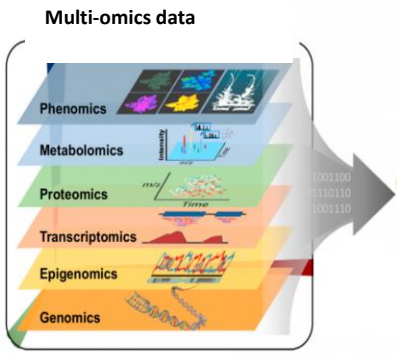
Evaluation of integrative clustering methods for the analysis of multi-omics data

Cécile Chauvel, Alexei Novoloaca, Pierre Veyre, Frédéric Reynier and Jérémie Becker

cecile.chauvel@bioaster.org

StatOmique workshop, November 5th 2019

Integrative clustering methods for multi-omics data



Four different strategies of integration:

1. Analyze each omics separately and combine results at the interpretation step
2. Clustering on each omics separately before applying consensus clustering
3. Concatenation into a single matrix before applying standard clustering approaches
4. Search for common variations across omics by specific models

Several questions to be adressed:

- How are omics data integrated?
- How is clustering performed?
- How are data pre-processed?
- How are the model parameters tuned?
- What are the performances of the methods?

OUTLINE

① Presentation of the methods

② Simulation study

③ Application on the TCGA breast cancer dataset

④ Conclusion

PRESENTATION OF THE METHODS

Presentation of the methods

- The dataset is composed of K matrices $X_1, \dots, X_K$
- Each matrix X_k is of size $p_k \times n$ (p_k variables/features, n samples)
- All matrices contain measurements on the same n samples

  ➔ The goal is to perform clustering on the samples

- Focus on approaches that
  - can be **applied to any omics,**
  - **do not require** any prior biological knowledge (e.g., pathways)
  - and **give an insight** to omics variables.

Non-integrative

- (2.) Gaussian mixture models on each omic + consensus clustering
- (3.) Concatenation + Gaussian mixture models

Matrix factorization

- iCluster
- moCluster
- JIVE
- iNMF

Bayesian

- BCC
- MDI

iCluster

Shen, R., Wang, S., and Mo, Q. (2013).*The annals of applied statistics*.

iCluster is a Gaussian joint latent variable model:

$$X_k = W_k Z + \epsilon_k,$$
$$Z \sim N_q(0, I),$$

W_k $(p_k \times q)$ data-specific loading matrix
Z $(q \times N)$ shared latent variable matrix
$\epsilon_k \sim N(0, \Sigma_k)$, with Σ_k diagonal

PARAMETERS

- Number of clusters determined by the Proportion of Deviance or the Rand Index.
- Number of latent variables = Number of clusters -1

DATA PRE-PROCESSING

Centering of the X_k

ESTIMATION

EM algorithm

CLUSTERING

K-means on $E(Z|X_1, \ldots, X_K)$

moCluster

Meng, C., Helm, D., Frejno, M., and Kuster, B. (2015) *Journal of proteome research*.

Model close to iCluster:

$$X_k = W_k Z + \epsilon_k,$$

W_k $(p_k \times q)$ data-specific loading matrix
Z $(q \times N)$ shared latent variable matrix
$\epsilon_k \sim N(0, \sigma^2 I)$

Same noise variance across variables and data types

➔ shared and specific variations no longer separable

PARAMETERS

- Number of clusters determined by the gap statistic
- Number of latent variables determined by inspection of eigen values (scree plot or permutation test)

DATA PRE-PROCESSING

X_k standardized and scaled by the inverse of the largest eigen value

ESTIMATION

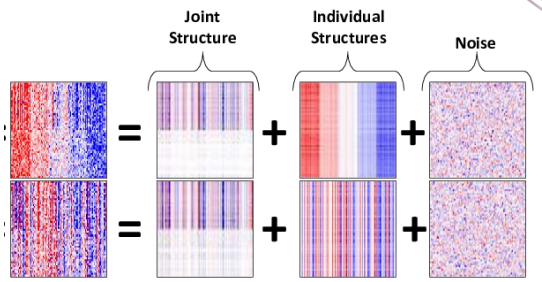
Consensus PCA (NIPALS algorithm)

CLUSTERING

HCA on the latent variable matrix Z

JIVE - JOINT AND INDIVIDUAL VARIATION EXPLAINED

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013)*The annals of applied statistics*.



Addition of a data-specific term:

$$X_k = W_k Z + W_k^s Z_k^s + \epsilon_k$$

W_k^s $(p_k \times q_k)$ data-specific loading matrix

Z_k^s $(\mathrm{q_k} \times N)$ data-specific latent variable matrix

Constraint of orthogonality for identifiability: $W_k Z . (W_k^s Z_k^s)^T = 0$

PARAMETERS

Number of latent variables estimated by permutation approach on the eigen values

DATA PRE-PROCESSING

X_k centered, and scaled by their Frobenius norm

ESTIMATION

Iterative error minimization by fixing one term (shared or specific) at a time
+ SVD decomposition

CLUSTERING

No guidelines

iNMF – integrative Non-negative Matrix Factorization

Yang, Z. and Michailidis, G. (2016) *Bioinformatics*.

The model is a particular case of JIVE in which the shared and specific loadings are equal:

$$X_k = (Z + Z_k^s)W_k + \epsilon_k,$$

with a non-negativity constraint: $Z, Z_k^s, W_k \geq 0$

ESTIMATION

Minimization of the penalized loss function:

$$\min_{\substack{Z,Z_1^s,\cdots,Z_K^s \\ W_1,\cdots,W_K}} \sum_{k=1}^{K} ||\mathbf{X}_k - (\mathbf{Z} + \mathbf{Z}_k^s)\mathbf{W}_k||^2 + \lambda \sum_{k=1}^{K} ||\mathbf{Z}_k^s \mathbf{W}_k||^2.$$

λ controls for the homogeneity between shared and specific structure:
High λ ➔ more emphasis on the shared structure.

DATA PRE-PROCESSING

Variance stabilization (log – transformation), non-negativity transformation and scaling by the Frobenius norm.

CLUSTERING

No guidelines

PARAMETERS

- Number of latent variables maximizing stability (consensus approach)
- λ : ad hoc procedure attributing as much weight as possible to the specific structure

MDI - Multiple Dataset Integration

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). *Bioinformatics*.

- Bayesian method: Dirichlet multinomial allocation mixture model
- Cluster assignments are dependant across datasets:

Mixture proportion associated with cluster c_{ik} in dataset k

$$P(c_{i1}, c_{i2}, \cdots c_{iK}|\phi) \propto \prod_{k=1}^{K} \pi_{c_{ik}k} \prod_{k=1}^{K-1} \prod_{l=k+1}^{K} (1 + \phi_{kl} \mathbb{1}(c_{ik} = c_{il})),$$

Cluster allocation of sample i in dataset k

Association strength between datasets k and l

DATA PRE-PROCESSING	ESTIMATION	GLOBAL CLUSTERING	PARAMETERS
None	Gibbs sampling	Maximization of the posterior expected adjusted Rand Index accross source-specific clusterings	Maximal number of clusters. The authors' recommendation: n/2 but instable in our simulations (n was chosen)

BCC - Bayesian Consensus Clustering

Lock, E. F. and Dunson, D. B. (2013) *Bioinformatics*.

Dirichlet mixture model, aiming at uncovering a single clustering across sources by:

Shared cluster allocation of sample n

Adherence parameter of dataset k

$$P(L_{kn} = l|C_n) = \begin{cases} \alpha_k \text{ if } C_n = L_{kn} \\ \frac{1-\alpha_k}{1-q} \text{ otherwise,} \end{cases}$$

Source-specific cluster allocation of sample n in dataset k

Maximal number of clusters

DATA PRE-PROCESSING	ESTIMATION	CLUSTERING	PARAMETERS
None	Gibbs sampling	Estimated by C (shared clustering) or L (source-specific)	Maximal number of clusters ***q*** maximizing the mean adherence

SIMULATION STUDY

Main simulation study

- K=3 data matrices with
  - 180, 210 and 240 variables
  - 60 samples
  - 3 shared clusters of 20 samples each
  - 2 levels of Signal to Noise Ratio (SNR)

- 3 simulation strategies
  - iNMF-derived scenario with overlaps between the shared and specific blocks
  - iNMF-derived scenario without overlaps between the shared and specific blocks
  - BCC-derived scenario with 3 to 5 specific clusters

- Evaluation (100 repetitions of each scenario):
  - Estimated number of shared clusters
  - Clustering performance (Adjusted Rand Index)

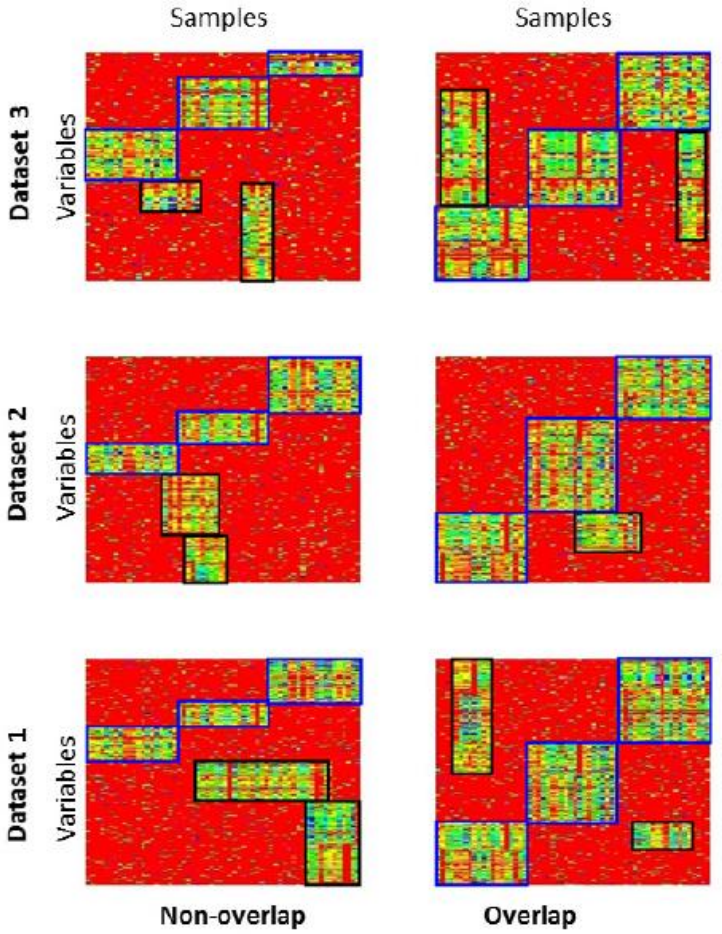
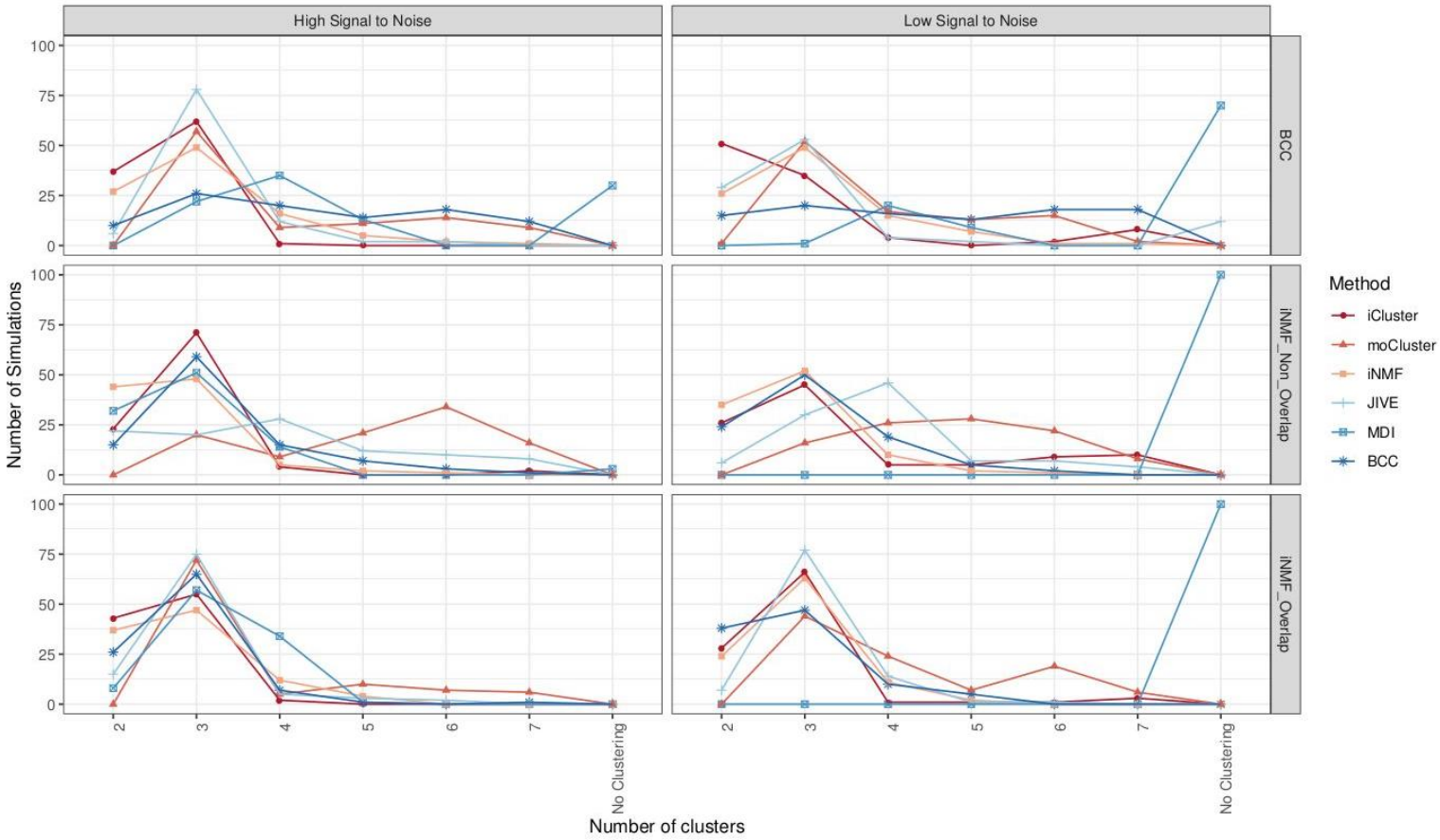


Illustration of iNMF simulations

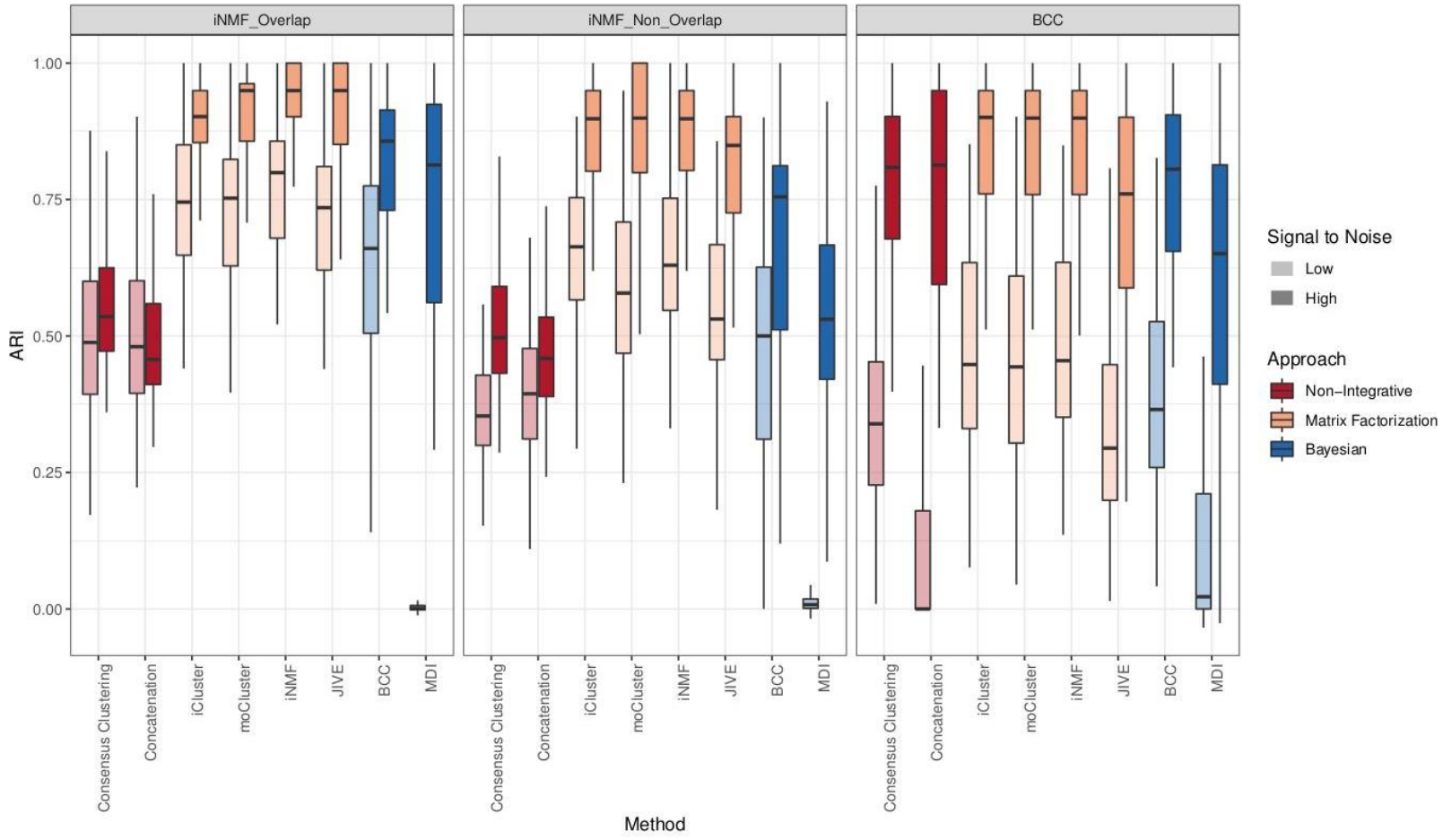
Number of shared clusters

- 3 clusters chosen on average
- Sharp peak around 3 for high SNR and iNMF overlap scenarios
- Ranking of the methods (by % of times 3 clusters are found):
  1. iNMF
  2. iCluster
  3. JIVE
  4. BCC
  5. moCluster
  6. MDI



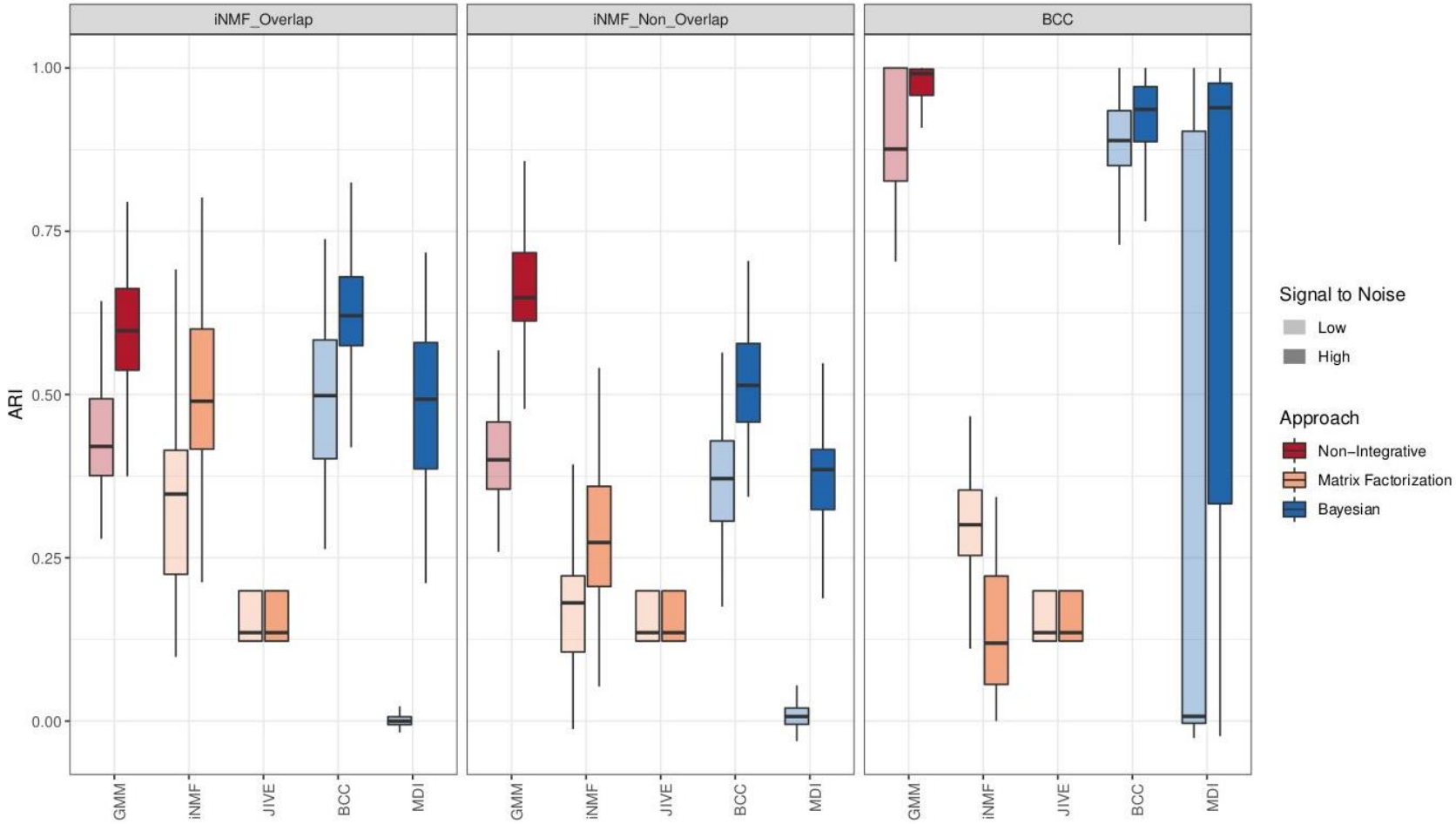
Clustering on shared structures

- ARI of integrative methods are higher than those of non-integrative ones
- SNR + simulation design have a great impact on clustering.
- Ranking of the methods:
  1. iCluster, moCluster, iNMF
  2. JIVE, BCC
  3. MDI (extremely sensitive to noise)



Clustering on specific structures

- Not central in our study, classical clustering methods apply such as GMM
- GMM slightly outperforms BCC
- JIVE underperforms ➔ identifiability issues
- MDI sensitive to noise



High-dimension simulation study

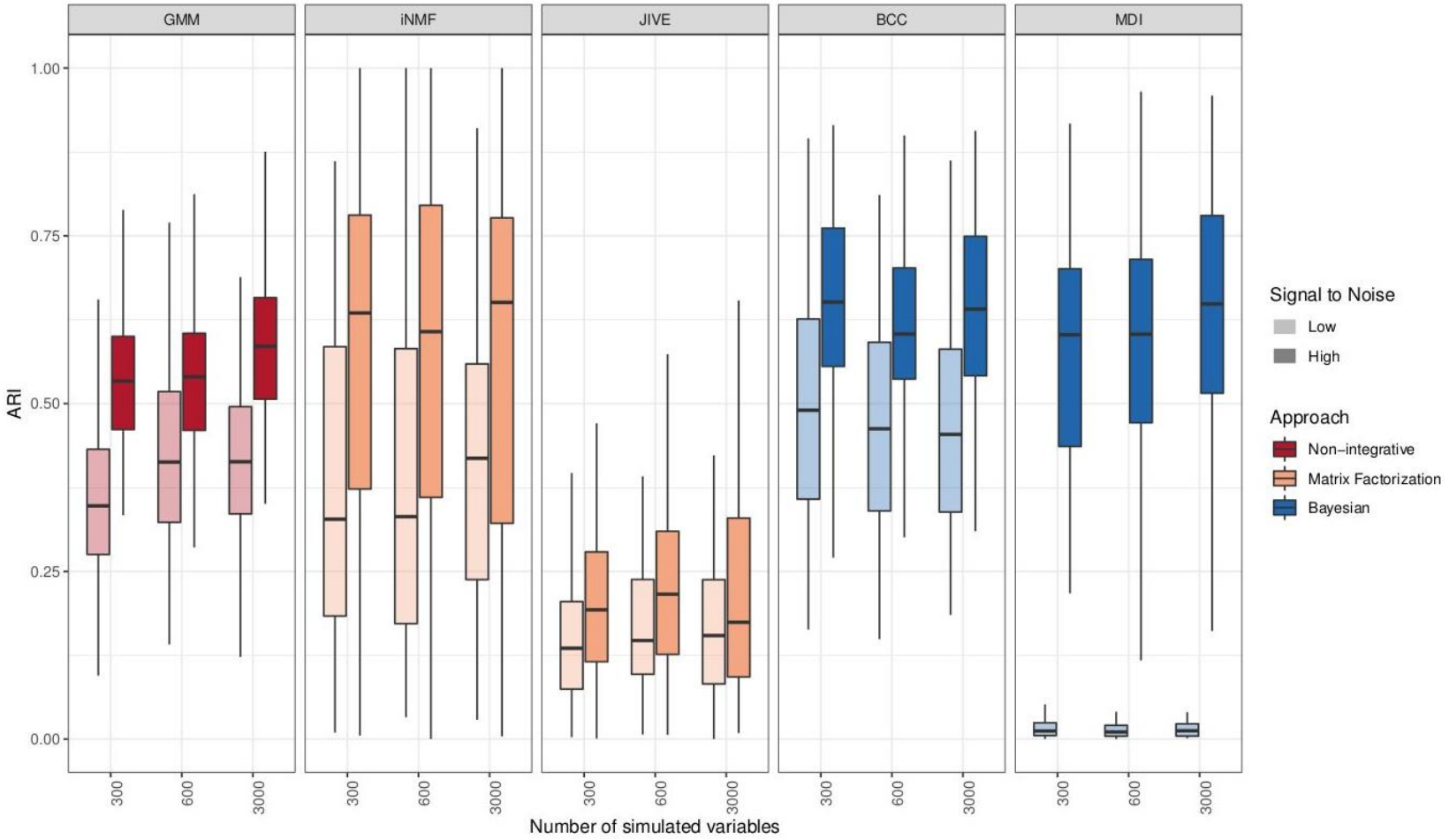
- Design of the high dimension study

On the iNMF-overlap scenario

  - 300, 600 and 3000 variables
  - 60 samples
  - 3 common clusters of 20 samples each
  - 2 levels of signal to noise ratio
  - 100 repetitions

High dimension study – no impact of the data set size

- Shared clustering: same performances and ranking as in the low dimension case (not shown here)
- Specific clustering not impacted by the sample size of the data set:



Run times

Method	Time (sec)
moCluster	0.5
Consensus clustering	1.3
GMM	1.3
Concatenation	1.6
iCluster	16.0
JIVE	111.9
iNMF	102.6
BCC	1441.4
MDI	3810.6

Main study

Method	Time (sec)
moCluster	0.1
Consensus clustering	34.7
GMM	34.7
Concatenation	58.9
iCluster	194.6
JIVE	20.3
iNMF	1056.3
BCC	14300.6
MDI	63153.4

High-dimension study

Conclusion on the simulations

- For shared structures, iCluster, moCluster and iNMF have good clustering performances
- For specific structures, only BCC reaches the performances of non-integrative methods
- No method can well detect both shared and specific structures at the same time
- No impact of the number of features in the datasets
- Ranking supported as well by a sensitivity simulation study.

APPLICATION ON TCGA DATA

TCGA breast cancer data

- Four omics measured on 348 patients:
  - mRNA
  - miRNA
  - DNA methylation
  - proteins

- Practitioners divide patients into 4 subtypes based on:
  - expression of proliferating protein Ki67
  - receptor status for estrogen (ER)
  - receptor status for progesterone (PR)
  - receptor status for human epidermal growth factor 2 (HER2)

- Comparison of these classes with clusters from integrative methods

Subtype	Markers Status
Basal	ER- PR- HER2-
HER2	ER- PR- HER2+
Luminal A	ER+ and/or PR+ HER2-
Luminal B	ER+ and/or PR+ HER2+ or High Ki67

TCGA breast cancer data

Method	%ER	%PR	%HER2	ARI
Consensus clustering	97	89	11	0.52
	66	45	38	
	11	5	2	
	96	80	16	
Concatenation	97	89	7	0.52
	98	77	22	
	13	6	2	
	63	44	41	
iCluster	95	71	26	0.42
	96	85	8	
	90	81	18	
	12	5	9	
moCluster	13	6	2	0.57
	98	83	8	
	64	40	56	
	96	89	9	
iNMF	8	56	41	0.56
	97	9	6	
	13	6	7	
	100	87	8	
JIVE	96	84	9	0.40
	12	4	7	
	99	84	6	
	79	63	39	
BCC	70	49	43	0.51
	18	9	4	
	97	84	10	
	98	89	9	
MDI	94	87	10	0.55
	14	8	3	
	100	94	19	
	99	83	10	

mRNAs	-	0.50
DNA methylations	-	0.41
miRNAs	-	0.30
Proteins	-	0.00

- Performances of the single omics vary: impact of the number of features or biological explanation?
- All integrative methods but iCluster and JIVE overpass single omics
- Ranking of the methods :
  1. moCluster, iNMF (consistent with simulations)
  2. MDI, BCC, Non-integrative
  3. iCluster, JIVE (different from simulations)
- Limit of the comparison:
  - Classification used as gold standard in clinics but no « true » classes
  - Very low prevalence of HER2 subclass ➔ difficult to detect

CONCLUSIONS

Key points

- The integration of multiple omics shows a clear improvement in clustering performance as compared to non-integrative methods
- Matrix factorization methods are on average better at identifying shared clusters (especially moCluster and iNMF).
- Although iNMF showed a lack of sensitivity, it can finely be tuned to recover either common or specific clusters.
- Despite moderate performances on shared clusters, BCC displayed the best ability to recover both structures.
- MDI highly impacted by noise.
- Bayesian methods easier to parametrize, but longer to run.
- It would be interesting to study variable selection (available in iCluster, moCluster, JIVE and iNMF)

References

1. Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings in Bioinformatics* 2019;bbz015

2. Tini G, Marchetti L, Priami C, et al. Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings in Bioinformatics* 2017;167.

3. Shen R, Wang S, Mo Q. Sparse integrative clustering of multiple omics data sets. *Annals Appl Stat* 2013;7(1):269.

4. Meng C, Helm D, Frejno M, *et al.* moCluster: identifying joint patterns across multiple omics data sets. *J Proteome Res* 2016;15(3):755–65.

5. Lock EF, Hoadley KA, Marron JS, *et al.* Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat* 2013;7(1):523–42.

6. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 2016;32(1):1–8.

7. Kirk P, Griffin JE, Savage RS, *et al.* Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 2012;28(24):3290–7.

8. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics* 2013;29(20):2610–6.

Thank you for your attention

Sensitivity study

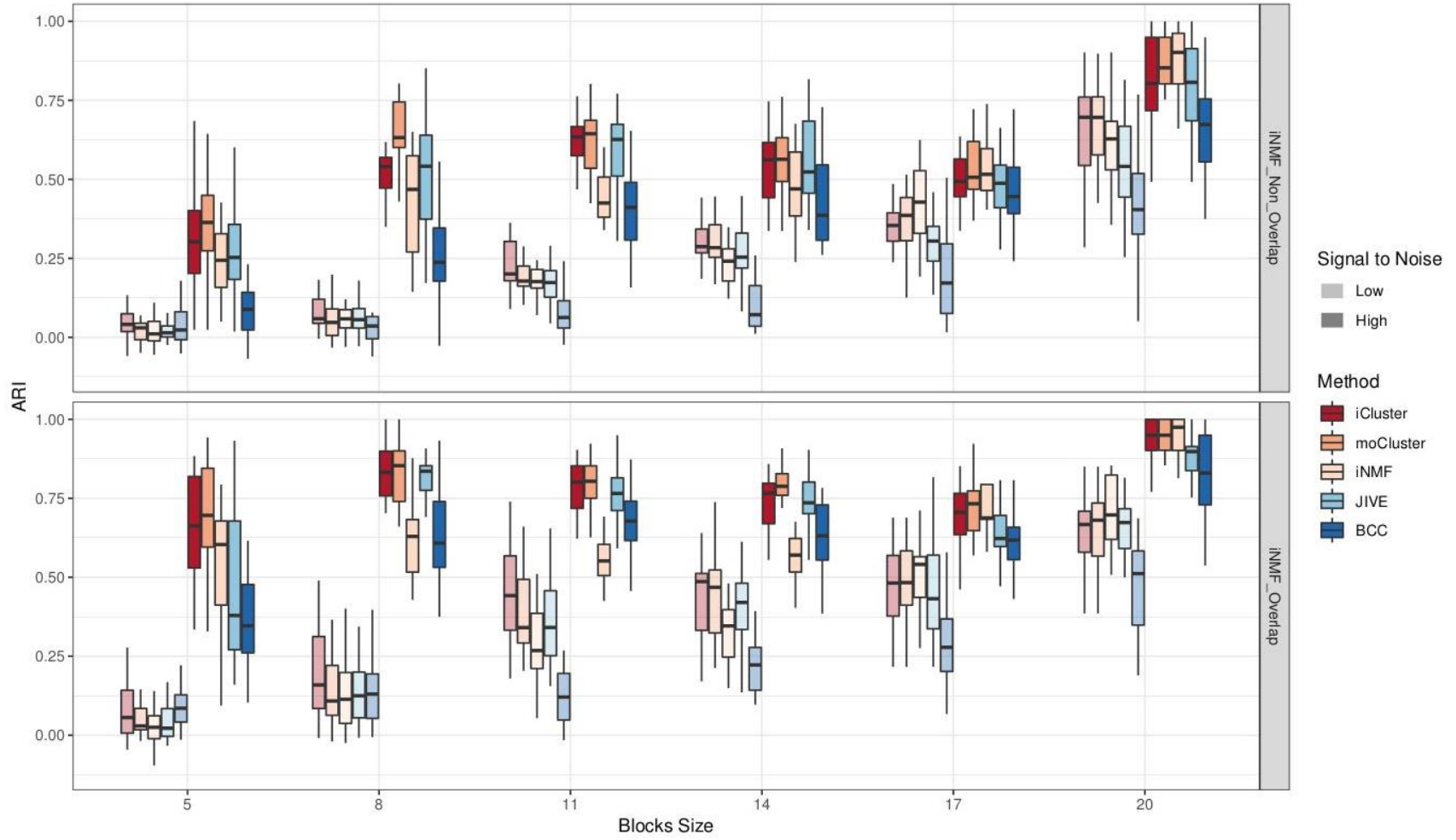
- Design of the sensitivity study

On iNMF scenarios

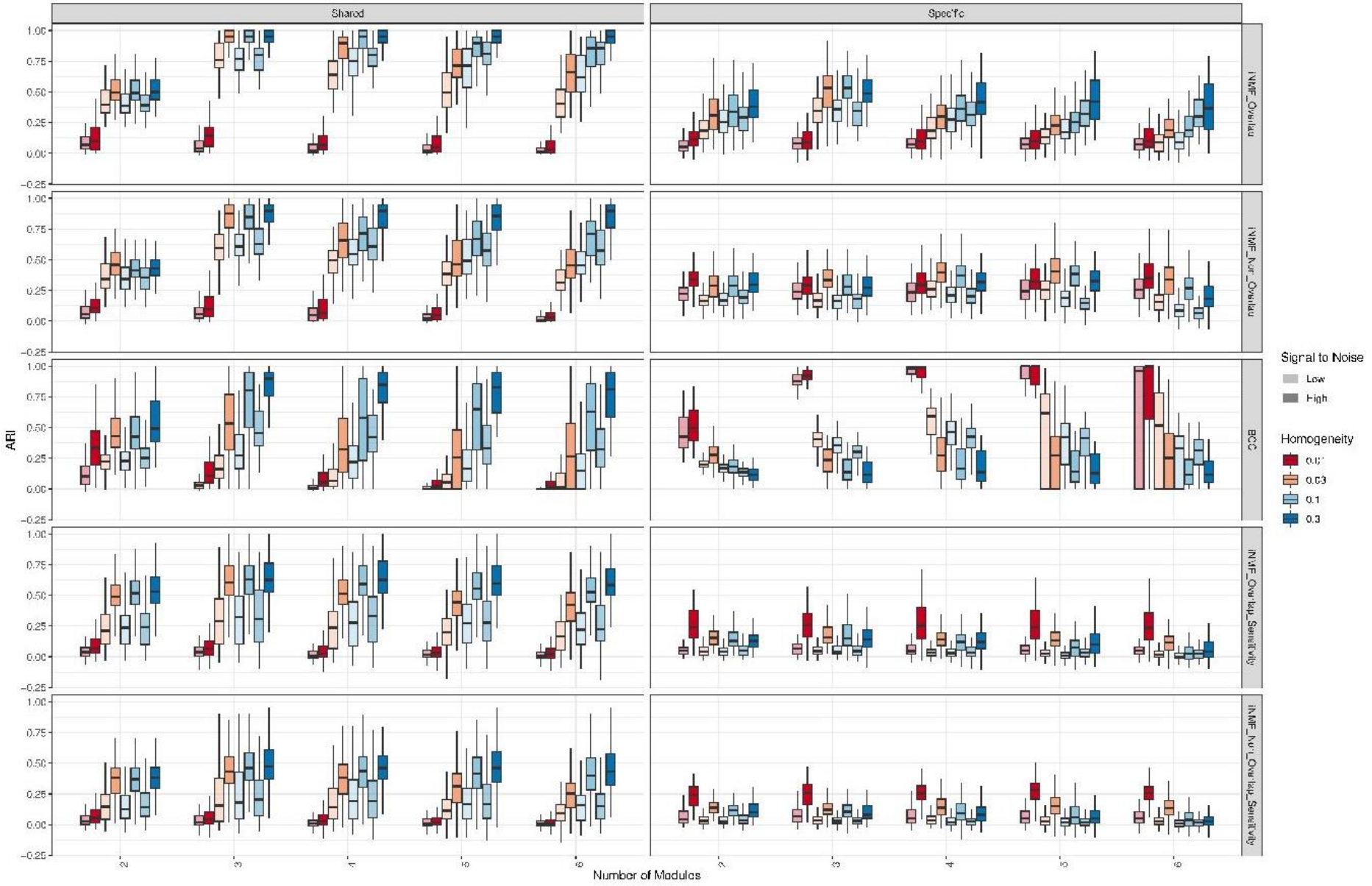
- 180, 210 and 240 variables
- 60 samples, in which 3 blocks of {15, 12, 9, 6, 3, 0} samples are noise
- 3 common clusters of 20 samples each
- 2 levels of signal to noise ratio
- 20 repetitions

Sensitivity study

- SNR, simulation design (overlaps or not) and cluster sizes impact ARI
- Methods ranking:
  1. iCluster, moCluster, JIVE
  2. iNMF, BCC
  3. No results for MDI (too sensitive to noise)



Grid search on parameters for iNMF



BIOASTER
MICROBIOLOGY TECHNOLOGY INSTITUTE

www.bioaster.org