# Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data

Matthew Stone[1,2], Jiaxin Li[1,3], Sunnie Grace McCalla[1,3], Alireza Fotuhi Siahpirani[1], Viswesh Periyasamy[1,4], Junha Shin[1], and Sushmita Roy[1,2,4*]

[1]Wisconsin Institute for Discovery, University of Wisconsin-Madison

[2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison

[3]Laboratory of Genetics, University of Wisconsin-Madison

[4]Department of Computer Sciences, University of Wisconsin-Madison

[*]To whom correspondence should be addressed.

**Abstract**

Single-cell RNA-sequencing (scRNA-seq) offers unparalleled insight into the transcriptional programs of different cellular states by measuring the transcriptome of thousands individual cells. An emerging problem in the analysis of scRNA-seq is the inference of transcriptional gene regulatory networks and a number of methods with different learning frameworks have been developed. Here we present an expanded benchmarking study of eleven recent network inference methods on six published single-cell RNA-sequencing datasets in human, mouse, and yeast considering different types of gold standard networks and evaluation metrics. We evaluate methods based on their computing requirements as well as on their ability to recover the network structure. We find that while most methods have a modest recovery of experimentally derived interactions based on global metrics such as AUPR, methods are able to capture targets of regulators that are relevant to the system under study. Based on overall performance we grouped the methods into three main categories and found a combination of information-theoretic and regression-based methods to have a generally high performance. We also evaluate the utility of imputation for gene regulatory network inference and find that a small number of methods benefit from imputation, which further depends upon the dataset. Finally, comparisons to inferred networks for comparable bulk conditions showed that networks inferred from scRNA-seq datasets are often better or at par to those from bulk suggesting that scRNA-seq datasets can be a cost-effective way for gene regulatory network inference. Our analysis should be beneficial in selecting algorithms for performing network inference but also argues for improved methods and better gold standards for accurate assessment of regulatory network inference methods for mammalian systems.

## Introduction

Inference of genome-scale regulatory networks from global mRNA profiles has been a long-standing problems in systems biology and gene regulation. These networks are important for understanding the mechanisms of cell type specification and disease. The availability of single-cell omic data opens up new opportunities to reverse engineer transcriptional regulatory networks, enabling us to study the regulatory network from multiple cell types. Single-cell RNA-seq (scRNA-seq) measures genome-wide expression profiles of tens of thousands of cells, which greatly reduces the cost of generating datasets with the large sample sizes needed for computational network inference. Furthermore, it offers the potential to infer cell-type-specific regulatory networks for both known and novel cell populations in the same experiment, offering an efficient way to dissect heterogeneous populations.

The availability of scRNA-seq datasets has fueled the development of a number of methods for network inference from these data that use different types of models ranging from Gaussian graphical models [1], information theoretic methods [2–4], random forests [5], ordinary differential equations [6], and Boolean networks [7] (See [8, 9] for comprehensive reviews). Methods also vary in their inclusion of pseudotime [3, 6, 10] or imputed, de-noised signals [4]. Some of these methods are specifically modeling the statistical properties of scRNAseq data [11, 12], while others are adaptations of existing methods for bulk data [5]. In parallel with method development, there have been two benchmarking studies [13, 14] to compare existing network inference algorithms. One study showed that existing methods perform poorly on both simulated and real data [14], while the second showed that methods that perform well on simulated data can be relatively poor on real data [13]. While these studies have been useful, a number of open questions need to be addressed to gain a comprehensive understanding of the strengths and weaknesses of existing network inference algorithms from scRNA-seq datasets. In particular, existing benchmarking efforts have studied relatively small number of networks from real data ($<$2000 genes), while in reality genome-scale regulatory networks can have between 5-10k genes. Another unknown is the extent to which different benchmarks (e.g. ChIP-chip/seq versus regulator perturbation) can affect performance and to what extent do methods agree with each other on novel predictions and parts of the gold standard networks. In addition, imputation of sparse scRNA-seq count matrices has been suggested to improve downstream analysis including identification functional relationships among genes [4]. However, it is not clear whether this benefits network

inference methods more generally.

To address these gaps, we compared 11 network inference methods on six published scRNA-seq datasets from human, mouse, and yeast (**Figure 1**). Benchmarking the algorithms for time and memory consumption on datasets of different sizes identified several algorithms that are unlikely to scale to genome-wide gene regulatory networks. We compare the algorithm performance using different global network topology metrics such as area under the precision recall curve and F-score as well as local metrics such as the number of regulators whose targets can be predicted accurately. Based on global metrics, network inference methods had modest performance compared to random, however the local metrics show that methods can predict targets of several relevant regulators. Generally, the top performing algorithms tend to agree with each other on predicted edges, although this depends upon the depth of the dataset. Imputation of scRNA-seq datasets did not offer substantial benefits for most network inference approaches. To our surprise, simple correlation based methods performed as well or better on these metrics compared to most network inference algorithms. Finally, we compared the performance of these algorithms on single cell datasets to those on bulk datasets for matched conditions and find comparable performance for both classes of algorithms. Overall, our study presents an extended comparison of recent algorithms of network inference from scRNA-seq datasets and identifies common strengths and weaknesses on different types of gold standards, which should be beneficial to the community for improved network inference.

## Results

### Computing requirements of regulatory network inference algorithms for single cell RNA-seq data

We started with 11 algorithms specifically designed to infer regulatory networks from single-cell RNA-seq data, along with a bulk inference method (**Figure 1**, **Additional File 1**, **Table S1**). In addition to these published algorithms, we used Pearson's correlation to infer a network of gene pairs with edges weighted by the experiment-wide correlation of expression. We first benchmarked these algorithms for their computing requirements, namely memory consumption and runtime, using different numbers of genes from 10 to 8000 genes (**Methods**, **Figure 2**). Of the compared methods, SCHiRM and BTR did not complete in a reasonable

amount of time and were excluded from further analysis. SCRIBE and HurdleNormal could be run upto 2000 genes, however beyond that took much longer to run. Most algorithms that were runnable for upto 8k genes and consumed upto 15G of RAM. Exceptions to this were SCHiRM, Inferelator and PIDC which took higher memory. SCHiRM memory consumption grew exponentially and was considered infeasible for larger gene sets. For the subsequent analysis we excluded SCHiRM, BTR and HurdleNormal from our analysis.

### Assessing scRNA-seq regulatory network inference algorithms based on global metrics

We next compared the performance of algorithms on real gold standard networks from three species, yeast, mouse and human (**Figure 1**). These datasets included two yeast stress response datasets, four mouse datasets, one for dendritic cells, and three for cellular reprogramming to pluripotent stem cells on different growth media, and one human dataset for embryonic stem cell differentiation (**Additional File 1**, **Table S2**). The datasets had a wide range of numbers of cells, ranging from 163 cells from Gasch et al [15] to the largest dataset with 36,199 cells from one of the mouse reprogramming studies [16]. For each species dataset we had three types of gold standards, those derived from knock down or knock out of the regulator followed by global mRNA profiling (Perturb, **Figure 3**), those derived from ChIP-chip or ChIP-seq experiments (ChIP, **Figure 3**), and those derived from the intersection of these two gold standards (Perturb+ChIP, **Figure 3**). We assessed the performance of the algorithms using standard area under the precision-recall curve (AUPR) as well as F-score (**Figure 3**, **Additional File 2**, **Figure S1**) considering each algorithm on each of the datasets (**Method**). While AUPR considers all edges estimated by a network inference algorithm, F-score is computed using the top edges. To select the number of edges for F-score computation, we considered different number of edges and computed the stability of the metrics as a function of the number of edges (**Additional File 2**, **Figure S2**). We selected 5k edges for F-score computation as this resulted in the most stable results across algorithms.

We first examined the performance of each algorithm for individual datasets from the different species and ranked each method based on its F-score and AUPR (**Figure 3A**). The ranking of methods was consistent based on both F-score and AUPR. There was more variation in ranking due to different gold standards (e.g. ChIP versus Perturb) than across datasets. Across datasets, no single method had consistently the highest

scores, however the ranking of methods across datasets remained largely consistent, that is, methods that ranked high in one dataset ranked high in another dataset. A few exceptions were SILGGM and knnDREMI which had a relatively lower performance on most datasets with the exception of Sridharan A2S/FBS Perturb where it ranked highly. LEAP performed poorly on the Shalek dataset, but was generally among the middle ranking methods. Among the different methods, SCENIC, MERLIN and Pearson correlation were most stably high performing across datasets based on F-score and AUPR.

We next examined the performance of algorithms with respect to the different types of gold standards. Both F-score and AUPR were generally better for the Perturb gold standard datasets compared to ChIP gold standards (**Figure 3B**). Exception was SCODE which was better for ChIP compared to Perturb. We aggregated the metrics across all datasets for each gold standard based on their median rank (**Figure 3C**). Based on median F-score on the Perturb gold standard, the top 3 methods were Pearson with a median rank of 3 and PIDC and SCENIC with a median rank of 4. When using ChIP, the top three methods were MERLIN, SCENIC and PIDC. Finally based on Perturb+ChIP, Pearson, PIDC were the best with a median rank of 3, followed by SCENIC and MERLIN with a median rank of 4. AUPR-based median rank varied depending upon the gold standard, which is consistent with overall lower correlation across method ranks when using AUPR (**Additional File 2**, **Figure S3**). In particular, using Perturb, the top three methods were SILGGM, Pearson and LEAP, using ChIP the top three methods were SCENIC, SCODE and Pearson and based on Perturb+ChIP, they were Pearson, SCENIC and PIDC.

Finally, we collated the metrics across gold standards and datasets (**Figure 3D**) and ranked each method based on its overall median rank across both types of gold standards (**Figure 3E**). Based on the median rank computed using F-score, Pearson, MERLIN and PIDC had the best median rank of 3, followed by SCENIC (rank of 4) and LEAP (rank of 5.5). Based on AUPR, Pearson, SCENIC and MERLIN were the top three ranking methods.

## Assessing scRNA-seq regulatory network inference algorithms based on local network metrics

The AUPR and F-score provide a quantification of the global accuracy of a regulatory network comparing the predicted and gold standard networks one edge at a time. However, from a biological application point of

view, it might be beneficial to examine if there are some smaller components of the regulatory network that are predicted better than others. Therefore, we next focused on a more granular view of the inferred networks by measuring the ability to predict the targets of individual transcription factors based on the fold enrichment of true targets of a transcription factor (TF) in the predicted set as described by Siahpirani et al [17] (**Methods**). We quantified the overlap using an FDR-corrected hypergeometric test P-value and called a TF "predictable" if it had a significant P-value ($< 0.05$). We used the number of significantly predictable TFs as another quantification of network performance (**Figure 4A**). Unlike AUPR and F-score where the performance of the algorithms was often close to random, the number of predictable TFs for the random baseline was no more than 1, and typically 0. Across datasets, the methods ranked consistently, with the exception of LEAP performing much worse in Shalek and Tran (FBS) datasets, and SCODE performing much better on the Gasch and Han datasets based on the ChIP gold standard. Between Perturb and ChIP gold standards, the rankings of the algorithms were more correlated than AUPR or F-score (Pearson correlation of 0.55 for predictable TFs vs correlation of 0.41 for F-score and 0.07 for AUPR (**Additional File 2**, **Figure S3**)), but recapitulated some preference of methods to perform better on Perturb versus ChIP datasets. We aggregated the metrics per gold standard type across datasets by computing the median rank (**Figure 4B**). Based on the Perturb gold standard, the top three methods were Pearson, PIDC and MERLIN, while based on ChIP, Peason and MERLIN were best followed by SCENIC and SCODE. Inspection of the number of predictable TFs showed there is a substantial variation across datasets and gold standards (**Figure 4C**). Finally, as in the AUPR and F-score comparison, we ranked the methods based on the predictable TFs from the three types of gold standards. Based on this ranking, the top performing methods were Pearson, SCENIC, MERLIN, and PIDC (**Figure 4D**), which was consistent with the rankings based on AUPR and F-score.

We next examined for each dataset which specific TFs were predicted by the different methods (**Figure 5**). This information can help determine if there are TFs that can be predicted by specific category of methods or by all methods. For the Han dataset which profiled transcriptional programs during lineage specification from the embryonic stem cell (ESC) state, we found several ESC-specific regulators such POU5F1, SOX2 and NANOG [18], as predictable TFs in multiple methods using the ChIP gold-standard (**Figure 5A**). We also found some lineage-specific TFs like CDX2 [19] and TBX3 [20] which serve as major regulators for different lineages predicted by a number of methods. When comparing the Perturb gold standard, we found

similar regulators as in the ChIP gold standard, but additionally regulators such as OTX2 and GATA3 for retinal [21] and hematopoietic lineages [22], respectively. SCODE had a number of predictable TFs in the Han dataset, however, many of them are general regulators such as SP1, YY1, POL2RA, ATF2. For the mouse cellular reprogramming datasets (Tran (FBS), Tran (A2S), Zhao), we observed a similar behavior with many of the methods identifying several developmental regulators such as Pou5f1, Esrrb and Sox2 in both ChIP and Perturb gold standards (**Figure 5B**, **Additional File 2**, **Figure S4**, **S5**). Importantly, when compared to the dendritic cell dataset (Shalek), we found a consistent over-representation of a different set of regulators across methods.This included Rel, Nfkb, Stat1 and Stat3 that are associated with immune response and were identified by several methods including SCENIC, MERLIN, PIDC, Scribe, Inferelator and Pearson (**Additional File 2**, **igure S6**). We found a similar behavior for the yeast datasets (**Figure 5C**, **Additional File 2**, **Figure S7**), where the methods uniformly were able to recover key regulators associated with stress response such as HAP4 (oxidative stress) and GCN4 (amino acid starvation, [23]). The overall fold enrichment for predictable TFs was higher for the yeast datasets compared to the mammalian datasets. Between the two yeast datasets, Jackson exhibited a greater enrichment of targets compared to Gasch, likely because of the fewer number of cells in the Gasch dataset. Overall, this shows that although the performance of methods based on global metrics such as F-score and AUPR is modest, methods are able to consistently recover relevant regulators for a particular system.

Finally, as large-scale gold standards themselves may be noisy, we curated a small gold standard network from regulatory interactions reported in literature for the ESC state [24–30]. The regulatory network has 35 regulators and 90 target genes (some of which included regulators as well) connected by 267 edges (**Additional File 3**, gold_standard_datasets.zip). We next used the top 5000 edges from the networks inferred on the Tran A2S dataset and depicted true positives (found by a method) and false negatives (missed by a method) among 199 of the 267 interactions that remained after removing edges due to missing expression (**Methods**). Most of the methods recovered cross regulatory interactions between Nanog, Sox2 and Pou5f1, which are master regulators for establishing the ESC state (**Figure 5D**). Methods that recovered the fewest number of true positives included SCODE, knnDREMI and PIDC. Interestingly SCRIBE which like PIDC is based on information theory was able to infer many more true positives (14) compared to PIDC (9). Overall the performance based not the recovery of true positive edges was consistent with our global and predictable

TF metrics but also highlighted additional properties of the network inference methods in terms of recovery of curated regulatory interactions.

### Similarity of inferred networks from different methods depends upon the dataset

We next asked the extent to which the algorithms agree in their predictions by measuring the Jaccard similarity and F-score between the top ranked edges of each pair of inferred networks (**Figure 6**). The magnitude of the similarity increased as more edges were considered but the overall trend of pairwise similarity remained the same (**Additional File 2**, **Figure S8**, **S9**). We focused on the Jaccard scores obtained on the top 5k edges for consistency with our other results; similar behavior was observed for F-score as well (**Additional File 2**, **Figure S10**). We grouped the methods in two ways. First, we ordered the methods in the same way in each dataset based on the median similarity of inferred networks and visualized the pairwise similarity for each dataset (**Figure 6A**). This allowed us to examine the change in similarity for a pair of methods as a function of datasets. Scribe's similarity to other methods depended upon the dataset, however, it is most similar to PIDC, which is also based on information theoretic metrics. LEAP was most similar to Correlation but this also depended upon the dataset; e.g. high Jaccard score in Gasch, Jackson and Zhao and low scores in other datasets. Some datasets resulted in more similar networks compared to others. In particular, methods were most similar (high Jaccard coefficient) in the Han and Jackson datasets and were least similar for the Shalek and Gasch. These two datasets were also the smallest of the two datasets, which could explain the lower similarity. Methods that tended to have the highest similarity with other methods were Pearson, SCENIC, Inferelator and PIDC.

We next ordered the methods based on the Jaccard similarity per dataset producing dataset-specific orderings (**Figure 6B**). SILGGM and knnDREMI generally learned different networks compared to other methods, including those based on the same class of models, e.g. information theory for knnDREMI (PIDC, Scribe) and probabilistic models for SILGGM (MERLIN, Inferelator). The methods that exhibited the highest similarity across datasets were LEAP, PIDC and Pearson, with the exception of Tran (FBS), Tran (A2S) and Shalek where LEAP learned more different networks and Gasch where PIDC was different from Pearson and LEAP. This was followed by SCENIC, MERLIN and Inferelator, which also tended to group together in multiple datasets. In summary these comparisons revealed that the similarities between methods

depended upon the dataset, however, several of the methods such as PIDC and Correlation tended to infer similar networks across datasets. Furthermore, the networks inferred from the same family of algorithms were not necessarily more similar than from different families of algorithms.

### Imputation of scRNA-seq datasets has a nominal effect on network inference

Single-cell expression data is characterized by a large number of zero expression counts which can arise due to technical (low depth) or biological reasons (cell type-specific expression). To address this problem, several methods have been developed to impute the value of missing expression counts [31]. We applied the MAGIC algorithm to each dataset [4], which was one of the top imputation methods from a recent benchmarking study [31]. MAGIC computes pairwise similarity between cells and creates a Markov transition graph based on these similarities. It then "diffuses" expression counts among similar cells based on this Markov transition graph. We inferred networks using the imputed data from each experiment and compared their AUPR, F-score and predictable TF metrics with those of the networks inferred from the original sparse datasets (**Figure 7**). Based on F-score (**Figure 7A**), most methods did not benefit from imputation across datasets. An exception to this were the Shalek and Gasch datasets, which as noted before are among the smaller datasets. We examined the ratio of F-score after and before imputation across different gold standards and datasets and found SCODE and kNN-DREMI to improve when using imputed data (**Figure 7B**). However, imputation generally did not benefit the network inference procedure with most algorithms exhibiting a ratio $<1$ across datasets and gold standards. We repeated these comparisons based on AUPR (**Figure 7C**, **D**) and predictable TFs (**Figure 7E**, **F**). AUPR did not change much with imputation, however, we noted a similar boost in performance for the Shalek dataset. Finally, based on predictable TFs, performance did not generally improve with the exception of the Tran (A2S) and Tran (FBS) datasets where algorithms such as SCODE and knnDREMI showed a benefit (**Figure 7E**, **F**). Taken together, our experiments across different datasets and gold standards showed that imputation could benefit in a handful of cases for algorithms and datasets, but generally did not help with network inference for most methods compared.

### Comparison of network inference from single cell versus bulk expression datasets

Before the availability of scRNA-seq datasets, expression-based network inference was performed with bulk expression datasets, which required large sample sizes. Such datasets are available from a single study in yeast [23] or can be created by collecting studies from public gene expression omnibus. Among the conditions (species and cell states) we examined in this study, we had matching bulk expression datasets in yeast (stress conditions), mouse ESC (mESC) and hESC (hESC). We therefore applied network inference on these bulk expression datasets using five state of the art algorithms and compared the performance of the methods based on the same gold standards and metrics of F-score, AUPR and predictable TFs (**Figure 8**, **Methods**). For the single cell comparison, we used the inferred networks from the Gasch dataset for yeast, Tran (A2S) for mESC and Han dataset for hESCs.

Based on F-score scRNA-seq-based networks exhibited a moderate improvement on the Han (human ESC) dataset and Gasch dataset when using either the Perturb+ChIP or Perturb gold standards respectively. The results were largely consistent for AUPR and F-score. However, when comparing predictable TFs, we observed a marked decrease in the number of predictable TFs when comparing the results for mouse ESC and human ESC datasets using ChIP and ChIP+Perturb. Overall this suggests that individual scRNA-seq datasets capture meaningful variation that offers comparable performance to bulk expression datasets collected from a large number of experiments.

## Discussion

The rapid growth of single cell RNA-seq datasets has opened up remarkable opportunities in the field of expression-based gene regulatory network inference. Accordingly a number of efforts have been invested for developing and applying regulatory network inference algorithms to single cell datasets. Here we benchmarked the computing requirements, and overall performance of methods in network structure recovery across a large number of datasets spanning different species and model cell lines. Our work expands on previous efforts by considering larger scRNA-seq datasets across multiple model systems, additional classes of methods, different types of metrics and gold standards that can provide biological insight into the predictions made by the different methods.

We evaluated methods based on their computing requirements as well as performance on different gold standard networks. Due to excessive computing requirements we could not rank several methods that were specifically geared towards handling the statistical properties of single cell datasets, such as SCHiRM, HurdleNormal and BTR. As single cell datasets grow, efficient implementation of algorithms would important. Our evaluation metrics included both global metrics such as AUPR and F-score as well as local metrics such as the number regulators whose targets could be predicted significantly. Based on AUPR and F-score, the overall performance of methods remains modestly better than random, however, we find predictable TFs as a more sensitive metric that highlights the strengths of network inference methods. Importantly, different methods were able to recapitulate relevant regulators for the system of interest, for example key stem cell regulators in the developmental datasets and immune response regulators in the dendritic cell dataset. No single method was a winner across all datasets and gold standards, however, based on the overall performance and computing requirements of the methods, the methods PIDC, SCENIC, MERLIN and PEARSON are among the top methods.

One challenge with scRNA-seq datasets is the high proportion of zeros, which can be due to both biological as well as technical reasons. As imputation has been proposed as an approach to handle sparse datasets such as scRNA-seq datasets, we studied the impact of imputation on the quality of our inferred networks. As such imputation did not improve the performance of most methods, however, the datasets where it did benefit tended to be those with a relatively smaller number of cells. One caveat in our analysis is that we considered only one imputation method, MAGIC, which was shown to be one of the top imputing methods. A direction of future work is to consider additional imputation methods to examine the impact of imputation on network inference.

Single cell transcriptomic datasets have the advantage that a single experiment can produce large number of samples that are comparable or larger than existing bulk datasets that have been used for network inference. Therefore, we compared the quality of the inferred networks from bulk and single cell RNA-seq datasets for yeast, mouse and human using methods for both bulk and single cell datasets. We find that scRNA-seq datasets, despite being sparse are able to capture sufficient meaningful biological variation and perform at par when using bulk RNA-seq datasets. Our study however is not perfect since the bulk and single cell datasets were collected from different sources. Generating controlled datasets capturing bulk

and single cell profiles for the same system could provide additional insight into the relative advantage of scRNA-seq datasets for inferring gene regulatory networks.

The current study compared methods that rely solely on expression. Experiments from bulk data have shown that incorporating prior information for both constraining the network structure [17, 32] as well as estimating hidden regulatory activity [33] can benefit network inference. A direction of future work is to both develop as well benchmark methods that incorporate prior knowledge into the estimation of inferred networks. The gold standards and datasets that we have collected in our work should be beneficial for these future studies. Another direction of future work is to leverage the inherent heterogeneity and population structure of scRNA-seq datasets. Methods based multi-task learning is a promising framework to model population and network heterogeneity [34, 35]. One surprising finding that we made was the relatively good performance of a simple Pearson's correlation based metric. This could be an artifact of our gold standards which are admittedly imperfect. Generation of improved gold standards, especially for mammalian systems, based on newer high throughput perturbation studies such as Perturb-seq [36] could significantly benefit our ability to infer genome-scale gene regulatory networks.

# Methods

## Dataset pre-processing

We obtained a total of seven scRNA-seq datasets from three different species, human, mouse and yeast (see **Additional File** 1, **Table S1** for details). For each dataset, we filtered out genes that appeared in fewer than 50 cells and cells with fewer than 2000 total UMIs, then square-root transformed the expression counts and depth-normalized the matrix. For performing the memory and time profiling experiments, we used the Han human pluripotent stem cell dataset [37], which had a total of 5,520 cells and selected number of genes, $n \in \{10, 25, 50, 100, 250, 500, 1000, 2000, 5000, 8000\}$. Each algorithm was applied on each dataset and profiled for memory requirements and execution time. Runtime and memory were captured using the `/usr/bin/time -v` command with Linux.

## Regulatory network inference algorithms for single cell RNA-seq datasets

To assess algorithms for regulatory network inference, we considered recently published methods that span a variety of modeling formalisms including ordinary differential equations, probabilistic models and information theoretic methods. We considered algorithms developed for both bulk and single cell RNA-seq datasets for network inference. Here we provide a brief summary of each algorithm included in our analyses (see **Additional File** 1, **Table S2** for more details about the algorithm implementation).

**BTR.** BTR is based on a Boolean model representation of a gene regulatory network [7]. Here a regulatory interaction is inferred by assuming a gene has a value of either 0 or 1 and the state of a target gene is specified by a boolean function of the state of its regulators.

**HurdleNormal.** This approach is based on a multi-variate Hurdle model which is a probabilistic graphical model [12] suited for handling zero-inflated datasets. The graph structure, which represents the gene regulatory network is estimated by first defining the model in terms of a set of conditional distributions and solving these regression problems by solving a set of penalized neighborhood selection problems.

**scHiRM.** Single cell hierarchical regression model (SCHiRM) is a probabilistic model based framework suited for handing sparse datasets such as those from scRNA-seq experiments [11]. The expression level of a gene is modeled as a Poisson log distribution with overdispersion. A regulatory network is estimated by a regression function that links the expected expression level of a gene to the expected expression level of target genes using a hierarchical framework.

**kNN-DREMI.** kNN-DREMI recovers gene-gene relationships by adapting the DREMI algorithm [38] to scRNA-seq data [4]. The DREMI makes use of a conditional density to estimate mutual information between pairs of genes. It makes use of a heat-diffusion based kernel density estimator to estimate the joint density, followed by conditional density estimation. kNN-DREMI replaces the heat diffusion kernel with a knn-graph based density estimator to handle the high-dimensionality and sparsity of the scRNA-seq data.

**LEAP.** LEAP [10] requires cells be ordered along a pseudotime trajectory, and computes pairwise correlation between genes $i$ and $j$ at various lags along this trajectory. The algorithm chooses a window size $s$, and for all possible start times $t$ and lags $l$, computes the Pearson's correlation coefficient between the first $s$ observations of $i$ beginning at time $t$ and the $s$ observations of $j$ beginning at time $t + l$. The score for a regulatory relationship from $i$ to $j$ is the maximum of all computed correlation coefficients $\rho_{ij}$.

**PIDC.** PIDC [2] is based on an information-theoretic framework and estimates a specific type of multivariate information called Partial Information Decomposition (PID) between three random variables, each representing a gene. PID is defined using the "unique", "synergistic" and "redundant" information between two variables, $X$ and $Y$ and a third variable $Z$. The "unique" information is the contribution to the multi-information specifically from two of the variables, "redundant" is the contribution from either $X$ or $Y$, while "synergistic" is the contribution from both $X$ and $Y$ to $Z$. In the context of network inference, PIDC makes use of the ratio of the unique information to the mutual information between two variables and establishes an edge based on the magnitude of this ratio.

**SCENIC.** SCENIC is based on the GENIE3 [39] network inference algorithm, which infers a gene regulatory network by solving a set of regression problems using ensemble regression trees. SCENIC also offers two additional steps, RcisTarget for enriched motif identification in the target set of a TF and AUCell to determine the activity of a set of targets of a TF in each cell. For our experiments we only used the GENIE3 outputs from SCENIC.

**SCODE.** SCODE is based on an Ordinary Differential Equation based formulation of a gene regulatory network and requires pseudotime as input in addition to the expression matrix for learning the regulatory network [6]. SCODE makes use of a lower dimensional representation of the major expression dynamics of regulators, $\mathbf{z}$ and uses a linear transformation matrix, $\mathbf{W}$ to estimate the observed expression dynamics of all genes from $\mathbf{z}$. The gene regulatory network is represented as a function of $\mathbf{W}$ and estimated by solving a set of linear regressions.

**Scribe.** Scribe [3] is based on an information theoretic framework that uses pseudotime as an additional input for regulatory network inference. Scribe is based on a specific type of information measure called Restricted Directed Information that measures the mutual information between the current state of a target gene and a previous state of a regulator, conditioned on the previous state of the target gene. Scribe additionally uses a bias correction scheme of the samples such that cells capturing transitioning states are weighted more towards establishing the regulatory relationships.

**SILGGM.** SILGGM makes use of a Gaussian graphical model [1] representation of a gene regulatory network and offers several recent and efficient implementations for estimating conditional dependence between random variables (genes): bivariate nodewise scaled Lasso, de-sparsified nodewise scaled Lasso, de-sparsified graphical Lasso, and GGM estimation with false discovery rate (FDR) control using Lasso or scaled Lasso. For our experiments, we applied SILGGM with de-sparsified nodewise scaled Lasso implementation of neighborhood selection, which is the default algorithm for SILGGM.

**Bulk dataset network inference methods.** In addition to above methods which are specifically geared for scRNA-seq datasets, we also considered methods developed for bulk RNA-seq datasets: MERLIN [40] and Inferelator [32]. MERLIN is based on a dependency network learning framework and employs a probabilistic graphical model prior to impose modularity prior. We ran MERLIN with stability selection and default parameter settings for the sparsity $p = -5$, modularity $r = 4$, and clustering $h = 0.6$. Inferelator is also based on a dependency network learning framework which uses regularized regression for network inference. We used the BBSR algorithm within Inferelator, v0.3.0 downloaded from `https://github.com/simonsfoundation/inferelator_ng` and used Inferelator's internal stability selection to obtain edge confidence.

**Pearson and random networks.** Our analyses included two inferred networks as baselines. The "Pearson" network was an undirected fully connected network, with edges weighted by the Pearson correlation between each pair of genes over all cells. The "random" network was an undirected fully connected network, with edge weights drawn from a uniformly distribution $\text{Unif}(0, 1)$.

## Comparison to network inference on bulk expression data

Methods used to infer networks on bulk data include GENIE3 [39], TIGRESS [41], MERLIN [40]. To run GENIE3 we used $K = \texttt{sqrt}$ and $\texttt{nb_trees} = 1000$, to run TIGRESS we used $\texttt{R}$=$1,000$, $\alpha$=0.3, and $\texttt{L}$=30, and to run MERLIN we used default configuration of sparsity $p = -5$, modularity $r = 4$, and clustering cut-off $h = 0.6$. MERLIN was run in a stability selection framework, where 100 subsets were created by randomly selecting half the samples in the input expression matrix, a network was inferred on each subset, and a consensus network was made by calculating the frequency of observing each edge in the 100 inferred networks. As in the single cell case, we created a network where edges were ranked by absolute value of Pearson correlation of expression profile of genes. PGG (**P**er **G**ene **G**reedy) is MERLIN without modularity prior and was run using the same parameters and in a stability selection framework similar to MERLIN. To run LARS-EN we used `lambda2=` 1E-6 and the L1 penalty was set in a 10 fold cross validation. We ran LARS-EN in a stability selection procedure similar to MERLIN. We used the MATLAB implementation of LARS-EN from the Imm3897 package downloaded from `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3897`.

We performed these comparisons on three datasets that came from comparable conditions for bulk and single cell. In yeast, we used the environmental stress response data from Gasch et al [23] for bulk to match our single cell RNA-seq dataset (Gasch) also generated by Gasch and colleagues [15]. For human and mouse, we curated publicly available datasets from the gene expression omnibus (**Additional File 1**, **Table S3**) for human and mouse embryonic stem cell state (Siahpirani in preparation). We compared the performance of bulk human dataset to that from the Han dataset [37] and the performance of the bulk mouse dataset to that from the A2S (Tran) dataset [42]. Because the gene sets of the bulk and single cell data were different, we filtered the gold standards to include TFs and targets that were in both datasets to make the results comparable.

## Application of network inference algorithms

**Execution and post-processing** Each algorithm was applied to the processed expression matrices on a local high-performance computing cluster using HTCondor `https://chtc.cs.wisc.edu`. Where possible, we deployed each algorithm in a stability selection mode. Specifically, we selected 50% of the

cells from each expression matrix randomly, applied the algorithm to this subset, and repeated this process 100 times, averaging the interaction scores over the 100 subset networks to obtain a final inferred network. As some algorithms perform a similar downsampling internally (SCENIC, Inferelator), we did not run these algorithms within our external stability selection framework. Other algorithms could not be run 100 times due to excessive runtimes (PIDC, Scribe). After completing the runs of each algorithm, we formatted the variety of network formats to a single three-column file (transcription factor, target, score). We restricted all networks to include only edges where the "regulator" was a known transcription factor from one of our experimentally derived gold standard networks. The final networks were used in the evaluations of network accuracy and similarity (detailed below) and are available in **Additional File 3**.

**Estimation of pseudo time.** Some algorithms (LEAP, SCODE, Scribe) incorporate knowledge of cellular trajectory and pseudotime, which was learned using Monocle 2 [43]. To avoid biasing the trajectories with a choice of marker genes, we followed the best practices in the Monocle documentation to learn trajectories in an unsupervised approach. In brief, we clustered the cells using Monocle's density-peak clustering algorithm after dimensionality reduction with t-SNE. We next tested for genes differentially expressed across these clusters, and selected the top thousand most significantly differential genes (as measured by q-value) to use as marker genes. Finally, we learned a DDRTree embedding of the expression in each cell and applied Monocle's cell ordering function to infer the position of each cell on a trajectory and their corresponding pseudotime assignments. We used Monocle trajectories for all datasets except the yeast Jackson dataset for which we used Diffusion Pseudo Time (DPT) [44] due to high runtime of the Monocle software.

## Description of gold standards

We curated multiple experimentally derived networks of regulatory interactions from published databases and the literature (**Additional File 3**) to serve as gold standards for our network inference algorithms. These experiments are typically based on ChIP-chip, ChIP-seq or regulator perturbation followed by global transcriptome profiling. We obtained multiple networks based on ChIP and TF perturbation experiments for each organism and cell type. When multiple ChIP or perturbation interactions were available we took a union of the networks. We refer to the ChIP derived gold standard as "ChIP" and the perturbation derived

gold standard as "Perturb". Finally, we took the intersection of these two unions, as the third primary gold standard networks for our evaluations of network accuracy (ChIP+Perturb). Finally, we created a fourth embryonic stem cell (ESC) specific gold standard network from the primary literature [24–30].

## Evaluation metrics

To evaluate the accuracy of the inferred networks, we used three metrics - Area under the Precision Recall curve (AUPR), F-score, and predictable TFs - between each network and each relevant gold standard network. Prior to computing each accuracy measure, we filtered both the inferred network and the gold standard to include only edges that contained transcription factors and target genes from the intersection of the gold standard's gene set with the original expression matrix's gene set. This was to eliminate any penalty for missing edges included in the gold standard that could not be inferred by the algorithm due to lack of expression in the original data.

**Area Under the Precision Recall curve (AUPR).** We used AUPR to measure the accuracy of the global network structure. Specifically, we sorted the inferred edges in descending order of confidence, and computed the precision and recall with respect to a gold standard as edges were sequentially included. We report the area under this curve as AUPR. We used the script `auc.java` from `https://github.com/Arun-George-Zachariah/AUC` [45].

**F-score.** We used F-score of the top $x$ edges, $x \in \{100, 300, 500, 1000, 3000, 5000, 10000, 30000, 50000\}$ to measure the accuracy of the most confident edges in an inferred network. Specifically, the F-score is computed as the harmonic mean of the precision and recall of the top $x$ edges with respect to a gold standard.

**Predictable TFs.** We used predictable TFs to obtain a granular measure of network accuracy at the individual transcription factor (TF)'s target set. Specifically, for each transcription factor in the inferred network, we use the hypergeometric test to assess whether the overlap between the predicted targets and the targets in the gold standard was significantly higher than random. The P-values were corrected using FDR for multiple hypothesis correction. We considered a P-value $< 0.05$ as significant overlap. Each algorithm was ranked

based on the number of predictable TFs.

## Availability of data and materials

Datasets used in our experiments and analyses have been deposited with DOI `10.5281/zenodo.5909090`, along with the following groups of files which were too large to include in the manuscript as supplementary materials:

- Gold standard networks used as ground truth to measure accuracy of the inferred networks.
- Networks generated from network inference methods.

All scripts for executing and evaluating the algorithms are available at `https://github.com/Roy-lab/scRNAseq_NetInference`.

## Competing interests

The authors declare that they have no competing interests.

## Author's contributions

M.S. and S.R. conceptualized the project. M.S., S.G.M, A.F.S, V.P. performed data download and preprocessing. M.S., S.G.M., A.F.S, V.P., J.L. performed benchmarking experiments. M.S., A.F.S., J.L, and J.S. ran evaluation scripts for examining results. J.L., S.G.M, M.S. and S.R. wrote the paper.

## Acknowledgements

This work was supported by the NIH grants R01GM117339 and U01EB029371, a UW Data Science Foundation grant. M.S. acknowledges support from an NIH T15 training fellowship (T15 LM007359) and S.G.M acknowledges support from an NIH T32 training fellowship (T32 GM007133). The authors thank the Center for High-throughput Computing for enabling the large-scale computational experiments of this project.

## References


[1] Zhang, R., Ren, Z., Chen, W.: SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. PLOS Computational Biology **14**(8), 1006369 (2018). doi:10.1371/journal.pcbi.1006369

[2] Chan, T.E., Stumpf, M.P.H., Babtie, A.C.: Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. Cell Systems **5**(3), 251–2673 (2017). doi:10.1016/j.cels.2017.08.014

[3] Qiu, X., Rahimzamani, A., Wang, L., Mao, Q., Durham, T., McFaline-Figueroa, J.L., Saunders, L., Trapnell, C., Kannan, S.: Towards inferring causal gene regulatory networks from single cell expression Measurements. bioRxiv, 426981 (2018). doi:10.1101/426981

[4] Dijk, D.v., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., al, e.: Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell **174**(3), 716–72927 (2018). doi:10.1016/j.cell.2018.05.061

[5] Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., al, e.: SCENIC: single-cell regulatory network inference and clustering. Nature Methods **14**(11), 1083–1086 (2017). doi:10.1038/nmeth.4463

[6] Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S.H., Ko, S.B.H., Gouda, N., Hayashi, T., Nikaido, I.: SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics **33**(15), 2314–2321 (2017). doi:10.1093/bioinformatics/btx194

[7] Lim, C.Y., Wang, H., Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J., Göttgens, B.: BTR: training asynchronous Boolean models using single-cell expression data. BMC Bioinformatics **17**(1), 355 (2016). doi:10.1186/s12859-016-1235-y

[8] Fiers, M.W.E.J., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., Aerts, S.: Mapping gene regulatory networks from single-cell omics data. Briefings in Functional Genomics **17**(4), 246–254 (2018). doi:10.1093/bfgp/elx046. Accessed 2019-09-02

[9] Nguyen, H., Tran, D., Tran, B., Pehlivan, B., Nguyen, T.: A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. Briefings in Bioinformatics **22**(3), 190 (2021). doi:10.1093/bib/bbaa190. Accessed 2022-01-27

[10] Specht, A.T., Li, J.: LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. Bioinformatics **33**(5), 764–766 (2017). doi:10.1093/bioinformatics/btw729

[11] Intosalmi, J., Mannerström, H., Hiltunen, S., Lähdesmäki, H.: SCHiRM: Single Cell Hierarchical Regression Model to detect dependencies in read count data. bioRxiv, 335695 (2018). doi:10.1101/335695

[12] McDavid, A., Gottardo, R., Simon, N., Drton, M.: Graphical Models for Zero-Inflated Single Cell Gene Expression. The Annals of Applied Statistics **13**(2), 848–873 (2019). doi:10.1214/18-AOAS1213

[13] Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., Murali, T.M.: Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature Methods **17**(2), 147–154 (2020). doi:10.1038/s41592-019-0690-6

[14] Chen, S., Mar, J.C.: Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC bioinformatics **19**(1), 232 (2018). doi:10.1186/s12859-018-2217-z

[15] Gasch, A.P., Yu, F.B., Hose, J., Escalante, L.E., Place, M., Bacher, R., Kanbar, J., Ciobanu, D., Sandor, L., Grigoriev, I.V., al, e.: Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. PLOS Biology **15**(12), 2004050 (2017). doi:10.1371/journal.pbio.2004050

[16] Zhao, T., Fu, Y., Zhu, J., Liu, Y., Zhang, Q., Yi, Z., Chen, S., Jiao, Z., Xu, X., Xu, J., etal.: Single-Cell RNA-Seq Reveals Dynamic Early Embryonic-like Programs during Chemical Reprogramming. Cell Stem Cell **23**(1), 31–457 (2018). doi:10.1016/j.stem.2018.05.025

[17] Siahpirani, A.F., Roy, S.: A prior-based integrative framework for functional transcriptional regulatory network inference. Nucleic Acids Research **45**(4), 21–21 (2017). doi:10.1093/nar/gkw963. Accessed 2021-06-01

[18] Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al.: Core transcriptional regulatory circuitry in human embryonic stem cells. Cell **122**(6), 947–956 (2005). doi:10.1016/j.cell.2005.08.020

[19] Strumpf, D., Mao, C.-A., Yamanaka, Y., Ralston, A., Chawengsaksophak, K., Beck, F., Rossant, J.: Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. Development **132**(9), 2093–2102 (2005). doi:10.1242/dev.01801

[20] Esmailpour, T., Huang, T.: Tbx3 promotes human embryonic stem cell proliferation and neuroepithelial differentiation in a differentiation stage-dependent manner. Stem Cells **30**(10), 2152–2163 (2012). doi:10.1002/stem.1187

[21] Yamamoto, H., Kon, T., Omori, Y., Furukawa, T.: Functional and Evolutionary Diversification of Otx2 and Crx in Vertebrate Retinal Photoreceptor and Bipolar Cell Development. Cell Reports **30**(3), 658–6715 (2020). doi:10.1016/j.celrep.2019.12.072. Accessed 2021-06-01

[22] Ku, C.-J., Hosoya, T., Maillard, I., Engel, J.D.: GATA-3 regulates hematopoietic stem cell maintenance and cell-cycle entry. Blood **119**(10), 2242–2251 (2012). doi:10.1182/blood-2011-07-366070. Accessed 2021-06-01

[23] Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., Brown, P.O.: Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. Mol. Biol. Cell **11**(12), 4241–4257 (2000)

[24] Zhou, Q., Chipperfield, H., Melton, D.A., Wong, W.H.: A gene regulatory network in mouse embryonic stem cells. Proceedings of the National Academy of Sciences **104**(42), 16438–16443 (2007). doi:10.1073/pnas.0701014104. http://www.pnas.org/content/104/42/16438.full.pdf

[25] Kim, J., Chu, J., Shen, X., Wang, J., Orkin, S.H.: An extended transcriptional network for pluripotency of embryonic stem cells. Cell **132**(6), 1049–1061 (2008). doi:10.1016/j.cell.2008.02.039

[26] Young, R.A.: Control of the embryonic stem cell state. Cell **144**(6), 940–954 (2011). doi:10.1016/j.cell.2011.01.032

[27] Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., Jaenisch, R.: Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell **150**(6), 1209–1222 (2012). doi:10.1016/j.cell.2012.08.023

[28] Dunn, S.J., Martello, G., Yordanov, B., Emmott, S., Smith, A.G.: Defining an essential transcription factor program for naïve pluripotency. Science **344**(6188), 1156–1160 (2014). doi:10.1126/science.1248882

[29] Xu, H., Ang, Y.-S., Sevilla, A., Lemischka, I.R., Ma'ayan, A.: Construction and validation of a regulatory network for pluripotency and Self-Renewal of mouse embryonic stem cells. PLoS Comput Biol **10**(8), 1003777 (2014). doi:10.1371/journal.pcbi.1003777

[30] Malleshaiah, M., Padi, M., Rué, P., Quackenbush, J., Martinez-Arias, A., Gunawardena, J.: Nac1 coordinates a sub-network of pluripotency factors to regulate embryonic stem cell differentiation. Cell Reports **14**(5), 1181–1194 (2016). doi:10.1016/j.celrep.2015.12.101

[31] Hou, W., Ji, Z., Ji, H., Hicks, S.C.: A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biology **21**(1), 218 (2020). doi:10.1186/s13059-020-02132-x. Accessed 2021-05-31

[32] Greenfield, A., Hafemeister, C., Bonneau, R.: Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics (Oxford, England) **29**(8), 1060–1067 (2013). doi:10.1093/bioinformatics/btt099

[33] Miraldi, E.R., Pokrovskii, M., Watters, A., Castro, D.M., Veaux, N.D., Hall, J.A., Lee, J.-Y., Ciofani, M., Madar, A., Carriero, N., Littman, D.R., Bonneau, R.: Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. Genome Research (2019). doi:10.1101/gr.238253.118. Company: Cold Spring Harbor Laboratory Press Distributor: Cold

Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Accessed 2021-06-01

[34] Castro, D.M., Veaux, N.R.d., Miraldi, E.R., Bonneau, R.: Multi-study inference of regulatory networks for more accurate models of gene regulation. PLOS Computational Biology **15**(1), 1006591 (2019). doi:10.1371/journal.pcbi.1006591. Publisher: Public Library of Science. Accessed 2021-06-01

[35] Pierson, E., Consortium, t.G., Koller, D., Battle, A., Mostafavi, S.: Sharing and Specificity of Co-expression Networks across 35 Human Tissues. PLOS Computational Biology **11**(5), 1004220 (2015). doi:10.1371/journal.pcbi.1004220. Publisher: Public Library of Science. Accessed 2021-06-01

[36] Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T.M., Lander, E.S., Weissman, J.S., Friedman, N., Regev, A.: Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell **167**(7), 1853–186617 (2016). doi:10.1016/j.cell.2016.11.038. Accessed 2021-06-01

[37] Han, X., Chen, H., Huang, D., Chen, H., Fei, L., Cheng, C., Huang, H., Yuan, G.-C., Guo, G.: Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. Genome Biology **19**(1), 47 (2018). doi:10.1186/s13059-018-1426-0

[38] Krishnaswamy, S., Spitzer, M.H., Mingueneau, M., Bendall, S.C., Litvin, O., Stone, E., Pe'er, D., Nolan, G.P.: Systems biology. Conditional density-based analysis of T cell signaling in single-cell data. Science (New York, N.Y.) **346**(6213), 1250689 (2014). doi:10.1126/science.1250689

[39] Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLOS ONE **5**(9), 12776 (2010). doi:10.1371/journal.pone.0012776

[40] Roy, S., Lagree, S., Hou, Z., Thomson, J.A., Stewart, R., Gasch, A.P.: Integrated Module and Gene-Specific Regulatory Inference Implicates Upstream Signaling Networks. PLOS Computational Biology **9**(10), 1003252 (2013). doi:10.1371/journal.pcbi.1003252

[41] Haury, A.-C.C., Mordelet, F., Vera-Licona, P., Vert, J.-P.P.: TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. BMC systems biology **6**(1), 145 (2012). doi:10.1186/1752-0509-6-145

[42] Tran, K.A., Pietrzak, S.J., Zaidan, N.Z., Siahpirani, A.F., McCalla, S.G., Zhou, A.S., Iyer, G., Roy, S., Sridharan, R.: Defining Reprogramming Checkpoints from Single-Cell Analyses of Induced Pluripotency. Cell Reports **27**(6), 1726–17415 (2019). doi:10.1016/j.celrep.2019.04.056

[43] Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., Trapnell, C.: Reversed graph embedding resolves complex single-cell trajectories. Nature Methods **14**(10), 979–982 (2017). doi:10.1038/nmeth.4402. Number: 10 Publisher: Nature Publishing Group. Accessed 2021-06-01

[44] Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, Fabian J Theis: Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods **13**(10), 3971 (2016). doi:10.1038/nmeth.3971

[45] Davis, J., Goadrich, M.: The Relationship Between Precision-Recall and ROC Curves. In: Proceedings of the 23rd International Conference on Machine Learning. ICML '06, pp. 233–240. ACM, Pittsburgh, Pennsylvania, USA (2006). doi:10.1145/1143844.1143874. http://dx.doi.org/10.1145/1143844.1143874

# Figures

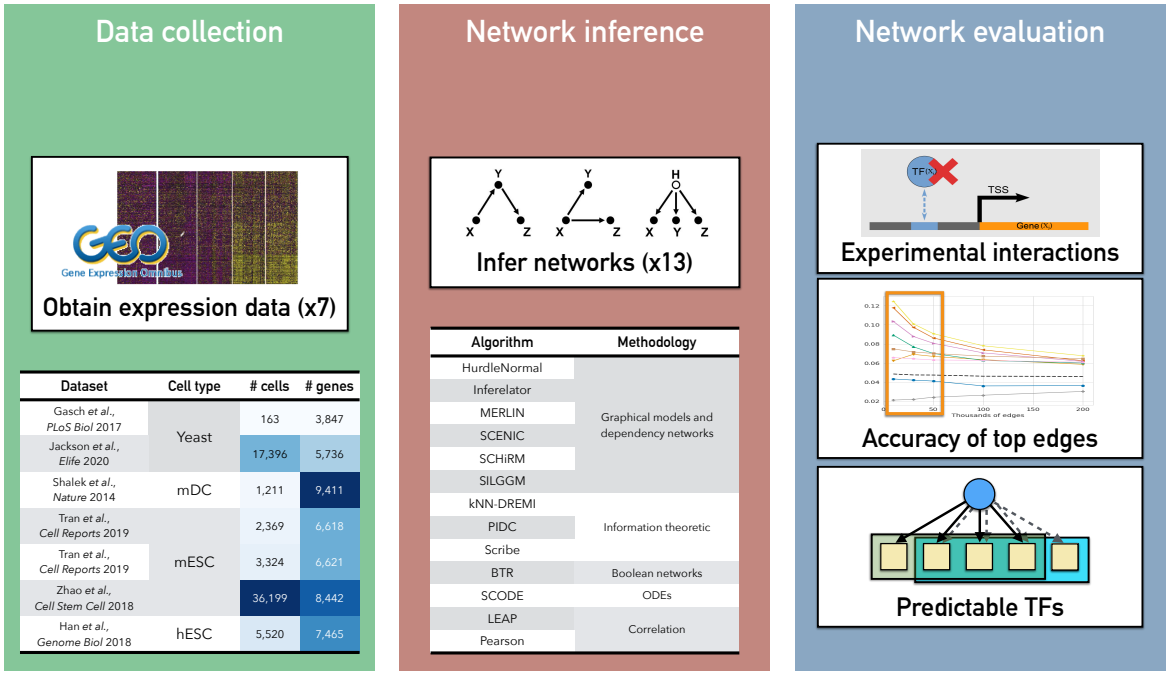



**Fig 1.** Overview of network inference benchmarking study.

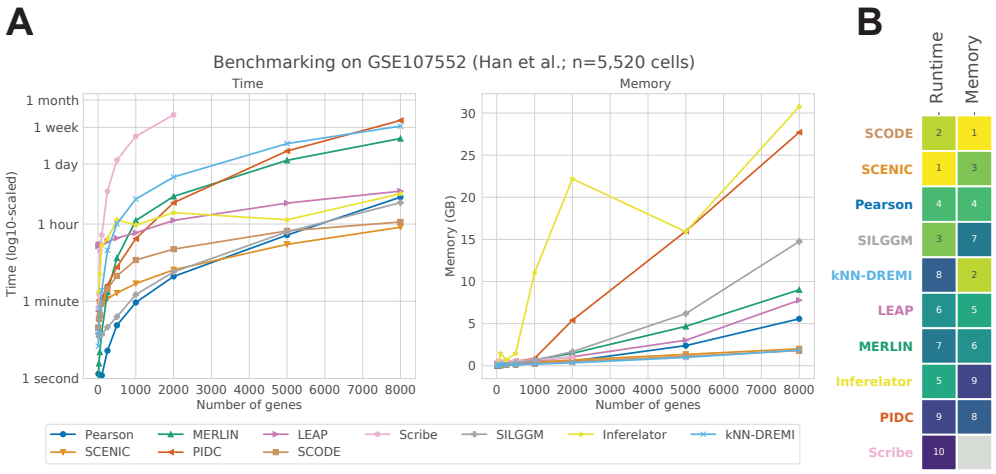



**Fig 2.** Benchmarking runtime and memory requirements. The x-axis shows the number of genes and the y-axis denotes the memory or time needed to run the algorithm. Some algorithms such as SCHiRM, HurdleNormal and BTR did not complete in a reasonable amount of time and were excluded from downstream analyses. These benchmarks of resource consumption were computed using all 5,520 cells in the Han et al. dataset.

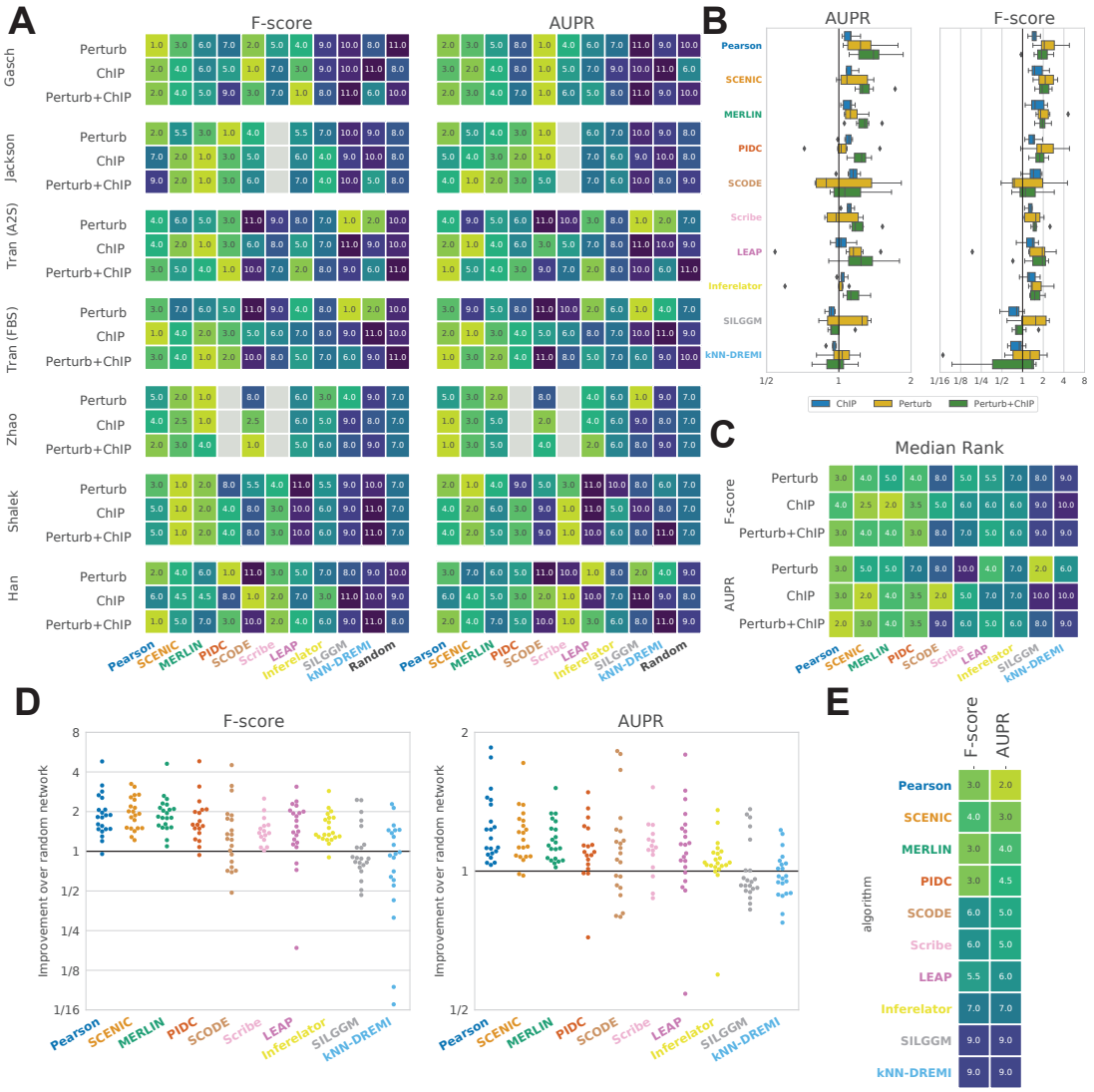



**Fig 3.** Algorithm performance across different global metrics of network structure recovery. **A.** Algorithm ranks in each of the 21 comparisons (7 datasets x 3 gold standard networks) based on AUPR and F-score. Gray cells represent the missing experiments due to scalability issue of algorithms on large datasets. **B.** Box plots showing distribution of relative performance (AUPR or F-score) compared to random (defined by a randomly weighted network with the same set of nodes and edges as the dataset) across different datasets for each type of gold standard: ChIP, Perturb and their intersection, Perturb+ChIP. **C.** Median rank of algorithm performance with respect to each type of gold standard (median over 7 datasets). **D.** Distribution of AUPR and F-score fold improvement over random performance in the same comparison. Algorithms are ordered as shown in panel **E.E.** Median rank of algorithms across all 21 comparisons using F-score and AUPR. Rows are ordered based on the mean of rank of F-score and AUPR.

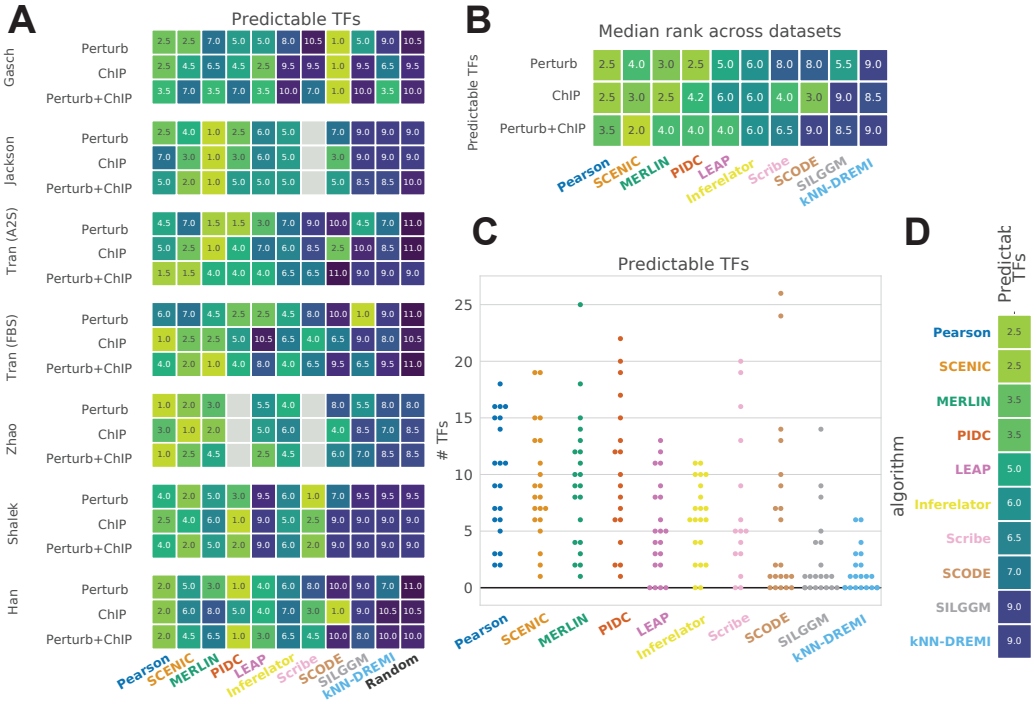



**Fig 4.** Algorithm performance based on predictable TFs. **A.** Rankings of algorithms based on the number of predictable TFs in 21 comparisons (7 datasets x 3 experimentally derived networks). Gray cells represent the missing experiments due to scalability issue of algorithms on large datasets. **B.** Median rank for each method with respect to each type of gold standard: ChIP, Perturb and Perturb+ChIP. **C.** Predictable TFs are shown as counts as the random network generated no predictable TFs. **D.** Overall rank of algorithms across all 21 comparisons using predictable TFs.

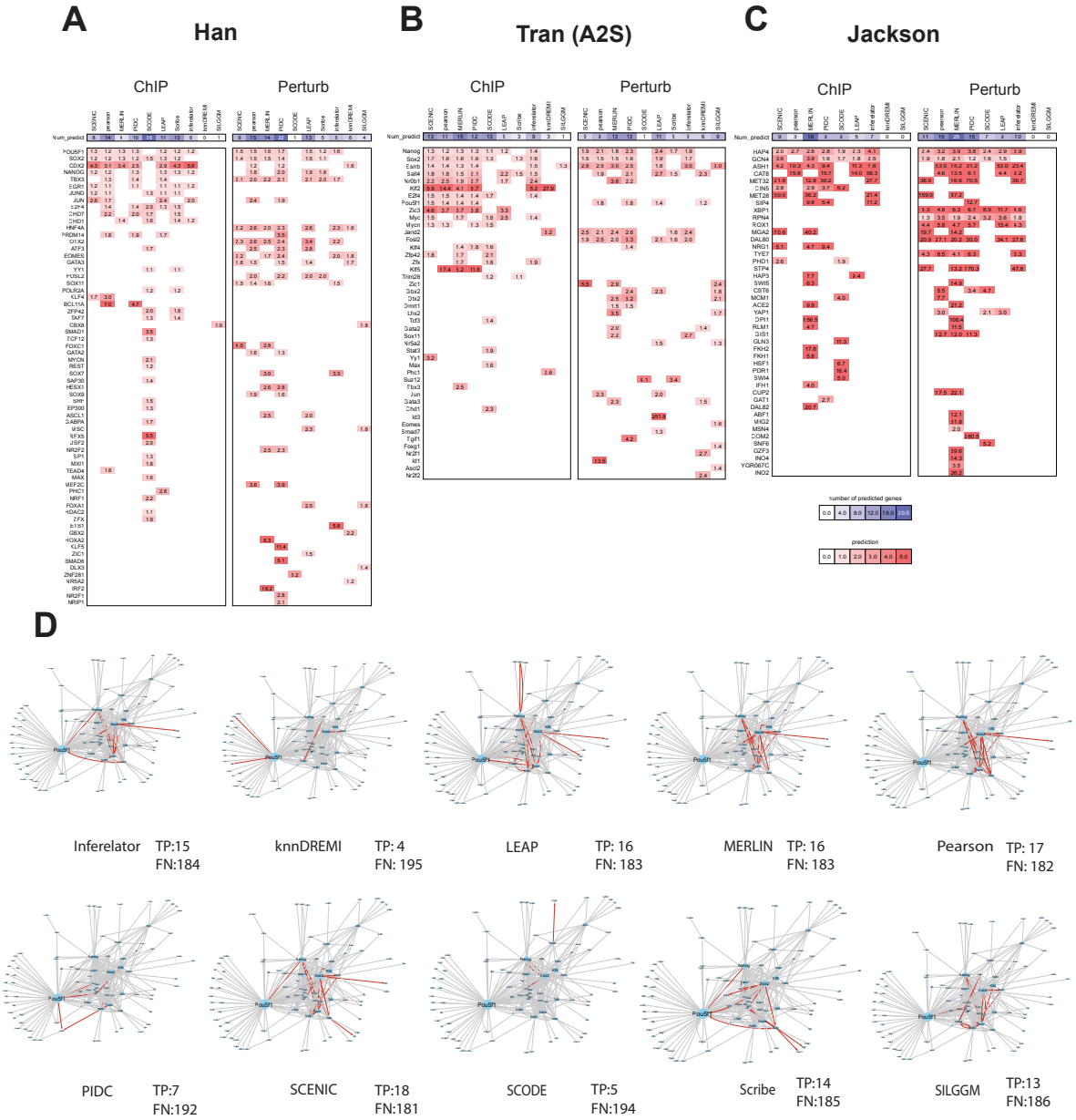
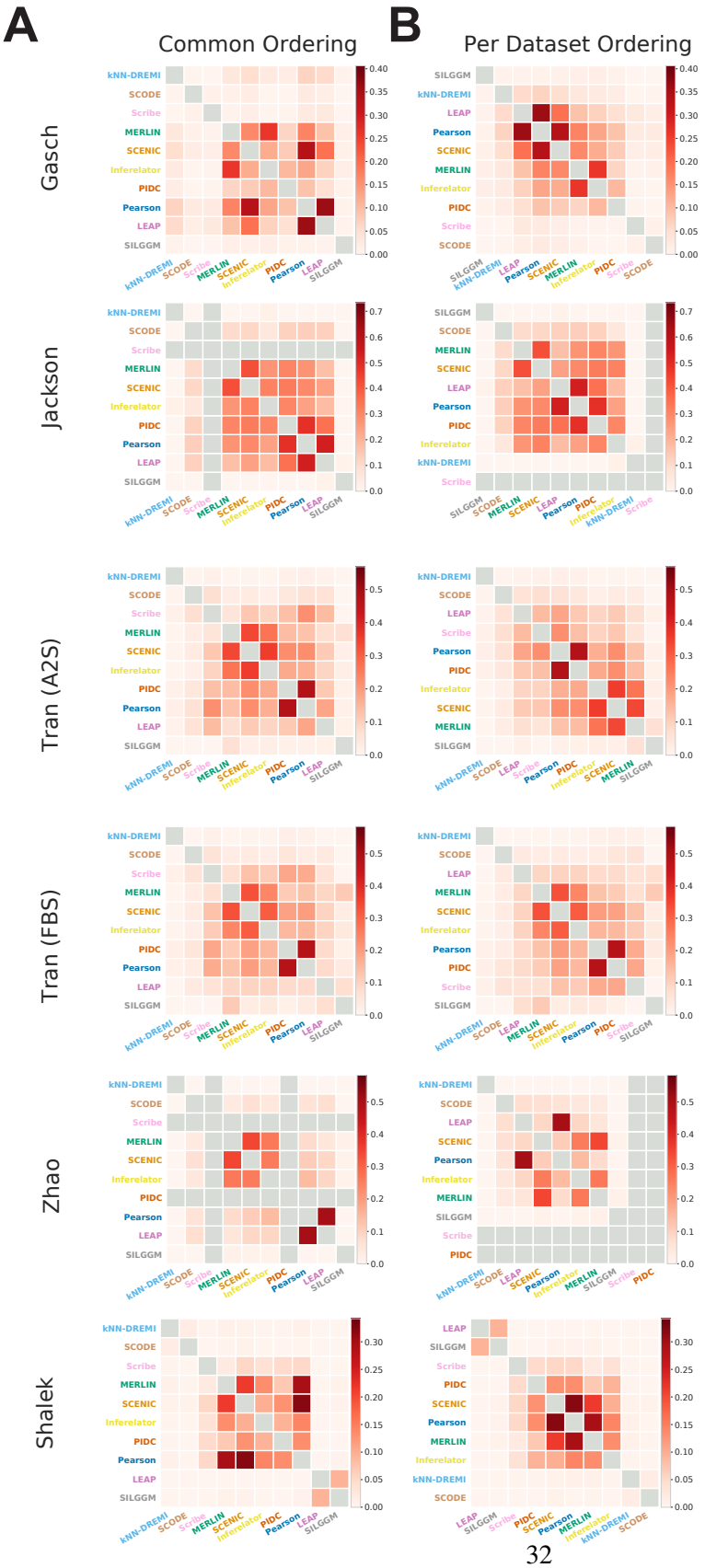



**Fig 5.** Predictable TFs identified in specific datasets. **A** Predictable transcription factors in the human embryonic stem cell state dataset from Han et al. Rows are ordered by the number of algorithms the TF is found predictable. The color is proportional to fold enrichment of true targets in the predicted target set of the TF. Heatmaps show the enrichment of each transcription factor's target set from a ChIP-based (left) or perturbation-based (right) gold standard network in an inferred network.**B**. Columns are ordered with hierarchical clustering. Grey cells indicate transcription factors that did not appear in one of the two experimentally derived networks. Predictable TFs in the mouse cellular reprogramming dataset from Tran et al. **C**. Predictable TFs for the yeast stress dataset from Jackson et al.
**D.** Recovery of a literature curated regulatory network for the embryonic stem cell state. The network has 199 interactions curated from the literature. Shown is the overlap of this network with that inferred by each algorithm on the Tran et al. (A2S) dataset. True positive interactions are in red and false negative interactions are in grey. The number of true positives and false negatives are also listed with each method.


A
Common Ordering
B
Per Dataset Ordering
Gasch
Jackson
Tran (A2S)
Tran (FBS)
Zhao
Shalek
Han
kNN-DREMI
SCODE
Scribe
MERLIN
SCENIC
Inferelator
PIDC
Pearson
LEAP
SILGGM

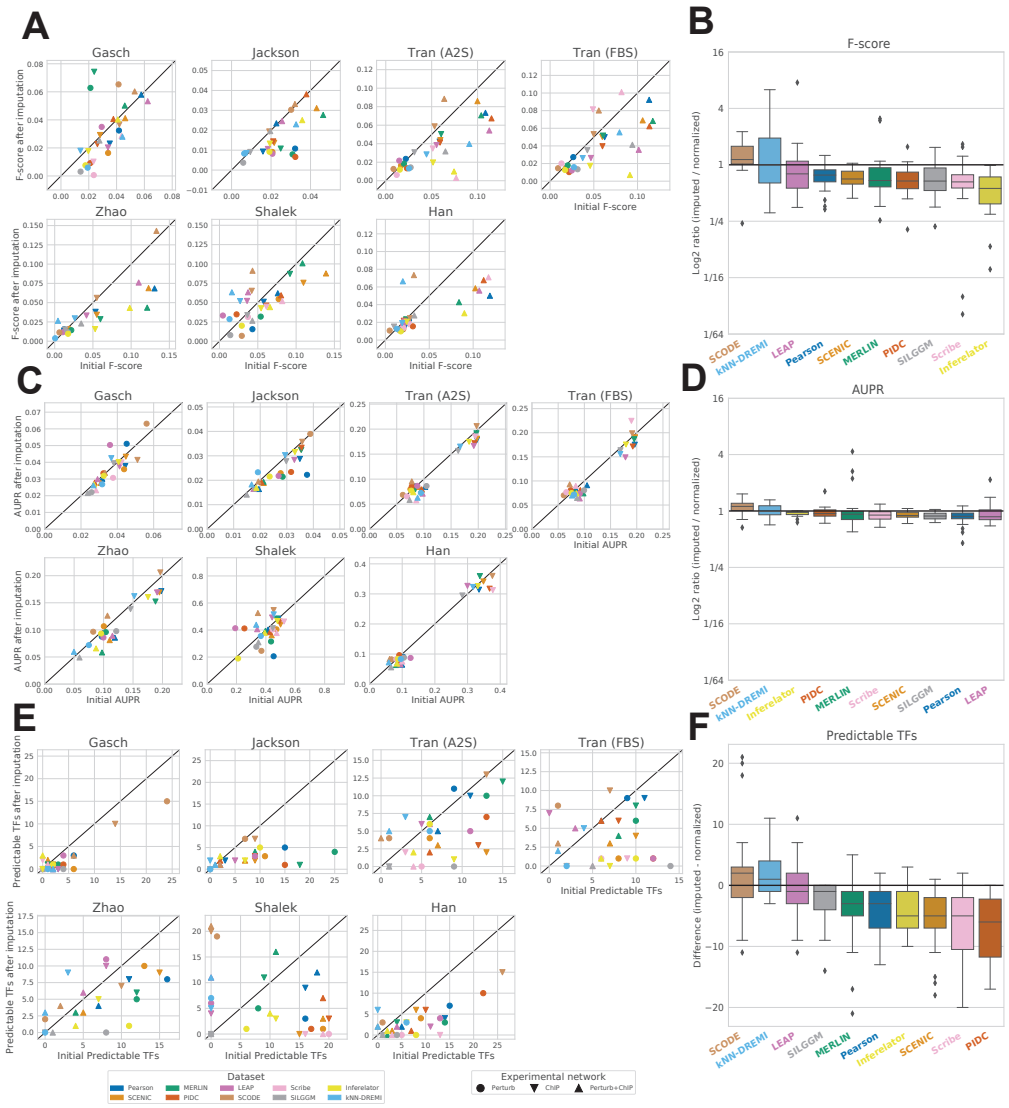



**Fig 7.** Effect of imputation on inferred networks. **A.** Each separate plot corresponds to a dataset, the shape of the marker denotes the type of gold standard and the color depicts the method of network inference. F-score before imputation is on the x-axis and F-score after imputation is on the y-axis. A marker above the diagonal corresponds to improved performance after imputation. **B.** The box plots summarize the relative performance after imputation compared to before imputation over the 21 comparisons: 7 datasets x 3 gold standard networks. Algorithms are ordered based on median performance gain. **C.** Similar to **A**, showing AUPR. **D.** Similar to **B** reporting relative AUPR after and before imputation. **E.** Similar to **A**, showing the number of predictable TFs. **F.** Similar to **B**, reporting the difference in the predictable TFs between the two networks after and before imputation.

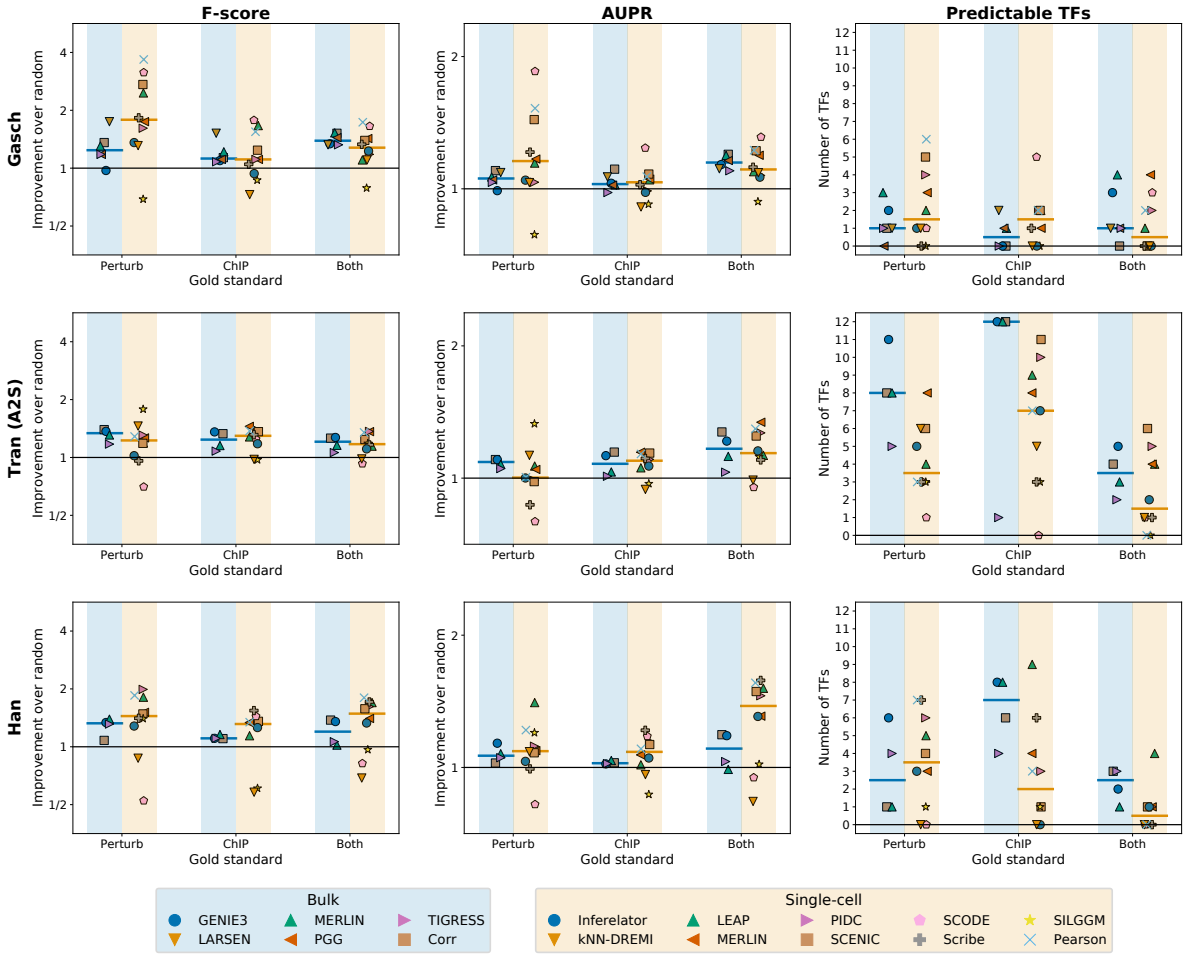



**Fig 8.** Comparison of network inference on bulk versus single cell RNA-seq data.Shown are the relative F-score and AUPR scores compared to random and the number of predictable TFs on comparable bulk and single cell RNA-seq datasets. Methods used for network inference for bulk and single cell datasets are listed at the bottom. Each marker on each plot corresponds to a method for bulk (blue shading) or single cell data (yellow shading).

## Additional Files

### Additional file 1 — Supplementary Tables

This file has all the supplementary tables, S1-S3. **Table S1** has the summary and statistics of datasets used for algorithm comparisons listing reference, GEO ID, species, cell type, number of cells before and after filtering was applied, number of genes before and after filtering was applied, and brief dataset description. **Table S2** has the summary of algorithms used in comparisons and benchmarking. **Table S3** has the description and data source of bulk RNA-seq datasets.

### Additional file 2 — Supplemental Figures

All supplemental figures are documented in this pdf file.

### Additional file 3 — Supplementary Datasets

Expression datasets, regulator sets and gold standard networks for each of the biological conditions considered.