



## OPEN ACCESS

## EDITED BY

Richard R Rodrigues,  
Oregon State University, United States

## REVIEWED BY

Tiyash Parira,  
Florida International University, United States  
Kalpana Singh,  
Guru Angad Dev Veterinary and Animal  
Sciences University, India  
Maitreya Das,  
Sanofi Pasteur, United States

## \*CORRESPONDENCE

Rommel Thiago Jucá Ramos,  
✉ rommelramos@ufpa.br

RECEIVED 01 August 2024

ACCEPTED 30 October 2024

PUBLISHED 13 November 2024

## CITATION

Alves SIA, Dantas CWD, Macedo DB and  
Ramos RTJ (2024) What are microsatellites and  
how to choose the best tool: a user-friendly  
review of SSR and 74 SSR mining tools.  
*Front. Genet.* 15:1474611.  
doi: 10.3389/fgene.2024.1474611

## COPYRIGHT

© 2024 Alves, Dantas, Macedo and Ramos. This  
is an open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# What are microsatellites and how to choose the best tool: a user-friendly review of SSR and 74 SSR mining tools

Sandy Ingrid Aguiar Alves<sup>1,2,3</sup>, Carlos Willian Dias Dantas<sup>1,2,3</sup>,  
Daralyns Borges Macedo<sup>2,3</sup> and Rommel Thiago Jucá Ramos<sup>2,3\*</sup>

<sup>1</sup>Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, <sup>2</sup>Laboratory of Simulation and Computational Biology – SIMBIC, High Performance Computing Center – CCAD, Federal University of Pará, Belém, Pará, Brazil, <sup>3</sup>Laboratory of Bioinformatics and Genomics of Microorganisms, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil

Microsatellites, also known as SSR or STR, are essential molecular markers in genomic research, playing crucial roles in genetic mapping, population genetics, and evolutionary studies. Their applications range from plant breeding to forensics, highlighting their diverse utility across disciplines. Despite their widespread use, traditional methods for SSR analysis are often laborious and time-consuming, requiring significant resources and expertise. To address these challenges, a variety of computational tools for SSR analysis have been developed, offering faster and more efficient alternatives to traditional methods. However, selecting the most appropriate tool can be daunting due to rapid technological advancements and the sheer number of options available. This study presents a comprehensive review and analysis of 74 SSR tools, aiming to provide researchers with a valuable resource for SSR analysis tool selection. The methodology employed includes thorough literature reviews, detailed tool comparisons, and in-depth analyses of tool functionality. By compiling and analyzing these tools, this study not only advances the field of genomic research but also contributes to the broader scientific community by facilitating informed decision-making in the selection of SSR analysis tools. Researchers seeking to understand SSRs and select the most appropriate tools for their projects will benefit from this comprehensive guide. Overall, this study enhances our understanding of SSR analysis tools, paving the way for more efficient and effective SSR research in various fields of study.

## KEYWORDS

microsatellites, short tandem repeats, simple sequence repeats, SSR tools, molecular markers, computational tools

## 1 Introduction

Microsatellites (MS) are highly polymorphic regions of DNA widely employed in fields such as oncology, forensics, plant breeding, comparative genomics, and microorganism research (Chambers and MacAvoy, 2000; Das et al., 2019). However, initiating studies on these repetitive elements can be challenging due to the fragmented nature of the existing knowledge. This fragmentation may arise from several factors, including the use of different synonymous terms, the variety of available methodologies, and the inherent limitations of

the technique, all of which can pose challenges for researchers in the field (Vieira et al., 2016).

It has been approximately 35 years since the term Microsatellite (MS) was first introduced (Litt and Luty, 1989). Since then, these repetitive sequences have been referred to by diverse names. The most recurrent synonyms used almost interchangeably to MS in the literature and throughout this paper are Simple Sequence Repeats (SSRs) and Short Tandem Repeats (STRs). However, they are also mentioned in studies regarding Variable Number Tandem Repeats (VNTR), Simple Sequence Tandem Repeats (SSTR), Inter Simple Sequence Repeats (ISSR), Simple Sequence Length Polymorphisms (SSLP), and Sequence-Tagged Microsatellites (STMS) (Mudunuri et al., 2010b; Marwal et al., 2014; Jilani and Ali, 2022; Park et al., 2022).

Traditional methodologies for SSR exploration involve genomic fragmentation, microsatellite enrichment, and clone library construction, utilizing techniques such as PCR amplification in biological samples, gel electrophoresis, and Sanger sequencing. These methods have long been the cornerstone of wet laboratory experiments and are still widely employed in fields such as forensic identification, where length polymorphisms of certain STR markers in alleles are analyzed to compare individuals and determine, for example, paternity (Haddrill, 2021). While these techniques have significantly advanced the field, they often face practical challenges, including the need for laboratory infrastructure, equipment, reagents, and specific primers for SSR analysis. These factors contribute to high costs and labor-intensive procedures, particularly in large-scale studies (Thiel et al., 2003; Oliveira et al., 2008; Churbanov et al., 2012; Metz et al., 2016; Das et al., 2019; Guang et al., 2019; Luo et al., 2020).

The rise of Next-Generation Sequencing (NGS) platforms has been progressively shifting the emphasis from studying a few markers towards whole-genome analysis and population genetics (Vieira et al., 2016; Alves et al., 2022). Genome sequencing has become faster and more affordable, enabling the sequencing of hundreds of genomes and transcriptomes of key organisms (Guang et al., 2019), while generating large amounts of publicly available sequence data in databases (Alves et al., 2023). The direct sequence to SSR approach offers a distinct advantage over enrichment-based strategies, as it eliminates the need for prior selection of specific motifs or prior knowledge of the genomic SSR content (Castoe et al., 2012).

*In silico* analysis, which involves exploring microsatellites directly from sequence data through computational techniques, has emerged as a promising approach to manage the significant volume of data and expedite processing while maintaining precision (Oliveira et al., 2008; Vieira et al., 2016). This method is becoming increasingly prevalent for microsatellite discovery and marker development, as it is more efficient and cost-effective (Sharma et al., 2007; Churbanov et al., 2012; Vieira et al., 2016; Umang et al., 2022). A key advantage is that users can freely download genomes from databases such as NCBI and use free computational tools - referred to here as SSR tools - to perform microsatellite identification (also known as SSR mining or prospection), analysis, and even develop new markers, design primers, and simulate primer amplification *in silico* (da Costa Pinheiro et al., 2022). However, it is important to recognize that despite the potential of *in silico* SSR analysis, wet lab methodologies continue to play a crucial role,

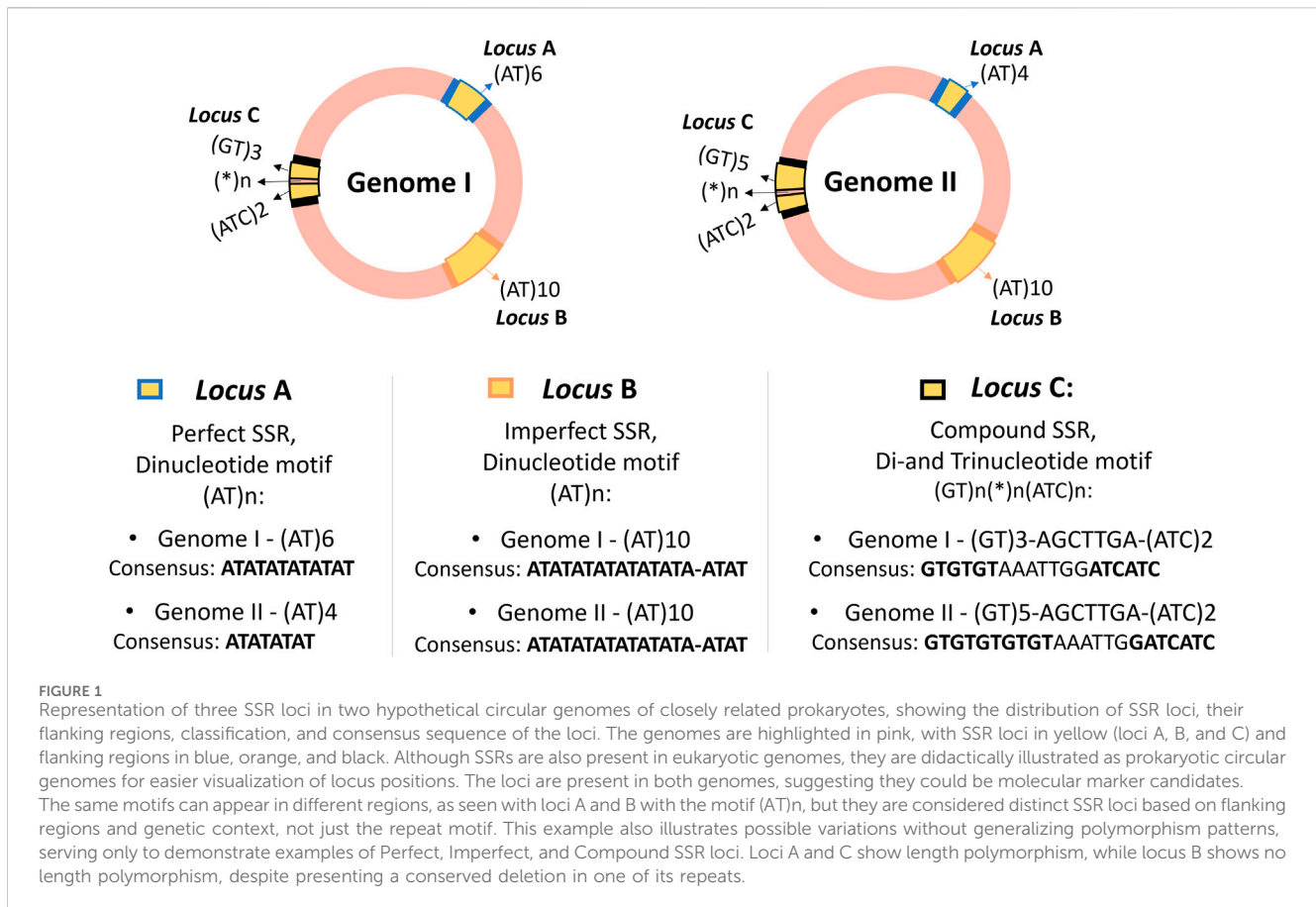
particularly in validating computational predictions and providing critical biological insights (Li et al., 2020).

Regarding the identification of microsatellites through computational methods, numerous tools have been developed over the years, primarily to address gaps identified in existing software, allowing for more sensitive and efficient analysis of these repetitive elements (Pickett et al., 2016). Some older but well-established tools continue to be widely cited in the literature, such as TRF (Benson, 1999) and RepeatMasker (Tarailo-Graovac and Chen, 2009), while new ones are continually emerging, including EasySSR (Alves et al., 2023) and MegaSSR (Mokhtar et al., 2023). The diversity of these tools is evident in various aspects such as execution, input type, and outputs, offering researchers a broad range of options tailored to different datasets and experimental needs (Merkel and Gemmell, 2008; Mathur et al., 2020).

The abundance of tools available for SSR analysis presents a challenge for researchers seeking the most suitable option for their specific needs. This situation often initiates a cycle: a researcher searching for a tool may encounter numerous options but feel uncertain about which to choose. Consequently, many tend to select tools based on their visibility in methodologies or high citation rates. While these tools might be perfectly suitable, some researchers may find them lacking, necessitating adaptations to their work or the development of new tools, often unaware that alternatives with the desired functions may already exist. The fragmentation of the SSR literature further complicates this process, making it challenging to identify these alternative tools. This cycle not only boosts the citation counts of popular tools but also leads to the continual emergence of new tools, many of which are innovative and more efficient, while others may offer redundant functions and performance (Mudunuri et al., 2010b; Lim et al., 2013).

Despite the availability of insightful reviews on microsatellite prediction software over the past decades, the accelerating pace of development in the field leads to rapid information obsolescence (Ledercq et al., 2007; Sharma et al., 2007; Merkel and Gemmell, 2008; Mudunuri et al., 2010b; Lim et al., 2013; Zribi et al., 2016; Mathur et al., 2020). For example, new tools have been introduced, and some of the listed tools are no longer operational, emphasizing the need for ongoing reviews to ensure that the information presented remains current and comprehensive regarding the maximum number of available tools (Mathur et al., 2020).

This paper presents an exhaustive examination of the current state of the art microsatellites mining tools. To provide guidance to users of SSR tools, it comprises two main sections: “Section 1: A guide to what are microsatellites” – the SSR section, and “Section 2: A guide to tools for identifying microsatellites” – the tools section. Advanced readers may focus on the section of interest without detriment to understanding if they skip one of the sections. However, for readers seeking a comprehensive understanding, reading the SSR section is recommended to grasp the main concepts that will aid in understanding what SSRs are and assist in interpreting many of the terms used in the parameters and outputs of the SSR tools. In the tools section, the goal was to gather as many SSR tools as possible, group these tools into subgroups to facilitate analysis, and provide informational tables highlighting various criteria that can influence tool selection. Lastly,



the discussion highlights key factors influencing tool selection, addressing the question, “How to choose the best tool?”. The purpose of this article is not to indicate the best SSR tools but to serve as a guiding resource for users, helping them understand what microsatellites are and assisting them in the conscientious selection of the most suitable tool for their specific research requirements.

## 2 Section 1: a guide to what are microsatellites

Microsatellites are short repetitions in tandem of motifs consisting of 1–6 base pair (bp), that may appear with or without interruptions, distributed across the genomes of all known organisms, including eukaryotes, prokaryotes, and viruses, as well as in some organelles (Chambers and MacAvoy, 2000; Sahu et al., 2020). A single genome can contain thousands of distinct microsatellite loci, as illustrated in Figure 1, where three SSR loci were compared in two circular genomes of prokaryotes. SSR loci are particularly useful in fields where the polymorphism of specific loci can be compared and analyzed to establish close relationships, such as in paternity testing in forensics (Vieira et al., 2016).

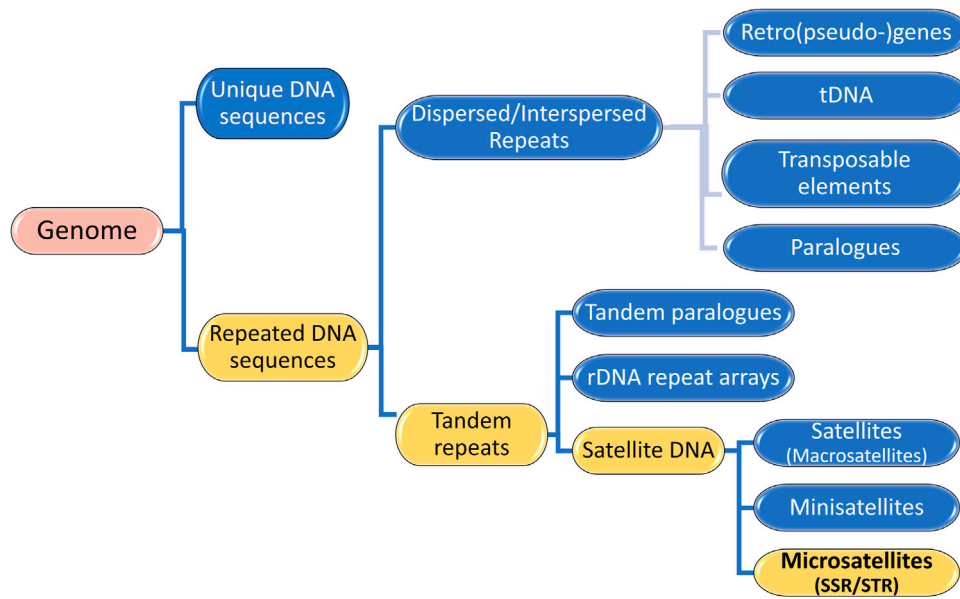
To truly understand what microsatellites are, however, this section aims to go beyond this traditional definition. It is anticipated that Figure 1 may prompt questions in the reader’s mind, as they may not yet be familiar with the specific terminology. To fully comprehend the advanced concepts illustrated in Figure 1, it

is advisable to first understand that SSRs are polymorphic repetitive elements and to grasp their importance and classifications. Readers are encouraged to revisit this figure after reading this section. This deeper understanding will allow not only a clearer interpretation of SSRs but also aid in making sense of the outputs from SSR tools and extracting biological meaning from the computational predictions.

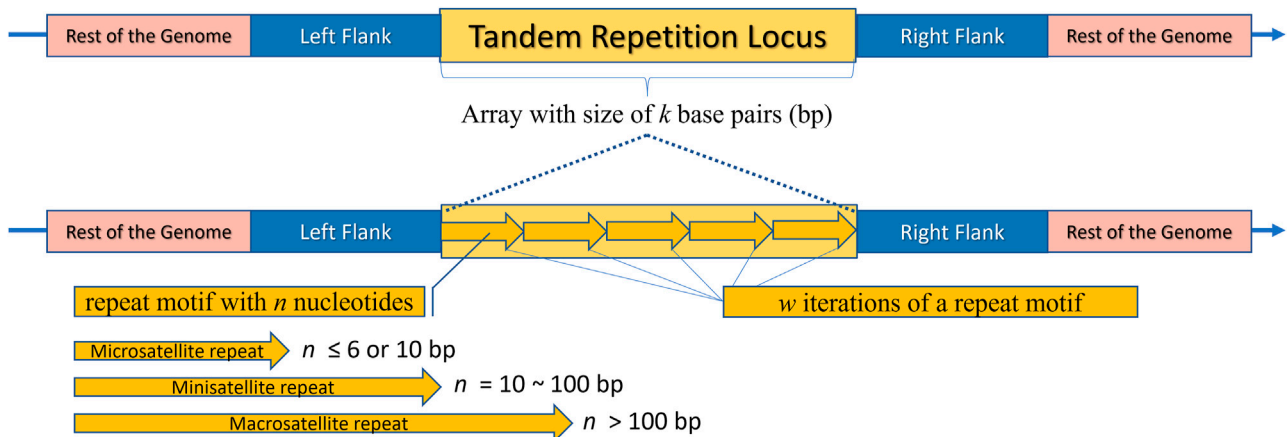
### 2.1 Repetitive elements, tandem repeats and microsatellites

Genomes consist of numerous DNA sequences organized into arrays of different sizes. As depicted in Figure 2, a continuous segment of DNA is called a sequence and is categorized into (i) unique DNA sequences and (ii) repeated DNA sequences, also known as Repetitive Elements or repeats (Richard et al., 2008). Unique segments are non-repetitive, while repetitive elements appear multiple times in the genome (Agarwal and States, 1994; Chambers and MacAvoy, 2000; Richard et al., 2008; da Silva Lopes et al., 2015). Repetitive Elements (RE) can be classified into two groups: (i) Dispersed or Interspersed repeats and (ii) Tandem repeats (Figure 2) (Richard et al., 2008; Lerat, 2010; Girgis and Sheetlin, 2013; Dumbovic et al., 2017; Srivastava et al., 2019).

Tandem repeats (TR) comprise repetitive sequences that occur head-to-tail arrangement and their classification includes (i) gene tandems, (ii) ribosomal DNA (rDNA) repeat arrays, and (iii) satellite DNA (Figure 2) (Richard et al., 2008; Lerat, 2010). A



**FIGURE 2** Genome Composition: Unique DNA sequences and Repeated DNA sequences. The illustration highlights that Microsatellites are Satellite DNA, a subcategory of Tandem Repeats, which are Repetitive Elements in a Genome.



**FIGURE 3** Schematic representation of a Tandem Repeat locus structure. The rest of the genome is highlighted in pink, flanking regions in blue, and the repeat locus itself in yellow. The locus consists of an array with a size of “k” base pairs. The bottom part of the image shows an enlarged view of the locus, where arrows are repeated side by side “w” times, representing the tandemly repeated motifs. It is noteworthy that each arrow represents an illustrative motif, which is a pattern composed of “n” nucleotides. Depending on “n”, the locus can be classified as a Microsatellite, Minisatellite, or Macrosatellite.

tandem repetition locus is an array with size of “k” base pairs (bp), consisting of “w” iterations of a repeat motif with “n” nucleotides, being mutable regions flanked by sequences that are usually conserved (Figure 3) (Alves et al., 2023). Satellite DNA, a type of tandem repeat, can be further classified as (i) satellite or macrosatellite, (ii) minisatellite or (iii) microsatellite (Figure 2) (Lim et al., 2013; Avvaru et al., 2018). As illustrated in Figure 3, this classification is based on the size of the repeated nucleotide pattern, designated as “n” or “motif”. If the repeat motif has large periods, exceeding 100 bp, it is named macrosatellite; those with periods exceeding 10 bp are known as minisatellites, while

microsatellites are short tandem repeats with motifs of “n” ≤ 6 bp (Buschiazzo and Gemmell, 2006; Kelkar et al., 2010; Chen et al., 2011a; Lim et al., 2013; Saeed et al., 2016; Dumbovic et al., 2017; Guang et al., 2019; Alves et al., 2023).

Nevertheless, there is no consensus regarding the classification of microsatellites and minisatellites, which has led to distinct categorizations among researchers (Chambers and MacAvoy, 2000; Lim et al., 2013; Alves et al., 2023). While most papers define SSRs as 1 to 6 base pair repeats, some consider larger motifs up to 10 bp still as SSR. This results in a microsatellite threshold ranging from 6 to 10 bp (Richard et al., 2008; Churbanov



et al., 2012; Korotkov et al., 2021). The terminology surrounding the variable number of tandem repeats (VNTRs) also faces challenges in terms of consensus. Some authors classify these sequences as synonymous with minisatellites, while other researchers categorize them within the broader group of microsatellites and minisatellites (Karaca et al., 2005; Zribi et al., 2016; Lu et al., 2021).

The repetitive nature of these regions often results in sequencing errors, misalignment, and incomplete assemblies, especially when using short-read sequencing. This is a critical limitation, as many pathogenic STR alleles are longer than short reads (Uguen et al., 2024). Frequently, this leads to the absence of these regions from reference genomes or their misplacement within the genomic context. However, despite once being regarded as non-functional or “junk” DNA, numerous repeats have since been identified as important structural or evolutionary markers (Dumbovic et al., 2017). Recent advances in long-read sequencing and alignment tools have enhanced SSR detection by capturing full repeat regions, overcoming some limitations inherent to short-read (Editorial Nature, 2024).

## 2.2 Importance of microsatellites

SSR loci are highly polymorphic, prone to genetic mutations due to errors in DNA replication, recombination, or defective mismatch repair (dMMR), leading to microsatellite instability (MSI) (Yamamoto et al., 2024). This instability can result in the addition or deletion of SSR motifs (Figure 1), leading to length polymorphism and the generation of new inheritable SSR alleles (Jäger, 2022; Mokhtar et al., 2023). An allele with a frequency exceeding 1% within a population is considered polymorphic (Xia et al., 2016; Das et al., 2019; Guang et al., 2019).

Due to their significant role in genetic variation, SSRs have emerged as valuable molecular markers for genetic analysis. Their high variability and polymorphism, along with co-dominant inheritance and non-random distribution in the genome, contribute to their utility (Palliyarakkal et al., 2011). Moreover, the reproducibility of SSRs and the specific design of primers facilitate their amplification (Untergasser et al., 2012), enabling differentiation within and between populations (Marwal et al., 2014; Alves et al., 2023). Researchers explore various aspects of MS, including their incidence, frequency, prevalence, abundance, distribution, polymorphisms, composition, information content, localization, transferability, and associations with other sequence elements (Thiel et al., 2003; Sharma et al., 2007; Jilani and Ali, 2022).

SSRs widespread presence allows for comprehensive studies of DNA, transcribed sequences, and their corresponding proteins, as they can be found in both coding and non-coding regions, and are identifiable in various contexts, including sequenced DNA, assembled genomes, genes, and expressed sequence tags (ESTs) (Thiel et al., 2003; Karaca et al., 2005; Mathur et al., 2020).

In addition, MS have diverse applications across various fields. They are associated with over 30 human genetic diseases (Marwal et al., 2014; Lu et al., 2021) and very important in oncology (Baudrin et al., 2018). For instance, Indels in coding microsatellites (cMS) within tumor suppressor genes like *TGFBR2* and *ACVR2* act as key drivers of cancer progression in mismatch repair-deficient (MMRd)

cells, generating immunogenic frameshift peptide (FSP) neoantigens. Darwinian selection favors cMS mutations that enhance cell survival and tumor growth, resulting in their accumulation. Thus, MMRd cancers are immunogenic not only due to a high number of somatic mutations but also the abundance of FSP-derived epitopes generated by these indels (Hernandez-Sanchez et al., 2022).

In forensics, STRs can be used for identification and parentage determination, as they form a “genetic fingerprint” for each individual (Haddrill, 2021; Jäger, 2022). By analyzing the distribution of specific STR alleles across populations, researchers can uncover group relationships and trace migration patterns, providing insights into human evolutionary history (Editorial Nature, 2024). Moreover, MS have been linked to influencing virulence in pathogens (Reneker et al., 2004) and can serve as biomarkers in fungi (Sokolova et al., 2022), protozoa (Durigan et al., 2018), bacteria (da Costa Pinheiro et al., 2022), and viruses (Laskar et al., 2022). They can also be applied in diagnostics, as exemplified by the investigation of leprosy transmission utilizing microsatellite typing through amplification of compound SSR loci (Mohanty et al., 2019).

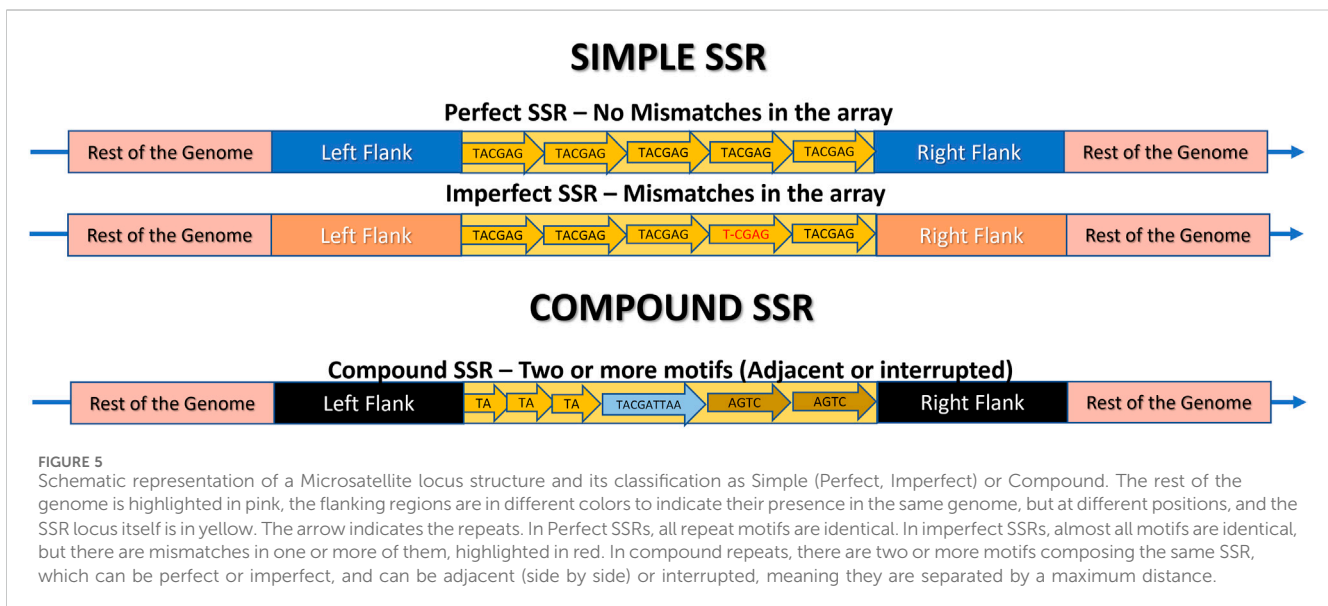
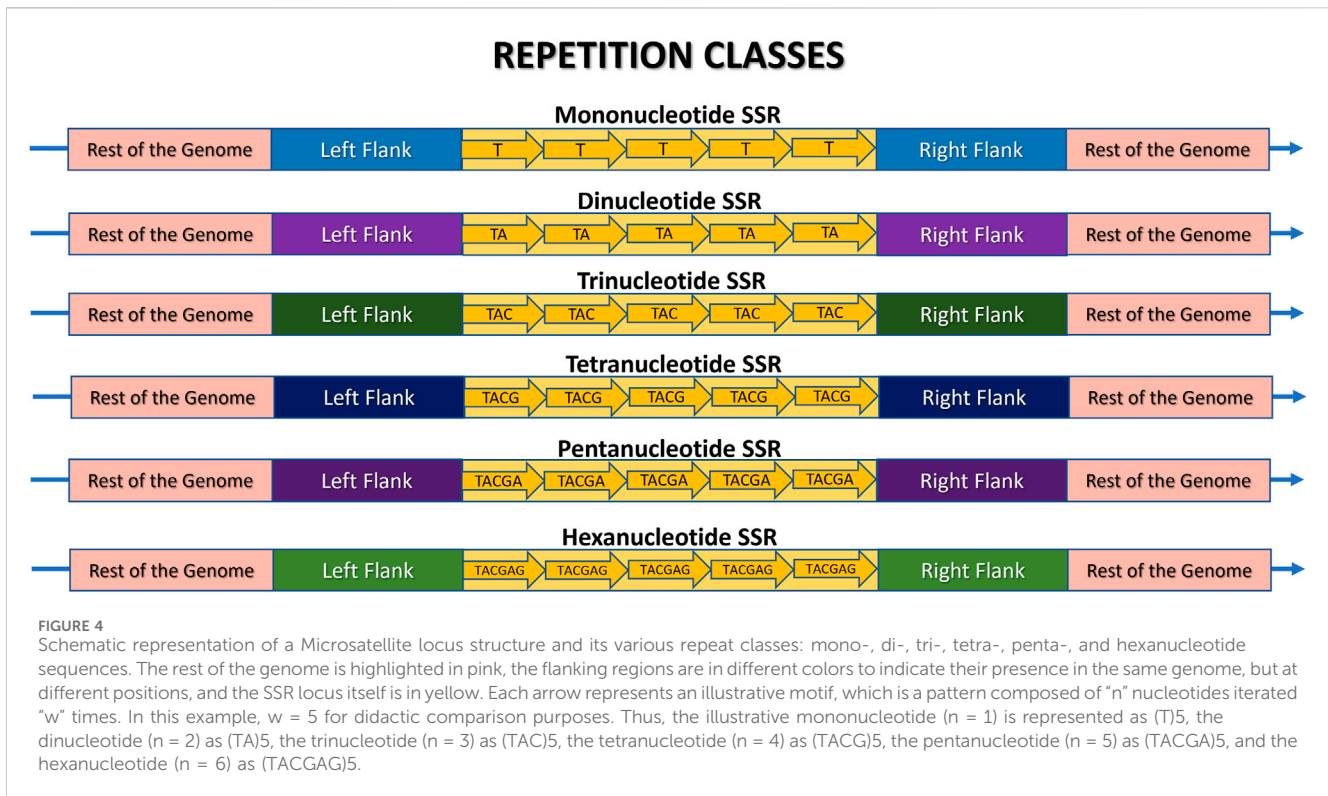
Furthermore, SSRs are widely applied in plant research and breeding, providing insights into genetic diversity, population structure, and evolutionary patterns (Biswas et al., 2018). They aid in crop improvement by identifying alleles linked to desirable traits (Oliveira et al., 2008), and can play a crucial role in conservation biology by assessing genetic diversity in endangered plants (Yuan et al., 2018). Additionally, SSRs help trace the evolutionary history of plant species, providing insights into their adaptation and divergence (Morgante et al., 2002; Oliveira et al., 2008; Yuan et al., 2018).

## 2.3 Classifications of microsatellites

In the literature it is usual to see SSR categorized based on various parameters (Figure 1). Common classifications include those based on repeat classes, perfection level, and tract composition, as illustrated in Figures 4, 5. Additional sub-classifications may be applied depending on specific research objectives (Mudunuri and Nagarajaram, 2007; Chen et al., 2011b; Ledenyova et al., 2019; da Costa Pinheiro et al., 2022).

### 2.3.1 Based on repetition classes

Microsatellites can be classified into repetition classes based on the size “n” of their repeated nucleotide pattern, which is commonly referred to as the motif (Figure 3). As illustrated in Figure 4, the repetition classes of SSRs include: (i) Mononucleotide, which consists of “w” repetitions of a single nucleotide ( $n = 1$ ), for example, (T)<sub>5</sub>; (ii) Dinucleotide, composed of “w” repetitions of a pair of nucleotides ( $n = 2$ ), such as (TA)<sub>9</sub>; (iii) Trinucleotide, characterized by “w” repetitions of three nucleotides ( $n = 3$ ), such as (TAC)<sub>12</sub>; (iv) Tetranucleotide, consisting of “w” repetitions of four nucleotides ( $n = 4$ ), e.g., (TACG)<sub>8</sub>; (v) Pentanucleotide, formed by “w” repetitions of five nucleotides ( $n = 5$ ), for instance (TACGA)<sub>4</sub>; and (vi) Hexanucleotide, consisting of “w” repetitions of six nucleotides ( $n = 6$ ), such as (TACGAG)<sub>5</sub> (Sharma et al., 2007; Sahu et al., 2020).



Usually, SSR tools permit users to define the minimum number of repetitions (*w*) that must occur for a motif (*n*) to be considered an SSR. In instances where the tool identifies a repetition of the motif with fewer than the specified repetitions, that SSR will not be included in the final output (Alves et al., 2023). It is worth noting that there is no standard value for this parameter, as it can vary depending on the specific organism under study. Therefore, it is advisable for users to consult relevant literature on the organism in question to identify the most suitable parameters for their study (Oliveira et al., 2008).

### 2.3.2 Based on perfection level

As SSRs are prone to suffer mutations, they can be classified as (i) Perfect SSR (p-SSR or pSSR) or (ii) Imperfect SSR (i-SSR or iSSR) (Figure 5) (Mokhtar and Atia, 2019; Song et al., 2021).

Perfect microsatellite tracts, also referred as Pure or Exact, consist of motifs that are replicated multiple times with precise replication of the exact pattern, such as (AT)<sub>20</sub> and the examples in Figure 4, without any deviations. However, as mutational events occur, mismatches can arise, leading to disruptions in the p-SSR through base substitutions or nucleotide insertions or deletions

(INDELs) (Figure 5). Consequently, the p-SSR is transformed into an imperfect SSR when the copies deviate by at least one base pair. The i-SSR may also be referred to as Approximate or Interrupted (Sharma et al., 2007; Lim et al., 2013; da Costa Pinheiro et al., 2022). For example, if the p-SSR (AT)<sub>20</sub> was interrupted due to an insertion of a “G” it would be classified as an i-SSR (AT)<sub>12</sub>G (AT)<sub>8</sub> by some SSR mining tools (Alves et al., 2023).

Most research in this field has concentrated on perfect repeats, given their association with selective forces and higher length polymorphism. Consequently, the prevalence of SSR tools with algorithms that can only identify perfect SSR (Mudunuri and Nagarajaram, 2007; Behura and Severson, 2015; Ledenyova et al., 2019). In contrast, imperfect SSRs are more stable and less susceptible to slippage mutations, resulting in less length polymorphism but featuring INDELs, which can be valuable for studying single nucleotide polymorphisms (SNPs) (Metz et al., 2016; Guang et al., 2019). It is worth noting that when mismatches are allowed in SSR tools, some SSRs previously identified as p-SSR may be elongated and reclassified as i-SSR arrays (Alves et al., 2023).

### 2.3.3 Based on tract composition

Microsatellite tracts may exhibit two distinct compositions. They may either comprise a single motif or a combination of motifs. In this way they can be categorized as (i) Simple SSRs or (ii) Compound SSRs (c-SSRs, cSSRs) (Figure 5) (George et al., 2015; Sahu et al., 2020; Song et al., 2021).

Simple SSRs comprise p-SSR and i-SSR and consist of loci with a unique motif repeated in tandem, such as (TA)<sub>7</sub>. Compound microsatellites, also referred as Fuzzy or Interrupted by authors, arise from mutations or imperfections in SSR (George et al., 2015). They are loci composed of two or more simple SSR motifs, that might be adjacent or interrupted, separated by a nucleotide sequence (Chen et al., 2011b; Ledenyova et al., 2019). For instance, in Figure 5, “X” represents the sequence interrupting the two parts of the SSR in (TA)<sub>3</sub>-X-(AGTC)<sub>2</sub>. In Figure 1 simple and compound SSR loci are illustrated and compared in two hypothetical circular genomes.

In the context of SSR mining tools, whether two simple microsatellites are regarded as a c-SSR is contingent upon the distance between motifs resulting from the interrupting sequence. For instance, two SSRs separated by distances falling within a specified range (dMAX) may be regarded as a single c-SSR tract. However, depending on the dMAX set by the user, it is possible that some C-SSRs could be considered as two distinct SSRs (Alam et al., 2019).

### 2.3.4 Based on genomic context

Microsatellites can be classified as either (i) Coding SSRs, if they are fully or partially situated within coding regions, or (ii) Non-coding SSRs, if they are located in regions that do not encode proteins (Mudunuri and Nagarajaram, 2007; da Costa Pinheiro et al., 2022; Alves et al., 2023).

The presence of SSRs in coding regions results in the emergence of these repetitive patterns in transcribed sequences and in their proteins, reflecting potential associations with genes and phenotypes (Vieira et al., 2016).

In general, the most SSRs are present at intergenic and non-coding regions, and less frequent in exons and genic regions (Srivastava et al., 2019). This is primarily attributed to the high

mutation rate of microsatellites, which could potentially disrupt gene expression (Vieira et al., 2016). Most coding regions are composed of SSRs with tri- and hexanucleotide motifs. This is likely due to the selective pressure against mutations that could alter the reading frame (Li et al., 2002; Vieira et al., 2016). The length variations of SSRs within exons have been associated with various diseases, including Huntington’s and Spinocerebellar Ataxia (Srivastava et al., 2019).

Some SSR mining tools can ascertain whether an SSR resides in a coding region, although this requires the input of an additional file by the user, in which the regions that are and are not genetic should be indicated (Alves et al., 2023).

### 2.3.5 Based on mutability and array length

Regarding their mutability, microsatellites can be categorized as: (i) Hypermutable SSRs, which consist of multiple repeat units that exhibit high rates of INDELs; (ii) Mutable SSRs, which have intermediate-length repeat tracts and therefore lower mutation rates; and (iii) Proto-mutable SSRs, which consist of a small number of repeat units and exhibit mutation rates slightly higher than the average for the genome (Bidmos and Bayliss, 2014).

There is an additional classification that considers mutability and array length, designated “k” This classification is as follows: (i) Class I, hypervariable markers, with arrays exceeding 20 base pairs (bp); (ii) Class II, potentially variable markers, with arrays from 12 to 20 bp; and (iii) Class III, SSRs, less variable markers, characterized by smaller arrays with less than 11 bp (Temnykh et al., 2001; Saeed et al., 2016).

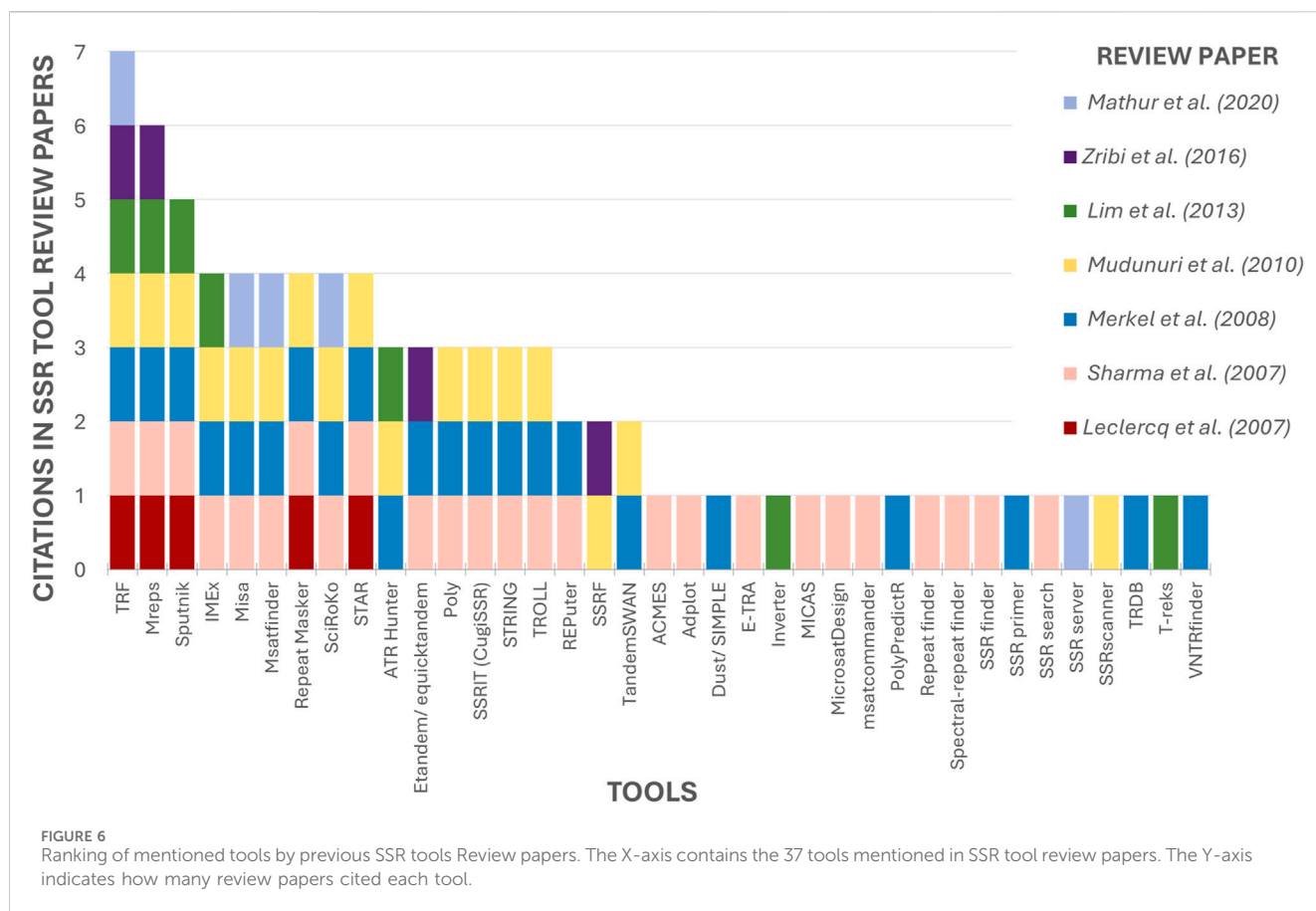
Although the classifications in question might not be commonly employed in a general context, they do exist in certain tools designed to study polymorphic SSRs (PolySSRs). Consequently, this classification should be carefully considered, particularly by those engaged in the development of polymorphic SSR markers (Xia et al., 2016).

At this point, the reader has been introduced to the fundamental concepts and significance of microsatellites, which are crucial for understanding their applications in various fields. With this foundational knowledge, Figure 1 should now be more comprehensible, offering clarity on the discussed principles.

## 3 Section 2: a guide to tools for identifying microsatellites

The standard procedure for *in silico* SSR analysis involves obtaining the DNA data to be analyzed and using it as input for an SSR tool. These tools typically identify SSRs and their positions in the sequence, enabling various analyses. However, the abundance of available tools may challenge users in identifying the most suitable one for their *in silico* studies (Alves et al., 2023). Considering that the reader now possesses understanding of the key concepts of MS, this section aims to provide a comprehensive analysis of the tools utilized for tandem repeat prospecting, focusing on microsatellite mining software, and provide guidance on selecting the most appropriate tool for their specific research needs.

To achieve this, a literature review was conducted on Pubmed with the terms “Review” and “SSR” or “STR” or “Microsatellites”, and after careful evaluation were included in the study all review



papers that focused on tools for SSR mining. To find more tools, this was complemented by a comprehensive survey of the broader SSR literature, with particular attention to references made in papers that released new softwares. These articles often compare the newly launched tools with existing ones, enabling the identification of most tools released to date. A total of 74 tools were identified, with every SSR tool found included in the analysis. If a tool is not listed, it was likely not discovered by the authors at the time this paper was written. The citation index for each tool's publication was obtained from Google Scholar, and citations were compared to determine which tools were most widely adopted by the scientific community. Their availability was assessed, and an in-depth analysis was conducted for each functional tool, including parameters, inputs, outputs, and other relevant aspects, which are discussed in the following sections.

### 3.1 Overview of all tools identified in this paper

#### 3.1.1 Previous reviews of microsatellite search tools

Seven reviews regarding SSR tools have been retrieved (Leclercq et al., 2007; Sharma et al., 2007; Merkel and Gemmell, 2008; Mudunuri et al., 2010b; Lim et al., 2013; Zribi et al., 2016; Mathur et al., 2020). The maximum number of tools mentioned in a single article was 25 (Sharma et al., 2007), and by combining the data from the authors, a total of 37 tools were identified, some of

which were cited by most of the papers (Figure 6). The Tandem Repeats Finder (TRF) (Benson, 1999) stands out as the most frequently cited tool, having been cited by all the review papers. Second is Mreps (Kolpakov et al., 2003) mentioned in 06 papers, followed by Sputnik (La Rota et al., 2005), cited by 05 authors. Fourth on the list are IMEx (Mudunuri and Nagarajaram, 2007), Misa (Thiel et al., 2003) and STAR (Delgrange and Rivals, 2004), all of which are cited by the majority (04 out of 07 reviews). The remaining tools are mentioned in Tables 1, 2.

Only 18 tools, approximately 49% of the total, were cited by more than one author (Figure 6), indicating that despite the numerous tools that exist, previous review articles have been limited to examining only a few of them. This shows that the current state of knowledge is scattered and that a comprehensive review such as the present study is needed to collect and analyze the maximum possible number of tools.

In addition to the 37 tools referenced in previous reviews, an exhaustive examination of microsatellite literature and an analysis of references cited in other published tools identified an additional 37 tools, resulting in a total of 74 tools that will be examined in the following sections. These tools were identified and summarized in Tables 1, 2, with further details provided in the subsequent sections.

#### 3.1.2 Analysis of availability of the tools

The first aspect that was considered was the tool's availability for present-day usage by users. Given the dynamic nature of the field, there is a risk that individuals relying on previous review articles to



TABLE 1 Total tools identified part 1–42 tools available in 2024.

Name	Cited in reviews	Availability in 2024	Reference/Year	Citations	Tool's link
BatchPrimer3	No	Yes	You 2008	961	<a href="http://wheat.pw.usda.gov/demos/BatchPrimer3/">http://wheat.pw.usda.gov/demos/BatchPrimer3/</a>
CandiSSR	No	Yes	Xia 2016	56	<a href="https://github.com/xiaenhua/CandiSSR">https://github.com/xiaenhua/CandiSSR</a>
Dot2Dot	No	Yes	Genovese 2019	15	<a href="https://github.com/Gege7177/Dot2dot">https://github.com/Gege7177/Dot2dot</a>
EasySSR	No	Yes	Alves 2023	1	<a href="https://computationalbiology.ufpa.br/easyssr/">https://computationalbiology.ufpa.br/easyssr/</a>
Etandem/Equicktandem	Yes	Yes	Rice 2000	N/A	<a href="https://www.bioinformatics.nl/cgi-bin/emboss/help/equicktandem">https://www.bioinformatics.nl/cgi-bin/emboss/help/equicktandem</a>
FullSSR	No	Yes	Metz 2016	18	<a href="https://sourceforge.net/projects/fullsr/">https://sourceforge.net/projects/fullsr/</a>
GMATa	No	Yes	Wang 2016	211	<a href="https://sourceforge.net/projects/gmata/?source=navbar">https://sourceforge.net/projects/gmata/?source=navbar</a>
GMATo	No	Yes	Wang 2013	119	<a href="https://sourceforge.net/projects/gmato/files/?source=navbar">https://sourceforge.net/projects/gmato/files/?source=navbar</a>
IDSSR	No	Yes	Guang 2019	15	<a href="https://github.com/Allsummerking/IDSSR">https://github.com/Allsummerking/IDSSR</a>
IMEx	Yes	Yes	Mudunuri 2007	231	<a href="http://www.mcr.org.in/IMEX/download.html">http://www.mcr.org.in/IMEX/download.html</a>
Imperfect SSR Finder	No	Yes	Stieneke 2007	10	<a href="https://ssr.nwisrl.ars.usda.gov/">https://ssr.nwisrl.ars.usda.gov/</a>
Kmer-SSR	No	Yes	Pickett 2017	25	<a href="https://github.com/ridgelab/Kmer-SSR">https://github.com/ridgelab/Kmer-SSR</a>
Krait	No	Yes	Du 2018	132	<a href="https://github.com/lmdu/krait">https://github.com/lmdu/krait</a>
MegaSSR	No	Yes	Mokhtar 2023	1	<a href="https://bioinformatics.um6p.ma/MegaSSR">https://bioinformatics.um6p.ma/MegaSSR</a>
MICAS	Yes	Yes	Sreenu 2003	27	<a href="http://www.mcr.org.in/micas/">http://www.mcr.org.in/micas/</a>
Micro-Primers	No	Yes	Alves 2022	2	<a href="https://github.com/FilAlves/micro-primers">https://github.com/FilAlves/micro-primers</a>
Microsatellite repeats finder	No	Yes	Bikandi 2006	5	<a href="http://insilico.ehu.es/mini_tools/microsatellites/?info">http://insilico.ehu.es/mini_tools/microsatellites/?info</a>
MiMi	No	Yes	Fox 2019	16	<a href="https://github.com/graemefox/mimi">https://github.com/graemefox/mimi</a>
MISA	Yes	Yes	Thiel 2003	2602	<a href="https://webblast.ipk-gatersleben.de/misa/">https://webblast.ipk-gatersleben.de/misa/</a>
Mreps	Yes	Yes	Kolpakov 2003	477	<a href="https://mreps.univ-mlv.fr/">https://mreps.univ-mlv.fr/</a>
msatcommander	Yes	Yes	Faircloth 2008	1071	<a href="https://code.google.com/archive/p/msatcommander/">https://code.google.com/archive/p/msatcommander/</a>
PALfinder	No	Yes	Castoe 2012	300	<a href="https://sourceforge.net/projects/palfinder/">https://sourceforge.net/projects/palfinder/</a>
PERF	No	Yes	Avvaru 2018	41	<a href="https://github.com/rkmlab/perf">https://github.com/rkmlab/perf</a>
PolyMorphPredict	No	Yes	Das 2019	17	<a href="http://webtom.cabgrid.res.in/polypred/">http://webtom.cabgrid.res.in/polypred/</a>
PolySSR	No	Yes	Tang 2008	102	<a href="http://www.bioinformatics.nl/tools/polyssr/">http://www.bioinformatics.nl/tools/polyssr/</a>
RepeatMasker	Yes	Yes	Smit 1996 apud Tarailo-Graovac 2009	2443	<a href="https://www.repeatmasker.org/cgi-bin/WEBRepeatMasker">https://www.repeatmasker.org/cgi-bin/WEBRepeatMasker</a>
REPuter	Yes	Yes	Kurtz 2001	1859	<a href="http://bibiserv.techfak.uni-bielefeld.de/reputer/">http://bibiserv.techfak.uni-bielefeld.de/reputer/</a>
SA-SSR	No	Yes	Pickett 2016	15	<a href="https://github.com/ridgelab/SA-SSR">https://github.com/ridgelab/SA-SSR</a>
SciRoKo	Yes	Yes	Kofler 2007	399	<a href="https://kofler.or.at/bioinformatics/SciRoKo/index.html">https://kofler.or.at/bioinformatics/SciRoKo/index.html</a>
Spectral-repeat finder	Yes	Yes	Sharma 2004	196	<a href="http://www.imtech.res.in/raghava/srf">http://www.imtech.res.in/raghava/srf</a>
Sputnik	Yes	Yes	Morgante 2002	1465	<a href="http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/">http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/</a>
SSR_pipeline	No	Yes	Miller 2013	57	<a href="https://pubs.usgs.gov/ds/778/">https://pubs.usgs.gov/ds/778/</a>
SSR2Marker	No	Yes	Yue 2022	1	<a href="https://github.com/aaranyue/SSR2Marker">https://github.com/aaranyue/SSR2Marker</a>

(Continued on following page)

TABLE 1 (Continued) Total tools identified part 1–42 tools available in 2024.

Name	Cited in reviews	Availability in 2024	Reference/Year	Citations	Tool's link
SSRenricher	No	Yes	Luo 2020	3	<a href="https://github.com/byemaxx/SSREnricher">https://github.com/byemaxx/SSREnricher</a>
SSRIT (CugiSSR)	Yes	Yes	Temnykh 2001	2056	<a href="https://archive.gramene.org/db/markers/ssrtool">https://archive.gramene.org/db/markers/ssrtool</a>
SSRMMD	No	Yes	Gou 2020	25	<a href="https://github.com/GouXiangJian/SSRMMD">https://github.com/GouXiangJian/SSRMMD</a>
SSRpoly	No	Yes	Duran 2013	8	<a href="https://appliedbioinformatics.com.au/Edwards/index.php/SSRPoly">https://appliedbioinformatics.com.au/Edwards/index.php/SSRPoly</a>
STAR	Yes	Yes	Delgrange 2004	142	<a href="http://atgc.lirmm.fr/star/">http://atgc.lirmm.fr/star/</a>
T-reks	Yes	Yes	Jorda 2009	196	<a href="https://bioinfo.crbm.cnrs.fr/index.php?route=tools&amp;tool=3">https://bioinfo.crbm.cnrs.fr/index.php?route=tools&amp;tool=3</a>
TRF	Yes	Yes	Benson 1999	7903	<a href="https://tandem.bu.edu/trf/trf.html">https://tandem.bu.edu/trf/trf.html</a>
TROLL	Yes	Yes	Castelo 2002	265	<a href="https://sourceforge.net/projects/finder/">https://sourceforge.net/projects/finder/</a>
Websat (WebTROLL)	No	Yes	Martins 2009	365	<a href="https://bioinfo.inf.ufg.br/websat/">https://bioinfo.inf.ufg.br/websat/</a>

"Yes" has been highlighted in the columns for visualization purposes.

guide their tool selection for methodologies might find their chosen tool non-operational. All 74 tools were assessed regarding their accessibility status. Multiple attempts were made to access the tools between 2023 and 2024 using the links provided in the papers. Figure 7A illustrates that only half of the total 74 tools had previously been cited in reviews. Of these 37 tools referenced in the review articles, only 16 are still available. This implies that 57% of the previously reviewed tools are no longer functional. Conversely, 70% of the tools identified in this article alone are available. Overall, out of the total of 74 tools initially identified, 42 remained functional and were summarized in Table 1, while the 32 tools not accessible at the time of analysis were labeled as not available and grouped in Table 2.

### 3.1.3 Analysis of temporality of tool's release and citations

The historical development of microsatellite analysis can be traced through the papers of each tool (Tables 1, 2). These indicate a temporal range of release years between 1994 and 2023, as evidenced in Figure 7B. This suggests a dynamic evolution within the field, with the introduction of innovative tools and methodologies occurring continuously (Alves et al., 2022). Prior to 2005, the prevailing trend in the field was the development of basic tools that introduced new algorithms or approaches, performed specific functions, or served as a basis for integration into pipelines. From 2006 onwards, the focus has shifted to the release of pipeline tools, demonstrating a pattern where modern tools build upon established frameworks and incorporate additional features, to enhance analytical capabilities for various objectives.

The continuous release of new tools prompted an investigation into the relative prominence of these tools compared to older ones. The investigation also sought to determine whether the high citation counts for a tool could be attributed to the fact that it was mentioned in review articles. To test this hypothesis, the citation counts for each tool's original papers were assessed using Google Scholar in early 2024 (Tables 1, 2).

A discrepancy was observed between the order of the tools most frequently mentioned in review articles (Figure 6) and their position

in the citation ranking depicted in Figure 8. Nevertheless, a considerable number of the most frequently cited tools were referenced in review articles. At the time of analysis, TRF (Benson, 1999) remained the most frequently cited tool, with a substantial margin of citations over the second-ranked tool. Furthermore, some tools, despite being less frequently mentioned in review articles such as msatcommander (Faircloth, 2007), have received considerable citation counts. Conversely, the emergence of previously unmentioned tools with considerable citation indices, such as BatchPrimer3 (You et al., 2008) and GMATa (Wang and Wang, 2016), underscores the dynamic nature of the field.

Figure 8 also reveals that the top 20 most cited tools have remained in use for a minimum of a decade, indicating a concentration of citations in long-established tools over time. Conversely, despite their versatile features, more contemporary tools released in the last 5 years, such as EasySSR (Alves et al., 2023), MegaSSR (Mokhtar et al., 2023), Micro-primers (Alves et al., 2022), and SSR2Marker (Yue and Liu, 2022), appear not to have yet received substantial attention from the scientific community.

Although the tools on the top concentrate an impressive number of citations (Figure 8), this does not necessarily indicate that they are superior to others in the middle to bottom part of the citation ranking. If this were the case, there would be no incentive to develop new tools to address the limitations of existing ones (Sharma et al., 2007). Users often have specific expectations for tools that align with their research goals. While a suitable tool with the necessary features may exist, users may be unaware of it and therefore not use or cite it. Instead researchers may encounter difficulties in adapting their project to utilize a popular tool (Alves et al., 2023). Thus, the subsequent sections aim to disseminate knowledge regarding the tools, irrespective of the citation rate.

## 3.2 Comparison of the available tools

In the previous subsection, the availability status of every tool was identified, providing a clear understanding of which ones are operational. To help users concentrate on viable options, this section

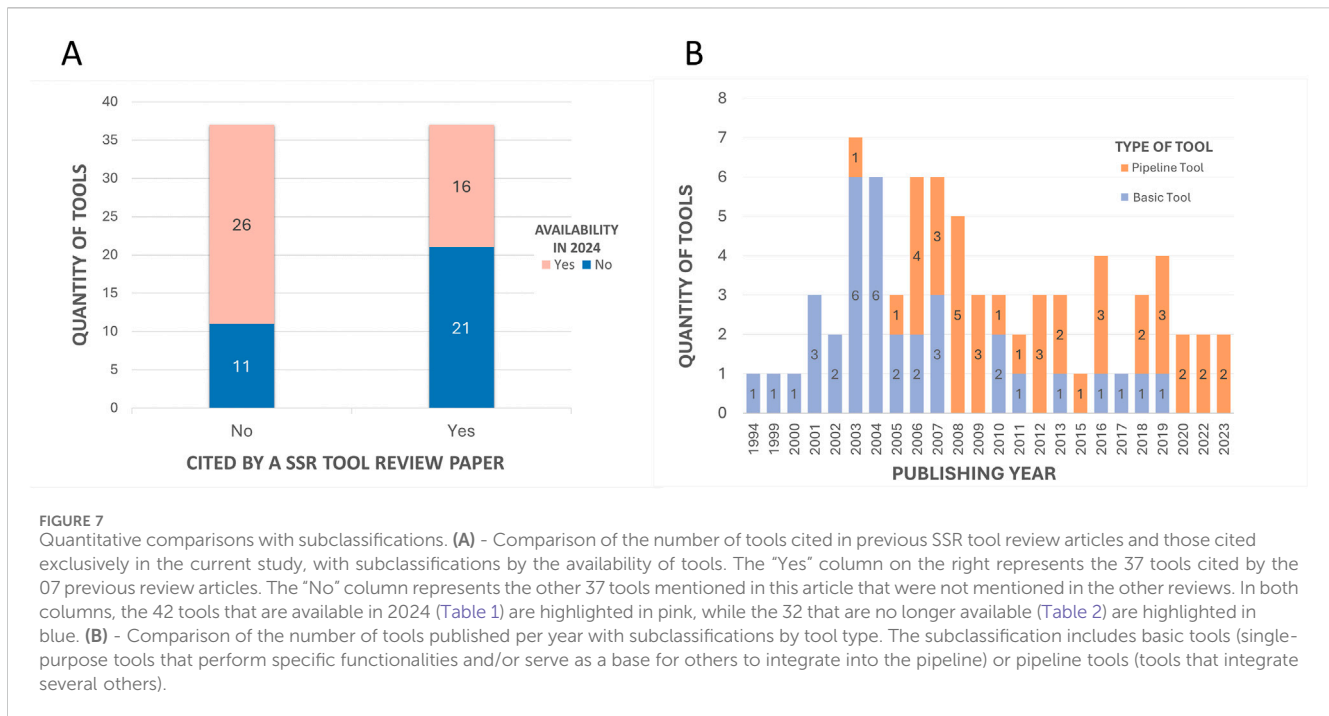
TABLE 2 Total tools identified part 2–32 tools not available in 2024.

Name	Cited in reviews	Availability in 2024	Reference/Year	Citations	Tool's link
ACMES	Yes	No	Reneker 2004	16	<a href="http://acmes.rnet.missouri.edu/">http://acmes.rnet.missouri.edu/</a>
Adplot	Yes	No	Taneda 2004	25	Email
ATR Hunter	Yes	No	Wexler 2004	141	<a href="http://www.bioinfo.cs.technion.ac.il/ATRHunter">http://www.bioinfo.cs.technion.ac.il/ATRHunter</a>
Dust/SIMPLE	Yes	No	Hancock 1994	170	<a href="http://life.anu.edu.au/">http://life.anu.edu.au/</a>
E-TRA	Yes	No	Karaca 2005	51	<a href="ftp.akdeniz.edu.tr/Araclar/e-TRA">ftp.akdeniz.edu.tr/Araclar/e-TRA</a>
Inverter	Yes	No	Wirawan 2010	14	<a href="http://bmsserver.sce.ntu.edu.sg/INVERTER">http://bmsserver.sce.ntu.edu.sg/INVERTER</a>
LSAT	No	No	Biswas 2018	3	<a href="http://210.110.86.160/Lsat/Lsat.html">http://210.110.86.160/Lsat/Lsat.html</a>
MfSAT	No	No	Chen 2011	3	<a href="http://hudacm11.mysinamail.com/hunan.html">http://hudacm11.mysinamail.com/hunan.html</a>
MicrosatDesign	Yes	No	Singan 2005	4	<a href="http://daphnia.cgb.indiana.edu/wfleafbase/software">http://daphnia.cgb.indiana.edu/wfleafbase/software</a>
Msatfinder	Yes	No	Thurston 2005	123	<a href="http://www.genomics.ceh.ac.uk/msatfinder/">http://www.genomics.ceh.ac.uk/msatfinder/</a>
Poly	Yes	No	Bizzaro 2003	36	<a href="http://bioinformatics.org/poly/wiki/">http://bioinformatics.org/poly/wiki/</a>
PolyPredictR	Yes	No	Odushlaine 2006	16	<a href="http://bioinformatics.rcsi.ie/vntrfinder/">http://bioinformatics.rcsi.ie/vntrfinder/</a>
ProGeRF	No	No	Lopes 2015	11	<a href="http://64.79.105.19/lipg/">http://64.79.105.19/lipg/</a>
QDD	No	No	Meglecz 2010	694	<a href="http://www.univ-provence.fr/gsite/Local/egee/dir/meglecz/QDD.html">http://www.univ-provence.fr/gsite/Local/egee/dir/meglecz/QDD.html</a>
RepeatFinder	Yes	No	Volfovsky 2001	203	<a href="http://www.tigr.org/softlab/">http://www.tigr.org/softlab/</a>
ReRep	No	No	Otto 2008	15	<a href="http://bioinfo.pdtis.fiocruz.br/ReRep/">http://bioinfo.pdtis.fiocruz.br/ReRep/</a>
Risa	No	No	Kim 2012	5	<a href="http://sol.kribb.re.kr/RISA/">http://sol.kribb.re.kr/RISA/</a>
SAT	No	No	Dereeper 2007	40	<a href="http://sat.cirad.fr/sat">http://sat.cirad.fr/sat</a>
SSRF	Yes	No	Sreenu 2003	27	Email
SSRfinder	Yes	No	Gao 2003	388	<a href="https://www.fresnostate.edu/ssrfinder/">https://www.fresnostate.edu/ssrfinder/</a>
SSRlocator	No	No	Da Maia 2008	276	<a href="http://microsatellite.org/ssr.php">http://microsatellite.org/ssr.php</a>
SSRprimer	Yes	No	Jewell 2006	105	<a href="http://bioinformatics.pcbasc.la.trobre.edu.au/ssrdiscovery.html">http://bioinformatics.pcbasc.la.trobre.edu.au/ssrdiscovery.html</a>
SSR-Primer Generator	No	No	Hong 2011	1	<a href="http://168.188.15.158:8080/ssrpg/">http://168.188.15.158:8080/ssrpg/</a>
SSRscanner	Yes	No	Anwar 2006	5	Email
SSRsearch	Yes	No	Nicot 2004	286	<a href="ftp://ftp.gramene.org/pub/gramene/software/scripts/ssr.pl">ftp://ftp.gramene.org/pub/gramene/software/scripts/ssr.pl</a>
SSRserver	Yes	No	Jung 2007	267	<a href="https://www.rosaceae.org/node/55">https://www.rosaceae.org/node/55</a>
STRING	Yes	No	Parisi 2003	79	<a href="http://www.caspur.it/~castr/STRING/">http://www.caspur.it/~castr/STRING/</a>
TandemSWAN	Yes	No	Boeva 2006	115	<a href="http://bioinform.genetika.ru/projects/swan/www/">http://bioinform.genetika.ru/projects/swan/www/</a>
TRDB	Yes	No	Gelfand 2007	218	<a href="https://tandem.bu.edu/cgi-bin/trdb/trdb.exe">https://tandem.bu.edu/cgi-bin/trdb/trdb.exe</a>
TReaDS	No	No	Pellegrini 2012	14	<a href="http://bioalgo.iit.cnr.it/treads/">http://bioalgo.iit.cnr.it/treads/</a>
TRStalker	No	No	Pellegrini 2010	40	<a href="http://bioalgo.iit.cnr.it/">http://bioalgo.iit.cnr.it/</a>
VNTRfinder	Yes	No	Odushlaine 2006	16	<a href="http://bioinformatics.rcsi.ie/vntrfinder/">http://bioinformatics.rcsi.ie/vntrfinder/</a>

“Yes” has been highlighted in the columns for visualization purposes.

compares the 42 functional programs featured in the main paper, as summarized in Table 1. However, although they were not included in the main text discussion, the 32 non-functional programs (listed in Table 2) were also evaluated and their data is available in

Supplementary Table S1, acknowledging the possibility that some may become operational again in the future. The complete gathered data for all 74 tools is also provided in Supplementary Table S1, which includes not only the analyzed information but also filters to



facilitate deeper analysis, along with more specific data for each tool reviewed, such as the parameters utilized, input and output formats, features, and availability of each tool. It is advisable that users download the [Supplementary Table S1](#) and take it as a resource to consult while selecting the tool to use in their research.

To facilitate comparison, the tools were summarized based on their focus of analysis—Repetitive Elements (RE), Tandem Repeats (TR), and Short Tandem Repeats (STR). These categories were further divided by tool type (Basic or Pipeline Tool) and by their ability to identify only perfect SSRs (p-SSR) or both perfect and imperfect SSRs (i-SSR). The main data was presented in tables to enhance user access to the comparison of specific features, accompanied by a brief analysis of select attributes, including advantages, disadvantages, and the applicability of each tool to SSR research. The tools were analyzed and discussed in alphabetical order, without bias toward citations. However, if readers wish to consider citation frequency as a criterion, they can refer to [Figures 6, 8](#) to identify the most popular tools in the literature. Additionally, advanced users with specific research questions or familiarity with SSR tools can focus primarily on the tables in the text and [Supplementary Table S1](#) to compare relevant aspects. Users may also filter for suitable tools, in the tables and [Supplementary Table S1](#), then compare and read the descriptions of those selected tools, rather than reviewing all descriptions, as many tools share similar functionalities.

### 3.2.1 Tools for detecting repetitive elements and tandem repeats

The analysis identified nine tools with a generalist focus, comprising three tools for detecting RE and six tools for TR (Table 3). As illustrated in [Figure 2](#), short tandem repeats (STRs) are a type of tandem repeat, which belongs to the broader category of repetitive elements. These tools were included because, although STR are not their sole focus, they represent a significant subgroup

within the broader spectrum of repetitive elements and tandem repeats, thus are present in their outputs ([Sharma et al., 2007](#)). Many studies and reviews have applied these tools for STR analysis, given that they usually include parameters that can be personalized for identifying MS ([Gao et al., 2003](#); [Leclercq et al., 2007](#)). Additionally, it is essential to highlight that their citation rates observed in [Table 1](#) and [Figure 8](#) may not exclusively reflect citations in SSR-related projects, as these tools can also detect other types of repeats.

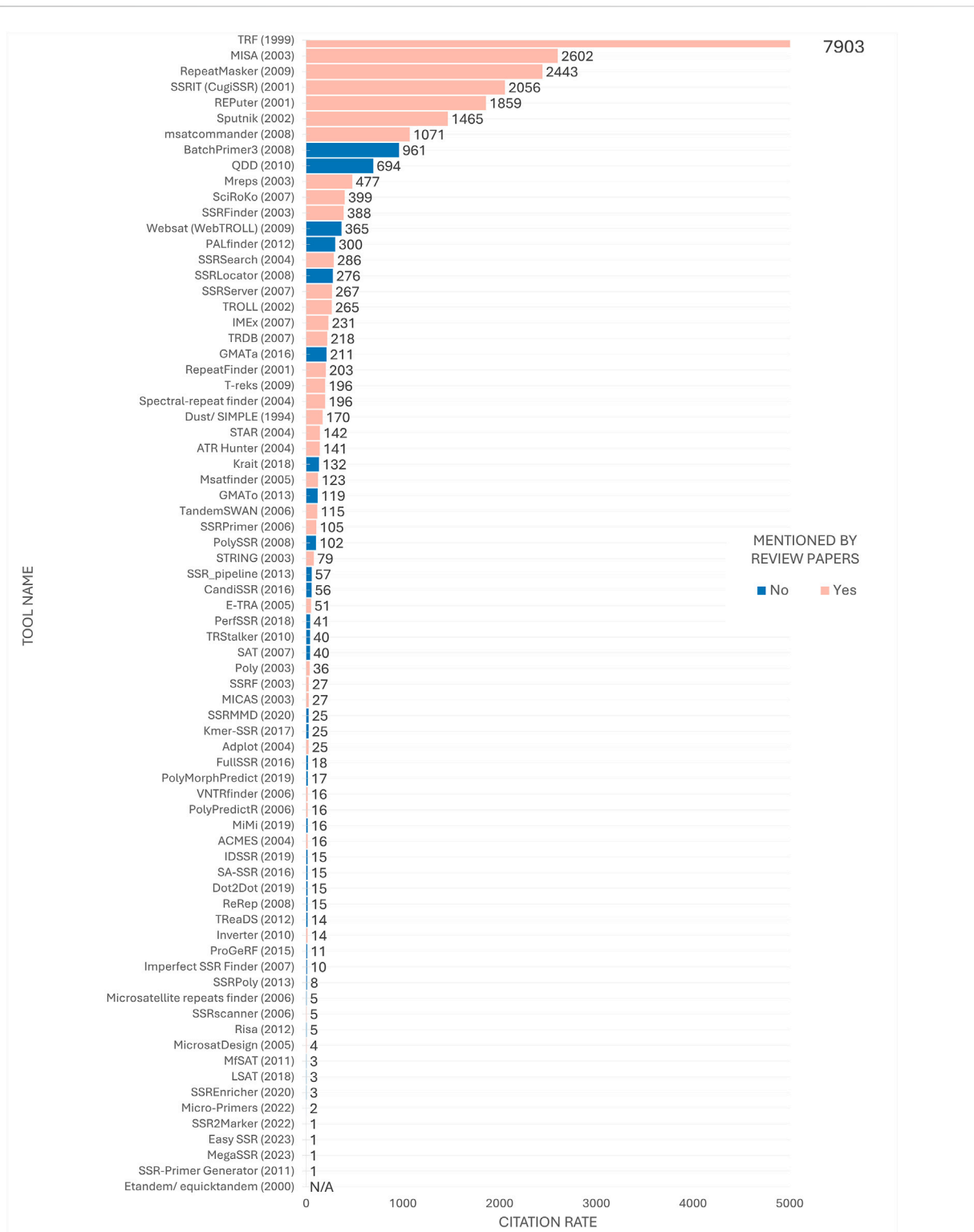
Upon comparison of the tools, as depicted in [Table 3](#), it was observed that none of them integrated primer design functionality natively. These tools employ a variety of algorithms, with some identifying only p-SSRs, both p-SSRs and i-SSRs, while only REPuter ([Kurtz et al., 2001](#)) allows the identification of compound SSRs and provides Graphs/Charts as outputs. Except for Dot2Dot ([Genovese et al., 2019](#)) and Etandem/Equicktandem ([Rice et al., 2000](#)), which can only be executed locally through the command line, the remaining tools offer a web server for graphical interface analysis. Dot2Dot stands out as the only tool that accepts FastQ files as input, while the others generally accept Fasta files.

#### 3.2.1.1 RE tools

RepeatMasker ([Tarailo-Graovac and Chen, 2009](#)) is a software developed by Smit and Green (A.F.A. Smit, R. Hubley and P. Green, unpublished data). It annotates repeats, replacing them with Ns, and returns masked sequences, a table of repeat content, and optional alignments. The web version has a 100 kb sequence limit, with longer sequences queued. It is important to note that RepeatMasker lacks specific parameters for defining the size of SSR motifs or the minimum number of iterations, offering options for selecting which repetitive elements to mask.

REPuter ([Kurtz et al., 2001](#)) offers comprehensive and efficient detection of various repeat types, coupled with a robust evaluation of their significance and interactive visualization capabilities. However, the online version of the software has a 5 Mb data size limit for





**FIGURE 8** Citation ranking of SSR tools. Number of citations retrieved from Google Scholar in early 2024 by tool's papers. The year of release is indicated in parenthesis next to the tool's name for temporal comparison purposes. Tools mentioned in the review articles under study are highlighted in pink, while tools studied solely in the present study are highlighted in blue. The citation rate for the Etandem/equicktandem tools was considered N/A (Not Available) as their citations were combined with other tools citing EMBOSS, making it currently inviable to estimate the number of citations specifically for these tools.

TABLE 3 Tools for detecting repetitive elements (RE) and tandem repeats (TR).

Name	Reference/Year	Type	Mining tool	Integrated toolset	Platform	Execution	Focus	Max motif length	Input files	Output graphs/Charts	p-SSR	i-SSR	c-SSR	Primer design	Tool's link
RepeatMasker	Smit 1996 apud Tarailo-Graovac 2009	Pipeline Tool	RepeatMasker	cross_match, WU-BLAST	Linux, Web	Graphical Interface and Command Line	RE	No limit	FASTA	No	Yes	Yes	No	No	<a href="https://www.repeatmasker.org/cgi-bin/WEBRepeatMasker">https://www.repeatmasker.org/cgi-bin/WEBRepeatMasker</a>
REPuter	Kurtz 2001	Basic tool	REPuter	No	Linux, Mac OS, Web	Graphical Interface and Command Line	RE	<5 000 bp	DNA sequence	Yes	Yes	Yes	Yes	No	<a href="http://bibiserv.techfak.uni-bielefeld.de/reputer/">http://bibiserv.techfak.uni-bielefeld.de/reputer/</a>
Spectral-repeat finder	Sharma 2004	Basic tool	Spectral-repeat finder	No	Linux, Mac OS, Windows, Web	Graphical Interface	RE	No limit	FASTA; GBK	No	Yes	Yes	No	No	<a href="http://www.imtech.res.in/raghava/srf">http://www.imtech.res.in/raghava/srf</a>
Dot2Dot	Genovese 2019	Basic tool	Dot2Dot	No	Linux, Mac OS	Command Line	TR	No limit	FASTA; FASTQ	No	Yes	Yes	No	No	<a href="https://github.com/Gege7177/Dot2dot">https://github.com/Gege7177/Dot2dot</a>
Etandem/Equicktandem	Rice 2000	Basic tool	Etandem/Equicktandem	No	Linux	Command Line	TR	1–600 bp	FASTA; GBK; GFF; EMBL and Others	No	Yes	No	No	No	<a href="https://www.bioinformatics.nl/cgi-bin/emboss/help/equicktandem">https://www.bioinformatics.nl/cgi-bin/emboss/help/equicktandem</a>
Mreps	Kolpakov 2003	Basic tool	Mreps	No	Linux, Windows, Web	Graphical Interface and Command Line	TR	No limit	FASTA	No	Yes	Yes	No	No	<a href="https://mreps.univ-mlv.fr/">https://mreps.univ-mlv.fr/</a>
STAR	Delgrange 2004	Basic tool	STAR	No	Linux, Mac OS, Windows, Web	Graphical Interface and Command Line	TR	No limit	FASTA	No	Yes	Yes	No	No	<a href="http://atgc.lirmm.fr/star/">http://atgc.lirmm.fr/star/</a>
T-reks	Jorda 2009	Pipeline Tool	T-reks	CLUSTALW, MUSCLE	Linux, Mac OS, Windows, Web	Graphical Interface and Command Line	TR	No limit	FASTA	No	Yes	Yes	No	No	<a href="https://bioinfo.crbm.cnrs.fr/index.php?route=tools&amp;tool=3">https://bioinfo.crbm.cnrs.fr/index.php?route=tools&amp;tool=3</a>
TRF	Benson 1999	Basic tool	TRF	No	Linux, Mac OS, Windows, Web	Graphical Interface and Command Line	TR	1–5, 10–2000 bp	FASTA	No	Yes	Yes	No	No	<a href="https://tandem.bu.edu/trf/trf.html">https://tandem.bu.edu/trf/trf.html</a>

"Yes" and "Pipeline" have been highlighted in the columns for visualization purposes. "RE" = repetitive elements, "TR" = Tandem Repeats. "BP" = Base pair. "p-SSR" = Perfect SSR. "i-SSR" = Imperfect SSR. "c-SSR" = Compound SSR.

uploaded data and a 5000-repeat cap due to server capacity constraints. For users who wish to analyze only SSRs, this program has limitations, as it only allows the definition of parameters such as minimal repeat size, maximum computed repeats, edit distance and hamming distance.

Spectral-repeat finder (SRF) (Sharma et al., 2004) employs an *ab initio* methodology, not relying on any prior assumptions regarding the length, fidelity, or arrangement of repeats. For users who seek to utilize RE tools for the study of SSR, this tool offers an appropriate level of flexibility, with a comprehensive range of parameters accessible via their web server, enabling users to identify repeats with varying sizes and specify a minimum number of repetitions. However, it should be noted that the web tool has a limitation in that it can only process one file at a time. Consequently, for larger projects, it is necessary to use a multifasta file or to perform individual genome analysis.

### 3.2.1.2 TR tools

Dot2Dot (Genovese et al., 2019) employs a distinctive methodology for the identification of pure and fuzzy TRs through dot-plot matrices. As the only tool in this group to accept both assembled genomes (FASTA) and NGS data (FASTQ) as input, it allows the analysis of NGS sequences and generates tabular outputs in the formats “.bed” and “.dot”. While capable of identifying TRs of any size, this tool may be particularly valuable for researchers studying microsatellites, as it allows for specifying the minimum and maximum motif sizes in the parameters.

Etrandem/Equicktandem programs (Rice et al., 2000) are components of the EMBOSS package, developed by Richard Durbin, utilized in tandem repeat identification in DNA. Equicktandem identifies repeats for each pattern size up to a specified limit, whereas Etrandem, which should be employed after the other tool, computes a potential consensus of the repeated pattern. The tools support a wide range of output formats, and they apply to SSRs, as they allow for the definition of maximum and minimum motif sizes and enable the study of perfect and/or imperfect SSRs by including a parameter for mismatches.

Mreps (Kolpakov et al., 2003) employs a mixed combinatorial/heuristic approach to identify repeats of all possible sizes within a single program run. The output includes detailed information about each repeat, such as start and end positions, size, period, exponent, error level, and the repeat sequence. The web interface provides visualizations of repeat alignments. For SSR analyses, mreps allows users to define motif size and iteration parameters.

STAR (Delgrange and Rivals, 2004) is capable of identifying significant approximate (imperfect) tandem repeats of a given motif in DNA sequences. It distinguishes between exact tandem repeats (ETRs), which result from the tandem duplication of the motif, and approximate tandem repeats (ATRs), which arise from ETRs through point mutations. While STAR can be utilized to investigate microsatellites, its primary objective is to identify specific motifs pre-defined by the user within sequences, rather than to identify all SSRs.

T-reks (Jorda and Kajava, 2009) is designed primarily for the analysis of TR in proteins, although it also works with DNA sequences. It offers a standalone mode with a user-friendly

graphical interface for local use and a web interface version. However, the web interface is limited to approximately 100,000 amino acids as input, does not allow file upload, or the definition of parameters such as minimum and maximum repeat sizes. This may be an issue for studies that focus on SSRs.

Tandem Repeats Finder (TRF) (Benson, 1999) is a software program that can be used to locate and display TR in DNA sequences. TRF allows users to restrict their search to small motifs of 1–5 period sizes, which is useful for SSR research. However, TRF does not offer an option to search for perfect arrays only, as it considers mismatches in its default parameters. One advantage of TRF is that users are not required to specify the pattern, pattern size, or any other parameter.

## 3.2.2 Tools specific for detecting microsatellites

A total of 33 tools specifically designed for SSR identification were identified, including 11 basic tools and 22 pipeline tools.

### 3.2.2.1 Basic tools

**3.2.2.1.1 Basic tools for p-SSR.** A comparison of the seven tools that focus on perfect SSRs (Table 4) revealed that none of them have the integrated primer design function. All of them can be executed via command line on Linux; however, only GMATo (Wang et al., 2013) and PERF (Avvaru et al., 2018) can also be executed on Windows and MacOS, with GMATo being the only one with a graphical interface for local executions. Additionally, only SSRIT (Temnykh et al., 2001) and Misa (Thiel et al., 2003) have web versions. Despite focusing on p-SSRs, GMATo and Kmer-SSR (Pickett et al., 2017) do not define a limit for the maximum motif size identified. Others, such as Misa, SA-SSR (Pickett et al., 2016), and SSRIT (Temnykh et al., 2001), allow the analysis of motifs larger than 6 bp. Among these, only PERF (Avvaru et al., 2018) can use FastQ files as input and generate user-friendly chart outputs.

GMATo (Wang et al., 2013) stands out by providing graphical interfaces for SSR mining through local execution. Also, c-SSR and i-SSR can be derived from the SSR loci output using additional scripts. GMATo outputs p-SSR reports and statistical distribution files in tab-delimited plain text format for easy import into other applications. An interesting point to note is that the developers of GMATo identified the need for future development to include additional functions. To address this, they launched a new pipeline tool named GMATa (Wang and Wang, 2016), with a novel algorithm, integrating features such as graphical display of statistical data and primer designing.

Kmer-SSR (Pickett et al., 2017) and SA-SSR (Pickett et al., 2016) are tools developed by the same author that share similar execution, parameters, and output format. However, they utilize different algorithms for SSR discovery in large genetic sequences. The SA-SSR employs a suffix array-based algorithm, while the Kmer-SSR's algorithm is based on k-mer decomposition. In its tool validation section, the Kmer-SSR demonstrated better benchmark test performance than the SA-SSR. Moreover, Kmer-SSR offers a range of filters for analysis, including those based on atomicity, cyclic duplicates, enclosed SSRs, minimum SSR length, and specific SSR period sizes.

Misa (Thiel et al., 2003) identifies p-SSRs and c-SSRs in fasta sequences, providing the SSRs identified and statistics. Despite not having an integrated primer design feature, MISA provides

TABLE 4 Tools for detecting SSR part 1 – basic tools for perfect SSR analysis.

Name	Reference/ Year	Mining tool	Integrated toolset	Platform	Execution	Focus	Max motif length	Input files	Output graphs/ Charts	p-SSR	i-SSR	c-SSR	Primer design	Tool's link
GMATo	<a href="#">Wang 2013</a>	GMATo	No	Linux, Mac OS, Windows	Graphical Interface and Command Line	SSR	No limit	FASTA	No	Yes	No	No	No	<a href="https://sourceforge.net/projects/gmato/files/?source=navbar">https:// sourceforge.net/ projects/gmato/ files/?source= navbar</a>
Kmer-SSR	<a href="#">Pickett 2017</a>	K-mer-SSR	No	Linux	Command Line	SSR	No limit	FASTA	No	Yes	No	No	No	<a href="https://github.com/ridgelab/Kmer-SSR">https://github. com/ridgelab/ Kmer-SSR</a>
Misa	<a href="#">Thiel 2003</a>	MISA	No	Linux, Web	Graphical Interface and Command Line	SSR	1–6 bp	FASTA	No	Yes	No	Yes	No	<a href="https://webblast.ipk-gatersleben.de/misa/">https://webblast. ipk-gatersleben. de/misa/</a>
PERF	<a href="#">Avvaru 2018</a>	PerfSSR	No	Linux, Windows, Mac OS	Command Line	SSR	1–6 bp	FASTA; FASTQ	Yes	Yes	No	No	No	<a href="https://github.com/rkmlab/perf">https://github. com/rkmlab/perf</a>
SA-SSR	<a href="#">Pickett 2016</a>	SA-SSR	No	Linux	Command Line	SSR	1–7 bp	FASTA	No	Yes	No	No	No	<a href="https://github.com/ridgelab/SA-SSR">https://github. com/ridgelab/ SA-SSR</a>
SSRIT (CugiSSR)	<a href="#">Temnykh 2001</a>	SSRIT	No	Linux, Web	Graphical Interface and Command Line	SSR	2–10 bp	DNA Sequence	No	Yes	No	No	No	<a href="https://archive.gramene.org/db/markers/ssrtool">https://archive. gramene.org/db/ markers/ssrtool</a>
TROLL	<a href="#">Castelo 2002</a>	Troll	No	Linux	Command Line	SSR	1–6 bp	DNA sequence	No	Yes	No	No	No	<a href="https://sourceforge.net/projects/finder/">https:// sourceforge.net/ projects/finder/</a>

"Yes" has been highlighted in the columns for visualization purposes. "SSR" = Simple Sequence Repeats or microsatellite. "BP" = Base pair. "p-SSR" = Perfect SSR. "i-SSR" = Imperfect SSR. "c-SSR" = Compound SSR.



supplementary scripts for command line execution to process outputs in Primer3 (Untergasser et al., 2012). The web version (Beier et al., 2017) limits input files to 2 Mb and does not provide online result visualization, emailing the results instead. Misa does not support batch processing, requiring users to handle files individually or as a multifasta file. Advanced users often create custom batch scripts to address this limitation.

PERF (Avvaru et al., 2018) compares direct strings to repeat to avoid missing overlapping or mid-motif SSRs in Fasta or FastQ sequences. It accepts both GFF and GTF format files and can classify SSRs as Genic, Exonic, Intronic, or Intergenic based on their position. Despite being a command-line tool, PERF provides user-friendly, post-processed outputs. It can produce interactive HTML reports, including a bar chart displaying the 10 most frequent repeats, a depiction of the relative distribution of repeats categorized by motif length, and a line chart illustrating the relationship between length and frequency of selected repeats.

SSRIT (Temnykh et al., 2001) provides a command-line and web-based tool. However, the command-line version was not accessible via the provided link on their website. The webtool, though user-friendly, presented limitations. It only accepts pasted sequences in fasta format and allows the selection of the maximum motif length (2–10 nucleotides) and minimum number of repeats as parameters, without the option to set different minimum repeats for each motif length. The results are displayed in tabular format, showing the motif, number of repeats, SSR start, and SSR end. Despite these constraints, the platform is efficient and straightforward, suitable for rapid online analyses.

TROLL (Castelo et al., 2002), despite its name, is primarily focused on identifying p-SSRs using a pre-defined dictionary containing all possible SSR motifs. Its execution requires input arguments specifying the minimum length of SSRs, the maximum motif length, and files containing the motif list and DNA sequence. The output includes the starting position of each repeat, the motif, and the repeat length in base pairs (bp). The resulting file can be readily imported into a variety of applications and spreadsheets. Additionally, customized filter scripts can be created to process TROLL's output.

**3.2.2.1.2 Basic tools for p-SSR and i-SSR analysis.** Among the four basic tools capable of identifying i-SSRs, summarized in Table 5, none integrated primer design. Although IMEx (Mudunuri and Nagarajaram, 2007; Mudunuri et al., 2010a) claims to have this functionality in its web version, this version is currently non-functional. However, IMEx (Mudunuri et al., 2010a) provides a graphical interface for local executions and identifies SSRs with motifs ranging from 1 to 6 bp. In contrast, Imperfect SSR Finder (Stieneke and Eujayl, 2007) and Microsatellite Repeats Finder (Bikandi, 2006) are web tools that do not detect mononucleotides, but can analyze motifs up to 10 bp. None of the tools accept FastQ input files. Only IMEx can determine if the SSR is in a coding or non-coding position by accepting PTT files. Furthermore, the only two tools that can identify c-SSRs are IMEx and Imperfect SSR Finder. Additionally, none of the tools generate graphs or charts.

IMEx (Mudunuri and Nagarajaram, 2007) is an effective tool for identifying p-SSR, i-SSR, and c-SSR. Its graphical interface for local execution, inclusion of flanking regions in the output for primer design, support for batch mode input files, and generation of HTML

and text outputs distinguish it from other similar software. IMEx aligns each repeat with its consensus sequence and can classify SSRs as coding or non-coding. The primary limitations include the numerous customizable parameters and the requirement for a PTT format file for analyzing coding regions, which might pose challenges for those accustomed to GenBank format annotations. However, these limitations are addressed by the user-friendly pipeline web tool EasySSR (Alves et al., 2023), which runs IMEx as a mining algorithm.

Imperfect SSR Finder (Stieneke and Eujayl, 2007) is a webtool based on the SSRIT algorithm, that has been modified to identify p-SSR, i-SSR, and c-SSR. Despite its user-friendly interface and capacity to seemingly handle inputs of any size, it lacks mononucleotide repeat detection and may initially seem complex due to its extensive parameter range. For i-SSR and c-SSR, the user can define a non-repeating region (NRR), which is the maximum distance of seemingly random nucleotides separating the SSRs. In this context, NRR might be considered synonymous with mismatches or with dMAX. One drawback is that it does not include flanking regions in its output.

Sputnik repeat-finder tool, developed by Chris Abajian in 1994 (unpublished data), is designed to identify microsatellites with a customizable deviation from a perfect repeat. Two modified versions of the tool have been developed: Modified Sputnik I (Morgante et al., 2002) and Modified Sputnik II (La Rota et al., 2005). The most recent version identified was that of La Rota, but some pipelines tools like PolySSR (Tang et al., 2008) and SSRpoly (Duran et al., 2013) also employ Sputnik as a basic tool for SSR mining with potential additional modifications. Modified Sputnik II has versatile parameters, adjusts FASTA sequence header parsing, and formats reports for direct database import. Its command-line interface, however, may pose challenges for inexperienced users.

Microsatellite repeats finder (Bikandi, 2006) is a web-based tool designed for quick analysis of perfect p-SSR and i-SSR. While the source code is available for users, the platform primarily operates online, limiting sequence analysis to 100,000 bp. The tool is straightforward, with few parameters. The output is also simple, displaying the SSR start position, their class, the number of iterations, and the repeated sequence. This tool may be of benefit to researchers seeking a rapid assessment of SSR presence within their sequences.

### 3.2.2.2 SSR pipeline tools

This study identified 22 functional pipeline tools for SSR analysis, with 18 of these incorporating primer design functionality. Among the 22 tools identified, 14 are exclusively focused on the analysis of p-SSRs, while the remaining 8 are designed to analyze both perfect and imperfect SSRs.

**3.2.2.2.1 Pipeline tools for p-SSR analysis.** Regarding the 14 pipeline tools with focus on the identification of perfect SSRs, as summarized in Table 6, it was observed that all the pipeline tools integrate an SSR mining algorithm with other functions. It should be noted that most integrate Primer3 (Untergasser et al., 2012) and have the native primer design function, except MICAS (Sreenu et al., 2003) and SSRenricher (Luo et al., 2020). Six tools employ original algorithms for SSR mining, while eight utilize consolidated basic tools as mining algorithms. Among these, MICAS (Sreenu et al.,

TABLE 5 Tools for detecting SSR part 2 – basic tools for perfect and Imperfect SSR analysis.

Name	Reference/ Year	Mining tool	Integrated toolset	Platform	Execution	Focus	Max motif length	Input files	Output graphs/ Charts	p-SSR	i-SSR	c-SSR	Primer design	Tool's link
IMEx	<a href="#">Mudunuri 2007</a>	IMEx	No	Linux, Web	Graphical Interface and Command Line	SSR	1–6 bp	FASTA; PTT	No	Yes	Yes	Yes	No	<a href="http://www.mcr.org.in/IMEX/download.html">http://www.mcr.org.in/IMEX/download.html</a>
Imperfect SSR Finder	<a href="#">Stieneke 2007</a>	Imperfect SSR Finder	No	Web	Graphical interface	SSR	2–10 bp	FASTA	No	Yes	Yes	Yes	No	<a href="https://ssr.nwisr.ars.usda.gov/">https://ssr.nwisr.ars.usda.gov/</a>
Microsatellite repeats finder	<a href="#">Bikandi 2006</a>	Microsatellite repeats finder	No	Web	Graphical Interface	SSR	2–10 bp	DNA Sequence	No	Yes	Yes	No	No	<a href="http://insilico.ehu.es/mini_tools/microsatellites/?info">http://insilico.ehu.es/mini_tools/microsatellites/?info</a>
Sputnik	<a href="#">Morgante 2002</a>	Sputnik	No	Linux, Windows	Command Line	SSR	1–5 bp	FASTA	No	Yes	Yes	No	No	<a href="http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/">http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/</a>

“Yes” has been highlighted in the columns for visualization purposes. “SSR” = Simple Sequence Repeats or microsatellite. “BP” = Base pair. “p-SSR” = Perfect SSR. “i-SSR” = Imperfect SSR. “c-SSR” = Compound SSR.

TABLE 6 Tools for detecting SSR part 3 – pipeline tools for perfect SSR analysis.

Name	Reference/Year	Mining tool	Integrated toolset	Platform	Execution	Focus	Max motif length	Input files	Output graphs/Charts	p-SSR	i-SSR	c-SSR	Primer design	Tool's link
CandiSSR	Xia 2016	MISA	BLAST, Primer3 and Clustalw	Linux	Command Line	SSR	2–6 bp	FASTA	No	Yes	No	Yes	Yes	<a href="https://github.com/xiaenhua/CandiSSR">https://github.com/xiaenhua/CandiSSR</a>
FullSSR	Metz 2016	FullSSR	Primer3	Linux	Command Line	SSR	1–5 bp	FASTA	Yes	Yes	No	No	Yes	<a href="https://sourceforge.net/projects/fullssr/">https://sourceforge.net/projects/fullssr/</a>
GMATa	Wang 2016	GMATa	Primer3, e-PCR	Linux, Mac OS, Windows	Graphical Interface and Command Line	SSR	No limit	FASTA; FASTQ	Yes	Yes	No	No	Yes	<a href="https://sourceforge.net/projects/gmata/?source=navbar">https://sourceforge.net/projects/gmata/?source=navbar</a>
MegaSSR	Mokhtar 2023	MISA	Primer3, Primersearch	Linux, Web	Graphical Interface and Command Line	SSR	1–6 bp	FASTA; GFF	Yes	Yes	No	Yes	Yes	<a href="https://bioinformatics.um6p.ma/MegaSSR">https://bioinformatics.um6p.ma/MegaSSR</a>
MICAS	Sreenu 2003	IMEx	IMEx, MICdb3.0	Web	Graphical Interface	SSR	1–10 bp	FASTA; PTT	No	Yes	No	No	No	<a href="http://www.mcr.org.in/micas/">http://www.mcr.org.in/micas/</a>
Micro-Primers	Alves 2022	MISA	Trimmomatic, Cutadapt, FLASH, CD-HIT, Primer3	Linux, Mac OS	Graphical Interface and Command Line	SSR	No limit	FASTQ	No	Yes	No	Yes	Yes	<a href="https://github.com/FilAlves/micro-primers">https://github.com/FilAlves/micro-primers</a>
MiMi	Fox 2019	PALfinder	Muscle, Primer3	Linux, Mac OS	Command Line	SSR	1–6 bp	FASTA; FASTQ	No	Yes	No	No	Yes	<a href="https://github.com/graemefox/mimi">https://github.com/graemefox/mimi</a>
msatcommander	Faircloth 2008	msatcommander	Primer3	Linux, Mac OS, Windows	Graphical Interface and Command Line	SSR	1–6 bp	FASTA	No	Yes	No	No	Yes	<a href="https://code.google.com/archive/p/msatcommander/">https://code.google.com/archive/p/msatcommander/</a>
PALfinder	Castoe 2012	PALfinder	Primer3, RepBase	Linux, Mac OS	Command Line	SSR	2–6 bp	FASTA; FASTQ	No	Yes	No	No	Yes	<a href="https://sourceforge.net/projects/palfinder/">https://sourceforge.net/projects/palfinder/</a>
PolyMorphPredict	Das 2019	MISA	Primer3, e-PCR	Web	Graphical Interface	SSR	1–6 bp	FASTA	No	Yes	No	Yes	Yes	<a href="http://webtom.cabgrid.res.in/polypred/">http://webtom.cabgrid.res.in/polypred/</a>
SSR2Marker	Yue 2022	MISA	BLAST, MAFFT, Primer3 and e-PCR	Linux	Command Line	SSR	1–6 bp	FASTA; MISA	No	Yes	No	Yes	Yes	<a href="https://github.com/aaranyue/SSR2Marker">https://github.com/aaranyue/SSR2Marker</a>

(Continued on following page)

TABLE 6 (Continued) Tools for detecting SSR part 3 – pipeline tools for perfect SSR analysis.

Name	Reference/Year	Mining tool	Integrated toolset	Platform	Execution	Focus	Max motif length	Input files	Output graphs/Charts	p-SSR	i-SSR	c-SSR	Primer design	Tool's link
SSRenricher	Luo 2020	MISA	MISA, CD-HIT, Muscle	Linux, Mac OS	Graphical Interface and Command Line	SSR	1–6 bp	FASTQ	No	Yes	No	Yes	No	<a href="https://github.com/byemaxx/SSRenricher">https://github.com/byemaxx/SSRenricher</a>
SSRMMD	Gou 2020	SSRMMD	Primer3	Linux, Mac OS, Windows	Command Line	SSR	1–6 bp	FASTA	No	Yes	No	Yes	Yes	<a href="https://github.com/GouXiangjian/SSRMMD">https://github.com/GouXiangjian/SSRMMD</a>
Webat (WebTROLL)	Martins 2009	Troll	Troll, Primer3	Web	Graphical Interface	SSR	1–6 bp	FASTA	No	Yes	No	No	Yes	<a href="https://bioinfo.inf.ufg.br/websat/">https://bioinfo.inf.ufg.br/websat/</a>

"Yes" has been highlighted in the columns for visualization purposes. "SSR" = Simple Sequence Repeats or microsatellite. "BP" = Base pair. "p-SSR" = Perfect SSR. "i-SSR" = Imperfect SSR. "c-SSR" = Compound SSR.

2003) and WebSat (Martins et al., 2009) are the only ones that do not integrate Misa (Thiel et al., 2003), employing instead IMEx (Mudunuri and Nagarajaram, 2007) and Troll (Castelo et al., 2002), respectively, as their basic tools.

In terms of execution, as detailed in table 6, there is a diverse range of profiles among the tools. Some, like WebSat (Martins et al., 2009), are solely web-based, while others, such as SSRMMD (Gou et al., 2020) and MiMi (Fox et al., 2019), are exclusively for local execution via command lines. Additionally, tools like Micro-primers and GMATa (Wang and Wang, 2016) offer graphical user interfaces for local executions, and tools like Misa (Thiel et al., 2003) and MegaSSR (Mokhtar et al., 2023) provide web servers and execution via command line.

Most tools target SSRs ranging from 1 to 6 base pairs. However, notable exceptions include GMATa (Wang and Wang, 2016) and Micro-primers (Alves et al., 2022), which do not impose size restrictions on the repeats, and PALfinder (Castoe et al., 2012) and CandiSSR (Xia et al., 2016), which specifically exclude mononucleotide repeats. Moreover, an increase in the number of analytical tools capable of processing FastQ files, identifying c-SSRs, and generating graphical representations such as charts and graphs as part of their output was observed. Following the structure used so far, some unique features of each tool will be highlighted.

CandiSSR (Xia et al., 2016) is a pipeline that integrates Misa (Thiel et al., 2003) with BLAST (Altschul et al., 1997), Primer3 (Untergasser et al., 2012), and Clustalw (Thompson et al., 2003). Its purpose is to automatically identify candidate polymorphic SSRs (polySSRs) from multiple assembled sequences. It can identify putative polySSRs from transcriptome datasets and assembled genome sequences. The pipeline provides two confidence metrics (standard deviation and missing rate of SSR repetitions) to assess polySSR feasibility. It automatically designs primer pairs and evaluates primer-binding region similarities for successful marker development. The output includes detailed information on candidate PolySSRs, primer pairs, and flanking sequences for further genetic studies.

FullSSR (Metz et al., 2016) is a pipeline designed to identify microsatellite loci by using their algorithm and design primers with primer3 (Untergasser et al., 2012). The tool produces three types of output: a clear HTML report, Primer3 native outputs, and individual text files for each identified SSR and their respective primers. The publication that released the FullSSR tool highlights that in comparison to MISA (Thiel et al., 2003), FullSSR has superior capability in identifying SSRs suitable for primer design, as many SSRs identified by Misa may be unsuitable for primer design due to their proximity to sequence ends, precluding primer design for both sides.

GMATa (Wang and Wang, 2016) is a pipeline tool developed by the same authors as GMATo (Wang et al., 2013), but with its own novel SSR mining algorithm and integrating Primer3 (Untergasser et al., 2012), and e-PCR (Schuler, 1997). It accepts Fasta and FastQ files, with a workflow of six modules that can be seamlessly integrated and executed automatically or independently. GMATa provides graphical display within a genome browser and produces various user-friendly outputs, such as SSR loci results, SSR statistical analysis and graphic plotting, genome wide SSR marker design, marker mapping results, and polymorphisms these outputs facilitate



the generation of high-quality statistical data that is suitable for direct integration into publications.

MegaSSR (Mokhtar et al., 2023) is a novel tool that is available as a web server and a standalone pipeline. It utilizes Misa (Thiel et al., 2003) as the basic tool and integrates the results with Primer3 (Untergasser et al., 2012) and Primersearch (Rice et al., 2000), which enables the design of PCR-based primers and *in silico* PCR validation. The tool accepts as input fasta files, and additional annotations in gff or gff3 format if the user wants to locate coding or non-coding regions. The results of MegaSSR can be accessed through an interactive HTML page that includes graphs and tables containing various aspects of SSR markers and corresponding PCR primers. A limitation of the tool, however, is its single-file analysis capability.

MICAS (Sreenu et al., 2003) is a web server designed for the study of non-redundant microsatellites in bacterial or archaeal genomes. Their integrated webtool for SSR mining is no longer fully functional, preventing users from identifying new SSRs in MICAS. Nevertheless, MICAS 3.0 also integrates MICdb3.0 (Mudunuri et al., 2014), which contains SSR information from over 5,000 prokaryotic genome sequences. This functionality enables users to explore and compare SSR distributions and perform pairwise genome comparisons. Users can export data in various formats for further analysis. Despite the limitations in new SSR identification, MICAS remains a valuable resource for studying microsatellites in prokaryotic genomes, providing access to a vast database of pre-extracted data.

Micro-primers (Alves et al., 2022) integrates Misa with Trimmomatic (Bolger et al., 2014), Cutadapt (Martin, 2011), FLASH (Magoč and Salzberg, 2011), CD-HIT (Li and Godzik, 2006), and Primer3 (Untergasser et al., 2012) to identify and design PCR primers for amplifying SSR loci. It takes as input a FASTQ file containing sequences (reads) from NGS (next-generation sequencing). The output is a text file with information about the microsatellite markers, including the number of alleles, melting temperature, and corresponding primer set products, which facilitates marker selection. It can be executed via command line or a graphical interface. Therefore, it could be a valuable tool for researchers analyzing newly sequenced data, for metagenomic projects, or integration into other pipelines.

Msatcommander (Faircloth, 2007) is a tool designed to facilitate the identification of SSR by integrating their original algorithm with primer3 (Untergasser et al., 2012) for primer design and primer tagging through an automated process. It offers a local graphical user interface or command line for execution. The results and primers are outputted in CSV format for easy integration with spreadsheet or database programs. Primer-specific files generated by Primer3 are provided in TXT format, compatible with spreadsheet, database, or text-editor applications.

PALfinder (Castoe et al., 2012) is a command-line pipeline for the automated analysis of assembled genomes and sequencing reads, to identify SSRs and design PCR primers for potentially amplifiable SSR loci (PALs). It integrates an original SSR mining algorithm with Primer3 (Untergasser et al., 2012) and RepBase (Jurka et al., 2005), accepting Fasta and FastQ, formats for input, including Illumina paired-end reads and 454 single-end reads. PALfinder identifies SSR-containing reads and selects flanking sequences suitable for PCR primer sites, resulting in PALs presented in a tab-delimited

format with primer information. Controlled through a customizable parameter settings file, PALfinder is valuable for research groups analyzing SSRs directly from sequencing reads and can be integrated into other pipelines, as demonstrated by its integration into the MiMi pipeline (Fox et al., 2019).

MiMi (Fox et al., 2019) enhances SSR mining efficiency by incorporating the PALfinder (Castoe et al., 2012) pipeline and the Muscle alignment tool (Edgar, 2004). It can use Fasta and FastQ files as input and aims to compare genomic data from multiple individuals of the same species, a departure from using data from a single individual. The software allows for the *in silico* identification of polymorphic loci and other key characteristics of potential microsatellite markers. It also enables marker amplification by PCR and reduces the number of markers requiring laboratory testing for polymorphism, thereby improving overall marker development success. Furthermore, MiMi enables clear visualization and avoidance of insertion/deletions in flanking regions when designing microsatellite panels, which is a significant source of error.

PolyMorphPredict (Das et al., 2019) integrates Misa (Thiel et al., 2003) with Primer3 (Untergasser et al., 2012) and e-PCR (Schuler, 1997) to mine microsatellite loci and compute primers from genome/transcriptome data of any species, focusing however in agriculture research. It performs e-PCR using published primers for polymorphism discovery and across species transferability of microsatellite loci. The tool also compares various whole genome sequences and their genotypes to discover microsatellite loci, identify polymorphic loci, and design primers for rapid genotyping. The tool is only available as a web service, which occasionally malfunctions. However, it is a valuable tool for designing primers for polymorphic SSR markers.

SSR2marker (Yue and Liu, 2022) is a new pipeline that integrates Misa (Thiel et al., 2003) with BLAST (Altschul et al., 1997), MAFFT (Nakamura et al., 2018), Primer3 (Untergasser et al., 2012), and e-PCR (Schuler, 1997) to explore polymorphic SSR markers between any two given sequences. It identifies monomorphic and dimorphic SSR markers, providing detailed information for genetic analyses and marker-assisted breeding, such as SSR motifs, primer pairs, amplified fragments, sequence sizes, length polymorphisms, and statistical calculations, to facilitate subsequent genetic analyses and marker-assisted breeding for example. SSR2marker is a versatile tool that can significantly reduce processing time for scientists aiming to identify markers and develop primers from sequence comparisons.

SSRenricher (Luo et al., 2020) is optimized for polymorphic SSR enrichment in transcripts, rather than primer design and has a graphical user interface. The integration of Misa (Thiel et al., 2003) and CD-HIT (Li and Godzik, 2006) with additional scripts enables the execution of six core analysis steps: SSR mining, sequence clustering, sequence modification, enrichment containing polymorphic SSR sequences, false-positive removal, and results output and multiple sequence alignment. Collectively, these steps facilitate the identification and analysis of polymorphic SSRs in transcripts. This tool may be particularly pertinent for groups engaged in transcriptomics studies.

SSRMMD (Gou et al., 2020) is a command-line pipeline with a proprietary algorithm for mining p-SSR loci and potential polymorphic SSRs from assembled sequences like genomes or

transcriptomes. It facilitates primer design by providing a native tool called `connectorToPrimer3` to link SSRMMD with Primer3 (Untergasser et al., 2012). To identify candidate polymorphic SSRs, at least two assembled sequences are required. The SSRMMD algorithm assesses the conservatism and uniqueness of SSR flanking sequences for the identification of polymorphic SSRs, generating comprehensive information records and statistical analyses of SSRs, along with records of candidate polymorphic SSRs.

Websat (Webtroll) (Martins et al., 2009) is a user-friendly webtool that integrates TROLL (Castelo et al., 2002) for SSR mining and Primer3 (Untergasser et al., 2012) for primer design. Users submit sequences in raw or FASTA format (max upload 150,000 characters). The output highlights SSRs in yellow and underlines them in a table with numbered lines and spaces to aid SSR coordinate location. Clicking on an SSR invokes primer design to generate flanking primers. Users can modify primer3 parameters for custom design. Successful primers are green, SSRs blue, and failed designs are reported. All primers can be downloaded in a .csv file with SSR info, product size, sequences, and melting temperatures.

**3.2.2.2.2 Pipeline tools for p-SSR and i-SSR analysis.** This topic will analyze eight pipeline tools that, in addition to mining p-SSRs, also provide options to identify i-SSRs. Table 7 indicates that five pipeline tools utilize consolidated basic tools as SSR mining algorithms, while the remaining three use proprietary algorithms. It was further observed that two tools employ Sputnik (La Rota et al., 2005), while the others each use a different algorithm for SSR mining. All tools, except for EasySSR (Alves et al., 2023) and SSR\_pipeline (Miller et al., 2013), incorporated integrated primer design functionality.

Only SciRoKo (Kofler et al., 2007), Krait (Du et al., 2018), EasySSR (Alves et al., 2023), and BatchPrimer3 (You et al., 2008) were executable through a graphical interface, with the latter two being the sole tools to offer web servers. Most tools focused on SSRs with 1-6 base pairs, although some excluded mononucleotides. Regarding file formats, only IDSSR (Guang et al., 2019) and SSR\_pipeline (Miller et al., 2013) accepted FastQ files as input, while only Krait (Du et al., 2018) and EasySSR (Alves et al., 2023) allowed the use of annotation files to analyze SSRs in the coding/non-coding context. Outputs containing graphics and charts were provided exclusively by EasySSR (Alves et al., 2023) and IDSSR (Guang et al., 2019).

BatchPrimer3 (You et al., 2008) while primarily designed for primer design, was also considered as an SSR pipeline tool. It can identify SSRs within a genome provided in FASTA format by integrating with Primer3 (Untergasser et al., 2012) with SSR search (Nicot et al., 2004) as the basic tool to identify SSRs. This basic tool, however, was not detailed in this paper due to the current unavailability of its source code (Table 2). SSR primers are selected from flanking regions, with flexible screening criteria typically detecting motifs from dinucleotide to hexanucleotide repeats. Output includes a main HTML page summarizing primer design, an HTML table of designed primers, a tab-delimited text file with the same information, and a detailed primer view page for each sequence.

EasySSR (Alves et al., 2023) is a user-friendly web server for large-scale batch analyses, enabling the comparison of genomic data

from multiple individuals in a single run. It accepts one or more fasta files as input, with optional GBK files for coding region identification, which are converted to PTT files for output. EasySSR executes IMEx in batches (Mudunuri and Nagarajaram, 2007) for p-SSR identification in every file, and optionally detects i-SSR and c-SSRs. The IMEx raw results are made available to users and are automatically processed and compared via additional scripts. In the HTML page displaying the results, EasySSR offers customizable flanking regions for use in primer design and provides additional processed outputs such as interactive tables and graphs, which show statistical and comparative results among genomes, along with suggested SSR markers.

IDSSR (Guang et al., 2019) is a tool with focus at identifying polymorphic SSRs by integrating SSRs with nucleotide insertions/deletions (INDEL) solely based on a single genome sequence and paired-end reads. Input files include assembled genome sequences in FASTA format and sequenced clean reads in FASTQ format. The tool employs SSRIT (Temnykh et al., 2001) as its SSR mining tool, which is focused on identifying perfect SSRs. However, the tool also integrates SOAPindel (Li et al., 2013), BLAST (Altschul et al., 1997), and Primer3 (Untergasser et al., 2012), which enables it to identify i-SSR as well. A limitation is that INDEL markers require short insert sizes and 25× genome coverage, potentially limiting IDSSR's use. Nonetheless, IDSSR provides valuable data for further analysis, including repeat motifs, positions, chromosome locations, annealing temperatures, and primer sequences.

Krait (Du et al., 2018) is a user-friendly tool that can be executed directly from a desktop application with a graphical interface. It can identify p-SSR, i-SSR and c-SSR in whole genomic sequences, and design primers by integrating their proprietary algorithm with Primer3 (Untergasser et al., 2012). Krait can also locate SSRs within gene coding regions if GTF or GFF annotation files are provided as input. Additionally, Krait provides statistical data for interpretation and the outputs can be exported in formats such as Fasta, GFF3, or CSV. The tables provided by the tool are easy to navigate and include filters, making it an intuitive tool.

PolySSR (Tang et al., 2008) and SSRpoly (Duran et al., 2013), despite their similar names and attributes, are distinct tools with no shared publication teams. Both tools have Sputnik (La Rota et al., 2005) as their SSR mining tool, focusing on identifying and designing primers for polymorphic SSRs in EST sequences using cluster-based strategies. PolySSR (Tang et al., 2008) offers an online database for examining polymorphic SSRs from preprocessed EST data and integrates `cross_match` (Ewing and Green, 1998), CAP3 (Huang and Madan, 1999), Primer3 (Untergasser et al., 2012) and CheckSSR (Tang et al., 2008), to conduct sequence alignment, vector removal with EST, sequence clustering, polymorphic SSR prediction, and primer design. However, accessing PolySSR is limited, requiring users to request access to the download link via email. In contrast, SSRpoly (Duran et al., 2013) is easily downloadable from its website, functioning similarly to PolySSR (Tang et al., 2008), by using a custom method for polymorphic SSR prediction and SSRPrimer (Robinson et al., 2004) for primer design.

SciRoKo (Kofler et al., 2007) is a tool for identifying perfect, imperfect, and compound SSRs. It offers a standalone user-friendly tool with a graphical interface, including an SSR search module with five search modes and an SSR statistics module with three classification and statistics options. Although SciRoKo does not

TABLE 7 Tools for detecting SSR part 4 – pipeline Tools for perfect and imperfect SSR analysis.

Name	Reference/Year	Mining tool	Integrated toolset	Platform	Execution	Focus	Max motif length	Input files	Output graphs/Charts	p-SSR	i-SSR	c-SSR	Primer design	Tool's link
BatchPrimer3	You 2008	SSR search	Primer3	Linux, Mac OS, Windows, Web	Graphical Interface and Command Line	SSR	2–6 bp	FASTA	No	Yes	Yes	No	Yes	<a href="http://wheat.pw.usda.gov/demos/BatchPrimer3/">http://wheat.pw.usda.gov/demos/BatchPrimer3/</a>
EasySSR	Alves 2023	IMEx	EasySSR	Linux, Web	Graphical Interface	SSR	1–6 bp	FASTA; GBK	Yes	Yes	Yes	Yes	No	<a href="https://computationalbiology.ufpa.br/easyssr/">https://computationalbiology.ufpa.br/easyssr/</a>
IDSSR	Guang 2019	SSRIT	SOAPindel, BLAST, Primer3	Linux	Command Line	SSR	2–6 bp	FASTA; FASTQ	Yes	Yes	Yes	No	Yes	<a href="https://github.com/Allsummerking/IDSSR">https://github.com/Allsummerking/IDSSR</a>
Krait	Du 2018	Krait	Primer3	Linux, Mac OS, Windows	Graphical Interface	SSR	1–6 bp	FASTA; GTF; GFF	No	Yes	Yes	Yes	Yes	<a href="https://github.com/lmdu/krait">https://github.com/lmdu/krait</a>
PolySSR	Tang 2008	Sputnik	Primer3, Cross_match, CAP3, CheckSSR	Linux	Command Line	SSR	2–6 bp	FASTA	No	Yes	Yes	No	Yes	<a href="http://www.bioinformatics.nl/tools/polysrr/">http://www.bioinformatics.nl/tools/polysrr/</a>
SciRoKo	Kofler 2007	SciRoKo	Primer3	Linux, Mac OS, Windows	Graphical Interface and Command Line	SSR	2–6 bp	FASTA	No	Yes	Yes	Yes	Yes	<a href="https://kofler.or.at/bioinformatics/SciRoKo/index.html">https://kofler.or.at/bioinformatics/SciRoKo/index.html</a>
SSR_pipeline	Miller 2013	SSR_pipeline	SSR_pipeline	Linux, Mac OS, Windows	Command Line	SSR	1–6 bp	FASTA; FASTQ	No	Yes	Yes	Yes	No	<a href="https://pubs.usgs.gov/ds/778/">https://pubs.usgs.gov/ds/778/</a>
SSRpoly	Duran 2013	Sputnik	Sputnik, MySQL, SSRPrimer	Linux	Command Line	SSR	2–6 bp	FASTA	No	Yes	Yes	No	Yes	<a href="https://appliedbioinformatics.com.au/Edwards/index.php/SSRPoly">https://appliedbioinformatics.com.au/Edwards/index.php/SSRPoly</a>

“Yes” has been highlighted in the columns for visualization purposes. “SSR” = Simple Sequence Repeats or microsatellite. “BP” = Base pair. “p-SSR” = Perfect SSR. “i-SSR” = Imperfect SSR. “c-SSR” = Compound SSR.

directly integrate Primer3, it provides a DesignPrimers module for automated PCR primer design using Primer3 (Untergasser et al., 2012) and SciRoKo. Additionally, SciRoKoCo is the command line version for the SciRoKo SSR-search module.

SSR\_pipeline (Miller et al., 2013) is a command-line tool that uses original algorithm for identifying SSRs from Illumina DNA sequencing data and other platforms like Roche 454 and Ion Torrent. The program was designed for analyses of paired-end sequences from low-coverage whole genomic DNA libraries. It comprises three analysis modules and a control module for automating large data analyses. These modules are used to identify high-quality paired-end sequences, align paired-end reads into composite DNA sequences, and identify microsatellite-containing sequences based on user-defined parameters. While the pipeline does not include a primer design function, it provides flanking regions in its output for primer design in other tools.

## 4 Discussion

### 4.1 Why are there so many different tools?

The proliferation and diversity of tools aimed at discovering simple sequence repeats (SSRs) are intricately linked to the complex nature of genomic research (Lerat, 2010). Various factors contribute to this diversity, with each tool designed to address specific limitations of existing tools (Avvaru et al., 2018). Here, will be discussed key aspects to consider when selecting an SSR mining tool. All data is available in [Supplementary Table S1](#).

#### 4.1.1 Availability

When users search for a tool, they expect it to be functional and accessible. However, some tools are no longer available due to various reasons, such as discontinued web servers, broken links, or lack of maintenance (Alves et al., 2023). Among the 74 tools identified in this study, 43% were no longer functional. Therefore, the remaining 42 active tools were filtered and analyzed to help users.

#### 4.1.2 Basic tools and pipelines

Another important aspect analyzed for the variety of tools is the existence of basic tools and pipeline tools. Although they perform the function of SSR identification, many variations of algorithms are made to obtain more accurate results in less time (Pickett et al., 2017). Furthermore, many authors feel the need to go beyond SSR identification (Bizzaro and Marx, 2003), creating new basic tools and pipelines with new features such as batch analysis, support for a variety of inputs, graphical outputs, statistics, primer design, and polymorphism studies (Mudunuri et al., 2010b). This study identified 38 pipeline tools and 36 basic tools, reflecting the diverse range of functionalities available in SSR research tools.

#### 4.1.3 Execution

The method of tool execution is a crucial factor for users. Some prefer command line interfaces for speed and integration capabilities, especially for batch analyses and pipeline integration (Yue and Liu, 2022). Others find graphical interfaces more intuitive

and user-friendly (Alves et al., 2023; Mokhtar et al., 2023). Language barriers and limited command line knowledge can hinder tool use (Alves et al., 2023). Additionally, some users may face compatibility issues with their machines, leading them to prefer web-based tools (Das et al., 2019). Among the tools evaluated, 41% were command line only, 19% were graphical interface only, and 40% offered both options. Furthermore, only 43% of the available tools offered web-based analysis.

#### 4.1.4 Polymorphic SSRs and primer design

Depending on the user's research objective, the identification of polymorphic SSRs and primer design can be crucial (Abdul-Muneer, 2014). Thus, finding tools that already have these functions integrated can save a lot of time in their research and optimize their results (Alves et al., 2022). While none of the basic tools analyzed had these functions integrated, 75% of the pipelines design primers, and some focus on polymorphic SSR, saving time and optimizing results (You et al., 2008; Xia et al., 2016). Other pipelines are focused on batch analyses, generating alignments, or providing statistics (Alves et al., 2023).

#### 4.1.5 Tool's focus

Some tools are generalists, detecting all repetitive elements, while others specialize in tandem repeats, including not only SSRs but also, mini- and macrosatellites (You et al., 2008). Tools vary in their ability to refine analysis and may include parameters for restricting to microsatellites (Benson, 1999; Genovese et al., 2019). Users may benefit from acquiring SSR data alongside other repeats or elements, but some might prefer tools exclusively for SSRs (Sharma et al., 2007). Among the total number of tools identified, 68% were specific SSR tools, while 23% focused on Tandem Repeats and 9% on repetitive elements.

#### 4.1.6 Type of SSR

The type of SSR identified can also be an important factor in choosing the best tool because while some authors focus exclusively on the study of perfect SSRs for length polymorphism studies (Kelkar et al., 2010; Avvaru et al., 2020), others consider the inclusion of imperfect and compound SSRs to be of fundamental relevance in their results, for indel and SNP studies, for example., (Chen et al., 2011b; George et al., 2015; Ledenyova et al., 2019). Regarding SSR tools, it was observed that 64% of them focus exclusively on perfect SSRs, being unable to identify SSRs with mismatches (i-SSR or c-SSR) or excluding them from their results.

#### 4.1.7 Motif size

Most SSR tools focus on motifs with 1–6 bp, for being the widely accepted size criterion (Alves et al., 2023). However, some adopt alternative ranges, highlighting the lack of consensus (Bikandi, 2006). Some tools may extend to motifs up to 2000 bp or 5,000 bp or have no preset limits (Benson, 1999; Tarailo-Graovac and Chen, 2009). Users should choose tools based on the specific size of repetition they are analyzing.

#### 4.1.8 Algorithm and parameters

Due to algorithmic differences in the SSR mining, it is unlikely that any tools will always provide identical results (Sharma et al., 2007). The algorithm and user-defined parameters determine what



the tool considers to be an SSR, leading to false positives and false negatives. Some methods cannot detect repeats that end abruptly in the middle of the motif sequence (Avvaru et al., 2018). Allowing mismatches within a tract can extend previously interrupted tracts (Behura and Severson, 2015; Alves et al., 2023). For example, the average number of perfect SSRs per genome decreased from 53.5 in an analysis that permitted only perfect SSRs to 40 in an analysis that included imperfections. These tracts were initially considered perfect but were re-labeled when mismatches were permitted (Alves et al., 2023). The parameters used by each tool were compiled in Supplementary Table S1. However, this paper does not delve into the parameters, algorithms, or benchmark testing as we believe discussions on how parameter settings of specific algorithms can impact empirical microsatellite distributions have already been made whether by the tool's launch paper or specific review papers (Leclercq et al., 2007; Lerat, 2010). Instead, it focuses on elucidating how the programs are practically utilized, to assist users in selecting the most suitable tool for their requirements.

#### 4.1.9 Input file–type and size

Depending on the data available, researchers may find certain tools more suitable. A significant variety of input formats was observed, with FASTA being the most widely accepted format, followed by plain DNA sequences and FastQ. This highlights the availability of tools capable of identifying SSRs from sequencing reads and assembled sequences (Alves et al., 2022). Tools capable of processing specific formats like GenBank, PTT, and GFF are beneficial for researchers interested in SSRs within gene regions (Mudunuri and Nagarajaram, 2007). In addition to input type, the maximum input size accepted by a tool is crucial. For example, some web tools limit input to 150,000 nucleotides (Martins et al., 2009) or 2 MB (Beier et al., 2017). This can be limiting for researchers with larger datasets. To overcome this, users should seek tools without input size restrictions or use command-line versions (Yue and Liu, 2022). Another factor is the number of genomes analyzed in a single run. Many tools analyze one file at a time, but researchers often need to analyze multiple genomes simultaneously. Using a multifasta file or custom scripts can address this. Researchers can also look for programs that accept multiple files as input or offer batch analysis mode (Fox et al., 2019; Alves et al., 2023).

#### 4.1.10 Output–format, content and delivery

Each researcher approaches SSR research with specific questions and objectives, seeking tools that provide efficient solutions tailored to their needs. The analysis of 74 tools revealed a diverse array of output formats, such as text, HTML, Excel tables, and graphs, each serving different user requirements (Beier et al., 2017; Alves et al., 2023). These formats enable researchers to visualize patterns and trends, integrate data, and perform gene annotation analyses. Moreover, the content of the outputs varies, including identified SSRs, polymorphic SSRs, SSR positions, statistics, abundance graphs, and primer information, allowing users to select tools that align with their data analysis and interpretation needs (Pellegrini et al., 2012). The method of delivering output is also crucial in tool selection. Some users prefer web-based tools that provide quick on-screen results, while others require data for further analysis and may prefer tools that allow for data download or access (Mokhtar et al., 2023). Additionally, considerations such as user

anonymity and data delivery mechanisms, such as email, influence tool suitability. Overall, researchers should choose tools that align with their specific research questions, objectives, and preferences to ensure effective SSR analysis (Beier et al., 2017).

## 5 Conclusion

This study provides a comprehensive overview of microsatellites, aiding researchers in selecting the most appropriate tools for SSR analysis. Previously, the maximum number of tools cited in a single article was 25, while a total of 37 tools were identified from past reviews. The current study presents a comparative analysis of 74 tools for the first time, creating a significant resource for users. Additionally, a detailed supplementary table is provided, which encompasses all the data discussed in the article, as well as extra information, available for download and offline analysis with filtering capabilities. A detailed evaluation of each tool reveals the diversity of approaches and functionalities, reflecting ongoing innovation in genomic research. However, an in-depth analysis shows that each tool possesses unique characteristics, highlighting that no single tool can address all project requirements. The dynamic nature of genomics, coupled with the specific demands of various research objectives, complicates the development of a universal solution for SSR mining. Thus, this article offers multiple options for users to select tools based on their preferences, whether by literature citations or innovative functionalities. Ultimately, the selection of a tool should be guided by the study's specific context, considering factors such as data type, research focus, and computational resources. Overall, this compilation serves as a reference for SSR tool selection and enhances the understanding of available SSR tools, contributing to more informed decision-making and promoting accurate and efficient SSR studies across diverse research areas, while underscoring the necessity for continuous refinement and innovation in this field.

## Author contributions

SA: Conceptualization, Data curation, Formal Analysis, Methodology, Project administration, Supervision, Validation, Writing–original draft, Writing–review and editing. CD: Data curation, Formal Analysis, Validation, Visualization, Writing–original draft, Writing–review and editing. DB: Data curation, Formal Analysis, Investigation, Validation, Visualization, Writing–original draft, Writing–review and editing. RR: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Visualization, Writing–review and editing, Writing–original draft.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work has been supported by National Council for Scientific and Technological Development (CNPq) project #312316/2022–4, Coordination for the Improvement of Higher Education Personnel (CAPES), Secretary of State for Science, Technology

and Professional and Technological Education (SECTET), and the Dean's Office for Research and Graduate Studies/Federal University of Pará (PROPEP/UFPA).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

## References

- Abdul-Muneer, P. M. (2014). Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. *Genet. Res. Int.* 2014, 691759. doi:10.1155/2014/691759
- Agarwal, P., and States, D. J. (1994). The Repeat Pattern Toolkit (RPT): analyzing the structure and evolution of the *C. elegans* genome. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 1–9. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/7584377>.
- Alam, C. M., Iqbal, A., Sharma, A., Schulman, A. H., and Ali, S. (2019). Microsatellite diversity, complexity, and host range of mycobacteriophage genomes of the siphoviridae family. *Front. Genet.* 10, 207–211. doi:10.3389/fgene.2019.00207
- Altschul, S., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389
- Alves, F., Martins, F. M. S., Areias, M., and Muñoz-Mérida, A. (2022). Automating microsatellite screening and primer design from multi-individual libraries using Micro-Primers. *Sci. Rep.* 12, 295–299. doi:10.1038/s41598-021-04275-8
- Alves, S. I. A., Ferreira, V. B. C., Dantas, C. W. D., da Costa Silva, A. L., and Ramos, R. T. J. (2023). EasySSR: a user-friendly web application with full command-line features for large-scale batch microsatellite mining and samples comparison. *Front. Genet.* 14, 1228552. doi:10.3389/fgene.2023.1228552
- Avvaru, A. K., Sharma, D., Verma, A., Mishra, R. K., and Sowpati, D. T. (2020). MSDB: a comprehensive, annotated database of microsatellites. *Nucleic Acids Res.* 48, D155–D159. doi:10.1093/nar/gkz886
- Avvaru, A. K., Sowpati, D. T., and Mishra, R. K. (2018). PERF: an exhaustive algorithm for ultra-fast and efficient identification of microsatellites from large DNA sequences. *Bioinformatics* 34, 943–948. doi:10.1093/bioinformatics/btx721
- Baudrin, L. G., Deleuze, J., and How-Kit, A. (2018). Molecular and computational methods for the detection of microsatellite instability in cancer. *Front. Oncol.* 8, 621. doi:10.3389/fonc.2018.00621
- Behura, S. K., and Severson, D. W. (2015). Motif mismatches in microsatellites: insights from genome-wide investigation among 20 insect species. *DNA Res.* 22, 29–38. doi:10.1093/dnares/dsu036
- Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi:10.1093/bioinformatics/btx198
- Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi:10.1093/nar/27.2.573
- Bidmos, F., and Bayliss, C. (2014). Genomic and global approaches to unravelling how hypermutable sequences influence bacterial pathogenesis. *Pathogens* 3, 164–184. doi:10.3390/pathogens3010164
- Bikandi, J. (2006). Microsatellite repeats finder. Available at: [http://insilico.ehu.es/mini\\_tools/microsatellites/](http://insilico.ehu.es/mini_tools/microsatellites/).
- Biswas, M. K., Natarajan, S., Biswas, D., Nath, U. K., Park, J.-I., and Nou, I. (2018). LSAT: liliaceae simple sequences analysis tool, a web server. *Bioinformatics* 14, 181–182. doi:10.6026/97320630014181
- Bizzaro, J. W., and Marx, K. A. (2003). Poly: a quantitative analysis tool for simple sequence repeat (SSR) tracts in DNA. *BMC Bioinforma.* 4, 22. doi:10.1186/1471-2105-4-22

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2024.1474611/full#supplementary-material>

### SUPPLEMENTARY TABLE S1

Contains the data analyzed for this study and additional data. (A) contains the integrated data from the Tables 1 and 2 of the paper, with filters for easy navigation. (B) contains the data from Tables 1 to 8, in addition to the data from the 32 unavailable tools. (C) contains additional data for all 74 tools, including information on input, parameters, outputs, and limitations.

- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170
- Buschiazzo, E., and Gemell, N. J. (2006). The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays* 28, 1040–1050. doi:10.1002/bies.20470
- Castelo, A. T., Martins, W., and Gao, G. R. (2002). TROLL—tandem repeat occurrence locator. *Bioinformatics* 18, 634–636. doi:10.1093/bioinformatics/18.4.634
- Castoe, T. A., Poole, A. W., de Koning, A. P. J., Jones, K. L., Tomback, D. F., Oyler-McCance, S. J., et al. (2012). Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLoS One* 7, e30953. doi:10.1371/journal.pone.0030953
- Chambers, G. K., and MacAvoy, E. S. (2000). Microsatellites: consensus and controversy. *Comp. Biochem. Physiol. Part B Biochem. Mol. Biol.* 126, 455–476. doi:10.1016/S0305-0491(00)00233-9
- Chen, M., Tan, Z., and Zeng, G. (2011a). MfSAT: detect simple sequence repeats in viral genomes. *Bioinformatics* 6, 171–172. doi:10.6026/97320630006171
- Chen, M., Zeng, G., Tan, Z., Jiang, M., Zhang, J., Zhang, C., et al. (2011b). Compound microsatellites in complete *Escherichia coli* genomes. *FEBS Lett.* 585, 1072–1076. doi:10.1016/j.febslet.2011.03.005
- Churbanov, A., Ryan, R., Hasan, N., Bailey, D., Chen, H., Milligan, B., et al. (2012). HighSSR: high-throughput SSR characterization and locus development from next-generation sequencing data. *Bioinformatics* 28, 2797–2803. doi:10.1093/bioinformatics/bts524
- da Costa Pinheiro, K., Gois, B. V. A., Nogueira, W. G., Araújo, F. A., Queiroz, A. L. C., Cardenas-Alegria, O., et al. (2022). *In silico* approach to identify microsatellite candidate biomarkers to differentiate the biovar of *Corynebacterium pseudotuberculosis* genomes. *Front. Bioinforma.* 2, 931583. doi:10.3389/fbinf.2022.931583
- Das, R., Arora, V., Jaiswal, S., Iquebal, M., Angadi, U., Fatma, S., et al. (2019). PolyMorphPredict: a universal web-tool for rapid polymorphic microsatellite marker discovery from whole genome and transcriptome data. *Front. Plant Sci.* 9, 1966. doi:10.3389/fpls.2018.01966
- da Silva Lopes, R., Moraes, W. J. L., Rodrigues, T. D. S., and Bartholomeu, D. C. (2015). ProGeRF: proteome and genome repeat finder utilizing a fast parallel hash function. *Biomed. Res. Int.* 2015, 394157. doi:10.1155/2015/394157
- Delgrange, O., and Rivals, E. (2004). STAR: an algorithm to search for tandem approximate repeats. *Bioinformatics* 20, 2812–2820. doi:10.1093/bioinformatics/bth335
- Du, L., Zhang, C., Liu, Q., Zhang, X., Yue, B., and Hancock, J. (2018). Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics* 34, 681–683. doi:10.1093/bioinformatics/btx665
- Dumbovic, G., Forcales, S.-V., and Perucho, M. (2017). Emerging roles of microsatellite repeats in genome organization and disease development. *Epigenetics* 12, 515–526. doi:10.1080/15592294.2017.1318235
- Duran, C., Singhania, R., Raman, H., Batley, J., and Edwards, D. (2013). Predicting polymorphic EST-SSRs *in silico*. *Mol. Ecol. Resour.* 13, 538–545. doi:10.1111/1755-0998.12078
- Durigan, M., Cardoso-Silva, C. B., Ciampi-Guillard, M., Toledo-Silva, G., Mori, G. M., Franco, R. M. B., et al. (2018). Molecular genotyping, diversity studies and high-resolution molecular markers unveiled by microsatellites in *Giardia duodenalis*. *PLoS Negl. Trop. Dis.* 12, e0006928. doi:10.1371/journal.pntd.0006928



- Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinforma.* 5, 113. doi:10.1186/1471-2105-5-113
- Editorial Nature (2024). Tandem repeats in the long-read sequencing era. *Nat. Rev. Genet.* 25, 449. doi:10.1038/s41576-024-00751-9
- Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194. doi:10.1101/gr.8.3.186
- Faircloth, B. C. (2007). msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Mol. Ecol. Resour.* 8, 92–94. doi:10.1111/j.1471-8286.2007.01884.x
- Fox, G., Preziosi, R. F., Antwis, R. E., Benavides-Serrato, M., Combe, F. J., Harris, W. E., et al. (2019). Multi-individual microsatellite identification: a multiple genome approach to microsatellite design (MiMi). *Mol. Ecol. Resour.* 19, 1672–1680. doi:10.1111/1755-0998.13065
- Gao, L., Tang, J., Li, H., and Jia, J. (2003). Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Mol. Breed.* 12, 245–261. doi:10.1023/A:1026346121217
- Genovese, L. M., Mosca, M. M., Pellegrini, M., and Geraci, F. (2019). Dot2dot: accurate whole-genome tandem repeats discovery. *Bioinformatics* 35, 914–922. doi:10.1093/bioinformatics/bty747
- George, B., Alam, C. M., Kumar, R. V., Gnanasekaran, P., and Chakraborty, S. (2015). Potential linkage between compound microsatellites and recombination in geminiviruses: evidence from comparative analysis. *Virology* 482, 41–50. doi:10.1016/j.virol.2015.03.003
- Girgis, H. Z., and Shehtin, S. L. (2013). MsDetector: toward a standard computational tool for DNA microsatellites detection. *Nucleic Acids Res.* 41, e22. doi:10.1093/nar/gks881
- Gou, X., Shi, H., Yu, S., Wang, Z., Li, C., Liu, S., et al. (2020). SSRMMD: a rapid and accurate algorithm for mining SSR feature loci and candidate polymorphic SSRs based on assembled sequences. *Front. Genet.* 11, 706–710. doi:10.3389/fgene.2020.00706
- Guang, X.-M., Xia, J.-Q., Lin, J.-Q., Yu, J., Wan, Q.-H., and Fang, S.-G. (2019). IDSSR: an efficient pipeline for identifying polymorphic microsatellites from a single genome sequence. *Int. J. Mol. Sci.* 20, 3497. doi:10.3390/ijms20143497
- Hadrill, P. R. (2021). Developments in forensic dna analysis. *Emerg. Top. Life Sci.* 5, 381–393. doi:10.1042/ETLS20200304
- Hernandez-Sanchez, A., Grossman, M., Yeung, K., Sei, S. S., Lipkin, S., and Kloor, M. (2022). Vaccines for immunoprevention of DNA mismatch repair deficient cancers. *J. Immunother. Cancer* 10, e004416. doi:10.1136/jitc-2021-004416
- Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877. doi:10.1101/gr.9.9.868
- Jäger, R. (2022). New perspectives for whole genome amplification in forensic STR analysis. *Int. J. Mol. Sci.* 23, 7090. doi:10.3390/ijms23137090
- Jilani, M. G., and Ali, S. (2022). Assessment of simple sequence repeats signature in hepatitis E virus (HEV) genomes. *J. Genet. Eng. Biotechnol.* 20, 73. doi:10.1186/s43141-022-00365-w
- Jorda, J., and Kajava, A. V. (2009). T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* 25, 2632–2638. doi:10.1093/bioinformatics/btp482
- Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi:10.1159/000084979
- Karaca, M., Bilgen, M., Onus, A. N., Ince, A. G., and Elmasulu, S. Y. (2005). Exact tandem repeats analyzer (E-TRA): a new program for DNA sequence mining. *J. Genet.* 84, 49–54. doi:10.1007/BF02715889
- Kelkar, Y. D., Strubczewski, N., Hile, S. E., Chiaromonte, F., Eckert, K. A., and Makova, K. D. (2010). What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biol. Evol.* 2, 620–635. doi:10.1093/gbe/evq046
- Kofler, R., Schlotterer, C., and Lelley, T. (2007). SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics* 23, 1683–1685. doi:10.1093/bioinformatics/btm157
- Kolpakov, R., Bana, G., and Kucherov, G. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.* 31, 3672–3678. doi:10.1093/nar/gkg617
- Korotkov, E. V., Kamionskya, A. M., and Korotkova, M. A. (2021). Detection of highly divergent tandem repeats in the rice genome. *Genes (Basel)* 12, 473. doi:10.3390/genes12040473
- Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi:10.1093/nar/29.22.4633
- La Rota, M., Kantety, R. V., Yu, J.-K., and Sorrells, M. E. (2005). Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* 6, 23. doi:10.1186/1471-2164-6-23
- Laskar, R., Jilani, M. G., Nasrin, T., and Ali, S. (2022). Microsatellite signature of reference genome sequence of SARS-CoV-2 and 32 species of coronaviridae family. *Int. J. Infect.* 9. doi:10.5812/iji-122019
- Leclercq, S., Rivals, E., and Jarne, P. (2007). Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinforma.* 8, 125. doi:10.1186/1471-2105-8-125
- Ledenyova, M. L., Tkachenko, G. A., and Shpak, I. M. (2019). Imperfect and compound microsatellites in the genomes of burkholderia pseudomallei strains. *Mol. Biol.* 53, 142–153. doi:10.1134/S0026898419010087
- Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Hered. (Edinb.)* 104, 520–533. doi:10.1038/hdy.2009.165
- Li, S., Li, R., Li, H., Lu, J., Li, Y., Bolund, L., et al. (2013). SOAPindel: efficient identification of indels from short paired reads. *Genome Res.* 23, 195–200. doi:10.1101/gr.132480.111
- Li, T. T., Tang, B., Bai, X., Wang, X. L., Luo, X. N., Yan, H. B., et al. (2020). Development of genome-wide polymorphic microsatellite markers for *Trichinella spiralis*. *Parasites Vectors* 13, 58–11. doi:10.1186/s13071-020-3929-2
- Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi:10.1093/bioinformatics/btl158
- Li, Y., Korol, A. B., Fahima, T., Beiles, A., and Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.* 11, 2453–2465. doi:10.1046/j.1365-294X.2002.01643.x
- Lim, K. G., Kwok, C. K., Hsu, L. Y., and Wirawan, A. (2013). Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. *Brief. Bioinform.* 14, 67–81. doi:10.1093/bib/bbs023
- Litt, M., and Luty, J. A. (1989). A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* 44, 397–401.
- Lu, T.-Y., Munson, K. M., Lewis, A. P., Zhu, Q., Tallon, L. J., Devine, S. E., et al. (2021). Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nat. Commun.* 12, 4250. doi:10.1038/s41467-021-24378-0
- Luo, W., Wu, Q., Yang, L., Chen, P., Yang, S., Wang, T., et al. (2020). SSREnricher: a computational approach for large-scale identification of polymorphic microsatellites based on comparative transcriptome analysis. *PeerJ* 8, e9372. doi:10.7717/peerj.9372
- Magoč, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi:10.1093/bioinformatics/btr507
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi:10.14806/ej.17.1.200
- Martins, W. S., Lucas, D. C. S., Neves, K. F. S., and Bertioli, D. J. (2009). WebSat - a web software for MicroSatellite marker development. *Bioinformation* 3, 282–283. doi:10.6026/97320630003282
- Marwal, A., Sahu, A. K., and Gaur, R. K. (2014). “Molecular markers,” in *Animal biotechnology* (Elsevier), 289–305. doi:10.1016/B978-0-12-416002-6.00016-X
- Mathur, M., Tyagi, S., and Kataria, P. (2020). A comparative study of various simple sequence repeats identification tools using *Aspergillus fumigatus* genome. *jbcg* 3, 1–13. doi:10.17303/jbcg.2020.3.102
- Merkel, A., and Gemmill, N. (2008). Detecting short tandem repeats from genome data: opening the software black box. *Brief. Bioinform.* 9, 355–366. doi:10.1093/bib/bbn028
- Metz, S., Cabrera, J. M., Rueda, E., Giri, F., and Amavet, P. (2016). FullSSR: microsatellite finder and primer designer. *Adv. Bioinforma.* 2016, 6040124. doi:10.1155/2016/6040124
- Miller, M. P., Knaus, B. J., Mullins, T. D., and Haig, S. M. (2013). SSR\_pipeline: a bioinformatic infrastructure for identifying microsatellites from paired-end Illumina high-throughput DNA sequencing data. *J. Hered.* 104, 881–885. doi:10.1093/jhered/est056
- Mohanty, P. S., Bansal, A. K., Naaz, F., Arora, M., Gupta, U. D., Gupta, P., et al. (2019). Multiple strain infection of *Mycobacterium leprae* in a family having 4 patients: a study employing tandem repeats. *PLoS One* 14, e0214051. doi:10.1371/journal.pone.0214051
- Mokhtar, M. M., Alsamman, A. M., and El Allali, A. (2023). MegaSSR: a web server for large scale microsatellite identification, classification, and marker development. *Front. Plant Sci.* 14, 1219055. doi:10.3389/fpls.2023.1219055
- Mokhtar, M. M., and Atia, M. A. M. (2019). SSRome: an integrated database and pipelines for exploring microsatellites in all organisms. *Nucleic Acid. Res.* 47, D244–D252. doi:10.1093/nar/gky998
- Morgante, M., Hanafey, M., and Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* 30, 194–200. doi:10.1038/ng822
- Mudunuri, S. B., Kumar, P., Rao, A. A., Pallamsetty, S., and Nagarajaram, H. A. (2010a). G-IMEx: a comprehensive software tool for detection of microsatellites from genome sequences. *Bioinformation* 5, 221–223. doi:10.6026/97320630005221
- Mudunuri, S. B., and Nagarajaram, H. A. (2007). IMEx: imperfect microsatellite extractor. *Bioinformatics* 23, 1181–1187. doi:10.1093/bioinformatics/btm097

- Mudunuri, S. B., Patnana, S., and Nagarajaram, H. A. (2014). MICdb3.0: a comprehensive resource of microsatellite repeats from prokaryotic genomes. *Database* 2014, bau005. doi:10.1093/database/bau005
- Mudunuri, S. B., Rao, A. A., Pallamsetty, S., and Nagarajaram, H. A. (2010b). "Comparative analysis of microsatellite detecting software," in *Proceedings of the international symposium on bioinformatics* (New York, NY, USA: ACM), 1–7. doi:10.1145/1722024.1722068
- Nakamura, T., Yamada, K. D., Tomii, K., and Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 34, 2490–2492. doi:10.1093/bioinformatics/bty121
- Nicot, N., Chiquet, V., Gandon, B., Amilhat, L., Legeai, F., Leroy, P., et al. (2004). Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). *Theor. Appl. Genet.* 109, 800–805. doi:10.1007/s00122-004-1685-x
- Oliveira, E. J., Dantas, J. L. L., da Silva Castellen, M., and Machado, M. D. (2008). Identificação de microsatélites para o mamoeiro por meio da exploração do banco de dados de DNA. *Rev. Bras. Frutic.* 30, 841–845. doi:10.1590/s0100-29452008000300049
- Palliyarakkal, M. K., Ramaswamy, M., and Vadivel, A. (2011). Microsatellites in palm (Arecaceae) sequences. *Bioinformatics* 7, 347–351. doi:10.6026/97320630007347
- Park, J. H., Kim, E. Y., Luchini, C., Eccher, A., Tizaoui, K., Shin, J., et al. (2022). Artificial intelligence for predicting microsatellite instability based on tumor histomorphology: a systematic review. *Int. J. Mol. Sci.* 23, 2462. doi:10.3390/ijms23052462
- Pellegrini, M., Renda, M. E., and Vecchio, A. (2012). Tandem repeats discovery service (TReADS) applied to finding novel cis-acting factors in repeat expansion diseases. *BMC Bioinforma.* 13, S3. doi:10.1186/1471-2105-13-S4-S3
- Pickett, B. D., Karlinsky, S. M., Penrod, C. E., Cormier, M. J., Ebbert, M. T. W., Shiozawa, D. K., et al. (2016). SA-SSR: a suffix array-based algorithm for exhaustive and efficient SSR discovery in large genetic sequences. *Bioinformatics* 32, 2707–2709. doi:10.1093/bioinformatics/btw298
- Pickett, B. D., Miller, J. B., and Ridge, P. G. (2017). Kmer-SSR: a fast and exhaustive SSR search algorithm. *Bioinformatics* 33, 3922–3928. doi:10.1093/bioinformatics/btx538
- Reneker, J., Shyu, C.-R., Zeng, P., Polacco, J. C., and Gassmann, W. (2004). ACMES: fast multiple-genome searches for short repeat sequences with concurrent cross-species information retrieval. *Nucleic Acids Res.* 32, W649–W653. doi:10.1093/nar/gkh455
- Rice, P., Longden, I., and Bleasby, A. (2000). EMBOS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi:10.1016/s0168-9525(00)02024-2
- Richard, G.-F., Kerrest, A., and Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.* 72, 686–727. doi:10.1128/MMBR.00011-08
- Robinson, A. J., Love, C. G., Batley, J., Barker, G., and Edwards, D. (2004). Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics* 20, 1475–1476. doi:10.1093/bioinformatics/bth104
- Saeed, A. F., Wang, R., and Wang, S. (2016). Microsatellites in pursuit of microbial genome evolution. *Front. Microbiol.* 6, 1462. doi:10.3389/fmicb.2015.01462
- Sahu, B. P., Majee, P., Singh, R. R., Sahoo, A., and Nayak, D. (2020). Comparative analysis, distribution, and characterization of microsatellites in Orf virus genome. *Sci. Rep.* 10, 13852. doi:10.1038/s41598-020-70634-6
- Schuler, G. D. (1997). Sequence mapping by electronic PCR. *Genome Res.* 7, 541–550. doi:10.1101/gr.7.5.541
- Sharma, D., Issac, B., Raghava, G. P. S., and Ramaswamy, R. (2004). Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics* 20, 1405–1412. doi:10.1093/bioinformatics/bth103
- Sharma, P. C., Grover, A., and Kahl, G. (2007). Mining microsatellites in eukaryotic genomes. *Trends Biotechnol.* 25, 490–498. doi:10.1016/j.tibtech.2007.07.013
- Sokolova, O., Moročko-Bičevska, I., and Lácis, G. (2022). Genetic diversity of venturia inaequalis in Latvia revealed by microsatellite markers. *Pathogens* 11, 1165. doi:10.3390/pathogens11101165
- Song, X., Yang, T., Zhang, X., Yuan, Y., Yan, X., Wei, Y., et al. (2021). Comparison of the microsatellite distribution patterns in the genomes of euarchontoglires at the taxonomic level. *Front. Genet.* 12, 622724. doi:10.3389/fgene.2021.622724
- Sreenu, V. B., Alevoor, V., Nagaraju, J., and Nagarajaram, H. A. (2003). MICdb: database of prokaryotic microsatellites. *Nucleic Acids Res.* 31, 106–108. doi:10.1093/nar/kgk002
- Srivastava, S., Avvaru, A. K., Sowpati, D. T., and Mishra, R. K. (2019). Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics* 20, 153. doi:10.1186/s12864-019-5516-5
- Stieneke, D. L., and Eujayl, I. A. (2007). Imperfect SSR finder. Available at: <http://ssr.nwisr.ars.usda.gov/>.
- Tang, J., Baldwin, S. J., Jacobs, J. M. E., van der Linden, C. G., Voorrips, R. E., Leunissen, J. A. M., et al. (2008). Large-scale identification of polymorphic microsatellites using an *in silico* approach. *BMC Bioinforma.* 9, 374. doi:10.1186/1471-2105-9-374
- Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* 25, Unit 4.10–14. doi:10.1002/0471250953.bi0410s25
- Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., and McCouch, S. (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.* 11, 1441–1452. doi:10.1101/gr.184001
- Thiel, T., Michalek, W., Varshney, R., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* 106, 411–422. doi:10.1007/s00122-002-1031-0
- Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2003). Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinforma.* 2, Unit 2.3–22. doi:10.1002/0471250953.bi0203s00
- Uguen, K., Michaud, J. L., and Génin, E. (2024). Short Tandem Repeats in the era of next-generation sequencing: from historical loci to population databases. *Eur. J. Hum. Genet.* 32, 1037–1044. doi:10.1038/s41431-024-01666-z
- Umang, U., Bharti, P. K., and Husai, A. (2022). Mining and analysis of microsatellites in human coronavirus genomes using the in-house built Java pipeline. *Genomics Inf.* 20, e35. doi:10.5808/gi.20033
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Res.* 40, e115. doi:10.1093/nar/gks596
- Vieira, M. L. C., Santini, L., Diniz, A. L., and Munhoz, C. D. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genet. Mol. Biol.* 39, 312–328. doi:10.1590/1678-4685-GMB-2016-0027
- Wang, X., Lu, P., and Luo, Z. (2013). GMATo: a novel tool for the identification and analysis of microsatellites in large genomes. *Bioinformatics* 9, 541–544. doi:10.6026/97320630009541
- Wang, X., and Wang, L. (2016). GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. *Front. Plant Sci.* 7, 1350. doi:10.3389/fpls.2016.01350
- Xia, E.-H., Yao, Q.-Y., Zhang, H.-B., Jiang, J.-J., Zhang, L.-P., and Gao, L.-Z. (2016). CandiSSR: an efficient pipeline used for identifying candidate polymorphic SSRs based on multiple assembled sequences. *Front. Plant Sci.* 6, 1171–1179. doi:10.3389/fpls.2015.01171
- Yamamoto, H., Watanabe, Y., Arai, H., Umemoto, K., Tateishi, K., and Sunakawa, Y. (2024). Microsatellite instability: a 2024 update. *Cancer Sci.* 115, 1738–1748. doi:10.1111/cas.16160
- You, F. M., Huo, N., Gu, Y. Q., Luo, M., Ma, Y., Hane, D., et al. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinforma.* 9, 253. doi:10.1186/1471-2105-9-253
- Yuan, S., Zeng, G., and Zhang, D. (2018). Development of microsatellite markers for *Primula oreodoxa* (Primulaceae), a distylous-homostylous species. *Appl. Plant Sci.* 6, e01150. doi:10.1002/aps3.1150
- Yue, J., and Liu, Y. (2022). SSR2Marker: an integrated pipeline for identification of SSR markers within any two given genome-scale sequences. *Mol. Horticult.* 2, 11. doi:10.1186/s43897-022-00033-0
- Zribi, S., Oueslati, A. E., and Lachiri, Z. (2016). "Tandem repeat search tools performance for the *Arabidopsis thaliana* genome," in 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, Tunisia, 21–23 March 2016 (IEEE), 330–335. doi:10.1109/ATSIP.2016.7523101