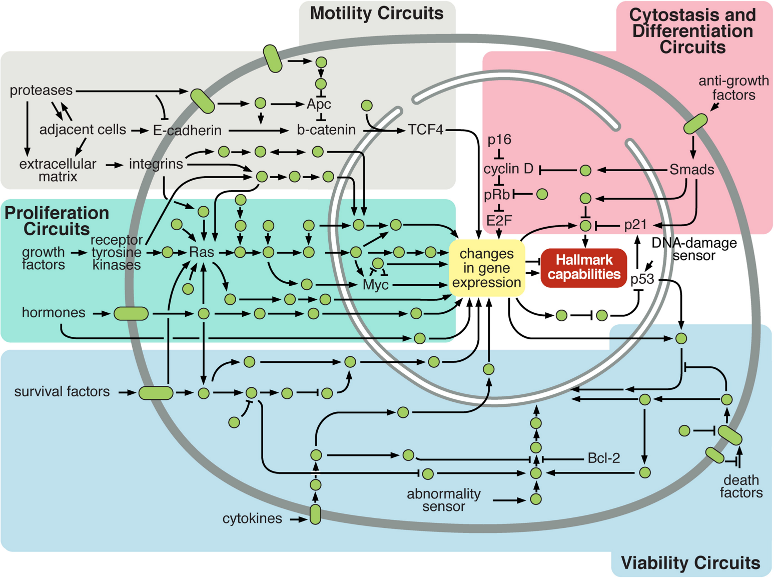
CS 5854: Supervised Inference of Gene Regulatory Networks from Single-Cell Gene Expression Data

T. M. Murali

February 16, 18, 2021

Signaling Pathways and Gene Expression

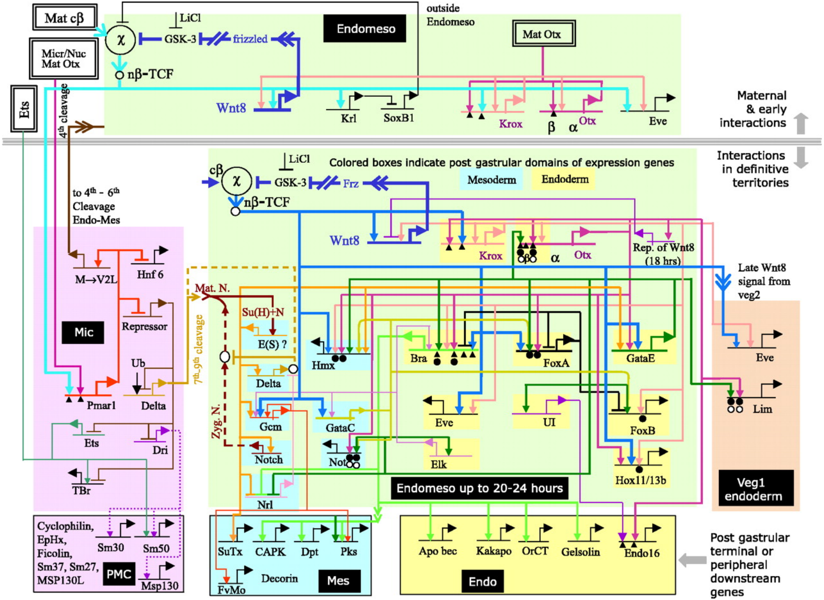


Gene Expression is a Dynamic Process

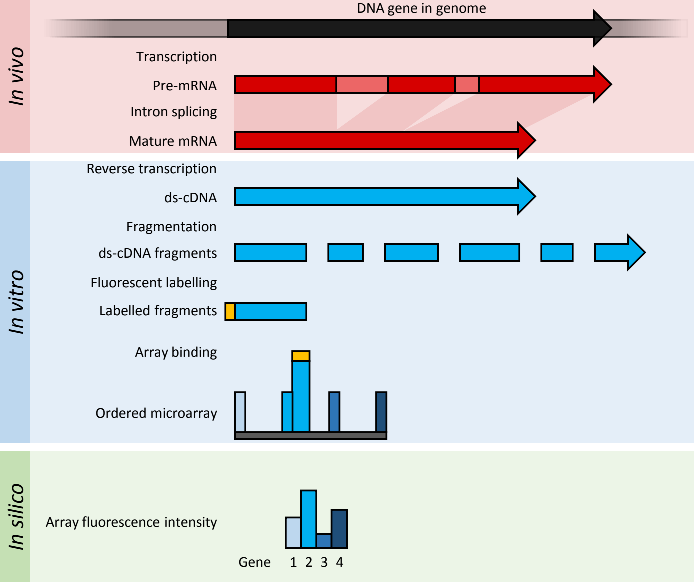
B

if ($F = 1$ or $E = 1$ or $CD = 1$) and ($Z = 1$)	Repression functions of modules F, E, and DC mediated by Z site
$\alpha = 1$	
else $\alpha = 0$	
if ($P = 1$ and $CG_1 = 1$)	Both P and CG_1 needed for synergistic link with module B
$\beta = 2$	
else $\beta = 0$	
if ($CG_2 = 1$ and $CG_3 = 1$ and $CG_4 = 1$)	Final step up of system output
$\gamma = 2$	
else $\gamma = 1$	
$\delta(t) = B(t) + G(t)$	Positive input from modules B and G
$\varepsilon(t) = \beta^*\delta(t)$	Synergistic amplification of module B output by CG_1-P subsystem
if ($\varepsilon(t) = 0$)	Switch determining whether Otx site in module A, or upstream modules (i.e., mainly module B), will control level of activity
$\xi(t) = Otx(t)$	
else $\xi(t) = \varepsilon(t)$	
if ($\alpha = 1$)	Repression function inoperative in endoderm but blocks activity elsewhere
$\eta(t) = 0$	
else $\eta(t) = \xi(t)$	
$\Theta(t) = \gamma^*\eta(t)$	Final output communicated to BTA

Gene Expression is a Dynamic Process

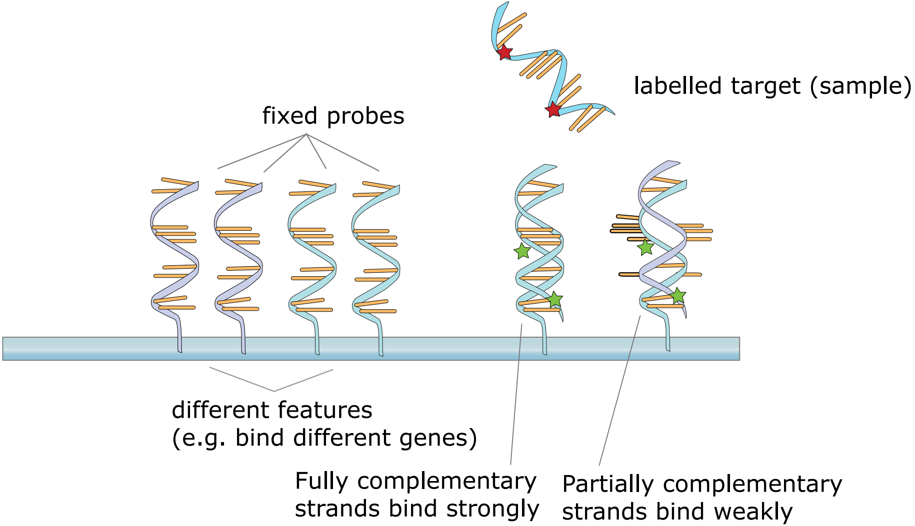


Measuring Genomewide Gene Expression: DNA Microarrays



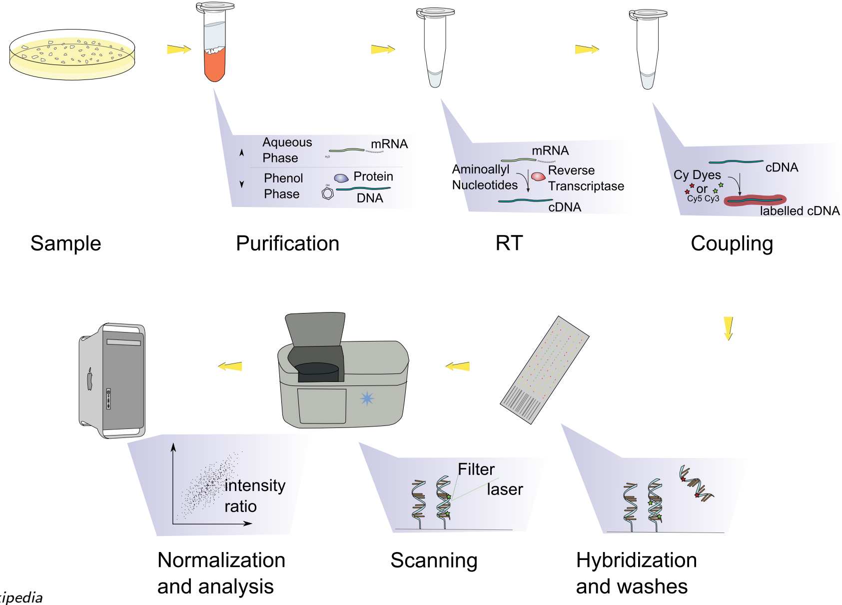
Wikipedia

Measuring Genomewide Gene Expression: DNA Microarrays



Wikipedia

Measuring Genomewide Gene Expression: DNA Microarrays



Wikipedia

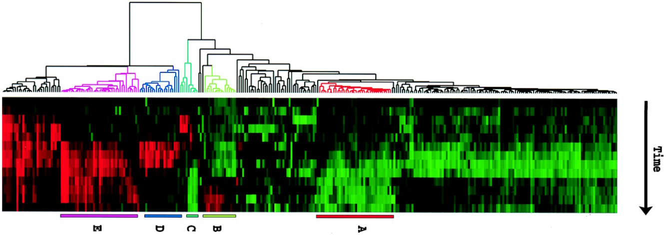
Applications of DNA Microarray Data

RESEARCH ARTICLE

Cluster analysis and display of genome-wide expression patterns

Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein

PNAS December 8, 1998 95 (25) 14863-14868;



Applications of DNA Microarray Data

RESEARCH ARTICLE

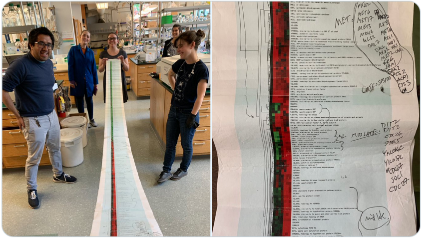
Cluster analysis and display of genome-wide expression patterns

Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein

PNAS De

Michael Eisen
@mbeisen

Inspired by @UCSDCooperLab's question about origins of the red/green color scheme in microarray clustering, I present THE FIRST dna microarray cluster analysis made by me in 1997 for ncbi.nlm.nih.gov/m/pubmed/97841... w/handwritten notes from Pat Brown and the late Ira Herskowitz.



6:27 PM · Jun 4, 2019 · Twitter for iPhone

Applications of DNA Microarray Data

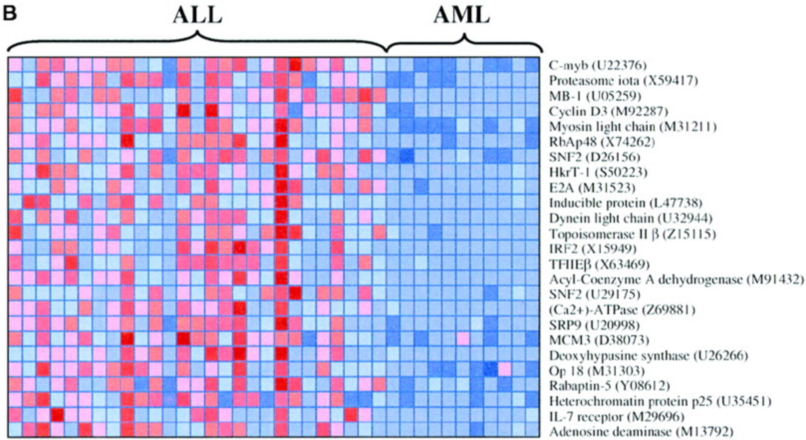
REPORT

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

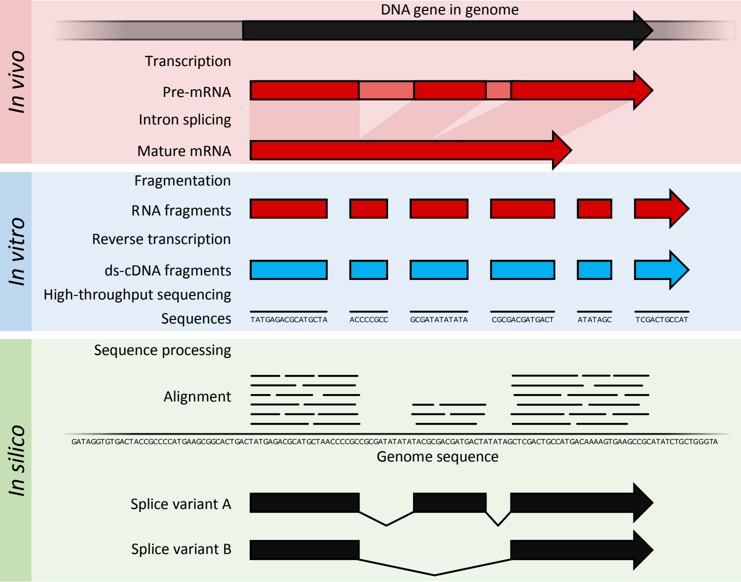
T. R. Golub[1,2,*,†], D. K. Slonim[1,†], P. Tamayo[1], C. Huard[1], M. Gaasenbeek[1], J. P. Mesirov[1], H. Coller[1], M. L. Loh[2], J. R. Downin...

+ See all authors and affiliations

Science 15 Oct 1999:
Vol. 286, Issue 5439, pp. 531-537
DOI: 10.1126/science.286.5439.531

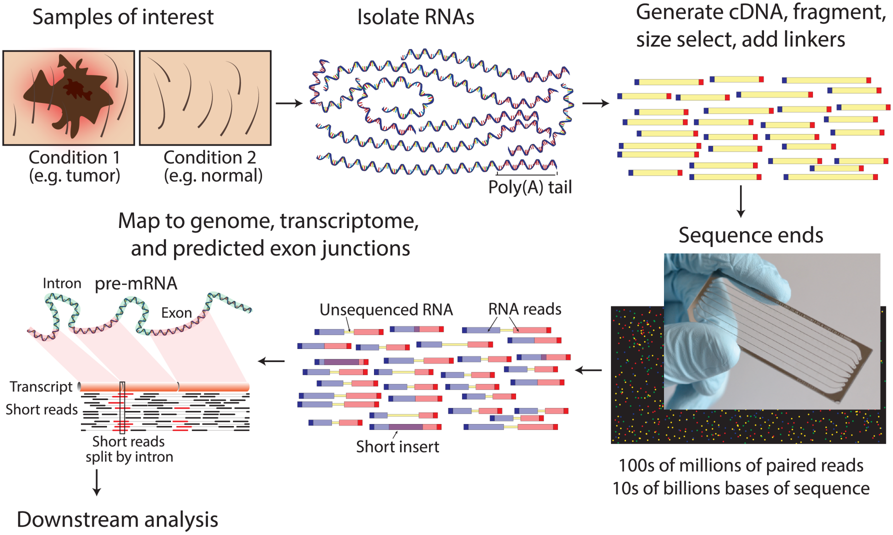


Measuring Genomewide Gene Expression: RNA-seq



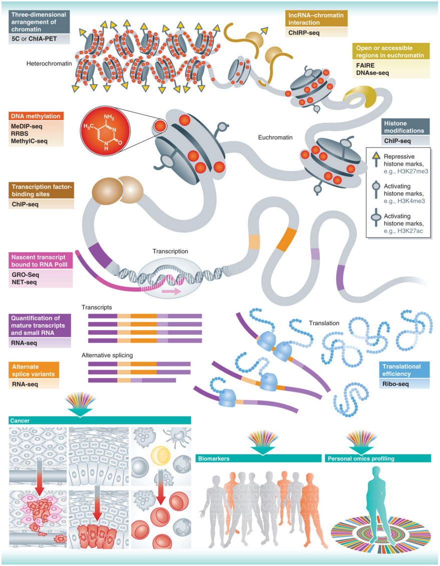
Wikipedia

Measuring Genomewide Gene Expression: RNA-seq



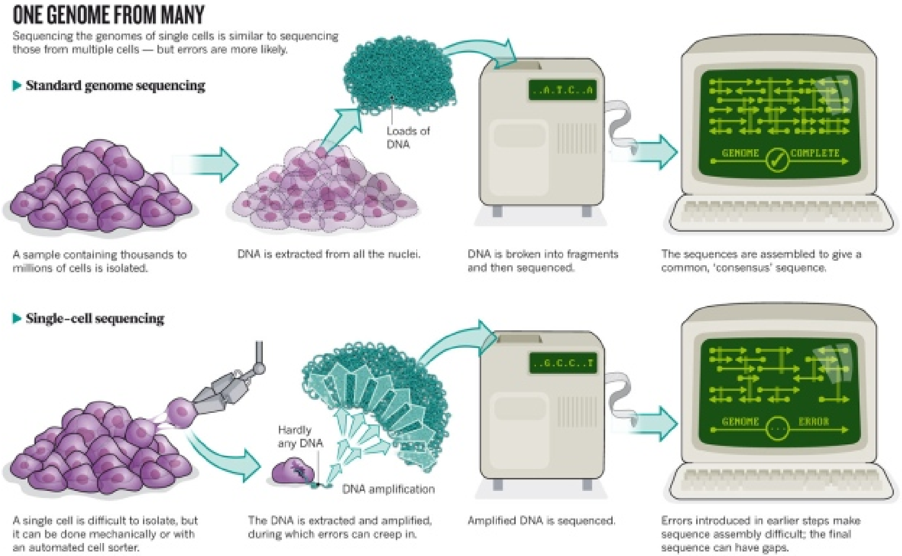
Griffith et al., Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. PLoS Comput Biol, 2015

*-Seq Techniques



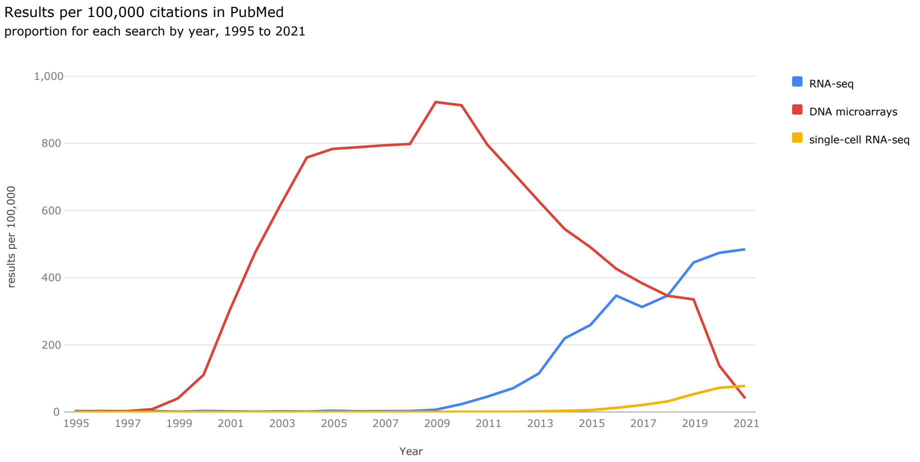
Soon, Hariharan, and Snyder. High-throughput sequencing for biology and medicine. Mol. Sys. Bio, 2013.

Single-Cell RNA-Seq

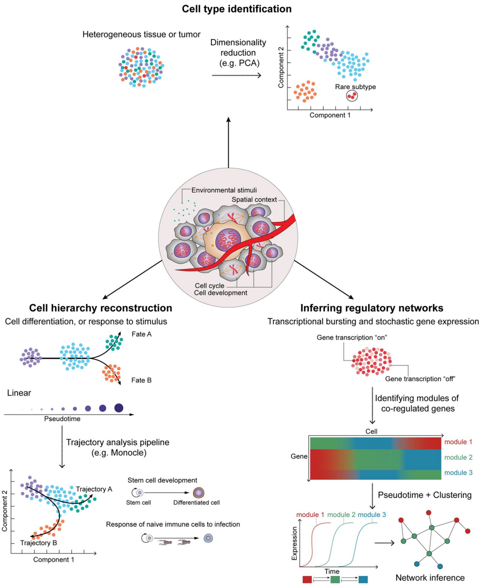


Owens, Genomics: The single life, Nature, 2012.

Technology Trends

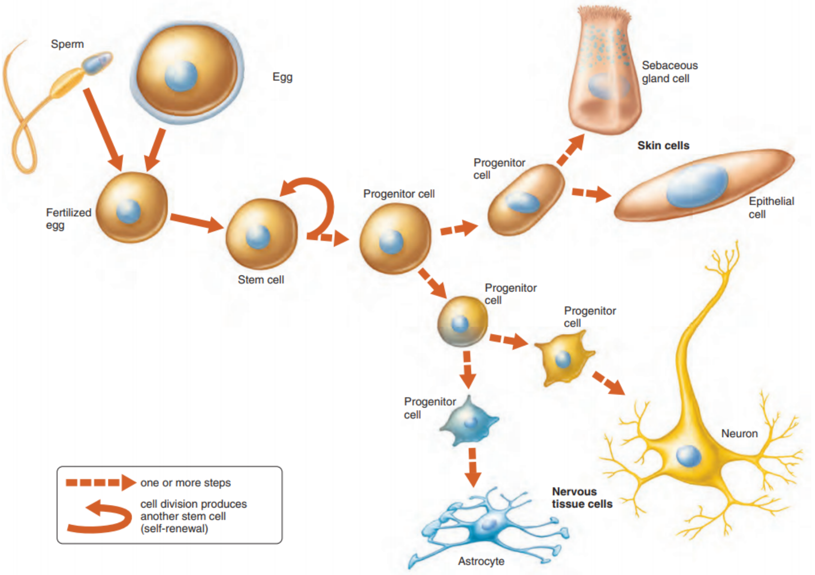


Applications of scRNA-seq Data



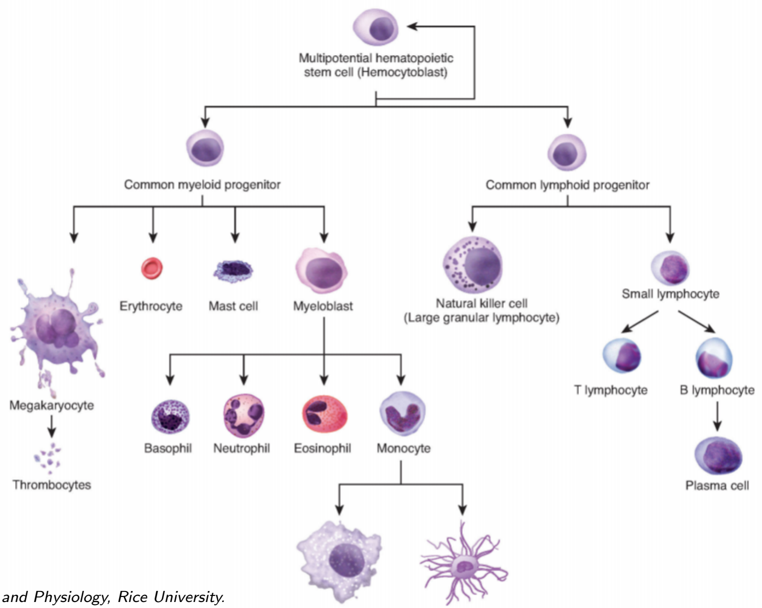
Hwang, Lee, and Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines, Exp. Mol. Med., 2018

Cellular Differentiation



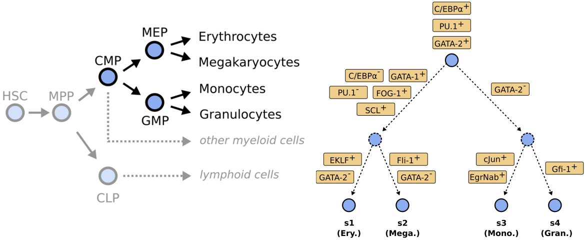
Shier, *et al.*, (2015) "*Hole's Essentials of Human Anatomy and Physiology*", McGraw-Hill

Cellular Differentiation



Anatomy and Physiology, Rice University.

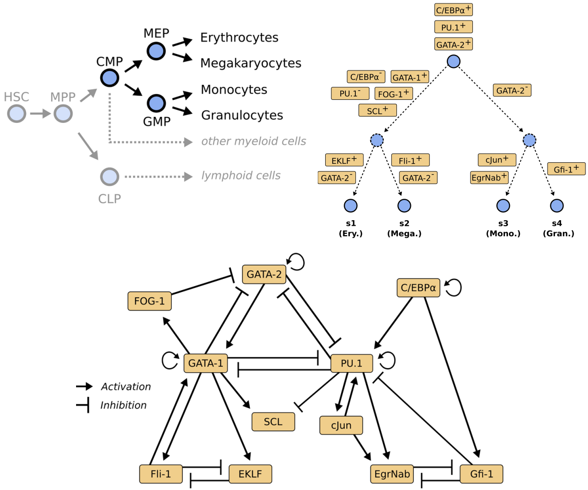
Cellular Differentiation



- 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

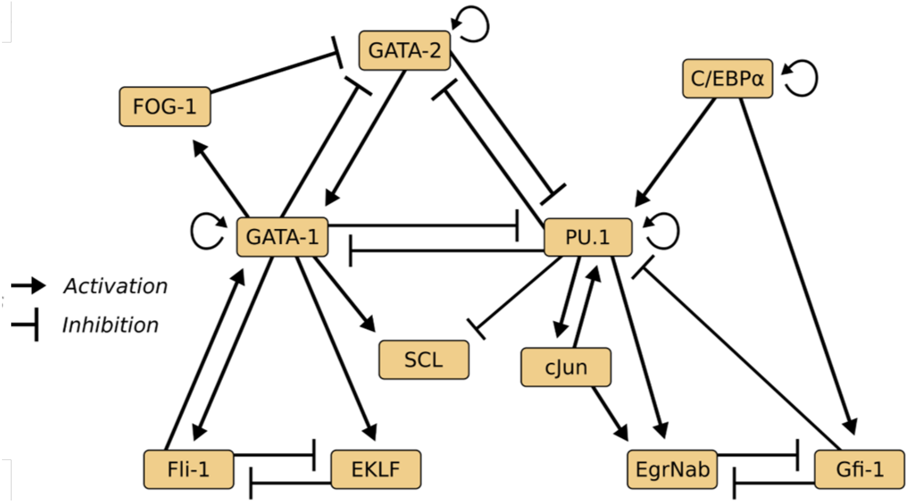
Cellular Differentiation



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

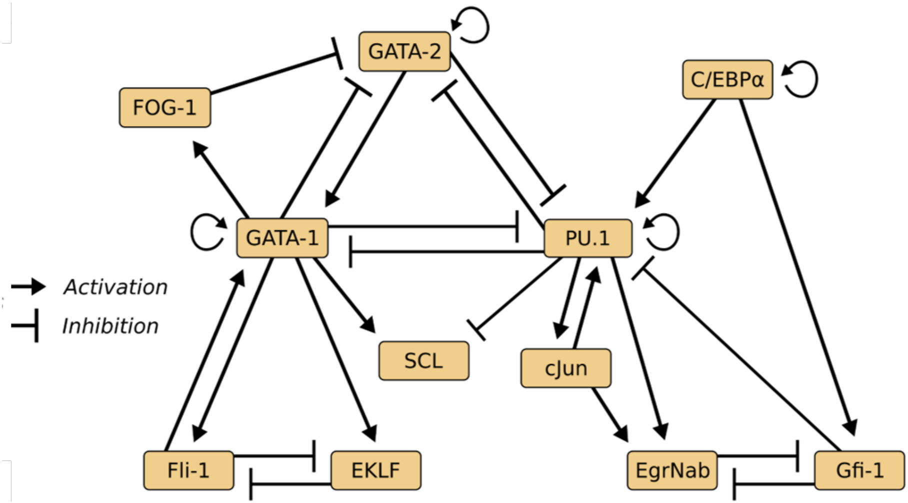
Krumsiek et al. (2010). "Hierarchical Differentiation of Myeloid Progenitors..." PLoS ONE

Gene Regulatory Network (GRN)



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

Gene Regulatory Network (GRN)



How do we build GRNs using computational techniques?

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

BEELINE Results for Human Datasets

TFs + 500 genes

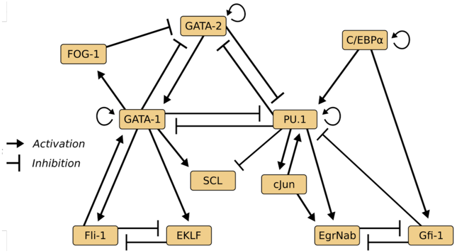
		Network Statistics			EPR					
		#TFs	#Genes	Density	PIDC	GENI	GRNB	SCODE	PCOR	SINC
STRING	hESC	343	517	0.02	3.8				1.1	
	hHep	409	656	0.03	3.5					1.3
Non-specific ChIP-Seq	hESC	283	760	0.02	1.9			1.0		
	hHep	322	832	0.02	2.5				1.0	
Cell-type specific ChIP-Seq	hESC	34	815	0.16				1.0	1.1	
	hHep	30	874	0.38	1.0			1.2		

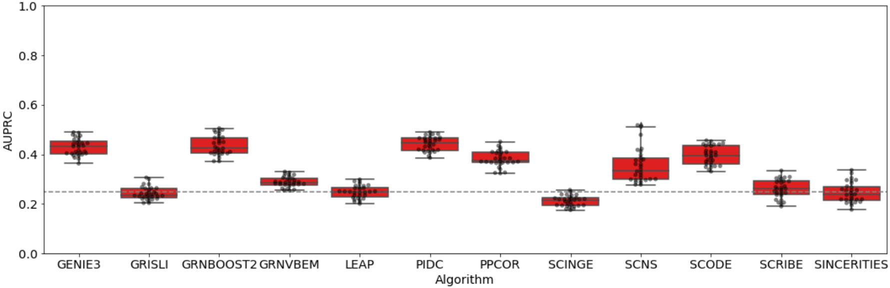
TFs + 1000 genes

		Network Statistics			EPR					
		#TFs	#Genes	Density	PIDC	GENI	GRNB	SCODE	PCOR	SINC
STRING	hESC	351	709	0.02	4.1			1.1		
	hHep	414	889	0.02	3.6		2.8			
Non-specific ChIP-Seq	hESC	292	1149	0.01					1.4	2.2
	hHep	332	1224	0.01				2.5	1.3	
Cell-type specific ChIP-Seq	hESC	34	1260	0.17					1.1	1.7
	hHep	31	1331	0.38	1.0	1.1				

Random Predictor Low/Poor High/Good

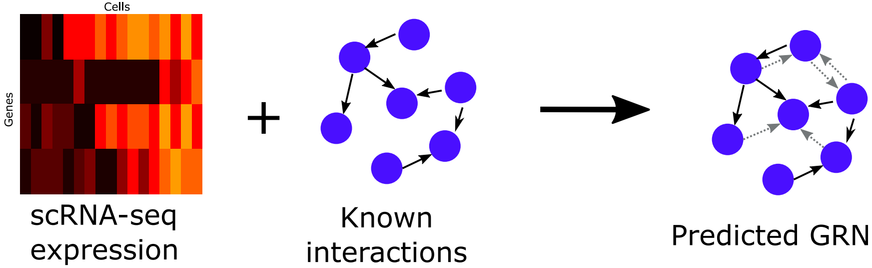
Poor AUPRC performance





Supervised GRN inference

Can supervised learning methods take advantage of known regulatory interactions for GRN inference from scRNA-seq data?



Existing approaches for supervised GRN

- Generate TF-gene features and build a classifier (e.g., SVM)
  - Concatenate expression vectors[1]
  - Outer product[2]
  - Kernels[3]

[1] *Cerulo et al. (2010) "Learning gene regulatory ..." BMC Bioinfo., 11(1):228*
[2] *Maetschke et al. (2014) "Supervised, semi- ..." Brief. Bioinfo., 15(2):195–211*
[3] *Cuong et al. (2008) "Supervised inference ..." BMC Bioinfo., 9(1):2*

Drawbacks

Drawbacks of fixed TF-gene feature representation:

1. Dropouts + noise in the input expression data
   - Dropout: where a gene is observed in one cell but is not detected in another cell of the same cell type
   - Unclear how fixed feature representation can overcome these problems

Drawbacks

Drawbacks of fixed TF-gene feature representation:

1. Dropouts + noise in the input expression data
   - Dropout: where a gene is observed in one cell but is not detected in another cell of the same cell type
   - Unclear how fixed feature representation can overcome these problems
2. Do not scale well for datasets with large number of cells

Proposed solutions

1. Dropouts + noise in the input expression data → Denoise and impute data using network propagation[1] [2] [3]

[1]Ronen *et al.* (2018) "netSmooth ..." *F1000 Res.* 7.
[2]Ye *et al.* (2019) "scNPF " *BMC Genomics* 20, 347.
[3]Elyanow *et al.* (2020) "netNMF-sc ..." *Gen Res* 30.2: 195-204.

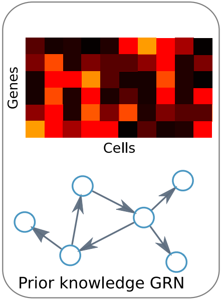
Proposed solutions

1. Dropouts + noise in the input expression data → Denoise and impute data using network propagation[1] [2] [3]
2. Do not scale well for datasets with large number of cells → Dimensionality reduction

[1] Ronen *et al.* (2018) "netSmooth ..." *F1000 Res.* 7.
[2] Ye *et al.* (2019) "scNPF " *BMC Genomics* 20, 347.
[3] Elyanow *et al.* (2020) "netNMF-sc ..." *Gen Res* 30.2: 195-204.

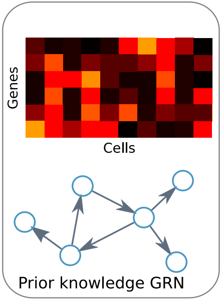
Denoising the input data

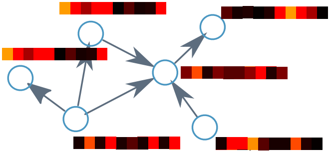
Inputs



Denoising the input data

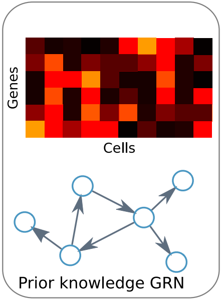
Inputs

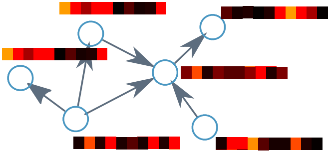




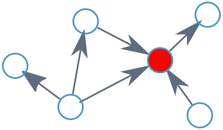
Denoising the input data

Inputs



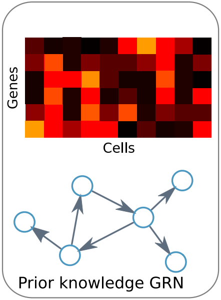


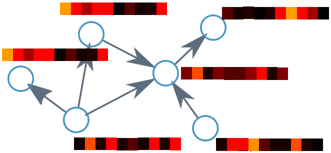
Update expression vector for gene in red



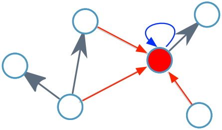
Denoising the input data

Inputs



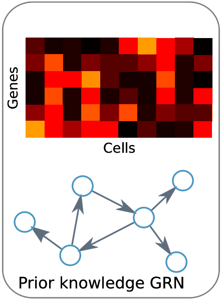


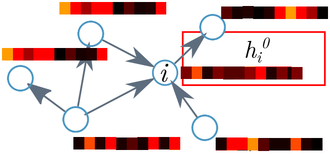
Update expression vector for gene in red



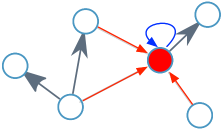
Denoising the input data

Inputs



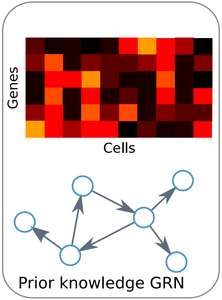


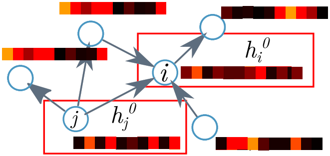
Update expression vector for gene in red



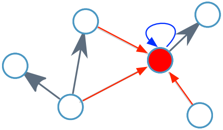
Denoising the input data

Inputs



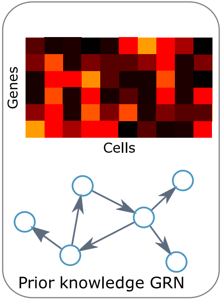


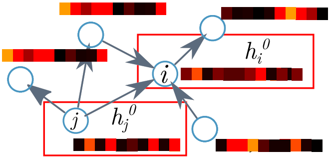
Update expression vector for gene in red



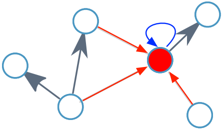
Denoising the input data

Inputs





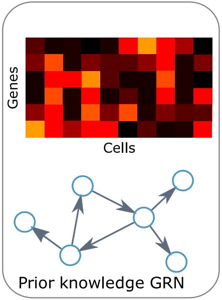
Update expression vector for gene in red

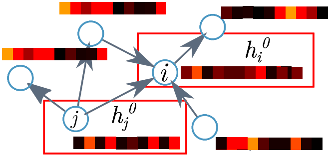


Update: $h_i = h_i^0 w_0 + \sum_{j \in N_i} h_j^0 w_1$

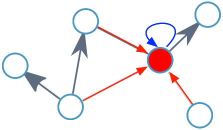
Denoising the input data

Inputs





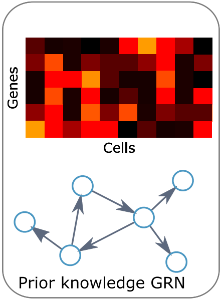
Update expression vector for gene in red

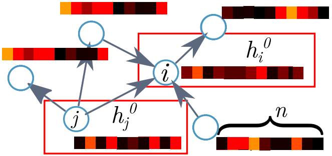


Update: $h_i^1 = h_i^0 W_0 + \sum_{j \in N_i} h_j^0 W_1$

Reducing dimensions of the input data

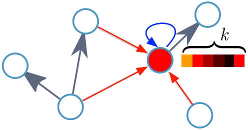
Inputs





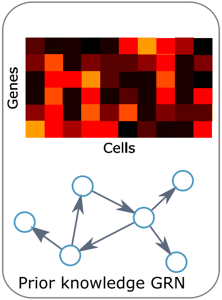
Update: $h_i^1 = h_i^0 W_0 + \sum_{j \in N_i} h_j^0 W_1$

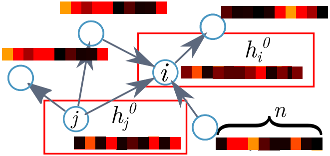
Update expression vector for gene in red



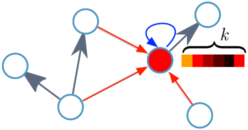
$\boxed{|W| = n \times k, k < n}$

Inputs





Update expression vector for gene in red



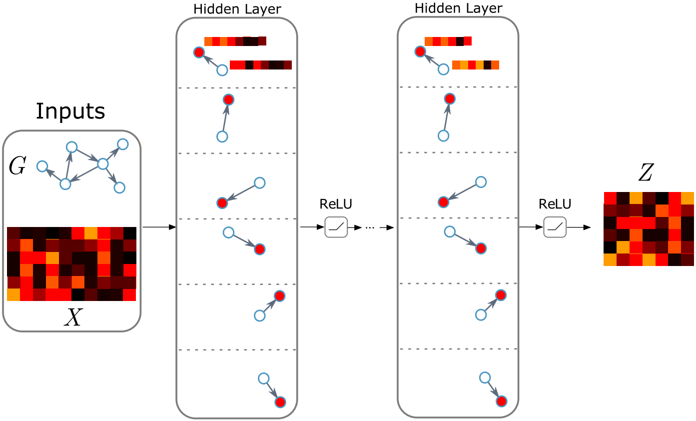
Update: $h_i^1 = h_i^0 W_0 + \sum_{j \in N_i} h_j^0 W_1$

$\boxed{|W| = n \times k, k < n}$

Graph convolutional networks (GCNs)

Kipf et al. (2016) "Semi-supervised classification with graph ...", CoRR, 1609.02907.

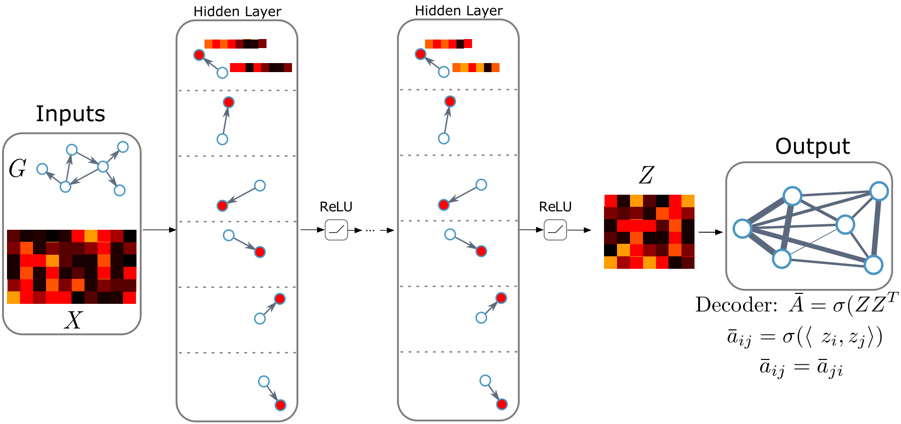
GCN-based Autoencoders



$$\text{Encoder: } Z = DGCN(X, A)$$

Kipf et al. (2016) "Semi-supervised classification with graph ...", CoRR, 1609.02907.

GCN-based Autoencoders



Decoder: $\bar{A} = \sigma(ZZ^T)$

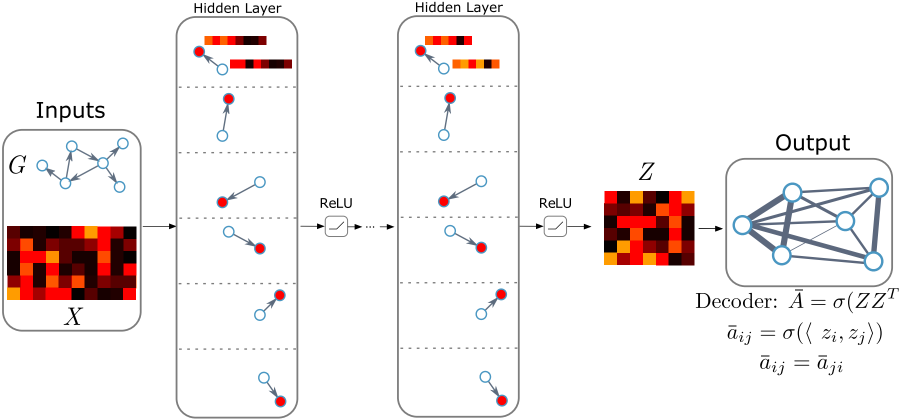
$\bar{a}_{ij} = \sigma(\langle z_i, z_j \rangle)$

$\bar{a}_{ij} = \bar{a}_{ji}$

Encoder: $Z = DGCN(X, A)$

Kipf et al. (2016) "Semi-supervised classification with graph ...", CoRR, 1609.02907.

GCN-based Autoencoders



Decoder: $\bar{A} = \sigma(ZZ^T)$

$\bar{a}_{ij} = \sigma(\langle z_i, z_j \rangle)$

$\bar{a}_{ij} = \bar{a}_{ji}$

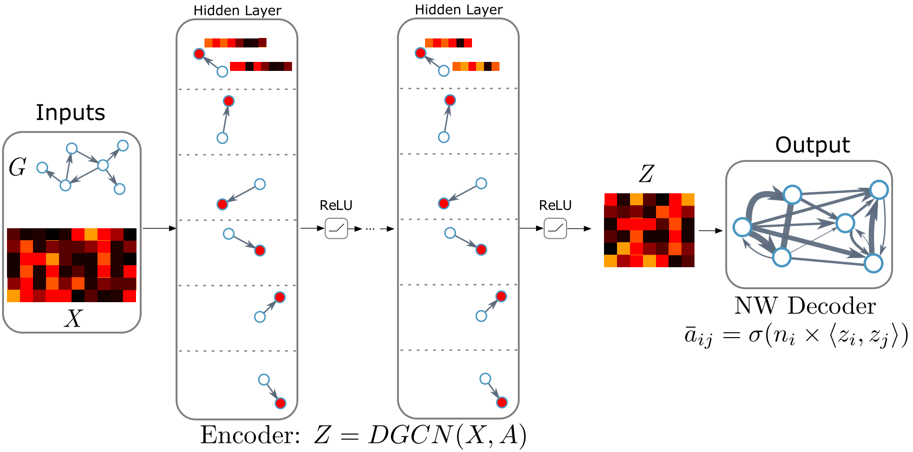
Encoder: $Z = DGCN(X, A)$

Learning objective: minimize the cross-entropy loss between A and $\bar{A}$

$$\mathcal{L} = \frac{1}{|E|}\left(- \sum_{(i,j)\in E} \log \bar{a}_{ij} - \sum_{(p,q)\in \bar{E}} \log(1 - \bar{a}_{pq}) \right)$$

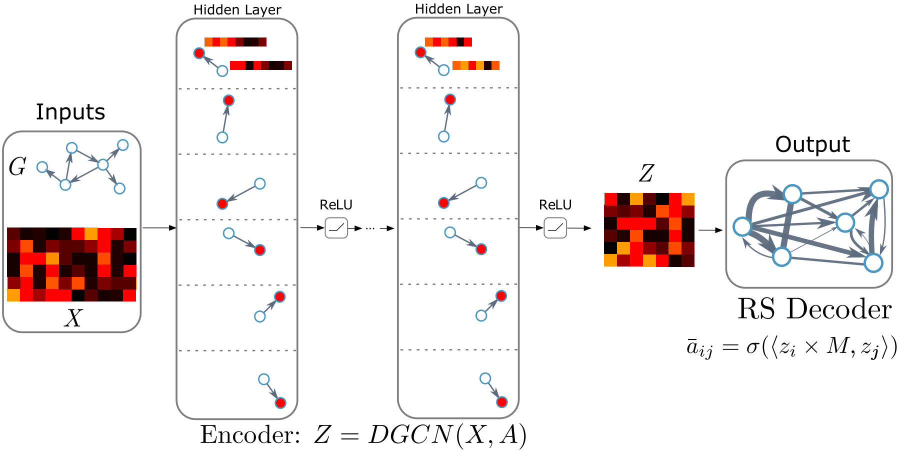
Kipf et al. (2016) "Semi-supervised classification with graph ...", CoRR, 1609.02907.

Directed GCN + Node Weight Decoder (NW)



n_i: Learned node weight for node i

Directed GCN + RESCAL Decoder (RS)[1]



M: Learned weight matrix

[1] Nickel *et al.* (2012) "Factorizing YAGO... ", *In Proc. WWW*, pp. 271–280

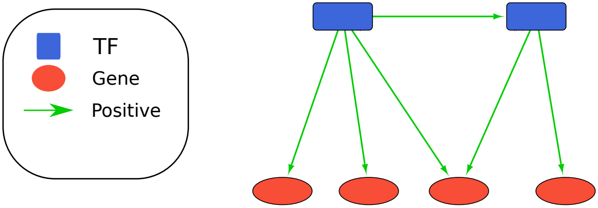
Variants of GCN-based Autoencoders

Six encoder-decoder combinations:

- Encoders
  - Undirected GCN-based encoder (**GCN**)
  - Directed GCN-based encoder (**DGCN**)
- Decoders
  - Inner Product decoder (**IP**)
  - Node Weight decoder (**NW**)
  - RESCAL Decoder (**RS**)

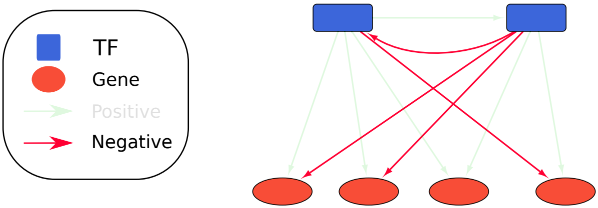
Evaluation

- *k*-fold cross validation: Edges, TFs
- Positive edges: TF-gene edges present in the input GRN $G = (V, E)$



Evaluation

- k-fold cross validation: Edges, TFs
- Positive edges: TF-gene edges present in the input GRN $G = (V, E)$
- Negative edges: TF-gene edges not in G



10-Fold Edge Holdout Cross-Validation

- Randomly partition positive edges into 10 sets
- Holdout one set of edges as testing positives
- Use the remaining edges as training positives
- Sample uniformly at random as many training (testing) negatives as there are training (testing) positives

10-Fold TF Holdout Cross-Validation

- Randomly partition TF nodes in the GRN into 10 sets
- Holdout one set of TFs and all the edges adjacent to them in the GRN as testing positives
- Use the remaining edges in the GRN as training positives
- How do we sample negatives?

10-Fold TF Holdout Cross-Validation

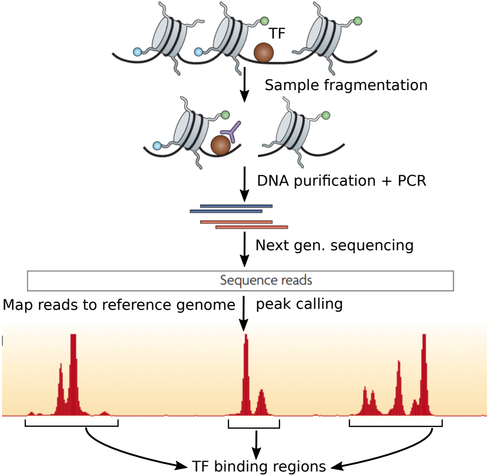
- Randomly partition TF nodes in the GRN into 10 sets
- Holdout one set of TFs and all the edges adjacent to them in the GRN as testing positives
- Use the remaining edges in the GRN as training positives
- How do we sample negatives? For each TF, we randomly sample as many negatives as there are positives adjacent to that TF, once for the set training and once for the testing set.

Ground-Truth Networks

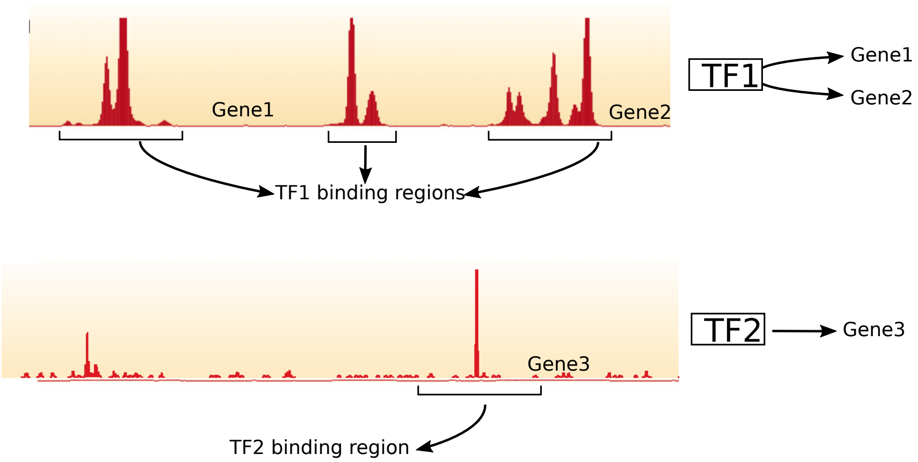
- Cell-type specific ChIP-seq network
- Non-specific ChIP-seq network

ChIP-seq

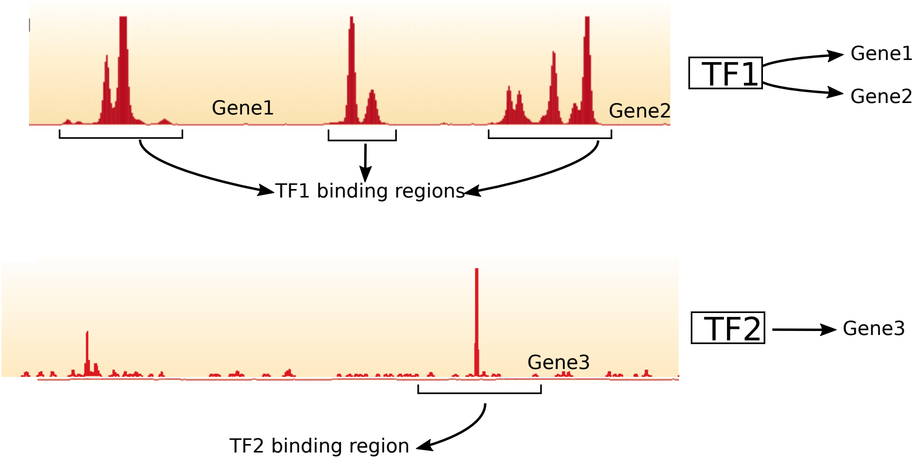
Chromatin: any protein interacting with DNA, e.g., TF
Immuno**P**recipitation: enrichment of DNA bound to the protein of interest



ChIP-seq


Gene1
Gene2
TF1 binding regions
TF1
Gene1
Gene2
Gene3
TF2 binding region
TF2
Gene3

Cell-type specific ChIP-seq network



Very few TFs tested for each cell-type

Non-specific ChIP-seq network

- Collected curated TF-gene interactions from
  1. RegNetwork[1]
  2. TRRUST[2]
  3. DoRothEA[3]

1. *Liu et al. (2015) "RegNetwork: an integrated ..." Database, 2015*
2. *Han et al. (2018) "TRRUSTv2 ..." NAR, 46(D1):D380–D386*
3. *Garcia-Alonso et al. (2019) "Benchmark ..." Gen. Res., 29:1363–1375*

Gene Expression Datasets

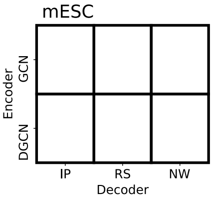
Name	#Cells	#Nodes	#Edges	# TFs
mESC[1]	471	896	6,893	516
mHSC[2]	3,175	4,158	17,309	445
mMac[3]	6,283	7,428	35,347	747
hESC[4]	758	1,142	4,597	292

[1] Hayashi *et al.* (2018) "Single-cell full-length..." *Nat Comm*, 9, 619
[2] Nestorowa *et al.* (2016) "A Single-Cell Resolution..." *Blood*, 128(8):e20-31
[3] Alavi *et al.* (2018) "A web server for..." *Nat Comm*, 9, 4768
[4] Chu *et al.* "Single-cell RNA-seq reveals ... *Genome Biology*, 17(1), 173

Best GCN-based autoencoder architecture

- Median test early precision from 10-fold evaluations

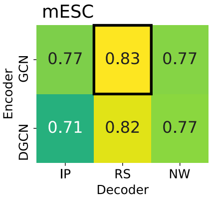
(a) 10-fold edge CV



Best GCN-based autoencoder architecture

- Median test early precision from 10-fold evaluations
- **GCN-RS** performs the best for 10-fold edge cross-validation

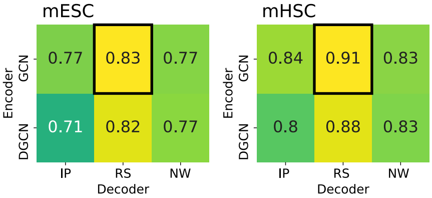
(a) 10-fold edge CV



Best GCN-based autoencoder architecture

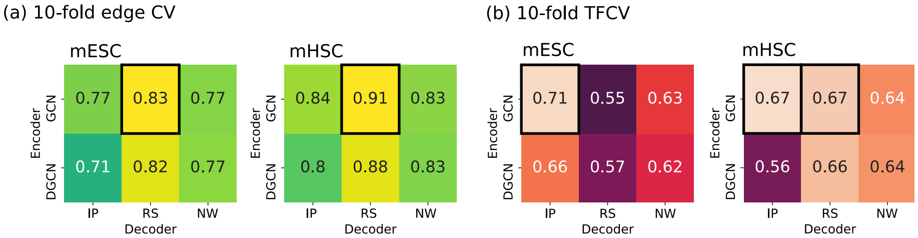
- Median test early precision from 10-fold evaluations
- **GCN-RS** performs the best for 10-fold edge cross-validation

(a) 10-fold edge CV



Best GCN-based autoencoder architecture

- Median test early precision from 10-fold evaluations
- **GCN-RS** performs the best for 10-fold edge cross-validation
- **GCN-IP** performs the best for 10-fold TF cross-validation



Methods evaluated

- For 10-fold edge CV: **GCN-RS** autoencoder
- For 10-fold TF CV: **GCN-IP** autoencoder

[1]Yuan (2020) "Deep learning for inferring ..." *PNAS*, 116 (52) 27151-27158

Methods evaluated

- For 10-fold edge CV: **GCN-RS** autoencoder
- For 10-fold TF CV: **GCN-IP** autoencoder
- **CNNC**[1]: CNN-based method that uses normalized empirical probability function (NEPDF) as features for every pair of genes
- **MLP-C**: a multi-layer perceptron with as many hidden layers as in the GCN and with concatenated expression vectors as input features
- **SVM-C**: Linear SVM with concatenated expression vectors as input features

[1] Yuan (2020) "Deep learning for inferring . . ." *PNAS*, 116 (52) 27151-27158

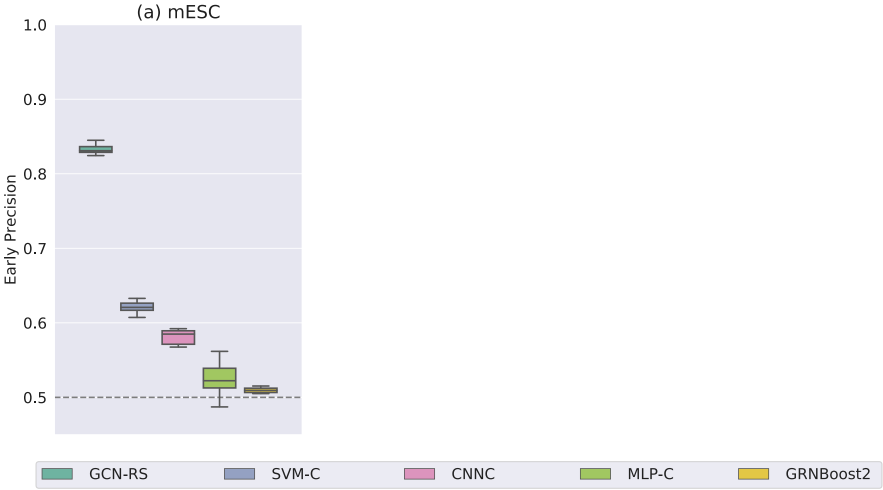
Methods evaluated

- For 10-fold edge CV: **GCN-RS** autoencoder
- For 10-fold TF CV: **GCN-IP** autoencoder
- **CNNC**[1]: CNN-based method that uses normalized empirical probability function (NEPDF) as features for every pair of genes
- **MLP-C**: a multi-layer perceptron with as many hidden layers as in the GCN and with concatenated expression vectors as input features
- **SVM-C**: Linear SVM with concatenated expression vectors as input features
- **GRNBoost2**: One of the top performing unsupervised learning methods from BEELINE (baseline)

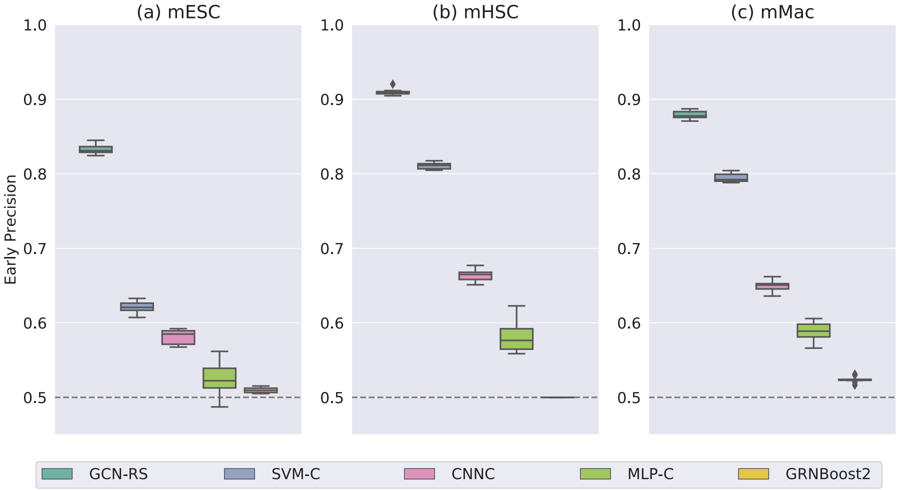
[1] Yuan (2020) "Deep learning for inferring ..." *PNAS*, 116 (52) 27151-27158

Evalutation: 10-fold edge cross-validation

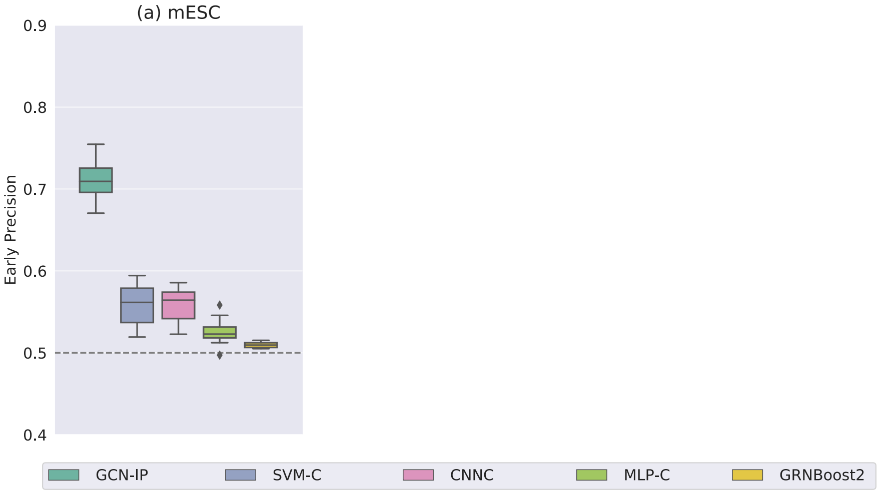
Evalutation: 10-fold edge cross-validation



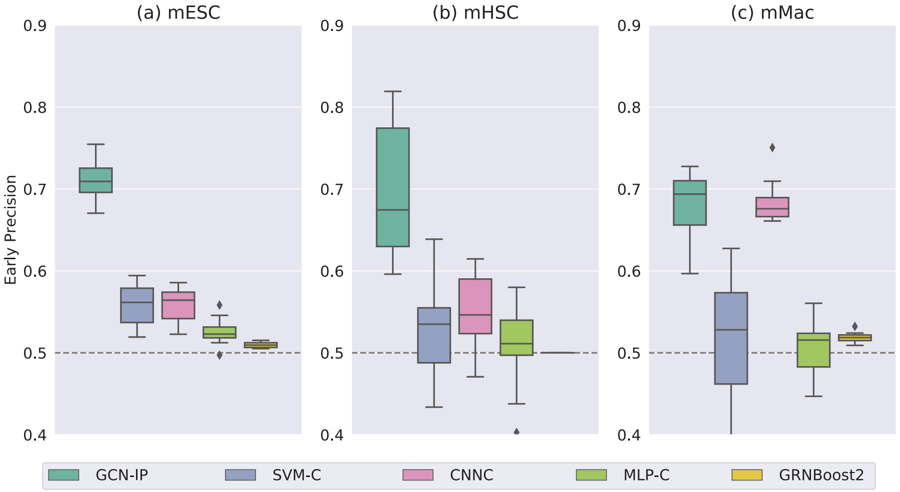
Evalutation: 10-fold edge cross-validation



Evalutation: 10-fold TF-holdout cross-validation

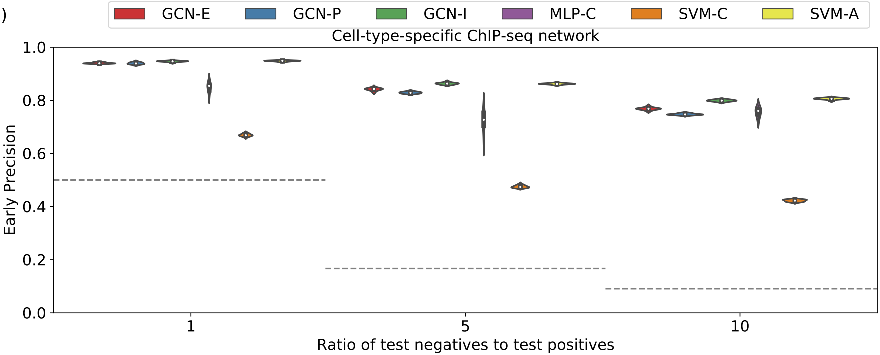


Evalutation: 10-fold TF-holdout cross-validation



Evaluation: mESC ChIP-seq network

(c)



Case Study: hESC scRNA-seq dataset [3]

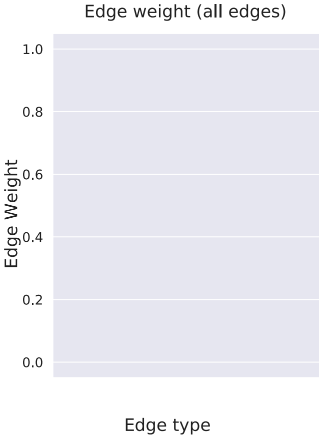
- Human embryonic stem cell dataset (hESC)

Dataset	#Cells	#Nodes	#Edges	# TFs
hESC	758	1,142	4,597	292

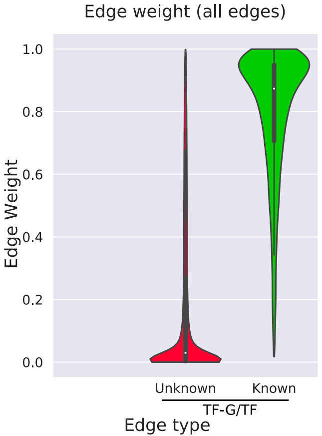
- Ground-truth network: Non-cell-type specific ChIP-seq network[1] [2]
- Training set-up:
  - **GCN-RS-E**
  - Positives: Edges in the human non-specific ChIP-seq network
  - Negatives: All possible TF-gene edges that not in the human non-specific ChIP-seq network

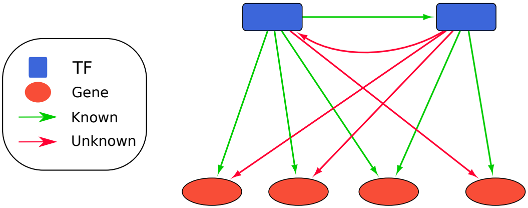
[1] *Liu et al. (2015) "RegNetwork: an integrated ..." Database, 2015*
[2] *Han et al. (2108) "TRRUSTv2 ..." Nucleic Acids Res., 46(D1):D380–D386*
[3] hesc-single-cell-genbio-december-2016

Predicted network: Edge weight distribution

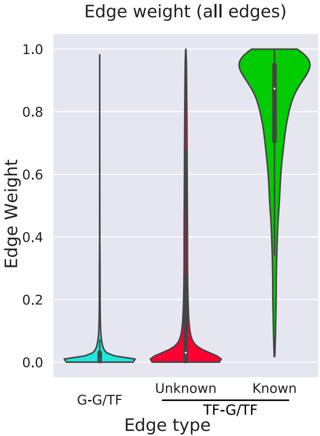


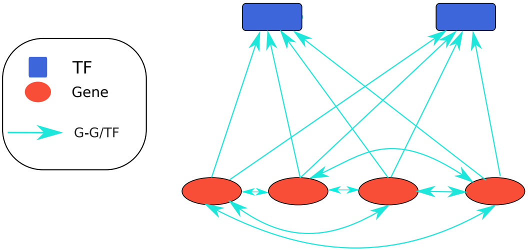
Predicted network: Edge weight distribution



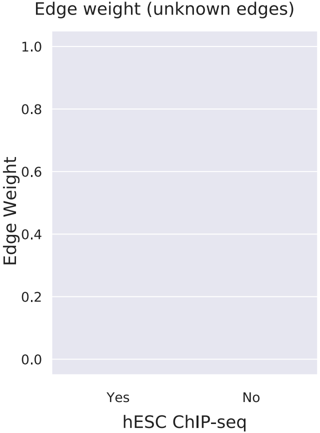


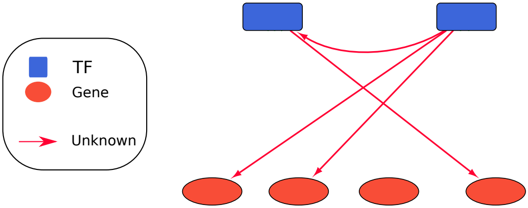
Predicted network: Edge weight distribution



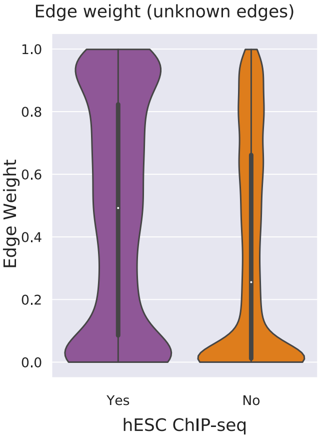


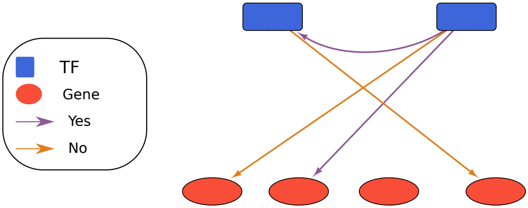
Unknown Edges: hESC cell-type specific network





Unknown Edges: hESC cell-type specific network





Summary

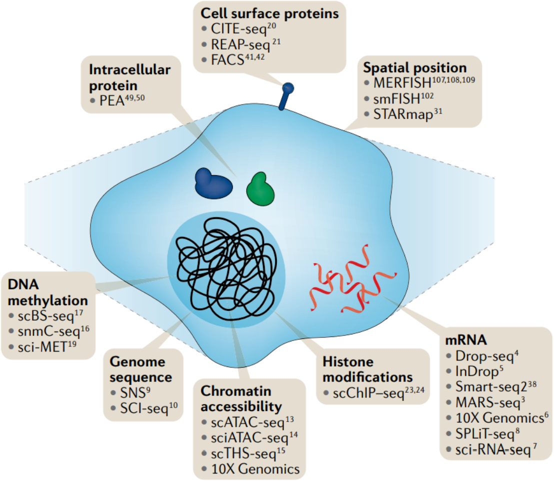
- GCN-based autoencoders are useful for denoising and reducing dimensions
- We use GCN-based autoencoders to train model for supervised GRN inference

Summary

- GCN-based autoencoders are useful for denoising and reducing dimensions
- We use GCN-based autoencoders to train model for supervised GRN inference
- GCN-autoencoder outperforms other methods for supervised GRN inference
- Can identify cell-type specific regulatory interactions even when trained on non-cell type specific GRN

Future Research

- Integrative single-cell analysis



Stuart et al..(2019) "Integrative single-cell..." Nat Rev Genet 20, 257–272.