

# CS 5854: Projects

T. M. Murali

February 15, 2023

# Class Projects that Resulted in Papers

- 1 VIRGO: Computational Prediction of Gene Functions, Naveed Massjouni, Corban Rivera, and T. M. Murali, *Nucleic Acids Research*, 2006.
- 2 Network Legos: Building Blocks of Cellular Wiring Diagrams, T. M. Murali and Corban G. Rivera, *RECOMB 2007*, *JCB* 2008.
- 3 Computational Prediction of Interactions between Host and Pathogen Proteins, Matthew Dyer, T. M. Murali, and Bruno Sobral, *ISMB 2007*.
- 4 Divergence of Gene Expression Profiles in Tandemly Arrayed Genes in Human and Mouse, Valia Shoja, T. M. Murali, and Liqing Zhang, *Comparative and Functional Genomics*, 2007.
- 5 Network-Based Prediction and Analysis of HIV Dependency Factors, T. M. Murali, Matthew D. Dyer, David Badger, Brett M. Tyler, and Michael G. Katze, *PLoS Computational Biology*, 2011.
- 6 Top-Down Network Analysis to Drive Bottom-Up Modeling of Physiological Processes, Christopher L. Poirel, Richard R. Rodrigues, Katherine C. Chen, John J. Tyson, and T. M. Murali, *JCB*, 2013.
- 7 Pathways on Demand: Automatic Reconstruction of Human Signaling Networks, Anna Ritz, Christopher L. Poirel, Allison N. Tegge, Nicholas Sharp, Allison Powell, Kelsey Simmons, Shiv D. Kale, and T. M. Murali, *npj: Systems Biology and Applications*, 2016.
- 8 Computational Construction of Toxicant Signaling Networks, Jeffrey Law, Sophia M. Orbach, Bronson Weston, Peter Steele, Padmavathy Rajagopalan, and T. M. Murali, revision in preparation, *Chemical Research in Toxicology*, 2023.

# List of Projects

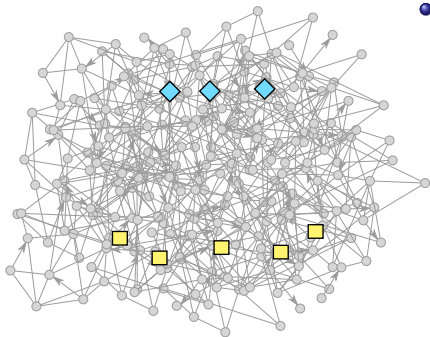
- 1 Develop PathLinker 2.0
- 2 Develop BEELINE 2.0
- 3 Predict cell types
- 4 Predict protein complexes
- 5 Predict virus-host interactions

# Overview

- 1 Develop PathLinker 2.0
- 2 BEELINE 2.0
- 3 Cell Type Prediction
- 4 Predict Structures of Virus-Host Protein Complexes
- 5 Predict Virus-Host Interactions

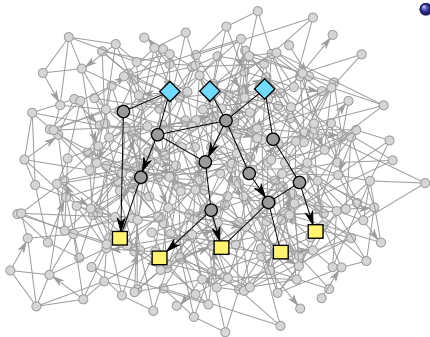


# 1 PathLinker 2.0



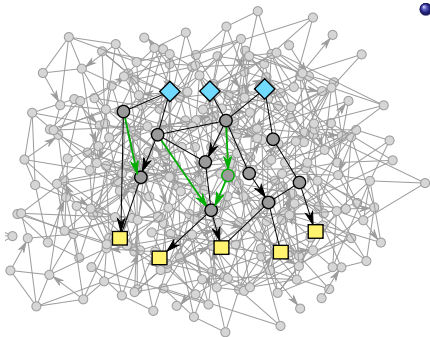
- Goal: develop a new algorithm to reconstruct signaling pathways.
  - ▶ Supervised algorithm that can use information on the edges in a pathway to predict new edges.
  - ▶ Permit paths with cycles.
  - ▶ Propose alternative, biologically meaningful formulations of the problem.

# 1 PathLinker 2.0



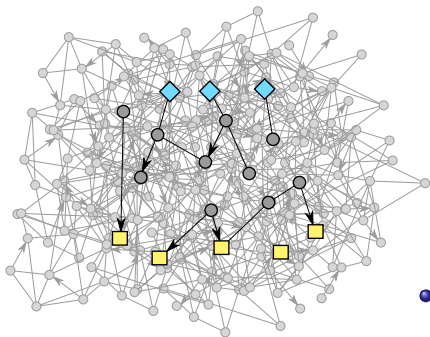
- Goal: develop a new algorithm to reconstruct signaling pathways.
  - ▶ Supervised algorithm that can use information on the edges in a pathway to predict new edges.
  - ▶ Permit paths with cycles.
  - ▶ Propose alternative, biologically meaningful formulations of the problem.

# 1 PathLinker 2.0



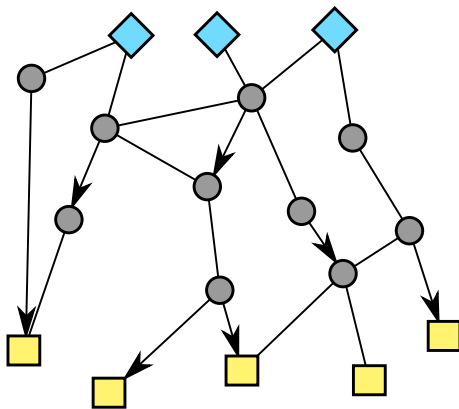
- Goal: develop a new algorithm to reconstruct signaling pathways.
  - ▶ Supervised algorithm that can use information on the edges in a pathway to predict new edges.
  - ▶ Permit paths with cycles.
  - ▶ Propose alternative, biologically meaningful formulations of the problem.

# 1 PathLinker 2.0



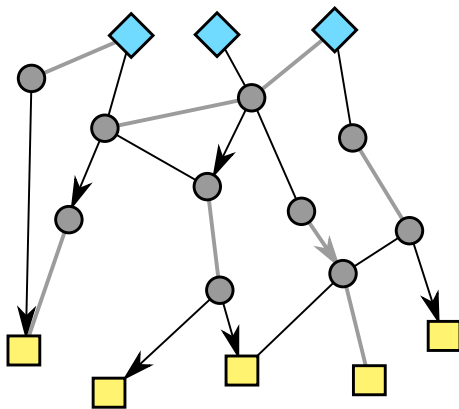
- Goal: develop a new algorithm to reconstruct signaling pathways.
  - ▶ Supervised algorithm that can use information on the edges in a pathway to predict new edges.
  - ▶ Permit paths with cycles.
  - ▶ Propose alternative, biologically meaningful formulations of the problem.
- Use cross-validation to test performance: develop meaningful ways of deleting nodes/edges.

# Evaluating Reconstructions with Cross Validation



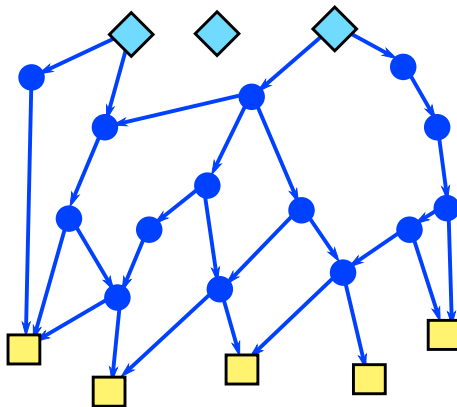
Curated pathway

# Evaluating Reconstructions with Cross Validation



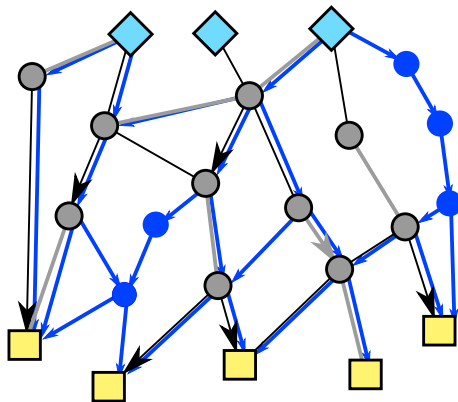
Edges removed for cross validation

# Evaluating Reconstructions with Cross Validation



Proposed reconstruction

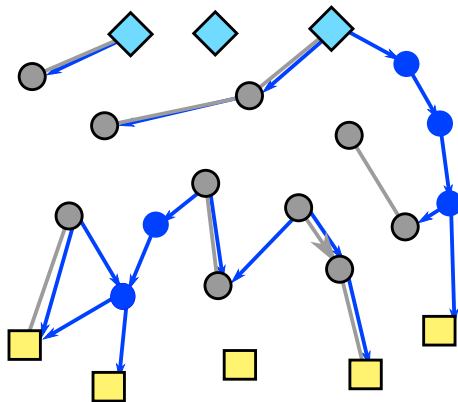
# Evaluating Reconstructions with Cross Validation



Curated pathway and proposed reconstruction

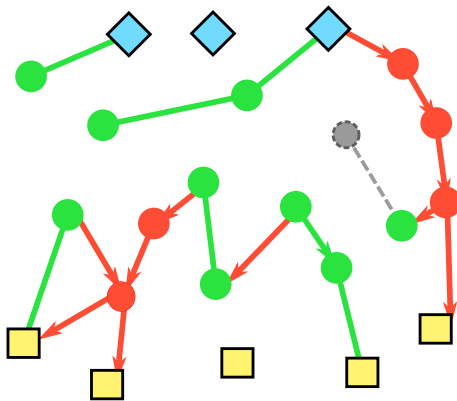


# Evaluating Reconstructions with Cross Validation



Cross validation edges and proposed reconstruction

# Evaluating Reconstructions with Cross Validation



Precision and recall

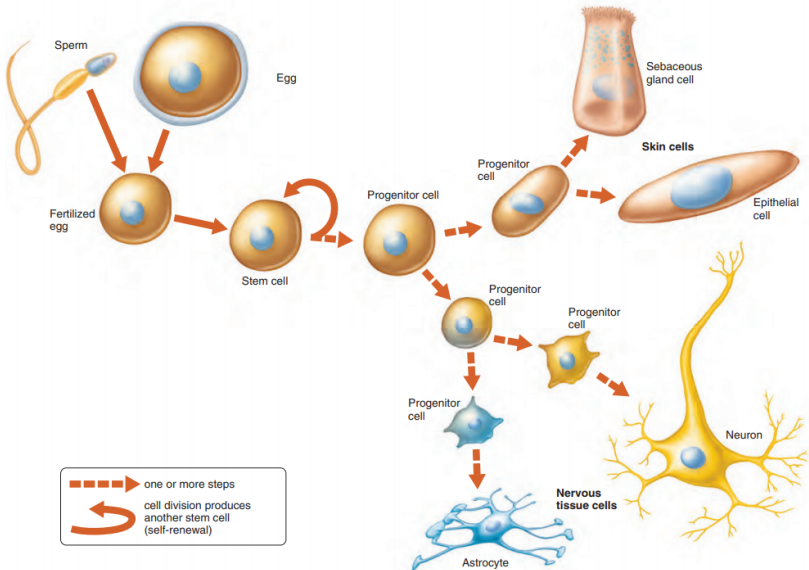
# Project Details

- Paper: **Pathways on Demand: Automatic Reconstruction of Human Signaling Pathways**, Ritz et al., *Systems Biology and Applications*, a Nature partner journal, 2016
- Ideas published in the literature since the PathLinker paper.
- **PathLinker code**
- **SPRAS software** for comparing pathway reconstruction algorithms.
- Update signaling pathways network dataset used in PathLinker paper. What does SPRAS do?
- Find several meaningful alternatives methods to compare your algorithm with.
- **Create computational analyses that are different from the PathLinker paper.**
- **Do literature analysis of predicted interactions.**

# Overview

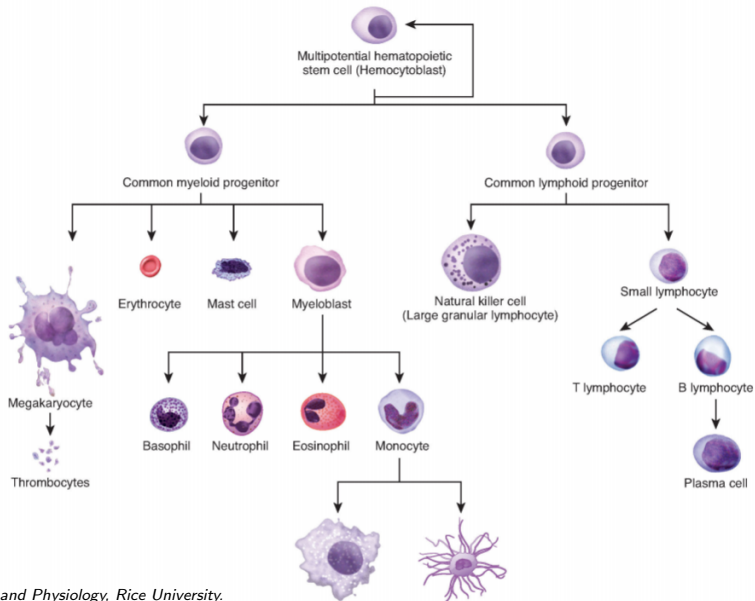
- 1 Develop PathLinker 2.0
- 2 BEELINE 2.0**
- 3 Cell Type Prediction
- 4 Predict Structures of Virus-Host Protein Complexes
- 5 Predict Virus-Host Interactions

# Cellular Differentiation

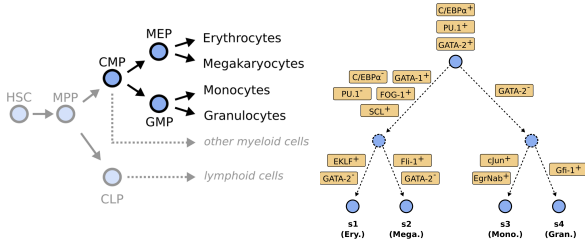


Shier et al., (2015) "Hole's Essentials of Human Anatomy and Physiology", McGraw-Hill

# Cellular Differentiation



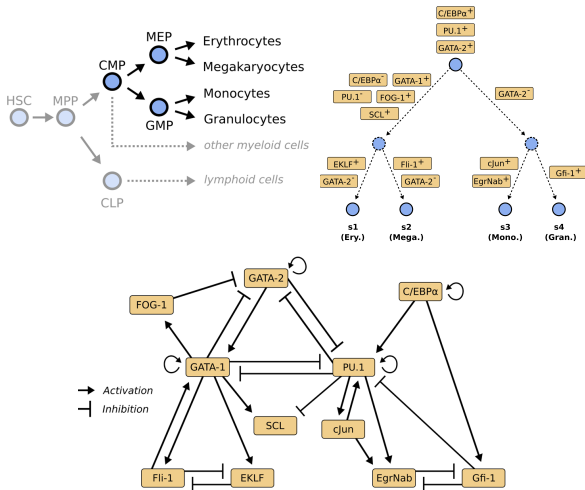
# Cellular Differentiation



- Cells in different states express different sets of genes.
- Cells move from one “state” to another.

Krumsiek et al. (2010). “Hierarchical Differentiation of Myeloid Progenitors...” PLoS ONE

# Cellular Differentiation

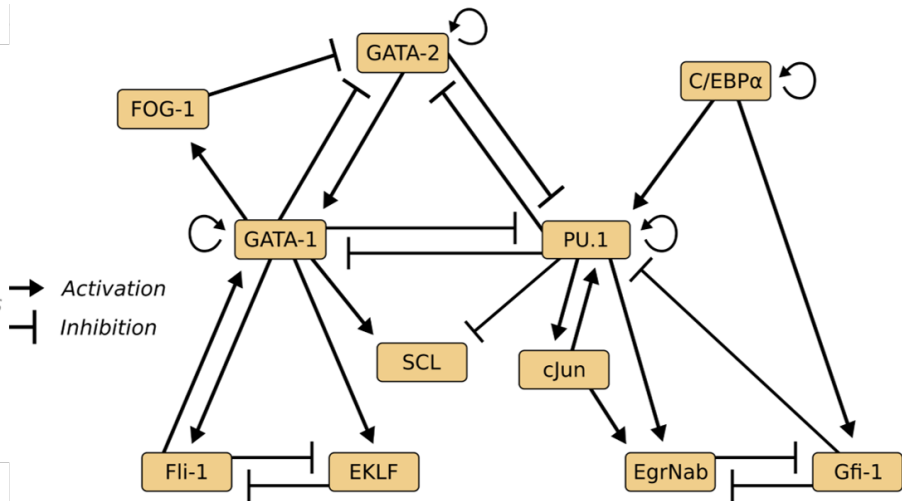


- Transcription factors activate/inhibit genes to effect cell transition from one state to another.

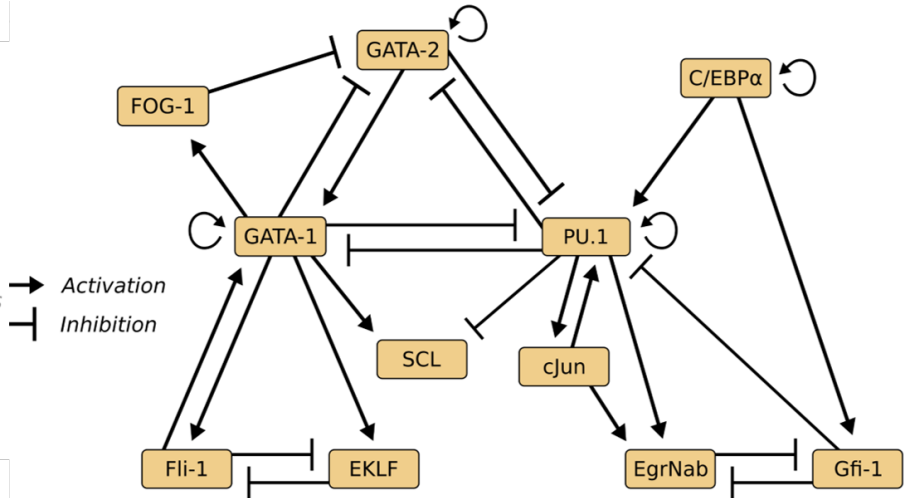
Krumsiek et al. (2010). "Hierarchical Differentiation of Myeloid Progenitors..." *PLoS ONE*



# Gene Regulatory Network (GRN)



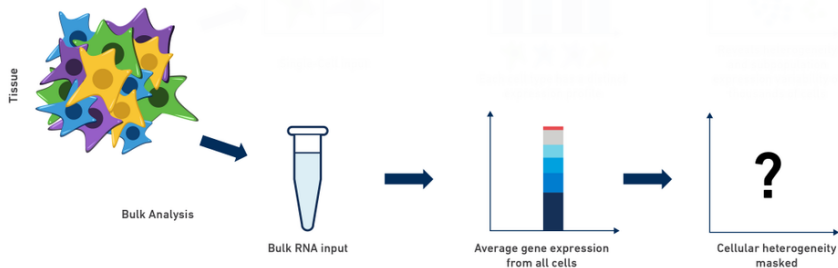
# Gene Regulatory Network (GRN)



How do we build GRNs using computational techniques?

# Bulk RNA Sequencing

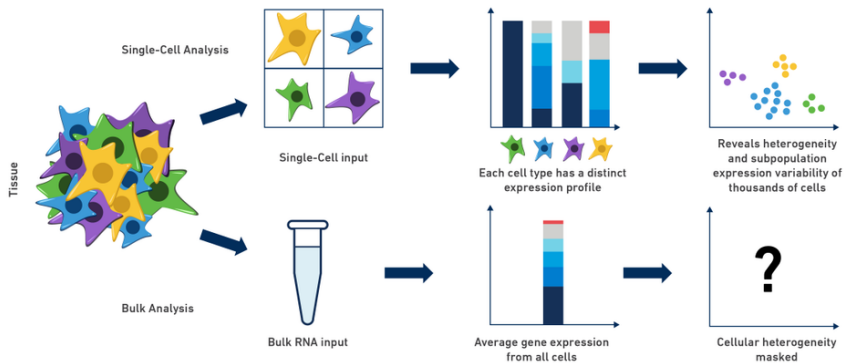
- A population of cells isolated at the same time may correspond to multiple, distinct intermediate differentiation states.
- Averages gene expression and masks cellular heterogeneity.
- Difficult to experimentally purify cells in intermediate states.



10x Genomics

# Single-cell RNA Sequencing (scRNA-seq)

- Produce thousands of independent measurements.
- Computational ordering of cells along “lineages” provide a high-resolution “pseudotemporal” view of gene expression kinetics.
- Richness of these datasets may facilitate inference.



10x Genomics

Trapnell et al., "The dynamics and regulators of cell fate decisions ...", *Nat. Biotech.*, 2014.

# Over a Dozen Methods Have Already Been Developed

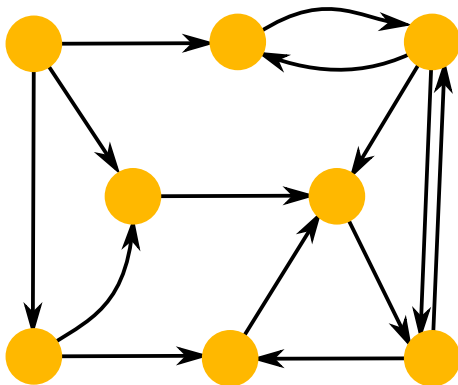
	Properties	
	Category	Addl. Inputs
PIDC	MI	-
GENIE3	RF	-
GRNBOOST2	RF	-
SCODE	ODE+Reg	ODE parameters
PPCOR	Corr	-
SINCERITIES	Reg	-
SCRIBE	MI	Type of RDI
SINGE	GC	Regression parameters
LEAP	Corr	Lag
GRISLI	ODE+Reg	Regression parameters
GRNVBEM	Reg	-
SCNS	Bool	Boolean model

# Over a Dozen Methods Have Already Been Developed

	Properties	
	Category	Addl. Inputs
PIDC	MI	-
GENIE3	RF	-
GRNBOOST2	RF	-
SCODE	ODE+Reg	ODE parameters
PPCOR	Corr	-
SINCERITIES	Reg	-
SCRIBE	MI	Type of RDI
SINGE	GC	Regression parameters
LEAP	Corr	Lag
GRISLI	ODE+Reg	Regression parameters
GRNVBEM	Reg	-
SCNS	Bool	Boolean model

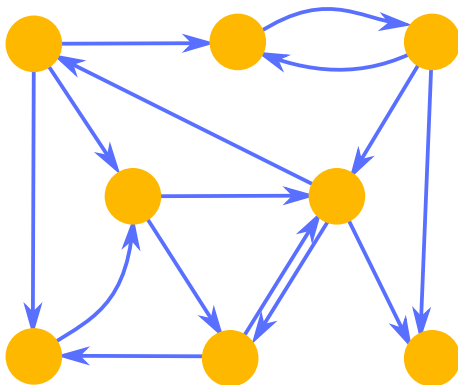
How accurately do these methods infer GRNs?

# Evaluation of Inferred GRNs



Ground-truth GRN

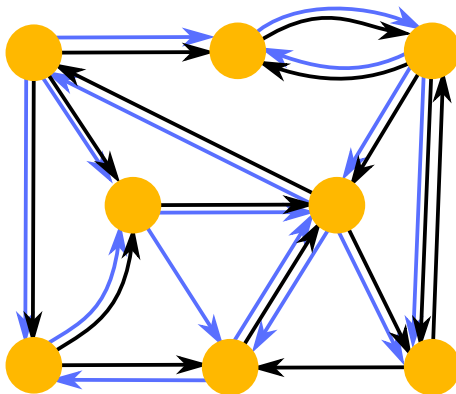
# Evaluation of Inferred GRNs



Inferred Reconstruction

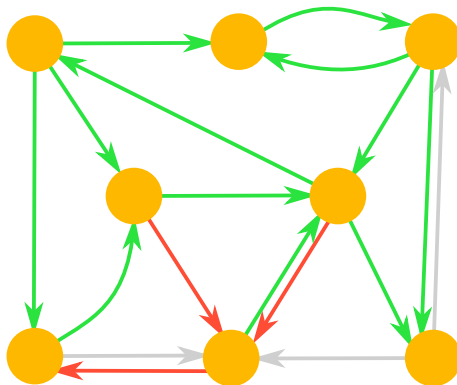


# Evaluation of Inferred GRNs



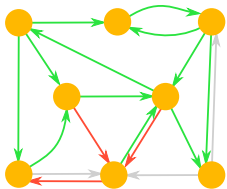
Ground-truth GRN and Inferred Reconstruction

# Evaluation of Inferred GRNs

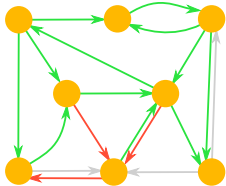


Overlap

# Evaluation of Inferred GRNs



# Evaluation of Inferred GRNs

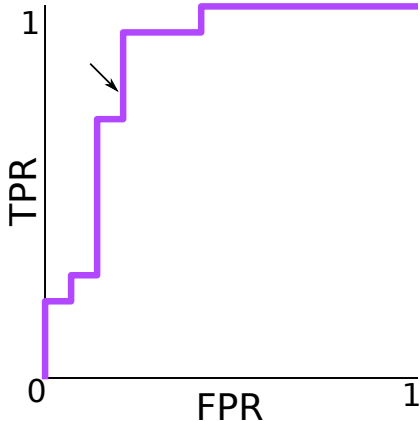


Specificity/False positive rate:

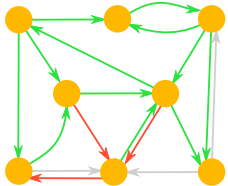
$$sp_i = \frac{\# \text{false positives up to } i}{\# \text{negatives}}$$

Recall/True positive rate:

$$r_i = \frac{\# \text{true positives up to } i}{\# \text{positives}}$$



# Evaluation of Inferred GRNs

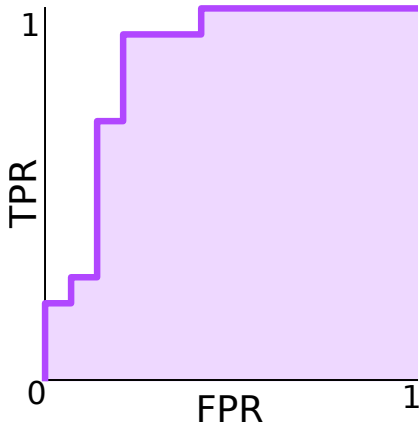


Specificity/False positive rate:

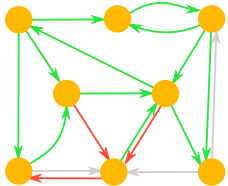
$$sp_i = \frac{\# \text{false positives up to } i}{\# \text{negatives}}$$

Recall/True positive rate:

$$r_i = \frac{\# \text{true positives up to } i}{\# \text{positives}}$$



# Evaluation of Inferred GRNs

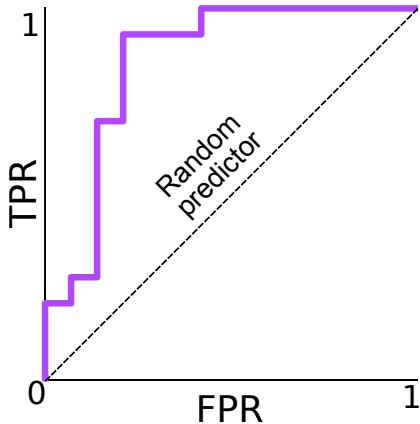


Specificity/False positive rate:

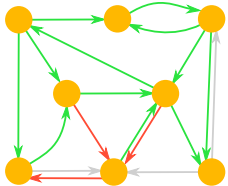
$$sp_i = \frac{\# \text{false positives up to } i}{\# \text{negatives}}$$

Recall/True positive rate:

$$r_i = \frac{\# \text{true positives up to } i}{\# \text{positives}}$$



# Evaluation of Inferred GRNs



Specificity/False positive rate:

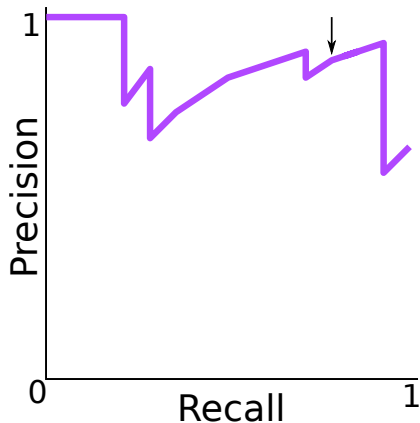
$$sp_i = \frac{\# \text{false positives up to } i}{\# \text{negatives}}$$

Recall/True positive rate:

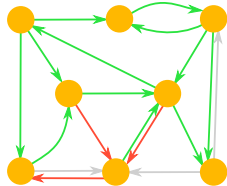
$$r_i = \frac{\# \text{true positives up to } i}{\# \text{positives}}$$

Precision:

$$p_i = \frac{\# \text{true positives up to } i}{i}$$



# Evaluation of Inferred GRNs



Specificity/False positive rate:

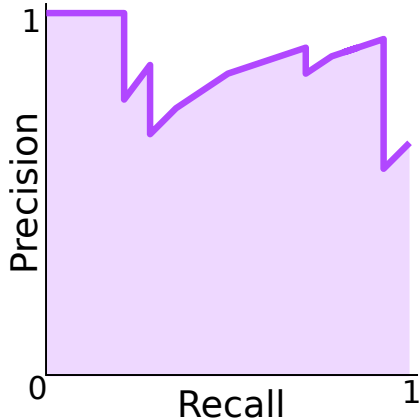
$$sp_i = \frac{\# \text{false positives up to } i}{\# \text{negatives}}$$

Recall/True positive rate:

$$r_i = \frac{\# \text{true positives up to } i}{\# \text{positives}}$$

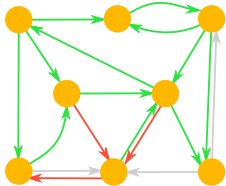
Precision:

$$p_i = \frac{\# \text{true positives up to } i}{i}$$





# Evaluation of Inferred GRNs



Specificity/False positive rate:

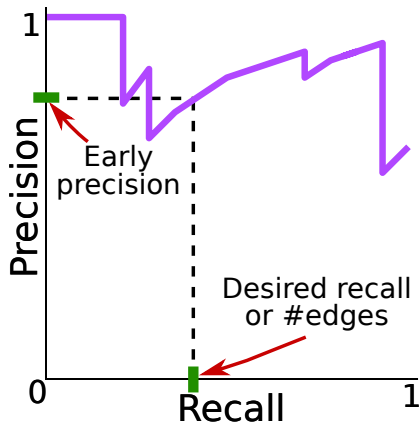
$$sp_i = \frac{\# \text{false positives up to } i}{\# \text{negatives}}$$

Recall/True positive rate:

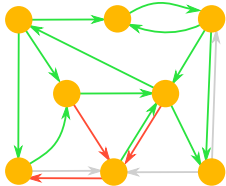
$$r_i = \frac{\# \text{true positives up to } i}{\# \text{positives}}$$

Precision:

$$p_i = \frac{\# \text{true positives up to } i}{i}$$



# Evaluation of Inferred GRNs



Specificity/False positive rate:

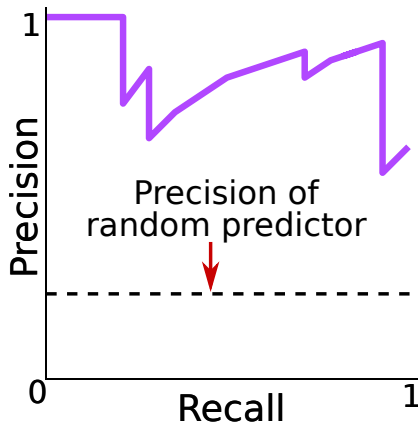
$$sp_i = \frac{\# \text{false positives up to } i}{\# \text{negatives}}$$

Recall/True positive rate:

$$r_i = \frac{\# \text{true positives up to } i}{\# \text{positives}}$$


Precision:

$$p_i = \frac{\# \text{true positives up to } i}{i}$$



# Performance of Current Algorithms

**SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation** 

Hirotaaka Matsumoto , Hisanori Kiryu, Chikara Furusawa, Minoru S H Ko, Shigeru B H Ko, Norio Gouda, Tetsutaro Hayashi, Itoshi Nikaido

*Bioinformatics*, Volume 33, Issue 15, 01 August 2017, Pages 2314–2321,  
<https://doi.org/10.1093/bioinformatics/btx194>

**Published:** 04 April 2017 **Article history** 

## Table 1

The AUC values of each method for each dataset

	<b>SCODE</b>	<b>lm</b>	<b>msgps</b>	<b>Cor</b>	<b>GENIE3</b>	<b>Jump3</b>
Data1	0.536	0.480	0.510	0.505	0.474	0.504
Data2	0.581	0.489	0.516	0.492	0.472	0.492
Data3	0.523	0.480	0.499	0.524	0.522	0.501

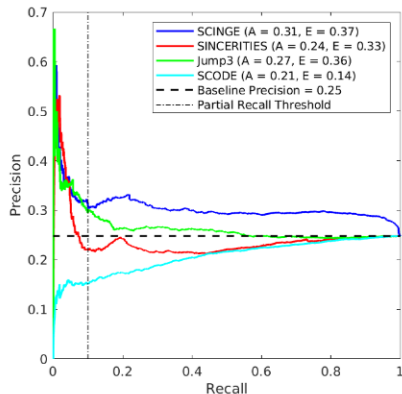
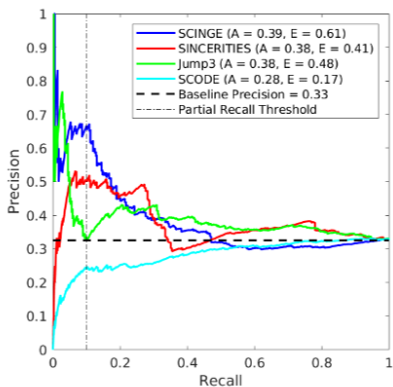
*Note:* Cor is the correlation network.

# Performance of Current Algorithms

## Network Inference with Granger Causality Ensembles on Single-Cell Transcriptomic Data

Atul Deshpande, Li-Fang Chu, Ron Stewart, Anthony Glitter

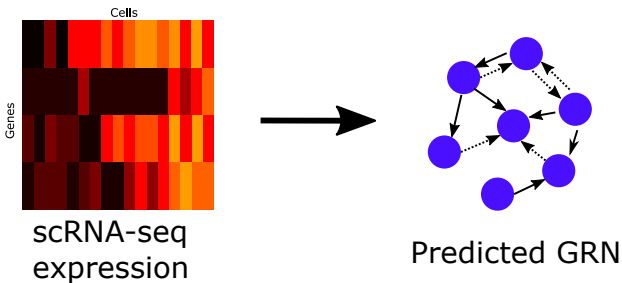
doi: <https://doi.org/10.1101/534834>



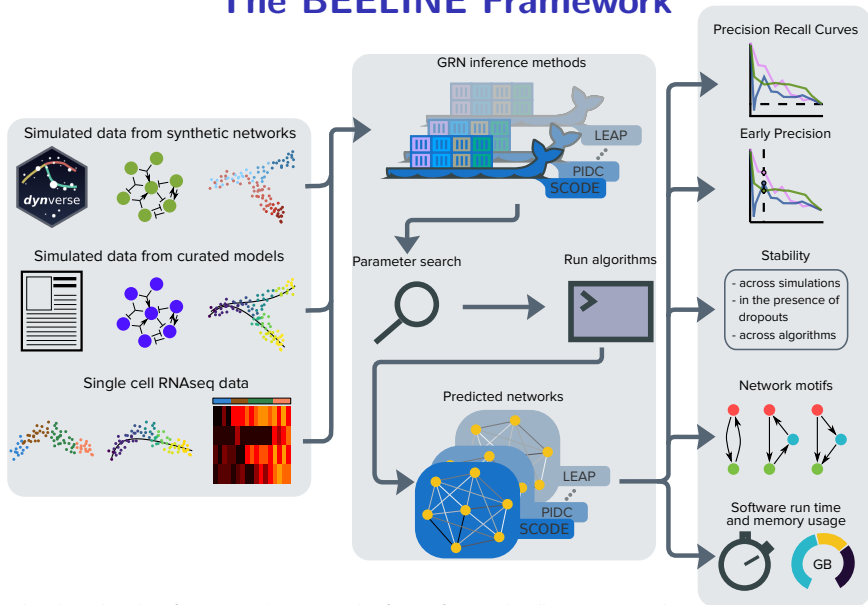
Performance is close to that of a random predictor!

# Motivation for Benchmarking GRN Inference Methods

- Criteria for evaluation and comparison of methods vary from one paper to another.
- No existing framework for systematic comparison of methods.

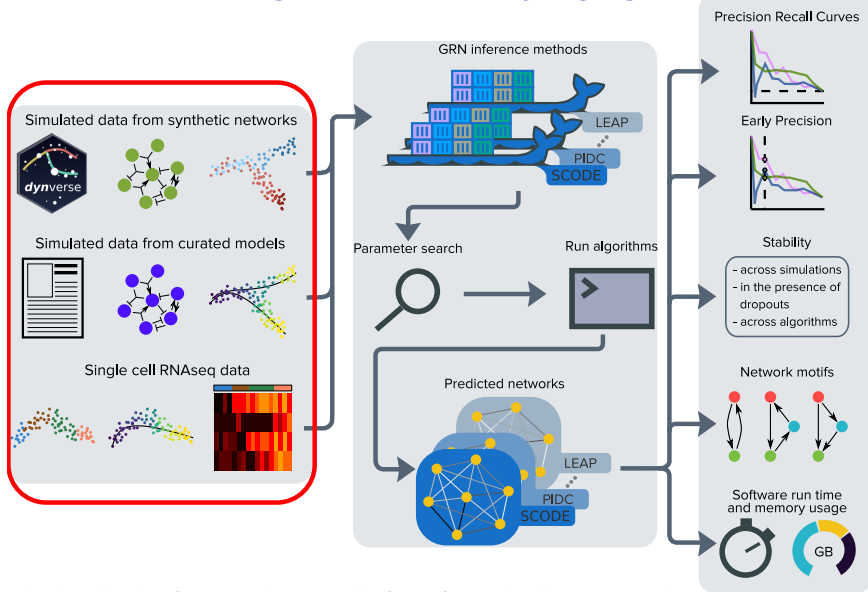


# The BEELINE Framework



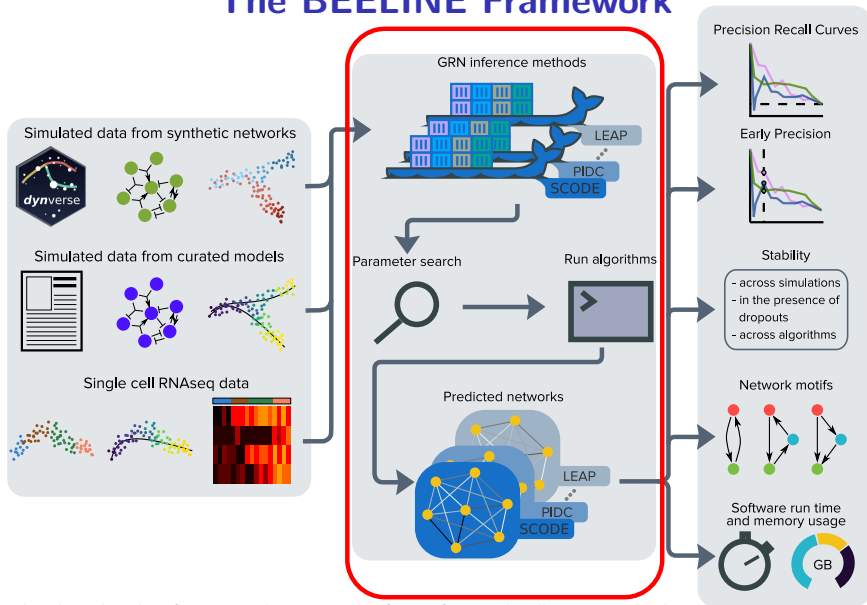
*Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data, Pratapa, Jalihal, Law, Bharadwaj, and Murali, Nature Methods, 2020.*

# The BEELINE Framework



*Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data, Pratapa, Jalihal, Law, Bharadwaj, and Murali, Nature Methods, 2020.*

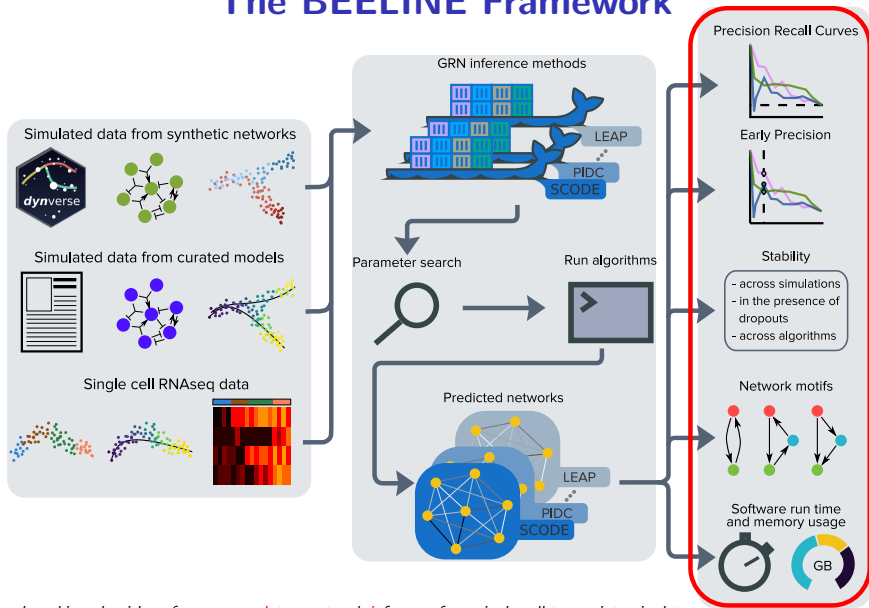
# The BEELINE Framework



*Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data, Pratapa, Jalihal, Law, Bharadwaj, and Murali, Nature Methods, 2020.*

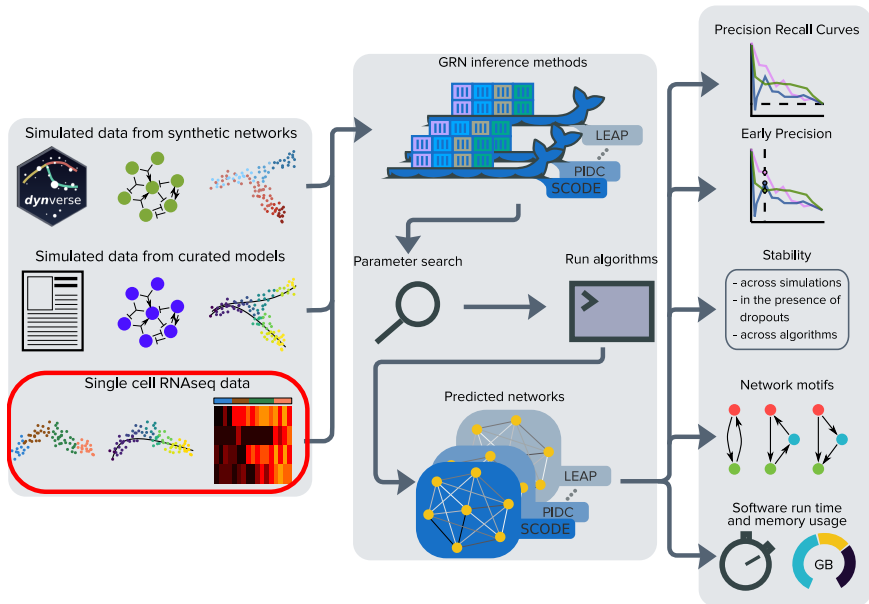


# The BEELINE Framework

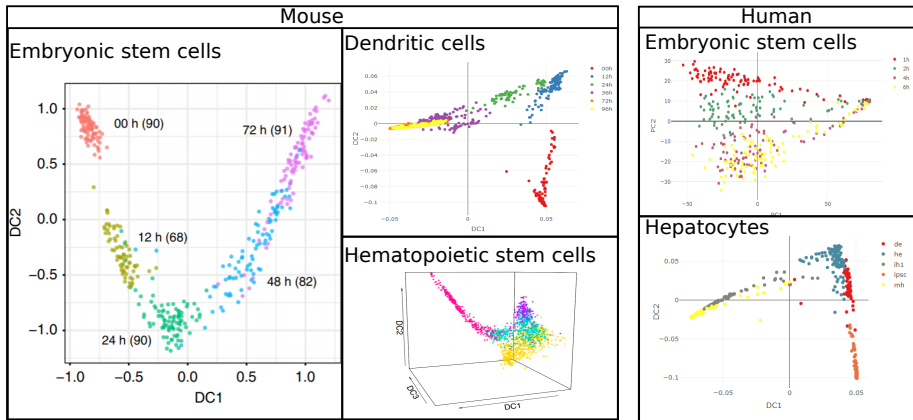


*Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data, Pratapa, Jalihal, Law, Bharadwaj, and Murali, Nature Methods, 2020.*

# Input Type 3: Experimental scRNA-seq Datasets



# Input Type 3: Experimental scRNA-seq Datasets



<sup>1</sup> Nestorowa, et al. (2016) "A single-cell ...". *Blood*, 128, 20–31.

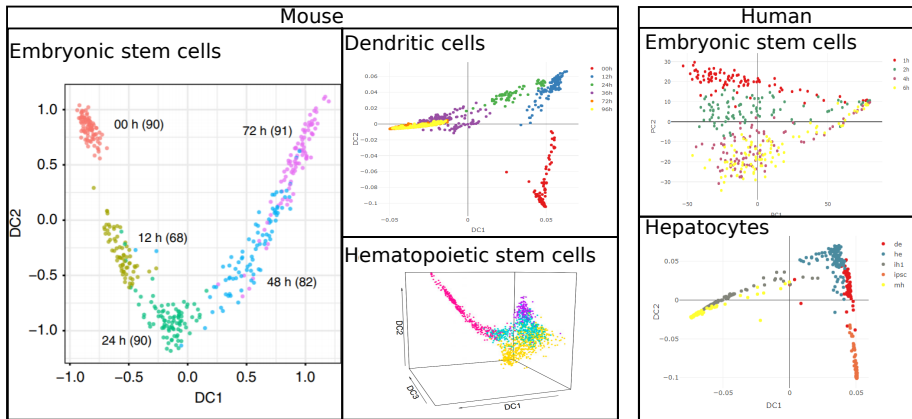
<sup>2</sup> Hayashi et al. (2018) "Single-cell ..." *Nat. Commun.*, 9, 619.

<sup>3</sup> Shalek et al. (2014) "Single-cell RNA-seq ..." *Nature*, 510, 363–369.

<sup>4</sup> Camp et al. (2017) "Multilineage communication ..." *Nature*, 546, 533–538.

<sup>5</sup> Chu et al. (2016) "Single-cell ..." *Genome Biol*, 17, 173.

# Input Type 3: Experimental scRNA-seq Datasets



No standard ground-truth networks

# Ground-Truth Networks

- Cell-type specific ChIP-seq network<sup>1</sup>
- Non-specific ChIP-seq network<sup>2,3,4</sup>
- STRING network<sup>5</sup>

1. Oki et al. (2018) "ChIP-Atlas: ..." *EMBO Rep.* e46255

2. Liu et al. (2015) "RegNetwork: an integrated ..." *Database*, 2015

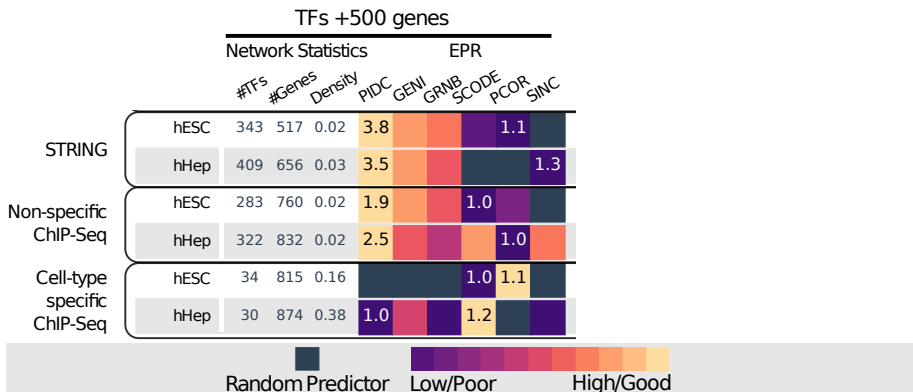
3. Han et al. (2018) "TRRUSTv2 ..." *NAR*, 46(D1):D380–D386

4. Garcia et al. (2019) "Benchmark ..." *Gen. Res.*, 29:1363–1375

5. Szklarczyk et al. (2019) "STRING v11..." *NAR*, 47:D607–613

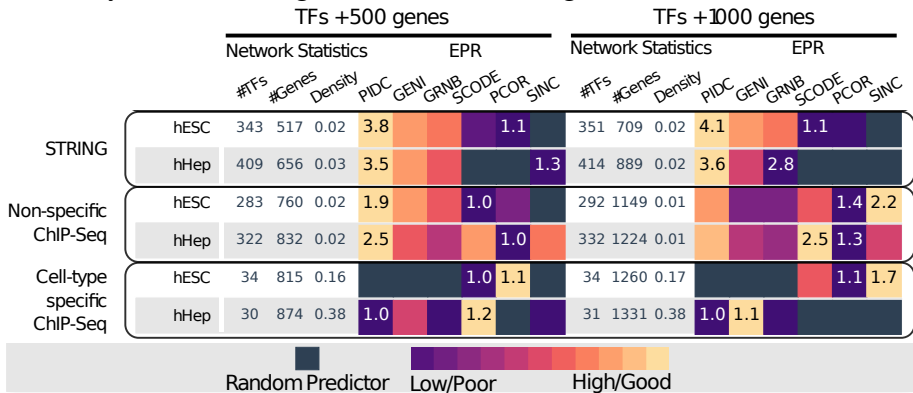
## Results on Human Datasets

- EPR: predicted interactions of higher confidence will be more interesting to experimentalists.
- Only considered edges between TFs and genes.



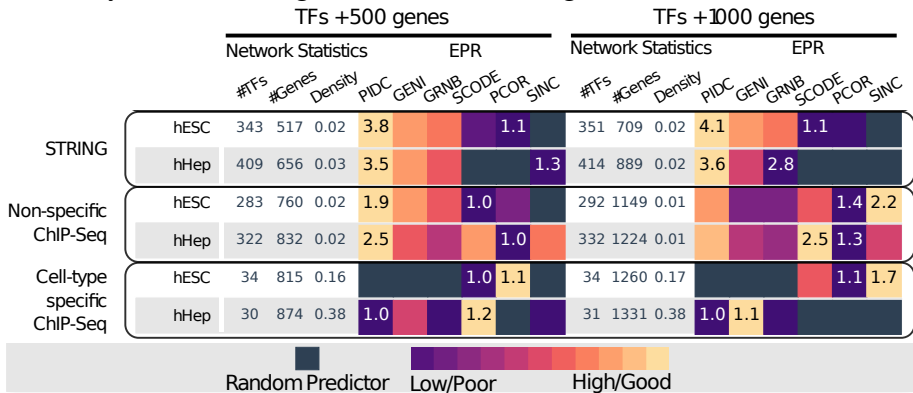
## Results on Human Datasets

- EPR: predicted interactions of higher confidence will be more interesting to experimentalists.
- Only considered edges between TFs and genes.



## Results on Human Datasets

- EPR: predicted interactions of higher confidence will be more interesting to experimentalists.
- Only considered edges between TFs and genes.



Substantial fraction of the edges in the inferred GRNs were indirect.



# Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data

Aditya Pratapa<sup>1</sup>, Amogh P. Jalihal<sup>2</sup>, Jeffrey N. Law<sup>2</sup>, Aditya Bharadwaj<sup>1</sup> and T. M. Murali<sup>1\*</sup>

We present a systematic evaluation of state-of-the-art algorithms for inferring gene regulatory networks from single-cell transcriptional data. As the ground truth for assessing accuracy, we use synthetic networks with predictable trajectories, literature-curated Boolean models and diverse transcriptional regulatory networks. We develop a strategy to simulate single-cell transcriptional data from synthetic and Boolean networks that avoids pitfalls of previously used methods. Furthermore, we collect networks from multiple experimental single-cell RNA-seq datasets. We develop an evaluation framework called BEELINE. We find that the area under the precision-recall curve and early precision of the algorithms are moderate. The methods are better in recovering interactions in synthetic networks than Boolean models. The algorithms with the best early precision values for Boolean models also perform well on experimental datasets. Techniques that do not require pseudotime-ordered cells are generally more accurate. Based on these results, we present recommendations to end users. BEELINE will aid the development of gene regulatory network inference algorithms.

Single-cell RNA-sequencing technology has made it possible to trace cellular lineages during differentiation and to identify new cell types<sup>1,2</sup>. A central question that arises now is whether we can discover the gene regulatory networks (GRNs) that control cellular differentiation and drive transitions from one cell type to another. In such a GRN, each edge connects a transcription factor (TF) to a gene it regulates. Ideally, the edge is directed from the TF to the target gene, represents direct rather than indirect regulation and corresponds to activation

## Results

**Overview of algorithms.** We surveyed the literature and bioRxiv preprints for papers that either published a new GRN inference algorithm or used an existing approach. We ignored methods that did not assign weights or ranks to the interactions, required additional datasets or supervision, or sought to discover cell-type-specific networks. We selected 12 algorithms using these criteria (Methods).

We used BEELINE to evaluate these approaches on over 400 simulated datasets (across six synthetic networks and four curated

## Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data

58

A Pratapa, AP Jalihal, JN Law, A Bharadwaj, TM Murali  
Nature methods 17 (2), 147-154

Pratapa, A., et al. "Benchmarking algorithms for gene regulatory ..." *Nat. Methods* (2020), 17(2), pp. 147-154.

## 2 BEELINE 2.0

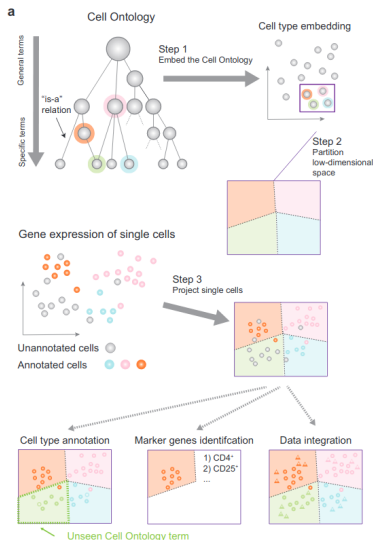
- Goal: Improve usefulness of BEELINE for experimental scRNA-seq datasets.
- High priority
  - ▶ Implement continuous integration.
  - ▶ Add GRN inference methods and test them.
  - ▶ Develop alternative gene selection strategies.
  - ▶ Implement additional evaluation measures developed in GRN inference papers.
- Medium priority
  - ▶ Add experimental scRNA-seq datasets.
  - ▶ Find better ground truth datasets. Automate selection of cell type.
- Low priority
  - ▶ Try **imputation of missing data** first.
  - ▶ Add denoising methods, e.g., **molecular cross validation** paper.
  - ▶ For real datasets, run parameter search on each type of network when using that type as ground truth.

# Overview

- 1 Develop PathLinker 2.0
- 2 BEELINE 2.0
- 3 Cell Type Prediction**
- 4 Predict Structures of Virus-Host Protein Complexes
- 5 Predict Virus-Host Interactions

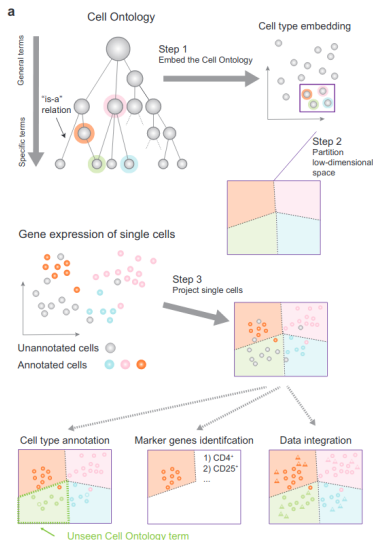
# 3 Cell Type Prediction

- A comparison of automatic cell identification methods for single-cell RNA sequencing data
- Leveraging the Cell Ontology to classify unseen cell types uses a 2-layer perceptron in combination with the Cell Ontology.



# 3 Cell Type Prediction

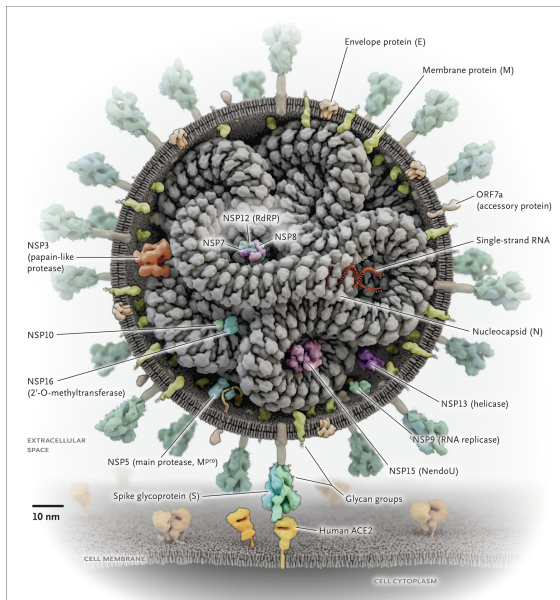
- A comparison of automatic cell identification methods for single-cell RNA sequencing data
- Leveraging the Cell Ontology to classify unseen cell types uses a 2-layer perceptron in combination with the Cell Ontology.
- Use network-based algorithms for predicting cell types.
  - ▶ Two networks: one is among cells and the other is the Cell Ontology.
  - ▶ Evaluate network propagation algorithms that respect the ontology structure.
  - ▶ Alternative is to develop improved deep learning methods.



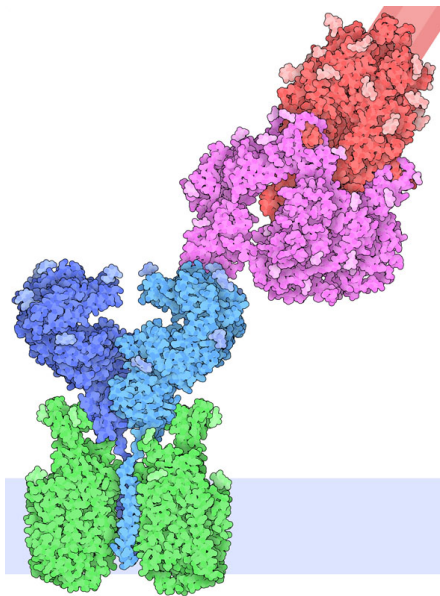
# Overview

- 1 Develop PathLinker 2.0
- 2 BEELINE 2.0
- 3 Cell Type Prediction
- 4 Predict Structures of Virus-Host Protein Complexes**
- 5 Predict Virus-Host Interactions

# S and Ace2

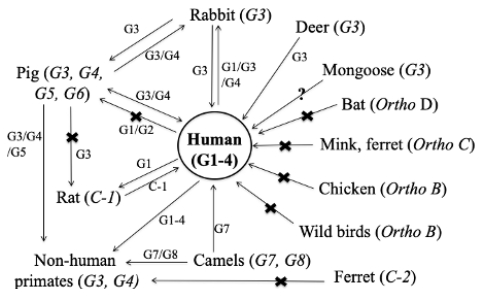


# S and Ace2



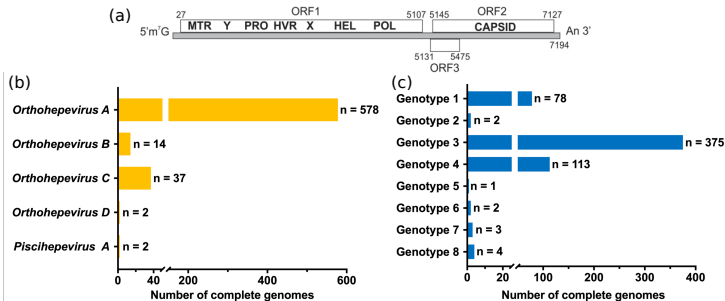


# Hepatitis E Virus



- Causes a public health disease worldwide with more than 20 million new human infections and 40,000 deaths annually.
- HEV-specific antiviral drug is not available.
- Zoonotic pathogen with more than a dozen animal hosts.
- Viral genetic element(s) responsible for species jumping and adaptation in humans remain unknown.

# Hepatitis E Virus



- Many HEV sequences that infect humans and other animal species are available.
- HEV-human protein interactions are known via yeast 2-hybrid screens.
- **Human receptor for Capsid protein is unknown!**

## 4 Predict Host Receptor for a Virus

- Given a viral gene sequence  $v$  and a human receptor sequence  $r$ , predict the structure of the complex.  $v$  and  $r$  may have experimentally-determined structures or you can predict their structures.
- Given a viral gene sequence  $v$  and two human receptors  $r_1$  and  $r_2$ , does  $v$  bind better to  $r_1$  than to  $r_2$ ?
- Given two viral gene sequences  $v_1$  and  $v_2$  and a human receptor  $r$ , does  $v_1$  bind better to  $r$  than  $v_2$  to  $r$ ?

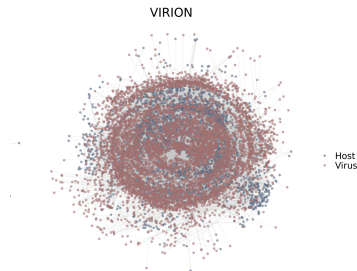
## 4 Predict Host Receptor for a Virus

- Given a viral gene sequence  $v$  and a human receptor sequence  $r$ , predict the structure of the complex.  $v$  and  $r$  may have experimentally-determined structures or you can predict their structures.
- Given a viral gene sequence  $v$  and two human receptors  $r_1$  and  $r_2$ , does  $v$  bind better to  $r_1$  than to  $r_2$ ?
- Given two viral gene sequences  $v_1$  and  $v_2$  and a human receptor  $r$ , does  $v_1$  bind better to  $r$  than  $v_2$  to  $r$ ?
- Use any techniques that are appropriate:
  - ▶ Predict structures of individual proteins and use simulations to test if they interact.
  - ▶ Use existing algorithms to predict and score structures of protein complexes.
  - ▶ Develop your own ideas.

# Overview

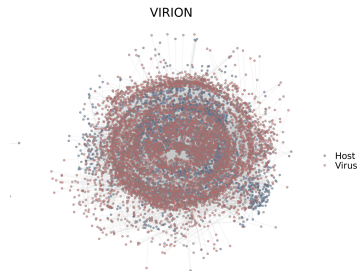
- 1 Develop PathLinker 2.0
- 2 BEELINE 2.0
- 3 Cell Type Prediction
- 4 Predict Structures of Virus-Host Protein Complexes
- 5 Predict Virus-Host Interactions**

# Virus-Host Interactions



- Global virome (set of all viruses in the biosphere) is highly underdocumented.
- $> 40,000$  species of viruses may infect mammals and thousands can probably infect humans.
- **The Global Virome in One Network (VIRION) database** records 23,147 unique interactions between 9,521 viruses and 3,692 vertebrate hosts.
- Discovering even one such interaction requires extensive wildlife surveillance and testing.

## 5 Predict Virus-Host Interactions



- Goal: Develop, implement, and test algorithms to predict links between viruses and hosts.
  - ▶ Read papers on link prediction, general as well as host-parasite networks.
  - ▶ Create BEELINE-style benchmark.
  - ▶ Implement cross-validation to avoid data leakage.
  - ▶ Design your own algorithm.

# Hardware Support for Projects

- Research virtual machines maintained by the Department of Computer Science.
- Obtain accounts on `bioinformatics.cs.vt.edu` from Rob Hunter (rhunter at vt dot edu).
- Can get accounts on ARC, if necessary.



# Ground Rules for Projects

- Send me project choices by Monday, February 20.
- I will schedule 1 hour meetings with each group every 2 weeks.
- Maintain Google docs describing your project and your progress.
- Preliminary project reports (motivation, background, related and previous research, approach, data, any preliminary results) due on Monday, March 27.
- Final project presentations on May 1 and May 3.
- Final project reports due on 5pm, Friday, May 5: 11pt font, 10 pages (not counting references), formatted like a journal paper.

# List of Projects

- 1 Develop PathLinker 2.0
- 2 Develop BEELINE 2.0
- 3 Predict cell types
- 4 Predict protein complexes
- 5 Predict virus-host interactions