BLOCKCLUST

CLUSTERING ET CLASSIFICATION EFFICACE DES ARNS NON CODANT TIRÉS DES PROFILS D'RNA-SEQ

UNIVERSITE PAUL SABATIER TOULOUSE III

Chazalviel
Maxime

INTRODUCTION

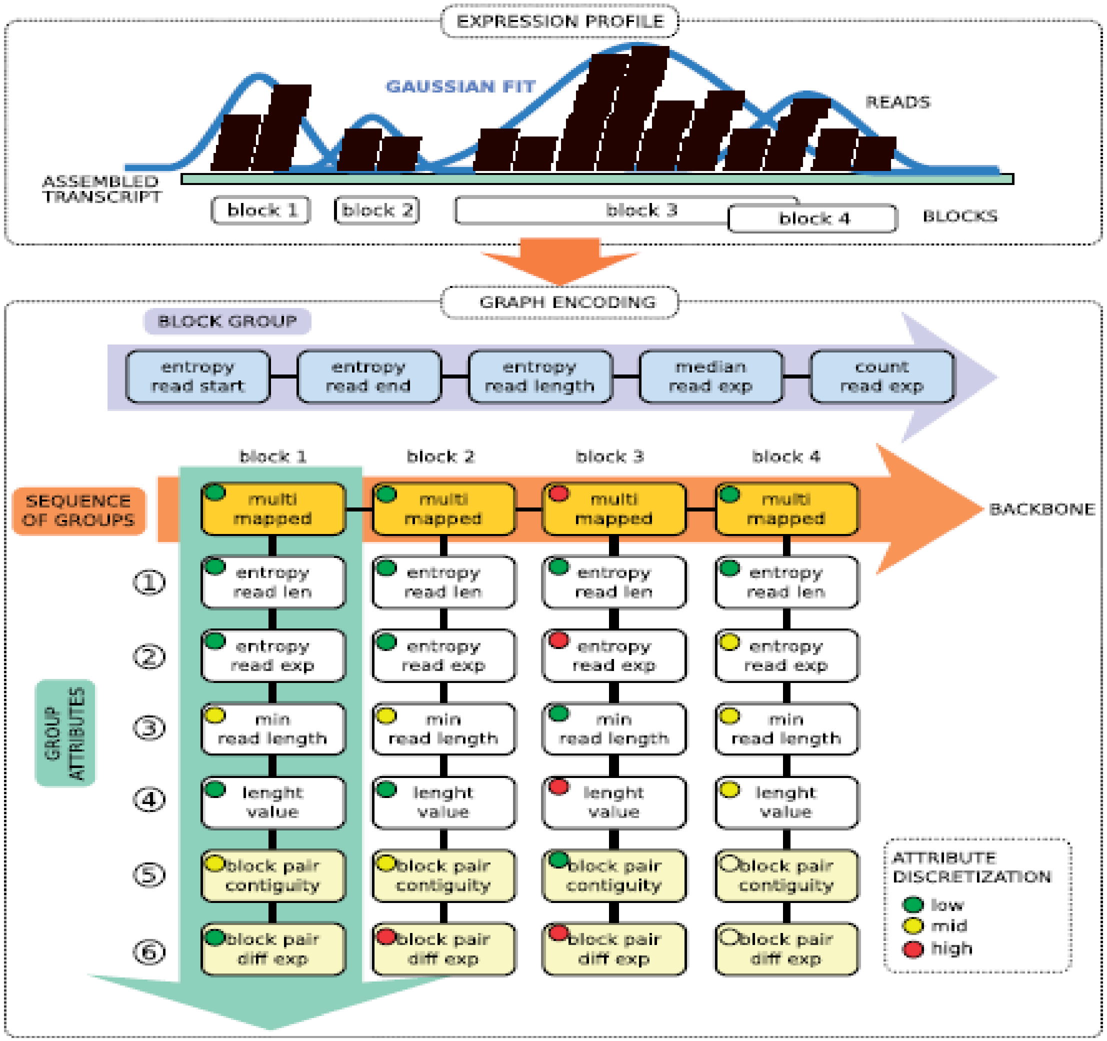
- ARNs non codant
- Comment trouver leur fonction?
  - clustering des transcrits selon la séquence et la structure secondaire
- L'approche BlockClust

DÉMARCHE

- Features
- Graph-kernel
- Encodé les profile d’expression
- Génération de caractéristiques combinatoires

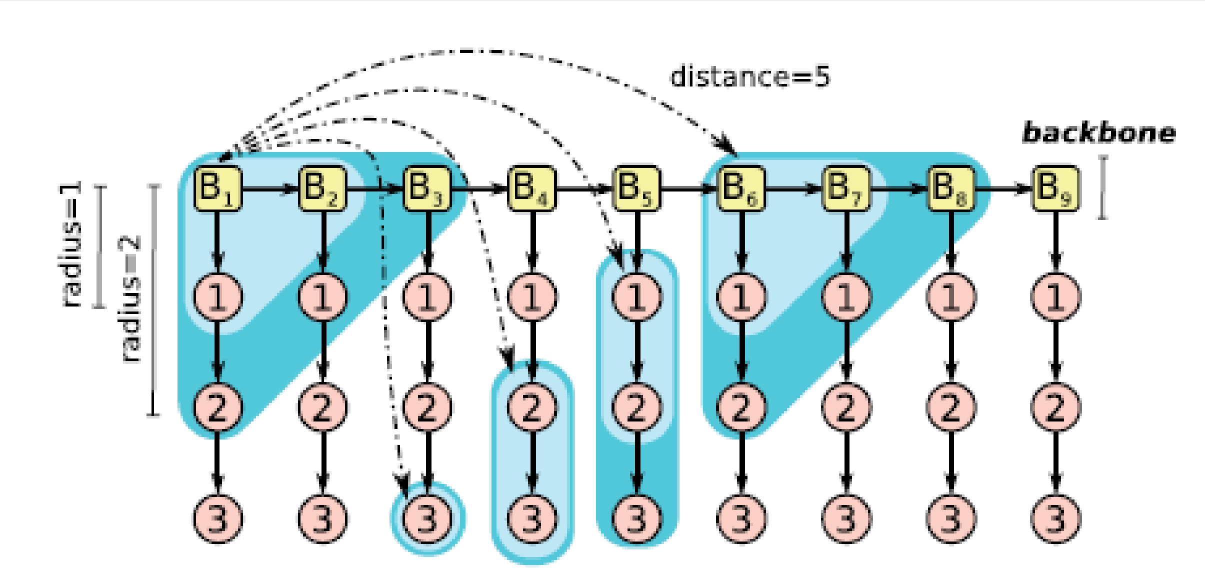
CODAGE DES PROFILS D’EXPRESSION

- Read profile encoding



GÉNÉRATION DE CARACTÉRISTIQUES COMBINATOIRES

- Caractéristiques combinatoires



CLUSTERING ET CLASSIFICATION DES PROFILS D'EXPRESSIONS DES ARN NON-CODANT

- Clustering
  - NSPDK utiliser par l'algorithme de clustering
  - Markov Cluster Process

- Classification
  - Technique linéaire modulable

RÉSULTATS & DISCUSSIONS

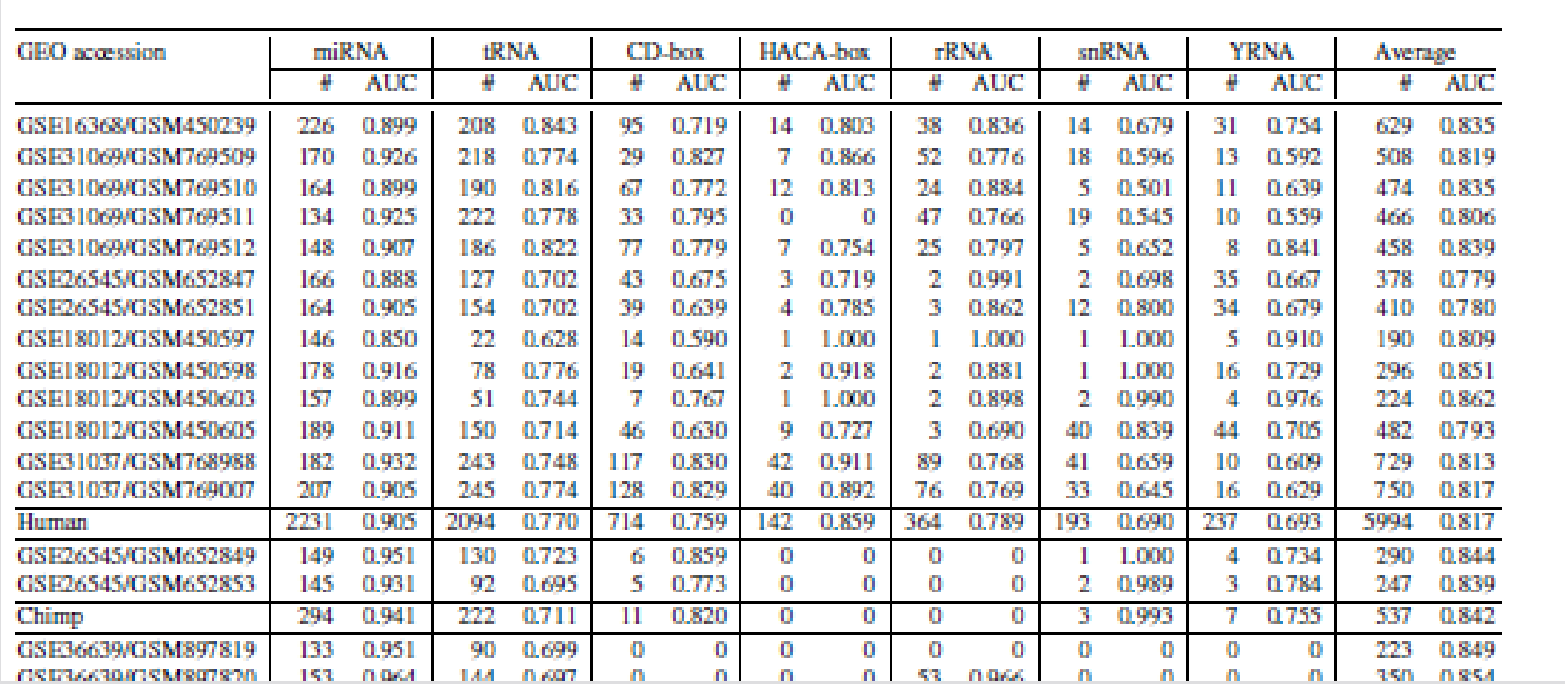
- Q1 : Clustering d’ARNs non codant avec des profils d’expression encodé
- Q2 : Robustesse et modularité
- Q3 : Annotation des ARNs nc avec des profils d’expression encodé
- Q4 : Comparaison des performances
- Analyse des clusters d’ARNnc connus

Q1 : CLUSTERING D'ARNS NON CODANT AVEC DES PROFILS D'EXPRESSION ENCODÉ

- Différentes mesures de performances et résultats
  - Score de similarité
  - AUC ROC
  - Pureté
- Ensemble de données
  - NGS data generated
- Optimisation des paramètres
  - BlockClust
    - Block identification
    - Codage du graphe
    - Clustering ou classification
  - Blockbuster
    - Grain resolution
    - Résolution de discrétisation

Q2 : ROBUSTNESS AND RANGE OF APPLICABILITY

- Performance du clustering



GEO accession	miRNA		tRNA		CD-box		HACA-box		rRNA		snRNA		YRNA		Average	
	#	AUC	#	AUC	#	AUC	#	AUC	#	AUC	#	AUC	#	AUC	#	AUC
GSE16368/GSM450239	226	0.899	208	0.843	95	0.719	14	0.803	38	0.836	14	0.679	31	0.754	629	0.835
GSE31069/GSM769509	170	0.926	218	0.774	29	0.827	7	0.866	52	0.776	18	0.596	13	0.592	508	0.819
GSE31069/GSM769510	164	0.899	190	0.816	67	0.772	12	0.813	24	0.884	5	0.501	11	0.639	474	0.835
GSE31069/GSM769511	134	0.925	222	0.778	33	0.795	0	0	47	0.766	19	0.545	10	0.559	466	0.806
GSE31069/GSM769512	148	0.907	186	0.822	77	0.779	7	0.754	25	0.797	5	0.652	8	0.841	458	0.839
GSE26545/GSM652847	166	0.888	127	0.702	43	0.675	3	0.719	2	0.991	2	0.698	35	0.667	378	0.779
GSE26545/GSM652851	164	0.905	154	0.702	39	0.639	4	0.785	3	0.862	12	0.800	34	0.679	410	0.780
GSE18012/GSM450597	146	0.850	22	0.628	14	0.590	1	1.000	1	1.000	1	1.000	5	0.910	190	0.809
GSE18012/GSM450598	178	0.916	78	0.776	19	0.641	2	0.918	2	0.881	1	1.000	16	0.729	296	0.851
GSE18012/GSM450603	157	0.899	51	0.744	7	0.767	1	1.000	2	0.898	2	0.990	4	0.976	224	0.862
GSE18012/GSM450605	189	0.911	150	0.714	46	0.630	9	0.727	3	0.690	40	0.839	44	0.705	482	0.793
GSE31037/GSM768988	182	0.932	243	0.748	117	0.830	42	0.911	89	0.768	41	0.659	10	0.609	729	0.813
GSE31037/GSM769007	207	0.905	245	0.774	128	0.829	40	0.892	76	0.769	33	0.645	16	0.629	750	0.817
Human	2231	0.905	2094	0.770	714	0.759	142	0.859	364	0.789	193	0.690	237	0.693	5994	0.817
GSE26545/GSM652849	149	0.951	130	0.723	6	0.859	0	0	0	0	1	1.000	4	0.734	290	0.844
GSE26545/GSM652853	145	0.931	92	0.695	5	0.773	0	0	0	0	2	0.989	3	0.784	247	0.839
Chimp	294	0.941	222	0.711	11	0.820	0	0	0	0	3	0.993	7	0.755	537	0.842
GSE36639/GSM897819	133	0.951	90	0.699	0	0	0	0	0	0	0	0	0	0	223	0.849
GSE36639/GSM897820	153	0.964	144	0.697	0	0	0	0	53	0.966	0	0	0	0	350	0.854

Q3 : ANNOTATION DES ARNS NC AVEC DES PROFILS D'EXPRESSION ENCODÉ

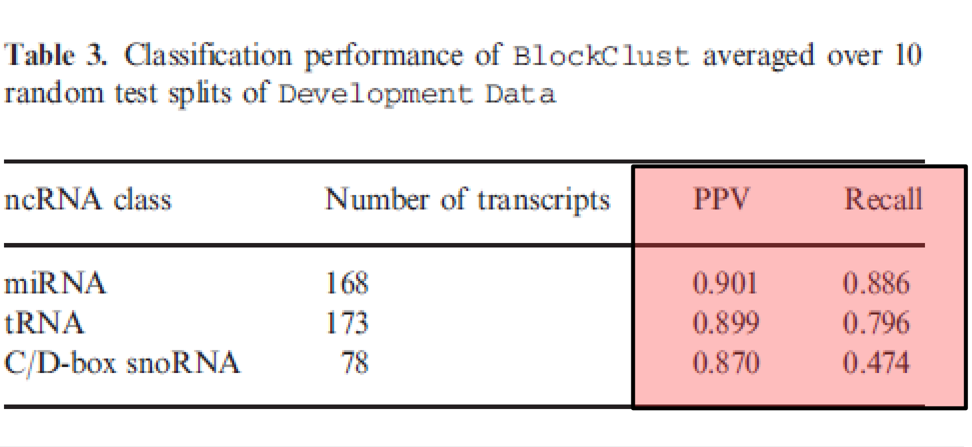


Table 3. Classification performance of BlockClust averaged over 10 random test splits of Development Data

ncRNA class	Number of transcripts	PPV	Recall
miRNA	168	0.901	0.886
tRNA	173	0.899	0.796
C/D-box snoRNA	78	0.870	0.474

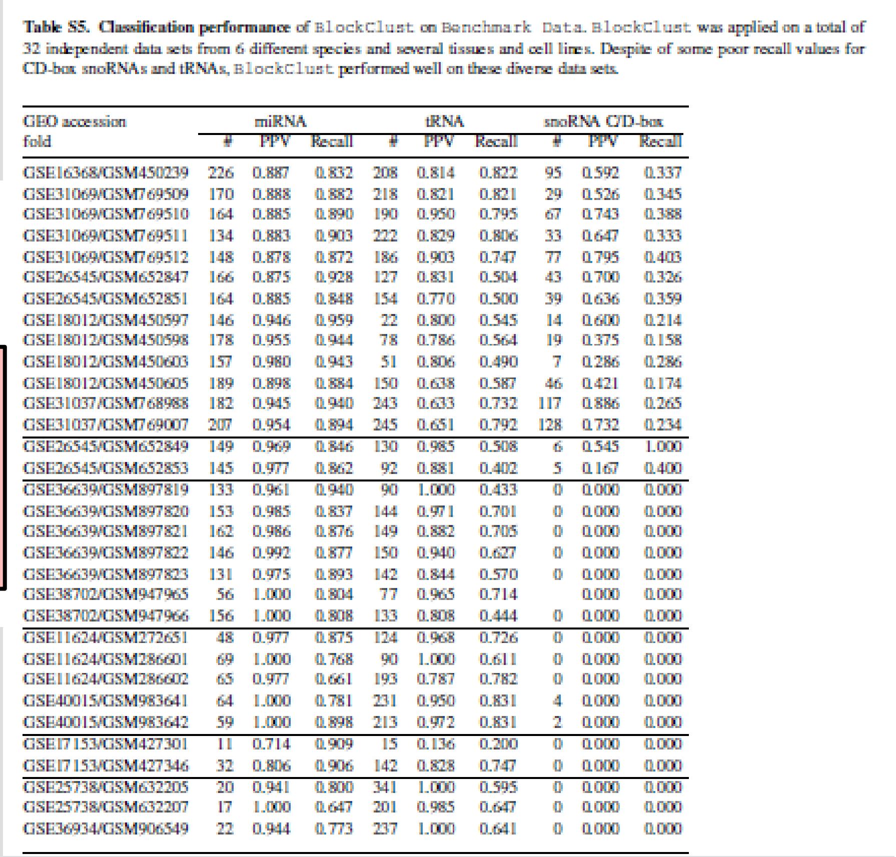


Table S5. Classification performance of BlockClust on Benchmark Data. BlockClust was applied on a total of 32 independent data sets from 6 different species and several tissues and cell lines. Despite of some poor recall values for CD-box snoRNAs and tRNAs, BlockClust performed well on these diverse data sets.

GEO accession	miRNA			tRNA			snoRNA C/D-box		
fold	#	PPV	Recall	#	PPV	Recall	#	PPV	Recall
GSE16368/GSM450239	226	0.887	0.832	208	0.814	0.822	95	0.592	0.337
GSE31069/GSM769509	170	0.888	0.882	218	0.821	0.821	29	0.526	0.345
GSE31069/GSM769510	164	0.885	0.890	190	0.950	0.795	67	0.743	0.388
GSE31069/GSM769511	134	0.883	0.903	222	0.829	0.806	33	0.647	0.333
GSE31069/GSM769512	148	0.878	0.872	186	0.903	0.747	77	0.795	0.403
GSE26545/GSM652847	166	0.875	0.928	127	0.831	0.504	43	0.700	0.326
GSE26545/GSM652851	164	0.885	0.848	154	0.770	0.500	39	0.636	0.359
GSE18012/GSM450597	146	0.946	0.959	22	0.800	0.545	14	0.600	0.214
GSE18012/GSM450598	178	0.955	0.944	78	0.786	0.564	19	0.375	0.158
GSE18012/GSM450603	157	0.980	0.943	51	0.806	0.490	7	0.286	0.286
GSE18012/GSM450605	189	0.898	0.884	150	0.638	0.587	46	0.421	0.174
GSE31037/GSM768988	182	0.945	0.940	243	0.633	0.732	117	0.886	0.265
GSE31037/GSM769007	207	0.954	0.894	245	0.651	0.792	128	0.732	0.234
GSE26545/GSM652849	149	0.969	0.846	130	0.985	0.508	6	0.545	1.000
GSE26545/GSM652853	145	0.977	0.862	92	0.881	0.402	5	0.167	0.400
GSE36639/GSM897819	133	0.961	0.940	90	1.000	0.433	0	0.000	0.000
GSE36639/GSM897820	153	0.985	0.837	144	0.971	0.701	0	0.000	0.000
GSE36639/GSM897821	162	0.986	0.876	149	0.882	0.705	0	0.000	0.000
GSE36639/GSM897822	146	0.992	0.877	150	0.940	0.627	0	0.000	0.000
GSE36639/GSM897823	131	0.975	0.893	142	0.844	0.570	0	0.000	0.000
GSE38702/GSM947965	56	1.000	0.804	77	0.965	0.714		0.000	0.000
GSE38702/GSM947966	156	1.000	0.808	133	0.808	0.444	0	0.000	0.000
GSE11624/GSM272651	48	0.977	0.875	124	0.968	0.726	0	0.000	0.000
GSE11624/GSM286601	69	1.000	0.768	90	1.000	0.611	0	0.000	0.000
GSE11624/GSM286602	65	0.977	0.661	193	0.787	0.782	0	0.000	0.000
GSE40015/GSM983641	64	1.000	0.781	231	0.950	0.831	4	0.000	0.000
GSE40015/GSM983642	59	1.000	0.898	213	0.972	0.831	2	0.000	0.000
GSE17153/GSM427301	11	0.714	0.909	15	0.136	0.200	0	0.000	0.000
GSE17153/GSM427346	32	0.806	0.906	142	0.828	0.747	0	0.000	0.000
GSE25738/GSM632205	20	0.941	0.800	341	1.000	0.595	0	0.000	0.000
GSE25738/GSM632207	17	1.000	0.647	201	0.985	0.647	0	0.000	0.000
GSE36934/GSM906549	22	0.944	0.773	237	1.000	0.641	0	0.000	0.000

Q4 : COMPARAISON DES PERFORMANCES

- Performance : BlockClust vs. deepBlockAlign

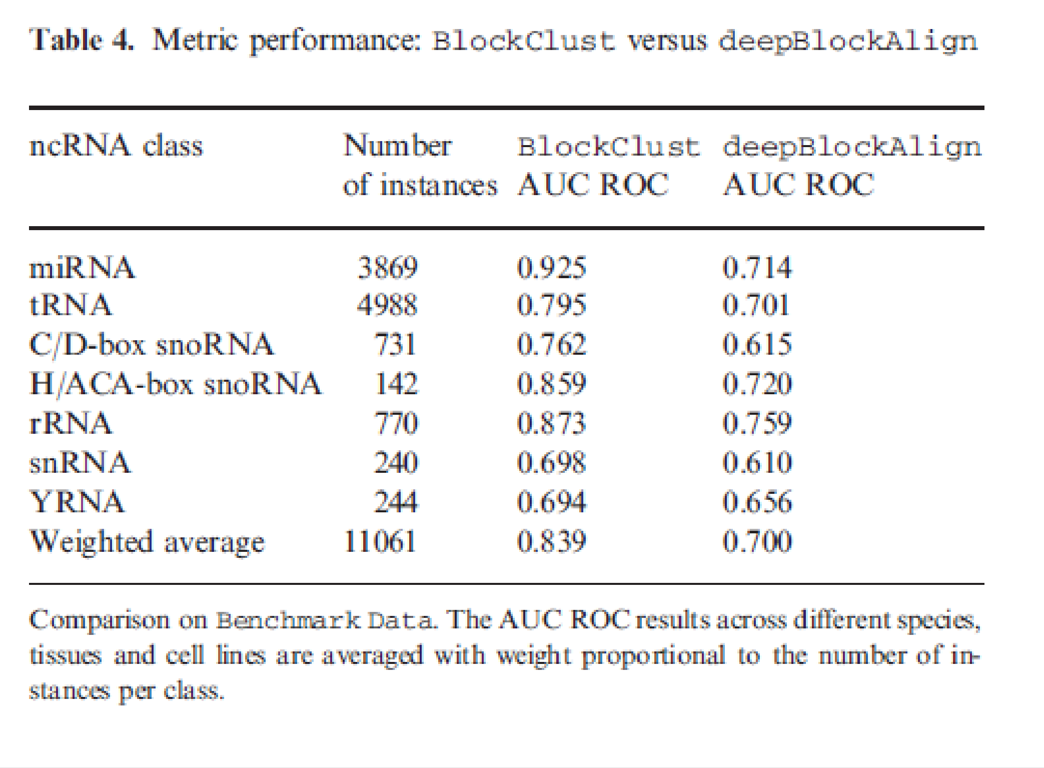


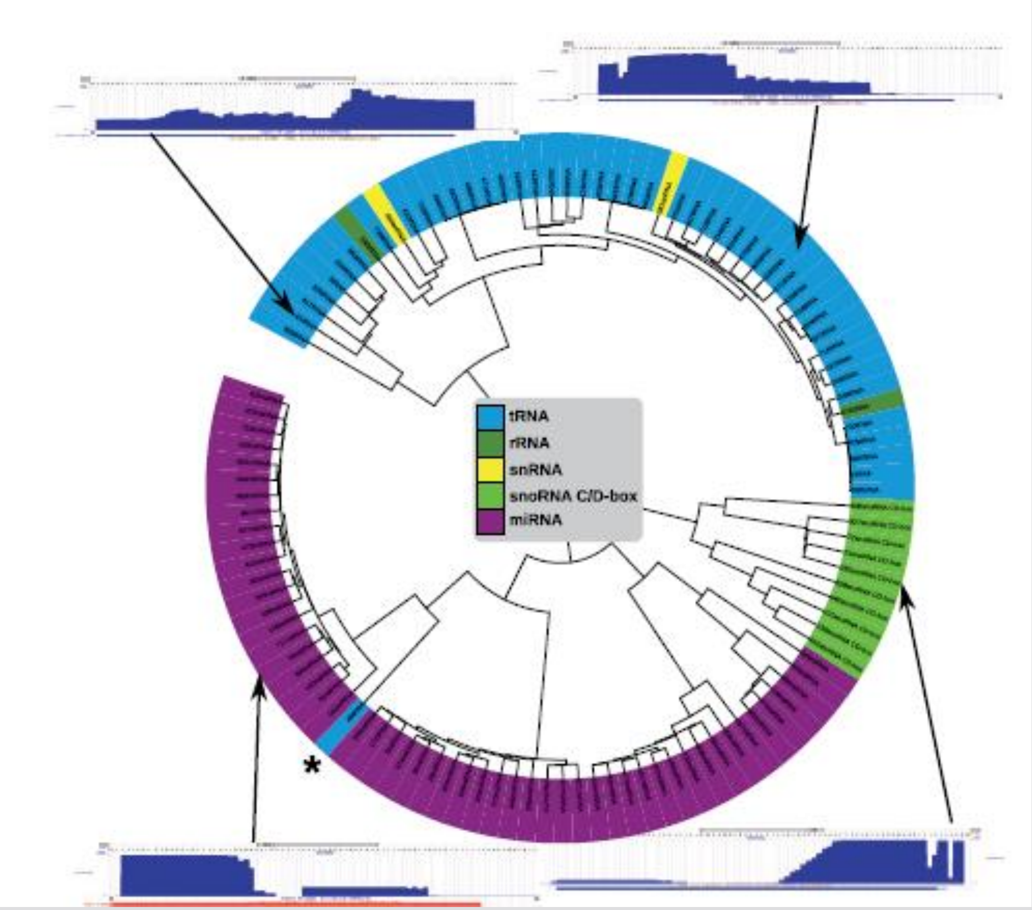
Table 4. Metric performance: `BlockClust` versus `deepBlockAlign`

ncRNA class	Number of instances	`BlockClust` AUC ROC	`deepBlockAlign` AUC ROC
miRNA	3869	0.925	0.714
tRNA	4988	0.795	0.701
C/D-box snoRNA	731	0.762	0.615
H/ACA-box snoRNA	142	0.859	0.720
rRNA	770	0.873	0.759
snRNA	240	0.698	0.610
YRNA	244	0.694	0.656
Weighted average	11061	0.839	0.700

Comparison on `Benchmark Data`. The AUC ROC results across different species, tissues and cell lines are averaged with weight proportional to the number of instances per class.

ANALYSE DES CLUSTERS D'ARN NC CONNUS

- Dendrogram des clusters



CONCLUSION

- Approche efficace pour la détection de transcrit
- Codez les profils d'expression en structures discrète
- Robuste
- Modulable
- Fiable
- Nouveaux packages