


SatXplor—a comprehensive pipeline for satellite DNA analyses in complex genome assemblies

Marin Volarić, Nevenka Meštrović, Evelin Despot-Slade *

Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia

*Corresponding author. Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia. E-mail: evelin.despot.slade@irb.hr

Abstract

Satellite DNAs (satDNAs) are tandemly repeated sequences that make up a significant portion of almost all eukaryotic genomes. Although satDNAs have been shown to play an important role in genome organization and evolution, they are relatively poorly analyzed, even in model organisms. One of the main reasons for the current lack of in-depth studies on satDNAs is their underrepresentation in genome assemblies. Due to complexity, abundance, and highly repetitive nature of satDNAs, their analysis is challenging, requiring efficient tools that ensure accurate annotation and comprehensive genome-wide analysis. We present a novel pipeline, named satellite DNA Exploration (SatXplor), designed to robustly characterize satDNA elements and analyze their arrays and flanking regions. SatXplor is benchmarked against other tools and curated satDNA datasets from diverse species, including mice and humans, showcase its versatility across genomes with varying complexities and satDNA profiles. Component algorithms excel in the identification of tandemly repeated sequences and, for the first time, enable evaluation of satDNA variation and array annotation with the addition of information about surrounding genomic landscape. SatXplor is an innovative pipeline for satDNA analysis that can be paired with any tool used for satDNA detection, offering insights into the structural characteristics, array determination, and genomic context of satDNA elements. By integrating various computational techniques, from sequence analysis and homology investigation to advanced clustering and graph-based methods, it provides a versatile and comprehensive approach to explore the complexity of satDNA organization and understand the underlying mechanisms and evolutionary aspects. It is open-source and freely accessible at <https://github.com/mvolar/SatXplor>.

Keywords: SatXplor; pipeline; satellite DNA; genome

Introduction

Repetitive DNA sequences comprise a substantial portion of the eukaryotic genomes and play pivotal roles in genome stability, evolution, and functional diversification [1]. These sequences are broadly classified into tandemly repeated satellite DNAs (satDNAs) and dispersed repetitive DNAs such as transposable elements (TEs) (reviewed in [2]). SatDNA is a unique class of repetitive elements with tandemly arranged monomers, typically 150–400 bp in length (reviewed in [3]), forming arrays that can extend up to 1 Mb. SatDNAs are an integral part of the (peri)centromeric region in eukaryotes but have also been found in euchromatin [4–6]. Among the fastest-evolving genome parts, satDNAs show species-specific profiles, with some species containing hundreds of distinct satDNA families [7]. Also, satDNA can form a higher-order repeat (HOR) organization, where different monomers create structured repeated patterns, further adding complexity to these regions [8]. Due to the variability of the monomers, the presence of long satDNA arrays, and the high genome abundance of diverse satDNA families, satDNAs pose challenges for genome assembly using Illumina sequencing, often resulting in insufficient representation. However, advancements in long-read sequencing by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) now enable the capture of extensive satDNA

regions [9], offering a powerful platform for detailed satDNA analysis across various species.

Today, many different tools have been developed specifically for detection and identification of repetitive DNA in sequenced genomic data. For example, the assembly-free algorithm TAREAN performs graph-based clustering for high throughput detection of satDNAs on short Illumina reads [10]. The data obtained from this analysis contains only basic information about satDNA, such as the variability and abundance of monomers but without any information on genome organization. The other frequently used program, Tandem Repeat Finder (TRF), offers detection of tandem repeats on assembled sequences by recognizing periodicity and providing monomers and their consensus [11]. Similarly, ULTRA enables tandem repeat annotation with improved sensitivity and identification of longer and degenerated repeats [12]. mreps utilizes the Hamming distance approach for satDNA detection on whole genomes [13], but it has limitations when investigating complex genomic regions. There is also a RepBase database [14] of several classes of repetitive elements, including satDNA with curated annotations and classifications. It is used by RepeatMasker [15] to accurately identify previously described repeats relying on sequence similarity and evolutionary conservation. Recently several tools such as TRASH [16] and NanoTRF [17] have

Received: August 12, 2024. Revised: October 31, 2024. Accepted: December 4, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

been developed to provide improved detection and annotation of satDNA in long-read genome assemblies or primarily for centromere visualization [18]. However, what all these programs for satDNA detection lack is the capability of finding a complete set of satDNA monomer variants on the genome scale, accurate detection of arrays, and surrounding regions. Therefore, proper genome-wide annotation of satDNA monomers and arrays is a prerequisite for disclosing the organization, mechanisms of propagation, and evolutionary processes for sequences in question.

Going past the detection step, any comprehensive analysis requires an extensive understanding of both bioinformatics and repetitive DNA behavior. This is especially true for satDNAs, which are often investigated only at the monomer level with little insight into their array organization and genomic environment. The absence of a unified methodology for satDNA analyses on the genome scale leads to diverse approaches across literature, often focusing only on specific organisms, satDNA families, or even one satDNA sequence [19–21].

The flour beetle *Tribolium castaneum* is an ideal model for satDNA analyses due to its well-defined satDNA families in euchromatin [5] within a highly repetitive genome, including large centromeric satDNA [22] and a well-described satellitome [23]. Our recent study disclosing mechanisms of euchromatic satDNA spread in *T. castaneum* is based on an improved genome assembly using nanopore sequencing and its enrichment in the repetitive part of satDNA in particular [24]. Our intention to perform genome-wide analyses of subsets of satDNA has revealed a significant lack of available tools for satDNA analysis, in particular the lack of a standardized pipeline for downstream analysis of satDNA. Therefore, there is an urgent need for dedicated resources to enable comprehensive analyses of satDNA dynamics.

In this work, we developed SatXplor, an integrated pipeline that greatly improves and automates the approach initially used in *T. castaneum* analyses. It encompasses multiple algorithms for thorough satellite DNA Exploration (SatXplor) and analyses validated on different genomes utilizing multiple publicly available long read-based assemblies with well-described satDNAs. To validate and further develop the SatXplor pipeline, we selected moderately repetitive and smaller genomes (150–350 Mb) of *Meloidogyne nematodes* (*Megalaima incognita*, *M. arenaria*) and *Drosophila melanogaster*, highly repetitive and very large genome (6.3 Gb) of the locust *Locusta migratoria*, while the genome of *Arabidopsis thaliana* served as a plant model. We have also validated SatXplor on human (*Homo sapiens*) and mouse (*Mus musculus*) complete genome assemblies which also contain satDNA-rich (peri)centromeric regions. SatXplor represents a novel and versatile pipeline for the complete satDNA analysis in different genome assemblies, allowing detailed exploration of all satDNA regions, including the most abundant ones. This tool provides new insights into genome organization and the mechanisms governing the satDNAs across diverse species.

Materials and methods

Genome and satDNA data

SatXplor algorithms were developed and tested on genomes of *T. castaneum*, *D. melanogaster*, *L. migratoria*, *M. incognita*, *M. arenaria*, *A. thaliana*, *H. sapiens*, and *M. musculus* with accessions provided in [Supplementary Table 1](#). Information on analyzed satDNA sequences, as well as the sources are provided in [Supplementary Table 2](#).

SatXplor pipeline overview

Detection and extraction of monomers

The process of monomer detection based on query satDNAs is performed by NCBI BLAST+ [25], running from a Python subprocess command. The program also creates the subject database and deletes it since it is a fast process for most genomes. The main parameters for the BLAST search are:

```
– evaluate 10 – outfmt 6 – max_target_seqs 10000 – task blastn –
num_threads 2 – dust no – soft_masking false
```

Dusk and soft masking are turned off to ensure detection of low-complexity satDNA sequences.

Array creation

Arrays for a single satDNA family were created using a custom Python script. First, all BLAST hits from a single satDNA family were ordered by chromosome and their respective start position. Next, absolute distances were calculated between consecutive monomers. The resulting distances were then graphically visualized, and a histogram of the data was constructed. Since the histogram is a one-dimensional array, peaks were identified using `scipy.findpeaks()` function, keeping only peaks which amount to >5% of total monomer number in the genome (meaning that at least 5% of monomers are peak distance apart). Afterwards the max peak value for each family was used as an “extension factor” by which each monomer annotation end position was elongated and finally each overlapping groups of monomer annotations were found and annotated as arrays. Up to the development of this method, array creation often necessitated manual adjustment and often creation of arrays by hand.

Monomer density profiles

The 2D density approximations from BLAST output were created using a custom Python script. First, a 2D probability density Gaussian kernel was estimated for approximate query coverage and percentage identity values using `scipy.stats.gaussian_kde()`. For plotting, the density approximation was binned in `NKERNEL_BINS` on a 2D grid and plotted.

K-mer-based profiling of arrays

To find the exact edges of individual arrays, SatXplor employs the following algorithm for each satDNA under examination:

1. Extract all monomers from the genome to serve as a database of all possible k-mers for a satDNA family.
2. Create synthetic monomer-dimers. Dimers were created to enrich the database with all possible transitive k-mer states.
3. Extract extended arrays (arrays with +/- 500 bp flanks).
4. Create a hash table of all 32 k-mers from both the synthetic dimers and the extended arrays.
5. For each 32 k-mer in the extended array, find the closest 32 k-mer from the synthetic dimer table by Hamming distance.
6. Calculate a rolling sum score by averaging +/- 5 bp scores for each position.
7. Define new array edges as the positions where first/last k-mer in the window has a score lower/higher than five (meaning an 87% similarity to a hit in the database).

Alignment and distance calculations

The main algorithm used in steps of monomer, flank, and microhomology alignment is MAFFT [26], which is run through a Python

subprocess command with the main settings:

```
'mafft --adjustdirection-reorder --threads str(half_threads)'
```

Where by default *half_threads* are half of total CPU processors available. Subsequent genetic distances are calculated by *ape* package in R using the “F81” genetic distance model.

Dimensionality reduction

Since phylogenetic trees often proved lacking when analyzing a lot of highly similar sequences from the same genome, the PCA and UMAP algorithms and their implementation in R [27] were used. First step was loading the MAFFT alignments into R, after which the distance matrix of all sequences from the alignments was calculated by *ape* package in R using the “F81” genetic distance model since it presumes a non-equal rate of evolution across nucleotides. The distance matrix is then processed for missing values or outliers and internally min-max scaled. For the PCA implementation in R, the *PCA()* function from the *FactoMineR* [28] package was used. The resulting principal components are then visualized in reduced-dimensional space using the *ggplot2* [29] package, and the *Scree* plots for the first 10 principal components are also created. For the UMAP implementation in R, the *umap()* function from the *umap* package was used to find the UMAP embedding of the distance matrices with default parameters controlling aspects such as the number of dimensions in the reduced space and the number of neighbors to be considered for the local structure. The resulting UMAP embedding is also visualized directly with the *ggplot2* package.

Network and distance graphs

For generating heatmaps, the MAFFT alignment of extracted junction regions was loaded into R, the distance matrix was created again using the “F81” model, and the matrix was scaled to values from 0 to 1. Finally, the matrix was then processed using the ‘*heatmap*’ package in R to cluster and visualize the junction regions.

For creating the undirected graph networks, we used the following steps:

1. Uniquely label all monomers as well as the arrays they originate from and create the alignment using MAFFT.
2. Calculate the distance matrix from the alignment of all monomers in the genome using the “F81” genetic distance model.
3. From the distance matrix, find the closest N monomers not belonging to the same array.
4. Create a table of all possible array-array connections.
5. Remove redundant or duplicate connections.
6. Create a regular network using ‘*igraph*’.
7. Create HTML interactive networks using *network3D*.

Performance

The performance of SatXplor is mainly defined by two key factors: number of input satDNA sequences and their respective number of monomers, with the greatest emphasis being on the total number of monomers/arrays present in the genome. All the examples were run on an Intel(R) Core(TM) i9-9900 CPU with 128GB RAM, but all the programs are configured to run on any machine >8GB of RAM and a multicore processor; however, the run times may differ (Table 1).

In general, it is advisable not to use the most abundant satDNA sequences in large genomes with this pipeline, as they may contain hundreds of thousands of monomers in the assembly and are likely to take a long time to complete. It is important to note that most of the pipeline is multithreaded at points where the likelihood of an I/O bottleneck is low. However, when using a large machine (e.g. a cluster) with a relatively slow drive, it is best to limit the number of threads used.

Results

The SatXplor pipeline contains several algorithms for SatXplor and analysis (Fig. 1), requiring only FASTA sequences of potential satDNAs and a long read-based assembly. The first step is the accurate and complete detection of satDNA monomers in respective assemblies using a BLAST search followed by a homology-based inspection. The next algorithm recognizes satDNA organization followed by array definition and precise profiling to disclose potential patterns in arrays’ organization. Next, individual satDNA monomers, arrays, and their flanking regions are extracted and serve as input for subsequent analysis. In satDNA monomer analyses principal component analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) are used to cluster satDNA monomers to disclose their relationship. Finally, SatXplor uses the alignments to construct undirected graphs to comprehensively visualize the relationships between monomers in satDNA arrays. A k-mer-based analysis increases the precision in determining precise array boundaries, allowing extraction of flanking regions. Array boundaries and flanking regions serve as a basis for the detection of potential micro- (using *seqLogo*) and macrohomologies (using distance maps). Whole output is stored within the results folder that is presented in the structure depicted in Supplementary Fig. 1.

The robustness of SatXplor toolbox for analysis of satDNAs was tested in several species with genomes of various sizes and repetitiveness (Supplementary Table 1) and structurally different satDNAs (Supplementary Table 2). In *D. melanogaster* well-described 1.688 satDNA was investigated [30]; in *Meloidogyne* we used the database of tandem repeats [31, 32]; in *L. migratoria* we utilized well-described satellitome [19] and explored satDNAs of *A. thaliana* described in [33]. For validation of SatXplor on mammalian genomes we used consensus of the human (*H. sapiens*) major alpha satellite together with other described pericentromeric satellites (HSAT3, HSAT4, HSAT5). For the mouse genome (*M. musculus*), we used the major pericentromeric satDNA sequence (MaSat). The detection of satDNA consensus monomer sequence for each species was previously performed and experimentally validated using molecular methods and bioinformatics tools (TRF, TAREAN) (Supplementary Table 2).

SatDNA exploration

SatDNA detection

First step in SatXplor pipeline is the detection of all potential monomers in the genome, which is performed using NCBI BLAST+ executable and satDNA consensus sequences as a query. A python-based BLAST wrapper with parameters best suited for satDNA detection (see Materials and methods section) controls the process and generates a result table with all potential hits in the assembly (Supplementary Table 3). This table serves as the basis for subsequent analyses within the satDNA toolbox, enabling the extraction of essential satDNA information. This includes the localization of each satDNA monomer on the target sequence, as well as rough estimates of percentage identity and

Table 1. Examples of run times, CPU, and memory usage for the five organisms we used in testing the pipeline

Genome (size)	Total monomers	Peak memory (GB)	Wall time (seconds)
<i>Arabidopsis thaliana</i> (119 Mb)	12 107	0.6	230
<i>Drosophila melanogaster</i> (143 Mb)	426	0.5	120
<i>Meloidogyne incognita</i> (199 Mb)	10 400	0.8	344
<i>Meloidogyne arenaria</i> (297 Mb)	83 460	8	626
<i>Tribolium castaneum</i> (225 Mb)	11 712	1	362
<i>Locusta migratoria</i> (6.2 Gb)	148 132	25	5238

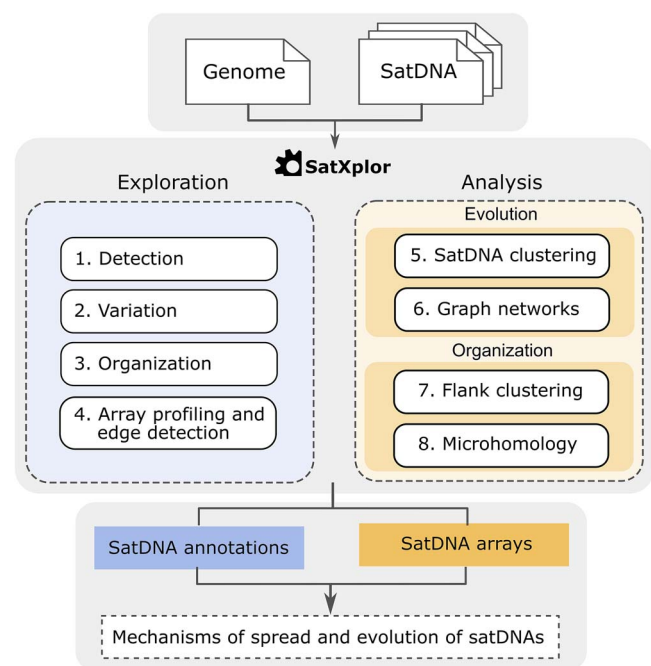


Figure 1. Workflow of SatXplor pipeline for comprehensive satDNA annotation and analysis in long read-based genome assemblies. SatXplor pipeline consists of four exploration algorithms for finding structural characteristics and variants of satDNA elements together with four analysis algorithms for evolutionary insight and organizational patterns dissecting further the satDNA genomic landscape. The output of SatXplor provides detailed annotations and visualizations, shedding light on the mechanisms underlying their spread and evolution within the genome.

query coverage. It is noteworthy that, at this stage, no filtering step is applied, allowing for a comprehensive downstream exploration of satDNAs.

Finding homology between annotated satDNA

The first output of SatXplor is to visualize the variability of satDNAs. For this purpose, we have developed a visualization technique that quickly and efficiently analyses all detected monomers in the genome. Using 2D relative density approximation on data from the BLAST output table (Supplementary Table 3), we visualize potential groupings of satDNA monomers, enabling a clear assessment of intragenomic variability of a satDNA family. This analysis generated distinct scenarios of monomer conservation of particular satDNA within the genome under investigation (Fig. 2A). For instance, satDNA that is characterized by low monomer variability is visible as a single spot in the graph (Fig. 2A, Example 1). Slight variation is visible as light areas on the graph (Fig. 2A, Example 2), whereas higher variation is depicted as several spots of similar intensity (Fig. 2A,

Example 3). Finally, some satDNAs may exhibit high variability and degeneracy in the monomer sequence, which becomes visible as a large region of intensive signal that extends toward lower percentages (Fig. 2A, Example 4). The information gained from this graph is used to define specific sequence identity and reference sequence coverage parameters to annotate and extract all monomers for particular satDNA from the genome. This approach creates a fundamental basis for further satDNA assessments and facilitates novel whole genome analyses in subsequent pipeline algorithms.

Detection of complex satDNA organization

We have successfully identified parameters for satDNA detection using the approach described above (Fig. 2A); however, many satDNAs do not form pure tandem arrays and are often interspersed with other sequences. Therefore, the crucial step before we start the subsequent determination and analysis of satDNA arrays is the detection of possible interspersed sequences between satDNA monomers. For each satDNA family, SatXplor outputs a plot with the approximate distance to the nearest monomer in a strand-independent manner, finds the peaks, and then elongates both the satDNAs and the newly found interspersed sequences into larger arrays, outputting both the annotation and general statistics of the arrays found. For example, in satDNA, which occurs in homogeneous tandem repeat arrays, the graph shows a dominant single peak at distance values ~ 0 (Fig. 2B, Examples 1 and 2). However, when a sequence is inserted between satDNA monomers, it is represented by the presence of one or more peaks at distances greater than monomer length (Fig. 2B, Example 3 and 4). An additional feature of this algorithm is the capability to detect and extract all sequences that come in between satDNA monomers of interest, even if they are not tandemly repeated or if there are multiple types of complex organization for a single satDNA family.

K-mer-based profiling and edge detection

Defining the ends of satDNA arrays is crucial for downstream analyses of propagation mechanisms and evolutionary dynamics within the genome. The main problem for proper arrays' edge definition is the tendency of arrays to degenerate in sequence at their edges. For example, first or last monomers of an array are often truncated [21] and subject to higher sequence variability. This limits the finding of potential micro- and macrohomologies in flanking regions (Supplementary Fig. 2). To solve this problem, we developed a novel method for precise edge detection employing a per-array k-mer-distance calculating method followed by postprocessing. The result of this analysis is precisely defined array edges subsequently used in the analysis, as well as per-array profiles. Utilizing the k-mer-counting method, we generated monomer distance profiles for each detected array, offering visual representations of intricate organization patterns (Fig. 2C). These

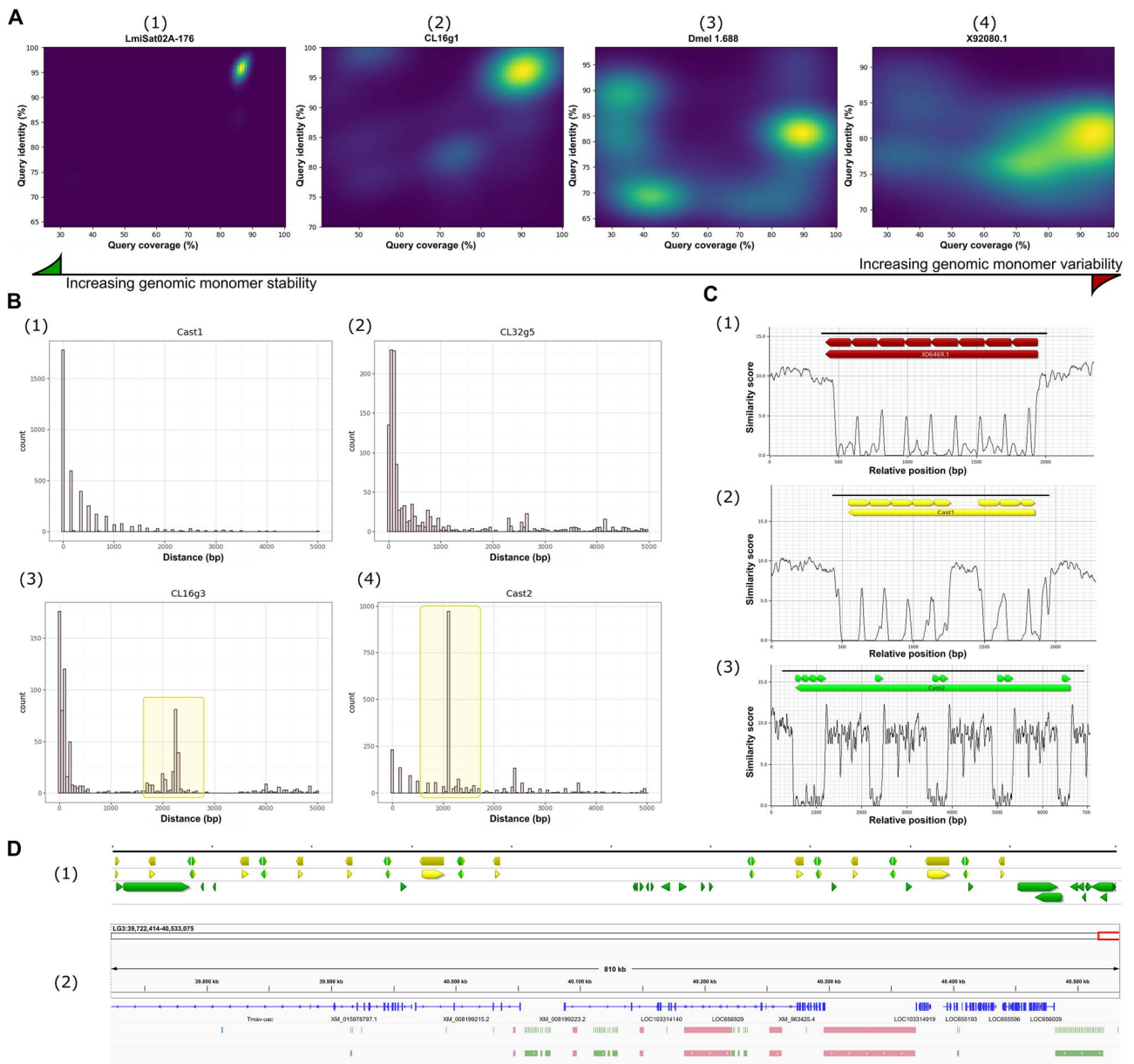


Figure 2. Preparation and annotation of satDNA data with SatXplor. (A) Representative scenarios of different results for BLAST based monomer detection in genomes of interest. Depending on the satDNA characteristics different levels of monomer variability can be observed from extremely low (1), through moderate (2), and highly variable (3), to extremely variable and fragmented monomers (4). (B) SatDNA organization. Depicted are certain scenarios of tandem organization for specific satDNA monomers. SatDNAs can have low distances between monomers (1,2) with values often being random, indicating homogenous array. This approach can also detect complex organization of satDNA monomers characterized by the interspersion with another specific sequence (highlighted spike in 3, 4). (C) Precise array profiling. Per-array monomer distance profiles generated by the k-mer-counting method give powerful visualizations to study patterns of specific arrays' organization. For example, a homogenous tandem array (1), a slightly discontinuous array (2), and a discontinuous array (3) can be observed. Colored arrows represent satDNA monomers and arrays corresponding to graph profiles beneath. (D) Annotation and visualization. Output of SatXplor can easily be used in different bioinformatics visualization tools such as Geneious Prime 2023.2.1 (<https://www.geneious.com>) (1) and IGV [34] (2). Here, each satDNA monomer and accompanying array are shown in separated tracks and can then later be separated between different satDNA families. All satDNA names and species used (A–D) are listed in [Supplementary Table 2](#).

profiles can reveal distinct features, such as uniform head-to-tail organization of monomers within an array (Example 1), to a slightly irregular (Example 2), and further to a highly discontinuous structure (Example 3). This approach elucidates the variability of array organization, and the annotations are compatible with any genome browser, such as Geneious and IGV (Fig. 2D), to help explain and understand the different patterns of satDNA organization.

SatDNA analysis

Clustering analysis for uncovering satDNA monomer relationships

In the identification of satDNA variants, the standard phylogeny tree approach encounters limitations, particularly in distinguishing very subtle differences among satDNA monomers together with slow processing speed in drawing and analyzing trees as well as often found low bootstrap values, long run times and

complex dendrogram organizations (Supplementary Fig. 3). To address these challenges, we employ a dimensionality reduction-based method of satDNA monomer analysis (Fig. 3A). This method facilitates a thorough exploration of satDNA variation by processing all monomers from the same genome simultaneously with huge efficiency. Two main algorithms used are PCA (Fig. 3A, upper panels) and UMAP (Fig. 3A, bottom panels), operating on the distance matrices generated from the alignments. The output is graphed with satDNA monomers represented with individual dots that can be colored based on different attributes, such as different arrays, chromosomes, or species. The information contained in clusters, or lack thereof, provides valuable insight into evolution and genome organization of specific satDNA families. These visualizations reveal specific satDNA characteristics based on the clustering patterns or their absence. Certain satDNA monomers may show distinct segregation (Fig. 3A, Example 1), indicating sequence variability and putative evolutionary events. Furthermore, it is possible to observe satDNA with a high degree of intrachromosomal similarity that nevertheless exhibits some degree of interchromosomal mixing (Fig. 3A, Example 2), indicating several distinct expansion events. There are also instances where satDNA has undergone pronounced intrachromosomal exchanges (Fig. 3A, Example 3), indicating active and dynamic genomic interactions within and between chromosomes. SatXplor provides an interface for running either statistical (PCA) or geometric approaches (UMAP) for dimensionality reduction, showing similar results; however, the results vary between satDNA families, thus it is best to run them both.

Unveiling evolutionary trends through satDNA network analysis

Finally, SatXplor uses undirected graph networks to explain satDNA-specific evolutionary history. Using this approach, it is possible to find shortest evolutionary paths of particular satDNA monomers in a family since each array gets linked to only its closest neighbors. Expanding this for all arrays in the genome, clear dispersion centers emerge, the arrays from which rapid expansion probably occurred. Thus, it is possible to construct several key patterns of satDNA evolution. (Fig. 3B). Arrays exhibiting a high degree of mixing are depicted by one cohesive cluster (Fig. 3B, Example 1). Conversely, there are clusters distinct from the central region (Fig. 3B, Example 2) or even separated (Fig. 3B, Example 3), indicating divergence of arrays over time with mixed scenarios characterized by numerous interconnected clusters (Fig. 3B, Example 4).

Distance mapping of surrounding regions

SatDNA may be associated with other repetitive elements or embedded in gene-rich regions. Therefore, it is essential to assess satDNA features and recognize potential genomic association with different repeats or specific sequences. After the precise determination of the array edges (see Results section 1.4), the contiguous regions of 500 bp up- and downstream surrounding each array are systematically extracted. The size of investigated regions can be varied during explorative phase and later adjusted depending on the attributes of a particular satDNA array. A detailed distance clustered heatmap is then generated based on the multiple alignments for each of these regions. Analyzing the distance maps can reveal various patterns in satDNA surrounding regions (Fig. 3C). One of them shows satDNAs with arrays that have almost no similarity between their neighboring regions (Fig. 3C, Example 1). The other pattern of satDNA arrays surrounding

regions shows high similarity regions at array edges representing occasional mixing with some other repeats (Fig. 3C, Example 2). Interestingly, some satDNA arrays exhibit more frequent mixing with other highly similar or repetitive regions (Fig. 3C, Example 3), indicating the presence of several distinct sequences in the vicinity of satDNA arrays. Finally, some arrays show large, conserved blocks (Fig. 3C, Example 4) that originate from highly conserved sequences in the vicinity. In some cases, this algorithm can also detect the presence of sequences with shared homology only on one side of the array, suggesting a possible association with other repeats and not a uniform embedding within repetitive regions.

Microhomology detection in flanking regions of satDNA arrays

Additionally, a precise edge detection algorithm allows SatXplor to focus on microhomologies near satDNA arrays. Microhomologies have been proposed to mediate both microhomology-induced break repair and eccDNA genome reintegration (reviewed in [35]) and, as such, represent a vital pathway in evolution of satDNA in the genome [21]. The sequence logo graphs can be used to spot potential conserved microhomology regions (Fig. 3D) as targets for investigation and manual curation or to reject certain mechanisms in organisms of interest. It allows finding specific satDNA characteristics, such as high GC content (Fig. 3D, Example 1) within repetitive environments those are typically AT-rich. Additionally, microhomology analysis can detect regions containing polyN (A, T, G, C) stretches (Fig. 3D, Example 2) indicative of potential regulatory role of sequence motifs.

Benchmarking

To compare the performance between SatXplor and the established algorithms in the field of satDNA research, we performed a benchmarking analysis against TRASH, TRF, ULTRA, and Stained-Glass. These tools were selected due to their comparable features with SatXplor, relevance, and capabilities of detecting tandem repeats in genomic sequence. Firstly, tools offering annotation were compared in the amount and coverage of detected satDNA on the same region (Fig. 4A), where SatXplor showed best result both in number and coverage, using a large satDNA database. Next, we visually inspected mapped monomers and array regions on two different satDNA organization types (Fig. 4B). While all tools performed similarly for a homogenous array, greatest difference was observed for satDNA that have a complex organization such as interspersed repeats in discontinuous arrays where only SatXplor managed to accurately span the whole repetitive region. Finally, general features of SatXplor and other algorithms essential for effective satDNA investigation were compared (Fig. 4C). It can be observed that only SatXplor offers a complete suite of analytical features that significantly extend those of its counterparts, surpassing all but basic satDNA detection of TRF in run-time. While tools like TRASH, TRF, and ULTRA provide capabilities for *de novo* and genomic detection, they primarily emphasize detection rather than in-depth analysis. StainedGlass, on the other hand, offers mostly visualization. In contrast, SatXplor is specifically designed for satDNA analyses, providing detailed monomer variability insight, array creation with precise definitions of array edges of different complexity, junction region exploration and advanced visualization options, all with genome-wide comparisons.

To validate the performance of SatXplor on complex telomere-to-telomere (T2T) assemblies of eukaryotic genomes containing complex (peri)centromeric regions, we processed the data from mouse and human assemblies. The analyses demonstrated its

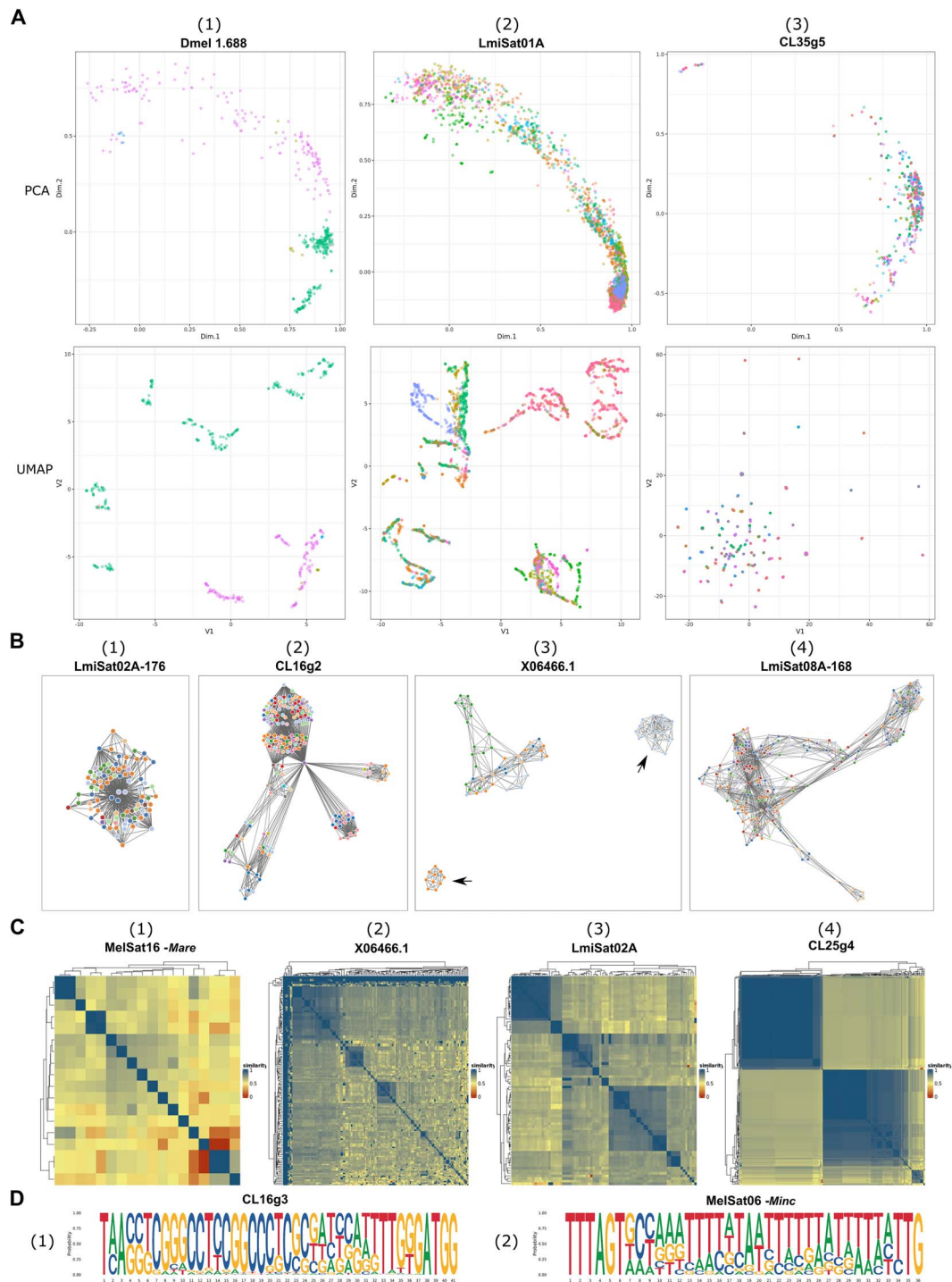


Figure 3. Evolution and organization analysis outputs of SatXplor. (A) SatDNA monomer clustering analysis using PCA and UMAP. The monomers are colored according to the origin of the chromosome. Specific satDNA monomers can show clear separation on both PCA (upper panel) and UMAP (bottom panel) (1); intrachromosomal similarity with a certain degree of mixing with other chromosomes can sometimes be visualized with only one of the algorithms, in this case, UMAP (2), and extremely high degree of intra- and intrachromosomal exchange that prevents efficient clustering by the algorithms (3). (B) Graph networks. Graph networks of arrays for four different satDNAs based on their sequence similarity relationship. Each dot on these graphs represents an array colored based on the chromosome they originate from. SatDNA arrays with a high degree of mixing represented by dense clusters of closely connected arrays (1), several arrays' clusters that are distant from the central region (2), separated arrays' clusters (marked with black arrows) (3), and example where there are many arrays' clusters almost all linked together (4). (C) SatDNA array surrounding region analysis. Pairwise sequence similarity matrices of 500 bp neighboring regions satDNA arrays visualized with heatmap. Neighboring regions of satDNA arrays without similarity (1) examples of several small blocks indicating limited correlation with several sequences (i.e. transposons, other satDNAs), (2) mixture of different conserved sequences in the neighboring regions represented by blocks of different sizes (3), and satDNA flanking regions that exhibit conserved large blocks indicating a conserved pattern of colocalization (4). (D) microhomology analysis. Analysis of flanking regions can reveal the presence of high GC content in a usually AT-high environment of satDNA (1) or regions that have poly(a/T) stretches. All satDNA names and species used (A–D) are listed in Supplementary Table 2.

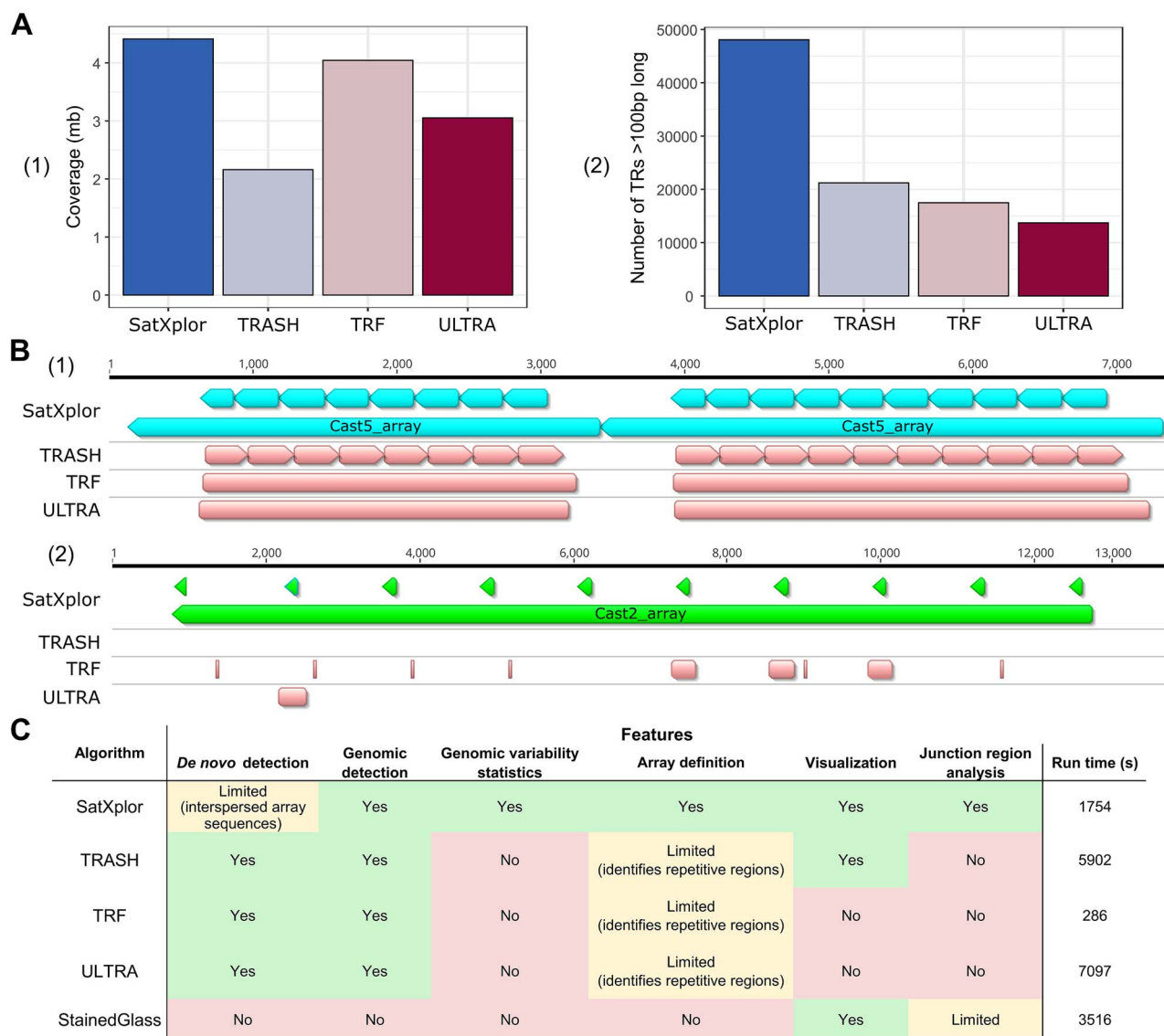


Figure 4. Benchmarking SatXplor against algorithms used in satDNA research that have comparable features. All comparisons were run on *T. castaneum* LG3 chromosome with the published satDNA database [5, 23]: (A) SatDNA annotation coverage (1) and the total number of annotations (2). (B) Annotation precision between different tools on different types of satDNA array organization: homogenous satDNA array (1) and complex discontinued array (2). (C) Qualitative comparison of SatXplor and other algorithms (TRASH [16], TRF [11], ULTRA [12] and StainedGlass [18]) in defining general features important for satDNA detection and analysis.

capability to effectively analyze satDNA within complex genomic contexts, revealing crucial patterns of variation and structural differences in highly repetitive regions of even the most complex genomes (Fig. 5). In the human genome, it effectively annotated centromeric (alpha satDNA) and pericentromeric satDNAs (HSAT4, HSAT5 and HSAT6), capturing both monomer variation and partial chromosome-specific clustering, while flanking region analysis revealed high sequence similarity surrounding the arrays (Fig. 5A). In the mouse genome, SatXplor uncovered key patterns of monomer variation of the pericentromeric major satDNA (MaSat), highlighting the presence of interspersed sequences, and revealed distinct structural differences between homogeneous and irregular satellite DNA arrays (Fig. 5B). This multifaceted approach positions SatXplor as a novel tool specifically designed for analyzing diverse satDNA families, providing deeper insights into the roles and dynamics of satDNAs in genomic architecture.

Discussion

The study of repetitive DNA poses a particular challenge, especially in the context of long stretches of satDNA, which have major structural and evolutionary implications [36]. There is a huge struggle to discern and accurately represent repetitive regions, leading to gaps and misrepresentations in genome assemblies and often requiring the implementation of multiple tailored approaches to obtain suitable platforms for their investigation [37]. While achieving contiguous assemblies represents a significant milestone in satDNA research, the formidable challenge lies in the development of algorithms capable of accurately describing and deeply analyzing these repetitive elements within the complexities of the genome.

SatXplor pipeline is a set of algorithms developed for precise definition and comprehensive analysis of satDNA in the assemblies of complex genomes. An assembly based on long reads and enriched in satDNAs is used as a platform in analyses

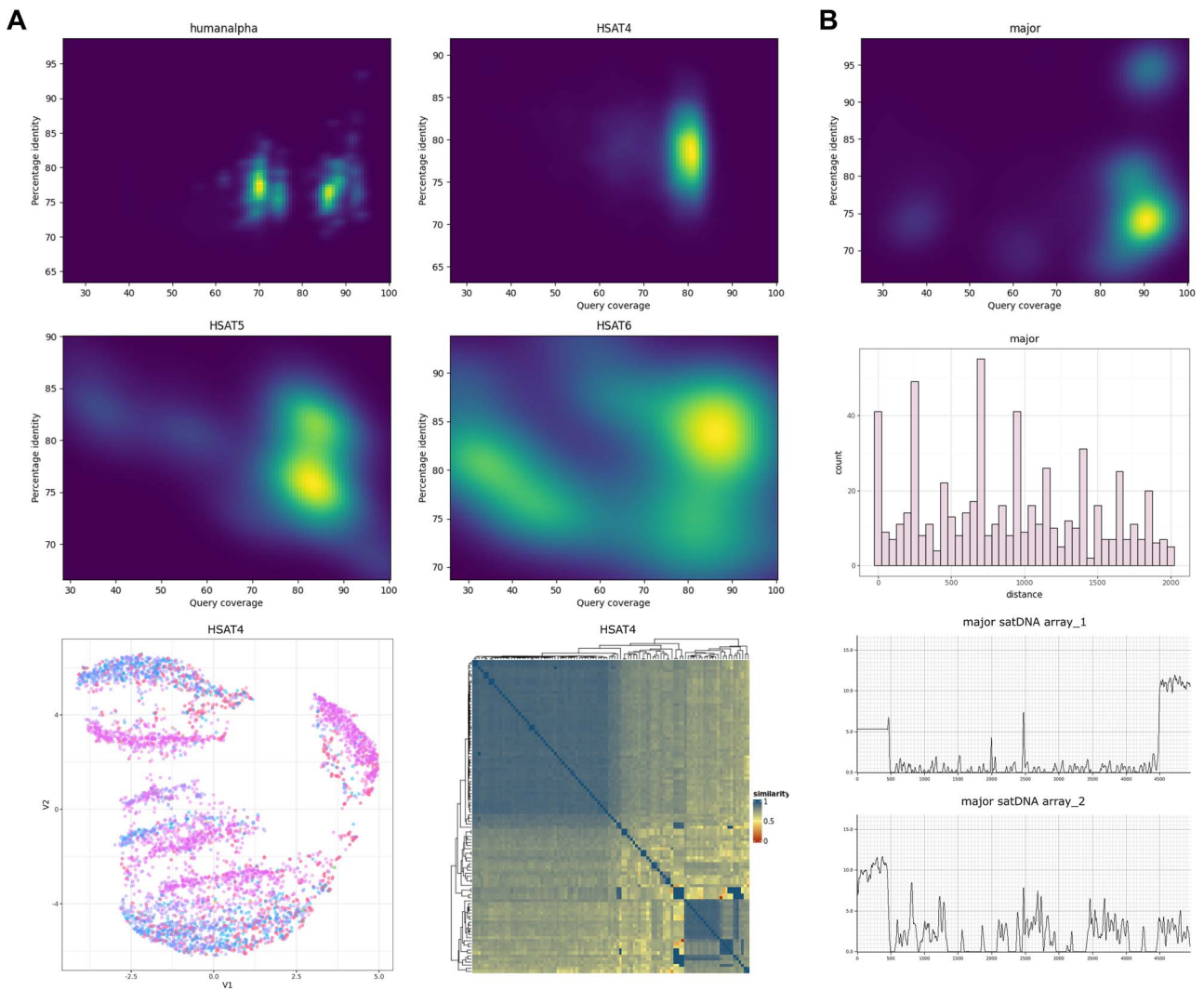


Figure 5. Validation of SatXplor on human and mouse genome assemblies with highly repetitive (peri)centromeric satDNAs. (A) A analysis of human T2T genome. Density graphs of monomer variation are shown (top) for human centromeric alpha satDNA and pericentromeric satDNAs (HSAT4–6). UMAP clustering of large number of HSAT4 monomers (bottom left) reveals partial chromosome specificity. Flanking region analysis of HSAT4 arrays (bottom right) demonstrates high similarity between surrounding sequences. (B) Analysis on mouse genome. For mouse major pericentromeric satDNA, monomer variation is shown (top). The presence of interspersed sequences between MaSat monomers is indicated by several taller bars at specific lengths (middle). An example of varying array organization (bottom) is shown for two arrays of the same satDNA, where upper k-mer count graph illustrates a homogenous array, while the lower one reveals a discontinuous, irregular array, suggesting major satDNA array divergence across the genome.

with the SatXplor pipeline. This approach enables the precise characterization of satDNA monomer variability, which can range from very homogeneous to highly variable, providing specific parameters for their annotation. It facilitates the identification of how the satDNA monomers are arranged within arrays, distinguishing between typical tandems, complex organization of satDNA arrays, and the presence of monomers in different orientations. In addition, SatXplor uses high-throughput analysis of all monomers of particular satDNAs with PCA and UMAP visualization to uncover underlying patterns in satDNA evolution. It is the first tool to define and precisely annotate both satDNA arrays and their edges, enabling further assessment of the genomic environment, presence of conserved sequences, and association with mobile elements or euchromatic regions. Moreover, SatXplor provides insights into the exact sites of satDNA insertion, leveraging microhomology evaluation to pinpoint insertion sites with precision. Graph networks show connections and distances among the arrays, offering valuable insights into evolutionary

dynamics of satDNAs on the genome scale. SatXplor pipeline can be used as a complementary pipeline following any algorithm utilized for the detection of repetitive elements.

SatXplor is optimized for satDNA with low to moderate genome abundancies (<5%) and monomer lengths 80–1000 bp, but is also capable of processing highly abundant (peri)centromeric satDNAs in human and mice (Fig. 5). Running the pipeline for micro- and mini-satDNA may require manual parameter adjustments and verification. This pipeline proved particularly efficient in analyses of euchromatic satDNA, a distinct class of tandem repeats mainly localized in gene-rich regions of the genome, which is characterized by moderate array lengths, variable copy numbers, and a propensity for interspersions with protein-coding genes [24]. SatXplor emerges as a versatile tool for satDNA analysis, featuring compatibility with various detection tools and requiring only satDNA monomer consensus sequence and genome assembly sequences as input. A notable strength of SatXplor is the efficient detection of satDNA variation optimized for fast processing

of all monomers within the genome. The pipeline streamlines the entire process by unifying the explorative and evolutionary analysis of satDNA in a single pipeline. It introduces a fast k-mer counting algorithm for array edge definition, which contributes to a vastly improved accuracy (max 5 bp, previously +/- monomer length) in satDNA array definition. Additionally, SatXplor allows exploration of neighboring array regions and reveals both micro- and macrohomologies together with interactive graph networks, providing insights into possible mechanisms of both intra- and interchromosomal spread and evolution.

Running SatXplor on different genome assemblies has provided valuable insights that can serve as a great starting point for in-depth analysis of satellitomes. It managed to successfully analyze even centromeric and highly abundant satDNA, as demonstrated by identifying chromosome specificity of satDNA in *D. melanogaster* and *L. migratoria*. It was able to detect conserved flanking regions around satDNA arrays in *M. incognita* and *A. thaliana*. In the human genome, SatXplor confirmed the large variation among alpha satDNA monomers, reflecting their tendency for a high degree of sequence divergence within centromeric regions [38, 39]. Pericentromeric HSAT4 showed substantial sequence conservation in its flanking regions, likely due to its mixing with the alpha satellite [40]. Conversely, the observed differences may be attributed to its proximity to the centromeric transition region [41]. Finally, SatXplor successfully identified both the pattern of interspersed sequences within the mouse MaSat arrays and the relatively high level of sequence variation, which have only recently been described [42].

Although there are numerous tools for the detection of satDNA sequences, only a handful of algorithms are specifically designed for intragenomic satDNA analysis. However, most of these programs lack the versatility offered by SatXplor. For example, TRAP is an algorithm with both detection and analysis capabilities but limited to only parsing TRF outputs [43]. SatDNA Analyzer recognizes intraspecies variation, detects polymorphisms, and generates consensus sequences that provide a valuable starting point for phylogenetic studies [44], but requires multispecies input to perform analysis. There are also programs like RepeatAnalyzer, focused only on intergenic microsatellite tracking, management, analysis, and cataloging [45]. However, it is developed exclusively on bacterial data. There are also some programs with specific functions, such as TRTTools with a focus on TR genotyping [46] or RepeatOBserver [47] and StainedGlass [18] with focus on visualization of centromeric repeats. As shown, customized programs are usually created only for specific targets, underscoring the need for a unified, versatile tool that encompasses both detection and downstream analysis. Our benchmarking and comparison with similar tools show that only SatXplor offers an integrated solution for SatXplor and analyses, introducing a new approach in the field of repetitive DNA and setting a standard for comprehensive satDNA studies. Importantly, its utility extends to functional satDNA analysis, enabling the identification of potential mechanisms and evolutionary trends, thereby advancing our understanding of genome dynamics.

Key Points

- We have developed SatXplor, the first pipeline of its kind for comprehensive annotation and analysis of satellite DNA, introducing novel approaches for studying these complex and unexplored parts of the genome.

- SatXplor requires only the assembly of interest and a collection of previously detected satellite DNA repeats to perform the entire analysis.
- SatXplor is efficient and fast, easily processing complex entire eukaryote genomes and satDNA libraries, resulting in publication-ready visualizations and tables.

Funding

This work has been fully supported by Croatian Science Foundation under the project IP-2019-04-4910.

Conflict of interest: None declared.

Author contributions

Conceptualization and Validation: M.V., N.M. and E.D.-S.; Formal Analysis, Investigation, Methodology, Visualization: M.V. and E.D.-S.; Resources: N.M.; Software: M.V.; Supervision: E.D.-S. and N.M.; Writing—Original Draft Preparation: E.D.-S.; Writing—Review and Editing, M.V., N.M. and E.D.-S.; Funding Acquisition: N.M. All authors have read and agreed to the published version of the manuscript.

Data availability statement

The SatXplor source code, documentation and example test data is openly available at <https://github.com/mvolar/SatXplor>. All other information on the datasets used can be found in the Supplementary Data.

Biographical note

Authors are from the laboratory of non-coding DNA at the Ruđer Bošković Institute in Zagreb. Our focus is on the study of satellite DNA and long-range genome organization of repeats.

References

1. Biscotti MA, Olmo E, Pat Heslop-Harrison JS. Repetitive DNA in eukaryotic genomes. *Chromosome Res* 2015;**23**:415–20. <https://doi.org/10.1007/s10577-015-9499-z>.
2. Liao X, Zhu W, Zhou J. et al. Repetitive DNA sequence detection and its role in the human genome. *Commun Biol* 2023;**6**:954.
3. Garrido-Ramos MA. Satellite DNA: an evolving topic. *Genes (Basel)* 2017;**8**:230.
4. Cabral-de-Mello DC, Mora P, Rico-Porras JM. et al. The spread of satellite DNAs in euchromatin and insights into the multiple sex chromosome evolution in Hemiptera revealed by repeatome analysis of the bug *Oxycarenus hyalinipennis*. *Insect Mol Biol* 2023;**32**:725–37. <https://doi.org/10.1111/imb.12868>.
5. Pavlek M, Gelfand Y, Pohl M. et al. Genome-wide analysis of tandem repeats in *Tribolium castaneum* genome reveals abundant and highly dynamic tandem repeat families with satellite DNA features in euchromatic chromosomal arms. *DNA Res* 2015;**22**:387–401. <https://doi.org/10.1093/dnares/dsv021>.
6. Rico-Porras JM, Mora P, Palomeque T. et al. Heterochromatin is not the only place for satDNAs: the high diversity of satDNAs in the euchromatin of the beetle *Chrysolina americana* (Coleoptera, Chrysomelidae). *Genes (Basel)* 2024;**15**:395.
7. Utsunomia R, de Andrade Silva DMZ, Ruiz-Ruano FJ. et al. Satellitome landscape analysis of *Megaloporus macrocephalus*

- (Teleostei, Anostomidae) reveals intense accumulation of satellite sequences on the heteromorphic sex chromosome. *Sci Rep* 2019;**9**:1–10.
8. Sujiwattanarat P, Thapana W, Srikulnath K. *et al.* Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. *Sci Rep* 2015;**5**:10315.
 9. Nurk S, Koren S, Rhie A. *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53.
 10. Novák P, Robledillo LÁ, Kobližková A. *et al.* TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res* 2017;**45**:e111.
 11. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;**27**:573–80.
 12. Olson D, Wheeler T. ULTRA: A Model Based Tool to Detect Tandem Repeats. In: *Proc. 2018 ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics*, Association for Computing Machinery, New York, US, pp. 37–46, 2018. <https://doi.org/10.1145/3233547.3233604>.
 13. Kolpakov R, Bana G, Kucherov G. Mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* 2003;**31**:3672–8.
 14. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 2015;**6**:4–9.
 15. Smit AFA, Hubley R, Green P. *RepeatMasker Open-4.0*. 2013–2015. <http://www.repeatmasker.org>.
 16. Wlodzimierz P, Hong M, Henderson IR. TRASH: tandem repeat annotation and structural hierarchy. *Bioinformatics* 2023;**39**:btad308.
 17. Kirov I, Kolganova E, Dudnikov M. *et al.* A pipeline NanoTRF as a new tool for *de novo* satellite DNA identification in the raw nanopore sequencing reads of plant genomes. *Plan Theory* 2022;**11**:2103.
 18. Vollger MR, Kerpedjiev P, Phillippy AM. *et al.* StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* 2022;**38**:2049–51. <https://doi.org/10.1093/bioinformatics/btac018>.
 19. Ruiz-Ruano FJ, López-León MD, Cabrero J. *et al.* High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Sci Rep* 2016;**6**:28333.
 20. Vondrak T, Ávila Robledillo L, Novák P. *et al.* Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. *Plant J* 2020;**101**:484–500. <https://doi.org/10.1111/tbj.14546>.
 21. Sproul JS, Khost DE, Eickbush DG. *et al.* Dynamic evolution of euchromatic satellites on the x chromosome in drosophila melanogaster and the simulans clade. *Mol Biol Evol* 2020;**37**:2241–56. <https://doi.org/10.1093/molbev/msaa078>.
 22. Gržan T, Despot-Slade E, Meštrović N. *et al.* CenH3 distribution reveals extended centromeres in the model beetle *Tribolium castaneum*. *PLoS Genet* 2020;**16**:e1009115.
 23. Gržan T, Dombi M, Despot-Slade E. *et al.* The low-copy-number satellite DNAs of the model beetle *Tribolium castaneum*. *Genes (Basel)* 2023;**14**:999.
 24. Volarić M, Despot-Slade E, Veseljak D. *et al.* Long-read genome assembly of the insect model organism *Tribolium castaneum* reveals spread of satellite DNA in gene-rich regions by recurrent burst events. *Genome Res* 2024;**34**:1878–94. <https://doi.org/10.1101/gr.279225.124>.
 25. Camacho C, Coulouris G, Avagyan V. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.
 26. Katoh K, Misawa K, Kuma KI. *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66.
 27. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. <https://www.R-project.org/>.
 28. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *J Stat Softw* 2008;**25**:1–18.
 29. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. 2016. ISBN 978-3-319-24277-4. <https://ggplot2.tidyverse.org>.
 30. de Lima LG, Hanlon SL, Gerton JL. Origins and evolutionary patterns of the 1.688 satellite DNA family in drosophila phylogeny. *G3 Genes Genomes Genet* 2020;**10**:4129–46.
 31. Despot-Slade E, Mravinac B, Širca S. *et al.* The centromere histone is conserved and associated with tandem repeats sharing a conserved 19-bp box in the holocentromere of *Meloidogyne* Nematodes. *Mol Biol Evol* 2021;**38**:1943–65. <https://doi.org/10.1093/molbev/msaa336>.
 32. Despot-Slade E, Širca S, Mravinac B. *et al.* Satellitome analyses in nematodes illuminate complex species history and show conserved features in satellite DNAs. *BMC Biol* 2022;**20**:1–19.
 33. Simoens CR, Gielen J, van Montagu M. *et al.* Characterization of highly repetitive sequences of *Arabidopsis thaliana*. *Nucleic Acids Res* 1988;**16**:6753–66.
 34. Robinson JT, Thorvaldsdóttir H, Winckler W. *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011;**29**:24–6. <https://doi.org/10.1038/nbt.1754>.
 35. Yang L, Jia R, Ge T. *et al.* Extrachromosomal circular DNA: biogenesis, structure, functions and diseases. *Signal Transduct Target Ther* 2022;**7**:342.
 36. Louzada S, Lopes M, Ferreira D. *et al.* Decoding the role of satellite DNA in genome architecture and plasticity—an evolutionary and clinical affair. *Genes (Basel)* 2020;**11**:72. <https://doi.org/10.3390/genes11010072>.
 37. Peona V, Blom MPK, Xu L. *et al.* Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour* 2021;**21**:263–86. <https://doi.org/10.1111/1755-0998.13252>.
 38. Altemose N, Logsdon GA, Bzikadze AV. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* 2022;**376**:eabl4178.
 39. McNulty SM, Sullivan BA. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res* 2018;**26**:115–38. <https://doi.org/10.1007/s10577-018-9582-3>.
 40. Warburton PE, Hasson D, Guillem F. *et al.* Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* 2008;**9**:533.
 41. Gershman A, Sauria MEG, Guitart X. *et al.* Epigenetic patterns in a complete human genome. *Science* 2022;**376**:eabj5089.
 42. Packiaraj J, Thakur J. DNA satellite and chromatin organization at mouse centromeres and pericentromeres. *Genome Biol* 2024;**25**:1–24.
 43. Sobreira TJP, Durham AM, Gruber A. TRAP: automated classification, quantification and annotation of tandemly repeated sequences. *Bioinformatics* 2006;**22**:361–2.
 44. Navajas-Pérez R, Rubio-Escudero C, Aznarte JL. *et al.* SatDNA analyzer: a computing tool for satellite-DNA evolutionary analysis. *Bioinformatics* 2007;**23**:767–8.
 45. Catanese HN, Brayton KA, Gebremedhin AH. RepeatAnalyzer: a tool for analysing and managing short-sequence repeat data. *BMC Genomics* 2016;**17**:422.

46. Mousavi N, Margoliash J, Pusarla N. et al. TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics* 2021;**37**: 731–3. <https://doi.org/10.1093/bioinformatics/btaa736>.
47. Elphinstone C, Elphinstone R, Todesco M. et al. *RepeatObserver: tandem repeat visualization and centromere detection*. bioRxiv 2023.12.30.573697. <https://doi.org/10.1101/2023.12.30.573697>.