Genome Medicine

RESEARCH Open Access

Detection of primary cancer types via fragment size selection in circulating cell-free extrachromosomal circular DNA

Jingwen Fang[1,2†], Songwen Luo[1†], Shouzhen Li[1†], Yehong Xu[3†], Jing Wang[4†], Benjie Shan[1†], Mingjun Hu[5†], Qiaoni Yu[1,6], Wen Zhang[1,7], Ke Liu[1], Yunying Shao[2], Jiaxuan Yang[2], YouYang Zhou[2], Guangtao Xu[1], Xinfeng Yao[1], Ruoming Sun[1], Mengyuan Zhang[4], Kun Li[8], Xihai Xu[8], Yongliang Zhang[5*], Zhihong Zhang[3*], Xinghua Han[1*], Yueyin Pan[1*], Chuang Guo[9,10*] and Kun Qu[1,6,7*]

Abstract

Background Extrachromosomal circular DNA has emerged as a pivotal factor in tumor biology, contributing to intratumor heterogeneity, oncogene amplification, and tumor evolution. Despite its relevance, the presence and molecular characteristics of plasma-derived cell-free extrachromosomal circular DNA (eccDNA) in cancer patients remain insufficiently explored.

Methods In this study, we profiled plasma-derived cell-free eccDNA from a multi-cancer cohort consisting of 413 cancer patients and 239 healthy individuals. We analyzed eccDNA fragment size distributions using a patient-derived xenograft (PDX) mouse model and identified fragment size features to distinguish tumor-derived eccDNA from non-tumor-derived eccDNA. We further performed in silico fragment size selection and developed a gene-based annotation method to characterize the gene content carried by cell-free eccDNA across different cancer types. By utilizing these cell-free eccDNA signatures, we developed ScanTecc (screening cancer types with cell-free eccDNA), a machine learning-based approach for cancer detection and tissue-of-origin classification. The classification

Background

Circulating tumor DNA (ctDNA) has attracted widespread attention in clinical oncology as a minimally invasive liquid biopsy biomarker for detecting target mutations and monitoring cancer recurrence or persistence [1–3]. However, most ctDNA clinical research focuses on genomic variations [4], epigenomic characteristics [5] or fragmentation patterns [6], limiting its ability to detect early cancer due to dilution by much larger quantities of DNA of noncancerous origins. To address this issue, researchers have attempted to improve ctDNA detection by increasing sequencing depth or combining other blood biomarkers, such as histone modifications, proteomics, or metabolomics [7, 8]. However, these

[†] Jingwen Fang, Songwen Luo, Shouzhen Li, Yehong Xu, Jing Wang, Benjie Shan and Mingjun Hu contributed equally to this work.

*Correspondence:
Yongliang Zhang
zyl2020@ustc.edu.cn
Zhihong Zhang
zhangzhihope@126.com
Xinghua Han
hxhmail@ustc.edu.cn
Yueyin Pan
panyueyin@ustc.edu.cn
Chuang Guo
gchuang@ustc.edu.cn
Kun Qu
qukun@ustc.edu.cn
Full list of author information is available at the end of the article

performance of ScanTecc was further evaluated at the individual sample level using multiple machine learning classifiers, including adaptive boosting and logistic regression.

Results We observed a significantly higher abundance and longer fragment lengths of eccDNA in cancer patients' plasma. Analysis of the PDX mouse model revealed a distinct fragment size threshold of approximately 1,000 base pairs that effectively differentiates tumor-derived eccDNA from non-tumor-derived counterparts. Following fragment size selection, gene-based annotation of large-sized eccDNA revealed cancer type–specific enrichment of tumor-associated genes. Leveraging these features, ScanTecc achieved an overall AUC of 0.92 for distinguishing cancer patients from healthy individuals, with consistently high performance across disease stages, including stage I (AUC = 0.92) and stage IV (AUC = 0.93). ScanTecc also enabled accurate tissue-of-origin classification and achieved an overall AUC of 0.79 in identifying specific cancer types, with AUC values ranging from 0.70 for gastric cancer to 0.81 for ovarian cancer.

Conclusions Our study establishes a multi-cancer plasma cell-free eccDNA landscape and introduces a non-invasive cancer screening framework based on cell-free eccDNA features, highlighting the potential of plasma cell-free eccDNA for early cancer detection and tumor classification.

Keywords Extrachromosomal circular DNA, Plasma-derived cell-free EccDNA, Machine learning-based approach; early cancer detection

approaches are still constrained by cost and clinical feasibility, hindering the development, evaluation, and implementation of early cancer detection [4]. These limitations highlight the need to explore novel biomarkers capable of enhancing early detection and offering a more comprehensive genomic view of tumors.

Extrachromosomal circular DNA is a double-stranded circular DNA molecule that exists independently of conventional chromosomes [9]. With the rapid development of high-throughput sequencing and super-resolution imaging technologies, extrachromosomal circular DNA has been widely found in both normal tissues and various tumor tissues, garnering significant attention in biological and clinical research [10–12]. Depending on its fragment size and biological characteristics, extrachromosomal circular DNA can be classified into two categories: (1) large, oncogene-enriched, copy number-amplified extrachromosomal circular DNA (ecDNA), and (2) the predominant small extrachromosomal circular DNA (eccDNA). Studies have shown that ecDNA amplification, prevalent in various human cancers and precancerous lesions, is linked to accelerated cancer evolution and poor prognosis due to the overexpression of amplified oncogenes [13, 14]. Additionally, ecDNA exhibits high mobility, clustered mutations, and non-homogeneous segregation, functioning as "mobile enhancers" that accelerate tumor evolution [15–17]. In contrast, eccDNA primarily originates from apoptotic or replication error processes [18], exhibiting immune-stimulatory activity via its circular structure and cytosolic DNA sensors [12]. Notably, the abundance of both ecDNA and eccDNA in cancerous tissues is significantly higher than in normal tissues [11, 19], underscoring their potential as diagnostic and therapeutic biomarkers.

Studies have further reported that cell-free eccDNA, referring to eccDNA released into the circulation, can be stably present in various body fluids, such as plasma [20] and urine [21]. This type of DNA retains its circular structure, and both long-read and short-read sequencing have revealed that the size and number of cell-free eccDNA in the plasma of cancer patients are significantly greater than those in healthy individuals [22, 23]. These differences are especially pronounced in presurgery plasma samples, which are enriched in larger eccDNA fragments, further demonstrating the potential of eccDNA in cancer monitoring [24]. Although cell-free DNA (cfDNA), including ctDNA, is widely used in oncology for tumor monitoring [25], ctDNA only accounts for a small proportion (0.1–1.0%) of total cfDNA and is difficult to isolate [26]. Additionally, compared with cfDNA, cell-free eccDNA exhibits higher stability and lower contamination from white blood cells, providing distinct advantages for liquid biopsy diagnostics. However, comprehensive characterization of cell-free eccDNA in cancer patients remains limited, which significantly hinders its widespread clinical application.

In this study, we analyzed the cell-free eccDNA profiles of 652 plasma samples, encompassing 413 patients diagnosed with eighteen distinct primary cancer types and 239 healthy individuals. We verified the presence of cell-free eccDNA in human plasma using outward PCR, Sanger sequencing, and atomic force microscopy (AFM). Leveraging plasma samples from patient-derived xenograft mouse model, we identified a clear fragment size threshold near 1,000 base pairs (bp) that effectively distinguishes tumor-derived eccDNA from non-tumor counterparts, both in terms of size distribution and genomic localization. We further developed ScanTecc, an innovative machine learning framework that combines eccDNA-specific feature selection with a random forest classifier to enable robust cancer detection and tissue-of-origin prediction. This model achieved an accuracy of 0.92 for distinguishing cancer patients from healthy individuals and an overall prediction rate of 0.79 for identifying specific cancer types based on eccDNA-associated gene features. Collectively, our study establishes the presence and diagnostic relevance of cell-free eccDNA in peripheral blood, underscoring its potential as a minimally invasive biomarker for early cancer detection and precision oncology application.

Methods

Cohort inclusion and exclusion criteria

A total of 413 cancer patients and 239 healthy individuals were enrolled in this study. All participants were recruited from The First Affiliated Hospital of the University of Science and Technology of China and Anhui Provincial Cancer Hospital. Peripheral blood samples were collected from all participants and plasma was isolated for subsequent eccDNA sequencing. All collected plasma samples were used for cell-free eccDNA profiling as well as for cancer screening model training and testing. Participants meeting the following criteria were considered eligible for enrollment in the cancer group: (1) individuals aged between 18 and 75 years old, all genders eligible; (2) individuals with a clinically confirmed diagnosis of cancer or identified as having a solid mass by imaging examinations (computed tomography (CT), magnetic resonance imaging (MRI), or endoscopic ultrasonography (EUS)), with no prior treatment; (3) individuals whose lesion diameter was greater than 1 cm; (4) individuals who provided informed consent. Participants meeting the following criteria were considered eligible for enrollment in the healthy group: (1) individuals aged between 18 and 75 years old, all genders eligible; (2) individuals who showed no evidence of tumor or suspected tumor-related diseases based on imaging examinations (CT, MRI, or EUS) and laboratory tests; (3) individuals who provided informed consent. Participants meeting any of the following criteria were excluded from the cancer group and the healthy group: (1) individuals with severe cardiopulmonary diseases or poor general physical condition; (2) individuals with contraindications to blood collection; (3) individuals who had active infection or experienced persistent fever within 14 days prior to screening; (4) individuals who were unable or unwilling to provide informed consent; (5) individuals with any other condition that, in the opinion of the investigator, made them unsuitable for inclusion in the study. The detailed criteria for cohort inclusion and exclusion are additionally provided in Additional file 1, Figure S1. Detailed clinical information for the cohort is described in Additional file 2, Table S1.

In addition to the primary cohort, an independent validation cohort consisting of 21 lung cancer patients and 14 healthy individuals (Additional file 2: Table S2) was included to evaluate the generalizability of ScanTecc. These participants were enrolled using the same inclusion and exclusion criteria, sample collection procedures, and plasma processing protocols as those applied to the primary cohort. Samples from this independent cohort were used exclusively for external validation analyses and were not included in model training.

Furthermore, to compare the performance of ScanTecc with conventional tumor markers, we analyzed CEA and CA19-9 data from a previously published gastric cancer cohort [27]. This external cohort was used solely for comparative performance evaluation and was not involved in model training or validation.

Human samples

Peripheral blood samples were collected from cancer patients and healthy individuals at the First Affiliated Hospital of University of Science and Technology of China and the First Affiliated Hospital of Anhui Medical University. All patients were treatment-naive and had no serious complications. Whole blood was collected in EDTA tubes and processed either immediately or within one day after storage at 4 °C. Informed consent was obtained from the patients. Study procedures were followed in accordance with protocols approved by the ethics committee of the University of Science and Technology of China.

Sample processing

Peripheral blood samples were collected and centrifuged at 1,600 g for 10 min at 4 °C. The plasma portion was further centrifuged at 16,000 g for 10 min at 4 °C to remove residual cells and debris. Plasma samples were then stored at – 80 °C for subsequent DNA extraction. Plasma DNA extractions were performed using the QIAamp Circulating Nucleic Acid Kit (Qiagen, 55114). Quality control of extracted cfDNA was assessed with the Bioanalyzer 2100 (Agilent Technologies).

EccDNA library Preparation and sequencing

For elimination of linear DNA and enrichment of eccDNA, 25 ng of plasma DNA was treated with 5 units of exonuclease V (New England Biolabs) in a 50-μL reaction system at 37 °C for 30 min, followed by column purification using MinElute Reaction Cleanup Kit (Qiagen, 28206) [20]. Exonuclease digestion and column purification were then performed twice to further eliminate linear DNA. The exonuclease V was then heat-inactivated by incubating at 70 °C for 30 min. Illumina sequencing libraries for eccDNA were prepared by Tn5-transposon-based tagmentation using the TruePrep DNA Library

Prep Kit V2 (Vazyme, TD501) according to the manufacturer's instructions. DNA libraries were sequenced with the Illumina NovaSeq6000 platform in 2 × 150-bp paired-end reads.

Atomic-force microscopy (AFM) imaging

AFM imaging of DNA was performed in dry mode. Briefly, $\frac{1}{10}$ volume of 10× imaging buffer (100 mM NiCl2 and 100 mM Tris-HCl, pH 8.0) was added to the sample with a final DNA concentration of 0.6–1.0 ng/μl. 5 μl of the mixture was then spread on a freshly cleaved mica (Ted Pella) surface. After a 2-minute incubation, the specimen was rinsed twice with 200 μL of 2 mM magnesium acetate, and dried with compressed air before and after rinsing. Images were acquired using ScanAsyst-Air probe on a Dimension icon MultiMode V atomic-force microscope in 'ScanAsyst-Air mode' and were processed with NanoScope Analysis 2.0.

Polymerase chain reaction (PCR) for validation

Rolling Circle Amplification (RCA) was performed before PCR validation. Purified eccDNAs were added to 5 μL of 100 μM random hexamer primers. The samples were denatured at 95 °C for 5 min, followed by annealing at 50 °C for 15 s, 30 °C for 15 s and 20 °C for 10 min, and then held on ice for 5 min. A reaction mix was added so that the final concentrations were 1 × phi29 DNA polymerase reaction buffer, 0.2 mg/mL BSA, 2 mM dNTP and 5 U of phi29 DNA polymerase (NEB). RCA was performed at 30 °C for 24 h. The products were incubated with 10 U of T7 endonuclease I (NEB) at 37 °C for 30 min. Using the post-RCA eccDNA as a template, outward divergent primer sets (Additional file 2: Table S3) were designed to detect the target small eccDNAs by PCR. PCR products were loaded on the 2% agarose gel for electrophoresis. The base composition of PCR products was confirmed by Sanger sequencing.

Data preprocessing

Raw reads were aligned to the human reference assembly hg19 by Burrows-Wheeler Aligner MEM v.0.7.12 [28] with default parameters. The bam file was sorted using samtools v.0.1.19 [29]. The samtools flagstat tool was used to generate the statistical comparison results of bam files. PCR and optical duplicates were marked using Sambamba markdup v.0.6.6 [30] with default parameters. An in-house Python script was used to separate alignments into split reads, discordant and concordant reads. Candidate eccDNAs were first identified based on split reads (high-confidence ones). If the total length of two sub-alignments of split reads exceeded the read length, homologous sequences were searched. When homologous sequences were found, we recorded the coordinates of the leftmost form of eccDNA and an offset corresponding to the length of homologous sequences to represent all possible eccDNA variants. Potential split reads that failed to be mapped as split reads in the first place (low-confidence ones) as well as discordant reads were identified and counted using in-house Python scripts. The average coverages (in terms of RPK) for candidate eccDNAs and surrounding regions were then calculated based on all different types of reads. Any eccDNA supported by at least two high-confidence split reads or discordant reads, with its 95% region covered by at least one read, and with an average coverage at least twice that of its surrounding region, was considered a high-confidence eccDNA. For each sample, normalized eccDNA counts were measured in units of eccDNA per million mappable reads (EPM).

In Silico size selection

Paired-end sequencing was performed to generate reads from both ends of the eccDNA fragments present in the library. The start and end coordinates of cell-free eccDNA were identified using our optimized bioinformatic pipeline. Cell-free eccDNA longer than 1,000 bp was selected based on the inferred original fragment size, which was calculated from the mapping positions of the paired-end reads. These size-selected cell-free eccDNA fragments were subsequently used as input for our prediction model.

Genomic feature analysis

Genetic element annotations were obtained from the hg19 reference genome GTF file, and repetitive element annotations were retrieved from RepeatMasker (https://www.repeatmasker.org/) [31]. Identified eccDNA fragments were grouped into different length bins and annotated using BEDTools [32] intersect command. To be included in the analysis, eccDNA must have at least 50% overlap with the annotated feature. We calculated the proportion of eccDNA overlapping annotated elements in each length bin using the following formula:

$$\begin{aligned}&\mathit{Proportion\ of\ eccDNA\ overlapping\ annotated\ elements\ in\ each\ length\ bin}\\&=\frac{\mathit{Number\ of\ eccDNA\ overlapping\ with\ annotated\ elements\ in\ each\ length\ bin}}{\mathit{Total\ number\ of\ eccDNA\ across\ all\ length\ bins}}\end{aligned}$$

Cancer type-specific feature extraction

To compare the differences in eccDNA fragments across cancer types, each eccDNA fragment was functionally annotated to a gene. Specifically, the gene body together with a 10 kb upstream region of the promoter was used as the reference for annotation. The total number of bases from eccDNA fragments mapped to the gene body was calculated for each gene. To account for gene length and ensure comparability across genes, the gene cyclization probability was defined as the ratio of the total mapped bases within the gene body to the gene length. The normalized procedure was defined as:

$$Gene\ cyclization\ probability = \frac{Total\ number\ of\ bases\ mapped\ to\ the\ gene\ body}{Gene\ length}$$

This metric reflects the relative abundance and distribution of eccDNA fragments across genes while correcting for gene size.

Machine learning-based evaluation of individual classification performance

Two classical machine learning models, adaptive boosting (AdaBoost) and logistic regression (LR), were applied to evaluate sample-level classification performance. The normalized eccDNA count for each sample was used as the input. Sample labels were binarized, by assigning 1 to cancer and − 1 to healthy individuals. Classification probabilities for each sample were estimated using the "cross_val_predict" function from the scikit-learn 1.5.2 Python package [33]. The output probabilities represent the likelihood of each sample being classified as "cancer" by the model and were visualized using scatter plots across cancer types.

Statistics

For data presented in Figs. 1c, d and f and 2a and b, Figure S2f, Figure S3b, S3c, P-values were calculated via the two-tailed t-test in Python. For data presented in Fig. 2c, Figure S3e, P-values were calculated via the Mann-Whitney-Wilcoxon test in Python. For data presented in Figure S4, P-values were calculated via linear regression analysis using the "linregress" function from the "scipy.stats" module in Python, which performs a two-sided t-test to assess the significance of the slope.

Application of large Language models in the writing process

During the preparation of this work the authors used ChatGPT-4o in order to improve the language and readability. After using the tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Results

Profiling of cell-free EccDNA in cancer patients and healthy individuals

We obtained the profiles of cell-free eccDNA from a total of 652 plasma samples, including 413 cancer patients and 239 healthy individuals (Fig. 1a, Additional file 1: Fig. S1, Additional file 2: Table S1). Based on our recent benchmarking of methodologies for detecting eccDNA [34], we chose and optimized a previously reported library construction protocol using exonuclease V and Tn5 enzyme [20]. Specifically, we introduced a three-step digestion process to facilitate more efficient removal of linear DNA. This optimized method could efficiently purify cell-free eccDNA and reduce bias towards small fragments by omitting the rolling circle amplification (RCA) step (Additional file 1: Fig. S2a). We randomly selected 10 eccDNA fragments ranging from 0.3 kb to 1.6 kb from cancer patients and successfully verified the breakpoint locations and surrounding sequences using outward PCR and Sanger sequencing (Additional file 1: Fig. S2b and Additional file 2: Table S3). To directly validate the existence and reliability of cell-free eccDNA, we employed atomic force microscopy (AFM) to visualize the circular structure of cell-free eccDNA (Fig. 1b). Additionally, we verified the presence of long eccDNA fragments using Oxford Nanopore long-read sequencing and deep Illumina short-read sequencing, thereby confirming the reliability of eccDNA length estimation (Additional file 1: Fig. S2c, S2d). These results were consistent with previous tissue-based eccDNA studies [11, 35], confirming that our method effectively and reliably enriched and detected cell-free eccDNA from plasma. We further confirmed the advantage of omitting the RCA step through comparative analysis. Excluding RCA increased the number of cell-free eccDNA, with markedly enhanced coverage particularly for long fragments (> 1 kb) (Additional file 1: Fig. S2e). We also evaluated the digestion efficiency under different enzymatic conditions and found that three consecutive rounds of digestion effectively eliminated linear DNA contamination, with nearly complete removal observed in our spike-in assays (Additional file 1: Fig. S2f).

Each eccDNA library was sequenced to yield an average of approximately 83 million reads that were mapped to the human genome (Additional file 2: Table S4). High-confidence eccDNA was identified with our adapted bioinformatics pipeline based on split reads, discordant reads, and sequence similarity information (Additional file 1: Fig. S3a). We compared eccDNA differences between cancer patients and healthy individuals utilizing the metric of eccDNA number per million mappable reads (EPM). The EPM in the plasma of cancer patients was approximately twice that of healthy individuals (Additional file 1: Fig. S3b, t-test, $P < 0.0001$). This observation remained consistent across various cancer types, where EPM values were significantly higher than in healthy individuals (Fig. 1c, t-test).

To analyze the fragment length characteristics of cell-free eccDNA, we first divided fragment lengths into five logarithmic intervals and calculated the normalized eccDNA count for each group. The results showed that most eccDNA fragments were distributed in the

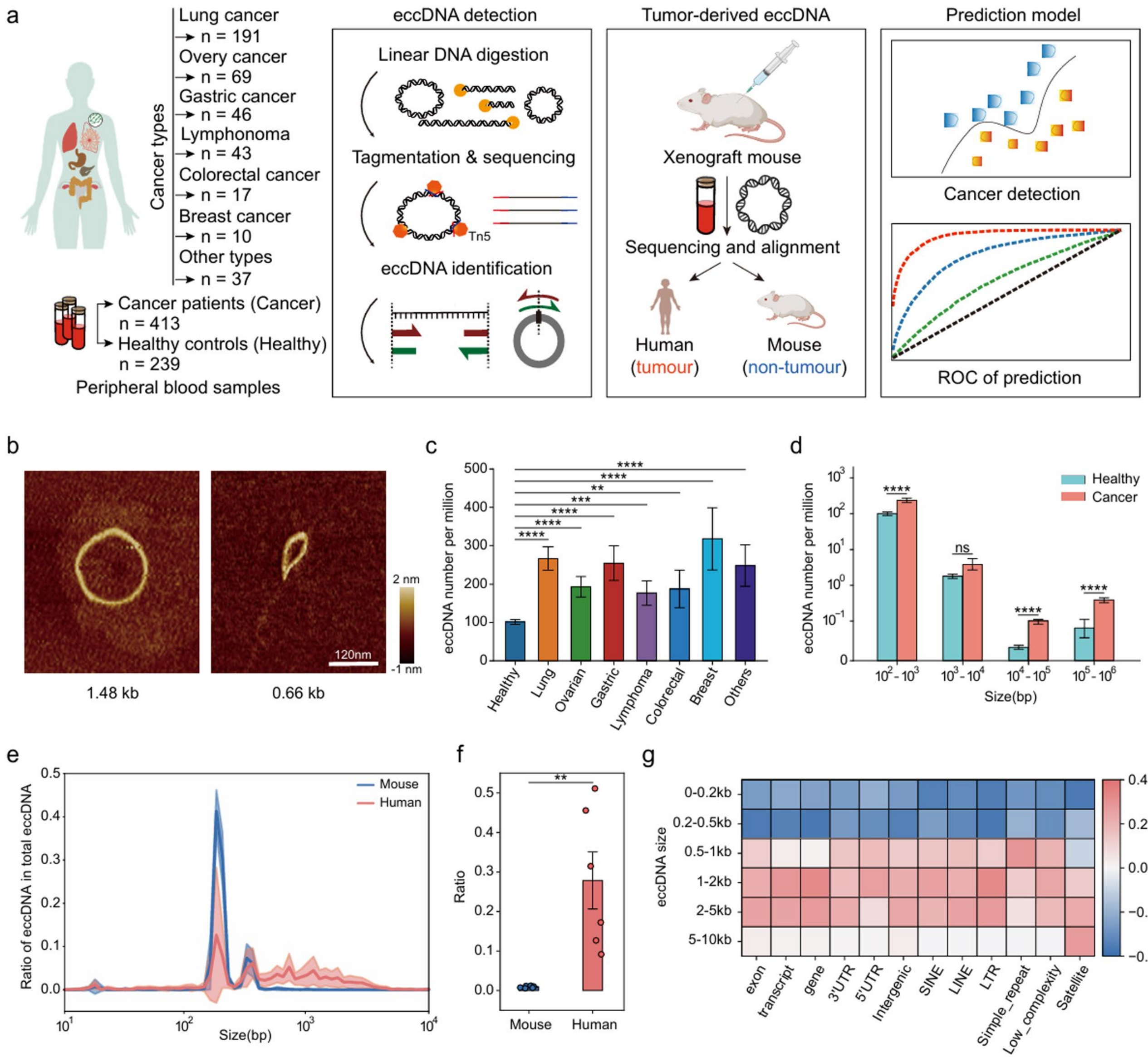

Fig. 1 Characteristics of cell-free extrachromosomal circular DNA in plasma samples. **a** Overview of study profiles. Workflow illustrating the cohort inclusion, eccDNA detection, tumor-derived eccDNA identification and model-based cancer prediction. **b** Atomic force microscopy (AFM) image of cell-free eccDNA purified from human plasma. The observed circular DNA structures are estimated to be approximately 1.48 kb and 0.66 kb in length, based on the assumption of A-form DNA. Scale bar, 120 nm. **c** Normalized cell-free eccDNA counts in plasma samples from healthy individuals and patients across multiple cancer types. Data are presented as mean values ± SEM. **d** Normalized cell-free eccDNA counts detected in healthy individuals and cancer patients, stratified by 10-fold genomic length bins. Data are presented as mean values ± SEM. **e** Size distribution of tumor-derived versus non-tumor-derived cell-free eccDNA identified from plasma of patient-derived xenograft (PDX) mouse. **f** Proportion of long eccDNA (> 1,000 bp) among tumor-derived and non-tumor-derived cell-free eccDNA identified from PDX mouse plasma. **g** Genomic feature enrichment of tumor-derived and non-tumor-derived cell-free eccDNA identified from PDX mouse plasma

100-1,000 bp range (Fig. 1d, t-test). We further stratified fragment lengths into two size bins (< 10^3 bp and ≥ 10^3 bp), revealing a statistically significant enrichment of long eccDNA fragments (≥ 10^3 bp) in cancer patients (Additional file 1: Fig. S3c, t-test). Further analysis of size distribution revealed that the majority of cell-free eccDNA identified in plasma (approximately 80%) was smaller than 500 bp, with two prominent peaks at approximately 200 bp and 360 bp (Additional file 1: Fig. S3d). This distribution pattern was consistent with previous studies on eccDNA in plasma and urine [20, 21].

We then investigated the genomic features of the eccDNA identified from plasma and found the eccDNA distribution was not restricted to specific genomic regions. Compared to healthy individuals, eccDNA in the plasma of cancer patients was significantly enriched in intergenic regions (Mann-Whitney-Wilcoxon test, P < 0.0001) and exonic regions (Mann-Whitney-Wilcoxon

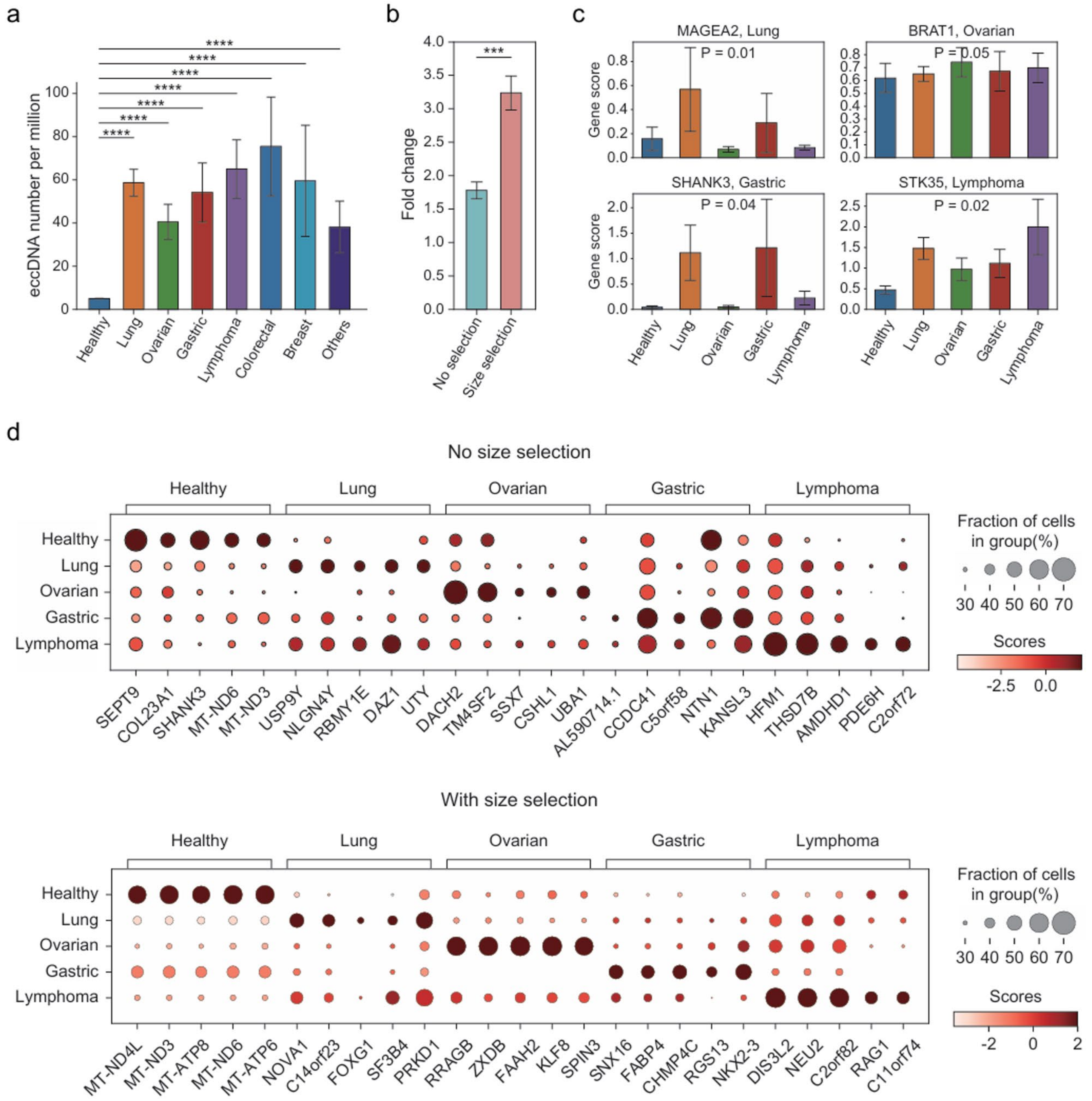

Fig. 2 Size selection of cell-free extrachromosomal circular DNA reveals fragmentomic features and enhances tumor detection sensitivity. **a** Normalized counts of large-size cell-free eccDNA (length >1,000 bp) detected in healthy individuals and patients across multiple cancer types. Data are presented as mean values ± SEM. **b** Comparison of fold changes in normalized eccDNA levels between cancer patients and healthy individuals before and after size selection. **c.** Enrichment of canonical cancer driver genes in specific cancer types based on large-size cell-free eccDNA. **d** Size selection enhances detection sensitivity of tumor-associated genes in large-size cell-free eccDNA

test, $P < 0.05$), whereas its proportion in intronic regions was markedly reduced (Additional file 1: Fig. S3e, Mann-Whitney-Wilcoxon test, $P < 0.0001$). This distribution pattern is strongly consistent with the characteristics of circular DNA fragments observed in tumor tissues in previous studies [36, 37], suggesting that the cell-free eccDNA detected in plasma may originate from the release of tumor tissues.

Characterization of tumor-derived cell-free EccDNA

While cell-free eccDNA showed significant differences between cancer patients and healthy individuals, these results may be affected by blood cell-derived eccDNA, as the majority of cell-free DNA (cfDNA) originates from white blood cells [38–40]. To further characterize the signature of tumor-derived cell-free eccDNA, we generated eccDNA profiles from the plasma of mice bearing patient-derived xenograft (PDX) tumors (Additional file 2: Table S5, N = 6). Using the same library construction and sequencing strategy as applied to human samples, sequencing reads were aligned to both mouse and human reference genomes to identify high-confidence eccDNA. EccDNA mapped to the mouse genome was classified as non-tumor-derived, whereas that mapped to the human genome was considered tumor-derived.

We first compared the size distribution of tumor-derived and non-tumor-derived eccDNA fragments using the sliding window with a step size of 50 bp. A notable shift in fragment size was observed around 1,000 bp, marking the point at which tumor-derived eccDNA became predominant (Fig. 1e). Specifically, eccDNA fragments longer than 1,000 bp accounted for approximately 30% (range 9%–51%) of tumor-derived eccDNA, whereas they only accounted for approximately 1% of non-tumor-derived eccDNA (Fig. 1f, t-test, P < 0.01). These results were consistent with previous tissue eccDNA studies that normal tissues randomly generate small eccDNA while large ecDNA is only observed in cancer cells [10]. Although cell-free eccDNA fragments exceeding 1,000 bp have rarely been reported, recent long-read sequencing studies have identified cfDNA over 10,000 bp in peripheral blood [41–43], further supporting the existence of long cell-free eccDNA.

We observed a significant positive correlation between eccDNA density and protein-coding gene density in both tumor-derived and non-tumor-derived cell-free eccDNA (Additional file 1: Fig. S4), consistent with previous research [35, 44]. Motivated by this genome-wide correlation, we analyzed the enrichment of eccDNA fragments of varying lengths across different genomic features. We further visualized the variation by calculating the differences in genomic distribution between tumor-derived and non-tumor-derived cell-free eccDNA. Notably, large tumor-derived eccDNA fragments (> 1,000 bp) exhibited significantly greater enrichment in exons, transcripts, and gene regions compared to non-tumor-derived counterparts (Fig. 1g). Overall, these results indicate that long tumor-derived eccDNAs may originate from biologically relevant gene regions and hold potential for distinguishing cancer patients from healthy individuals.

Incorporating fragment size selection of EccDNA to distinguish cancer patients from healthy individuals

Given the results from the PDX model, we assessed the utility of selecting large-sized (>1,000 bp) cell-free eccDNA fragments for cancer detection. After in silico size selection of large-sized eccDNA, our multi-cancer cohort displayed a much greater difference in normalized eccDNA number between cancer patients and healthy individuals (Fig. 2a, t-test). Specifically, the average fold change of normalized eccDNA between cancer patients and healthy individuals increased from 1.8-fold to 3.2-fold after size selection, indicating a substantial enrichment of tumor-derived signals (Fig. 2b, t-test, P<0.001).

As eccDNA can carry distinct gene content in various cancer types, we then assessed the variation in genes carried by cell-free eccDNA among different tumors. We developed a gene-based annotation method (Additional file 1: Fig. S5) and analyzed the genes carried by eccDNA from four cancer types (lung cancer, n=191; ovarian cancer, n=69; gastric cancer, n=46; lymphoma, n=43) compared to healthy individuals (n=239). Specifically, each identified eccDNA fragment was mapped to the human reference genome. For each gene, we defined the normalized number of eccDNA fragments mapped to the gene body as the "gene score", assessing the potential gene regulatory role of eccDNA (see Methods).

The gene annotation results of large-sized eccDNA showed significant enrichment of several well-known cancer-related driver genes, including *MAGEA2*[45] (lung cancer), *BRAT1*[46] (ovarian cancer), *SHANK3*[47] (gastric cancer), and *STK35*[48] (lymphoma), in their corresponding cancer types (Fig. 2c, Mann-Whitney-Wilcoxon test). Notably, while mitochondria-related genes were enriched in healthy individuals across both total and large-sized eccDNA, the tumor groups displayed a broader set of tumor-associated genes with stronger enrichment patterns under the large-sized eccDNA context (Fig. 2d). Furthermore, we compared these genes carried by cell-free eccDNA with TCGA's cancer type specific highly copy number amplification gene (HCNA) and ecDNA containing gene identified by a recent study [49]. There is a highly correlation between the frequency of pan-lung cancer HCNA gene from TCGA and genes carried by cell-free eccDNA (Additional file 1: Fig. S6a, r ~ 0.5, P < 0.01). We also performed gene set enrichment analysis (GSEA) and found that ecDNA containing gene was significantly enriched in plasma cell-free eccDNA from lung cancer patients compared with healthy individuals (running score = 0.68) (Additional file 1: Fig. S6b). Taken together, we concluded that selecting large-sized cell-free eccDNA fragments in plasma can enrich tumor content on a genome-wide scale. These findings suggest that large eccDNA fragments in plasma reflect the molecular characteristics of specific tumor types, underscoring their promising potential for clinical applications.

ScanTecc: integrating feature selection and machine learning for cancer diagnosis

To investigate the potential of cell-free eccDNA characteristics in plasma for cancer detection and tissue-of-origin classification, we developed ScanTecc (screening cancer types with cell-free eccDNA), a machine learning-based framework that integrates eccDNA feature selection, to distinguish cancer patients from healthy

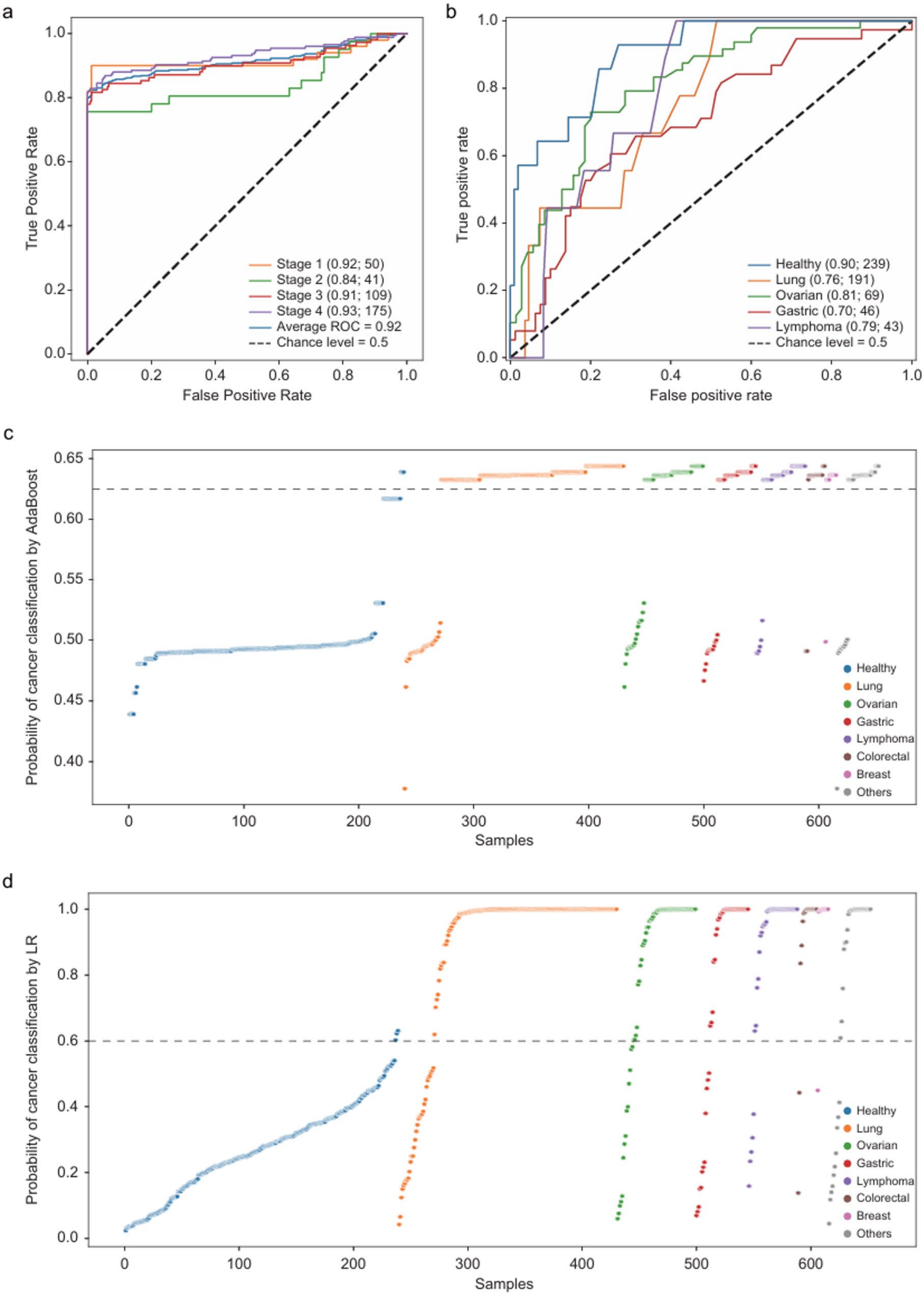

Fig. 3 (See legend on next page.)

(See figure on previous page.)

Fig. 3 Performance of ScanTecc in cancer detection across disease stages and cancer types. **a** ROC analysis of ScanTecc performance for cancer detection across different disease stages. **b** ROC analysis of ScanTecc performance for tissue-of-origin classification in cancer patients versus healthy controls. **c** Plot representing the probability of classification as cancer with the adaptive boosting (AdaBoost) model for all samples. Samples are separated by cancer type and sorted within each by the prediction probability of classification as cancer. The dashed horizontal line indicates 62.5% probability (achieving specificity 99%). **d** Plot representing the probability of classification as cancer with the logistic regression (LR) model for all samples. Samples are separated by cancer type and sorted within each by the prediction probability of classification as cancer. The dashed horizontal line indicates 60% probability (achieving specificity 99%)

individuals and determine cancer types. Receiver operating characteristic (ROC) analyses based on the normalized number of eccDNA showed that size-selected cell-free eccDNA fragments (>1,000 bp) achieved an AUC of 0.92 for overall cancer detection, markedly outperforming the diagnostic performance of total eccDNA without size selection (AUC = 0.72). Specifically, the AUC for stage I cancer patients was 0.92, and that for stage IV patients reached 0.93 (Fig. 3a, Additional file 1: Fig. S7a). To further enhance the reliability of the prediction results, we applied ScanTecc to an independent validation cohort comprising 14 healthy individuals and 21 lung cancer patients, achieving a high prediction accuracy (AUC = 0.88) (Additional file 1: Fig. S8). We also randomly excluded 10% of tumor samples as a holdout test set, using the remaining 90% of samples for model training and internal validation to assess the generalizability of ScanTecc. Our ScanTecc model achieved a high AUC of 0.94 on the holdout test set, with only a minimal difference (ΔAUC = 0.03) compared to the training set, indicating strong generalization ability and low risk of over-training (Additional file 1: Fig. S9a, S9b).

In addition, we applied ScanTecc to predict the tissue of origin and evaluate classification performance across multiple cancer types. Through three-fold cross-validation, ScanTecc achieved variable predictive performance, with AUC values ranging from 0.70 for gastric cancer to 0.81 for ovarian cancer across all stages (Fig. 3b). However, integrating gene annotation information did not significantly improve the cancer diagnosis performance compared to using the normalized eccDNA number only (Additional file 1: Fig. S7b). We next compared the classification performance of ScanTecc with that of traditional tumor markers by evaluating CEA and CA19-9 in a published gastric cancer cohort [27]. Analyzing data from the cohort achieved an AUC of 0.58 (Additional file 1: Fig. S10a). However, using their healthy controls and our gastric cancer samples, the AUCs dropped to 0.34 and 0.35 for CEA and CA19-9, respectively (Additional file 1: Fig. S10b, S10c), far below the performance of ScanTecc. We further analyzed their positive rates and observed increasing positive rates with tumor stage (Additional file 1: Fig. S10d).

To evaluate the classification performance of ScanTecc at the individual sample level, we introduced two widely used machine learning models, adaptive boosting (AdaBoost) and logistic regression (LR), to assess cancer classification probabilities across all 652 samples. At a specificity of 99%, the AdaBoost model identified 331 out of 413 cancer samples (80%) as positive, with most tumor samples exhibiting predicted probabilities exceeding the 0.625 predict probability (Fig. 3c). Similarly, the LR model demonstrated comparable performance, recognizing 338 cancer cases (82%) at 99% specificity, using a probability threshold of 0.6 (Fig. 3d). To assess whether ScanTecc exhibits any sex-specific performance bias, we stratified our cohort by gender and compared the prediction results. ScanTecc achieved an AUC of 0.91 in females (*n* = 334) and 0.92 in males (*n* = 318), showing robust performance across both subgroups (Additional file 1: Fig. S11). The difference in AUC of 0.01 between the two subgroups was minimal, suggesting that gender does not significantly impact the model's classification ability. Collectively, these results further confirm the robustness and generalizability of the ScanTecc framework across different machine learning algorithms, highlighting its potential as a non-invasive and clinically applicable cancer screening tool.

Discussion

Circulating eccDNA has the potential to serve as a ctDNA biomarker that can be used to complement studies of linear ctDNA, since eccDNA is commonly characterized in cancer cell lines and somatic tissues. A previous study reported that small eccDNA originating from specific genes had great diagnostic value for multiple cancers in both tissues and plasma [22]. However, these findings are based on limited sample sizes, and the diagnostic performance still needs to be combined with traditional tumor markers CEA/CA19-9 to achieve better results. Therefore, research on the molecular characteristics, functions, and clinical application potential of cell-free eccDNA in large-scale cohorts remains limited.

In our study, we demonstrated that cell-free eccDNA is commonly present in the plasma of both cancer patients and healthy individuals, with significantly higher abundance and longer fragment lengths observed in cancer patients. We found that eccDNA is preferentially enriched in exonic regions and reduced in intronic regions, consistent with a previous study that reported eccDNA tends to overlap more with exons than with introns [50]. This distinct distribution pattern likely reflects the eccDNA biogenesis mechanism linked to transcriptional activity and RNA splicing, both of which are typically upregulated in cancer. We also observed an enrichment of driver genes on eccDNA, likely due to the close association between eccDNA generation and transcriptional activity. Since each tumor type has a distinct transcriptional profile, this leads to tumor-specific eccDNA landscapes. Consequently, eccDNAs associated with tumor-specific driver genes are prominently identified in our analysis, reflecting the underlying transcriptional specificity of different tumor types.

By integrating machine learning algorithms, we developed ScanTecc and demonstrated its high accuracy in distinguishing cancer patients from healthy individuals. Furthermore, existing cfDNA-based tissue-of-origin prediction models often suffer from limited performance, due to the low abundance of tumor-derived cfDNA and the substantial background of non-tumor-derived DNA in circulation [51–53]. To address this challenge, ScanTecc employed size selection for effectively enrichment of tumor-derived components and utilized the gene score to further differentiate cancer types, thereby enhancing the accuracy of tissue-of-origin prediction. Our results suggest that cell-free eccDNA in plasma may serve as a more stable and sensitive liquid biopsy biomarker, providing an efficient, reliable, and non-invasive approach for early cancer detection. However, the performance of our model ScanTecc remains limited by its dependence on predefined cancer types during supervised learning, restricting its ability to recognize unseen tumor classes. Future work could focus on integrating open-set classification strategies to address this limitation.

Current strategies for cell-free eccDNA purification and identification still require urgent optimization, despite the development of various experimental methods and analysis pipelines for eccDNA detection in recent years, such as WGS + AmpliconArchitect (AA) [54] and Circle-seq [35] + Circle-Map [55]. However, our recent benchmarking study [34] has reported the performance and biases of these strategies. This is because the environment in the circulating system differs from that in cells. For example, there are no large fragments of chromosomes to degrade, and the ratio of circular DNA to linear DNA is lower. These differences impair the performance of existing methods, and WGS + AA is almost unable to identify cell-free eccDNA, even at 50x sequencing depth. Additionally, the RCA step is known to preferentially amplify small eccDNA, which affects the detection efficiency of large fragments [56]. To obtain a more representative and unbiased profile of cell-free eccDNA, we selected the experimental protocol that omits RCA, despite its relatively low overall eccDNA detection rate. A critical step in this protocol is the enzymatic digestion of linear DNA during library construction. Extending the digestion time and increasing the number of digestion rounds significantly improved the removal efficiency of linear DNA, with nearly complete elimination achieved after three rounds. More importantly, second-generation sequencing methods may not accurately resolve the true structure of eccDNA, potentially leading to false positives for long eccDNA [57]. Future detection strategies incorporating high-precision, high-enrichment efficiency, and long-read sequencing technologies like Nanopore + CReSIL [58] or 3SEP + eccDNA_RCA_nanopore [12, 59], may provide a more comprehensive eccDNA profile and advance its precise application in clinical diagnostics.

Previous studies have proposed using whole-genome sequencing [4, 60], whole-exome sequencing [7, 61], or whole-genome bisulfite sequencing [39, 62] of plasma DNA at multiple time points during cancer treatment, to monitor tumor evolution and identify potential resistance mechanisms. However, these technologies are costly and typically applicable only to samples with the tumor DNA fraction greater than 5%–10% [63–65]. Our work provides evidence that tumor-derived cell-free eccDNA in many cancer types is larger than 1,000 bp. This finding suggests that fragment size selection strategies, either in vitro or in silico, could effectively enrich tumor components in plasma samples and offer a viable approach for improving detection sensitivity in low-abundance samples. However, the generation mechanism of eccDNA in cancer cells and its release into peripheral blood circulation remains unclear. Further studies on eccDNA profiling from primary tumor tissues and paired plasma may provide insights into its biological origins and circulation mechanisms.

In summary, our investigation has revealed distinct molecular characteristics of cell-free eccDNA in plasma between cancer patients and healthy individuals. Notably, our findings underscore the significance of size selection and gene annotation in identifying tumor-derived eccDNA. Our integrated analytical approach, exemplified by ScanTecc, has demonstrated the clinical potential of cell-free eccDNA as a non-invasive liquid biopsy biomarker. These findings offer theoretical and practical support for the clinical utility of cell-free eccDNA in early cancer diagnosis and personalized medicine.

Conclusions

In conclusion, we have found that plasma from cancer patients contains significantly higher levels of eccDNA and longer fragment lengths compared to healthy individuals. We have also identified a tumor-specific eccDNA size threshold and developed a machine learning-based approach that utilizes eccDNA profiles to distinguish cancer patients from healthy individuals and classify cancer types with high accuracy. These results provide the feasibility of applying peripheral blood-derived eccDNA signatures for non-invasive cancer diagnosis.

Abbreviations

AdaBoost	Adaptive boosting
AFM	Atomic force microscopy
AUC	Area under the ROC curve
ctDNA	Circulating tumor DNA
cfDNA	Cell-free DNA
eccDNA	Extrachromosomal circular DNA
EPM	eccDNA number per million mappable reads
LR	Logistic regression
PCR	Polymerase chain reaction
PDX	Patient-derived xenograft
RCA	Rolling circle amplification
ROC	Receiver operating characteristic
WGS	Whole-genome sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-025-01595-6.

Additional file 1: Figures S1-S11

Additional file 2: Tables S1-S5

Acknowledgements

We thank all members in the Qu laboratory for helpful discussions. We thank the USTC supercomputing center and the School of Life Science Bioinformatics Center for providing computing resources for this project.

Authors' contributions

K. Q., J. F., and C. G. conceived and supervised the project. J. F. and S. L. designed the framework and performed data analysis with the help of Q. Y., W. Z., K. L., J. Y., K. L., G. X., and X. Y. S. L., and Y. S. performed atomic-force microscopy and high throughput sequencing experiments with the help from R. S. and Y. Z. Y. X., J. W., B. S., M. H., M. Z., K. L., X. X., Y. Z., Z. Z., X. H., and Y. P. provided clinical blood samples from multiple cancer patients and healthy individuals. C. G., J. F., S. L., and K. Q. wrote the manuscript with inputs from all authors. All authors read and approved the final manuscript.

Funding

This work was supported by the National Key R&D Program of China (2022YFA1303200 to K.Q.), the National Natural Science Foundation of China grants (T2125012 to K.Q., and 32270978 to C.G.), USTC Research Funds of the Double First-Class Initiative (YD9100002032 to K.Q. and YD9990006004 to S.L.), the Strategic Priority Research Program of Chinese Academy of Sciences (XDB0940301 to K.Q.), Anhui Province Natural Science Foundation (2508085QC071 to S.L.) and the Fundamental Research Funds for the Central Universities (WK9100000086 to C.G.).

Data availability

All the raw sequencing data generated in this study are available and have been deposited in the Genome Sequence Archive for Human (GSA-Human) database [66] in National Genomics Data Center [67], China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences under accession code HRA002648, (https://ngdc.cncb.ac.cn/gsa-human/browse/HRA002648) [68]. All other data supporting the findings described in this paper are available in the article and its supplementary information files. Source code and analysis scripts supporting the findings of this study are available on the Github repository (https://github.com/QuKunLab/ScanTecc) [69] and have been archived on Zenodo with a DOI (https://doi.org/10.5281/zenodo.17971506).

Declarations

Ethics approval and consent to participate

This study was approved by the ethics committee of The First Affiliated Hospital of the University of Science and Technology of China (No. 2022-KY-264). All participants provided written informed consent to participate in this study. All procedures involving human participants conformed to the principles of the Helsinki Declaration. The animal model study scheme was approved by the Institutional Animal Care and Use Committee (IACUC) of Precedo Pharmaceuticals Co., Ltd. (Approval No. IACUC-20220801). During the study, animal breeding and use were strictly conducted in accordance with the regulations and guidelines of the International Management Committee for Assessment and Certification of Laboratory Animals.

Consent for publication

Not applicable.

Competing interests

Jingwen Fang is the chief executive officer of HanGene Biotech. Kun Qu and Chuang Guo are science advisers of HanGene Biotech. Authors Yunying Shao, Jiaxuan Yang, YouYang Zhou are employees of HanGene Biotech. The remaining authors declare that they have no competing interests.

Author details

[1]Department of Oncology, The First Affiliated Hospital of USTC, State Key Laboratory of Eye Health, School of Basic Medical Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230021, Anhui, China
[2]HanGene Clinical Laboratory, Hefei, Anhui, China
[3]Department of Respiratory Oncology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China
[4]Department of Obstetrics & Gynecology, The First Affiliated Hospital of Anhui Medical University, Hefei 230022, Anhui, China
[5]Department of Health Management Center, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China
[6]Anhui Province Key Laboratory of Biomedical Imaging and Intelligent Processing, Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, Anhui, China
[7]CAS Center for Excellence in Molecular Cell Sciences, the CAS Key Laboratory of Innate Immunity and Chronic Disease, University of Science and Technology of China, Hefei 230027, Anhui, China
[8]Department of Health Management Center, The First Affiliated Hospital of Anhui Medical University, Hefei 230022, Anhui, China
[9]Department of Rheumatology and Immunology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China
[10]School of Pharmacy, Bengbu Medical University, Bengbu, China

Received: 16 July 2025 / Accepted: 24 December 2025
Published online: 06 January 2026

References

1. Tivey A, Church M, Rothwell D, Dive C, Cook N. Circulating tumour DNA - looking beyond the blood. Nat Rev Clin Oncol. 2022;19:600–12. https://doi.org/10.1038/s41571-022-00660-y.
2. Sanz-Garcia E, Zhao E, Bratman SV, Siu LL. Monitoring and adapting cancer treatment using circulating tumor DNA kinetics: current research, opportunities, and challenges. Sci Adv. 2022;8:eabi8618. https://doi.org/10.1126/sciadv.abi8618.
3. Pantel K, Alix-Panabieres C. Liquid biopsy and minimal residual disease - latest advances and implications for cure. Nat Rev Clin Oncol. 2019;16:409–24. https://doi.org/10.1038/s41571-019-0187-3.
4. Budhraja KK, et al. Genome-wide analysis of aberrant position and sequence of plasma DNA fragment ends in patients with cancer. Sci Transl Med. 2023;15:eabm6863. https://doi.org/10.1126/scitranslmed.abm6863.
5. Li S, et al. Comprehensive tissue deconvolution of cell- free DNA by deep learning for disease diagnosis and monitoring. Proc Natl Acad Sci U S A. 2023;120:e2305236120. https://doi.org/10.1073/pnas.2305236120.
6. Cristiano S, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570:385–9. https://doi.org/10.1038/s41586-019-1272-6.
7. Murtaza M, et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature. 2013;497:108–12. https://doi.org/10.1038/nature12065.
8. Crosby D, et al. Early detection of cancer. Science. 2022;375:eaay9040. https://doi.org/10.1126/science.aay9040.

9. Yang L, et al. Extrachromosomal circular DNA: biogenesis, structure, functions and diseases. Signal Transduct Target Ther. 2022;7:342. https://doi.org/10.1038/s41392-022-01176-8.
10. Turner KM, et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature. 2017;543:122–5. https://doi.org/10.1038/nature21356.
11. Kim H, et al. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nat Genet. 2020;52:891–7. https://doi.org/10.1038/s41588-020-0678-2.
12. Wang Y, et al. eccDNAs are apoptotic products with high innate immunostimulatory activity. Nature. 2021;599:308–14. https://doi.org/10.1038/s41586-021-04009-w.
13. Wu S, et al. Circular ecDNA promotes accessible chromatin and high oncogene expression. Nature. 2019;575:699–703. https://doi.org/10.1038/s41586-019-1763-5.
14. Luebeck J, et al. Extrachromosomal DNA in the cancerous transformation of Barrett's oesophagus. Nature. 2023;616:798–805. https://doi.org/10.1038/s41586-023-05937-5.
15. Zhu Y, et al. Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. Cancer Cell. 2021;39:694–707. https://doi.org/10.1016/j.ccell.2021.03.006.
16. Bergstrom EN, et al. Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. Nature. 2022;602:510–7. https://doi.org/10.1038/s41586-022-04398-6.
17. Yi E, et al. Live-cell imaging shows uneven segregation of extrachromosomal DNA elements and transcriptionally active extrachromosomal DNA hubs in cancer. Cancer Discov. 2022;12:468–83. https://doi.org/10.1158/2159-8290.Cd-21-1376.
18. Yang F, et al. Retrotransposons hijack alt-EJ for DNA replication and eccDNA biogenesis. Nature. 2023;620:218–25. https://doi.org/10.1038/s41586-023-06327-7.
19. Paulsen T, Kumar P, Koseoglu MM, Dutta A. Discoveries of extrachromosomal circles of DNA in normal and tumor cells. Trends Genet. 2018;34:270–8. https://doi.org/10.1016/j.tig.2017.12.010.
20. Sin STK, et al. Identification and characterization of extrachromosomal circular DNA in maternal plasma. Proc Natl Acad Sci U S A. 2020;117:1658–65. https://doi.org/10.1073/pnas.1914949117.
21. Lv W, et al. Circle-Seq reveals genomic and disease-specific hallmarks in urinary cell-free extrachromosomal circular DNAs. Clin Transl Med. 2022;12:e817. https://doi.org/10.1002/ctm2.817.
22. Luo X, et al. Small extrachromosomal circular DNAs as biomarkers for multicancer diagnosis and monitoring. Clin Transl Med. 2023;13:e1393. https://doi.org/10.1002/ctm2.1393.
23. Hansen LB, et al. Methods for the purification and detection of single nucleotide KRAS mutations on extrachromosomal circular DNA in human plasma. Cancer Med. 2023;12:17679–91. https://doi.org/10.1002/cam4.6385.
24. Kumar P, et al. Normal and cancerous tissues release extrachromosomal circular DNA (eccDNA) into the circulation. Mol Cancer Res. 2017;15:1197–205. https://doi.org/10.1158/1541-7786.Mcr-17-0095.
25. Ma LW, et al. Liquid biopsy in cancer: current status, challenges and future prospects. Signal Transduct Target Ther. 2024. https://doi.org/10.1038/s41392-024-02021-w.
26. Campos-Carrillo A, et al. Circulating tumor DNA as an early cancer detection tool. Pharmacol Ther. 2020. https://doi.org/10.1016/j.pharmthera.2019.107458.
27. Johnson AEW, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023. https://doi.org/10.1038/s41597-022-01899-x.
28. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
29. Danecek P, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10:giab008. https://doi.org/10.1093/gigascience/giab008.
30. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4. https://doi.org/10.1093/bioinformatics/btv098.
31. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. http://www.repeatmasker.org (2013–2015).
32. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.
33. Pedregosa F, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
34. Gao X, et al. Comparative analysis of methodologies for detecting extrachromosomal circular DNA. Nat Commun. 2024;15:9208. https://doi.org/10.1038/s41467-024-53496-8.
35. Moller HD, et al. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. Nat Commun. 2018;9:1069. https://doi.org/10.1038/s41467-018-03369-8.
36. Lv W, et al. Extrachromosomal circular DNA orchestrates genome heterogeneity in urothelial bladder carcinoma. Theranostics. 2024;14:5102–22. https://doi.org/10.7150/thno.99563.
37. Sheng Z, et al. Genome-wide characterization of extrachromosomal circular DNA in breast cancer and its potential role in carcinogenesis and cancer progression. Cell Rep. 2024;43:114845. https://doi.org/10.1016/j.celrep.2024.114845.
38. Kustanovich A, Schwartz R, Peretz T, Grinshpun A. Life and death of circulating cell-free DNA. Cancer Biol Ther. 2019;20:1057–67. https://doi.org/10.1080/15384047.2019.1598759.
39. Sun K, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A. 2015;112:E5503–12. https://doi.org/10.1073/pnas.1508736112.
40. Moss J, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun. 2018;9:5068. https://doi.org/10.1038/s41467-018-07466-6.
41. Yu SCY, et al. Single-molecule sequencing reveals a large population of long cell-free DNA molecules in maternal plasma. Proc Natl Acad Sci U S A. 2021;118:e2114937118. https://doi.org/10.1073/pnas.2114937118.
42. Yu SCY, Choy LYL, Lo YMD. "Longing" for the next generation of liquid biopsy: the diagnostic potential of long cell-free DNA in oncology and prenatal testing. Mol Diagn Ther. 2023;27:563–71. https://doi.org/10.1007/s40291-023-00661-2.
43. Choy LYL, et al. Single-molecule sequencing enables long cell-free DNA detection and direct methylation analysis for cancer patients. Clin Chem. 2022;68:1151–63. https://doi.org/10.1093/clinchem/hvac086.
44. Henriksen RA, et al. Circular DNA in the human germline and its association with recombination. Mol Cell. 2022;82:209–17. https://doi.org/10.1016/j.molcel.2021.11.027.
45. Ujiie H, et al. Overexpression of MAGEA2 has a prognostic significance and is a potential therapeutic target for patients with lung cancer. Int J Oncol. 2017;50:2154–70. https://doi.org/10.3892/ijo.2017.3984.
46. Srivastava S, et al. BRAT1 mutations present with a spectrum of clinical severity. Am J Med Genet A. 2016;170:2265–73. https://doi.org/10.1002/ajmg.a.37783.
47. Zhao L-L, et al. The effect of LNCRNA SHANK3 on the malignant development of gastric cancer cells by regulating the miR-4530/MNX1. Transl Oncol. 2024;46:102000. https://doi.org/10.1016/j.tranon.2024.102000.
48. Polyanskaya SA, et al. SCP4-STK35/PDIK1L complex is a dual phospho-catalytic signaling dependency in acute myeloid leukemia. Cell Rep. 2022;38:110233. https://doi.org/10.1016/j.celrep.2021.110233.
49. Khandekar A et al. Examining the Role of Extrachromosomal DNA in 1,216 Lung Cancers. bioRxiv, 2025.2006. 2003.657117. 2025.
50. Dillon LW, et al. Production of extrachromosomal microDNAs is linked to mismatch repair pathways and transcriptional activity. Cell Rep. 2015;11:1749–59. https://doi.org/10.1016/j.celrep.2015.05.020.
51. Doan NNT, et al. Tissue of origin detection for cancer tumor using low-depth cfDNA samples through combination of tumor-specific methylation atlas and genome-wide methylation density in graph convolutional neural networks. J Clin Oncol. 2024;42:224–224. https://doi.org/10.1200/JCO.2024.42.23_suppl.224.
52. Liu MC, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020;31:745–59. https://doi.org/10.1016/j.annonc.2020.02.011.
53. Sharma M, Verma RK, Kumar S, Kumar V. Computational challenges in detection of cancer using cell-free DNA methylation. Comput Struct Biotechnol J. 2022;20:26–39. https://doi.org/10.1016/j.csbj.2021.12.001.
54. Deshpande V, et al. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nat Commun. 2019;10:392. https://doi.org/10.1038/s41467-018-08200-y.
55. Prada-Luengo I, Krogh A, Maretty L, Regenberg B. Sensitive detection of circular DNAs at single-nucleotide resolution using guided realignment of partially aligned reads. BMC Bioinformatics. 2019;20:663. https://doi.org/10.1186/s12859-019-3160-3.

56. Norman A, et al. An improved method for including upper size range plasmids in metamobilomes. PLoS One. 2014;9:e104405. https://doi.org/10.1371/journal.pone.0104405.
57. Noer JB, Horsdal OK, Xiang X, Luo Y, Regenberg B. Extrachromosomal circular DNA in cancer: history, current knowledge, and methods. Trends Genet. 2022;38:766–81. https://doi.org/10.1016/j.tig.2022.02.007.
58. Wanchai V, et al. CReSIL: accurate identification of extrachromosomal circular DNA from long-read sequences. Brief Bioinform. 2022;23:bbac422. https://doi.org/10.1093/bib/bbac422.
59. Wang Y, Wang M, Zhang Y. Purification, full-length sequencing and genomic origin mapping of eccDNA. Nat Protoc. 2023;18:683–99. https://doi.org/10.1038/s41596-022-00783-7.
60. Mouliere F, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. 2018;10:eaat4921. https://doi.org/10.1126/scitranslmed.aat4921.
61. Manier S, et al. Whole-exome sequencing of cell-free DNA and circulating tumor cells in multiple myeloma. Nat Commun. 2018;9:1691. https://doi.org/10.1038/s41467-018-04001-5.
62. Cui P, et al. Prediction of methylation status using WGS data of plasma cfDNA for multi-cancer early detection (MCED). Clin Epigenetics. 2024;16:34. https://doi.org/10.1186/s13148-024-01646-6.
63. Adalsteinsson VA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. 2017;8:1324. https://doi.org/10.1038/s41467-017-00965-y.
64. Heitzer E, et al. Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. Genome Med. 2013;5:30. https://doi.org/10.1186/gm434.
65. Belic J, et al. Rapid identification of plasma DNA samples with increased ctDNA levels by a modified FAST-SeqS approach. Clin Chem. 2015;61:838–49. https://doi.org/10.1373/clinchem.2014.234286.
66. Chen T, et al. The genome sequence archive family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics. 2021;19:578–83. https://doi.org/10.1016/j.gpb.2021.08.001.
67. Xue Y, et al. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2022. Nucleic Acids Res. 2022;50:D27–38. https://doi.org/10.1093/nar/gkab951.
68. Fang J et al. Detection of primary cancer types via fragment size selection in Circulating cell-free extrachromosomal circular DNA. Genome Sequence Archive for Human (HRA002648). https://ngdc.cncb.ac.cn/gsa-human/browse/HRA002648 .2025.
69. Fang J et al. Detection of primary cancer types via fragment size selection in Circulating cell-free extrachromosomal circular DNA. Github. https://github.com/QuKunLab/ScanTecc .2025.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.