https://doi.org/10.1186/s12864-023-09948-4

SOFTWARE Open Access

High-fidelity (repeat) consensus sequences from short reads using combined read clustering and assembly

Ludwig Mann[1], Kristin Balasch[1], Nicola Schmidt[1] and Tony Heitkam[1,2*]

Abstract

Background Despite the many cheap and fast ways to generate genomic data, good and exact genome assembly is still a problem, with especially the repeats being vastly underrepresented and often misassembled. As short reads in low coverage are already sufficient to represent the repeat landscape of any given genome, many read cluster algorithms were brought forward that provide repeat identification and classification. But how can trustworthy, reliable and representative repeat consensuses be derived from unassembled genomes?

Results Here, we combine methods from repeat identification and genome assembly to derive these robust consensuses. We test several use cases, such as (1) consensus building from clustered short reads of non-model genomes, (2) from genome-wide amplification setups, and (3) specific repeat-centred questions, such as the linked vs. unlinked arrangement of ribosomal genes. In all our use cases, the derived consensuses are robust and representative. To evaluate overall performance, we compare our high-fidelity repeat consensuses to RepeatExplorer2-derived contigs and check, if they represent real transposable elements as found in long reads. Our results demonstrate that it is possible to generate useful, reliable and trustworthy consensuses from short reads by a combination from read cluster and genome assembly methods in an automatable way.

Conclusion We anticipate that our workflow opens the way towards more efficient and less manual repeat characterization and annotation, benefitting all genome studies, but especially those of non-model organisms.

Keywords Repetitive DNA, Transposable elements, Consensus sequences, Repeat assembly, Repeat clustering, eccDNA, Ribosomal DNA, rDNA, Non-model organisms

Background

In the last ten years, the amount of publicly available read data has been growing exponentially from Terabytes in 2012 to over 50 Petabytes in 2023 alone, as published in the European Nucleotide Archive (ENA statistics, accessed 08/2023). This aptly reflects the recent advances in sequencing techniques. Especially the costs of short read sequencing methods are dropping drastically and long read technologies are advancing and getting more and more accurate. This results in an unprecedented volume of sequencing data and the needs for automated or at least semi-automated ways to analyse the amount of data are constantly rising. The gold standard of genomics has become to assemble a reference genome sequence; however, this can be very challenging and costly depending on the organism of interest. For species with large and complex genomes, huge amounts of data are necessary,

*Correspondence:
Tony Heitkam
tony.heitkam@tu-dresden.de
[1] Faculty of Biology, Technische Universität Dresden, D-01069 Dresden, Germany
[2] Institute of Biology, NAWI Graz, Karl-Franzens-Universität, Graz A-8010, Austria

including long reads, HiC sequencing reads or similar [1–3]. In addition to the high costs, very capable computational resources are required. Hence, especially prior to full genome sequencing projects, it is helpful to gain insight into the complexity of the target species genome in advance. One major aspect of the complexity of a genome is the fraction of repetitive DNA, which is also a main genome size determining factor [4]. Even some of the latest telomere-to-telomere genome assemblies using long read techniques lack the complete resolving of repetitive DNA. In contrast to full genome assemblies, to gain an overview of the repetitive DNA of a genome, only short reads and genome skimming are needed [5, 6]. Today, short read sequencing is still by a magnitude cheaper than long read sequencing. For genome skimming typically a low coverage of the target genome is randomly sequenced [7]. Both measures help keeping the costs fairly low, while already gaining deep knowledge on the repetitive DNA fraction.

The repetitive DNA is divided into two main fractions – tandem repeats and dispersed repeats. Tandem repeats such as satellite DNA or ribosomal DNA can form long arrays with up to thousands of copies spanning megabases in length [8–10]. Disperse repeats on the other hand are typically spread throughout the genome and are mobile. They can be further divided into two main groups – DNA transposons and retrotransposons with sizes ranging from 100 bp up to 20 kbp per copy. Current research on repetitive DNAs highlights their roles in providing genome structure, regulating transcription and pushing evolution [11–15]. One widely applied tool to identify, classify and, to a certain extent, quantify repetitive DNA from genome skimming data is the RepeatExplorer2 pipeline (RE2) [16]. Although it produces an excellent overview of the repetitive DNA fraction, building trustworthy repeat consensuses from its output using short reads is still a very manual, time-consuming and not reproducible process, relying mainly on the experience of the user. The RE2 output includes highly informative cluster graphs, but rather fragmented contigs of different repeat families such as transposable elements. Thus, building a comprehensive repeat consensus database with high-fidelity consensuses spanning complete or nearly complete transposable elements would be very beneficial to reduce the workload for manual curation.

Moreover, consensus reconstruction is not only applicable for whole genome shotgun sequencing data. It can also be applied to detect enriched DNA in genome-wide amplification setups. Here, we further explore our consensus building pipelines to identify and reconstruct extrachromosomal circular DNAs (eccDNAs). EccDNAs are ring-like DNAs that are physically separated from the chromosomes and have received much interest in recent years. There is still only little known about their function, but eccDNAs have been assumed to be related to aging, cancer and transposable element activation [18–21]. Here, we explore consensus-building in eccDNA circle reconstruction, building on our eccDNA identification pipeline, the ECCsplorer [17]. Similar to the default usage of RE2, the resulting ECCsplorer contigs represent rather fragmented than complete circular sequences. Therefore, reconstructing a consensus would improve the analysis of eccDNAs, since the ECCsplorer pipeline is the only published method for the *de novo* identification of eccDNAs from only short reads so far. To overcome the fragmentation of contigs, we here combine repeat clustering by RE2/ECCsplorer with assembly tools and visualization to derive high-fidelity repeat/eccDNA consensuses.

Furthermore, to complement the combination of tools, we investigate the usage of genome skimming data with only assembly tools to analyse structural features of repetitive DNA, such as linkage and separation of the highly abundant rDNA. Typically, the tandem repeats of 5S rDNA and 35S rDNA are organized in separate arrays, often on different chromosomes. In some species, however, the 5S rDNA gene is integrated into the 35S rDNA spacers, forming a so-called linked arrangement [22].

To demonstrate all mentioned strategies, we set up three different use cases. First, we use assembly tools on the default RE2 output to build comprehensive repeat consensus sequences from *Beta corolliflora*. *Beta corolliflora* is closely related to cultivated beet varieties (*Beta vulgaris*) with a well-studied genome and a deeply curated repeat annotation. Second, we reconstruct mitochondrial mini-circles from the *Beta vulgaris* ssp. *vulgaris* (*Beta vulgaris*) genome derived from a previous ECCsplorer output with enriched sequencing data (eccDNASeq). The mitochondrial minicircles are a well characterized positive control for existing eccDNA harbouring repetitive and non-repetitive sequence features. Last, we investigate the presence or absence of rDNA linkage in the two Asteraceae species *Artemisia annua* and *Tragopogon porrifolius*. The linkage of rDNA is a species-specific, structural feature which is here analysed exemplarily. For all of these use cases there is public data available. We examine all three use cases and assess the performance of the suggested workflow. Based on this, we conclude that our explorative analysis is a useful tool to semi-automatize the analysis of repetitive DNA in non-model genomes.

Methods and implementation

The presented repeat assembly workflow uses clustering and assembly tools to create informed consensus sequences from repeats to answer a wide variety of questions. We use the term "informed consensus" to suggest that the derived sequences are not mere averages or sequence profiles, but that they have been carefully constructed using relevant data and analysis. The workflow can be used to generate initial repeat databases with comprehensive consensus sequences, reconstruct eccDNA sequences from eccDNA-Seq or provide insights into structural features of highly abundant repeats (Fig. 1).

Create comprehensive consensus sequences from repeat clusters

This part of the repeat assembly workflow (Fig. 1, green) uses genome-skimming data (short or long reads, example data: short reads from *Beta corolliflora*, 0.1× genome representation) that already has been clustered by RE2 [4, 16] to create comprehensive consensus sequences using a combination of MEGAHIT [23] and SPAdes [24].

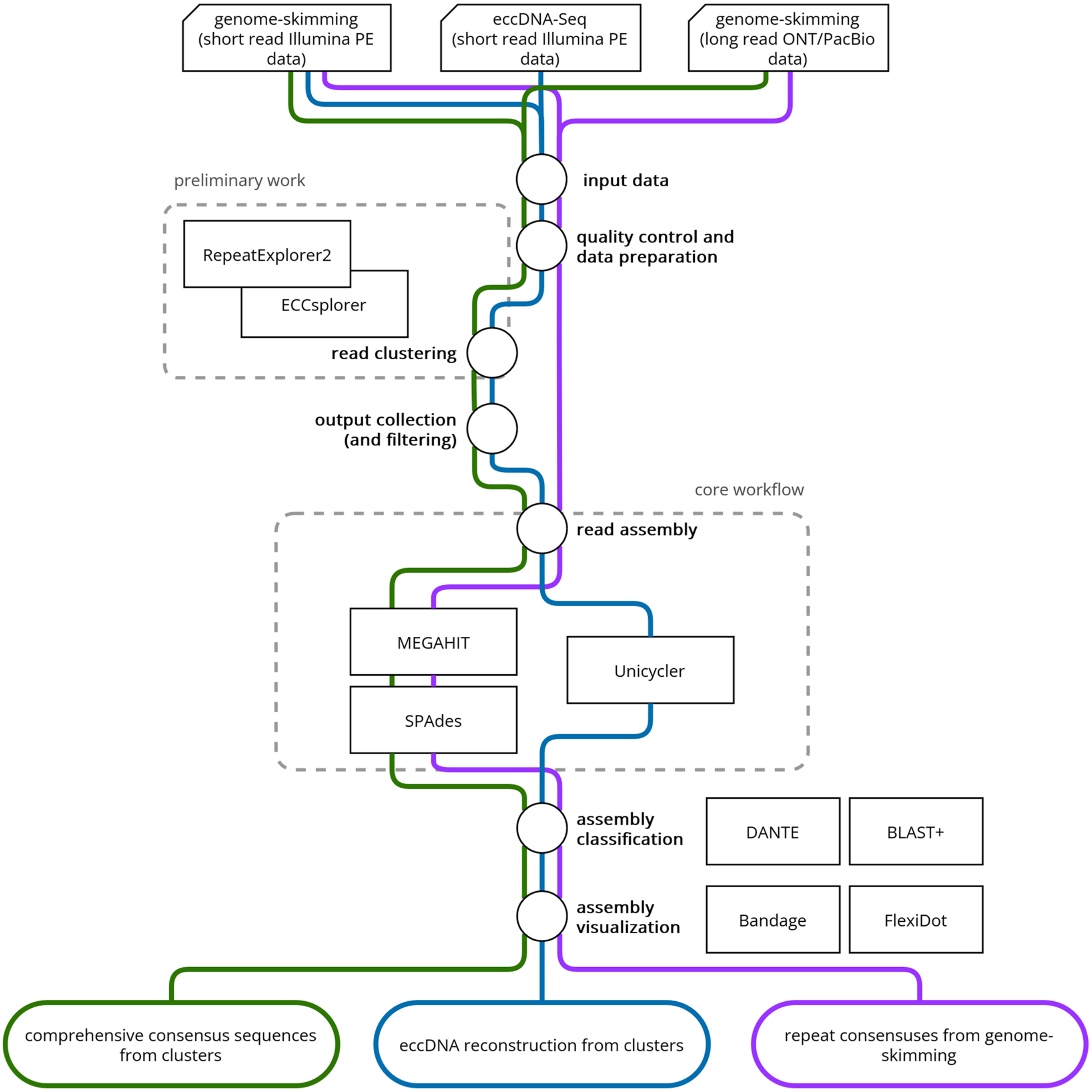
genome-skimming (short read Illumina PE data)
eccDNA-Seq (short read Illumina PE data)
genome-skimming (long read ONT/PacBio data)
input data
preliminary work
quality control and data preparation
RepeatExplorer2
ECCsplorer
read clustering
output collection (and filtering)
core workflow
read assembly
MEGAHIT
SPAdes
Unicycler
assembly classification
DANTE
BLAST+
assembly visualization
Bandage
FlexiDot
comprehensive consensus sequences from clusters
eccDNA reconstruction from clusters
repeat consensuses from genome-skimming

Fig. 1 Overview of the repeat assembly workflow. Guided workflows for the creation of comprehensive consensus sequences from repeat clusters (green), the reconstruction of circular sequences from eccDNA candidate clusters from whole genome amplification methods (blue) and the assembly of consensus sequences from highly abundant repeats to explore structural features to solve specific repeat-derived questions (purple)

After quality control and trimming, the data was processed according to the RE2 protocol using the web-based (galaxy) version or a local installation (--keep-names enabled). Manual correction of the annotation was performed according to the RE2 online resources (repeatexplorer.org online tutorials). Reads from the superclusters were collected as fasta, fastq, or as 'contigs-based' fastq reads in the following way: Reads in fasta format were used from superclusters as is. Reads in fastq format have been collected from the raw reads using the supercluster reads as reads list. Contig-based reads have been collected from raw read data using only reads that mapped against the supercluster contigs with bowtie2 [25]. Each supercluster read set in each format has been assembled using MEGAHIT (meta-large preset) in a first assembly round. The second assembly round was done with SPAdes (--isolate, --cov-cutoff auto) using the final MEGAHIT contigs and the supercluster contigs as trusted contigs. For a more detailed and guided workflow and access to custom python scripts see the GitHub repository (https://github.com/crimBubble/repeats_and_circles_assembly).

For the example data of *Beta corolliflora*, the statistical data of assembled sequences was collected with seqkit [26] and visualized with ggplot2 [27]. Assembled sequences from reads in fasta format were compared graphically using FlexiDot [28] and sequence-wise using blast+ [29] with the original RE2-contigs, ONT long reads (representing real repeat sequences) and a repeat data base (see section Availability of data). Protein domains have been annotated with DANTE [30] according to the domains specified in the REXdb [31]. Additionally, a score has been calculated for each consensus sequence, defined as length (bp) multiplied by coverage (as reported by SPAdes or RE2). For each repetitive element (represented by a cluster), the ten consensus sequences with the highest scores were considered the most representative ones and have been selected for comparative statistics and visualizations.

Reconstruct circular sequences from eccDNA candidate clusters

The second part of the repeat assembly workflow (Fig. 1, blue) reconstructs complete circular sequences following the ECCsplorer pipeline [17] with Unicycler [32]. For this, eccDNA-Seq data is used as input (short read, example data: short reads from *Beta vulgaris*, circle-enriched and control data). After quality control and trimming, data were processed according to the ECCsplorer protocol. Manual correction of the annotation and filtering was performed (note that filtering of mitochondrial clusters was omitted compared to the detailed instructions). From the eccDNA candidate superclusters, reads were collected as fastq reads and as contigs-based fastq reads in two ways: Reads in fastq format have been collected from the raw reads using the supercluster reads as reads list. Contig-based reads have been collected from raw reads, using only those that mapped against the supercluster contigs with bowtie2 [25]. For each eccDNA candidate supercluster, the reads in both formats and a combination of both read sets have been assembled with Unicycler in normal mode (--min_fasta_length 1, --keep 2). Additionally, circles were detected with a custom python script using the networx package. For a more detailed and guided workflow and access to custom python scripts see the GitHub repository (https://github.com/crimBubble/repeats_and_circles_assembly).

For the example data of *Beta vulgaris*, the assembled sequences were compared graphically using FlexiDot [28] and the Bandage assembly viewer [33] with the NCBI reference sequences of the *Beta vulgaris* mitochondrial mini-circles a, d, pO.

Assemble consensus sequences from highly abundant repeats and explore structural features

The last part of the repeat assembly workflow (Fig. 1, purple) aims at exploring specific questions regarding repeat organization and structure, such understanding linkage or separation of certain repetitive DNAs. Here, we combine MEGAHIT [23] and SPAdes [24] to directly run on genome-skimming reads (short or long reads possible; example data: short reads from *Artemisia annua* and *Tragopogon porrifolius*). After quality control and trimming, data were directly assembled with MEGAHIT (meta-large preset) in a first assembly round. A second assembly round was done with SPAdes (--isolate, --cov-cutoff 20) using the final MEGAHIT contigs as trusted contigs. For a more detailed and guided workflow and access to custom python scripts see the GitHub repository (https://github.com/crimBubble/repeats_and_circles_assembly).

Results and discussion

Use case 1: Repeat assembly from *Beta corolliflora* superclusters reveals improved continuity of consensus sequences representing real repetitive elements

The first use case in this study represents the capability of the described workflow to create high-confidence and comprehensive consensus sequences from repetitive elements. Here, we use available data from the recently sequenced *Beta corolliflora*, a wild beet species and close relative of the crop plant sugar beet (*Beta vulgaris*). The consensus sequences have been assembled after clustering of the read data with RE2 [16].

There are multiple ways to retrieve reads from RE2 superclusters for use in our workflow (also see Methods and Implementation). These three ways extract different levels of read information, as follows: (i) "fasta": read information only; (ii) "fastq": read and quality information; (iii) "contig-based": read and quality information based on RE2-contigs from original data (non-sub-sampled data). All three ways have been tested and compared against each other (Fig. 2a; with the colour gradient from orange to green representing repeat abundance). Overall, reads from superclusters in fasta and fastq format performed very similarly, with the fastq reads showing a slight advantage for large superclusters (high read count; Fig. 2a, shades of orange). The contig-based reads showed a higher number of sequences (count of contigs per supercluster) along with the related low N50 values. However, for smaller superclusters (low read count; Fig. 2a, shades of green), the contig-based reads resulted in better consensuses as the standard reads sometimes failed to produce consensus sequences at all. In some cases, SPAdes failed to produce appropriate (at least 100 bp long) consensus sequences (NODEs, e.g. Figure 2b SCL012). In such cases, the MEGAHIT output usually contained consensuses that instead may be used for downstream analysis (not shown). Taken together, depending on the abundance and characteristics of the repeats of interest, either the supercluster reads in fastq or the contig-based reads show the most useful results. A combination of both read sets can yield additional information (see use case 2). Hence, we conclude that while read input does strongly influence the results, the appropriate choice of reads can be different for individual repetitive elements. As our workflow can be heavily parallelized, all setups can be tested in short time providing the optimal results for each genomic element. In the following, for evaluation of the results, we use read input option (i) "fasta".

To assess overall performance of our assembly workflow, we compare the newly generated consensus sequences, in the following referred to as NODEs, with the corresponding RE2-contigs. The Top Ten best-scoring NODEs were overall longer than the corresponding RE2-contigs. Only if NODE generation failed, RE2-contigs were longer (Fig. 2b, side-by-side comparison). Usually, both showed sequence overlaps with a variable length and high sequence similarity (0.4–17.7 kb; 88–100% identity; see Fig. 2b, length and shading of the central blue bar). For most superclusters, at least one long NODE was produced that was significantly longer than the longest comparable RE2-contig, hence, resulting in a higher continuity of the consensus sequence. Comparing all of the 100 most abundant repetitive DNAs, represented by the Top 100 superclusters, we find that 63 produced more continuous NODEs (Fig. 2b, left, indicated with a red dot) as opposed to 37 more continuous RE2-contigs (Fig. 2b, right, indicated with a green dot).

Exemplarily, we illustrate both, NODEs and RE2-contigs for supercluster 8 (SCL8), representing an Angela-type LTR retrotransposon family (Fig. 2c). In the dotplots, the high fragmentation of the RE2-contigs becomes visible, whereas, for the NODEs, only two sequences are needed to represent the complete retrotransposon (NODE_1_r and NODE_9, Fig. 2c). To further verify the accuracy of the NODEs, we compared them to real repeat representatives (derived from ONT long reads) and to consensus sequences from a *Beta vulgaris* repeat library. In the representative example of SCL8, a real Angela (Tork)-type retrotransposon is highly similar to the calculated NODE sequence (Fig. 2c). Similar results were observed for most superclusters especially for the more continuous NODEs (indicated by a red dot in Fig. 2b). More examples including different types of repetitive elements are collected in the Supplementary dotplots file (Additional file 1).

Investigating the architecture of the NODEs more closely, we find that the longest NODEs mostly represent the more conserved parts of a repetitive element, which are usually the protein-coding domains in the case of transposable elements, and the genes in case of rDNAs. The long terminal repeats (LTRs) of the most abundant repetitive elements in plant genomes are less conserved over single retrotransposon families and are present in NODEs with lower scores (i.e., higher NODE numbers, such as NODE_9 of SCL008, Fig. 2c). However, even these less conserved parts are assembled using the presented workflow and usually can be visualised in the bandage graph representations as regions with more branching.

Overall, the assembled NODEs can serve as an additional layer of information when analysing the repetitive DNA of non-model organisms and can be used as a database for transposable elements in following genome assemblies or in the analysis of closely related species. The consensus sequences are especially useful for identifying and classifying individual copies of repetitive elements in subsequent analyses. The consensus sequences can serve as a repeat database reducing the amount of manual work by (a) providing more continuous sequences and by (b) reducing the amount of sequences which need manual curation. Moreover, a deeper knowledge on the repetitive fraction of a genome might also help to design whole genome assembly strategies: Since this fraction is a major hitch in most assembly approaches, educated decisions on suited technologies and appropriate sequencing coverages are profound and can be addressed with the presented workflow. There is a huge difference in needed

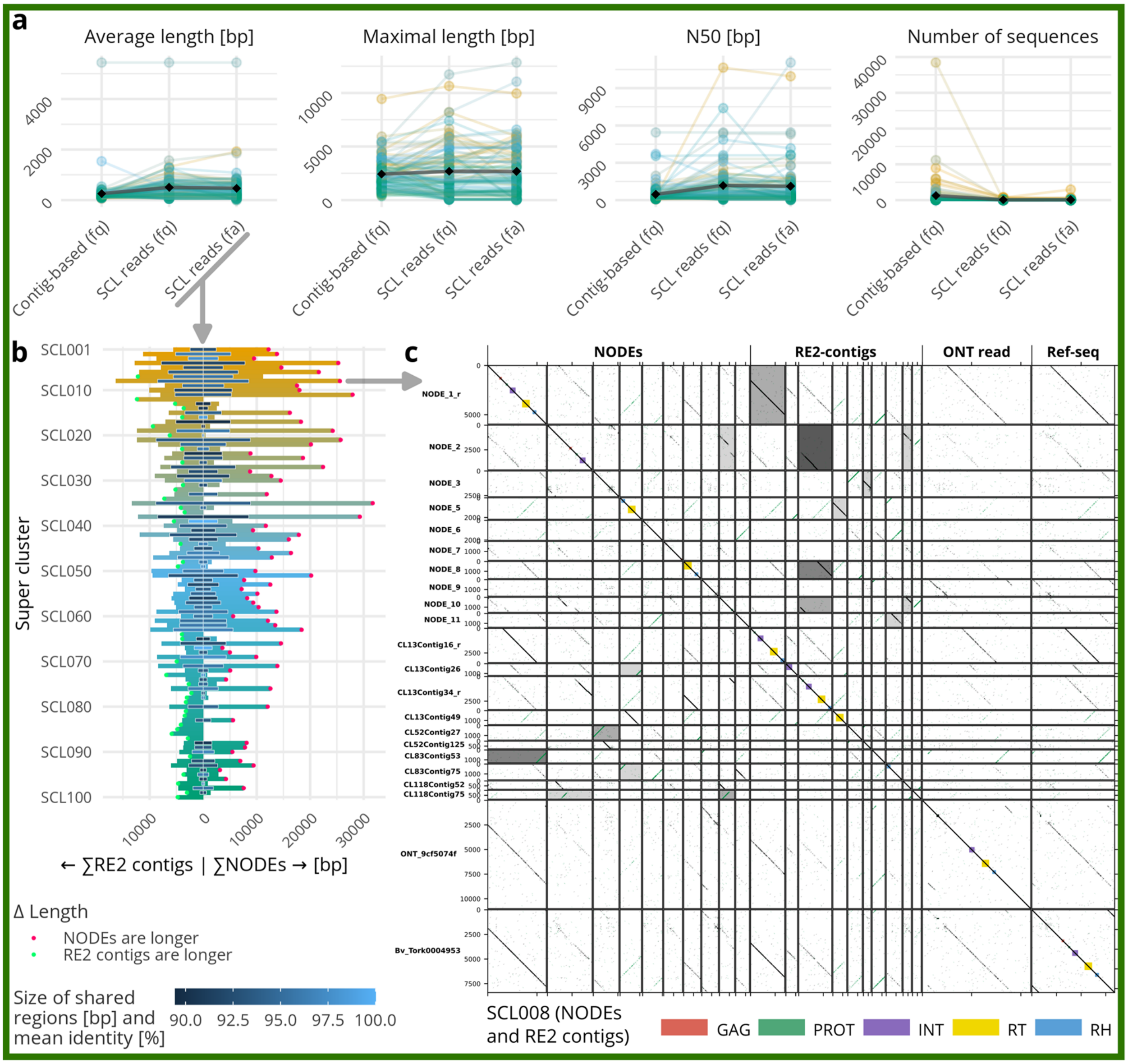

Fig. 2 Automated repeat assembly from RepeatExplorer2 (RE2)-derived superclusters (SCLs) leads to long, continuous and accurate repeat consensus sequences. **a** Impact of read input: Comparison of repeat assembly quality measures, when using different input reads for the MEGAHIT/SPAdes assembly workflow. There are three different ways to collect input reads for the assembly workflow: (i) Reads in fasta format can be directly used from the superclusters (SCL reads; fa); (ii) reads in fastq format can be selected from the original read data using the reads listed in the supercluster (SCL reads; fq); (iii) or the reads in fastq format can be selected from the original data based on their similarity to the supercluster contigs (contig-based; fq). Each colour represents a supercluster, displaying the 100 largest superclusters, with colour being on a continuous scale as outlined in Fig. 2b. Black diamonds show mean values. **b** Impact of the new MEGAHIT/SPAdes workflow on consensus length: For the most abundant 100 superclusters (SCLs), we compared the combined length of the assembled consensuses (NODEs; right-facing bars) and RE2-contigs (left-facing bars) using the ten best-scoring contigs/NODEs of each supercluster. The central blue bar indicates the length of the shared sequences between both, NODEs and RE2-contigs, whereas the depth of the blue shade indicates their mean sequence identity. If the NODE assembly produces longer consensuses, this was indicated by a red dot, whereas a superior RE2 assembly was marked by a green dot. **c** The accuracy of the generated consensus is illustrated by an in-depth view into a selected repeat family, an Angela-type retrotransposon, represented by supercluster 8 (SCL8): Dotplot comparison of the 10 highest-scored NODEs and RE2-contigs from supercluster 8 (SCL008), as well as an ONT long read with an actual Angela copy and a reference sequence for a more detailed sequence-wise comparison. The shading refers to the longest common subsequence (LCS), in which darker grey indicates a longer sequence overlap. In the upper part (above the main diagonal) the forward LCS and in the lower part (below the main diagonal) the reverse LCS is used for shading. The ending "_r" indicates sequences that are displayed as reverse complement

data for a genome with, e.g., many conserved sequences in tandem versus a genome with, e.g., many dispersed and less conserved transposon sequences. By providing a global repeat analysis, the RE2 combined with assembly tools can be used to make an optimal use of sequencing budgets by leveraging the different advantages of the latest sequencing technologies such as whale-long ONT reads, high-fidelity PacBio reads, or structural advanced Hi-C conformation capturing.

Use case 2: Automated reconstruction of mitochondrial mini-circles from *Beta vulgaris* following the analysis with the ECCsplorer pipeline

The ECCsplorer pipeline is a tool for the detection of eccDNA that is useful for non-model-organisms as it does not rely on a reference genome assembly. Instead, two short read datasets are compared – one is enriched in eccDNAs, whereas the other is a canonical whole genome shotgun read set. To determine the enriched eccDNA candidates, the ECCsplorer uses RE2 for clustering, similar to use case 1. Hence, it mostly lacks the discovery of complete circular sequences that represent closed DNA rings. Using the presented repeat assembly workflow in addition to the ECCsplorer pipeline solves this problem. To demonstrate this, we use an exemplary dataset with well-characterized eccDNAs [17]. This dataset is from sugar beet (a *Beta vulgaris* cultivar) and is enriched in so-called mitochondrial mini-circles [17]. There are three beet mini-circles in total, named a, d, and p0 that contain shared regions [34, 35]. Running the ECCsplorer on this sample dataset yields a single supercluster with sequences from all three mini-circles, though reconstruction of full circular sequences is failing. Now, with the new MEGAHIT-SPAdes assembly workflow, we could fully retrieve all three complete circular sequences from just this one single supercluster (Fig. 3). The assembled circles are almost identical to the references, sequenced in the 1980s [34, 35], only varying by a few SNPs. The origin of the SNPs cannot be traced back exactly but may be a biological variation.

Overall, this shows that the presented workflow is capable of discovering complete circular sequences from eccDNA-Seq data that has been clustered with the ECCsplorer (and RE2) without the need of any reference sequence. This will be especially helpful for the analysis of chimeric eccDNAs (single eccDNAs with sequences parts from multiple origins), which have been reported recently [36, 37]. Further, we believe that also the bandage graphs and coverage information produced during the workflow can assist in the understanding and differentiation of real chimeric eccDNAs versus template switches of the phi29 polymerase (commonly used during the amplification step of eccDNA-Seq) [38–40], without the immediate need for long read sequencing. Template switches might show similar clustering results as the overlapping mini-circles, but our workflow was still able to differentiate between all three individual circular sequences. Therefore, we predict that using the ECCsplorer followed by the presented workflow is an improvement over current eccDNA assembly methods.

Use case 3: Exploring the linkage and separation of rDNA in Asteraceae species

The repeat assembly workflow introduced here, may also be helpful in resolving unusual repeat organization patterns. For example, among the Asteraceae, some species harbour an unusual linkage of the 5S and 35S rDNAs [22]. This linkage is often confirmed by fluorescence in situ hybridisation (FISH), a powerful, but time-consuming and complex method, relying on specialized equipment and trained staff. With the recent unprecedented volume of sequencing data, the question of rDNA linkage or separation can also be answered using our workflow. The repeat assembly workflow represents a quick and reliable method to confirm linkage or separation of rDNA including the possibility to create consensus sequences. Here, we used data from two Asteraceae species with one of them showing a known linkage of rDNA (*Artemisia annua*, L-type) and the other showing a separation of 5S and 35S rDNA (*Tragopogon porrifolius*, S-type). From the resulting bandage graphs the linkage in *Artemisia annua* and separation in *Tragopogon porrifolius* is clearly visible (Fig. 4). Additionally, the rDNA genes are highly conserved, whereas the spacer regions show some higher variability, indicated by branching. The variability in *Tragopogon porrifolius* is even higher and it might be possible that there are some TE insertions in some 35S rDNA copies which would not be unusual. Furthermore, there is also some variation in the rDNA genes of *Tragopogon porrifolius* that may reflect the additional rDNA loci of both 35S and 5S rDNA [41]. Of course, many derived questions that target homogenized repeat co-occurrences can be targeted as well by the proposed repeat assembly and visualization algorithm (e.g. transposable elements in rDNA, or similar).

Advantages and limitations

Overall, the present workflows build on existing and well-established methods and extend them in a reproducible manner to provide guidance for follow-up analyses. Manual workloads are reduced wherever possible, allowing automation of the time-consuming manual repeat curation processes.

The new repeat assembly building on RE2 supercluster information assists and automates the manual work that is usually carried out after an RE2 run. Our workflow

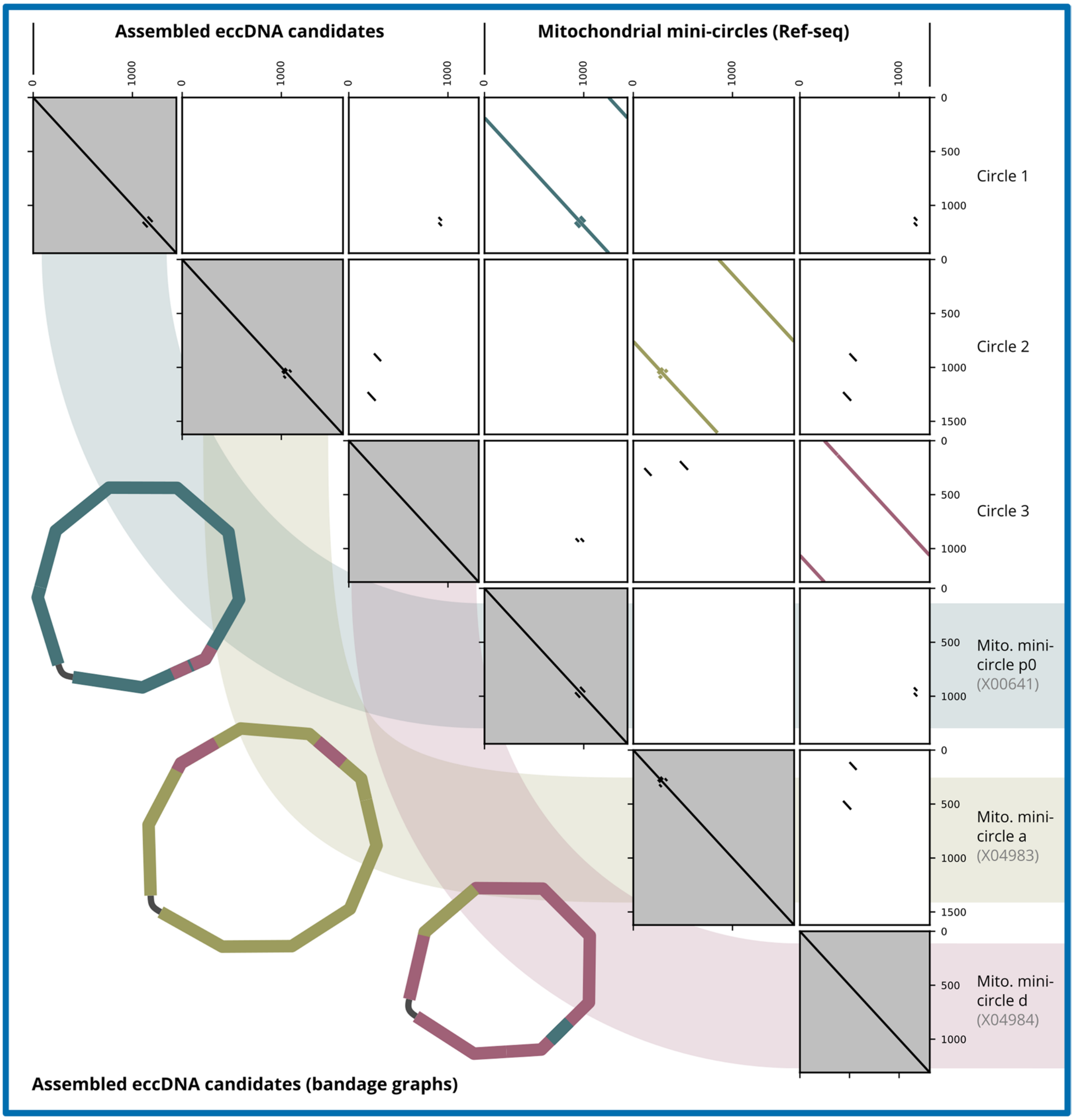

Fig. 3 Mitochondrial mini-circles assembled using ECCsplorer (clustering module) output. The dotplots display very high similarity between reconstructed circles and the published reference sequences. The bandage graph representations show that the full circular sequences could be assembled despite sequence similarities between the three mini-circles (indicated by colours)

provides a consistent methodology and is less dependent on previous knowledge about the analysis. Therefore, the presented workflow represents an easy and time-saving way to gain more information from existing data. However, even with these advances, the presented workflow may not always provide finalised, complete consensus sequences, but rather more continuous sequences compared to the original contigs. To reach the best consensus sequences, it may be necessary to try different read collection methods as instructed in use case 1 (Fig. 2a). The completeness of individual elements (represented in one supercluster) is, as to be expected, very dependent on the initial clustering and the characteristics the individual repeat. Therefore, less preserved repeats

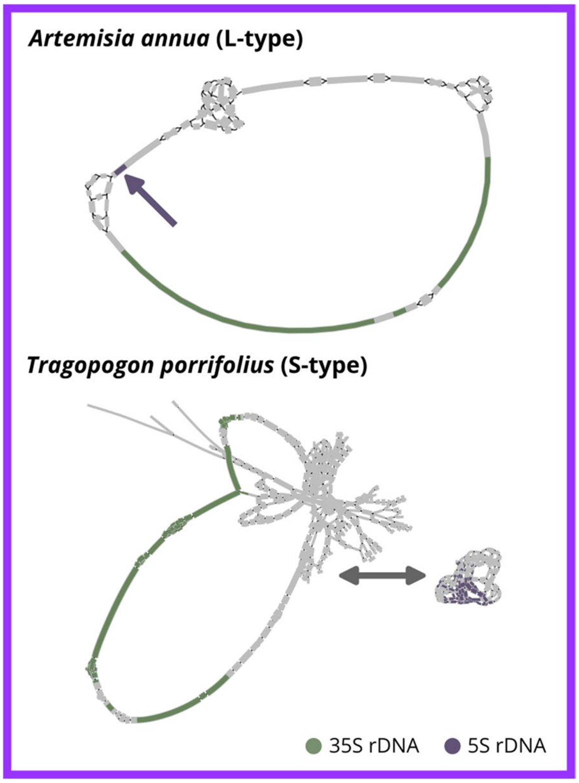

Fig. 4 Graphical representation of the repeat consensus assembly (bandage graphs) from genome-skimming data confirms the linkage or separation of rDNA in *Artemisia annua* and *Tragopogon porrifolius*

e.g., ancient repeats, will be less present in the workflow results. Further, the workflow does not provide a one-for-all solution for building a repeat database. It rather outputs a collection of consensus sequences and semi-automated classifications for each repetitive element (represented in one supercluster) which can be used as-is in a database or be further refined. The refinement can be based on the calculated scores, other metrics e.g., longest NODEs, element representation or depending on individual needs.

The reconstruction of circular sequences from the clustering workflow provides the only described method for the *de novo* detection of eccDNA sequences from short read eccDNA-Seq data so far. Nevertheless, if multiple circles cluster together, the reconstruction still can be challenging and might need an experienced user. However, this is, as demonstrated, still possible and opens up new possibilities for the eccDNA detection and classification.

Exploring structural, chromosomal or chromatin-related features of highly abundant repeats can be quite challenging and often needs time- and cost-intensive experimental methods such as FISH. Recently Garcia et al. [42] showed how structural repeat features can also be detected by RE2, however, using the presented workflow speeds up the process. Bypassing the RE2 approach completely means that the automated repeat classification is missing. However, this allows for a larger read data input (up to 1× coverage without any observed issues), and thus enables the generation of even the most complex rDNA consensus sequences [43]. Also, the use of long reads is possible without any limitations compared to the RE2 approach.

Conclusion

In this study, three related workflows are presented to assemble repeats and DNA circles from genome skimming and enrichment strategies, such eccDNA-Seq.

Repeats remain one of the major challenges in genome assembly, despite their analysis harboring great potential for understanding genome organization, regulation and evolution. With the presented workflow, we offer a method to generate robust and useful repeat consensuses without extensive manual work. Furthermore, it complements the already existing methods for repeat identification, classification and annotation such as RepeatMasker, LTR_finder, or REPET, which often focus on analysing already assembled reference genome sequences [44–46]. We conclude that our repeat assembly approach can add large value to read clustering methods that usually only provide an assortment of shorter contigs (use cases 1 and 2). Additionally, repeat assembly without read clustering (use case 3) serves as a faster alternative to answering specific repeat-related questions and to giving insights into structural features. We predict that this workflow will be even more reliable with the upcoming highly accurate sequencing techniques such as PacBio's Onso short read system.

Extending the usefulness of our repeat assembly approach, we also test it for the *de novo* detection of complete DNA circles. This is useful to understand chimeric eccDNAs and to analyse eccDNAs in non-model organisms. As for the other use cases, the presented workflow builds on existing *de novo* identification and provides examples for advanced uses of the ECCsplorer pipeline. Hence, this approach will generate new insights by providing complete circular consensus sequences for eccDNA candidates.

Overall, the three presented approaches are useful to automate workloads in repeat identification, characterization and annotation, and to manage the recent surge in data volume.

Availability and requirements

Project name: Repeats and circles assembly workflow.

Project home page: https://github.com/crimBubble/repeats_and_circles_assembly.

Operating system(s): Linux (tested on Ubuntu 20.04 LTS and 22.04 LTS).

Programming language: Shell, Python.
Other requirements: see GitHub repository.
License: GLP-3.0.

Abbreviations

(r)RCA	(random) rolling circle amplification
A. annua	*Artemisia annua*
B. corolliflora	*Beta corolliflora*
B. vulgaris	*Beta vulgaris* ssp. *vulgaris*
chimeric DNA	Structural observation in eccDNA where multiple sequences with different genomic origins are recombined in a single circle of DNA
cluster	Conglomeration of reads based on sequence similarities. Also see: Novák *et al.*, 2020 [16]
eccDNA	Extrachromosomal circular DNA
eccDNA-Seq	Illumina-sequencing of RCA amplified circular DNA (synonyms circSeq, mobilome-seq)
FISH	Fluorescence *in situ* hybridisation
genome-skimming	Untargeted, shallow (low coverage: 0.1–10× coverage) sequencing of genomic DNA of a target species. Results in comparatively deep representation of high-copy genome fractions (plastome, mitogenome and repetitive elements). Also see: Dodsworth, 2015 [7]
LTR	Long terminal repeat
ONT	Oxford nanopore technologies (long read technology)
PacBio	Pacific Biosciences (long read technology)
PE	Paired-end (sequence reads)
RE2	RepeatExplorer2
read	Sequenced DNA fragment
raw read	Here: unchanged DNA fragment subsequent to sequencing (typically in fastq format, no trimming or other processing)
long read	Long DNA fragment (typically > 1500 bp). Commonly produced by PacBio or Oxford Nanopore Technologies sequencing
short read	Short DNA fragment (typically < 250 bp) Commonly produced by Illumina sequencing
T. porrifolius	*Tragopogon porrifolius*
TE	Transposable element
template switching	Spontaneous and erroneous exchange of DNA strands during amplification with a polymerase. Also see [38–40]
WGS	Whole genome sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09948-4.

Additional file 1. Collection of dotplots similar to Fig. 2c from various repetitive elements (each represented by a supercluster).

Additional file 2. Underlying data from Figs. 2, 3 and 4. Contents: Data-Fig. 2A_B-corolliflora_inputs.tsv, Data-Fig. 2B_B-corolliflora_contigs.tsv, Data-Fig. 2C_B-corolliflora-SCL008.fasta, Data-Fig. 2C_B-corolliflora-SCL008.gff, Data-Fig. 3_B-vulgaris-mini-circles_assembly.gfa, Data-Fig. 4_A-annua-rDNA.gfa, Data-Fig. 4_T-porrifolius-rDNA.gfa.

Acknowledgements

We thank the laboratory of B. Weisshaar, and D. Holtgräwe (CeBiTec, Bielefeld) for the publicly available long and short read data used in use case 1. We thank S. Garcia, S. Maiwald, and L. Baader for support and data acquisition of the analysed *Asteraceae* species. Computational resources were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure.

Authors' contributions

LM wrote the manuscript. LM and KB developed the workflows and wrote the bioinformatic scripts. NS performed the RE2 analysis and provided insights on the repeat landscape. LM and TH conceived the study and designed the experiments. All authors were involved in discussion and editing of the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. Open Access funding was enabled and organized by the Sächsische Staats- und Universitätsbibliothek and the Technische Universität Dresden. This work was partially funded by Bundesministerium für Bildung und Forschung (BMBF 031B1221A) and Deutsche Forschungsgemeinschaft (DFG 433081887), awarded to TH (HE 7194/2 – 1). The funders had no role in the design of the study, data collection, data analysis, interpretation of results, or writing of the paper.

Availability of data and materials

The datasets together with the accession codes are as follows: *Beta corolliflora* (WGS, Illumina short reads): ERR6110441 (PRJEB45680) [47], the according RepeatExplorer2 run is available at: doi: https://doi.org/10.5281/zenodo.7821055 [4]; *Beta corolliflora* (WGS, ONT long reads): ERR10684093 – ERR10684133 (PRJEB56520) [48]; *Beta vulgaris* (repeat data base): https://doi.org/10.5281/zenodo.8255813; *Beta vulgaris* (eccDNA-Seq, Illumina short reads): ERR6004146 (PRJEB45524) [17]; *Beta vulgaris* (WGS, Illumina short reads): SRR869631 (PRJNA41497) [49]; *Artemisia annua* & *Tragopogon porrifolius* (WGS, Illumina short reads): ERR11535563 and ERR11535566 (PRJEB63080) [43].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 26 October 2023 Accepted: 28 December 2023
Published online: 24 January 2024

References

1. Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, et al. The complete sequence of a human Y chromosome. Nature. 2023;621:344–54.
2. Sun H, Jiao W-B, Krause K, Campoy JA, Goel M, Folz-Donahue K, et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat Genet. 2022;54:342–7.
3. Jayakodi M, Golicz AA, Kreplak J, Fechete LI, Angra D, Bednář P, et al. The giant diploid faba genome unlocks variation in a global protein crop. Nature. 2023;615:652–9.
4. Schmidt N, Sielemann K, Breitenbach S, Fuchs J, Pucker B, Weisshaar B, et al. Repeat turnover meets stable chromosomes: repetitive DNA sequences mark speciation and gene pool boundaries in sugar beet and wild beets. Plant J. 2023. https://doi.org/10.1111/tpj.16599.
5. Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot. 2012;99:349–64.
6. Vitales D, Garcia S, Dodsworth S. Reconstructing phylogenetic relationships based on repeat sequence similarities. Mol Phylogenet Evol. 2020;147:106766.
7. Dodsworth S. Genome skimming for next-generation biodiversity analysis. Trends Plant Sci. 2015;20:525–7.
8. Heslop-Harrison JS, Schwarzacher T. Organisation of the plant genome in chromosomes. Plant J. 2011;66:18–33.
9. Garrido-Ramos MA. Satellite DNA in plants: more than just Rubbish. Cytogenet Genome Res. 2015;146:153–70.

10. Garcia S, Kovařík A, Leitch AR, Garnatje T. Cytogenetic features of rRNA genes across land plants: analysis of the plant rDNA database. Plant J. 2017;89:1020–30.
11. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14:49–61.
12. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19:199–211.
13. Wells JN, Feschotte C. A Field Guide to eukaryotic transposable elements. Annu Rev Genet. 2020;54:539–61.
14. Mhiri C, Borges F, Grandbastien M-A. Specificities and dynamics of transposable elements in land plants. Biology. 2022;11:488.
15. Gebrie A. Transposable elements as essential elements in the control of gene expression. Mob DNA. 2023;14:9.
16. Novák P, Neumann P, Macas J. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nat Protoc. 2020;15:3745–76.
17. Mann L, Seibt KM, Weber B, Heitkam T. ECCsplorer: a pipeline to detect extrachromosomal circular DNA (eccDNA) from next-generation sequencing data. BMC Bioinformatics. 2022;23:40.
18. Noer JB, Hørsdal OK, Xiang X, Luo Y, Regenberg B. Extrachromosomal circular DNA in cancer: history, current knowledge, and methods. Trends Genet. 2022;38:766–81.
19. Peng H, Mirouze M, Bucher E. Extrachromosomal circular DNA: a neglected nucleic acid molecule in plants. Curr Opin Plant Biol. 2022;69:102263.
20. Koche RP, Rodriguez-Fos E, Helmsauer K, Burkert M, MacArthur IC, Maag J, et al. Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. Nat Genet. 2020;52:29–34.
21. Meinema AC, Marzelliusardottir A, Mirkovic M, Aspert T, Lee SS, Charvin G, et al. DNA circles promote yeast ageing in part through stimulating the reorganization of nuclear pore complexes. Elife. 2022;11:e71196.
22. Garcia S, Panero JL, Siroky J, Kovarik A. Repeated reunions and splits feature the highly dynamic evolution of 5S and 35S ribosomal RNA genes (rDNA) in the Asteraceae family. BMC Plant Biol. 2010;10:176.
23. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. Bioinforma Oxf Engl. 2015;31:1674–6.
24. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol J Comput Mol Cell Biol. 2012;19:455–77.
25. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
26. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11:e0163962.
27. Wickham H. ggplot2: elegant graphics for data analysis. Springer; 2016.
28. Seibt KM, Schmidt T, Heitkam T. FlexiDot: Highly customizable, ambiguity-aware dotplots for visual sequence analyses. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty395.
29. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
30. Hostakova N, Novak P, Neumann P, Macas J. Domain based annotation of transposable elements - DANTE. 2023.
31. Neumann P, Novák P, Hoštáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob DNA. 2019;10:1.
32. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLOS Comput Biol. 2017;13:e1005595.
33. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31:3350–2.
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. Thomas CM. The nucleotide sequence and transcription of minicircular mitochondrial DNA's associated with male-fertile and cytoplasmic male-sterile lines of sugarbeet. Nucleic Acids Res. 1986;14:9353–70.
36. Henriksen RA, Jenjaroenpun P, Sjøstrøm IB, Jensen KR, Prada-Luengo I, Wongsurawat T, et al. Circular DNA in the human germline and its association with recombination. Mol Cell. 2022;82:209-217e7.
37. Zhang P, Mbodj A, Soundiramourtty A, Llauro C, Ghesquière A, Ingouff M, et al. Extrachromosomal circular DNA and structural variants highlight genome instability in Arabidopsis epigenetic mutants. Nat Commun. 2023;14:5236.
38. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. 2001;11:1095–9.
39. Nelson JR. Random-Primed, Phi29 DNA polymerase-based whole genome amplification. Curr Protoc Mol Biol. 2014;105:15.13.1-15.13.16.
40. Lasken RS, Stockwell TB. Mechanism of chimera formation during the multiple displacement amplification reaction. BMC Biotechnol. 2007;7:19.
41. Pires JC, Lim KY, Kovarík A, Matyásek R, Boyd A, Leitch AR, et al. Molecular cytogenetic analysis of recently evolved *Tragopogon* (Asteraceae) allopolyploids reveal a karyotype that is additive of the diploid progenitors. Am J Bot. 2004;91:1022–35.
42. Garcia S, Pascual-Díaz JP, Krumpolcová A, Kovarík A. Analysis of 5S rDNA genomic organization through the RepeatExplorer2 Pipeline: a simplified protocol. Methods Mol Biol Clifton NJ. 2023;2672:501–12.
43. Maiwald S, Mann L, Garcia S, Heitkam T. Evolving together: Cassandra retrotransposons gradually mirror promoter mutations of the 5S rRNA genes. bioRxiv. 2023;2023.07.14.548913.
44. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35 Web Server:W265-268.
45. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in *de novo* annotation approaches. PLoS ONE. 2011;6:e16526.
46. Smit A, Hubley R, Green P. RepeatMasker software program (computer program), ver. 3.1.8. Seattle: Institute for Systems Biology; 2007.
47. Sielemann K, Pucker B, Schmidt N, Viehöver P, Weisshaar B, Heitkam T, et al. Complete pan-plastome sequences enable high resolution phylogenetic classification of sugar beet and closely related crop wild relatives. BMC Genomics. 2022;23:113.
48. Sielemann K, Schmidt N, Guzik J, Kalina N, Pucker B, Viehöver P, et al. Pangenome of cultivated beet and crop wild relatives reveals parental relationships of a tetraploid wild beet. bioRxiv. 2023;2023.06.28.546919.
49. 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.

Publisher's Note

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