OXFORD

Briefings in Bioinformatics, 2023, **24(3)**, 1–11
https://doi.org/10.1093/bib/bbad147
Advance access publication date 21 April 2023
Problem Solving Protocol

Short human eccDNAs are predictable from sequences

Kai-Li Chang†, Jia-Hong Chen†, Tzu-Chieh Lin, Jun-Yi Leu, Cheng-Fu Kao, Jin Yung Wong and Huai-Kuang Tsai

Corresponding authors: Huai-Kuang Tsai, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan. Tel.: +886-2-27883799-1718; E-mail: hktsai@iis.sinica.edu.tw; Jin Yung Wong, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan. Tel.: +886-2-27883799-1473; E-mail: wongjinyung@gmail.com

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Abstract

Background: Ubiquitous presence of short extrachromosomal circular DNAs (eccDNAs) in eukaryotic cells has perplexed generations of biologists. Their widespread origins in the genome lacking apparent specificity led some studies to conclude their formation as random or near-random. Despite this, the search for specific formation of short eccDNA continues with a recent surge of interest in biomarker development. Results: To shed new light on the conflicting views on short eccDNAs' randomness, here we present DeepCircle, a bioinformatics framework incorporating convolution- and attention-based neural networks to assess their predictability. Short human eccDNAs from different datasets indeed have low similarity in genomic locations, but DeepCircle successfully learned shared DNA sequence features to make accurate cross-datasets predictions (accuracy: convolution-based models: $79.65 \pm 4.7\%$, attention-based models: $83.31 \pm 4.18\%$). Conclusions: The excellent performance of our models shows that the intrinsic predictability of eccDNAs is encoded in the sequences across tissue origins. Our work demonstrates how the perceived lack of specificity in genomics data can be re-assessed by deep learning models to uncover unexpected similarity.

Keywords: bidirectional encoder representations from transformers, convolutional neural network, deep learning, extrachromosomal circular DNA

INTRODUCTION

Recently, there has been renewed interest in the research of extrachromosomal circular DNA (eccDNA), as new findings have shown that they might be more prevalent than previously thought, in both tumor and normal cells [1, 2]. EccDNAs are derived from chromosomes, and range from several hundred base pairs (bp) to megabase pairs (Mbp). The eccDNAs that are long enough to span entire gene bodies, termed as double minutes or extrachromosomal DNAs (ecDNAs), were found to be prevalent in tumor tissues but rare in normal tissues. EcDNAs promote tumor progression and drug resistance through copy number amplification of oncogene, leading to shortened survival and poor outcomes in cancer patients [3–7]. In addition, distal enhancers could be co-amplified with oncogenes in ecDNAs [8]. The enhancer-carrying ecDNA could also act as mobile enhancers to promote genome-wide chromosomal gene expression [9]. The properties of ecDNAs that make them potent drivers of massive gene expression has become the research focus in treating aggressive tumors.

On the other hand, the eccDNAs with shorter length are less understood, despite the fact that they are much more abundant in both tumor and normal cells than ecDNAs. The short eccDNAs, termed as small polydispersed circular DNA (spcDNA) or microDNA in some literature, are typically shorter than 1000 bp [10, 11]. Short eccDNAs seldom contain complete genes due to their lengths but recent studies have found that they can regulate gene expression or stimulate immune response [12, 13]. Although short eccDNAs can be linked to biological functions, their generation was suggested to be random because their origins spread across the genome lacking apparent specificity [13, 14]. However, the proposition of random biogenesis contradicts the reports of overrepresented features in short eccDNAs [7, 15–18], which favor non-random production mechanisms. In addition to the implications in fundamental biology, the randomness of short

Kai-Li Chang received the BS degree in Life Sciences and the MS degree in Physiology from National Cheng Kung University, Tainan, Taiwan in 2018 and 2020 respectively. He is a research assistant at the Institute of Information Science, Academia Sinica. His research interests include genomics, cancer biology, and machine learning.
Jia-Hong Chen received the BS degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan in 2022. He is an MS student at the Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan. His research interests include transformer models and deep learning.
Tzu-Chieh Lin received the BS degree in Life Sciences from National Cheng Kung University, Tainan, Taiwan in 2018. He is a research assistant at the Institute of Information Science, Academia Sinica. His research interests include genomics, genetics, and machine learning.
Jun-Yi Leu received the BS degree in Chemical Engineering from National Tsing Hua University, Hsinchu City, Taiwan and the PhD degree in Molecular, Cellular and Developmental Biology from Yale university, New Haven, CT, USA in 1988, 1999 respectively. He is a distinguished research fellow at the Institute of Molecular Biology, Academia Sinica. His research interests include endosymbiosis, phenotypic robustness, and genetic incompatibility.
Cheng-Fu Kao received the PhD degree in Biochemistry from the University of Edinburgh, Edinburgh, United Kingdom in 2002. He is a research fellow at the Institute of Cellular and Organismic Biology, Academia Sinica. His research interests include chromatin structure, DNA replication, and genome stability.
Jin Yung Wong received the PhD degree in Biodiversity from the National Taiwan Normal University, Taipei, Taiwan, in 2020. He is a postdoctoral research fellow at the Institute of Information Science, Academia Sinica. His research interests include evolution, genomics, machine learning, and biomechanics.
Huai-Kuang Tsai received the BS, the MS, and the PhD degrees in Computer Science and Information Engineering from the National Taiwan University, Taipei, Taiwan, in 1996, 1998, and 2003, respectively. He is a Research Fellow at the Institute of Information Science, Academia Sinica. His research interests include computational biology, bioinformatics, gene regulation, and machine learning.
Received: November 3, 2022. **Revised:** March 23, 2023. **Accepted:** March 27, 2023

eccDNAs also concerns the active field of applied research in using them as biomarkers [19–21]. If they are mostly random, most will not be good biomarker candidates. Given the importance of the question on the short eccDNAs' randomness, a better assessment is needed.

The randomness question can be approached by quantifying the predictability of short eccDNAs, a task that machine learning methods are well suited for. Existing computational tools for short eccDNA studies focused on detection from experimental sequencing data enriched for eccDNAs, by detecting circular junction from reads mapping [13, 22–25]. By utilizing these tools, a significant amount of eccDNA data has been accumulated, which can be used to train prediction models. Deep learning has achieved prediction performance close to that of humans in various tasks in the domains of image and natural language processing [26, 27]. While deep learning might still be not as good as humans in everyday tasks, it beats us in specialized tasks that deal with input data which humans struggle to learn from [28]. One such example is the field of genomics, which deep learning can help decoding the biological insights hidden in the DNA sequences consisting of only four letters [29]. Deep learning models can achieve this without the requirement of any prior knowledge, as integrated feature extraction removes the need of feature definition. For example, convolutional neural network (CNN), a type of model widely used for image processing, can learn the local dependency of data derived from convolutional layers. By combining multiple convolutional layers, CNN is capable of learning complex patterns from raw data, and makes predictions better than conventional machine learning models.

To overcome the CNN's limitation in locality, architectures that can model long-term dependency of sequences have been combined with CNN's representation power to achieve state-of-the-art performance in predicting genomic elements [30, 31]. Bidirectional Encoder Representations of Transformers (BERT) is capable of modeling global dependency of data through self-attention mechanisms [32]. Originally developed for natural language processing, BERT models have been demonstrated to be capable of learning DNA sequences as a language too, after some tweaks (DNABERT) [33]. DNABERT pre-trained on the human genome is suitable for transfer learning in various tasks. Through fine-tuning the pre-trained DNABERT model on labeled data, it outperformed other machine learning models in prediction of promoters, splice sites and transcription factor binding sites. Therefore, DNABERT has great potential in complementing CNN models in learning complex sequence features. Together, the two state-of-the-art methods will provide a good estimate of short eccDNAs' predictability.

We aimed to make the assessment on short human eccDNAs' predictability as general as possible. Therefore, we focused on the intrinsic feature that is available for all datasets regardless of cell types, the DNA sequences. To achieve this aim, we developed DeepCircle, an end-to-end bioinformatics pipeline adopting CNN and DNABERT to predict eccDNAs. Using DeepCircle, we demonstrated that eccDNAs could be accurately predicted by both CNN and DNABERT. We also demonstrated that eccDNAs derived from various tissue origins exhibited similar DNA sequence properties, despite having little overlap in genomic locations. With the help from deep learning models, we established the general predictability of short eccDNAs, which is a necessary condition for the search of specific short eccDNAs and their formation mechanisms. Our work serves as a basis for further exploration of sufficient feature sets to predict cell- or condition-specific eccDNAs.

RESULTS

Short human eccDNAs are predictable using sequence-based deep learning models

To assess the predictability of short eccDNAs, we developed a deep learning-based bioinformatics pipeline called DeepCircle, which adopted CNN and DNABERT as underlying deep learning models for predicting eccDNAs and incorporated DNA motif analysis for interpretation of prediction results (Figure 1). We applied DeepCircle on eleven datasets covering several human cancer cell lines and tissues to obtain as general an assessment as possible. We first evaluated the performance of DeepCircle by training on each dataset and validating using eccDNAs unseen during the training process (see Methods). Both CNN and DNABERT models achieved promising prediction performance, reaching a mean accuracy around 0.8 across different datasets (CNN $= 0.797 \pm 0.047$ and DNABERT $= 0.833 \pm 0.042$; mean $\pm$ SD; Figure 2). The good predictive power from the models were achieved with a balanced performance in recall and precision, as shown by a mean performance around 0.8 in both metrics across datasets (recall: CNN $= 0.792 \pm 0.063$ and DNABERT $= 0.872 \pm 0.052$; precision: CNN $= 0.800 \pm 0.047$ and DNABERT $= 0.809 \pm 0.036$; Figure 2). The balanced prediction performance of our models is further supported by the composite metrics, F1 score, with CNN $= 0.795 \pm 0.049$ and DNABERT $= 0.839 \pm 0.042$. Overall, the performance is comparable between two models and across the datasets. Our results indicate that deep learning models could reliably predict short human eccDNAs from the information of DNA sequence only.

DeepCircle is robust: models trained on one dataset can predict eccDNAs from other tissue origins

The similar prediction performance across datasets prompted us to hypothesize that eccDNAs from different tissue origins might share similar sequence features useful for prediction. To test this hypothesis, we performed cross prediction across datasets, i.e. use the model trained on a dataset to predict on another dataset for all pairwise combinations. But first we needed to preclude the possibility that good cross prediction performance comes from shared origins in genomic locations of eccDNAs between two datasets. We used the Jaccard index to quantify the pairwise similarity of each dataset pair in terms of genomic coordinates of eccDNAs (Figure 3A, see Methods). The results showed that most of the dataset pairs have the Jaccard index less than 10% (with a minimum of 0.2% and maximum of 12%), indicating that eccDNAs identified in every dataset are dissimilar in their origins of genomic locations, ideal for use in testing model robustness.

By testing all the models on all the datasets, we showed that the deep learning models within DeepCircle achieved good performance in predicting test data from datasets that are not the one used for training. Since CNN and DNABERT obtained similar performances, for simplicity, in the remaining results we will show the results from CNN model in the main text and leave the DNABERT part in the Supplementary Data available online at http://bib.oxfordjournals.org/. Surprisingly, some of these cross-testing achieved even better performance than the dataset they are originally trained for, indicating that there is probably no tissue-specificity regarding the prediction of eccDNAs (Figure 3B–E, Supplementary Figure S1A–D available online at http://bib.oxfordjournals.org/). However, we found that there was some distinction between models trained on the cell

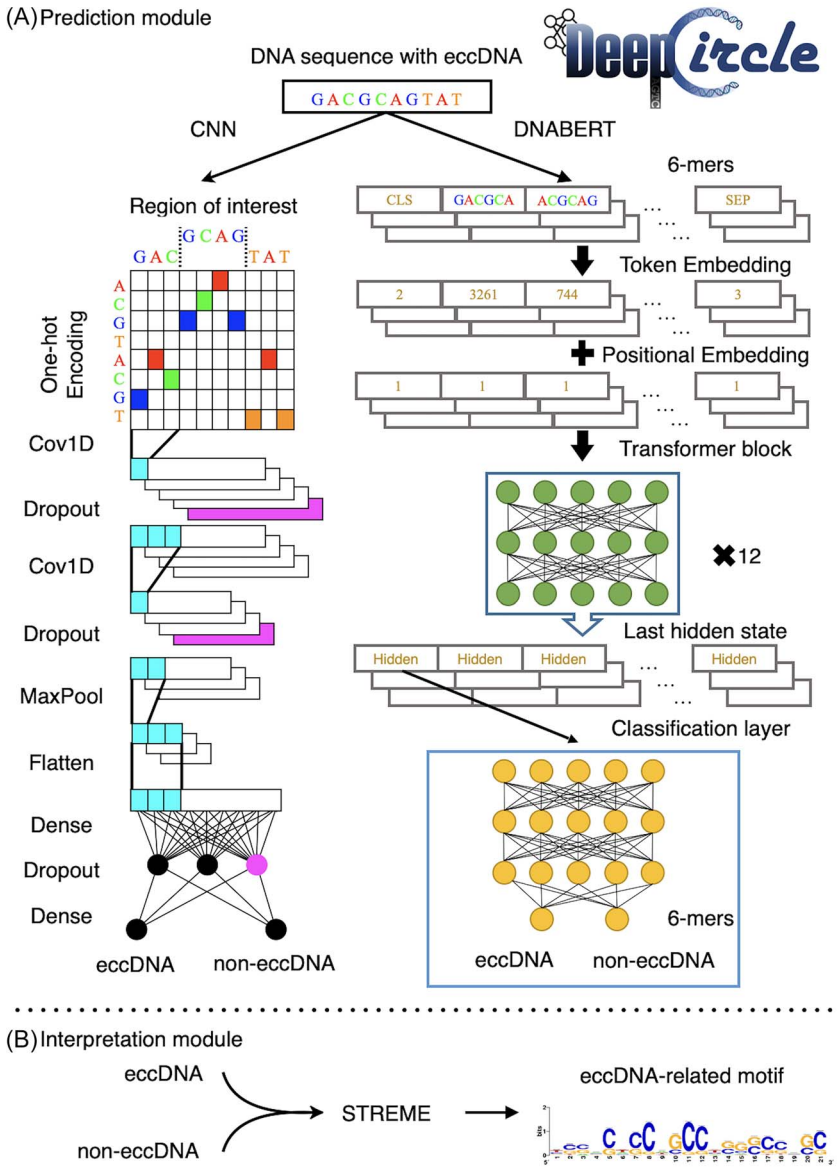

Figure 1. Workflow of DeepCircle. (**A**) DNA sequences containing eccDNA were inputted to either CNN or DNABERT for predicting eccDNAs. For CNN, DNA sequences were converted into numerical tensors by onehot encoding and used for predicting eccDNAs. For DNABERT, DNA sequences were split into series of 6- mers and used for fine-tuning on a pre-trained DNABERT model for predicting eccDNAs. Following prediction, the predicted eccDNAs and non-eccDNAs were analyzed with the interpretation module (**B**) of DeepCircle to identify eccDNA-related motifs.

lines datasets (cell line models) and models trained on the tissue datasets (tissue models). Cell line models have worse recall when tested on the tissue datasets than on the cell line datasets (Supplementary Figure S2C and G available online at http://bib.oxfordjournals.org/). In contrast, tissue models have no significant difference in the performance between testing on the cell line datasets and on the tissue datasets (Supplementary Figure S2 available online at http://bib.oxfordjournals.org/). In summary, cross-testing results confirmed that the deep learning models within DeepCircle have robust predictive power. Our models managed to predict eccDNAs of different tissue origins with dissimilar origins in genomic locations that they were not exposed to during training, indicating that there might be some common sequence patterns in eccDNAs useful for prediction. These results also indicate that tissue and cell line datasets might share some common eccDNA sequence features, but tissue datasets might have more diverse eccDNA sequence features than cell line datasets, as a result, higher values in recall were obtained by the tissue models.

General sequence features used in predicting eccDNAs includes GC content and dinucleotide frequencies

To further elucidate the sequence difference between those eccDNAs (TP and FP) and non-eccDNAs (TN and FN) predicted by DeepCircle, we first compared their GC content of their 1000-bp-long sequences. The results showed that the sequences predicted to be eccDNAs have a higher GC content than those predicted to be non-eccDNAs, in all the datasets we have analyzed (Figure 4A, F and Supplementary Figure S3 available online at http://bib.oxfordjournals.org/). We further performed an in-depth characterization of DNA composition by computing the dinucleotide frequencies of the DNA sequences that are predicted

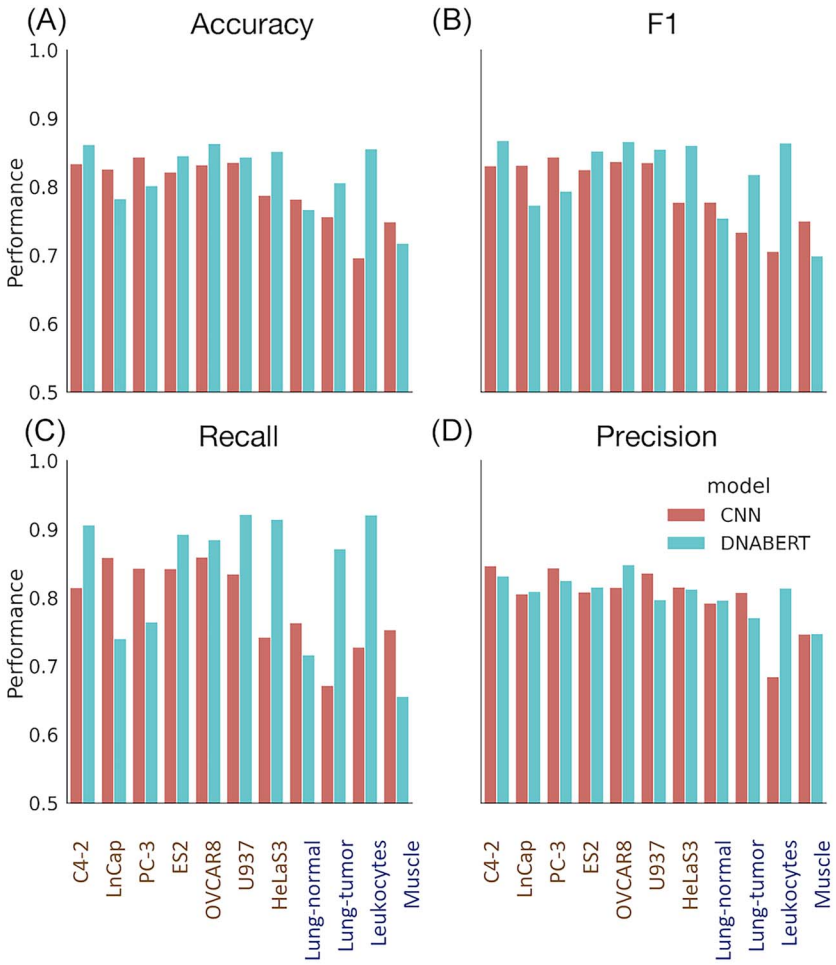

Figure 2. Short human eccDNAs are predictable using sequence-based deep learning models. Deep learning models implemented within DeepCircle including CNN and DNABERT models were evaluated by testing the models on the test data derived from the same dataset separately. Accuracy (**A**), F1 score (**B**), recall (**C**) and precision (**D**) were used to evaluate the performance of models. Labels are colored by the datasets on which models were trained (cell lines in orange and tissues in blue).

to be eccDNAs and non-eccDNAs. The results showed that the sequences predicted to be eccDNAs have higher frequencies of dinucleotides containing at least one cytosine or guanine (CC, CG, GC and GG) and lower frequencies of dinucleotides containing adenine or thymine (AA, AT, TA and TT), than those predicted to be non-eccDNAs (Figure 4B–E, G–J and Supplementary Figure S3 available online at http://bib.oxfordjournals.org/). We further examined the potential bias of the lengths of eccDNAs on the prediction of eccDNAs. We showed that the eccDNAs that are 100 to 500-bp long, which are well-represented in the datasets, are accurately predicted (Supplementary Figure S4 available online at http://bib.oxfordjournals.org/). Additionally, there was no significant difference between the length distribution of DNA sequences that are predicted to be eccDNAs and non-eccDNAs (Supplementary Figure S5 available online at http://bib.oxfordjournals.org/), indicating that predictions were not dictated by the coverage ratio of eccDNA to flanking regions in genomic intervals.

Motif analysis revealed common motifs utilized by models to discern eccDNAs from non-eccDNAs

The results showed that DeepCircle was capable of predicting eccDNAs of different tissue origins and had distinct robustness in performance of cell line models and tissue models. These results motivated us to characterize the sequence features that are shared by all correct predictions (TP or TN) or correct predictions exclusive to the cell line models/tissue models. To this end, for each dataset, we partitioned the prediction results into three disjoint sets: 'High-conf' set of TP, 'Cell' set of TP and 'Tissue' set of TP, where 'High-conf' set represents consistent predictions made by all models, 'Cell' set represents consistent predictions made exclusively by all cell line models and 'Tissue' set represents consistent predictions made exclusively by all tissue models. Afterwards, we calculated the proportion of TP in these three sets over all the TP in each dataset. Results showed that TP within the 'High-conf' set were consistently predicted by all the models (CNN = 58.39 ± 10.27%; DNABERT = 61.23 ± 10.46%), but the ratio is higher for the cell line datasets than the tissue datasets (Figure 5A and Supplementary Figure S6A available online at http://bib.oxfordjournals.org/). We also performed the same analysis described above for the prediction results of TN for every non-eccDNAs in each dataset. A large portion of TN was also consistently predicted by all the models ('High-conf' set of TN, CNN = 64.67 ± 1.33%; DNABERT = 62.74 ± 0.76%), but no distinction between cell lines and tissue datasets was observed (Figure 5B, Supplementary Figure S6B available online at http://bib.oxfordjournals.org/). The results that the tissue models predicted more exclusive TP but almost equal proportion of

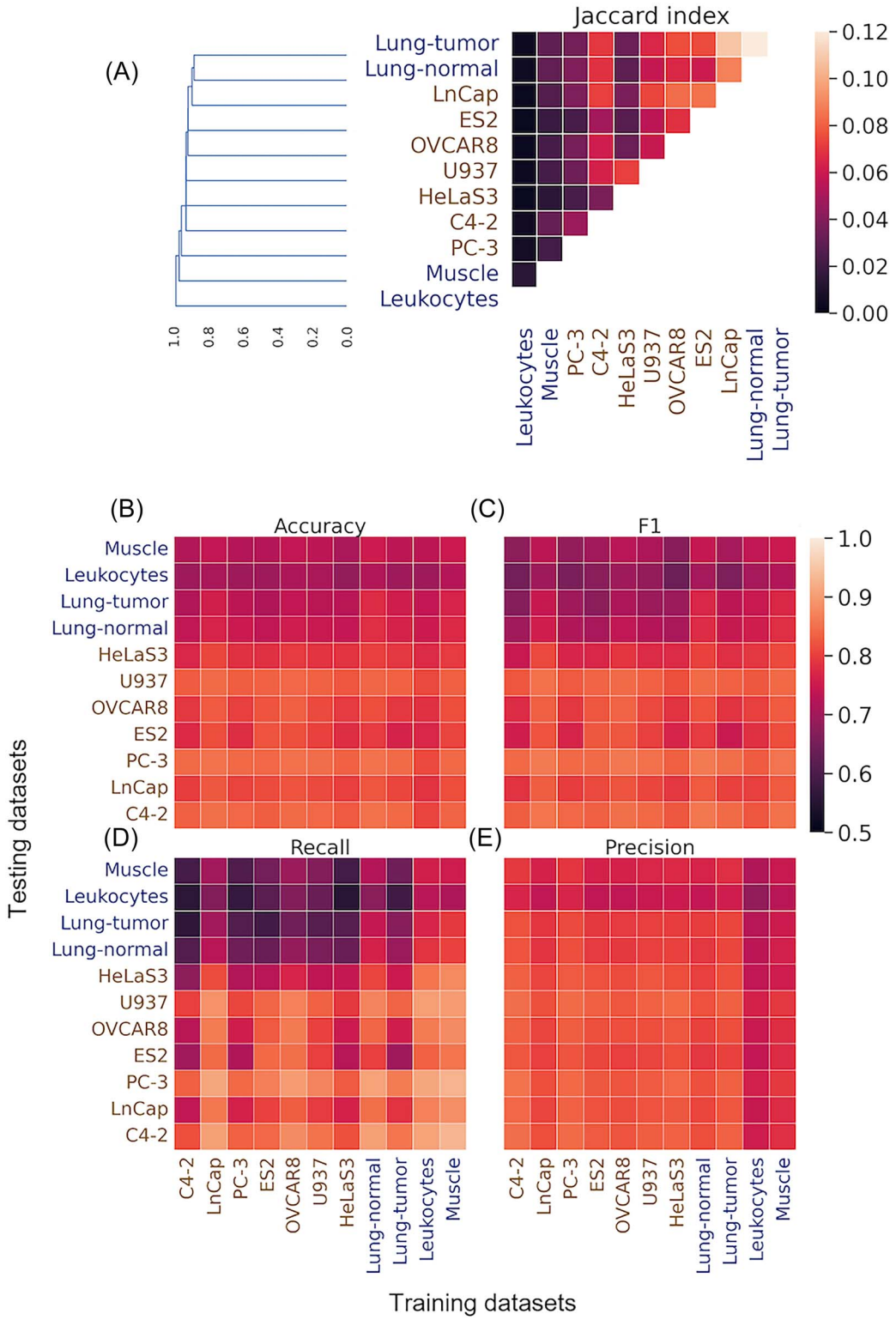

Figure 3. DeepCircle is robust: models trained on one dataset can predict eccDNAs from other tissue origins. (**A**) The Jaccard index showed that eccDNAs derived from various datasets exhibited very low similarity. (**B**)–(**E**) respectively represent the accuracy, F1 score, recall and precision of CNN models trained from one dataset and test on other unseen datasets. If the model was trained and tested using 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. The performance of DNABERT models was demonstrated in Supplementary Figure S1 available online at http://bib.oxfordjournals.org/.

TN compared to the cell line models agree with the previous observation that tissue models generalized better in terms of recall.

Based on the previous results, we hypothesized that DNA sequence patterns that are common in eccDNAs might be enriched in the DNA sequences of consistently predicted

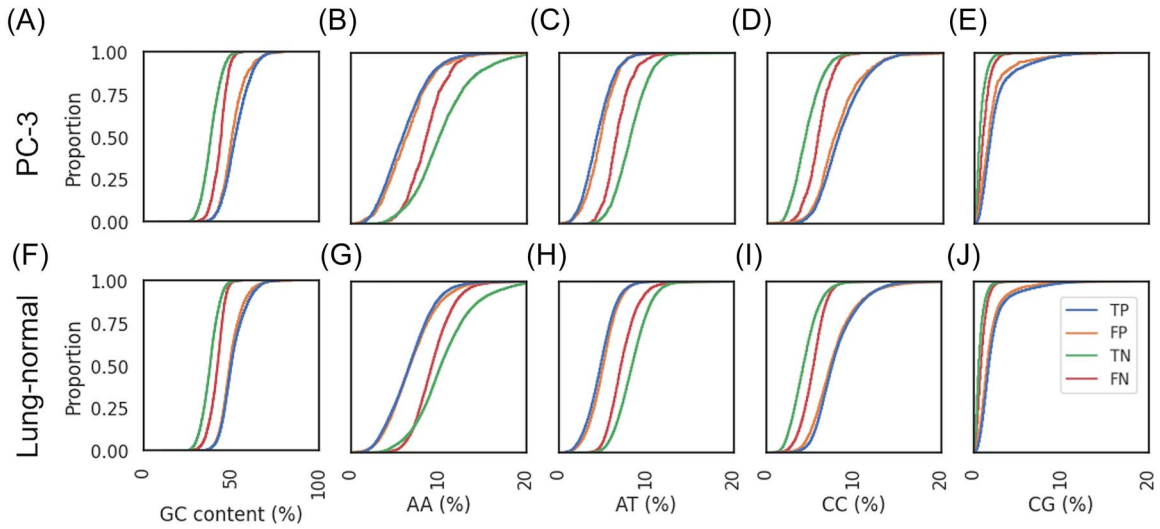

Figure 4. General sequence features used in predicting eccDNAs includes GC content and dinucleotide frequencies. The prediction results of the CNN models were analyzed to obtain the distribution of GC content and dinucleotide (AA, AT, CC and CG) frequencies. Since PC-3 and Lung-normal dataset obtained best performance respectively in the cell line datasets and tissue datasets, here we only showed the results of these two datasets. The results of other datasets are shown in Supplementary Figure S3 available online at http://bib.oxfordjournals.org/. (**A**), (**F**) Distribution of GC content of each DNA sequence in the PC-3 and Lung-normal dataset, respectively. DNA sequences predicted to be eccDNAs exhibited higher GC content compared to the ones predicted to be noneccDNAs. DNA sequences predicted to be eccDNAs contained higher frequencies of dinucleotides with at least one guanine or cytosine (**D**), (**E**), (**I**), (**J**) and lower frequencies for dinucleotides with at least one adenine or thymine (**B**), (**C**), (**G**), (**H**). P and N denotes positive (eccDNAs) and negative (non-eccDNAs) label data, respectively.

TP. To test this hypothesis, we applied the interpretation module of DeepCircle to identify consensus eccDNA-related motifs (see Supplementary Figure S7A, B available online at http://bib.oxfordjournals.org/). Notably, there was a consensus eccDNA-related motif identified in all the datasets, indicating that it might represent the common features of short human eccDNAs (the first motif within the 'High-conf' set in Figure 5C and Supplementary Figure S6C available online at http://bib.oxfordjournals.org/). In contrast to the results that no consensus motifs within the 'Cell' set were identified in tissue datasets, every consensus motif within the 'Tissue' set was identified in some of the cell line datasets. This result indicated that the consensus motifs inferred from the 'Tissue' set might more readily recapitulate the common features of eccDNAs compared to the ones from the 'Cell' set, and corroborated previous results that tissue models generalized better than the cell line models. To further obtain a general understanding regarding the biological functions of these motifs, we used the most representative consensus motif (the first motif within the 'High-conf' set in Figure 5C) to query for similar human DNA motifs (see Methods). It was found that the top 10 most similar DNA binding motifs all belong to the zinc-finger protein family, suggesting their potential roles in the biology underlying short human eccDNAs (Supplementary Table S1 available online at http://bib.oxfordjournals.org/).

DISCUSSION

Short human eccDNAs are abundant in cells and have widespread origins in the genome, leading to the proposition that they are generated by random processes such as spontaneous mutation and apoptosis [13, 14]. Such a random nature puts the predictability of short human eccDNAs into question. We showed that publicly available short human eccDNA datasets do appear to be random: they shared little overlaps in their origins in genomic locations. But the prediction results from DeepCircle showed that short human eccDNAs can be predicted, even by models trained on eccDNAs from different tissues. While the genomic regions of where the short eccDNAs derived from could be relatively random in different cell types, the sequences within are not. Our study demonstrated that learning eccDNA features is feasible, which has ramifications for research intending to use the predictability of eccDNA to gain biological insights or to develop clinical applications.

To our knowledge, DeepCircle provides the first machine learning models to predict the presence of eccDNAs. As a proof of concept on the predictability of short human eccDNAs, we used only the DNA sequences for model training. Sequence-based deep learning models have shown excellent predicting performance on genomic elements known to have conserved sequences such as promoters [33, 34]. EccDNAs have no known conserved sequences, except for telomeric eccDNAs [35]. But even in absence of any prior knowledge of common eccDNA sequences, DeepCircle achieved good predicting performance. Our results demonstrated the effectiveness of deep learning models in learning complex sequence features from a non-conserved genomic element with possibly random distribution, without any input on the knowledge of features.

Although deep learning has exceptional predictive power compared to traditional machine learning methods, the model itself offers no explanation on what is learned. Getting caught with the pitfall of low 'explainability' of deep learning models can result in nonsensical predictions based on data artifacts instead of data features [36]. To verify the validity of DeepCircle's models, we analyzed the prediction results with the interpretation module. Regions predicted to be containing eccDNAs have higher GC content and higher frequencies of dinucleotides containing G or C, consistent with previous reports [11, 15, 18, 23, 37]. Further motif analysis on the predicted results showed that there were common motifs utilized by DeepCircle's models in predicting the presence of eccDNAs across various datasets. Notably, the predictions made by CNN and DNABERT are highly consistent and the two

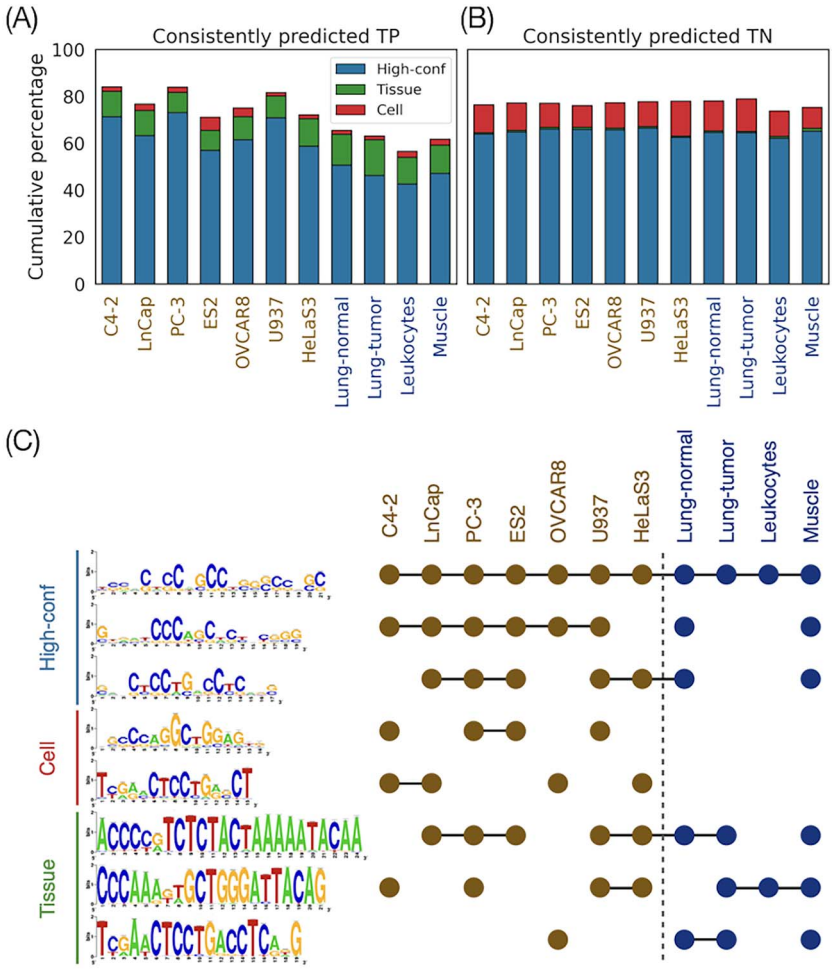

Figure 5. Motif analysis revealed common motifs utilized by models to discern eccDNAs from non-eccDNAs. (**A**) and (**B**) respectively represent the percentages of DNA sequences that were consistently predicted to be TP and TN among models trained on various datasets. (**C**) The consensus eccDNA-related motifs were identified in three sets (as described in Supplementary Figure S7 available online at http://bib.oxfordjournals.org/). Motifs identified in the 'High-conf' set occurred in most of the datasets. Interestingly, motifs identified in the 'Cell' set were only identified in the cell line datasets but the ones in the 'Tissue' set were identified in both the human cell line datasets and tissue datasets. The consensus eccDNA-related motifs identified in the DNABERT were shown in Supplementary Figure S6C available online at http://bib.oxfordjournals.org/.

models utilized similar motifs in their predictions, indicating that shared features were learnt successfully by models with different learning mechanisms (Figure 5C and Supplementary Figure S6C, Table S2 available online at http://bib.oxfordjournals.org/). By querying the motif database, representative motifs used to predict eccDNAs were found to be similar to the binding motifs of the zinc-finger protein family (Supplementary Table S1 available online at http://bib.oxfordjournals.org/). It raises the possibility that the binding of some zinc finger proteins can facilitate the formation of eccDNAs. Moreover, extra copies of eccDNAs may work as a sponge to regulate the available levels of specific zinc-finger proteins, further impacting cell physiology. Through identifying sequence features that contributed to the prediction of eccDNAs, we confirmed the validity of DeepCircle and found common motifs that can be the subject for future studies.

Several studies have shown that the distribution of short eccDNAs have cell type specificity [15, 38–40]. We found little evidence of such specificity at the DNA sequence level. Most models performed well when tested on datasets that they were not trained for, except when testing cell line models on the tissue datasets. The low generalizability of cell line models on the tissue datasets could be due either to the cell lines having less diverse eccDNAs than tissues, or the high cell heterogeneity in the tissue datasets. Nonetheless, a high proportion of eccDNAs were consistently predicted by most models, suggesting that the intrinsic propensity of a genomic region to form eccDNA may be encoded in the DNA sequences [41]. In addition, we did not find any highly enriched dataset-specific motifs (Supplementary Figure S8 available online at http://bib.oxfordjournals.org/). Production of specific eccDNAs have been reported to be induced by increased transcription related to specific cell functions or physiology [1, 42, 43]. Therefore, additional extrinsic factors likely determine whether the sequence-encoded intrinsic potential will be realized, resulting in cell-specific eccDNA distribution.

Although DNA sequence features alone already offer a strong support to the general predictability of short eccDNAs, additional features are likely required to further fine-tune models to predict cell- or condition-specific presence of eccDNAs. DNABERT can only accept DNA sequences for its input, but CNN model can be easily extended to accept other features. For example, epigenetic features [44] can be included to predict specific eccDNAs in different biological or clinical conditions. Currently the availability of specific epigenetic data is limited (Supplementary Figure S9 available online at http://bib.oxfordjournals.org/) but we anticipate

their inclusion in the future extension of DeepCircle. Although we focus only on human cells in this study, short eccDNAs are ubiquitous in eukaryotes [14, 42]. It is possible that the universal predictive power of DNA sequences on eccDNA's presence can be extended to other species to variable degrees, which will offer insights into the evolution of eccDNAs. Proof-of-concept on cross-species prediction has already been shown for DNABERT [33], but we will need to wait for more data to become available to perform a comprehensive study. Our work lays the foundation for exciting potential applications of deep learning models on eccDNA studies, for both the fundamental and clinical aspects.

METHODS

Data sources and eccDNA calling

High-throughput sequencing data enriched in eccDNAs through exonuclease digestion and rolling circle amplification were used in this study [11, 14, 15, 37]. These sequencing data derived from human cell lines (C4–2, PC-3, LnCap, ES2, OVCAR8, HeLaS3 and U937) or tissues (muscle tissues, leukocytes, lung cancer tumor tissues and lung cancer normal tissues) were downloaded from Sequence Read Archive (SRA, leukocyte and muscle tissues: PRJNA419440; human cancer cell lines: PRJNA283289, PRJNA151905; lung tumor tissues and matched normal lung tissues derived from lung cancer patients: PRJNA657416). The genomic coordinates where the eccDNAs derive from were called using Circle-Map (version 1.1.4) [22]. To ensure high sensitivity of the models, we applied the following filters on called eccDNAs to retain eccDNAs of high confidence: eccDNAs that were supported by at least one split read, with read depth $\geq$2 times of that of neighboring regions of equal lengths, with read breadth covered by $\geq$0.8 fraction, and with $\leq$0.9 fraction overlapped with other eccDNAs were considered as candidates for further analysis. To compare the similarities of the eccDNAs in different datasets used in this study, we employed Jaccard index to compute pairwise similarity across the datasets. The Jaccard index was defined in terms of the genomic coordinates of all the eccDNAs derived from any two datasets used for the comparison. Jaccard index was computed by dividing the intersection of all eccDNA base pair counts by the union of all eccDNA base pair counts using bedtools (version 2.29.2) [45].

Training data preparation

To train deep learning models, we used DNA sequences of 1000 base pairs (bp) containing eccDNA and without eccDNA as our positive label and negative label training data respectively. EccDNAs with length $\geq$ 1000 bp, which represented <9% of all eccDNAs in all the datasets, were excluded from the training data since our focus is on the prediction of short eccDNAs (Supplementary Figure S10, Table S3 available online at http://bib.oxfordjournals.org/). To generate the positive label data (eccDNA), coordinates of 1000-bp-long genomic intervals centered on the midpoints of each eccDNA were first computed. In each interval, regions not occupied by the eccDNA were padded with the flanking regions (for eccDNAs with length of exactly 1000 bp length, there was no flanking region). To generate the negative label data (non-eccDNA), we randomly sampled n genomic intervals of 1000 bp length outside of any eccDNA regions and assembly gaps for each dataset, where n is the number of intervals of the positive label data. Each interval was split into a 'region of interest' centered on the midpoint of the interval and flanking regions such that the length distribution of the 'region of interest' in the negative label data mirrors that of the positive label data. After obtaining the genomic coordinates of positive and negative label data, we used bedtools (version 2.29.2) to retrieve corresponding DNA sequences from the human reference genome (hg38, UCSC Genome Browser). To partition training and testing data, we randomly sampled 80% of genomic sequences for each dataset as training data and the remaining 20% as testing data. In both training and testing data, we used a ratio of 1:1 for positive and negative label data.

Overview of DeepCircle

DeepCircle consists of two modules: prediction module and interpretation module (Figure 1). The prediction module is at the core of DeepCircle, which can preprocess input data, train two types of deep learning models, CNN and DNABERT, and make predictions with trained models. During training or testing, depending on the choice of deep learning model, input data will be converted into acceptable formats via the corresponding workflow. The two alternative models in DeepCircle represent different learning approaches: CNN makes predictions based on local features while DNABERT makes predictions based on global features. Following the prediction module is the interpretation module, which infers eccDNA-related motifs by identifying overrepresented motifs in DNA sequences predicted to be eccDNAs.

Data preprocessing

For CNN models, to convert genomic intervals containing DNA sequences into tensors, we utilized one-hot encoding to convert DNA sequences into numerical values. One-hot encoding encodes categorical variables of DNA sequence into binary variables of 0 and 1 corresponding to A, C, G and T. To retain the positional information of eccDNA breakpoints, we split each genomic interval matrix at eccDNA breakpoints into two separate matrices, thus producing a 'region of interest' matrix and a flanking region matrix. To concatenate the two matrices, each matrix was first padded with zeros to have a shape of 1000×4 (i.e. zeros were padded on the flanking region of 'region of interest' matrix and padded on the 'region of interest' of the flanking region matrix). Following padding, the two matrices were stacked together, making a matrix with a shape of 1000×8 for each genomic interval. Finally, for each dataset, all the genomic sequences were stacked, making a matrix with a shape of number of eccDNAs $\times$ 1000 $\times$ 8. For DRNABERT models, we turned the eccDNA data into acceptable format by converting DNA sequences of the genomic sequences into 6-mer sequences (e.g. sequence GATACCC will be converted into GATACC and ATACCC). Due to the limitation from the available pre-trained DNABERT model, we slightly increased the interval size used for DNABERT training into 1024 bp by extending the interval bidirectionally. Six-mer was chosen over other k-mers as it was shown to have the best performance out of all tested k-mers in the previous work [33]. The 6-mer sequence is the final input format for training and testing data. To facilitate an unbiased comparison of DNABERT to CNN, the training data and testing data of DNABERT is essentially the same as the ones used by CNN except the format (6-mers versus one-hot encoding).

Deep learning model architecture and model training

In the CNN models, we assumed each nucleotide in DNA is a single word and short consecutive sequences of words are features that can be used for predicting eccDNAs. Therefore, we implemented a 1D CNN model, which is better suited for sequence classification. To obtain optimal performance, we fine-tuned several CNN's

hyperparameters including number of filters, kernel size of each convolutional layer, number of convolutional layers, learning rate, batch size, number of neurons in the fully connected layer, and number of training epochs. We evaluated the CNN models with two, three, four, and five convolutional layers and found out the CNN models with two convolutional layers had comparable performance in precision and recall (Supplementary Figure S11 available online at http://bib.oxfordjournals.org/). The final CNN model consists of two convolutional layers, each of which has 16 filters with a kernel size of 3, followed by a rectified linear unit (ReLU) activation function and a dropout layer with dropout rate of 0.2. To prevent the model from overfitting, we added L2 regularizers (regularization factor of 0.0001) to apply penalty on layer weights and dropout layers immediately following each convolutional layer. Following convolutional layers are a max pooling layer with pool size of 2 and a flatten layer. Finally, the flatten layer is connected to a fully connected layer with 100 nodes followed by a dropout layer with dropout rate of 0.2 and another fully connected layer with 2 nodes representing eccDNA or non-eccDNA. To train the CNN models, the number of training epochs was set to 100, batch size was set to 32, and Adam was used as an optimizer with a learning rate of 0.001. The model with the lowest binary cross entropy during training was used for testing. The CNN models were implemented using Tensorflow 2.2.0 and Keras 2.3.1 as backend.

DNABERT adopted the mechanism of self-attention to model DNA as a language [33]. Training a DNABERT model to predict a specific genomic element requires a two-step process including pre-training and fine-tuning. To train DNABERT models for eccDNA prediction, we used a DNABERT model pre-trained on the human genome and fine-tuned the model for the task of eccDNA prediction using human eccDNA data. During the training process, we set the training batch size to 4, warmup percentage to 0.1, dropout probability to 0.1, and weight decay to 0.01. Other hyperparameters including training epoch, logging steps, save steps, and learning rate were customized based on the size of each dataset as listed in Supplementary Table S4 available online at http://bib.oxfordjournals.org/.

Model evaluation

To evaluate the performance of the deep learning models, we employed the commonly used metrics including precision, recall, F1 score and accuracy to evaluate model performances, with precision = TP/(TP + FP), recall = TP/(TP + FN), F1 = 2∗Precision∗Recall/(Precision + Recall), and accuracy = (TP + TN)/(TP + TN + FP + FN); where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

Calculation of dinucleotide frequency

To compare the sequence properties predicted to be eccDNA or non-eccDNAs, we computed the dinucleotide frequency of 1000-bp-long DNA sequences derived from various cases of prediction including TP, FP, TN and FN. To quantify the difference in dinucleotide frequencies across various classes, for each type of dinucleotide, we used Kruskal-Wallis test to infer whether the four classes are derived from the same population. The result of the test with a P-value ≤ 0.05 and eta-square ≥ 0.14 were deemed as significant.

Inference of eccDNA-related motifs

To infer eccDNA-related motifs, we performed motif analysis with STREME in the MEME suite [46] to identify overrepresented motifs in the predicted eccDNAs. Within the prediction results of 1000-bp-long sequences, TP was used as primary sequences and TN was used as background sequences for inference of eccDNA-related motifs. The criteria for identification of motifs include a minimum motif width of 8, maximum motif width of 15, and P-value ≤ 0.05. To facilitate a concise representation, the top 10 motifs with lowest P-values were selected for further analysis.

Inference of consensus motifs and database query

To infer consensus motifs, we classified the motifs inferred from individual motifs into different categories and clustered them using RSAT [47] (Supplementary Figure S6 available online at http://bib.oxfordjournals.org/). We used the consensus motifs identified in this study to query similar motif in CIS-BP 2.00 human database with Tomtom [48, 49]. For the sake of brevity, we selected the top 10 motifs with lowest E-values (Supplementary Table S1 available online at http://bib.oxfordjournals.org/).

Key Points

- Short eccDNAs were suggested to be randomly generated but we show that they can be predicted from DNA sequences.
- We show that eccDNAs from diverse tissue origins have little overlap in genomic locations but share sequence features.
- Our study showcases how deep learning methods can help answer a key biological question.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

ACKNOWLEDGEMENTS

The authors would like to show their gratitude to Ru-In Jian and Hung-Yeh Lee for their technical support. In addition, the authors also thank Chia-Wei Liao, Chien-Fu Jeff Liu, Huei-Wen Chen, and Sheng-Fang Su for their kind suggestions.

FUNDING

National Science and Technology Council (Taiwan) (110-2811-E-001-502); and Academia Sinica Grand Challenge Program (AS-GC-110-L15). Funding for open access charge: Academia Sinica Grand Challenge Program (AS-GC-110-L15).

DATA AVAILABILITY

All of the data used for eccDNA calling in this study is publicly available [11, 14, 15, 37]. The source code used for replicating this study, bed files of eccDNAs used in this study and the trained CNN models are available at the GitHub repository of DeepCircle (https://github.com/bio-it-station/DeepCircle). The pre-trained, fine-tuned DNABERT models and the reference genome sequence used in this study were deposited at the Open Science Framework (https://osf.io/wmjdg/).

References

1. Hull RM, Houseley J. The adaptive potential of circular DNA accumulation in ageing cells. *Curr Genet* 2020;**66**:889–94.
2. Verhaak RGW, Bafna V, Mischel PS. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat Rev Cancer* 2019;**19**:283–8.
3. Kim H, Nguyen N-P, Turner K, *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet* 2020;**52**:891–7.
4. Robert M, Crasta K. Breaking the vicious circle: extrachromosomal circular DNA as an emerging player in tumour evolution. *Semin Cell Dev Biol* 2022;**123**:140–50.
5. Ståhl F, Wettergren Y, Levan G. Amplicon structure in multidrug-resistant murine cells: a nonrearranged region of genomic DNA corresponding to large circular DNA. *Mol Cell Biol* 1992;**12**:1179–87.
6. Turner KM, Deshpande V, Beyter D, *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* 2017;**543**:122–5.
7. Wu S, Turner KM, Nguyen N, *et al.* Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* 2019;**575**:699–703.
8. Helmsauer K, Valieva ME, Ali S, *et al.* Enhancer hijacking determines extrachromosomal circular MYCN amplicon architecture in neuroblastoma. *Nat Commun* 2020;**11**:5823.
9. Zhu Y, Gujar AD, Wong C-H, *et al.* Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. *Cancer Cell* 2021;**39**:694–707.e7.
10. Cohen S, Regev A, Lavi S. Small polydispersed circular DNA (spcDNA) in human cells: association with genomic instability. *Oncogene* 1997;**14**:977–85.
11. Shibata Y, Kumar P, Layer R, *et al.* Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. *Science* 2012;**336**:82–6.
12. Paulsen T, Shibata Y, Kumar P, *et al.* Small extrachromosomal circular DNAs, microDNA, produce short regulatory RNAs that suppress gene expression independent of canonical promoters. *Nucleic Acids Res* 2019;**47**:4586–96.
13. Wang Y, Wang M, Djekidel MN, *et al.* eccDNAs are apoptotic products with high innate immunostimulatory activity. *Nature* 2021;**599**:308–14.
14. Møller HD, Ramos-Madrigal J, Prada-Luengo I, *et al.* Near-random distribution of chromosome-derived circular DNA in the condensed genome of pigeons and the larger, more repeat-rich human genome. *Genome Biol Evol* 2020;**12**:3762–77.
15. Dillon LW, Kumar P, Shibata Y, *et al.* Production of extrachromosomal MicroDNAs is linked to mismatch repair pathways and transcriptional activity. *Cell Rep* 2015;**11**:1749–59.
16. Henriksen RA, Jenjaroenpun P, Sjøstrøm IB, *et al.* Circular DNA in the human germline and its association with recombination. *Mol Cell* 2022;**82**:209–217.e7.
17. Mehanna P, Gagné V, Lajoie M, *et al.* Characterization of the microDNA through the response to chemotherapeutics in lymphoblastoid cell lines. *PloS One* 2017;**12**:e0184365.
18. Zhu J, Zhang F, Du M, *et al.* Molecular characterization of cell-free eccDNAs in human plasma. *Sci Rep* 2017;**7**:10968.
19. Sin STK, Jiang P, Deng J, *et al.* Identification and characterization of extrachromosomal circular DNA in maternal plasma. *Proc Natl Acad Sci U S A* 2020;**117**:1658–65.
20. Zhu Y, Liu Z, Guo Y, *et al.* Whole-genome sequencing of extrachromosomal circular DNA of cerebrospinal fluid of medulloblastoma. *Front Oncol* 2022;**12**:934159.
21. Lv W, Pan X, Han P, *et al.* Circle-Seq reveals genomic and disease-specific hallmarks in urinary cell-free extrachromosomal circular DNAs. *Clin Transl Med* 2022;**12**:e817.
22. Prada-Luengo I, Krogh A, Maretty L, *et al.* Sensitive detection of circular DNAs at single-nucleotide resolution using guided realignment of partially aligned reads. *BMC Bioinformatics* 2019;**20**:663.
23. Kumar P, Kiran S, Saha S, *et al.* ATAC-seq identifies thousands of extrachromosomal circular DNA in cancer and cell lines. *Sci Adv* 2020;**6**:eaba2489.
24. Mann L, Seibt KM, Weber B, *et al.* ECCsplorer: a pipeline to detect extrachromosomal circular DNA (eccDNA) from next-generation sequencing data. *BMC Bioinformatics* 2022;**23**:40.
25. Zhang P, Peng H, Llauro C, *et al.* ecc_finder: a robust and accurate tool for detecting extrachromosomal circular DNA from sequencing data. *Front Plant Sci* 2021;**12**:743742.
26. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–44.
27. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform* 2017;**18**:851–69.
28. Iuchi H, Matsutani T, Yamada K, *et al.* Representation learning applications in biological sequence analysis. *Comput Struct Biotechnol J* 2021;**19**:3198–208.
29. Alipanahi B, Delong A, Weirauch MT, *et al.* Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8.
30. Yang B, Liu F, Ren C, *et al.* BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics* 2017;**33**:1930–6.
31. Li J, Pu Y, Tang J, *et al.* DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences. *Brief Bioinform* 2021;**22**:bbaa159.
32. Devlin J, Chang M-W, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint, arXiv:1810.04805, 2018.
33. Ji Y, Zhou Z, Liu H, *et al.* DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:btab083.
34. Oubounyt M, Louadi Z, Tayara H, *et al.* DeePromoter: robust promoter predictor using deep learning. *Front Genet* 2019;**10**:286.
35. Hartig JS, Kool ET. Small circular DNAs for synthesis of the human telomere repeat: varied sizes, structures and telomere-encoding activities. *Nucleic Acids Res* 2004;**32**:e152.
36. Dincer AB, Janizek JD, Lee S-I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* 2020;**36**:i573–82.
37. Kumar P, Dillon LW, Shibata Y, *et al.* Normal and cancerous tissues release extrachromosomal circular DNA (eccDNA) into the circulation. *Mol Cancer Res* 2017;**15**:1197–205.
38. Gaubatz JW, Flores SC. Tissue-specific and age-related variations in repetitive sequences of mouse extrachromosomal circular DNAs. *Mutat Res* 1990;**237**:29–36.
39. Shoura MJ, Gabdank I, Hansen L, *et al.* Intricate and cell type-specific populations of endogenous circular DNA (eccDNA) in Caenorhabditis elegans and Homo sapiens. *G3 (Bethesda)* 2017;**7**:3295–303.
40. Wang K, Tian H, Wang L, *et al.* Deciphering extrachromosomal circular DNA in Arabidopsis. *Comput Struct Biotechnol J* 2021;**19**:1176–83.
41. Kaplan N, Moore IK, Fondufe-Mittendorf Y, *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 2009;**458**:362–6.

42. Hull RM, King M, Pizza G, *et al.* Transcription-induced formation of extrachromosomal DNA during yeast ageing. *PLoS Biol* 2019;**17**:e3000471.
43. Møller HD, Mohiyuddin M, Prada-Luengo I, *et al.* Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. Nature. *Communications* 2018;**9**:9.
44. Baisya DR, Lonardi S. Prediction of histone post-translational modifications using deep learning. *Bioinformatics* 2020;**36**: btaa1075.
45. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.
46. Bailey TL. STREME: accurate and versatile sequence motif discovery. *Bioinformatics* 2021;**37**:2834–40.
47. 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.
48. Gupta S, Stamatoyannopoulos JA, Bailey TL, *et al.* Quantifying similarity between motifs. *Genome Biol* 2007;**8**:R24.
49. Weirauch MT, Yang A, Albu M, *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 2014;**158**:1431–43.