PERIODICUM BIOLOGORUM
VOL. 121–122, No 1–2, 7–14, 2020
DOI: 10.18054/pb.v121-122i1-2.10571

UDC 57:61
CODEN PDBIAD
ISSN 0031-5362

Review Article

Tools and databases for solving problems in detection and identification of repetitive DNA sequences

EVA ŠATOVIĆ*
MONIKA TUNJIĆ CVITANIĆ
MIROSLAV PLOHL

Division of Molecular Biology, Ruđer Bošković Institute, Bijenička 54, 10 000 Zagreb, Croatia

***Correspondence:**
Eva Šatović
esatovic@irb.hr

Keywords: repetitive DNA; transposable elements; satellite DNA; bioinformatics tools; software; repeatome

Abbreviations:

LINE	– Long Interspersed Nuclear Element
LTR	– Long Terminal Repeat
MITE	– Miniature Inverted-repeat Transposable Element
ORF	– Open Reading Frame
satDNA	– Satellite DNA
SINE	– Short Interspersed Nuclear Element
TDR	– Terminal Direct Repeat
TE	– Transposable Element
TIR	– Terminal Inverted Repeat

Received March 23, 2020
Revised May 22, 2020
Accepted May 22, 2020

Abstract

Genome compartments known to carry out very important biological functions (e.g. centromeres and telomeres) are mostly constituted of repetitive sequences. At the same time, regions of the genomes enriched in repetitive sequences have always presented great technical challenges for sequence alignments and genome assemblies. Fast evolving sequencing technologies and the increasing accessibility of genomic datasets have opened the opportunity to gain new insights into poorly explored genome fractions, built of repetitive DNA. Comprehensive and accurate annotation and characterization of these sequences is therefore an important contribution to the understanding of genomic architecture and function as a whole. In order to attend the emerging needs in repeat analysis and characterization, many bioinformatics tools, databases and pipelines have been generated. This review is intended to draw attention to the problems encountered in the genomic studies of repetitive sequences and to provide an overview of a spectrum of most prominent bioinformatics tools used for gaining better insight into these important genomic components. Some of the described assets are focused on detection of a wide range of repeats while the others are focused on a specific type of repetitive DNA sequences, generated as an answer to specific research interests and needs of the scientific community.

REPETITIVE SEQUENCES IN EUKARYOTIC GENOMES

Two classes of highly abundant repeats present in eukaryotic genomes are sequences repeated in tandem and interspersed sequences (Figure 1). Tandem repeats can be divided into satellite, minisatellite, microsatellite and telomeric DNA sequences, differing in repeat unit length, the mechanisms of their origin, and the length of the arrays they build. The most prominent among them, satellite DNAs (satDNAs) are abundant genomic sequences commonly localized in heterochromatic genome compartments near centromeres and telomeres, as well as at interstitial chromosomal positions, reviewed in *(1–4)*. SatDNA repeats typically form long arrays, although short ones or individual monomers can also be found dispersed in euchromatic genome compartments. Many different satDNAs usually interlace in the genome, distinct in sequence and length of their monomers, abundance, and chromosomal distribution. Because of the random non-reciprocal exchanges between sequences in arrays, satDNA monomers evolve in concert, maintaining low sequence variability of satDNA within the genome (usually 2–3%), and promoting rapid alterations in the copy number of satDNA monomers *(3, 5, 6)*. Concerning structural and/or functional roles, satDNAs are, for example, considered to be important in centromeres *(7, 8)*, and in raising reproductive barriers between species *(9)*, while their transcripts trigger heterochromatin formation or can be involved in processes leading to tumor transformation *(4, 10–12)*.

Transposable elements (TEs) are sequences capable of moving to the new genomic locations and forming interspersed repeats. They are grouped into two main classes, based on mechanisms of transposition. Class I elements transpose by RNA-mediated mechanisms, while class II elements propagate through DNA-mediated processes *(13–15)*. In each class, there are autonomous copies, coding for all the products needed for their own transposition, and non-autonomous ones, which depend on the enzymes produced by the autonomous counterparts *(16)*. Further subdivision is based on structural features of TEs. Class I elements with coding capacity and long terminal repeats at their ends are called LTR retrotransposons. Their central part codes for structural and enzymatic components required for retrotransposition via *gag* and *pol* open reading frames (ORFs). The *pol* gene is composed of several domains, PR-RT-RH-IN, coding for: protease, reverse transcriptase, RNAse H and integrase. The order of these domains in the *pol* gene is used to define superfamilies within this class of elements *(17)*. Retroviruses structurally resemble LTR retrotransposons, with the main difference in the presence of an active envelope (*env*) gene in retroviruses *(17)*. Small, non-autonomous LTR retrotransposons called Terminal Repeat Retrotransposons in Miniature (TRIMs) also belong to class I. They contain terminal direct repeats (TDRs) flanking an internal domain which starts with a primer binding site, complementary to a tRNA, and ends with a polypurine tract *(18)*. Class I also contains non-LTR retrotransposons, further divided into Long Interspersed Nuclear Elements (LINE) and Short Interspersed Nuclear Elements (SINE) (Figure 1). LINE harbor an internal polymerase II promoter and encode two ORFs, one with RNA-binding capability and the other for endonuclease and reverse transcriptase. SINEs contain an internal polymerase III promoter boxes A and B but their mobility is dependent on products of LINEs *(19)*. LINE retrotransposition can also produce new chimeric retrogenes and retropseudogenes through reverse transcriptase template switching from LINE RNA to other nuclear RNAs *(20)*.

Class II DNA transposons have terminal inverted repeats (TIRs) at their ends, and encode for transposase that binds to sequence segments residing in the terminal regions of autonomous and non-autonomous elements during transposition process *(21)*. Non-autonomous elements, called Miniature Inverted-repeat Transposable Elements (MITEs), usually arise from autonomous elements by internal deletions, preserving similarities in TIR sequences *(22)*. One type of DNA transposons, Helitrons, use rolling-circle replication in their spread. These elements contain two modules which can include subterminal inverted repeats. In addition, left module (at 5' element side) holds a microsatellite sequence, while the right module contains a short palindromic sequence at its 3' end. An array of tandem repeats is frequently found between the two modules *(23)*.

It has been observed that satDNAs and TEs are connected in many different ways, reviewed in *(24)*. For example, satDNA can be formed by tandemization of a

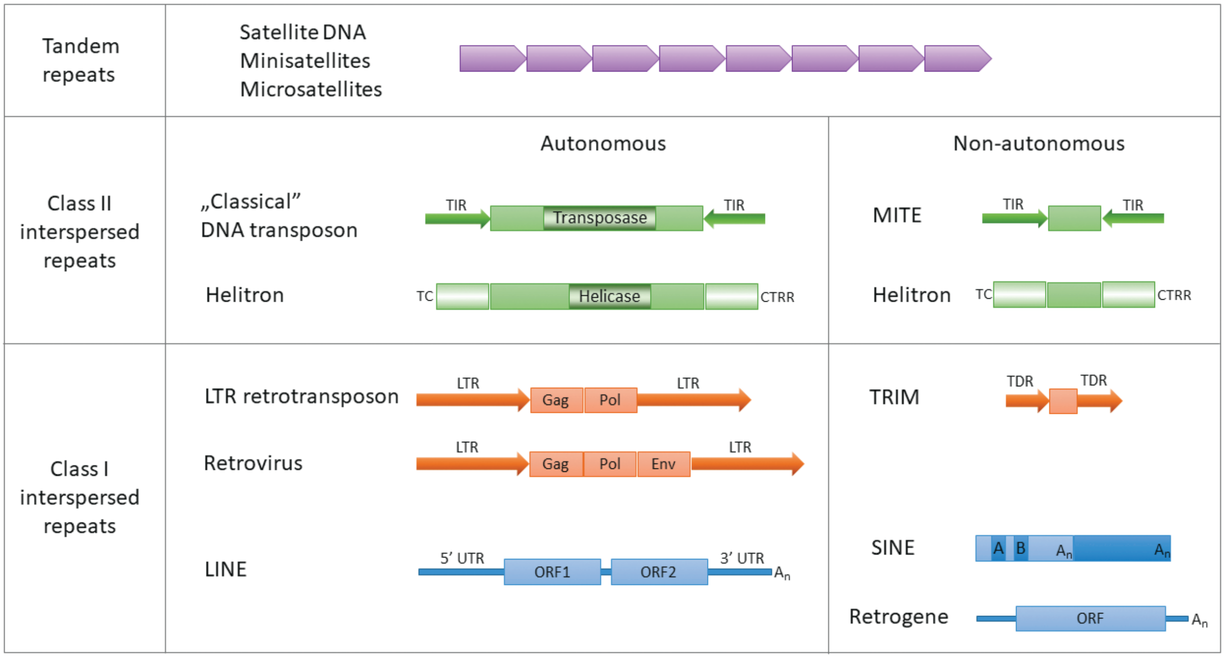

Figure 1. *Main types of tandem and interspersed repetitive sequences found in eukaryotic genomes.*

complete TE or its segments *(25–27)*. As already mentioned for Helitrons, some TEs have an internal region composed of sequences repeated in tandem. In some cases, TE-incorporated tandem repeats can be a source of monomers used in the formation of classical satDNA arrays *(28)*. Altogether, the fact that genome evolution is predominantly a consequence of DNA sequence rearrangements and changes in copy-numbers promotes TEs, satDNAs and their transitional forms into crucial players that shape the genomes' architecture *(3, 29)*.

Genome sequencing and assembly is significantly challenged by repetitive DNA sequences

Genomic regions enriched in repetitive sequences cause significant technical problems in DNA sequencing and assembly. The difficulties arise in reconstructing exact sequential order and length of segments composed of highly similar repeat units present in high copy-numbers. Consequently, repetitive sequences are frequently omitted from datasets available in public databases, or even introduce significant errors in genome annotation *(30)*. For example, major satDNA residing in centromeric chromosomal regions of the beetle *Tribolium castaneum* was estimated to build 17% of the genome *(31)*. While satDNA monomer sequences were found abundant in unassembled genomic scaffolds, they occupied only 0.3% of the genome assembly *(32)*. Next-generation sequencing projects, producing huge number of short reads, have made assembly challenges even more complex *(33)*. Human genome assemblies that employed only Illumina technology and small insert libraries forced to leave out hundreds of megabases of genomic sequences, leading to the conclusion that long-range sequencing approaches must be combined with high-throughput sequencing for comparative genomics analyses and studies of genome evolution *(34)*. In order to deal with high individual polymorphism and abundant repetitive sequences of the oyster *Crassostrea gigas* genome (36% in total; but only 4.8% of tandem repeats), the platform combining fosmid pooling, next-generation sequencing, and hierarchical assembly was used in the genome project, but still leaving >60% of detected repeats unclassified *(35)*.

At the same time, improvements in sequencing technologies and the growing accessibility of genomic datasets in public databases within the last decade have opened the possibility to gain new insights into the poorly explored repetitive fraction of the genomes *(36)*. There has been an explosion of software and database resources specifically targeted at advancing our ability to assess repeat detection and characterization within genomic data, reviewed in *(37, 38)*. In particular, high-throughput strategies combining low-coverage short-read DNA sequencing and specialized bioinformatic tools enabled identification of a complete inventory of repetitive DNAs in the genome, the repeatome and the satellitome *(39–41)*. These approaches are particularly useful in exploring the content of repetitive DNAs in non-model species lacking sequenced genomes, while in assembled genomes they can help in filling the gaps left because of the repetitive sequences. Knowledge about the whole-genome composition and distribution of repetitive DNAs is a valuable step towards better understanding of the entire repetitive landscape, genome architecture and functioning as a whole, for example *(42)*. However, it cannot address the second question regarding the repetitive DNAs in a genome, namely, how to determine precise sequential order of tandem repeats in long arrays. The appropriate solution to this problem may be offered by *de novo* sequencing using single-molecule long-read methodology, which also appeared useful in studies of long satDNA arrays *(43)*.

While more and more tools are focused on detection of all types of repeats, trying to give a comprehensive repeatome analysis, there are still specialized tools focused on a specific type of repetitive DNA sequences that enable more detailed insight into some of these important genomic components. In the following paragraphs we provide a brief overview of currently most prominent bioinformatic tools used in repeat detection and classification. Readers must be aware that this list is not exhaustive, and that many other programs exist, as some of the older ones are slowly being abandoned and new programs are constantly being published.

Tools and databases used for detection and classification of different types of repetitive DNA sequences

One of the assets, covering the wide spectrum of repeats, is RepeatExplorer *(44)*, a collection of software tools accessible via the web interface. It is a computational pipeline that uses a sequence-clustering algorithm to enable novel repeat identification, with no need for a reference database of known elements. Ideal input is NGS data of low genome coverage, preferably <0.5x. The system can process up to several millions of short sequence reads. Implemented tools enable classification of identified repeats, determine phylogenetic relationships among retroelements, allow extraction of repetitive sequences associated specifically with the centromeric region, and perform comparative analysis of repeat composition between multiple species.

The next one is Repbase Update *(45)*, which is a comprehensive database of repetitive elements from diverse eukaryotic organisms. It allows searches against annotated sequences representing different families and subfamilies of repeats, many of which are not reported anywhere else. It is being used in genome sequencing projects worldwide as a reference collection for masking and annotation of repetitive DNA sequences. Part of it is a separate electronic journal, Repbase Reports, which publishes information on all new data deposited to Repbase.

In continuation is RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker), a web server that screens DNA sequences in search for interspersed repeats and low complexity repetitive DNA sequences. The output of this program is a detailed annotation of repeats that are present in the query sequence as well as a modified version of the query sequence in which all of the annotated repeats have been masked. RepeatMasker makes use of curated Repbase library of repeats and currently supports Dfam classification system that combines concepts from several classification systems with phylogenies based on reverse transcriptase and transposases. This web server also hosts RepeatModeler (http://www.repeatmasker.org/RepeatModeler), a package used for *de novo* repeat family identification in sequenced genomes. The latter is based on two repeat finding programs, RECON (http://selab.janelia.org/recon.html) and RepeatScout *(46)*, which employ complementary computational methods for identifying repeat-element boundaries and repeat-family relationships. RepeatModeler uses their output to generate, refine and classify consensus sequences of putative interspersed repeats.

Another tool oriented towards wide spectrum of repeats is Red *(47)*. It focuses on transposons and simple repeats. The input to the system is FASTA format, with program being capable of processing assembled or unassembled genomes. Red outputs contain detected repeats, masked sequences and the genomic locations of the regions of interest.

Repeat-finding tool aimed at prokaryotic genome can also be found. Prokaryotic Repeats Annotation Program software package (PRAP) *(48)* is oriented to automated *ab initio* identification of wide spectrum of repeats within the prokaryotic genome, working on completed and draft genomes.

Softwares dedicated to detection of interspersed repeats

Among the most prominent tools focused mainly towards transposable elements are: REPCLASS *(49)*, TE-Locate *(50)*, TESeeker *(51)*, REPET (https://urgi.versailles.inra.fr/Tools/REPET), Generic Repeat Finder *(52)*, already mentioned RECON, and similar.

REPCLASS processes single FASTA files, and the entry passes through the three classification modules that are based on: homology, structure, and target site duplications. In the final step, results of the three modules are compared, ranked, and integrated, yielding a single tentative classification, supplemented with a description of the characterized structural features.

TE-Locate uses paired-end next-generation sequencing data reads to identify novel locations of known TEs. It utilizes either a database of TE sequences, or annotated TEs within the reference sequence.

TESeeker approach also begins with BLAST searches against the genome using representative TEs for the chosen family. Resulting BLAST hits are extracted and the next step is CAP3 assembly *(53)* in order to obtain a coding sequence. CAP3 results are used for another BLAST search against the genome and hits are processed in the same manner, this time with adding the flanking regions.

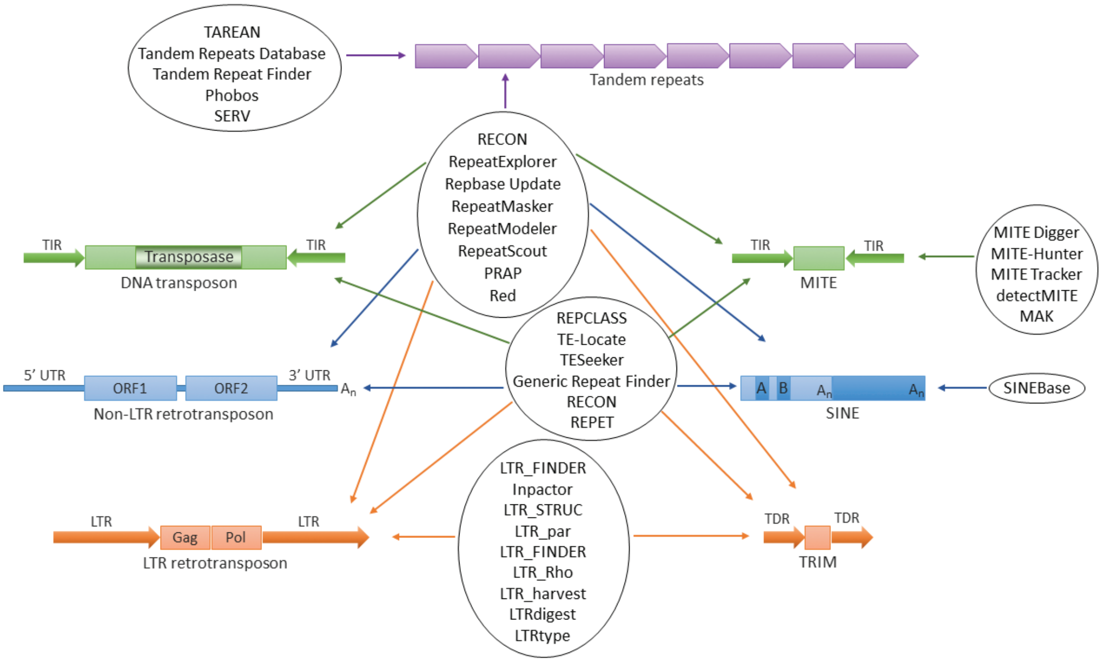

Figure 2. *Graphical overview of tools and databases for repeat detection and characterization, and types of repetitive DNA sequences they focus on.*

This approach does not enable detection of elements without coding regions, like MITE or SINE.

REPET starts with all-to-all alignment of genomic sequences, followed by clustering of matches and generating multiple sequence alignment for each cluster. This step is followed by classification of the consensus sequences, redundancy elimination, and comparison of novel annotations with those from existing TE databases. Later steps include identification of structural variations within identified TE families and their manual inspection.

Generic Repeat Finder (GRF) is a high-sensitivity tool for genome-wide *de novo* repeat detection, integrated with optimized dynamic programming strategies. In that respect, GRF sensitively identifies interspersed repeats that bear both inverted and direct repeats based on the fast and exhaustive numerical calculation algorithms. GRF also helps improve the annotation for various DNA transposons and retrotransposons, such as MITEs, LTR retrotransposons, and non-LTR retrotransposons, including LINEs and SINEs.

Tools and databases aiming at the detection of tandem repeats

While many tools are focused on detection of different types of repetitive DNA sequences, there are those specialized for a specific type of repeats (Figure 2).

Most prominent tools and softwares focused specifically on detection of tandem repeats are: Tandem Repeats Database *(54)*, Tandem Repeat Finder *(55)*, Phobos (http://www.ruhr-uni-bochum.de/ecoevo/cm/cm_phobos.htm), and TAREAN *(56)*.

Tandem Repeats Database (TRDB) works on assembled genomes and contains a set of tools for repeat analysis, also implementing the Tandem Repeats Finder (TRF) program. TRDB provides many filtering options for finding particular repeats of interest, runs similarity-based repeat clustering, does polymorphism prediction based on shared patterns of mutation, enables PCR primer selection and data download in a variety of formats.

TRF program is simple and fast, yielding two output files: the table file and alignment file. The table file contains information about each repeat, including its location, size, number of copies, percent of matches and indels between adjacent copies, and nucleotide composition.

Phobos uses assembled reads and has high detection and alignment quality for repeats of small monomer range (1–50 bp). That capacity is reduced for repeats of bigger pattern size, with detection limit of 10,000 bp-long monomers. It also detects and reports overlapping satellites and shows flanking regions of the repeats.

Tandem Repeat Analyzer (TAREAN) pipeline uses NGS reads of low genome coverage and reconstructs consensus monomer sequences of tandem repeats by using improved RepeatExplorer protocol *(56)*. It uses input data to perform graph-based repeat clustering, followed by examination of obtained clusters for the presence of circular structures, distinctive for tandem repeats. Most frequent multimer fractions reconstructed during this process are used for generating consensus sequences of each satDNA. Information from paired-end reads is used to distinguish clusters belonging to potential satellite repeats from other types of sequences repeated in tandem.

Additional toolkit based on RepeatExplorer is satMiner, specialized for detection of low-copy tandem repeats in the genome by filtering out reads of satDNAs detected in each of several consecutive cycles of RepeatExplorer analysis. This modification enables detection of tandem repeats present in low copy numbers in large genomes *(40)*.

In addition, a specialized program repeatConnector has been developed for screening next-generation datasets in order to find particular satDNAs in different species and to analyze them *(57)*.

A tool that can be used for estimation of minisatellite and microsatellite repeat variability in the genomes has also been developed, named SERV *(58)*. It uses three parameters (number of repeated units, repeat length, and identity) to produce a numeric "VARscore", which can be used for genotyping and forensic purposes.

Tools and databases focused on specific types of repeats

Specialized tools that are focused specifically on a certain type of interspersed elements also exist. Among them is LTR_FINDER *(59)*, web server developed for *de novo* detection of LTR retrotransposons. It predicts locations and structure of full-length LTR retrotransposons by recognizing structural features common for these elements, such as long terminal repeats, primer binding sites, reverse transcriptase, integrase, and RNaseH domains. The output shows LTR sizes, element location, identity between two LTRs, sharpness (prediction reliability of LTR boundaries) etc.

Inpactor (Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons) *(60)* is also a pipeline aimed at classification of this type of retroelements. It identifies both autonomous and non-autonomous retrotransposons, generates phylogenetic trees based on RT genes, and analyzes element's time of insertion based on the divergence between two LTR sequences of each copy. Inpactor integrates previously mentioned LTR_FINDER and several other external bioinformatics softwares.

LTR_STRUC program *(61)* has advantages over conventional search methods in the case of LTR retrotransposon families with low sequence homology to queries, or in the case of non-autonomous elements lacking canonical retroviral ORFs. For each LTR retrotransposon found, LTR_STRUC automatically generates an analysis of a variety of structural features of biological interest. Output file contains: name and length of a source contig, location and orientation of the element within the contig, length of the element, LTRs, and largest ORF. It also shows nucleotide sequences for the whole element, LTRs, primer binding site, polypurine tract, dinucleotides terminating the LTRs and ORFs, as well as intra-element percent identity of LTRs, and alignment of putative LTRs. Several other programs with similar functions were developed, starting with *de novo* prediction of LTRs, and taking into consideration other features and constitutive components of LTR retrotransposons in later processing steps to enhance the quality or sensitivity of the predictions. Examples are: LTR_par *(62)*, LTR_Rho *(63)*, LTR_harvest *(64)*, LTRdigest *(65)*, and others.

More recently developed software, LTRtype *(66)*, is intended to characterize different types of structurally complex LTR retrotransposon elements, in addition to the canonical LTR retrotransposons. Such include: solo-LTR elements, elements with three or more LTRs, fragmented forms of all mentioned elements, and nested events where TEs served as a hotspot for further insertions.

Website oriented towards non-autonomous non-LTR retrotransposons is SINEBase *(67)*. It can be used for exploring the existing database of different SINE families, or to analyze individual modules of a candidate SINE sequence. Four databases can be included in SINESearch: SINEBank, RNABank, LINEBank, and COREBank (holding consensus sequences of SINE central domains).

In searches oriented specifically towards non-autonomous DNA transposons (MITEs), several programs can be employed. Among most frequently used are MITE Digger *(68)*, MITE-Hunter *(69)*, MAK *(70)*, detectMITE *(71)* and MITE Tracker *(72)*. All programs have the ability to process genome-scale inputs, splicing them into shorter fragments and starting with the search for inverted repeats. In continuation, algorithms of different complexity are employed in these programs. Comparative analyses have shown that detectMITE is one of the most efficient, precise and comprehensive in detecting MITEs, while meticulous filtering of false positives makes MITE Tracker the most accurate. Of course, additional programs performing similar functions can be found, in addition to the abovelisted.

CONCLUSION

Quickly evolving sequencing technologies are rapidly advancing the availability of genomic data across many taxa. In respect to that, comprehensive and accurate annotation and characterization of repetitive sequences is an important contribution to the understanding of genomic architecture and function as a whole. For that purpose, many bioinformatics tools, databases and pipelines have been created to attend the emerging needs in repeat analysis. Some of these programs are focused on a specific type of analysis or on a specific repeat type, and intended to answer specific scientific questions. Others try to give a broader overview, although not a single program so far has proven to be sufficiently exhaustive, making the employment and improvement of others unnecessary. For that purpose, new programs are constantly being published, co-evolving with sequencing strategies, available data and specific needs of the research community.

Acknowledgments: *This review paper was prepared in the frame of Croatian Science Foundation project IP-09-2014-3183.*

REFERENCES

1. PLOHL M, MEŠTROVIĆ N, MRAVINAC B 2012 Satellite DNA evolution. In: Garrido- Ramos M (ed) Genome dynamics, Karger AG, Basel, p 126
2. LÓPEZ-FLORES I, GARRIDO-RAMOS MA 2012 The repetitive DNA content of eukaryotic genomes. In: Garrido- Ramos M (ed) Genome Dynamics, Karger AG, Basel, p 1
3. GARRIDO-RAMOS MA 2017 Satellite DNA: An evolving topic. Genes (Basel) 8(230): 1–41 https://doi.org/10.3390/genes8090230
4. LOUZADA S, LOPES M, FERREIRA D, ADEGA F, ESCUDEIRO A, GAMA-CARVALHO M, CHAVES R 2020 Decoding the role of satellite dna in genome architecture and plasticity — an evolutionary and clinical affair. Genes (Basel) 11:72 https://doi.org/10.3390/genes11010072
5. DOVER GA 1986 Molecular drive in multigene families: How biological novelties arise, spread and are assimilated. Trends Genet 2: 159–165 https://doi.org/10.1016/0168-9525(86)90211-8
6. PLOHL M, LUCHETTI A, MEŠTROVIĆ N, MANTOVANI B 2008 Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409(1–2): 72–82 https://doi.org/10.1016/j.gene.2007.11.013
7. HENIKOFF S, AHMAD K, MALIK HS 2001 The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293(5532): 1098–1102 https://doi.org/10.1126/science.1062939
8. HARTLEY G, O'NEILL R 2019 Centromere Repeats: hidden gems of the genome. Genes (Basel) 10(3): 223 https://doi.org/10.3390/genes10030223
9. FERREE PM, BARBASH DA 2009 Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. PLoS Biol 7(10): e1000234 https://doi.org/10.1371/journal.pbio.1000234
10. FELICIELLO I, AKRAP I, UGARKOVIĆ Đ 2015 Satellite DNA modulates gene expression in the beetle *Tribolium castaneum* after heat stress. PLoS Genet 11(8): 1–18 https://doi.org/10.1371/journal.pgen.1005466
11. BISCOTTI MA, OLMO E, HESLOP-HARRISON JS 2015 Repetitive DNA in eukaryotic genomes. Chromosom Res 23(3): 415–420 https://doi.org/10.1007/s10577-015-9499-z
12. FERREIRA D, ESCUDEIRO A, ADEGA F, CHAVES R 2019 DNA Methylation patterns of a satellite non-coding sequence – FA-SAT in cancer cells: Its expression cannot be explained solely by DNA methylation. Front Genet 10(101): 1–10 https://doi.org/10.3389/fgene.2019.00101
13. FINNEGAN DJ 1989 Eukaryotic transposable elements and genome evolution. Trends Genet 5: 103–107 https://doi.org/10.1016/0168-9525(89)90039-5

14. JURKA J, KAPITONOV VV, KOHANY O, JURKA MV 2007 Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet 8: 241–59 https://doi.org/10.1146/annurev.genom.8.080706.092416

15. KOJIMA KK 2019 Structural and sequence diversity of eukaryotic transposable elements. Genes Genet Syst 94(6): 233–252 https://doi.org/10.1266/ggs.18-00024

16. CRAIG NL 1995 Unity in transposition reactions. Science 270(5234): 253–253 https://doi.org/10.1126/science.270.5234.253

17. CAPY P 2005 Classification and nomenclature of retrotransposable elements. Cytogenet Genome Res 110(1–4): 457–61 https://doi.org/10.1159/000084978

18. WITTE CP, LE QH, BUREAU T, KUMAR A 2001 Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 98(24): 13778–13783 https://doi.org/10.1073/pnas.241341898

19. WONG LH, CHOO KHA 2004 Evolutionary dynamics of transposable elements at the centromere. Trends Genet 20(12): 611–616 https://doi.org/10.1016/j.tig.2004.09.011

20. KAZAZIAN HH 2004 Mobile elements: drivers of genome evolution. Science 303(5664): 1626–32 https://doi.org/10.1126/science.1089670

21. FESCHOTTE C, ZHANG X, WESSLER S 2002 Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. In: Craig N (ed) Mobile DNA II, ASM Press, Washington, p 1147

22. FESCHOTTE C, JIANG N, WESSLER SR 2002 Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3(5): 329–341 https://doi.org/10.1038/nrg793

23. THOMAS J, PRITHAM EJ 2015 Helitrons, the eukaryotic rolling-circle transposable elements. Microbiol Spectr 3(4): 1–32 https://doi.org/10.1128/microbiolspec.MDNA3-0049-2014

24. MEŠTROVIĆ N, MRAVINAC B, PAVLEK M, VOJVODA-ZELJKO T, ŠATOVIĆ E, PLOHL M 2015 Structural and functional liaisons between transposable elements and satellite DNAs. Chromosom Res 23(3): 583–596 https://doi.org/10.1007/s10577-015-9483-7

25. MACAS J, KOBLÍZKOVÁ A, NAVRÁTILOVÁ A, NEUMANN P 2009 Hypervariable 3' UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 448(2): 198–206 https://doi.org/10.1016/j.gene.2009.06.014

26. MCGURK MP, BARBASH DA 2018 Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. Genome Res 28(5): 714–725 https://doi.org/10.1101/gr.231472.117

27. VONDRAK T, ÁVILA ROBLEDILLO L, NOVÁK P, KOBLÍŽKOVÁ A, NEUMANN P, MACAS J 2020 Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. Plant J 101(2): 484–500 https://doi.org/10.1111/tpj.14546

28. DIAS GB, SVARTMAN M, DELPRAT A, RUIZ A, KUHN GCS 2014 *Tetris* is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*. Genome Biol Evol 6(6): 1302–1313 https://doi.org/10.1093/gbe/evu108

29. ESCUDEIRO A, FERREIRA D, MENDES-DA-SILVA A, HESLOP-HARRISON JS, ADEGA F, CHAVES R 2019 Bovine satellite DNAs–a history of the evolution of complexity and its impact in the Bovidae family. Eur Zool J 86(1): 20–37 https://doi.org/10.1080/24750263.2018.1558294

30. TØRRESEN OK, STAR B, MIER P, ANDRADE-NAVARRO MA, BATEMAN A, JARNOT P, GRUCA A, GRYNBERG M, KAJAVA AV, PROMPONAS VJ, ANISIMOVA M, JAKOBSEN KS, LINKE D 2019 Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res 47(21): 10994–11006 https://doi.org/10.1093/nar/gkz841

31. UGARKOVIĆ Đ, PODNAR M, PLOHL M 1996 Satellite DNA of the red flour beetle *Tribolium castaneum*-comparative study of satellites from the genus *Tribolium*. Mol Biol Evol 13(8): 1059–66

32. WANG S, LORENZEN MD, BEEMAN RW, BROWN SJ 2008 Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. Genome Biol 9(3): 1–14 https://doi.org/10.1186/gb-2008-9-3-r61

33. TREANGEN TJ, SALZBERG SL 2012 Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13(1): 36–46 https://doi.org/10.1038/nrg3117

34. ALKAN C, SAJJADIAN S, EICHLER EE 2011 Limitations of next-generation genome sequence assembly. Nat Methods 8(1): 61–65 https://doi.org/10.1038/nmeth.1527

35. ZHANG G, FANG X, GUO X, LI L, LUO R, XU F, YANG P, ZHANG L, WANG X, QI H, XIONG Z, QUE H, XIE Y, HOLLAND PWH, PAPS J, ZHU Y, WU F, CHEN Y, WANG J, PENG C, MENG J, YANG L, LIU J, WEN B, ZHANG N, HUANG Z, ZHU Q, FENG Y, MOUNT A, HEDGECOCK D, XU Z, LIU Y, DOMAZET-LOŠO T, DU Y, SUN X, ZHANG S, LIU B, CHENG P, JIANG X, LI J, FAN D, WANG W, FU W, WANG T, WANG B, ZHANG J, PENG Z, LI Y, LI N, WANG J, CHEN M, HE Y, TAN F, SONG X, ZHENG Q, HUANG R, YANG H, DU X, CHEN L, YANG M, GAFFNEY PM, WANG S, LUO L, SHE Z, MING Y, HUANG W, ZHANG S, HUANG B, ZHANG Y, QU T, NI P, MIAO G, WANG J, WANG Q, STEINBERG CEW, WANG H, LI N, QIAN L, ZHANG G, LI Y, YANG H, LIU X, WANG J, YIN Y, WANG J 2012 The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490(7418): 49–54 https://doi.org/10.1038/nature11413

36. LOWER SS, MCGURK MP, CLARK AG, BARBASH DA 2018 Satellite DNA evolution: old ideas, new approaches. Curr Opin Genet Dev 49(i): 70–78 https://doi.org/10.1016/j.gde.2018.03.003

37. LERAT E 2010 Identifying repeats and transposable elements in sequenced genomes: How to find your way through the dense forest of programs. Heredity (Edinb) 104(6): 520–533 https://doi.org/10.1038/hdy.2009.165

38. SEDLAZECK FJ, LEE H, DARBY CA, SCHATZ MC 2018 Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. Nat Rev Genet 19(6): 329–346 https://doi.org/10.1038/s41576-018-0003-4

39. WEISS-SCHNEEWEISS H, LEITCH AR, MARY Q, MCCANN J, JANG T 2015 Employing next generation sequencing to explore the repeat landscape of the plant genome. In: Hörandl E, Appelhans M (eds) Next Generation Sequencing in Plant Systematics, International Association for Plant Taxonomy, Königstein, Germany, p 1

40. RUIZ-RUANO FJ, LÓPEZ-LEÓN MD, CABRERO J, CAMACHO JPM 2016 High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep 6(1): 28333 https://doi.org/10.1038/srep28333

41. PITA S, PANZERA F, MORA P, VELA J, CUADRADO Á, SÁNCHEZ A, PALOMEQUE T, LORITE P 2017 Comparative repeatome analysis on *Triatoma infestans* Andean and Non-Andean lineages, main vector of Chagas disease. PLoS One 12(7): e0181635 https://doi.org/10.1371/journal.pone.0181635

42. BELYAYEV A, JOSEFIOVÁ J, JANDOVÁ M, KALENDAR R, KRAK K, MANDÁK B 2019 Natural history of a satellite DNA Family: From the ancestral genome component to species-specific sequences, concerted and non-concerted evolution. Int J Mol Sci 20(5):1201 https://doi.org/10.3390/ijms20051201

43. KHOST D, EICKBUSH D, LARRACUENTE A 2017 Single molecule long read sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. Genome Res 27: 1–13 https://doi.org/10.1101/gr.213512.116

44. NOVÁK P, NEUMANN P, PECH J, STEINHAISL J, MACAS J 2013 RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29(6): 792–793 https://doi.org/10.1093/bioinformatics/btt054

45. JURKA J, KAPITONOV V, PAVLICEK A, KLONOWSKI P, KOHANY O, WALICHIEWICZ J 2005 Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110(1–4): 462–7 https://doi.org/10.1159/000084979

46. PRICE AL, JONES NC, PEVZNER PA 2005 *De novo* identification of repeat families in large genomes. Bioinformatics 21(Suppl. 1): i351–i358 https://doi.org/10.1093/bioinformatics/bti1018

47. GIRGIS HZ 2015 Red: An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics 16(1): 1–19 https://doi.org/10.1186/s12859-015-0654-5

48. CHEN GL, CHANG YJ, HSUEH CH 2013 PRAP: An *ab initio* software package for automated genome-wide analysis of DNA repeats for prokaryotes. Bioinformatics 29(21): 2683–2689 https://doi.org/10.1093/bioinformatics/btt482

49. FESCHOTTE C, KESWANI U, RANGANATHAN N, GUIBOTSY ML, LEVINE D 2009 Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol 1(1): 205–220 https://doi.org/10.1093/gbe/evp023

50. PLATZER A, NIZHYNSKA V, LONG Q 2012 TE-Locate: A tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. Biology (Basel) 1(3): 395–410 https://doi.org/10.3390/biology1020395

51. KENNEDY RC, UNGER MF, CHRISTLEY S, COLLINS FH, MADEY GR 2011 An automated homology-based approach for identifying transposable elements. BMC Bioinformatics 12: 130 https://doi.org/10.1186/1471-2105-12-130

52. SHI J, LIANG C 2019 Generic repeat finder: A high-sensitivity tool for genome-wide de novo repeat detection. Plant Physiol 180(4): 1803–1815 https://doi.org/10.1104/pp.19.00386

53. HUANG X, MADAN A 1999 CAP3: A DNA sequence assembly program. Genome Res 9: 868–877

54. GELFAND Y, RODRIGUEZ A, BENSON G 2007 TRDB - the tandem repeats database. Nucleic Acids Res 35(Suppl 1): D80–D87 https://doi.org/10.1093/nar/gkl1013

55. BENSON G 1999 Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2): 573–580 https://doi.org/10.1093/nar/27.2.573

56. NOVÁK P, ROBLEDILLO LÁ, KOBLÍŽKOVÁ A, VRBOVÁ I, NEUMANN P, MACAS J 2017 TAREAN: A computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Res 45(12): e111 https://doi.org/10.1093/nar/gkx257

57. SMALEC BM, HEIDER TN, FLYNN BL, O'NEILL RJ 2019 A centromere satellite concomitant with extensive karyotypic diversity across the *Peromyscus* genus defies predictions of molecular drive. Chromosom Res 27: 237–252 https://doi.org/10.1007/s10577-019-09605-1

58. LEGENDRE M, POCHET N, PAK T, VERSTREPEN KJ 2007 Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res 17(12): 1787–1796 https://doi.org/10.1101/gr.6554007

59. XU Z, WANG H 2007 LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35(Suppl. 2): W265–W268 https://doi.org/10.1093/nar/gkm286

60. OROZCO-ARIAS S, LIU J, TABARES-SOTO R, CEBALLOS D, SILVA DOMINGUES D, GARAVITO A, MING R, GUYOT R 2018 Inpactor, integrated and parallel analyzer and classifier of ltr retrotransposons and its application for pineapple LTR retrotransposons diversity and dynamics. Biology 7(2): 32 https://doi.org/10.3390/biology7020032

61. MCCARTHY EM, MCDONALD JF 2003 LTR STRUC: A novel search and identification program for LTR retrotransposons. Bioinformatics 19(3): 362–367 https://doi.org/10.1093/bioinformatics/btf878

62. KALYANARAMAN A, ALURU S 2005 Efficient algorithms and software for detection of full-length LTR retrotransposons. Proc IEEE Comput Syst Bioinform Conf 56–64 https://doi.org/10.1109/csb.2005.31

63. RHO M, CHOI JH, KIM S, LYNCH M, TANG H 2007 *De novo* identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics 8: 1–16 https://doi.org/10.1186/1471-2164-8-90

64. ELLINGHAUS D, KURTZ S, WILLHOEFT U 2008 LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9: 18 https://doi.org/10.1186/1471-2105-9-18

65. STEINBISS S, WILLHOEFT U, GREMME G, KURTZ S 2009 Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res 37(21): 7002–7013 https://doi.org/10.1093/nar/gkp759

66. ZENG FC, ZHAO YJ, ZHANG QJ, GAO LZ 2017 LTRtype, an efficient tool to characterize structurally complex LTR retrotransposons and nested insertions on genomes. Front Plant Sci 8(April): 1–9 https://doi.org/10.3389/fpls.2017.00402

67. VASSETZKY NS, KRAMEROV DA 2013 SINEBase: A database and tool for SINE analysis. Nucleic Acids Res 41(D1): 83–89 https://doi.org/10.1093/nar/gks1263

68. YANG G 2013 MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. BMC Bioinformatics 14: 186 https://doi.org/10.1186/1471-2105-14-186

69. HAN Y, WESSLER SR 2010 MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38(22): 1–8 https://doi.org/10.1093/nar/gkq862

70. YANG G 2003 MAK, a computational tool kit for automated MITE analysis. Nucleic Acids Res 31(13): 3659–3665 https://doi.org/10.1093/nar/gkg531

71. YE C, JI G, LIANG C 2016 DetectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes. Sci Rep 6: 19688 https://doi.org/10.1038/srep19688

72. CRESCENTE JM, ZAVALLO D, HELGUERA M, VANZETTI LS 2018 MITE Tracker: An accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics 19: 348. https://doi.org/10.1186/s12859-018-2376-y