LncCat: An ORF Attention Model to Identify LncRNA Based on Ensemble Learning Strategy and Fused Sequence Information

Hongqi Feng[1], Shaocong Wang[1], Yan Wang[3, 4], Xinye Ni[2], Zexi Yang[1], Xuemei Hu[3], Sen Yang[1, 2, *]

[1]School of Computer Science and Artificial Intelligence Aliyun School of Big Data School of Software, Changzhou University, Changzhou 213164, China;

[2]The Affiliated Changzhou No.2 People's Hospital of Nanjing Medical University, Changzhou 213164, China;

[3]Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China;

[4]School of Artificial Intelligence, Jilin University, Changchun 130012, China;

*Correspondence: ys@cczu.edu.cn (Sen. Yang);

Table S1. Detailed description of all features used in experiment

Type	No.	variable	Features	dimension	Description
codon-related	1	codon_num	stop codon count	1	The number of stop codons in transcript.
	2	codon_frequency	stop codon frequency	1	The number of stop codons divided by transcript length.
	3	fickett	Fickett TESTCODE Score	1	Position value of X (A, T, C, G) is calculated as: X_1 = Number of Xs in position 0, 3, 6, … X_2 = Number of Xs in position 1, 4, 7, … X_3 = Number of Xs in position 2, 5, 8, … $X_{pos} = \frac{MAX(X_1, X_2, X_3)}{MIN(X_1, X_2, X_3)+1}$ A_{pos} , T_{pos} , C_{pos} , G_{pos} and the percentage composition of each base are calculated. These eight values are converted into probabilities (p) of coding using a lookup table in the original article. Each probability is multiplied by a weight (w) for the respective base. Finally, the Fickett score is calculated as: $\text{Fickett Score} = \sum_{i=1}^{8} p_i w_i$.
GC	4	gc_content	GC content	3	GC content in the 1st, 2nd and 3rd position of

					codons.
	5	gc_frame_list	GC1_frame_score GC2_frame_score GC3_frame_score	3	The variance of GC-content in the 1st, 2nd and 3rd among three reading frames.
	6	CTD	CTD	30	The nucleotide composition (C) describes the percent composition of each nucleotide (A, T, C, G) in a transcript sequence. The nucleotide distribution (D) describes five relative positions along the transcript sequence of each nucleotide, with the 0 (first one), 25%, 50%, 75% and 100% (last one). The nucleotide transition (T) describes the percent frequency with the conversion of four nucleotides between adjacent positions.
sequence-related	7	distance_list	Euclidean-distance to lncRNA Euclidean-distance to PCT Logarithm-distance to lncRNA Logarithm-distance to PCT Euclidean-distance ratio Logarithm-distance ratio	6	$\text{EucDist.LNC}=\sqrt{\sum(\text{freq.seq(i)}-\text{freq.lnc(i)})^2}$, $\text{EucDist.PCT}=\sqrt{\sum(\text{freq.seq(i)}-\text{freq.pct(i)})^2}$, $\text{logDist.LNC}=\frac{1}{n}\sum\frac{\text{freq.seq(i)}}{\text{freq.lnc(i)}},\text{i}=1,2,\ldots,4^k$, $\text{logDist.PCT}=\frac{1}{n}\sum\frac{\text{freq.seq(i)}}{\text{freq.pct(i)}},\text{i}=1,2,\ldots,4^k$, $\text{EucDist.Ratio}=\frac{\text{EucDist.LNC}}{\text{EucDist.PCT}}$, $\text{logDist.Ratio}=\frac{\text{logDist.LNC}}{\text{logDist.PCT}}$, where freq.seq are the k-adjoining base(s) frequencies of an unevaluated sequence; freq.lnc are the average frequencies of lncRNAs' k-adjoining base(s); freq.pct are the average frequencies of PCTs' k-adjoining base(s); *i* denotes the different types of k-adjoining base(s), and n is the total number of the k-adjoining base(s) in one sequence.
	8	hexmaer	Hexamer score	1	For a given DNA sequence, the probability of the sequence under the coding and noncoding models is calculated. $F_c(h_i)$ and $F_{nc}(h_i)$ are calculated from coding and non-coding training sets respectively, refer to in-frame hexamer frequency (*i*=1, 2, ...,4096). For a given hexamer sequences

					$S=H_1, H_2, \ldots, H_m$, the hexamer score can be calculated as: $\text{Hexamer Score} = \frac{1}{m}\sum_{i=1}^{m}\log(\frac{F_c(h_i)}{F_{nc}(h_i)})$.
	9	3_mer	3-mer	64	The occurrences of k-length contiguous subsequences.
	10	2_mer	2-mer	16	
structure-related	11	mw	Molecular weight (Mw)	1	Molecular weight of the peptide.
	12	pi	Isoelectric point (pI)	1	Theoretical isoelectric point of the peptide.
	13	gravy	Gravy	1	Average hydrophilicity of the peptide.
	14	ii	Instability index	1	Average stability of the peptide.
ORF-related	15	ORF_length	The longest ORF length	1	Longest ORF length.
	16	ORF_cover	ORF coverage	1	The ratio of longest ORF and transcript lengths.
ORF-attention1	17	ORF1_length	ORF length	1	Length of ORF1.
	18	ORF1_fickett	Fickett TESTCODE Score	1	Fickett score of ORF1.
	19	ORF1_gc	GC content	3	GC content of ORF1.
	20	ORF1_gc_frame_list	GC_frame_score	3	GC frame score of ORF1.
	21	ORF1_CTD	CTD	30	CTD of ORF1.
	22	ORF1_3_mer	3-mer	64	3_mer of ORF1.
	23	ORF1_hexamer	Hexamer score	1	Hexamer score of ORF1.
	24	ORF1_mw	Molecular weight (Mw)	1	Mw of ORF1.
	25	ORF1_gravy	Gravy	1	Gravy of ORF1.
	26	ORF1_ss	Secondary Structure	3	Calculate fraction of helix, turn and sheet. Amino acids in sheet: E, M, A, L. Amino acids in Turn: N, P, G, S. Amino acids in helix: V, I, Y, F, W, L.
	27	ORF1_EIIP		5	Nucleotides can be replaced with the following EIIP values: {A: 0.1260; C: 0.1340; G: 0.0806, T: 0.1335}. Let $X_e[n]$ be the EIIP indicator sequence of $Seq[n]$. Using FFT on $X_e[n]$ can get the corresponding power spectrum

					$\{S_e[k]\}(k=0, 1, 2, \dots, N-1)$: $X_e[k]=\sum_{n=0}^{N-1} X_e[k]\, e^{-j\frac{2\pi kn}{N}}, S_e[k]=\lvert X_e[k]\rvert^2$ We extract quantile statistics features by calculating the top 0%, top 25%, top 50%, top 75% top 100% of the sorted power spectrum.
ORF-attention2	28	ORF2_length	ORF length	1	Length of ORF2.
	29	ORF2_fickett	Fickett TESTCODE Score	1	Fickett score of ORF2.
	30	ORF2_gc	GC content	3	GC content of ORF2.
	31	ORF2_gc_frame_list	GC_frame_score	3	GC frame score of ORF2.
	32	ORF2_CTD	CTD	30	CTD of ORF2.
	33	ORF2_3_mer	3-mer	64	3_mer of ORF2.
	34	ORF2_hexamer	Hexamer score	1	Hexamer score of ORF2.
	35	ORF2_mw	Molecular weight (Mw)	1	Mw of ORF2.
	36	ORF2_gravy	Gravy	1	Gravy of ORF2.
	37	ORF2_ss	Secondary Structure	3	Secondary Structure of ORF2.
	38	ORF2_EIIP		5	EIIP values of ORF1.
ORF-attention3	39	ORF3_length	ORF length	1	Length of ORF3.
	40	ORF3_fickett	Fickett TESTCODE Score	1	Fickett Score of ORF3.
	41	ORF3_gc	GC content	3	GC content of ORF3.
	42	ORF3_gc_frame_list	GC_frame_score	3	GC frame score of ORF3.
	43	ORF3_CTD	CTD	30	CTD of ORF3.
	44	ORF3_3_mer	3-mer	64	3_mer of ORF3.
	45	ORF3_hexamer	Hexamer score	1	Hexamer score of ORF3.
	46	ORF3_mw	Molecular weight (Mw)	1	Mw of ORF3.
	47	ORF3_gravy	Gravy	1	Gravy of ORF3.
	48	ORF3_ss	Secondary Structure	3	Secondary structure of ORF3.

	49	ORF3_EIIP		5	EIIP values of ORF1.
	50	BERT	BERT	128	BERT model.
Total dimensions				599	

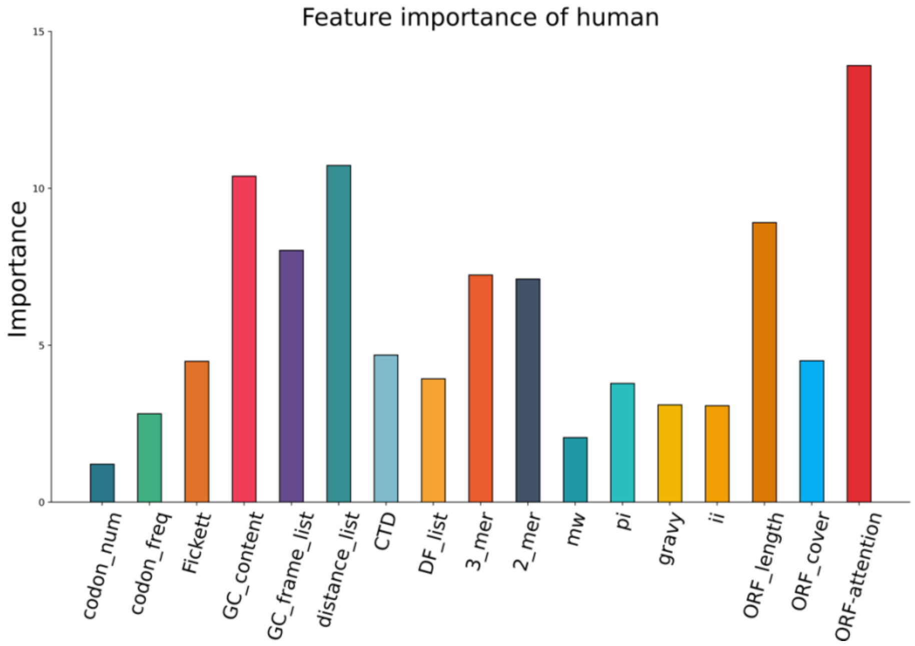

Figure S1. Feature importance on human data set

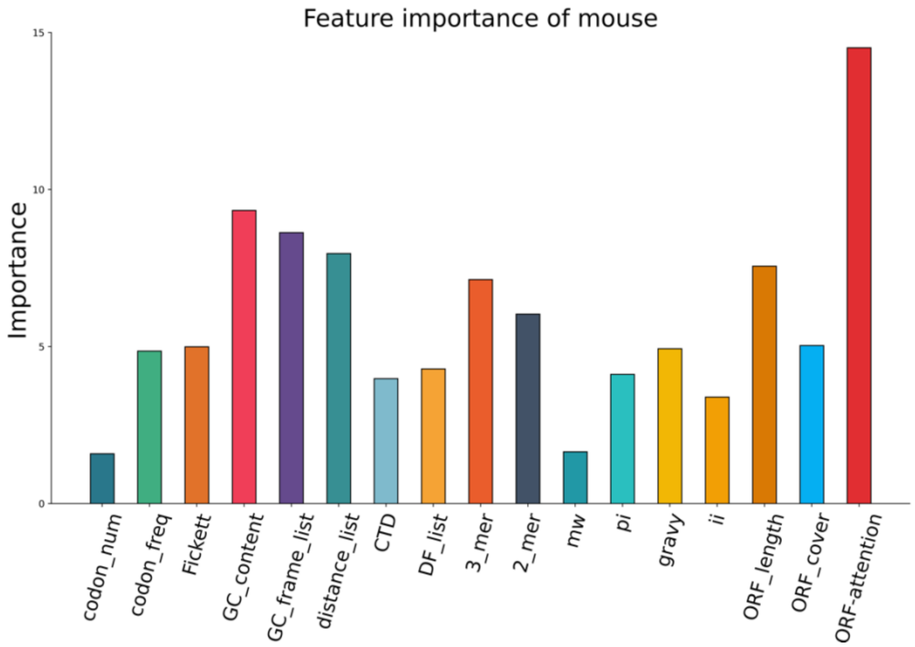

Figure S2. Feature importance on mouse data set

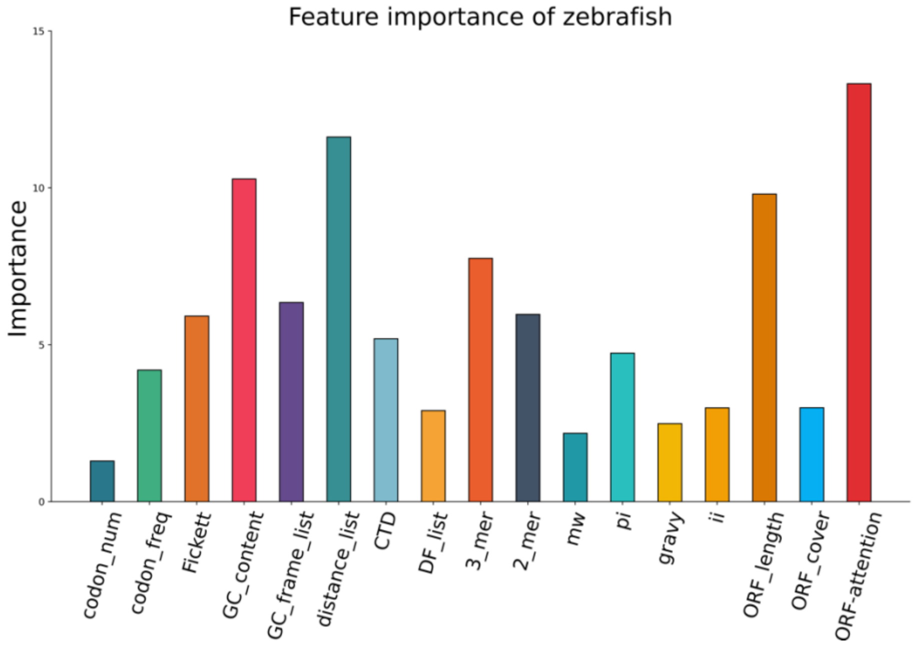

Figure S3. Feature importance on zebrafish dataset

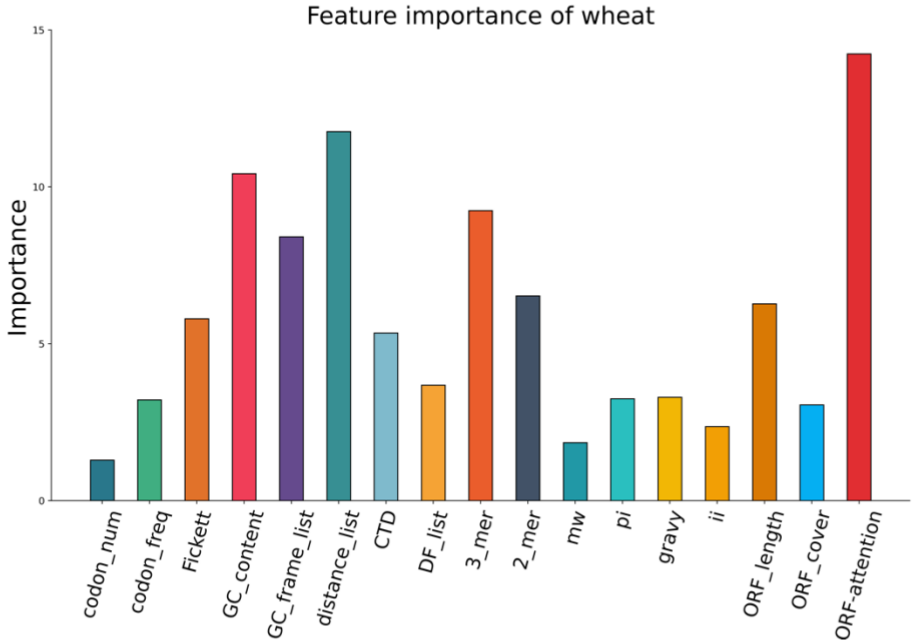

Figure S4. Feature importance on wheat dataset

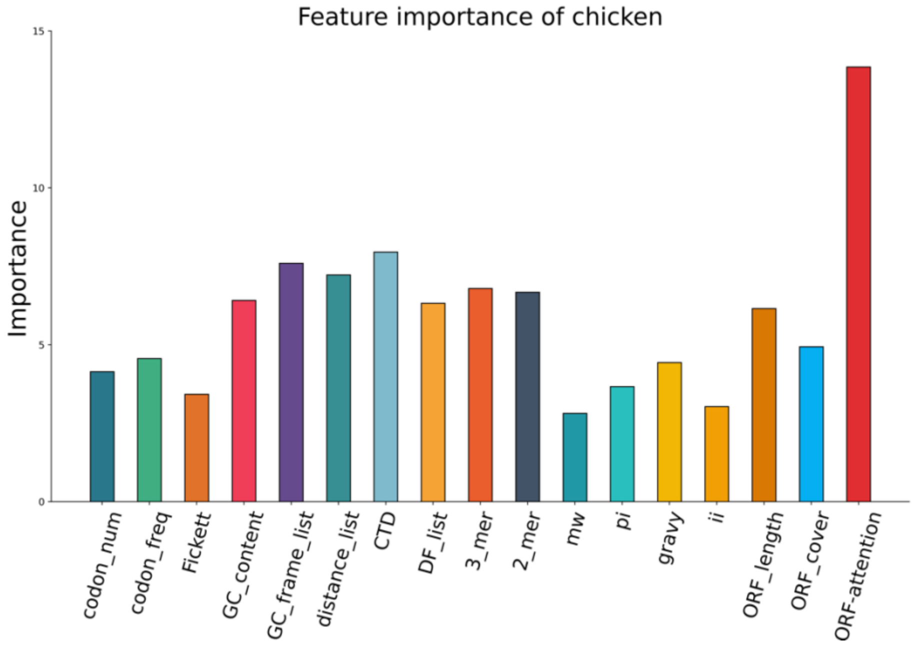

Figure S5. Feature importance on chicken dataset

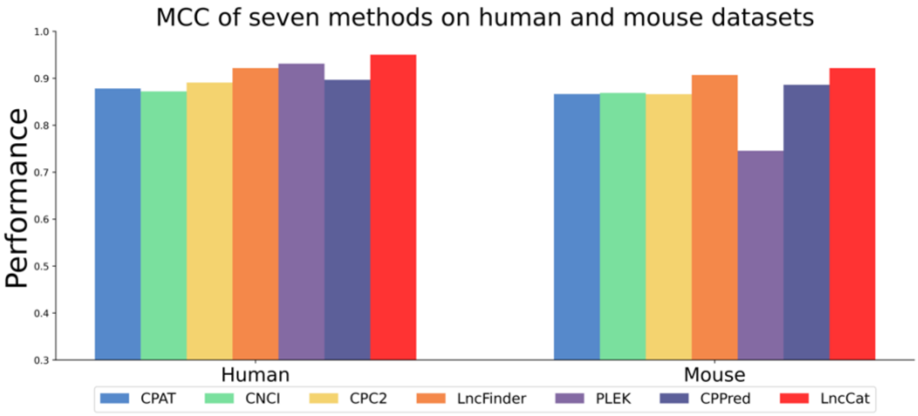

Figure S6. MCC of seven methods on human and mouse

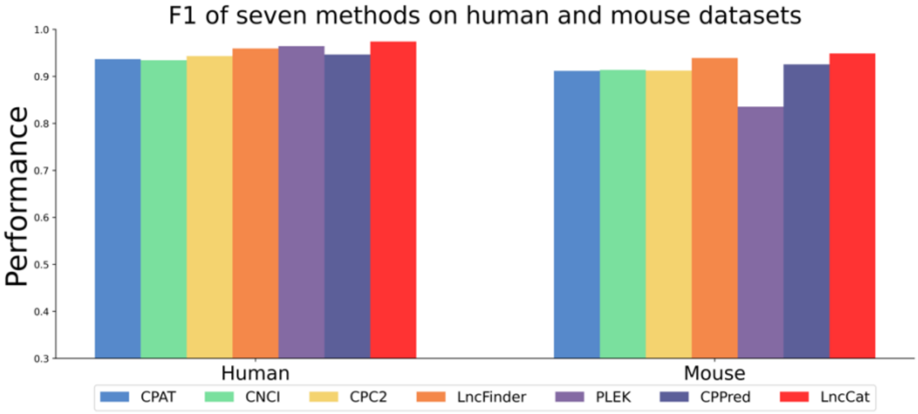

Figure S7. F1 of seven methods on human and mouse

Table S2. Performances of different methods on human dataset

Methods	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC	AUC
CPAT	4307	3966	285	251	0.9330	0.9405	0.9379	0.9392	0.9367	0.8782	0.9835
CNCI	4118	4113	474	104	0.8967	0.9753	0.8968	0.9344	0.9343	0.8721	0.9271
CPC2	4333	3996	259	221	0.9391	0.9476	0.9436	0.9455	0.9433	0.8909	0.9865
LncFinder	4384	4080	208	137	0.9515	0.9675	0.9547	0.9608	0.9594	0.9217	0.9873
PLEK	4360	4143	231	74	0.9472	**0.9825**	0.9497	0.9654	0.9645	0.9313	0.9918
CPPred	4330	4025	262	192	0.9389	0.9545	0.9429	0.9485	0.9466	0.8969	0.9872
LncCat	4449	4141	143	76	**0.9666**	0.9819	**0.9688**	**0.9751**	**0.9742**	**0.9503**	**0.9966**

Bold numbers indicate the highest value of the metrics.

Table S3. Performances of different methods on mouse dataset

Methods	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC	AUC
CPAT	3957	1950	184	193	0.9138	0.9099	0.9556	0.9400	0.9119	0.8664	0.9796
CNCI	3830	2065	312	77	0.8687	0.9641	0.9247	0.9381	0.9139	0.8687	0.9473
CPC2	3886	2012	256	130	0.8871	0.9393	0.9382	0.9386	0.9125	0.8661	0.9805
LncFinder	3959	2058	183	84	0.9183	0.9608	0.9558	0.9575	0.9391	0.9070	0.9822
PLEK	3580	1940	562	202	0.7754	0.9057	0.8643	0.8784	0.8355	0.7456	0.9532
CPPred	3902	2052	240	90	0.8953	0.9580	0.9421	0.9475	0.9256	0.8863	0.9843
LncCat	3985	2075	157	67	**0.9296**	**0.9687**	**0.962**	**0.9643**	**0.9487**	**0.9219**	**0.9920**

Bold numbers indicate the highest value of the metrics.

Table S4. Performances of different methods on zebrafish dataset

Methods	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC	AUC
CPAT	3049	679	130	198	0.8393	0.7742	0.9591	0.9191	0.8055	0.7555	0.9378
CNCI	2997	726	182	151	0.7996	0.8278	0.9427	0.9179	0.8134	0.7610	0.9203
CPC2	3007	743	172	134	0.8120	0.8472	0.9459	0.9246	0.8292	0.7811	0.9509
LncFinder	3088	698	91	179	**0.8847**	0.7959	**0.9714**	0.9334	0.8379	0.7980	0.9365
PLEK	2804	711	375	166	0.6547	0.8107	0.8820	0.8666	0.7244	0.6441	0.9325
CPPred	3066	699	113	178	0.8608	0.7970	0.9645	0.9283	0.8277	0.7834	0.9531
LncCat	3057	801	122	76	0.8678	**0.9133**	0.9616	**0.9511**	**0.8899**	**0.8591**	**0.9827**

Bold numbers indicate the highest value of the metrics.

Table S5. Performances of different methods on wheat dataset

Methods	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC	AUC
CPAT	826	880	113	196	0.8862	0.8178	0.8797	0.8467	0.8507	0.6960	0.9152
CNCI	803	957	136	119	0.8756	0.8894	0.8552	0.8734	0.8824	0.7455	0.8626
CPC2	861	730	78	346	0.9035	0.6784	0.9169	0.7896	0.7749	0.6060	0.9180
LncFinder	819	1031	120	45	0.8957	0.9582	0.8722	0.9181	0.9259	0.8370	0.9564
PLEK	685	668	254	408	0.7245	0.6208	0.7295	0.6715	0.6687	0.3508	0.7349
CPPred	872	746	67	330	**0.9176**	0.6933	**0.9286**	0.8030	0.7898	0.6324	0.9263
LncCat	844	1037	95	39	0.9160	**0.9637**	0.8988	**0.9334**	**0.9393**	**0.8672**	**0.9780**

Bold numbers indicate the highest value of the metrics.

Table S6. Performances of different methods on chicken dataset

Methods	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC	AUC
CPAT	1272	670	71	112	0.9042	0.8568	0.9471	0.9139	0.8798	0.8135	0.9575
CNCI	1224	687	121	93	0.8502	0.8808	0.9100	0.8993	0.8652	0.7852	0.9020
CPC2	1269	571	76	209	0.8825	0.7321	0.9435	0.8659	0.8003	0.7076	0.9348
LncFinder	1270	708	75	72	0.9042	0.9077	0.9442	0.9308	0.9060	0.8512	0.9718
PLEK	1068	672	277	108	0.7081	0.8615	0.7941	0.8188	0.7773	0.6356	0.8929
CPPred	1285	633	60	147	0.9134	0.8115	**0.9554**	0.9026	0.8595	0.7885	0.9587
LncCat	1285	745	60	35	**0.9254**	**0.9551**	0.9553	**0.9552**	**0.9400**	**0.9047**	**0.9882**

Bold numbers indicate the highest value of the metrics.

Table S7. Metrics corresponding to different cutoffs on human dataset

CutOff	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC
0.1	4342	4194	250	23	0.9437	0.9945	0.9456	0.9690	0.9685	0.9393
0.2	4390	4179	202	38	0.9539	0.9910	0.9560	0.9728	0.9721	0.9462
0.3	4411	4170	181	47	0.9584	0.9889	0.9606	0.9741	0.9734	0.9486
0.4	4436	4155	156	62	0.9638	0.9853	0.9660	**0.9753**	**0.9744**	**0.9507**
0.5	4449	4141	143	76	0.9666	0.9820	0.9689	0.9751	0.9742	0.9503
0.6	4462	4124	130	93	0.9694	0.9779	0.9717	0.9747	0.9737	0.9493
0.7	4482	4098	110	119	0.9739	0.9718	0.9760	0.9740	0.9728	0.9479
0.8	4495	4075	97	142	0.9767	0.9663	0.9789	0.9729	0.9715	0.9457
0.9	4513	4017	79	200	0.9807	0.9526	0.9828	0.9683	0.9664	0.9368

Table S8. Metrics corresponding to different cutoffs on mouse dataset

CutOff	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC
0.1	3899	2112	243	30	0.8968	0.9860	0.9413	0.9566	0.9393	0.9081
0.2	3936	2105	206	37	0.9109	0.9827	0.9503	0.9613	0.9454	0.9171
0.3	3954	2098	188	44	0.9178	0.9795	0.9546	0.9631	0.9476	0.9203
0.4	3973	2083	169	59	0.9250	0.9725	0.9592	0.9637	0.9481	0.9209
0.5	3985	2075	157	67	0.9297	0.9687	0.9621	0.9644	0.9488	0.9219
0.6	4002	2068	140	74	0.9366	0.9655	0.9662	**0.9659**	**0.9508**	**0.9250**
0.7	4017	2048	125	94	0.9425	0.9561	0.9698	0.9651	0.9492	0.9228
0.8	4038	2024	104	118	0.9511	0.9449	0.9749	0.9647	0.9480	0.9213
0.9	4052	1980	90	162	0.9565	0.9244	0.9783	0.9599	0.9402	0.9103

Table S9. Metrics corresponding to different cutoffs on zebrafish dataset

CutOff	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC
0.1	2984	828	195	49	0.8094	0.9441	0.9387	0.9398	0.8716	0.8368
0.2	3007	818	172	59	0.8263	0.9327	0.9459	0.9430	0.8763	0.8421
0.3	3026	809	153	68	0.8410	0.9225	0.9519	0.9455	0.8798	0.8462
0.4	3043	805	136	72	0.8555	0.9179	0.9572	0.9487	**0.8856**	**0.8535**
0.5	3056	793	123	84	0.8657	0.9042	0.9613	**0.9490**	0.8846	0.8521
0.6	3061	780	118	97	0.8686	0.8894	0.9629	0.9470	0.8789	0.8450
0.7	3079	767	100	110	0.8847	0.8746	0.9685	0.9482	0.8796	0.8466
0.8	3095	753	84	124	0.8996	0.8586	0.9736	0.9487	0.8786	0.8465
0.9	3115	731	64	146	0.9195	0.8335	0.9799	0.9482	0.8744	0.8435

Table S10. Metrics corresponding to different cutoffs on wheat dataset

CutOff	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC
0.1	791	1062	148	14	0.8777	0.9870	0.8424	0.9196	0.9291	0.8447
0.2	808	1056	131	20	0.8896	0.9814	0.8605	0.9251	0.9333	0.8536
0.3	821	1052	118	24	0.8991	0.9777	0.8743	0.9295	0.9368	0.8613
0.4	834	1042	105	34	0.9085	0.9684	0.8882	0.9310	0.9375	0.8629
0.5	844	1037	95	39	0.9161	0.9638	0.8988	**0.9335**	**0.9393**	**0.8672**
0.6	848	1031	91	45	0.9189	0.9582	0.9031	0.9325	0.9381	0.8649
0.7	858	1020	81	56	0.9264	0.9480	0.9137	0.9320	0.9371	0.8634
0.8	868	1009	71	67	0.9343	0.9377	0.9244	0.9315	0.9360	0.8624
0.9	881	979	58	97	0.9441	0.9099	0.9382	0.9231	0.9266	0.8465

Table S11. Metrics corresponding to different cutoffs on chicken dataset

CutOff	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC
0.1	1237	762	108	18	0.8759	0.9769	0.9197	0.9407	0.9236	0.8789
0.2	1261	757	84	23	0.9001	0.9705	0.9375	0.9496	0.9340	0.8950
0.3	1270	755	75	25	0.9096	0.9679	0.9442	0.9529	0.9379	0.9012
0.4	1281	749	64	31	0.9213	0.9603	0.9524	**0.9553**	**0.9404**	**0.9051**
0.5	1285	745	60	35	0.9255	0.9551	0.9554	**0.9553**	0.9401	0.9047
0.6	1290	740	55	40	0.9308	0.9487	0.9591	0.9553	0.9397	0.9043
0.7	1294	733	51	47	0.9349	0.9397	0.9621	0.9539	0.9373	0.9009
0.8	1304	722	41	58	0.9463	0.9256	0.9695	0.9534	0.9358	0.8994
0.9	1310	702	35	78	0.9525	0.9000	0.9740	0.9468	0.9255	0.8851

Table S12. Performances of different methods on human-sORFs dataset

Methods	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC	AUC
CPAT	150	615	491	24	0.5561	0.9624	0.2340	0.5977	0.7049	0.2866	0.8164
CNCI	207	625	434	14	0.5902	0.9781	0.3229	0.6500	0.7362	0.3982	0.5847
CPC2	81	607	560	32	0.5201	0.9499	0.1264	0.5375	0.6722	0.1344	0.8004
LncFinder	108	622	533	17	0.5385	0.9734	0.1685	0.5703	0.6934	0.2390	0.8378
PLEK	499	621	140	18	0.8160	0.9718	0.7809	0.8764	0.8871	0.7668	0.9553
CPPred	266	614	375	25	0.6208	0.9609	0.4150	0.6875	0.7543	0.4484	0.8575
LncCat	530	637	111	2	**0.8516**	**0.9968**	**0.8268**	**0.9117**	**0.9185**	**0.8356**	**0.9769**

Bold numbers indicate the highest value of the metrics.

Table S13. Performances of different methods on mouse-sORFs dataset

Methods	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC	AUC
CPAT	279	897	567	96	0.6127	0.9033	0.3298	0.6395	0.7302	0.2884	0.5093
CNCI	291	962	555	31	0.6341	0.9688	0.3440	0.6813	0.7665	0.4101	0.6221
CPC2	111	942	735	51	0.5617	0.9486	0.1312	0.5726	0.7056	0.1404	0.7905
LncFinder	365	956	481	37	0.6653	0.9627	0.4314	0.7183	0.7868	0.4754	0.7844
PLEK	374	901	469	92	0.6577	0.9074	0.4437	0.6944	0.7626	0.4019	0.7828
CPPred	383	952	463	41	0.6728	0.9587	0.4527	0.7259	0.7907	0.4868	0.8865
LncCat	584	977	262	16	**0.7885**	**0.9838**	**0.6903**	**0.8488**	**0.8754**	**0.7166**	**0.9550**

Bold numbers indicate the highest value of the metrics.

Table S14. Performances of different methods on zebrafish-sORFs dataset

Methods	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC	AUC
CPAT	166	376	221	121	0.6298	0.7565	0.4289	0.6131	0.6874	0.1965	0.6476
CNCI	141	403	246	94	0.6210	0.8109	0.3643	0.6154	0.7033	0.1968	0.5876
CPC2	45	408	342	89	0.5440	0.8209	0.1163	0.5124	0.6544	-0.0869	0.6575
LncFinder	233	387	154	113	0.7153	0.7740	0.6021	0.6990	0.7435	0.3824	0.7339
PLEK	208	399	179	98	0.6903	0.8028	0.5375	0.6867	0.7423	0.3549	0.7597
CPPred	193	382	194	115	0.6632	0.7686	0.4987	0.6505	0.7120	0.2783	0.7259
LncCat	327	460	60	37	**0.8846**	**0.9255**	**0.8449**	**0.8902**	**0.9046**	**0.7767**	**0.9475**

Bold numbers indicate the highest value of the metrics.

Table S15. Performances of different methods on integrated dataset

Methods	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC	AUC
CPAT	13147	8110	1047	985	0.8857	0.8917	0.9262	0.9127	0.8887	0.8169	0.9627
CNCI	12972	8548	1225	544	0.8747	0.9402	0.9137	0.9240	0.9062	0.8441	0.9285
CPC2	13356	8052	841	1040	0.9054	0.8856	0.9408	0.9192	0.8954	0.8298	0.9709
LncFinder	13497	8437	700	655	0.9234	0.9280	0.9507	0.9418	0.9257	0.8779	0.9741
PLEK	12497	8134	1699	958	0.8272	0.8946	0.8803	0.8859	0.8596	0.7654	0.9510
CPPred	13455	8155	742	937	0.9166	0.8969	0.9477	0.9279	0.9067	0.8481	0.9751
LncCat	13662	8775	535	317	**0.9425**	**0.9651**	**0.9623**	**0.9634**	**0.9537**	**0.9236**	**0.9915**

Bold numbers indicate the highest value of the metrics.

Table S16. Performances of different methods on sORFs integrated dataset

Methods	TN	TP	FP	FN	PRE	SEN	SPE	ACC	F1	MCC	AUC
CPAT	429	1959	1445	170	0.5755	0.9202	0.2289	0.5966	0.7081	0.2085	0.7482
CNCI	639	1990	1235	139	0.6171	0.9347	0.3410	0.6568	0.7434	0.3477	0.6050
CPC2	237	1957	1637	172	0.5445	0.9192	0.1265	0.5481	0.6839	0.0753	0.7617
LncFinder	647	1990	1227	139	0.6186	0.9347	0.3453	0.6588	0.7445	0.3517	0.7640
PLEK	1081	1921	788	208	0.7091	0.9023	0.5784	0.7509	0.7941	0.5131	0.8412
CPPred	842	1948	1032	181	0.6537	0.9150	0.4493	0.6970	0.7626	0.4167	0.8385
LncCat	1587	2077	287	52	**0.8786**	**0.9756**	**0.8469**	**0.9153**	**0.9245**	**0.8346**	**0.9717**

Bold numbers indicate the highest value of the metrics.

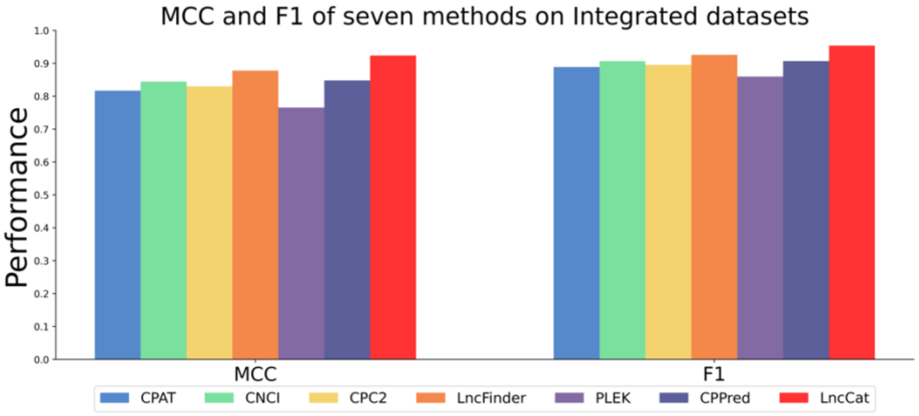

Figure S6. MCC and F1 of CPAT, CNCI, CPC2, LncFinder, PLEK, CPPred and LncCat on Integrated datasets

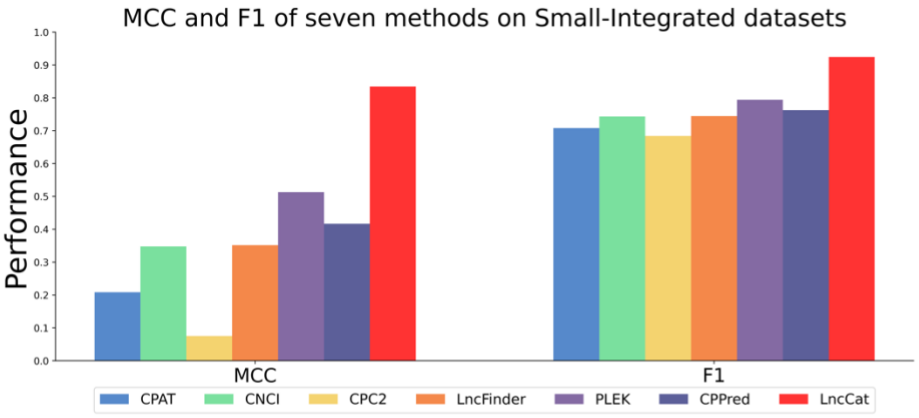

Figure S7. MCC and F1 of CPAT, CNCI, CPC2, LncFinder, PLEK, CPPred and LncCat on Small-Integrated datasets