# Comparison study of sparse biclustering algorithms in gene expression datasets

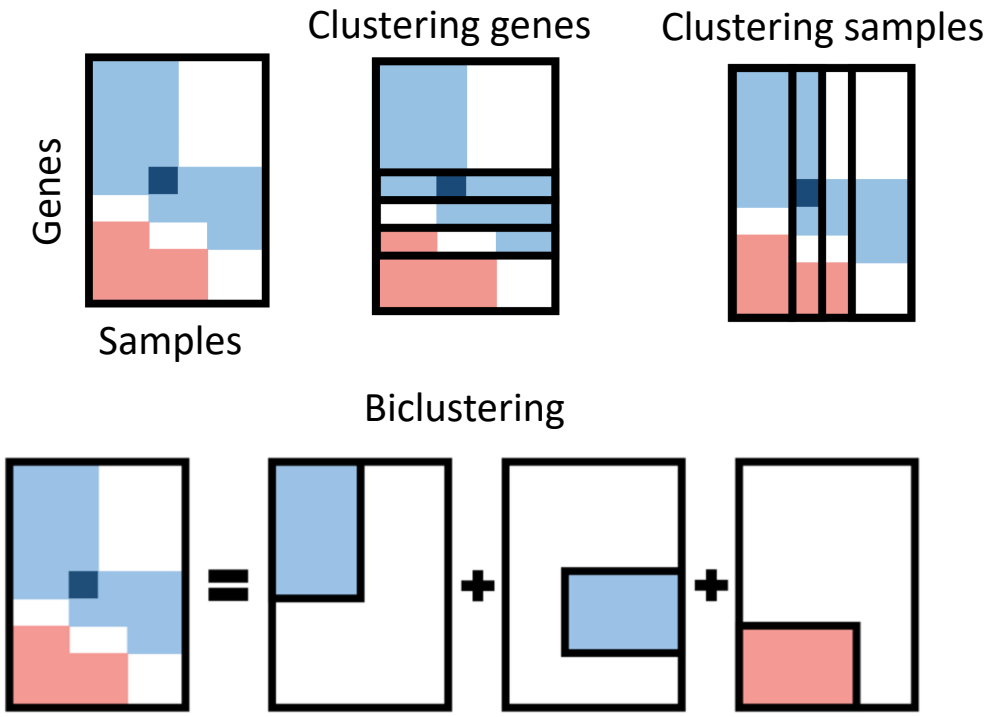
Kath Nicholls, MGM Seminar Day

24th November 2020

**CITIID and MRC BSU**

# Motivation: why biclustering?

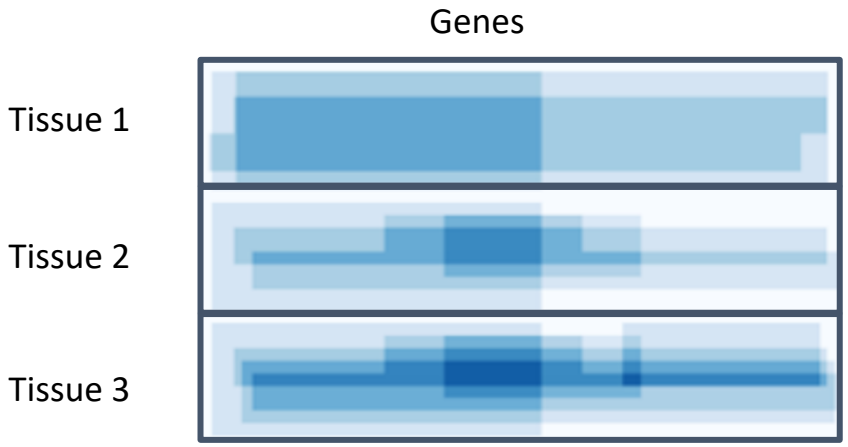
- Measure expression of 20,000 genes across 1000 samples
- Bicluster: subset of samples, subset of genes
- Links between sample sets and gene sets
- Sum of effects, allowing adjustment for confounders

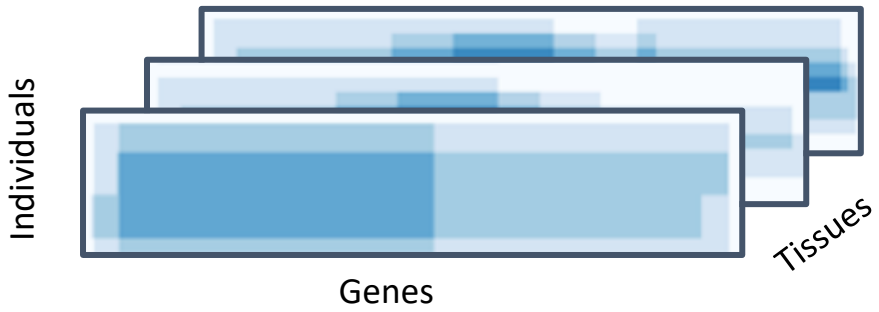


# Motivation: why a comparison study?

- More realistic simulated datasets
- New classes of method
  - *Popular* (FABIA, Plaid)
  - *NMF* (nsNMF, SNMF) – faster?
  - *Tensor* (SDA, MultiCluster) – exploits similarity between tissues?
  - *Adaptive* (BicMix, SSLB) – sparser?

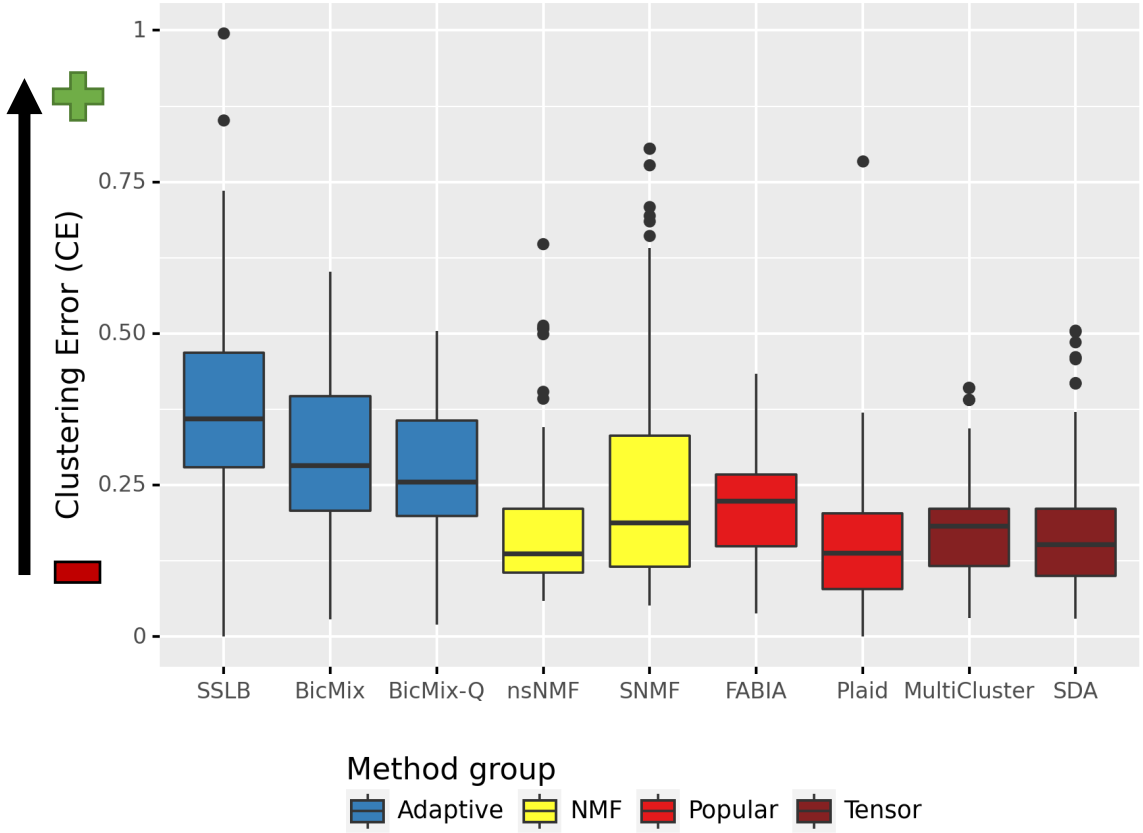
- Sparsity aids robustness and interpretability



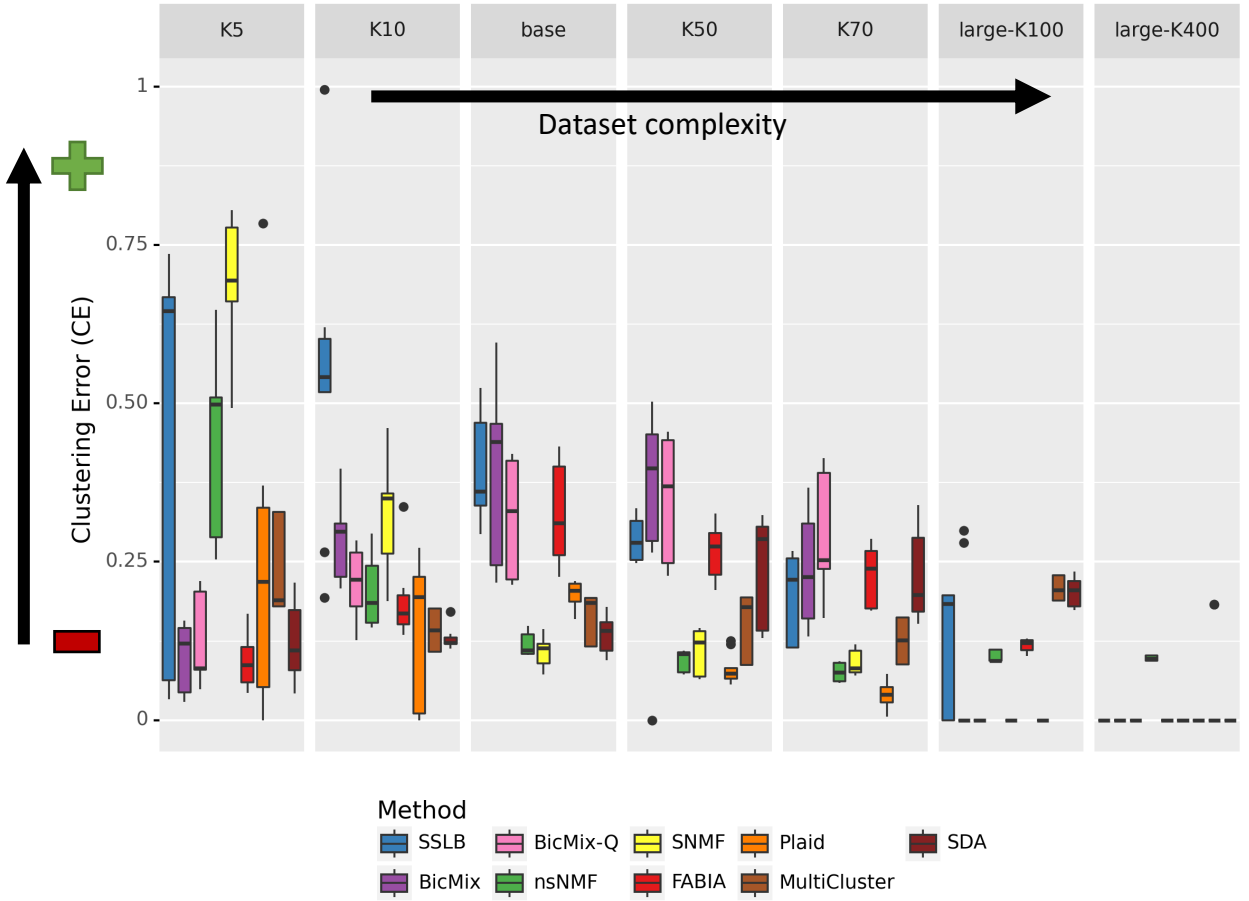



# Accuracy in simulated datasets

## *Adaptive* (SSLB) best on simulated datasets

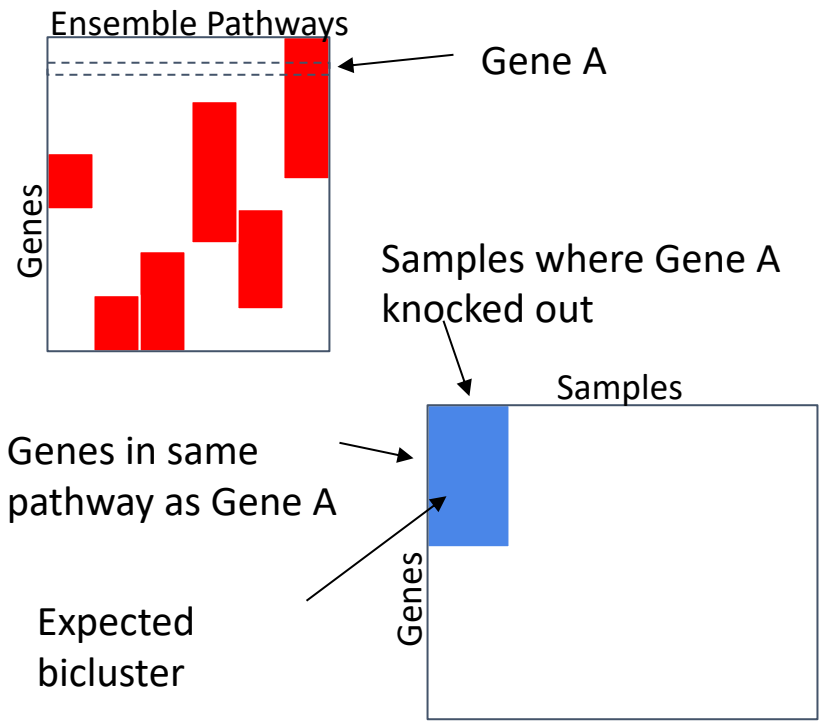



## Performance decreased on more complex datasets

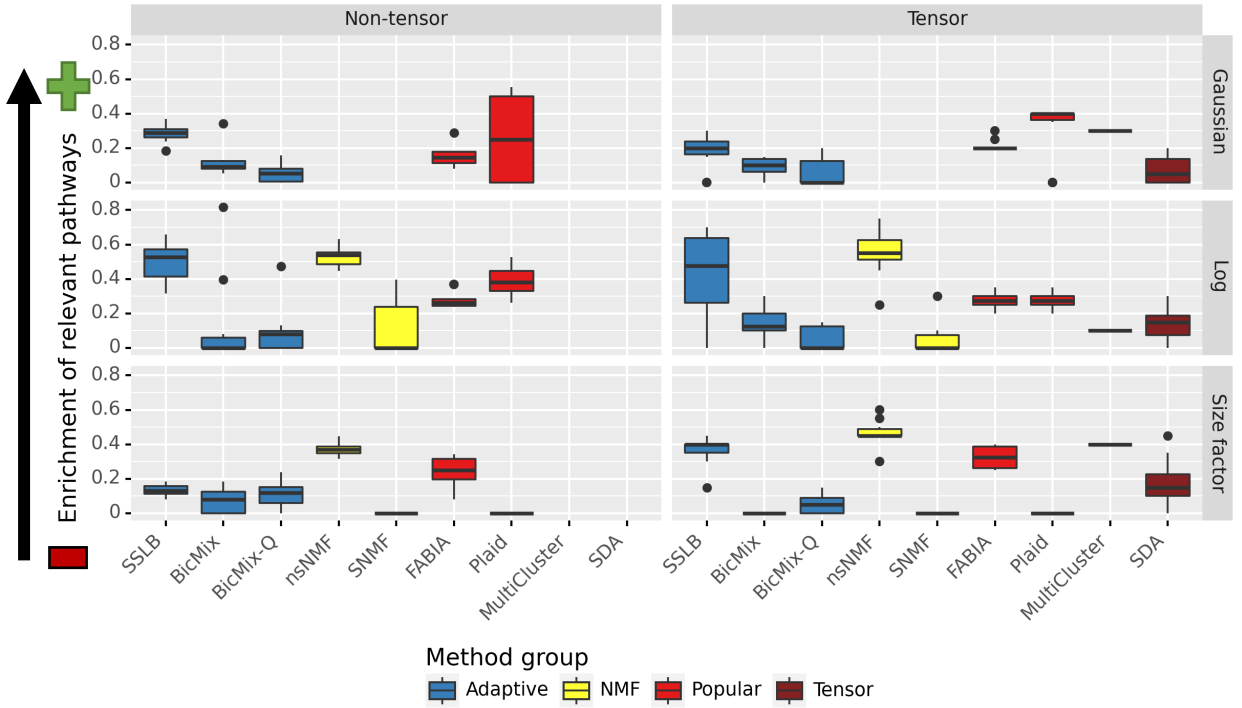


# Knockout mouse dataset allows evaluation of biclustering in real datasets

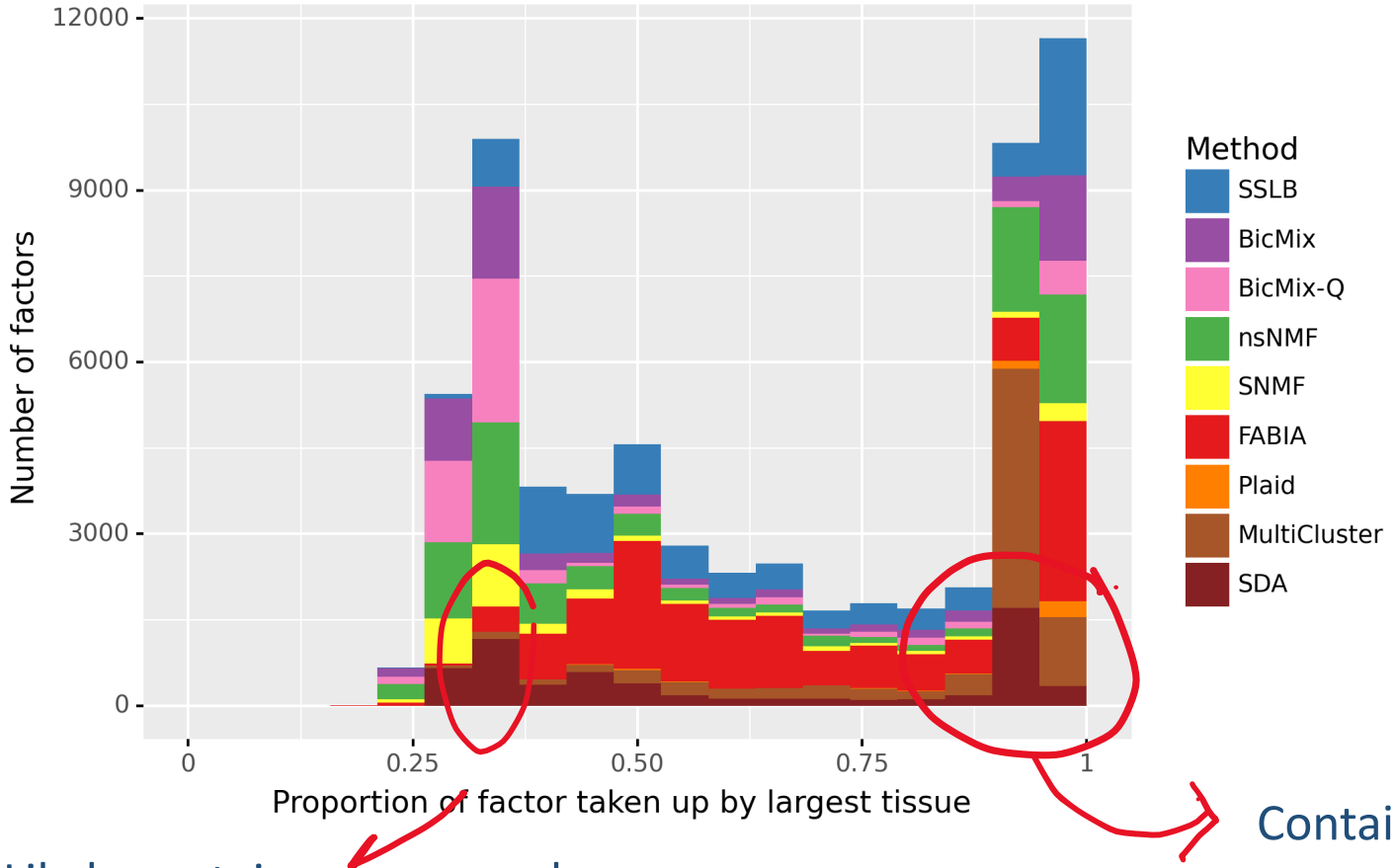
- International Mouse Phenotype Consortium dataset
- 1143 samples from 7 tissues
- 106 knockout genotypes + wildtype




SSLB, nsNMF best



# Poor clustering across tissues




- Many algorithms failed to cluster samples from multiple tissues

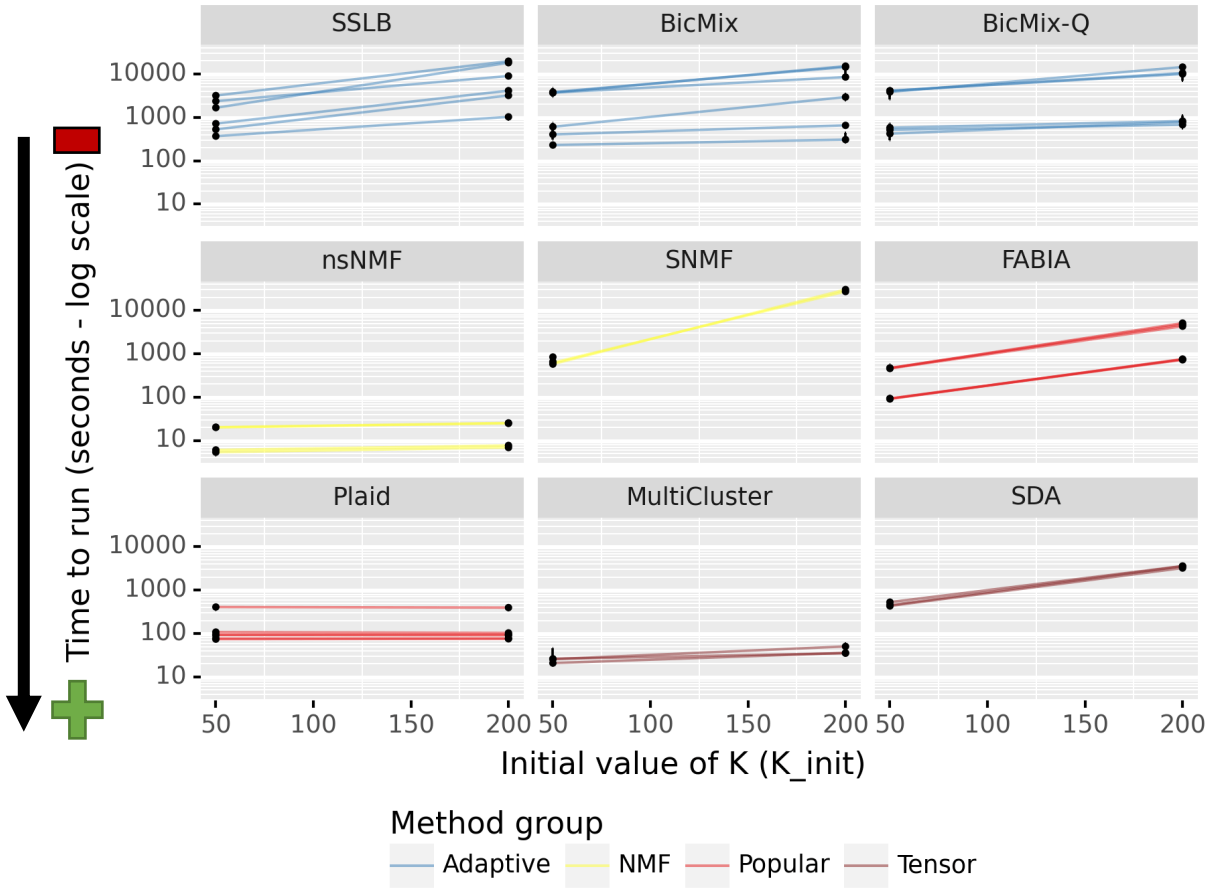
Contain samples from only one tissue

Likely contain every sample

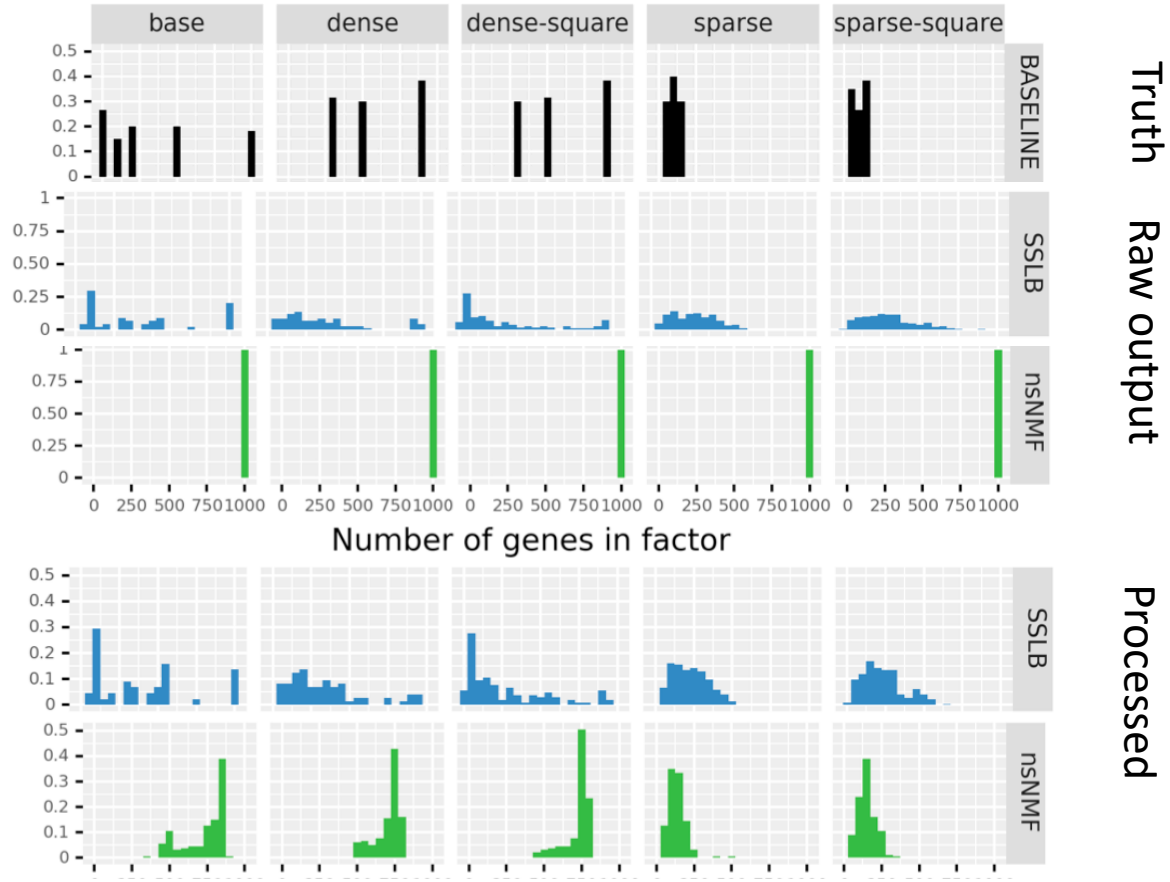
0.29 is proportion of samples in largest tissue in the dataset (liver)

# Practical issues – computational time, post-processing

## nsNMF, MultiCluster significantly faster




## Post-processing required to reveal biclusters (except *Adaptive*)



# Conclusions

- Improvements needed to deal with complex datasets
- Better normalisation required for multi-tissue datasets
- *NMF* methods promising (nsNMF)
    - Significant computational benefit
- *Adaptive* methods best overall
    - Good performance on simulated and real datasets
    - Did not require post-processing
    - Do not require exact number of factors