nature portfolio

Peer Review File

Comparative analysis of methodologies for detecting extrachromosomal circular DNA

Corresponding Author: Professor Kun Qu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Gao et al present a manuscript titled “Comparative analysis of methodologies for detecting extrachromosomal circular DNA” describe a comparison of different bioinformatics tools to identified extrachromosomal circular DNAs (eccDNAs) using simulated data, which has similar characteristic of some published eccDNA results as the gold standard. Based on the vary of sequencing depth of simulated datasets with chimeric sequences contaminations of both short and long reads, the authors reported that Circle-Map and CReSIL are the top performer for short and long read respectively, with a stable computational resource consumption. Then the authors further applied these two bioinformatic tools for further investigation of different 7 eccDNA enrichment methods. The authors compared the identified eccDNAs of the 7 enrichment methods with many aspects using both short and long read and conclude that Circle-Seq method with long reads give highest detection efficiency for longer eccDNA, while SEP with long reads is more effective for shorter eccDNA. The study is straight forward yet important, and the conclusion is valid and give good guidance for eccDNA research field. However, there are lacking comparisons that will improve the quality of the manuscript and missing aspect that need to be reported.

Comments

- Couple published papers show that there is extremely high heterogeneity in eccDNA profile, leading to a difficult to detect the high percentage of common eccDNAs across different technical and biological replicates. I know that; this point will create many questions on study eccDNA as low commonality will present across replicates. The study pooled the DNA extracted from Hela cells and aliquoted for the 7 different enrichment methods, the authors can compare the percentage of commonly detect eccDNA across different technical replicates of individual enrichment methods and across different enrichment methods. I know that it will be LOW (I think the authors did that already), don’t be afraid to put the results because it is important to the filed to know. It will be the research area that need the field to pay attention and possible develop a better strategy to study eccDNA together. I highly recommend the author to work on the results and present in the manuscript. Figure 4 is still missing . Adding a section around the results would be highly appropriate.

- The ecDNA is an important molecule esp., for cancer biology. The authors need to investigate and present more results. For example, the association of copy number identified from WGS-LR data with the sequencing depth of the overlapped ecDNA. What’s known oncogenes of Hela cell are harbored by ecDNAs or eccDNA.

- Some basic statistic of identified eccDNAs that associated with the known genomic features should be reported.

- Fig. 1 especially panel C and is very hard to read, all things are squeezed. Please improve.

- What the is x axis of the Fig 2C ?. It is not mentioned on figure legend.

- Notebook for compare statistics between given data and simulated data on github is missing

Reviewer #2

(Remarks to the Author)

Comparative analysis of methodologies for detecting extrachromosomal circular DNA

This is a timely benchmarking paper on the eccDNA method. However, there is a lot more that needs to be done to make a fair comparison between various tools. Below are some suggestions that should be considered to improve the benchmarking process.

1. Define the problem clearly: Are you benchmarking the tools that identify simple eccDNA or complex eccDNA? Simple eccDNA refers to circular DNAs generated by end-to-end ligation of one linear fragment. On the other hand, complex eccDNA is composed of two or more linear fragments and may be considered chimeric. Some eccDNA tools claim to identify only simple eccDNA, and therefore should not be evaluated for their ability to identify chimeric eccDNA.
2. Parameter and final output filtering: This text "Realign to identify eccDNA and used recommended filters (circle score > 50, split reads > 2, discordant reads > 2, coverage increase in the start coordinate > 0.33 and coverage increase in the end coordinate > 0.33)." is from the manuscript. The default parameters for Circle-Map seem to differ from those used for other tools. The authors should make an effort to optimize parameters for other tools as well if they choose to optimize specifically for Circle-Map.
3. Why AmpliconArchitect is not part of Figure 1?
4. The Circle-Map outperforms other compared tools when the coverage exceeds 35X. This should be included in the abstract.
5. What was the rationale behind selecting the similarity threshold? Why weren't the tools evaluated for detecting eccDNA at the basepair resolution?
6. There is ample evidence that multiple eccDNA are generated from the same locus. As long as there is evidence of different junctional sequences, especially in the case of multiple junctional reads, they should be considered as different circles.
7. True positive identification was defined as having over 90% sequence identity for the simulated dataset. Why was this not set to near 100% identity, or to be at the level of base pair resolution?
8. Creating a table of circular DNA finding tools, along with their development dates and supported sequencing protocols, would be beneficial for providing a comprehensive overview of available tools in the field.
9. Impact of eccDNA enrichment steps on eccDNA identification: Circle_finder has a distinct pipeline (microDNA.lnOne.sh) tailored for an exhaustive search, making it more suitable for circular DNA-enriched samples (PMID: 28550083). In fact, Circle-Map was initially compared to Circle_finder upon its first publication.
10. Detection efficiency of specific type of eccDNA: What was the coverage of eccDNA identified in real data, and how random was the selection of eccDNA for verification by PCR method?
11. here are two Circle_finder pipelines available on GitHub: "microDNA.lnOne.sh" and "circle_finder-pipeline-bwa-mem-samblaster.sh". The developer mentions on their GitHub page that the "microDNA.lnOne.sh" has higher specificity but takes longer to run. Can the author explain why they chose the "circle_finder-pipeline-bwa-mem-samblaster.sh" script to compare it with other tools? "microDNA.lnOne.sh" has been used for multiple publications where eccDNA were enriched. I suggest it would be a good idea to also evaluate the more sensitive version of Circle_finder.
12. Computational resources consumed by different analysis pipelines: The author states that "Notably, Circle-Map and CReSIL kept stable computational resources consumption across all sequencing depths (Supplementary Figure 1F & 1G).". This statement seems a bit odd in this context. Instead of commenting on runtime and memory usage, they are commenting on which pipeline has stable computational resource consumption. Looking at Supplementary Figure 1F & 1G, Circle-Map and CReSIL are not the winners compared to other tools. This should be mentioned very clearly.

Reviewer #3

(Remarks to the Author)
The authors compared seven developed sequencing methods from multiple perspectives, including accuracy, similarity, duplication rate, and computational resource consumption, to help researchers identify eccDNAs in future experiments.

The authors innovatively constructed a Python script to simulate possible circular and linear DNA in human sperm cells, EJM cell lines, and JJN3 cell lines after analyzing existing data. The eccDNA identified by actual sequencing methods was compared with the simulated dataset results, thereby demonstrating the advantages and disadvantages of different sequencing methods. However, eccDNA is widely present in various species in nature, and its presence has been confirmed in various diseases of human cells. Except for human sperm cells, the EJM cell line and JJN3 cell line used are closely related to hematological diseases. If they can be replaced with other types of tumor cells, it can improve the universality of the article.

The simulation of eccDNA in the article is generated through random movement, and the existing literature does not fully understand the randomness of eccDNA production. In addition, the existence of retrotransposon hijacking mechanisms in viral infection and tumor cells implies that eccDNA production is not completely random. These results seem to reduce the authenticity of simulated datasets.

Meanwhile, a significant portion of the existing understanding of eccDNA comes from existing sequencing methods, which are also included in the article. This will inevitably lead to the inclusion of the experimental methods to be identified in the simulation results when simulating the length distribution of eccDNA, chromosome origin, etc. in the simulation dataset. The widespread use of some sequencing methods will inevitably result in the evaluation results of the simulation dataset being biased towards these sequencing methods, and the authors did not seem to exclude this bias in the article.

In the discussion section, the authors mentioned that the digestion of linear DNA has a marginal effect. In the future, it can be

considered to skip the digestion of linear DNA and directly use solution A to purify RCA and circular DNA products. However, in current understanding, the digestion of linear DNA using DNA enzymes is a necessary step to obtain high-purity circular DNA, which is included in various sequencing methods and is crucial for obtaining accurate results of eccDNA. The results of the article does not seem to provide a detailed explanation of this marginal effect, nor does it provide relevant data to prove that skipping linear DNA digestion does not affect subsequent sequencing results.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The revised manuscript is improved substantially. Thank you go a good and comprehensive assessment.This will be a very good paper for the eccDNA research.

Reviewer #2

(Remarks to the Author)
The authors have considered the suggestions and significantly upgraded the manuscript.

Responses to reviewers

We thank the reviewers for their positive assessments and helpful suggestions regarding our study, "**Comparative analysis of methodologies for detecting extrachromosomal circular DNA**". Your feedback has provided us with valuable insights and opportunities to substantially improve the quality and scope of our study and manuscript. Following a short summary narrative, we provide our point-by-point responses below.

Briefly, we have undertaken the reviewers' suggestions and revised our manuscript as follows:

1. **Increase the Universality:** We included 4 additional published eccDNA sequencing datasets and processed them with diverse analysis pipelines to generate more representative simulated datasets (R3#1 and R3#3, revised **Figure 1**, revised **Supplementary Fig. 1-2**). We also incorporated the Circle_finder (microDNA.InOne.sh) mode in the comparison (R2#9 and R2#11, revised **Figure 1**, revised **Supplementary Fig. 3**).
2. **Strengthen the Fairness:** In addition to the 90% sequence identity threshold, we included a 250 bp threshold for true positive event selection to increase the selection power for long eccDNA and replaced the Similarity score with base pair difference to provide a straightforward assessment (R2#5, R2#6, and R2#7, revised **Figure 1** and **Supplementary Fig. 3**). We also excluded pipelines incapable of identifying chimeric eccDNA from the performance comparison for chimeric eccDNA identification (R2#1, revised **Figure 1** and **Supplementary Fig. 3**).
3. **Deepen the Investigation:** We added a section to show the heterogeneities of the eccDNA profiles obtained from different experimental methods, including the heterogeneity in the eccDNA sequence, enriched oncogenes and enriched repeat elements (R1#1, R1#2 and R1#3, added **Figure 4** and **Supplementary Fig. 7**).
4. **Enhanced Clarity and Reproducibility:** We refined figures to make them easy to read (R1#4 and R1#5, revised **Figure 1** and **Figure 2c**). We made a table to outline available pipelines and experimental methods in this field (R2#8, **Table 1**), which also clarifies our comparison scope for analysis pipelines (R2#1 and R2#3). We revised the methods section and updated the notebook on GitHub to increase the clarity and reproducibility of our study (R1#5, R1#6, R1#7, R2#10, and R3#2).
5. We have also responded to all other comments raised by reviewers in a point-by-point manner. Please find the details below.

We have taken the reviewers' suggestions seriously and devoted tremendous time and effort to revising this work. Numerous additional analyses were performed. 1 new main figure (**Figure 4**), 1 main table (**Table 1**), and 4 new supplementary figures

(**Supplementary Fig. 1**, **Supplementary Fig. 2**, **Supplementary Fig. 5**, and **Supplementary Fig. 7**) have been added to the revised manuscript. Moreover, 3 figures/ supplementary figures and 3 suplementary tables have been updated in the revised manuscript. We appreciate the reviewers' constructive feedback and are confident that our work will provide valuable insights and contributions to the field.

Reviewer #1

Remarks to the author:

Gao et al present a manuscript titled “Comparative analysis of methodologies for detecting extrachromosomal circular DNA” describe a comparison of different bioinformatics tools to identified extrachromosomal circular DNAs (eccDNAs) using simulated data, which has similar characteristic of some published eccDNA results as the gold standard. Based on the vary of sequencing depth of simulated datasets with chimeric sequences contaminations of both short and long reads, the authors reported that Circle-Map and CReSIL are the top performer for short and long read respectively, with a stable computational resource consumption. Then the authors further applied these two bioinformatic tools for further investigation of different 7 eccDNA enrichment methods. The authors compared the identified eccDNAs of the 7 enrichment methods with many aspects using both short and long read and conclude that Circle-Seq method with long reads give highest detection efficiency for longer eccDNA, while SEP with long reads is more effective for shorter eccDNA. The study is straight forward yet important, and the conclusion is valid and give good guidance for eccDNA research field. However, there are lacking comparisons that will improve the quality of the manuscript and missing aspect that need to be reported.

Response: We appreciate the reviewer’s positive assessments and constructive comments. We have carefully revised our manuscript based on the reviewer’s suggestions.

1. Couple published papers show that there is extremely high heterogeneity in eccDNA profile, leading to a difficult to detect the high percentage of common eccDNAs across different technical and biological replicates. I know that; this point will create many questions on study eccDNA as low commonality will present across replicates. The study pooled the DNA extracted from HeLa cells and aliquoted for the 7 different enrichment methods, the authors can compare the percentage of commonly detect eccDNA across different technical replicates of individual enrichment methods and across different enrichment methods. I know that it will be LOW (I think the authors did that already), don’t be afraid to put the results because it is important to the filed to know. It will be the research area that need the field to pay attention and possible develop a better strategy to study eccDNA together. I highly recommend the author to

work on the results and present in the manuscript. Figure 4 is still missing. Adding a section around the results would be highly appropriate.

Response: We thank the reviewer for the insightful suggestion. We have incorporated a new section and added **Figure 4** and **Supplementary Fig. 7** to our revised manuscript to present a comprehensive analysis of the heterogeneity of eccDNA profiles obtained from different experimental methods and technical replicates. This new addition includes the following aspects:

(1) **eccDNA sequence heterogeneity**: We assessed sequence correlation of eccDNA profiles across various experimental methods. Our results showed that highly-correlated eccDNA (HC eccDNA, over 90% sequence identity) was observed more frequently in WGS-SR (>50%) and WGS-LR (>54.19%) compared to other experimental methods, such as ATAC-Seq-SR (<10%), 3SEP-SR (<2%), 3SEP-LR (<1%), Circle-Seq-SR (<1%), and Circle-Seq-LR (<1%)). This suggests a significant heterogeneity in detected eccDNA sequence across experimental methods and technical replicates (revised Figure 4a and 4b).

(2) **Genomic feature heterogeneity**: Despite the overall heterogeneity of eccDNA profiles obtained from various experimental methods, we suspected that different experimental methods might share common genomic features, including common oncogenes in HeLa cell line and common repeat elements detected in eccDNA. By cross-referencing with the OnGene database *(*Liu, Sun and Zhao, *Journal of genetics and genomics*, 2017), we mapped ecDNA (copy number amplified eccDNA) sequences to 125 distinct oncogenes (revised Supplementary Fig. 7). The ecDNA detected from WGS-SR and ATAC-Seq-SR did not harbor any oncogenes in our sequencing data. Although no oncogene was universally detected across all experimental methods, 87 out of 125 oncogenes were present in ecDNA from at least two different experimental methods, highlighting the potential commonality in oncogene presence despite methodological differences (revised Figure 4c and 4d). Notably, several oncogenes known to influence HeLa cell viability and metastasis (*e.g.*, *ZNF217* (Thollet, Vendrell et al., *Mol Cancer*, 2010)*, TRIO* (Hou, Zhuang et al., *Oncol Rep*, 2018), *and CULA4* (Mei, Ye et al., *Eur Rev Med Pharmacol Sci*, 2020)) were detected by 4 experimental methods, primarily in eccDNA-enriched experimental methods. We also examined the proportion of reads mapped to various repeat elements (*e.g.*, LTRs, SINEs, LINEs, and satellite elements), which are commonly detected in eccDNA. The analysis revealed significant method-dependent disparities in read distribution among repeat categories. WGS-LR showed the highest proportions of reads mapped to all examined repeat elements (revised Figure 4e). The eccDNA profiles generated by long-read based experimental methods showed a higher proportion of reads mapped to repeat elements compared to their short-read

counterparts. This demonstrates the heterogeneity of genomic features within the eccDNA profiles generated by different experimental methods.

The analysis results were added to the **Results** section of the revised manuscript (from line 251 to 304, highlighted in yellow), which reads as follows:

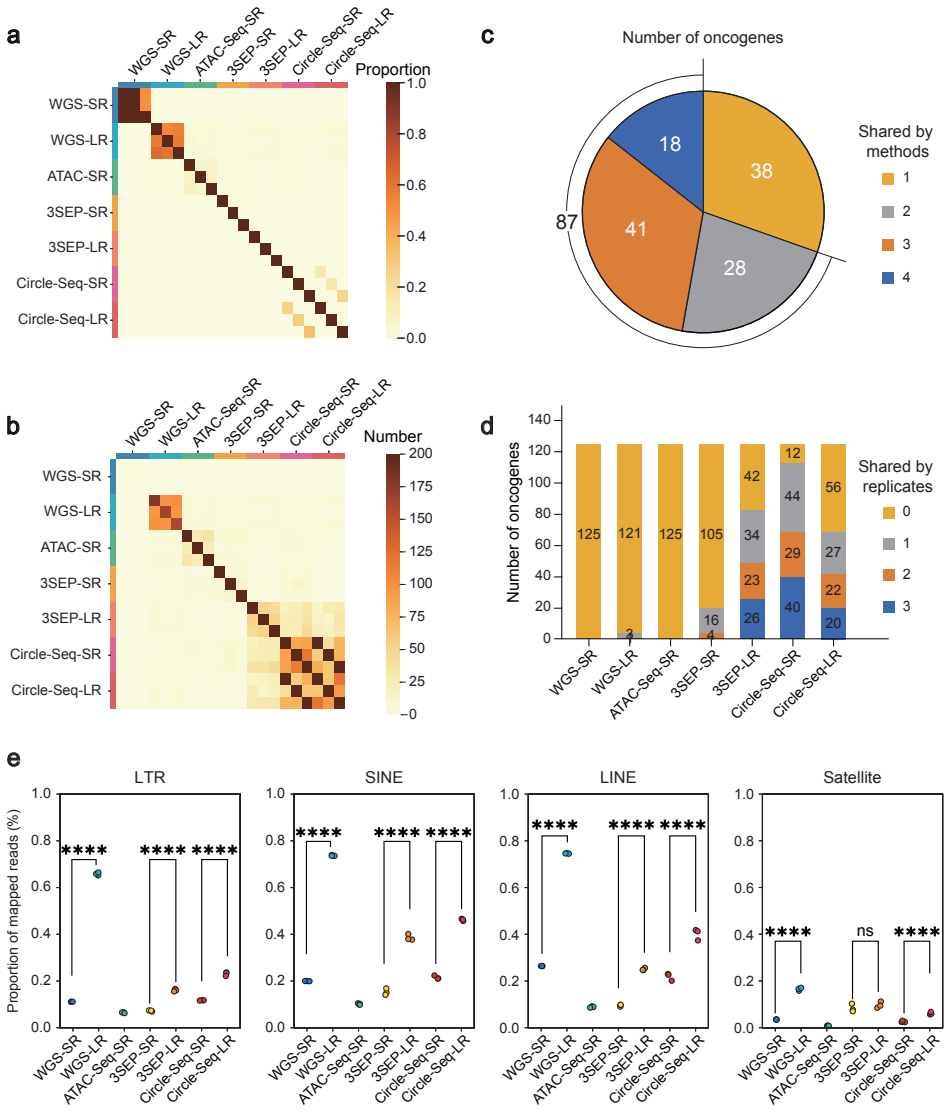
"EccDNA profiles showed heterogeneity across experimental methods

We investigated the correlation of eccDNA profiles across various technical replicates from different experimental methods. We considered eccDNA from different technical replicates to be highly-correlated (HC) when their eccDNA shared over 90% sequence identity. Our analysis indicated that, among the methods examined, eccDNA profiles detected from different experimental replicates within WGS-SR (>50%) or WGS-LR (>54.19%) exhibited higher correlations compared to other methods such as ATAC-Seq-SR (<10%), 3SEP-SR (<2%), 3SEP-LR (<1%), Circle-Seq-SR (<1%), and Circle-Seq-LR (<1%) (Figure 4a and Supplementary Table 4). Specifically, compared to WGS-SR (≤2), WGS-LR replicates showed more shared highly-correlated eccDNA (>95) (Figure 4b). Furthermore, we observed higher correlations between paired Circle-Seq-SR/LR replicates (HC eccDNA proportion >15%, number of HC eccDNA >4900) (e.g., Circle-Seq-SR1/Circle-Seq-LR1) compared to unpaired Circle-Seq replicates (e.g., Circle-Seq-SR1/Circle-Seq-SR2 or Circle-Seq-SR1/Circle-Seq-LR2) (Figure 4a and 4b). Despite the DNA material for Circle-Seq-SR/LR pairs originating from the same debranched RCA product, the HC eccDNA proportions within these pairs were below 35% (Supplementary Table 4), indicating that the choice of sequencing platform and analysis pipelines can influence the final eccDNA profiles.

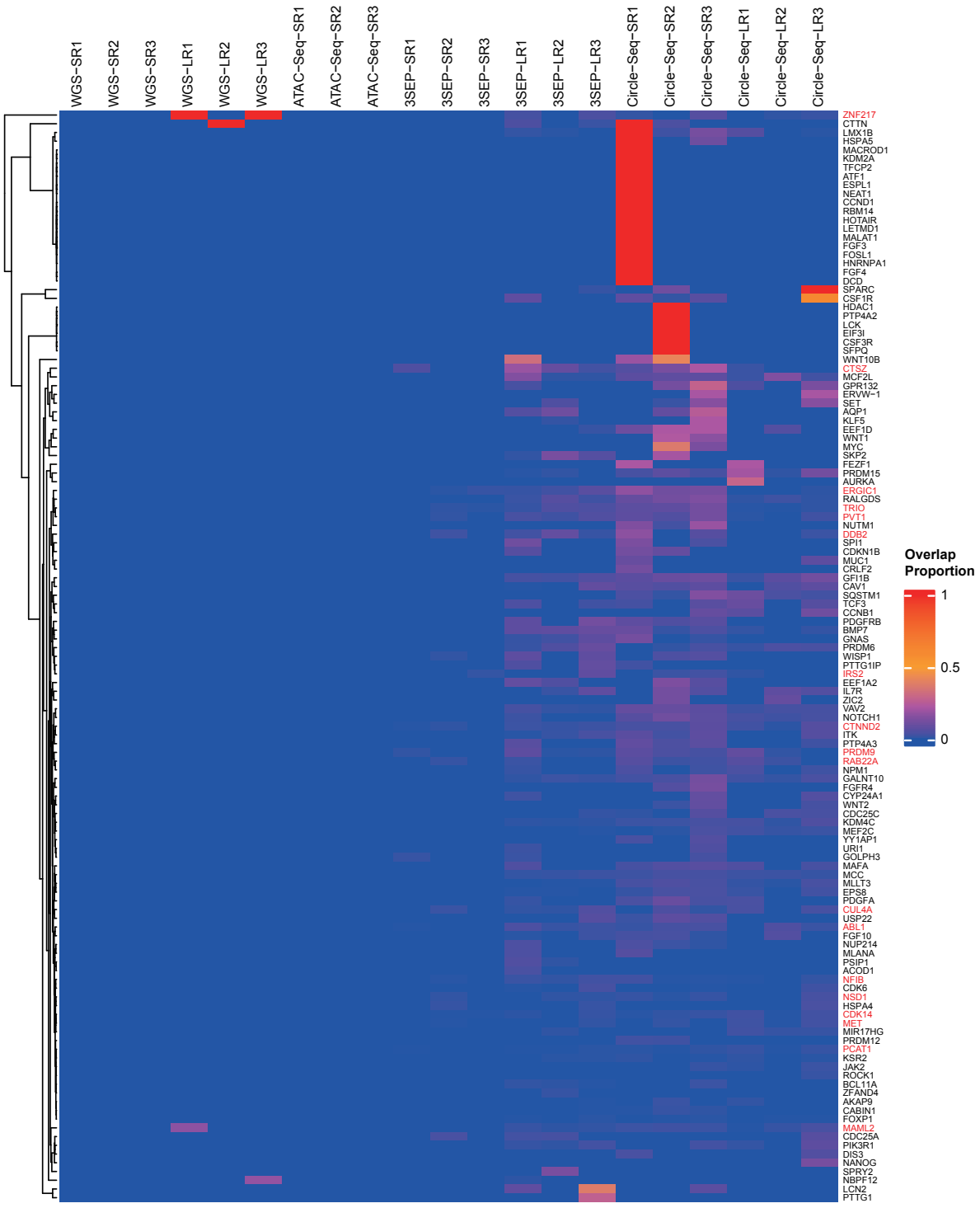
We speculated that though there existed high heterogeneity across different experimental methods or technical replicates, the copy number amplified eccDNA profiles (ecDNA profiles) of different methods might share common oncogenes. To explore this, we compiled a list of oncogenes from OnGene database[1], and compared the detected oncogenes by different methods. Our analysis revealed that the ecDNA sequences obtained from all examined experimental methods mapped to a total of 125 oncogenes (Supplementary Fig. 7 and Supplementary Table 4). No oncogene was detected by all the examined experimental methods. 87 out of 125 oncogenes could be detected by at least two different experimental methods (Figure 4c and Supplementary Table 4). A total of 18 oncogenes were detected by 4 experimental methods (Figure 4c and Supplementary Table 4). For example, ZNF217, reported to promote HeLa cell viability[2], was detected by Circle-Seq-SR, Circle-Seq-LR, 3SEP-LR, WGS-LR. Further, 16 of the 18 oncogenes were detected by eccDNA-enriched experimental methods. For example, TRIO[3] and CULA4[4], reported to promote metastasis and invasion of HeLa cells, were detected by Circle-Seq-SR/LR and 3SEP-SR/LR. PVT1, a long non-coding RNA that can enhance proliferation[5] and promote the cancer progress[6,7] of cervical cancer cells, was also detected by Circle-Seq-SR/LR and 3SEP-SR/LR. Notably, experimental methods employing the RCA step demonstrated a higher capacity for oncogene detection (>69, data from Circle-Seq-LR, Figure

4d) compared to those lacking this step (<20, data from 3SEP-SR, Figure 4d). Furthermore, experimental replicates utilizing RCA exhibited a greater overlap in detected oncogenes (at least 42 oncogenes were detected in more than 2 replicates) compared to those without RCA (Figure 4d and Supplementary Table 4).

Repeat elements are commonly detected in the sequencing data that are used for identifying eccDNA[8-10]. Considering that our sequencing data originated from the same DNA pool, we postulated that the proportion of reads mapping to repeat elements would remain consistent across different experimental methods. However, our findings revealed notable disparities. WGS-LR exhibited the highest proportion of reads mapping to the examined repeat elements (Figure 4e and Supplementary Table 4), including long terminal repeats (LTRs, 65.84%), short interspersed nuclear elements (SINEs, 73.70%), long interspersed nuclear elements (LINEs, 74.61%), and satellite elements (16.5%). Furthermore, WGS-LR, 3SEP-LR, and Circle-Seq-LR displayed significantly elevated proportions of reads mapping to LTRs, SINEs, and LINEs compared to their short-read counterparts (Figure 4e and Supplementary Table 4). This suggests that sequencing results from different experimental methods inherently exhibit heterogeneity. Consequently, when comparing results across different studies, it is important to consider the experimental methods used. ”



Revised Figure 4. eccDNA profile heterogeneity across experimental methods. a. Proportion of highly-correlated eccDNA between each pair experimental replicates in vertical replicates; **b.** Number of highly-correlated eccDNA between each pair of experimental replicates; **c.** Number of detected oncogenes shared across different methods. **d.** Number of oncogenes shared by different replicates. **e.** Proportion of read mapped to repeat elements, (n=3). Statistical analyses were performed using one-way ANOVA with Tukey correction; ns, not significant, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, and $****p < 0.0001$.



Revised Supplementary Fig. 7. Enriched oncogene of each replicate. Genes, marked in red, were detected by 4 experimental methods.

2. The ecDNA is an important molecule esp., for cancer biology. The authors need to investigate and present more results. For example, the association of copy number

identified from WGS-LR data with the sequencing depth of the overlapped ecDNA. What's known oncogenes of HeLa cell are harbored by ecDNAs or eccDNA.

Response: We thank the reviewer for providing this insightful suggestion, which has indeed enhanced the depth of our study. In response, we have expanded the **Results** section of our revised manuscript to include detailed analyses regarding (1) the correlation between copy number identified from WGS-LR data and the sequencing depth of the overlapped ecDNA, and (2) the known oncogenes in HeLa cell that were harbored by ecDNA.

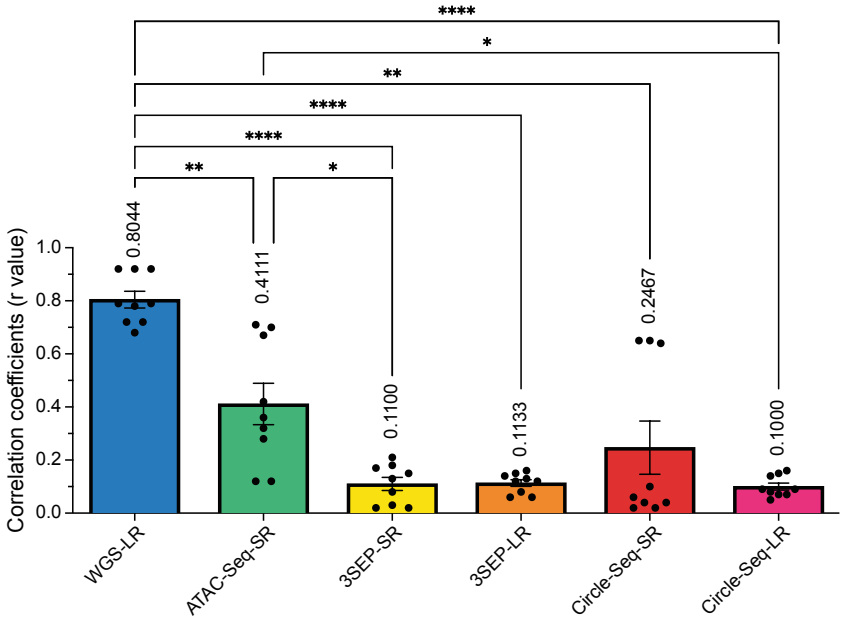
(1) **Correlation between copy number and eccDNA sequencing depth**: We applied Control-FREEC (Boeva, Popova et al., *Bioinformatics*, 2012) to obtain the copy number from WGS-LR data and employed BEDTool Coverage (Quinlan and Hall, *Bioinformatics*, 2010) to calculate the sequencing depth of eccDNA. Our analysis revealed a notable positive correlation in WGS-LR data, indicating that higher copy numbers are generally associated with increased eccDNA detection depth (revised Supplementary Fig. 5). However, in eccDNA-enriched experimental methods, particularly those utilizing rolling circle amplification (RCA), the correlation was low ($r < 0.25$, revised Supplementary Fig. 5). This indicates that RCA-based experimental methods can obtain the eccDNA profile with high coverage regardless of their copy number amplification status. However, this introduces a level of heterogeneity due to the presence of unamplified eccDNAs across different experimental replicates. We have incorporated associated descriptions in the **Results** section of our revised manuscript (lines 209 to 217, highlighted in yellow), which reads as follows:

"Due to the notable efficiency of circular DNA enrichment through RCA and the use of solution A, we hypothesized that eccDNA-enriched experimental methods could effectively detect such DNA entities without the need for copy number amplification. We investigated the association between genome copy numbers and the sequencing depth of overlapped eccDNA. Our analysis revealed a positive correlation between genome copy numbers and the sequencing depth of overlapped eccDNA. Notably, the correlation coefficients (r values) derived from the WGS-LR (0.80) and ATAC-Seq-SR (0.41) datasets were higher than those obtained from eccDNA-enriched experimental methods (<0.25) (Supplementary Fig. 5)."

(2) **Known oncogenes harbored by ecDNA**: Further, we have conducted a comprehensive analysis to identify oncogenes within detected ecDNA. 87 out of 125 oncogenes were detected by at least two different experimental methods, highlighting the potential commonality in oncogene presence despite heterogeneity (revised Figure 4c and 4d, see above R1#1). Additionally, several oncogenes known to influence HeLa cell viability and metastasis (*e.g.*, *ZNF217* (Thollet, Vendrell et al., *Mol Cancer*, 2010), *TRIO* (Hou, Zhuang et al., *Oncol Rep*, 2018), *and CULA4* (Mei, Ye et al., *Eur Rev Med Pharmacol Sci*, 2020)) were detected by 4 experimental methods but

mainly harbored by ecDNA detected from eccDNA-enriched experimental methods (Supplementary Fig. 7, see above R1#1). These results highlight the method-dependent heterogeneity in eccDNA profiles. We have incorporated associated descriptions in the **Results** section of our revised manuscript (lines 269 to 290, highlighted in yellow), which reads as follows:

"*We speculated that though there existed high heterogeneity across different experimental methods or technical replicates, the copy number amplified eccDNA profiles (ecDNA profiles) of different methods might share common oncogenes. To explore this, we compiled a list of oncogenes from OnGene database*[1]*, and compared the detected oncogenes by different methods. Our analysis revealed that the ecDNA sequences obtained from all examined experimental methods mapped to a total of 125 oncogenes (Supplementary Fig. 7 and Supplementary Table 4). No oncogene was detected by all the examined experimental methods. 87 out of 125 oncogenes could be detected by at least two different experimental methods (Figure 4c and Supplementary Table 4). A total of 18 oncogenes were detected by 4 experimental methods (Figure 4c and Supplementary Table 4). For example, ZNF217, reported to promote HeLa cell viability*[2]*, was detected by Circle-Seq-SR, Circle-Seq-LR, 3SEP-LR, WGS-LR. Further, 16 of the 18 oncogenes were detected by eccDNA-enriched experimental methods. For example, TRIO*[3] *and CULA4*[4]*, reported to promote metastasis and invasion of HeLa cells, were detected by Circle-Seq-SR/LR and 3SEP-SR/LR. PVT1, a long non-coding RNA that can enhance proliferation*[5] *and promote the cancer progress*[6,7] *of cervical cancer cells, was also detected by Circle-Seq-SR/LR and 3SEP-SR/LR. Notably, experimental methods employing the RCA step demonstrated a higher capacity for oncogene detection (>69, data from Circle-Seq-LR, Figure 4d) compared to those lacking this step (<20, data from 3SEP-SR, Figure 4d). Furthermore, experimental replicates utilizing RCA exhibited a greater overlap in detected oncogenes (at least 42 oncogenes were detected in more than 2 replicates) compared to those without RCA (Figure 4d and Supplementary Table 4).*"



Revised Supplementary Fig. 5. Correlation coefficients (r value) between copy number and sequencing depth of eccDNA. Copy number data were obtained from 3 WGS-LR replicates. The correlation between the eccDNA coverage from each replicate and genome copy number from each WGS-LR replicate was analyzed, resulting in a total of 9 data points for each experimental method.

3. Some basic statistic of identified eccDNAs that associated with the known genomic features should be reported.

Response: Thank you for your suggestion. We have now added content to show the statistics results of identified eccDNA with associated genomic features. Specifically:

(1) The ecDNA is related to oncogene amplification, so we analyzed the oncogenes harbored by ecDNA detected from different experimental methods.

(2) The repeat elements, especially the **LTR** (Yang, Su et al., *Nature*, 2023), **SINE** (Møller, Mohiyuddin et al., *Nature Communications*, 2018), **LINE** (Dillon, Kumar et al., *Cell reports*, 2015) and **Satellite** (Cohen, Agmon et al., *Mob DNA*, 2010, Ling, Han et al., *Mol Cancer*, 2021) were commonly detected in the sequencing data that are used for identifying eccDNA (Sun, Lu and Zou, *Computational and Structural Biotechnology Journal*, 2023). Therefore, we anlayzed the proportion of reads mapped to these repeat elements.

This revison demonstrated the impact of the selection of experimental methods on the statistics of specific genomic features, and also emphasized the importance of verifying the experimental methods used in different studies when comparing their statistics data on genomic features.

(1) **Known oncogenes harbored by ecDNA**: We analyzed the oncogenes harbored by ecDNA and found that 87 oncogenes were harbored by ecDNA from at least two different experimental methods (revised Figure 4c and 4d, see above R1#1). Additionally, several oncogenes known to influence HeLa cell viability and metastasis (*e.g.*, *ZNF217* (Thollet, Vendrell et al., *Mol Cancer*, 2010), *TRIO* (Hou, Zhuang et al., *Oncol Rep*, 2018), *and CULA4* (Mei, Ye et al., *Eur Rev Med Pharmacol Sci*, 2020)) were detected by 4 experimental methods (Supplementary Fig. 6, see above). These results showed the heterogeneity of oncogene detection across different experimental methods.

(2) **Proportion of reads mapped to repeat elements (LTRs, SINEs, LINEs, and satellite elements):** The results showed that WGS-LR had the highest proportions of read mapped to all examined repeat elements (revised Figure 4e, see above R1#1). Long-read-based experimental methods showed higher proportion of reads mapped to repeat elements compared to their short-read counterparts. This indicates that sequencing results from different experimental methods inherently exhibit heterogeneity. Therefore, when comparing results across different studies, especially regarding the proportions of different genomic features, it is important to consider the experimental methods used.

We have incorporated associated descriptions in the **Results** section (from line 269 to 304, highlighted in yellow), which reads as follows:

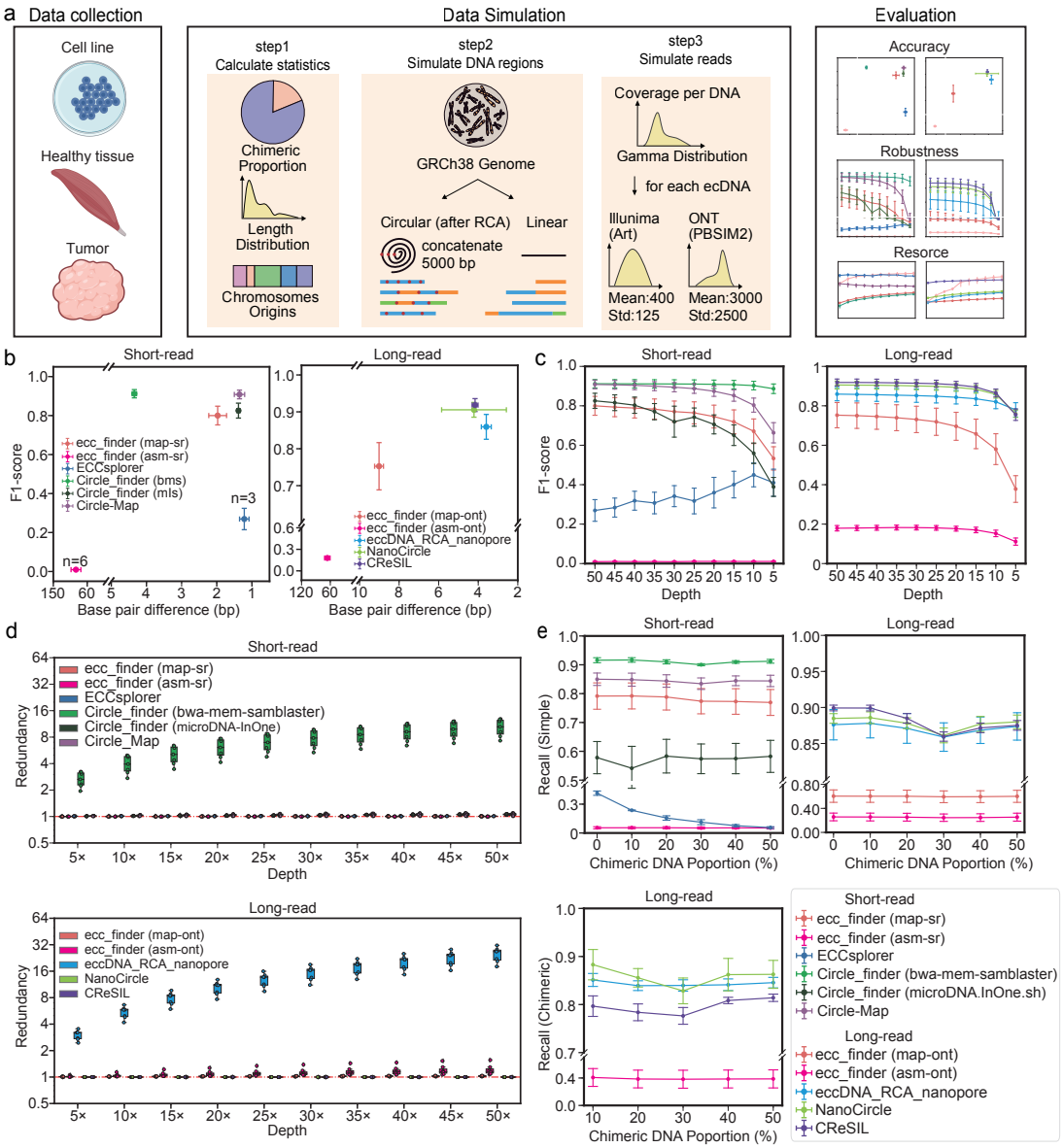
"*We speculated that though there existed high heterogeneity across different experimental methods or technical replicates, the copy number amplified eccDNA profiles (ecDNA profiles) of different methods might share common oncogenes. To explore this, we compiled a list of oncogenes from OnGene database[1], and compared the detected oncogenes by different methods. Our analysis revealed that the ecDNA sequences obtained from all examined experimental methods mapped to a total of 125 oncogenes (Supplementary Fig. 7 and Supplementary Table 4). No oncogene was detected by all the examined experimental methods. 87 out of 125 oncogenes could be detected by at least two different experimental methods (Figure 4c and Supplementary Table 4). A total of 18 oncogenes were detected by 4 experimental methods (Figure 4c and Supplementary Table 4). For example, ZNF217, reported to promote HeLa cell viability[2], was detected by Circle-Seq-SR, Circle-Seq-LR, 3SEP-LR, WGS-LR. Further, 16 of the 18 oncogenes were detected by eccDNA-enriched experimental methods. For example, TRIO[3] and CULA4[4], reported to promote metastasis and invasion of HeLa cells, were detected by Circle-Seq-SR/LR and 3SEP-SR/LR. PVT1, a long non-coding RNA that can enhance proliferation[5] and promote the cancer progress[6,7] of cervical cancer cells, was also detected by Circle-Seq-SR/LR and 3SEP-SR/LR. Notably, experimental methods employing the RCA step demonstrated a higher capacity for oncogene detection (>69, data from Circle-Seq-LR, Figure 4d) compared to those lacking this step (<20, data from 3SEP-SR, Figure 4d). Furthermore, experimental replicates utilizing RCA exhibited a greater overlap in detected oncogenes (at*

least 42 oncogenes were detected in more than 2 replicates) compared to those without RCA (Figure 4d and Supplementary Table 4).

Repeat elements are commonly detected in the sequencing data that are used for identifying eccDNA[8-10]. Considering that our sequencing data originated from the same DNA pool, we postulated that the proportion of reads mapping to repeat elements would remain consistent across different experimental methods. However, our findings revealed notable disparities. WGS-LR exhibited the highest proportion of reads mapping to the examined repeat elements (Figure 4e and Supplementary Table 4), including long terminal repeats (LTRs, 65.84%), short interspersed nuclear elements (SINEs, 73.70%), long interspersed nuclear elements (LINEs, 74.61%), and satellite elements (16.5%). Furthermore, WGS-LR, 3SEP-LR, and Circle-Seq-LR displayed significantly elevated proportions of reads mapping to LTRs, SINEs, and LINEs compared to their short-read counterparts (Figure 4e and Supplementary Table 4). This suggests that sequencing results from different experimental methods inherently exhibit heterogeneity. Consequently, when comparing results across different studies, it is important to consider the experimental methods used."

4. Fig. 1 especially panel C and is very hard to read, all things are squeezed. Please improve.

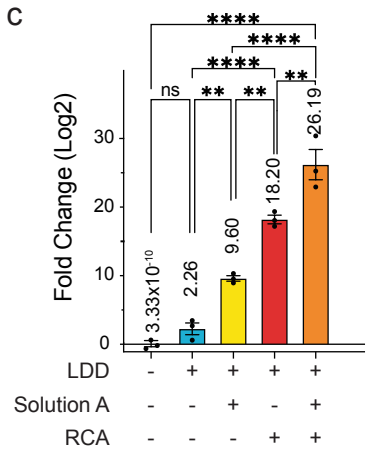
Response: Thank you for your careful examination and useful suggestion. We divided the axes into two segments, creating a more dispersed distribution of squeezed data points in our figures. Consequently, we have revised Figure 1 to make the figure easier to read.



Revised Figure 1. Assessment of analysis pipelines in eccDNA identification. a. Schematic overview of the benchmarking workflow used to compare the performance of bioinformatic pipelines. **b.** Performance comparison of analysis pipelines at a simulated sequencing depth of 50X (bms, bwa-mem-samblaster; mls, microDNA.InOne.sh). **c.** Impact of simulated sequencing depth on eccDNA identification accuracy. **d.** Impact of simulated sequencing depth on eccDNA identification duplication rates. **e.** Impact of chimeric DNA proportion on eccDNA identification recall. For panels **b, c** and **d**, n=7, except for ecc_finder (asm-sr) with n=6 and ECCsplorer with n=6 when depth ≤10, n=5 for depths between 15X and 25X, n=4 for depths between 30X and 40X, and n=3 for depths higher than 40X. For panel **e**, n=4.

5. What the is x axis of the Fig 2C ?. It is not mentioned on figure legend.

Response: Thank you for your comment. The X-axis of the Figure 2c represents the experimental steps. Specifically, "original" indicates the extracted DNA material containing the spiking-in of both pUC19 and *egfr* fragment, the "D" represents the linear DNA digestion, "S" denotes the application of Solutation A to purify circular DNA, and "R" signifies the application of rolling circle amplification (RCA). We realized the original Figure 2c was difficult to read, so we revised both the figure and its legend to enhance clarity and readability. The revised figure is shown below:



Revised Figure 2. Impact of eccDNA enrichment operations on eccDNA identification. c. Circular DNA enrichment efficiency. LDD, Linear DNA Digestion; Solution A: using Solution A for circular DNA purification; RCA, Rolling Cycle Amplification. n = 3 for all experiments. Statistical analyses were performed using one-way ANOVA with Tukey correction; error bars represent SEM. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, and $****p < 0.0001$.

6. Notebook for compare statistics between given data and simulated data on github is missing.

Response: Thank you for your kind reminder. The notebook for comparing statistics between given data and simulated data has been updated and archived in https://github.com/QuKunLab/eccDNABenchmarking.

7. Remarks on code availability: Notebook for compare statistics between given data and simulated data on github is missing.

Response: Thank you for suggestions. The notebook for comparing statistics between given data and simulated data has been updated and archived in https://github.com/QuKunLab/eccDNABenchmarking.

Reviewer #2

Remarks to the author:

This is a timely benchmarking paper on the eccDNA method. However, there is a lot more that needs to be done to make a fair comparison between various tools. Below are some suggestions that should be considered to improve the benchmarking process.

Response: We would like to express our gratitude to the reviewer for taking the time to review our manuscript, and appreciate your valuable feedback and insightful comments, which have helped us to improve the clarity and impact of our work. In this response letter, we address the concerns and questions raised by the reviewer.

1. Define the problem clearly: Are you benchmarking the tools that identify simple eccDNA or complex eccDNA? Simple eccDNA refers to circular DNAs generated by end-to-end ligation of one linear fragment. On the other hand, complex eccDNA is composed of two or more linear fragments and may be considered chimeric. Some eccDNA tools claim to identify only simple eccDNA, and therefore should not be evaluated for their ability to identify chimeric eccDNA.

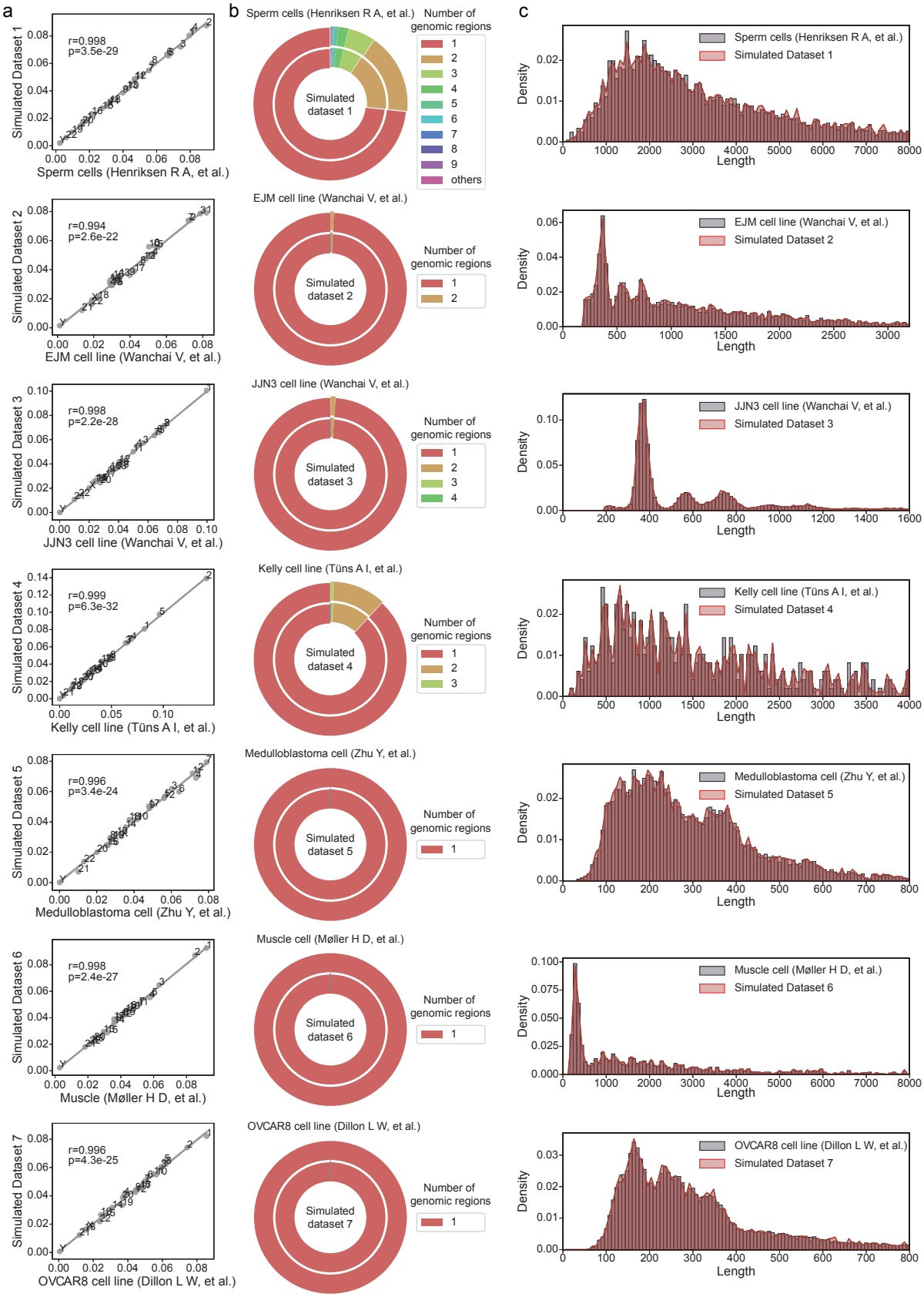
Response:

We thank the reviewer for emphasizing the need to clarify the scope of our eccDNA tool benchmarking regarding simple or complex eccDNA. In our study, we have specifically focused on evaluating analysis pipelines capable of identifying eccDNA from sequencing data derived from eccDNA-enriched experimental methods. Our primary aim was not to distinguish between analysis pipelines based on their abilities to identify simple versus complex eccDNA. Given the presence of both simple eccDNA and chimeric eccDNA in the real datasets and possible influence of chimeric eccDNA sequence on the performance of the analysis pipeline, we generated simulated datasets with different proportions of chimeric eccDNA to examine the performance of different analysis pipelines (revised Supplementary Fig. 2).

We agree that some eccDNA analysis pipelines can not identify chimeric eccDNA. Therefore, in the revised manuscript, we examined the performance of the analysis pipelines in chimeric eccDNA identification only for those pipelines that are capable for identifying chimeric eccDNA, including CReSIL, NanoCircle, eccDNA_RCA_nanopore, and ecc_finder (asm-ont) (revised Figure 1e). While ecc_finder (asm-sr) can identify chimeric eccDNA from short-read data, no other short-read-based analysis pipeline is capable for identifying chimeric eccDNA, so no comparison can be performed. Although AmpliconArchitect can also identify chimeric eccDNA, it is designed for WGS-SR which is beyond the scope of our comparison.

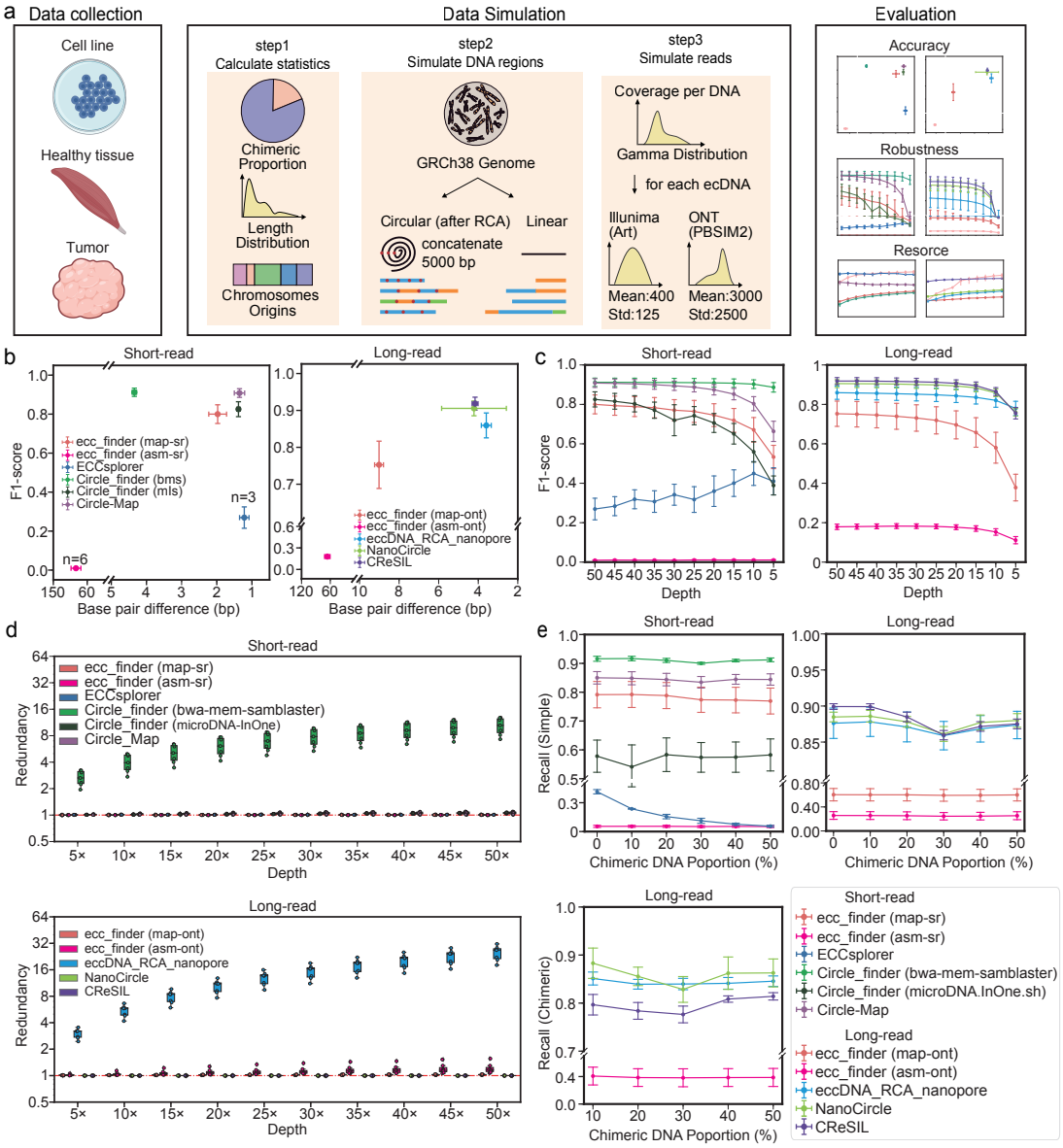
The revised content can be found in the **Results** section (from line 160 to 173, highlighted in yellow), and reads as follows:

"For short-read data analysis, the change of chimeric DNA proportion did not affect the recall for simple eccDNA identification of Circle-finder (bwa-mem-samblaster), Circle-Map, and ecc_finder (map-sr). However, the recall for simple eccDNA identification of ECCsplorer decreased from 0.414 at 0% to 0.056 at 50%. (Figure 1e). ecc_finder (asm-sr) showed the lowest recall. The base pair differences between identified eccDNA and simulated eccDNA for these pipelines remained relatively stable except ecc_finder (asm-sr) (Supplementary Fig. 3c). Among long-read data analysis pipelines, most maintained consistent recall (with changes less than 0.1) for both simple eccDNA and chimeric eccDNA identification (Figure 1e). The base pair differences of CReSIL, eccDNA_RCA_nanopore and ecc_finder (map-ont) showed a relatively slight increase compared to NanoCircle, of which the base pair difference increased from 2.492 at 0% to 16.899 at 50% (Supplementary Fig. 3d). Unlike the other pipelines, ecc_finder (asm-ont) showed a decreased base pair difference as the chimeric eccDNA proportion increased (Supplementary Fig. 3d)."

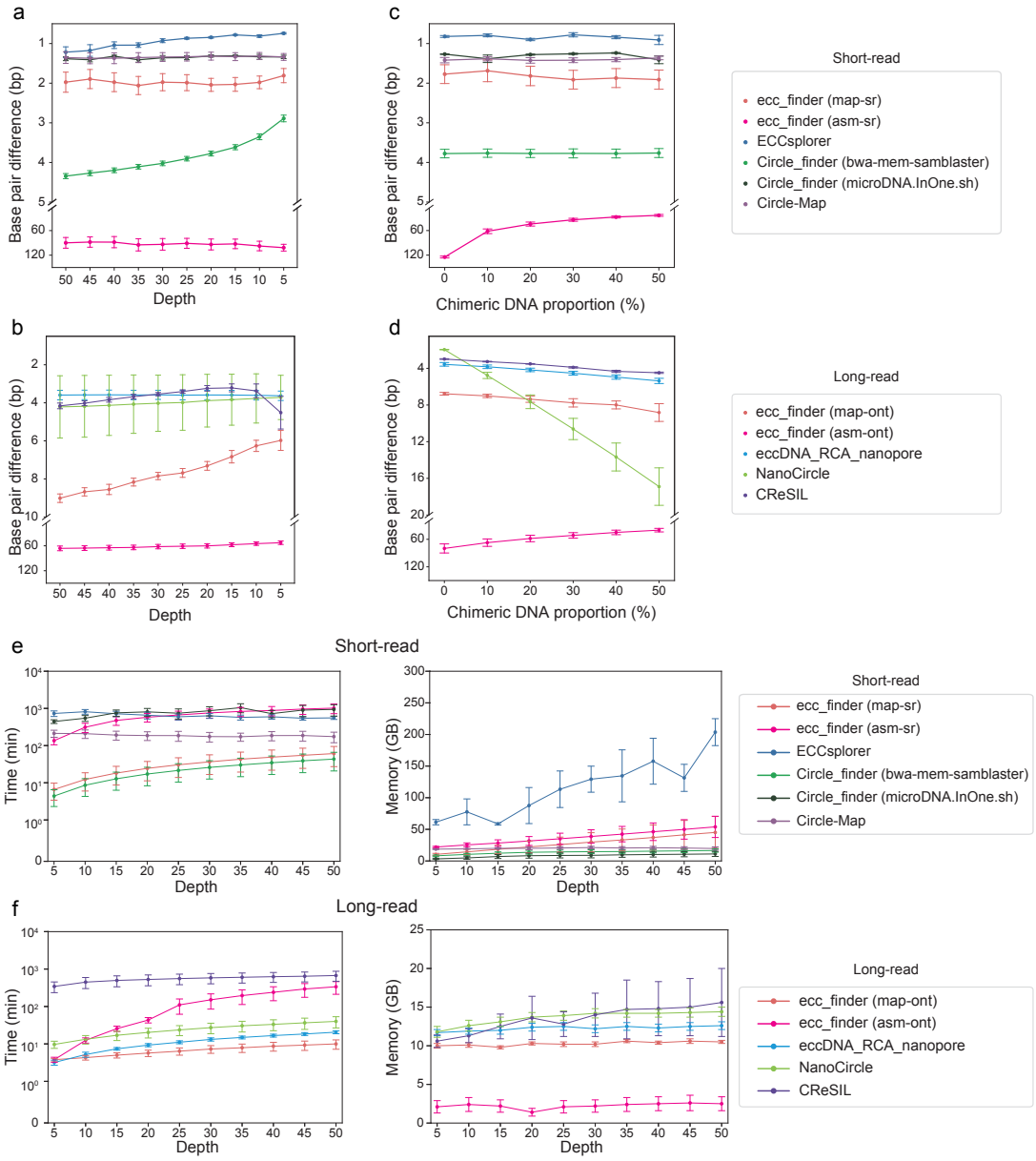


Revised Supplementary Fig. 2. eccDNA profiles of different simulated datasets. a. Chromatin distribution comparison between the simulated dataset and corresponding real dataset. **b**.

Chimeric eccDNA composition of the simulated dataset and corresponding real dataset. **c**. Length distribution comparison between the simulated dataset and corresponding real dataset.



Revised Figure 1. Assessment of analysis pipelines in eccDNA identification. a. Schematic overview of the benchmarking workflow used to compare the performance of bioinformatic pipelines. **b.** Performance comparison of analysis pipelines at a simulated sequencing depth of 50X (bms, bwa-mem-samblaster; mls, microDNA.InOne.sh). **c.** Impact of simulated sequencing depth on eccDNA identification accuracy. **d.** Impact of simulated sequencing depth on eccDNA identification duplication rates. **e.** Impact of chimeric DNA proportion on eccDNA identification recall. For panels **b, c** and **d**, n=7, except for ecc_finder (asm-sr) with n=6 and ECCsplorer with n=6 when depth ≤10, n=5 for depths between 15X and 25X, n=4 for depths between 30X and 40X, and n=3 for depths higher than 40X. For panel **e**, n=4.



Revised Supplementary Fig. 3. **Base pair difference and computational resources cost of different analysis pipelines. a.** Base pair difference of short-read data analysis pipelines under different sequencing depths. **b.** Base pair difference of long-read data analysis pipelines under different sequencing depths. **c.** Base pair difference of short-read data analysis pipelines under different chimeric DNA proportions. **d.** Base pair difference of short-read data analysis pipelines under different chimeric DNA proportions. **e.** Time and computer memory used by different pipelines in simulated short-read data analysis under different simulated sequencing depths. **f.** Time and computer memory used by different pipelines in simulated long-read data analysis under different simulated sequencing depths. For panels **a**, **b**, **e**, **f**, n=7, except for ecc_finder (asm-sr) with n=6 and ECCsplorer with n=6 when depth ⩽10, n=5 for depths between 15X and 25X, n=4 for depths between 30X and 40X, and n=3 for depths higher than 40X. For panels **c** and **d**, n=4.

2. Parameter and final output filtering: This text "Realign to identify eccDNA and used recommended filters (circle score > 50, split reads > 2, discordant reads > 2, coverage

increase in the start coordinate > 0.33 and coverage increase in the end coordinate > 0.33)." is from the manuscript. The default parameters for Circle-Map seem to differ from those used for other tools. The authors should make an effort to optimize parameters for other tools as well if they choose to optimize specifically for Circle-Map.

Response: We appreciate the reviewer's attention to this matter. In our study, we used the default settings of all pipelines, including Circle-Map, to identify eccDNA. We did not optimize parameters for any pipeline. The parameters we listed in the paper indicate the default settings we used.

3. Why AmpliconArchitect is not part of Figure 1?

Response: Thank you for your question. The scope of our comparative analysis for bioinformatics analysis pipelines is to compare the performance of those pipelines that can identify eccDNA from the sequencing data generated by eccDNA-enriched experimental methods (revised Table 1). However, AmpliconArchitect is designed for analyzing short-read WGS (WGS-SR) data, and the WGS-SR does not include the eccDNA enrichment procedure, so AmpliconArchitect is out of scope.

Revised Table 1. Summary of eccDNA analysis pipelines and supported experimental methods.

Pipeline (mode)	Published date	Supported experimental methods	Can identify chimeric eccDNA	eccDNA enrichment status	Read format
AmpliconArchitect	2019	WGS-SR	Y	Non-enriched	Short-read
Circle-Map	2019	Circle-Seq-SR 3SEP-SR	N	Enriched	Short-read
Circle_finder (bwa-mem-samblaster, microDNA.InOne.sh)	2020	ATAC-Seq-SR Circle-Seq-SR 3SEP-SR	N	Both	Short-read
ECCsplorer	2022	Circle-Seq-SR 3SEP-SR	N	Enriched	Short-read
ecc_finder (asm, map)	2021	Circle-Seq-SR 3SEP-SR Circle-Seq-LR 3SEP-LR	Y	Enriched	Short-read Long-read
eccDNA_RCA_nanopore	2021	Circle-Seq-LR 3SEP-LR	Y	Enriched	Long-read
NanoCircle	2022	Circle-Seq-LR 3SEP-LR	Y	Enriched	Long-read
CReSIL	2023	WGS-LR Circle-Seq-LR 3SEP-LR	Y	Both	Long-read

4. The Circle-Map outperforms other compared tools when the coverage exceeds 35X. This should be included in the abstract.

Response: We thank the reviewer for pointing out this issue. After adding more datasets as suggested by Reviewer 3, we found that Circle-Map and Circle_finder (bwa-mem-samblaster) outperformed other analysis pipelines across all the sequencing depths (even those deeper than 35X, revised Figure 1b, and 1c, see above R2#1). However, Circle_finder (bwa-mem-samblaster) displayed high redundancy in its results (revised Figure 1d, see above R2#1). The revised content can be found in the **Abstract** section (from line 24 to 27, highlighted in yellow) and reads as follows:

"*Here we show that Circle-Map and Circle_finder (bwa-mem-samblaster) outperform the other short-read pipelines. However, Circle_finder (bwa-mem-samblaster) exhibits notable redundancy in its outcomes. CReSIL is the most effective pipeline for eccDNA detection in long-read sequencing data at depths higher than 10X.*"

5. What was the rationale behind selecting the similarity threshold? Why weren't the tools evaluated for detecting eccDNA at the basepair resolution?

Response: We appreciate the reviewer for thoughtful consideration of this important topic. We selected the true positive event by using 90% sequence identity, according to a previous study (Wanchai, Jenjaroenpun et al., *Brief Bioinform*, 2022), and then we compared the sequence similarity between the true positive event and simulated eccDNA by using the Similarity score (derived from PCC, RMSE, JSD), which is referenced from an earlier published benchmark paper (Li, Zhang et al., *Nat Methods*, 2022) from our lab.

It is a good point to use base pair resolution, which is more straightforward than using the similarity score. In the revised manuscript, we used the F1-score (Accuracy) and base pair difference between the identified eccDNA and the simulated eccDNA to evaluate the performance of analysis pipelines. We then updated all the Similarity score to the base pair (bp) difference in the revised manuscript. The top-ranked pipelines Circle_finder (bwa-mem-samblaster) and Circle-Map outperformed the others in short-read data analysis, while CReSIL was the top analysis pipeline for analyzing the long-read data (revised Figure 1b, see above R2#1). These results are consistent with the conclusions using the Similarity score. The evaluation metrics are revised in the **Results** section (from line 97 to 100, highlighted in yellow), which reads as follows:

"*True positive identification was defined as having over 90% sequence identity and less than 250 base pair (bp) difference with the simulated eccDNA. Performance metrics included F1-*

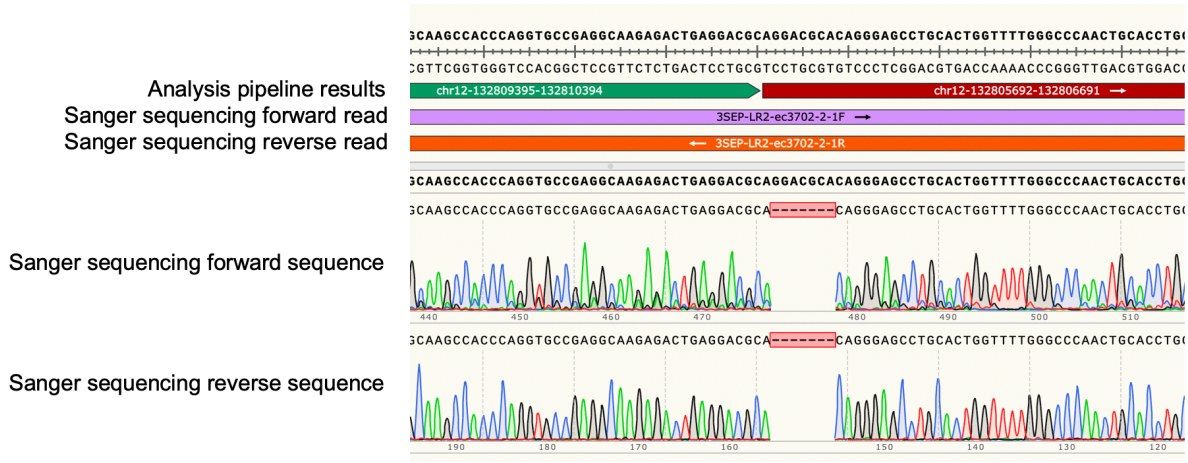
score and base pair difference between the identified eccDNA and the simulated eccDNA (see Methods)."

6. There is ample evidence that multiple eccDNA are generated from the same locus. As long as there is evidence of different junctional sequences, especially in the case of multiple junctional reads, they should be considered as different circles

Response: Thank you for focusing our attention on this important issue. The published papers used 90% sequence identity (Wanchai, Jenjaroenpun et al., *Brief Bioinform*, 2022) and less than 250 bp difference (Luebeck, Ng et al., *Nature*, 2023) for eccDNA identification, which showed a tolerance of 10% sequence or 250 bp difference for calling the same eccDNA. Indeed, we found that the upstream and downstream regions of the breakpoint may have homologous sequences, causing difficulty in identifying the exact coordinates of the breakpoint for the analysis pipelines. For example (Graph 1), Sanger sequencing showed that the real breakpoint region only had one copy of "GGACGCA", but the identified breakpoint by the analysis pipeline had two copies of "GGACGCA". This suggests that there can be several base pair differences between the real eccDNA and the eccDNA identified by analysis pipelines. Under current technological conditions, we cannot determine the origins of these homologous sequences, even manually.

Further, in our simulated datasets, the exact breakpoint of each eccDNA were known. The mean base pair difference between the identified eccDNA and the simulated eccDNA of all the analysis pipelines were larger than 1 (revised Figure 1b, see above R2#1) , which showed a systematic difference between the simulated eccDNA sequences and the identified eccDNA sequences of the results from all the analysis pipelines. If we considered the different junctional sequences as different eccDNA, this will misclassify identified eccDNA as false negative events.

After consideration, we used 90% sequence identity (Wanchai, Jenjaroenpun et al., *Brief Bioinform*, 2022) and less than 250 bp difference as the threshold to classify the same eccDNA. The eccDNA from the same locus but with different junctional reads were classified as true positive events but redundant results.



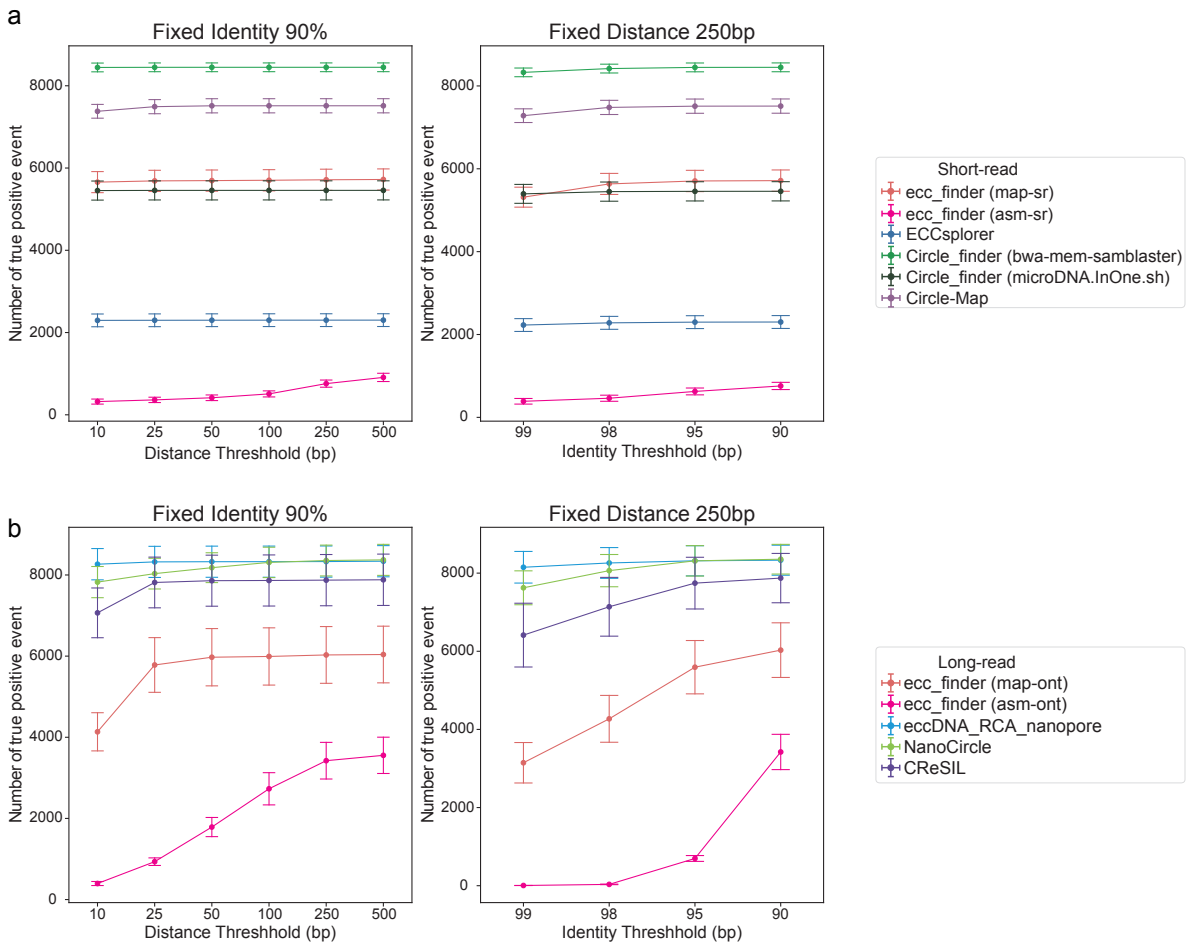
Graph 1. Example for Sanger sequencing validation.

7. True positive identification was defined as having over 90% sequence identity for the simulated dataset. Why was this not set to near 100% identity, or to be at the level of base pair resolution?.

Response: As shown in our response to your Comment 6, setting the threshold too high, such as near 100% identity, can misclassify true positive events as false negatives. We referenced the published study (Wanchai, Jenjaroenpun et al., *Brief Bioinform*, 2022) that used 90% sequence identity as the threshold for selecting true positive events. Additionally, another study used a 250 bp difference (Luebeck, Ng et al., *Nature*, 2023) as the threthold for identifying the shared breakpoint between different eccDNA.

To determine the appropriate threshold for selecting true positive events, we performed a series of analyses to examine the influence of different sequence identities and base pair difference thresholds on eccDNA identification. We found that changing the base pair difference threshold from 10 bp to 500 bp (under 90% sequence identity) and changing the sequence identity from 90% to 99% (under 250 bp difference) had no significant impact on the ranking of different analysis pipelines for short-read data analysis (Graph 2a). For long-read analysis pipelines, when the base pair difference threshold reached 250 bp, most pipelines showed stable performance (Graph 2b). Although loosening the sequence identity threshold can improve the performance of long-read analysis pipelines (Graph 2b), we decided to use the threshold from the published study, setting it at 90% (Wanchai, Jenjaroenpun et al., *Brief Bioinform*, 2022).

Based on your suggestion, the published studies, and our analysis results shown in Graph 2, we adopted a 90% sequence identity (Wanchai, Jenjaroenpun et al., *Brief Bioinform*, 2022) and less than a 250 bp difference (Luebeck, Ng et al., *Nature*, 2023) as the threshold for selecting the true positive events in the revised manuscript.



Graph 2. Influence of base pair difference threshold and sequence identity threshold on eccDNA identification. a. Influence of base pair difference threshold and sequence identity threshold on eccDNA identification from short-read data. **b.** Influence of base pair difference threshold and sequence identity threshold on eccDNA identification from long-read data.

8. Creating a table of circular DNA finding tools, along with their development dates and supported sequencing protocols, would be beneficial for providing a comprehensive overview of available tools in the field.

Response: We have now provided a table (revised Table 1, see above R2#3) to summarize the eccDNA identification pipelines, their published dates, supported experimental methods, chimeric eccDNA identification abilities, eccDNA enrichment status in the datasets, and supported sequencing data format. This table aims to assist readers to better understand this manuscript.

9. Impact of eccDNA enrichment steps on eccDNA identification: Circle_finder has a distinct pipeline (microDNA.InOne.sh) tailored for an exhaustive search, making it

more suitable for circular DNA-enriched samples (PMID: 28550083). In fact, Circle-Map was initially compared to Circle_finder upon its first publication

Response: Thank you for your suggestion. We have now included the microDNA.InOne.sh in all related comparisons throughout the revised manuscript. Circle_finder (microDNA.InOne.sh) performed better than Circle_finder (bwa-mem-samblaster) in terms of the "base pair difference" (revised Figure 1b and revised Supplementary Fig. 3, see above R2#1) and "Redundancy" (revised Figure 1d, see above R2#1). However, Circle_finder (bwa-mem-samblaster) showed a higher F1-score than Circle_finder (microDNA.InOne.sh) across all investigated sequencing depths (revised Figure 1b and 1c, see above R2#1). Including Circle_finder (microDNA.InOne.sh) did not change the top two analysis pipelines for analyzing short-read data. Furthermore, Circle_finder (microDNA.InOne.sh) had the least memory usage among all the analysis pipelines designed for short-read data analysis. We have now included the Circle_finder (microDNA.InOne.sh) mode in our **Study design** (from line 94 to 97, highlighted in yellow), which reads as follows, and revised all associated figures:

"We evaluated 11 modes of 7 pipelines, including Circle-Map, Circle_finder (bwa-mem-samblaster and microDNA.InOne.sh), ECCsplorer, and ecc_finder (map-sr/asm-sr) for short-read data analysis, and CReSIL, eccDNA_RCA_nanopore, NanoCircle, and ecc_finder (map-ont/asm-ont) for long-read data analysis."

10. Detection efficiency of specific type of eccDNA: What was the coverage of eccDNA identified in real data, and how random was the selection of eccDNA for verification by PCR method?.

Response: Thank you for your thoughtful comment. To obtain the eccDNA profile without the interference from redundant results, we selected Circle-Map for analyzing the Circle-Seq-SR and 3SEP-SR data for experimental methods comparison, and CReSIL for long-read data analysis. We calculated the average coverage (sequencing depth) of the identified eccDNA (Table R1).

Table R1. Mean coverage of identified eccDNA

	Circle-Seq-SR	Circle-Seq-LR	3SEP-SR	3SEP-LR
Mean Coverage	31.745204	13.147103	23.952322	10.767805
Standard Error	2.211889	0.150032	1.305185	0.135532

To select the eccDNA for validation, we adopted an *in silico* random selection strategy. The corresponding revision can be found in the **Methods** section (from line 569 to 573, highlighted in yellow), which reads as follows:

"*We created a numerical index for each eccDNA from each sample and used the random number generating formula in EXCEL (=randbetween(start index:end index)) to select the eccDNA. For the eccDNA that we could not design primers (potentially due to repeat sequences or low sequence complexity), we added 1 to the rolled random number and redesigned the primer for the newly indexed eccDNA.*"

11. Here are two Circle_finder pipelines available on GitHub: "microDNA.InOne.sh" and "circle_finder-pipeline-bwa-mem-samblaster.sh". The developer mentions on their GitHub page that the "microDNA.InOne.sh" has higher specificity but takes longer to run. Can the author explain why they chose the "circle_finder-pipeline-bwa-mem-samblaster.sh" script to compare it with other tools? "microDNA.InOne.sh" has been used for multiple publications where eccDNA were enriched. I suggest it would be a good idea to also evaluate the more sensitive version of Circle_finder.

Response: Thank you very much for your insightful suggestion. As detailed in the response to your Comment 9, we have now included Circle_finder (microDNA.InOne.sh) in the revised manuscript and updated all the comparison results. The inclusion of Circle_finder (microDNA.InOne.sh) did not significantly alter the ranking of top-performing analysis pipelines. Compared to Circle_finder (bwa-mem-samblaster), Circle_finder (microDNA.InOne.sh) showed a better performance in terms of the "base pair difference" and "Redundancy". Meanwhile, it had the least memory usage among all the analysis pipelines designed for short-read data analysis.

12. Computational resources consumed by different analysis pipelines: The author states that "Notably, Circle-Map and CReSIL kept stable computational resources consumption across all sequencing depths (Supplementary Figure 1F & 1G).". This statement seems a bit odd in this context. Instead of commenting on runtime and memory usage, they are commenting on which pipeline has stable computational resource consumption. Looking at Supplementary Figure 1F & 1G, Circle-Map and CReSIL are not the winners compared to other tools. This should be mentioned very clearly.

Response: Thank you for your helpful suggestion. In the revised manuscript, we mentioned the pipelines that had the least time cost and memory usage (revised Supplementary Fig. 3e and 3f, see above R2#1), instead of focusing on which pipeline had stable computational resource consumption. The revised content can be found in the **Results** section (from line 177 to 187, highlighted in yellow), which reads as follows:

"We observed that both the time and memory consumption of most pipelines increased with mean coverage rising (Supplementary Fig. 3e & 3f). For identifying eccDNA from short-read data, Circle_finder (bwa-mem-samblaster) was the fastest pipeline to identify eccDNA and Circle_finder (microDNA-InOne) used the least memory across all the investigated sequencing depths (Supplementary Fig. 3e). When considering the long-read data analysis pipelines, ecc_finder used the shortest time (map-ont) and the least memory (asm-ont) (Supplementary Fig. 3f). For dataset 5, ecc_finder (asm-sr) and ECCsplorer experienced memory errors on our platform (Supplementary Table 1). Besides, ECCsplorer also encountered memory errors in analyzing dataset 1 (depth over 25X), dataset 4 (depth over 40X), and dataset 6 (depth over 10X) (Supplementary Table 1)."

Reviewer #3

Remarks to the author:

The authors compared seven developed sequencing methods from multiple perspectives, including accuracy, similarity, duplication rate, and computational resource consumption, to help researchers identify eccDNAs in future experiments..

Response: We would like to express our gratitude to the reviewer for taking the time to review our manuscript, and appreciate your valuable feedback and insightful comments, which have helped us to improve the universality and impact of our work.

1. The authors innovatively constructed a Python script to simulate possible circular and linear DNA in human sperm cells, EJM cell lines, and JJN3 cell lines after analyzing existing data. The eccDNA identified by actual sequencing methods was compared with the simulated dataset results, thereby demonstrating the advantages and disadvantages of different sequencing methods. However, eccDNA is widely present in various species in nature, and its presence has been confirmed in various diseases of human cells. Except for human sperm cells, the EJM cell line and JJN3 cell line used are closely related to hematological diseases. If they can be replaced with other types of tumor cells, it can improve the universality of the article..

Response: Thank you very much for your valuable suggestions. We have now added 4 new eccDNA profiles generated from the Kelly cell line (neuroblastoma), medulloblastoma, muscle cells and the OVCAR8 cell line (ovarian cancer) to enhance the universality of our study (revised Supplementary Fig. 1 and Supplementary Fig. 2). We have updated the performance comparison results of the investigated analysis pipelines in the revised manuscript.

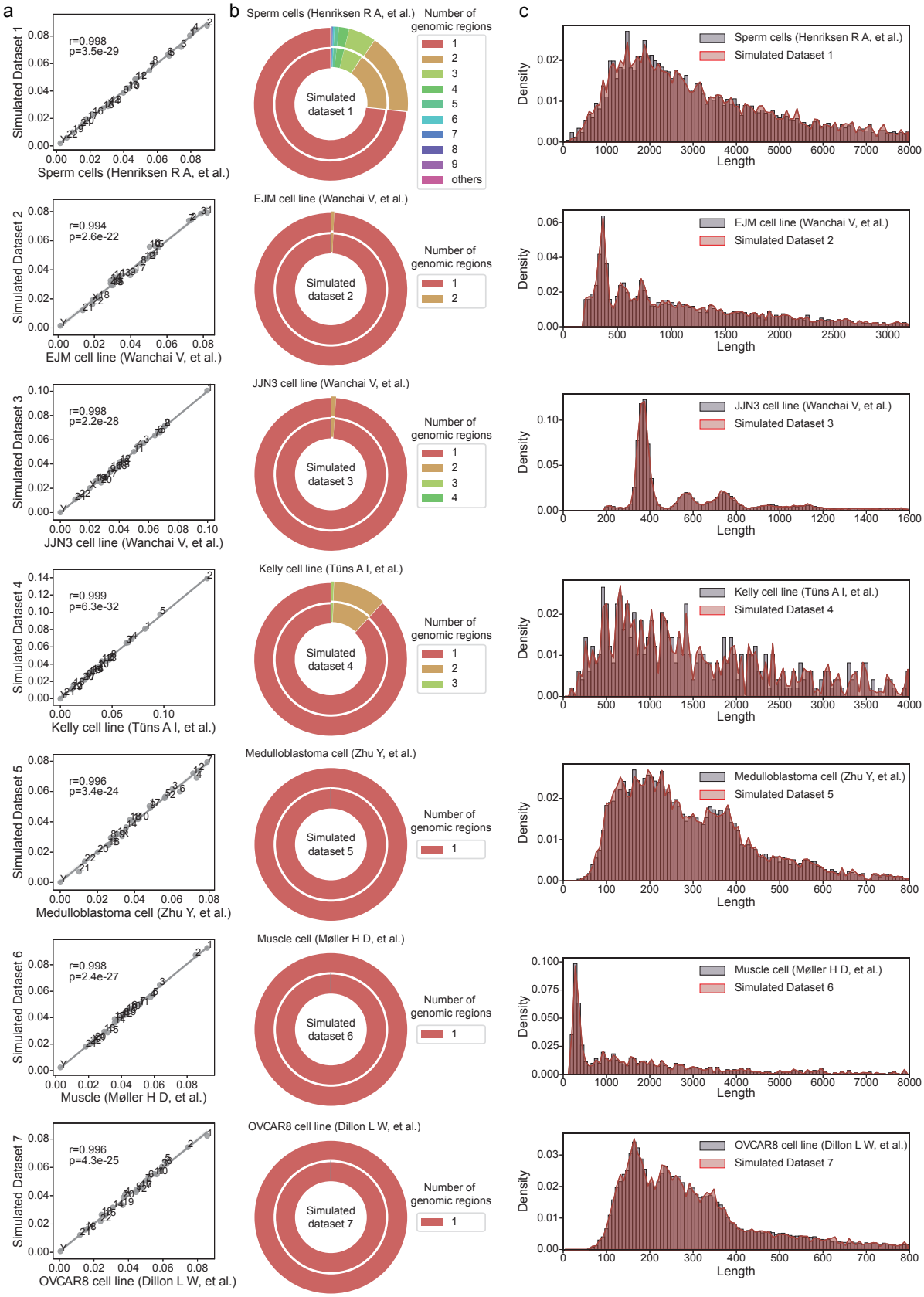
Overall, the inclusion of more datasets did not significantly alter the ranking of the top-performing analysis pipelines. Circle_finder (bwa-mem-samblaster) and Circle-Map outperformed the other analysis pipelines in short-read data analysis, while CReSIL showed the best performance in long-read data analysis (revised Figure 1b and 1c). However, Circle_finder (bwa-mem-samblaster) still exhibited notable redundancy in its results (revised Figure 1d). The proportion of chimeric eccDNA did not affect the recall of simple eccDNA identification for most short-read-based analysis pipelines, except for ECCsplorer (revised Figure 1e).

The added dataset information can be found in the **Results** section (line 89 to 92, highlighted in yellow), which reads as follows:

"Seven simulated datasets were produced, mirroring eccDNA identified in human sperm cells[19] , EJM cell line[16], JJN3 cell line[16], Kelly cell line[20], medulloblastoma[21], muscle cells[8] and OVCAR8 cell line[13] (Supplementary Fig. 1 and Supplementary Fig. 2),"

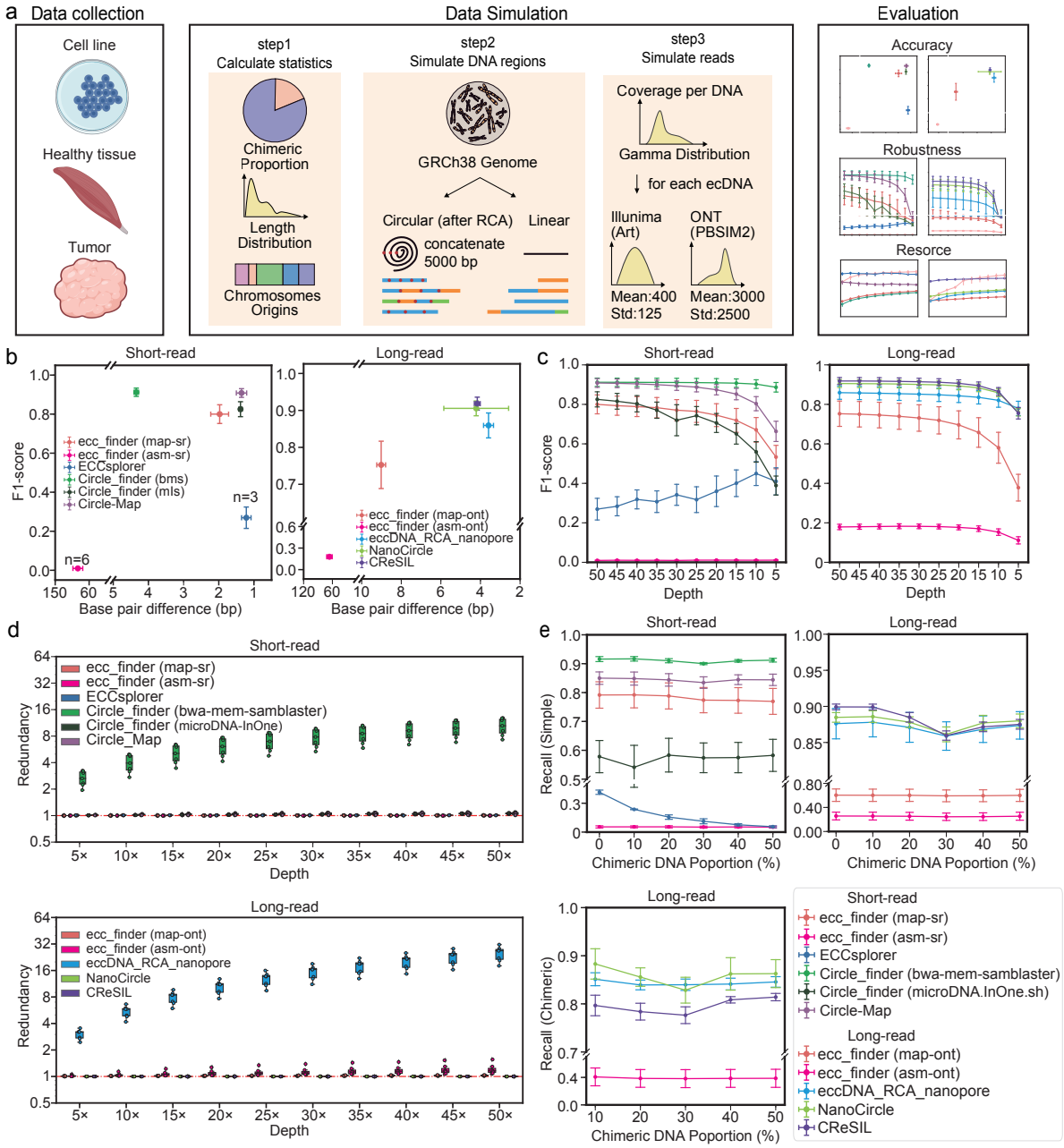
Dataset ID	Sample	Tissue/ Disease	Sequencing Platform	Pipeline	Number of Simple eccDNA	Number of Chimeric eccDNA
Dataset 1	Sperm cells	Semen	Nanopore	NanoCircle	8306	3014
Dataset 2	EJM cell line	Multiple myeloma	Nanopore	CReSIL	4635	33
Dataset 3	JJN3 cell line	Multiple myeloma	Nanopore	CReSIL	90661	988
Dataset 4	Kelly cell line	Neuroblastoma	Nanopore	eccDNA_RCA_nanopore	431	58
Dataset 5	Medulloblastoma	Medulloblastoma	Illumina	Circle-Map	12518	/
Dataset 6	Muscle cells	Muscle	Illumina	ecc_finder (map-sr)	1421	/
Dataset 7	OVCAR8 cell line	Ovarian cancer	Illumina	Circle_finder (bwa-mem-samblaster)	57231	/

Revised Supplementary Fig. 1. Dataset information. Column charts depicting simple eccDNA numbers and chimeric eccDNA numbers were logarithmically transformed using a base 10 logarithm.



Revised Supplementary Fig. 2. eccDNA profiles of different simulated datasets. a. Chromatin distribution comparison between the simulated dataset and corresponding real dataset. **b**.

Chimeric eccDNA composition of the simulated dataset and corresponding real dataset. **c**. Length distribution comparison between the simulated dataset and corresponding real dataset.



Revised Figure 1. Assessment of analysis pipelines in eccDNA identification. a. Schematic overview of the benchmarking workflow used to compare the performance of bioinformatic pipelines. **b.** Performance comparison of analysis pipelines at a simulated sequencing depth of 50X (bms, bwa-mem-samblaster; mls, microDNA.InOne.sh). **c.** Impact of simulated sequencing depth on eccDNA identification accuracy. **d.** Impact of simulated sequencing depth on eccDNA identification duplication rates. **e.** Impact of chimeric DNA proportion on eccDNA identification recall. For panels **b, c** and **d**, n=7, except for ecc_finder (asm-sr) with n=6 and ECCsplorer with

n=6 when depth ≤10, n=5 for depths between 15X and 25X, n=4 for depths between 30X and 40X, and n=3 for depths higher than 40X. For panel **e**, n=4.

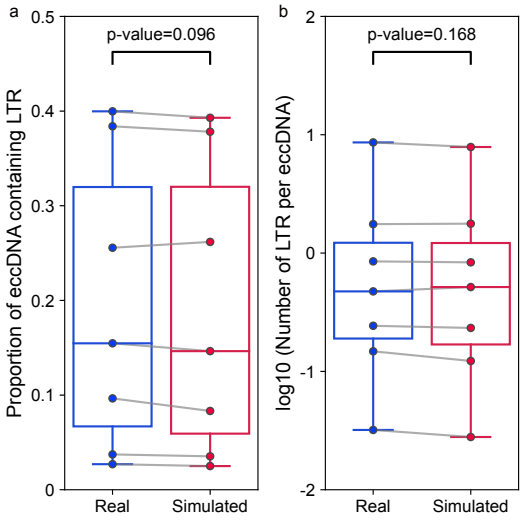
2. The simulation of eccDNA in the article is generated through random movement, and the existing literature does not fully understand the randomness of eccDNA production. In addition, the existence of retrotransposon hijacking mechanisms in viral infection and tumor cells implies that eccDNA production is not completely random. These results seem to reduce the authenticity of simulated datasets.

Response: Thank you for raising this important point. Indeed, we agree that the generation of eccDNA in biological systems may not be entirely random, especially considering mechanisms such as retrotransposon hijacking (Yang, Su et al., *Nature*, 2023). However, retrotransposon hijacking is only one of the mechanisms for eccDNA biogenesis. Wang et. al. have shown that eccDNAs can be generated from apoptosis and the detected eccDNA are randomly mapped across the entire genome (Wang, Wang et al., *Nature*, 2021). Additionally, the generation of double minutes (one type of eccDNA) by the ligation of chromosome fragments generated from chromothripsis occurs in a near-random manner (Stephens, Greenman et al., *Cell*, 2011).

Besides, studies have used random methods to generate simulated eccDNA datasets (Prada-Luengo, Krogh et al., *BMC Bioinformatics*, 2019, Wanchai, Jenjaroenpun et al., *Brief Bioinform*, 2022) to compare the performance of their developed analysis pipelines with others. Therefore, we generated the simulated eccDNA randomly but with the given length distribution, chromatin distribution and chimeric eccDNA proportions according to published datasets. We have now added the referenced studies in the **Methods** section of the revised manuscript to support this simulation method (from line 370 to 372, highlighted in yellow), which reads as follows:

"Because the biogenesis of eccDNA has not been fully known, we considered findings or eccDNA simulating methods from previously published papers[16,22,24] and created a python script to generate simulated eccDNA datasets for evaluation."

Despite using a random simulation strategy to generate simulated datasets in our study, the presence of retrotransposon hijacking mechanisms may undermine the usage of random simulation to simplify the eccDNA biogenesis. We further examined the proportion of long terminal repeat (LTR) elements in our simulated datasets, which are involved in the retrotransposon hijacking related eccDNA biogenesis. No significant differences were found between the proportions of LTR-containing eccDNA in the real and their paired simulated datasets (Graph 3a). We also examined the average number of LTR elements per eccDNA and still no significant differences were found between the real and their paired simulated datasets (Graph 3b). These results show that our simulated datasets can mimic the eccDNA profile from the real data.



Graph 3. Comparison of LTR elements between given datasets and paired simulated datasets. a. Proportion of eccDNA containing LTR elements. **b.** Average number of LTR elements per eccDNA.

3. Meanwhile, a significant portion of the existing understanding of eccDNA comes from existing sequencing methods, which are also included in the article. This will inevitably lead to the inclusion of the experimental methods to be identified in the simulation results when simulating the length distribution of eccDNA, chromosome origin, etc. in the simulation dataset. The widespread use of some sequencing methods will inevitably result in the evaluation results of the simulation dataset being biased towards these sequencing methods, and the authors did not seem to exclude this bias in the article.

Response: We thank the reviewer for focusing our attention on this important concern. Given our emphasis on evaluating analysis pipelines for eccDNA-enriched sequencing data, our generation of simulated datasets primarily relied on published sequencing data derived from eccDNA-enriched experimental methods. To minimize bias:

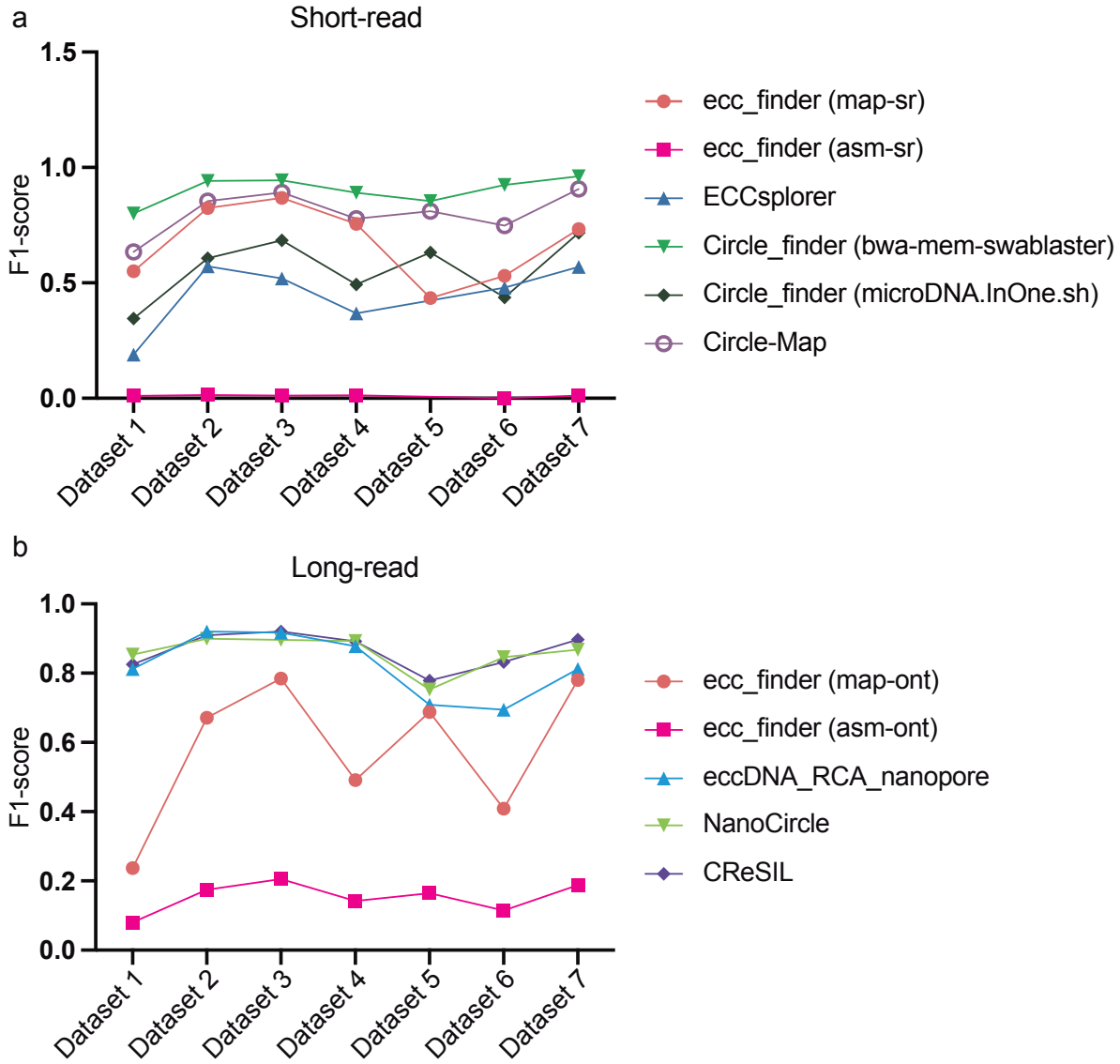
(1) we added 4 more published sequencing datasets from different sequencing methods (Circle-Seq-SR, Circle-Seq-LR) to reduce the bias caused by sequencing methods;

(2) We utilized different analysis pipelines to process these published datasets (revised Supplementary Fig. 1, see above R3#1), ensuring a more representative chromatin distribution, chimeric eccDNA composition, and length distribution (revised Supplementary Fig. 2, see above R3#1) compared to the original manuscript. This approach aimed to reduce bias introduced by varied analysis methods.

Further, we examined whether there exists any dataset bias towards a specific analysis pipeline (depth = 10X). We found that most of the pipelines showed only slightly different F1-scores in analyzing different datasets, except for ecc_finder (map-sr/map-ont). ECCsplorer and ecc_finder (asm-sr) encountered memory errors in analyzing dataset 5. We found that the ranking of top-performing pipelines remained consistent (Graph 4).

The associated contents can be found in the **Results** section of the revised manuscript (from line 374 to 377, highlighted in yellow), and reads as follows:

"We collected the eccDNA profiles identified by different analysis pipelines (Supplementary Table 1) from human sperm cells[19], EJM cell line[16], JJN3 cell line[16], Kelly cell line[20], medulloblastoma[21], muscle cells[8] and OVCAR8 cell line[13], and used these 7 datasets as input."

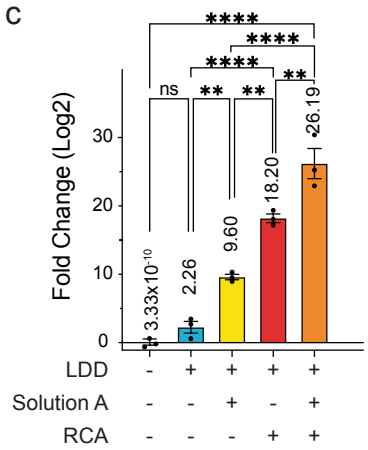


Graph 4. Dataset bias for each pipeline. a. F1-scores of short-read pipelines across different datasets. **b.** F1-scores of long-read pipelines across different datasets.

4. In the discussion section, the authors mentioned that the digestion of linear DNA has a marginal effect. In the future, it can be considered to skip the digestion of linear DNA and directly use solution A to purify RCA and circular DNA products. However, in current understanding, the digestion of linear DNA using DNA enzymes is a necessary step to obtain high-purity circular DNA, which is included in various sequencing methods and is crucial for obtaining accurate results of eccDNA. The results of the article does not seem to provide a detailed explanation of this marginal effect, nor does it provide relevant data to prove that skipping linear DNA digestion does not affect subsequent sequencing results.

Response: Thank you for pointing out this issue. Our study found that a 16-hour linear DNA digestion step can increase the ratio of circular DNA to linear DNA by approximately 4.79-fold ($2^{2.26}$). After linear DNA digestion, using Solution A for purification and RCA can further improve the circular DNA enrichment to 776-fold ($2^{9.60}$) and 301,124-fold ($2^{18.20}$), respectively(revised Figure 2c). We revised the content in the **Dissussion** section (from line 360 to 365, highlighted in yellow), which reads as follows :

"Our findings showed that a 16-hour linear DNA digestion step can increase the ratio of circular DNA to linear DNA by approximately 4.79-fold ($2^{2.26}$), while further using Solution A for purification and RCA, the circular DNA enrichment can be improved to 776-fold ($2^{9.60}$) and 301,124-fold ($2^{18.20}$), respectively. Increasing the efficiency of linear DNA digestion will be beneficial for enhancing the enrichment of circular DNA, and further efforts in this direction will be appreciated."



Revised Figure 2. Impact of eccDNA enrichment operations on eccDNA identification. c. Circular DNA enrichment efficiency. LDD, Linear DNA Digestion; Solution A: using Solution A for circular DNA purification; RCA, Rolling Cycle Amplification. n = 3 for all experiments. Statistical analyses were performed using one-way ANOVA with Tukey correction. error bars represent SEM. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, and $****p < 0.0001$.

References

1. Liu Y, Sun J, Zhao M. ONGene: A literature-based database for human oncogenes. *Journal of genetics and genomics* **44**, 119-121 (2017).
2. Thollet A*, et al.* ZNF217 confers resistance to the pro-apoptotic signals of paclitaxel and aberrant expression of Aurora-A in breast cancer cells. *Mol Cancer* **9**, 291 (2010).
3. Hou C, Zhuang Z, Deng X, Xu Y, Zhang P, Zhu L. Knockdown of Trio by CRISPR/Cas9 suppresses migration and invasion of cervical cancer cells. *Oncol Rep* **39**, 795-801 (2018).
4. Mei Q, Ye LJ, Lin H, Chen CY. CUL4A promotes the invasion of cervical cancer cells by regulating NF-κB signaling pathway. *Eur Rev Med Pharmacol Sci* **24**, 10403-10409 (2020).
5. Wang C*, et al.* C-Myc-activated long non-coding RNA PVT1 enhances the proliferation of cervical cancer cells by sponging miR-486-3p. *J Biochem* **167**, 565-575 (2020).
6. Wang C*, et al.* Long non-coding RNA plasmacytoma variant translocation 1 gene promotes the development of cervical cancer via the NF-κB pathway. *Mol Med Rep* **20**, 2433-2440 (2019).
7. Gao YL, Zhao ZS, Zhang MY, Han LJ, Dong YJ, Xu B. Long Noncoding RNA PVT1 Facilitates Cervical Cancer Progression via Negative Regulating of miR-424. *Oncol Res* **25**, 1391-1398 (2017).
8. Møller HD*, et al.* Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. *Nature Communications* **9**, 1069 (2018).
9. Yang F*, et al.* Retrotransposons hijack alt-EJ for DNA replication and eccDNA biogenesis. *Nature* **620**, 218-225 (2023).
10. Sun H, Lu X, Zou L. EccBase: A high-quality database for exploration and characterization of extrachromosomal circular DNAs in cancer. *Computational and Structural Biotechnology Journal* **21**, 2591-2601 (2023).
11. Boeva V*, et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423-425 (2012).
12. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
13. Dillon LW*, et al.* Production of extrachromosomal microDNAs is linked to mismatch repair pathways and transcriptional activity. *Cell reports* **11**, 1749-1759 (2015).
14. Cohen S, Agmon N, Sobol O, Segal D. Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells. *Mob DNA* **1**, 11 (2010).
15. Ling X*, et al.* Small extrachromosomal circular DNA (eccDNA): major functions in evolution and cancer. *Mol Cancer* **20**, 113 (2021).
16. Wanchai V*, et al.* CReSIL: accurate identification of extrachromosomal circular DNA from long-read sequences. *Brief Bioinform* **23**, (2022).
17. Li B*, et al.* Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat Methods* **19**, 662-670 (2022).
18. Luebeck J*, et al.* Extrachromosomal DNA in the cancerous transformation of Barrett's oesophagus. *Nature* **616**, 798-805 (2023).
19. Henriksen RA*, et al.* Circular DNA in the human germline and its association with recombination. *Mol Cell* **82**, 209-217.e207 (2022).
20. Tüns AI*, et al.* Detection and validation of circular DNA fragments using Nanopore sequencing. *Frontiers in Genetics* **13**, 867018 (2022).

21. Zhu Y*, et al.* Whole-genome sequencing of extrachromosomal circular DNA of cerebrospinal fluid of medulloblastoma. *Frontiers in Oncology* **12**, 934159 (2022).
22. Wang Y*, et al.* eccDNAs are apoptotic products with high innate immunostimulatory activity. *Nature* **599**, 308-314 (2021).
23. Stephens PJ*, et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).
24. 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* **20**, 663 (2019).