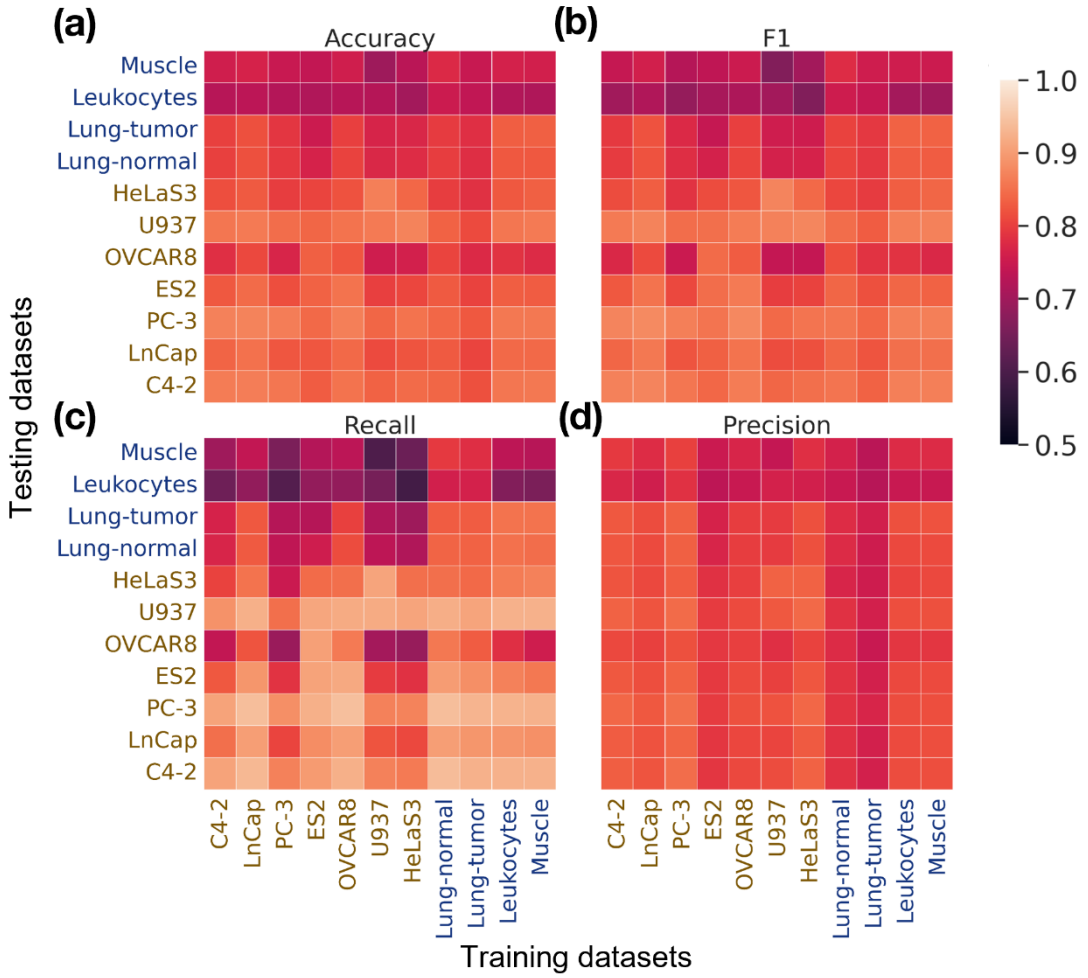


Figure S1. DNABERT models generalized well on different datasets. (a), (b), (c), and (d) respectively represent the accuracy, F1 score, recall, and precision of DNABERT models trained from one dataset and tested on other unseen datasets. If the model was trained and tested using an identical dataset (the elements on the diagonal line starting from the lower left corner to the upper right corner), only testing data was used for evaluating the model.

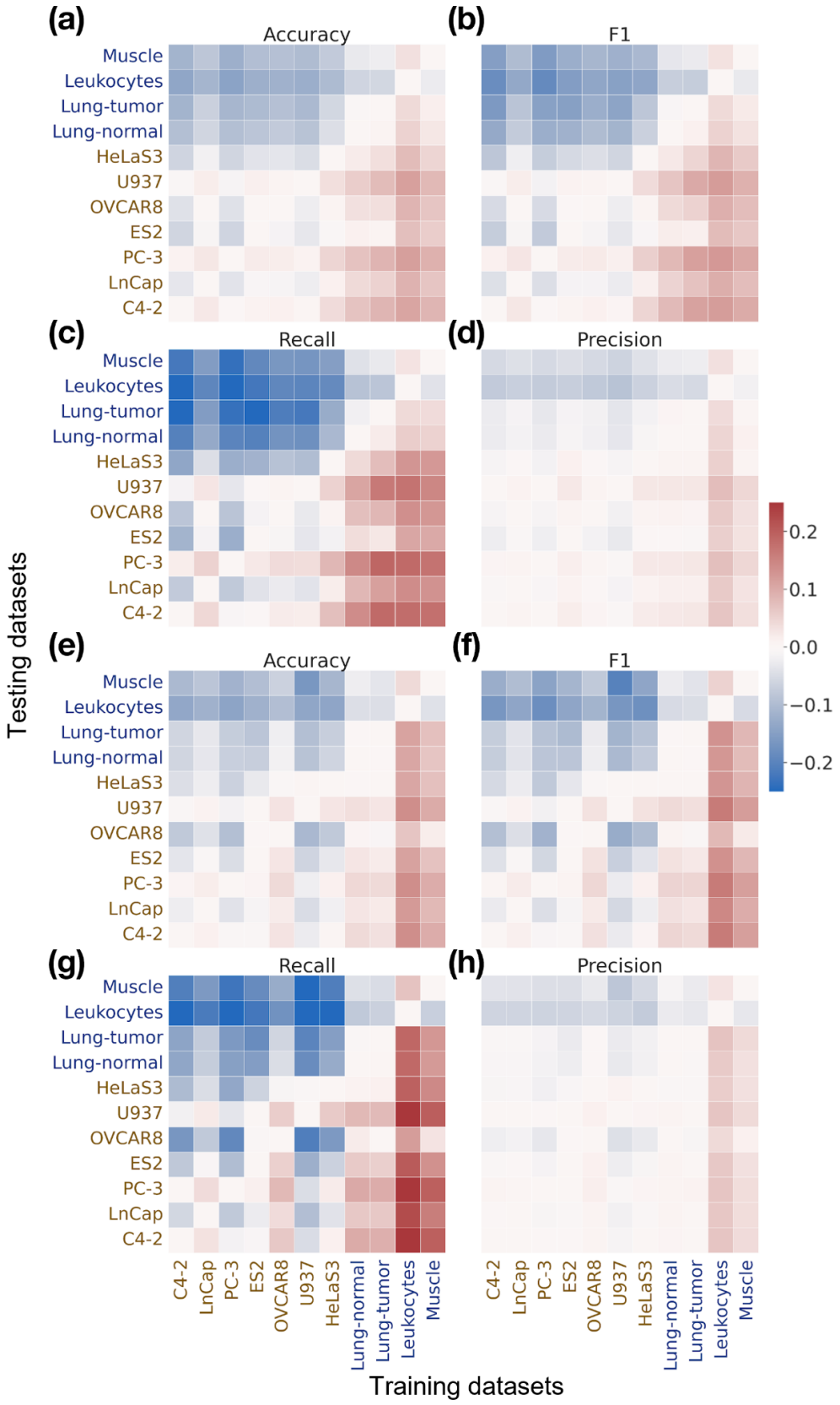


Figure S2. Performance differences between the original test dataset and all the other datasets. (a)-(d) shows the performance differences of CNN models, while (e)-(h) shows the performance differences of DNABERT models. The performance difference is the difference relative to the base performance, where the base performances are the performance of models trained and tested using partitions derived from the same dataset.

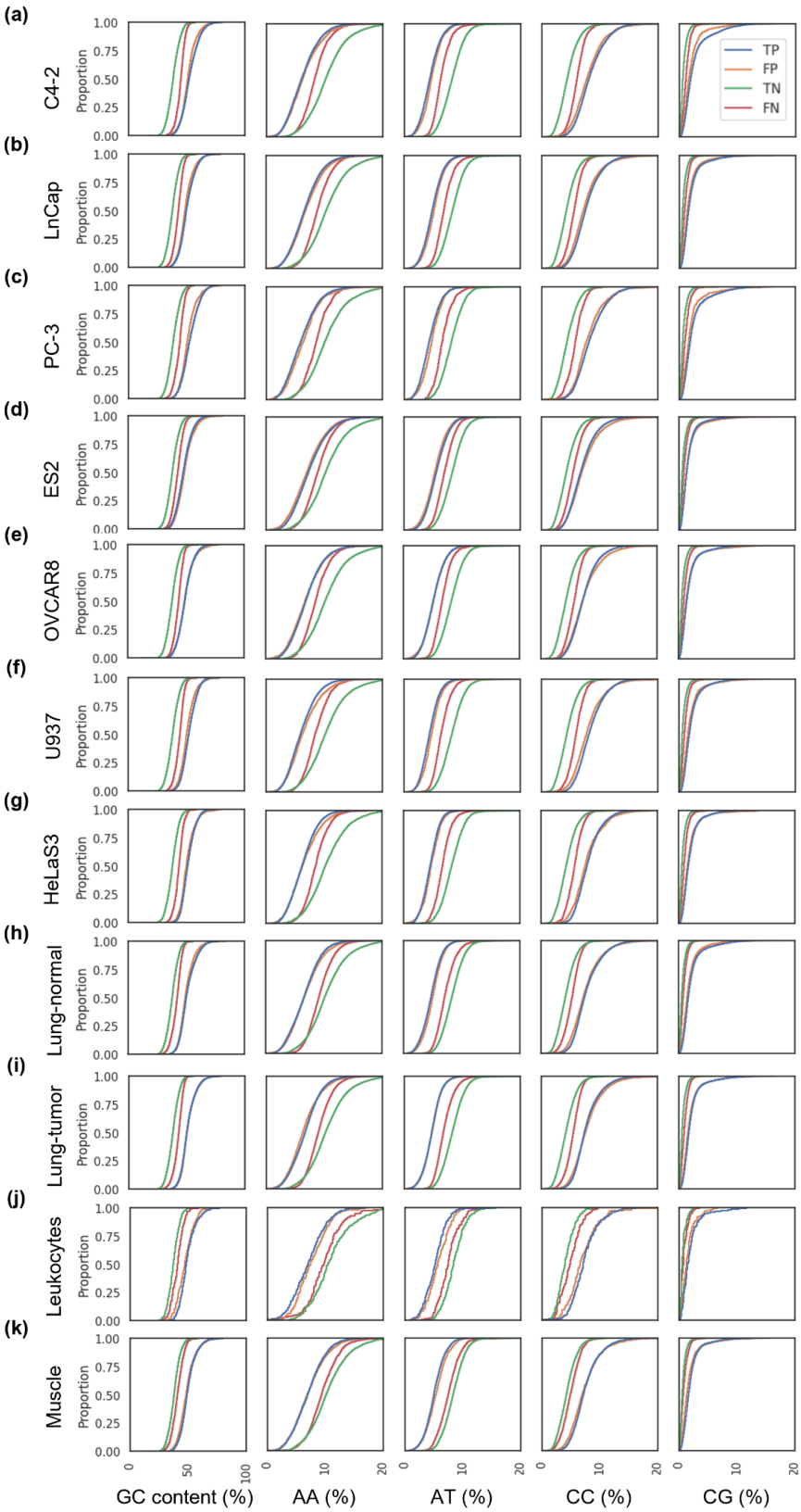


Figure S3. General sequence features used in predicting eccDNAs include GC content and dinucleotide frequencies. The distribution of GC content and dinucleotide (AA, AT, CC and CG) frequencies were obtained for every DNA sequence predicted by the CNN models across every dataset used in this study. The results showed that DNA sequences predicted to be eccDNAs exhibited higher GC content, compared to the ones predicted to be non-eccDNAs. In addition, DNA sequences predicted to be eccDNAs contained higher frequencies of dinucleotides with at least one guanine or cytosine and lower frequencies for dinucleotides with at least one adenine or thymine.

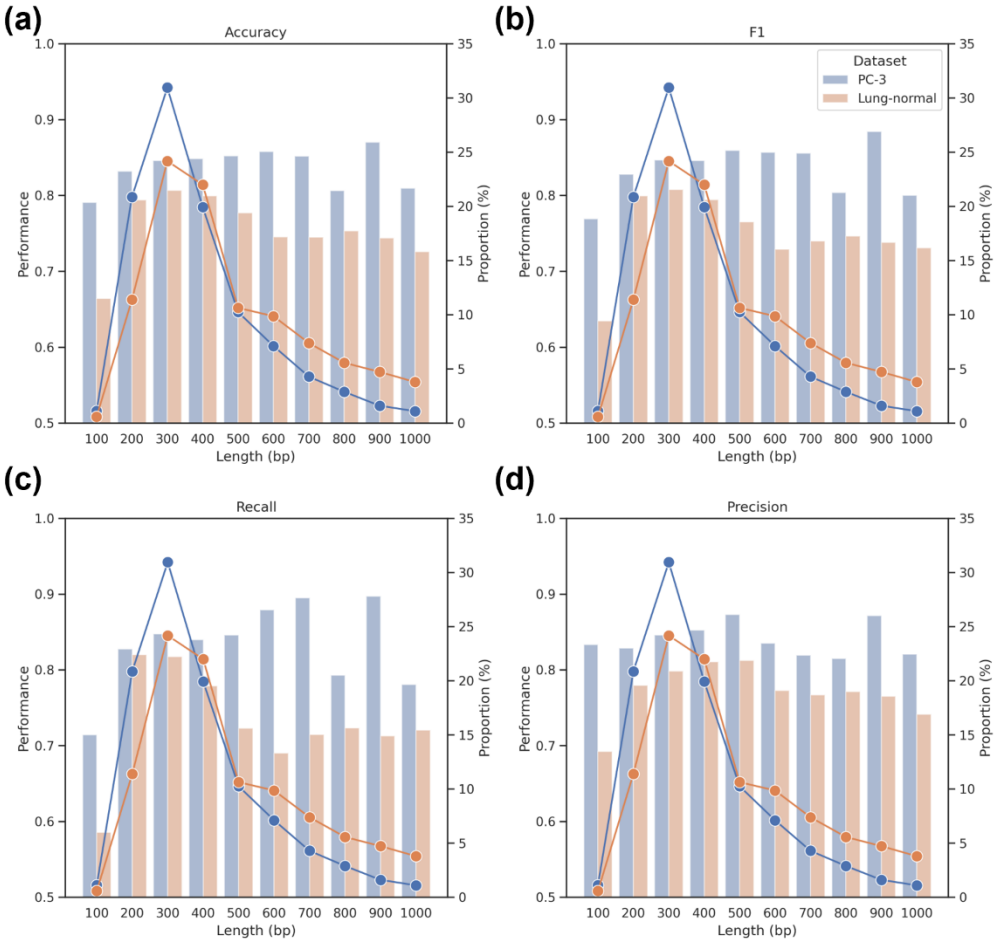


Figure S4. Performance of the CNN in predicting eccDNAs of various lengths. (a), (b), (c) and (d) respectively represent the accuracy, F1 score, recall, and precision of the CNN (bar plots, labels on the left) and proportions of eccDNAs within bins of lengths (line plots, labels on the right).

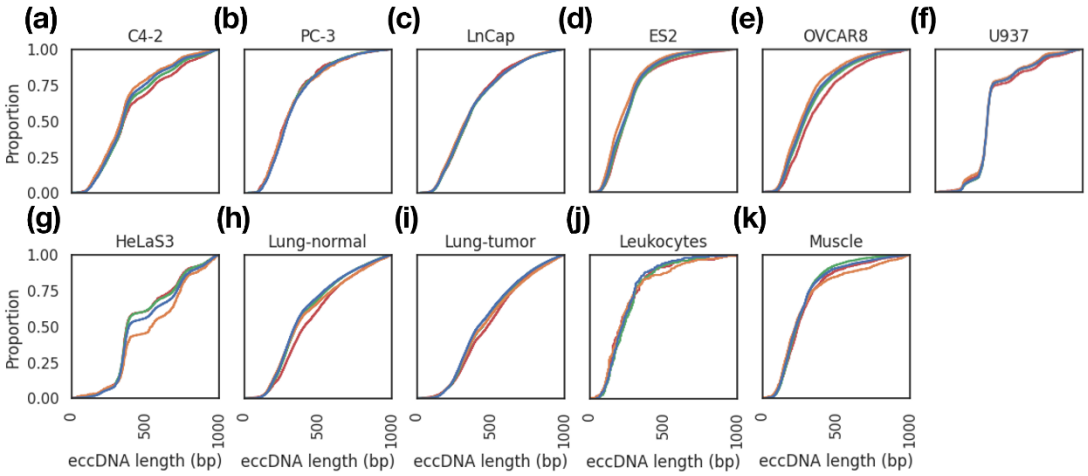


Figure S5. Length distribution of DNA sequences that are predicted to be eccDNAs and non-eccDNAs by CNN. The length distribution of eccDNAs and non-eccDNAs predicted by CNN models showed that the prediction was not biased by the coverage ratio of eccDNA to flanking regions.

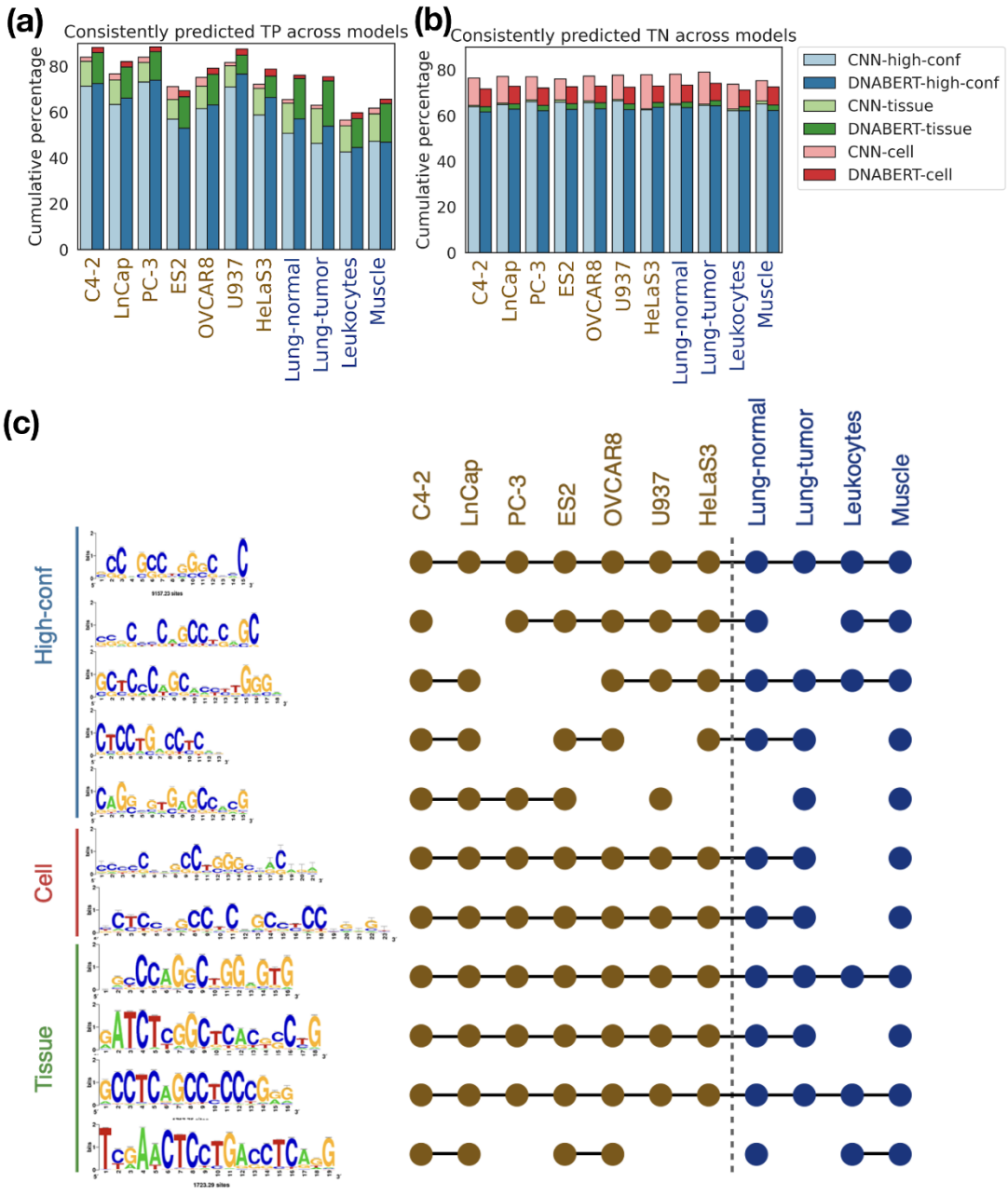


Figure S6. Consensus eccDNA-related motifs identified in DNABERT models. (a) and (b) respectively represent the percentages of DNA sequences that were consistently predicted to be TP and TN among models trained on various datasets. (c) The consensus eccDNA-related motifs from the three sets were inferred for DNABERT models. Motifs identified in the “High-conf” set occurred in most of the datasets. The dot on the right of the motif indicates the source dataset from which the consensus motif was derived.

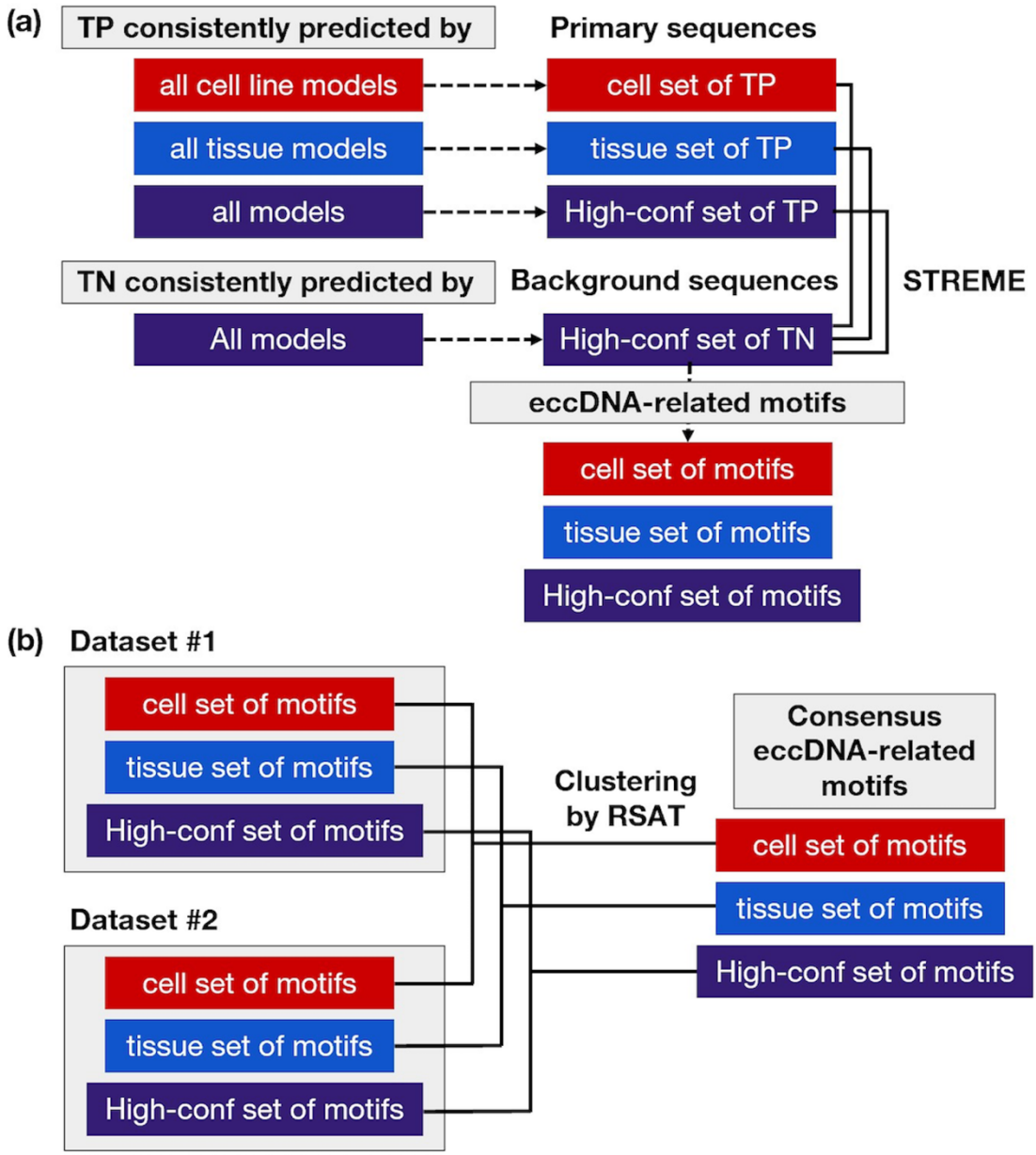


Figure S7. Workflow for the identification of consensus eccDNA-related motifs. (a) Workflow for the identification of eccDNA-related motifs. For each dataset, we used the "High-conf" set of TN as background sequences and inferred three sets of motifs that are overrepresented in the "High-conf" set of TP, "Cell" set of TP, and "Tissue" set of TP, respectively [1]. Afterwards, we kept the top 5 motifs with the highest number of matching sequences within each set of motifs. (b) Workflow for identification of consensus eccDNA-related motifs. As an example, here we showed that the "Cell" set of motifs, "Tissue" set of motifs, and "High-conf" set of motifs from two different datasets was combined and clustered by RSAT to obtain consensus eccDNA-related motifs representing each category of motifs [2]. The hierarchical clustering was performed with average linkage using width-normalized scores (Ncor), as the motif comparison matrices. Ncor ≥ 0.4 was used as a threshold to partition the tree generated from clustering. To obtain the non-redundant set of motifs that encompasses all motifs within each set, we inferred the consensus motif of each representative cluster by averaging the frequencies of the descendant motifs. The representative clusters were defined as clusters encompassing motifs derived from more than half of the corresponding datasets, i.e. ≥ 6, ≥ 4, ≥ 3 different datasets for "High-conf", "Cell" and "Tissue" sets, respectively.

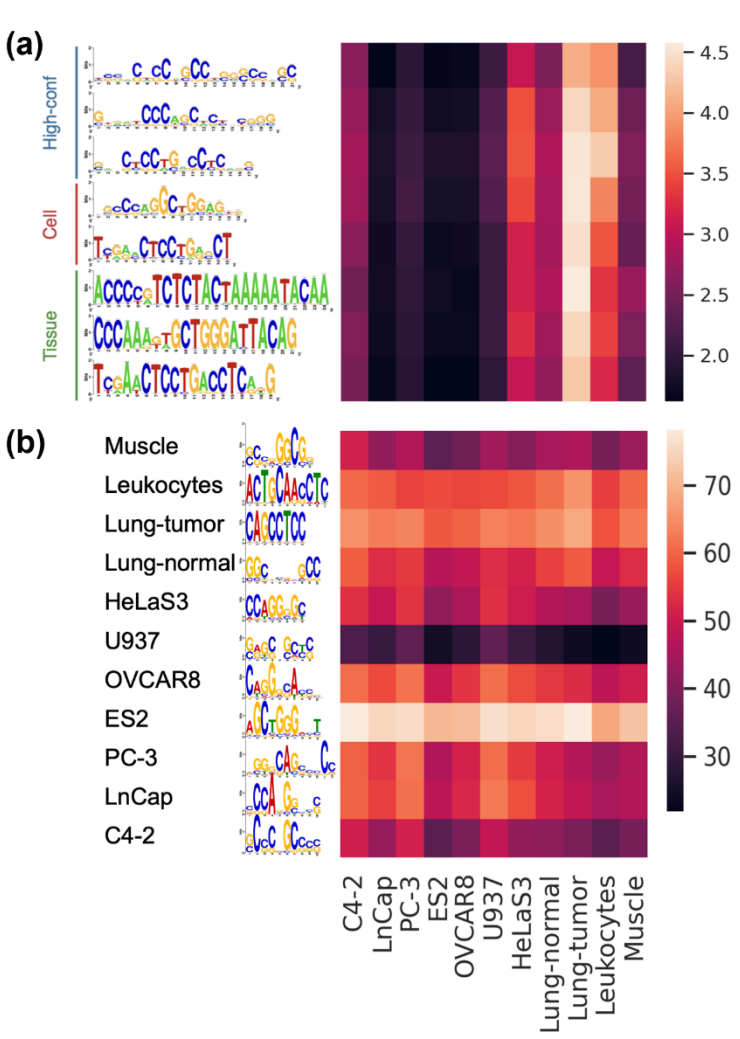


Figure S8. Results of motif scanning in eccDNA sequences. EccDNA sequences were scanned using motifs identified in this study by using FIMO [3]. The result with a p-value ≤ 0.0001 was deemed as significant. The values of heatmaps show the proportion (%) of sequences with ≥ 1 occurrence of the specific motif. (a) False negatives predicted by individual model were scanned with consensus motifs. The heatmap shows that consensus motifs occur in < 4.58% of false negatives for all datasets. (b) We selected the dataset-specific motifs by choosing motifs identified in the CNN models that maximize differences in occurrences (motifs that have the highest occurrences in the corresponding dataset but lowest occurrences in other datasets). By scanning eccDNAs in the testing data, we showed that the dataset-specific motifs are not enriched in the corresponding datasets, indicating there is little tissue-specificity in eccDNAs.

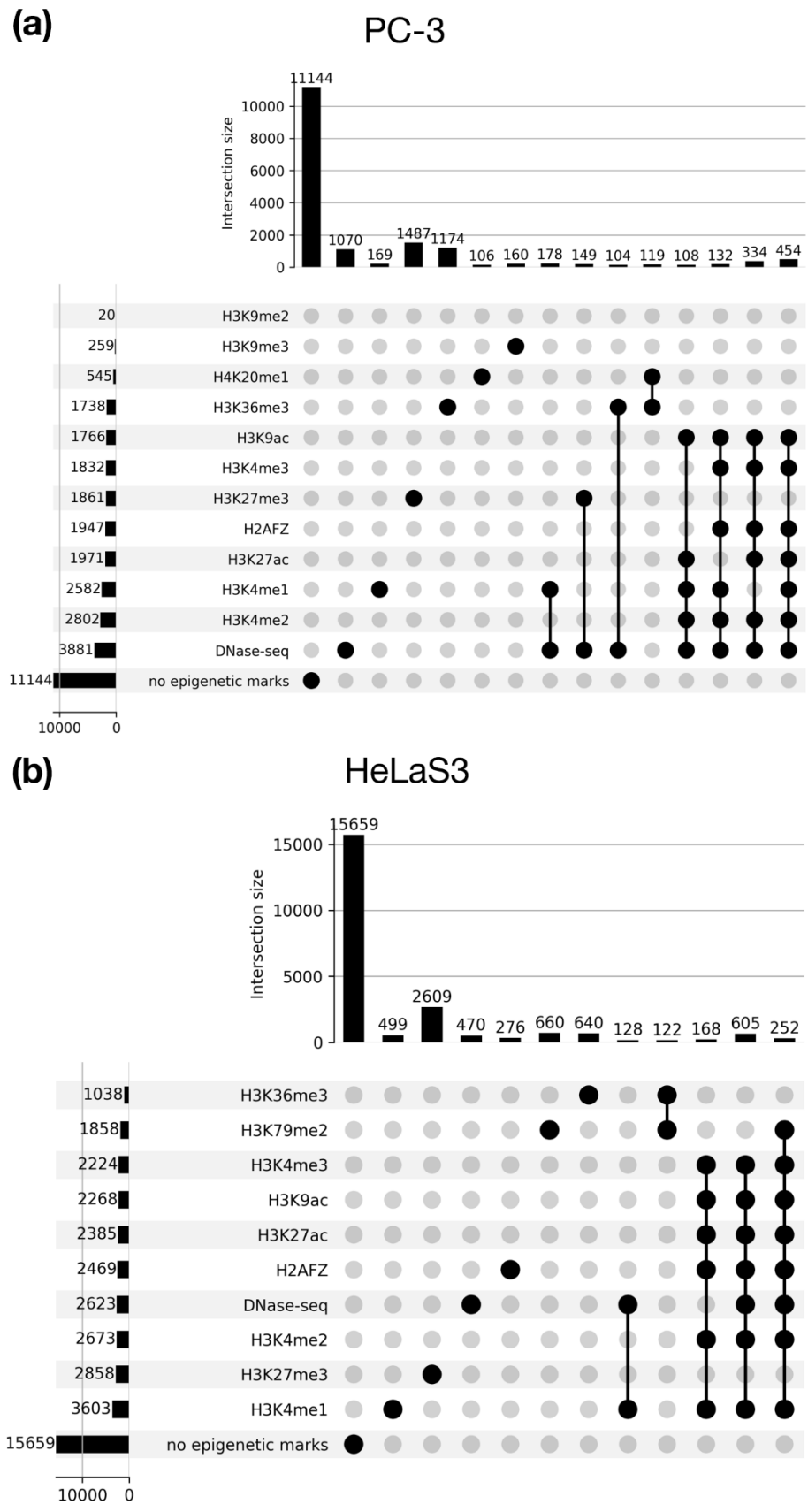


Figure S9. Numbers of eccDNAs overlapping epigenetic marks. To identify the epigenetic features associated with eccDNAs, firstly, we downloaded the bed files of the epigenetic marks including peaks identified in the histone chromatin immunoprecipitation followed by

sequencing (ChIP-seq) and DNase I hypersensitive sites sequencing (DNase-seq) of PC-3 (accession number: ENCSR052AWE, ENCSR946QFD, ENCSR826UTD, ENCSR881TWJ, ENCSR849APH, ENCSR566UMF, ENCSR788EQL, ENCSR275NCH, ENCSR197QXT, ENCSR943LZX, ENCSR339ZMJ and ENCSR764KFK) and HeLaS3 (accession number: ENCSR000ENO, ENCSR959ZXU, ENCSR000EJS, ENCSR000EJT, ENCSR000AQN, ENCSR000AOC, ENCSR000DTY, ENCSR000AOD, ENCSR000APW, ENCSR000AOE, ENCSR000DUA, ENCSR000AOG, ENCSR000AOH) from the ENCyclopedia Of DNA Elements (ENCODE, https://www.encodeproject.org/)[4]. We further counted the number of eccDNAs overlapping epigenetic marks with at least one base pair. After obtaining the numbers of eccDNAs overlapping each epigenetic mark, we partitioned the eccDNAs into disjoint sets and displayed the results of PC-3 (a) and HeLaS3 (b) dataset on the above figure. Every dot on the figure represents the subset of eccDNAs fitting specific criteria (not overlapping any epigenetic mark or overlapping a specific type of epigenetic mark) and only the subset with over 100 eccDNAs was displayed on the plot. The columns of the plot correspond to the subsets and the rows correspond to the types of epigenetic marks. The numbers at the top panel of the plot show the numbers of eccDNAs corresponding to the subsets and the numbers at the left panel show the numbers of eccDNAs overlapping the corresponding epigenetic mark. Here shows that over half of the eccDNAs within the two datasets do not overlap any epigenetic marks (PC-3, 57.48%; HeLaS3, 62.83%).

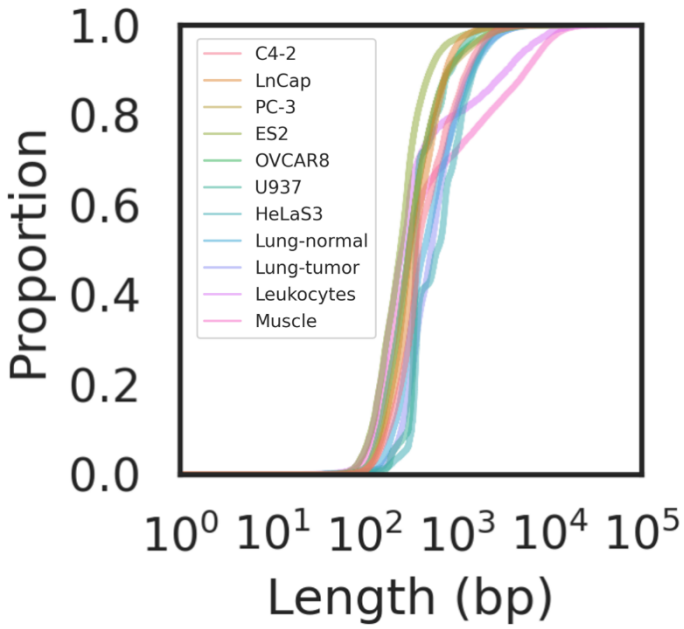


Figure S10. Length distribution of eccDNAs used in this study. 85.53% eccDNAs have length ≤1000 bp on average per dataset.

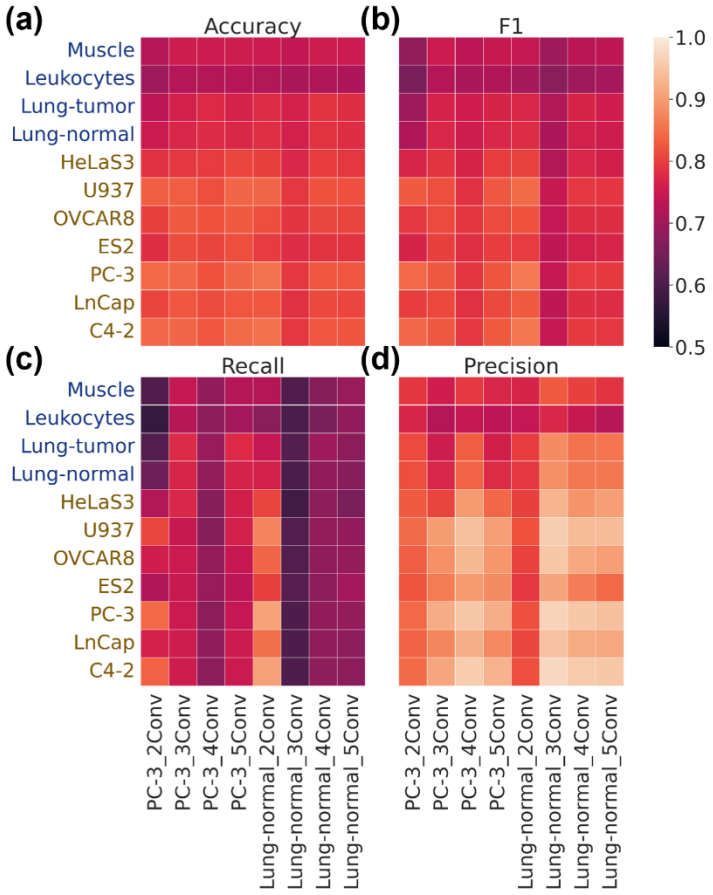


Figure S11. The performance of CNN models with 2, 3, 4 and 5 convolutional layers evaluated across all datasets. (a), (b), (c), and (d) represent the accuracy, F1 score, recall, and precision, respectively.

Table S1. Human DNA binding motifs in CIS-BP were queried to identify similar motifs. The first consensus eccDNA-related motif in the "High-conf" set of CNN models was used to query CIS-BP 2.00 human database for identification of similar motifs using Tomtom [5,6]. It was found that the top 10 most similar motifs (according to the *E-value*, where the *E-value* indicates the expected number of false positives in the database) all belong to the zinc-finger protein family.

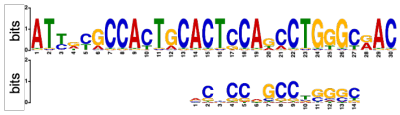
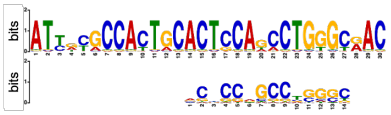
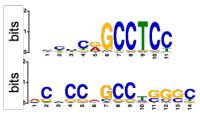
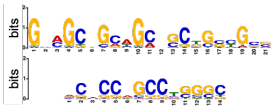
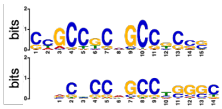
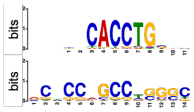
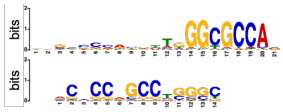
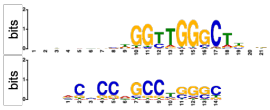
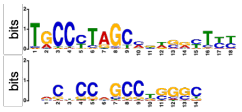
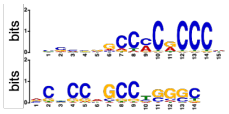
Motif ID	Alignment of query motif to target motifs (target motif is shown as the logo above the query motif)	*E*-value
M07609_2.00 (ZNF496)		8.82e-06
M07587_2.00 (ZNF304)		3.90e-01
M08388_2.00 (ZNF770)		4.13e-01
M07734_2.00 (ZNF311)		2.27e+00
M07658_2.00 (ZNF571)		2.29e+00
M08080_2.00 (ZEB2)		2.40e+00
M07639_2.00 (ZNF417)		3.07e+00
M08344_2.00 (ZNF449)		3.07e+00
M07772_2.00 (ZNF550)		3.22e+00
M08293_2.00 (KLF1)		3.45e+00

Table S2. The comparison of prediction results of CNN and DNABERT models. The TP, FN, TN, and FP predicted by CNN and DNABERT models were compared to assess consistency of the two models. By taking the union or intersection of the prediction results, performance could be further improved.

	Prediction results								Mode of integration					
	% positive				% negative				Union			Intersection		
CNN	TP	TP	FN	FN	TN	TN	FP	FP						
DNABERT	TP	FN	TP	FN	TN	FP	TN	FP	Acc	Rec	Pre	Acc	Rec	Pre
C4-2	74	8	17	2	70	15	12	3	98	98	77	97	74	96
LnCap	77	9	13	1	63	16	17	4	97	99	73	96	77	95
PC-3	74	10	14	2	71	14	13	2	98	98	77	97	74	97
OVCAR8	79	7	13	1	64	17	15	4	97	99	73	96	79	95
ES2	76	8	14	2	61	19	15	5	97	98	72	96	76	94
HeLaS3	65	9	23	3	68	16	13	3	97	97	75	96	65	95
Muscle	60	15	20	5	56	18	19	6	94	95	68	91	60	90
Leukocytes	56	17	21	7	48	19	24	10	92	93	64	86	56	85
Lung-normal	65	11	20	4	64	16	16	4	96	96	73	95	65	94
Lung-tumor	57	10	28	5	68	16	13	3	96	95	75	94	57	95

Table S3. Number of sequences of datasets.

Dataset	Number of sequences		Proportion (%)
	Total	Length ≤ 1000 bp	
C4-2	59713	49720	83.26
LnCap	129220	123350	95.46
PC-3	21030	19387	92.19
ES2	171943	166919	97.08
OVCAR8	77409	71886	92.87
U937	62982	58266	92.51
HeLaS3	33808	24921	73.71
Lung-normal	139864	111786	79.92
Lung-tumor	208330	169685	81.45
Leukocytes	4144	3328	80.31
Muscle	35092	25286	72.06

Table S4. Parameters of DNABERT models. Training batch size was fixed to 4 because of limited memory. Training epoch, logging steps, and save steps were adjusted with the size of datasets. Other parameters including warmup percentage, dropout probability, and weight decay were fixed to 0.1, 0.1, and 0.01, respectively, as they have little effect on performance. We tuned the learning rate between 0.00001 to 0.00005 and picked the one with the best performance.

Dataset	Parameters			
	training epoch	logging steps	save steps	learning rate
C4-2	2	4000	10000	0.00004
ES2	1	7000	20000	0.00002
HeLaS3	2	3000	9000	0.00003
leukocytes	4	1300	5000	0.00003
LnCap	1	6000	15000	0.00003
muscle	2	3000	9000	0.00003
OVCAR8	2	6000	15000	0.00002
PC-3	2	3000	7000	0.00004
Lung-normal	1	4000	21500	0.00003
Lung-tumor	1	5000	30000	0.00002
U937	2	5000	12000	0.00002

Supplementary references

1. Bailey TL. STREME: accurate and versatile sequence motif discovery. Bioinformatics 2021; 37:2834–2840
2. Castro-Mondragon JA, Jaeger S, Thieffry D, et al. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. 2017; 45:e119
3. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics 2011; 27:1017–1018
4. . An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature 2012; 489:57–74
5. Gupta S, Stamatoyannopoulos JA, Bailey TL, et al. Quantifying similarity between motifs. Genome Biol. 2007; 8:R24
6. Weirauch MT, Yang A, Albu M, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 2014; 158:1431–1443