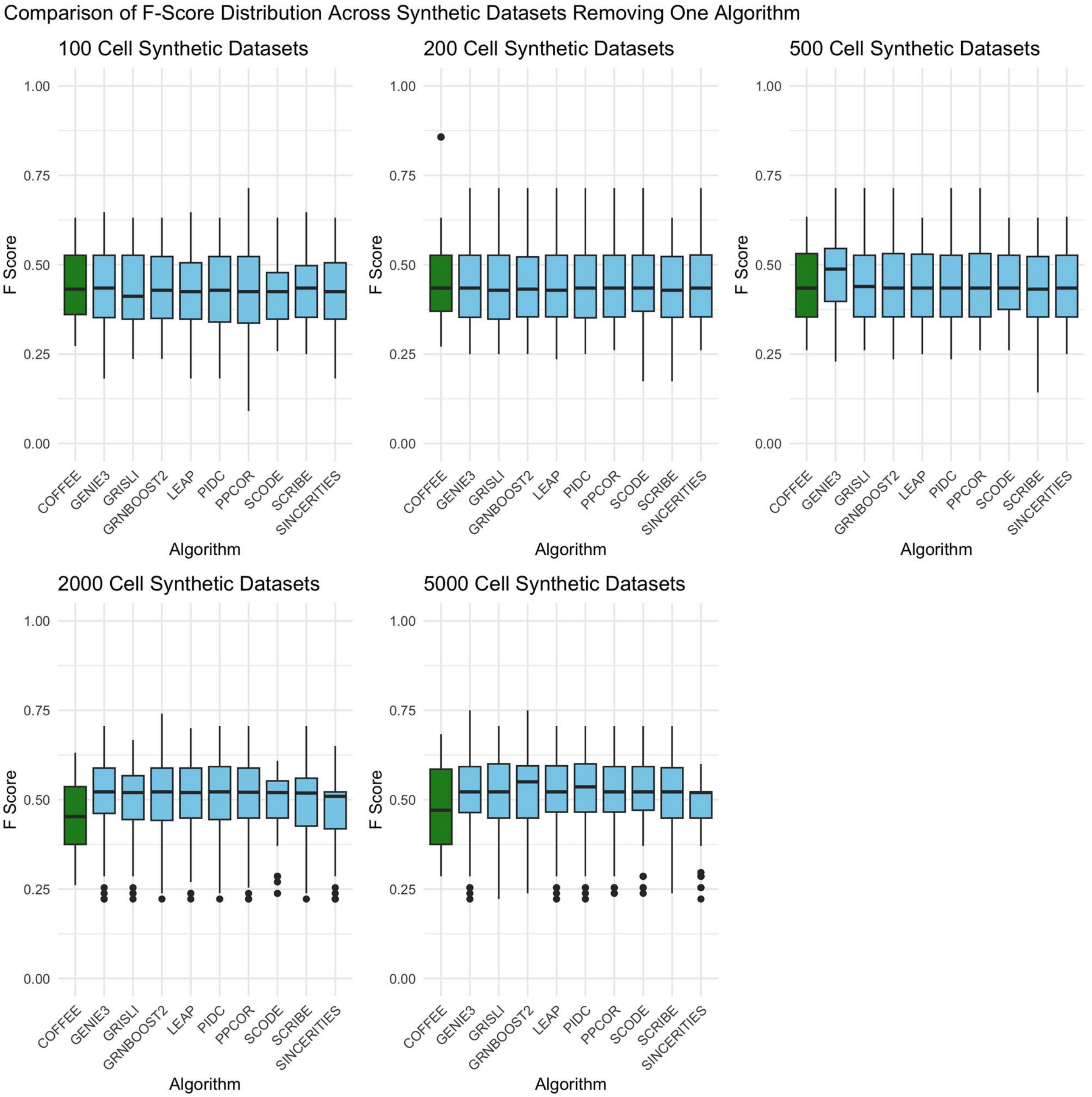


Figure 8. Comparison of F-score distribution across synthetic datasets: removing one algorithm; the label corresponds to an algorithm being removed from the consensus; for example, the box labeled ”GENIE3” represents a consensus algorithm performance for the other nine algorithms in the pipeline.

The GO BP analysis importantly revealed enrichment for MHC protein complex assembly-related and adaptive immune response pathways. The MHC protein complex assembly is a significant finding, as this pathway has previously been linked to CD8+ T cells. The MHC protein is what allows CD8+ T cells to identify pathological cells that express mutated proteins. A common mechanism for cancers is to target the MHC protein and therefore weaken the effectiveness of the CD8+ T cells [42]. The pathway enrichment reveals that STAT1 potentially has a key role in regulating the MHC protein complex, specifically in the CD8+ TRM subcluster. Further analysis could be performed to determine individual gene importance to these pathways, and knockout experiments could be performed to confirm STAT1’s role in this regulatory process.

Discussion

From our analysis, we are able to conclude that a consensus approach for single cell GRN inference has several advantages. Across various types of datasets, COFFEE demonstrated improved performance in determining true interactions when compared with gold standard datasets.

For Synthetic datasets, we determined the ideal threshold to select edges from the consensus network generated by COFFEE. We found that for larger datasets, including more edges improved the performance of the algorithm, while for smaller datasets, a higher threshold maximized the F-score when compared with the gold standard. However, users may want to adjust the threshold value based on the statistic they want to maximize; increasing the consensus threshold will lead to a higher precision, while

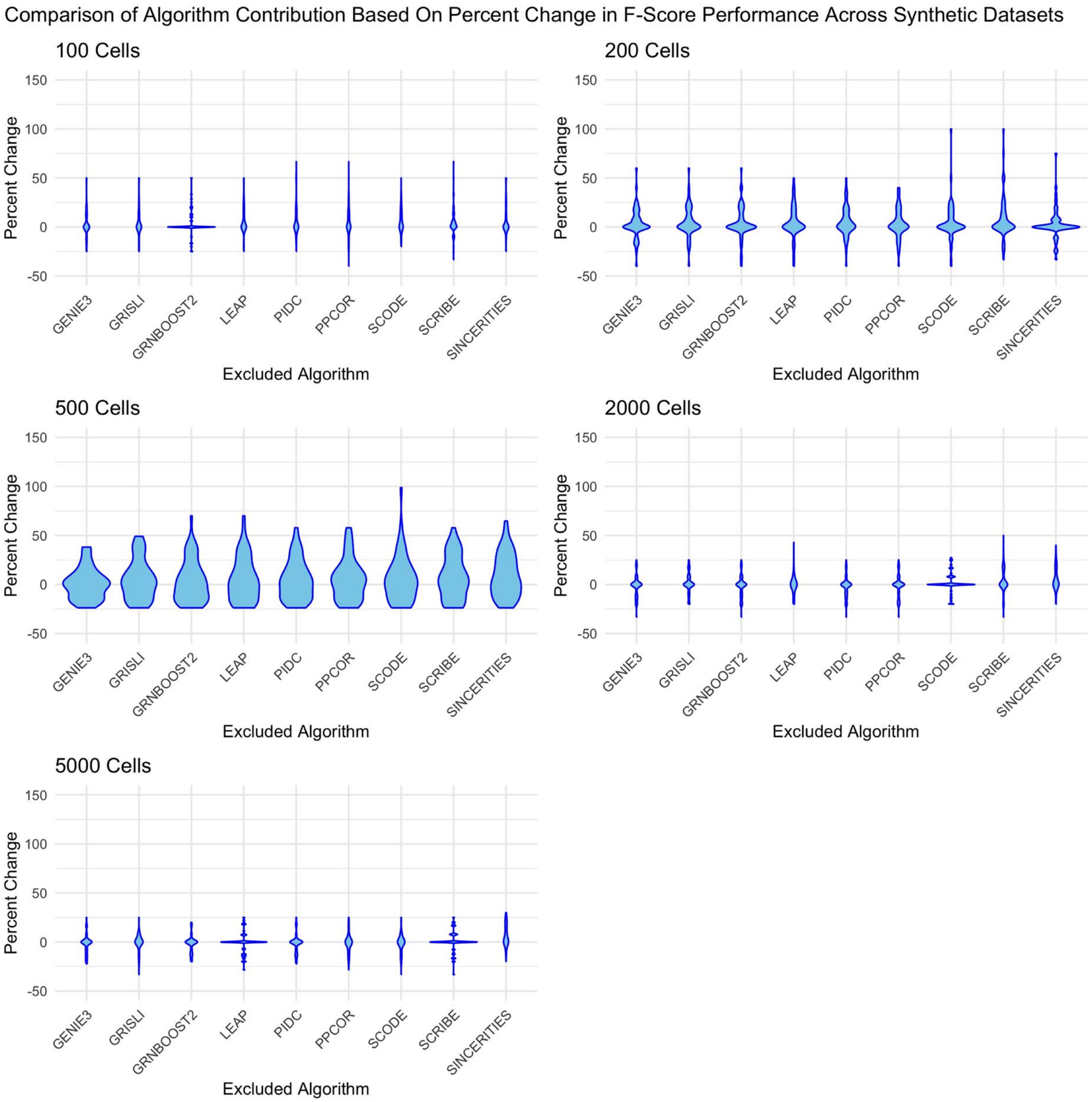


Figure 9. Percent change of F-score distribution across synthetic datasets: removing one algorithm; the label corresponds to an algorithm being removed from the consensus; for example, the box labeled "GENIE3" represents a consensus algorithm performance for the other nine algorithms in the pipeline.

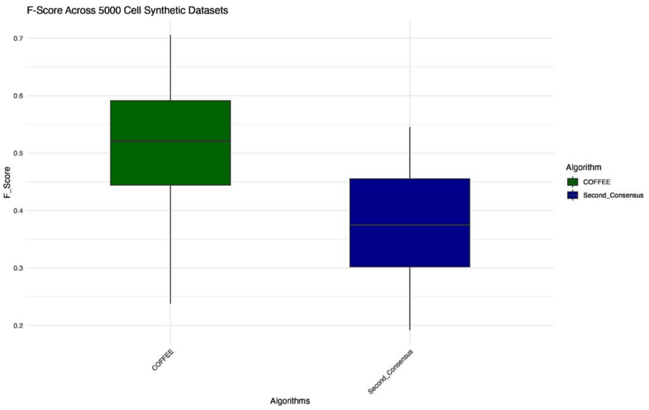


Figure 10. F-score performance for Second Level Consensus vs COFFEE on Synthetic datasets.

decreasing the threshold will lead to a higher recall. The suggested thresholds in Table 6 are based on the F-score. Users ideally should run COFFEE with various threshold values and evaluate the performance on their particular dataset.

We also found that there was significant variation between algorithm performance on dataset size, which is an effect also noticed in previous scRNAseq-based GRN inference benchmarking studies [43]. For example, SINCERITIES did not perform as well as the other baseline algorithms on smaller datasets but performed extremely well on larger datasets in terms of F-score. The primary advantage of COFFEE is that the algorithm has consistent high performance across multiple types of datasets; this is inherent in the algorithm design, as edges appearing in multiple GRN inferred networks will have higher confidence scores in the final consensus network. This approach establishes that the consensus approach will work across several different kinds of datasets, particularly with a large selection of algorithms to

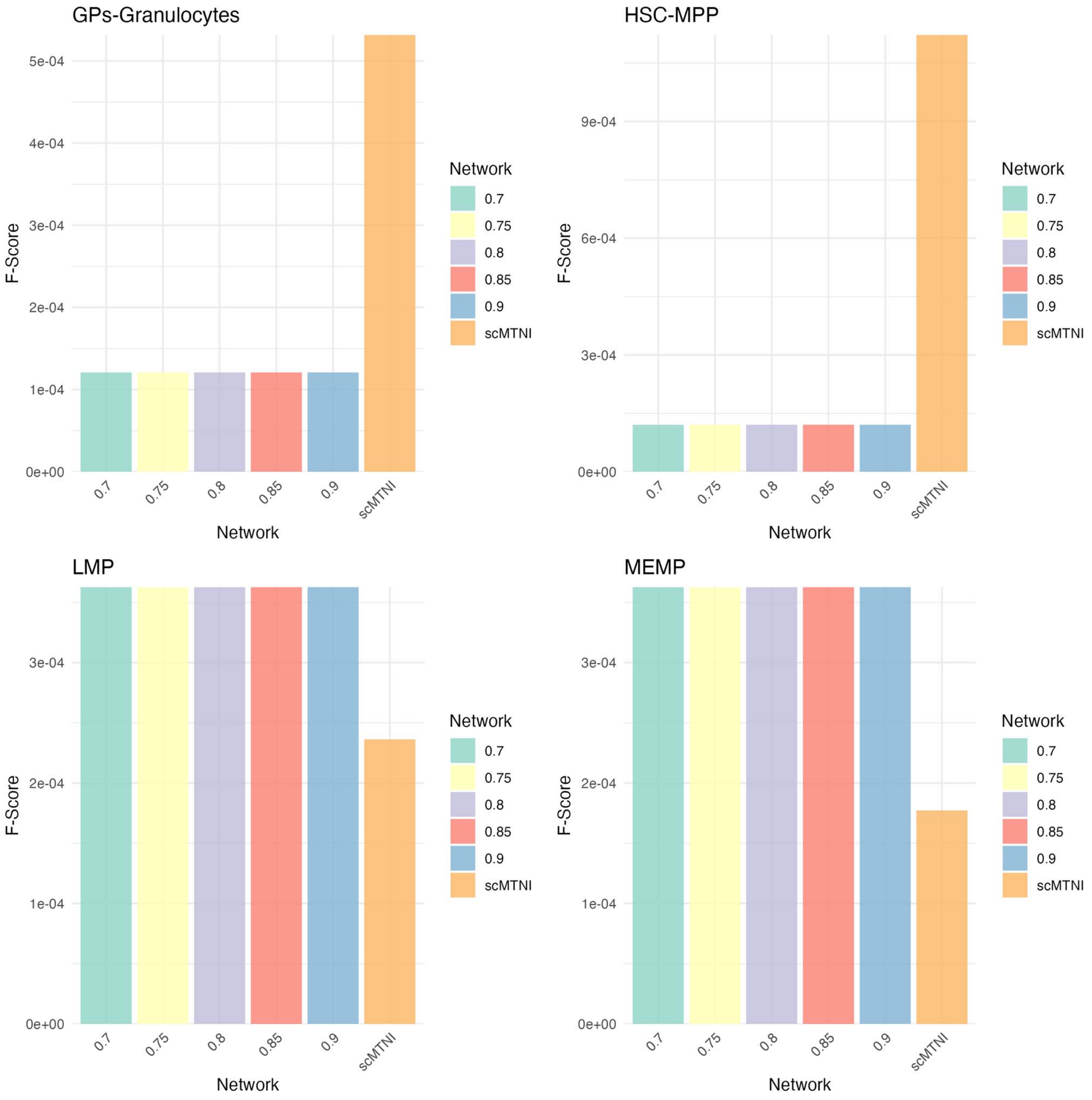


Figure 11. F-score for each cell type across COFFEE thresholds and scMTNI.

integrate. COFFEE is a generalizable tool, where any number or type of algorithm results may be used for the downstream consensus integration. Therefore, as methods improve for scRNAseq GRN inference, COFFEE will remain valuable as a tool for the community.

This effect is even more pronounced when analyzing Curated and Experimental datasets. Despite the number of cells being the same across Curated datasets, different algorithms performed better on some datasets than others. This establishes that users cannot go by dataset size alone when choosing an optimal algorithm. Additionally, algorithms that require pseudotime data as input are more likely to suffer in their performance if the pseudotime data are inaccurate, or incorrectly computed [2]. While Pratapa *et al.* in their benchmarking study make suggestions for which algorithms to use for various datasets, a far safer approach would be to use a consensus-based method. We see that across Curated datasets, COFFEE has high F-score values. In contrast, we see that each Curated dataset had a different top performer. Without considering COFFEE, the top performer for VSC was GENIE3, PPCOR for HSC, and SINCERITIES for GSD. The variation in algorithm performance across datasets makes a consensus method such as COFFEE a safe choice to consistently infer high-quality networks.

On the Experimental datasets, we chose to evaluate on Cell-Type–specific and Non-Specific ground truths, as was done in the BEELINE evaluation framework [2]. From this analysis, we noticed that while the baseline algorithms performed poorly on the Non-Specific networks, COFFEE performed significantly better. However for the Cell-Type–Specific ground truth data, COFFEE performed better generally, but there was less difference between the baseline and consensus methods in terms of F-score. This suggests that COFFEE without modifications is not as well

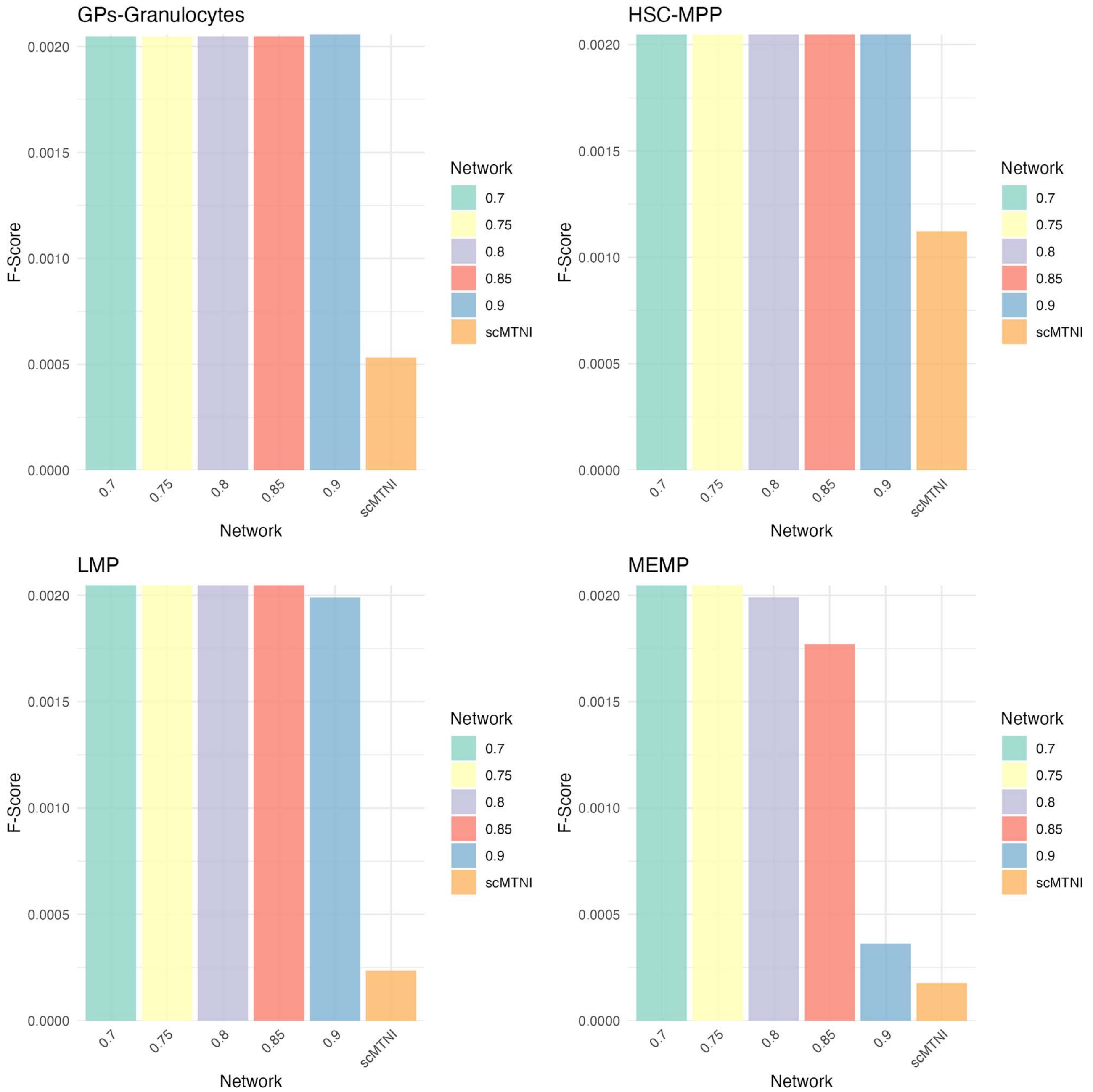
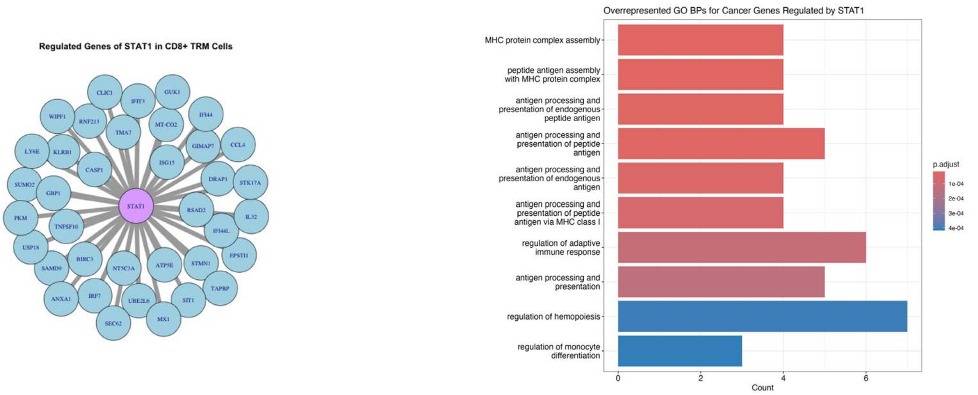


Figure 12. F-score for each cell-type across COFFEE thresholds and COFFEE+scMTNI.

equipped to infer more specific interactions, rather than general interactions collected by organism level Chip-Seq networks. This trend is also seen when evaluating COFFEE based on EPR. EPR is defined as the precision of the top *k* edges when compared with the ground truth, where *k* is the number of interactions present in the filtered ground truth. The reasoning behind this was that experimental groups would primarily be interested in only high confidence edges from a network [2]. COFFEE demonstrates a higher EPR, and a much better EPR for the Non-Specific networks when compared with the Cell-Type–Specific networks. We show this result in Supplementary Figs S2 and S3. This finding motivated us to adapt the COFFEE algorithm for cell-type–specific inference, by developing a prioritized consensus method. We see that across dataset types and conditions, COFFEE generally demonstrates higher precision but a lower recall. This is due to the fact that it is giving more precedence to high confidence edges predicted by other algorithms. Therefore, COFFEE is more likely to predict true edges but may potentially not be able to predict all edges present in a GRN. Therefore, we can conclude that when COFFEE predicts an edge, it is likely to be a true edge but may not predict all correct edges in a GRN.

We see that a modified COFFEE algorithm is able to improve the performance of a well-established cell-type specific GRN inference algorithm. Despite the lack of the additional information that scMTNI takes as input, COFFEE with four algorithms had improved performance on two out of four cell-types on the benchmarking dataset. We were also able to substantially improve the performance of scMTNI by incorporating a prioritized consensus approach, where interactions predicted by scMTNI were all initialized with a raw count of 1. We anticipate that


Regulated Genes of STAT1 in CD8+ TRM Cells

Overrepresented GO BPs for Cancer Genes Regulated by STAT1

(a) Highest confidence (EdgeWeight > 0.9) interactions of STAT1 within the CD8+ TRM cells. The network visualization was generated using the igraph R package.

(b) Overrepresented GO BPs of cancer associated genes regulated by STAT1 within CD8+ TRM cells

Figure 13. COFFEE driven STAT1 regulatory analysis within CD8+ TRM cells in human breast cancer.

the modified consensus approach has the potential to be very effective for cell-type–specific GRN inference as more methods are made available to the community.

We also note that COFFEE can be integrated with other ”omics” and database sources. For example, Transcriptome Wide Association Studies (TWAS) are a valuable tool for determining key regulators for particular phenotypes. If regulators are confirmed from TWAS studies, then they can be prioritized within the COFFEE algorithm, in a similar way as shown in our Cell-Type–specific inference method. However, rather than giving higher priority to weights predicted from a certain algorithm, we can give higher weightage to interactions where the regulator of that interaction is predicted from TWAS studies. Tools such as kTWAS, mkTWAS, and webTWAS tool in particular seems very useful to this end [44–46]. Additionally, the integration of information from databases such as String or Wikipathways can be beneficial to the COFFEE workflow [47, 48]. The most straightforward way to integrate the results from String or Wikipathway databases for consensus would be to integrate the results from a inference method that predicts GRNs from these databases and then have this network within the COFFEE pipeline for consensus integration [47, 48]. GRN inference methods that utilize data from these databases show promise for this approach, such as the method from Abbaszadeh *et al.* [49].

Another way to use the String or Wikipathway information would be to identify regulators for a disease or biological context (such as searching for regulators of breast cancer) and give higher weightage to edges predicted from COFFEE that contain those regulators [47, 48]. This would require some manual work from the user and also some prior knowledge regarding the biological context of the network. However, this has the potential to be a way to integrate additional knowledge sources within the COFFEE framework.

We conclude that while a regular consensus approach for bulk RNA-seq data has performed well in prior studies, a modified consensus approach is warranted for scRNA-seq data. The best practice method for bulk RNA-seq GRN inference is to incorporate as many algorithms as possible, as this demonstrates improved performance [1]. However, we see that a consensus approach for scRNA-seq data requires more careful selection of the algorithms incorporated for integration. We demonstrate that across Synthetic datasets of differing sizes, distinct algorithm combinations leads to variable performance. From this result, we establish that simply including every GRN algorithm available will not necessarily lead to the best performance in the consensus approach. Therefore, the user needs to exercise discretion in choosing the best algorithm combination for their individual needs, and more benchmarking in this area needs to be done in order to make specific recommendations.

The primary limitation of a consensus-based approach is that it is only as strong as its underlying algorithms. As more improved GRN inference algorithms emerge, it will be difficult to determine what will be the best algorithm for any given dataset. With this point in mind, we continue to see a consensus-based approach to be beneficial for the community when there is uncertainty in choosing the best algorithms to use.

Conclusion

In this paper, we present COFFEE, a Borda voting-based consensus algorithm for GRN inference on scRNA-seq data. COFFEE has demonstrated improved and consistent performance across Synthetic, Curated, and Experimental datasets. Additionally, a modified consensus-based approach for cell-type–specific GRN inference has shown a promising ability to improve performance on existing state-of-the-art methods, by augmenting important gene interactions. As future work improves the landscape of cell-type specific GRN inference, this will necessitate a weighted consensus algorithm approach to merge the predictions of the sets of cell-type–specific and non-cell-type–specific algorithms for robust GRN inference.

Key Points

- Several GRN inference methods for scRNAseq data have been developed, each with distinct underlying methodologies and capabilities; extensive benchmarking has determined that there is no singular method that is the best in all cases.
- We present COFFEE, a Borda count voting-based consensus algorithm that uses a wisdom of crowds approach to integrate the results from several different GRN inference algorithms into one composite network.
- COFFEE demonstrates improved performance on several benchmarking datasets from the BEELINE evaluation

framework, across Synthetic, Curated, and Experimental datasets.
- We also highlight the effectiveness of a prioritized consensus algorithm; methods that are shown to perform better or incorporate other data modalities can be prioritized in the COFFEE setup so that predictions from them are given more importance. To demonstrate this, we chose scMTNI for cell-type–specific GRN inference and demonstrate that a prioritized COFFEE with scMTNI performs better than just using predictions from scMTNI.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was partially supported by 5R21MH128562-02 (PI: Roberson-Nay), 5R21AA029492-02 (PI: Roberson-Nay), CHRB-2360623 (PI: Das), NSF-2316003 (PI: Cano), VCU Quest (PI: Das), and VCU Breakthroughs (PI: Ghosh) funds awarded to P.G.

Code availability

A Python implementation of COFFEE may be found on GitHub: https://github.com/lodimk2/coffee.

Data availability

To perform the evaluation for COFFEE against the 10 baseline algorithms, we used data from the BEELINE evaluation framework [2]. The BEELINE data may be downloaded here: https://zenodo.org/records/3378976.

To perform COFFEE experiments for the cell-type specific datasets, we used the data from scMTNI's evaluation [5]. The scMTNI data may be downloaded here: https://zenodo.org/records/7879228.

References

1. Nalluri JJ, Barh D, Azevedo V. *et al.* Mirsig: a consensus-based network inference methodology to identify pan-cancer mirna-mirna interaction signatures. *Sci Rep* 2017;**7**. https://doi.org/10.1038/srep39684.
2. Pratapa A, Jalihal AP, Law JN. *et al.* Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* 2020;**17**:147–54. https://doi.org/10.1038/s41592-019-0690-6.
3. Chaitankar V, Ghosh P, Perkins E. *et al.* A novel gene network inference algorithm using predictive minimum description length approach. *BMC Syst Biol* 2010;**4**. https://doi.org/10.1186/1752-0509-4-S1-S7.
4. Chaitankar V, Ghosh P, Perkins E. *et al.* Time lagged information-theoretic approaches to the reverse engineering of gene regulatory networks. *BMC Bioinformatics* 2010;**11**. https://doi.org/10.1186/1471-2105-11-S6-S19.
5. Zhang S, Pyne S, Pietrzak S. *et al.* Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nature. Communications* 2023;**14**:3064. https://doi.org/10.1038/s41467-023-38637-9.
6. Lähnemann D, Köster J, Szczurek E. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* 2020;**21**. https://doi.org/10.1186/s13059-020-1926-6.
7. Specht AT, Li J. Leap: Constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering. *Bioinformatics* 2016;**33**:764–6.
8. Kim S. Ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods* 2015;**22**:665–74. https://doi.org/10.5351/CSAM.2015.22.6.665.
9. Street K, Risso D, Fletcher RB. *et al.* Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 2018;**19**:477. https://doi.org/10.1186/s12864-018-4772-0.
10. Ly L-H, Vingron M. Effect of imputation on gene network reconstruction from single-cell rna-seq data. *Patterns* 2022;**3**:100414. https://doi.org/10.1016/j.patter.2021.100414.
11. Hamada D, Nakayama M, Saiki J. Wisdom of crowds and collective decision-making in a survival situation with complex information integration. *Cogn Res Princ Implic* 2020;**5**. https://doi.org/10.1186/s41235-020-00248-z.
12. Marbach D, Costello JC, Küffner R. *et al.* Wisdom of crowds for robust gene network inference. *Nat Methods* 2012;**9**:796–804. https://doi.org/10.1038/nmeth.2016.
13. Davies J, Katsirelos G, Narodytska N. *et al.* Complexity of and algorithms for the manipulation of borda, nanson's and baldwin's voting rules. *Artif Intell* 2014;**217**:20–42. https://doi.org/10.1016/j.artint.2014.07.005.
14. Nalluri J, Rana P, Barh D. *et al.* Determining causal mirnas and their signaling cascade in diseases using an influence diffusion model. *Sci Rep* 2017;**7**:8133. https://doi.org/10.1038/s41598-017-08125-4.
15. Zeng Y, He Y, Zheng R. *et al.* Inferring single-cell gene regulatory network by non-redundant mutual information. *Brief Bioinform* 2023;**24**. https://doi.org/10.1093/bib/bbad326.
16. Wang J, Chen Y, Zou Q. Inferring gene regulatory network from single-cell transcriptomes with graph autoencoder model. *PLoS Genet* 2023;**19**:e1010942. https://doi.org/10.1371/journal.pgen.1010942.
17. Huynh-Thu V, Irrthum A, Wehenkel L. *et al.* Inferring regulatory networks from expression data using tree-based methods. *PloS One* 2010;**5**:e12776. https://doi.org/10.1371/journal.pone.0012776.
18. Chan TE, Stumpf MPH, Babtie AC. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems* 2017;**5**:251–267.e3. https://doi.org/10.1016/j.cels.2017.08.014.
19. Gao NP, Minhaz Ud-Dean SM, Gandrillon O. *et al.* Sincerities: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* 2018;**34**:258–66. https://doi.org/10.1093/bioinformatics/btx575.
20. Sanchez-Castillo M, Blanco D, Tienda-Luna IM. *et al.* A bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* 2017;**34**:964–70. https://doi.org/10.1093/bioinformatics/btx605.
21. Qiu X, Rahimzamani A, Wang L. *et al.* Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell Systems* 2020;**10**:265–274.e11. https://doi.org/10.1016/j.cels.2020.02.003.
22. Moerman T, Santos SA, Carmen BG. *et al.* Grnboost2 and arboreto: Efficient and scalable inference of gene regulatory networks. *Bioinformatics* 2018;**35**:2159–61.

23. Aubin-Frankowski P-C, Vert J-P. Gene regulation inference from single-cell rna-seq data with linear differential equations and velocity inference. *Bioinformatics* 2020;**36**:4774–80. https://doi.org/10.1093/bioinformatics/btaa576.
24. Deshpande A, Chu L-F, Stewart R. *et al.* Network inference with granger causality ensembles on single-cell transcriptomics. *Cell Rep* 2022;**38**:110333. https://doi.org/10.1016/j.celrep.2022.110333.
25. Ranzoni AM, Tangherloni A, Berest I. *et al.* Integrative single-cell rna-seq and atac-seq analysis of human developmental hematopoiesis. *Cell Stem Cell* 2021;**28**:472–487.e7. https://doi.org/10.1016/j.stem.2020.11.015.
26. Schaffter T, Marbach D, Floreano D. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* 2011;**27**:2263–70. https://doi.org/10.1093/bioinformatics/btr373.
27. Anna, Lovrics Y, Gao BJ. *et al.* Boolean modelling reveals new regulatory connections between transcription factors orchestrating the development of the ventral spinal cord. *PloS One* 2014;**9**. https://doi.org/10.1371/journal.pone.0111430.
28. Krumsiek J, Marr C, Schroeder T. *et al.* Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PloS One* 2011;**6**:e22649. https://doi.org/10.1371/journal.pone.0022649.
29. Rios O, Frias S, Rodriguez A. *et al.* A boolean network model of human gonadal sex determination. *Theoretical Biology and Medical Modelling* 2015;**12**:26. https://doi.org/10.1186/s12976-015-0023-0.
30. Giacomantonio CE, Goodhill GJ. A boolean model of the gene regulatory network underlying mammalian cortical area development. *PLoS Comput Biol* 2010;**6**:e1000936. https://doi.org/10.1371/journal.pcbi.1000936.
31. Gray Camp J, Sekine K, Gerber T. *et al.* Multilineage communication regulates human liver bud development from pluripotency. *Nature* 2017;**546**:533–8. https://doi.org/10.1038/nature22796.
32. Chu L-F, Leng N, Zhang J. *et al.* Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol* 2016;**17**:173. https://doi.org/10.1186/s13059-016-1033-x.
33. Nestorowa S, Hamey FK, Sala BP. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* 2016;**128**:e20–31. https://doi.org/10.1182/blood-2016-05-716480.
34. Hayashi T, Ozaki H, Sasagawa Y. *et al.* Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. *Nat Commun* 2018;**9**:619. https://doi.org/10.1038/s41467-018-02866-0.
35. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**:296. https://doi.org/10.1186/s13059-019-1874-1.
36. Hao Y, Stuart T, Kowalski MH. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2023;**42**:293–304. https://doi.org/10.1038/s41587-023-01767-y.
37. Cusanovich DA, Pavlovic B, Pritchard JK. *et al.* The functional consequences of variation in transcription factor binding. *PLoS Genet* 2014;**10**:e1004226. https://doi.org/10.1371/journal.pgen.1004226.
38. Savas P, Virassamy B, Ye C. *et al.* Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nature News* 2018;**24**:986–93. https://doi.org/10.1038/s41591-018-0078-7.
39. Kok L, Masopust D, Schumacher TN. The precursors of cd8+ tissue resident memory t cells: From lymphoid organs to infected tissues. *Nat Rev Immunol* 2021;**22**:283–93. https://doi.org/10.1038/s41577-021-00590-3.
40. Dressler L, Bortolomeazzi M, Keddar MR. *et al.* Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: An update of the network of cancer genes (ncg) resource. *Genome Biol* 2022;**23**:35. https://doi.org/10.1186/s13059-022-02607-z.
41. Tianzhi W, Erqiang H, Shuangbin X. *et al.* Clusterprofiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation* 2021;**2**:100141. https://doi.org/10.1016/j.xinn.2021.100141.
42. Dhatchinamoorthy K, Colbert JD, Rock KL. Cancer immune evasion through loss of mhc class i antigen presentation. *Front Immunol* 2021;**12**. https://doi.org/10.3389/fimmu.2021.636568.
43. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* 2018;**19**:232. https://doi.org/10.1186/s12859-018-2217-z.
44. Cao C, Kwok D, Edie S. *et al.* Ktwas: Integrating kernel machine with transcriptome-wide association studies improves statistical power and reveals novel genes. *Brief Bioinform* 2020;**22**. https://doi.org/10.1093/bib/bbaa270.
45. Cao C, Kossinna P, Kwok D. *et al.* Disentangling genetic feature selection and aggregation in transcriptome-wide association studies. *Genetics* 2021;**220**. https://doi.org/10.1093/genetics/iyab216.
46. He J, Wen W, Beeghly A. *et al.* Integrating transcription factor occupancy with transcriptome-wide association analysis identifies susceptibility genes in human cancers. *Nature News* 2022, 7118. https://doi.org/10.1038/s41467-022-34888-0.
47. Szklarczyk D, Kirsch R, Koutrouli M. *et al.* The string database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 2022;**51**:D638–46. https://doi.org/10.1093/nar/gkac1000.
48. Agrawal A, Balci H, Hanspers K. *et al.* Wikipathways 2024: next generation pathway database. *Nucleic Acids Res* 2023;**52**:D679–89. https://doi.org/10.1093/nar/gkad960.
49. Abbaszadeh O, Azarpeyvand A, Khanteymoori A. *et al.* Data-driven and knowledge-based algorithms for gene network reconstruction on high-dimensional data. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:1545–57. https://doi.org/10.1109/TCBB.2020.3034861.