Computational and Structural Biotechnology Journal 21 (2023) 2591–2601

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Database article

EccBase: A high-quality database for exploration and characterization of extrachromosomal circular DNAs in cancer

Haiyang Sun [b,c], Xinyi Lu [b], Lingyun Zou [a,c,*]

[a] Department of Clinical Data Research, Chongqing Emergency Medical Center, Chongqing Key Laboratory of Emergency Medicine, Chongqing University Central Hospital, Chongqing University, 400014 Chongqing, China
[b] State Key Laboratory of Medicinal Chemical Biology, Nankai University, 300350 Tianjin, China
[c] Shenzhen Baoan Women's and Children's Hospital, Jinan University, 518102 Shenzhen, China

ARTICLE INFO

Article history:
Received 18 January 2023
Received in revised form 13 April 2023
Accepted 13 April 2023
Available online 14 April 2023

Keywords:
Extrachromosomal circular DNA
Database
Sequence alignment
Cancer
Tissue specificity
Machine learning

ABSTRACT

Extrachromosomal circular DNAs (eccDNAs) are widely observed in eukaryotes. Previous studies have demonstrated that eccDNAs are essential to cancer progression, and found that they can not only express in normal cells to regulate RNA, but also function differently in different tissues. It is of major interest to conduct computational or experiments assay to elucidate the mechanisms of eccDNA function, uncover key eccDNAs associated with diseases, and even develop related algorithms for liquid biopsy. Naturally, a comprehensive eccDNAs data resource is urgently needed to provide annotation and analysis more in-depth research. In this study, we constructed the eccBase (http://www.eccbase.net) in literature curation and database retrieval, which was the first database mainly collecting eccDNAs from Homo sapiens (n = 754,391) and Mus musculus (n = 481,381). Homo sapiens eccDNAs were taken from 50 kinds of cancer tissue and/or cell line, and 5 kinds of healthy tissues. The Mus musculus eccDNAs were sourced from 13 kinds of healthy tissue and/or cell line. We thoroughly annotated all eccDNA molecules in terms of basic information, genomic composition, regulatory elements, epigenetic modifications, and raw data. EccBase provided users with the ability to browse, search, download for targets of interest, as well as similarity alignment by the integrated BLAST. Further, comparative analysis suggested the cancer eccDNA is composed of nucleosomes and is prominently derived from the gene-dense regions. We also initially revealed that eccDNAs are strongly tissue-specific. In short, we have started a robust database for eccDNA resource utilization, which may facilitate studying the role of eccDNA in cancer development and therapy, cell function maintenance, and tissue differentiation.

1. Introduction

Extrachromosomal circular DNA (eccDNA) are closed-circle molecules within the nucleus, derived from chromosomal DNA. In 1965, eccDNA was first found in tumor-derived cells via cytogenetic examination [1,2]. Subsequently, the application of electron microscopy and high-throughput sequencing showed that eccDNAs are ubiquitous in eukaryotes, such as *Mus musculus* (*mmu*) and *Saccharomyces cerevisiae* (*sce*) [3,4]. Regarding how eccDNA is generated, previous studies have proposed multiple potential models, including the DSBs (Double-Strand Breaks), BFB (Breakage-Fusion-Bridge) cycle, translocation-deletion-amplification, and Chromothripsis [5,6]. Recently, Yuangao Wang et.al. reported that eccDNA are products of cell apoptosis, that is apoptotic DNA fragmentation followed by end-to-end ligation by DNA ligase 3 [7]. The sizes of eccDNAs vary from tens of bases to millions of bases [5,8]. In comparison, the sizes of eccDNAs sourced from healthy individuals are usually smaller, whereas eccDNAs detected in tumors are substantially larger [9,10]. The circle structure of eccDNAs leads to the rewiring of enhancers and genes, as well as strong accessibility, together conferring higher transcriptional capacity on eccDNAs [10,11]. Further, eccDNAs are acentric, so their segregation during mitosis and meiosis does not comply with Mendel's laws of inheritance [12].

The significance of eccDNA can be summarized as follows. (1) The eccDNAs are pivotal for interpreting cell dysfunction in cancers. First, random distribution of eccDNAs into daughter cells inevitably

* Corresponding author at: Department of Clinical Data Research, Chongqing Emergency Medical Center, Chongqing Key Laboratory of Emergency Medicine, Chongqing University Central Hospital, Chongqing University, 400014 Chongqing, China.
E-mail address: zoulingyun@cqu.edu.cn (L. Zou).

https://doi.org/10.1016/j.csbj.2023.04.012

leads to intrapulmonary heterogeneity [13,14]. Next, high transcriptional activity of eccDNA causes over expression of focal oncogenes located in the eccDNAs [10,15]. Third, the eccDNAs equip cells with highly adaptable to the environment, which is indispensable for understanding drug resistance [14]. For example, in the presence of methotrexate, tumor cells accumulate high levels of eccDNAs containing the resistance gene DHFR [16–18]. In addition, the eccDNAs have strong immunostimulatory activity and can be sensed by the STING pathway [7]. Undoubtedly, cancer eccDNAs are of profound significance for elucidating cancer metastasis, therapeutic resistance, tumor heterogeneity evolution, and poor prognoses. (2) In normal cells, the eccDNAs are mostly smaller than in cancer cells, but have been shown to express functional small regulatory RNA including microRNA and novel si-like RNA [19]. (3) The distribution and function of eccDNAs are clearly tissue-specific. Laura W. Dillon et. al. revealed that the eccDNAs in mouse sperm are highly enriched for L1Md_T (repetitive elements of the LINE-1) relative to other tissues [20]. This indicated the value of exploring different characteristics and functions of eccDNAs among various types of tissues and samples. It is therefore significant to pave the way for a solution to the problem of eccDNAs. A few studies have recently reported the results of collation and analysis of certain types of eccDNA molecules. For example, the eccDNAdb database (http://www.eccdnadb.org/) integrated 1450 eccDNAs in human cancers that were predicted from publicly available WGS data and provided annotation and retrieval of these molecules [21]. However, researchers in different fields need a comprehensive database to utilize and analyze all specifically determined eccDNA molecules, but such a resource is still lacking. Here, we have purposefully created a more comprehensive eccDNAs database for studies on cancer and various other diseases.

Herein we developed the eccBase (the extrachromosomal circular DNA database, http://www.eccbase.net/) for the collection and annotation of all identified eccDNAs from *Homo sapiens* (*hsa*) and *mmu*. These molecules were drawn from a rich variety of cancer tissues and/or cell lines, as well as normal tissues. To comprehensively profile all eccDNAs, a total of 50 features were annotated, including basic features, molecular biology properties, and source data. We provide users with functional modules such as Browse, Search, Download, BLAST, and Submission in eccBase, which are conducive to exploring the eccDNA world conveniently and effectively. Although efforts of presenting more effective annotation are still needed, we highlight the significance of eccBase advancing the clarification of the role of eccDNAs in the development and clinical treatment of cancer, as well as in embryonic development and tissue differentiation.

2. Materials and methods

2.1. Data collection

The development pipeline of eccBase was illustrated in Fig. 1. First, we conducted an extensive resource search in the PubMed, SRA, ENA, and GEO databases until January 20, 2023 using the keywords ‘extrachromosomal circular DNA’, ‘extrachromosomal DNA’, ‘extra chromosomal DNA’, ‘ecDNA’, and ‘eccDNA’. As a result, 963 articles were recovered, from which experimentally and putatively supported eccDNAs were collected. Considering the species in which eccDNA was frequently reported, and the biological features of eccDNA could be well characterized, we mainly collect eccDNAs from *hsa* , *mmu, Gallus gallus* (*gga*) and *sce*, and only *hsa* and *mmu* eccDNAs were used for comprehensive annotation. Then, Mimicking eccDNA, cell-free DNA, ribosome DNA, mitochondrial DNA, and incomplete eccDNA were excluded in this study. The eccDNAs consisting of distinct chromosomal fragments were very few and therefore have not been incorporated into the current version. Finally, a total of 21 publications providing eccDNA information were retained [3,9,10,20,22–38]. Among them, 19, 2, 1 and 2 are related to *hsa*, *mmu*, *gga* and *sce* eccDNAs, respectively. The eccDNA data will be updated every three months if relevant literature is published. The new data will be formatted using an automated annotation pipeline and then imported into the database from the backend.

2.2. Aggregating eccDNA information

We read through each article as well as supplementary material to obtain the key information, including source species, genomic position, cell lines and/or tissues of origin, associated diseases (if any), identification methods, sequencing methods, accession number of raw data, biological function descriptions, and publication information. Unique eccDNAs were preserved based on genomic locations and their pooled characteristics. Since the number of published datasets is not huge, we did not use natural language processing models to analyze the literature and web texts in order to guarantee accuracy.

2.3. Annotating eccDNA characteristics

The process of annotating eccDNA was automatized with shell script. All the eccDNAs received an identification code (ecc_*species*_xxx) and further annotated with more detailed information as follows (Fig. S3). (1) The genomic position of eccDNA was re-labeled in the specific genome version (the hg19 and hg38 for *hsa*, the mm9 and mm10 for *mmu*) using LiftOver [39]. (2) The length of the eccDNAs was calculated from their starting and ending positions in the genome. (3) The reference genomes of *hsa* and *mmu* were downloaded from UCSC [39]. Then, the DNA sequence of the eccDNA was extracted from the corresponding fasta file by the samtools (version 1.10) [40]. (4) The GC content (keep three decimal places) of eccDNAs was calculated using bedtools (version 2.29.2) [41]. (5) The GTF files of the *hsa* and *mmu* were downloaded from GENECODE (release 33) [42]. For eccDNAs of *hsa* and *mmu* that harbor genes or gene segments, we annotated gene symbols, gene numbers, and the proportion of the portion that covered by known genes using bedtools and bedops (version 2.4.39) [43]. We also separately annotated eccDNAs with the information of gene types, including protein-coding genes, long noncoding RNA genes, noncoding RNA genes, and pseudogenes. (6) The transcript start sites were obtained from the Ensembl database (release 104) and were counted in the eccDNAs. (7) Repetitive sequence information was extracted from the UCSC RepeatMasker table (Data last updated for hg19, hg38, mm9, and mm10 were 2020–02–20, 2021–09–03, 2007–07–20, 2021–04–08, respectively) to label the number and type of repeats in eccDNAs. (8) Exons, introns, 5′UTRs and 3′UTRs information was obtained from the UCSC knownGene and ncbiRefSeq (Data last updated for hg19, hg38, mm9, and mm10 was 2013–06–14, 2021–12–08, 2011–03–02, 2021–04–23, respectively). Based on this information, the number of exons, introns, 5′UTRs, and 3′UTRs in eccDNAs was calculated, as well as the proportion of the portion covered by them. (9) CpG island library files were downloaded from the UCSC cpgIslandExt (Data last updated for hg19, hg38, mm9, mm10 were 2020–03–13, 2020–02–20, 2007–10–25, 2021–04–08, respectively) to annotating the CpG islands distribution in eccDNAs. (10) The enhancer regions were acquired from EnhancerAtlas (version 2.0), and the promoter information was retrieved from UCSC epdNewPromoter (Data last updated for *hsa* and *mmu* were 2018–05–15 and 2018–06–14, respectively) to callout enhancers and promoters in eccDNAs. (11) The transcripts factor binding sites were retrieved from UCSC encRegTfbsClustered for *hsa* (Data last updated was 2019–05–16), and from ORegAnno database for *mmu* (Data last updated was 2016–01–19). (12) The histone modification peaks as well as the DNase I hypersensitive sites were sourced from ENCODE (version 4)

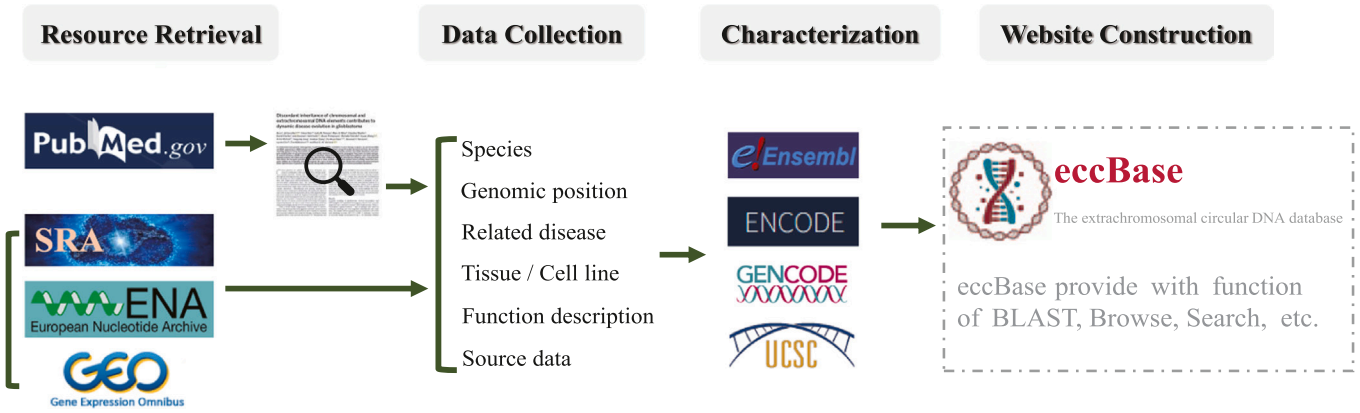

Fig. 1. The workflow of developing eccBase.

[44]. Finally, each eccDNA in the database was annotated with up to 50 features covering the vast majority of known sequence knowledge (Fig. S3).

2.4. Web construction

EccBase is running on an Alibaba Cloud Linux 3.2104 64-bit (Linux kernel 5.10.23) server, with 2 Intel(R) Xeon(R) Platinum 8269CY CPUs, 2 TB HDD and 8 GB RAM. Data are managed using MySQL (Ver 14.14 Distrib 5.7.34). Web interface was designed by LayUI (version 2.56) and jQuery (1.7.2) for data retrieve, visualization, and online interactive services. The BLAST alignment service was constructed by WebShell based on the NCBI BLAST+ algorithm (version 2.12.0 +) [45]. The web application was developed with the Zend (PHP framework), and hosted on Nginx (version 1.18.0) web server. Accordingly, eccBase is supported by principal standard-compliant web browsers such as Firefox, Google Chrome, Internet Explorer and Safari. The background management system of eccBase supports automatic updating of structured eccDNA data.

2.5. Bioinformatics and statistical analysis

The bioinformatic and statistical analysis was performed by in-house R and Python scripts. Wilcoxon rank-sum test and Kruskal-Wallis (K-W) test were employed for comparison between two groups and multiple groups, respectively (Figs. 5 and 6). Statistical significance was only accepted when the p-value < 0.05 (Fig. 6, Fig. S5, Table S1). The clusterProfiler package was used for KEGG enrichment analysis (Fig. S4). Based on available annotations, the eccDNA genomic regions constitutions herein were roughly considered as comprising of promoters, exons, introns, UTRs, and intergenic regions which are beyond that (Fig. 5). The proportion of the genome covered by repetitive sequences was calculated using UCSC RepeatMasker table and marked in Fig. 6.

2.6. Machine learning

To build the machine learning model, eccDNAs from gastric cancer (n = 10,216) and healthy human blood (n = 8989) were selected as positive training samples and negative training samples, respectively. To effectively and systematically represent an eccDNA, 16 features based on sequence were extracted from the above annotation results and then scaled by Z-score. Sixteen features include sequence length, GC content, the proportion of the portion of eccDNA covered by genes, protein-coding genes, lncRNAs, ncRNAs, pseudogenes, exons, introns, 5′ UTRs, 3′ UTRs, repetitive sequences, CpG islands, promoters, enhancers, and DNase I sites. Next, the performance of algorithms commonly used in binary classification was tested and evaluated [46–48]. The following 6 classifiers were constructed using scikit-learn (version 0.23.1): eXtreme Gradient Boosting (XGB), Random Forest (RF), Support Vector Machine (SVM), Neutral Network (NNet), Naive Bayes (Bayes), Logistic Regression (LR) [49]. These models were trained by using 10-fold cross-validation tests, where their performance was evaluated by using the area under the curve (AUC). The hyperparameters of all binary classification models are determined by grid search[50]. We take RF as an instance, whose key parameter of 'n_estimators were given a range from 100 to 900 by a step size of 200, key parameter of 'max_depth' were given a range from 2 to 10 by a step size of 2. Obviously, that would yield 25 parameter combinations, then grid search would output the one with highest AUC. Table S2 shows the range of parameters used to optimize all classifiers, as well as the final parameters used.

3. Results

3.1. Statistics of eccBase data

A total of 1235,772 eccDNAs were collected into eccBase via literature curation and public database retrieval, including 754,391 records sourced from *hsa* and 481,381 records sourced from *mmu*. For those eccDNAs from *hsa*, 79.32% (n = 598,395) were distributed in up to 28 kinds of cancers, and the left 20.68% (n = 156,007) were from healthy tissues (Fig. 2A). Specifically, 15 kinds of cancers all produced more than 5000 eccDNAs, while only 2 kinds of cancers produced less than 10 eccDNAs. The total number of experimentally identified eccDNAs from 33 kinds of tissues and/or 17 kinds of cell lines was 325,635 and/or 428,721, respectively (Fig. 2B-C). The percentage of eccDNAs from tissues ranged from 1.52% (Testis) to 41.38% (Muscle), except for the 'Other' category consisting of 18 kinds of tissue, which had fewer than 5000 eccDNAs. The percentage of eccDNAs from cell lines ranged from 1.04% (FaDu_DDP cell lines) to 26.48% (ES2). All the *mmu* eccDNAs were derived from healthy lineages, where 2.16% (n = 10,517) detected in the (NIH3T3) cell line and 97.84% (n = 470,920) detected in 13 types of tissue (Fig. 2D). The proportion of eccDNAs in these tissues varied from 1.08% (adult mouse sperm) to 18.68% (adult mouse thymus).

The size of *hsa* eccDNAs ranges from < 100 bp to ~243 Mb, but most (96.34%, n = 726,815) are shorter than 1 Mb (Fig. 3A). Notably, 63.38% (n = 478,109) of eccDNAs in *hsa* contained at least 1 gene, mainly protein-coding genes. As the number of genes carried increases, the number of eccDNAs decreases exponentially. (Fig. 3C, Fig. S1A). Distribution of repeats in eccDNAs varied widely, for example, 70.60% (n = 532,585) of *hsa* eccDNAs have at least 1 repetitive sequence, mostly SINE repeats (Fig. 3E, Fig. S2A). Size of eccDNAs in *mmu* ranges from < 100 bp to 5074 bp, with most are around 200 bp

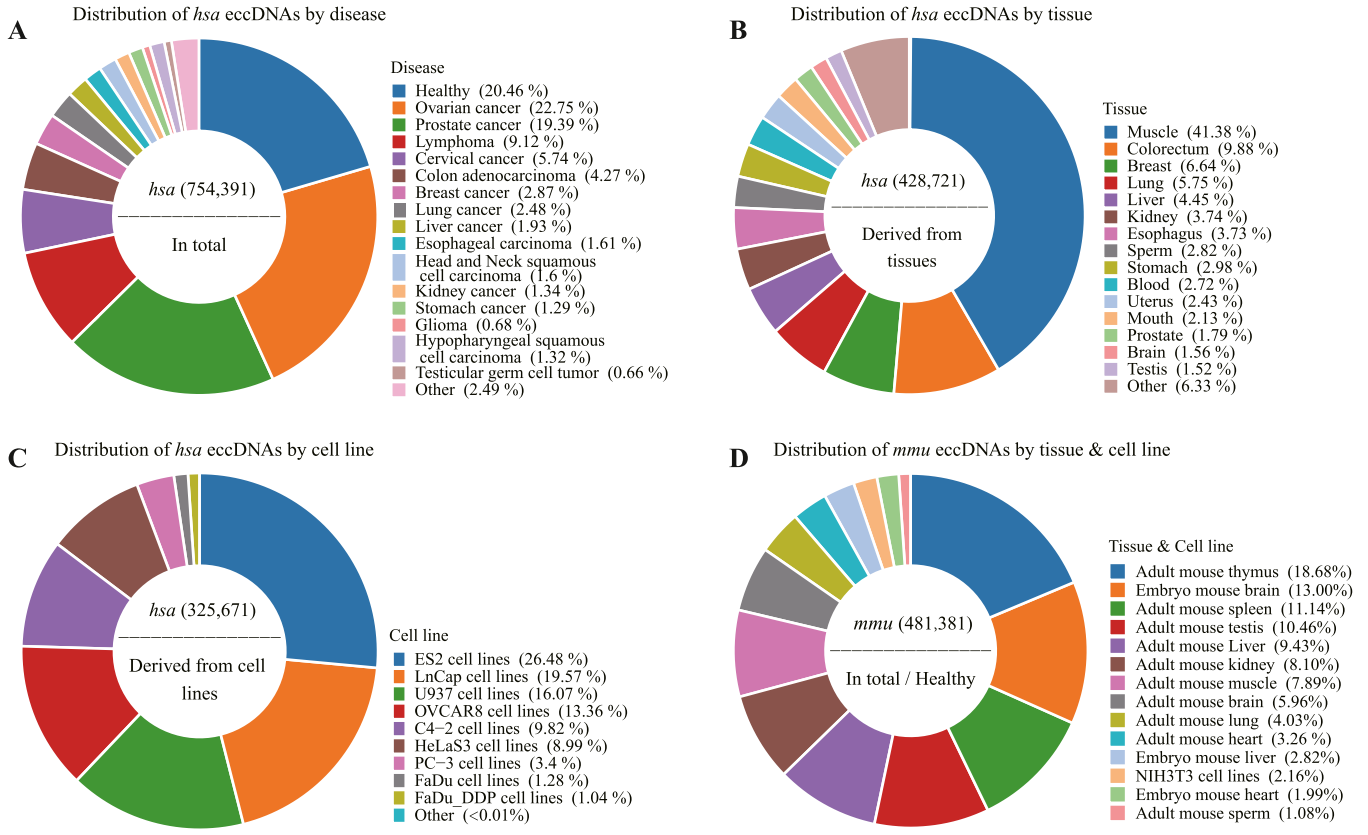

Fig. 2. Percentage of *hsa* and *mmu* eccDNAs distributed into categories of associated disease, categories of sourced tissue, and/or of the cell line. (A) Percentage of *hsa* eccDNAs in different diseases. (B) Percentage of *hsa* eccDNAs in different tissues. (C) Percentage of *hsa* eccDNAs in different cell lines. (D) Percentage of *mmu* eccDNAs in different tissues and/or cell lines.

(Fig. 3B). Specifically, 60.35% (n = 290,513) of *mmu* eccDNAs hold at least 1 gene and are predominantly protein-coding genes (Fig. 3D, Fig. S1B). Unlike *hsa* eccDNAs, only 41.62% (n = 200,547) of *mmu* eccDNAs contain repetitive sequences, in which the SINE repeat is the dominant (Fig. 3F, Fig. S2B).

3.2. Design of eccBase interface

Four functional modules including Browse, Search, BLAST and Download were constructed in the front end of eccBase web, which will promote users to efficiently explore the features and functions of eccDNA molecules (Fig. 4A). The Browse module was designed as display matching eccDNA entries by selecting keywords in the directory tree, which are logical combinations of related diseases, derived lineages, and identification methods (Fig. 4B). EccDNAs from Browse retrieval will list as a brief table and can be selected and exported for local analysis. The full molecular information is accessible via the hyperlink to the eccDNA ID, including basic features, genetic sequences, gene annotation, DNA elements, and source data (Figure S3). We developed “Keyword search” and “Advanced search” in the Search page for different search preferences. “Advanced Search” is customizable to the user. This function not only provides options for common species and diseases, but also telegraphs the search range by entering constraints such as gene name, transcription factor name, proportion of carrying genes, the number of transcript start sites or CpG islands (Fig. 4C). We have further developed an online BLAST server that accepts sequences and parameters input from web pages(Fig. 4D). Users can be helped in discovering and annotating novel eccDNA candidates via similarity alignment and homology identification. Additionally, eccBase supports users to download all the eccDNAs and their compiled information. Note that the eccDNAs sourced from *sce* (n = 2010) and *gga* (n = 700,087) with only basic and source information are also available on the Download page (Fig. 4E).

3.3. Characterization of cancer eccDNAs

The eccDNAs of *hsa* were categorized into the cancer-related group and the normal group (health-related) to perform the comparison of molecular biological characteristics. First, in the length < 3 kb range that contains most eccDNAs, the normal eccDNAs showed an unimodal size distributions peaking at ~100 bp, but interestingly, the cancer eccDNAs showed a multimodal size distribution which peaking at ~192, ~359, ~536, and ~722 bp (Fig. 5A-B). Second, as shown in Fig. 5C, the distribution of fractions of eccDNA molecules covered by repeats clearly indicates that cancer eccDNAs contain fewer repetitive sequences than normal eccDNAs (Wilcoxon test, P-value < 0.05). The corresponding median values in the cancer group and the normal group were 0.349 and 0.676 respectively. Next, as shown in Fig. 5D, the GC content of cancer eccDNA sequences was significantly higher than that of normal sequences (Wilcoxon test, P-value < 0.05). The corresponding median values in the cancer and the normal group were 0.518 and 0.452 respectively. Finally, the genomic region composition analysis revealed that eccDNA molecules in cancer were enriched in the exonic region, 5′-untranslated regions (UTRs), 3′-UTRs, and promoter regions, but relatively less from the intronic regions, and intergenic regions (Fig. 5E, Table S1).

Built on the frequency statistics of eccDNA genes, we performed KEGG pathway enrichment analysis and comparison of the top 1000 and 500 genes in the cancer and normal groups, and a corresponding number of genes randomly selected from UniProt, respectively [51].

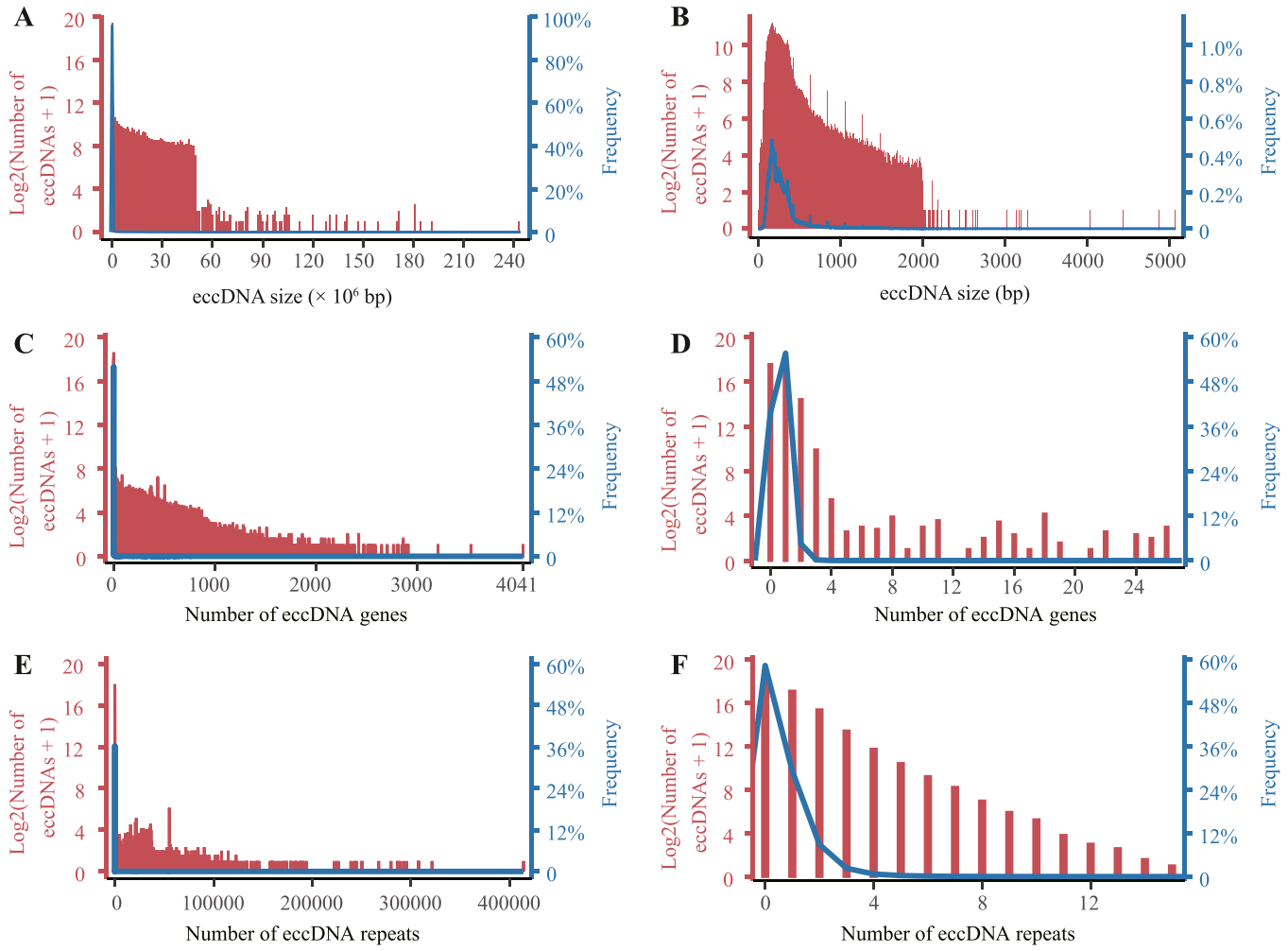

Fig. 3. Basic statistics of eccDNAs in eccBase. (A) Distribution of *hsa* eccDNAs by eccDNA size. (B) Distribution of mmu eccDNAs by eccDNA size. (C) Distribution of *hsa* eccDNAs by eccDNA gene. (D) Distribution of *mmu* eccDNAs by eccDNA gene. (E) Distribution of *hsa* eccDNAs by eccDNA repetitive elements. (F) Distribution of *mmu* eccDNAs by eccDNA repetitive elements.

Notably, genes from cancer eccDNAs were preferred to be enriched in cancer-specific pathways, although genes from normal eccDNAs were also enriched in one cancer-specific pathway (Fig. S4A-B). Identifying eccDNA markers that have a central regulatory role in malignancy progression is a challenging problem. As a case study, we prepared a machine learning model to identify cancer-associated eccDNAs. The model treats eccDNAs in gastric cancer as positive training samples and eccDNAs in the healthy human blood as negative samples because the two are the closest in number. Subsequently, the performance of six algorithms (XGB, RF, SVM, NNet, Bayes, and LR) commonly used in binary classification were evaluated, and the mean AUCs of 10-fold cross-validations were showed in Fig. S5A, XGB has achieved an outstanding score of 0.946, and the AUCs of other classifiers (RF, MLP, SVC, LR, and Bayes) are 0.942, 0.886, 0.891, 0.842, and 0.820, respectively. Feature ranking indicated that sequence length and genic region fractions seem more supportive of the predictive ability of XGB (Fig. S5B).

3.4. Disclosure of eccDNA tissue specificity

We explored differences in repeat sequences between eccDNAs from different *mmu* lineages and tissues. Fig. 6A showed the percentage of unique eccDNAs from each lineage type which covered by three major classes of repetitive elements, namely LTRs (Long Terminal Repeats), LINEs (Long Interspersed Nuclear Elements), and SINEs (Short Interspersed Nuclear Elements). Such percentages presented a clear disparity between 14 *mmu* lineages, especially for the LINEs, where unique eccDNAs in different lineages ranged from 5.28% (adult mouse kidney) to 15.16% (adult mouse sperm). As depicted in Fig. 6B-D, the proportion of eccDNAs covered by LTRs, LINEs, and SINEs displayed significant differences between lineages (K-W test, P-value < 0.001). In addition, the proportion of eccDNAs covered by LTR elements was all nearly consistent with the proportion of genomes covered by LTR elements (11.67%) (Fig. 6B). Although eccDNAs carrying LINE and SINE elements accounted for only 14% and 11% of the lineage, respectively, they showed differences opposed to the proportion of genomes containing these two classes (20.04% and 7.49%, respectively) (Fig. 6C-D). It is worth mentioning that the LINE elements with less content in eccDNAs can almost all belong to the L1 family, while the abundant SINE elements mainly belong to the Alu, B2 and B4 families.

4. Discussion

The eccDNAs are notorious for promoting cancer progress and contributing to drug resistance, and even are influential upon tissue development [6,20,52]. In this study, we constructed eccBase as the first database to date integrating all known eccDNAs from *hsa*, *mmu*, *gga* and *sce* (Fig. 1). Several recently published eccDNA databases, such as CircleBase and eccDNAdb, integrate subsets of eccDNA from

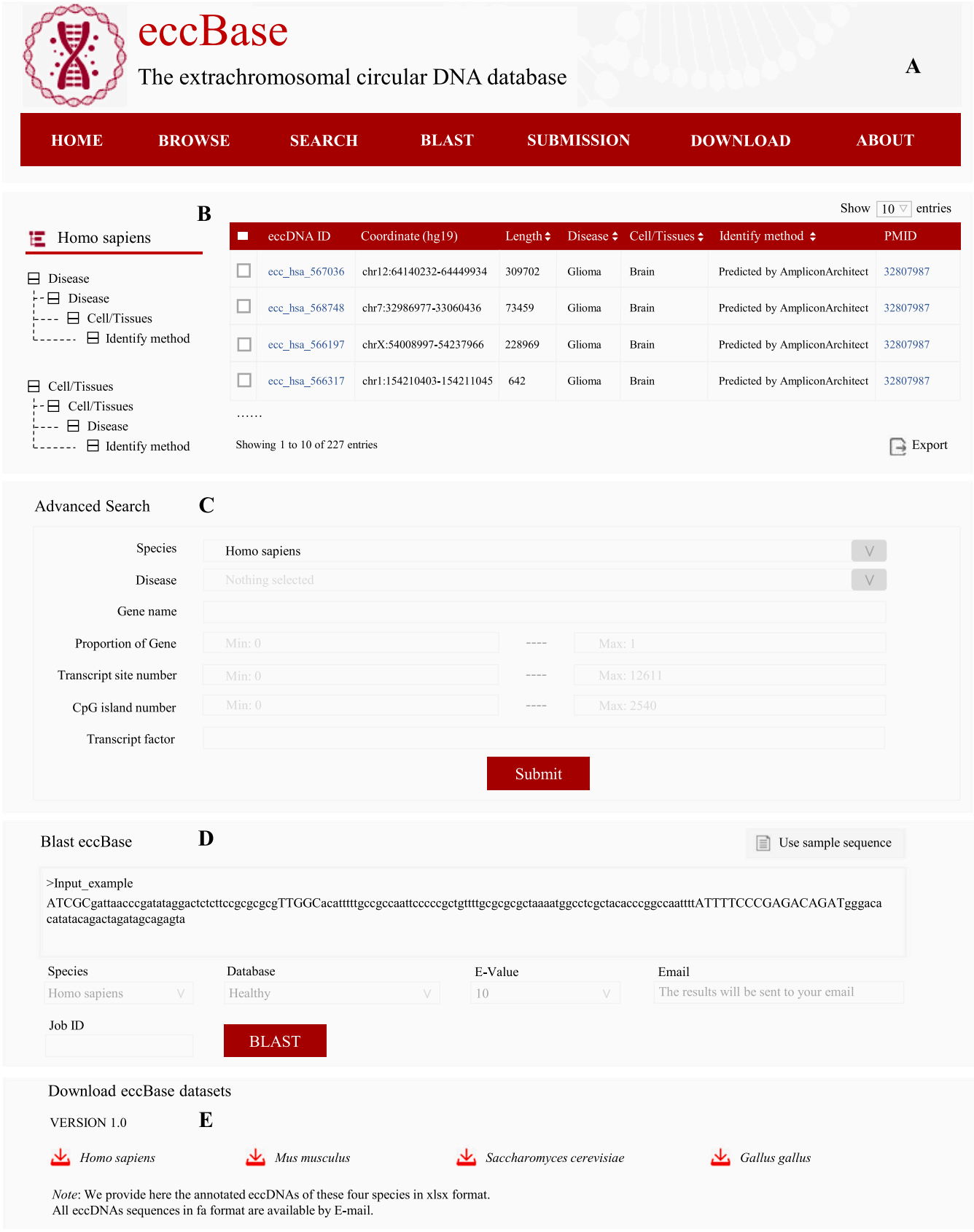

Fig. 4. The essential introduction to eccBase. (A) Navigation bar. (B) Browse module. (C) Advanced search module. (D) BLAST module. (E) Download module.

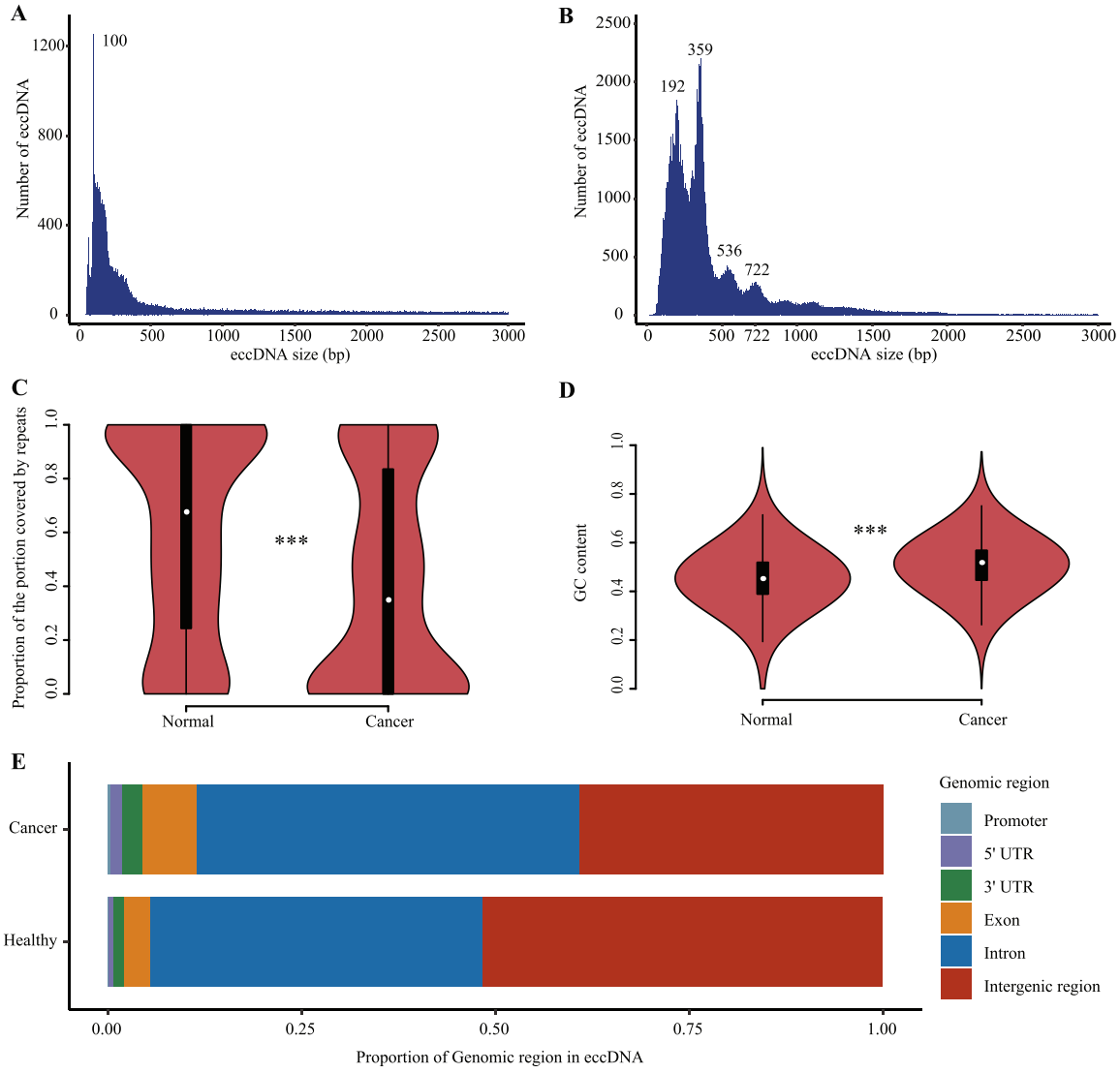

Fig. 5. Comparison analysis of *hsa* cancer and normal eccDNAs. (A) Distribution of *hsa* normal eccDNAs, which are less than 3000 bp, by eccDNA size. (B) Distribution of *hsa* cancer eccDNAs, which are less than 3000 bp, by eccDNA size. (C) Proportions comparison of the portion covered by repetitive elements. (D) GC content comparison. (E) Proportion comparison of the genomic regions constituted eccDNA. Note: the symbol of '***' represent the P-value between the groups was less than 0.001 (Wilcoxon test).

a retrospective perspective [53]. In contrast, eccBase performs better not only in terms of comprehensiveness and accessibility of data, but also in terms of richness of functional modules (Table S3). EccBase has the following advantages in facilitating eccDNA research.

First, the eccBase provides a knowledge atlas of eccDNAs in *hsa* cancers. The cancer eccDNAs were recorded from 50 lineages, consisting of 33 kinds of tissues and 17 kinds of cell lines, across 28 cancer types. And the eccDNAs from 5 kinds of healthy individual tissues were also collected for controlled studies (Fig. 2A-C). Second, eccBase supports discovering the tissue specificity of eccDNAs. For instance, eccDNAs in healthy mice are derived from 13 different tissues and have their own characteristics (Fig. 2D). Third, all eccDNAs were manually annotated and proofread, including basic features, nucleotide sequences, gene annotations, DNA elements, and source data (see Materials and methods for details), which will assist users in molecular characterization (Fig. S3). Compiled eccDNAs are available for download and local analysis (Fig. 4B-E). Next, Search and Browse modules facilitate the user to obtain eccDNA entries from the database according to the specified criteria. The search options are broad and include species, disease, tissue and molecular features. (Fig. 4C). The group browsing function organizes the corresponding eccDNA into a tree through nodes such as disease name, pedigree name, and identification method, so as to clearly present the content of each entry (Fig. 4B). Sequence similarity alignment is a preferred method for identifying potential novel eccDNAs. Therefore, we deployed the BLAST service in eccBase to assist user-defined sequences and parameters (Fig. 4D). Certainly, alignment results from BLAST will contribute to guiding a successful biological experiment. Finally, more species and diseases are considered to update the underlying data and release a new version in the future. To further expand the research on eccDNAs of model species, we also collected the eccDNAs of *gga* (n = 700,087) and *sce* (2010) and annotated them briefly, and also opened the download. (Fig. 4E).

Molecular features comparison showed divergence between cancer and normal eccDNAs. First, cancer eccDNAs with ultra-long sizes are indicative because they may carry oncogenes, but yet know little about tiny cancer eccDNAs [10,14]. Fig. 5A-B depicted that under the size < 3000 bp, cancer eccDNAs exhibited a unique model that the size profiles are multimodal distribution which peaks and peaks with ~177 bp intervals. It suggests that cancer eccDNAs originate from nucleosomes, i.e., the nucleosome cores plus part of the linker DNA, and might speculate that the breakpoints of cancer eccDNAs tend to locate in the linkers. Yuangao Wang et.al.

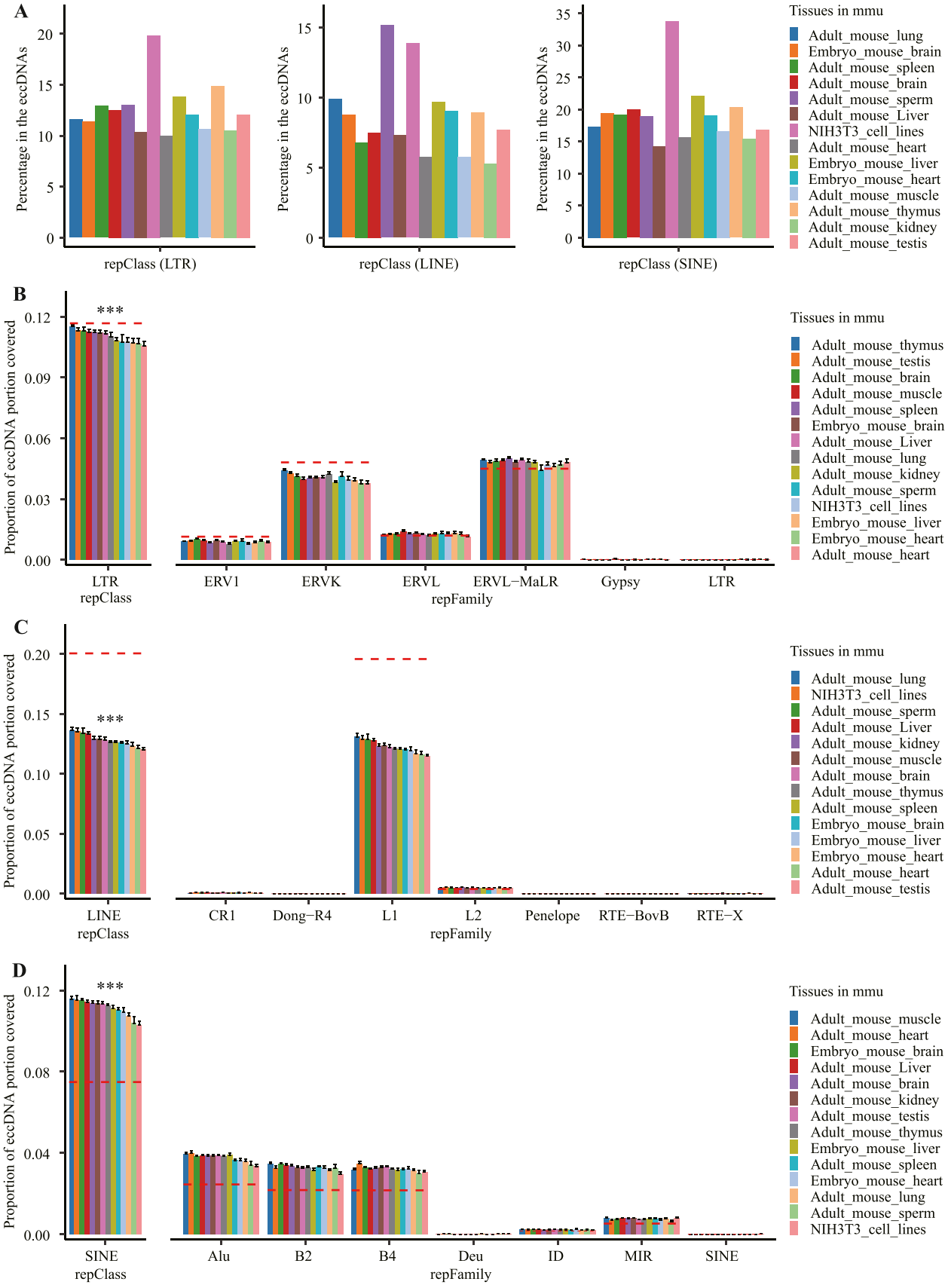

Fig. 6. Statistics of *mmu* eccDNA repetitive elements. To eccDNAs which harbor repetitive elements of LTR, LINE, and SINE, separately, percentage in different tissues and/or cell lines. (B) To eccDNAs which are covered by the LTRs, as well as which are covered by their subfamilies separately, proportions of the covered portion in different tissues and/or cell lines. (C) To eccDNAs which are covered by the LINEs, as well as which are covered by their subfamilies separately, proportions of the covered portion in different tissues and/or cell lines. (D) To eccDNAs which are covered by the SINEs, as well as which are covered by their subfamilies separately, proportions of the covered portion in different tissues and/or cell lines. Note: the symbol of '***' means the P-value between the groups was less than 0.001 (K-W test). The red dot lines represent the proportions of the portion covered by the corresponding repetitive elements in the reference genome.

demonstrated that apoptosis-inducing agents can increase eccDNA production, i.e. DNase γ triggers apoptotic DNA fragmentation followed by circularization/ligation by Lig3 [7]. This indicated the formation mechanism of cancer eccDNAs. Furthermore, cancer eccDNAs clearly tended to be derived from promoter, UTR, exon dense, and GC-rich regions, rather than from the intergenic region (Fig. 5C-E; Table S1). It represents the production preference of cancer eccDNAs and infers cancer eccDNA functioning by transcribing carried genes directly rather than as repetitive elements [54]. Interestingly, top genes occurring on eccDNAs in the cancer group showed preference of enriching in cancer-related KEGG pathways, reflecting the role of eccDNAs as pathogenic and deteriorated contributors in cancer (Fig. S4) [55,56]. Note that the enrichment of specific cancer-related KEGG pathways, such as Lung cancer and Melanoma, may be attributed to the relative abundance of eccDNAs in specific cancers, but this does not contradict the above inferences. Taken together, sequence features and genomic origin of cancer eccDNAs indicated their formation mechanisms and production tendency, and specific roles in supporting the tenacious survival of cancer cells. Top genes occurring in healthy-related eccDNAs also enriched in one cancer-related pathway, it highlighted the complexity of eccDNA, suggested the importance of distinguishing vital eccDNA via comprehensive information.

As a case study, we attempted to distinguish vital eccDNAs of gastric cancer via machine learning. Fig. S5A, suggested XGBoost as a suitable algorithm for identifying key eccDNA candidates in cancer. XGBoost is an ensemble machine learning method and has been declared to have excellent performance in biological classification problems [48,57]. Feature ranking suggested that the eccDNA size, the portion covered by genes in eccDNA contributed the most to the predictive power, also implying a significant distinction between cancer against normal eccDNAs (Fig. S5B).

As shown in Table S4, the cancer eccDNAs were mainly identified by the methods of 'Island method and split read method', 'ATAC-seq', while the normal eccDNAs were mainly identified by 'Circle-Seq' [4,28,58]. Actually, 'Island method and split read method' and 'Circle-Seq' are similar approaches, both are purified by digestion of linear DNA and/or RNA, mtDNA, and then perform NGS using the products of multiple displacement amplification or rolling circle amplification. The circles identified therein depend on directional inward junctional paired reads that may represent breakage/ligation points. 'ATAC-seq' exploits the properties of open chromatin structure of eccDNAs, that is, detects eccDNAs from ATAC-seq data, but adopts the same algorithm (circle_finder) as 'Island method and split read method', and the results usually are validated by inverse PCR and metaphase FISH [59]. Therefore, the results of our analysis are affected by different identification methods, but the main conclusions are testable. In the future, more data will be needed in order to strengthen the robustness of the conclusions, but here we highlight the significance of such a strategy for developing effective prediction tools and then applied in the clinical diagnosis and prognosis evaluation. We also systematically summarized the eccDNA identify methods (Table S5). As mentioned above, the eccDNAs number showed distributional difference across the methods. Then, eccDNAs from 'Circle-seq' achieved a relative balance between cancer and healthy group, that is conducive to more precisely disclose cancer eccDNAs. Then, most eccDNAs are function as changing gene/oncogene expression pattern, and leading to genomic instability and intracellular heterogeneity. But, eccDNA functions are variety also that mean eccBase could be service to users have varied objectives.

We further deciphered the eccDNAs of *mmu* are tissue specific. Repetitive elements are major components of eukaryotic genomes, which not are junk DNA but have structural and regulatory functions on genomic function and gene expression [60,61]. Hence interrogating eccDNA repetitive elements across tissues can help reveal their concrete biological functions and even contribute to innovative explanations of embryonic development and tissue differentiation. As Fig. 6 delineated, both the percentage of eccDNAs with repetitive elements and the proportions covered by repetitive elements is varied evidently in tissues. In addition, Xiaohua Shen et.al. reported the transposons embedded in mammalian genomes, with L1 elements preferentially enriched in the heterochromatic region, while Alu/B1 elements in gene-dense euchromatic regions [62]. Combined with preferences for eccDNAs from LINE and SINE and their subfamilies, it indicated that chromatin accessibility may be positively associate with the formation of eccDNAs, i.e. exposed regions may be vulnerable and mutable (Fig. 6C-D). It enlightened the importance of investigating genes believed to increase chromatin accessibility, which further may boost eccDNA yields.

Currently, we do not use natural language processing models or web semantic search models to assist in data collection. With the increase of eccDNA-related literature and web texts, the efficiency of manual review will not meet the needs of data collection and analysis. We are considering using web crawlers to crawl more data and building natural language processing models based on transformer and bert to automatically parse the relationship between eccDNA, genes, pathways and diseases in the text. Based on this, we will build various ontologies related to eccDNA, and build a more comprehensive eccDNA knowledge graph through ontology extraction and association, which will help users to obtain various types of knowledge related to eccDNAs beyond sequence features.

5. Conclusion

The eccBase we developed provides peers with the most comprehensive extrachromosomal DNA resources available today. It annotated 50 features for each molecular sequence and has an easy-to-operate user interface, which not only facilitates user browsing and retrieval, but also supports data download for localized analysis. With the help of the integrated BLAST, users can discover novel homologous eccDNAs and quickly map their genomic locations. We analyzed the characteristics of *hsa* cancer eccDNA that distinguish it from normal eccDNA, and speculated that cancer eccDNA molecules may be composed of nucleosomes and originate from gene-dense regions. Furthermore, we established the potential of machine learning methods to screen cancer eccDNAs by the example of predicting key eccDNAs in gastric cancer. We also performed statistical analysis on *mmu* eccDNA sequences to reveal their tissue specificity and propensity of origin for open chromatin regions. In conclusion, eccBase will advance the exploration of eccDNA in cancer development, tissue differentiation and clinical applications, and will continue to be maintained and updated.

CRediT authorship contribution statement

Haiyang Sun: Investigation, Resources, Formal analysis, Writing – original draft. **Xinyi Lu:** Formal analysis. **Lingyun Zou:** Conceptualization, Writing – original draft, Supervision, Funding acquisition. All authors read and agreed to the final manuscript.

Declaration of Competing Interest

The authors declare that there is not any conflict of interest regarding the publication of this article.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Nos. 31771468) and Fundamental Research Foundation of Shenzhen (Nos. JCYJ20190809, 182411369).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.04.012.

References

[1] Cox D, Yuncken C, Spriggs AI. Minute chromatin bodies in malignant tumours of childhood. Lancet 1965;1(7402):55–8.
[2] LuBs HA, SALMON JH. The chromosomal complement of human solid tumors. J Neurosurg 1965;22(2):160–8.
[3] Moller HD, Parsons L, Jorgensen TS, Botstein D, Regenberg B. Extrachromosomal circular DNA is common in yeast. Proc Natl Acad Sci USA 2015;112(24):E3114–22.
[4] Shibata Y, Kumar P, Layer R, Willcox S, Gagan JR, Griffith JD, et al. Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. Science 2012;336(6077):82–6.
[5] Yan Y, Guo G, Huang J, Gao M, Zhu Q, Zeng S, et al. Current understanding of extrachromosomal circular DNA in cancer pathogenesis and therapeutic resistance. J Hematol Oncol 2020;13(1):124.
[6] Ashique S, Upadhyay A, Garg A, Mishra N, Hussain A, Negi P, et al. Impact of ecDNA: A mechanism that directs tumorigenesis in cancer drug Resistance-A review. Chem Biol Inter 2022;363:110000.
[7] Wang Y, Wang M, Djekidel MN, Chen H, Liu D, Alt FW, et al. eccDNAs are apoptotic products with high innate immunostimulatory activity. Nature 2021;599(7884):308–14.
[8] Ott CJ. Circles with a point: new insights into oncogenic extrachromosomal DNA. Cancer Cell 2020;37(2):145–6.
[9] Moller HD, Mohiyuddin M, Prada-Luengo I, Sailani MR, Halling JF, Plomgaard P, et al. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. Nat Commun 2018;9(1):1069.
[10] Wu S, Turner KM, Nguyen N, Raviram R, Erb M, Santini J, et al. Circular ecDNA promotes accessible chromatin and high oncogene expression. Nature 2019;575(7784):699–703.
[11] Morton AR, Dogan-Artun N, Faber ZJ, MacLeod G, Bartels CF, Piazza MS, et al. Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. Cell 2019;179(6):1330–41. e13.
[12] Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature 2017;543(7643):122–5.
[13] Lundberg G, Rosengren AH, Hakanson U, Stewenius H, Jin Y, Stewenius Y, et al. Binomial mitotic segregation of MYCN-carrying double minutes in neuroblastoma illustrates the role of randomness in oncogene amplification. PLoS One 2008;3(8):e3099.
[14] Cao X, Wang S, Ge L, Zhang W, Huang J, Sun W. Extrachromosomal Circular DNA: Category, Biogenesis, Recognition, and Functions. Front Vet Sci 2021;8:693641.
[15] Zhao XK, Xing P, Song X, Zhao M, Zhao L, Dang Y, et al. Focal amplifications are associated with chromothripsis events and diverse prognoses in gastric cardia adenocarcinoma. Nat Commun 2021;12(1):6489.
[16] Alt FW, Kellems RE, Bertino JR, Schimke RT. Selective multiplication of dihydrofolate reductase genes in methotrexate-resistant variants of cultured murine cells. J Biol Chem 1978;253(5):1357–70.
[17] Shoshani O, Brunner SF, Yaeger R, Ly P, Nechemia-Arbely Y, Kim DH, et al. Chromothripsis drives the evolution of gene amplification in cancer. Nature 2021;591(7848):137–41.
[18] Li R, Wang Y, Li J, Zhou X. Extrachromosomal circular DNA (eccDNA): an emerging star in cancer. Biomark Res 2022;10(1):53.
[19] Paulsen T, Shibata Y, Kumar P, Dillon L, Dutta A. Small extrachromosomal circular DNAs, microDNA, produce short regulatory RNAs that suppress gene expression independent of canonical promoters. Nucleic Acids Res 2019;47(9):4586–96.
[20] Dillon LW, Kumar P, Shibata Y, Wang YH, Willcox S, Griffith JD, et al. Production of extrachromosomal MicroDNAs is linked to mismatch repair pathways and transcriptional activity. Cell Rep 2015;11(11):1749–59.
[21] Peng L, Zhou N, Zhang CY, Li GC, Yuan XQ. eccDNAdb: a database of extrachromosomal circular DNA profiles in human cancers. Oncogene 2022;41(19):2696–705.
[22] Cen Y, Fang Y, Ren Y, Hong S, Lu W, Xu J. Global characterization of extrachromosomal circular DNAs in advanced high grade serous ovarian cancer. Cell Death Dis 2022;13(4):342.
[23] deCarvalho AC, Kim H, Poisson LM, Winn ME, Mueller C, Cherba D, et al. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. Nat Genet 2018;50(5):708–17.
[24] Gibaud A, Vogt N, Hadj-Hamou NS, Meyniel JP, Hupe P, Debatisse M, et al. Extrachromosomal amplification mechanisms in a glioma with amplified sequences from multiple chromosome loci. Hum Mol Genet 2010;19(7):1276–85.
[25] Henriksen RA, Jenjaroenpun P, Sjostrom IB, Jensen KR, Prada-Luengo I, Wongsurawat T, et al. Circular DNA in the human germline and its association with recombination. Mol Cell 2022;82(1):209–17. e7.
[26] Hung KL, Yost KE, Xie L, Shi Q, Helmsauer K, Luebeck J, et al. ecDNA hubs drive cooperative intermolecular oncogene expression. Nature 2021;600(7890):731–6.
[27] Kim H, Nguyen NP, Turner K, Wu S, Gujar AD, Luebeck J, et al. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nat Genet 2020;52(9):891–7.
[28] Kumar P, Kiran S, Saha S, Su Z, Paulsen T, Chatrath A, et al. ATAC-seq identifies thousands of extrachromosomal circular DNA in cancer and cell lines. Sci Adv 2020;6(20):eaba2489.
[29] Lin C, Chen Y, Zhang F, Liu B, Xie C, Song Y. Encoding gene RAB3B exists in linear chromosomal and circular extrachromosomal DNA and contributes to cisplatin resistance of hypopharyngeal squamous cell carcinoma via inducing autophagy. Cell Death Dis 2022;13(2):171.
[30] Purshouse K, Friman ET, Boyle S, Dewari PS, Grant V, Hamdan A, et al. Oncogene expression from extrachromosomal DNA is driven by copy number amplification and does not require spatial clustering in glioblastoma stem cells. Elife 2022:11.
[31] Schulz C, Gomez Perdiguero E, Chorro L, Szabo-Rogers H, Cagnard N, Kierdorf K, et al. A lineage of myeloid cells independent of Myb and hematopoietic stem cells. Science 2012;336(6077):86–90.
[32] Song K, Minami JK, Huang A, Dehkordi SR, Lomeli SH, Luebeck J, et al. Plasticity of extrachromosomal and intrachromosomal BRAF amplifications in overcoming targeted therapy dosage challenges. Cancer Disco 2022;12(4):1046–69.
[33] Sun Z, Ji N, Zhao R, Liang J, Jiang J, Tian H. Extrachromosomal circular DNAs are common and functional in esophageal squamous cell carcinoma. Ann Transl Med 2021;9(18):1464.
[34] Xu G, Shi W, Ling L, Li C, Shao F, Chen J, et al. Differential expression and analysis of extrachromosomal circular DNAs as serum biomarkers in lung adenocarcinoma. J Clin Lab Anal 2022;36(6):e24425.
[35] Yang H, He J, Huang S, Yang H, Yi Q, Tao Y, et al. Identification and Characterization of Extrachromosomal Circular DNA in Human Placentas With Fetal Growth Restriction. Front Immunol 2021;12:780779.
[36] Zeng T, Huang W, Cui L, Zhu P, Lin Q, Zhang W, et al. The landscape of extrachromosomal circular DNA (eccDNA) in the normal hematopoiesis and leukemia evolution. Cell Death Disco 2022;8(1):400.
[37] Zhu Y, Gujar AD, Wong CH, Tjong H, Ngan CY, Gong L, et al. Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. Cancer Cell 2021;39(5):694–707. e7.
[38] Moller HD, Bojsen RK, Tachibana C, Parsons L, Botstein D, Regenberg B. Genome-wide Purification of Extrachromosomal Circular DNA from Eukaryotic Cells. J Vis Exp 2016;110:e54239.
[39] Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 2019;47(D1):D853–8.
[40] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078–9.
[41] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–2.
[42] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47(D1):D766–73.
[43] Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics 2012;28(14):1919–20.
[44] Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res 2020;48(D1):D882–9.
[45] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinforma 2009;10:421.
[46] Zhang S, Zhao L, Zheng CH, Xia J. A feature-based approach to predict hot spots in protein-DNA binding interfaces. Brief Bioinform 2020;21(3):1038–46.
[47] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019;47(D1):D886–94.
[48] Chen X, Huang L, Xie D, Zhao Q. EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. Cell Death Dis 2018;9(1):3.
[49] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(85):2825–30.
[50] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al. Scikit-learn: Machine Learning in Python. 2011; abs/1201.0490.
[51] UniProt C. UniProt: a hub for protein information. Nucleic Acids Res 2015;43(Database issue). D204-12.
[52] Wu S, Bafna V, Mischel PS. Extrachromosomal DNA (ecDNA) in cancer pathogenesis. Curr Opin Genet Dev 2021;66:78–82.
[53] Zhao X, Shi L, Ruan S, Bi W, Chen Y, Chen L, et al. CircleBase: an integrated resource and analysis platform for human eccDNAs. Nucleic Acids Res 2022;50(D1):D72–82.
[54] Zhou Y, Bizzaro JW, Marx KA. Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+C)% composition. BMC Genom 2004;5:95.
[55] Hong J, Zheng S, Jiang D. The contributions of extrachromosomal DNA elements in neoplasm progression. Am J Cancer Res 2021;11(6):2417–29.
[56] Koche RP, Rodriguez-Fos E, Helmsauer K, Burkert M, MacArthur IC, Maag J, et al. Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. Nat Genet 2020;52(1):29–34.
[57] Tianqi C. (2016) XGBoost: A Scalable Tree Boosting System. In: Krishnapuram B, Shah M, editors. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM. pp. 785–94.

[58] Deshpande V, Luebeck J, Nguyen ND, Bakhtiari M, Turner KM, Schwab R, et al. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nat Commun 2019;10(1):392.
[59] Kumar P, Dillon LW, Shibata Y, Jazaeri AA, Jones DR, Dutta A. Normal and Cancerous Tissues Release Extrachromosomal Circular DNA (eccDNA) into the Circulation. Mol Cancer Res 2017;15(9):1197–205.
[60] Billingsley KJ, Lattekivi F, Planken A, Reimann E, Kurvits L, Kadastik-Eerme L, et al. Analysis of repetitive element expression in the blood and skin of patients with Parkinson's disease identifies differential expression of satellite elements. Sci Rep 2019;9(1):4369.
[61] Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol 2018;19(1):199.
[62] Lu JY, Chang L, Li T, Wang T, Yin Y, Zhan G, et al. Homotypic clustering of L1 and B1/Alu repeats compartmentalizes the 3D genome. Cell Res 2021;31(6):613–30.