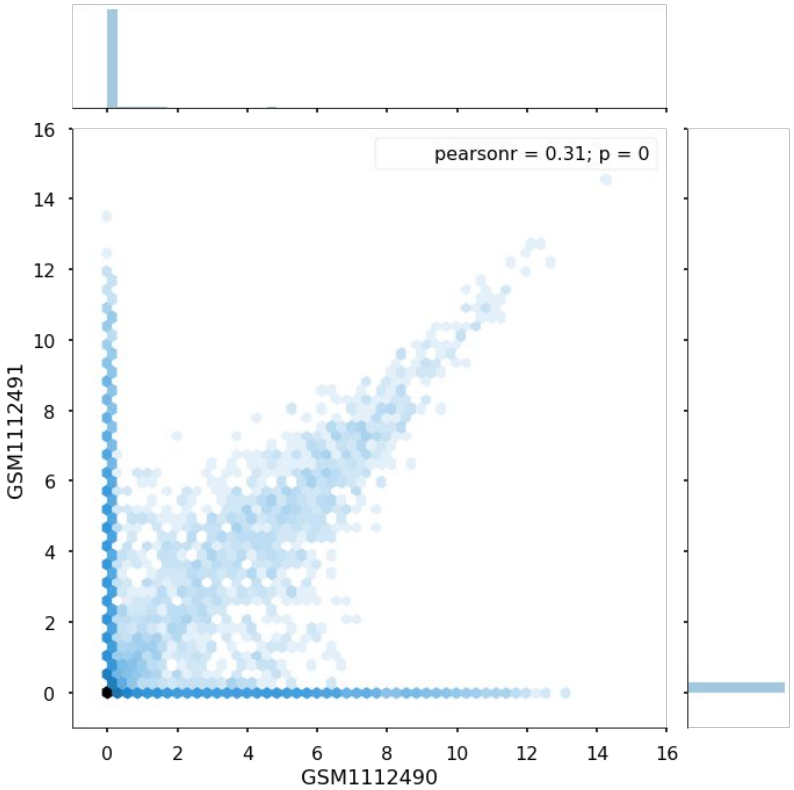
Single Cell RNA seq

Imputation methods

Feb 20 2018, MDC Berlin

The “drop-out” problem

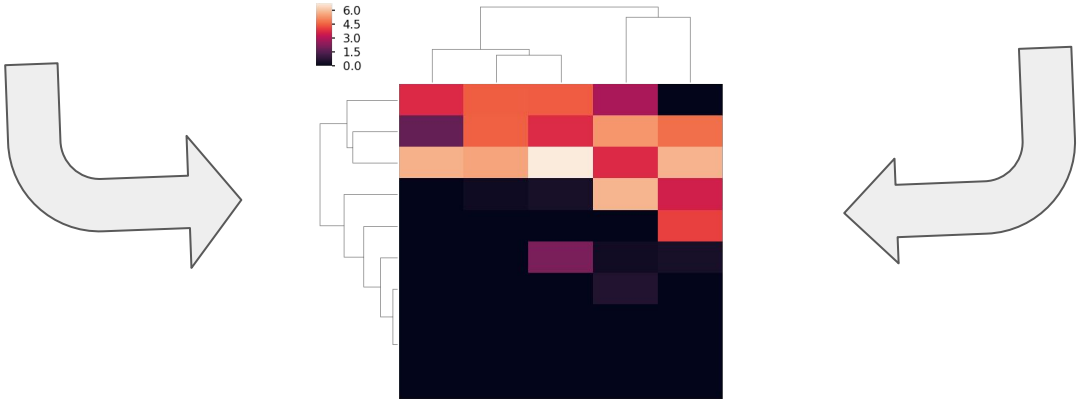


- “drop-out” = a gene with 0 counts which we’re not sure is true
- Rate varies with
  - Cell
  - Gene
  - Technology / platform

The high variance problem

Expression variability has many causes

- Technical variation
  - Batch effect
  - Library quality
  - Cell-specific capture efficiency
  - Amplification bias
- Biological variation
  - “Bursty” / stochastic transcription
  - Varying rate of RNA processing
  - Cell identity
  - Temporal progression / oscillation



Drop-outs and high variance complicate analysis

Whether they are biological or technical

Session outline

- Describe three algorithms that deal with these issues

- MAGIC
- scImpute
- netSmooth

MAGIC (Markov Affinity-based Graph Imputation of Cells)

Assumptions:

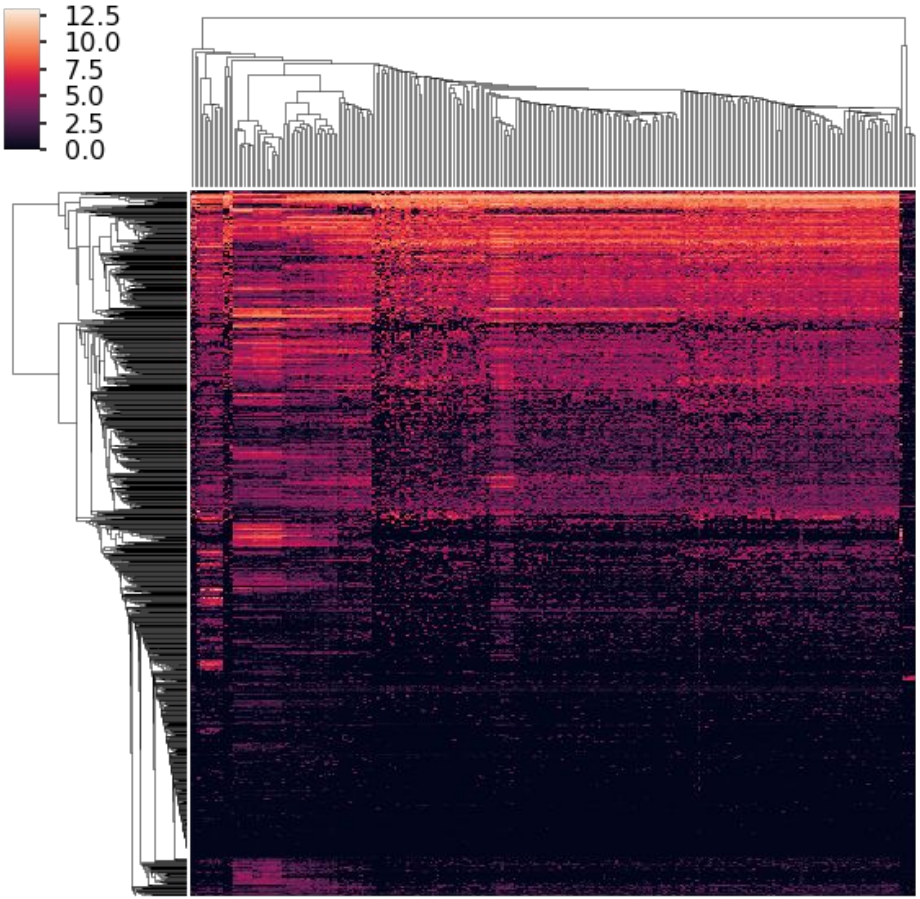
- Gene expression is highly collinear (many genes co-regulated)
- Dropout makes expression sparse - but we can impute based on similar cells
  - Similarity is distance on manifold

MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data

David van Dijk, Juozas Nainys, Roshan Sharma, Pooja Kathail, Ambrose J Carr, Kevin R Moon, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, Dana Pe'er

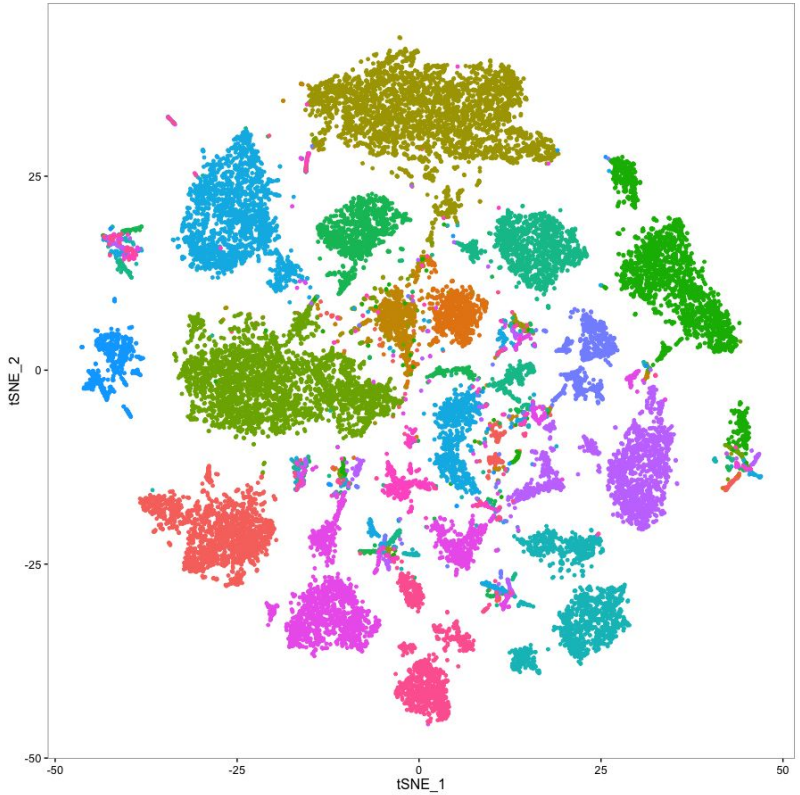
bioRxiv 111591; doi: https://doi.org/10.1101/111591

Collinearity



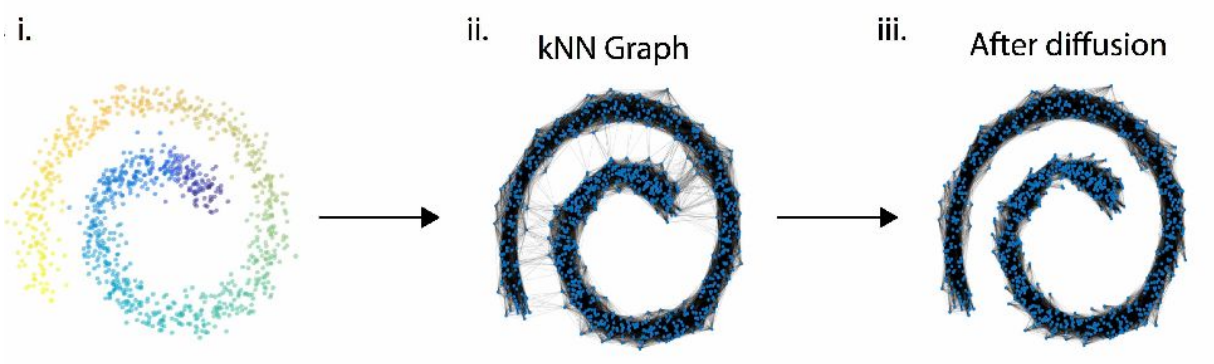
Distance on a manifold

Distance on a manifold



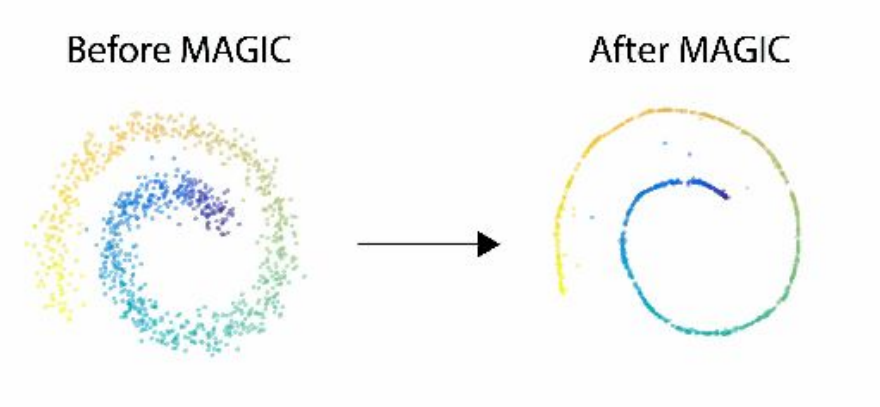
MAGIC - the algorithm

1. Finding the manifold



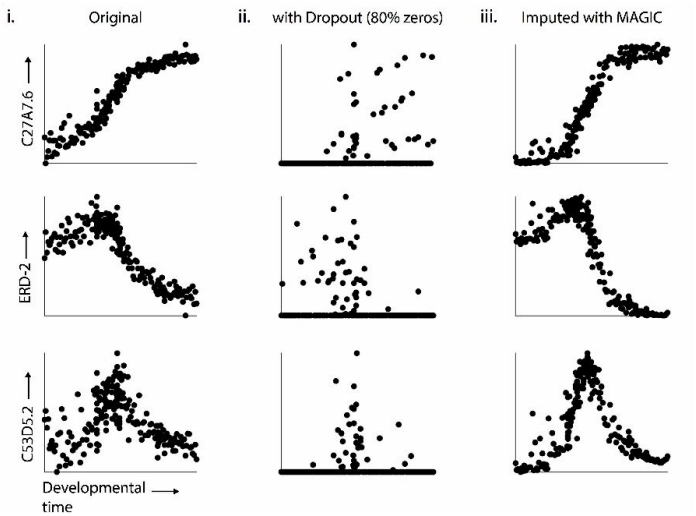
MAGIC - the algorithm

1. Finding the manifold
2. Local averaging



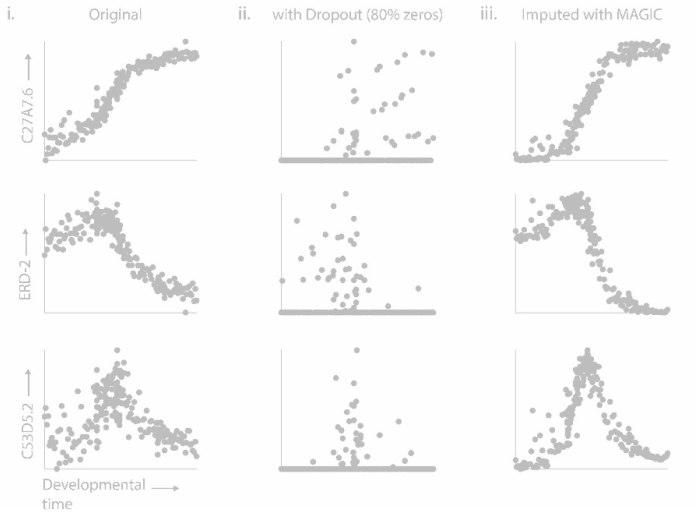
MAGIC - when does it work?

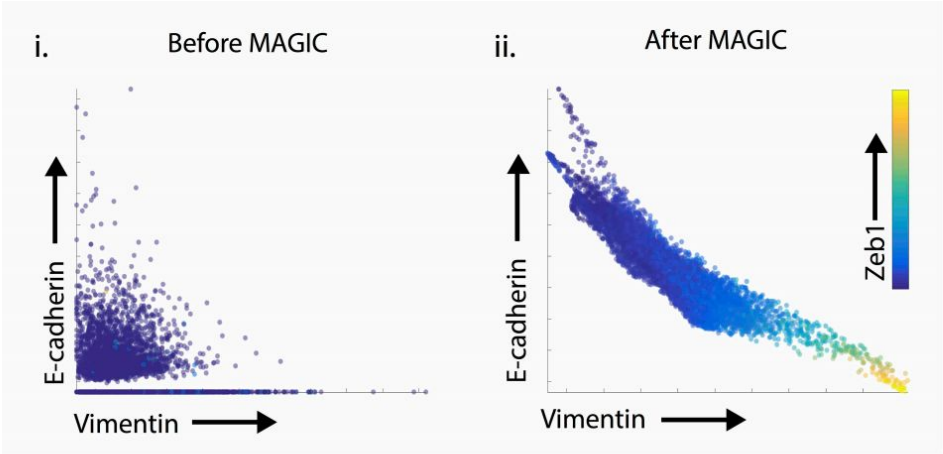
- Good for temporal relationships



MAGIC - when does it work?

- Good for temporal relationships
- Good for picking up gene-gene relationships





MAGIC - weaknesses

- Imputes non-dropouts (is this a strength or a weakness?)
- Requires large number of cells (paper says >7,000)
- Cluster structure

MAGIC - how to use

https://github.com/KrishnaswamyLab/magic/blob/develop/notebooks/Magic_single_cell_RNAseq.ipynb

scImpute

Assumptions:

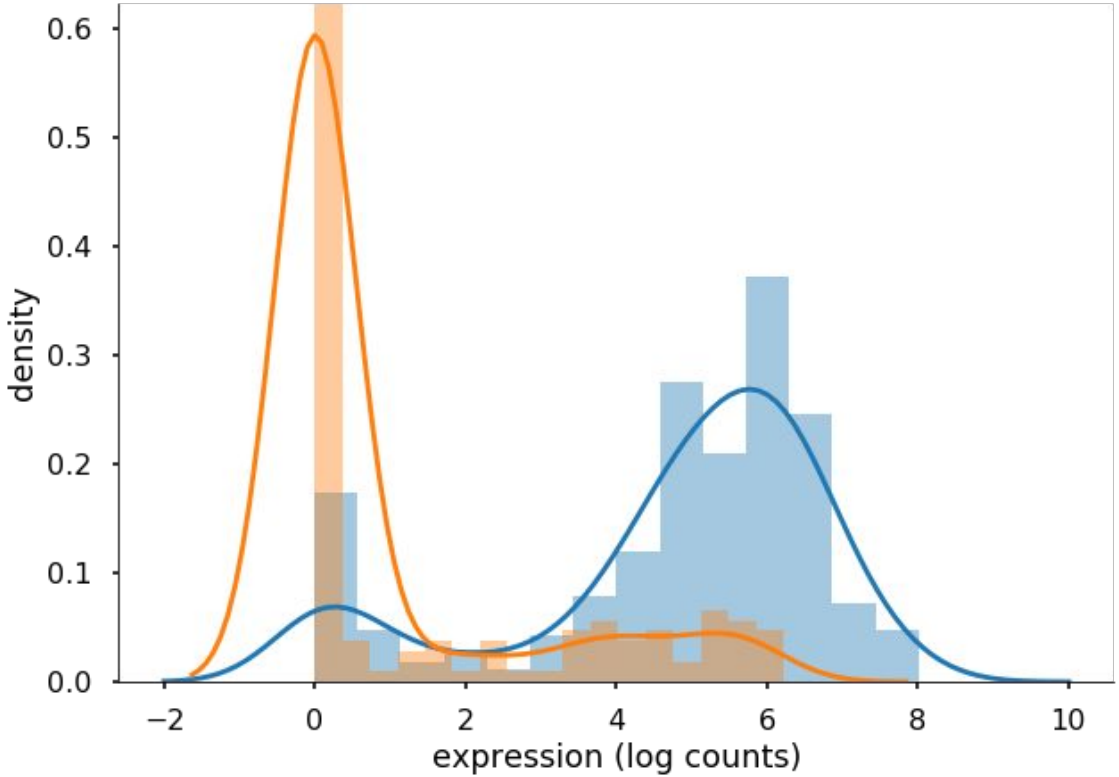
- Genes have their own drop-out likelihoods
- Cells are differently affected by drop-outs
- Non drop-outs have more information than drop-outs

scImpute: Accurate And Robust Imputation For Single Cell RNA-Seq Data

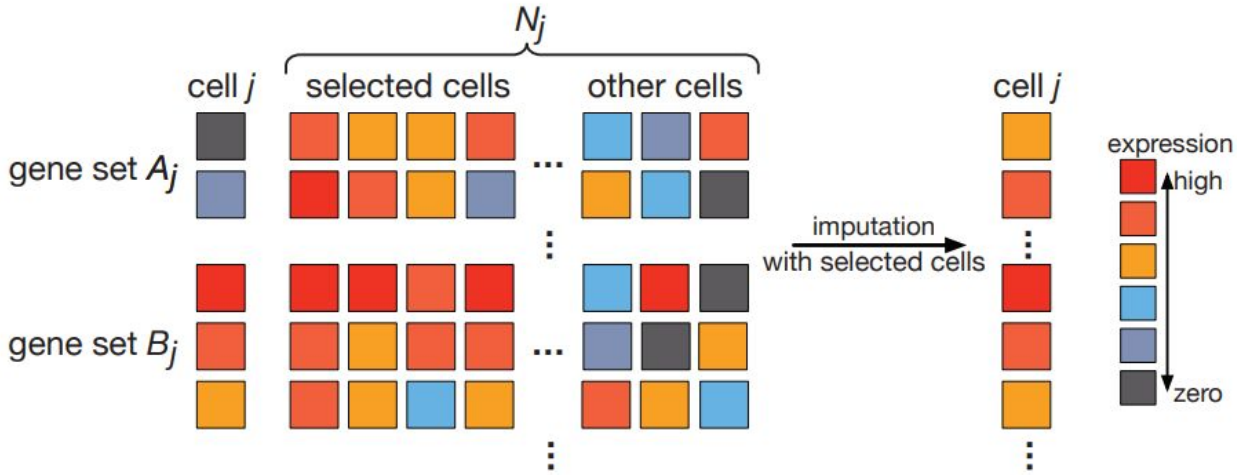
Wei Vivian Li, Jingyi Jessica Li

bioRxiv 141598; doi: https://doi.org/10.1101/141598

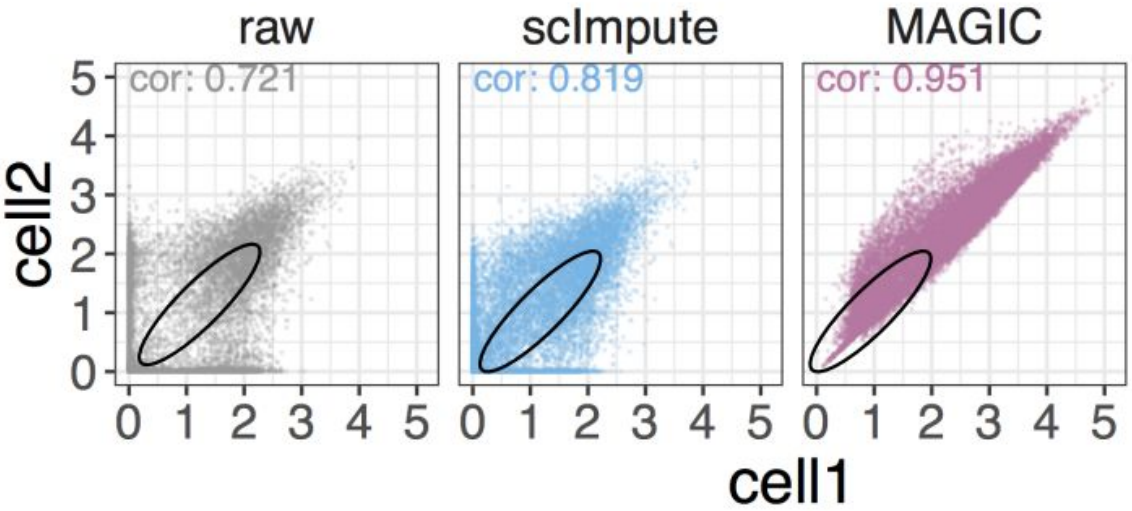
Genes have different dropout likelihoods



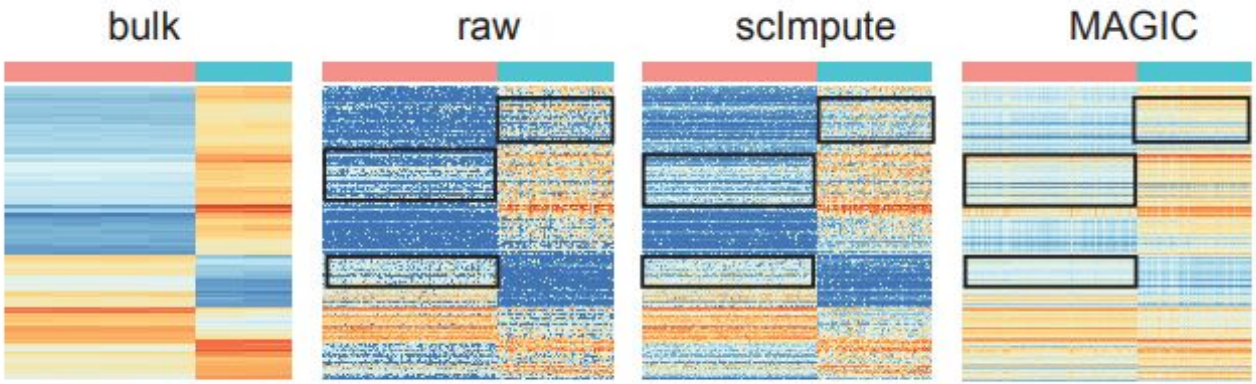
scImpute - the algorithm



scImpute - example



scImpute - example



scImpute - weaknesses

- Orders of magnitude slower

scImpute - how to use

```
scimpute(# full path to raw count matrix
         count_path = system.file("extdata", "raw_count.csv", package = "scImpute"),
         infile = "csv",           # format of input file
         outfile = "csv",          # format of output file
         out_dir = "./",           # full path to output directory
         labeled = FALSE,          # cell type labels not available
         drop_thre = 0.5,          # threshold set on dropout probability
         Kcluster = 2,             # 2 cell subpopulations
         ncores = 10)              # number of cores used in parallel computation
```

netSmooth (Network-smoothing based imputation for single cell RNA-seq)

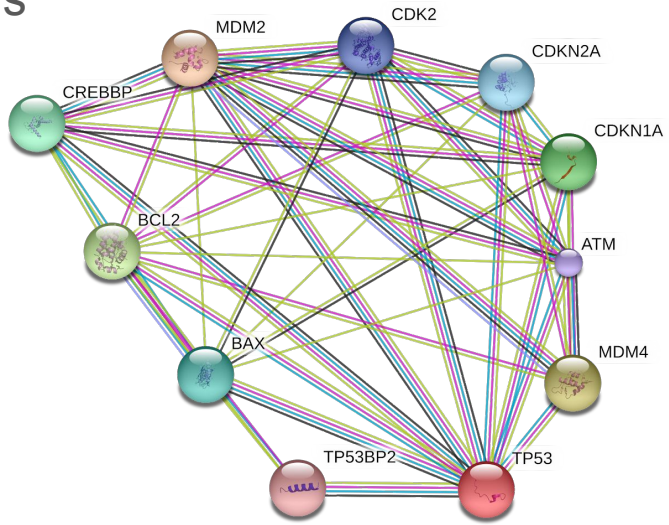
Assumptions:

- We can use information in previous experiments to impute (priors)
  - Interacting proteins are co-expressed
  - Protein-protein interaction networks encode co-expression expectations

Ronen J and Akalin A. *netSmooth:* Network-smoothing based imputation for single cell RNA-seq. *F1000Research* 2018, **7**:8
(doi: 10.12688/f1000research.13511.2)

On using priors

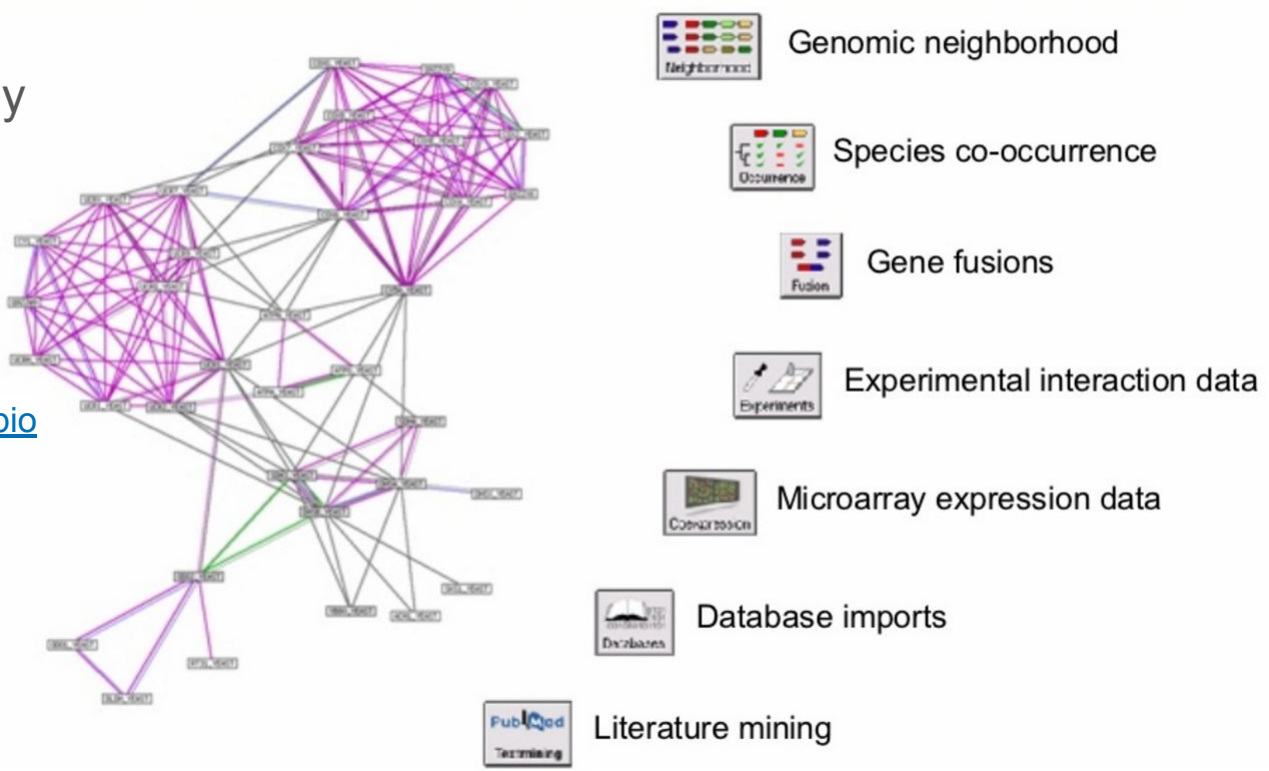
- Tenet of bayesian inference to temper evidence using priors
- We measure genes (transcripts) solo
- But they never act solo - always interacting with other genes
- Our prior isn’t individual genes’ expression levels
    - It’s the relations between them



https://string-db.org

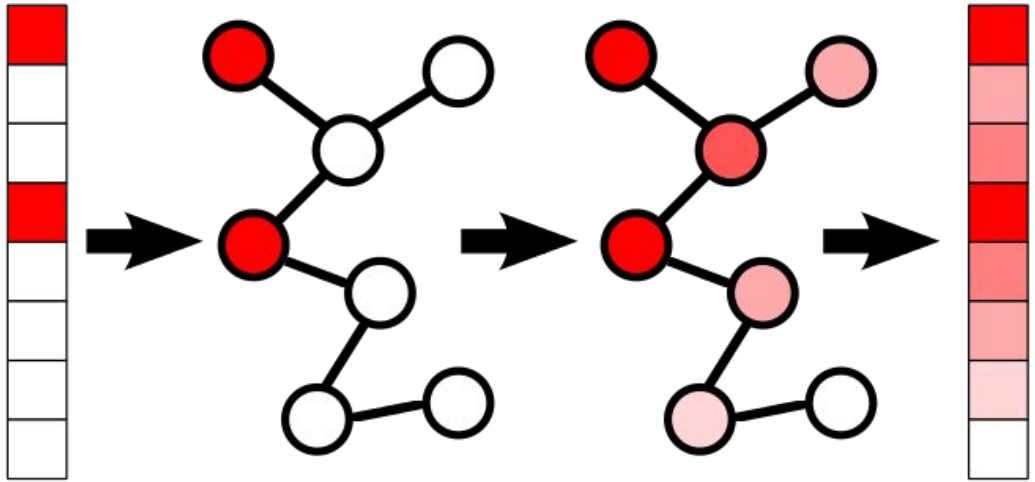
Protein-Protein interactions

- PPI database lists interactions from many sources
- Interacting proteins likely to be co-expressed
  - https://doi.org/10.1093/bioinformatics/bti398

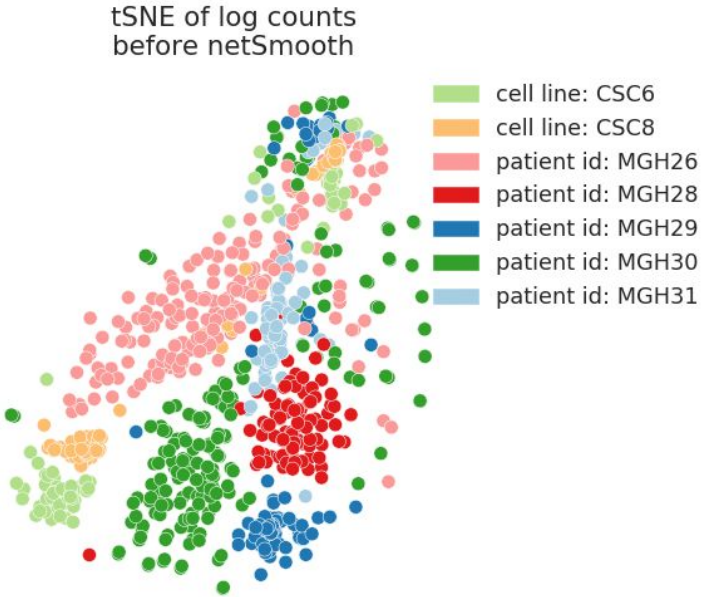


https://string-db.org

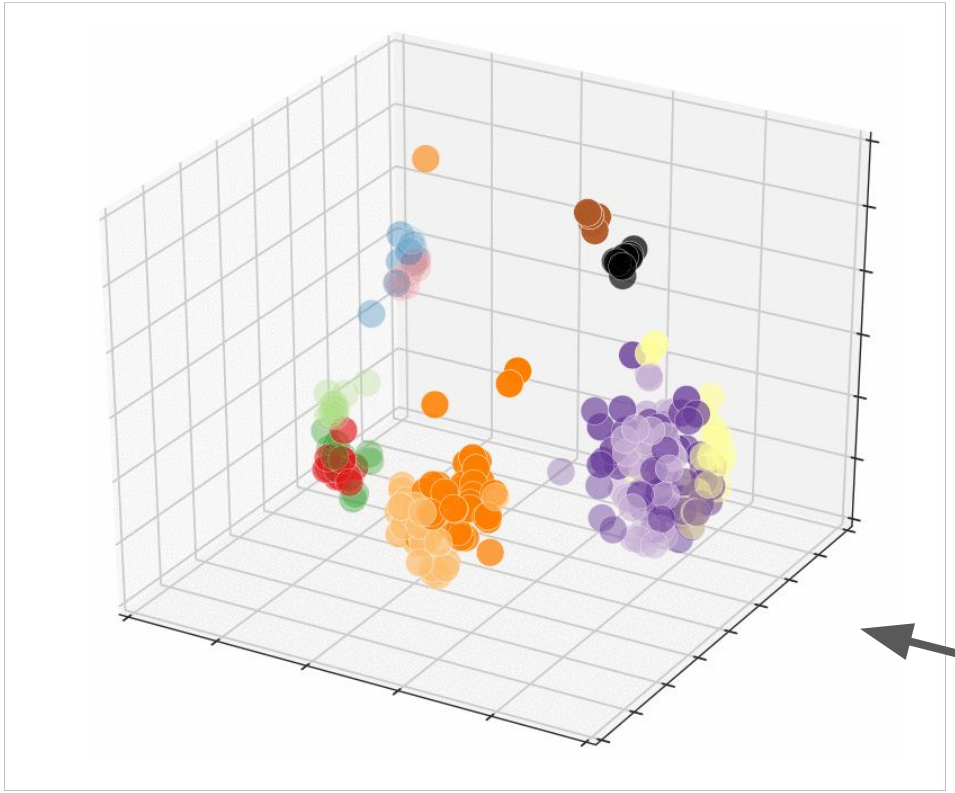
Network smoothing for RNAseq



netSmooth - glioblastoma tumors and cell lines



netSmooth - PCA of embryonic cells



- Zygote
- Early 2-cell stage blastomere
- Mid 2-cell stage blastomere
- Late 2-cell stage blastomere
- 2-cell stage blastomere
- 4-cell stage blastomere
- 8-cell stage blastomere
- 16-cell stage blastomere
- Early blastocyst cell
- Mid blastocyst cell
- Late blastocyst cell
- Liver cell
- fibroblast

netsmooth in action

netSmooth - weaknesses

- Dependent on a good gene network for your organism
- Gene networks may be context specific
  - Novel cell types with previously uncharacterized interactions
- Speed
  - MAGIC (30 seconds) > netSmooth (20 minutes) > scImpute (3 hours)

Network smoothing parameters

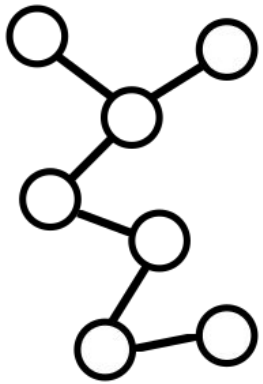
Random walk with restarts

$$E_{t+1} = \alpha A E_t + (1 - \alpha) E_0$$

Network smoothing parameters

Random walk with restarts

$$E_{t+1} = \alpha A E_t + (1 - \alpha) E_0$$



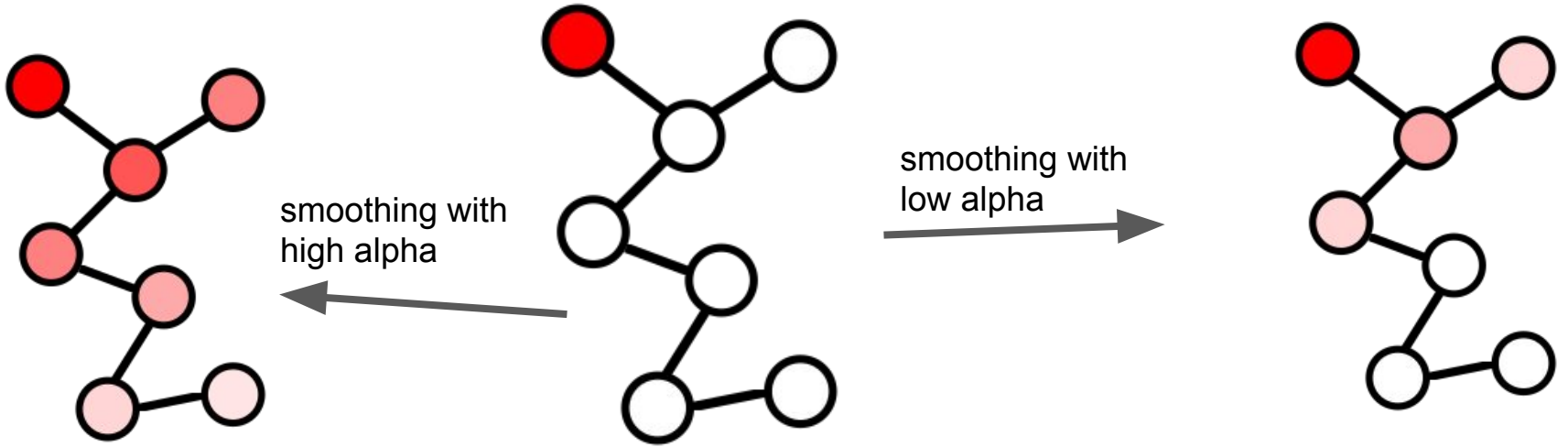
Gene network

Restart probability of random walk

Network smoothing parameters

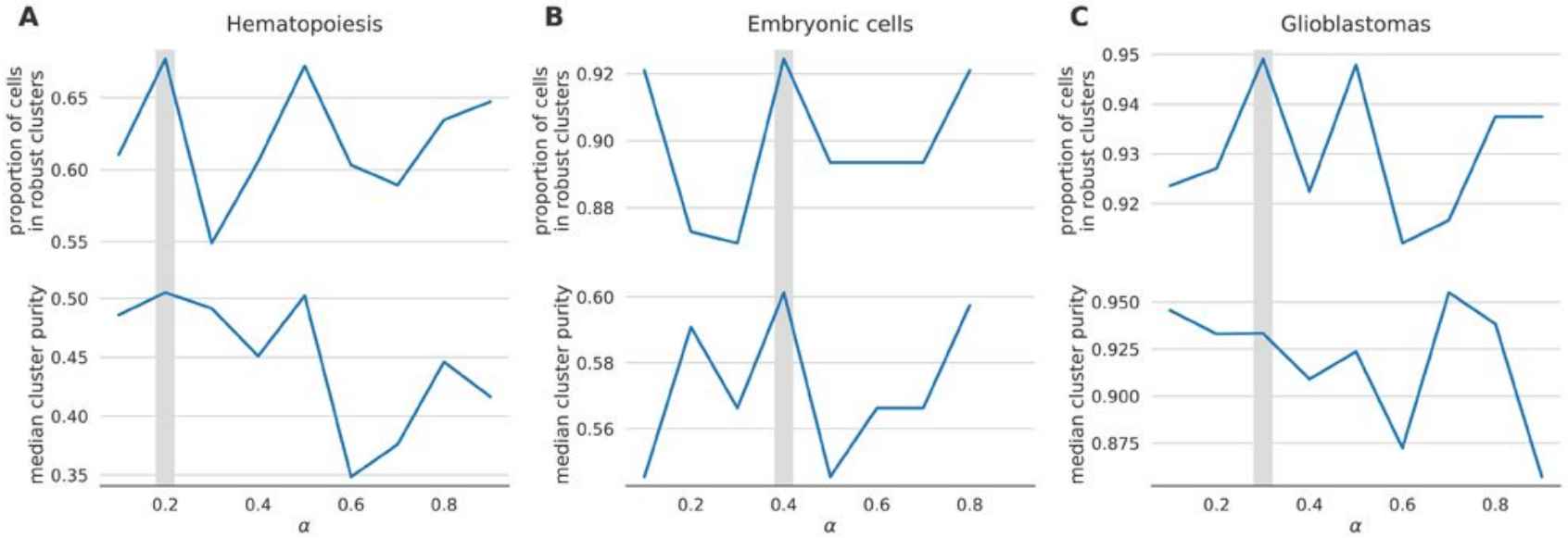
Random walk with restarts

$$E_{t+1} = \alpha A E_t + (1 - \alpha) E_0$$

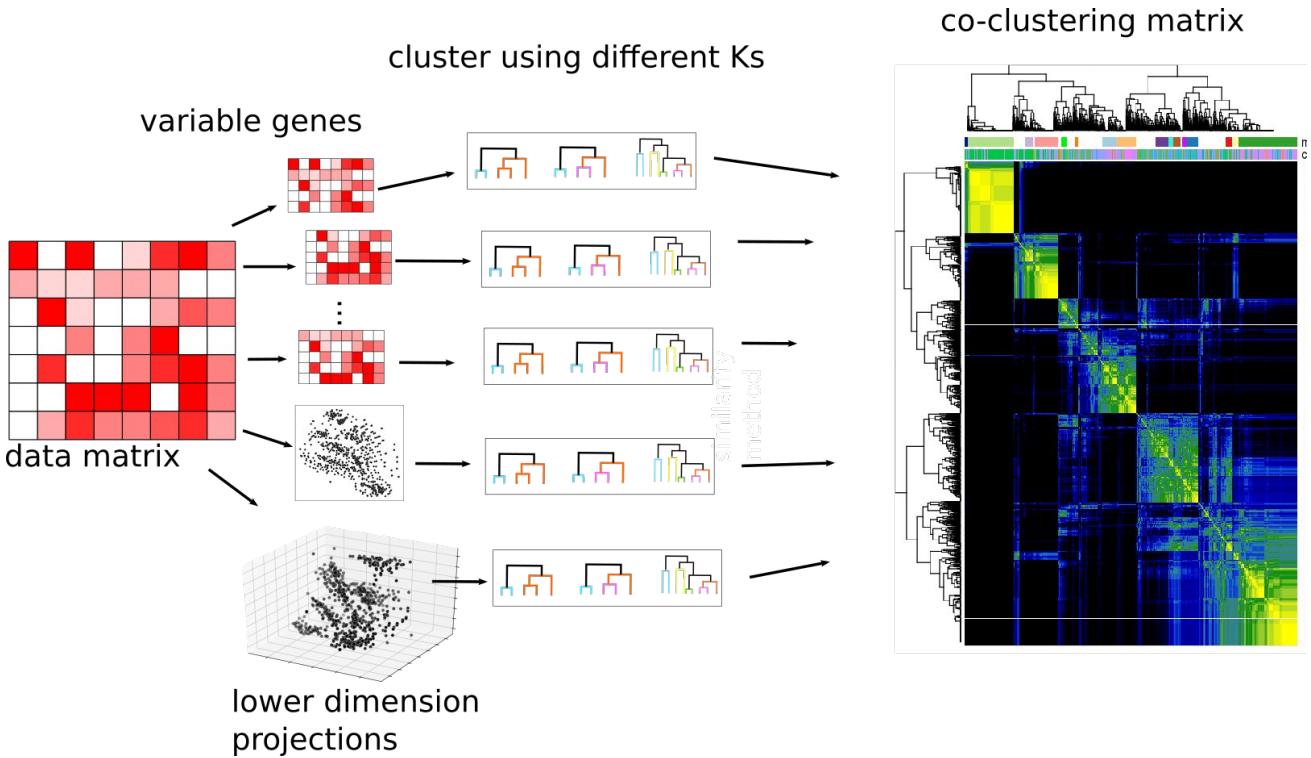


netSmooth - optimizing the smoothing parameters

- With no ground-truth labeling of cells
- Based on statistical robustness of clustering results



netSmooth - clustering single cells



Purdom E and Risso D (2017). clusterExperiment: Compare Clusterings for Single-Cell Sequencing

netSmooth - how to use

Walk through vignette