BMC Genomics

SOFTWARE

Open Access

Krait2: a versatile software for microsatellite investigation, visualization and marker development

Lianming Du[1*], Jiahao Chen[1], Dalin Sun[1], Kelei Zhao[1], Qianglin Zeng[1] and Nan Yang[2*]

Abstract

Background Microsatellites are highly polymorphic repeat sequences ubiquitously interspersed throughout almost all genomes which are widely used as powerful molecular markers in diverse fields. Microsatellite expansions play pivotal roles in gene expression regulation and are implicated in various neurological diseases and cancers. Although much effort has been devoted to developing efficient tools for microsatellite identification, there is still a lack of a powerful tool for large-scale microsatellite analysis.

Results We present Krait2, a user-friendly graphical tool for investigating perfect, imperfect and compound microsatellites from FASTA and FASTQ formatted genomic datasets. Krait2 not only provides features such as primer design, repeat filtering, repeat annotation and statistical analysis, but also offers various output formats to support customized downstream analysis. Moreover, it has capability of analyzing multiple genomes simultaneously and conducting comparative analysis.

Conclusions Krait2 is a versatile and easy-to-use software for both novices and experts to identify and analyze microsatellites. The installer and source code are available at https://github.com/lmdu/krait2.

Keywords Microsatellite, Tandem repeats, Genomes, Genetic marker, Primer design

Background

Microsatellites, also termed as simple sequence repeats (SSRs) or short tandem repeats (STRs), consist of continuously repeated short DNA sequences with length varying from one to six base pairs [1]. Microsatellites are ubiquitous occurrence in nearly all genomes and highly abundant in eukaryote genomes, covering around 3% of the human genome [2]. During DNA replication, microsatellites are prone to slippage with the addition or deletion of repeat units leading to widespread length polymorphisms also known as microsatellite instability [3]. Microsatellites represent a large source of genetic variation and a powerful molecular tool to estimate genetic diversity and differentiation among populations, particularly in conservation genetics [4, 5]. In addition to genotypes, microsatellites can also affect phenotypes by altering the gene expression, alternative splicing, transcription factor binding and methylation [6–8]. Moreover, their instability or expansion are significantly implicated in various human genetic diseases, even in several cancers [9, 10].

In recent years, the rapid development of next-generation sequencing (NGS) technology makes single

*Correspondence:
Lianming Du
dulianming@cdu.edu.cn
Nan Yang
yangnan0204@126.com
[1] Antibiotics Research and Re-Evaluation Key Laboratory of Sichuan Province, School of Pharmacy, Chengdu University, Chengdu 610106, China
[2] Key Laboratory of Qinghai-Tibetan Plateau Animal Genetic Resource Reservation and Utilization, Sichuan Province and Ministry of Education, Southwest Minzu University, Chengdu 610225, China

nucleotide polymorphisms (SNPs) to be attractive genetic makers. Although several studies have demonstrated that SNPs outperform microsatellites in genetic analysis owing to their higher abundance, better accuracy and well-understood mutational mechanism [11–13], SNPs cannot completely replace microsatellites in genetic diversity assessment [14]. Compared to typically biallelic SNPs, microsatellites possess a higher mutation rate and are mostly multiallelic per locus making them more sensitive in detecting genetic variations within populations [15, 16]. Because of their greater per-locus information content, microsatellites have been proved to be preferable markers for parentage and assignment studies as well as forensic identification, regardless of sample size [17, 18]. In addition, recent studies tend to combine microsatellites with SNPs to leverage the strengths of both marker types, enhancing the resolution and accuracy of genetic analyses [19, 20]. Moreover, SNP data analysis often requires more sophisticated bioinformatics tools and computational resources, whereas microsatellite datasets are relatively small, making them easier to process using well-established software solutions [21]. Therefore, microsatellites remain an economical, informative and easy-to-use technology for population and conservation genetics.

Over the past two decades, much effort has been devoted to developing microsatellite identification tools and designing tandem repeat search algorithms [22, 23]. There are several tools with graphical user interfaces that have attracted a wide range of users across various fields due to their ease of use, flexibility and powerful functionality, such as web-based tools including SSRIT [24], MISA-web [25], EasySSR [26], MegaSSR [27] and desktop applications involving SciRoKo [28], msatcommander [29], GMATA [30], MSDB [31], Krait [32]. However, these tools are only suitable for detecting microsatellites from small amounts of sequence data or small genomes. The boom of NGS has uncovered vast complex genomes and generates unprecedented amounts of sequencing data which brings a big challenge to microsatellite identification. To meet the urgent demand, we developed Krait2, an updated version of the Krait. Krait2 is a versatile tool for exploiting microsatellites from large-scale genome sequences. Krait2 also provides an intuitive graphical user interface for facilitating microsatellite identification, visualization, annotation, primer design and comparative analysis.

Implementation

Krait2 is written in Python programming language and its intuitive graphical interface is developed using PySide6 (https://doc.qt.io/qtforpython-6). It is designed to run as standalone desktop application on Windows, MacOS and Linux operating systems. The functionality and workflow of Krait2 are shown in Fig. 1. The application utilizes pyfastx [33] to parse and index FASTA and FASTQ sequence files. The indexed sequence files allow fast random access to subsequence without loading entire sequence, for example retrieval of microsatellite flanking sequences. We employ pytrf (https://github.com/lmdu/pytrf) to identify perfect, imperfect and compound microsatellites as well as generic tandem repeats. The pytrf is a Python library built on the Krait algorithm, which has improved the accuracy of searching for exact and approximate tandem repeats. We have integrated primer3 [34] into the software along with primer3-py (https://github.com/libnano/primer3-py) for designing primers. The pygros, a Python binding to cgranges [35], is applied to find coordinate overlaps between microsatellites and gene features to annotate microsatellites. Finally, the front-end frameworks including tabler (https://tabler.io), datatables (https://datatables.net/) and echarts [36] are adopted to generate HTML statistics analysis report. All the input and output datasets are saved into different tables in an SQLite database file which can be shared by any other machines and systems.

Results

Comparison with other tools

We have performed comprehensive comparison between Krait2 and several other similar graphic tools in certain aspects of features. The comparison results are depicted in Fig. 2. Although web-based tools are simple and easy to use with well cross-platform compatibility, they are generally restricted by the requirement of stable internet environment and server-side shared computational resources, resulting in limited processing power for large datasets. In contrast, desktop applications like Krait and Krait2 can leverage local hardware resources to effectively deal with large-scale genome datasets. As shown in Fig. 2, Krait and Krait2 have more specific functionalities than other tools. Both of them can accept gzipped FASTA formatted sequence files as input, which significantly reduces disk usage, especially for ultra-large genomes. They not only are able to identify perfect, imperfect and compound microsatellites but also are extended to find generic tandem repeats with any motif size. In addition to statistical function found in most tools, they provide advanced features that are rarely supported by other tools, such as filtering microsatellites according to motif, position, repeat number, length, etc., designing primer sequences, locating to gene regions and viewing microsatellite sequences.

Compared with the previous version, Krait2 has added some new features and improved performance in many aspects. It enables users to directly find microsatellites

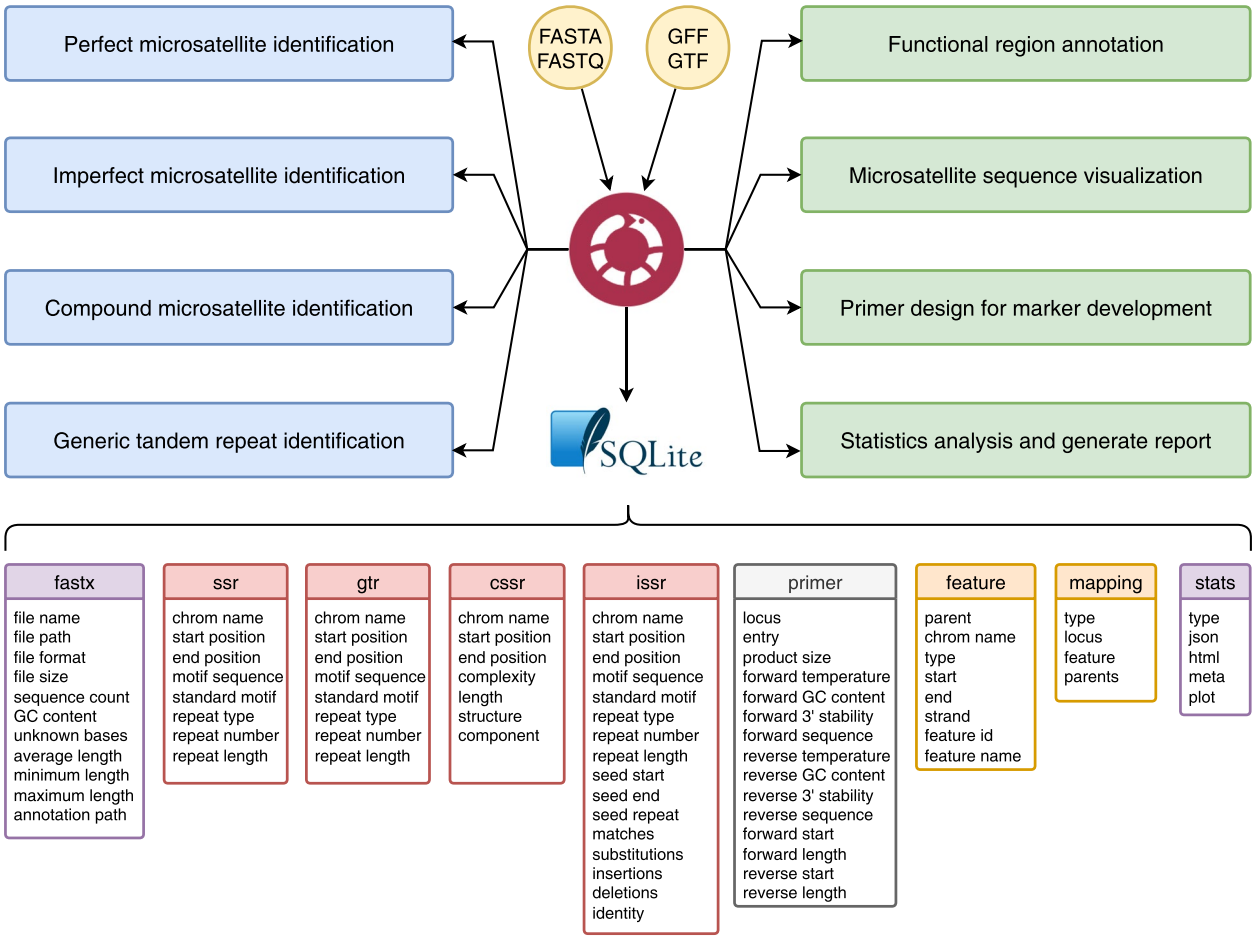

Fig. 1 The workflow of the Krait2 and the structure of the data backend

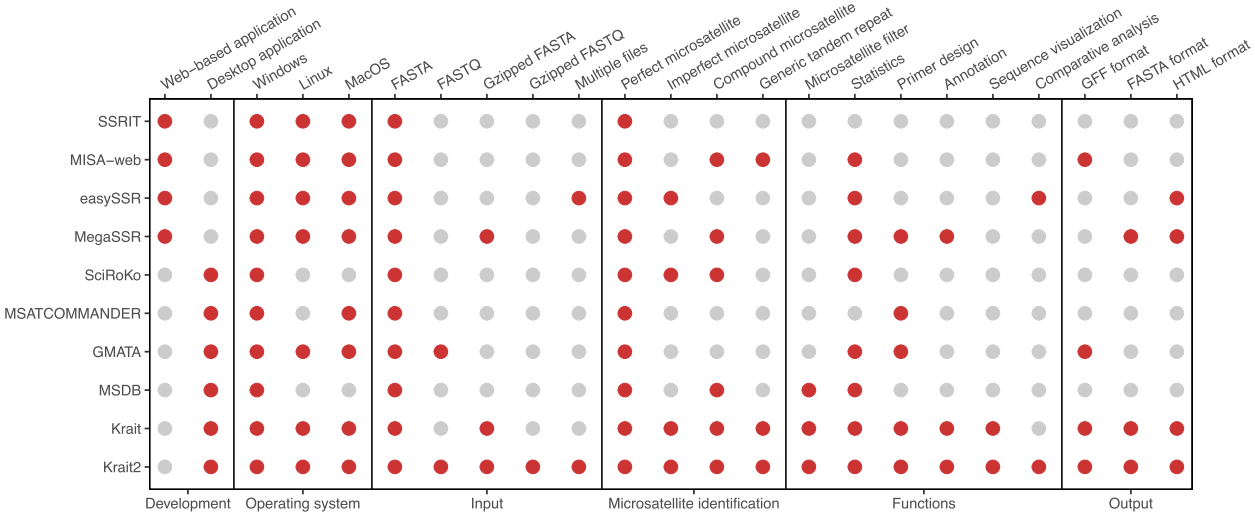

Fig. 2 Comparison of features between Krait2 and other tools. The red dots indicate feature support

from FASTQ or gzipped FASTQ files. Krait2 has the capability to detect microsatellites from plenty of genomes at once and conduct comparative analysis. Krait2 offers comparison of microsatellite distribution patterns, including microsatellite frequency and density, motif abundance and distribution in different gene regions (See Additional File 1). Except for showing microsatellites with flanking sequence, the sequence viewer has been extended to display the highlighted imperfect microsatellites and primer sequence locations. After annotation, the Krait2 also allows users to examine which genes and gene features the microsatellites are located in. Moreover, the Krait2 has greatly improved the performance of annotation and the accuracy for finding imperfect microsatellites.

Software overview and usage

The main window of Krait2 is composed of a toolbar, one fixed area for data view and three removable panels which can be floated as an independent window (Fig. 3). All the imported sequence files will be separately listed in file panel (Fig. 3A). When the annotation file with the same name as the sequence file is imported, the corresponding sequence file name will be displayed in bold in the file panel. Then, you can click the toolbar buttons to perform various microsatellite analysis (Fig. 3B). After microsatellite analysis, Krait2 allows users to obtain detailed information of sequence files, identified microsatellites, designed primers and statistical results in data view panel by clicking sequence file name (Fig. 3C). If a microsatellite is located to a gene region, you can click on it to view the specific feature of that gene in the annotation panel (Fig. 3D). Simultaneously, the repeat sequence and flanking sequence of the clicked microsatellite will be shown in sequence panel (Fig. 3E). If the clicked microsatellite is imperfect, you can go to the alignment tab to examine the pattern alignment result between it and its perfect counterpart (Fig. 3F).

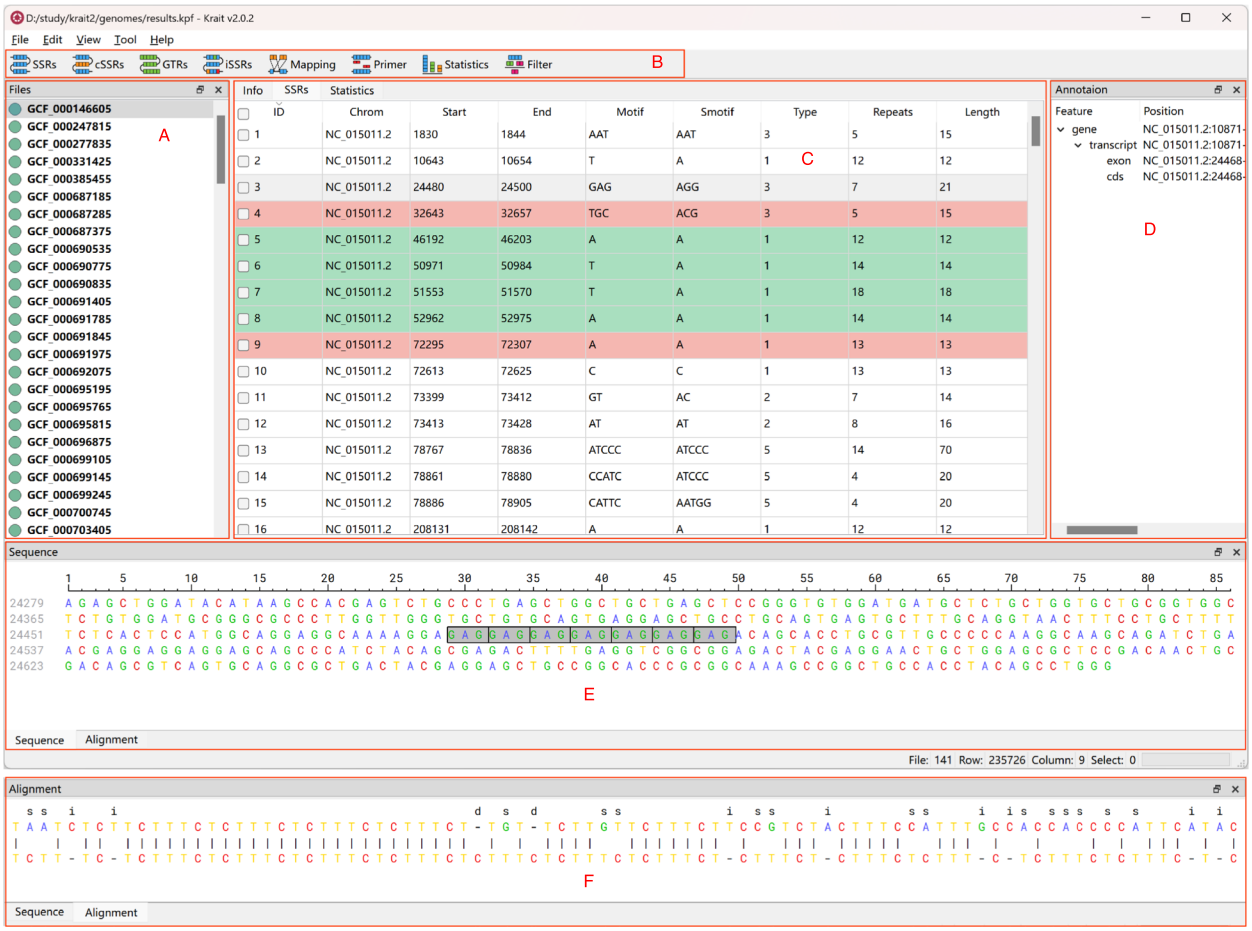

Fig. 3 The overview of the Krait2 main window. **A** Input file list. **B** Tools for performing analysis. **C** Tables for showing results. **D** Repeat annotation information. **E** Repeat sequence visualization. **F** Alignment patterns between imperfect microsatellite and its perfect counterpart

Input and Output

Krait2 can read DNA sequences from FASTA and FASTQ formatted files. In order to save disk space, Krait2 is also capable of parsing sequences directly from gzipped FASTA and FASTQ files. For FASTQ file, the long-read sequencing dataset is recommended to analyze microsatellites due to its unparalleled accuracy in identifying tandem repeats [37]. Krait2 can extract gene features from GFF and GTF formatted files to annotate microsatellites. The microsatellite search results can be exported into CSV, TSV and GFF formatted table files. The repeat sequence and flanking sequence of microsatellites can be exported as FASTA formatted files. Moreover, Krait2 has the ability to generate HTML formatted statistical report files which offer interactive charts and data tables. Finally, all results can be saved to a project file with.kpf extension, allowing them to be easily reused by Krait2.

Case Study

We have screened 141 annotated avian genomes across 36 orders and 69 families from NCBI RefSeq database. The corresponding genome FASTA files and GTF annotation files have been downloaded using NCBI Datasets [38]. We have identified microsatellites from these genomes using Krait2 with default parameters and performed distribution comparative analysis. The size of these genomes ranges from 0.93 Gb to 1.54 Gb and the GC content varies from 40.76% to 46.81% (See Additional File 2). In total, we identified 38,208,452 perfect microsatellites covering about 0.49% ~ 5.05% of the genome sequence. The relative abundance or frequency ranged from 102.53 to 648.08 loci/Mb and the relative density ranged between 1891.42 and 50,457.82 bp/Mb (See Additional File 3). We observed that both relative abundance and relative density have no significant correlation with genome size (See Additional File 4), which is consistent with previous study in birds [39]. We have calculated the z-score of density and frequency which showed no obvious taxon-specific variation (Fig. 4A). Mono-nucleotide

Fig. 4 SSR distribution analysis of avian genomes across multiple orders. **A** Z-score of SSR density and frequency. **B** SSR counts distribution of different types. **C** SSR counts distribution in different gene regions

microsatellites are the most abundant in almost all avian genomes, except for *Patagioenas fasciata* and *Falco peregrinus* which are dominated by tetra-nucleotides and hexa-nucleotides, respectively (Fig. 4B). More than 90% of microsatellites are located in non-coding regions, most of which are located in intergenic regions, followed by introns (Fig. 4C). These findings are in line with previous studies in eukaryotic genomes [40], especially in avian genomes [41].

Conclusions

In this study, we introduce Krait2, a user-friendly graphical tool for investigating microsatellites from genomic datasets. Krait2 enables researchers to search for perfect, imperfect and compound microsatellites with custom parameters from both FASTA and FASTQ files as well as gzipped genomic data. To our knowledge, Krait2 is the most versatile microsatellite processing tool with functions including primer design, sequence visualization, repeat filter, annotation and comparative analysis. In addition, all data in Krait2 can be saved to a project file for repurposing and exported to various output format files for downstream analysis. In summary, these features make Krait2 easy to use for both novices and experts to detect microsatellites and perform comprehensive analysis.

Availability and requirements

Project name: Krait2.
Project home page: https://github.com/lmdu/krait2
Operating system(s): Windows, Linux, MacOS.
Programming language: Python.
Other requirements: none.
License: MIT.
Any restrictions to use by non-academics: No.

Abbreviations

SSRs	Simple sequence repeats
STRs	Short tandem repeats
NGS	Next-generation sequencing
SNPs	Single nucleotide polymorphisms
GFF	Generic feature format
GTF	Gene transfer format
TSV	Tab-separated values
CSV	Comma-separated values
HTML	Hypertext markup language

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11252-2.

Supplementary Material 1.
Supplementary Material 2.
Supplementary Material 3.
Supplementary Material 4.

Acknowledgements

Not applicable.

Authors' contributions

LD and NY conceived the project. LD implemented the software and was a major contributor in writing the manuscript. JC, DS and QZ test the software. KZ revised the manuscript. NY supervised the project. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China [No. 32200525].
National Natural Science Foundation of China,32200525

Data availability

The accession number of avian genomes used in case study can be found in supplementary files. Project name: Krait2 Project home page: https://github.com/lmdu/krait2 Operating system(s): Windows, Linux, MacOS Programming language: Python Other requirements: none License: MIT Any restrictions to use by non-academics: No.
Project name: Krait2.
Project home page: https://github.com/lmdu/krait2
Operating system(s): Windows, Linux, MacOS.
Programming language: Python.
Other requirements: none.
License: MIT.
Any restrictions to use by non-academics: No.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 25 November 2024 Accepted: 16 January 2025
Published online: 25 January 2025

References

1. Du L, Liu Q, Zhao K, Tang J, Zhang X, Yue B, Fan Z. PSMD: An extensive database for pan-species microsatellite investigation and marker development. Mol Ecol Resour. 2020;20(1):283–91.
2. Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol. 2003;4(2):R13.
3. Aska EM, Zagidullin B, Pitkänen E, Kauppi L. Single-Cell Mononucleotide Microsatellite Analysis Reveals Differential Insertion-Deletion Dynamics in Mouse T Cells. Front Genet. 2022;13: 913163.
4. Abdul-Muneer PM. Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. Genet Res Int. 2014;2014: 691759.
5. Casado-Amezúa P, García-Jiménez R, Kersting DK, Templado J, Coffroth MA, Merino P, Acevedo I, Machordom A. Development of microsatellite markers as a molecular tool for conservation studies of the Mediterranean reef builder coral Cladocora caespitosa (*Anthozoa, Scleractinia*). J Hered. 2011;102(5):622–6.
6. Wright SE, Todd PK. Native functions of short tandem repeats. Elife. 2023;20(12):e84043.
7. Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, Goren A, Gymrek M. The impact of short tandem repeat variation on gene expression. Nat Genet. 2019;51(11):1652–9.

8. Martin-Trujillo A, Garg P, Patel N, Jadhav B, Sharp AJ. Genome-wide evaluation of the effect of short tandem repeat variation on local DNA methylation. Genome Res. 2023;33(2):184–96.
9. Yoon JG, Lee S, Cho J, Kim N, Kim S, Kim MJ, Kim SY, Moon J, Chae JH. Diagnostic uplift through the implementation of short tandem repeat analysis using exome sequencing. Eur J Hum Genet. 2024;32(5):584–7.
10. Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, Tchan M, Fung V, Ng K, Cortese A, Houlden H, Dobson-Stone C, Fitzpatrick L, Halliday G, Ravenscroft G, Davis MR, Laing NG, Fellner A, Kennerson M, Kumar KR, Deveson IW. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. Sci Adv. 2022;8(9):eabm5386.
11. Kaiser SA, Taylor SA, Chen N, Sillett TS, Bondra ER, Webster MS. A comparative assessment of SNP and microsatellite markers for assigning parentage in a socially monogamous bird. Mol Ecol Resour. 2017;17(2):183–93.
12. Camacho-Sanchez M, Velo-Antón G, Hanson JO, Veríssimo A, Martínez-Solano Í, Marques A, Moritz C, Carvalho SB. Comparative assessment of range-wide patterns of genetic diversity and structure with SNPs and microsatellites: A case study with Iberian amphibians. Ecol Evol. 2020;10(19):10353–63.
13. Skey ED, Ottewell KM, Spencer PB, Shaw RE. Empirical landscape genetic comparison of single nucleotide polymorphisms and microsatellites in three arid-zone mammals with high dispersal capacity. Ecol Evol. 2023;13(5):e10037.
14. Zimmerman SJ, Aldridge CL, Oyler-McCance SJ. An empirical comparison of population genetic analyses using microsatellite and SNP data for a species of conservation concern. BMC Genomics. 2020;21(1):382.
15. Tsykun T, Rellstab C, Dutech C, Sipos G, Prospero S. Comparative assessment of SSR and SNP markers for inferring the population genetic structure of the common fungus Armillaria cepistipes. Heredity. 2017;119(5):371–80.
16. Vyhnánek T, Nevrtalová E, Bjelková M, Balgová B. SSR loci survey of technical hemp cultivars: The optimization of a cost-effective analyses to study genetic variability. Plant Sci. 2020;298: 110551.
17. García C, Guichoux E, Hampe A. A comparative analysis between SNPs and SSRs to investigate genetic variation in a juniper species (Juniperus phoenicea ssp. turbinata). Tree Genetics & Genomes. 2018;14:87.
18. Hauser SS, Athrey G, Leberg PL. Waste not, want not: Microsatellites remain an economical and informative technology for conservation genetics. Ecol Evol. 2021;11(22):15800–14.
19. Koontz AC, Schumacher EK, Spence ES, Hoban SM. Ex situ conservation of two rare oak species using microsatellite and SNP markers. Evol Appl. 2024;17(3): e13650.
20. Zhang J, Yang J, Lv Y, Zhang X, Xia C, Zhao H, Wen C. Genetic diversity analysis and variety identification using SSR and SNP markers in melon. BMC Plant Biol. 2023;23(1):39.
21. Mauger S, Baud A, Le Corguillé G, Tanguy G, Legeay E, Creis E, Valero M, Potin P, Destombe C. Genetic resources of macroalgae: Development of an efficient method using microsatellite markers in non-model organisms. Algal Res. 2023;75: 103251.
22. Grover A, Aishwarya V, Sharma PC. Searching microsatellites in DNA sequences: approaches used and tools developed. Physiol Mol Biol Plants. 2012;18(1):11–9.
23. Lim KG, Kwoh CK, Hsu LY, Wirawan A. Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. Brief Bioinform. 2013;14(1):67–81.
24. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 2001;11(8):1441–52.
25. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. Bioinformatics. 2017;33(16):2583–5.
26. Alves SIA, Ferreira VBC, Dantas CWD, da Silva ALDC, Ramos RTJ. EasySSR: a user-friendly web application with full command-line features for large-scale batch microsatellite mining and samples comparison. Front Genet. 2023;14:1228552.
27. Mokhtar MM, Alsamman AM, El Allali A. MegaSSR: a web server for large scale microsatellite identification, classification, and marker development. Front Plant Sci. 2023;14:1219055.
28. Kofler R, Schlötterer C, Lelley T. SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics. 2007;23(13):1683–5.
29. Faircloth BC. msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. Mol Ecol Resour. 2008;8(1):92–4.
30. Wang X, Wang L. GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. Front Plant Sci. 2016;7:1350
31. Du L, Li Y, Zhang X, Yue B. MSDB: a user-friendly program for reporting distribution and building databases of microsatellites from genome sequences. J Hered. 2013;104(1):154–7.
32. Du L, Zhang C, Liu Q, Zhang X, Yue B. Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. Bioinformatics. 2018;34(4):681–3.
33. Du L, Liu Q, Fan Z, Tang J, Zhang X, Price M, Yue B, Zhao K. Pyfastx: a robust Python package for fast random access to sequences from plain and gzipped FASTA/Q files. Briefings in Bioinformatics. 2021;22(4):bbaa368.
34. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3—new capabilities and interfaces. Nucleic Acids Res. 2012;40(15): e115.
35. Li H, Rong J. Bedtk: finding interval overlap with implicit interval tree. Bioinformatics. 2021;37(9):1315–6.
36. Li D, Mei H, Shen Y, Su S, Zhang W, Wang J, Zu M, Chen W. ECharts: A declarative framework for rapid construction of web-based visualization. Visual Informatics. 2018;2(2):136–46.
37. Tandem repeats in the long-read sequencing era. Nat Rev Genet. 2024;25(7):449.
38. O'Leary NA, Cox E, Holmes JB, Anderson WR, Falk R, Hem V, Tsuchiya MTN, Schuler GD, Zhang X, Torcivia J, Ketter A, Breen L, Cothran J, Bajwa H, Tinne J, Meric PA, Hlavina W, Schneider VA. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. Sci Data. 2024;11(1):732.
39. Huang J, Li W, Jian Z, Yue B, Yan Y. Genome-wide distribution and organization of microsatellites in six species of birds. Biochem Syst Ecol. 2016;67:95–102.
40. Srivastava S, Avvaru AK, Sowpati DT, Mishra RK. Patterns of microsatellite distribution across eukaryotic genomes. BMC Genomics. 2019;20(1):153.
41. Feng K, Zhou C, Wang L, Zhang C, Yang Z, Hu Z, Yue B, Wu Y. Comprehensive Comparative Analysis Sheds Light on the Patterns of Microsatellite Distribution across Birds Based on the Chromosome-Level Genomes. Animals (Basel). 2023;13(4):655.

Publisher's Note