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 no method is a universal winner, and 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 [36], information theoretic methods [4, 6, 27] , random forests [1], ordinary differential equations [21], and Boolean networks [19] (**Fig 1**). Methods also vary in their inclusion of pseudotime [21, 27, 31] or imputed, de-noised signals [6]. Some of these methods are specifically modeling the statistical properties of scRNAseq data [16, 22], while others are adaptations of existing methods for bulk data [1]. In parallel with method development, there have been two benchmarking studies [5, 25] to compare existing network inference algorithms. One study showed that either existing methods perform poorly on both simulated and real data [5], while another showed that methods that perform well on simulated data can be relatively poor on real data [25]. While these studies have been useful, there are a number of open questions that need be addressed to gain a comprehensive understanding of the strengths and weaknesses of existing network inference algorithms from scRNA-seq datasets. Existing benchmarking efforts have also 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 to what extent different benchmarks (e.g. those from ChIP versus regulator perturbation) can affect performance and to what extent do methods agree with each other as well on specific components of the gold standard networks. In addition, imputation of the sparse count matrices obtained from scRNA-seq has been suggested to improve downstream analysis of scRNA-seq datasets including identifying functional relationships among genes [6]. However, it is not clear whether

this benefits network inference methods more generally. Finally, it is currently not clear how well networks inferred from single cell versus bulk datasets for the same system compare against each other.

Here, we compared 11 network inference methods on six datasets from published scRNA-seq experiments on human, mouse, and yeast samples (**Fig 1**). Benchmarking the algorithms for computing requirements 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 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. To our surprise, simple correlation based methods performed as well or better on these metrics. 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 algorithms that perform well compared to others tend to agree with each other, although this depends upon the depth of the dataset. Imputation of scRNA-seq datasets did not offer substantial benefits for most network inference approaches. Finally, we compared the performance of these algorithms on single cell datasets to those on bulk and find several cases where inference on scRNA-seq datasets is associated with improved inference. Overall, our study presents an extended analytical comparison of recent algorithms of network inference from scRNA-seq datasets, identifies common strengths and weaknesses on different types of gold standards that should be beneficial to the network inference community for improved network inference.

Results

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

We started with 11 algorithms specifically designed to infer regulatory networks from single-cell sequencing data, along with a bulk inference method (**Fig 1**, **Supplementary Table S2**). In addition to these published algorithms, we built a network where edges between pairs of genes were weighted by their experiment-wide correlation of expression. We first benchmarked these algorithms using different numbers of genes from 10 to 8000 genes (**Methods**). We estimated the memory and runtime for each network inference algorithm

(**Fig 2**). Of the compared methods, SCHiRM andd 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 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 the performance of 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 (**Fig 1**). These datasets included two yeast stress related datasets, four mouse datasets, one for dendritic cells, and three from cellular reprogramming to embryonic stem cells generated from different media, and one human dataset for human embryonic stem cell differentiation (**Supplementary Table S1**). The datasets had a range of numbers of cells, ranging from 163 cells from Gasch et al [10] to the largest dataset with 36,199 cells from Zhao et al [37]. 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, **Fig 3**), those derived from ChIP-chip or ChIP-seq experiments (ChIP, **Fig 3**), and those derived from the intersection of these two gold standards (Perturb+ChIP, **Fig 3**). We assessed the performance of the algorithms using standard Area under the precision-recall curve (**Fig 3**, **Supplementary Fig 1**) as well as F-score (**Method**), considering each algorithm on each of the datasets. While area under the precision-recall curve (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 selected 5k edges as this resulted in the most stable results across algorithms (**Supplementary Fig 2**).

We first examined the performance of each algorithm for individual datasets from the different organisms and ranked each method based on its F-score (**Fig 3A**). Most methods exhibit some variation in their rankings, however the performance was overall consistent with a few exceptions. For example, SILGGM had a relatively lower performance on most datasets with the exception of Sridharan A2S/FBS Perturb+ChIP

where it ranked highly. SCODE performed well on the yeast datasets, but relatively poorly on the mammalian datasets. LEAP performed poorly on the Shalek dataset, but was generally among the middle ranking methods. The relative performance of algorithms was generally consistent between the two metrics F-score and AUPR. Among different methods, PIDC, SCENIC, MERLIN and Pearson correlation were most stable in their performance across datasets based on F-score and AUPR.

We examined the performance of algorithms with respect to the different types of gold standards. Both F-score and AUPR were generally higher than random for the Perturb gold standard datasets compared to ChIP gold standards (**Fig 3D**). Exception was SCODE which was better for ChIP compared to Knockdown. We next aggregated the algorithms across all datasets for each gold standards based on their median rank (**Fig 3C**) and ranked each method based on its overall median rank across both types of gold standards (**Fig 3C**, E). Based on the median rank of all algorithms across datasets Correlation, SCENIC, MERLIN, and PIDC were also among the top four methods when using F-score or AUPR. This trend was clear when examining the individual fold improvements over random network performance using both F-score and AUPR (**Fig 3E**).

Assessing the performance of scRNA-seq regulatory network inference algorithms based on subnetwork metrics

The AUPR and F-score provide a quantification of the global accuracy of a regulatory network and compares one edge at a time. However, from a biological point of view, it might be beneficial to understand if there are some transcription factors (TF) whose targets are easier to predict compared to others. Furthermore, comparison across methods could identify if specific methods are more sensitive to certain TFs. 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 TF in the predicted set as described by Siahpirani et al [30] (**Methods**). We assigned a p-value to the overlap using an FDR-corrected hypergeometric test and used the number of significantly predictable TFs as another quantification of network performance (**Fig 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 SCODE performing much better on Gasch. When aggregating over the dataset using the median rank, we found SCENIC and Pearson as the top methods followed by MERLIN and PIDC (**Fig 4B, D**). Taken together, considering all three metrics of AUPR, F-score and predictable TFs, top performing methods were SCENIC, MERLIN and Correlation (**Fig 4B, D**).

We next examined for each dataset which specific TFs were predicted by the different methods (**Fig 5**). This information can help determine if there are parts of the network 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 state, we found several ESC specific regulators such POU5F1, SOX2 and NANOG as a predictable TF in multiple methods using the ChIP gold-standard (**Fig 5**). These are key ESC TFs that are responsible for establishing this state. We also found some lineage-specific TFs like CDX2 and TBX3 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 [34] and hematopoeitic lineages [18]. For the mouse cellular reprogramming dataset, we find a similar behavior with many of the methods able to identify several developmental regulators such as Pou5f1, Esrrb and Sox2 in both ChIP and Perturb gold standards (**Fig 5B**). SCODE had a number of predictable TFs in the Han dataset, however, many of them are general regulators such as SP1, YY1, POL2RA, ATF2. Importantly, when compared to the dendritic cell dataset (Shalek), we found a consistent over-representation of a different set of regulators across methods. In particular this included regulators such as Rel, Nfkb, Stat1 and Stat3 that are associated with the immune response and were identified by several methods including SCENIC, MERLIN, PIDC, Scribe, Inferelator and Pearson. We found a similar behavior for the yeast datasets (**Fig 5C**, **Supplemental Fig 3**), where the methods uniformly were able to recover key regulators associated with stress response such as HAP4 (oxidative stress) and GCN4 (amino acid starvation, [9]). The overall fold enrichment for predictable TFs was higher for the yeast datasets compared to the mammalian datasets. Between the two 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 AUPR of the methods is modest, these methods are able to consistently recover relevant regulators for a particular system (**Supplemental Fig 6**).

Finally to gain more insight into each method's ability to recover mammalian regulatory networks, we

curated small gold standard network from regulatory interactions reported in literature [2, 8, 17, 20, 33, 35, 38]. The regulatory network had 35 regulators and 90 target genes (some of which included regulators as well) connected by 267 edges. 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. 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.

Defining common and method-specific network components

We next asked the extent to which the algorithms agree in their predictions. To this end we measured the Jaccard similarity and F-score between the top 500, 5,000 and 50,000 edges of each pair of inferred networks (**Fig 6**, **Supplementary Fig 7**, **Supplementary Fig 8**). We found that patterns of similarity were generally similar regardless of how many top edges were included and similarity measure (**Supplementary Fig 9**). The magnitude of the similarity generally increased as more edges were considered. We focused on the Jaccard scores obtained on the top 5k edges for consistency with our other results. We grouped the methods using two orderings. First, we clustered the methods based on the median similarity of networks inferred for each dataset (**Fig 6A**). This allowed us to compare all the methods against each other across datasets. Some datasets resulted in more similar networks compared to others. For example, on the Han hESC dataset, the maximum Jaccard coefficient was highest compared to Shalek and Gasch which had among the lower similarities. These two datasets were also the smallest of the two datasets, which could explain the lower similarity. 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). Scribe's similarity to other

methods depended upon the dataset, however, in most it most similar to PIDC, which is consistent with both methods relying on information theoretic metrics. LEAP was most similar to Correlation but generally identified different networks. Methods that tended to have the highest similarity among them were Pearson, SCENIC, Inferelator and PIDC.

We next clustered the methods based on the network Jaccard similarity on each dataset (**Fig 6B**). The latter revealed more dataset-specific groupings. In particular, we found LEAP and Pearson were the most similar methods in Gasch and Zhao datasets, while PIDC is most similar to Pearson in the four of the seven datasets (Jackson, Tran (A2S), Tran (FBS), and Han). We also found LEAP and SILGGM to form a cluster on Shalek compared to another group comprising SCRIBE, PIDC, SCENIC, Pearson, MERLIN and Inferelator. In summary these comparisons revealed that top performing algorithms tend to learn similar networks, however, methods such PIDC and Pearson often learned similar networks. Furthermore, agreement and differences between methods could be determined by the nature of the dataset, including sample size and the experimental conditions.

Examining the impact of imputation on scRNA-seq 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, which are gene truly not expressed in the cell. To address this problem, several methods have been developed to impute the value of missing expression counts [14]. We applied the MAGIC algorithm [6] to each dataset, which was one of the top imputation methods from a recent benchmarking study [14]. 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 sparse data (**Fig 7**). Based on F-score (**Fig 7A**), most methods did not benefit from imputation across datasets. An exception to this was the Shalek and Gasch datasets, which as we noted before are among the smaller datasets. We aggregated the F-score values across different gold standards and datasets and compared the ratio compared to before imputation (**Fig 7B**) and found that SCODE and kNN-DREMI appeared to improve when using imputed data. However, we found that imputation generally did not benefit the network

inference procedure. We repeated these comparisons based on AUPR (**Fig 7C, D**) and also predictable TFs. AUPR generally did not change much between the with and without imputation, however, we noted a similar boost in performance for the Shalek dataset. Finally, based on predictable TFs, we found that most algorithms did not benefit from smoothing across datasets, the Tran datasets offered moderate benefit to the SCODE and knnDREMI methods. Taken together, our experiments across different experimental 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 from bulk expression datasets, which required large sample sizes. Such datasets are available in yeast 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 human ESC (Siahpirani in preparation). 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 (Fig 8). For the single cell comparison, we used the inferred networks from the Gasch dataset for yeast, Tran (A2S) for mESC and Han dataset for human ESCs.

Based on F-score, both scRNA-seq and bulk expression-based networks were comparable and 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 using the Tran A2S datasets compared to the mouse ESC datasets and the human ESC dataset for ChIP and ChIP+Perturb.

Overall this suggests that individual single cell scRNA-seq datasets capture meaningful variation that offers comparable or better performance compared to bulk expression datasets collected from a large number of experiments. However, using different metrics also revealed that bulk expression data can still be more useful when considering specific components of the gene regulatory network.

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 efforts expand on previous efforts by considering more datasets, additional classes of methods and multiple local and global network assessment metrics that can provide more biological insight into the predictions made by the different methods and also reveal potential directions of methodological and resource generation advancements for improved network recovery.

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 be of importance. 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, we group them into three main categories: top (PIDC, MERLIN, SCENIC, PEARSON), middle (Inferelator, SCODE, LEAP, Scribe) and bottom (knnDREMI, SILGGM) set of 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 address highly 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 and perform additional experiments 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. Generated 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 [11, 30] as well as estimating hidden regulatory activity [23] 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 [3, 24]. 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 [7] and Perturb-ATAC [29] 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 Supplementary 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 [12], 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 **Supplementary Table S2** for more details about the algorithm implementation).

BTR BTR (is based on a Boolean model representation of a gene regulatory network. 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 [22] 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. 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 first computing joint density using a k-nearest neighbor based density estimator. Then on a coarsened grid the discrete mutual information is calculated. The conditional density is computed to estimate the gene-gene functional relationship. Finally, entropy and mutual information is calculated based on the coarsened grid.

LEAP. LEAP [31] 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 some 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 ρ_{ij}.

PIDC. PIDC [4] 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 [15] 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. 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 [27] 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 [36] 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 [28] and Inferelator [11]. 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 networks as controls that we hoped would be out-performed by the other algorithms. 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 Unif$(0, 1)$.

Comparison to network inference on bulk expression data

Methods used to infer networks on bulk data include GENIE3 [15], TIGRESS [13], MERLIN [28]. To run GENIE3 we used $K =$ `sqrt` and `nb_trees` $= 1000$, to run TIGRESS we used `R`=$1,000$, α=0.3, and `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 ran 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 [9] for bulk to match our single cell RNA-seq dataset (Gasch) also generated by Gasch and colleagues [10]. For human and

mouse, we curated publicly available datasets from the gene expression omnibus (**Supplemental 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 [12] and the performance of the bulk mouse dataset to that from the A2S (Tran) dataset [32].

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 the HTCondor scheduler. 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. Some algorithms perform a similar downsampling internally (SCENIC, Inferelator) and so 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 (transciption 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 (details on these networks below). The final networks were used in the evaluations of network accuracy and similarity (detailed below) and are available in the Supplementary Material.

Estimation of pseudo time. Some algorithms (LEAP, SCODE, Scribe) incorporate knowledge of cellular trajectory and pseudotime, which was learned using Monocle 2 [26]. 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 (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.

Description of gold standards

We curated multiple experimentally derived networks of regulatory interactions from published databases and the literature (**Supplementary Dataset S1**) 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).

Evaluation metrics

To evaluate the accuracy of the inferred networks, we performed a series of pairwse comparisons using three accuracy measures - AUPR, F-score, and predictable TFs - between each network and a cell type-specific 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.

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.

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 inferd 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. Most of our experiments are with the top 5000 edges.

Predictable TFs. We used predictable TFs to obtain a granular measure of network accuracy at the individual TF/target set level. Specifically, for each transcription factor in the inferred network, we use the hypergeometric test to test 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.

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] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, and et al. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, November 2017.

[2] Yosef Buganim, Dina A. Faddah, Albert W. Cheng, Elena Itskovich, Styliani Markoulaki, Kibibi Ganz, Sandy L. Klemm, Alexander van Oudenaarden, and Rudolf Jaenisch. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, 150(6):1209–1222, September 2012.

[3] Dayanne M. Castro, Nicholas R. de Veaux, Emily R. Miraldi, and Richard Bonneau. Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLOS Computational Biology*, 15(1):e1006591, January 2019. Publisher: Public Library of Science.

[4] Thalia E. Chan, Michael P. H. Stumpf, and Ann C. Babtie. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Systems*, 5(3):251–267.e3, September 2017.

[5] Shuonan Chen and Jessica C. Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC bioinformatics*, 19(1):232, June 2018.

[6] David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer, Diwakar Pattabiraman, and et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3):716–729.e27, July 2018.

[7] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17, December 2016.

[8] S. J. Dunn, G. Martello, B. Yordanov, S. Emmott, and A. G. Smith. Defining an essential transcription factor program for naïve pluripotency. *Science*, 344(6188):1156–1160, June 2014.

[9] Audrey P. Gasch, Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, and Patrick O. Brown. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Mol. Biol. Cell*, 11(12):4241–4257, December 2000.

[10] Audrey P. Gasch, Feiqiao Brian Yu, James Hose, Leah E. Escalante, Mike Place, Rhonda Bacher, Jad Kanbar, Doina Ciobanu, Laura Sandor, Igor V. Grigoriev, and et al. Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. *PLOS Biology*, 15(12):e2004050, December 2017.

[11] Alex Greenfield, Christoph Hafemeister, and Richard Bonneau. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics (Oxford, England)*, 29(8):1060–1067, April 2013.

[12] Xiaoping Han, Haide Chen, Daosheng Huang, Huidong Chen, Lijiang Fei, Chen Cheng, He Huang, Guo-Cheng Yuan, and Guoji Guo. Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biology*, 19(1):47, April 2018.

[13] Anne-Claire C. Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe P. Vert. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC systems biology*, 6(1):145+, 2012.

[14] Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C. Hicks. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biology*, 21(1):218, December 2020.

[15] Vân A. Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE*, 5(9):e12776+, September 2010.

[16] Jukka Intosalmi, Henrik Mannerström, Saara Hiltunen, and Harri Lähdesmäki. SCHiRM: Single Cell Hierarchical Regression Model to detect dependencies in read count data. *bioRxiv*, page 335695, May 2018.

[17] Jonghwan Kim, Jianlin Chu, Xiaohua Shen, Jianlong Wang, and Stuart H. Orkin. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 132(6):1049–1061, March 2008.

[18] Chia-Jui Ku, Tomonori Hosoya, Ivan Maillard, and James Douglas Engel. GATA-3 regulates hematopoietic stem cell maintenance and cell-cycle entry. *Blood*, 119(10):2242–2251, March 2012.

[19] Chee Yee Lim, Huange Wang, Steven Woodhouse, Nir Piterman, Lorenz Wernisch, Jasmin Fisher, and Berthold Göttgens. BTR: training asynchronous Boolean models using single-cell expression data. *BMC Bioinformatics*, 17(1):355, September 2016.

[20] Mohan Malleshaiah, Megha Padi, Pau Rué, John Quackenbush, Alfonso Martinez-Arias, and Jeremy Gunawardena. Nac1 coordinates a sub-network of pluripotency factors to regulate embryonic stem cell differentiation. *Cell Reports*, 14(5):1181–1194, January 2016.

[21] Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru S. H. Ko, Shigeru B. H. Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, 33(15):2314–2321, August 2017.

[22] Andrew McDavid, Raphael Gottardo, Noah Simon, and Mathias Drton. Graphical Models for Zero-Inflated Single Cell Gene Expression. *The Annals of Applied Statistics*, 13(2):848–873, June 2019.

[23] Emily R. Miraldi, Maria Pokrovskii, Aaron Watters, Dayanne M. Castro, Nicholas De Veaux, Jason A. Hall, June-Yong Lee, Maria Ciofani, Aviv Madar, Nick Carriero, Dan R. Littman, and Richard Bonneau. Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. *Genome Research*, January 2019. 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.

[24] Emma Pierson, the GTEx Consortium, Daphne Koller, Alexis Battle, and Sara Mostafavi. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLOS Computational Biology*, 11(5):e1004220, May 2015. Publisher: Public Library of Science.

[25] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, February 2020.

[26] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–982, October 2017. Number: 10 Publisher: Nature Publishing Group.

[27] Xiaojie Qiu, Arman Rahimzamani, Li Wang, Qi Mao, Timothy Durham, José L. McFaline-Figueroa, Lauren Saunders, Cole Trapnell, and Sreeram Kannan. Towards inferring causal gene regulatory networks from single cell expression Measurements. *bioRxiv*, page 426981, September 2018.

[28] Sushmita Roy, Stephen Lagree, Zhonggang Hou, James A. Thomson, Ron Stewart, and Audrey P. Gasch. Integrated Module and Gene-Specific Regulatory Inference Implicates Upstream Signaling Networks. *PLOS Computational Biology*, 9(10):e1003252, October 2013.

[29] Adam J. Rubin, Kevin R. Parker, Ansuman T. Satpathy, Yanyan Qi, Beijing Wu, Alvin J. Ong, Maxwell R. Mumbach, Andrew L. Ji, Daniel S. Kim, Seung Woo Cho, Brian J. Zarnegar, William J. Greenleaf, Howard Y. Chang, and Paul A. Khavari. Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell*, 176(1):361–376.e17, January 2019.

[30] Alireza F. Siahpirani and Sushmita Roy. A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Research*, 45(4):e21–e21, February 2017.

[31] Alicia T. Specht and Jun Li. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics*, 33(5):764–766, March 2017.

[32] Khoa A. Tran, Stefan J. Pietrzak, Nur Zafirah Zaidan, Alireza Fotuhi Siahpirani, Sunnie Grace McCalla, Amber S. Zhou, Gopal Iyer, Sushmita Roy, and Rupa Sridharan. Defining Reprogramming Checkpoints from Single-Cell Analyses of Induced Pluripotency. *Cell Reports*, 27(6):1726–1741.e5, May 2019.

[33] Huilei Xu, Yen-Sin Ang, Ana Sevilla, Ihor R. Lemischka, and Avi Ma'ayan. Construction and validation of a regulatory network for pluripotency and Self-Renewal of mouse embryonic stem cells. *PLoS Comput Biol*, 10(8):e1003777+, August 2014.

[34] Haruka Yamamoto, Tetsuo Kon, Yoshihiro Omori, and Takahisa Furukawa. Functional and Evolutionary Diversification of Otx2 and Crx in Vertebrate Retinal Photoreceptor and Bipolar Cell Development. *Cell Reports*, 30(3):658–671.e5, January 2020.

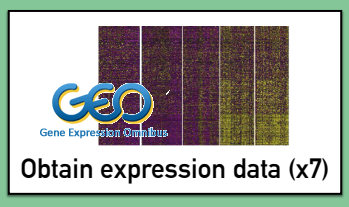
[35] Richard A. Young. Control of the embryonic stem cell state. *Cell*, 144(6):940 – 954, 2011.

[36] Rong Zhang, Zhao Ren, and Wei Chen. SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. *PLOS Computational Biology*, 14(8):e1006369, August 2018.

[37] Ting Zhao, Yao Fu, Jialiang Zhu, Yifang Liu, Qian Zhang, Zexuan Yi, Shi Chen, Zhonggang Jiao, Xiaochan Xu, Junquan Xu, and etal. Single-Cell RNA-Seq Reveals Dynamic Early Embryonic-like Programs during Chemical Reprogramming. *Cell Stem Cell*, 23(1):31–45.e7, July 2018.

[38] Qing Zhou, Hiram Chipperfield, Douglas A. Melton, and Wing Hung Wong. A gene regulatory network in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences*, 104(42):16438–16443, 2007.

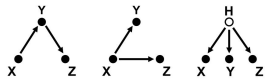
Data collection



Obtain expression data (x7)

Dataset	Cell type	# cells	# genes
Gasch *et al.*, *PLoS Biol* 2017	Yeast	163	3,847
Jackson *et al.*, *Elife* 2020		17,396	5,736
Shalek *et al.*, *Nature* 2014	mDC	1,211	9,411
Tran *et al.*, *Cell Reports* 2019	mESC	2,369	6,618
Tran *et al.*, *Cell Reports* 2019		3,324	6,621
Zhao *et al.*, *Cell Stem Cell* 2018		36,199	8,442
Han *et al.*, *Genome Biol* 2018	hESC	5,520	7,465

Network inference

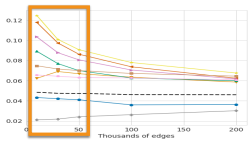


Infer networks (x13)

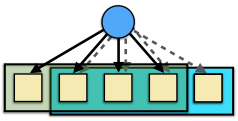
Algorithm	Methodology
HurdleNormal	Graphical models and dependency networks
Inferelator	
MERLIN	
SCENIC	
SCHiRM	
SILGGM	
kNN-DREMI	Information theoretic
PIDC	
Scribe	
BTR	Boolean networks
SCODE	ODEs
LEAP	Correlation
Pearson	

Network evaluation

Experimental interactions



Accuracy of top edges



Predictable TFs

Fig 1. Overview of network inference benchmarking study.

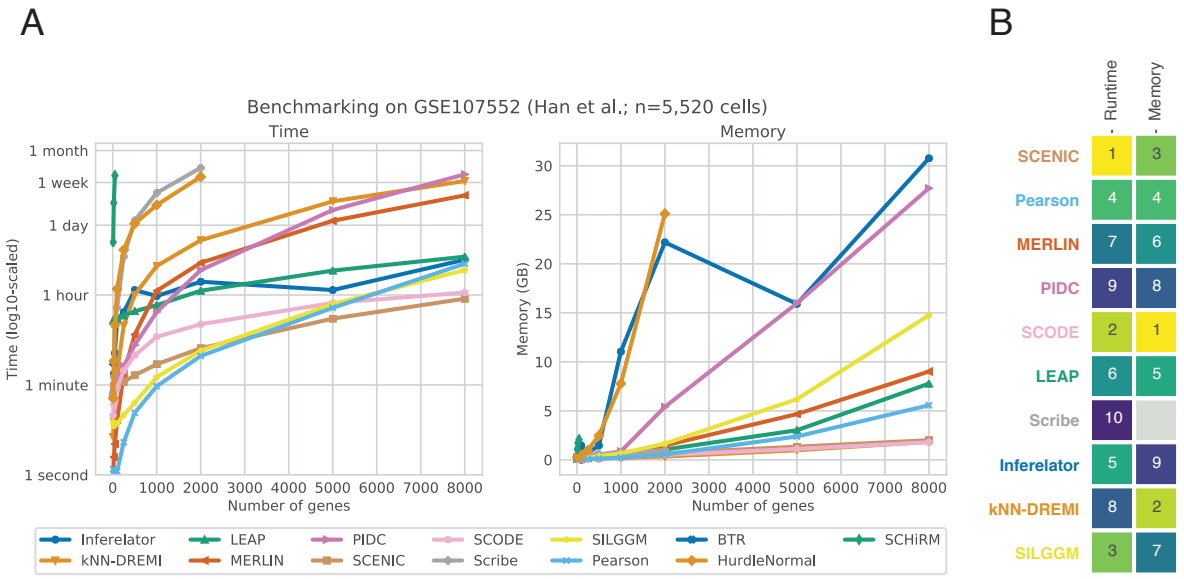

A
B
Benchmarking on GSE107552 (Han et al.; n=5,520 cells)
Time
Memory
Time (log10-scaled)
1 month
1 week
1 day
1 hour
1 minute
1 second
Memory (GB)
Number of genes
Inferelator
kNN-DREMI
LEAP
MERLIN
PIDC
SCENIC
SCODE
Scribe
SILGGM
Pearson
BTR
HurdleNormal
SCHiRM
- Runtime
- Memory
SCENIC 1 3
Pearson 4 4
MERLIN 7 6
PIDC 9 8
SCODE 2 1
LEAP 6 5
Scribe 10
Inferelator 5 9
kNN-DREMI 8 2
SILGGM 3 7

Fig 2. Benchmarking of 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.

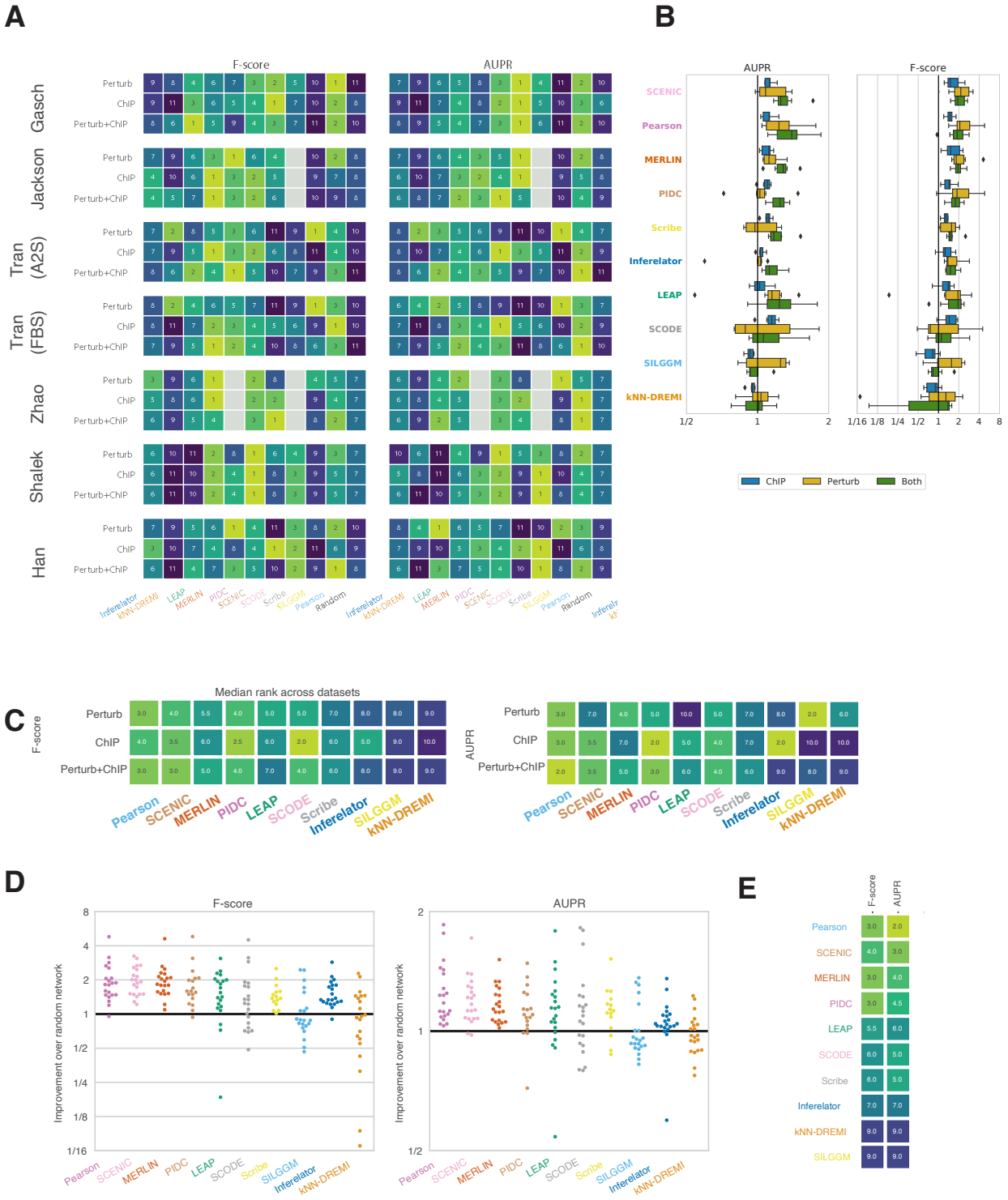

A
F-score
AUPR
Gasch
Jackson
Tran (A2S)
Tran (FBS)
Zhao
Shalek
Han
Perturb
ChIP
Perturb+ChIP
Inferelator
kNN-DREMI
LEAP
MERLIN
PIDC
SCENIC
SCODE
Scribe
SILGGM
Pearson
Random
B
AUPR
F-score
SCENIC
Pearson
MERLIN
PIDC
Scribe
Inferelator
LEAP
SCODE
SILGGM
kNN-DREMI
1/2
1
2
1/16
1/8
1/4
4
8
ChIP
Perturb
Both
C
Median rank across datasets
F-score
AUPR
Perturb
ChIP
Perturb+ChIP
Pearson
SCENIC
MERLIN
PIDC
LEAP
SCODE
Scribe
Inferelator
SILGGM
kNN-DREMI
D
F-score
AUPR
Improvement over random network
8
4
2
1
1/2
1/4
1/8
1/16
E
F-score
AUPR
Pearson 3.0 2.0
SCENIC 4.0 3.0
MERLIN 3.0 4.0
PIDC 3.0 4.5
LEAP 5.5 6.0
SCODE 6.0 5.0
Scribe 6.0 5.0
Inferelator 7.0 7.0
kNN-DREMI 9.0 9.0
SILGGM 9.0 9.0

Fig 3. Algorithm performance across different global metrics of structure recovery. **A.** Overall algorithm ranks in performance in 21 comparisons (7 datasets x 3 experimentally derived networks), measured using AUPR, F-score, and the number of predictable TFs (see methods). Gray cells represent the missing experiments due to scalability issue of algorithms on large datasets. Scribe is not scalable enough for Jackson and Zhao datasets, and PIDC is not scalable enough for Zhao dataset. **B.** Comparison of algorithm performance with respect to each class of experimentally derived network - ChIP, peturbation, and their intersection. **C.** Ranked algorithm performance with respect to each class of experimentally derived network (median over 7 datasets). **E.** Overall ranks of algorithms across all 21 comparisons using F-score and AUPR. **D.** AUPR and F-score are shown as log-fold improvement over the performance of a randomly weighted network in the same comparison (raw values in Supplementary Data). **E.** Overall ranks of algorithms across all 21 comparisons using F-score

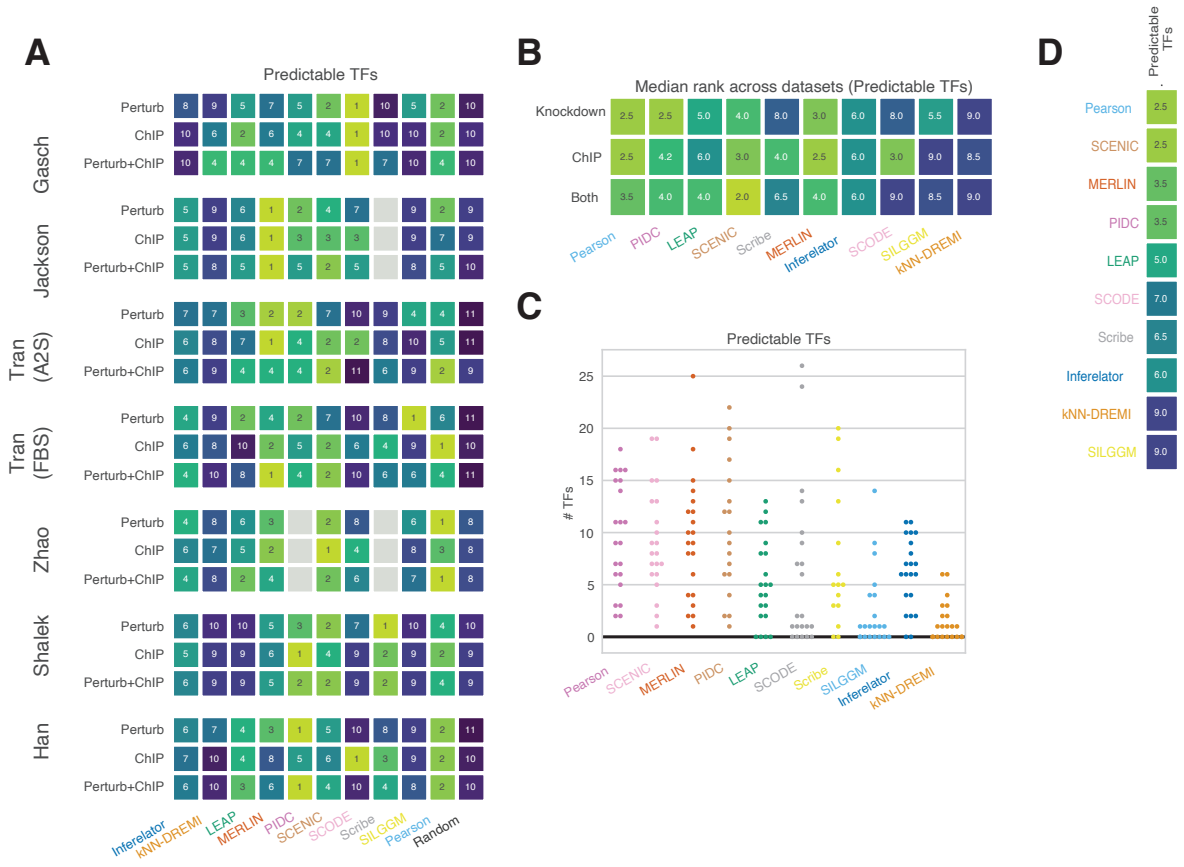
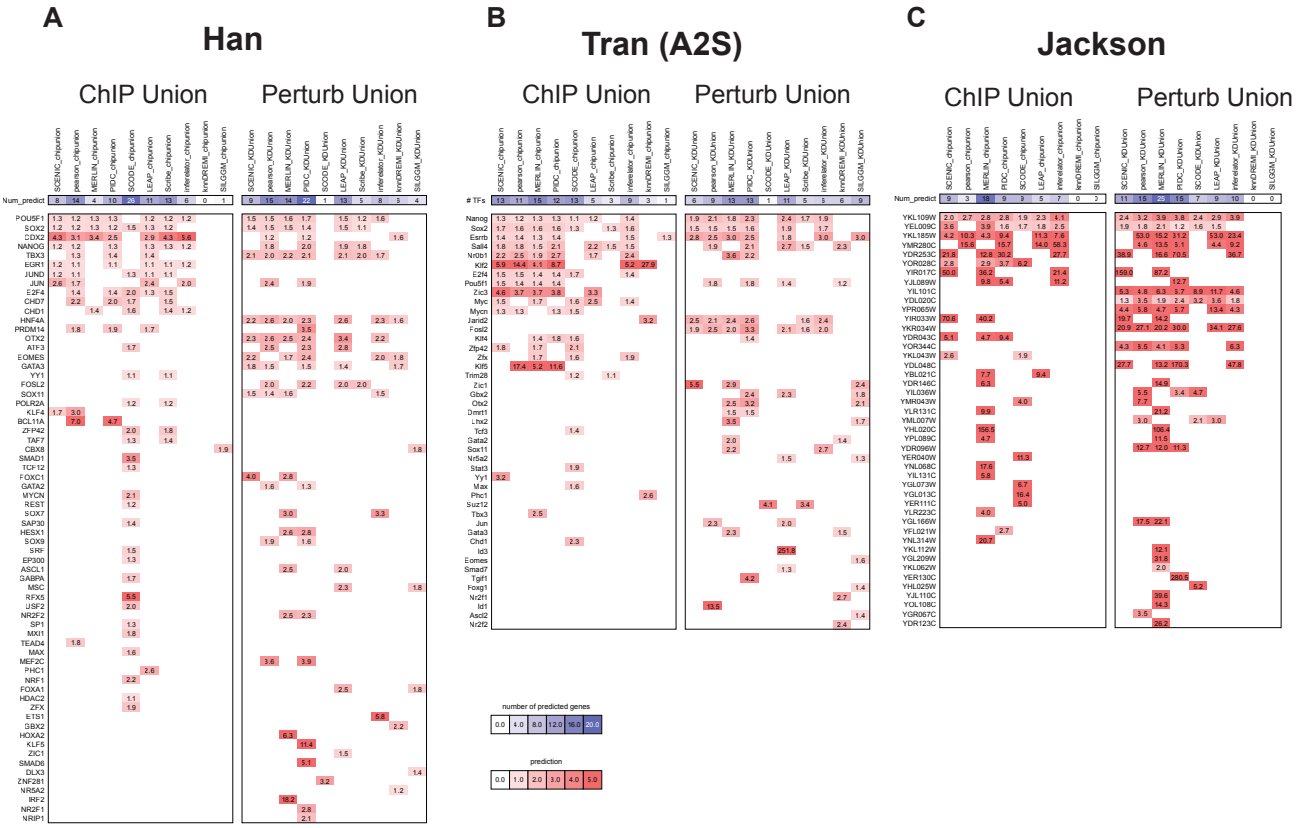

A
Predictable TFs
Gasch
Perturb 8 9 5 7 5 2 1 10 5 2 10
ChIP 10 6 2 6 4 4 1 10 10 2 10
Perturb+ChIP 10 4 4 4 7 7 1 7 10 4 10
Jackson
Perturb 5 9 6 1 2 4 7 9 2 9
ChIP 5 9 6 1 3 3 3 9 7 9
Perturb+ChIP 5 8 5 1 5 2 5 8 5 10
Tran (A2S)
Perturb 7 7 3 2 2 7 10 9 4 4 11
ChIP 6 8 7 1 4 2 2 8 10 5 11
Perturb+ChIP 6 9 4 4 4 2 11 6 9 2 9
Tran (FBS)
Perturb 4 9 2 4 2 7 10 8 1 6 11
ChIP 6 8 10 2 5 2 6 4 9 1 10
Perturb+ChIP 4 10 8 1 4 2 10 6 6 4 11
Zhao
Perturb 4 8 6 3 2 8 6 1 8
ChIP 6 7 5 2 1 4 8 3 8
Perturb+ChIP 4 8 2 4 2 6 7 1 8
Shalek
Perturb 6 10 10 5 3 2 7 1 10 4 10
ChIP 5 9 9 6 1 4 9 2 9 2 9
Perturb+ChIP 6 9 9 5 2 2 9 2 9 4 9
Han
Perturb 6 7 4 3 1 5 10 8 9 2 11
ChIP 7 10 4 8 5 6 1 3 9 2 10
Perturb+ChIP 6 10 3 6 1 4 10 4 8 2 10
Inferelator kNN-DREMI LEAP MERLIN PIDC SCENIC SCODE Scribe SILGGM Pearson Random
B
Median rank across datasets (Predictable TFs)
Knockdown 2.5 2.5 5.0 4.0 8.0 3.0 6.0 8.0 5.5 9.0
ChIP 2.5 4.2 6.0 3.0 4.0 2.5 6.0 3.0 9.0 8.5
Both 3.5 4.0 4.0 2.0 6.5 4.0 6.0 9.0 8.5 9.0
Pearson PIDC LEAP SCENIC Scribe MERLIN Inferelator SCODE SILGGM kNN-DREMI
C
Predictable TFs
TFs
0 5 10 15 20 25
Pearson SCENIC MERLIN PIDC LEAP SCODE Scribe SILGGM Inferelator kNN-DREMI
D
Predictable TFs
Pearson 2.5
SCENIC 2.5
MERLIN 3.5
PIDC 3.5
LEAP 5.0
SCODE 7.0
Scribe 6.5
Inferelator 6.0
kNN-DREMI 9.0
SILGGM 9.0

Fig 4. Performance based on predictable TFs for each method across different gold standards and seven datasets. **A.** Rankings of algorithms based on the number of predictable TFs in 21 comparisons (7 datasets x 3 experimentally derived networks), measuring the number of predictable TFs (see methods). Gray cells represent the missing experiments due to scalability issue of algorithms on large datasets. Scribe is not scalable enough for Jackson and Zhao datasets, and PIDC is not scalable enough for Zhao dataset. **B.** Median rank for each method with respect to each type of experimentally derived network - ChIP, perturbation, and their intersection. **C.** Predictable TFs are shown as counts, as the random network generally has no predictable TFs. Algorithms are ordered by their median performance with respect to F-score in **Fig 3**. **D.** Overall ranks of algorithms across all 21 comparisons using predictable TFs.


A
Han
ChIP Union
Perturb Union
B
Tran (A2S)
ChIP Union
Perturb Union
C
Jackson
ChIP Union
Perturb Union
number of predicted genes
prediction

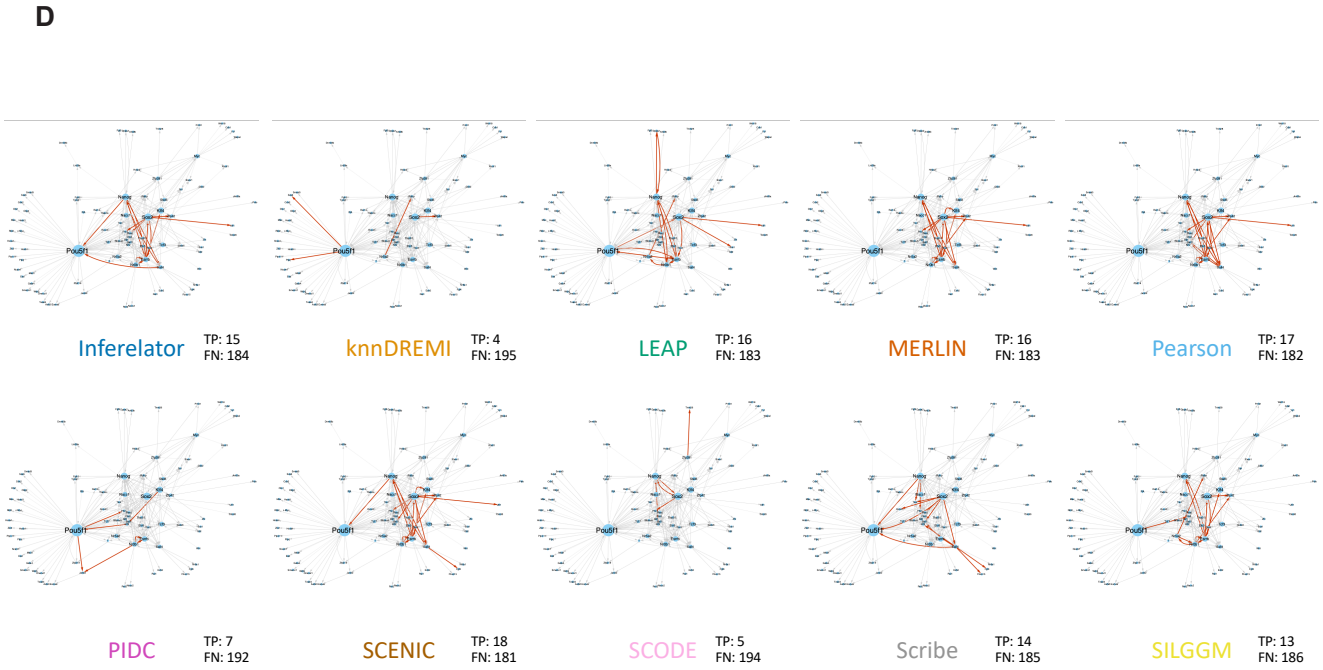

D
Inferelator TP: 15 FN: 184
knnDREMI TP: 4 FN: 195
LEAP TP: 16 FN: 183
MERLIN TP: 16 FN: 183
Pearson TP: 17 FN: 182
PIDC TP: 7 FN: 192
SCENIC TP: 18 FN: 181
SCODE TP: 5 FN: 194
Scribe TP: 14 FN: 185
SILGGM TP: 13 FN: 186

Fig 5. Method performance with respect to individual transcription factors. **A-C.** Enrichment of predictable transcription factors in Han et al. (**A**), Tran et al. (**B**), and Jackson et al. (**C**). Heatmaps show the enrichment of each transcription factor's target set from a perturbation-based (left) or ChIP-based (right) experimentally derived network in an inferred network (Methods). Columns are ordered with hierarchical clustering. Grey cells indicate transcription factors that did not appear in one of the two experimentally derived networks. **D.** Highlighting recovery of a literature sub-network. Here we depict a network of 199 interactions curated from the literature. Using the network inferred by each algorithm on the Tran et al. (A2S) dataset, we highlight successfully inferred true positive interactions in red, and false negative interactions the algorithm failed to infer in grey.

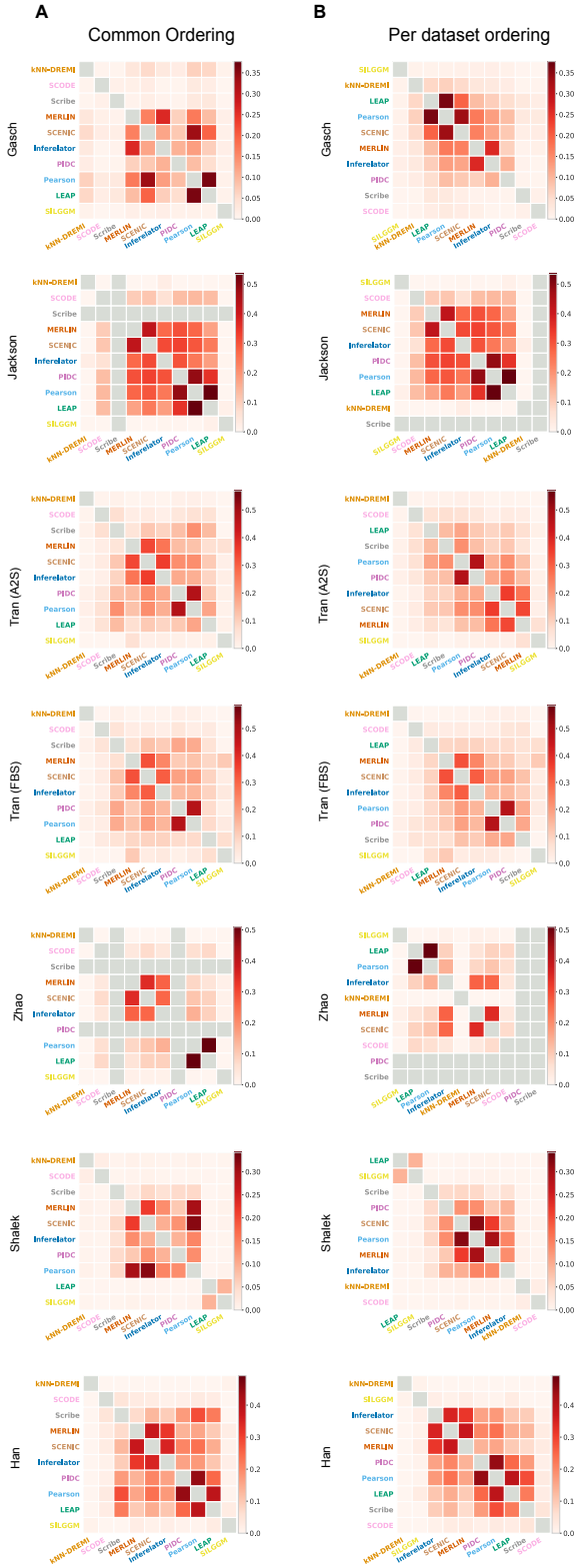

A
Common Ordering
B
Per dataset ordering
Gasch
Jackson
Tran (A2S)
Tran (FBS)
Zhao
Shalek
Han
kNN-DREMI
SCODE
Scribe
MERLIN
SCENIC
Inferelator
PIDC
Pearson
LEAP
SILGGM

Fig 6. Pairwise network similarity between inferred networks from each method on each of the datasets. Shown here are the Jaccard similarity between the top 5,000 edges in each pair of networks inferred on each dataset. **A.** The columns and rows of each matrix are ordered with respect to a hierarchical clustering of the median similarity over datasets. **B.** The columns and rows are ordered with respect to each dataset.

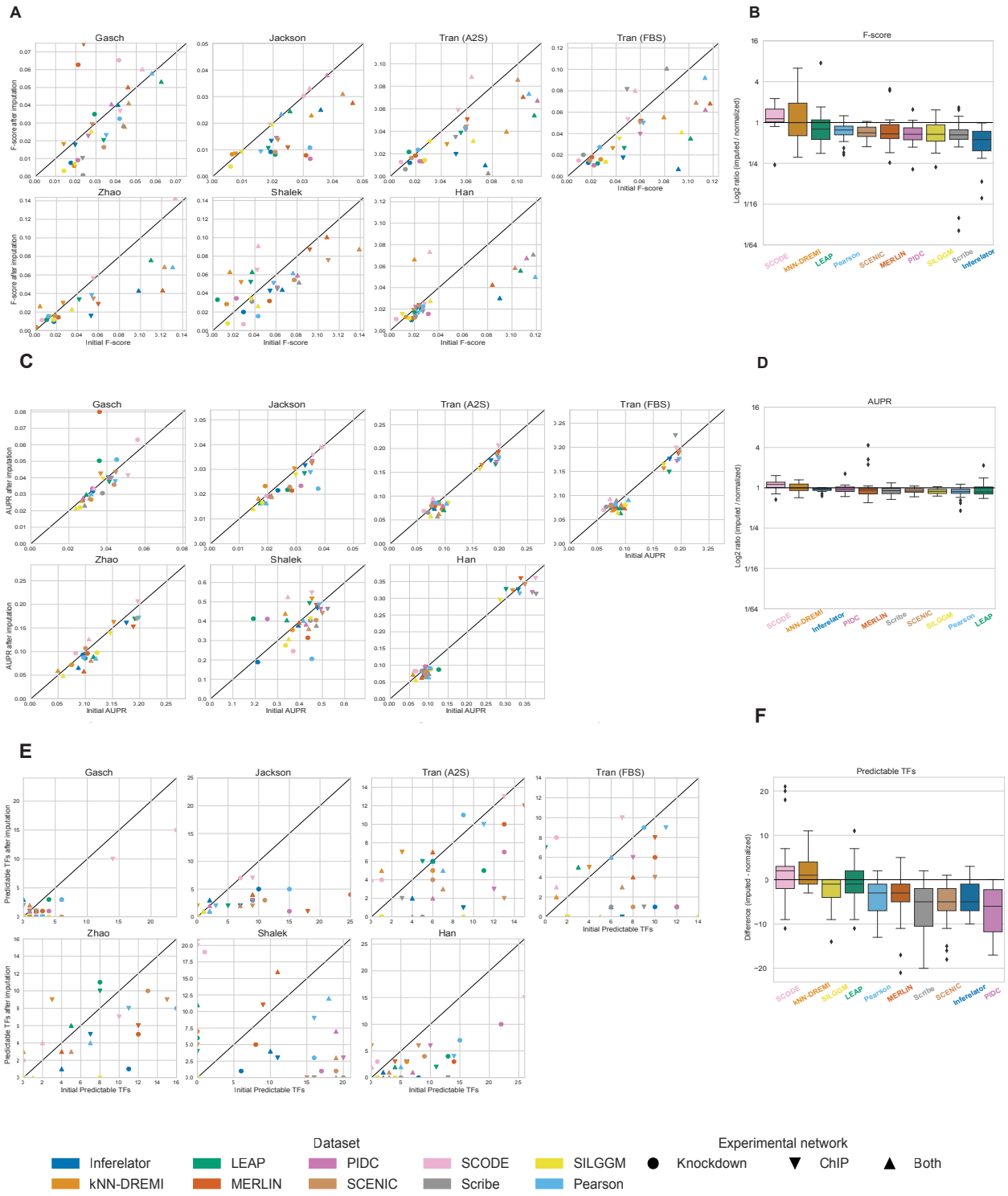

A
Gasch
Jackson
Tran (A2S)
Tran (FBS)
Zhao
Shalek
Han
F-score after imputation
Initial F-score
B
F-score
Log2 ratio (imputed / normalized)
SCODE
kNN-DREMI
LEAP
Pearson
SCENIC
MERLIN
PIDC
SILGGM
Scribe
Inferelator
C
AUPR after imputation
Initial AUPR
D
AUPR
E
Predictable TFs after imputation
Initial Predictable TFs
F
Predictable TFs
Difference (imputed - normalized)
Dataset
Inferelator
kNN-DREMI
LEAP
MERLIN
PIDC
SCENIC
SCODE
Scribe
SILGGM
Pearson
Experimental network
Knockdown
ChIP
Both

Fig 7. Effect of imputation on inferred networks using F-score, AUPR and predictable TFs. **A.** Each panel corresponds to a dataset, the shape of the marker corresponds to the type of gold standard and the color corresponds to the method. The x-axis correspond to the F-score before imputation and y-axis corresponds to after imputation. A marker above the diagonal corresponds to improved performance. **B.** The box plots summarize the performance over the primary 21 comparisons: 7 datasets x 3 experimentally derived networks. Performance based on imputed data is shown relative to the original performance without imputation using log ratio of F-score. Algorithms are ordered based on median performance change. **C.** Similar to **A**, reporting AUPR. **D.** Similar to **B** reporting log ratio of AUPR after and before imputation. **E.** Similar to **A**, reporting the number of predictable TFs. **F.** Similar to **B**, reporting the difference in the predictable TFs between the two networks after and before imputation.

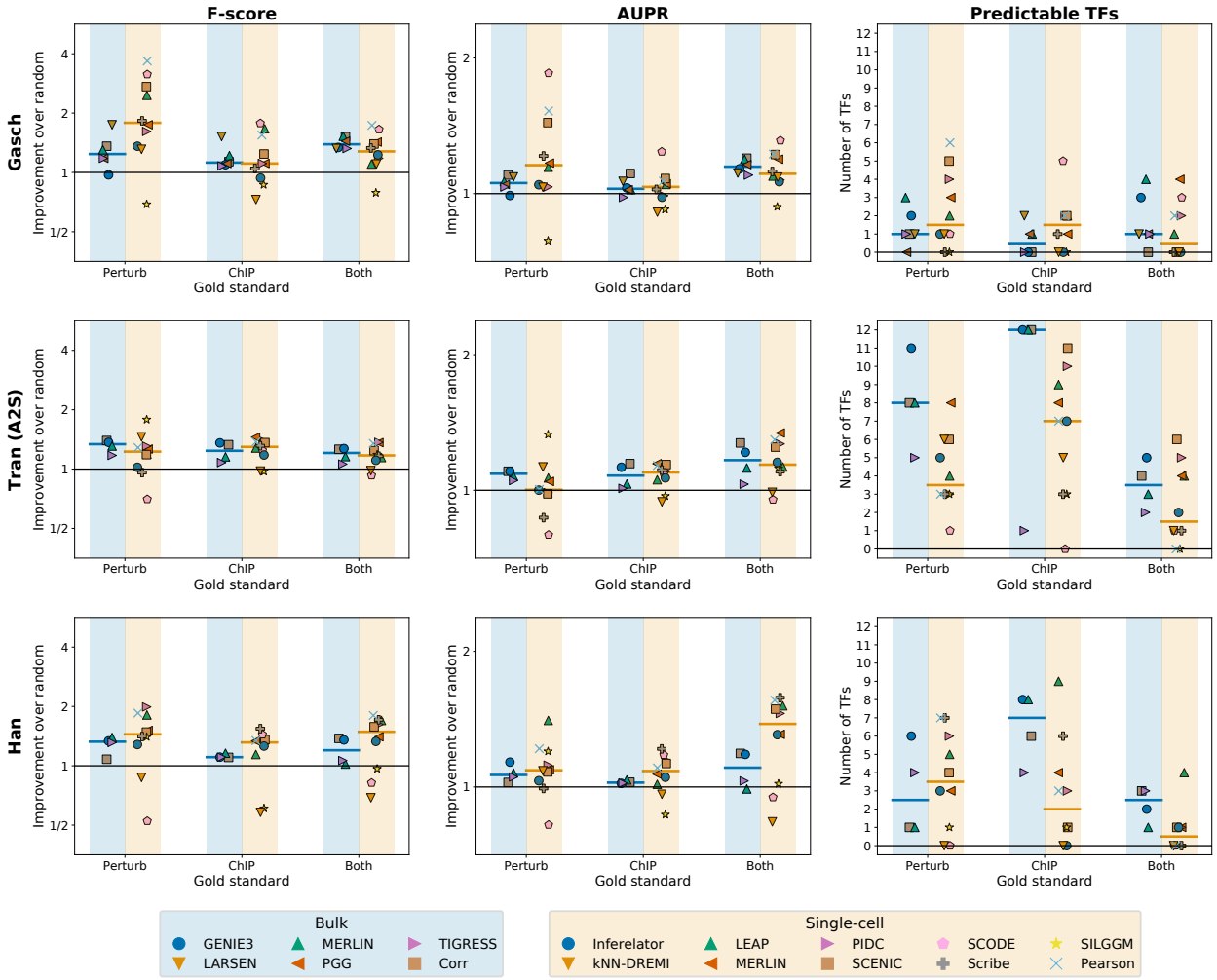

F-score
AUPR
Predictable TFs
Gasch
Tran (A2S)
Han
Improvement over random
Number of TFs
Perturb
ChIP
Both
Gold standard
Bulk
GENIE3
LARSEN
MERLIN
PGG
TIGRESS
Corr
Single-cell
Inferelator
kNN-DREMI
LEAP
MERLIN
PIDC
SCENIC
SCODE
Scribe
SILGGM
Pearson

Fig 8. Comparison of inferred networks on single cell and bulk RNA-seq datasets. Shown are the relative F-score and AUPR scores compared to random and the number of predictable regulators 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).