OMTN, Volume 22

Supplemental Information

m5UPred: A Web Server for the Prediction of RNA 5-Methyluridine Sites from Sequences

Jie Jiang, Bowen Song, Yujiao Tang, Kunqi Chen, Zhen Wei, and Jia Meng

Table S1. Overlapped Positive m5U Sites Between Different Techniques and Cell Types

Separation Method	Mode	Condition	Site #	Overlap #	Total #
Technique	Full	miCLIP-Seq	2,225	521	3,696
		FICC-Seq	1,471		
Cell type	Full	HEK293	2,467	732	3,696
		HAP1	1,229		

Table S2. Positive m5U Sites in Different RNAs Families

Gene region	CDS	Intergenic	Intronic	ncRNA_intronic	Ipstream	3'UTR	5'UTR
Number of sites	175	1,880	654	232	277	50	29

Table S3. Performance Evaluation of m5UPred by Cross-technique and Cross-cell Type Validation Using Different Machine Learning Classifiers

	Mode	Classifier	Sn (%)	Sp (%)	ACC (%)	MCC	AUC
Cross-technique validation	Full transcript	SVM	75.87%	85.47%	80.67%	0.616	**0.896**
		RF	69.61%	85.08%	77.35%	0.554	0.870
		NB	80.99%	57.76%	69.37%	0.400	0.785
		GLM	77.51%	81.07%	79.29%	0.586	0.876
	Mature mRNA	SVM	88.48%	89.05%	88.77%	0.775	**0.962**
		RF	81.42%	92.69%	87.05%	0.746	0.955
		NB	91.83%	52.81%	72.32%	0.485	0.855
		GLM	89.63%	84.43%	87.03%	0.742	0.949
Cross-cell type validation	Full transcript	SVM	80.13%	85.98%	83.06%	0.662	**0.918**
		RF	73.05%	85.63%	79.34%	0.592	0.890
		NB	81.64%	59.12%	70.38%	0.420	0.805
		GLM	80.31%	81.64%	80.98%	0.620	0.894
	Mature mRNA	SVM	93.87%	86.15%	90.01%	0.803	**0.970**
		RF	89.15%	92.36%	90.76%	0.816	0.969
		NB	91.54%	53.08%	72.31%	0.483	0.860
		GLM	93.63%	81.04%	87.34%	0.753	0.953

Note: We randomly selected 80% of experimentally validated m5U sites as training dataset and the performance of predictors were evaluated by the rest of 20% of m5U sites as independent testing data.

Table S4. Whole Dataset Performance evaluation

Full	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.960	0.960	0.955	0.953	0.950	0.957	0.950	0.954	0.961	0.956	0.956
MCC	0.784	0.781	0.770	0.755	0.747	0.769	0.748	0.773	0.781	0.763	0.767
SEN	89.72%	87.82%	88.23%	87.69%	86.20%	89.45%	86.60%	87.01%	89.17%	87.14%	87.90%
SPE	88.63%	90.26%	88.77%	87.82%	88.50%	87.42%	88.23%	90.26%	88.90%	89.17%	88.80%
ACC	89.17%	89.04%	88.50%	87.75%	87.35%	88.43%	87.42%	88.63%	89.04%	88.16%	88.35%

Table S5. Whole Dataset Performance evaluation

Mature	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.957	0.959	0.937	0.956	0.960	0.954	0.943	0.960	0.955	0.961	0.954
MCC	0.786	0.794	0.765	0.793	0.773	0.805	0.789	0.814	0.819	0.811	0.795
SEN	86.18%	86.99%	85.37%	89.84%	87.40%	89.02%	87.40%	87.80%	87.40%	86.99%	87.44%
SPE	92.28%	92.28%	91.06%	89.43%	89.84%	91.46%	91.46%	93.50%	94.31%	93.90%	91.95%
ACC	89.23%	89.63%	88.21%	89.63%	88.62%	90.24%	89.43%	90.65%	90.85%	90.45%	89.70%

Table S6. miCLIP_F_train&FICC_F_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.908	0.915	0.907	0.912	0.912	0.908	0.910	0.916	0.904	0.911	0.910
MCC	0.632	0.650	0.651	0.663	0.649	0.659	0.637	0.683	0.637	0.662	0.652
SEN	74.03%	76.00%	75.80%	75.32%	73.62%	75.39%	74.64%	77.36%	75.12%	76.27%	75.36%
SPE	88.51%	88.51%	88.72%	90.21%	90.35%	89.80%	88.44%	90.35%	88.04%	89.33%	89.23%
ACC	81.27%	82.26%	82.26%	82.77%	81.99%	82.60%	81.54%	83.85%	81.58%	82.80%	82.29%

[a] miCLIP_F dataset as train dataset and FICC_F dataset as an independent dataset

Table S7. miCLIP_M_train&FICC_M_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.972	0.971	0.967	0.978	0.966	0.969	0.970	0.984	0.961	0.966	0.970
MCC	0.824	0.822	0.822	0.854	0.800	0.792	0.797	0.844	0.743	0.798	0.809
SEN	89.98%	91.93%	91.93%	91.20%	90.22%	89.98%	88.75%	92.42%	86.31%	88.02%	90.07%
SPE	92.42%	90.22%	90.22%	94.13%	89.73%	89.24%	90.95%	91.93%	88.02%	91.69%	90.86%
ACC	91.20%	91.08%	91.08%	92.67%	89.98%	89.61%	89.85%	92.18%	87.16%	89.85%	90.46%

[a] miCLIP_M dataset as train dataset and FICC_M dataset as an independent dataset

Table S8. FICC_F_train&miCLIP_F_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.852	0.853	0.851	0.852	0.859	0.844	0.860	0.858	0.849	0.851	0.853
MCC	0.487	0.504	0.494	0.493	0.494	0.476	0.509	0.500	0.488	0.502	0.495
SEN	56.58%	57.75%	56.40%	55.87%	56.76%	55.24%	56.45%	58.11%	54.97%	56.63%	56.48%
SPE	89.44%	89.93%	90.07%	90.38%	89.89%	89.48%	91.24%	89.39%	90.61%	90.61%	90.10%
ACC	73.01%	73.84%	73.24%	73.12%	73.33%	72.36%	73.84%	73.75%	72.79%	73.62%	73.29%

[a] FICC_F dataset as train dataset and miCLIP_F dataset as an independent dataset

Table S9. FICC_M_train&miCLIP_M_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.881	0.884	0.866	0.877	0.868	0.883	0.881	0.857	0.868	0.863	0.873
MCC	0.441	0.460	0.424	0.475	0.438	0.457	0.448	0.432	0.468	0.445	0.449
SEN	40.10%	38.40%	37.67%	42.77%	35.72%	40.58%	36.33%	38.40%	40.10%	38.03%	38.81%
SPE	96.35%	98.42%	96.60%	97.08%	98.42%	97.08%	98.66%	96.72%	98.06%	97.69%	97.51%
ACC	68.23%	68.41%	67.13%	69.93%	67.07%	68.83%	67.50%	67.56%	69.08%	67.86%	68.16%

[a] FICC_M dataset as train dataset and miCLIP_M dataset as an independent dataset

Table S10. HEK293_F_train&HAP1_F_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.940	0.940	0.933	0.940	0.940	0.947	0.946	0.942	0.943	0.943	0.941
MCC	0.715	0.729	0.721	0.720	0.729	0.737	0.734	0.717	0.720	0.737	0.726
SEN	83.48%	82.18%	82.83%	83.40%	83.65%	82.91%	82.51%	81.61%	80.55%	84.78%	82.79%
SPE	87.96%	90.48%	89.10%	88.53%	89.18%	90.56%	90.64%	89.83%	91.05%	88.85%	89.62%
ACC	85.72%	86.33%	85.96%	85.96%	86.41%	86.74%	86.57%	85.72%	85.80%	86.82%	86.20%

[a] HEK293_F dataset as train dataset and HAP1_F test dataset as an independent dataset

Table S11. HEK293_M_train&HAP1_M_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.981	0.985	0.981	0.982	0.979	0.981	0.982	0.980	0.978	0.976	0.981
MCC	0.842	0.830	0.863	0.871	0.829	0.848	0.848	0.851	0.827	0.842	0.845
SEN	95.33%	95.88%	94.51%	95.05%	95.05%	95.33%	95.88%	96.15%	95.60%	95.33%	95.41%
SPE	88.74%	86.81%	91.76%	92.03%	87.64%	89.29%	88.74%	88.74%	86.81%	88.74%	88.93%
ACC	92.03%	91.35%	93.13%	93.54%	91.35%	92.31%	92.31%	92.45%	91.21%	92.03%	92.17%

[a] HEK293_M dataset as train dataset and HAP1_M test dataset as an independent dataset

Table S12. HAP1_F_train&HEK293_F_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.854	0.848	0.860	0.848	0.850	0.857	0.855	0.861	0.861	0.872	0.857
MCC	0.504	0.493	0.505	0.487	0.510	0.509	0.508	0.504	0.524	0.530	0.507
SEN	55.17%	58.94%	56.51%	56.06%	57.80%	57.44%	58.82%	59.02%	59.99%	58.01%	57.77%
SPE	91.77%	88.20%	90.88%	89.83%	90.43%	90.56%	89.50%	89.06%	89.99%	91.85%	90.21%
ACC	73.47%	73.57%	73.69%	72.94%	74.12%	74.00%	74.16%	74.04%	74.99%	74.93%	73.99%

[a] HAP1_F dataset as train dataset and HEK293_F test dataset as an independent dataset

Table S13. HAP1_M_train&HEK293_M_test

	1	2	3	4	5	6	7	8	9	10	Average
AUROC	0.863	0.877	0.870	0.867	0.875	0.869	0.894	0.854	0.861	0.878	0.871
MCC	0.450	0.470	0.456	0.478	0.454	0.463	0.457	0.448	0.446	0.483	0.461
SEN	41.24%	38.02%	39.06%	42.28%	39.06%	40.44%	36.87%	36.52%	37.21%	39.40%	39.01%
SPE	96.31%	99.19%	97.81%	97.58%	97.70%	97.58%	98.96%	98.62%	98.16%	99.31%	98.12%
ACC	68.78%	68.61%	68.43%	69.93%	68.38%	69.01%	67.91%	67.57%	67.68%	69.35%	68.57%

[a] HAP1_M dataset as train dataset and HEK293_M test dataset as an independent dataset

Table S14. Performance Evaluation by FDR and FOR at Different Thresholds

Mode	Threshold	TPR	FDR	FOR
Full Transcript	0.1	0.985	0.322	0.027
	0.2	0.970	0.262	0.043
	0.3	0.958	0.212	0.053
	0.4	0.916	0.181	0.095
	0.5	0.876	0.140	0.127
	0.6	0.838	0.119	0.155
	0.7	0.773	0.089	0.197
	0.8	0.685	0.061	0.248
	0.9	0.463	0.037	0.354
Mature mRNA	0.1	0.972	0.269	0.042
	0.2	0.931	0.196	0.082
	0.3	0.911	0.164	0.098
	0.4	0.862	0.142	0.139
	0.5	0.846	0.107	0.147
	0.6	0.821	0.082	0.162
	0.7	0.776	0.073	0.192
	0.8	0.720	0.048	0.225
	0.9	0.557	0.021	0.310

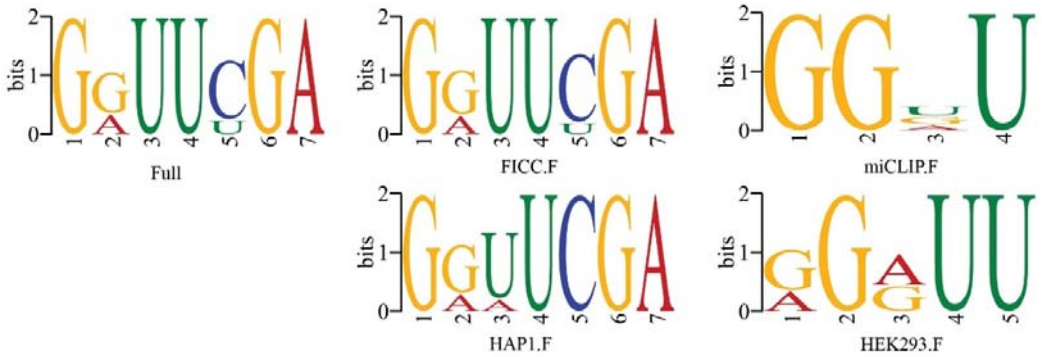

Figure S1. Motif analysis of positive m5U sites generated from different cell types and sequencing methods