BMC Bioinformatics

SOFTWARE

Open Access

ECCsplorer: a pipeline to detect extrachromosomal circular DNA (eccDNA) from next-generation sequencing data

Ludwig Mann, Kathrin M. Seibt, Beatrice Weber and Tony Heitkam*

*Correspondence:
tony.heitkam@tu-dresden.de
Institute of Botany,
Technische Universität
Dresden, 01069 Dresden,
Germany

Abstract

Background: Extrachromosomal circular DNAs (eccDNAs) are ring-like DNA structures physically separated from the chromosomes with 100 bp to several megabasepairs in size. Apart from carrying tandemly repeated DNA, eccDNAs may also harbor extra copies of genes or recently activated transposable elements. As eccDNAs occur in all eukaryotes investigated so far and likely play roles in stress, cancer, and aging, they have been prime targets in recent research—with their investigation limited by the scarcity of computational tools.

Results: Here, we present the ECCsplorer, a bioinformatics pipeline to detect eccDNAs in any kind of organism or tissue using next-generation sequencing techniques. Following Illumina-sequencing of amplified circular DNA (circSeq), the ECCsplorer enables an easy and automated discovery of eccDNA candidates. The data analysis encompasses two major procedures: first, read mapping to the reference genome allows the detection of informative read distributions including high coverage, discordant mapping, and split reads. Second, reference-free comparison of read clusters from amplified eccDNA against control sample data reveals specifically enriched DNA circles. Both software parts can be run separately or jointly, depending on the individual aim or data availability. To illustrate the wide applicability of our approach, we analyzed semi-artificial and published circSeq data from the model organisms *Homo sapiens* and *Arabidopsis thaliana*, and generated circSeq reads from the non-model crop plant *Beta vulgaris*. We clearly identified eccDNA candidates from all datasets, with and without reference genomes. The ECCsplorer pipeline specifically detected mitochondrial mini-circles and retrotransposon activation, showcasing the ECCsplorer's sensitivity and specificity.

Conclusion: The ECCsplorer (available online at https://github.com/crimBubble/ECCsplorer) is a bioinformatics pipeline to detect eccDNAs in any kind of organism or tissue using next-generation sequencing data. The derived eccDNA targets are valuable for a wide range of downstream investigations—from analysis of cancer-related eccDNAs over organelle genomics to identification of active transposable elements.

Keywords: eccDNA, Extrachromosomal circular DNA, circDNA, circSeq, Mobilome-seq, TE activity, Mitochondrial minicircle

Background

Although first described over 50 years ago, the physiological role of most extrachromosomal circular DNAs (eccDNAs) remains debated [1–4]. Nevertheless, eccDNAs are often associated with many different biological roles and processes. In animals and human they have been reported as drivers for aging [5, 6] and cancer [7–9], whereas in plants they also play a role in gradual development of glyphosate resistance [10]. Many of these physiological consequences can be traced to extra gene copies residing extrachromosomally within the DNA circles. Beside including genes, eccDNAs often consist of repetitive DNAs such as tandem repeats (satellite DNAs) or (active) transposable elements [3, 11–13]. Thus, due to their relationship with transposable elements and genes, eccDNAs may also be involved in transcriptional regulation, e.g. mediated by small RNAs [14].

Although few functions of eccDNAs are known so far, there is a growing interest from medicine (e.g. early disease detection or therapy) and economy (e.g. new traits for crop plants through transposon activity or gene accumulation). Therefore, many emerging studies focus on eccDNAs in model and non-model organisms. This increasing interest comes along with new challenges for eccDNA detection, both experimentally and bioinformatically. Most recent approaches rely on next-generation sequencing after experimental enrichment of eccDNA [3, 12, 13, 15]. To enrich double-stranded circular DNA, phi29 polymerases are usually used for a random rolling circle amplification (rRCA) after linear was removed by an exonuclease. Although single long-read approaches are described [16], short-read techniques are widely common to generate so called circSeq data (in special cases also referred to as mobilome-Seq data [17]). Computationally, new eccDNA candidates are currently identified based on typical read signatures, commonly detected by the mapping against a reference genome [12, 18]. These characteristic signatures include split and discordant mapping reads, which result when mapping a circular sequence against a linear reference. Based on analysis of those reads, the most likely eccDNAs are commonly called. Yet, these properties are not unique to eccDNAs—repetitive DNAs, structural variants and assembly errors in the reference yield a similar read spectrum—thus, possibly resulting in many false positives.

To date, only very few software solutions are available to analyze the growing amount of circSeq data, and currently no solution is able to address the aforementioned challenges in a single approach. In addition, there is no software solution, yet, to analyze eccDNAs from short reads if a reference genome sequence is lacking.

Here, we present the ECCsplorer pipeline—a bioinformatics approach to specifically detect candidates with high confidence from eccDNA-enriched sequence datasets. Our pipeline is modular to guarantee maximal flexibility and reliability and represents a reproducible way for the detection of eccDNAs in both model and non-model organisms. It supports standard input formats (un-/compressed FASTQ and FASTA) as provided by most sequencing services and is available at https://github.com/crimBubble/ECCsplorer. We demonstrate its wide applicability using a combination of real and simulated sequence reads from three organisms, including model- and non-model organisms from two kingdoms. This allows us to showcase the ECCsplorer pipeline's specificity and sensitivity by targeting a broad range of eccDNA candidates: partial gene copies

(*Homo sapiens*), active transposable elements (*Arabidopsis thaliana*), and mitochondrial minicircles (*Beta vulgaris*).

Methods

ECCsplorer pipeline overview

The ECCsplorer pipeline (Fig. 1a; Additional file 1: Fig. S1) is modular and implemented in Python 3 and partly in R [19]. The ECCsplorer pipeline provides a framework for the automated detection of eccDNA candidates using well established tools including data transfer between tools, data summarization and assessment as well as data visualization in publication-ready figures for a convenient and reproducible user experience.

The ECCsplorer pipeline is allowing three input configurations (running modes: all/map/clu) to answer a variety of scientific questions (Fig. 1a). First, circSeq data from rRCA enriched DNA is mandatory to start the pipeline. Second, either a control dataset (mode: clu) or a reference genome sequence (mode: map) is necessary. As control data a variety of sequencing data can be provided for use (e.g. rRCA enriched data from a different ecotype or tissue, non-enriched data, or WGS (whole genome sequencing) data). Using both control data and a reference genome sequence, the ECCplorer yields

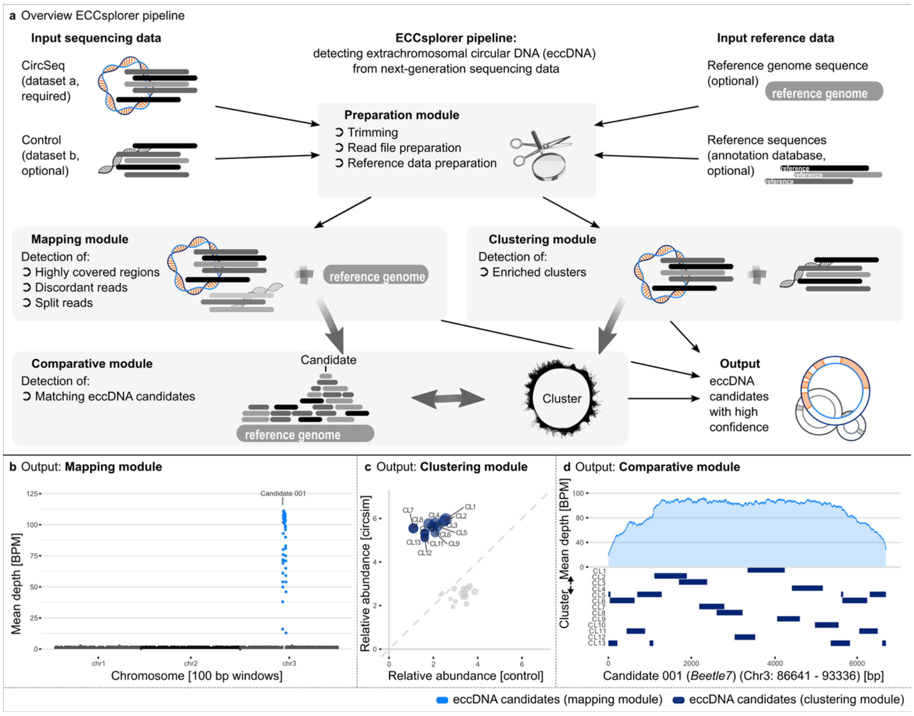
a Overview ECCsplorer pipeline
Input sequencing data
CircSeq (dataset a, required)
Control (dataset b, optional)
ECCsplorer pipeline:
detecting extrachromosomal circular DNA (eccDNA) from next-generation sequencing data
Input reference data
Reference genome sequence (optional)
reference genome
Reference sequences (annotation database, optional)
reference
reference
reference
Preparation module
Trimming
Read file preparation
Reference data preparation
Mapping module
Detection of:
Highly covered regions
Discordant reads
Split reads
reference genome
Clustering module
Detection of:
Enriched clusters
Comparative module
Detection of:
Matching eccDNA candidates
Candidate
reference genome
Cluster
Output
eccDNA candidates with high confidence
b Output: Mapping module
Candidate 001
125
100
75
50
25
0
Mean depth [BPM]
chr1
chr2
chr3
Chromosome [100 bp windows]
c Output: Clustering module
CL1
CL2
CL3
CL4
CL5
CL6
CL7
CL8
CL9
CL10
CL11
CL12
CL13
Relative abundance [circsim]
0
2
4
6
Relative abundance [control]
d Output: Comparative module
100
80
40
0
Mean depth [BPM]
Cluster
CL1
CL2
CL3
CL4
CL5
CL6
CL7
CL8
CL9
CL10
CL11
CL12
CL13
0
2000
4000
6000
Candidate 001 (Beetle7) (Chr3: 86641 - 93336) [bp]
eccDNA candidates (mapping module)
eccDNA candidates (clustering module)

Fig. 1 Overview of the ECCsplorer pipeline (**a**) and it's output (**b**, **c**). **a** Our pipeline consists of four main modules: one for the data preparation module and three for the analysis modules. The execution of the different modules depends on the given input data. Thick grey arrows represent multi-step bioinformatics computation processes, thin black arrows show data transfer. Semi-artificial circSeq and control data were simulated from a circularized retrotransposon and a random reference sequence to illustrate a typical pipeline output. A more detailed overview of the ECCsplorer pipeline is available in Additional file 1: Figure S1. **b** Mapping against the reference sequence revealed an eccDNA candidate region. **c** The comparative read clustering identified 13 clusters containing potential eccDNA candidates. **d** The comparative module links the retrotransposon candidate region from the mapping module to all 13 clusters from the clustering module

the most reliable results and represents the third input configuration (mode: all). Additionally, a user-defined file of reference sequences (containing e.g. known repeats, genes, etc.) might be provided for candidate annotation.

The ECCsplorer workflow can directly use raw sequence data as input and covers the steps for read file and reference preparation. The so-called, preparation module includes read trimming (using Trimmomatic [20]) and sampling as well as file conversion and indexing of the reference sequence.

The main part of the pipeline controls the analysis and consists of three additional modules: the mapping module, the clustering module and the comparative module. Depending on the given input data either the mapping module, the clustering module, or both are started (Fig. 1a):

- The mapping module maps the circSeq data to a reference sequence (e.g. genome assembly) with the tool segemehl [21]. Use of an additional control dataset will improve the module's results, but is not obligatory for the mapping itself.
- The clustering module is only executed, if control data are provided in addition to the circSeq dataset. This module conducts a comparative read clustering analysis with the RepeatExplorer2 pipeline [22, 23].
- If all three input datasets are available, the results of both modules will be combined by the comparative module using a similarity search with the BLAST + package [24].

The ECCsplorer pipeline will output the resulting eccDNA candidates for each analysis module summarized in up to three HTML files including publication-ready figures. In addition, the R script underlying all visualizations is available to the user for customization.

All major pipeline parameters as well as settings for the used 3rd party tools can be edited in the configuration file (for more details see the Additional file 2: Supplementary Methods). In the following, the main modules of the ECCsplorer pipeline are surveyed with their default settings. A detailed step-by-step description of all modules, their functionality and adjustable parameters can be found in the Additional file 2: Supplementary Methods.

The preparation module

The preparation module processes all input data and checks for the required software packages, libraries and third-party tools. If the recommended trimming option is enabled, Trimmomatic [20] is used to ensure high quality input for the ECCsplorer pipeline. Then, the read files are converted from FASTQ to FASTA if necessary using either seqtk [25] (if pre-installed) or the converter from the SeqIO [26] python package. In the next steps the read files are prepared for the read clustering with RepeatExplorer2 [22, 23]. In short, read files are sub-sampled, interlaced, concatenated, read names are modified with pre- and suffix, and the read lengths are equalized over all reads and files. For the mapping, the reference genome file is indexed. To enable the annotation of detected eccDNA candidates, a BLAST + database is created, later referred to as annotation database. By default, this annotation database corresponds to the RepeatExplorer2's internal database and can be extended by a user-defined FASTA file of reference sequences (containing e.g. known repeats, genes, etc.). The user-extended annotation database will be available to the mapping/comparative module. After the data preparation, the analysis modules are started.

The mapping module

The mapping module maps circSeq and control reads (if provided) against a reference sequence and identifies eccDNA candidate regions. For mapping, we use the segemehl (bioinf.uni-leipzig.de/Software/segemehl/, [21]) algorithm. This tool is a read mapper that was specifically developed for split and circular mappings and is consequently well-suited for detecting eccDNAs. The mapping relies on a reference genome sequence and enables the detection of eccDNA candidate regions. To achieve this, the ECCsplorer relies on three characteristics of amplified eccDNAs mapped to a linear reference, namely (1) split reads, (2) discordantly mapping reads, and (3) locally high coverages. These are analyzed in the following way: first, potentially circular regions flanked by at least five split reads are determined using the haarz [21] algorithm included within the segemehl tool. Second, regions containing at least one pair of discordantly mapping reads are identified based on the corresponding SAM-flags using SAMtools [27] and BEDtools [28]. And third, regions with a significantly high coverage, caused by the experimental enrichment of the eccDNA, are detected using the peak finder algorithm of SciPy [29] with a minimum peak prominence value of one. Finally, all three files with candidate regions are compared for matching or overlapping regions using BEDtools [28]. The boundaries of the final eccDNA candidate regions are primarily based on the split read regions. Regions meeting all three criteria are considered as candidate regions with high confidence, whereas those complying with any two out of the three criteria are marked as candidate regions with low confidence. Regions with low confidence are only considered in further analysis, if regions with high confidence are absent.

After the final eccDNA candidate regions are determined, a similarity search with BLASTn [24] is performed against the annotation database that was previously built. In addition, the candidates' enrichment scores are calculated using normalized mapping coverages. All coverages are normalized to bases per million bases (BPM), which allow a direct comparison of read inputs of different read lengths. The enrichment score reflects the enrichment by rRCA of a certain eccDNA candidate and is used to rank candidates by their estimated abundance.

The clustering module

For the clustering module, we assume that the presence of eccDNAs leads to the over-representation of the encoded sequences. This is additionally enhanced by the experimental amplification. To determine the over-representation of eccDNA sequences, this module performs a comparative read clustering using RepeatExplorer2 (http://repeatexplorer.org/, [22, 23]) and requires circSeq and control data as input. With RepeatExplorer2, multiple datasets are comparatively clustered and their repetitive fractions are detected, classified, and quantified. Due to the experimental rRCA, all eccDNAs in the circSeq data exhibit repetitive characteristics, although they might not have originated from true repetitive elements. In contrast to the read mapping the clustering approach is not dependent on a reference genome sequence and less susceptible to background noise. For example, non-circular genomic DNA might still be present in the sample, but in a much lower copy number and will, thus, be filtered by this approach. The key outputs of the RepeatExplorer2 [22, 23] utilized within the ECCsplorer pipeline are the cluster-derived contigs and each cluster's comparative read counts (read count in a cluster per dataset). Clusters with a circSeq read count proportion of more than 80% are selected for further analysis.

The comparative module

If both the clustering and the mapping modules are executed, the detected eccDNA candidates are further analyzed within the comparative module. For this, contigs of candidate read clusters serve as query for a similarity search with BLASTn [24] against the candidate regions detected by the mapping module. Candidates detected by both approaches are considered eccDNA candidates supported with very high confidence and are reported in a conclusive summary.

Results and discussion

Case studies demonstrating the ECCsplorer's functionality

To evaluate and demonstrate the ECCsplorer's functionality, we first outlined the results using simulated enrichment of the eccDNA fraction (Fig. 1b–d). Next, we reanalyzed publicly available data of the model organisms *A. thaliana* (Fig. 2a–c) and *H. sapiens* (Fig. 2d), as well as newly generated circSeq data for the non-model crop *B. vulgaris* (Fig. 2e). The three datasets illustrate three different research questions, three different input configurations, and a variety of different eccDNA candidates; thus together fully demonstrate the ECCsplorer pipeline's wide applicability and functionality.

Case study I: proof of concept using simulated data

To illustrate the ECCsplorer pipeline functionality, we applied it to simulated semi-artificial data (Fig. 1b–d; Additional file 3: Tables S1–S3). For this, three regions of the *B. vulgaris* reference genome sequence [30] were randomly selected to represent chromosomes with a total length of about 0.6 Mb, including a copy of the well-described long terminal repeat (LTR) retrotransposon *Beetle7* [31]. A potential enrichment of the 6695 bp long *Beetle7* circular DNA was simulated by introducing a total nine copies of this retrotransposon in tandem arrangement (including solo-LTR and LTR-LTR junctions). To simulate the short-read data, the tool dwgsim (github.com/nh13/DWGSIM) was used. As circSeq data 50,000 paired-end (PE) reads with a length of 200 bp each were simulated using the reference sequence and the tandemly arranged retrotransposon. For the control data the same amount of 200 bp PE reads was generated from the reference only. The reference itself served as reference genome sequence. In addition, the reference sequences of *Beetle1* to *Beetle7* [31] were used as annotation database.

After trimming, 46,950 PE reads of the circSeq dataset and 46,754 PE reads of the control dataset were piped into the subsequent bioinformatics steps, respectively. Of those reads 12,000 PE reads (6000 per dataset) were processed further for the clustering analysis, which corresponds to a 4 × coverage of the reference. The results from the mapping module as well as those from the clustering module yielded *Beetle7* as the only eccDNA candidate with high confidence. The mapping module reported one candidate

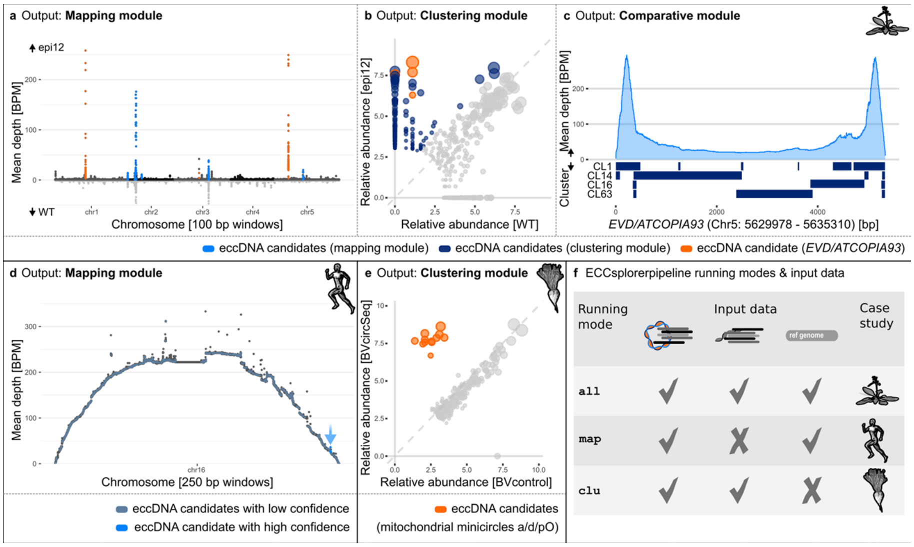

Fig. 2 Outputs of ECCsplorer for different case studies representing the running modes and input data configurations. **a**–**c** An ECCsplorer run using *A. thaliana* circSeq (epi12) and control data (WT) [12] detected *EVD/ATCOPIA93* as key eccDNA candidate (orange). **a** A mapping against the reference genome sequence (TAIR10) revealed eight eccDNA candidate regions. These were identified by applying three criteria: the region is flanked by split reads (at least five), contains discordant mapping reads (at least one pair), and shows a high coverage compared to its neighboring regions (peak prominence of at least one). **b** The comparative read clustering identified 41.04% of the clustered epi12 reads as potential eccDNA candidates. **c** To report eccDNA candidates with high confidence the outputs of the mapping and clustering strategies are compared by the ECCsplorer pipeline. The comparative module links the *EVD/ATCOPIA93* candidate region from the mapping module to four clusters from the clustering module. **d** An ECCsplorer run using *H. sapiens* circSeq data [32] detected one eccDNA candidate region containing genic content (light blue) and many candidate region with low confidence (grey blue). **e** ECCsplorer analysis using *B. vulgaris* circSeq data and available sequenceing data [30] as control data detected multiple candidate clusters that were in the following annotated as mitchondrial minicircles a/d/pO (orange). **f** Overview on the running modes, their required input data and the corresponding case studies

region on the artificial reference chromosome 3 with a length of 6695 bp, an enrichment score of 46.5, and a BLAST+annotation as *Beetle7*, respectively (Fig. 1b; Additional file 3: Table S1). The clustering module reported 13 candidate clusters (containing 3475 reads in total, 3399 reads from the circSeq data and 76 from the control data) adding up to one circular supercluster (Fig. 1c; Additional file 3: Table S2). The observed split of the eccDNA candidate into multiple clusters is relatively common after read clustering with RepeatExplorer2. Based on paired end read information, the clusters are commonly linked into superclusters. The comparative module reported that all 13 candidate clusters showed sequence similarities to the eccDNA candidate region (Fig. 1d; Additional file 3: Table S3). Four of those candidate clusters the clusters (namely 5, 6, 11, and 13) cover the circular break-point.

Taken together, the ECCsplorer pipeline successfully retrieved the artificial eccDNA candidates and revealed no other false-positives.

Case study II: identification of active retrotransposons using circSeq data of A. thaliana

To demonstrate the ECCsplorer's functionality to detect the activation of LTR retrotransposons, we re-analyzed public data from *A. thaliana* [12] (Fig. 2a–c; Additional file 3: Tables S4–S6). As a reference genome and both circSeq and control data are available, all ECCsplorer modules were used (running mode: all). Lanciano et al. [12] amplified eccDNAs of epigenetically impaired *A. thaliana* plants and used the wild-type as control. We expected that our pipeline shows enrichment of the *EVD/ATCOPIA93* retrotransposon, as originally reported. The full ECCsplorer pipeline was started with default settings and the trimming option enabled using the Nextera™ adapter. A total of 501,558 circSeq PE reads and 143,372 control PE reads were mapped against the TAIR10 (The Arabidopsis Information Resource, http://www.arabidopsis.org) reference genome.

The ECCsplorer mapping module retrieved the well-known LTR retrotransposon *EVD/ATCOPIA93* as eccDNA candidate with the highest probability (Fig. 2a, orange; Additional file 3: Table S4) as well as other potential eccDNA candidates, e.g. originating from the ribosomal genes (Fig. 2a, blue; Additional file 3: Table S4). In detail, the mapping module reported 13 highly confident candidate regions and 673 candidates with low confidence. The candidate region with the highest enrichment score of 56 was 5332 bp long and annotated as *EVD* by the pipeline's BLASTn analysis—in line with our expectation. We conclude that the ECCsplorer's mapping module produces results that compare well with the original findings and that it is able to detect active retrotransposons from current datasets.

For the comparative read clustering, equal read counts of circSeq and control data were used (i.e. 96,016 PE reads per dataset). The clustering module identified 41.04% of the clustered circSeq and 1.41% of the control data reads as potential eccDNA candidates, respectively (39,402/1371 reads from circSeq/control), pointing to a 28.7 × overall increase (Fig. 2b). Four of the largest clusters containing about 10.3% of all reads in candidate clusters (4058/2 reads from circSeq/control), were reported as Ty1_copia/Ale, and correspond to the *EVD/ATCOPIA93* reference (Fig. 2b, orange circles; Additional file 3: Table S5). We conclude that the eccDNA candidates can be readily detected with the clustering module as well as with the mapping module.

To be able to report highly confident eccDNA candidates, the pipeline compares the outputs of the mapping and clustering strategies. Here, the comparative module links eight candidate regions from the mapping module (Fig. 2a) to eleven clusters from the clustering module (Fig. 2b) with the best results for the eccDNA candidate *EVD/ATCOPIA93* (Fig. 2c; Additional file 3: Table S6). After removing duplicates, this leads to three target retrotransposon regions for potential experimental verification. These results are in line with the original study [12], and lead us to highlight the ECCsplorer's potential for the fast and reliable identification of retrotransposon mobilization.

Case study III: detection of (genic) eccDNAs using circSeq data from healthy humans (H. sapiens)

Typically, eccDNAs in *H. sapiens* are associated with cancer and other diseases. However, also in healthy tissues eccDNAs may arise [32]: The study by Møller et al. generated multiple circSeq datasets from healthy blood and muscle tissues. They detected about 100,000 unique eccDNAs including genic eccDNAs. As circSeq enrichment data and the reference genome sequence were available, we re-analyzed these data using the ECCsplorer's mapping module (running mode: map). As the original publication reported high eccDNA candidate density on chromosome 16 of the hg38 assembly (*H. sapiens* genome assembly GRCh38) [33], we used this chromosome as reference, along with the corresponding mRNA database as annotation database (UCSC Genome Browser, https://genome.ucsc.edu/).

The mapping module detected eccDNA candidates with high (n = 1) and low confidence (n = 840) on the analyzed chromosome 16 (Fig. 2d). The highly confident candidate region was 22,772 bp long and was located on the distal end of chromosome 16 (Fig. 2d, light blue, arrowed). It contains several gene annotations with the highest BLAST score observed for the gene mitofusin (MFN1, Additional file 3: Table S8). Due to the arch-like distribution of the read coverage over the whole chromosome the candidate region with high confidence only showed an enrichment score of 0.16 as it is calculated globally. Although the ECCsplorer detected only a single eccDNA candidate region with high confidence, our manual analysis of regions with lower confidence showed also promising results. In total, 29 of those low-confidence candidate regions were supported by Møller et al.'s approach ([32], Fig. 3f; Additional file 4: Supplementary Data).

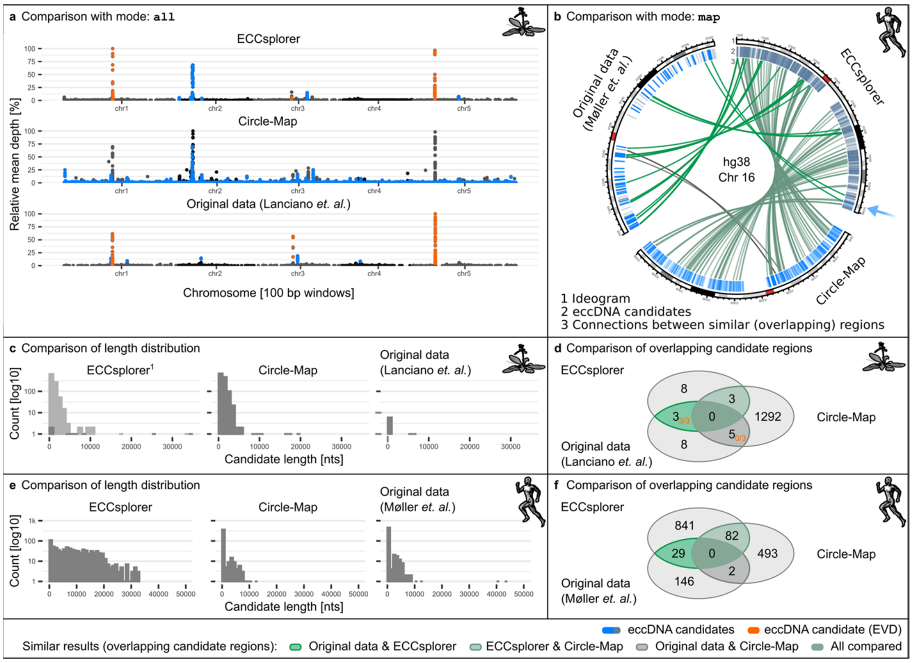

Fig. 3 Comparison of results from ECCsplorer with Circle-Map [18] and originally published eccDNA candidates. We reanalyzed data from *A. thaliana* [12] and *H. sapiens* [32] using available eccDNA tools, our ECCsplorer and Circle-Map. **a** Mapping plots from all three approaches (ECCsplorer, Circle-Map and original published data by Lanciano et al. 2017) of (re-) analyzed *A. thaliana* circSeq data. For each approach, the detected eccDNA candidates have been highlighted in blue with the key eccDNA candidate *EVD/ATCOPIA93* shown in orange (experimentally validated: "ground truth"; maximal coverage depth of each individual plot corresponds to 100% relative depth). **b** Comparison of all three approaches (ECCsplorer, Circle-Map and original published data by Møller et al. 2018) of (re-)analyzed *H. sapiens* circSeq data. Bars in track 2 represent the detected eccDNA candidate regions by each approach. The connections highlight similar (overlapping) genomic regions detected on chromosome 16 of hg38 (arrow points to eccDNA candidate with high confidence). **c** Length distribution of by each approach detected eccDNA candidates using *A. thaliana* circSeq data ([1] candidates with low confidence additionally shown in light grey). The ECCsplorer's high-confidence hits show large overlap to the "ground truth". **d** Venn-diagram of similarly detected eccDNA candidates using *A. thaliana* circSeq data with *EVD/ATCOPIA93* candidate overlaps shown in orange. **e** Length distribution of by each approach detected eccDNA candidates using *H. sapiens* circSeq data. **f** Venn-diagram of similarly detected eccDNA candidates using *H. sapiens* circSeq data. The ECCsplorer shows a larger overlap to the original data than the CircleMap tool; yet "ground truth" data is still lacking

A main difficulty for this specific read dataset was the high background noise, presumably from linear DNAs that remained after incomplete exonuclease treatments. Nevertheless, the ECCsplorer pipeline was able to detect eccDNA candidates from genic regions, confirming the presence of eccDNAs in healthy *H. sapiens*.

Case study IV: detection of eccDNAs absent from the reference genome using circSeq data from B. vulgaris

To test whether our ECCsplorer pipeline is able to detect eccDNAs absent from reference genome assemblies, we queried sugar beet (*B. vulgaris*), a non-model organism. The *B. vulgaris* genome harbors small extrachromosomal circular stretches of mitochondrial DNAs [34] that are absent from the published reference assembly [30, 35]. We generated circSeq data from inflorescences of *B. vulgaris* after experimental enrichment of the eccDNA fraction according to Lanciano et al. (2017) and Diaz-Lara et al. (2016). As control, we used publicly available data from the same genotype (KWS2320, see Data Availability).

To test the ECCsplorer's usability for the detection of extrachromosomal mitochondrial DNAs absent from the reference genome, we used the comparative read clustering as embedded in the clustering module (running mode: clu). For the clustering, 322,580 PE reads (161,290 PE reads per dataset) were used, respectively. The clustering module revealed twelve candidate clusters (combined in one supercluster) containing 30,792 reads (30,636/156 in circSeq/control). All were clearly enriched (Fig. 2e, orange circles; Additional file 3: Table S9) and of mitochondrial origin. A manual BLAST assigned all twelve clusters to the *B. vulgaris*-typical mitochondrial minicircles, termed a, d and pO [34].

These results clearly show that the ECCsplorer is capable of reference-free eccDNA detection at low sequencing coverages. We further want to highlight that already existing sequencing runs can be used for the comparative clustering analysis. Ideally, however, we recommend preparing enriched and non-enriched DNA from the same samples.

In contrast to other eccDNA detection methods, the ECCsplorer can be used without a high-quality reference genome sequence and is therefore much less vulnerable to assembly errors. This makes the ECCsplorer the current method of choice when working with non-model organisms and low-coverage, short-read sequencing data (see also 3.2 Comparison with other tools).

Comparison with existing tools

Although the amount of available circSeq data and the interest in eccDNAs has been growing lately, there is yet no standardized way of analyzing such data. To date, only few attempts for software solutions are available and currently no solution is able to address different approaches in a single tool. Most of the available tools are aimed at a very specific use-case and are not applicable for non-model organisms at all.

To our knowledge there are currently only a few software solutions to detect eccDNAs. Circle-Map [18] is a realigning-based pipeline to detect eccDNA from circSeq datasets already mapped to a reference genome sequence. Circle_finder [15] is a script collection written for the detection of eccDNA (there called microDNA) from *H. sapiens* samples. CIRCexplorer2 [36] is intended to detect circular RNA. Two further tools available are AmpliconArchitect [37] and CIDER-Seq [16]. Whereas the AmpliconArchitect aims very specifically at detecting eccDNAs from *H. sapiens* cancer tissue, the CIDER-Seq approach relies on long-read (Pac-Bio/SMRT) sequences, only. At last there is the Circulome-Seq [38], which is an RCA-free, column-free approach that enriches eccDNA by prolonged exonuclease V treatment and consequently library construction with the transposase Tn5. Though the very different application ranges of AmpliconArchitect, CIDER-Seq and Circulome-Seq, a direct comparison with our ECCsplorer pipeline is unfortunately prevented. For human data, Circle-Map has been compared to Circle_finder and CIRCexplore2 before and shows similar or better performance overall [18].

To evaluate the performance of our ECCsplorer pipeline, we compared its output with published results and those that we retrieved after re-running Circle-Map. For the comparison, the datasets analyzed with the ECCsplorer (Fig. 2a–d) were reanalyzed with Circle-Map and bwa-mem [39] as mapping tool, as recommended (https://github.com/iprada/Circle-Map/wiki). In addition, we compared the results obtained from both tools with the originally published results. As Circle-Map relies on a gold-standard reference genome, only *A. thaliana* and *H. sapiens* data are suitable as input. To our knowledge, non-model organisms that lack high-quality draft genomes are only analyzable with our ECCsplorer pipeline when using short-read sequencing data at low coverage (< 0.2 ×).

First, we compared results of both tools and the originally published data from the enrichment of eccDNA in *A. thaliana* [12]. In summary, the ECCsplorer detected eight eccDNA candidate regions and 673 regions with low confidence (Fig. 3a, c, d). Three of the eight candidates were annotated as *EVD/ATCOPIA93* (Fig. 3a, first skyline plot, orange peaks), an LTR retrotransposon previously reported as transpositionally active [40]. The five remaining candidates appeared to be organellar DNA or tandem repeats. The original publication [12] also reported eight candidate regions with three of them being *EVD/ATCOPIA93* (Fig. 3a, third skyline plot, orange peaks). The original data reported no organelle-derived eccDNAs as those reads had been filtered out before their analysis. Re-analysis with the Circle-Map tool reported 1292 candidate regions, whereas regions greater than 50 kb have been manually filtered out as they were very likely false-positives or overlapped other candidates (Fig. 3a, second skyline plot, blue peaks and Fig. 3e, f).

The ECCsplorer output and the original data shared three high-confidence candidate regions with all of them being *EVD/ATCOPIA93* (Fig. 3d; Additional file 3: Table S7). Four additional originally reported regions occurred in the low confidence output of the ECCsplorer pipeline. The ECCsplorer output and the Circle-Map output share three candidates as well, but none of them were *EVD/ATCOPIA93.* The Circle-Map output and the original data share five candidate regions, but again none of them being *EVD/ATCOPIA93.* There was no candidate region found by all three approaches (Fig. 3d; Additional file 3: Table S7). The comparison of the length distributions of the eccDNA candidates showed similar profiles between the ECCsplorer candidates (with high confidence) and the originally published data as well as between the ECCsplorer candidates (with low confidence) and the Circle-Map results (Fig. 3c). This comparison demonstrates that the ECCsplorer pipeline provides more accurate results than the Circle-Map tool using default settings and considering the originally published results [12] as ground truth with experimental validation. Quite surprising was the absence of an *EVD/ATCOPIA93* eccDNA candidate in the Circle-Map results despite the large number of output candidates. The ECCsplorer pipeline outperforms Circle-Map in this case study using low-coverage (~ 1 ×) input data.

Second, we compared the results of ECCsplorer and Circle-Map with the originally published results of the *H. sapiens* circSeq study focusing on candidates on chromosome 16 from the hg38 assembly (Fig. 3b). Chromosome 16 was reported by the original study [32] to have a high per Mb eccDNA count and is also one of the shorter *H. sapiens* chromosomes allowing all tools to run on a desktop-grade computer. The ECCsplorer pipeline found one candidate region with high and 840 with low confidence using default setting (Fig. 3b, e, f), respectively. The original study [32] reported 70 highly and 75 lowly confident candidate regions. Re-analysis with the Circle-Map tool found 493 candidates using recommended settings, and manual filtering of candidates greater than 50 kb. The outputs of all approaches were compared with BEDtools intersect (for details see Supp. Info 2 Methods). The ECCsplorer pipeline had 29 results in common with the originally published data and 82 overlaps to candidates reported by the Circle-Map tool (Fig. 3b, f; Additional file 4: Supplementary Data). Curiously, only two similar candidates were shared between the original data and the Circle-Map results (Fig. 3f), despite the length distribution profiles appearing to be similar (Fig. 3e). No candidates were shared across all approaches. This demonstrates the complexities of the detection of eccDNA candidates from circSeq data and the need of a unified, and reproducible software solution. The ECCsplorer pipeline shared similar candidates with both other approaches (Fig. 3f), while detecting longer candidates overall (Fig. 3e). This finding makes the ECCsplorer a viable starting point for the analysis of eccDNAs. This is especially true in the light that the published candidates from *H. sapiens* [32] have not entirely been experimentally validated. Additionally, our findings demonstrate that an approach based on mapping only may result in many false-positive eccDNA candidates. Therefore, we recommend running the ECCsplorer pipeline with all implemented approaches, including the mapping, clustering, and comparative modules.

Current limitations

The main limitations of the current version of the ECCsplorer pipeline are justified in the implemented tools and algorithms, which can heavily occupy computing resources. First, segemehl is using very high amounts of RAM (random-access memory) for rather medium-sized reference genome sequences (128 GB of RAM are recommended for usage with *H. sapiens* reference genome). Nevertheless, its high accuracy in split-read detection make it a very valuable choice for the analysis of eccDNA candidates. Second, the peak_finder algorithm is comparably slow scanning chromosome scale data, but needed to generate eccDNA candidates. Third, the all-against-all BLAST performed by the RepeatExplorer2 limits the number of analyzable reads. However, implementation of RepeatExplorer2 read clustering offers a fundamentally different way to detect eccDNA amplification, hence greatly increasing the usefulness of the pipeline. Last, the amplification bias in the experimental procedure prevents a quantitative analysis of the eccDNA fraction, and instead focuses on the qualitative detection of eccDNAs. The current limitations can be overcome by splitting the individual dataset and performing multiple runs. Summarizing, the ECCsplorer pipeline aims for robust, qualitative results over speed, producing highly confident eccDNAs by integrating different detection approaches.

Conclusion

Due to the wide relevance in medicine and breeding, eccDNA analyses are gaining popularity. Nevertheless, the detection of eccDNAs from next-generation enrichment sequencing is far from being a solved problem, and there is a high demand for new tools. In particular, organisms lacking gold-standard genome assemblies are virtually impossible to analyze to date. To meet this need, we introduce our ECCsplorer pipeline that modularly combines reference-free and reference-guided procedures to identify eccDNA candidates from rolling-circle amplification protocols such as circSeq and mobilome-seq. Using simulated data (case study I) as well as real world data (case studies II–IV) from various organisms, we were able to demonstrate the ECCsplorer pipeline's functionality, showcase its wide applicability, and highlight its advantages over existing tools. Finally, the usage of a bioinformatics pipeline for the analysis of data examining a fairly new research field allows advanced comparability between different studies and ensures their reproducibility.

Abbreviations

A. thaliana: *Arabidopsis thaliana*; all: Running mode: run all modules; *B. vulgaris*: *Beta vulgaris* ssp. *vulgaris*; bp/kb: Base-pairs/kilo base-pairs; BPM: Bases per million mapped bases; circDNA: Circular DNA; circSeq: Illumina-sequencing of RCA amplified circular DNA; clu: Running mode: run clustering (and preparation) module; eccDNA: Extrachromosomal circular DNA; EVD: Evade, LTR-retrotransposon ATCOPIA93; *H. sapiens*: *Homo sapiens*; hg38: *H. sapiens* (human) genome assembly GRCh38; KWS2320: Double-haploid accession of *B. vulgaris*; LTR: Long terminal repeat; map: Running mode: run mapping (and preparation) module; MFN1: Mitofusin; mobilome-seq: Mobilome sequencing; PE: Paired-end; RAM: Random-access memory; RCA: Rolling circle amplification; rRCA: Random rolling circle amplification; SAM: Sequence alignment/map; TE: Transposable element; WGS: Whole genome sequencing.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04545-2.

Additional file 1. Figure S1: Extended version of Figure 1a with detailed pipeline overview. Our pipeline consists of four main modules: one for the data preparation module and three for the analysis modules. The execution of the different modules depends on the given input data.

Additional file 2. Methods: Detailed methods for data generation (including circSeq), analysis using the ECCsplorer pipeline (including specific commands), comparison with other tools, hardware requirements, and a detailed description of each ECCsplorer pipeline modules.

Additional file 3. Tables S1–S9: Results from eccDNA detection using the ECCsplorer pipeline. (1) Case study I: Top results from mapping module. (2) Case study I: Top results from clustering module. (3) Case study I: Top results from comparative module. (4) Case study II: Top results from mapping module. (5) Case study II: Top results from clustering module. (6) Case study II: Top results from comparative module. (7) Case study II: Comparison of tools. (8) Case study III: Top results from mapping module. (9) Case study IV: Top results from clustering module.

Additional file 4. Data: Data underlying the plots in Figures 1–3. Results from comparison of tools using *H. sapiens* data can be found in Supp. Info. 4: Supplementary Data (Data_Fig_3b_comparison… 2 files).

Acknowledgements

The authors are grateful for support and mentorship from the late Thomas Schmidt. Furthermore, we thank Stefan Wanke for supporting our research initiative. We thank the laboratory of Marie Mirouze (IRD, Marseille) for kindly providing the mobilome-Seq protocol and the publicly available circSeq (mobilome-Seq) data shown in Figs. 2 and 3. Similarly, we acknowledge the laboratory of Birgitte Regenberg (University of Copenhagen) for the publicly available circSeq data shown in Figs. 2 and 3. The KWS Saat SE & Co. KGaA are acknowledged for providing plant material. We thank N. Schmidt for beta testing of the software, and G. Menzel for his support.

Authors' contributions

LM developed ECCsplorer, and wrote the manuscript. KMS provided input in the development of ECCsplorer. BW and TH conceived the study and designed the experiments. All authors were involved in discussion of the manuscript, in editing, and approved the final manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. We recognize the financial support of the German Research Foundation (DFG) granted to TH (HE 7194/2-1). In addition, KMS was supported by the German Federal Ministry of Education and Research (KMU-innovativ-18 grant 031B0224B), managed by TH. BW acknowledges funding from the German Federal Ministry of Food and Agriculture (Fachagentur nachwachsende Rohstoffe e. V.; grant 22002216). The funders had no role in the design of the study, data collection, data analysis, interpretation of results, or writing of the paper.

Availability and requirements

Project name: ECCsplorer: pipeline to detect eccDNAs; Project home page: https://github.com/crimBubble/ECCsplorer; Operating system(s): Linux; Programming language: Python (and R); Other requirements: Python 3.6 or higher, and R 3.6.3 or higher; 3rd party tools: NCBI BLAST+, Trimmomatic, seqtk, segemehl, SAMtools, BEDtools, RepeatExplorer2 (recommended installation with Anaconda3, see GitHub page); License: GNU GPL-3.0 license; Any restrictions to use by non-academics: terms stated in GNU GPL-3.0

Availability of data materials

The semi-artificial test data is available on the ECCsplorers GitHub page (https://github.com/crimBubble/ECCsplorer/testdata). The datasets together with the accession codes are as follows: circSeq (mobilome-Seq) ***A. thaliana*** (epi12) [12], accession no. ERR1830501; circSeq (mobilome-Seq) ***A. thaliana*** (WT) [12], accession no. ERR1830499; circSeq ***H. sapiens*** muscle tissue [32], accession no. SRR6315430; circSeq ***B. vulgaris*** KWS2320 (inflorescences), accession no. ERR6004146; WGA ***B. vulgaris*** KWS2320 [30], accession no. SRR869631. Source data for plots shown are included with this paper's supplementary information (Additional file 3: Tables S1–S9, or included in the original articles).

Code availability

All source code for the ECCsplorer pipeline is available at https://github.com/crimBubble/ECCsplorer under the GNU license.

Availability of data and materials

The methods for the generation of circSeq data as well as the explicit commands used in the case studies are available in the supplementary information online (Additional file 2: Supplementary Methods).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 10 August 2021 Accepted: 20 December 2021
Published online: 14 January 2022

References

1. Hotta Y, Bassel A. Molecular size and circularity of DNA in cells of mammals and higher plants. Proc Natl Acad Sci. 1965;53:356–62.
2. Liao Z, Jiang W, Ye L, Li T, Yu X, Liu L. Classification of extrachromosomal circular DNA with a focus on the role of extrachromosomal DNA (ecDNA) in tumor heterogeneity and progression. Biochimica Biophysica Acta (BBA). 2020;1874:188392.
3. Møller HD, Ramos-Madrigal J, Prada-Luengo I, Gilbert MTP, Regenberg B. Near-random distribution of chromosome-derived circular DNA in the condensed genome of pigeons and the larger, more repeat-rich human genome. Genome Biol Evol. 2019;12:3762–77.
4. Molin WT, Yaguchi A, Blenner M, Saski CA. The EccDNA replicon: a heritable, extranuclear vehicle that enables gene amplification and glyphosate resistance in *Amaranthus palmeri* [OPEN]. Plant Cell. 2020;32:2132–40.
5. Gaubatz JW, Flores SC. Tissue-specific and age-related variations in repetitive sequences of mouse extrachromosomal circular DNAs. Mutation Res DNAging. 1990;237:29–36.
6. Sinclair DA, Guarente L. Extrachromosomal rDNA circles: a cause of aging in yeast. Cell. 1997;91:1033–42.
7. Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature. 2017;543:122–5.
8. Kumar P, Dillon LW, Shibata Y, Jazaeri AA, Jones DR, Dutta A. Normal and cancerous tissues release extrachromosomal circular DNA (eccDNA) into the circulation. Mol Cancer Res. 2017;15:1197–205.
9. 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.

10. Benoit M. Glyphosate resistance decoded: the reference sequence of the extrachromosomal DNA replicon in amaranth. Plant Cell. 2020;32:2059–60.
11. Flavell AJ, Ish-Horowicz D. Extrachromosomal circular copies of the eukaryotic transposable element copia in cultured *Drosophila* cells. Nature. 1981;292:591–5.
12. Lanciano S, Carpentier M-C, Llauro C, Jobet E, Robakowska-Hyzorek D, Lasserre E, et al. Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. PLoS Genet. 2017;13:e1006630–50.
13. Diaz-Lara A, Gent DH, Martin RR. Identification of extrachromosomal circular DNA in hop via rolling circle amplification. Cytogenet Genome Res. 2016;148:237–40.
14. Paulsen T, Shibata Y, Kumar P, Dillon L, Dutta A. Small extrachromosomal circular DNAs, microDNA, produce short regulatory RNAs that suppress gene expression independent of canonical promoters. Nucleic Acids Res. 2019;47:4586–96.
15. Shibata Y, Kumar P, Layer R, Willcox S, Gagan JR, Griffith JD, et al. Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. Science (New York, NY). 2012;336:82–6.
16. Mehta D, Cornet L, Hirsch-Hoffmann M, Zaidi SSA, Vanderschuren H. Full-length sequencing of circular DNA viruses and extrachromosomal circular DNA using CIDER-Seq. Nat Protocols. 2020;15:1673–89.
17. Lanciano S, Zhang P, Llauro C, Mirouze M. Identification of extrachromosomal circular forms of active transposable elements using mobilome-seq. In: Cho J, editor. Plant transposable elements: methods and protocols. New York: Springer US. 2021. p. 87–93. https://doi.org/10.1007/978-1-0716-1134-0_7.
18. 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 Bioinform. 2019;20:663.
19. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation forStatistical Computing; 2013. https://www.R-project.org/.
20. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
21. Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, et al. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. Genome Biol. 2014;15:R34-50.
22. Novák P, Neumann P, Macas J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinform. 2010;11:378–90.
23. Novák P, Neumann P, Macas J. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nat Protoc. 2020;15:3745–76.
24. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinform. 2009;10:421.
25. Li H, Bufallo V, Murray K, Langhorst B, Klötzl F, Jain C. Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences. 2013. https://github.com/lh3/seqtk.
26. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
28. Quinlan AR. BEDTools: The Swiss-Army Tool for genome feature analysis. Curr Protoc Bioinform. 2014;47:114.
29. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72.
30. Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, et al. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2014;505:546–9.
31. Weber B, Heitkam T, Holtgräwe D, Weisshaar B, Minoche AE, Dohm JC, et al. Highly diverse chromoviruses of *Beta vulgaris* are classified by chromodomains and chromosomal integration. Mob DNA. 2013;4:8–23.
32. Møller HD, Mohiyuddin M, Prada-Luengo I, Sailani MR, Halling JF, Plomgaard P, et al. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. Nat Commun. 2018;9:1069–81.
33. The 1000 Genomes Project Consortium, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
34. Munk Hansen B, Marcker KA. DNA sequence and transcription of a DNA minicircle isolated from male-fertile sugar beet mitochondria. Nucleic Acids Res. 1984;12:4747–56.
35. Funk A, Galewski P, McGrath JM. Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome. Plant J. 2018. https://doi.org/10.1111/tpj.13977.
36. Zhang X-O, Dong R, Zhang Y, Zhang J-L, Luo Z, Zhang J, et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res. 2016;26:1277–87.
37. Deshpande V, Luebeck J, Nguyen NPD, Bakhtiari M, Turner KM, Schwab R, et al. Exploring the landscape of focal amplifications in cancer using Amplicon. Architect Nat Commun. 2019;10:392.
38. Shoura MJ, Gabdank I, Hansen L, Merker J, Gotlib J, Levene SD, et al. Intricate and cell type-specific populations of endogenous circular DNA (eccDNA) in *Caenorhabditis elegans* and *Homo sapiens*. G3 Genes Genomes Genetics 2017;7:3295–303.
39. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics (Oxford, Engl). 2009;25:1754–60.
40. Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, et al. Selective epigenetic control of retrotransposition in Arabidopsis. Nature. 2009;461:427–30.

Publisher's Note

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