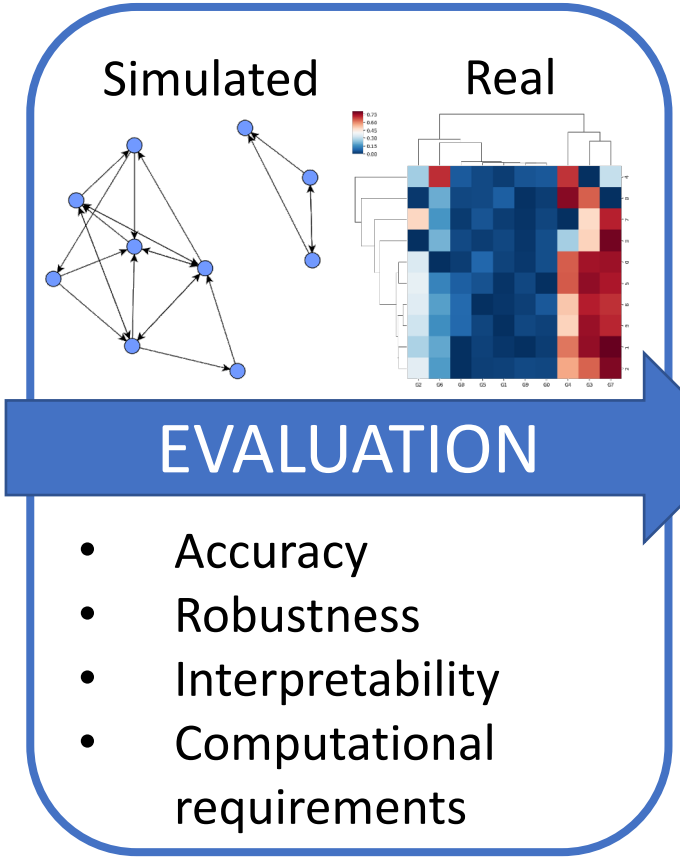
Evaluation of biclustering methods for complex RNA-seq data

Kath Nicholls

Smith lab meeting 5/2/2020

Overview

8 biclustering methods



Method	Strengths	Weaknesses
BicMix	…	…
nsNMF	…	…
…	…	…

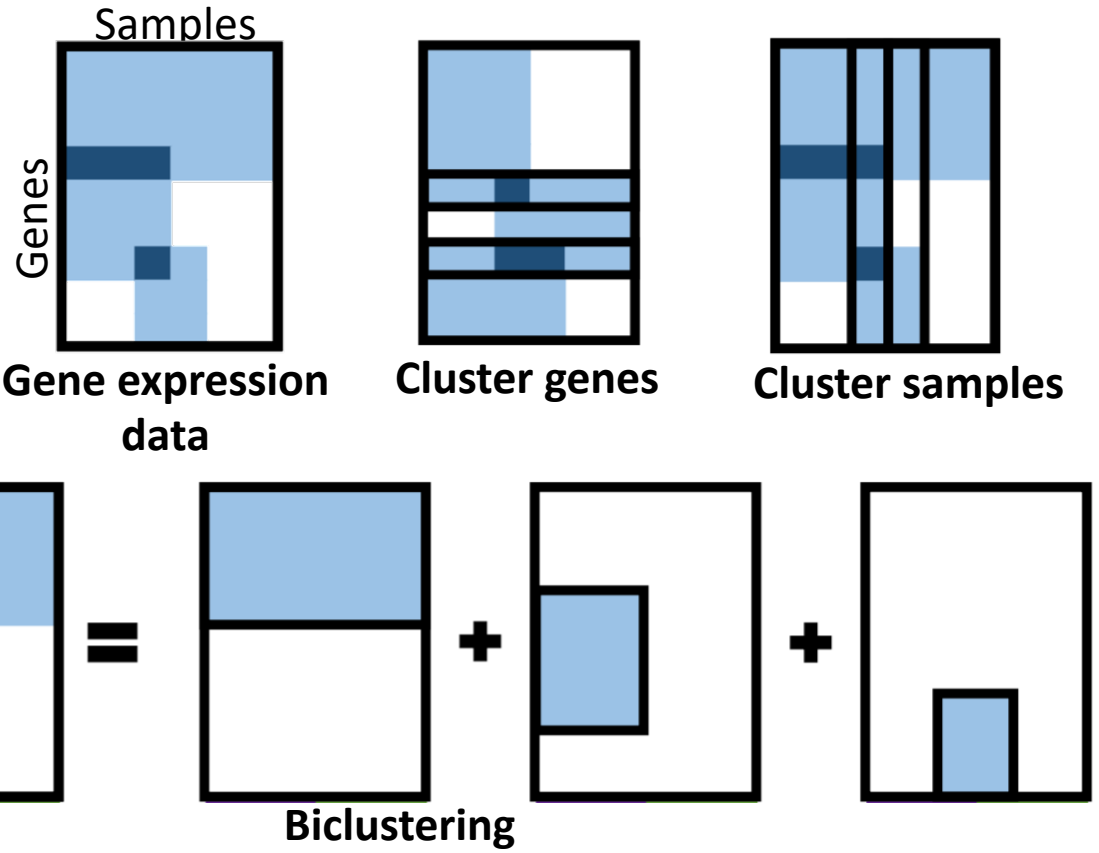
Apply best method to TRIAD II dataset

Overview

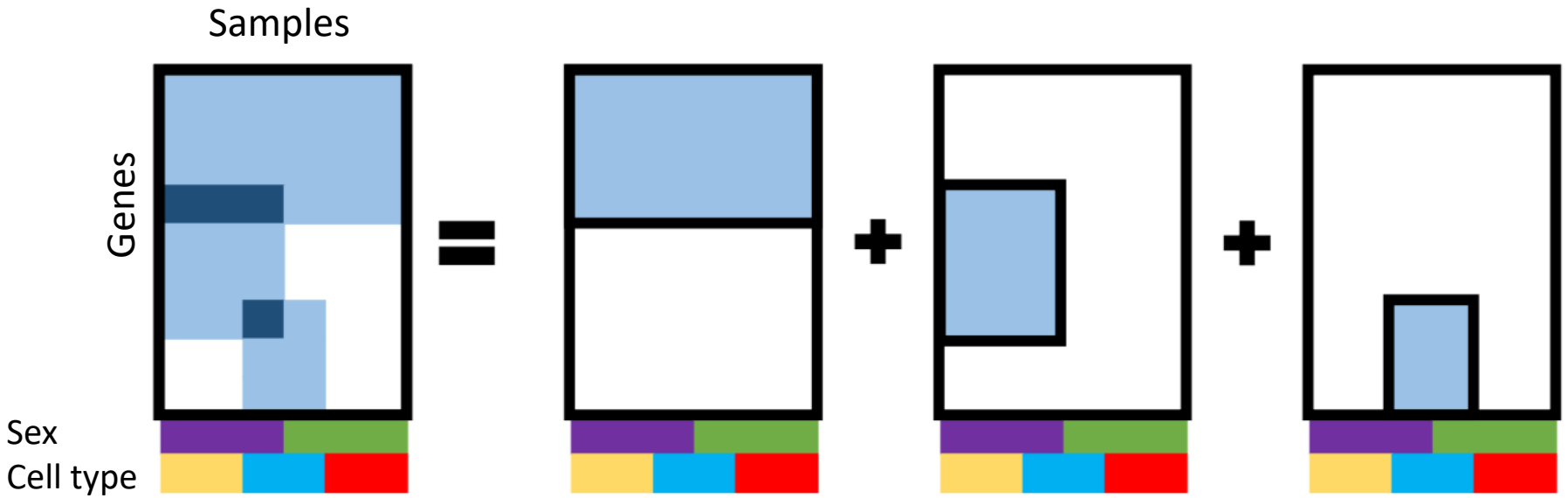
- What is biclustering?
- How to evaluate methods
  - Simulated datasets
  - Real datasets

Biclustering - finding patterns in gene expression data

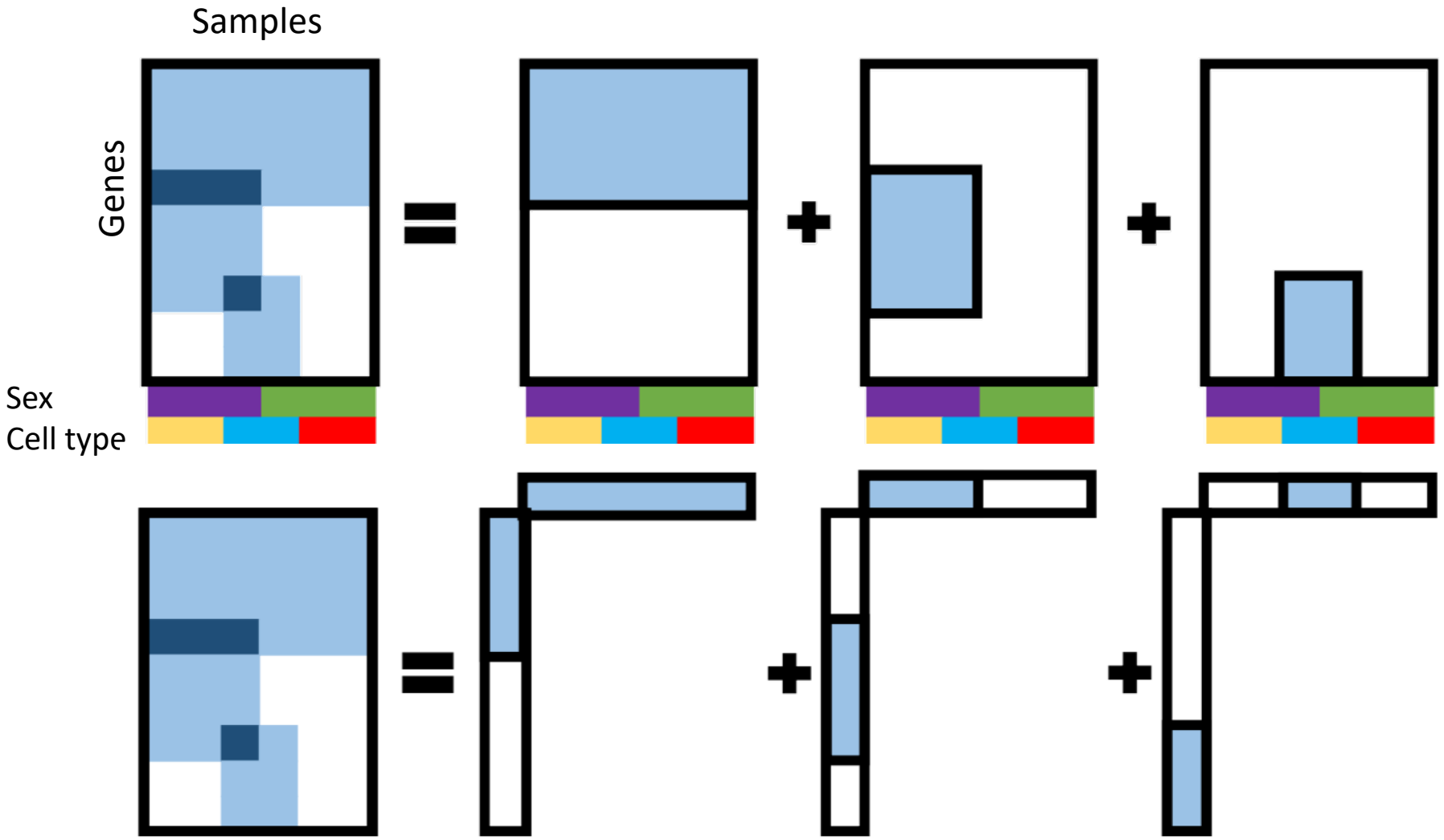
- Gene cluster: group of genes correlated across all samples
- Bicluster: group of genes correlated in a **subset** of samples
  - E.g. only in certain cell types or only in disease
  - Overlaps allowed



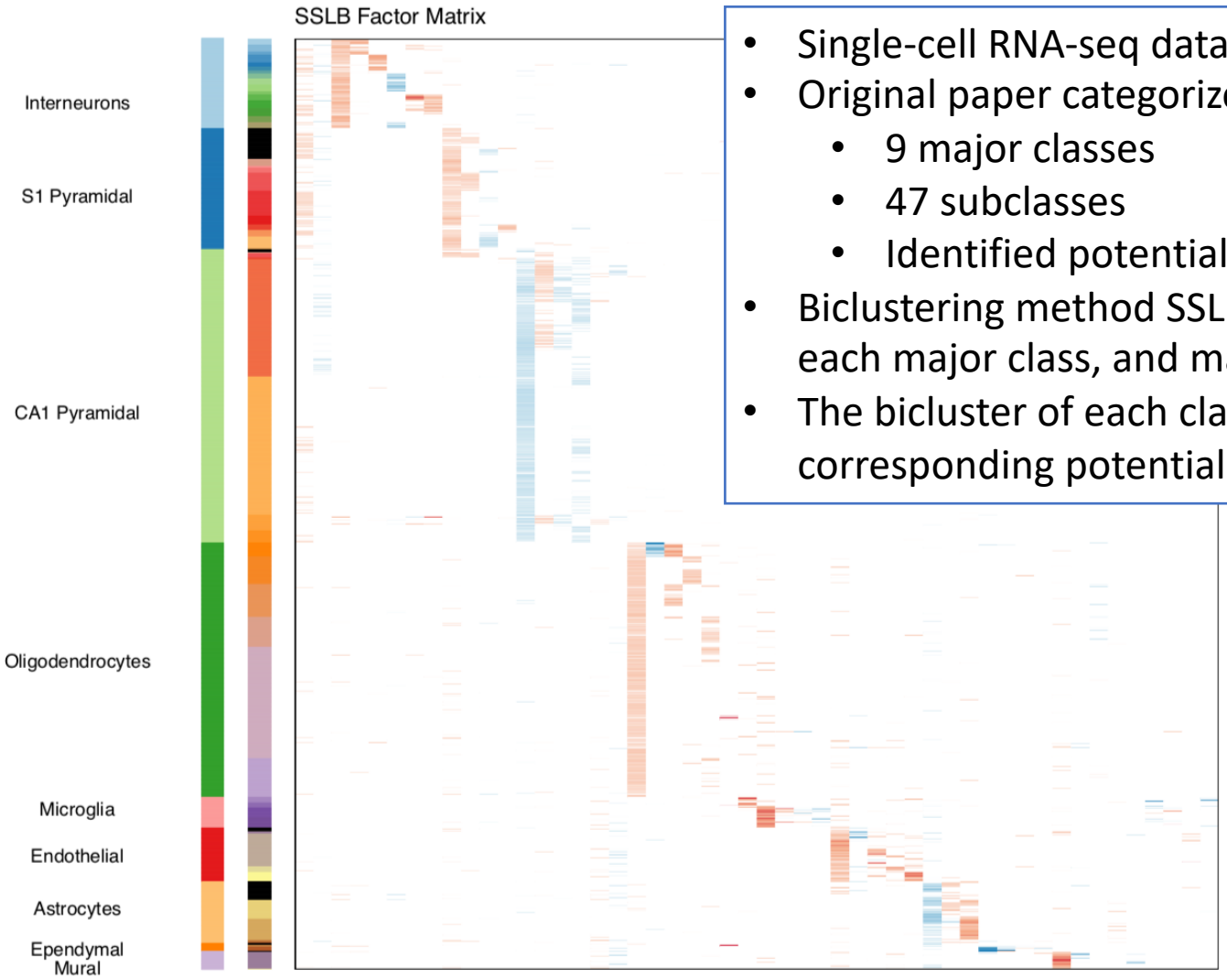
Biclustering - finding patterns in gene expression data



Biclustering - finding patterns in gene expression data



Application to cell-types in mouse brain



- Single-cell RNA-seq data (Zeisel et al. 2015)
- Original paper categorized cells
  - 9 major classes
  - 47 subclasses
  - Identified potential markers for classes
- Biclustering method SSLB found a cluster for each major class, and many subclasses
- The bicluster of each class contained the corresponding potential marker gene

SSLB, Moran et al. 2019 (preprint)

Advantages

- Adjust for **confounding effects** at the same time as biologically interesting effects
  - Expect factors related to confounders such as sex, batch

- **Find links** between groups of genes and sample traits such as disease and cell type

Overview

- What is biclustering?
- How to evaluate methods
  - Simulated datasets
  - Real datasets

Evaluating biclustering methods

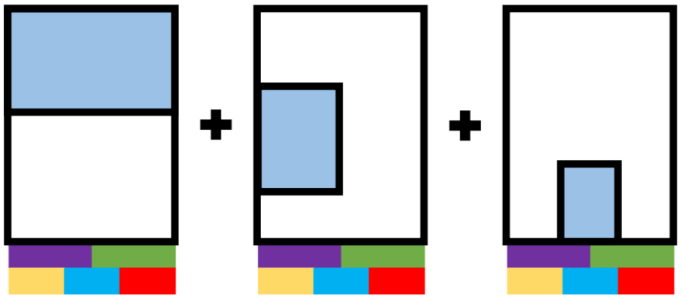
- Aim: know which method we can trust and how to use it best

- **Interpretability** – are the biclusters easy to interpret?
- **Robustness** – if you run it multiple times do you get similar results? Does parameter choice matter?
- **Computational requirements** – how long does it take to run?

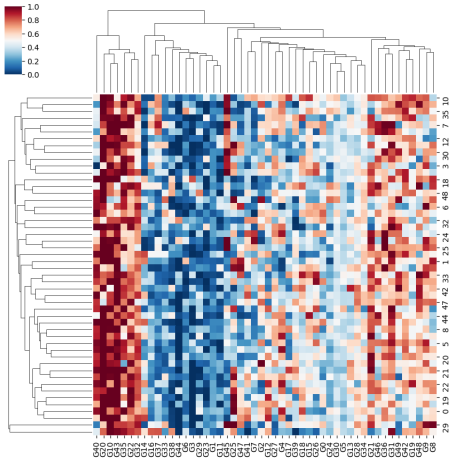
- **Accuracy** – simulated and real datasets

Accuracy on simulated datasets

True structure



(1) Generate gene expression data that has the 'true structure'

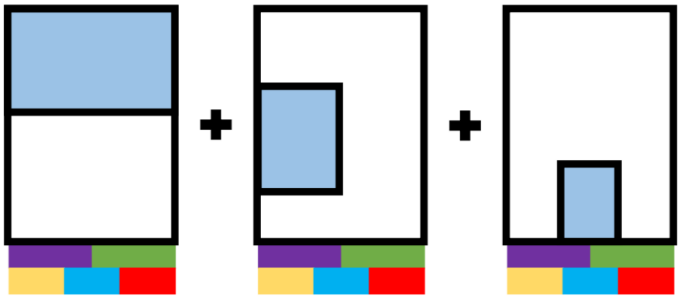


Samples (each involving perturbation)

Genes

(2) Run methods on dataset

Predicted structure



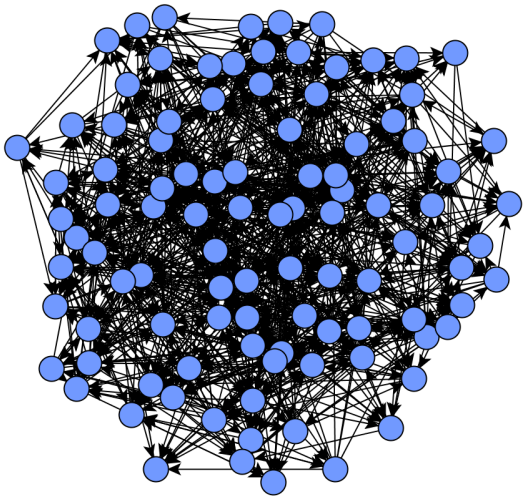
(3) Compare result to truth

GNW

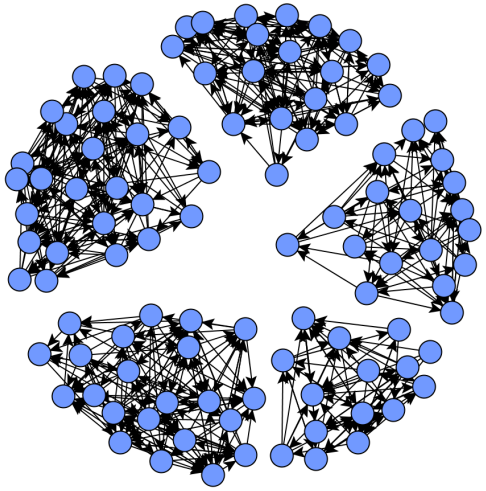
lis.epfl.ch/gnw

Accuracy on simulated datasets

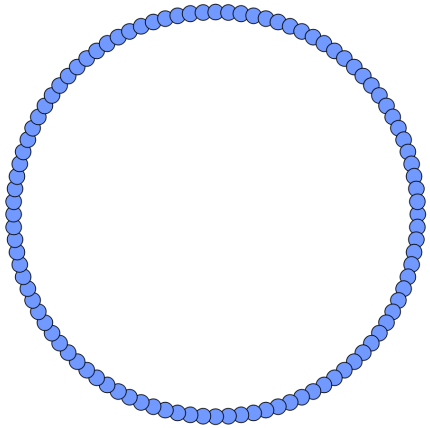
- Simulated data – generate from networks using Gene Net Weaver
  - Can simulate knockout experiments etc.
  - Want to tell difference between network structures



All connected



Modular

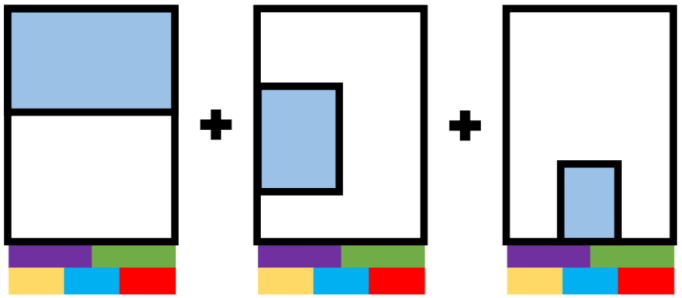


Disconnected

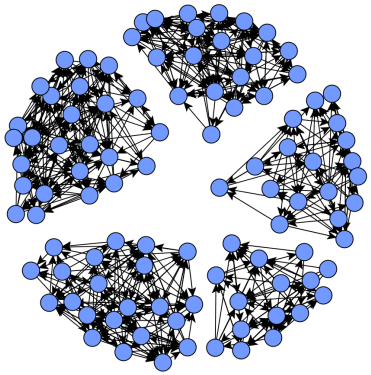
GNW
lis.epfl.ch/gnw

Accuracy on simulated datasets

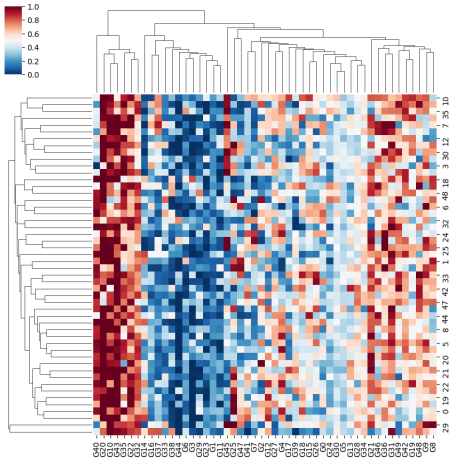
True structure



(1a) Generate graph(s) with structure



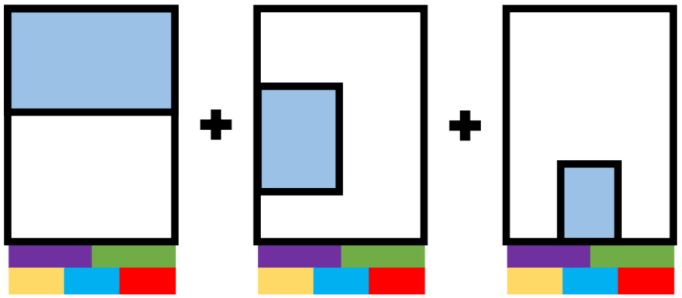
(1b) Simulate gene expression data with GNW



Samples (each involving perturbation)

Genes

(2) Run methods on dataset



Predicted structure

(3) Compare result to truth

GNW

lis.epfl.ch/gnw

Accuracy on simulated datasets

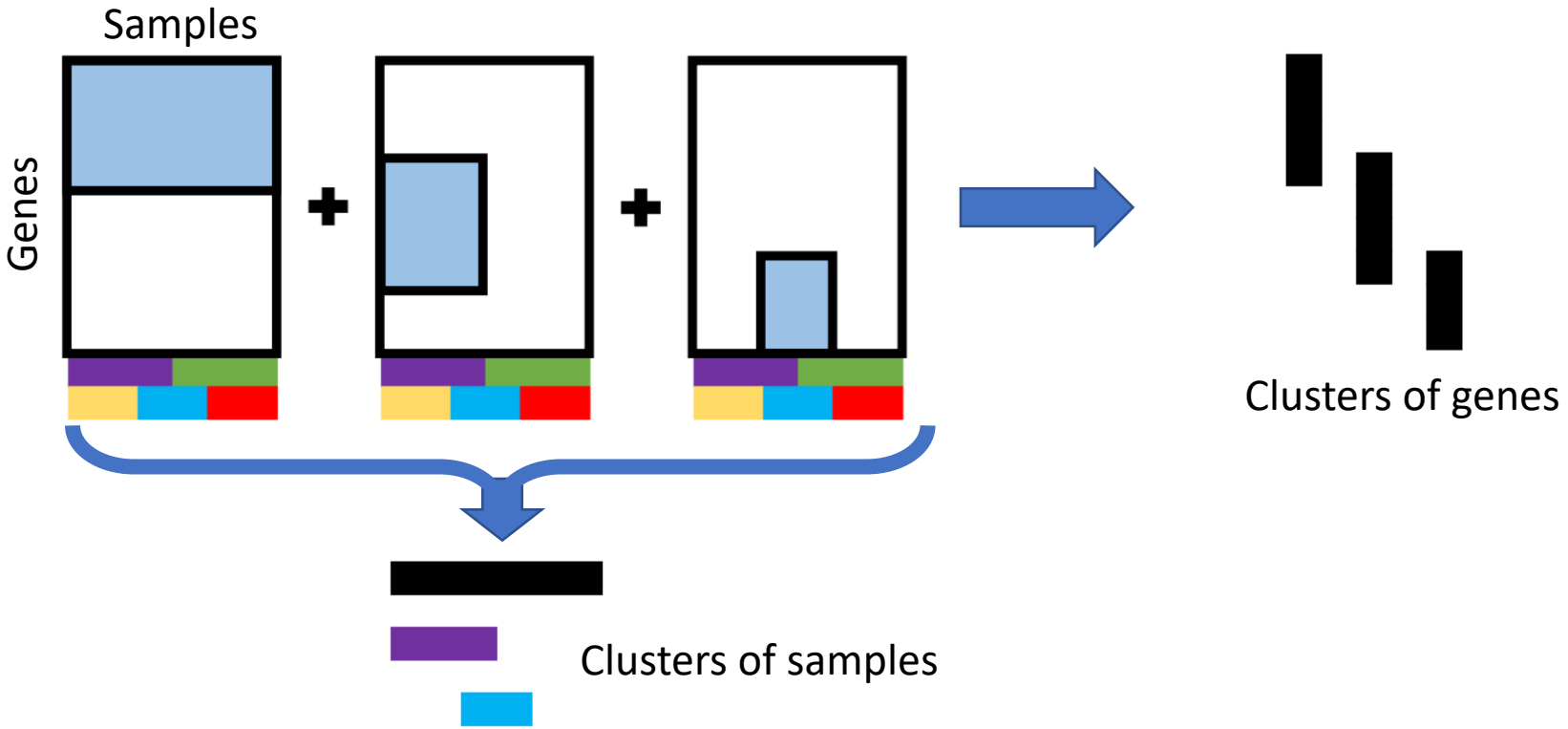
- Open question: what differences do we expect between gene networks?
  - Individual variation
  - Variation between two cell types
  - Variation between healthy and disease

Overview

- What is biclustering?
- How to evaluate methods
  - Simulated datasets
  - Real datasets

‘Accuracy’ on real datasets

- Common approaches to assessing accuracy:
  - Score clustering of samples e.g. cancer subtypes
  - Score clustering of genes e.g. by enrichment for GO/KEGG pathways



‘Accuracy’ on real datasets

- Plan: use known/expected biclusters in real data
- Open question: what biclusters can we expect? I.e. do we know a group of genes that should act differently in a subset of samples?
  - E.g. expect genes on X and Y chromosomes to act differently in male vs female patients
  - E.g. metabolism switch – are there good gene lists in KEGG? Are there experiments that would reveal this?
  - E.g. cell type specific genes

Conclusion

- Biclustering is a promising method for gene expression dataset over multiple cell types
- Important to evaluate methods before applying
  - Have confidence in the method
  - Guide choice of parameters
  - Understand how to interpret output
- Plan to use combination of simulated and real datasets